Statistics 500=Psychology 611=Biostatistics 550

Introduction to Regression and Anova

Paul R. Rosenbaum
Professor, Statistics Department, Wharton School

Description

Statistics 500/Psychology 611 is a second course in statistics for PhD students in the social, biological and business sciences. It covers multiple linear regression and analysis of variance. Students should have taken an undergraduate course in statistics prior to Statistics 500.

Topics

1-Review of basic statistics.
2-Simple regression.
3-Multiple regression.
4-General linear hypothesis.
5-Woes of Regression Coefficients.
6-Transformations.
7-Polynomials.
8-Coded variables.
9-Diagnostics.
10-Variable selection.
11-One-way anova.
12-Two-way and factorial anova.

How do I get R for free? http://cran.r-project.org/

Final exam date: http://www.upenn.edu/registrar/
Holidays, breaks, last class: http://www.upenn.edu/almanac/3yearcal.html

My web page: http://www-stat.wharton.upenn.edu/~rosenbap/index.html
Email: rosenbaum@wharton.upenn.edu
Phone: 215-898-3120
Office: 473 Huntsman Hall (in the tower, 4th floor)
Office Hours: Tuesday 1:30-2:30 and by appointment.

The bulk pack and course data in R are on my web page.

Overview

Review of Basic Statistics

Descriptive statistics, graphs, probability, confidence intervals, hypothesis tests.

Simple Regression

Simple regression uses a line with one predictor to predict one outcome.

Multiple Regression

Multiple regression uses several predictors in a linear way to predict one outcome.

General Linear Hypothesis

The general linear hypothesis asks whether several variables may be dropped from a multiple regression.

Woes of Regression Coefficients

Discussion of the difficulties of interpreting regression coefficients and what you can do.

Transformations

A simple way to fit curves or nonlinear models: transform the variables.

Polynomials

Another way to fit curves: include quadratics and interactions.

Coded Variables

Using nominal data (NY vs Philly vs LA) as predictors in regression.

Diagnostics

How to find problems in your regression model: residual, leverage and influence.

Variable Selection

Picking which predictors to use when many variables are available.

One-Way Anova

Simplest analysis of variance: Do several groups differ, and if so, how?

Two-Way Anova

Study two sources of variation at the same time.

Factorial Anova

Study two or more treatments at once, including their interactions.

Common Questions

Statistics Department Courses (times, rooms)
http://www.upenn.edu/registrar/roster/stat.html

Final Exams (dates, rules)
http://www.upenn.edu/registrar/finals/spring05_index.html

Computing and related help at Wharton
http://inside.wharton.upenn.edu/

Statistical Computing in the Psychology Department
http://www.psych.upenn.edu

When does the the course start? When does it end? Holidays?
http://www.upenn.edu/almanac/3yearcal.html

Does anybody have any record of this?
http://www.upenn.edu/registrar/

Huntsman Hall
http://www.facilities.upenn.edu/mapsBldgs/view_bldg.php3?id=146
http://www.facilities.upenn.edu/mapsBldgs/view_map.php3?id=393

Suggested reading

Box, G. E. P. (1966) Use and Abuse of Regression, Technometrics, 8, 625-629.
http://www.jstor.org/ or
http://www.jstor.org/stable/1266635?&Search=yes&term=abuse&term=box&list=hide&searchUri=%2Faction%2FdoAdvancedSearch%3Fq0%3Dbox%26f0%3Dau%26c0%3DAND%26q1%3Dabuse%26f1%3Dti%26c1%3DAND%26q2%3D%26f2%3Dall%26c2%3DAND%26q3%3D%26f3%3Dall%26wc%3Don%26sd%3D%26ed%3D%26la%3D%26jo%3D%26dc.Statistics%3DStatistics%26Search%3DSearch&item=1&ttl=1&returnArticleService=showArticle

Helpful articles from JSTOR http://www.jstor.org/

1. The Analysis of Repeated Measures: A Practical Review with Examples
   B. S. Everitt *The Statistician*, Vol. 44, No. 1. (1995), pp. 113-135.
2. The hat matrix in regression and anova. D. Hoaglin and R. Welsh, *American Statistician*, Vol 32, (1978), pp. 17-22.
3. The Use of Nonparametric Methods in the Statistical Analysis of the Two-Period Change-Over Design   Gary G. Koch
   Biometrics, Vol. 28, No. 2. (Jun., 1972), pp. 577-584.

Some Web Addresses

Web page for Sheather's text
http://www.stat.tamu.edu/~sheather/

Amazon for Sheather's text (required)
http://www.amazon.com/Modern-Approach-Regression-Springer-Statistics/dp/0387096078/ref=tmm_hrd_title_0/186-7302133-0606755?ie=UTF8&qid=1315493088&sr=1-1

Alternative text used several years ago (optional alternative, not suggested)
http://www.amazon.com/Applied-Regression-Analysis-Multivariable-Methods/dp/0495384968/ref=sr_1_1?s=books&ie=UTF8&qid=1315493363&sr=1-1

Good supplement about R (optional, suggested)
http://www.amazon.com/Data-Analysis-Graphics-Using-Example-Based/dp/0521762936/ref=sr_1_1?s=books&ie=UTF8&qid=1315493138&sr=1-1

Review basic statistics, learn basic R (optional, use if you need it)
http://www.amazon.com/Introductory-Statistics-R-Computing/dp/0387790535/ref=sr_1_1?s=books&ie=UTF8&qid=1315493184&sr=1-1

Excellent text, alternative to Sheather, more difficult than Sheather
http://www.amazon.com/Applied-Regression-Analysis-Probability-Statistics/dp/0471170828/ref=sr_1_1?s=books&ie=UTF8&qid=1315493220&sr=1-1

Good text, alternative/supplement to Sheather, easier than Sheather
http://www.amazon.com/Regression-Analysis-Example-Probability-Statistics/dp/0471746967/ref=tmm_hrd_title_0?ie=UTF8&qid=1315493316&sr=1-1

Free R manuals at R home page. Start with "An Introduction to R"
http://cran.r-project.org/
-->   Manuals --> An Introduction to R
--> Search --> Paradis --> R for Beginners

My web page (bulk pack, course data)
http://www-stat.wharton.upenn.edu/~rosenbap/index.html

Computing

How do I get R for free? http://cran.r-project.org/

After you have installed R, you can get the **course data** in the R-workspace on my web page: http://www-stat.wharton.upenn.edu/~rosenbap/index.html

I will probably add things to the R-workspace during the semester. So you will have to go back to my web page to **get the latest version**.

A common problem: You go to my web page and download the latest R-workspace, but it looks the same as the one you had before – the new stuff isn't there. This happens when your web browser thinks it has downloaded the file before and will save you time by not downloading it again. Bad web browser. You need to clear the cache; then it will get the new version.

Most people find an R book helpful. I recommend Maindonald and Braun, *Data Analysis and Graphics Using R*, published by Cambridge. A more basic book is Dalgaard, *Introductory Statistics with R*, published by Springer.

At http://cran.r-project.org/, click on **manuals** to get free documentation. "An Introduction to R" is there, and it is useful. When you get good at R, do a search at the site for Paradis' "R for Beginners," which is very helpful, but not for beginners.

Textbook

My sense is that students need a textbook, not just the lectures and the bulk pack.

The 'required' textbook for the course is Sheather (2009) *A Modern Approach to Regression with R*, NY: Springer. There is a little matrix algebra in the book, but there is none in the course. Sheather replaces the old text, Kleinbaum, Kupper, Muller and Nizam, *Applied Regression and other Multivariable Methods*, largely because this book has become very expensive. An old used edition of Kleinbaum is a possible alternative to Sheather – it's up to you. Kleinbaum does more with anova for experiments. A book review by Gudmund R. Iversen of Swathmore College is available at: http://www.jstor.org/stable/2289682?&Search=yes&term=kleinbaum&term=kupper&list=hide&searchUri=%2Faction%2FdoAdvancedSearch%3Fq0%3Dkleinbaum%26f0%3Dau%26c0%3DAND%26q1%3Dkupper%26f1%3Dau%26c1%3DAND%26q2%3D%26f2%3Dall%26c2%3DAND%26q3%3D%26f3%3Dall%26wc%3Don%26re%3Don%26sd%3D%26ed%3D%26la%3D%26jo%3D%26dc.Statistics%3DStatistics%26Search%3DSearch&item=6&ttl=7&returnArticleService=showArticle

Some students might prefer one of the textbooks below, and they are fine substitutes.

If you would prefer an easier, less technical textbook, you might consider *Regression by Example* by Chatterjee and Hadi. The book has a nice chapter on transformations, but it barely covers anova. An earlier book, now out of print, with the same title by Chatterjee and Price is very similar, and probably available inexpensively used.
http://www.amazon.com/Regression-Analysis-Example-Probability-Statistics/dp/0471746967/ref=sr_1_2?ie=UTF8&s=books&qid=1252524629&sr=1-2

If you know matrix algebra, you might prefer the text *Applied Regression Analysis* by Draper and Smith. It is only slightly more difficult than Kleinbaum, and you can read around the matrix algebra.
http://www.amazon.com/Applied-Regression-Analysis-Probability-Statistics/dp/0471170828/ref=sr_1_1?ie=UTF8&s=books&qid=1252524403&sr=1-1

If you use R, then as noted previously, I recommend the additional text Maindonald and Braun, *Data Analysis and Graphics Using R*, published by Cambridge. It is in its third edition, which is a tad more up to date than the first or second editions, but you might prefer an inexpensive used earlier edition if you can find one.

Graded Work

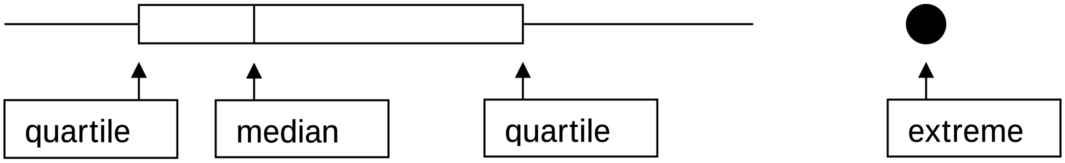
Your grade is based on three exams. Copies of old exams are at the end of this bulkpack. The first two exams are take-homes in which you do a data-analysis project. They are exams, so you do the work by yourself. The first exam covers the basics of multiple regression. The second exam covers diagnostics, model building and variable selection. The final exam is sometimes in-class, sometimes take home. The date of the final exam is determined by the registrar – see the page above for Common Questions. The decision about whether the final is in-class or take-home will be made after the first take-home is graded. That will be in the middle of the semester. If you need to make travel arrangements before the middle of the semester, you will need to plan around an in-class final.

The best way to learn the material is to practice using the old exams. There are three graded exams. If for each graded exam, you did two practice exams, then you would do nine exams in total, which means doing nine data analysis projects. With nine projects behind you, regression will start to be familiar.

Review of Basic Statistics – Some Statistics

- The review of basic statistics is a quick review of ideas from your first course in statistics.
- n measurements: $X_1, X_2, \ldots, X_n$
- **mean** (or average): $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$
- **order statistics** (or data sorted from smallest to largest): Sort $X_1, X_2, \ldots, X_n$ placing the smallest first, the largest last, and write $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$, so the smallest value is the first order statistic, $X_{(1)}$, and the largest is the nth order statistic, $X_{(n)}$. If there are n=4 observations, with values $X_1 = 5, X_2 = 4, X_3 = 9, X_4 = 5$, then the n=4 order statistics are $X_{(1)} = 4, X_{(2)} = 5, X_{(3)} = 5, X_{(4)} = 9$.
- **median** (or middle value): If n is odd, the median is the middle order statistic – e.g., $X_{(3)}$ if n=5. If n is even, there is no middle order statistic, and the median is the average of the two order statistics closest to the middle – e.g., $\dfrac{X_{(2)} + X_{(3)}}{2}$ if n=4. Depth of median is $\dfrac{n+1}{2}$ where a "half" tells you to average two order statistics – for n=5, $\dfrac{n+1}{2} = \dfrac{5+1}{2} = 3$, so the median is $X_{(3)}$, but for n=4, $\dfrac{n+1}{2} = \dfrac{4+1}{2} = 2.5$, so the median is $\dfrac{X_{(2)} + X_{(3)}}{2}$. The median cuts the data in half – half above, half below.
- **quartiles**: Cut the data in quarters – a quarter above the upper quartile, a quarter below the lower quartile, a quarter between the lower quartile and the median, a quarter between the median and the upper quartile. The **interquartile range** is the upper quartile minus the lower quartile.

- **boxplot**: Plots median and quartiles as a box, calls attention to extreme observations.



- **sample standard deviation**: square root of the typical squared deviation from the mean, sorta,

$$s = \sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}}$$

  however, you don't have to remember this ugly formula.
- **location**: if I add a constant to every data value, a measure of location goes up by the addition of that constant.
- **scale**: if I multiply every data value by a constant, a measure of scale is multiplied by that constant, but a measure of scale does not change when I add a constant to every data value.

Check your understanding: What happens to the mean if I drag the biggest data value to infinity? What happens to the median? To a quartile? To the interquartile range? To the standard deviation? Which of the following are measures of location, of scale or neither: median, quartile, interquartile range, mean, standard deviation? In a boxplot, what would it mean if the median is closer to the lower quartile than to the upper quartile?

Topic: Review of Basic Statistics – Probability

- **probability space**: the set of everything that can happen, Ω. Flip two coins, dime and quarter, and the sample space is Ω = {HH, HT, TH, TT} where HT means "head on dime, tail on quarter", etc.
- **probability**: each element of the sample space has a probability attached, where each probability is between 0 and 1 and the total probability over the sample space is 1. If I flip two fair coins: prob(HH) = prob(HT) = prob(TH) = prob(TT) = ¼.
- **random variable**: a rule **X** that assigns a number to each element of a sample space. Flip to coins, and the number of heads is a random variable: it assigns the number **X**=2 to HH, the number **X**=1 to both HT and TH, and the number **X**=0 to TT.
- **distribution of a random variable**: The chance the random variable **X** takes on each possible value, x, written prob(**X**=x). Example: flip two fair coins, and let **X** be the number of heads; then prob(**X**=2) = ¼, prob(**X**=1) = ½, prob(**X**=0) = ¼.
- **cumulative distribution of a random variable**: The chance the random variable **X** is less than or equal to each possible value, x, written prob(**X**≤x). Example: flip two fair coins, and let **X** be the number of heads; then prob(**X**≤ 0) = ¼, prob(**X**≤1) = ¾, prob(**X**≤2) = 1. Tables at the back of statistics books are often cumulative distributions.
- **independence of random variables**: Captures the idea that two random variables are unrelated, that neither predicts the other. The formal definition which follows is not intuitive – you get to like it by trying many intuitive examples, like unrelated coins and taped coins, and finding the definition always works. Two random variables, **X** and **Y**, are independent if the chance that simultaneously **X**=x and **Y**=y can be found by multiplying the separate probabilities

  prob(**X**=x and **Y**=y) = prob(**X**=x) prob(**Y**=y) for every choice of x,y.

Check your understanding: Can you tell exactly what happened in the sample space from the value of a random variable? Pick one: Always, sometimes, never. For people, do you think **X**=height and **Y**=weight are independent? For undergraduates, might **X**=age and **Y**=gender (1=female, 2=male) be independent? If I flip two fair coins, a dime and a quarter, so that prob(HH) = prob(HT) = prob(TH) = prob(TT) = ¼, then is it true or false that getting a head on the dime is independent of getting a head on the quarter?

Topic: Review of Basics – Expectation and Variance

- **Expectation**: The expectation of a random variable **X** is the sum of its possible values weighted by their probabilities,

$$E(\mathbf{X}) = \sum_x x \cdot prob(\mathbf{X} = x)$$

- **Example**: I flip two fair coins, getting **X**=0 heads with probability ¼, **X**=1 head with probability ½, and **X**=2 heads with probability ¼; then the expected number of heads is $E(\mathbf{X}) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$, so I expect 1 head when I flip two fair coins. Might actually get 0 heads, might get 2 heads, but 1 head is what is typical, or expected, on average.
- **Variance and Standard Deviation**: The standard deviation of a random variable **X** measures how far **X** typically is from its expectation *E*(**X**). Being too high is as bad as being too low – we care about errors, and don't care about their signs. So we look at the squared difference between **X** and *E*(**X**), namely $\mathbf{D} = \{\mathbf{X} - E(\mathbf{X})\}^2$, which is, itself, a random variable. The variance of **X** is the expected value of **D** and the standard deviation is the square root of the variance, $\mathrm{var}(\mathbf{X}) = E(\mathbf{D})$ and $st.dev.(\mathbf{X}) = \sqrt{\mathrm{var}(\mathbf{X})}$.
- **Example**: I independently flip two fair coins, getting **X**=0 heads with probability ¼, **X**=1 head with probability ½, and **X**=2 heads with probability ¼. Then *E*(**X**)=1, as noted above. So $\mathbf{D} = \{\mathbf{X} - E(\mathbf{X})\}^2$ takes the value **D** =

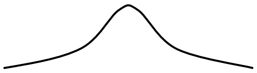
$(0-1)^2=1$ with probability ¼, the value **D** = $(1-1)^2=0$ with probability ½, and the value **D** = $(2-1)^2=1$ with probability ¼. The variance of **X** is the expected value of **D** namely: var(**X**) = $E(\mathbf{D})=1\cdot\frac{1}{4}+0\cdot\frac{1}{2}+1\cdot\frac{1}{4}=\frac{1}{2}$. So the standard deviaiton is $st.dev.(\mathbf{X})=\sqrt{\text{var}(\mathbf{X})}=\sqrt{\frac{1}{2}}=0.707$. So when I flip two fair coins, I expect one head, but often I get 0 or 2 heads instead, and the typical deviation from what I expect is 0.707 heads. This 0.707 reflects the fact that I get exactly what I expect, namely 1 head, half the time, but I get 1 more than I expect a quarter of the time, and one less than I expect a quarter of the time.

Check your understanding: If a random variance has zero variance, how often does it differ from its expectation? Consider the height **X** of male adults in the US. What is a reasonable number for *E*(**X**)? Pick one: 4 feet, 5’9”, 7 feet. What is a reasonable number for *st.dev.*(**X**)? Pick one: 1 inch, 4 inches, 3 feet. If I independently flip three fair coins, what is the expected number of heads? What is the standard deviation?

Topic: Review of Basics – Normal Distribution

- **Continuous random variable**: A continuous random variable can take values with any number of decimals, like 1.2361248912. Weight measured perfectly, with all the decimals and no rounding, is a continuous random variable. Because it can take so many different values, each value winds up having probability zero. If I ask you to guess someone’s weight, not approximately to the nearest millionth of a gram, but rather exactly to all the decimals, there is no way you can guess correctly – each value with all the decimals has probability zero. But for an interval, say the nearest kilogram, there is a nonzero chance you can guess correctly. This idea is captured in by the density function.

- **Density Functions**: A density function defines probability for a continuous random variable. It attaches zero probability to every number, but positive probability to ranges (e.g., nearest kilogram). The probability that the random variable **X** takes values between 3.9 and 6.2 is the area under the density function between 3.9 and 6.2. The total area under the density function is 1.
- **Normal density**: The Normal density is the familiar "bell shaped curve".



The standard Normal distribution has expectation zero, variance 1, standard deviation 1 = $\sqrt{1}$. About 2/3 of the area under the Normal density is between –1 and 1, so the probability that a standard Normal random variable takes values between –1 and 1 is about 2/3. About 95% of the area under the Normal density is between –2 and 2, so the probability that a standard Normal random variable takes values between –2 and 2 is about .95. (To be more precise, there is a 95% chance that a standard Normal random variable will be between –1.96 and 1.96.) If **X** is a standard Normal random variable, and μ and $\sigma > 0$ are two numbers, then $\mathbf{Y} = \mu + \sigma\mathbf{X}$ has the Normal distribution with expectation μ, variance σ^2 and standard deviation σ, which we write N(μ, σ^2). For example, $\mathbf{Y} = 3 + 2\mathbf{X}$ has expectation 3, variance 4, standard deviation 2, and is N(3,4).

- **Normal Plot**: To check whether or not data, $X_1, \ldots, X_n$ look like they came from a Normal distribution, we do a Normal plot. We get the order statistics – just the data sorted into order – or $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ and plot this ordered data against what ordered data from a standard Normal distribution should look like. The computer takes care of the details. A straight line in a

Normal plot means the data look Normal. A straight line with a couple of strange points off the lines suggests a Normal with a couple of strange points (called outliers). Outliers are extremely rare if the data are truly Normal, but real data often exhibit outliers. A curve suggest data that are not Normal. Real data wiggle, so nothing is ever perfectly straight. In time, you develop an eye for Normal plots, and can distinguish wiggles from data that are not Normal.

Topic: Review of Basics – Confidence Intervals

- Let $X_1, \ldots, X_n$ be n independent observations from a Normal distribution with expectation μ and variance σ^2. A compact way of writing this is to say $X_1, \ldots, X_n$ are iid from N(μ, σ^2). Here, iid means independent and identically distributed, that is, unrelated to each other and all having the same distribution.
- How do we know $X_1, \ldots, X_n$ are iid from N(μ, σ^2)? We don't! But we check as best we can. We do a boxplot to check on the shape of the distribution. We do a Normal plot to see if the distribution looks Normal. Checking independence is harder, and we don't do it as well as we would like. We do look to see if measurements from related people look more similar than measurements from unrelated people. This would indicate a violation of independence. We do look to see if measurements taken close together in time are more similar than measurements taken far apart in time. This would indicate a violation of independence. Remember that statistical methods come with a warrantee of good performance if certain assumptions are true, assumptions like $X_1, \ldots, X_n$ are iid from N(μ, σ^2). We check the assumptions to make sure we get the promised good performance of statistical methods. Using statistical methods when the assumptions are not

true is like putting your CD player in washing machine – it voids the warrantee.

- To begin again, having checked every way we can, finding no problems, assume $X_1,\ldots,X_n$ are iid from N(μ, σ^2). We want to estimate the expectation μ. We want an interval that in most studies winds up covering the true value of μ. Typically we want an interval that covers μ in 95% of studies, or a **95% confidence interval**. Notice that the promise is about what happens in most studies, not what happened in the current study. If you use the interval in thousands of unrelated studies, it covers μ in 95% of these studies and misses in 5%. You cannot tell from your data whether this current study is one of the 95% or one of the 5%. All you can say is the interval usually works, so I have confidence in it.
- If $X_1,\ldots,X_n$ are iid from N(μ, σ^2), then the confidence interval uses the sample mean, $\overline{X}$, the sample standard deviation, *s*, the sample size, *n*, and a critical value obtained from the t-distribution with *n-1* degrees of freedom, namely the value, $t_{0.025}$, such that the chance a random variable with a t-distribution is above $t_{0.025}$ is 0.025. If *n* is not very small, say n>10, then $t_{0.025}$ is near 2. The 95% confidence interval is:

$$\overline{X} \pm \text{(allowance for error)} \quad = \quad \overline{X} \pm \frac{t_{0.025} \cdot s}{\sqrt{n}}$$

Topic: Review of Basics – Hypothesis Tests

- **Null Hypothesis**: Let $X_1,\ldots,X_n$ be n independent observations from a Normal distribution with expectation μ and variance σ^2. We have a particular value of μ in mind, say μ_0, and we want to ask if the data contradict this value. It means something special to us if μ_0 is the correct value – perhaps it means the treatment has no effect, so the treatment should be discarded. We wish to test the null hypothesis, $H_0{:}\,\mu = \mu_0$. Is the null hypothesis plausible? Or do the data force us to abandon the null hypothesis?
- **Logic of Hypothesis Tests**: A hypothesis test has a long-winded logic, but not an unreasonable one. We say: Suppose, just for the sake of argument, not because we believe it, that the null hypothesis is true. As is always true when we suppose something for the sake of argument, what we mean is: Let's suppose it and see if what follows logically from supposing it is believable. If not, we doubt our supposition. So suppose μ_0 is the true value after all. Is the data we got, namely $X_1,\ldots,X_n$, the sort of data you would usually see if the null hypothesis were true? If it is, if $X_1,\ldots,X_n$ are a common sort of data when the null hypothesis is true, then the null hypothesis looks sorta ok, and we *accept* it. Otherwise, if there is no way in the world you'd ever see data anything remotely like our data, $X_1,\ldots,X_n$, if the null hypothesis is true, then we can't really believe the null hypothesis having seen $X_1,\ldots,X_n$, and we *reject* it. So the basic question is: Is data like the data we got commonly seen when the null hypothesis is true? If not, the null hypothesis has gotta go.
- **P-values or significance levels**: We measure whether the data are commonly seen when the null hypothesis is true using something called the P-value or significance level. Supposing the null hypothesis to be true, the P-value is the chance of data at least as inconsistent with the null hypothesis as

the observed data. If the P-value is ½, then half the time you get data as or more inconsistent with the null hypothesis as the observed data – it happens half the time by chance – so there is no reason to doubt the null hypothesis. But if the P-value is 0.000001, then data like ours, or data more extreme than ours, would happen only one time in a million by chance if the null hypothesis were true, so you gotta being having some doubts about this null hypothesis.

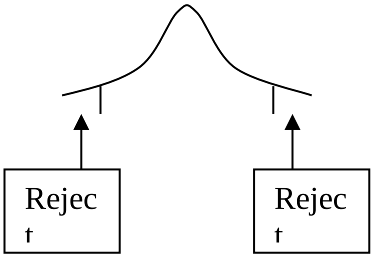
- **The magic 0.05 level:** A convention is that we "reject" the null hypothesis when the P-value is less than 0.05, and in this case we say we are testing at **level** 0.05. Scientific journals and law courts often take this convention seriously. It is, however, only a convention. In particular, sensible people realize that a P-value of 0.049 is not very different from a P-value of 0.051, and both are very different from P-values of 0.00001 and 0.3. It is best to report the P-value itself, rather than just saying the null hypothesis was rejected or accepted.
- **Example**: You are playing 5-card stud poker and the dealer sits down and gets 3 royal straight flushes in a row, winning each time. The null hypothesis is that this is a fair poker game and the dealer is not cheating. Now, there are or 2,598,960 five-card stud poker hands, and 4 of these are royal straight flushes, so the chance of a royal straight flush in a fair game is $\frac{4}{2{,}598{,}960} = 0.000001539$. In a fair game, the chance of three royal straight flushes in a row is 0.000001539x0.000001539x0.000001539 = 3.6×10^{-18}. (Why do we multiply probabilities here?) Assuming the null hypothesis, for the sake of argument, that is assuming he is not cheating, the chance he will get three royal straight flushes in a row is very, very small – that is the P-value or significance level. The data we see is highly improbable if the null hypothesis were true, so we doubt it is true. Either the dealer got very, very lucky, or he cheated. This is the logic of all hypothesis tests.

- **One sample t-test**: Let $X_1, \ldots, X_n$ be n independent observations from a Normal distribution with expectation μ and variance σ^2. We wish to test the null hypothesis, $H_0: \mu = \mu_0$. We do this using the one-sample t-test:

$$t = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{s}$$

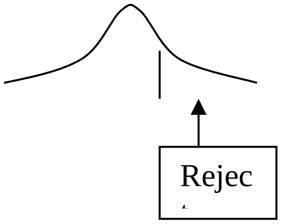
  looking this up in tables of the t-distribution with *n-1* degrees of freedom to get the P-value.

- **One-sided vs Two-sided tests**: In a two-sided test, we don't care whether $\overline{X}$ is bigger than or smaller than μ_0, so we reject at the 5% level when |t| is one of the 5% largest values of |t|. This means we reject for 2.5% of t's that are very positive and 2.5% of t's that are very negative:



In a two sided test we reject when t is big positive or big negative. If we reject when the P-value is less than 0.05, then each tail has probability 0.025.

In a one sided test, we do care, and only want to reject when $\overline{X}$ is on one particular side of μ_0, say when $\overline{X}$ is bigger than μ_0, so we reject at the 5% level when t is one of the 5% largest values of t. This means we reject for the 5% of t's that are very positive:



In a one sided test we reject on just one side, say big positive. If we reject when the P-value is less than 0.05, the tail on the right has probability 0.05.

- **Should I do a one-sided or a two-sided test**: Scientists mostly report two-sided tests.

REGRESSION ASSUMPTIONS

Assumption	If untrue:	How to detect:
Independent errors	95% confidence intervals may cover much less than 95% of the time. Tests that reject with $p<0.05$ may reject true hypotheses more than 5% of the time. You may think you have much more information than you do.	Often hard to detect. Questions to ask yourself: (i) Are the observations clustered into groups, such as several measurements on the same person? (ii) Are observations repeated over time?
Normal errors	Thick tails and outliers may distort estimates, and they may inflate the estimated error variance, so that confidence intervals are too long, and hypothesis tests rarely reject false hypotheses.	Do a Normal quantile plot. This is the one use of the Normal quantile plot. A more or less straight line in the plot suggests the data are approximately Normal.
Errors have constant variance.	Least squares gives equal weight to all observations, but if some observations are much more stable than others, it is not sensible to give equal weight to all observations.	Plot the residuals against the predicted values. A fan shape in the plot – narrow on one end, wide on the other – suggests unequal variances. Can also plot residuals against individual x's.
Model is linear.	Linear model may not fit, or may give the wrong interpretation of the data.	Plot the residuals against the predicted values. Curves, such as a U-shape, suggest the relationship is not linear. Can also plot residuals against individual x's.

Statistics 500: Basic Statistics Review

- **Reading**: In Kleinbaum, read chapter 3.
- **Practice**: The blood pressure data we discussed in class is given below. It is from MacGregor, et. al. (1979) British Medical Journal, 2, 1106-9. It is the change in systolic blood pressure two hours after taking Captopril, in mm Hg, after-before, so a negative number means a decline in blood pressure. Use JMP or another package to do a Normal plot, a boxplot and a t-test. Think about how you would describe what you see.

Patient #	Change in bp
1	-9
2	-4
3	-21
4	-3
5	-20
6	-31
7	-17
8	-26
9	-26
10	-10
11	-23
12	-33
13	-19
14	-19
15	-23

Homework: The following data are from Kaneto, Kosaka and Nakao (1969) Endocrinology, 80, 530-536. It is an experiment on 7 dogs. Question is whether stimulation of the vagus nerve increases levels of immunoreactive insulin in the blood. Two measurements were taken on each dog, one before, one five minutes after stimulation. The measurements are blood lead levels of immunoreative insulin ($\mu U\ /\ ml$).

Dog	Before	After
1	350	480
2	200	130
3	240	250
4	290	310
5	90	280
6	370	1450
7	240	280

Do an appropriate analysis.

Topic: Simple Regression

- **Simple regression**: Fitting a line a response Y using one predictor X.
- **Data**: 48 contiguous states in 1972, *i=1,...,48.* Y = FUEL = motor fuel consumption per person in gallons per person. X = TAX = motor fuel tax rate in cents per gallon.
- **First thing you do**: Plot the data.
- **Least squares**: Fit the line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ by minimizing the sum of the squares of the residuals $Y_i - \hat{Y}_i$ around the line, $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$.
- **Plot the residuals**: After you fit a line, you plot the residuals, $Y_i - \hat{Y}_i$. They tell you where and how the line fits poorly. The minimum is: (i) a boxplot of residuals, (ii) a Normal plot of residuals, (iii) a plot of residuals vs predicted values, $Y_i - \hat{Y}_i$ vs $\hat{Y}_i$.
- **Statistical Model**: The statistical model says:

  $Y_i = \alpha + \beta X_i + \varepsilon_i$ where the ε_i are iid $N(0, \sigma^2)$,

  so the Y's were generated by a true line, $\alpha + \beta X_i$, which we do not know, plus errors ε_i that are independent of each other and Normal with mean zero and constant variance σ^2. We use the residual plots to check whether the model is a reasonable description of the data. The line fitted by least squares, $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, is our estimate of the true line $\alpha + \beta X_i$.
- **Properties of least squares estimates**: The least squares estimators are great estimators – the best there are – when the model is correct, and not so great when the model is wrong. Checking the model is checking whether we are getting good estimates. When the model is true, least squares estimates are **unbiased**, that is, correct in expectation or on average, and they have **minimum variance** among all unbiased estimates, so they are the most stable, most accurate unbiased estimates (but only if the model is correct!).

They are not robust to outliers – one weird observation can move the fitted line anywhere it wants.

- **Basic Regression Output**

Variable	Estimated Coefficient	Estimated Standard Error of Estimated Coefficient	t-ratio
Constant	$\hat{\alpha}$	se($\hat{\alpha}$)	$\frac{\hat{\alpha}}{se(\hat{\alpha})}$
X	$\hat{\beta}$	se($\hat{\beta}$)	$\frac{\hat{\beta}}{se(\hat{\beta})}$

- **Hypothesis tests**: Use the t-ratio to test the null hypothesis $H_0{:}\beta = 0$. Under the model, the hypothesis $H_0{:}\beta = 0$ implies X and Y are unrelated.
- **Confidence intervals**: Under the model, a 95% confidence interval for β is:

$$estimate \pm allowance \quad = \quad \hat{\beta} \pm t_{0.025} \cdot se(\hat{\beta})$$

where $t_{0.025}$ is the upper 2.5% point of the t-distribution with n-2 degrees of freedom. When n-2 is not small, the interval is almost (but not quite) $\hat{\beta} \pm 2 \cdot se(\hat{\beta})$.

- **Points on a line vs Predictions**: Two problems look almost the same, but really are very different. One asks: Where is the line at X=8.5? That is, what is $\alpha + \beta 8.5$? That problem gets easier as I collect more data and learn where the line really is. The other asks: Where will a new observation on Y be if X=8.5? That is, what is $\alpha + \beta 8.5 + \varepsilon_{new}$? That problem always stays pretty hard, no matter how much data I collect, because I can't predict the new error, ε_{new}, for this new observation no matter how well I know where the line is. Important thing is to make sure you know which answer you want and to use the right method for that answer. They look similar, but they're not.

- **Regression Anova Table**: Partitions the total variation (or sum of squares) in the data about the mean, namely $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ into two parts that add back to the total, namely the variation fitted by the regression, $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$, and the variation in the residuals, $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. Degrees of freedom measure keep track of how many distinct numbers are really being described by a sum of squares. In simple regression, the variation fitted by the regression is just fitted by the slope, $\hat{\beta}$, which is just one number, so this sum of squares has 1 degree of freedom. A mean square is the ratio $\frac{\text{sum of squares}}{\text{degrees of freedom}}$. The F-ratio is the ratio of two mean squares, a signal to noise ratio. The F-ratio is used to test that all the slopes are zero.
- **Simple correlation**: If the data fall perfectly on a line tilted up, the correlation r is 1. If the data fall perfectly on a line tilted down, the correlation r is –1. If a line is not useful for predicting Y from X, the correlation r is 0. Correlation is always between –1 and 1. The correlation between Y and X is the regression coefficient of standardized Y on standardized X, that is, the regression of $\frac{Y_i - \overline{Y}}{st.dev(Y)}$ on $\frac{X_i - \overline{X}}{st.dev(X)}$. In simple, one-predictor regression, the square of the correlation, r^2, is the percent of variation fitted by the regression, so it summarizes the anova table. Correlation discards the units of measurement, which limits its usefulness.

Homework: Vocabulary Data

Homework: The following data are from M. E. Smith, (1926), "An investigation of the development of the sentence and the extent of vocabulary in young children." It relates the X=age of children in years to their Y=vocabulary size in words. I would like you to do a regression of Y and X, look closely at what you've done, and comment on what it all means. You should turn in (1) one paragraph of text, (2) linear regression output, (3) at most two plots you find interesting and helpful in thinking about what is special about these data. This is real data, so it is not a "trick question", but it does require some real thought about what makes sense and what is happening.

X=age	**Y=vocabulary**
0.67	0
0.83	1
1	3
1.25	19
1.5	22
1.75	118
2	272
2.5	446
3	896
3.5	1,222
4	1,540
4.5	1,870
5	2,072
5.5	2,289
6	2,562

Topic: Multiple Regression

- **Multiple regression**: Uses several predictor variables $X_1, X_2, \ldots X_k$ to fit a single response variable Y.
- **FUEL DATA**: Trying to predict Y = FUEL from X_1 = TAX and a second predictor, X_2 = LICENSES.
- **Least squares fit**: Multiple regression fits a plane $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$ making the residuals $Y_i - \hat{Y}_i$ small, in the sense that the sum of the squares of the residuals, namely, $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, is minimized.
- **Multiple Correlation**: The multiple correlation, R, is the ordinary correlation between the observed Y_i and the fitted $\hat{Y}_i$. The square of the multiple correlation, R^2, is the percent of variation fitted by the regression, that is, regression sum of squares in the ANOVA table divided by the total sum of squares of Y around its mean.
- **Fit vs Prediction**: Fit refers to how close the model is to the observed data. Predicition refers to how close the model is to new data one might collect. They are not the same. Adding variables, even junk variables, always improves the fit, but the predictions may get better or worse. R^2 is a measure of fit, not of prediction. We will develop a measure of prediction, C_P, later in the course.
- **Statistical Model**: The model underlying multiple regression says:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

where the ε_i are independent $N(0, \sigma^2)$

The true model is unknown, but the least squares fit is an estimate.

- **Hypothesis Tests and Confidence Intervals for a Coefficient**: Testing a hypothesis about a regression coefficient, say $H_0: \beta_5 = 0$, is done using the t-statistic as in simple regression. Confidence intervals are also done as in simple regression.

- **Testing that all coefficients are zero**: The F-test from the ANOVA table is used to test $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$.
- **Residual analysis**: One checks the model by plotting the residuals. The minimum is a plot of residuals against predicted, a boxplot of residuals, and a Normal plot of residuals, as in simple regression.

Topic: General Linear Hypothesis

- What is it? In model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

where the ε_i are independent $N(0,\sigma^2)$,

we know how to test a hypothesis about one coefficient, say $H_0: \beta_5 = 0$, (t-test) and we know how to test that all of the variables are unneeded, $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (F-test from regression anova table). The general linear hypothesis says that a particular subset of the coefficients is zero. For example, the hypothesis might say that the last *k–J* variables are not needed, $H_0: \beta_{J+1} = \beta_{J+2} = \ldots = \beta_k = 0$.

- Why do this? Generally, a hypothesis expresses an idea. Some ideas need to be expressed using more than one variable. For example, in the FUEL data, the 48 states might be divided into five regions, Northeast, Southeast, Midwest, Mountain, and Pacific, say. Later on, we will see how to code region into several variables in a regression. Testing whether "REGION" matters is testing whether all of these variables can be dropped from the model.
- Comparing Two Models: The test involves comparing two models, a reduced model which assumes the hypothesis is true, and a full model which assumes it is false. To test $H_0: \beta_{J+1} = \beta_{J+2} = \ldots = \beta_k = 0$, one fits the full model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

and the reduced model without variables $X_{J+1}, \ldots, X_k$,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_J X_{Ji} + \varepsilon_i,$$

and the test is based on comparing the ANOVA tables for these two models. Details in the textbook.

Topic: Woes of Regression Coefficients

- The phrase, "the woes of regression coefficients" is due to Fred Mosteller and John Tukey in a standard text, *Data Analysis and Regression.* The bulk pack contains another standard reading: George Box's paper "The use and abuse of regression". (The paper is mostly easy to read, but contains some technical material – just skip the technical material. The main points are not technical and not difficult.)
- The issue concerns the interpretation of regression coefficients. The bottom line is that it is hard to interpret regression coefficients. The reason is that whenever you add (or delete) a variable from a regression model, all of the other coefficients change to reflect the added (or deleted) variable. People want (but they can't have) a way of speaking of THE coefficient of a variable, but actually the coefficient of a variable always depends on what other variables are in the model with it. People want to say that β_j is the change in Y expected from a one unit change in X_j, but it simply isn't true. It can't be true, since β_j keeps changing as variables are added or deleted from a model, whereas changing X_j out in the world has nothing to do with which variables I put in the model.
- The bottom line is this: Whenever you hear people say that changing X_j will produce a particular change in Y, and they say they know this solely because they did a regression, you should be a little skeptical. There is more to knowing something like this than just running regressions.

Topic: Transformations

- **Key Idea**: Fit many kinds of curved (i.e., nonlinear) models by transforming the variables and fitting a linear model to the transformed variables.
- **Logs**. For b>0, the base b log has the property that:

  $$y = b^a \quad \text{is the same as} \quad \log_b(y) = a.$$

  Common choices of the base b are b=10, b=2 and b=e=2.71828… for natural logs. Outside high school, if no base is mentioned (e.g., log(y)) it usually means base e or natural logs. Two properties we use often are: log(xy)=log(x)+log(y) and $\log(y^a) = a \cdot \log(y)$.
- **Why transform?** (i) You plot the data and it is curved, so you can't fit a line. (ii) The Y's have boundaries (e.g., Y must be >0 or Y must be between 0 and 1), but linear regression knows nothing of the boundaries and overshots them, producing impossible $\hat{Y}'s$. (iii) The original data violate the linear regression assumptions (such as Normal errors, symmetry, constant variance), but perhaps the transformed variables satisfy the assumptions. (iv) If some Y's are enormously bigger than others, it may not make sense to compare them directly. If Y is the number of people who work at a restaurant business, the Y for McDonald's is very, very big, so much so that it can't be compared to the Y for Genji's (4002 Spruce & 1720 Samson). But you could compare log(Y).
- **Family of transformations**: Organizes search for a good transformation. Family is $\frac{Y^p - 1}{p}$ which tends to log(y) as p gets near 0. Often we drop the shift of 1 and the scaling of 1/p, using just $sign(p) \cdot Y^p$ for $p \neq 0$ and log(y) for p=0. Important members of this family are: (i)

p=1 for no transformation or Y, (ii) p=1/2 for $\sqrt{Y}$, (iii) p=1/3 for $\sqrt[3]{Y}$, (iv) p=0 for log(y), (v) p = –1 for 1/Y.

- **Straightening a scatterplot**: Plot Y vs X. If the plot looks curved, then do the following. Divide the data into thirds based on X, low, middle, high. In each third, find median Y and median X. Gives you three (X,Y) points. Transform Y and/or X by adjusting p until the slope between low and middle equals the slope between middle and high. Then plot the transformed data and see if it looks ok. You want it to look straight, with constant variance around a line.
- **Logit**: logit(a) =log{a/(1-a)} when a is between 0 and 1. If the data are between 0 and 1, their logits are unconstrained.
- **Picking Curves that Make Sense**: Sometimes we let the data tell us which curve to fit because we have no idea where to start. Other times, we approach the data with a clear idea what we are looking for. Sometimes we know what a sensible curve should look like. Some principles – (i) If the residuals show a fan pattern, with greater instability for larger Y's, then a log transformation may shift things to constant variance. (ii) If there is a naïve model based on a (too) simple theory (e.g., weight is proportional to volume), then consider models which include the naïve theory as a very special case. (iii) If outcomes Y must satisfy certain constraints (e.g., percents must be between 0% and 100%), consider families of models that respect those constraints.
- **Interpretable transformations**: Some transformations have simple interpretations, so they are easy to think and write about. Base 2 logs, i.e., $\log_2(y)$ can be interpreted in terms of doublings. Reciprocals, 1/Y, are often interpretable if Y is a ratio (like density) or a time. Squares and

square roots often suggest a relationship between area and length or diameter. Cubes and cube roots suggest a relationship between volume and diameter.

- **Transformations to constant variance**: A very old idea, which still turns up in things you read now and then. Idea is that certain transformations – often strange ones like the arcsin of the square root – make the variance nearly constant, and that is an assumption of regression.

Topic: Polynomials

Why fit polynomials? The transformations we talked about all keep the order of Y intact – big Y's have big transformed Y's. Often that is just what we want. Sometimes, however, we see a curve that goes down and comes back up, like a $\cup$, or goes up and comes back down, like a $\cap$, and the transformations we looked at don't help at all. Polynomials can fit curves like this, and many other wiggles. They're also good if you want to find the X that maximizes Y, the top point of the curve $\cap$.

- **Quadratic**: $y = a + bx + cx^2$ has a $\cup$ shape if c>0 and a $\cap$ shape if c<0 (why?) and is a line if c=0. Top of hill or bottom of valley is at $x = \frac{-b}{2c}$.
- **Fitting a Quadratic**: Easy – put two variables in the model, namely X and X^2.
- **Centering**: If X>0, then X is big at the same time X^2, so these two variables are highly correlated. Often a good idea to center, using X and $(X - \overline{X})^2$ instead of X and X^2. Fits the same curve, but is more stable as a computing algorithm.
- **Orthogonal polynomials**: Typically used in anova rather than in regression. Transforms X^2 so it is uncorrelated with X. Does this by regressing X^2 on X and using the residuals in place of X^2.
- **Cubics**: Can fit cubics using X, X^2 and X^3. Usually don't go beyond cubics. Usually center.
- **Polynomials in several predictors**: If I have two predictors, say x and w, the quadratic in x and w has squared terms, x^2 and w^2, but it adds something new, their crossproduct or interaction, xw:

$$y = a + b \cdot x + c \cdot w + d \cdot x^2 + f \cdot w^2 + h \cdot w \cdot x$$

- **Are quadratic terms needed?** You can judge whether you need several quadratic terms using a general linear hypothesis and its avova table.

Topic: Coded Variables (i.e., Dummy Variables)

- **Why use coded variables?** Coded or dummy variables let you incorporate nominal data (Philly vs New York vs LA) as predictors in regression.
- **Two categories**: If there are just two categories, say male and female, you include a single coded variable, say C=1 for female and C=0 for male. Fits a parallel line model. If you add interactions with a continuous variable, X, then you are fitting a two-line model, no longer a parallel line model.
- **More than Two Categories**: If there are 3 categories (Philly vs New York vs LA) then you need two coded variables to describe them (C=1, D=0 for New York; C=0, D=1 for LA; C=0, D=0 for Philly). Such a model compares each group to the group left out, the group without its own variable (here, Philly). When there are more than two categories – hence more than one coded variable – interesting hypotheses often involve several variables and are tested with the general linear hypothesis. Does it matter which group you leave out? Yes and no. Had you left out NY rather than Philly, you get the same fitted values, the same residuals, the same overall F-test, etc. However, since a particular coefficient multiplies a particular variable, changing the definition of a variable changes the value of the coefficient.

Topic: Diagnostics -- Residuals

- **Why do we need better residuals**?: We look at residuals to see if the model fits ok – a key concern for any model. But the residuals we have been looking at are not great. The problem is that least squares works very hard to fit data points with extreme X's – unusual predictors – so it makes the residuals small in those cases. A data point with unusual X's is called a high leverage point, and we will think about them in detail a little later. A single outlier (weird Y) at a high leverage point can pull the whole regression towards itself, so this point looks well fitted and the rest of the data looks poorly fitted. We need ways of finding outliers like this. We want regression to tell us what is typical for most points – we don't want one point to run the whole show.
- The model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

$$\text{where the } \varepsilon_i \text{ are independent } N(0, \sigma^2)$$

- By least squares, we estimate the model to be:

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$ with residuals $E_i = Y_i - \hat{Y}_i$

- Although the true errors, ε_i have constant variance, $\text{var}(\varepsilon_i) = \sigma^2$, the same for every unit i, the residuals have different variances, $\text{var}(E_i) = \sigma^2(1 - h_i)$ where h_i is called the leverage.

- The standardized residual we like best does two things: (i) it uses the leverages h_i to give each residual the right variance for that one residual, and (ii) it removes observation i when estimating the variance of the residual $E_i = Y_i - \hat{Y}_i$ for observation i. That is, $\text{var}(E_i) = \sigma^2(1 - h_i)$ is estimated by $\hat{\sigma}_{[-i]}{}^2(1-h_i)$, where $\hat{\sigma}_{[-i]}{}^2$ is the estimate of the residual variance we get by setting observation i aside and fitting the regression without it.

- **The residual we like has several names** – studentized, deleted, jacknife – no one of which is used by everybody. It is the residual divided by its estimated standard error:

$$r_{[i]} = \frac{E_i}{\sigma_{[-i]}\sqrt{1-h_i}}$$

- Another way to get this residual is to create a coded variable that is 1 for observation i and 0 for all other observations. Add this variable to your regression. The t-statistic for its coefficient equals $r_{[i]}$.

- We can test for outliers as follows. The null hypothesis says there are no outliers. If there are n observations, there are n deleted residuals $r_{[i]}$. Find the largest one in absolute value. To test for outliers at level 0.05, compute 0.025/n, and reject the hypothesis of no outliers if the largest absolute deleted residual is beyond the 0.025/n percentage point of the t-distribution with one less degree of freedom than the error line in the anova table for the regression. (You lose one degree of freedom for the extra coded variable mentioned in the last paragraph.)

Topic: Diagnostics -- Leverage

- **Three very distinct concepts**: An outlier is an observation that is poorly fitted by the regression – it has a response Y that is not where the other data points suggest its Y should be. A high leverage point has predictors, X's, which are unusual, so at these X's, least squares relies very heavily on this one point to decide where the regression plane should go – a high leverage point has X's that allow it to move the regression if it wants to. A high influence point is one that did move the regression – typically, such a point has fairly high leverage (weird X's) and is fairly poorly fitted (weird Y for these X's); however, it may not be the one point with the weirdest X or the one point with the weirdest Y. People often mix these ideas up without realizing it. Talk about a weird Y is outlier talk; talk about a weird X is leverage talk; talk about a weird Y for these X's is influence talk. We now will measure leverage and later influence.
- **Measuring Leverage**: Leverage is measured using the leverages h_i we encountered when we looked at the variance of the residuals. The leverages are always between 0 and 1, and higher values signify more pull on the regression.
- **When is leverage large?** If a model has k predictors and a constant term, using n observations, then the average leverage, averaging over the n observations is always $\frac{k+1}{n} = \frac{1}{n}\sum_{i=1}^{n} h_i$. A rule a thumb that works well is that leverage is large if it is at least twice the average, $h_i \geq \frac{2(k+1)}{n}$.
- **What do you do if the leverage is large?** You look closely. You think. Hard. You find the one or two or three points with $h_i \geq \frac{2(k+1)}{n}$ and you

look closely at their data. What is it about their X's that made the leverage large? How, specifically, are they unusual? Is there a mistake in the data? If not, do the X's for these points make sense? Do these points belong in the same regression with the other points? Or should they be described separately? Regression gives high leverage points a great deal of weight. Sometimes that makes sense, sometimes not. If you were looking at big objects in our solar system, and X=mass of object, you would find the sun is a high leverage point. After thinking about it, you might reasonably decide that the regression should describe the planets and the sun should be described separately as something unique. With the solar system, you knew this before you looked at the data. Sometimes, you use regression in a context where such a high leverage point is a discovery. (If you remove a part of your data from the analysis, you must tell people you did this, and you must tell them why you did it.)

- **Interpretation of leverage**: Leverage values h_i have several interpretations. You can think of them as the distance between the predictors X for observation i and the mean predictor. You can think of them as the weight that observation i gets in forming the predicted value $\hat{Y}_i$ for observation i. You can think of leverages as the fraction of the variance of Y_i that is variance of $\hat{Y}_i$. We will discuss these interpretations in class.

Topic: Diagnostics -- Influence

- **What is influence?** A measure of influence asks whether observation i *did* move the regression. Would the regression change a great deal if this one observation were removed? Not whether it *could* move the regression – that's leverage. Not whether it fits poorly – that's an outlier.
- **Measures of influence.** There are several measures of influence. They are all about the same, but no one has become the unique standard. Two common choices are DFFITS and Cook's Distance. Cook's distance is (almost) a constant times the square of DFFITS, so it makes little difference which one you use. It is easier to say what DFFITS does.
- **What is DFFITS?** Roughly speaking, DFFITS measures the change in the predicted value for observation i when observation it is removed from the regression. Let $\hat{Y}_i$ be the predicted value for observation i using all the data, and let $\hat{Y}_{i[i]}$ be the predicted value for observation i if we fit the regression without this one observation. Is $\hat{Y}_i$ close to $\hat{Y}_{i[i]}$? If yes, then this observation does not have much influence. If no, then it does have influence. DFFITS divides the difference, $\hat{Y}_i - \hat{Y}_{i[i]}$ by an estimate of the standard error of $\hat{Y}_i$, so a value of 1 means a movement of one standard erro. Recall that $\hat{\sigma}_{[-i]}{}^2$ is the estimated residual variance when observation i is removed from the regression. Then DFFITS is:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i[i]}}{\hat{\sigma}_{[-i]}\sqrt{h_i}}\,.$$

- **DFBETAS**: A related quantity is DFBETAS which looks at the standardized change in the regression coefficient $\hat{\beta}_j$ when observation i is removed. There is one DFBETAS for each observation and for each coefficient. DFFITS is always bigger than the largest DFBETAS, and there is only one DFFITS per observation, so many people look at DFFITS instead of all k DFBETAS.

Topic: Variable Selection

- **What is variable selection?** You have a regression model with many predictor variables. The model looks ok – you've done the diagnostic checking and things look fine. But there are too many predictor variables. You wonder if you might do just as well with fewer variables. Deciding which variables to keep and which to get rid of is variable selection.
- **Bad methods**. There are two bad methods you should not use. One bad method is to drop all the variables with small t-statistics. The problem is the t-statistic asks whether to drop a variable *providing you keep all the others*. The t-statistic tells you little about whether you can drop two variables at the same time. It might be you could drop either one but not both, and t can't tell you this. Another bad method uses the squared multiple correlation, R^2. The problem is R^2 always goes up when you add variables, and the size of the increase in R^2 is not a great guide about what to do. Fortunately, there is something better.
- **A good method**. The good method uses a quantity called C_P which is a redesigned R^2 built for variable selection. Suppose the model is, as before,

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

  where the ε_i are independent $N(0,\sigma^2)$, but now k is large (many predictors) and we think some β's might be zero. We fit this model and get the usual estimate $\hat{\sigma}^2$ of σ^2. A submodel has some of the k variables but not all of them, and we name the submodel by the set P of variables it contains. So the name of the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_9 X_{9i} + \varepsilon_i$ is

P={1,3,9}, and it has residual sum of squares SSE_P from the residual line in its Anova table, and p=3 variables plus one constant term or 4 parameters. (Note carefully – I let p=#variables, but a few people let p=#parameters.) We have n observations. Then the strange looking but simple formula for C_P is: $C_P = \frac{SSE_P}{\hat{\sigma}^2} - [n - 2(p+1)]$. Then C_P compares the model with all variable to the model with just the variables in P and asks whether the extra variables are worth it.

- Using C_P: The quantity C_P estimates the standardized total squared error of prediction when using model P in place of the model with all the variables. We like a model P with a small C_P. If a model P contains all the variables with nonzero coefficients, then C_P tends on average to estimate p+1, the number of variables plus 1 for the constant, so a good value of C_P is not much bigger than p+1. For instance, if $C_{\{1,3,9\}}$ = 8, then that is much bigger than p+1=3+1=4, so the model seems to be missing important variables, but if $C_{\{1,3,9,11\}}$ =5.1 then that is close to p+1=4+1=5 and smaller than 8, so that model predicts better and might have all important variables.
- **Searching**: If a model has k variables, then there are 2^k submodels formed by dropping variables, or about a billion models for k=30 variables. There are various strategies for considering these models: forward selection, backward elimination, stepwise, all subsets, best subsets.
- **Cautions**: Variable selection is an exploratory method, one that looks for interesting things, but because it searches so extensively, it may find some things that don't replicate. If we reject hypotheses when P-

value<0.05, then only 1 time in 20 do we reject a true hypothesis. But if we fit billions for regressions, calculating billions of P-values, then we reject many true hypotheses and make many mistakes. The results of variable selection need to be examined with caution avoiding overstatement. A good strategy is to split the sample, perform variable selection on one half, and confirm the results on the other. This is a simple type of cross-validation.

Topic: One Way Analysis of Variance

- **What is ANOVA**? Anova, or analysis of variance, is the decomposition of data into parts that add back up to the original data, and the summary of the parts in terms of their sizes measured by summing and squaring their numerical entries. At an abstract level, in statistical theory, anova and regression are not really different. In practice, however, they look very different. Most computer programs have separate routines for regression and anova. Center questions, issues and methods arise in anova that don't arise in regression. Anova tends to be used with structured data sets, often from carefully designed experiments, while regression is often used with data that arises naturally. However, by running enough regressions, knowing exactly what you are doing, and putting together the pieces very carefully, you can do even a complex anova using a regression program – it's easy to use an anova program. Anova has a nice geometry.
- **What is one-way anova?** One-way anova is the very simplest case. People fall into one of several groups and we want to understand the difference between the groups. Basic questions are: Do the groups differ? (F-test.) If so, how? (Multiple comparisons.) If I anticipate a specific pattern of differences between the groups, is this pattern confirmed by the data? (Contrasts.) Notation: There are I groups, i=1,…,I, and n_i people in group I, with $n = n_1 + \ldots + n_I$ people in total. Each person is in just one group and people in different groups have nothing to do with each other. Person j in group i, has response y_{ij}. The

mean response in group i is $\overline{y}_{i\bullet}$ and the mean response for everyone is $\overline{y}_{\bullet\bullet}$. The anova decomposition is:

$$y_{ij} = \overline{y}_{\bullet\bullet} + (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}) + (y_{ij} - \overline{y}_{i\bullet})$$

data = (grand mean) + (group difference) + (residual)

Model for One-Way Anova: The model for one-way anova says the observations are Normal with the same variance, are independent of each other, and have different means in the different groups. Specifically: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where the ε_{ij} are iid $N(0,\sigma^2)$ and $0 = \alpha_1 + \alpha_2 + \ldots + \alpha_I$.

- **Do the groups differ?** We test the hypothesis that the groups do not differ using the F-ratio from the one-way analysis of variance table. If F is large enough, judged by the F-table, we reject the null hypotheses and conclude there is strong evidence the groups differ. Otherwise, we conclude that we lack strong evidence that the groups differ (which is not the same thing as saying we know for certain they are the same).
- **If the groups differ, how do they differ?** It is not enough to know the groups differ – we need to understand what differences are present. There are two cases: (1) we have no idea what we are looking for, or (2) we have a clear and specific idea what we are looking for. Case 2 is the better case – if you know what you are looking for in statistics, then you can find it with a smaller sample size. We handle case 1 using multiple comparisons and case 2 using contrasts, described later. In multiple comparison, every group is compared to every other group. If there are I=10 groups, there are 45 comparisons of two groups, 1 with 2, 1 with 3, …, 9 with 10. If you did 45 t-tests to compare the groups, rejected for P-

values < 0.05, then you would falsely reject a true hypothesis of no difference in one out of 20 tests. This means that with I=10 groups and 45 comparisons, if the groups are really the same, you expect to get 45x0.05 = 2.25 significant (P-value<0.05) difference by chance alone. That's a problem – a big problem. It means you expect 2 mistakes – you expect to say a treatment worked when it didn't. Gotta do something to prevent this. There are many, many solutions to this problem – there are whole books of multiple comparison procedures. One of the first is due to John Tukey of Princeton. You can think of it as using essentially a t-statistic to compare groups in pairs, but as the number of groups, I, gets bigger, so more tests are being done, the procedure requires a bigger value of the t-statistic before declaring the difference significant. If you did this with I=10 groups and 45 comparisons of two groups, the procedure promises that if the groups are really the same, the chance of a significant difference anywhere in the 45 comparisons is less that 0.05 – if you find anything, you can believe it. For example, with just two groups and 30 degrees of freedom, we would reject at the 0.05 level if |t|>2.04, but using Tukey's method with I=10 groups, 45 comparisons, we would reject if |t|>3.41, and if t is that big, then it is unlikely to happen by chance even if you did 45 comparisons.

- **Planned Contrasts for Specific Hypotheses:** Ideally, when you do research, you have a clear idea what you are looking for and why. When this is true, you can build a test tailored to your specific hypothesis. You do this with contrasts among group means. You express your hypothesis using a set of contrast weights you pick, one weight for each group mean, summing to zero: $c_1, c_2, \ldots, c_I$ with $c_1 + c_2 + \ldots + c_I = 0$. For instance, consider a study with I=3 groups, with $n_1 = n_2 = n_3 = 100$ people in each group. The groups are two different drug treatments, A and B, and a placebo control. Then the contrast "drug vs placebo" is:

Contrast: Drug vs Placebo

Placebo	Drug A	Drug B
c_1	c_2	c_3
-1	$\frac{1}{2}$	$\frac{1}{2}$

whereas the contrast "drug A vs drug B" is:

Contrast: Drug A vs Drug B

Placebo	Drug A	Drug B
d_1	d_2	d_3
0	1	-1

- The value of contrast applies the contrast weights to the group means, $L = \sum_{i=1}^{I} c_i \cdot \overline{y}_{i\bullet}$, so for "Drug vs Placebo" it is $L = -1 \cdot \overline{y}_{1\bullet} + \frac{1}{2} \cdot \overline{y}_{2\bullet} + \frac{1}{2} \cdot \overline{y}_{3\bullet}$
- The t-test for a contrast tests the null hypothesis $H_0\colon\ 0 = \sum_{i=1}^{I} c_i \cdot \alpha_i$. Let $\hat{\sigma}^2$ be the residual mean square from the anova table, which estimates σ^2. The t-statistic is $t = \dfrac{L}{\sqrt{\hat{\sigma}^2 \cdot \sum \frac{c_i^2}{n_i}}}$ and the degrees of freedom are from the residual line in the anova table.
- The sum of squares for a contrast is $\dfrac{L^2}{\sum \frac{c_i^2}{n_i}}$. Two contrasts, $c_1, c_2, \ldots, c_I$ and $d_1, d_2, \ldots, d_I$ are orthogonal if $0 = \sum_{i=1}^{I} \frac{c_i \cdot d_i}{n_i}$. Example: "Drug vs Placebo" is orthogonal to "Drug A vs Drug B" because

$\sum_{i=1}^{I} \frac{c_i \cdot d_i}{n_i} = \frac{-1 \times 0}{100} + \frac{\frac{1}{2} \times 1}{100} + \frac{\frac{1}{2} \times -1}{100} = 0$. When contrasts are orthogonal, the sum of squares between groups may be partitioned into separate parts, one for each contrast. If there are I groups, then there are I-1 degrees of freedom between groups, and each degree of freedom can have its own contrast. Both of these formulas are mostly used in balanced designs where the sample sizes in the groups are the same, $n_1 = n_2 = \ldots = n_I$.

Topic: Two Way Analysis of Variance

- What is two-way ANOVA? In two-way anova, each measurement is classified into groups in two different ways, as in the rows and columns of a table. In the social sciences, the most common situation is to measure the same unit or person under several different treatments – this is the very simplest case of what is know as repeated measurements. Each person is a row, each treatment is a column, and each person gives a response under each treatment. The two-way's are person and treatment. Some people give higher responses than others. Some treatments are better than others. The anova measures both sources of variation. The units might be businesses or schools or prisons instead of people.

  Notation: There are I people, i=1,…,I, and J treatments, j=1,…,J, and person i gives response y_{ij} under treatment j. The mean for person i is $\bar{y}_{i\bullet} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}$ and the mean for treatment j is $\bar{y}_{\bullet j} = \frac{1}{I}\sum_{i=1}^{I} y_{ij}$, and the mean of everyone is $\bar{y}_{\bullet\bullet}$. The anova decomposition is:

  $y_{ij} = \bar{y}_{\bullet\bullet} + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})$.

- **Anova table**: The anova table now has “between rows”, “between columns” and “residual”, so the variation in the data is partitioned more finely.
- **Normal model**: The model is $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ where the errors are independent Normals with mean zero and variance σ^2. Under this model, F-statistics from the anova table may be used to test the hypotheses of no difference between rows and no difference between columns. Can do multiple comparisons and contrasts using the residual line from the anova table to obtain the estimate $\hat{\sigma}^2$.

Error Rates When Performing More Than One Hypothesis Test

Table Counts Null Hypotheses

	Accepted or untested null hypotheses	Rejected null hypotheses	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	m- m_0
Total	m-R=m-(U+T)	R	m

Family-wise error rate: Pr(V≥1), the probability of at least one false rejection in m tests.

The family-wise error rate is weakly controlled at α=0.05 if α=0.05≥Pr(V≥1) whenever m= m_0, that is, whenever all m null hypotheses are true.

The family-wise error rate is strongly controlled at α=0.05 if α=0.05≥Pr(V≥1) for all values of m_0, that is, no matter how many null hypotheses are true.

Weak control is not enough. Weak control means you are unlikely to find something when there is nothing, but you are still likely to find too much when there is something.

False discovery rate (FDR) is the expected number of false rejections, E(V/R) where E/R=0/0 is defined to be 0 (i.e., no false rejections if no rejections). This is a more lenient standard than the family-wise error rate, rejecting more true hypotheses.

If you do m tests at level α=0.05, you expect to falsely reject 0.05 x m_0 hypotheses, and if all hypotheses are true, this might be as high as 0.05 x m. The expected ratio of false rejections to tests, E(V/m), is called the per comparison error rate.

Example
Table Counts Null Hypotheses

	Accepted or untested null hypotheses	Rejected null hypotheses	Total
True null hypotheses	U	V	100
False null hypotheses	T	S	1
Total	m-R=m-(U+T)	R	101

In this example, there are 101 hypotheses and 100 are true.
If you test each hypothesis at level α=0.05, you expect 0.05 x 100 = 20 false rejections of true null hypotheses, plus if you are lucky a rejection of the one false null hypothesis, so you expect most rejections to be false rejections.

If you strongly control the family-wise error rate at α=0.05, then the chance of at least one false rejection is at most 5%.

If you weakly control the family-wise error rate at α=0.05, then there are no promises about false rejections in this case, as one null hypothesis is false.

What are adjusted P-values? (e.g. as produced by `pairwise.t.test()`)

A test of null hypothesis H_0 either rejects H_0 or it does not.
The level, α, of the test is such that $\alpha \geq \Pr(\text{Reject } H_0)$ when H_0 is true.
The P-value is the smallest α such that we reject H_0.
This definition of a P-value continues to work with multiple hypothesis testing.

Topic: Factorial Analysis of Variance

- **Two Factor Factorial Anova**: The simplest case of factorial anova involves just two factors – similar principles apply with more than two factors, but things get large quickly. Suppose you have two drugs, A and B – then “drug” is the first factor, and it has two levels, namely A and B. Suppose each drug has two dose levels, low and high – then “dose” is the second factor, and it too has two levels, low and high. A person gets one combination, perhaps drug B at low dose. Maybe I give 50 people each drug at each level, so I have 200 people total, 100 on A, 100 on B, 100 at low dose, 100 at high dose.

  Main effects and interactions: We are familiar with main effects – we saw them in two-way anova. Perhaps drug A is better than drug B – that’s a main effect. Perhaps high dose is more effective than low dose – that’s a main effect. But suppose instead that drug A is better than drug B at high dose, but drug A is inferior to drug B at low dose – that’s an interaction. In an interaction, the effect of one factor changes with the level of the other.

- **Anova table**: The anova table has an extra row beyond that in two-way anova, namely a row for interaction. Again, it is possible to do contrasts and multiple comparisons.
- **More Complex Anova**: Anova goes on and on. The idea is to pull apart the variation in the data into meaningful parts, each part having its own row in the anova table. There may be many factors, many groupings, etc.

Some Aspects of R

Script is my commentary to you. **`Bold Courier is what I type in R.`** `Regular Courier is what R answered.`

What is R?

R is a close relative of Splus, but R is available for free. You can download R from http://cran.r-project.org/ . *R is very powerful and is a favorite (if not the favorite) of statisticians; however, it is not easiest package to use. It is command driven, not menu driven, so you have to remember things or look them up – that's the only thing that makes it hard. You can add things to R that R doesn't yet know how to do by writing a little program. R gives you fine control over graphics. Most people need a book to help them, and so Mainland & Braun's Data Analysis and Graphics Using R, Cambridge University Press. Abnother book is Dalgaard's Introductory Statistics with R, NY: Springer. Dalgaard's book is better at teaching basic statistics, and it is good if you need a review of basic statistics to go with an introduction to R. R is similar to Splus, and there are many good books about Splus. One is: Venables and Ripley Modern Applied Statistics with S-Plus (NY: Springer-Verlag).*

Who should use R?

If computers terrify you, if they cause insomnia, cold sweats, and anxiety attacks, perhaps you should stay away from R. On the other hand, if you want a very powerful package for free, one you won't outgrow, then R worth a try. If you find you need lots of help to install R or make R work, then R isn't for you. Alternatives for Statistics 500 are JMP-IN, SPSS, Systat, Stata, SAS and many others. For Statistics 501, beyond the basics, R is clearly best.

You need to download R the first time from the webpage above.

You need to get the "Rst500" workspace for the course from
http://www-stat.wharton.upenn.edu/
going to "Course downloads" and the most recent Fall semester, Statistics 500, or in one step to
http://download.wharton.upenn.edu/download/pub/stat/Fall-2006/STAT-500/
For Statistics 501,
http://stat.wharton.upenn.edu/statweb/course/Spring-2007/stat501/

Start R.
From the File Menu, select "Load Workspace".
Select "Rst500"

To see what is in a workspace, type

```
     ls()
```

or type

```
     objects()
```

```
> ls()
[1] "fuel"
```

To display an object, type its name

```
> fuel
   ID state Fuel   Tax License   Inc   Road
1   1    ME  541  9.00    52.5 3.571  1.976
2   2    NH  524  9.00    57.2 4.092  1.250
3   3    VT  561  9.00    58.0 3.865  1.586
                     .
                     .
                     .
46 46    WN  510  9.00    57.1 4.476  3.942
47 47    OR  610  7.00    62.3 4.296  4.083
48 48    CA  524  7.00    59.3 5.002  9.794
```

Fuel is a data.frame.

```
> is.data.frame(fuel)
[1] TRUE
```

You can refer to a variable in a data frame as fuel$Tax, etc. It returns one column of fuel.

```
> fuel$Tax
 [1]  9.00  9.00  9.00  7.50  8.00 10.00  8.00  8.00  8.00  7.00  8.00  7.50
[13]  7.00  7.00  7.00  7.00  7.00  7.00  7.00  8.50  7.00  8.00  9.00  9.00
[25]  8.50  9.00  8.00  7.50  8.00  9.00  7.00  7.00  8.00  7.50  8.00  6.58
[37]  5.00  7.00  8.50  7.00  7.00  7.00  7.00  7.00  6.00  9.00  7.00  7.00
```

length() and dim() tell you how big things are. There are 48 states and seven variables.

```
> length(fuel$Tax)
[1] 48
> dim(fuel)
[1] 48  7
```

To get a summary of a variable, type summary(variable)

```
> summary(fuel$Tax)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.000   7.000   7.500   7.668   8.125  10.000
```

R has very good graphics. You can make a boxplot with

```
boxplot(fuel$Fuel)
```

or dress it up with

```
boxplot(fuel$Fuel,ylab="gallons per person",main="Figure 1:
Motor Fuel Consumption")
```

To learn about a command, type help(command)

```
help(boxplot)

help(plot)

help(t.test)

help(lm)
```

Optional Trick

It can get tiresome typing fuel$Tax, fuel$Licenses, etc. If you type attach(data.frame) then you don't have to mention the data frame. Type detach(data.frame) when you are done.

```
> summary(fuel$Tax)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.000   7.000   7.500   7.668   8.125  10.000 
> summary(Tax)
Error in summary(Tax) : Object "Tax" not found
> attach(fuel)
> summary(Tax)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.000   7.000   7.500   7.668   8.125  10.000 
> summary(License)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  45.10   52.98   56.45   57.03   59.52   72.40 
> detach(fuel)
```

HELP

R contains several kinds of help. Use help(keyword) to get documentation about keyword.

```
> help(boxplot)
```

Use help("key") to find the keywords that contain "key". The quotes are needed.

```
> apropos("box")
[1] "box"                "boxplot"            "boxplot.default"
     "boxplot.stats"
```

Use help.search("keyword") to search the web for R functions that you can download related to keyword. Quotes are needed.

```
> help.search("box")

> help.search("fullmatch")
```

At http://cran.r-project.org/ there is free documentation, some of which is useful, but perhaps not for first-time users. To begin, books are better.

Some R

A variable, "change" in a data.frame bloodpressure.

```
> bloodpressure$change
 [1]  -9  -4 -21  -3 -20 -31 -17 -26 -26 -10 -23 -33 -19 -19 -23
```

It doesn't know what "change" is.

```
> change
Error: Object "change" not found
```

Try attaching the data.frame

```
> attach(bloodpressure)
```

Now it knows what "change" is.

```
> change
 [1]  -9  -4 -21  -3 -20 -31 -17 -26 -26 -10 -23 -33 -19 -19 -23

> mean(change)
[1] -18.93333
> sd(change)
[1] 9.027471
> summary(change)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -33.00  -24.50  -20.00  -18.93  -13.50   -3.00 
> stem(change)

  The decimal point is 1 digit(s) to the right of the |

  -3 | 31
  -2 | 663310
  -1 | 9970
  -0 | 943

> hist(change)
> boxplot(change)
> boxplot(change,main="Change in Blood Pressure After
Captopril",ylab="Change mmHg")
> boxplot(change,main="Change in Blood Pressure After
Captopril",ylab="Change mmHg",ylim=c(-40,40))
> abline(0,0,lty=2)
> t.test(change)

        One Sample t-test

data:  change 
t = -8.1228, df = 14, p-value = 1.146e-06
alternative hypothesis: true mean is not equal to 0 
95 percent confidence interval:
 -23.93258 -13.93409 
sample estimates:
mean of x 
-18.93333 
```

Are the Data Normal?

```
> attach(bloodpressure)
> change
 [1]  -9  -4 -21  -3 -20 -31 -17 -26 -26 -10 -23 -33 -19 -19 -23
> par(mfrow=c(1,2))
> boxplot(change)
> qqnorm(change)
```

A straight line in a Normal quantile plot is consistent with a Normal distribution.

You can also do a Shapiro-Wilk test. A small p-value suggests the data are not Normal.

```
> shapiro.test(change)

        Shapiro-Wilk normality test

data:  change
W = 0.9472, p-value = 0.4821
```

The steps below show what the qqnorm() function is plotting

```
> round(ppoints(change),3)
 [1] 0.033 0.100 0.167 0.233 0.300 0.367 0.433 0.500 0.567 0.633
[11] 0.700 0.767 0.833 0.900 0.967
```

The plotting positions in the normal plot:

```
> round(qnorm(ppoints(change)),3)
 [1] -1.834 -1.282 -0.967 -0.728 -0.524 -0.341 -0.168  0.000  0.168
[10]  0.341  0.524  0.728  0.967  1.282  1.834
```

qqnorm(change) is short for

```
> plot(qnorm(ppoints(change)),sort(change))
```

Here are Normal quantile plots of several Normal and non-Normal distributions.
Can you tell from the plot which are Normal?

```
> qqnorm(rnorm(10))
> qqnorm(rnorm(100))
> qqnorm(rnorm(1000))
> qqnorm(rcauchy(100))
> qqnorm(rlogis(100))
> qqnorm(rexp(100))
```

Regression in R

Script is my commentary to you. **Bold Courier is what I type in R.** Regular Courier is what R answered.

```
> ls()
[1] "fuel"
```

To display an object, type its name

```
> fuel
   ID state Fuel   Tax License   Inc   Road
1   1    ME  541  9.00    52.5 3.571  1.976
2   2    NH  524  9.00    57.2 4.092  1.250
3   3    VT  561  9.00    58.0 3.865  1.586
                     .
                     .
                     .
46 46    WN  510  9.00    57.1 4.476  3.942
47 47    OR  610  7.00    62.3 4.296  4.083
48 48    CA  524  7.00    59.3 5.002  9.794
```

To do regression, use lm. lm stands for linear model.

To fit Fuel = $\alpha + \beta$ Tax + ε, type

```
> lm(Fuel~Tax)

Call:
lm(formula = Fuel ~ Tax)

Coefficients:
(Intercept)          Tax
     984.01       -53.11
```

To fit Fuel = $\beta_0 + \beta_1$ Tax + β_2 License + ε, type

```
> lm(Fuel~Tax+License)

Call:
lm(formula = Fuel ~ Tax + License)

Coefficients:
(Intercept)          Tax      License
     108.97       -32.08        12.51
```

To see more output, type

```
> summary(lm(Fuel~Tax))

Call:
lm(formula = Fuel ~ Tax)

Residuals:
     Min       1Q   Median       3Q      Max 
-215.157  -72.269    6.744   41.284  355.736 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   984.01     119.62   8.226 1.38e-10 ***
Tax           -53.11      15.48  -3.430  0.00128 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' 
' 1 

Residual standard error: 100.9 on 46 degrees of freedom
Multiple R-Squared: 0.2037,     Adjusted R-squared: 0.1863 
F-statistic: 11.76 on 1 and 46 DF,  p-value: 0.001285
```

You can save the regression in an object and then refer to it:

```
> reg1<-lm(Fuel~Tax+License)
```

Now the workspace has a new object, namely reg1:

```
> ls()
[1] "fuel" "reg1"
```

To see reg1, type its name:

```
> reg1

Call:
lm(formula = Fuel ~ Tax + License)

Coefficients:
(Intercept)          Tax      License  
     108.97       -32.08        12.51
```

To get residuals, type

```
> reg1$residuals
```

This works only because I defined reg1 above. To boxplot residuals, type:

```
>boxplot(reg1$residuals)
```

To plot residuals against predicted values, type

```
> plot(reg1$fitted.values,reg1$residuals)
```

To do a normal plot of residuals, type

```
> qqnorm(reg1$residuals)
```

To get deleted or jackknife residuals, type

```
> rstudent(reg1)
```

To get leverages or hats, type

```
>hatvalues(reg1)
```

To get dffits

```
> dffits(reg1)
```

To get Cook's distance

```
> cooks.distance(reg1)
```

Clean up after yourself. To remove reg1, type rm(reg1)

```
> ls()
[1] "fuel" "reg1"
> rm(reg1)
> ls()
[1] "fuel"
```

Predictions

Fit a linear model and save it.

```
> mod<-lm(Fuel~Tax)
```

A confidence interval for the line at Tax = 8.5

```
>  predict(mod,data.frame(Tax=8.5),interval="confidence")
          fit      lwr      upr
[1,] 532.6041 493.4677 571.7405
```

A prediction interval for a new observation at Tax = 8.5

```
> predict(mod,data.frame(Tax=8.5),interval="prediction")
          fit      lwr      upr
[1,] 532.6041 325.7185 739.4897
```

Same point estimate, 532.6 gallons, but a very different interval, because the prediction interval has to allow for a new error for the new observation.

Multiple Regression Anova in R

The standard summary output from a linear model in R contains the key elements of the anova table, which are underlined.

```
> summary(lm(Fuel~Tax+License))
Call:
lm(formula = Fuel ~ Tax + License)

Residuals:
     Min       1Q   Median       3Q      Max 
-123.177  -60.172   -2.908   45.032  242.558 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  108.971    171.786   0.634   0.5291    
Tax          -32.075     12.197  -2.630   0.0117 *  
License       12.515      2.091   5.986 3.27e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' 
' 1 

Residual standard error: 76.13 on 45 degrees of freedom
Multiple R-Squared: 0.5567,     Adjusted R-squared: 0.537 
F-statistic: 28.25 on 2 and 45 DF,  p-value: 1.125e-08
```

More explicitly, the model `lm(Fuel~1)` fits just the constant term, and the F test compares that model (with just the constant term) to the model with all the variables (here Tax & License).

```
> anova(lm(Fuel~1),lm(Fuel~Tax+License))
Analysis of Variance Table

Model 1: Fuel ~ 1
Model 2: Fuel ~ Tax + License
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     47 588366                                  
2     45 260834  2    327532 28.253 1.125e-08 ***
```

Most regression programs present an explicit anova table, similar to that above, rather than just the F-test.

Partial Correlation Example

Here are the first two lines of data from a simulated data set. We are interested in the relationship between y and x2, taking account of x1.

```
> partialcorEG[1:2,]
           y         x1         x2
1 -3.8185777 -0.8356356 -1.0121903
2  0.3219982  0.1491024  0.0853746
```

Plot the data. Always plot the data.

```
> pairs(partialcorEG)
```

Notice that y and x2 have a positive correlation.

```
> cor(partialcorEG)
           y        x1        x2
y  1.0000000 0.9899676 0.9535053
x1 0.9899676 1.0000000 0.9725382
x2 0.9535053 0.9725382 1.0000000
```

The partial correlation is the correlation between the residuals. Notice that y and x2 have a negative partial correlation adjusting for x1.

```
> cor(lm(y~x1)$residual,lm(x2~x1)$residual)
[1] -0.2820687
```

Notice that the multiple regression coefficient has the same sign as the partial correlation.

```
> summary(lm(y~x1+x2))
Call:
lm(formula = y ~ x1 + x2)
Residuals:
     Min       1Q   Median       3Q      Max
-1.13326 -0.27423 -0.02018  0.32216  1.07808
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.007177   0.048662   0.147  0.88305
x1           4.768486   0.243833  19.556  < 2e-16 ***
x2          -0.720948   0.248978  -2.896  0.00468 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.4866 on 97 degrees of freedom
Multiple R-Squared: 0.9816,     Adjusted R-squared: 0.9812
F-statistic:  2591 on 2 and 97 DF,  p-value: < 2.2e-16
```

Added Variable Plots

You have fit a model, mod1 say, $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$ where ε are iid $N(0,\sigma^2)$ and now want to ask about adding a new variable, x_{k+1}, to this model.

It can’t hurt to plot Y against x_{k+1}. However, that plot does not tell you what x_{k+1} will do in the model above. It could happen that Y increases with x_{k+1} but $\beta_{k+1} < 0$.

The added variable plot uses the idea of regression by stages. In regression by stages, you estimate β_{k+1} by regressing the residuals from mod1 on the residuals of x_{k+1} when regressed on x_1 , … , x_k. The added variable plot is simply the plot of these two sets of residuals, residuals of Y versus residuals of x_{k+1}. The slope in that plot estimates β_{k+1}. So the added variable plot lets you see what happens when x_{k+1} is added to mod1.

You can calculate the two set of residuals and plot them. That works fine. Or you can use **addedvarplot** in the course workspace.

```
> attach(fuel)
> head(fuel)
  ID state Fuel  Tax License   Inc  Road
1  1    ME  541  9.0    52.5 3.571 1.976
2  2    NH  524  9.0    57.2 4.092 1.250
3  3    VT  561  9.0    58.0 3.865 1.586
4  4    MA  414  7.5    52.9 4.870 2.351
5  5    RI  410  8.0    54.4 4.399 0.431
6  6    CN  457 10.0    57.1 5.342 1.333
> mod1<-lm(Fuel~Tax+License)
> addedvarplot(mod1,Inc)
```

The same plot is produced directly by:

```
> plot(lm(Inc~Tax+License)$resid,mod1$resid)
```

ADDED VARIABLE PLOTS IN THE car Package

```
> attach(fuel)
> pairs(cbind(Fuel,Tax,License))
> library(car)
> help(avPlots)
> avPlots(lm(Fuel~Tax+License))
> avPlots(lm(Fuel~Tax+License+Inc),term=~Inc)
> avPlots(lm(Fuel~Tax+License+Inc))
> summary(lm(Fuel~Tax+License+Inc))
```

car stands for “Companion to Applied Regression”. The book, *An R Companion to Applied Regression*” by John Fox and Sanford Weisberg discusses regression using this package.

Vocabulary Homework

```
> vocabulary
    Age Vocab
1  0.67     0
2  0.83     1
3  1.00     3
4  1.25    19
5  1.50    22
6  1.75   118
7  2.00   272
8  2.50   446
9  3.00   896
10 3.50  1222
11 4.00  1540
12 4.50  1870
13 5.00  2072
14 5.50  2289
15 6.00  2562

> attach(vocabulary)
```

Fit linear model (a line) and store results in "mod".

```
> mod<-lm(Vocab~Age)
```

Summary output for mod.

```
> summary(mod)
Call:
lm(formula = Vocab ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max 
-249.67 -104.98   13.14   78.47  268.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -621.16      74.04  -8.389 1.32e-06 ***
Age           526.73      22.12  23.808 4.17e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1 

Residual standard error: 148 on 13 degrees of freedom
Multiple R-Squared: 0.9776,     Adjusted R-squared: 0.9759 
F-statistic: 566.8 on 1 and 13 DF,  p-value: 4.170e-12
```

Plot the data. Does a line look appropriate?

```
>  plot(Age,Vocab,ylim=c(-1000,3000))
> abline(mod)
```

Plot residuals vs predicteds. Is there a pattern?

```
> plot(mod$fitted.values,mod$residuals)
```

Boxplot residuals. Unusual points? Skewness?

```
> boxplot(mod$residuals)
```

Normal plot of residuals. Do the residuals look Normal? (Is it a line?)

```
> qqnorm(mod$residuals)
```

Test of the null hypothesis that the residuals are Normal.

```
> shapiro.test(mod$residuals)

        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.9801, p-value = 0.9703
```

General Linear Hypothesis

```
> help(anova.lm)

> attach(fuel)
> fuel[1:2,]
  ID state Fuel Tax License   Inc  Road
1  1    ME  541   9    52.5 3.571 1.976
2  2    NH  524   9    57.2 4.092 1.250
```

Fit the full model.

```
> mod<-lm(Fuel~Tax+License+Inc)
> anova(mod)      Optional step – for your education only.
Analysis of Variance Table

Response: Fuel
          Df Sum Sq Mean Sq F value    Pr(>F)    
Tax        1 119823  119823  27.560 4.209e-06 ***
License    1 207709  207709  47.774 1.539e-08 ***
Inc        1  69532   69532  15.992 0.0002397 ***
Residuals 44 191302    4348                      
---
```

Fit the reduced model.

```
> mod2<-lm(Fuel~Tax)
> anova(mod2)       Optional step – for your education only.
Analysis of Variance Table

Response: Fuel
          Df Sum Sq Mean Sq F value   Pr(>F)   
Tax        1 119823  119823  11.764 0.001285 **
Residuals 46 468543   10186                    
```

Compare the models

```
> anova(mod2,mod)
Analysis of Variance Table

Model 1: Fuel ~ Tax
Model 2: Fuel ~ Tax + License + Inc
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     46 468543                                  
2     44 191302  2    277241 31.883 2.763e-09 ***
```

Notice the residual sum of squares and degrees of freedom in the three anova tables!

Polynomial Regression

```
> attach(cars)
```

Quadratic in size $y = \beta_0 + \beta_1 x + \beta_2 x^2$

```
> lm(mpg~size+I(size^2))

Call:
lm(formula = mpg ~ size + I(size^2))

Coefficients:
(Intercept)         size    I(size^2)
 39.3848313   -0.1485722    0.0002286
```

Centered quadratic in size $y = \beta_0 + \beta_1 x + \beta_2 \{x\text{-mean}(x)\}^2$

```
> lm(mpg~size+I((size-mean(size))^2))

Call:
lm(formula = mpg ~ size + I((size - mean(size))^2))

Coefficients:
             (Intercept)                       size  I((size -
mean(size))^2)
              28.8129567                 -0.0502460
0.0002286
```

Orthogonal Polynomial Quadratic in size

```
> lm(mpg~poly(size,2))

Call:
lm(formula = mpg ~ poly(size, 2))

Coefficients:
   (Intercept)  poly(size, 2)1  poly(size, 2)2
         20.74          -24.67           12.33
```

To gain understanding:

- *do all there regressions*
- *look at t-test for* β_2
- *type poly(size,2)*
- *plot poly(size,2)[,1] and poly(size,2)[,2] against size*

Centered Polynomial with Interaction

```
> fuel[1:2,]
  ID state Fuel Tax License   Inc  Road
1  1    ME  541   9    52.5 3.571 1.976
2  2    NH  524   9    57.2 4.092 1.250

> attach(fuel)
```

Construct the squared and crossproduct terms. Alternatives: use "" or ":" in model formula.*

```
> TaxC<-Tax-mean(Tax)
> LicC<-License-mean(License)
> TaxC2<-TaxC*TaxC
> LicC2<-LicC*LicC
> TaxLicC<-TaxC*LicC

> modfull<-lm(Fuel~Tax+License+TaxC2+LicC2+TaxLicC)
> summary(modfull)

Call:
lm(formula = Fuel ~ Tax + License + TaxC2 + LicC2 + TaxLicC)
Residuals:
       Min          1Q     Median          3Q         Max 
-121.52425  -51.08809   -0.01205   46.27051  223.28655 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 169.7242   179.6332   0.945   0.3501    
Tax         -32.4465    12.2906  -2.640   0.0116 *  
License      11.2776     2.3087   4.885 1.55e-05 ***
TaxC2         1.3171     8.6638   0.152   0.8799    
LicC2         0.2575     0.2868   0.898   0.3743    
TaxLicC      -2.5096     2.7343  -0.918   0.3640    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
Residual standard error: 76.42 on 42 degrees of freedom
Multiple R-Squared: 0.5831,     Adjusted R-squared: 0.5335 
F-statistic: 11.75 on 5 and 42 DF,  p-value: 3.865e-07 
```

Test whether the three squared and interaction terms are needed:

```
> modred<-lm(Fuel~Tax+License)
> anova(modred,modfull)
Analysis of Variance Table

Model 1: Fuel ~ Tax + License
Model 2: Fuel ~ Tax + License + TaxC2 + LicC2 + TaxLicC
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 260834                           
2     42 245279  3     15555 0.8879 0.4552
```

Understanding Linear Models with Interactions or Polynomials

NIDA data (DC*MADS) on birth weight of babies in DC and attributes of mom.

```
> DCBabyCig[1:2,]
    Age Married CIGS   BW
1    17       0    0 2385
2    23       1    0 4175
```

Age x Cigarettes interaction

```
> AC<-Age*CIGS
```

Model with interaction

```
> lm(BW~Age+CIGS+AC)

Call:
lm(formula = BW ~ Age + CIGS + AC)

Coefficients:
(Intercept)          Age         CIGS           AC
    2714.81        13.99       562.66       -28.04
```

How do you understand a model with interactions?

Let's create a new data.frame with 6 moms in it. Three moms are 18, three are 35. Some smoke 0, 1 or 2 packs.

```
> new[,1]<-c(18,35,18,35,18,35)
> new[,2]<-c(0,0,1,1,2,2)
> new[,3]<-new[,1]*new[,2]
> colnames(new)<-c("Age","CIGS","AC")
> new<-data.frame(new)

> new
  Age CIGS AC
1  18    0  0
2  35    1 35
3  18    0  0
4  35    1 35
5  18    0  0
6  35    1 35
```

Now, for these six moms, let's predict birth weight of junior. It is usually easier to talk about people than about coefficients, and that is what this table does: it talks about 6 moms.

```
> round(cbind(new,predict(lm(BW~Age+CIGS+AC),new,interval="confidence")))
  Age CIGS AC  fit  lwr  upr
1  18    0  0 2967 2865 3068
2  35    0  0 3204 3073 3336
3  18    1 18 3024 2719 3330
4  35    1 35 2786 2558 3013
5  18    2 36 3082 2474 3691
6  35    2 70 2367 1919 2814
```

Interpretation of an Interaction

```
> DCBabyCig[1:6,]
   Age Married CIGS   BW
1   17       0    0 2385
2   23       1    0 4175
3   25       0    0 3655
4   18       0    0 1855
5   20       0    0 3600
6   24       0    0 2820
```

Age = mother’s age
Married, 1=yes, 0=no
CIGS = packs per day, 0, 1, 2.
BW = birth weight in grams

```
> dim(DCBabyCig)
[1] 449   4
```

```
> mod<-lm(BW~Age+Married+CIGS+I(Married*CIGS))
> summary(mod)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2973.1866   152.5467  19.490  < 2e-16 ***
Age                  0.1699     6.5387   0.026  0.97928
Married            274.0662    89.2913   3.069  0.00228 **
CIGS               -88.4957    81.7163  -1.083  0.27941
I(Married * CIGS) -415.1501   160.4540  -2.587  0.00999 **

Residual standard error: 687.8 on 444 degrees of freedom
Multiple R-squared: 0.05337,    Adjusted R-squared: 0.04484
F-statistic: 6.258 on 4 and 444 DF,  p-value: 6.618e-05
```

Plot the data

```
> boxplot(BW~Married:CIGS)
```

A 25 year old mom in all combinations of Married and CIGS.

```
> DCBabyCigInter
  Age Married CIGS
1  25       0    0
2  25       0    1
3  25       0    2
4  25       1    0
5  25       1    1
6  25       1    2
```

Predicted birth weights for this mom, with confidence intervals.

```
> predict(mod,DCBabyCigInter,interval="conf")
       fit      lwr      upr
1 2977.434 2890.180 3064.688
2 2888.938 2738.900 3038.977
3 2800.443 2502.124 3098.761
4 3251.500 3114.163 3388.838
5 2747.854 2476.364 3019.345
6 2244.209 1719.423 2768.995
```

Let’s clean it up, converting to pounds (2.2 pounds per kilogram), and add the predictors:

```
> pr<-predict(mod,DCBabyCigInter,interval="conf")

> round(cbind(DCBabyCigInter,2.2*pr/1000),1)
  Age Married CIGS fit lwr upr
1  25       0    0 6.6 6.4 6.7
2  25       0    1 6.4 6.0 6.7
3  25       0    2 6.2 5.5 6.8
4  25       1    0 7.2 6.9 7.5
5  25       1    1 6.0 5.4 6.6
6  25       1    2 4.9 3.8 6.1
```

Using Restricted Cubic Splines (aka Natural Splines)

```
> library(Hmisc)
> head(cars)
       car  size  mpg group
1  ToyotaC  71.1 33.9     1
2   HondaC  75.7 30.4     1

> x<-rcspline.eval(size,nk=3) #Three knots, one additional variable
> plot(size,x) #What does the new variable look like?
> plot(size,mpg)
> m<-lm(mpg~size+x) #Add the new variable to the model.
> points(size,m$fit,pch=16,col="red") #What does the fit look like?
> summary(m)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.81737    1.91493  20.271  < 2e-16 ***
size        -0.11667    0.01419  -8.225 7.86e-09 ***
x            0.14618    0.02642   5.533 7.29e-06 ***
Residual standard error: 2.346 on 27 degrees of freedom
Multiple R-squared:  0.8395,    Adjusted R-squared:  0.8276
F-statistic: 70.62 on 2 and 27 DF,  p-value: 1.877e-11

> x<-rcspline.eval(size,nk=5) #Five knots, three additional variables
> m<-lm(mpg~size+x)
> summary(m)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.28095    3.61684  12.796 1.79e-12 ***
size        -0.19689    0.03655  -5.387 1.37e-05 ***
x1           2.17972    0.97629   2.233   0.0348 *
x2          -3.52907    1.75277  -2.013   0.0550 .
x3           1.45720    0.90924   1.603   0.1216
Residual standard error: 2.199 on 25 degrees of freedom
Multiple R-squared:  0.8694,    Adjusted R-squared:  0.8485
F-statistic: 41.61 on 4 and 25 DF,  p-value: 1.054e-10

> points(size,m$fit,pch=16,col="purple")
```

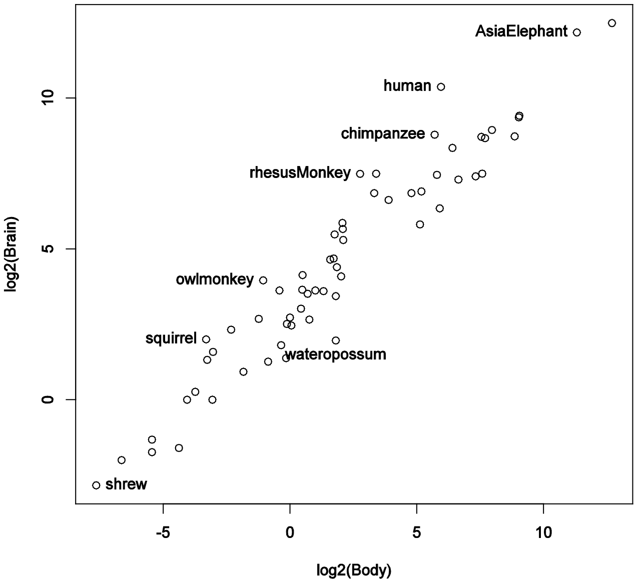
Reference: Harrell, F. (2015) *Regression Modeling Strategies*, New York: Springer, section 2.4.5.

Comment: There are many types of splines. Natural splines are linear beyond the final knots, so they wiggle less at the ends.

Dummy Variable in Brains Data

First two rows of "brains" data.

```
> brains[1:2,]
   Body  Brain    Animal Primate Human
1 3.385 44.500  articfox       0     0
2 0.480 15.499 owlmonkey       1     0

> attach(brains)

> plot(log2(Body),log2(Brain))
> identify(log2(Body),log2(Brain),labels=Animal)
```



```
> mod<-lm(log2(Brain)~log2(Body)+Primate)
> mod

Call:
lm(formula = log2(Brain) ~ log2(Body) + Primate)

Coefficients:
(Intercept)   log2(Body)      Primate
     2.8394       0.7402       1.6280
```

log2(Brain) ~ log2(Body) + Primate

is $2^{Brain} = 2^{\{\alpha + \beta log2(Body) + \gamma Primate + \varepsilon\}} = (2^{\alpha})(Body^{\beta})(2^{\gamma Primate})(2^{\varepsilon})$

$2^{1.628Primate}$ *= 3.1 for a primate, = 1 for a nonprimate*

Computing the Diagnostics in the Rat Data

```
> ratdata[1:3,]
  BodyWgt LiverWgt Dose Percent Rat3
1     176      6.5 0.88    0.42    0
2     176      9.5 0.88    0.25    0
3     190      9.0 1.00    0.56    1

> attach(ratdata)
> mod<-lm(Percent~BodyWgt+LiverWgt+Dose)
```

Standardized residuals (first 5)

```
> rstandard(mod)[1:5]
        1         2         3         4         5
 1.766047 -1.273040  0.807154 -1.377232 -1.123099
```

Deleted or jackknife or "studentized" residuals (first 5)

```
> rstudent(mod)[1:5]
         1          2          3          4          5
 1.9170719 -1.3022313  0.7972915 -1.4235804 -1.1337306
```

dffits (first 5)

```
> dffits(mod)[1:5]
         1          2          3          4          5
 0.8920451 -0.6087606  1.9047699 -0.4943610 -0.9094531
```

Cook's distance (first 5)

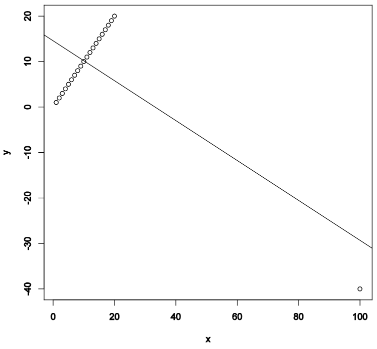
```
> cooks.distance(mod)[1:5]
         1          2          3          4          5
0.16882682 0.08854024 0.92961596 0.05718456 0.20291617
```

Leverages or 'hats' (first 5)

```
> hatvalues(mod)[1:5]
        1         2         3         4         5
0.1779827 0.1793410 0.8509146 0.1076158 0.3915382

> dfbeta(mod)[1:3,]
   (Intercept)       BodyWgt     LiverWgt       Dose
1 -0.006874698  0.0023134055 -0.011171761 -0.3419002
2  0.027118946 -0.0007619302 -0.008108905  0.1869729
3 -0.045505614 -0.0134632770  0.005308722  2.6932347

> dfbetas(mod)[1:3,]
  (Intercept)     BodyWgt   LiverWgt       Dose
1 -0.03835128  0.31491627 -0.7043633 -0.2437488
2  0.14256373 -0.09773917 -0.4817784  0.1256122
3 -0.23100202 -1.66770314  0.3045718  1.7471972
```

High Leverage Example

```
> t(highlev)    t() is transpose – make rows into columns and columns into rows – compact printing
  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  21
x 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 100
y 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -40

> mod<-lm(y~x)
> summary(mod)
Residuals:
     Min       1Q   Median       3Q      Max
-13.1343  -7.3790  -0.1849   7.0092  14.2034
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.57312    2.41264   6.040 8.24e-06 ***
x           -0.43882    0.09746  -4.503 0.000244 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.875 on 19 degrees of freedom
Multiple R-Squared: 0.5162,     Adjusted R-squared: 0.4908
F-statistic: 20.27 on 1 and 19 DF,  p-value: 0.0002437

> plot(x,y)
> abline(mod)    Puts the fitted line on the plot –– What a dumb model!
```



The bad guy, #21, doesn't have the biggest residual!

```
> mod$residual[21]
      21
-10.69070
```

Residuals:

```
     Min       1Q   Median       3Q      Max
-13.1343  -7.3790  -0.1849   7.0092  14.2034
```

But our diagnostics find him!

```
> rstudent(mod)[21]
       21
-125137800
> hatvalues(mod)[21]
      21
0.9236378
> dffits(mod)[21]
       21
-435211362
```

Outlier Testing

Use the Bonferroni inequality with the deleted/jackknife/"studentized" residuals.

Example uses random data – should not contain true outliers

```
> x<-rnorm(1000)
> y<-rnorm(1000)
> plot(x,y)
> summary(lm(y~x))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.005469   0.031685  -0.173    0.863
x           -0.044202   0.031877  -1.387    0.166

Residual standard error: 1.002 on 998 degrees of freedom
Multiple R-Squared: 0.001923,   Adjusted R-squared: 0.0009228
F-statistic: 1.923 on 1 and 998 DF,  p-value: 0.1659
```

Look at the deleted residuals (from rstudent)

```
> summary(rstudent(lm(y~x)))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-3.189e+00 -6.596e-01  2.186e-02 -7.077e-06  6.517e-01  3.457e+00
```

The big residual in absolute value is 3.457. Is that big for the biggest of 1000 residuals?

*The pt(value,df) command looks up value in the t-table with df degrees of freedom, returning Pr(t<value). You need the other tail, Pr(t>value), and you need to double it for a 2-tailed test. The degrees of freedom are **one less than the degrees of freedom in the error for the regression**, here 997.*

```
> 2*(1-pt(3.457,997))
[1] 0.0005692793
```

This is uncorrected p-value. Multiply by the number of tests, here 1000, to correct for multiple testing. (It's an inequality, so it can give a value bigger than 1.)

```
> 1000* 2*(1-pt(3.457,997))
[1] 0.5692793
```

As this is bigger than 0.05, the null hypothesis of no outliers is not rejected – it is plausible there are no outliers present.

IS WYOMING AN OUTLIER?

```
> attach(fuel)
> dim(fuel)
[1] 48  7
> mod<-lm(Fuel~Tax+License)
> which.max(abs(rstudent(mod)))
40
40
> fuel[40,]
   ID state Fuel Tax License   Inc  Road
40 40    WY  968   7    67.2 4.345 3.905
> rstudent(mod)[40]
      40
3.816379
> wy<-rep(0,48)
> wy[40]<-1
> wy
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0

> summary(lm(Fuel~Tax+License+wy))

Call:
lm(formula = Fuel ~ Tax + License + wy)

Residuals:
     Min       1Q   Median       3Q      Max 
-122.786  -55.294    1.728   46.621  154.557 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  198.651    152.405   1.303  0.19920    
Tax          -30.933     10.696  -2.892  0.00593 ** 
License       10.691      1.894   5.645 1.12e-06 ***
wy           267.433     70.075   3.816  0.00042 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 66.74 on 44 degrees of freedom
Multiple R-Squared: 0.6669,     Adjusted R-squared: 0.6442 
F-statistic: 29.37 on 3 and 44 DF,  p-value: 1.391e-10 

> 0.05/48
[1] 0.001041667
> 0.00042<=0.001041667
[1] TRUE
```

Testing Whether a Transformation of Y is Needed

Tukey's One Degree of Freedom for Nonadditivity

Tukey (1949) proposed testing whether a transformation of y is needed in the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \varepsilon \quad \varepsilon \sim \text{iid } N(0,\sigma^2)$$

by adding a scaled centered version of $\hat{y}^2$ to the model, specifically $\frac{(\hat{y}-\bar{y})^2}{2\bar{y}}$; see Atkinson (1985, p. 157). The function `tukey1df(mod)` in the class workspace does this, but you could easily do it yourself.

```
> mod<-lm(BW~Age+Married+CIGS)
> summary(mod)

Call:
lm(formula = BW ~ Age + Married + CIGS)

Residuals:
     Min       1Q   Median       3Q      Max 
-2408.30  -358.49    99.69   453.34  1952.79 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2936.618    152.859  19.211  < 2e-16 ***
Age            2.557      6.515   0.392  0.69488    
Married      200.615     85.198   2.355  0.01897 *  
CIGS        -196.644     70.665  -2.783  0.00562 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 692.2 on 445 degrees of freedom
Multiple R-squared: 0.0391,     Adjusted R-squared: 0.03262 
F-statistic: 6.036 on 3 and 445 DF,  p-value: 0.0004910 

> boxplot(mod$resid)
> qqnorm(mod$resid)
> shapiro.test(mod$resid)

        Shapiro-Wilk normality test

data:  mod$resid 
W = 0.9553, p-value = 1.996e-10

> plot(mod$fit,mod$resid)
> lines(lowess(mod$fit,mod$resid))
```

To do the test, add the transformed variable to the model and look at its t-statistic.

```
> summary(lm(BW~Age+Married+CIGS+tukey1df(mod)))

Call:
lm(formula = BW ~ Age + Married + CIGS + tukey1df(mod))

Residuals:
    Min      1Q  Median      3Q     Max 
-2301.3  -334.5   107.7   420.8  1981.2 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2962.751    151.775  19.521  < 2e-16 ***
Age              0.508      6.494   0.078  0.93769    
Married        104.750     90.366   1.159  0.24701    
CIGS          -489.699    120.693  -4.057 5.86e-05 ***
tukey1df(mod)   31.507     10.567   2.982  0.00302 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 686.2 on 444 degrees of freedom
Multiple R-squared: 0.05796,    Adjusted R-squared: 0.04947 
F-statistic:  6.83 on 4 and 444 DF,  p-value: 2.428e-05
```

Box – Cox Method

An alternative approach is due to Box and Cox (1964).

```
> library(MASS)
> help(boxcox)
> boxcox(mod)
```

Andrews, D. F. (1971) A note on the selection of data transformations. Biometrika, 58, 249-254.
Atkinson, A. C. (1985) Plots, Transformations and Regression. NY: Oxford.
Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). Journal of the Royal Statistical Society B, 26, 211–252.
Tukey, J. W. (1949) One degree of freedom for nonadditivity. *Biometrics*, 5, 232-242.
Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

Transformations in the car package

```
> attach(cars)
> plot(size,mpg)
> library(car)
```

This is looking for a transformation of size that would improve the fit. Because y=mpg is not transformed, only x=size, the procedure can compare residual sums of squares of y (RSS) for different transformations of x. Here, lambda is the power transformation, what we called p in class.

```
> invTranPlot(mpg~size)
     lambda      RSS
1 -1.262328 123.4974
2 -1.000000 126.3604
3  0.000000 192.1809
4  1.000000 317.0362
```

It likes the -1.26 power of x as a transformation, but -1 and -1.5 are in the confidence interval.

```
> invTranEstimate(size,mpg)
$lambda
[1] -1.262328
$lowerCI
[1] -1.737292
$upperCI
[1] -0.8237128
```

This is trying to transform y=mpg, not x=size. It uses the Box-Cox likelihood method. It likes y^(-0.99) or approximately 1/y.

```
> summary(powerTransform(lm(mpg~size)))
bcPower Transformation to Normality
   Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Y1   -0.9878   0.5553          -2.0763           0.1007
Likelihood ratio tests about transformation parameters
                            LRT df         pval
LR test, lambda = (0)  3.194518  1 0.0738855550
LR test, lambda = (1) 12.478121  1 0.0004117462
```

A graphical version.

```
> boxCox(lm(mpg~size))
```

The car package has an alternative to Tukey’s test for a transformation.

Atkinson’ method

```
> v<-boxCoxVariable(mpg)
> summary(lm(mpg~size+v))
Call:
lm(formula = mpg ~ size + v)
Residuals:
    Min      1Q  Median      3Q     Max 
-4.3645 -1.3910  0.0462  1.0028  6.4195 
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 76.651791   9.132471   8.393 5.27e-09 ***
size        -0.033455   0.004324  -7.737 2.55e-08 ***
v            2.519663   0.486271   5.182 1.87e-05 ***
```

Andrews/Tukey method

```
> summary(lm(mpg~size+tukey1df(md)))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.812957   1.008748   28.56  < 2e-16 ***
size         -0.050246   0.004508  -11.15 1.32e-11 ***
tukey1df(md)  5.577240   1.117584    4.99 3.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 2.471 on 27 degrees of freedom
Multiple R-squared:  0.8218,    Adjusted R-squared:  0.8087 
F-statistic: 62.28 on 2 and 27 DF,  p-value: 7.686e-11
```

References:

For Tukey’s method and related methods

Andrews, D. F. (1971) A note on the selection of data transformations. *Biometrika* 58, 249-254.

For the method in the car package:

Atkinson, A. C. (1973) Testing transformations to Normality. *JRSS-B* 473-479.

Calculating C_P for the Cathedral Data

```
> attach(cathedral)
> mod<-lm(length~height+gothic+GH)
> summary(mod)
Call:
lm(formula = length ~ height + gothic + GH)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  241.833    336.471   0.719    0.480
height         3.138      4.506   0.696    0.494
gothic      -204.722    347.207  -0.590    0.562
GH             1.669      4.641   0.360    0.723
Residual standard error: 79.11 on 21 degrees of freedom
Multiple R-Squared: 0.5412,     Adjusted R-squared: 0.4757
F-statistic: 8.257 on 3 and 21 DF,  p-value: 0.0008072

> drop1(lm(length~height+gothic+GH),scale=79.11^2)
Single term deletions

Model:
length ~ height + gothic + GH

scale:  6258.392

       Df Sum of Sq    RSS     Cp
<none>              131413 3.9979
height  1      3035 134448 2.4829
gothic  1      2176 133589 2.3455
GH      1       810 132223 2.1273

> drop1(lm(length~height+gothic),scale=79.11^2)
Single term deletions
Model:
length ~ height + gothic

scale:  6258.392

       Df Sum of Sq    RSS      Cp
<none>              132223  2.1273
height  1    119103 251326 19.1582
gothic  1     37217 169440  6.0740
```

Variable Selection

Highway data. First two rows of 39 rows. More details in the Variable Selection section of this bulkpack.

```
> highway[1:2,]
  ID rate   len adt trks slim lwid shld  itg sigs acpt lane fai pa ma
1  1 4.58  4.99  69    8   55   12   10 1.20    0  4.6    8   1  0  0
2  2 2.86 16.11  73    8   60   12   10 1.43    0  4.4    4   1  0  0
```

Highway data has 39 rows, 15 columns, of which y=rate, and columsn 3 to 15 or 3:15 are predictors. Want to select predictors.

```
> dim(highway)
[1] 39 15

> attach(highway)
```

To use "leaps" for best subsets regression, need to get it from the library. To get documentation, type help!

```
> library(leaps)
> help(leaps)
```

Easiest if you put the x's in a separate variable. These are columns 3:15 of highway, including all the rows.

```
> x<-highway[,3:15]
```

First three rows of 39 rows of x. Notices that the first two columns of highway are gone.

```
> x[1:3,]
    len adt trks slim lwid shld  itg sigs acpt lane fai pa ma
1  4.99  69    8   55   12   10 1.20    0  4.6    8   1  0  0
2 16.11  73    8   60   12   10 1.43    0  4.4    4   1  0  0
3  9.75  49   10   60   12   10 1.54    0  4.7    4   1  0  0
```

There are 13 predictors, hence 2^{13} = 8,192 possible models formed by including each variable or not.

```
> dim(x)
[1] 39 13
> 2^13
[1] 8192
```

Look at the names of your predictors: len = length of segment, ..., slim = speed limit, ..., acpt = number of access points per mile, ...

```
> colnames(x)
 [1] "len"  "adt"  "trks" "slim" "lwid" "shld" "itg"  "sigs" "acpt"
"lane" "fai"  "pa"   "ma"
```

A quick and easy, but not very complete, answer is obtained from regsubsets. Here, it gives the best model with 1 variable, the best with 2 variables, etc. Look for the *'s. The best 3 variable model is len, slim, acpt.

```
> summary(regsubsets(x=x,y=rate))
1 subsets of each size up to 8  Selection Algorithm: exhaustive
         len adt trks slim lwid shld itg sigs acpt lane fai pa  ma
1  ( 1 ) " " " " " "  " "  " "  " "  " " " "  "*"  " "  " " " " " "
2  ( 1 ) "*" " " " "  " "  " "  " "  " " " "  "*"  " "  " " " " " "
3  ( 1 ) "*" " " " "  "*"  " "  " "  " " " "  "*"  " "  " " " " " "
4  ( 1 ) "*" " " " "  "*"  " "  " "  " " "*"  "*"  " "  " " " " " "
5  ( 1 ) "*" " " " "  "*"  " "  " "  " " "*"  "*"  " "  " " "*" " "
6  ( 1 ) "*" " " "*"  "*"  " "  " "  " " "*"  "*"  " "  " " "*" " "
7  ( 1 ) "*" " " "*"  "*"  " "  " "  " " "*"  "*"  " "  " " "*" "*"
8  ( 1 ) "*" " " "*"  "*"  " "  " "  "*" "*"  "*"  " "  " " "*" "*"
```

To get the two best models of each size, type:

```
> summary(regsubsets(x=x,y=rate,nbest=2))
```

Variable Selection, Continued

Leaps does "best subsets regression". Here, x contains the predictors of y=rate. If you don't tell it the variable names, it uses 1, 2, …

```
mod<-leaps(x,rate,names=colnames(x))
```

The output from leaps, here "mod", has four parts, "which", "label", "size" and "Cp".

```
> summary(mod)
      Length Class  Mode
which 1573   -none- logical
label   14   -none- character
size   121   -none- numeric
Cp     121   -none- numeric
```

You refer to "which" for "mod" as "mod$which", etc. Also mod$size, mod$Cp.

The part, mod$which says which variables are in each model. It reports back about 121 models, and all 13 variables.

```
> dim(mod$which)
[1] 121  13
```

Here are the first 3 rows or models in mod$which. The first model has only acpt, while the second has only slim.

```
> mod$which[1:3,]
    len   adt  trks  slim  lwid  shld   itg  sigs  acpt  lane   fai    pa    ma
1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
```

Here are the last 3 rows or models in mod$which. The last model has all 13 variables.

```
> mod$which[119:121,]
    len  adt trks slim lwid shld  itg  sigs acpt lane  fai    pa   ma
12 TRUE TRUE TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE FALSE TRUE
12 TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE  TRUE TRUE
13 TRUE TRUE TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE
```

Here are the sizes of the 121 models. A 1-variable model has size 2, for constant-plus-one-slope. A 2-variable model has size 3. The final model, #121, with all 13 variables has size 14.

```
> mod$size
  [1]  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  4  4  4  4
 [25]  4  4  4  4  4  4  5  5  5  5  5  5  5  5  5  5  6  6  6  6  6  6  6  6
 [49]  6  6  7  7  7  7  7  7  7  7  7  7  8  8  8  8  8  8  8  8  8  8  9  9
 [73]  9  9  9  9  9  9  9  9 10 10 10 10 10 10 10 10 10 10 11 11 11 11 11 11
 [97] 11 11 11 11 12 12 12 12 12 12 12 12 12 12 13 13 13 13 13 13 13 13 13 13
[121] 14
```

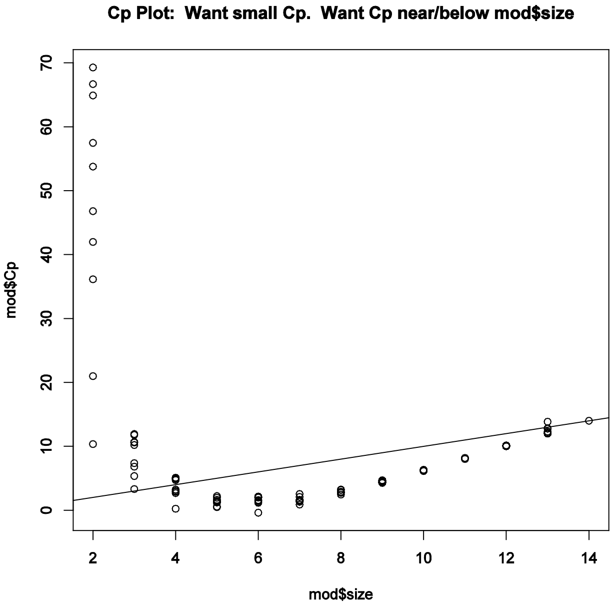
These are the C_p values for the 121 models.

```
> round(mod$Cp,2)
  [1] 10.36 20.98 36.13 41.97 46.79 53.77 57.48 64.90 66.67 69.28  3.31  5.35
 [13]  6.82  7.37 10.21 10.59 10.64 10.68 11.78 11.93  0.24  2.68  2.90  3.04
 [25]  3.24  4.72  4.88  4.99  5.05  5.10  0.49  0.56  1.21  1.26  1.33  1.39
 [37]  1.59  1.63  1.97  2.22 -0.38  1.17  1.34  1.51  1.60  2.00  2.00  2.02
 [49]  2.05  2.15  0.88  1.33  1.33  1.52  1.54  1.58  1.61  1.62  2.08  2.54
 [61]  2.45  2.73  2.74  2.82  2.82  2.87  2.88  3.18  3.19  3.23  4.27  4.33
 [73]  4.36  4.43  4.45  4.45  4.45  4.59  4.67  4.69  6.14  6.15  6.16  6.23
 [85]  6.24  6.26  6.26  6.30  6.32  6.32  8.02  8.07  8.12  8.12  8.14  8.14
 [97]  8.14  8.15  8.16  8.17 10.02 10.02 10.02 10.06 10.07 10.10 10.12 10.12
[109] 10.13 10.14 12.01 12.01 12.01 12.05 12.10 12.14 12.32 12.76 12.83 13.85
[121] 14.00
```

Variable Selection, Continued

This is the C_p plot.

```
> plot(mod$size,mod$Cp)
> abline(0,1)
```



Cp Plot: Want small Cp. Want Cp near/below mod$size

There is one pretty good 2 variable model (size=3), with C_p near the x=y line, and one very good 3 variable model (size=4), with C_p way below the line. The best model has 5 variables (size =6) but is only trivially better than the 3 variable model. R^2 is highest for the 14 variable model, but chances are it won't predict as well as the 3 variable model.

Let's put together the pieces.

```
> join<-cbind(mod$which,mod$Cp,mod$size)
```

Let's look at the 3 variable models (size=4). The best has C_p = 0.236 and variables len, slim, and acpt.

```
> join[mod$size==4,]
  len adt trks slim lwid shld itg sigs acpt lane fai pa ma
3   1   0    0    1    0    0   0    0    1    0   0  0  0 0.2356971 4
3   0   0    1    1    0    0   0    0    1    0   0  0  0 2.6805672 4
3   1   0    0    0    0    0   0    1    1    0   0  0  0 2.8975068 4
3   1   0    0    0    0    1   0    0    1    0   0  0  0 3.0404482 4
3   1   0    1    0    0    0   0    0    1    0   0  0  0 3.2366902 4
3   1   0    0    0    1    0   0    0    1    0   0  0  0 4.7193511 4
3   0   0    0    1    0    0   0    1    1    0   0  0  0 4.8847460 4
3   1   0    0    0    0    0   0    0    1    0   0  1  0 4.9933327 4
3   1   0    0    0    0    0   0    0    1    1   0  0  0 5.0489720 4
3   1   1    0    0    0    0   0    0    1    0   0  0  0 5.1013513 4
```

The full model has C_p = 14.

```
> join[mod$size==14,]
 len  adt trks slim lwid shld  itg sigs acpt lane  fai   pa   ma
   1    1    1    1    1    1    1    1    1    1    1    1    1   14   14
```

Cp thinks that the 14 variable model will have squared errors 59 times greater than the 3 variable model with len, slim, and acpt.

```
> 14/0.2356971
[1] 59.39827
```

A key problem is that variable selection procedures overfit. Need to cross-validate!

Variable Selection, Continued
(O2Uptake Example)

Load libraries

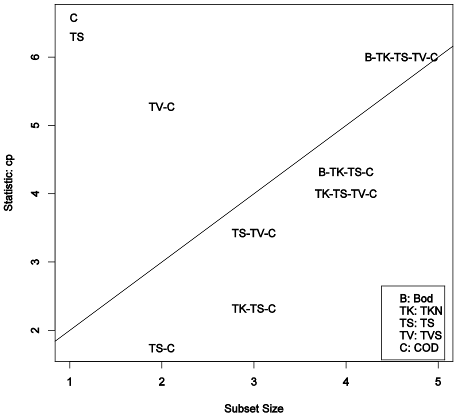
```
> library(leaps)
> library(car)

> O2Uptake[1:3,]
  Day  Bod TKN   TS  TVS  COD O2UP LogO2Up
1   0 1125 232 7160 85.9 8905 36.0  1.5563
2   7  920 268 8804 86.5 7388  7.9  0.8976
3  15  835 271 8108 85.2 5348  5.6  0.7482

> dim(O2Uptake)
[1] 20  8
```

Find best 2 models of each size.

```
> mod<-regsubsets(x=O2Uptake[,2:6],y=O2Uptake$LogO2Up,nbest=2)
> summary(mod)
2 subsets of each size up to 5
Selection Algorithm: exhaustive
         Bod TKN TS  TVS COD
1  ( 1 ) " " " " "*" " " " "
1  ( 2 ) " " " " " " " " "*"
2  ( 1 ) " " " " "*" " " "*"
2  ( 2 ) " " " " " " "*" "*"
3  ( 1 ) " " "*" "*" " " "*"
3  ( 2 ) " " " " "*" "*" "*"
4  ( 1 ) " " "*" "*" "*" "*"
4  ( 2 ) "*" "*" "*" " " "*"
5  ( 1 ) "*" "*" "*" "*" "*"
```

C_p plot

```
> subsets(mod,stat="cp")
> abline(1,1)
```



PRESS (and writing little programs in R)

We have seen many times in many ways that ordinary residuals, say E_i tend to be too small, because Y_i was used in fitting the model, so the model is too close to Y_i. Predicting Y_i having fitted the model using Y_i is called "in-sample-prediction," and it tends to suggest that a model is better than it is, because it saying you are making progress getting close to your current Y_i's, even if you could not do well in predicting a new Y_i.

If you left i out of the regression, and tried to predict Y_i from the regression without i, the error you would make is:

$$Y_i - \hat{Y}_{i[i]} = V_i \text{ say.}$$

Here, V_i is an "out-of-sample prediction," a true effort to predict a "new" observation, because i did not get used in fitting this equation. It gives me a fair idea as to how well a model can predict an observation not used in fitting the model.

The predicted error sum of squares or PRESS is

$$\text{PRESS} = \Sigma\ V_i^2.$$

It turns out that $V_i = E_i/(1-h_i)$ where E_i is the residual and h_i is the leverage or hatvalue.

```
> fuel[1:2,]
  ID state Fuel Tax License   Inc  Road
1  1    ME  541   9    52.5 3.571 1.976
2  2    NH  524   9    57.2 4.092 1.250
> attach(fuel)
> modMAX<-lm(Fuel~Tax+License+Inc+Road)
```

These are the out of sample prediction errors or V_i's:

```
> V<-modMAX$residual/(1-hatvalues(modMAX))
```

Let's look at Wyoming, WY. It's residual is about 235 gallons:

```
> modMAX$residual[state=="WY"]
      40
234.9472
```

but it's out of sample prediction error is about 26 gallons larger:

```
> V[state=="WY"]
      40
260.9721
```

PRESS (and writing little programs in R), continued

PRESS is the sum of the squares of the V_i

```
> sum(V^2)
[1] 235401.1
```

How does PRESS compare to R^2? Well R^2 is an in-sample measure, while press is an out-of-sample measure. For modMAX, R^2 is:

```
> summary(modMAX)$r.squared
[1] 0.6786867
```

Let's take Road out of the model, and see what happens to R^2 and PRESS.

```
> modSmall<-lm(Fuel~Tax+License+Inc)

> summary(modSmall)$r.squared
[1] 0.6748583
```

So R^2 went down, "got worse," which it always does when you delete variables;

```
> Vsmall<-modSmall$residual/(1-hatvalues(modSmall))
> sum(Vsmall^2)
[1] 229998.9
```

however, PRESS went down too, or "got better." In other words, adding Road to the model makes the residuals smaller, as adding variables always does, but it makes the prediction errors bigger. Sometimes adding a variable makes prediction errors smaller, sometimes it makes them bigger, and PRESS tells which is true in your model.

You could compute PRESS as above each time you fit a model, but it is easier to add a little program to R. Here is how you write a program called PRESS that computes PRESS.

```
> PRESS<-function(mod){
+ V<-mod$residual/(1-hatvalues(mod))
+ sum(V^2)}
```

If you type in the name of your program, here PRESS, it prints the program for you to look at.

```
> PRESS
function(mod){
     V<-mod$residual/(1-hatvalues(mod))
     sum(V^2)}
```

Your new program will compute PRESS for you:

```
> PRESS(modMAX)
[1] 235401.1
> PRESS(modSmall)
[1] 229998.9
```

Variance Inflation Factor (VIF)

Need library DAAG. You may have to install it the first time.

```
> library(DAAG)

> fuel[1:2,]
  ID state Fuel Tax License   Inc  Road
1  1    ME  541   9    52.5 3.571 1.976
2  2    NH  524   9    57.2 4.092 1.250

> attach(fuel)
```

Run a regression, saving results.

```
> mod<-lm(Fuel~Tax+License+Inc+Road)
```

Here are the VIF's

```
> vif(mod)
    Tax License     Inc    Road
 1.6257  1.2164  1.0433  1.4969
```

You can convert the VIF's to R^2

```
> 1-1/vif(mod)
       Tax    License        Inc       Road
0.38488036 0.17790201 0.04150292 0.33195270
```

This says: If you predict Tax from License, Inc and Road, the R^2 is 0.3849. You could do the regression and get the same answer; see below.

```
> summary(lm(Tax~License+Inc+Road))
Call:
lm(formula = Tax ~ License + Inc + Road)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.14672    1.35167   8.247 1.79e-10 ***
License     -0.05791    0.02058  -2.814  0.00728 **
Inc          0.15455    0.19881   0.777  0.44109
Road        -0.14935    0.03232  -4.621 3.34e-05 ***

Residual standard error: 0.7707 on 44 degrees of freedom
Multiple R-Squared: 0.3849,     Adjusted R-squared: 0.3429
F-statistic: 9.177 on 3 and 44 DF,  p-value: 7.857e-05
```

Spjotvoll's Method in Variable Selection

The maximum model has variables {1,2,…,k}, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$ where the errors ε are independent and $N(0,\sigma^2)$. Suppose that T is the true model, where T is a subset of {1,2,…,k}. That is, T is the model with exactly the nonzero β_js, so $\beta_j=0$ if and only if j is not in T. We could do a general linear hypothesis F-test to test model T against the maximum model, and if we did just this one test, the chance that we would falsely reject model T at level $\alpha=0.05$ would be 5%. The problem is that we don't know T, so we end up testing many models, and might make many mistakes in all those tests.

Spjotvoll (1977) defines an inadequate model as any model that omits a variable in T, and an adequate model as any model that includes all of the variables in T, perhaps including some extra variables whose coefficients are 0. If k=5 and T={1,2}, then {1,2} and {1,2,3} are adequate models, but {1,3} and {2,3,4,5} are inadequate models. Spjotvoll wants to reject some models as inadequate. He is not worried about having too many variables, and is only worried about omitting a needed variable.

Spjotvoll (1977) declares a model Q to be inadequate at level $\alpha=0.05$ if and only if the F-test rejects both Q and every model contained in Q. With k=5, to reject Q={1,3} as inadequate, you would have to reject Q={1,3}, and also {1}, {3} and {}, where {} is the model with no variables, that is, {} is $y = \beta_0 + \varepsilon$. So to reject Q={1,3}, four F-tests have to reject $\alpha=0.05$. A different way of saying the same thing is that the maximum of the four p-values must be less than or equal to $\alpha=0.05$; that is, the maximum of the F-test p-values for {1,3}, {1}, {3}, and {} must be less than or equal to 0.05. The chance that Spjotvoll's makes at least one mistake, saying that an adequate model is inadequate, is $\alpha=0.05$, despite doing lots of tests.

It is easy to see why this works. Model Q is adequate if and only if the true model T is a subset of Q, possibly T=Q; that's the definition of "adequate". Suppose Q is adequate, so it contains (or equals) T. To declare Q inadequate at the 0.05 level, you have to reject every model formed as a subset of Q – that's Spjotvoll's method – but as T is one of those subsets, you have to reject T, and the chance that the F-test falsely rejects the true model T is α=0.05.

Instead of focusing on α=0.05, we could do this for any α in just the same way. We can define an adjusted p-value for model Q as the maximum F-test p-value for all of the models that are subsets of Q, including Q itself and the empty model. This adjusted p-value rejects Q as inadequate at level α if and only if the adjusted p-value is at most α.

Spjotvoll’s Method in Variable Selection, continued.

```
> attach(O2Uptake)
> y<-LogO2Up
> x<-O2Uptake[,2:6]
```

Test of model lm(LogO2Up~TS+COD). Compare with row 7.

```
>anova(lm(LogO2Up~TS+COD),lm(LogO2Up~Bod+TKN+TS+TVS+COD))
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     17 1.08502
2     14 0.96512  3    0.1199 0.5797 0.6379
```

The p-value from the F-test is 0.638, whereas the adjusted p-value is the maximum of the p-values for the models contained in {TS,COD}, namely 0.638 for {TS,COD}, 0.139 for {TS}, 0.129 for {COD} and 0.000 for the empty model {}, so the adjusted p-value is 0.638. So this model is not declared inadequate.

Model #16, {TKN, TVS}, is declared inadequate, but its adjusted p-value comes from model #5, {TVS}, because the largest p-value for a model contained in {TKN,TVS} is from model {TVS}.

```
> spjotvoll(x,y)
   p     Cp     Fp  pval adjusted.pval inadequate Bod TKN TS TVS COD
1  1 55.463 11.893 0.000         0.000       TRUE   0   0  0   0   0
2  2  6.297  2.074 0.139         0.139      FALSE   0   0  1   0   0
3  2  6.576  2.144 0.129         0.129      FALSE   0   0  0   0   1
4  2 13.505  3.876 0.025         0.025       TRUE   1   0  0   0   0
5  2 20.331  5.583 0.007         0.007       TRUE   0   0  0   1   0
6  2 56.861 14.715 0.000         0.000       TRUE   0   1  0   0   0
7  3  1.739  0.580 0.638         0.638      FALSE   0   0  1   0   1
8  3  5.274  1.758 0.201         0.201      FALSE   0   0  0   1   1
9  3  6.872  2.291 0.123         0.129      FALSE   0   1  0   0   1
10 3  6.885  2.295 0.122         0.139      FALSE   1   0  1   0   0
11 3  7.165  2.388 0.113         0.139      FALSE   0   0  1   1   0
12 3  7.336  2.445 0.107         0.139      FALSE   0   1  1   0   0
13 3  7.705  2.568 0.096         0.129      FALSE   1   0  0   0   1
14 3  9.097  3.032 0.065         0.065      FALSE   1   1  0   0   0
15 3 11.331  3.777 0.036         0.036       TRUE   1   0  0   1   0
16 3 21.369  7.123 0.004         0.007       TRUE   0   1  0   1   0
17 4  2.319  0.160 0.854         0.854      FALSE   0   1  1   0   1
18 4  3.424  0.712 0.508         0.638      FALSE   0   0  1   1   1
19 4  3.439  0.720 0.504         0.638      FALSE   1   0  1   0   1
20 4  5.665  1.833 0.196         0.201      FALSE   0   1  0   1   1
21 4  6.253  2.126 0.156         0.156      FALSE   1   1  1   0   0
22 4  6.515  2.258 0.141         0.141      FALSE   1   1  0   0   1
23 4  7.152  2.576 0.112         0.201      FALSE   1   0  0   1   1
24 4  8.155  3.077 0.078         0.139      FALSE   1   0  1   1   0
25 4  8.165  3.082 0.078         0.139      FALSE   0   1  1   1   0
26 4  8.681  3.341 0.065         0.065      FALSE   1   1  0   1   0
27 5  4.001  0.001 0.972         0.972      FALSE   0   1  1   1   1
28 5  4.319  0.319 0.581         0.854      FALSE   1   1  1   0   1
29 5  5.068  1.068 0.319         0.638      FALSE   1   0  1   1   1
30 5  6.776  2.776 0.118         0.201      FALSE   1   1  0   1   1
31 5  7.697  3.697 0.075         0.156      FALSE   1   1  1   1   0
32 6  6.000     NA 1.000         1.000      FALSE   1   1  1   1   1
```

Spjotvoll’s method is a case of closted testing; see Marcus et al. (1976)

Spjotvoll, E. (1977) Alternatives to plotting C_P in multiple regression. Biometrika 64, 1-8. Correction: page 241.

Marcus R, Peritz E, Gabriel KR. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63, 655—60.

ANOVA

Memory data. 36 kids randomized to form 3 groups of 12, which were given different treatments. The 'data' are columns 1 and 2, for `group` *and Y=*`words`*. It is a "balanced design" because every group has the same sample size. The rest of memory consists of various ways of coding the 2 degrees of freedom between the three groups into two coded variables. The variables* `ten` *and* `five` *are "dummy variables" for two categories, leaving out the third category. The variables* `five_ten` *and* `nh_ten` *are used to produce effects that are deviations from a mean for all three groups. The best coding is* `hier` *and* `info` *which involve "orthogonal contrasts," discussed below. It is only with orthogonal contrasts that you partition the sum of squares between groups into single degree of freedom parts that add back to the total.*

```
> memory[1:3,]
> memory
    group words five_ten nh_ten ten five hier info
1     Ten    50       -1     -1   1    0  0.5    1
2     Ten    49       -1     -1   1    0  0.5    1
3     Ten    44       -1     -1   1    0  0.5    1
4     Ten    31       -1     -1   1    0  0.5    1
5     Ten    47       -1     -1   1    0  0.5    1
6     Ten    38       -1     -1   1    0  0.5    1
7     Ten    38       -1     -1   1    0  0.5    1
8     Ten    48       -1     -1   1    0  0.5    1
9     Ten    45       -1     -1   1    0  0.5    1
10    Ten    48       -1     -1   1    0  0.5    1
11    Ten    35       -1     -1   1    0  0.5    1
12    Ten    33       -1     -1   1    0  0.5    1
13   Five    44        1      0   0    1  0.5   -1
14   Five    41        1      0   0    1  0.5   -1
15   Five    34        1      0   0    1  0.5   -1
16   Five    35        1      0   0    1  0.5   -1
17   Five    40        1      0   0    1  0.5   -1
18   Five    44        1      0   0    1  0.5   -1
19   Five    39        1      0   0    1  0.5   -1
20   Five    39        1      0   0    1  0.5   -1
21   Five    45        1      0   0    1  0.5   -1
22   Five    41        1      0   0    1  0.5   -1
23   Five    46        1      0   0    1  0.5   -1
24   Five    32        1      0   0    1  0.5   -1
25 NoHier    33        0      1   0    0 -1.0    0
26 NoHier    36        0      1   0    0 -1.0    0
27 NoHier    37        0      1   0    0 -1.0    0
28 NoHier    42        0      1   0    0 -1.0    0
29 NoHier    33        0      1   0    0 -1.0    0
30 NoHier    33        0      1   0    0 -1.0    0
31 NoHier    41        0      1   0    0 -1.0    0
32 NoHier    33        0      1   0    0 -1.0    0
33 NoHier    38        0      1   0    0 -1.0    0
34 NoHier    39        0      1   0    0 -1.0    0
35 NoHier    28        0      1   0    0 -1.0    0
36 NoHier    42        0      1   0    0 -1.0    0
```

ANOVA

```
> attach(memory)
```

The anova can be done as a linear model with a factor as the predictor.

```
> anova(lm(words~group))
Analysis of Variance Table

Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
group      2 215.06  107.53  3.7833 0.03317 *
Residuals 33 937.92   28.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Or you can use the aov command. You get the same answer.

```
> summary(aov(words~group))
            Df Sum Sq Mean Sq F value  Pr(>F)
group        2 215.06  107.53  3.7833 0.03317 *
Residuals   33 937.92   28.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Comparisons using Tukey’s Method

```
> TukeyHSD(aov(words~group))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = words ~ group)

$group
                 diff        lwr       upr
NoHier-Five -3.750000 -9.0905714  1.590571
Ten-Five     2.166667 -3.1739047  7.507238
Ten-NoHier   5.916667  0.5760953 11.257238
```

These are simultaneous 95% confidence intervals for the difference in means between two groups. The promise is that all 3 confidence intervals will cover their population differences in 95% of experiments. This is a better promise than that each one, by itself, covers in 95% of uses, because then the first interval would have a 5% chance of error, and so would the second, and so would the third, and the chance of at least one error would be greater than 5%. If the interval includes zero, as the first two intervals do, then you can't declare the two groups significantly different. If the interval excludes zero, as the third interval does, you can declare the two groups significantly different.

Tukey, Bonferroni and Holm

```
> help(pairwise.t.test)
> help(p.adjust)

> TukeyHSD(aov(words~group))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = words ~ group)
$group
                 diff        lwr       upr
NoHier-Five -3.750000 -9.0905714  1.590571
Ten-Five     2.166667 -3.1739047  7.507238
Ten-NoHier   5.916667  0.5760953 11.257238

> pairwise.t.test(words,group,p.adj = "none")
        Pairwise comparisons using t tests with pooled SD
data:  words and group
       Five  NoHier
NoHier 0.094 -
Ten    0.327 0.010
P value adjustment method: none

>  pairwise.t.test(words,group,p.adj = "bonf")
        Pairwise comparisons using t tests with pooled SD
data:  words and group
       Five  NoHier
NoHier 0.283 -
Ten    0.980 0.031
P value adjustment method: bonferroni

> pairwise.t.test(words,group,p.adj = "holm")
        Pairwise comparisons using t tests with pooled SD
data:  words and group
       Five  NoHier
NoHier 0.189 -
Ten    0.327 0.031
```

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70. http://www.jstor.org/

Wright, S. P. (1992). Adjusted P-values for simultaneous inference. *Biometrics*, 48, 1005-1013. http://www.jstor.org/

ANOVA: Many Ways to Code the Same Anova

Here are three different codings with the same Anova table. Notice that much is the same, but some things differ. >

```
summary(lm(words~ten+five))
Call:
lm(formula = words ~ ten + five)
Residuals:
    Min      1Q  Median      3Q     Max 
-11.167  -3.479   0.875   4.771   7.833 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   36.250      1.539  23.554   <2e-16 ***
ten            5.917      2.176   2.718   0.0104 *  
five           3.750      2.176   1.723   0.0943 .  

Residual standard error: 5.331 on 33 degrees of freedom
Multiple R-Squared: 0.1865,     Adjusted R-squared: 0.1372 
F-statistic: 3.783 on 2 and 33 DF,  p-value: 0.03317 

> summary(lm(words~five_ten+nh_ten))
Call:
lm(formula = words ~ five_ten + nh_ten)
Residuals:
    Min      1Q  Median      3Q     Max 
-11.167  -3.479   0.875   4.771   7.833 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  39.4722     0.8885  44.424   <2e-16 ***
five_ten      0.5278     1.2566   0.420   0.6772    
nh_ten       -3.2222     1.2566  -2.564   0.0151 *  

Residual standard error: 5.331 on 33 degrees of freedom
Multiple R-Squared: 0.1865,     Adjusted R-squared: 0.1372 
F-statistic: 3.783 on 2 and 33 DF,  p-value: 0.03317 

> summary(lm(words~hier+info))
Call:
lm(formula = words ~ hier + info)
Residuals:
    Min      1Q  Median      3Q     Max 
-11.167  -3.479   0.875   4.771   7.833 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  39.4722     0.8885  44.424   <2e-16 ***
hier          3.2222     1.2566   2.564   0.0151 *  
info          1.0833     1.0882   0.996   0.3267    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.331 on 33 degrees of freedom
Multiple R-Squared: 0.1865,     Adjusted R-squared: 0.1372 
          F-statistic: 3.783 on 2 and 33 DF,  p-value: 0.03317
```

Contrasts in ANOVA: Better Coding of Nominal Variables

The third coding is best, because the predictors (contrasts) are uncorrelated, so the sums of squares partition.

```
> cor(memory[,3:4])
         five_ten nh_ten
five_ten      1.0    0.5
nh_ten        0.5    1.0

>  cor(memory[,5:6])
      ten five
ten   1.0 -0.5
five -0.5  1.0

>  cor(memory[,7:8])
     hier info
hier    1    0
info    0    1
```

Notice that hier and info have zero correlation: they are orthogonal. Because of this, you can partition the two degrees of freedom between groups into separate sums of squares.

```
> anova(lm(words~hier+info))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
hier       1 186.89  186.89  6.5756 0.01508 *
info       1  28.17   28.17  0.9910 0.32674
Residuals 33 937.92   28.42
---
```

Reverse the order of info and hier, and you get the same answer.

```
> anova(lm(words~info+hier))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
info       1  28.17   28.17  0.9910 0.32674
hier       1 186.89  186.89  6.5756 0.01508 *
Residuals 33 937.92   28.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can't do this with correlated predictors, because they overlap, and the order of the variables changes the sum of squares for the variable, so one can't really say that what portion of the sum of squares belongs to the variable.

```
> anova(lm(words~ten+five))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
ten        1 130.68  130.68  4.5979 0.03947 *
five       1  84.38   84.38  2.9687 0.09425 .
Residuals 33 937.92   28.42

> anova(lm(words~five+ten))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
five       1   5.01    5.01  0.1764 0.67720
ten        1 210.04  210.04  7.3902 0.01037 *
Residuals 33 937.92   28.42
```

Coding the Contrasts in ANOVA in R

This is about shortcuts to get R to convert a nominal variable into several contrasts. There's no new statistics here; just R details.

The group variable, memory$group, is a factor.

```
> is.factor(memory$group)
[1] TRUE
```

This factor has 3 levels. Notice that the levels are ordered and the order matters.

```
> levels(memory$group)
[1] "Five"   "NoHier" "Ten"

> memory$group
 [1] Ten    Ten    Ten    Ten    Ten    Ten    Ten    Ten    Ten
[10] Ten    Ten    Ten    Five   Five   Five   Five   Five   Five
[19] Five   Five   Five   Five   Five   Five   NoHier NoHier NoHier
[28] NoHier NoHier NoHier NoHier NoHier NoHier NoHier NoHier NoHier
Levels: Five NoHier Ten
```

If you do nothing, R codes a factor in a linear model using 'dummy coding'.

```
> contrasts(memory$group)
       NoHier Ten
Five        0   0
NoHier      1   0
Ten         0   1
```

You can change the coding. Essentially, you can replace the little table above by whatever you want. We will build an new 3x2 table and redefine the contrasts to be this new table.

```
> hier2<-c(.5,-1,.5)
> hier2
[1]  0.5 -1.0  0.5

> info2<-c(-1,0,1)
> info2
[1] -1  0  1

> cm<-cbind(hier2,info2)
> cm
     hier2 info2
[1,]   0.5    -1
[2,]  -1.0     0
[3,]   0.5     1
```

So cm is our new table, and we redefine the contrasts for memory$group.

```
> contrasts(memory$group)<-cm
```

This replaces the 'dummy coding' by our new coding.

```
> contrasts(memory$group)
       hier2 info2
Five     0.5    -1
NoHier  -1.0     0
Ten      0.5     1
```

Coding the Contrasts in ANOVA, Continued

If you ask R to extend the contrasts into variables, it will do this with "model.matrix". Notice that this is the coding in the original data matrix, but R is happy to generate it for you using the contrasts you specified.

```
> m<-model.matrix(memory$words~memory$group)
> m
   (Intercept) memory$grouphier2 memory$groupinfo2
1             1               0.5                  1
2             1               0.5                  1
3             1               0.5                  1
4             1               0.5                  1
5             1               0.5                  1
6             1               0.5                  1
7             1               0.5                  1
8             1               0.5                  1
9             1               0.5                  1
10            1               0.5                  1
11            1               0.5                  1
12            1               0.5                  1
13            1               0.5                 -1
14            1               0.5                 -1
15            1               0.5                 -1
16            1               0.5                 -1
17            1               0.5                 -1
18            1               0.5                 -1
19            1               0.5                 -1
20            1               0.5                 -1
21            1               0.5                 -1
22            1               0.5                 -1
23            1               0.5                 -1
24            1               0.5                 -1
25            1              -1.0                  0
26            1              -1.0                  0
27            1              -1.0                  0
28            1              -1.0                  0
29            1              -1.0                  0
30            1              -1.0                  0
31            1              -1.0                  0
32            1              -1.0                  0
33            1              -1.0                  0
34            1              -1.0                  0
35            1              -1.0                  0
36            1              -1.0                  0

> hcontrast<-m[,2]
> icontrast<-m[,3]
```

We now do the anova with single degree of freedom contrasts.

```
> anova(lm(memory$words~hcontrast+icontrast))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
hcontrast  1 186.89  186.89  6.5756 0.01508 *
icontrast  1  28.17   28.17  0.9910 0.32674
Residuals 33 937.92   28.42
```

ANOVA DECOMPOSITION

```
>  mod<-aov(words~group,projections=T)

> mod$projections
   (Intercept)      group     Residuals
1     39.47222  2.6944444  7.833333e+00
2     39.47222  2.6944444  6.833333e+00
3     39.47222  2.6944444  1.833333e+00
4     39.47222  2.6944444 -1.116667e+01
5     39.47222  2.6944444  4.833333e+00
6     39.47222  2.6944444 -4.166667e+00
7     39.47222  2.6944444 -4.166667e+00
8     39.47222  2.6944444  5.833333e+00
9     39.47222  2.6944444  2.833333e+00
10    39.47222  2.6944444  5.833333e+00
11    39.47222  2.6944444 -7.166667e+00
12    39.47222  2.6944444 -9.166667e+00
13    39.47222  0.5277778  4.000000e+00
14    39.47222  0.5277778  1.000000e+00
15    39.47222  0.5277778 -6.000000e+00
16    39.47222  0.5277778 -5.000000e+00
17    39.47222  0.5277778 -9.503032e-16
18    39.47222  0.5277778  4.000000e+00
19    39.47222  0.5277778 -1.000000e+00
20    39.47222  0.5277778 -1.000000e+00
21    39.47222  0.5277778  5.000000e+00
22    39.47222  0.5277778  1.000000e+00
23    39.47222  0.5277778  6.000000e+00
24    39.47222  0.5277778 -8.000000e+00
25    39.47222 -3.2222222 -3.250000e+00
26    39.47222 -3.2222222 -2.500000e-01
27    39.47222 -3.2222222  7.500000e-01
28    39.47222 -3.2222222  5.750000e+00
29    39.47222 -3.2222222 -3.250000e+00
30    39.47222 -3.2222222 -3.250000e+00
31    39.47222 -3.2222222  4.750000e+00
32    39.47222 -3.2222222 -3.250000e+00
33    39.47222 -3.2222222  1.750000e+00
34    39.47222 -3.2222222  2.750000e+00
35    39.47222 -3.2222222 -8.250000e+00
36    39.47222 -3.2222222  5.750000e+00
attr(,"df")
(Intercept)       group   Residuals
          1           2          33
```

Orthogonal and Non-orthogonal Predictors

```
> attach(memory)
> summary(aov(words~group))
            Df Sum Sq Mean Sq F value  Pr(>F)
group        2 215.06  107.53  3.7833 0.03317 *
Residuals   33 937.92   28.42
---
```

ten and five are not orthogonal predictors – so there is not a unique sum of squares for each

```
> anova(lm(words~ten+five))
Analysis of Variance Table
Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
ten        1 130.68  130.68  4.5979 0.03947 *
five       1  84.38   84.38  2.9687 0.09425 .
Residuals 33 937.92   28.42
---
> anova(lm(words~five+ten))
Analysis of Variance Table

Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
five       1   5.01    5.01  0.1764 0.67720
ten        1 210.04  210.04  7.3902 0.01037 *
Residuals 33 937.92   28.42
```

her and info are orthogonal predictors – so there is a unique sum of squares for each

```
> anova(lm(words~hier+info))
Analysis of Variance Table

Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
hier       1 186.89  186.89  6.5756 0.01508 *
info       1  28.17   28.17  0.9910 0.32674
Residuals 33 937.92   28.42

>  anova(lm(words~info+hier))
Analysis of Variance Table

Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
info       1  28.17   28.17  0.9910 0.32674
hier       1 186.89  186.89  6.5756 0.01508 *
Residuals 33 937.92   28.42
```

Simulating in R

Ten observations from the standard Normal distribution:

```
> rnorm(10)
 [1]  0.8542301 -1.3331572  1.4522862  0.8980641  0.1456334
 [6]  0.4926661 -0.4366962  0.6204263 -0.1582319 -0.6444449
```

Fixed integer sequences

```
> 1:2
[1] 1 2

> 1:5
[1] 1 2 3 4 5

> 0:1
[1] 0 1
```

20 coin flips

```
> sample(0:1,20,r=T)
 [1] 0 1 0 0 1 1 0 0 0 1 0 1 0 0 0 1 0 1 1 1
```

10 random numbers from 1 to 5

```
> sample(1:5,10,r=T)
 [1] 5 2 3 5 2 2 1 1 3 2
```

More information:

```
     help(sample)
     help(rnorm)
```

STATISTICS 500 FALL 2006 PROBLEM 1 DATA PAGE 1

Due in class Thusday 26 Oct 2006

This is an exam. Do not discuss it with anyone.

The data concern Y=sr=aggregate personal savings in 50 countries over ten years, as predicted by four predictors, the percentages of young and old people, per-capita disposable income, and the growth rate in per-capita disposable income. (A related paper, which you need not consult, is Modigliani (1988), The role of intergenerational transfers and life cycle saving in the accumulation of wealth, *Journal of Economic Perspectives*, 2, 15-40.)

In R, type:

```
> data(LifeCycleSavings)
```

and the data should enter your workspace as an object. In R, I would type:

```
> attach(LifeCycleSavings)
> nation<-rownames(LifeCycleSavings)
```

so the variable "nation" would have the country names.

The data are also available in JMP, Excel and text file formats publicly at
http://stat.wharton.upenn.edu/statweb/course/Fall2006/stat500/
or for Wharton accounts at the course download at:
http://www-stat.wharton.upenn.edu/

First 2 of 50 lines of data:

Country	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93

…

```
LifeCycleSavings          package:datasets          R Documentation
Intercountry Life-Cycle Savings Data
Description: Data on the savings ratio 1960-1970.
Usage:  LifeCycleSavings
Format:  A data frame with 50 observations on 5 variables.
       [,1]  sr     numeric  aggregate personal savings
       [,2]  pop15  numeric  % of population under 15
       [,3]  pop75  numeric  % of population over 75
       [,4]  dpi    numeric  real per-capita disposable income
       [,5]  ddpi   numeric  % growth rate of dpi
Details:
     Under the life-cycle savings hypothesis as developed by Franco
     Modigliani, the savings ratio (aggregate personal saving divided
     by disposable income) is explained by per-capita disposable
     income, the percentage rate of change in per-capita disposable
     income, and two demographic variables: the percentage of
     population less than 15 years old and the percentage of the
     population over 75 years old. The data are averaged over the
     decade 1960-1970 to remove the business cycle or other short-term
     fluctuations.
Source:
     Sterling, Arnie (1977) Unpublished BS Thesis. Massachusetts
     Institute of Technology.
     Belsley, D. A., Kuh. E. and Welsch, R. E. (1980) _Regression
     Diagnostics_. New York: Wiley.
```

STATISTICS 500 FALL 2006 PROBLEM 1 DATA PAGE 2

Due in class Thusday 26 Oct 2006

This is an exam. Do not discuss it with anyone.

The following models are mentioned on the answer page. Please note that different Greek letters are used so that different things have different symbols – there is no special meaning to θ or β -- they are just names. Notice carefully that the subscripts go from 0 to k in a model with k variables, but which variable is variable #1 changes from model to model.

Model 1: $sr = \theta_0 + \theta_1$ pop75 + η with $\eta \sim_{iid} N(0,\omega^2)$

Model 2: $sr = \beta_0 + \beta_1$ pop15 + β_2 pop75 + β_3 dpi + β_4 ddpi + ε with $\varepsilon \sim_{iid} N(0,\sigma^2)$

--This problem set is an exam. Do not discuss it with anyone. If you discuss it with anyone, you have cheated on an exam.

--Write your name and id# on BOTH sides of the answer page.

--Write answers in the spaces provided. Brief answers suffice. Do not attach additional pages. Do not turn in computer output. Turn in only the answer page.

--If a question asks you to circle an answer, then circle an answer. If you circle the correct answer you are correct. If you circle the incorrect answer you are incorrect. If you cross out an answer, no matter which answer you cross out, you are incorrect.

--If a question has several parts, answer every part. It is common to lose points by not answering part of a question.

Name: Last, First: ______________________________ **ID#**: ______________________

Statistics 500 Fall 2006 Problem 1 Answer Page 1

This is an exam. Do not discuss it with anyone.

1. Plot the data in various ways and answer the following questions.

Question	CIRCLE ONE
Which country has the highest savings rate (sr)?	US Japan Chile Bolivia Libya
Which country has the highest income (dpi)?	US Japan Chile Bolivia Libya
In which country is income rising at the fastest rate (ddpi)?	US Japan Chile Bolivia Libya
In these data, if a country has more than 40% of its population under 15 years old, then it has less than 3% of its population over 75 years old.	TRUE FALSE

2. Assume model #1 (on the data page) is true, and fit model #1, and use it to answer the following questions. Notice that θ_1 is the coefficient of pop75 in model #1.

Question	
Test the hypothesis that H_0: $\theta_1=0$ in model #1. What is the **name** of the test? What is the **numerical value** of the test statistic? What is the two-sided **P-value**? Is is the null hypothesis plausible (**Circle One**)?	Name of test: ___________________ Numerical value: ________________ P-value: _______________________ H_0 is PLAUSIBLE NOT PLAUSIBLE
What is the numerical value of the least squares estimate of θ_1 in model #1 and what is its estimated standard error?	Estimate of θ_1: ___________________________ Estimated standard error of θ_1: _______________
The fitted savings rates under model #1 are about 1% higher in a country with about 1% more people over 75 years of age. (Here "about" means "round to the nearest whole percent, so that 87.26% is about 87%.)	CIRCLE ONE TRUE FALSE
Given the 95% confidence interval for θ_1 in model #1.	[____________ , ______________]
What is the estimate of ω (not ω^2 !), the standard deviation of the η's?	Numerical estimate of ω: __________________

Name: Last, First: ___________________________ **ID#**: ____________________

Statistics 500 Fall 2006 Problem 1 Answer Page 2

This is an exam. Do not discuss it with anyone.

3. Assume that model #2 is true, and use its fit to answer the following questions.

In model #2, the coefficient of pop75 is β_2. What is the least squares estimate of β_2 in model #2? What is the two-sided P-value for testing H_0: β_2=0 in model #2?	Estimate of β_2: ____________________ P-value: ___________________________
Test the hypothesis H_0: β_1= β_2= β_3= β_4=0 in model #2. What is the name of the test? What is the value of the test statistic? What is the P-value? Is H_0 plausible? CIRCLE ONE	Name of test: ________________________ Test statistic: ________________________ P-value: ________________________ H_0 is PLAUSIBLE NOT PLAUSIBLE
Do a Normal plot of the residuals. Do the Shapiro-Wilk test applied to the residuals. Is there **clear evidence** that the residuals are not Normal? What is the **P-value** from the Shapiro-Wilk test?	CIRCLE ONE CLEAR EVIDENCE OTHER P-value: ______________________
Plot the residuals from model #2 against the predicted values. Is there a clear bend indicating a nonlinear relationship?	CIRCLE ONE CLEAR BEND NO CLEAR BEND
Which country has the largest absolute residual? What is the numerical value of this residual, including its sign? Did this country save more or less than the model predicted?	Country: ___________ Value: ___________ CIRCLE ONE MORE LESS

4. In model #2, **test H_0: β_1= β_2=0** which asserts that neither pop15 nor pop75 are needed in a model that includes dpi and ddpi. **Fill in table, F & P-value, CIRCLE ONE**.

	CIRCLE all variable names that apply	Sum of Squares	Degrees of freedom	Mean Square
Full model includes which variables?	pop15 pop75 dpi ddpi			
Reduced model includes which vars?	pop15 pop75 dpi ddpi			
Which variables add to the reduced model to give the full model	pop15 pop75 dpi ddpi			
Residual	XXXXXXXXXXX XXXXXXXXXXX			

F-value: ______ **P-value**: _______ H_0: β_1= β_2=0 is PLAUSIBLE NOT PLAUSIBLE

Statistics 500 Fall 2006 Problem 1 Answer Page 1

This is an exam. Do not discuss it with anyone.

1. Plot the data in various ways and use the data to answer the following questions.

Question	CIRCLE ONE
Which country has the highest savings rate (sr)?	US [✓ Japan] Chile Bolivia Libya
Which country has the highest income (dpi)?	[✓ US] Japan Chile Bolivia Libya
In which country is income rising at the fastest rate (ddpi)?	US Japan Chile Bolivia [✓ Libya]
In these data, if a country has more than 40% of its population under 15 years old, then it has less than 3% of its population over 75 years old.	[✓ TRUE] FALSE

2. Assume model #1 (on the data page) is true, and fit model #1, and use it to answer the following questions. Notice that θ_1 is the coefficient of pop75 in model #1.

Question	
Test the hypothesis that H_0: $\theta_1=0$ in model #1. What is the **name** of the test? What is the **numerical value** of the test statistic? What is the two-sided **P-value**? Is is the null hypothesis plausible (**Circle One**)?	Name of test: t-test Numerical value: $t = 2.31$ P-value: 0.025 H_0 is PLAUSIBLE [✓ NOT PLAUSIBLE]
What is the numerical value of the least squares estimate of θ_1 in model #1 and what is its estimated standard error?	Estimate of θ_1: 1.099 Estimated standard error of θ_1: 0.475
The fitted savings rates under model #1 are about 1% higher in a country with about 1% more people over 75 years of age. (Here "about" means "round to the nearest whole percent, so that 87.26% is about 87%.)	CIRCLE ONE [✓ TRUE] FALSE Bit of a surprise – you'd think people over 75 would be spending from their savings.
Given the 95% confidence interval for θ_1 in model #1.	[0.14 , 2.05]
What is the estimate of ω (not ω^2 !), the standard deviation of the η's?	Numerical estimate of ω: 4.3 (that is 4.3%)

Statistics 500 Fall 2006 Problem 1 Answer Page 2

3. Assume that model #2 is true, and use its fit to answer the following questions.

In model #2, the coefficient of pop75 is β_2. What is the least squares estimate of β_2 in model #2? What is the two-sided P-value for testing H_0: β_2=0 in model #2?	Estimate of β_2: -1.69 P-value: 0.13
Test the hypothesis H_0: β_1= β_2= β_3= β_4=0 in model #2. What is the name of the test? What is the value of the test statistic? What is the P-value? Is H_0 plausible? CIRCLE ONE	Name of test: F-test Test statistic: 5.76 on 4 and 45 degrees of freedom P-value: 0.0008 H_0 is PLAUSIBLE [✓ NOT PLAUSIBLE]
Do a Normal plot of the residuals. Do the Shapiro-Wilk test applied to the residuals. Is there **clear evidence** that the residuals are not Normal? What is the **P-value** from the Shapiro-Wilk test?	CIRCLE ONE CLEAR EVIDENCE [✓ OTHER] P-value: 0.85
Plot the residuals from model #2 against the predicted values. Is there a clear bend indicating a nonlinear relationship?	CIRCLE ONE CLEAR BEND [✓ NO CLEAR BEND]
Which country has the largest absolute residual? What is the numerical value of this residual, including its sign? Did this country save more or less than the model predicted?	Country: Zambia Value: 9.75% CIRCLE ONE [✓ MORE] LESS

4. In model #2, test H_0: β_1= β_2=0 which asserts that neither pop15 nor pop75 are needed in a model that includes dpi and ddpi. **Fill in table, F & P-value, CIRCLE ONE**.

	LIST all variable names that apply	Sum of Squares	Degrees of freedom	Mean Square
Full model includes which variables?	pop15 pop75 dpi ddpi	332.92	4	83.2
Reduced model includes which vars?	dpi ddpi	158.91	2	79.5
Which variables add to the reduced model to give the full model	pop15 pop75	174.01	2	87.0
Residual	XXXXXXXXXXX XXXXXXXXXXX	650.71	45	14.5

F-value: 6.02 **P-value**: 0.005 H_0: β_1= β_2=0 is PLAUSIBLE [✓ NOT PLAUSIBLE]

Doing the Problem Set in R

(Fall 2006, Problem Set 1)

This data set happens to be in R, so get it by typing:

```
> data(LifeCycleSavings)
```

Look at the first two rows:

```
> LifeCycleSavings[1:2,]
             sr pop15 pop75     dpi ddpi
Australia 11.43 29.35  2.87 2329.68 2.87
Austria   12.07 23.32  4.41 1507.99 3.93
```

If you attach the data set, the variable LifeCycleSavings$sr can be called sr.

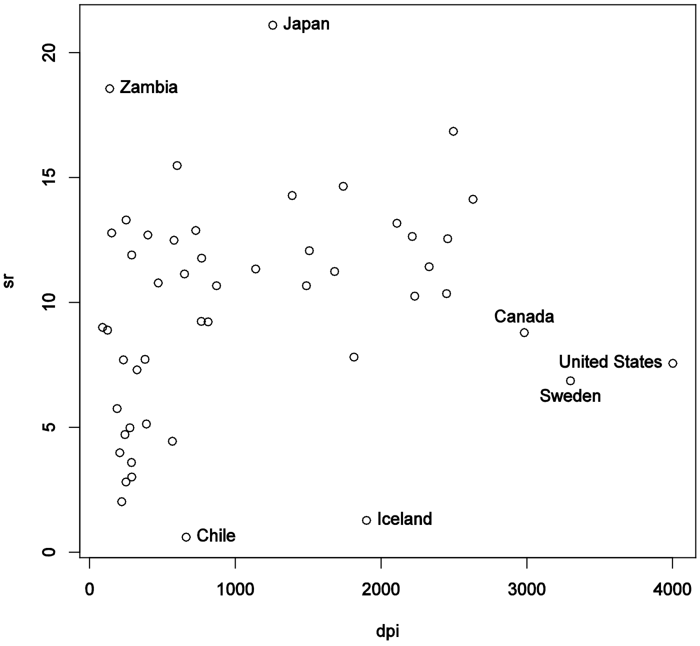
```
> attach(LifeCycleSavings)
```

In this data set, the country names are the row names. You can make them into a variable:

```
> nations<-rownames(LifeCycleSavings)
```

Plot the data. Identify the points.

```
> plot(dpi,sr)
> identify(dpi,sr,label=nations)
```



Doing the Problem Set in R, continued

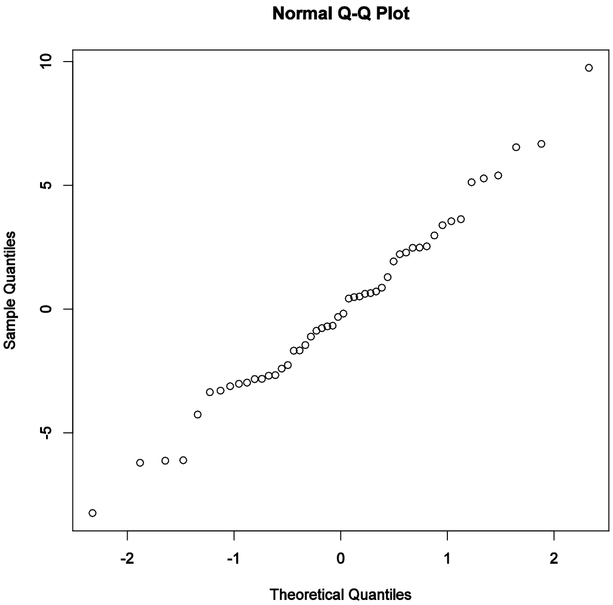
(Fall 2006, Problem Set 1)

```
> summary(lm(sr~pop75))

Call:
lm(formula = sr ~ pop75)

Residuals:
     Min       1Q   Median       3Q      Max
-9.26566 -3.22947  0.05428  2.33359 11.84979

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1517     1.2475   5.733  6.4e-07 ***
pop75         1.0987     0.4753   2.312   0.0251 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 4.294 on 48 degrees of freedom
Multiple R-Squared: 0.1002,     Adjusted R-squared: 0.08144
F-statistic: 5.344 on 1 and 48 DF,  p-value: 0.02513

> mod<-lm(sr~pop15+pop75+dpi+ddpi)
> summary(mod)
Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi)
Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-Squared: 0.3385,     Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Doing the Problem Set in R, continued (Fall 2006, Problem Set 1)

Normal quantile plot of residuals: Is it (more or less) a straight line?

```
> qqnorm(mod$residual)
```

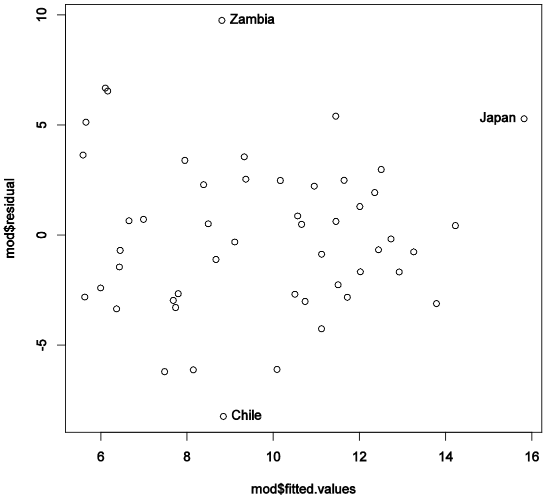


Shapiro-Wilk test of Normal distribution applied to residuals:

```
> shapiro.test(mod$residual)
        Shapiro-Wilk normality test
data:  mod$residual
W = 0.987, p-value = 0.8524
```

Plot residuals against predicted (or fitted) values:

```
> plot(mod$fitted.values,mod$residual)
> identify(mod$fitted.values,mod$residual,label=nations)
```



Doing the Problem Set in R, continued (Fall 2006, Problem Set 1)

General linear hypothesis: You've already fit the full model, called mod. Now fit the reduced model, called modReduce.

```
> modReduce<-lm(sr~dpi+ddpi)
```

Compare the two models:

```
> anova(modReduce,mod)

Analysis of Variance Table

Model 1: sr ~ dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     47 824.72
2     45 650.71  2    174.01 6.0167 0.004835 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

STATISTICS 500 FALL 2006 PROBLEM 2 DATA PAGE 1

Due in class Tuesday 28 November 2006

This is an exam. Do not discuss it with anyone.

The data set is the same as for the first problem set, although different issues will be examined.

In R, type:

```
> data(LifeCycleSavings)
```

and the data should enter your workspace as an object. In R, I would type:

```
> attach(LifeCycleSavings)
> nation<-rownames(LifeCycleSavings)
```

so the variable “nation” would have the country names.

The data are also available in JMP, Excel and text file formats publicly at

http://stat.wharton.upenn.edu/statweb/course/Fall2006/stat500/

or for Wharton accounts at the course download at:

http://www-stat.wharton.upenn.edu/

First 2 of 50 lines of data:

Country	**sr**	**pop15**	**pop75**	**dpi**	**ddpi**
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93

…

```
LifeCycleSavings          package:datasets          R Documentation
Intercountry Life-Cycle Savings Data
Description: Data on the savings ratio 1960-1970.
Usage:  LifeCycleSavings
Format:  A data frame with 50 observations on 5 variables.
       [,1]  sr     numeric  aggregate personal savings
       [,2]  pop15  numeric  % of population under 15
       [,3]  pop75  numeric  % of population over 75
       [,4]  dpi    numeric  real per-capita disposable income
       [,5]  ddpi   numeric  % growth rate of dpi
Details:
     Under the life-cycle savings hypothesis as developed by Franco
     Modigliani, the savings ratio (aggregate personal saving divided
     by disposable income) is explained by per-capita disposable
     income, the percentage rate of change in per-capita disposable
     income, and two demographic variables: the percentage of
     population less than 15 years old and the percentage of the
     population over 75 years old. The data are averaged over the
     decade 1960-1970 to remove the business cycle or other short-term
     fluctuations.
Source:
     Sterling, Arnie (1977) Unpublished BS Thesis. Massachusetts
     Institute of Technology.
     Belsley, D. A., Kuh. E. and Welsch, R. E. (1980) _Regression
                     Diagnostics_. New York: Wiley
```

STATISTICS 500 FALL 2006 PROBLEM 2 DATA PAGE 2

This is an exam. Do not discuss it with anyone.

The G7 or Group of 7 countries were Canada, France, Germany, Italy, Japan, the United Kingdom and the United States. **Construct a variable**, G7, which is 1 for the G7 countries and is zero for other countries.

The following models are mentioned on the answer page. Please note that different Greek letters are used so that different things have different symbols – there is no special meaning to θ or β -- they are just names. Notice carefully that the subscripts go from 0 to k in a model with k variables, but which variable is variable #1 changes from model to model.

Model 1: $\text{dpi} = \theta_0 + \theta_1\ \text{pop15} + \theta_2\ \text{pop75} + \eta$ with $\eta \sim_{\text{iid}} N(0,\omega^2)$

Model 2: $\log_2(\text{dpi}) = \beta_0 + \beta_1\ \text{pop15} + \beta_2\ \text{pop75} + \varepsilon$ with $\varepsilon \sim_{\text{iid}} N(0,\sigma^2)$

Model 3: $\log_2(\text{dpi}) = \gamma_0 + \gamma_1\ \text{pop15} + \gamma_2\ \text{pop75} + \gamma_3\ \text{G7} + \upsilon$ with $\upsilon \sim_{\text{iid}} N(0,\phi^2)$

Model 4: $\text{sr} = \lambda_0 + \lambda_1\ \text{pop15} + \lambda_2\ \text{pop75} + \lambda_3\ \text{dpi} + \lambda_4\ \text{ddpi} + \iota$ with $\iota \sim_{\text{iid}} N(0,\tau^2)$

--This problem set is an exam. Do not discuss it with anyone. If you discuss it with anyone, you have cheated on an exam.

--Write your name and id# on BOTH sides of the answer page.

--Write answers in the spaces provided. Brief answers suffice. Do not attach additional pages. Do not turn in computer output. Turn in only the answer page.

--If a question asks you to circle an answer, then circle an answer. If you circle the correct answer you are correct. If you circle the incorrect answer you are incorrect. If you cross out an answer, no matter which answer you cross out, you are incorrect.

--If a question has several parts, answer every part. It is common to lose points by not answering part of a question.

Name: Last, First: ____________________________ **ID#**: _____________________

Statistics 500 Fall 2006 Problem 2 Answer Page 1

This is an exam. Do not discuss it with anyone.

1. Fit model **#1** on the data page, and calculate its residuals and predicted values.

Question	CIRCLE ONE or FILL IN ANSWER
The Normal quantile plot of residuals suggests the residuals look Normal.	TRUE FALSE
Apply the Shapiro Wilk test is to the residuals. What is the p-value?	P-value = _____________________
The boxplot of residuals shows that the model fits well because the mean residual is close to zero.	TRUE FALSE
The plot of residuals vs predicted values has no fan shape suggesting the η in Model #1 have constant variance ω^2.	TRUE FALSE
Give the correlation between predicted values the **absolute values** of the residuals.	Correlation = __________________

2. Fit model **#2** on the data page, and calculate its residuals and predicted values.

Question	CIRCLE ONE or FILL IN ANSWER
The Normal quantile plot of residuals suggests the residuals look Normal.	TRUE FALSE
Apply the Shapiro Wilk test is to the residuals. What is the p-value?	P-value = _____________________
The boxplot of residuals shows that the model fits well because the mean residual is close to zero.	TRUE FALSE
The plot of residuals vs predicted values has a fan shape suggesting the ε in Model #2 do not have constant variance σ^2.	TRUE FALSE
Give the correlation between predicted values the **absolute values** of the residuals.	Correlation = _______________

Compare $\log_2$(dpi) = 7.97 for Tunisia with $\log_2$(dpi) = 11.97 for the United States, a difference. If the difference on the $\log_2$ scale were exactly 4, then the two dpi's, say dpiT and dpiUS, are related by which formula.	dpiUS = 4 dpiT dpiUS = exp(4dpiT)
	dpiUS = e^4 dpiT dpiUS = 10^4 dpiT
	dpiUS = 2^4 dpiT dpiUS = dpiT^4

Name: Last, First: ____________________________ **ID#**: _____________________

Statistics 500 Fall 2006 Problem 2 Answer Page 2

This is an exam. Do not discuss it with anyone.

3. Construct the variable G7, as described on the data page, and fit **Model #3**. This model assumes parallel planes, in the sense that the same slope is fitted for pop15 and pop75 in the G7 countries and in the remaining 43 countries. Test the null hypothesis of parallel regression planes against the alternative hypothesis that the slopes are different in G7 and other countries. Give the **name** of the test statistic, the **numerical value** of the test statistic, the **degrees of freedom** and the **p-value.**

Name of test:	Numerical value:
Degrees of freedom:	p-value:

4. Fit Model #4. The remaining questions refer to Model #4.

4a. In Model #4, which **nation** has the largest leverage or hat h_i? What is the numerical **value** of h_i for this nation?	**Nation**: | **Value**: | __________________ | ______________
4b. The nation you identified in 4a had a large value of h_i because its savings ratio sr is below the median but its ddpi is quite high – an unusual combination!	CIRCLE ONE TRUE FALSE
4c. In Model #4, our rule of thumb is to compare the leverage h_i to what **numerical value** to decide whether it is large?	**Numerical value**:
4d. In model #4, which nation has the largest **absolute** deleted or jackknife residual? What is the numerical value of this deleted residual including it sign (+-)?	**Nation**: | **Value**: | __________________ | ______________
4e. Test at the 0.05 level that the nation identified in 4d is an outlier. What is the **numerical value** of the test statistic? What is the p-value that **would have been** appropriate if you were not testing for outliers but knew in advance which nation to test? Given that you didn't know in advance which nation to test, is this nation an outlier at the 0.05 level?	Numerical value: _____________________ 'Would have been' p-value ____________ CIRCLE ONE Outlier Not an Outlier
4f. Which nation has the largest positive DFFITS? Which nation has the most negative DFFITS? What are the numerial values?	Positive Negative Nation: Nation: Value: Value:
4g. If the United States were added to Model #4 (49 states vs 50 states) the predicted value for the United States would go up by ¼ of its standard error in the 50 state fit.	CIRCLE ONE TRUE FALSE

Statistics 500 Fall 2006 Problem 2 Answer Page 1

This is an exam. Do not discuss it with anyone.

1 Fit model **#1** on the data page; calculate residuals and predicted values. **5 points each**

Question	CIRCLE ONE or FILL IN ANSWER
The Normal quantile plot of residuals suggests the residuals look Normal.	TRUE [✓ FALSE]
Apply the Shapiro Wilk test is to the residuals. What is the p-value?	P-value = 0.002. Doesn't look Normal!
The boxplot of residuals shows that the model fits well because the mean residual is close to zero.	TRUE [✓ FALSE] The mean residual is always zero – that tells you nothing about whether the model fits.
The plot of residuals vs predicted values has no fan shape suggesting the η in Model #1 have constant variance ω^2.	TRUE [✓ FALSE] Fan shape, suggesting variance is not constant.
Give the correlation between predicted values the **absolute values** of the residuals.	Correlation = 0.466, suggesting variance is not constant.

2. Fit model **#2**, and calculate its residuals and predicted values. **5 points each**

Question	CIRCLE ONE or FILL IN ANSWER
The Normal quantile plot of residuals suggests the residuals look Normal.	[✓ TRUE] FALSE
Apply the Shapiro Wilk test is to the residuals. What is the p-value?	P-value =0.755. Could be Normal.
The boxplot of residuals shows that the model fits well because the mean residual is close to zero.	TRUE [✓ FALSE] The mean residual is always zero – that tells you nothing about whether the model fits.
The plot of residuals vs predicted values has a fan shape suggesting the ε in Model #2 do not have constant variance σ^2.	TRUE [✓ FALSE] Not a fan shape.
Give the correlation between predicted values the **absolute values** of the residuals.	Correlation = -0.098, so variance looks fairly constant.

Compare $\log_2$(dpi) = 7.97 for Tunisia with $\log_2$(dpi) = 11.97 for the United States, a difference. If the difference on the $\log_2$ scale were exactly 4, then the two dpi's, say dpiT and dpiUS, are related by which formula.	dpiUS = 4 dpiT dpiUS = exp(4dpiT)
	dpiUS = e^4 dpiT dpiUS = 10^4 dpiT
	[✓ dpiUS = 2^4 dpiT] dpiUS = dpiT^4

Statistics 500 Fall 2006 Problem 2 Answer Page 2

3. Construct the variable G7, as described on the data page, and fit Model #3. This model assumes parallel planes, in the sense that the same slope is fitted for pop15 and pop75 in the G7 countries and in the remaining 43 countries. Test the null hypothesis of parallel regression planes against the alternative hypothesis that the slopes are different in G7 and other countries. Give the **name** of the test statistic, the **numerical value** of the test statistic, the **degrees of freedom** and the **p-value. 10 points total**

Name of test: F-test. Add two interaction terms to model, G7xPop15, G7xPop75. Are they needed?	Numerical value: F = 2.97
Degrees of freedom: 2 and 44	p-value: 0.061

4. Fit Model #4. The remaining questions refer to Model #4. **5 points each**

4a. In Model #4, which **nation** has the largest leverage or hat h_i? What is the numerical **value** of h_i for this nation?	**Nation**: **Value**: Libya 0.531
4b. The nation you identified in 4a had a large value of h_i because its savings ratio sr is below the median but its ddpi is quite high – an unusual combination!	TRUE [✓ FALSE] Leverage depends on X's, not on Y, but Y=sr!
4c. In Model #4, our rule of thumb is to compare the leverage h_i to what **numerical value** to decide whether it is large?	**Numerical value**: 0.2 = 2mean(h) = 2*5/50
4d. In model #4, which nation has the largest **absolute** deleted or jackknife residual? What is the numerical value of this deleted residual including it sign (+-)?	**Nation**: **Value**: Zambia 2.854
4e. Test at the 0.05 level that the nation identified in 4d is an outlier. What is the **numerical value** of the test statistic? What is the p-value that **would have been** appropriate if you were not testing for outliers but knew in advance which nation to test? Given that you didn't know in advance which nation to test, is this nation an outlier at the 0.05 level?	Numerical value: t = 2.854 'Would have been' p-value: 0.00657 CIRCLE ONE Outlier [✓ Not an Outlier] To be an outlier, the P-value had to be less than 0.05/50 = 0.001, and its not.
4f. Which nation has the largest positive DFFITS? Which nation has the most negative DFFITS? What are the numerial values?	Positive Negative Nation: Japan Nation: Libya Value: 0.86 Value: -1.16
4g. If the United States were added to Model #4 (49 states vs 50 states) the predicted value for the United States would go up by ¼ of its standard error in the 50 state fit.	CIRCLE ONE TRUE [✓ FALSE] Would go down, not up, by ¼ standard error. The sign of DFFITS talks about the impact of adding an observation to a regression which does not have it.

Doing the Problem Set in R

(Fall 2006, Problem Set 2)

This data set happens to be in R, so get it by typing:

```
> data(LifeCycleSavings)
```

Look at the first two rows:

```
> LifeCycleSavings[1:2,]
             sr pop15 pop75     dpi ddpi
Australia 11.43 29.35  2.87 2329.68 2.87
Austria   12.07 23.32  4.41 1507.99 3.93
```

If you attach the data set, the variable LifeCycleSavings$sr can be called sr.

```
> attach(LifeCycleSavings)
```

In this data set, the country names are the row names. You can make them into a variable:

```
> nations<-rownames(LifeCycleSavings)
```

Fit model #1.

```
> mod<-lm(dpi~pop15+pop75)
```

Normal quantile plot of residuals. Does not look like a line, so the residuals do not look Normal.

```
> qqnorm(mod$residual)
```

Shapiro Wilk test of the null hypothesis that the residuals are Normal. The p-value is 0.002, so there is strong evidence the residuals are not Normal. (Strictly speaking, the Shapiro Wilk test is an informal guide, not a formal test, when applied to residuals.)

```
> shapiro.test(mod$residual)

        Shapiro-Wilk normality test

data:  mod$residual
W = 0.9183, p-value = 0.002056
```

Plot residuals vs predicted. There is a fan shape, maybe also a bend.

```
> plot(mod$fitted.values,mod$residuals)
```

Correlation between absolute residuals and fitted values, to aid in spotting a fan shape.

```
> cor.test(abs(mod$residuals),mod$fitted.values)
        Pearson's product-moment correlation
      cor
0.4661606
```

Model #2 using base 2 logs.

```
> modl<-lm(log2(dpi)~pop15+pop75)
```

Fan shape is gone.

```
> plot(modl$fitted.values,modl$residuals)
>  cor.test(abs(modl$residuals),modl$fitted.values)
        Pearson's product-moment correlation
       cor
-0.0981661
```

Residuals now look plausibly Normal.

```
> qqnorm(modl$residual)
> shapiro.test(modl$residual)

        Shapiro-Wilk normality test

data:  modl$residual
W = 0.9846, p-value = 0.755
```

Create the dummy variable G7.

```
> G7<-rep(0,50)
> G7[c(6,14,15,22,23,43,44)]<-1
> G7
 [1] 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0
[26] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0

> nations[G7==1]
[1] "Canada"   "France"         "Germany"        "Italy"
[5] "Japan"    "United Kingdom" "United States"
```

Add the new variable to the data set, making it the last column.

```
> LifeCycleSavings<-cbind(LifeCycleSavings,G7)
> LifeCycleSavings[1:3,]
             sr pop15 pop75     dpi ddpi G7
Australia 11.43 29.35  2.87 2329.68 2.87  0
Austria   12.07 23.32  4.41 1507.99 3.93  0
Belgium   13.17 23.80  4.43 2108.47 3.82  0
```

Question 3: Testing whether the planes are parallel. Fit reduced and full model; compare by anova.

```
> mod<-lm(log2(dpi)~pop15+pop75+G7)
> mod2<-lm(log2(dpi)~pop15+pop75+G7+pop15*G7+pop75*G7)
> anova(mod,mod2)
Analysis of Variance Table

Model 1: log2(dpi) ~ pop15 + pop75 + G7
Model 2: log2(dpi) ~ pop15 + pop75 + G7 + pop15 * G7 +
pop75 * G7
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     46 29.4713
2     44 25.9624  2    3.5089 2.9734 0.06149
```

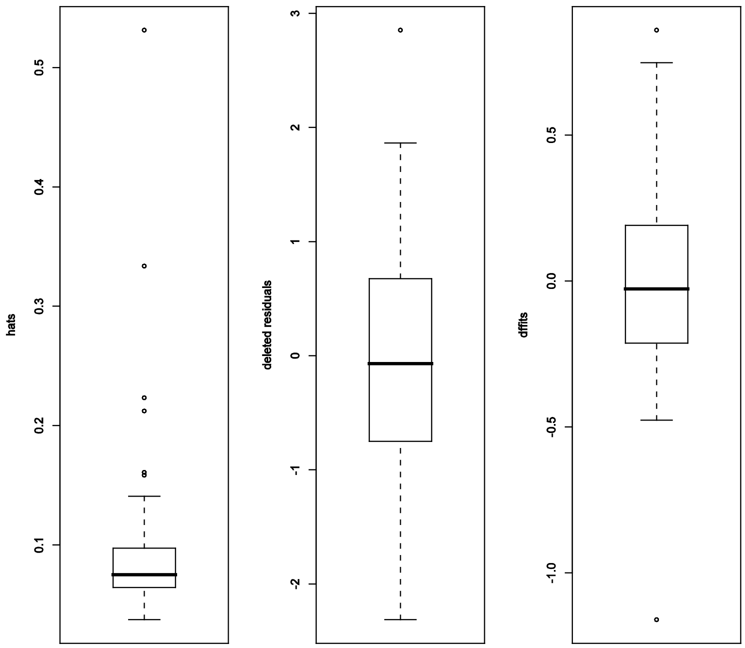
Fit and save Model #4.

```
> mod<-lm(sr~pop15+pop75+dpi+ddpi)
> mod

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi)

Coefficients:
(Intercept)        pop15        pop75          dpi         ddpi
 28.5660865   -0.4611931   -1.6914977   -0.0003369    0.4096949
```

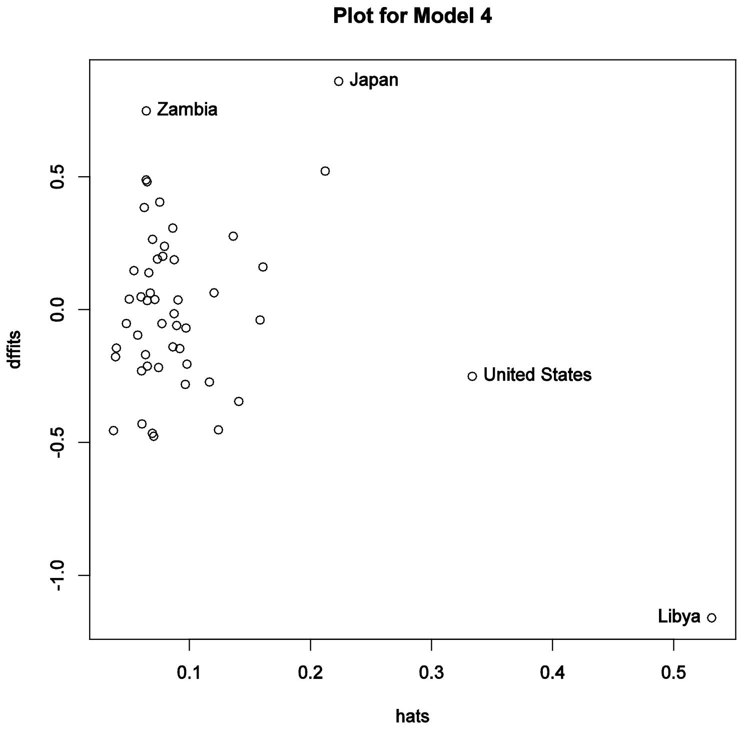
Key Step: *Compute diagnostics for Model #4.*

```
> h<-hatvalues(mod)
> deleted<-rstudent(mod)
> dft<-dffits(mod)

> summary(h)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03730 0.06427 0.07502 0.10000 0.09702 0.53150

> summary(deleted)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-2.313e+00 -7.400e-01 -6.951e-02 -4.207e-05  6.599e-01  2.854e+00

> summary(dft)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1.160000 -0.210800 -0.027100 -0.005891  0.189700  0.859700

> par(mfrow=c(1,3))
> boxplot(h,ylab="hats")
> boxplot(deleted,ylab="deleted residuals")
> boxplot(dft,ylab="dffits")
```

Boxplots of Diagnostics for Model #4



```
> plot(h,dft,ylab="dffits",xlab="hats",main="Plot for Model 4")
> identify(h,dft,labels=nations)
```



Who has the highest leverage?

```
> nations[h==max(h)]
[1] "Libya"

> max(h)
[1] 0.5314568
```

How is Libya different in terms of X's?

```
> LifeCycleSavings[nations=="Libya",]
         sr pop15 pop75    dpi  ddpi G7
Libya 8.89 43.69  2.07 123.58 16.71  0

> summary(LifeCycleSavings[,2:5])
     pop15           pop75            dpi               ddpi       
 Min.   :21.44   Min.   :0.560   Min.   :  88.94   Min.   : 0.220  
 1st Qu.:26.22   1st Qu.:1.125   1st Qu.: 288.21   1st Qu.: 2.002  
 Median :32.58   Median :2.175   Median : 695.66   Median : 3.000  
 Mean   :35.09   Mean   :2.293   Mean   :1106.76   Mean   : 3.758  
 3rd Qu.:44.06   3rd Qu.:3.325   3rd Qu.:1795.62   3rd Qu.: 4.478  
 Max.   :47.64   Max.   :4.700   Max.   :4001.89   Max.   :16.710  
```

Look at Libya's ddpi – 16.71 !

Deleted residuals and outlier tests. Use Bonferroni Inequality: 50 tests, split 0.05 into 50 parts

```
> max(deleted)
[1] 2.853558

> min(deleted)
[1] -2.313429

> nations[deleted==max(deleted)]
[1] "Zambia"
```

So Zambia is the candidate for being an outlier. There are 44 degrees of freedom, 50 observations, less 5 parameters in the model (constant and four variables) less 1 for Zambia which was deleted. Use t-distribution with 44 degrees of freedom. Want Prob(t>=2.853558) = 1-Prob(t<=2.853558) as the 1-sided p-value,

```
> 1-pt(2.853558,44)
[1] 0.003283335
```

but must double that to get the 2-sided p-value

```
> 2*(1-pt(2.853558,44))
[1] 0.006566669
```

Use Bonferroni Inequality: 50 tests, split 0.05 into 50 parts.

```
> 0.05/50
[1] 0.001
```

P-value is small, but not small enough – needed to be less than 0.05/50 = 0.001.

An equivalent way to test whether Zambia is an outlier is to add a dummy variable to the regression.

```
> nations[46]
[1] "Zambia"
> zambia<-rep(0,50)
> zambia[46]<-1

> summary(lm(sr~pop15+pop75+dpi+ddpi+zambia))
Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi + zambia)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 27.4482536  6.8434624   4.011 0.000231 ***
pop15       -0.4505547  0.1344223  -3.352 0.001658 ** 
pop75       -1.3502562  1.0137262  -1.332 0.189728    
dpi         -0.0004183  0.0008655  -0.483 0.631289    
ddpi         0.3680963  0.1828464   2.013 0.050242 .  
zambia      10.4213353  3.6520491   2.854 0.006567 ** 
---
Residual standard error: 3.533 on 44 degrees of freedom
Multiple R-Squared: 0.4418,     Adjusted R-squared: 0.3783 
F-statistic: 6.964 on 5 and 44 DF,  p-value: 7.227e-05
```

Same p-value as before, 0.006567. Use Bonferroni Inequality: 50 tests, split 0.05 into 50 parts.

```
> 0.05/50
[1] 0.001
> 0.006567<=(0.05/50)
[1] FALSE
```

P-value is small, but not small enough – needed to be less than 0.05/50 = 0.001.

The p-values you get from most regression programs are 2-sided, so you compare them to 0.05/50. When you use the t-distribution directly, you need to be careful to do a 2 sided test!

DFFITS

```
> max(dft)
[1] 0.8596508
> min(dft)
[1] -1.160133
> nations[dft==max(dft)]
[1] "Japan"
> nations[dft==min(dft)]
[1] "Libya"
> nations[44]
[1] "United States"
> dft[44]
        44 
-0.2509509
```

Goes down, not up, by ¼ of a standard error.

STATISTICS 500 FALL 2006 PROBLEM 3 DATA PAGE 1

Due Friday 15 December 2006 at 11:00 AM

This is an exam. Do not discuss it with anyone.

The data are adapted for this exam from Schoket, et al. (1991) ^{32}P-Postlabelling detection of aromatic DNA adducts in peripheral blood lymphocytes from aluminum production plant workers. *Mutation Research*, **260**, 89-98. You can obtain a copy from the library web page if you'd like, but that is not needed to do the problem set. There are four groups of 7 people. Half or 14 people worked in one of two aluminum production plants (A), which the other half were controls (C) without exposure to the production of aluminum. Half were nonsmokers (NS) and half reported smoking 20 cigarettes a day. (The original data contain many other people as well – I have selected 28 people to give a simple design.) For instance, NS_C refers to a nonsmoking control. The outcome is "Adducts." A blood sample was obtained from each person, and the DNA in lymphocytes was examined. An adduct is a molecule attached to your DNA, in this case, a polycyclic aromatic hydrocarbon (PAH), to which aluminum production workers are exposed. There are also PAHs in cigarette smoke. The unit is the number of adducts per 10^8 DNA nucleotides.

```
> ALdata
    grp Adducts Smoking Aluminum
1  NS_C    1.32       0        0
2  NS_C    1.26       0        0
3  NS_C    1.39       0        0
4  NS_C    1.38       0        0
5  NS_C    0.40       0        0
6  NS_C    1.24       0        0
7  NS_C    0.65       0        0
8  NS_A    1.15       0        1
9  NS_A    0.36       0        1
10 NS_A    0.31       0        1
11 NS_A    1.83       0        1
12 NS_A    1.34       0        1
13 NS_A    1.05       0        1
14 NS_A    1.05       0        1
15  S_C    1.20      20        0
16  S_C    0.62      20        0
17  S_C    1.30      20        0
18  S_C    1.30      20        0
19  S_C    0.79      20        0
20  S_C    2.42      20        0
21  S_C    2.03      20        0
22  S_A    2.95      20        1
23  S_A    4.66      20        1
24  S_A    2.18      20        1
25  S_A    2.32      20        1
26  S_A    0.96      20        1
27  S_A    0.81      20        1
28  S_A    2.90      20        1
```

Data are in the Rdata file, in a JMP-IN file called ALdata.jmp, and in a text file, ALdata.txt at http://stat.wharton.upenn.edu/statweb/course/Fall2006/stat500/

Tukey's method of multiple comparisons is sometimes called TukeyHSD or Tukey's honestly significant difference test, and it is one of several procedures using the studentized range.

STATISTICS 500 FALL 2006 PROBLEM 3 DATA PAGE 2

j=1 for NS_C, j=2 for NS_A, j=3 for S_C, j=4 for S_A.

Model 1: Adducts = $\zeta + \theta_j + \eta$ with $\eta \sim_{iid} N(0,\omega^2)$ with $0 = \theta_1 + \theta_2 + \theta_3 + \theta_4$

Model 2: $\log_2$(Adducts) = $\mu + \alpha_j + \varepsilon$ with $\varepsilon \sim_{iid} N(0,\sigma^2)$ with $0 = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$

Suggestion: Both JMP and R have features that help with setting up the analysis needed for question #6. These features allow groups to be coded into variables without entering numbers for each person. These features would be very helpful if there were thousands of observations and dozens of groups, but there are only 28 observations and four groups. If you find those features helpful, then use them. If you find them unhelpful, then don't use them. There are only four groups, only four means, and only 28 observations. You should not go a long distance to take a short-cut.

Turning in the exam: The registrar sets final exam dates, and even for take-homes, faculty cannot alter the exam date. For a T/Th noon class, the final exam is due Friday 15 Dec 06, 11:00am http://www.upenn.edu/registrar/pdf_main/06C_Exam_Schedule.pdf **Please make and keep a photocopy of your answer page.** You may turn in the exam early if you wish. Place your exam in a sealed envelop addressed to me, and either: (i) hand it to me in my office 473 Huntsman on the due date, or (ii) place it in my mailbox in Statistics, 4th floor of Huntsman, or (iii) leave it with the receptionist in Statistics. If you would like to receive your graded exam and an answer key by mail, please include a regular stamped, self-addressed envelop with your exam.

--This problem set is an exam. Do not discuss it with anyone. If you discuss it with anyone, you have cheated on an exam.

--Write your name and id# on BOTH sides of the answer page.

--Write answers in the spaces provided. Brief answers suffice. Do not attach additional pages. Do not turn in computer output. Turn in only the answer page.

--If a question asks you to circle an answer, then circle an answer. If you circle the correct answer you are correct. If you circle the incorrect answer you are incorrect. If you cross out an answer, no matter which answer you cross out, you are incorrect.

Name: Last, First: ____________________________ **ID#**: _____________________

Statistics 500 Fall 2006 Problem 3 Answer Page 1 (See also the Data Page)

This is an exam. Do not discuss it with anyone.

1. Do 4 parallel boxplots of Adducts for the 4 groups. Do 4 parallel boxplots of $\log_2$(Adducts) for the four groups. Use these plots to answer the following questions.

Question	**CIRCLE ONE**
Model #2 is more appropriate for these data than model #1 because of near collinearity for model #1.	TRUE FALSE
Model #2 is more appropriate for these data than model #1 because the dispersion of Adducts does not look constant over groups.	TRUE FALSE
Model #2 is more appropriate for these data than model #1 because of nested factors in model #1.	TRUE FALSE

2. Use model **#2** on the data page to test the null hypothesis H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$.

Question	**CIRCLE ONE or FILL IN ANSWER**
Name of test statistic is:	Numerical **value** of test statistic is:
p-value is:	H_0 is: (CIRCLE ONE) Plausible Not Plausible

3. Suppose you were to compare each pair of two of the four group means in a two-sided 0.05 level test using the ordinary t-test under model #2. That is, you will test the null hypothesis, H_0: $\alpha_i = \alpha_j$, for each i<j. **How many different tests** would you be doing? If it were true that $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, then what would be the **chance of a p-value less than or equal to 0.05 on each one single test** taken one at a time? If it were true that $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, then what is the **expected number of p-values** less than or equal to 0.05 in all the t-tests you are doing?

Question	**Fill in a Number**
How many different tests would you be doing?	
What is the chance of a p-value <= 0.05 on one single test? (Give a probability)	
What is the expected number of p-values <=0.05 in all the t-tests you would be doing? (enter an expected number)	

4. Use Tukey's method of multiple comparisons with experimentwise error rate of 0.05.

Under model #2, by Tukey's method:	**CIRCLE ONE**
H_0: $\alpha_{S\_C} = \alpha_{NS\_C}$ is rejected	TRUE FALSE
H_0: $\alpha_{S\_A} = \alpha_{NS\_A}$ is rejected	TRUE FALSE
H_0: $\alpha_{S\_A} = \alpha_{S\_C}$ is rejected	TRUE FALSE

Name: Last, First: ______________________________ **ID#**: ______________________

Statistics 500 Fall 2006 Problem 3 Answer Page 2 (See also the Data Page)

This is an exam. Do not discuss it with anyone.

5. Propose three **orthogonal** contrasts with the stated interpretations. Using integer weights. In other words, in each of 12 = 3 x 4 cells, place a positive or negative integer.

Group **Interpretation**	**NS_C**	**NS_A**	**S_C,**	**S_A**
a. Smoking versus nonsmoking				
b. Aluminum plant worker vs control				
c. Difference between aluminum plant workers and controls is different for smokers and for nonsmokers				

Show that contrasts b. and c. in your table are orthogonal:

---

6. Use your contrasts in question 5 to fill in the following ANOVA table for Model #2.

Source of variation	Sum of squares	Degrees of Freedom	Mean Square	F-statistic
Between Groups				
a. Smoking vs Nonsmoking				
b. Aluminum worker vs control				
c. Interaction				
Within Groups				XXXXXX XXXXXX

7. In model #2, test the null hypothesis: H_0: $0 = (\alpha_{S\_A} + \alpha_{S\_C}) - (\alpha_{NS\_A} + \alpha_{NS\_C})$ against a two-sided alternative. What is the name of the test? What is the numerical value of the test statistic? What is the p-value? Is the null hypothesis plausible?

Name of test:	Value of test statistic:
P-value:	CIRCLE ONE Plausible Not Plausible
Smokers tend to have more adducts than nonsmokers.	TRUE FALSE

Statistics 500 Fall 2006 Problem 3 Answer Page 1

This is an exam. Do not discuss it with anyone.

1. Do 4 parallel boxplots of Adducts for the 4 groups. Do 4 parallel boxplots of $\log_2$(Adducts) for the four groups. Use these plots to answer the following questions.

Question (10 points)	**CIRCLE ONE**
Model #2 is more appropriate for these data than model #1 because of near collinearity for model #1.	TRUE [✓ FALSE] Collinearity depends on x, not y, so it is the same for log(y)
Model #2 is more appropriate for these data than model #1 because the dispersion of Adducts does not look constant over groups.	[✓ TRUE] FALSE
Model #2 is more appropriate for these data than model #1 because of nested factors in model #1.	TRUE [✓ FALSE] There are no nested factors here.

2. Use model **#2** on the data page to test the null hypothesis H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$.

Question (15 points)	**CIRCLE ONE or FILL IN ANSWER**
Name of test statistic is: F-test	Numerical **value** of test statistic is: F = 3.08 on 3 and 24 degrees of freedom
p-value is: 0.0465	H_0 is: (CIRCLE ONE) Plausible [✓ Not Plausible]

3. Suppose you were to compare each pair of two of the four group means in a two-sided 0.05 level test using the ordinary t-test under model #2. That is, you will test the null hypothesis, H_0: $\alpha_i = \alpha_j$, for each i<j. **How many different tests** would you be doing? If it were true that $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, then what would be the **chance of a p-value less than or equal to 0.05 on each one single test** taken one at a time? If it were true that $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, then what is the **expected number of p-values** less than or equal to 0.05 in all the t-tests you are doing?

Question (15 points)	**Fill in a Number**
How many different tests would you be doing?	6 tests
What is the chance of a p-value <= 0.05 on one single test? (Give a probability)	On one test, 0.05
What is the expected number of p-values <=0.05 in all the t-tests you would be doing? (enter an expected number)	With 6 tests, 0.05 x 6 = 0.30 which is much more than 0.05

4. Use Tukey's method of multiple comparisons with experimentwise error rate of 0.05.

Under model #2, by Tukey's method:	**CIRCLE ONE** (15 points)
H_0: $\alpha_{S\_C} = \alpha_{NS\_C}$ is rejected	TRUE [✓ FALSE]
H_0: $\alpha_{S\_A} = \alpha_{NS\_A}$ is rejected	[✓ TRUE] FALSE
H_0: $\alpha_{S\_A} = \alpha_{S\_C}$ is rejected	TRUE [✓ FALSE]

Statistics 500 Fall 2006 Problem 3 Answer Page 2 (See also the Data Page)

This is an exam. Do not discuss it with anyone.

5. Propose three **orthogonal** contrasts with the stated interpretations. Using integer weights. In other words, in each of 12 = 3 x 4 cells, place a positive or negative integer.

(15 points) **Group** **Interpretation**	**NS_C**	**NS_A**	**S_C,**	**S_A**
a. Smoking versus nonsmoking	-1	-1	1	1
b. Aluminum plant worker vs control	-1	1	-1	1
c. Difference between aluminum plant workers and controls is different for smokers and for nonsmokers	1	-1	-1	1

Show that contrasts b. and c. in your table are orthogonal:

$$(-1x1) + (1x-1) + (-1x-1) + (1x1) = -1 + -1 + 1 + 1 = 0$$

6. Use your contrasts in question 5 to fill in the following ANOVA table for Model #2.

Source of variation (15 points)	Sum of squares	Degrees of Freedom	Mean Square	F-statistic
Between Groups	6.3388	3	2.1129	3.08
a. Smoking vs Nonsmoking	4.3734	1	4.3734	6.38
b. Aluminum worker vs control	0.4222	1	0.4222	0.62
c. Interaction	1.5433	1	1.5433	2.25
Within Groups	16.4570	24	0.6857	XXXXXX XXXXXX

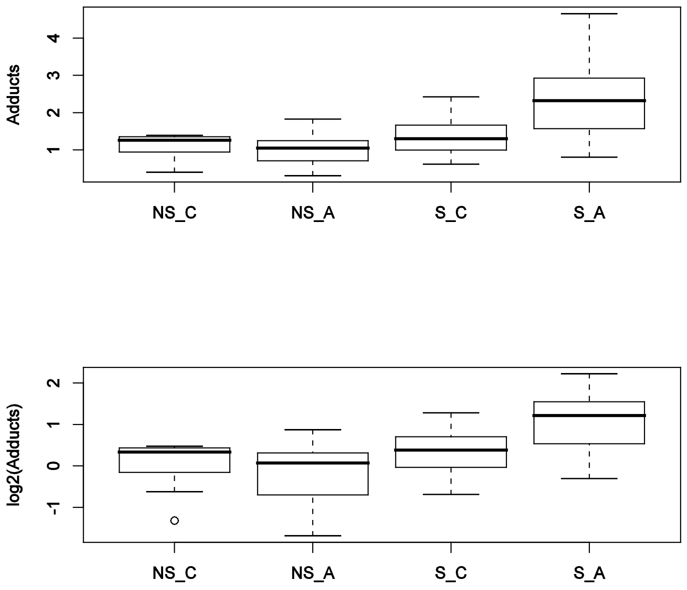
7. In model #2, test the null hypothesis: H_0: $0 = (\alpha_{S\_A} + \alpha_{S\_C}) - (\alpha_{NS\_A} + \alpha_{NS\_C})$ against a two-sided alternative. What is the name of the test? What is the numerical value of the test statistic? What is the p-value? Is the null hypothesis plausible? (15 points)

Name of test: F-test or t-test (F=t² with 1 df)	Value of test statistic: F = 6.38 from above
P-value: 0.0186	CIRCLE ONE Plausible [✓ Not Plausible]
Smokers tend to have more adducts than nonsmokers.	[✓ TRUE] FALSE

Analysis Done in R

Question 1:

```
> attach(ALdata)
> par(mfrow=c(2,1))
> boxplot(Adducts~grp,ylab="Adducts")
> boxplot(log2(Adducts)~grp,ylab="log2(Adducts)")
```



For adducts, the S_A boxplot is much more dispersed than the others. For log$_2$(Adducts), there is not a very definite pattern.

Question 2:

```
> summary(aov(log2(Adducts)~grp))
            Df  Sum Sq Mean Sq F value  Pr(>F)
grp          3  6.3388  2.1129  3.0814 0.04652 *
Residuals   24 16.4570  0.6857
---
```

Question 4:

```
> TukeyHSD(aov(log2(Adducts)~grp))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = log2(Adducts) ~ grp)
$grp
                diff         lwr       upr
NS_A-NS_C -0.2239629 -1.44499131 0.9970655
S_C-NS_C   0.3208815 -0.90014688 1.5419099
S_A-NS_C   1.0360018 -0.18502663 2.2570302
S_C-NS_A   0.5448444 -0.67618398 1.7658728
S_A-NS_A   1.2599647  0.03893628 2.4809931
S_A-S_C    0.7151203 -0.50590815 1.9361487
```

You can code up the contrast variables in any of several ways. With a small problem, like this one, you could enter the numbers by hand. Or you could create the contrast variables yourself. Here, I show you the "standard way" in R, which is to give the factor "grp" a set of contrasts, then have R create the contrast matrix. It is heavy handed for such a tiny problem, but in large problems, it is convenient.

A factor, here "grp", starts with dummy coding of categories. Here it codes the four groups in three variables, leaving out the first group. The remainder of this page just redefines contrasts(grp) to be the contrasts we want.

```
> contrasts(grp)
     2 3 4
NS_C 0 0 0
NS_A 1 0 0
S_C  0 1 0
S_A  0 0 1
```

Here are my three contrasts, which I defined as in smoke <- c(-1,-1,1,1), etc.

```
> smoke
[1] -1 -1  1  1
> alum
[1] -1  1 -1  1
> inter
[1]  1 -1 -1  1
```

Make them into a matrix or table by binding them as columns.

```
> m<-cbind(smoke,alum,inter)
> m
     smoke alum inter
[1,]    -1   -1     1
[2,]    -1    1    -1
[3,]     1   -1    -1
[4,]     1    1     1
```

Now the magic step: set the contrasts for "grp" equal to the matrix you just made.

```
> contrasts(grp)<-m
```

What this does is associate those contrasts with this factor forever, or until you specify different contrasts. It is a bit heavy handed for our small, one-time analysis. We now look at the contrasts.

```
> contrasts(grp)
     smoke alum inter
NS_C    -1   -1     1
NS_A    -1    1    -1
S_C      1   -1    -1
S_A      1    1     1
```

So all that has happened on this page is contrasts(grp) has be redefined to be our contrasts, not the dummy coded contrasts.

The model.matrix command extends the contrasts we just defined into variables for our 28 people. Again, in our small problem, you have to wonder whether this "short cut" was the long way around. In bigger problems, with more groups, people and variables, the short cut is helpful. If you had built this matrix "by hand", the analysis would be the same. Or you could use the formulas in the book applied to the four group means.

```
v<-model.matrix(log2(Adducts)~grp)
> v
   (Intercept) grpsmoke grpalum grpinter
1            1       -1      -1        1
2            1       -1      -1        1
3            1       -1      -1        1
4            1       -1      -1        1
5            1       -1      -1        1
6            1       -1      -1        1
7            1       -1      -1        1
8            1       -1       1       -1
9            1       -1       1       -1
10           1       -1       1       -1
11           1       -1       1       -1
12           1       -1       1       -1
13           1       -1       1       -1
14           1       -1       1       -1
15           1        1      -1       -1
16           1        1      -1       -1
17           1        1      -1       -1
18           1        1      -1       -1
19           1        1      -1       -1
20           1        1      -1       -1
21           1        1      -1       -1
22           1        1       1        1
23           1        1       1        1
24           1        1       1        1
25           1        1       1        1
26           1        1       1        1
27           1        1       1        1
28           1        1       1        1

> smk<-v[,2]
> al<-v[,3]
> int<-v[,4]
```

At long last, question 6 and 7:

```
>  anova(lm(log2(Adducts)~smk+al+int))
Analysis of Variance Table

Response: log2(Adducts)
          Df  Sum Sq Mean Sq F value  Pr(>F)
smk        1  4.3734  4.3734  6.3779 0.01857 *
al         1  0.4222  0.4222  0.6157 0.44034
int        1  1.5433  1.5433  2.2506 0.14660
Residuals 24 16.4570  0.6857
```

PROBLEM SET #1 STATISTICS 500 FALL 2007: DATA PAGE 1

Due in class Thusday 25 Oct 2007

This is an exam. Do not discuss it with anyone.

To learn about the dataset, type:

```
> help(BostonHousing2,package=mlbench)
BostonHousing            package:mlbench            R Documentation
     Housing data for 506 census tracts of Boston from the 1970 census.
     The dataframe 'BostonHousing' contains the original data by
     Harrison and Rubinfeld (1979), the dataframe 'BostonHousing2' the
     corrected version with additional spatial information (see
     references below).
Usage:
     data(BostonHousing)
     data(BostonHousing2)
Format:
     The original data are 506 observations on 14 variables, 'medv'
     being the target variable:
       crim     per capita crime rate by town
       zn       proportion of residential land zoned for lots over
                 25,000 sq.ft
       indus    proportion of non-retail business acres per town
       chas     Charles River dummy variable (= 1 if tract bounds
                 river; 0 otherwise)
       nox      nitric oxides concentration (parts per 10 million)
       rm       average number of rooms per dwelling
       age      proportion of owner-occupied units built prior to 1940
       dis      weighted distances to five Boston employment centres
       rad      index of accessibility to radial highways
       tax      full-value property-tax rate per USD 10,000
       ptratio  pupil-teacher ratio by town
       b        1000(B - 0.63)^2 where B is the proportion of blacks by
                 town
       lstat    percentage of lower status of the population
       medv     median value of owner-occupied homes in USD 1000's
     The corrected data set has the following additional columns:
       cmedv  corrected median value of owner-occupied homes in USD
              1000's
       town   name of town
       tract  census tract
       lon    longitude of census tract
       lat    latitude of census tract
References:
     Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the
     demand for clean air.  Journal of Environmental Economics and
     Management, 5, 81-102.
```

PROBLEM SET #1 STATISTICS 500 FALL 2007: DATA PAGE 2

To obtain the data, you can do one of several things:

Get it directly:

Go to the "packages" menu in R, click "load package" and click "mlbench" and type:

```
> library(mlbench)
> data(BostonHousing2)
```

Notice that you want BostonHousing2, NOT BostonHousing. You may wish to attach the data:

```
> attach(BostonHousing2)
```

The data are also in the latest version of Rst500.RData and in an Excel file Bostonhousing2.xls at::
http://stat.wharton.upenn.edu/statweb/course/Fall-2007/STAT-500/
or http://download.wharton.upenn.edu/download/pub/stat/Fall-2007/
and Rst500.RData is also on my web page:
http://www-stat.wharton.upenn.edu/~rosenbap/index.html
To obtain a Wharton username or password for course use, apply at:
http://apps.wharton.upenn.edu/accounts/class/
Use cmedv, **not** medv; here, cmedv contains the corrected values.

Model #1

cmedv $= \beta_0 + \beta_1$ nox $+ \varepsilon$ with ε iid $N(0,\sigma^2)$

Model #2

cmedv $= \gamma_0 + \gamma_1$ nox $+ \gamma_2$ crim $+ \gamma_3$ rm $+ \zeta$ with ζ iid $N(0,\omega^2)$

Follow instructions. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam.

Name: _____________________________ ID# _________________________

PROBLEM SET #1 STATISTICS 500 FALL 2007: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

1. Fit model #1 on the data page, and use the fit to answer the following questions.

Questions refer to Model #1	Answer
1.1 What is the name of the town containing the Census tract with the highest level of nox?	Name of town: _____________________
1.2 What is the full name of the town containing the Census tract with the lowest cmedv?	Name of town: _____________________
1.3 What is the least squares estimate of β_1? (Give the numerical value.)	Estimate: _______________
1.4 If you were to use model #1 to predict cmedv, and you were to compare predictions for two tracts, one with nox = .5 and the other with nox = .7, **how much higher** would the predicted value (in dollars) be for nox = .5?	Point estimate of **difference** in $. (Be careful with the sign and the units.) ______________________________
1.5 Give the 95% confidence interval for β_1 assuming model 1 is true. (Give two numbers, low endpoint, high endpoint.)	95% Interval: [,]
1.6 Test the null hypothesis, H_0: β_1=0. What is the **name** of the test statistic? What is the **value** of the test statistic? What is the two-sided **p-value**? Is H_0: β_1=0 **plausible**?	Name: __________ Value: _________ p-value: __________ CIRCLE ONE H_0 is Plausible Not Plausible
1.7 What is the unbiased estimate of σ^2 under model #1? What is the corresponding estimate of σ? What are the units (feet, pounds, whatever) for the estimate of σ?	Estimate of: σ^2 ___________ σ __________ Units:_________________________

2. Calculate the residuals and fitted values from model #1; base your answers on them.

Plot residuals vs fitted, Normal plot, boxplot, Shapiro.test	CIRCLE ONE
2.1 The 16 residuals for the highest pollution, nox==.8710, are all positive residuals.	TRUE FALSE
2.2 The residuals appear to be skewed to the right.	TRUE FALSE
2.3 The Shapiro-Wilk test suggests the residuals are not Normal	TRUE FALSE
2.4 The Normal quantile plot suggests the residuals are not Normal	TRUE FALSE

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2007: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

3. Fit model #2 on the data page and use it to answer the following questions.

Question	Answer
3.1 What is the least squares point estimate of the coefficient of nox, γ_1, in model #2? What is the least squares point estimate of the coefficient of nox, β_1, in model #1?	Estimate of: γ_1 ___________ β_1 ___________
3.2 Test the null hypothesis, H_0:γ_1=γ_2=γ_3=0 under model #2. What is the **name** of the test statistic? What is the **value** of the test statistic? What is the **p-value**? Is H_0 **plausible**?	Name: __________ Value: _________ p-value: __________ CIRCLE ONE H_0 is Plausible Not Plausible
3.3 What is the square of the correlation between observed and fitted `cmedv` in model #2? What is the square of the correlation between observed and fitted `cmedv` in model #1?	In model #2: In model #1:
3.4 What is the (ordinary Pearson) correlation between nox and crim? Does this correlation provide an adequate basis to assert that: (i) pollution causes crime or (ii) crime causes pollution?	Correlation: __________ CIRCLE ONE (i) Adequate basis Other (ii) Adequate basis Other
3.5 For Model 2, the plot of residuals against fitted values exhibits a pattern suggesting that a linear model is not an adequate fit.	CIRCLE ONE TRUE FALSE
3.6 The residuals do not look Normal.	CIRCLE ONE TRUE FALSE

4. In Model #2, test H_0:γ_2=γ_3=0. Which **variables** are in the full model? What is its **residual sum of squares** (RSS)? Which **variables** are in the reduced model? What is its **residual sum of squares** (RSS)? Give the numerical values of the **mean squares** in the numerator and denominator of the F ratio for testing H_0. What is the numerical value of **F**? What is the **p-value**? Is the null hypothesis **plausible**?

Full Model	Variables:	RSS:
Reduced Model	Variables:	RSS:
Numerator and denominator of F	Numerator=	Denominator:=
F= _____________	p-value= _____________________	CIRCLE ONE: Plausible Not Plausible

PROBLEM SET #1 STATISTICS 500 FALL 2007: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

1. Fit model #1 on the data page, and use the fit to answer the following questions.

Questions refer to Model #1	Answer (5 points per part)
1.1 What is the name of the town containing the Census tract with the highest level of nox?	Name of town: Cambridge
1.2 What is the full name of the town containing the Census tract with the lowest cmedv?	Name of town: Boston South Boston
1.3 What is the least squares estimate of β_1? (Give the numerical value.)	Estimate: -34.02
1.4 If you were to use model #1 to predict cmedv, and you were to compare predictions for two tracts, one with nox = .5 and the other with nox = .7, how much higher would the predicted value (in dollars) be for nox = .5?	Point estimate of difference in $\$$. (Be careful with the sign and the units.) $-34.02 \times (.5-.7) \times \$1000 = \$6,804$
1.5 Give the 95% confidence interval for β_1 assuming model 1 is true. (Give two numbers, low endpoint, high endpoint.)	95% Interval: [-40.3 , -27.8]
1.6 Test the null hypothesis, H_0: $\beta_1=0$. What is the **name** of the test statistic? What is the **value** of the test statistic? What is the two-sided p-value? Is H_0: $\beta_1=0$ plausible?	Name: t-test Value: $t = -10.67$ p-value: 2×10^{-16} CIRCLE ONE H_0 is Plausible [✓ Not Plausible]
1.7 What is the unbiased estimate of σ^2 under model #1? What is the corresponding estimate of σ? What are the units (feet, pounds, whatever) for the estimate of σ?	Estimate of: σ^2 $68.9 = 8.301^2$ σ 8.301 Units: $\$1,000$

2. Calculate the residuals and fitted values from model #1; base your answers on them.

Plot residuals vs fitted, Normal plot, boxplot, Shapiro.test	CIRCLE ONE (5 pts each)
2.1 The 16 residuals for the highest pollution, nox==.8710, are all positive residuals.	[✓ TRUE] FALSE
2.2 The residuals appear to be skewed to the right.	[✓ TRUE] FALSE
2.3 The Shapiro-Wilk test suggests the residuals are not Normal	[✓ TRUE] FALSE
2.4 The Normal quantile plot suggests the residuals are not Normal	[✓ TRUE] FALSE

PROBLEM SET #1 STATISTICS 500 FALL 2007: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

3. Fit model #2 on the data page and use it to answer the following questions.

Question	Answer (5 points each part)
3.1 What is the least squares point estimate of the coefficient of nox, γ_1, in model #2? What is the least squares point estimate of the coefficient of nox, β_1, in model #1?	Estimate of: γ_1 -13.3 β_1 -34.02
3.2 Test the null hypothesis, H_0:γ_1=γ_2=γ_3=0 under model #2. What is the **name** of the test statistic? What is the **value** of the test statistic? What is the p-value? Is H_0 plausible?	Name: F-test Value: 218 on 3 & 502 df p-value: 2×10^{-16} CIRCLE ONE H_0 is Plausible [✓ Not Plausible]
3.3 What is the square of the correlation between observed and fitted cmedv in model #2? What is the square of the correlation between observed and fitted cmedv in model #1?	In model #2: R^2 = 0.57 = 57% In model #1: R^2 = 0.18 = 18%
3.4 What is the (ordinary Pearson) correlation between nox and crim? Does this correlation provide an adequate basis to assert that: (i) pollution causes crime or (ii) crime causes pollution?	Correlation: 0.42 CIRCLE ONE (i) Adequate basis [✓ Other] (ii) Adequate basis [✓ Other]
3.5 For Model 2, the plot of residuals against fitted values exhibits a pattern suggesting that a linear model is not an adequate fit.	CIRCLE ONE [✓ TRUE] FALSE
3.6 The residuals do not look Normal.	CIRCLE ONE [✓ TRUE] FALSE

4. In Model #2, test H_0:γ_2=γ_3=0. Which **variables** are in the full model? What is its **residual sum of squares** (RSS)? Which **variables** are in the reduced model? What is its **residual sum of squares** (RSS)? Give the numerical values of the **mean squares** in the numerator and denominator of the F ratio for testing H_0. What is the numerical value of **F value**? What is the **p-value**? Is the null hypothesis **plausible**? (15 points)

Full Model	Variables: nox, crim, rm	RSS: 18,488
Reduced Model	Variables: nox	RSS: 34,731
Numerator and denominator of F	Numerator= 8,121	Denominator:= 36.83
F= 220.5	p-value= 2.2×10^{-16}	CIRCLE ONE: Plausible [✓ Not Plausible]

Doing the Problem Set in R

PROBLEM SET #1 STATISTICS 500 FALL 2007

```
> attach(BostonHousing2)
```

What is the first thing we do with data?

```
> pairs(cbind(cmedv,crim,nox,rm))
> boxplot(cmedv)
> boxplot(crim)
> boxplot(nox)
> boxplot(rm)
```

Question 1. Fit model #1.

```
> summary(lm(cmedv~nox))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   41.398      1.806   22.92   <2e-16 ***
nox          -34.018      3.188  -10.67   <2e-16 ***
Residual standard error: 8.301 on 504 degrees of freedom
Multiple R-Squared: 0.1843,     Adjusted R-squared: 0.1827
F-statistic: 113.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

Question 1.4

```
> -34.018*(.5-.7)
[1] 6.8036
```

Question 1.5. It is simple arithmetic to get the 95% CI; see Kleinbaum, et al (2008) section 5.7. I got tired of the arithmetic and wrote a little R-function to do it once and for all. All it does is the formula from Kleinbaum, et al. In the function, qt looks up the quantile (crit) in the t-table, and then it is just estimate +/- crit x standard error. You can do this by hand or use lmci, which is in Rst500.RData.

```
> lmci
function(mod){
co<-(summary(mod))$coefficients
k<-dim(co)[1]
out<-matrix(NA,k,2)
rownames(out)<-rownames(co)
crit<-qt(.975,(summary(mod))$df[2])
out[,2]<-co[,1]+crit*co[,2]
out[,1]<-co[,1]-crit*co[,2]
colnames(out)<-c("low","high")
out
}
> lmci(lm(cmedv~nox))
                  low      high
(Intercept)  37.84944  44.94734
nox         -40.28094 -27.75477
```

Doing the Problem Set in R: PROBLEM SET #1 STATISTICS 500 FALL 2007

Question 2: Residual Analysis.

```
> plot(lm(cmedv~nox)$fitted.values,lm(cmedv~nox)$residual)
```

Question 2.1: The highest nox values are all in Cambridge, and all the residuals are positive – higher cost than you would expect given nox.

```
> lm(cmedv~nox)$residual[nox==.8710]
```

Question 2.2: Both the boxplot and the qq-plot suggest the residuals are skewed right. A skewed distribution is not Normal. The Shapiro-Wilk test rejects the null hypothesis that the residuals are Normal.

```
> boxplot(lm(cmedv~nox)$residual)
> qqnorm(lm(cmedv~nox)$residual)
> shapiro.test(lm(cmedv~nox)$residual)
        Shapiro-Wilk normality test
data:  lm(cmedv ~ nox)$residual
W = 0.856, p-value < 2.2e-16
```

Question 3.

```
> summary(lm(cmedv~nox+crim+rm))
Call:
lm(formula = cmedv ~ nox + crim + rm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.02075    3.23054  -5.888 7.17e-09 ***
nox         -13.32759    2.64473  -5.039 6.53e-07 ***
crim         -0.19878    0.03481  -5.710 1.93e-08 ***
rm            7.90192    0.40551  19.486  < 2e-16 ***
Residual standard error: 6.069 on 502 degrees of freedom
Multiple R-Squared: 0.5658,     Adjusted R-squared: 0.5632
F-statistic:   218 on 3 and 502 DF,  p-value: < 2.2e-16
> res<-lm(cmedv~nox+crim+rm)$residual
> fit<-lm(cmedv~nox+crim+rm)$fitted.values
> boxplot(res)
> qqnorm(res)
> plot(fit,res)
> lines(lowess(fit,res))
> shapiro.test(res)
        Shapiro-Wilk normality test
data:  res
W = 0.8788, p-value < 2.2e-16
> cor(nox,crim)
[1] 0.4209717
```

Question 4.

```
> anova(lm(cmedv~nox),lm(cmedv~nox+crim+rm))
Analysis of Variance Table

Model 1: cmedv ~ nox
Model 2: cmedv ~ nox + crim + rm
  Res.Df   RSS  Df Sum of Sq      F    Pr(>F)
1    504 34731
2    502 18488   2     16242 220.50 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F = (16242/2)/(18488/502) = 8,121/36.83=220.5

PROBLEM SET #2 STATISTICS 500 FALL 2007: DATA PAGE 1

Due in class Tuesday 4 December 2007

This is an exam. Do not discuss it with anyone.

Same data set as Problem #1. To learn about the dataset, type:

```
> help(BostonHousing2,package=mlbench)
BostonHousing            package:mlbench            R Documentation
     Housing data for 506 census tracts of Boston from the 1970 census.
     The dataframe 'BostonHousing' contains the original data by
     Harrison and Rubinfeld (1979), the dataframe 'BostonHousing2' the
     corrected version with additional spatial information (see
     references below).
Usage:
     data(BostonHousing)
     data(BostonHousing2)
Format:
     The original data are 506 observations on 14 variables, 'medv'
     being the target variable:
       crim     per capita crime rate by town
       zn       proportion of residential land zoned for lots over
                 25,000 sq.ft
       indus    proportion of non-retail business acres per town
       chas     Charles River dummy variable (= 1 if tract bounds
                 river; 0 otherwise)
       nox      nitric oxides concentration (parts per 10 million)
       rm       average number of rooms per dwelling
       age      proportion of owner-occupied units built prior to 1940
       dis      weighted distances to five Boston employment centres
       rad      index of accessibility to radial highways
       tax      full-value property-tax rate per USD 10,000
       ptratio  pupil-teacher ratio by town
       b        1000(B - 0.63)^2 where B is the proportion of blacks by
                 town
       lstat    percentage of lower status of the population
       medv     median value of owner-occupied homes in USD 1000's
     The corrected data set has the following additional columns:
       cmedv  corrected median value of owner-occupied homes in USD
              1000's
       town   name of town
       tract  census tract
       lon    longitude of census tract
       lat    latitude of census tract
References:
     Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the
     demand for clean air.  Journal of Environmental Economics and
     Management, 5, 81-102.
```

PROBLEM SET #1 STATISTICS 500 FALL 2007: DATA PAGE 2

To obtain the data, you can do one of several things:

Get it directly:

Go to the "packages" menu in R, click "load package" and click "mlbench" and type:

```
> library(mlbench)
> data(BostonHousing2)
```

Notice that you want BostonHousing2, **NOT** BostonHousing. You may wish to attach the data:

```
> attach(BostonHousing2)
```

The data are also in the latest version of Rst500.RData and in an Excel file Bostonhousing2.xls at::
http://stat.wharton.upenn.edu/statweb/course/Fall-2007/STAT-500/
or http://download.wharton.upenn.edu/download/pub/stat/Fall-2007/
and Rst500.RData is also on my web page:
http://www-stat.wharton.upenn.edu/~rosenbap/index.html
To obtain a Wharton username or password for course use, apply at:
http://apps.wharton.upenn.edu/accounts/class/
Use cmedv, **not** medv; here, cmedv contains the corrected values.

Do the following plots. You may wish to enhance the plots using as `plot(x,y)` `lines(lowess(x,y))`
You should describe a plot a clearly bent if the lowess fit shows a clear bend that departs from a straight line. You should describe a plot as fairly straight if the lowess fit looks more or less straight, with either just very slight bends, or wiggles without clear pattern.

Plot A y=cmedv vs x=crim. Plot B y=log(cmedv) vs x=crim, Plot C y=log(cmedv) vs $(crim)^{1/3}$
Plot D y=cmedv vs log(crim)

Model #1

cmedv $= \beta_0 + \beta_1$ nox $+ \beta_2$ log(crim) $+ \beta_3$ rm $+ \beta_4$ ptratio $+ \beta_5$ chas $+ \varepsilon$
with ε iid $N(0,\sigma^2)$

Let rm2c be centered squared rm, `rm2c <- (rm-mean(rm))^2`

Model #2

cmedv $= \gamma_0 + \gamma_1$ nox $+ \gamma_2$ log(crim) $+ \gamma_3$ rm $+ \gamma_4$ rm2c $+ \gamma_5$ ptratio $+ \gamma_6$ chas $+ \zeta$
with ζ iid $N(0,\omega^2)$

Follow instructions. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam.

Name: ______________________________ ID# __________________________

PROBLEM SET #2 STATISTICS 500 FALL 2007: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

1. See instructions on data page.	CIRCLE ONE
Plot A y=cmedv versus x=crim	Fairly Straight Clearly Bent
Plot B y=log(cmedv) versus x=crim	Fairly Straight Clearly Bent
Plot C y=log(cmedv) versus $(\text{crim})^{1/3}$	Fairly Straight Clearly Bent
Plot D y=cmedv versus log(crim)	Fairly Straight Clearly Bent
If you want to straighten a plot using the log tranformation, you should try several bases, such as $\log_e(y)$, $\log_{10}(y)$, $\log_2(y)$,because one base may do a better job of straightening than another.	True False
The transformation $(y^p-1)/p$ has the advantage over y^p in that the former, but not the latter, can take the reciprocal 1/y by letting p tend to infinity.	True False
If $\log_2(y)- \log_2(x) = 4$, then you must double x four times to get y, that is, $y = 2^4 x = 16x$.	True False

2. Fit model #1 on data page and assume it is true for the purpose of question 2.

	CIRCLE ONE
2.1 The hypothesis H_0: $\beta_2=0$ is rejected in a conventional two-sided 0.05 level test, where variable 2 is log(crim).	TRUE FALSE
2.2 Which observation has the largest absolute studentized = jackknife = deleted residual? Give the **row #**, the **name** of the town, and the **value** of the studentized residual.	Row # ________ Value: _______ Name:
2.3 Continuing question 2.2, test the null hypothesis that there are no outliers in model #1. What is the **numerical value** of the test statistic? What is the **two-sided p-value corrected for testing multiple hypotheses**? What are the **degrees of freedom**? **How many hypotheses** were tested in testing for outliers? Is it **plausible** that there are no outliers?	Value: ________ P-value: __________ Degrees of Freedom: _________ How many: ______ Plausible Not Plausible

Name: ______________________________ ID# __________________________

PROBLEM SET #2 STATISTICS 500 FALL 2007: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

3.**Continue to use model #1 on data page.**	CIRCLE ONE
3.1 What is the **numerical value** of the largest leverage or hatvalue? What is the **row number** for this observation? What is the **name of this town**? In this data set, what is the **numerical cut off** for judging a leverage value to be large?	Value: _________ Row # ___________ Name of town: Numerical cut off:
3.2 In 3.1, the town has large leverage because cmedv is close to the median while rm is the maximum, but in model #1, more rooms usually raised the value of the home, leading to a negative residual.	TRUE FALSE
3.3 What is the **numerical value** of the largest absolute dffits? What is the **row number** for this observation? What is the **name of this town**? If one added this observation to the regression based on the 505 other observations, the fitted value would rise by more than one standard error.	Value: ________ Row # ___________ **Name** of town: TRUE FALSE
3.4 In question 3.3, the dffits was large because the homes were expensive (cmedv) despite fairly high pollution (nox), fairly high ptratio, and below median number of rooms (rm).	TRUE FALSE

4. Fit **model 2** on the data page, assume it to be true, and use it to answer question 4.

	CIRCLE ONE
4.1 Houses in tracts along the Charles River are estimated to be valued at 3.841 times more than houses in similar tracts not along the Charles River.	TRUE FALSE
4.2 The relationship between median home values (cmedv) and the average number of rooms is shaped like a hill, an upside down U, so tracts with middle values of rm have the highest median values.	TRUE FALSE
4.3 Test the hypothesis that the relationship is between cmedv and rm is linear, not quadratic, that is, test H_0: γ_4=0. What is the **name** of the test statistic? What is the **value** of the test statistic? Is the null hypothesis **plausible**?	Name: _________ Value:_________ Plausible Not plausible
4.4 If question 2.1 were asked about γ_2 in model #2, what would the answer be?	TRUE FALSE

PROBLEM SET #2 STATISTICS 500 FALL 2007: ANSWER PAGE 1

1. See the data page for instructions.	CIRCLE ONE (5 points each)
Plot A y=cmedv versus x=crim	Fairly Straight [✓ Clearly Bent]
Plot B y=log(cmedv) versus x=crim	Fairly Straight [✓ Clearly Bent]
Plot C y=log(cmedv) versus $(\text{crim})^{1/3}$	[✓ Fairly Straight] Clearly Bent
Plot D y=cmedv versus log(crim)	[✓ Fairly Straight] Clearly Bent
If you want to straighten a plot using the log tranformation, you should try several bases, such as $\log_e(y)$, $\log_{10}(y)$, $\log_2(y)$,because one base may do a better job of straightening than another.	True [✓ False]
The transformation $(y^p-1)/p$ has the advantage over y^p in that the former, but not the latter, can take the reciprocal 1/y by letting p tend to infinity.	True [✓ False]
If $\log_2(y)$- $\log_2(x) = 4$, then you must double x four times to get y, that is, $y = 2^4 x = 16x$.	[✓ True] False

2. Fit model #1 on data page and assume it is true for the purpose of question 2.

	CIRCLE ONE
2.1 The hypothesis H_0: $\beta_2=0$ is rejected in a conventional two-sided 0.05 level test, where variable 2 is log(crim). (5 points)	TRUE [✓ FALSE]
2.2 Which observation has the largest absolute studentized = jackknife = deleted residual? Give the **row #**, the **name** of the town, and the **value** of the studentized residual. (6 points)	Row # 369 Value: 7.46 Name: Boston Back Bay
2.3 Continuing question 2.2, test the null hypothesis that there are no outliers in model #1. What is the numerical **value** of the test statistic? What is the two-sided **p-value corrected for testing multiple hypotheses**? What are the **degrees of freedom**? **How many hypotheses** were tested in testing for outliers? Is it **plausible**	Value: 7.46 (same as studentized residual) Easy way to get p-value: add an outlier dummy to model! P-value: $1.89 \times 10^{-10} = 506 \times 3.74 \times 10^{-13}$ Degrees of Freedom: 499 How many: 506

that there are no outliers? (6 points)	Plausible Not Plausible

PROBLEM SET #2 STATISTICS 500 FALL 2007: ANSWER PAGE 2

3.**Continue to use model #1 on data page.**	CIRCLE ONE (6 points each)
3.1 What is the **numerical value** of the largest leverage or hatvalue? What is the **row number** for this observation? What is the **name of this town**? In this data set, what is the **numerical cut off** for judging a leverage value to be large?	Value: 0.070 Row # 365 Name of town: Boston Back Bay Numerical cut off: 0.0237 = 2(5+1)/506
3.2 In question 3.1, this town has large leverage because cmedv is close to the median while rm is the maximum, whereas in model #1, more rooms usually raised the value of the home, leading to a negative residual.	TRUE [✓ FALSE]
3.3 What is the **numerical value** of the largest absolute dffits? What is the **row number** for this observation? What is the **name of this town**? If one added this observation to the regression based on the 505 other observations, the fitted value would rise by more than one standard error.	Value: 1.045 Row # 373 Name of town: Boston Beacon Hill [✓ TRUE] FALSE
3.4 In question 3.3, the dffits was large because the homes were expensive (cmedv) despite fairly high pollution (nox), fairly high ptratio, and below median number of rooms (rm).	[✓ TRUE] FALSE

4. Fit **model 2** on the data page, assume it to be true, and use it to answer question 4.

	CIRCLE ONE (6 points each)
4.1 Houses in tracts along the Charles River are estimated to be valued at 3.841 times more than houses in similar tracts not along the Charles River.	TRUE [✓ FALSE]
4.2 The relationship between median home values (cmedv) and the average number of rooms is shaped like a hill, an upside down U, so tracts with middle values of rm have the highest median values.	TRUE [✓ FALSE]
4.3 Test the hypothesis that the relationship is between cmedv and rm is linear, not quadratic, that is, test H_0: γ_4=0. What is the **name** of the test statistic? What is the **value** of the test statistic? Is the null hypothesis **plausible**?	Name: t-statistic Value: 10.8 Plausible [✓ Not plausible]
4.4 If question 2.1 were asked about γ_2 in model #2, what would the answer be?	[✓ TRUE] FALSE

PROBLEM SET #2 STATISTICS 500 FALL 2007

Doing the Problem Set in R

```
> library(mlbench)
> data(BostonHousing2)
> d<-BostonHousing2
> attach(d)
```

Question 1:

```
> plot(crim^(1/3),log(cmedv))
> lines(lowess(crim^(1/3),log(cmedv)))

> plot(log(crim),cmedv)
> lines(lowess(log(crim),cmedv))
```

Question 2:

```
> mod<-lm(cmedv ~ nox+log(crim)  + rm + ptratio +chas)
> summary(mod)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.4403     5.0908   1.462    0.145
nox         -15.6894     3.7170  -4.221 2.89e-05 ***
log(crim)    -0.1982     0.2084  -0.951    0.342
rm            6.8292     0.4017  17.001  < 2e-16 ***
ptratio      -1.0606     0.1371  -7.734 5.77e-14 ***
chas          4.2289     1.0186   4.152 3.88e-05 ***
---
> which.max(abs(rstudent(mod)))
369
> max(rstudent(mod))
[1] 7.464907
> min(rstudent(mod))
[1] -3.075885
> dim(d)
[1] 506  19
```

The conceptually easy way is to do a dummy variable regression and adjust the p-value using the Bonferroni inequality.

```
> testout<-rep(0,506)
> testout[369]<-1
> summary(lm(cmedv ~ nox+log(crim)  + rm + ptratio +chas+testout))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.5657     4.8397   1.150    0.251
nox         -15.0840     3.5298  -4.273 2.31e-05 ***
log(crim)    -0.2470     0.1980  -1.248    0.213
rm            7.0345     0.3824  18.397  < 2e-16 ***
ptratio      -1.0537     0.1302  -8.093 4.47e-15 ***
chas          4.2572     0.9670   4.402 1.31e-05 ***
testout      40.6680     5.4479   7.465 3.74e-13 ***
```

Bonferroni inequality: multiply p-value by #tests; reject if <0.05. (Can be > 1; it's an inequality!)

```
> (3.74e-13)*506
[1] 1.89244e-10
```

PROBLEM SET #2 STATISTICS 500 FALL 2007
Doing the Problem Set in R, continued

Question 3:

```
> summary(hatvalues(mod))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002178 0.005978 0.008229 0.011860 0.012610 0.070160
> which.max(hatvalues(mod))
365
> d[365,c(1,6,7,10,11,12,16)]
               town cmedv    crim chas   nox   rm tax
365 Boston Back Bay  21.9 3.47428    1 0.718 8.78 666

> summary(dffits(mod))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.844900 -0.053880 -0.006664  0.004587  0.031590  1.045000
> which.max(dffits(mod))
373
> d[373,c(1,6,7,10,11,12,16)]
                  town cmedv    crim chas   nox    rm tax
373 Boston Beacon Hill    50 8.26725    1 0.668 5.875 666
```

Question 4:

```
> rm2c<-(rm-mean(rm))^2
> mod2<-lm( cmedv ~ nox + log(crim) + rm +rm2c+ ptratio + chas)
> summary(mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.4266     4.5917   1.182  0.23784
nox         -14.8874     3.3507  -4.443 1.09e-05 ***
log(crim)    -0.5486     0.1906  -2.878  0.00417 **
rm            6.0683     0.3688  16.453  < 2e-16 ***
rm2c          2.7110     0.2511  10.798  < 2e-16 ***
ptratio      -0.8023     0.1259  -6.373 4.21e-10 ***
chas          3.8412     0.9187   4.181 3.42e-05 ***
Residual standard error: 5.145 on 499 degrees of freedom
Multiple R-Squared: 0.6897,     Adjusted R-squared: 0.686
F-statistic: 184.9 on 6 and 499 DF,  p-value: < 2.2e-16
```

chas=1 adds $3,841; it doesn't multiply by 3.841.

In the quadratic in room, $a + bx + cx^2$, the quadratic term, c, or γ_4, is estimated at 2.71, positive, so U shaped.

PROBLEM SET #3 STATISTICS 500 FALL 2007: DATA PAGE 1

Due Thursday 13 December 2007 at noon in my office 473 Huntsman. In you wish to turn it in early, put it in sealed envelop addressed to me and leave it in my mail box in Statistics, Huntsman 4th floor. **Make & keep a photocopy of your answer page**. If you would like your exam + answer key, include a ordinary stamped self addressed envelope.

This is an exam. Do not discuss it with anyone.

Same data set as Problem #1. To learn about the dataset, type:

```
> help(BostonHousing2,package=mlbench)
Format:
       crim     per capita crime rate by town
       zn       proportion of residential land zoned for lots over
                 25,000 sq.ft
       nox      nitric oxides concentration (parts per 10 million)
       rm       average number of rooms per dwelling
       age      proportion of owner-occupied units built prior to 1940
       tax      full-value property-tax rate per USD 10,000
       ptratio  pupil-teacher ratio by town
     The corrected data set has the following additional columns:
       cmedv  corrected median value of owner-occupied homes in USD
              1000's
```

To obtain the data, you can do one of several things:

Get it directly:

Go to the "packages" menu in R, click "load package" and click "mlbench" and type:

```
      > library(mlbench)
      > data(BostonHousing2)
```

Notice that you want BostonHousing2, NOT BostonHousing. You may wish to attach the data:

```
      > attach(BostonHousing2)
```

The data are also in the latest version of Rst500.RData and in an Excel file Bostonhousing2.xls at::

http://stat.wharton.upenn.edu/statweb/course/Fall-2007/STAT-500/

and Rst500.RData is also on my web page:

http://www-stat.wharton.upenn.edu/~rosenbap/index.html

```
> X<-BostonHousing2[,c(7,8,11,12,13,16,17)]
> X[1:3,]
     crim zn   nox    rm  age tax ptratio
1 0.00632 18 0.538 6.575 65.2 296    15.3
2 0.02731  0 0.469 6.421 78.9 242    17.8
3 0.02729  0 0.469 7.185 61.1 242    17.8
```

Model #1:

$$\text{cmedv} = \beta_0 + \beta_1 \text{ crim} + \beta_2 \text{ zn} + \beta_3 \text{ nox} + \beta_4 \text{ rm} + \beta_5 \text{ age} + \beta_6 \text{ tax} + \beta_7 \text{ ptratio} + \varepsilon$$

with ε iid $N(0,\sigma^2)$

Model #2:

$$\text{cmedv} = \beta_0 + \beta_1 \text{ crim} + \beta_3 \text{ nox} + \beta_4 \text{ rm} + \beta_7 \text{ ptratio} + \varepsilon$$

Model #3:

$$\text{cmedv} = \beta_0 + \beta_1 \text{ crim} + \beta_2 \text{ zn} + \beta_3 \text{ nox} + \beta_4 \text{ rm} + \beta_7 \text{ ptratio} + \varepsilon$$

Model #4:

$$\text{cmedv} = \beta_0 + \beta_1 \text{ crim} + \beta_2 \text{ zn} + \beta_3 \text{ nox} + \beta_5 \text{ age} + \beta_6 \text{ tax} + \beta_7 \text{ ptratio} + \varepsilon$$

PROBLEM SET #3 STATISTICS 500 FALL 2007: DATA PAGE 2

The second data set, "SantaAna" is from Gonsebatt, et al. (1997) Cytogenetic effects in human exposure to arsenic, *Mutation Research*, 386, 219-228. The town of Santa Ana (in Mexico) has a fairly high level of arsenic in drinking water, whereas Nazareno has a much lower level. The data set has 14 males (M) and 14 females (F) from Nazareno (labeled Control) and 14 males (M) and 14 females (F) from Santa Ana (labeled Exposed). For these 56 individuals, samples of oral epithelial cells were obtained and the frequency (Y) of micronuclei per thousand cells upon cell division was determined, Y=Mnbuccal. You are to do an analysis of variance with four groups **MC**=(Male-control), **FC** = (Female-control), **ME** = (Male-exposed), **FE**=(Female-exposed). For groups g = MC, FC, ME, FE, and individuals i=1,2,…,14 the model is:

$$(Y_{gi})^{1/3} = \mu + \tau_g + \varepsilon_{gi} \text{ with } \varepsilon_{gi} \sim \text{iid } N(0,\sigma^2) \qquad \text{Model \#5}$$

The cube root, $(Y_{gi})^{1/3}$, is taken to make the variances more equal. **You must take the cube root in your analysis**.

SantaAna is in the latest Rst500.RData. You will need to download the latest copy. The data are in the latest version of Rst500.RData and in an plain text file SantaAna.txt at::

http://stat.wharton.upenn.edu/statweb/course/Fall-2007/STAT-500/

and Rst500.RData is also on my web page:

http://www-stat.wharton.upenn.edu/~rosenbap/index.html

```
> dim(SantaAna)
[1] 56  6
> SantaAna[1:3,]
    Group  Code Age Sex YearRes Mnbuccal
1 Control   236  36   M      36     0.58
2 Control 88629  37   M      37     0.49
3 Control 96887  38   M      38     0.00
```

Question 5 asks you to construct three orthogonal contrast, one for Santa Ana (exposed) vs Nazareno (control), one for Male vs Female, and one for the interaction (or difference in differences) which asks whether the male-female difference is different in Santa Ana and Nazareno.

Follow instructions. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam.

Name: ______________________________ ID# __________________________

PROBLEM SET #3 STATISTICS 500 FALL 2007: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

1. Refer to Models #1, 2, 3, 4 for the BostonHousing2 data to answer these questions. Assume Model #1 is true.	Fill in the answer or **CIRCLE** the correct choice.
1.1 If you include the model with no predictors and model #1, how many models can be formed as submodels of model #1 by removing some (none, all) predictor variables.	Number of models = ____________
1.2 What is the value of C_P when model #2 is viewed as a submodel of model #1? What is the value of C_P when model #3 is viewed as a submodel of model #1? What is the value of C_P when model #4 is viewed as a submodel of model #1?	C_P for model #2: _______________ C_P for model #3: _______________ C_P for model #4: _______________
1.3 Of models 1, 2, 3, and 4, which one does C_P estimate will make the smallest total squared prediction errors? Give one model number.	Model number _____________________
1.4 Of models 1, 2, 3, and 4, for which model or models is the value of C_P compatible with the claim that this model contains all the variables with nonzero coefficients? Write the number or numbers of the models. If none, write "none".	Model number(s):
1.5 C_P estimates that the total squared prediction errors from model 4 are more than 300 times greater than from model 1.	CIRCLE ONE TRUE FALSE
1.6 In model #1, which variable has the largest variance inflation factor (vif)? What is the value of this vif?	Name of one variable: ______________ Value of vif: __________________
1.7 For the variable you identified in question 1.6, what is the R^2 for this variable when predicted from the other 6 predictors in model #1? What is the Pearson (ordinary) correlation between this variable and its predicted values using the other 6 predictors in model #1?	R^2 = ________________________ Pearson correlation = _______________
1.8 C_P always gets smaller when a variable with a vif > 10 is added to the model.	TRUE FALSE
1.9 Because the variable in question 1.6 has the largest vif, it is the best single predictor of Y=cmedv.	TRUE FALSE

Name: _____________________________ ID# _________________________

PROBLEM SET #3 STATISTICS 500 FALL 2007: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

2. Use the SantaAna data and model #5 to perform the following analysis of variance describing the four groups of 14 subjects. Assume model #5 is true for all questions.

Source of Variation	Sum of Squares (**SS**)	Degrees of Freedom (**DF**)	Mean Square (**MS**)	**F**-statistic	**p**-value
Between Groups (regression)					
Within Groups (residual)				[redacted]	[redacted]

Based on the analysis above, is it plausible that there is no difference in $Y^{1/3}$ by town and gender? CIRCLE ONE

PLAUSIBLE NOT PLAUSIBLE

3. Use Tukey's two sided, 0.05 level multiple comparison method to compare the groups in all pairs. Circle all the pairs of two groups that differ significantly by this method. Example, if MC and FC differ significantly, circle (MC,FC).

(MC,FC) (MC,FE) (ME,FC) (ME,FE) (MC,ME) (FC,FE)

4. In Tukey's method as used in question 3, suppose that in model #5, unknown to us, $\tau_{MC} = \tau_{FC} < \tau_{ME} = \tau_{FE}$, so exposure matters but gender does not. Assuming this supposition is true, circle the correct answers.

The chance that Tukey's method finds a significant difference between any two groups is at most 0.05.	TRUE FALSE
The chance that Tukey's method rejects H_0: $\tau_{MC} = \tau_{FC}$ is at most 0.05.	TRUE FALSE
The chance that Tukey's method rejects H_0: $\tau_{MC} = \tau_{ME}$ is at most 0.05.	TRUE FALSE
The chance that Tukey's method rejects H_0: $\tau_{MC} = \tau_{FC}$ or rejects H_0: $\tau_{ME} = \tau_{FE}$ is at most 0.05.	TRUE FALSE

5. Use 3 orthogonal contrasts to partition the anova table in question 2, and fill in the following table. Also, give the variance inflation factor (vif) for each contrast.

Source	**SS**	**DF**	**MS**	**F**	**p-value**	**vif**
Santa Ana vs Nazareno						
Male vs Female						
Difference in differences						

PROBLEM SET #3 STATISTICS 500 FALL 2007: ANSWER PAGE 1

1. Refer to Models #1, 2, 3, 4 for the BostonHousing2 data to answer these questions. Assume Model #1 is true.	Fill in the answer or **CIRCLE** the correct choice. 3 points each, 27 total
1.1 If you include the model with no predictors and model #1, how many models can be formed as submodels of model #1 by removing some (none, all) predictor variables.	Number of models = $2^7 = 128$
1.2 What is the value of C_P when model #2 is viewed as a submodel of model #1? What is the value of C_P when model #3 is viewed as a submodel of model #1? What is the value of C_P when model #4 is viewed as a submodel of model #1?	C_P for model #2: 4.44 C_P for model #3: 6.00 C_P for model #4: 304.88
1.3 Of models 1, 2, 3, and 4, which one does C_P estimate will make the smallest total squared prediction errors? Give one model number.	Model number: #2
1.4 Of models 1, 2, 3, and 4, for which model or models is the value of C_P compatible with the claim that this model contains all the variables with nonzero coefficients? Write the number or numbers of the models. If none, write "none".	Model number(s): #1, 2, 3
1.5 C_P estimates that the total squared prediction errors from model 4 are more than 300 times greater than from model 1.	CIRCLE ONE TRUE [✓ FALSE]
1.6 In model #1, which variable has the largest variance inflation factor (vif)? What is the value of this vif?	Name of one variable: nox Value of vif: 3.46
1.7 For the variable you identified in question 1.6, what is the R^2 for this variable when predicted from the other 6 predictors in model #1? What is the Pearson (ordinary) correlation between this variable and its predicted values using the other 6 predictors in model #1?	$R^2 = 0.7108 = 1-1/3.4573 = 1-(1/vif)$ Pearson correlation = $0.843 = (0.7108)^{1/2}$
1.8 C_P always gets smaller when a variable with a vif > 10 is added to the model.	TRUE [✓ FALSE]
1.9 Because the variable in question 1.6 has the largest vif, it is the best single predictor of Y=cmedv.	TRUE [✓ FALSE]

PROBLEM SET #3 STATISTICS 500 FALL 2007: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

2. Use the SantaAna data and model #5 to perform the following analysis of variance describing the four groups of 14. Assume model #5 is true for all questions. 20pts

Source of Variation	Sum of Squares (**SS**)	Degrees of Freedom (**DF**)	Mean Square (**MS**)	**F**-statistic	**p**-value
Between Groups (regression)	2.008	3	0.669	1.8	0.16
Within Groups (residual)	19.356	52	0.372	[redacted]	[redacted]

Based on the analysis above, is it plausible that there is no difference in $Y^{1/3}$ by town and gender? CIRCLE ONE

[✓ PLAUSIBLE] NOT PLAUSIBLE

3. Use Tukey's two sided, 0.05 level multiple comparison method to compare the groups in all pairs. Circle all the pairs of two groups that differ significantly by this method. Example, if MC and FC differ significantly, circle (MC,FC). (12 points) None are significant!

(MC,FC) (MC,FE) (ME,FC) (ME,FE) (MC,ME) (FC,FE)

4. In Tukey's method as used in question 3, suppose that in model #5, unknown to us, $\tau_{MC} = \tau_{FC} < \tau_{ME} = \tau_{FE}$, so exposure matters but gender does not. Assuming this supposition is true, circle the correct answers. (20 points) Tukey's method controls the chance of falsely rejecting any true hypothesis – you don't want to reject true hypotheses – but it tries to reject false hypotheses. You cannot falsely reject a false hypothesis!

The chance that Tukey's method finds a significant difference between any two groups is at most 0.05.	TRUE [✓ FALSE]
The chance that Tukey's method rejects H_0: $\tau_{MC} = \tau_{FC}$ is at most 0.05.	[✓ TRUE] FALSE
The chance that Tukey's method rejects H_0: $\tau_{MC} = \tau_{ME}$ is at most 0.05.	TRUE [✓ FALSE]
The chance that Tukey's method rejects H_0: $\tau_{MC} = \tau_{FC}$ or rejects H_0: $\tau_{ME} = \tau_{FE}$ is at most 0.05.	[✓ TRUE] FALSE

5. Use 3 orthogonal contrasts to partition the anova table in question 2, and fill in the following table. Also, give the variance inflation factor (vif) for each contrast.(21 points)

Source	**SS**	**DF**	**MS**	**F**	**p-value**	**vif**
Santa Ana vs Nazareno	1.145	1	1.146	3.1	0.085	1
Male vs Female	0.504	1	0.504	1.4	0.24	1
Difference in differences	0.359	1	0.359	0.96	0.33	1

Sums of squares partition with orthogonal contrasts: 2.008 = 1.145 + 0.504 + 0.359. There is no variance inflation in a balanced design with orthogonal (uncorrelated) contrasts.

Doing the Problem Set in R (Problem 3, Fall 2007)

```
> library(mlbench)
> data(BostonHousing2)
> X<-BostonHousing2[,c(7,8,11,12,13,16,17)]
> X[1:3,]
     crim zn   nox    rm  age tax ptratio
1 0.00632 18 0.538 6.575 65.2 296    15.3
2 0.02731  0 0.469 6.421 78.9 242    17.8
3 0.02729  0 0.469 7.185 61.1 242    17.8
```

The first time you use leaps, you must install it from the web. Each time you use leaps, you must request it using library(.). You must do library(leaps) before help(leaps).

```
> library(leaps)
> help(leaps)
> modsearch<-leaps(x=X,y=cmedv,names=colnames(X))
> cbind(modsearch$which,modsearch$Cp)
  crim zn nox rm age tax ptratio
1    0  0   0  1   0   0       0 170.405285
1    0  0   0  0   0   0       1 469.502583
1    0  0   0  0   0   1       0 512.474253
1    0  0   1  0   0   0       0 562.680462
1    1  0   0  0   0   0       0 605.132128
1    0  0   0  0   1   0       0 616.737365  etc.
4    1  0   1  1   0   0       1   4.438924  Model #2
5    1  1   1  1   0   0       1   6.002240  Model #3
6    1  1   1  0   1   1       1 304.882166  Model #4
7    1  1   1  1   1   1       1   8.000000  Model #1
```

The last column is C_P.

```
> library(DAAG)
> help(vif)
> mod<-lm(cmedv~crim+zn+nox+rm+age+tax+ptratio)
> vif(mod)
   crim      zn     nox      rm     age     tax ptratio
 1.5320  1.8220  3.4573  1.2427  2.4461  2.8440  1.6969
> summary(lm(nox~ crim + zn + rm + age + tax + ptratio))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.534e-01  4.468e-02  14.624  < 2e-16 ***
crim         1.083e-04  4.014e-04   0.270 0.787513
zn          -9.809e-04  1.554e-04  -6.313 6.07e-10 ***
rm          -1.451e-02  4.378e-03  -3.313 0.000988 ***
age          1.720e-03  1.345e-04  12.786  < 2e-16 ***
tax          3.303e-04  2.368e-05  13.948  < 2e-16 ***
ptratio     -1.352e-02  1.566e-03  -8.637  < 2e-16 ***
Residual standard error: 0.06269 on 499 degrees of freedom
Multiple R-Squared: 0.7108,     Adjusted R-squared: 0.7073
F-statistic: 204.4 on 6 and 499 DF,  p-value: < 2.2e-16
> sqrt(0.7108)
[1] 0.8430896
```

Doing the Problem Set in R (Problem 3, Fall '07), Continued

See 2006, problem set 3, for text commentary; it's the same.

```
> attach(SantaAna)
> mn3<-Mnbuccal^(1/3)
> boxplot(mn3~Sex:Group)
> gr<-Sex:Group
> gr
 [1] M:Control M:Control M:Control M:Control M:Control M:Control etc.
> summary(aov(mn3~gr))
            Df  Sum Sq Mean Sq F value Pr(>F)
gr           3  2.0080  0.6693  1.7982  0.159
Residuals   52 19.3556  0.3722

> TukeyHSD(aov(mn3~gr))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = mn3 ~ gr)
$gr
                           diff        lwr       upr
F:Exposed-F:Control  0.12597723 -0.4860490 0.7380035
M:Control-F:Control  0.02965799 -0.5823682 0.6416842
M:Exposed-F:Control  0.47575824 -0.1362680 1.0877845
M:Control-F:Exposed -0.09631924 -0.7083455 0.5157070
M:Exposed-F:Exposed  0.34978101 -0.2622452 0.9618072
M:Exposed-M:Control  0.44610025 -0.1659260 1.0581265

> expo<-c(-1,1,-1,1)
> gend<-c(1,1,-1,-1)
> genexp<-expo*gend
> contrasts(gr)<-cbind(expo,gend,genexp)
> contrasts(gr)
          expo gend genexp
F:Control   -1    1     -1
F:Exposed    1    1      1
M:Control   -1   -1      1
M:Exposed    1   -1     -1
> h<-model.matrix(mn3~gr)
> h
> Expo<-h[,2]
> Gend<-h[,3]
> ExGe<-h[,4]
> summary(lm(mn3~Expo+Gend+ExGe))
> anova(lm(mn3~Expo+Gend+ExGe))
Analysis of Variance Table

Response: mn3
          Df  Sum Sq Mean Sq F value  Pr(>F)
Expo       1  1.1455  1.1455  3.0773 0.08528 .
Gend       1  0.5039  0.5039  1.3538 0.24993
ExGe       1  0.3587  0.3587  0.9636 0.33083
Residuals 52 19.3556  0.3722
> vif(lm(mn3~Expo+Gend+ExGe))
Expo Gend ExGe
   1    1    1
> hatvalues(lm(mn3~Expo+Gend+ExGe))
```

PROBLEM SET #1 STATISTICS 500 FALL 2008: DATA PAGE 1

Due in class Tuesday Oct 28 at noon.

This is an exam. Do not discuss it with anyone.

The data are from the Fed and concern subprime mortgages. You do not have to go to the Fed web page, but it is interesting: http://www.newyorkfed.org/mortgagemaps/ The data describe subprime mortgages in the US as of August 2008. The first two lines of data are below for Arkansas (AK) and Alabama (AL).

State-Level Subprime Loan Characteristics, August 2008

Column 1	Column 7	Column 10	Column 13	Column 36	Column 38	Column 48	
Property State	Average current interest rate	Average FICO score (b)	Average combined LTV at origination	Percent with no or low documentation	Percent of cash-out refinances	Percent ARM loans	
	_	Definition	Definition	Definition	Definition	Definition	Y
AK	8.50	614	87.36	28.0%	53.3%	75.2%	16.8%
AL	9.20	602	87.66	20.6%	52.2%	53.5%	21.5%

The following definitions are from the Fed spreadsheet:

-- **rate** is the current mortgage interest rate. For adjustable rate mortgages, the rate may reset to a higher interest rate, perhaps 6% higher.

-- **fico** is a credit bureau risk score. The higher the FICO score, the lower the likelihood of delinquency or default for a given loan. Also, everything else being equal, the lower the FICO score, the higher will be the cost of borrowing/interest rate.

-- **ltv** stands for the combined Loan to Value and is the ratio of the loan amount to the value of the property at origination. Some properties have multiple liens at origination because a second or "piggyback" loan was also executed. Our data capture only the information reported by the first lender. If the same lender originated and securitized the second lien, it is included in our LTV measure. Home equity lines of credit, HELOCS, are not captured in our LTV ratios.

-- **lowdoc Percent Loans with Low or No Documentation** refers to the percentage of owner-occupied loans for which the borrower provided little or no verification of income and assets in order to receive the mortgage.

-- **cashout Cash-Out Refinances** means that the borrower acquired a nonprime loan as a result of refinancing an existing loan, and in the process of refinancing, the borrower took out cash not needed to meet the underwriting requirements.

-- **arms** stands for **adjustable rate mortgages** and means that the loans have a **variable rate of interest** that will be reset periodically, in contrast to loans with interest rates fixed to maturity. All ARMs in this spreadsheet refer to owner-occupied mortgages.

-- **Y** the percent of subprime mortgages that are in one of the following categories: (i) a payment is at least 90 days past due, (ii) in the midst of foreclosure proceedings, or (iii) in REO meaning that the lender has taken possession of the property. (It is the sum of columns 25, 26 and 28 in the Fed's original spreadsheet.) In other words, Y is measures the percent of subprime loans that have gone bad.

Notice that some variables are means and others are percents.

PROBLEM SET #1 STATISTICS 500 FALL 2008: DATA PAGE 2

The data set is at
http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/
If you are using R, then it is in the **subprime** data.frame of the latest version of the **Rst500.Rdata** workspace; you need to download the latest version. There is also a text file **subprime.txt**, whose first line gives the variable names.

Model #1

$Y = \beta_0 + \beta_1 \text{rate} + \beta_2 \text{fico} + \beta_3 \text{ltv} + \beta_4 \text{lowdoc} + \beta_5 \text{cashout} + \beta_6 \text{arms} + \varepsilon$
with ε iid $N(0,\sigma^2)$

Model #2

$Y = \gamma_0 + \gamma_1 \text{ltv} + \gamma_2 \text{lowdoc} + \gamma_3 \text{cashout} + \gamma_4 \text{arms} + \zeta$
with ζ iid $N(0,\omega^2)$

Model 1 has slopes β (beta), while model 2 has slopes γ (gamma), so that different things have different names. The choice of Greek letters is arbitrary. A slope by any other name would tilt the same.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam.

Refer to states by their two-letter abbreviations.

Name: _____________________________ ID# _________________________

PROBLEM SET #1 STATISTICS 500 FALL 2008: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due in class Tuesday 28 Oct noon.

1. Which state has the largest Y? Which state has the smallest Y? What are the values of the predictors for these two states? What are the quartiles of Y?

	state	rate	fico	ltv	lowdoc	cashout	arms	Y
Max Y								
Min Y								

	Lower Quartile	Median	Upper Quartile
Y			

2. Fit model #1 defined on the data page.

Question	CIRCLE ONE or Fill in the Answer
2.1 When you plot Y (vertical) against X=lowdoc (horizontal), which state is in the upper right corner of the plot (high Y, high X)?	
2.2 When you plot Y (vertical) against X=lowdoc (horizontal), which state is in the lower right corner of the plot (low Y, high X)?	
2.3 In the fit of model #1, what is the two-sided p-value for testing the null hypothesis H_0: $\beta_1=0$, where β_1 is the coefficient of **rate**.?	
2.4 In model #1, what is the two-sided 95% confidence interval for β_1?	[,]
2.5 In model #1, there can be no plausible doubt that $\beta_1>0$, that is, no plausible doubt that higher rates of bad subprime loans (Y) are associated with higher current interest rates on those loans.	TRUE FALSE
2.6 What is the estimate of σ in model #1?	
2.7 What is the correlation between Y and the fitted value for Y in model #1? (Read this question carefully.)	
2.8 Suppose two states had identical predictors except that lowdoc was 2 units (2%) higher in state 1 than in state 2. Using the estimate of β_4 in model #1, the first state is predicted to have 1.11% more bad loans.	TRUE FALSE
2.9 Do a normal quantile plot and a Shapiro-Wilk test of the normality of the residuals in model #1. These clearly indicate the residuals are not Normally distributed.	TRUE FALSE

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2008: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

3. Fit model #2 on the data page. Question 3 refers to both model #1 and model #2, so make sure you use the correct model to answer each question.

Question	CIRCLE ONE or Fill in the Answer
3.1 Give the two-sided p-value for testing H_0: $\beta_2=0$ in model #1 (the coefficient of **fico**).	
3.2 Given the results of questions 2.3 and 3.1, it is reasonable to remove both **rate** and **fico** from model #1 and use model #2 instead.	TRUE FALSE
3.3 Give the two-sided p-value for testing H_0: $\beta_6=0$ in model #1 (the coefficient of **arms**).	
3.4 Give the two-sided p-value for testing H_0: $\gamma_4=0$ in model #2 (the coefficient of **arms**).	
3.5 What is the correlation between **rate** and **fico**?	

4. Test the hypothesis H_0: $\beta_1=\beta_2=0$ in model #1. Fill in the following table.

4.1	Variables	Sum of squares	Degrees of freedom	Mean square	F
Full Model		__________	__________	__________	Leave this space blank
Reduced model	ltv, lowdoc, cashout and arms alone	__________	__________	__________	Leave this space blank
	Added by rate and fico	__________	__________	__________	__________
Residual from full model		__________	__________	__________	Leave this space blank

Question	CIRCLE ONE or Fill in the Answer
4.2 The null hypothesis H_0: $\beta_1=\beta_2=0$ in model #1 is plausible.	TRUE FALSE
4.3 The current interest **rate** tends to be higher in states where the credit score **fico** is lower.	TRUE FALSE
4.4 The current interest **rate** tends to be lower in states where **arms** is higher.	TRUE FALSE

PROBLEM SET #1 STATISTICS 500 FALL 2008: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

1. Which state has the largest Y? Which state has the smallest Y? What are the values of the predictors for these two states? What are the quartiles of Y? (10 points)

	state	rate	fico	ltv	lowdoc	cashout	arms	Y
Max Y	CA	7.68	640	81.94	46.5	56.0	71.6	37.8
Min Y	WY	8.54	613	87.54	17.3	51.1	62.3	12.4

	Lower Quartile	Median	Upper Quartile
Y	17.9	22.1	27.5

2. Fit model #1 defined on the data page. (4 points each)

2.1 When you plot Y (vertical) against X=lowdoc (horizontal), which state is in the upper right corner of the plot (high Y, high X)?	CA
2.2 When you plot Y (vertical) against X=lowdoc (horizontal), which state is in the lower right corner of the plot (low Y, high X)?	HI
2.3 In the fit of model #1, what is the two-sided p-value for testing the null hypothesis H_0: β_1=0, where β_1 is the coefficient of **rate**.?	0.172
2.4 In model #1, what is the two-sided 95% confidence interval for β_1?	[-2.57, 13.98]
2.5 In model #1, there can be no plausible doubt that β_1>0, that is, no plausible doubt that higher rates of bad subprime loans (Y) are associated with higher current interest rates on those loans.	TRUE [✓ FALSE]
2.6 What is the estimate of σ in model #1?	4.165 That is, the typical state is estimated to deviate from the model by 4.2% bad loans.
2.7 What is the correlation between Y and the fitted value for Y in model #1? (Read this question carefully. It asks about R, not R^2)	The multiple R is 0.778. (However, R^2 = 0.778^2=0.6055)
2.8 Suppose two states had identical predictors except that lowdoc was 2 units (2%) higher in state 1 than in state 2. Using the estimate of β_4 in model #1, the first state is predicted to have 1.11% more bad loans.	TRUE [✓ FALSE] $2\beta_4$ is estimated to be 0.86834*2 = 1.73668, or 1.7% more bad loans with 2% more lowdoc loans!
2.9 Do a normal quantile plot and a Shapiro-Wilk test of the normality of the residuals in model #1. These clearly indicate the residuals are not Normally distributed.	TRUE [✓ FALSE]

PROBLEM SET #1 STATISTICS 500 FALL 2008: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

3. Fit model #2 on the data page. Question 3 refers to both model #1 and model #2, so make sure you use the correct model to answer each question. (4 points each)

Question	CIRCLE ONE or Fill in the Answer
3.1 Give the two-sided p-value for testing H_0: $\beta_2=0$ in model #1 (the coefficient of **fico**).	0.577
3.2 Given the results of questions 2.3 and 3.1, it is reasonable to remove both **rate** and **fico** from model #1 and use model #2 instead.	TRUE [✓ FALSE]
3.3 Give the two-sided p-value for testing H_0: $\beta_6=0$ in model #1 (the coefficient of **arms**).	0.0178
3.4 Give the two-sided p-value for testing H_0: $\gamma_4=0$ in model #2 (the coefficient of **arms**).	0.128
3.5 What is the correlation between **rate** and **fico**?	-0.92

4. Test the hypothesis H_0: $\beta_1=\beta_2=0$ in model #1. Fill in the following table. (22 points)

4.1	Variables	Sum of squares	Degrees of freedom	Mean square	F
Full Model		1171.23	6	195.205	Leave this space blank
Reduced model	ltv, lowdoc, cashout and arms alone	935.04	4	233.76	Leave this space blank
	Added by rate and fico	236.19	2	118.095	6.81
Residual from full model		763.16	44	17.34	Leave this space blank

Question (4 points each)	CIRCLE ONE or Fill in the Answer
4.2 The null hypothesis H_0: $\beta_1=\beta_2=0$ in model #1 is plausible.	TRUE [✓ FALSE]
4.3 The current interest **rate** tends to be higher in states where the credit score **fico** is lower.	[✓ TRUE] FALSE
4.4 The current interest **rate** tends to be lower in states where **arms** is higher.	[✓ TRUE] FALSE

Doing the Problem Set in R

Commands are in **bold**, comments in *script*, and needed pieces of output are underlined.

```
> attach(subprime)
```

Question 1

```
> which.max(Y)
[1] 5
> which.min(Y)
[1] 51
> subprime[c(5,51),]
   state rate fico   ltv lowdoc cashout arms    Y
5     CA 7.68  640 81.94   46.5    56.0 71.6 37.8
51    WY 8.54  613 87.54   17.3    51.1 62.3 12.4
> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.40   17.90   22.10   23.02   27.50   37.80
```

Question 2.1 and 2.2

```
> plot(lowdoc,Y)
> identify(lowdoc,Y,label=state)
```

Questions 2.3-2.9

```
> mod<-lm(Y~rate+fico+ltv+lowdoc+cashout+arms)
> summary(mod)
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -150.27452  131.81108  -1.140  0.26042
rate           5.70574    4.10621   1.390  0.17166
fico          -0.09873    0.17577  -0.562  0.57718
ltv            1.37314    0.57335   2.395  0.02095 *
lowdoc         0.86834    0.17702   4.905 1.32e-05 ***
cashout        0.54665    0.16731   3.267  0.00211 **
arms           0.20670    0.08393   2.463  0.01777 *
Residual standard error: 4.165 on 44 degrees of freedom
Multiple R-Squared: 0.6055,     Adjusted R-squared: 0.5517
F-statistic: 11.25 on 6 and 44 DF,  p-value: 1.391e-07
> lmci(mod)             low        high
(Intercept) -415.92229944 115.3732679
rate          -2.56978052  13.9812691
fico          -0.45298067   0.2555211
ltv            0.21762827   2.5286478
lowdoc         0.51158339   1.2250984
cashout        0.20945624   0.8838393
arms           0.03755305   0.3758557
```

Question 2.7

```
> cor(Y,mod$fitted.value)
[1] 0.7781248
> cor(Y,mod$fitted.value)^2
[1] 0.6054782
```

Question 2.8

```
> 0.86834*2
[1] 1.73668
```

Question 29

```
> qqnorm(mod$residual)
> shapiro.test(mod$residual)
        Shapiro-Wilk normality test
data:  mod$residual
W = 0.9867, p-value = 0.832
```

The plot looks reasonably straight. The p-value, 0.832, is large, much bigger than 0.05, so the null hypothesis that the residuals are Normal is not rejected.

Question 3

```
> mod2<-lm(Y ~ ltv + lowdoc + cashout + arms)
> summary(mod2)
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -146.22463   59.55846  -2.455  0.01792 *
ltv            1.35327    0.59458   2.276  0.02755 *
lowdoc         0.53532    0.15479   3.458  0.00118 **
cashout        0.53688    0.18577   2.890  0.00586 **
arms           0.14212    0.09174   1.549  0.12821
Residual standard error: 4.661 on 46 degrees of freedom
Multiple R-Squared: 0.4834,     Adjusted R-squared: 0.4385
F-statistic: 10.76 on 4 and 46 DF,  p-value: 3.065e-06
> cor(rate,fico)
[1] -0.9116614
```

Interest rates are higher, on average, in state where credit scores are lower, on average.

```
> plot(rate,fico)
> identify(rate,fico,label=state)
> anova(lm(Y~1),mod)
Analysis of Variance Table
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     50 1934.39
2     44  763.16  6   1171.23 11.255 1.391e-07 ***
> anova(lm(Y~1),mod2)
  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1     50 1934.39
2     46  999.35  4    935.04 10.76 3.065e-06 ***
---
> anova(mod2,mod)
Analysis of Variance Table
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     46 999.35
2     44 763.16  2    236.19 6.8089 0.002653 **
> 1-pf(6.81,2,44)
[1] 0.002650619
```

PROBLEM SET #2 STATISTICS 500 FALL 2008: DATA PAGE 1

Due in class Tuesday Nov 25 at noon.

This is an exam. Do not discuss it with anyone.

The data are as in Problem Set #1, except two new variables have been added. "Lower07" and "Upper07" indicate which political party, the Democrats (Dem) or Republicans (Rep) had a majority in the Lower and Upper houses of the state legislature. There is one exception, Nebraska, which no longer has parties in the state legislature – they are coded Rep to reflect their voting in most Presidential elections. The District of Columbia (Washington, DC) has been removed.

The data are from the Fed and concern subprime mortgages. You do not have to go to the Fed web page, but it is interesting: http://www.newyorkfed.org/mortgagemaps/

The data describe subprime mortgages in the US as of August 2008. The following definitions are from the Fed spreadsheet:

-- rate is the current mortgage interest rate. For adjustable rate mortgages, the rate may reset to a higher interest rate, perhaps 6% higher.

-- fico is a credit bureau risk score. The higher the FICO score, the lower the likelihood of delinquency or default for a given loan. Also, everything else being equal, the lower the FICO score, the higher will be the cost of borrowing/interest rate.

-- ltv stands for the combined Loan to Value and is the ratio of the loan amount to the value of the property at origination. Some properties have multiple liens at origination because a second or "piggyback" loan was also executed. Our data capture only the information reported by the first lender. If the same lender originated and securitized the second lien, it is included in our LTV measure. Home equity lines of credit, HELOCS, are not captured in our LTV ratios.

-- lowdoc Percent Loans with Low or No Documentation refers to the percentage of owner-occupied loans for which the borrower provided little or no verification of income and assets in order to receive the mortgage.

-- cashout Cash-Out Refinances means that the borrower acquired a nonprime loan as a result of refinancing an existing loan, and in the process of refinancing, the borrower took out cash not needed to meet the underwriting requirements.

-- arms stands for **adjustable rate mortgages** and means that the loans have a **variable rate of interest** that will be reset periodically, in contrast to loans with interest rates fixed to maturity. All ARMs in this spreadsheet refer to owner-occupied mortgages.

-- Y the percent of subprime mortgages that are in one of the following categories: (i) a payment is at least 90 days past due, (ii) in the midst of foreclosure proceedings, or (iii) in REO meaning that the lender has taken possession of the property. (It is the sum of columns 25, 26 and 28 in the Fed's original spreadsheet.) In other words, Y is measures the percent of subprime loans that have gone bad.

Notice that some variables are means and others are percents and others are nominal.

PROBLEM SET #2 STATISTICS 500 FALL 2008: DATA PAGE 2

The data set is at
http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/
If you are using R, then it is in the **subprime2** data.frame of the latest version of the **Rst500.Rdata** workspace; you need to download the latest version. There is also a text file **subprime2.txt**, whose first line gives the variable names.

Model #1

$Y = \beta_0 + \beta_1$rate + $\beta_2$fico + $\beta_3$ltv + $\beta_4$lowdoc + $\beta_5$cashout + $\beta_6$arms + ε
with ε iid $N(0,\sigma^2)$

Model #2

Model #2 is the same as model #1 except that an (uncentered) interaction between rate and arms is included as another variable, namely (arms x rate). In a fixed rate mortgage, it is good news to have a low rate, but in a subprime mortgage a low current rate is likely to be a teaser rate on an adjustable rate mortgage whose interest rate may soon rise by, perhaps, 6%, as in 8% now adjusts to 8+6 = 14% after the teaser rate ends. A state with high arms and low rate may have many mortgages with big increases coming soon. Would you struggle to pay your mortgage now if you knew it would soon adjust so that you could not pay it any more? Or might you walk away?

Model #3

Model #3 is model #2 with one more variable, namely **divided**. Model #3 also includes the interaction from model #2. Let **divided** = 1 if the upper and lower houses of the state legislature are of different parties (one Democrat, the other Republican) and **divided** = 0 if the upper and lower houses are of the same party (both Democrat or both Republican).

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam.

Special instructions:

1. **Refer to states by their two-letter abbreviations**.

2. It is important to use the data from **subprime2**, not from **subprime**. **subprime2** omits DC and includes additional variables.

3. One question asks about studentized residuals. This terminology is not standardized across statistical packages. These are called studentized residuals in R and jackknife residuals in your book. Do not assume that another package uses terminology in the same way.

Last name: ___________________ First name: ___________ ID# ________________

PROBLEM SET #2 STATISTICS 500 FALL 2008: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due in class Tuesday 28 Oct noon.

1. In the **fit of model #1 to subprime2**…	CIRCLE ONE or Fill in the value
1.1 **Which state** has the largest leverage or hat value? Give the two letter abbreviation of one state.	
1.2 What is the **numerical value** of the largest leverage or hat value for the state you identified in the previous question?	
1.3 For model #1, what is the numerical cut-point for a "large hat value"? Give one number.	
1.4 The state with the largest leverage or hat value has large leverage because the percent of subprime mortgages gone bad is one of the lowest in the 50 states.	TRUE FALSE
1.5 You should always remove from the regression the one observation with the largest leverage.	TRUE FALSE
1.6 **Which state** has the second largest leverage or hat value? Give the two letter abbreviation of one state.	

2. In the **fit of model #1 to subprime2**…	CIRCLE ONE or Fill in the value
2.1 **Which state** has the largest absolute studentized residual? Give the two letter abbreviation of one state.	
2.2 What is the **numerical value** of this most extreme studentized residual? Give a number with its sign, + or -.	
2.3 The state with the largest absolute studentized residual is largest because its percent of subprime mortgages gone bad is one of the lowest in the 50 states.	TRUE FALSE
2.4 Fit model #1 adding an indicator for the state you identified in 2.1 above. What is the t-statistic and p-value reported in the output for that indicator variable?	t = __________ p-value = ____________
2.5 For the state in 2.1 to be judged a statistically significant outlier at the 0.05 level, the p-value in 2.4 would need to be less than or equal to what number?	
2.6 The state in 2.1 is a statistically significant outlier at the 0.05 level.	TRUE FALSE

Last name: ___________________ First name: ___________ ID# ________________

PROBLEM SET #2 STATISTICS 500 FALL 2008: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Read the data page.

3. In the **fit of model #1 to subprime2**…	CIRCLE ONE or Fill in the value
3.1 **Which state** has the largest absolute dffits? Give the two letter abbreviation of one state.	
3.2 What is the **numerical value** of this most extreme dffits? Give a number with its sign, + or -.	
3.3 The addition of this state to a regression that did not include it reduces the coefficient of **arms** by about 1.6 standard errors.	TRUE FALSE
3.4 The addition of this state to a regression that did not include it will shift at least one of the 6 estimated slopes in model 1 by more than 1.6 standard errors in absolute value.	TRUE FALSE
3.5 If the Y for the state in identified in 3.1 were increased by 1, the fitted Y for this state in model #1 would increase by about 0.256.	TRUE FALSE

4. F**it of model #2 to subprime2**. Test the null hypothesis that rate and arms do not interact with each other in model #2.	CIRCLE ONE or Fill in the value
4.1 In this test, what is the name of the test statistic, the value of the test statistic, and the p-value?	Name: ___________ Value: __________ P-value: __________
4.2 Is it plausible that there is no interaction between rate and arms in model #2.	PLAUSIBLE NOT PLAUSIBLE
4.3 Give the observed Y and the fitted value for Y for Hawaii (HI) in model #1 and model #2.	Observed: In model 1: ________ In model 2: _____

5. F**it of model #3 to subprime2**.	CIRCLE ONE or Fill in the value
5.1 What is the **estimate** of the coefficient of “divided”? What is its estimated standard error (**se**)?	Estimate: _______ se: ___________
5.2 The model fits lower rates of subprime mortgages gone bad in states where control of the legislature is divided.	TRUE FALSE

PROBLEM SET #2 STATISTICS 500 FALL 2008 **5 points each, except as noted.**

1. In the **fit of model #1 to subprime2**…	CIRCLE ONE or Fill in the value
1.1 **Which state** has the largest leverage or hat value? Give the two letter abbreviation of one state. 1 point	TX = Texas
1.2 What is the **numerical value** of the largest leverage or hat value for the state you identified in the previous question?	0.467
1.3 For model #1, what is the numerical cut-point for a "large hat value"? Give one number.	0.28 = 2 x (1+6)/50
1.4 The state with the largest leverage or hat value has large leverage because the percent of subprime mortgages gone bad is one of the lowest in the 50 states.	TRUE [✓ FALSE]
1.5 You should always remove from the regression the one observation with the largest leverage.	TRUE [✓ FALSE]
1.6 **Which state** has the second largest leverage or hat value? Give the two letter abbreviation of one state. 1 point	HI = Hawaii

2. In the **fit of model #1 to subprime2**…	CIRCLE ONE or Fill in the value
2.1 **Which state** has the largest absolute studentized residual? Give the two letter abbreviation of one state. 1 point	CA = California
2.2 What is the **numerical value** of this most extreme studentized residual? Give a number with its sign, + or -.	2.957
2.3 The state with the largest absolute studentized residual is largest because its percent of subprime mortgages gone bad is one of the lowest in the 50 states.	TRUE [✓ FALSE]
2.4 Fit model #1 adding an indicator for the state you identified in 2.1 above. What is the t-statistic and p-value reported in the output for that indicator variable?	t = 2.957 p-value = 0.005081 Compare 2.2 and 2.4!
2.5 For the state in 2.1 to be judged a statistically significant outlier at the 0.05 level, the p-value in 2.4 would need to be less than or equal to what number?	0.001 = 0.05/50
2.6 The state in 2.1 is a statistically significant outlier at the 0.05 level.	TRUE [✓ FALSE]

PROBLEM SET #2 STATISTICS 500 FALL 2008: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due in class Tuesday 28 Oct noon.

3. In the **fit of model #1 to subprime2**…	CIRCLE ONE or Fill in the value
3.1 **Which state** has the largest absolute dffits? Give the two letter abbreviation of one state. 1 point	HI = Hawaii
3.2 What is the **numerical value** of this most extreme dffits? Give a number with its sign.	-1.612 Wow!
3.3 The addition of this state to a regression that did not include it reduces the coefficient of **arms** by about 1.6 standard errors.	TRUE [✓ FALSE] No, this is silly.
3.4 The addition of this state to a regression that did not include it will shift at least one of the 6 estimated slopes in model 1 by more than 1.6 standard errors in absolute value.	TRUE [✓ FALSE] No, the absolute dffits is an upper bound, not a lower bound, on the absolute dfbetas.
3.5 If the Y for the state in identified in 3.1 were increased by 1, the fitted Y for this state in model #1 would increase by about 0.256.	TRUE [✓ FALSE] Look at the hatvalue for HI, 0.456, not 0.256.

4. F**it of model #2 to subprime2**. Test the null hypothesis that rate and arms do not interact with each other in model #2.	CIRCLE ONE or Fill in the value
4.1 In this test, what is the name of the test statistic, the value of the test statistic, and the p-value?	Name: t-statistic Value: -2.21 P-value: 0.033
4.2 Is it plausible that there is no interaction between rate and arms in model #2.	PLAUSIBLE [✓ NOT PLAUSIBLE]
4.3 Give the observed Y and the fitted value for Y for Hawaii (HI) in model #1 and model #2.	Observed: 16.6% In model 1: 21.8% In model 2: 18.2% Whatever else, the interaction helped with HI.

5. F**it of model #3 to subprime2**.	CIRCLE ONE or Fill in the value
5.1 What is the **estimate** of the coefficient of "divided"? What is its estimated standard error (**se**)?	Estimate: 2.99 se: 1.32
5.2 The model fits lower rates of subprime mortgages gone bad in states where control of the legislature is divided.	TRUE [✓ FALSE] No, from 5.1, it is about 3% higher – obviously, this may not be the cause.

DOING THE PROBLEM SET IN R
(Fall 2008, Problem Set 2)

Question #1

```
> mod<-lm(Y~rate+fico+ltv+lowdoc+cashout+arms)
> dim(subprime2)
[1] 50 15
> mean(hatvalues(mod))
[1] 0.14
> (1+6)/50
[1] 0.14
> 2*.14
[1] 0.28
> subprime2[hatvalues(mod)>=.28,1:7]
   state rate fico   ltv lowdoc cashout arms
12    HI 7.45  646 80.37   44.9    62.0 46.6
44    TX 8.88  606 84.44   29.8    41.7 45.3
47    VT 8.65  612 80.29   30.3    67.6 59.7
> hatvalues(mod)[hatvalues(mod)>=.28]
       11        43        46
0.4561207 0.4674280 0.2938781
```

Question #2

```
> which.max(abs(rstudent(mod)))
5
> subprime2[5,1:8]
  state rate fico   ltv lowdoc cashout arms    Y
5    CA 7.68  640 81.94   46.5      56 71.6 37.8
> rstudent(mod)[5]
       5
2.956946
> ca<-rep(0,50)
> ca[5]<-1
> summary(lm(Y~rate+fico+ltv+lowdoc+cashout+arms+ca))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -157.55396  123.55951  -1.275 0.209273
rate           3.09684    3.81461   0.812 0.421463
fico          -0.21463    0.16338  -1.314 0.196087
ltv            2.43404    0.68514   3.553 0.000958 ***
lowdoc         0.99174    0.17504   5.666 1.20e-06 ***
cashout        0.77536    0.17560   4.415 6.93e-05 ***
arms           0.10337    0.08808   1.174 0.247192
ca            12.54794    4.24355   2.957 0.005081 **
> 0.05/50
[1] 0.001
```

Question #3

```
> boxplot(dffits(mod))
> which.max(dffits(mod))
5
> which.min(dffits(mod))
11
> subprime2[c(5,11),1:8]
   state rate fico   ltv lowdoc cashout arms    Y
5     CA 7.68  640 81.94   46.5      56 71.6 37.8
12    HI 7.45  646 80.37   44.9      62 46.6 16.6
> dffits(mod)[11]
```

```
       11
-1.612354
>  round(dfbetas(mod)[11,],2)
(Intercept)  rate   fico    ltv      lowdoc     cashout        arms
       0.82 -0.35  -0.83  -0.16       -0.02       -0.34        0.95
> max(abs(dffits(mod)))
[1] 1.612354
> max(abs(dfbetas(mod)))
[1] 0.9491895
```

Question #4

```
> interact<-rate*arms
> summary(lm(Y~rate+fico+ltv+lowdoc+cashout+arms+interact))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -463.25688  176.28667  -2.628  0.01195 *
rate          25.60511   10.18440   2.514  0.01585 *
fico          -0.01633    0.17304  -0.094  0.92524
ltv            2.29927    0.70883   3.244  0.00232 **
lowdoc         0.98048    0.18204   5.386 3.02e-06 ***
cashout        0.81987    0.19058   4.302 9.89e-05 ***
arms           2.83047    1.22124   2.318  0.02541 *
interact      -0.31779    0.14376  -2.211  0.03257 *
---
Residual standard error: 3.935 on 42 degrees of freedom
Multiple R-Squared: 0.6585,     Adjusted R-squared: 0.6016
F-statistic: 11.57 on 7 and 42 DF,  p-value: 4.396e-08
> subprime2[11,1:8]
   state rate fico   ltv lowdoc cashout arms    Y
12    HI 7.45  646 80.37   44.9      62 46.6 16.6
> lm(Y~rate+fico+ltv+lowdoc+cashout+arms)$fitted.values[11]
      11
21.81003
> lm(Y~rate+fico+ltv+lowdoc+cashout+arms+interactc)$fitted.values[11]
      11
18.16952
```

Question #5

```
>  summary(lm(Y~rate+fico+ltv+lowdoc+cashout+arms+interact+divided))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -507.84012  169.29768  -3.000  0.00458 **
rate          28.58942    9.80312   2.916  0.00572 **
fico           0.02464    0.16604   0.148  0.88274
ltv            2.22538    0.67692   3.288  0.00208 **
lowdoc         0.95104    0.17413   5.462 2.51e-06 ***
cashout        0.81514    0.18181   4.484 5.80e-05 ***
arms           3.21633    1.17724   2.732  0.00924 **
interact      -0.36180    0.13849  -2.612  0.01251 *
divided        2.98845    1.31563   2.271  0.02843 *
```

PROBLEM SET #3 STATISTICS 500 FALL 2008: DATA PAGE 1

Due in my office, 473 JMHH, Wednesday December 10, 2008 at 11:00am.

This is an exam. Do not discuss it with anyone.

The National Supported Work (NSW) project was a randomized experiment intended to provide job skills and experience to the long-term unemployed. The treatment consisted of gradual, subsidized exposure to regular work. The data are in the data.frame nsw in Rst500.Rdata – you need to get the latest version. There is a text file, nsw.txt. Both are at http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/

A coin was flipped to assign each person to treatment or control. A portion of the data is below. (All are men. Some sampling was used to create a balanced design to simplify your analysis.)

```
> nsw[1:3,c(1,3,8:12),]
   treat edu re74 re75   re78 change            group
7      1  12    0    0    0.0    0.0 Treated:Grade11+
55     1  11    0    0  590.8  590.8 Treated:Grade11+
50     1  12    0    0 4843.2 4843.2 Treated:Grade11+
> dim(nsw)
[1] 348  12
> table(group)
```

(Notice the group numbers, i=1,2,3,4.)

```
Group
    i=1                i=2               i=3             i=4
Control:Grade11+ Control:Grade10- Treated:Grade11+ Treated:Grade10-
              87               87               87               87
```

treat=1 for treated, =0 for control. **edu** is highest grade of education. The variable change is re78-(re75+re74)/2, where reYY is earnings in $ in year YY. For the men in nsw, re78 is posttreatment earnings and both re74 and re75 are pretreatment earnings. The variable “group” has four levels, based on treatment-vs-control and highest grade is 11th grade or higher vs 10th grade or lower. There are 87 men in each group. Obviously, the creators of the nsw treatment would have been happy to see large positive values of “change” among treated men.

You can read about the NSW in the paper by Couch (1992). The data are adapted from work by Dehjia and Wahba (1999). There is no need to go to these articles unless you are curious – they will not help in doing the problem set.

You are to do a one-way anova of change by group, so there are four groups.

Model #1 is $\text{change}_{ij} = \mu + \tau_i + \varepsilon_{ij}$ for i=1,2,3,4, j=1,2,…,87, with ε_{ij} iid $N(0,\sigma^2)$.

Concerning question 3: Create 3 orthogonal contrasts to represent a comparison of treatment and control (treatment), a comparison of grade 10 or less vs grade 11 or more (grade), and their interaction (interaction). Use integer weights.

Couch, K.A.: New evidence on the long-term effects of employment training programs. *J Labor Econ* 10, 380-388. (1992)

Dehejia, R.H., Wahba, W.: Causal effects in nonexperimental studies: reevaluating the evaluation of training programs causal effects in nonexperimental studies. *J Am Statist Assoc* 94, 1053-1062 (1999)

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam.

Special instructions:

1. **Make a photocopy of your answer page.**

2. **You may turn in the exam early**. You may leave it in my mail box in the statistics department, 4th floor of JMHH, in an envelop addressed to me.

3. One question asks about **studentized residuals**. This terminology is not standardized across statistical packages. These are called studentized residuals in R and jackknife residuals in your book. Do not assume that another package uses terminology in the same way.

Last name: __________________ First name: ___________ ID# ________________

PROBLEM SET #3 STATISTICS 500 FALL 2008: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due Wednesday 10-Dec-08 at 11:00am.

1. Do a one-way analysis of variance of y=change by the four groups defined by “group” in the nsw data. Use this to answer the following questions.

Question	CIRCLE ONE or Fill in Values
1.1 Test H_0: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ under model #1. What is the **name** of the test statistic? What is the numerical **value** of the test statistic? What is the **p-value**? Is the null hypothesis **plausible**?	Name:__________ Value: __________ p-value: ________ H_0 is: PAUSIBLE NOT PLAUSIBLE
1.2 What is the mean change in each of the four groups? Here Tr is treated, Co is control, Gr11+ is grade 11 or more, Gr10- is grade 10 or less.	TrGr11+ _________ TrGr10- _________ CoGr11+ _________ CoGr10- _________
1.3 What is the unbiased **estimate** of σ^2? What is the corresponding **estimate** of σ?	σ^2 : ________ σ :__________
1.4 If ε_{ij} were not Normal, then this could invalidate the test you did in 1.1.	TRUE FALSE

2. Use Tukey’s method to compare every pair of two groups. Use Tukey’s method in two-sided comparisons that control the experiment-wise error rate at 0.05.

Identify groups by number,
`i=1 Control:Grade11+`
`i=2 Control:Grade10-`
`i=3 Treated:Grade11+`
`i=4 Treated:Grade10-`
	CIRCLE ONE or Fill in Values

2.1 With four groups, there are how many pairwise tests done by Tukey’s method?	How many: __________
2.2 List all pairs (a,b) of null hypotheses, H_0: $\tau_a = \tau_b$ which are rejected by Tukey’s method. List as (a,b) where a and b are in {1,2,3,4}. If none, write “none”.	
2.3 It is logically possible that all of the null hypotheses H_0: $\tau_a = \tau_b$ you counted in 2.1 are true except for the rejected hypotheses in 2.2.	TRUE FALSE
2.4 If exactly one hypothesis H_0: $\tau_a = \tau_b$ were true and all the rest were false, then under model #1 the chance that Tukey’s method rejects the one true hypothesis is at most 0.05.	TRUE FALSE

Last name: __________________ First name: ___________ ID# ________________

PROBLEM SET #3 STATISTICS 500 FALL 2008: ANSWER PAGE 2

3. Create 3 orthogonal contrasts; **see the data page**.

	i=1 Control Grade11+	i=2 Control Grade10-	i=3 Treated Grade11+	i=4 Treated Grade10-
3.1 treatment				
3.2 grade				
3.3 interaction				

3.4 Demonstrate by a calculation that the contrast for grade is orthogonal to the contrast for interaction. Put the calculation in the space at the right.	
3.5 If the interaction contrast among the true parameter values, τ_i, were not zero, a reasonable interpretation is that the effect of the treatment on the change in earnings is different depending upon whether a man has completed 11th grade.	TRUE FALSE

4. Use the contrasts to fill in the following anova table.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-statistic	p-value
Between groups					
Treatment contrast					
Grade contrast					
Interaction Contrast					
Residual within groups				Leave blank	Leave blank

5. For model #1 in the nsw data to answer the following questions about model #1.

Question	CIRCLE ONE
5.1 There are three observations with high leverage (large hatvalues) by our standard.	TRUE FALSE
5.2 There is a statistically significant outlier in the Treated:Grade11+ group whose change in earnings was positive.	TRUE FALSE
5.3 Except perhaps for at most one outlier, the studentized residuals are plausibly Normal.	TRUE FALSE
5.4 Model #1 should be replaced by a similar model for $\log_2$(change)	TRUE FALSE

PROBLEM SET #3 STATISTICS 500 FALL 2008: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due Wednesday 10-Dec-08 at 11:00am.

1. Do a one-way analysis of variance of y=change by the four groups defined by "group" in the nsw data. Use this to answer the following questions.

Question	CIRCLE ONE or Fill in Values
1.1 Test H_0: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ under model #1. What is the **name** of the test statistic? What is the numerical **value** of the test statistic? What is the **p-value**? Is the null hypothesis **plausible**?	Name: F-statistic Value: 3.93 p-value: 0.0088 H_0 is: PAUSIBLE [✓ NOT PLAUSIBLE]
1.2 What is the mean change in each of the four groups? Here Tr is treated, Co is control, Gr11+ is grade 11 or more, Gr10- is grade 10 or less.	TrGr11+ \$6122 TrGr10- \$3385 CoGr11+ \$2387 CoGr10- \$3158
1.3 What is the unbiased **estimate** of σ^2? What is the corresponding **estimate** of σ?	σ^2 : 5.88×10^7 σ : \$7670
1.4 If ε_{ij} were not Normal, then this could invalidate the test you did in 1.1.	[✓ TRUE] FALSE

2. Use Tukey's method to compare every pair of two groups. Use Tukey's method in two-sided comparisons that control the experiment-wise error rate at 0.05.

Identify groups by number,
`i=1 Control:Grade11+`
`i=2 Control:Grade10-`
`i=3 Treated:Grade11+`
`i=4 Treated:Grade10-`
	CIRCLE ONE or Fill in Values

2.1 With four groups, there are how many pairwise tests done by Tukey's method?	How many: 6
2.2 List all pairs (a,b) of null hypotheses, H_0: $\tau_a = \tau_b$ which are rejected by Tukey's method. List as (a,b) where a and b are in {1,2,3,4} with a<b. If none, write "none".	(1,3)
2.3 It is <u>logically possible</u> that all of the null hypotheses H_0: $\tau_a = \tau_b$ you counted in 2.1 are true except for the rejected hypotheses in 2.2.	TRUE [✓ FALSE] If τ_1 does not equal τ_3 then τ_2 cannot equal both τ_1 and τ_3 because τ_2 would have to equal two different things.
2.4 If exactly one hypothesis H_0: $\tau_a = \tau_b$ were true and all the rest were false, then under model #1 the chance that Tukey's method rejects the one true hypothesis is <u>at most</u> 0.05.	[✓ TRUE] FALSE

Last name: ___________________ First name: ___________ ID# ________________

PROBLEM SET #3 STATISTICS 500 FALL 2008: ANSWER PAGE 2

3. Create 3 orthogonal contrasts; **see the data page**.

	i=1 Control Grade11+	i=2 Control Grade10-	i=3 Treated Grade11+	i=4 Treated Grade10-
3.1 treatment	-1	-1	1	1
3.2 grade	1	-1	1	-1
3.3 interaction	-1	1	1	-1

3.4 Demonstrate by a calculation that the contrast for grade is orthogonal to the contrast for interaction. Put the calculation in the space at the right.	$(1x-1)+(-1x1)+(1x1)+(-1x-1)$ $= -1 + -1 + 1 + 1$ $= 0$
3.5 If the interaction contrast among the true parameter values, τ_i, were not zero, a reasonable interpretation is that the effect of the treatment on the change in earnings is different depending upon whether a man has completed 11th grade.	[✓ TRUE] FALSE

4. Use the contrasts to fill in the following anova table.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-statistic	p-value
Between groups	$6.93 x 10^{8}$	3	$2.31 x 10^{8}$	3.93	0.0088
Treatment contrast	$3.4 x 10^{8}$	1	$3.4 x 10^{8}$	5.81	0.016
Grade contrast	$8.4 x 10^{7}$	1	$8.4 x 10^{7}$	1.43	0.232
Interaction Contrast	$2.7 x 10^{8}$	1	$2.7 x 10^{8}$	4.56	0.033
Residual within groups	$2.02 x 10^{10}$	344	$5.88 x 10^{7}$	Leave blank	Leave blank

5. For model #1 in the nsw data to answer the following questions about model #1.

Question	CIRCLE ONE
There are three observations with high leverage (large hatvalues) by our standard.	TRUE [✓ FALSE] All the leverages are equal.
There is a statistically significant outlier in the Treated:Grade11+ group whose change in earnings was positive.	[✓ TRUE] FALSE Wow! rudemt for 43 is 7.58!
Except perhaps for at most one outlier, the studentized residuals are plausibly Normal.	It's not just one outlier! Do a q-q-plot. TRUE [✓ FALSE]
Model #1 should be replaced by a similar model for $\log_2$(change)	Many changes are negative. Can't take log(x) for $x<0$ TRUE [✓ FALSE]

Doing the Problem Set in R

```
> summary(aov(change~group))
             Df   Sum Sq  Mean Sq F value Pr(>F)
group         3 6.93e+08 2.31e+08    3.93 0.0088 **
Residuals   344 2.02e+10 5.88e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> tapply(change,group,mean)
Control:Grade11+ Control:Grade10- Treated:Grade11+ Treated:Grade10-
            2387             3158             6122             3385
> TukeyHSD(aov(change~group))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = change ~ group)
$group
                                    diff      lwr    upr
Control:Grade10--Control:Grade11+   771.5 -2228.98 3771.9
Treated:Grade11+-Control:Grade11+  3735.3   734.88 6735.8
Treated:Grade10--Control:Grade11+   997.8 -2002.63 3998.3
Treated:Grade11+-Control:Grade10-  2963.9   -36.59 5964.3
Treated:Grade10--Control:Grade10-   226.3 -2774.11 3226.8
Treated:Grade10--Treated:Grade11+ -2737.5 -5737.96  262.9

> Treatment<-c(-1,-1,1,1)
> Grade<-c(1,-1,1,-1)
> Interact<-Treatment*Grade
> contrasts(nsw$group)<-cbind(Treatment,Grade,Interact)
> attach(nsw)
> contrasts(group)
                 Treatment Grade Interact
Control:Grade11+        -1     1       -1
Control:Grade10-        -1    -1        1
Treated:Grade11+         1     1        1
Treated:Grade10-         1    -1       -1
> summary(aov(change~group))
             Df     Sum Sq    Mean Sq F value   Pr(>F)
group         3 6.9324e+08 2.3108e+08  3.9327 0.008817 **
Residuals   344 2.0213e+10 5.8759e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> h<-model.matrix(change~group)
> dim(h)
[1] 348   4
> tr<-h[,2]
> grd<-h[,3]
> int<-h[,4]
> anova(lm(change~tr+grd+int))
Analysis of Variance Table
           Df     Sum Sq    Mean Sq F value  Pr(>F)
tr          1 3.4136e+08 3.4136e+08  5.8096 0.01646 *
grd         1 8.4071e+07 8.4071e+07  1.4308 0.23246
int         1 2.6781e+08 2.6781e+08  4.5577 0.03348 *
Residuals 344 2.0213e+10 5.8759e+07
```

```
>  summary(lm(change~tr+grd+int))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3763.0      410.9   9.158   <2e-16 ***
tr             990.4      410.9   2.410   0.0165 *
grd            491.5      410.9   1.196   0.2325
int            877.2      410.9   2.135   0.0335 *
---
Residual standard error: 7665 on 344 degrees of freedom
Multiple R-Squared: 0.03316,    Adjusted R-squared: 0.02473
F-statistic: 3.933 on 3 and 344 DF,  p-value: 0.008817

> which.max(abs(rstudent(lm(change~tr+grd+int))))
43
> nsw[43,]
    treat age edu black hisp married nodegree re74 re75  re78 change
group
132     1  28  11     1    0       0        1    0 1284 60308  59666
Treated:Grade11+
> rstudent(lm(change~tr+grd+int))[43]
  43
7.58
> dim(nsw)
[1] 348  12
> out<-rep(0,348)
> out[43]<-1
> summary(lm(change~tr+grd+int+out))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3607.3      381.4   9.458  < 2e-16 ***
tr             834.8      381.4   2.189   0.0293 *
grd            335.9      381.4   0.881   0.3791
int            721.6      381.4   1.892   0.0593 .
out          54166.3     7145.7   7.580 3.25e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7105 on 343 degrees of freedom
Multiple R-Squared: 0.1719,     Adjusted R-squared: 0.1622
F-statistic:  17.8 on 4 and 343 DF,  p-value: 2.731e-13

> 0.05/348
[1] 0.0001437


> qqnorm(rstudent(lm(change~group)))

> shapiro.test(rstudent(lm(change~group)))

        Shapiro-Wilk normality test

data:  rstudent(lm(change ~ group))
W = 0.8837, p-value = 1.401e-15
```

PROBLEM SET #1 STATISTICS 500 FALL 2010: DATA PAGE 1
Due in class Tuesday 26 October 2010 at noon.
This is an exam. Do not discuss it with anyone.

The data are from a paper: Redding and Strum (2008) The costs of remoteness: evidence from German division and reunification. *American Economic Review*, 98, 1766-1797. You can obtain the paper from the library web-page, but there is no need to do that to do the problem set.

The paper discusses the division of Germany into East and West following the Second World War. Beginning in 1949, economic activity that crossed the East/West divide was suppressed. So a West German city that was close to the East German border was geographically limited in commerce. Redding and Strum were interested in whether such cities had lower population growth than cities far from the East/West boarder.

The data are in the data.frame **gborder**. The outcome is **Y** = g3988, which is the percent growth in population from 1939 to 1988. (Germany reunified in 1990.) The variable dist is a measure of proximity to the East German border. Here, **D** = dist would be 1 if a city were on the border, it is 0 for cities 75 or more kilometers from the border, and in between it is proportional to the distance from the border, so dist=1/2 for a city 75/2 = 37.5 kilometers from the border. Redding and Strum would predict slow population growth for higher values of dist. The variables **Ru** = rubble, **F** = flats and **Re** = refugees describe disruption from World War II. Here, rubble is cubic meters of rubble per capita, flats is the number of destroyed dwellings as a percent of the 1939 stock of dwellings, and refugees is the percent of the 1961 city population that were refugees from eastern Germany. Finally, **G** = g1939 is the percent growth in the population of the city from 1919 to 1939. Also in gborder are the populations and distances used to compute the quantities growth rates and dist variables; for instance, dist_gg_border is the distance in kilometers to the East German border.

```
> dim(gborder)
[1] 122  11
           cities g3988 dist rubble flats refugees g1939
1          Aachen    43    0     21    48       16    11
2          Amberg    33    0      0     1       24    22
3         Ansbach    48    0      3     4       25    26
4   Aschaffenburg    36    0      7    38       19    41
4        Augsburg    32    0      6    24       20    20
```

If you are using R, the data are available on my webpage, http://www-stat.wharton.upenn.edu/~rosenbap/index.html in the object gborder. You will need to download the workspace again. You *may* need to clear your web browser’s cache, so that it gets the new file, rather that using the file already on your computer. In Firefox, this would be Tools -> Clear Private Data and check cache. If you cannot find the gborder object when you download the new R workspace, you probably have not downloaded the new file and are still working with the old one.

PROBLEM SET #1 STATISTICS 500 FALL 2010: DATA PAGE 2
This is an exam. Do not discuss it with anyone.

If you are not using R, the data are available in a .txt file (notepad) at
http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/
as benzene.txt, or
http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/gborder.txt The list of files here is case sensitive, upper case separate from lower case, so benzene.txt is with the lower case files further down. If you cannot find the file, make sure you are looking at the lower case files.

Model #1

$Y = \beta_0 + \beta_1 D + \beta_2 Ru + \beta_3 F + \beta_4 Re + \varepsilon$
or
$g3988 = \beta_0 + \beta_1 dist + \beta_2 Rubble + \beta_3 Flats + \beta_4 Refugees + \varepsilon$
with ε iid $N(0,\sigma^2)$

Model #2

$Y = \gamma_0 + \gamma_1 D + \gamma_2 Ru + \gamma_3 F + \gamma_4 Re + \gamma_5 G + \zeta$
or
$g3988 = \gamma_0 + \gamma_1 dist + \gamma_2 rubble + \gamma_3 flats + \gamma_4 refugees + \gamma_5\, g1939 + \zeta$
with ζ iid $N(0,\omega^2)$

Model #3

$Y = \lambda_0 + \lambda_1 D + \lambda_2 G + \eta$
or
$g3988 = \lambda_0 + \lambda_1 dist + \lambda_2\, g1939 + \eta$
with η iid $N(0,\kappa^2)$

Model 1 has slopes β (beta), while model 2 has slopes γ (gamma), so that different things have different names. The choice of Greek letters is arbitrary.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2010: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

Read the data page and fit model #1. Use model #1 to answer the following parts of question 1 and for this question assume the model is true.	Fill in or CIRCLE the correct answer
1.1 Give the least squares estimate of β_1 the coefficient of D = dist and also the estimated standard error of the estimate of β_1	Standard error: Estimate: ___________ ____________
1.2 Give the numerical value of the estimate of σ	Estimate: ___________
1.3 Do a two-sided test of the null hypothesis H_0: $\beta_1 = 0$. What is the name of the test? What is the value of the test statistic? What is the p-value? Is the null hypothesis plausible?	Name: __________ Value: __________ Circle one p-value: ______ PLAUSIBLE NOT
1.4 Test the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. What is the name of the test? What is the value of the test statistic? What is the p-value? Is the null hypothesis plausible?	Name: __________ Value: __________ Circle one p-value: ______ PLAUSIBLE NOT
1.5 What is the regression sum of squares? What is the residual sum of squares? What percent of the total sum of squares (around the mean) has been fitted by the regression?	Regression SS: _____________ Residual SS: _________ Percent: ____%
1.6 Consider two cities which are the same in terms of Ru = rubble, F=flats and Re=refugees. Suppose that one (near) was at the East/West border and the other (far) was more than 75 kilometers away. For these two cities, model 1 predicts a certain difference, near-minus-far, in their growth (in Y=g3988). What is that predicted difference? (Give a number.)	Difference, near-minus-far: _____________________
1.7 Give the 95% confidence interval for the quantity you estimated in question 1.6. Is it plausible that the difference is zero?	95% CI [,] Circle one PLAUSIBLE NOT
1.8 Which city is closest the East German border? What is the distance in kilometers from that city to the border? What is the actual growth and the fitted growth for that city (Y and fitted Y)?	City name:________ kilometers: ________ Actual Y: ________ fitted Y: __________

Name: _____________________________ ID# _________________________

PROBLEM SET #1 STATISTICS 500 FALL 2010: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

Use model 1 to answer the parts of question 2.	Fill in or CIRCLE the correct answer
2.1 Boxplot the residuals (do not turn in the plot.) Which city has the largest absolute residual? What is the numerical value of that residual and what is its Y?	City name: _____________ Residual: __________ Y:__________
2.2 Do a normal quantile plot of the residuals (do not turn in the plot) and a Shapiro-Wilk test. What is the p-value from the Shapiro-Wilk test? Is it plausible that the residuals are Normal?	P-value: _________________ Circle one PLAUSIBLE NOT

Use model 2 to answer the parts of question 3. For the purpose of question 3, assume model 2 is true.	Fill in or CIRCLE the correct answer
3.1 Model 2 provides strong evidence that cities whose populations grew substantially from 1919 to 1939 continued on to grow substantially more than other cities from 1939 to 1988.	TRUE FALSE
3.2 In model 2, cities with more rubble from the War typically grew more than cities with less rubble, among cities similar in terms of other variables in model 2.	TRUE FALSE
3.3 In model 2, test the hypothesis H_0: $\gamma_2 = \gamma_3 = \gamma_4 = 0$, that is, the coefficients of Ru, F and Re are zero, so these war related variables have zero coefficients. What is the name of the test statistic? What is the numerical value of the test statistic? Give the degrees of freedom for the test. What is the p-value. Is the null hypothesis plausible?	Name: __________ Value: __________ Degrees of freedom: ____________ p-value: ________________ Circle one PLAUSIBLE NOT

4. Fit model 3, assuming it to be true and give a 95% confidence interval for the coefficient λ_1 of D=dist. Is it plausible that this coefficient is zero?	95% CI [,] Circle one PLAUSIBLE NOT

ANSWERS

PROBLEM SET #1 STATISTICS 500 FALL 2010: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

Read the data page and fit model #1. Use model #1 to answer the following parts of question 1 and for this question assume the model is true.	Fill in or CIRCLE the correct answer 7 points each, except 3.3 for 9 points
1.1 Give the least squares estimate of β_1 the coefficient of D = dist and also the estimated standard error of the estimate of β_1	Standard error: Estimate: *-51.2* *20.5*
1.2 Give the numerical value of the estimate of σ	Estimate: *48.3*
1.3 Do a two-sided test of the null hypothesis H_0: $\beta_1 = 0$. What is the name of the test? What is the value of the test statistic? What is the p-value? Is the null hypothesis plausible?	Name: *t-test* Value: *-2.49* Circle one p-value: *0.014* PLAUSIBLE [✓ NOT]
1.4 Test the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. What is the name of the test? What is the value of the test statistic? What is the p-value? Is the null hypothesis plausible?	Name: *F-test* Value: *8.5* Circle one p-value: $4.6\text{x}10^{-6}$ PLAUSIBLE [✓ NOT]
1.5 What is the regression sum of squares? What is the residual sum of squares? What percent of the total sum of squares (around the mean) has been fitted by the regression?	Regression SS: *79,369* Residual SS: *272,827* Percent: *22.5%* *The percent is R^2*
1.6 Consider two cities which are the same in terms of Ru = rubble, F=flats and Re=refugees. Suppose that one (near) was at the East/West border and the other (far) was more than 75 kilometers away. For these two cities, model 1 predicts a certain difference, near-minus-far, in their growth (in Y=g3988). What is that predicted difference? (Give a number.)	Difference, near-minus-far: *-51.2*
1.7 Give the 95% confidence interval for the quantity you estimated in question 1.6. Is it plausible that the difference is zero?	95% CI [*-91.85* , *-10.55*] Circle one PLAUSIBLE [✓ NOT]
1.8 Which city is closest the East German border? What is the distance in kilometers from that city to the border? What is the actual growth and the fitted growth for that city (Y and fitted Y)?	City name: *Luebeck* kilometers: *5.4* Actual Y: *35.9%* fitted Y: *65.75%*

ANSWERS

PROBLEM SET #1 STATISTICS 500 FALL 2010: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

Use model 1 to answer the parts of question 2.	Fill in or CIRCLE the correct answer
2.1 Boxplot the residuals (do not turn in the plot.) Which city has the largest absolute residual? What is the numerical value of that residual and what is its Y?	City name: *Hamm* Residual: *144.5* Y: *191.4%*
2.2 Do a normal quantile plot of the residuals (do not turn in the plot) and a Shapiro-Wilk test. What is the p-value from the Shapiro-Wilk test? Is it plausible that the residuals are Normal?	P-value: *0.0000187* Circle one PLAUSIBLE [✓ NOT]

Use model 2 to answer the parts of question 3. For the purpose of question 3, assume model 2 is true.	Fill in or CIRCLE the correct answer
3.1 Model 2 provides strong evidence that cities whose populations grew substantially from 1919 to 1939 continued on to grow substantially more than other cities from 1939 to 1988.	TRUE [✓ FALSE]
3.2 In model 2, cities with more rubble from the War typically grew more than cities with less rubble, among cities similar in terms of other variables in model 2.	TRUE [✓ FALSE]
3.3 In model 2, test the hypothesis H_0: $\gamma_2=\gamma_3=\gamma_4=0$, that is, the coefficients of Ru, F and Re are zero, so these war related variables have zero coefficients. What is the name of the test statistic? What is the numerical value of the test statistic? Give the degrees of freedom for the test. What is the p-value. Is the null hypothesis plausible?	*An F statistic has both numerator and denominator degrees of freedom!* Name: *(partial)-F-test* Value: *11.19* Degrees of freedom: *3 and 116* p-value: *0.00000167* Circle one PLAUSIBLE [✓ NOT]

4. Fit model 3, assuming it to be true and give a 95% confidence interval for the coefficient λ_1 of D=dist. Is it plausible that this coefficient is zero?	95% CI [*-70.0* , *15.8*] Circle one [✓ PLAUSIBLE] NOT

PROBLEM SET #1 STATISTICS 500 FALL 2010:
Doing the problem set in R

```
Question 1.
> summary(lm(g3988~dist+rubble+flats+refugees))
Call:lm(formula = g3988 ~ dist + rubble + flats + refugees)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.6381    19.6162   0.950 0.344000
dist        -51.2013    20.5269  -2.494 0.014015 *
rubble       -2.2511     0.7224  -3.116 0.002304 **
flats         0.3618     0.2700   1.340 0.182845
refugees      2.5433     0.7121   3.572 0.000516 ***

Residual standard error: 48.29 on 117 degrees of freedom
Multiple R-squared: 0.2254,     Adjusted R-squared: 0.1989
F-statistic: 8.509 on 4 and 117 DF,  p-value: 4.616e-06
```

```
Question 1.5
> anova(lm(g3988~1),lm(g3988~dist+rubble+flats+refugees))
Analysis of Variance Table
Model 1: g3988 ~ 1
Model 2: g3988 ~ dist + rubble + flats + refugees
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    121 352196
2    117 272827  4     79369 8.5092 4.616e-06 ***
```

```
Question 1.7
> confint(lm(g3988~dist+rubble+flats+refugees))
                  2.5 %      97.5 %
(Intercept) -20.2107139  57.4869120
dist        -91.8538551 -10.5487643
rubble       -3.6817031  -0.8205187
flats        -0.1729454   0.8966347
refugees      1.1331156   3.9534813
```

```
Question 1.8
> which.min(dist_gg_border)
[1] 73
> gborder[73,]
    cities    g3988  dist rubble flats refugees    g1939 pop1988 pop1939
pop1919 dist_gg_border
73 Luebeck 35.90766 0.928    4.5  19.6     38.4 36.87975  210400  154811
113100            5.4
> mod<-lm(g3988~dist+rubble+flats+refugees)
> mod$fit[73]
     73
65.7481
```

```
Question 2.1
> boxplot(mod$resid)
> which.max(abs(mod$resid))
50
50
> mod$resid[50]
      50
144.4552
> gborder[50,]
   cities    g3988 dist rubble flats refugees    g1939 pop1988 pop1939 pop1919
dist_gg_border
50   Hamm 191.3526    0   20.3  60.3     20.5 28.89738  172000   59035   45800
152
```

PROBLEM SET #1 STATISTICS 500 FALL 2010:
Doing the problem set in R, continued

```
Question 2.2
> qqnorm(mod$resid)
> shapiro.test(mod$resid)
        Shapiro-Wilk normality test
data:  mod$resid
W = 0.9356, p-value = 1.873e-05

Question 3.
> summary(lm(g3988~dist+rubble+flats+refugees+g1939))
Call:
lm(formula = g3988 ~ dist + rubble + flats + refugees + g1939)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.7149    19.8866   1.142 0.255715
dist        -52.7980    20.5376  -2.571 0.011411 *
rubble       -2.2746     0.7214  -3.153 0.002058 **
flats         0.3774     0.2699   1.398 0.164728
refugees      2.7511     0.7324   3.756 0.000271 ***
g1939        -0.2559     0.2171  -1.179 0.240865

Residual standard error: 48.21 on 116 degrees of freedom
Multiple R-squared: 0.2345,     Adjusted R-squared: 0.2015
F-statistic: 7.108 on 5 and 116 DF,  p-value: 7.837e-06

Question 3.3
> anova(lm(g3988~dist+g1939),lm(g3988~dist+rubble+flats+refugees+g1939))
Analysis of Variance Table
Model 1: g3988 ~ dist + g1939
Model 2: g3988 ~ dist + rubble + flats + refugees + g1939
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    119 347615
2    116 269597  3     78018 11.190 1.669e-06 ***
```

This F-test has 3 and 116 degrees of freedom.

```
Question 4
> confint(lm(g3988~dist+g1939))
                  2.5 %     97.5 %
(Intercept)  42.2449698 80.1674037
dist        -70.0433787 15.8198535
g1939        -0.4506696  0.4818223
```

PROBLEM SET #2 STATISTICS 500 FALL 2010: DATA PAGE 1
Due in class Thursday 2 December 2010 at noon.
This is an exam. Do not discuss it with anyone.

The data are the same as in Problem 1, from Redding and Strum (2008) The costs of remoteness: evidence from German division and reunification. *American Economic Review*, 98, 1766-1797. You can obtain the paper from the library web-page, but there is no need to do that to do the problem set.

The paper discusses the division of Germany into East and West following the Second World War. Beginning in 1949, economic activity that crossed the East/West divide was suppressed. So a West German city that was close to the East German border was geographically limited in commerce. Redding and Strum were interested in whether such cities had lower population growth than cities far from the East/West boarder.

The data are in the data.frame **gborder**. The outcome is **Y** = g3988, which is the percent growth in population from 1939 to 1988. (Germany reunified in 1990.) The variable dist is a measure of proximity to the East German border. Here, **D** = dist would be 1 if a city were on the border, it is 0 for cities 75 or more kilometers from the border, and in between it is proportional to the distance from the border, so dist=1/2 for a city 75/2 = 37.5 kilometers from the border. Redding and Strum would predict slow population growth for higher values of dist. The variables **Ru** = rubble, **F** = flats and **Re** = refugees describe disruption from World War II. Here, rubble is cubic meters of rubble per capita, flats is the number of destroyed dwellings as a percent of the 1939 stock of dwellings, and refugees is the percent of the 1961 city population that were refugees from eastern Germany. Finally, **G** = g1939 is the percent growth in the population of the city from 1919 to 1939. Also in gborder are the populations and distances used to compute the quantities growth rates and dist variables; for instance, dist_gg_border is the distance in kilometers to the East German border.

```
> dim(gborder)
[1] 122  11
           cities g3988 dist rubble flats refugees g1939
1          Aachen    43    0     21    48       16    11
2          Amberg    33    0      0     1       24    22
3         Ansbach    48    0      3     4       25    26
4   Aschaffenburg    36    0      7    38       19    41
4        Augsburg    32    0      6    24       20    20
```

If you are using R, the data are available on my webpage, http://www-stat.wharton.upenn.edu/~rosenbap/index.html in the object gborder. You will need to download the workspace again. You *may* need to clear your web browser's cache, so that it gets the new file, rather that using the file already on your computer. In Firefox, this would be Tools -> Clear Private Data and check cache. If you cannot find the gborder object when you download the new R workspace, you probably have not downloaded the new file and are still working with the old one.

If you are not using R, the data are available in a .txt file (notepad) at
http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/
as benzene.txt, or
http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/gborder.txt The list of files here is case sensitive, upper case separate from lower case, so benzene.txt is with the lower case files further down. If you cannot find the file, make sure you are looking at the lower case files.

PROBLEM SET #2 STATISTICS 500 FALL 2010: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

In the current analysis, we will follow the paper more closely than we did in Problem 1. They used a coded variable for proximity to the East/West German border, specifically 1 if within 75 KM of the border, 0 otherwise. In R, create the variable as follows:

```
> border<-1*(gborder$dist_gg_border<=75)
> gborder<-cbind(gborder,border)
> border[1:10]
 [1] 0 0 0 0 0 0 0 1 1 0
> rm(border)
> attach(gborder)
```

Model #A

Y $= \beta_0 + \beta_1$border + β_2Ru + β_3F + β_4Re + ε with ε iid $N(0,\sigma^2)$

or g3988= $\beta_0 + \beta_1$ border + β_2Rubble + β_3Flats + β_4Refugees + ε

For question 1.2, the reasons are:

A: This city grew the most between 1939 and 1988.

B: This city was high on rubble but not high on flats or refugees.

C: This city has an unusual value of refugees.

D: The growth of this city from 1939 to 1988 does not fit with its value of refugees.

For question 1.7, the descriptions are:

a: This growth rate for this city lies above the regression plane and it raises its own predicted value by more than 1 standard error.

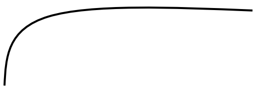
b: This growth rate for this city lies below the regression plane and it lowers its own predicted value by more than 1 standard error.

c: This growth rate for this city lies above the regression plane and it raises its own predicted value by less than 1 standard error.

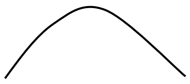
d: This growth rate for this city lies below the regression plane and it lowers its own predicted value by less than 1 standard error.

For question 2.1, the shapes are:

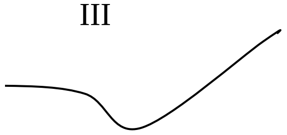
I



II



III



Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name: ______________________________ ID# __________________________

PROBLEM SET #2 STATISTICS 500 FALL 2010: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

Fit model A and use it to answer the following questions.	Fill in or circle the correct answer.
1.1 In model A, which city has the largest leverage (or hat value or h_i or Sheather's h_{ii})? (Give the name of the city.) What is the numerical value of h_i? What is the numerical value of the cut-off for judging whether h_i is large? Is it large?	City: ________________ h_i = ____________ cut-off = ___________ LARGE NOT LARGE
1.2 From the reasons listed on the data page, write in the letter (A or B or C or D) of the one best reason for what you found in 1.1.	Letter of one best reason: _______________________
1.3 Test the null hypothesis that the residuals of model A are Normal. What is the name of the test? What is the p-value? Is it plausible that the residuals are Normal?	Name:___________ P-value: __________ PAUSIBLE NOT PLAUSIBLE
1.4 In model A, which city has the largest absolute studentized residual? Give the name of the city and the numerical value with sign of this studentized residual.	City: ____________ Value: ___________
1.5 Is the city you identified in 1.4 a statistically significant outlier at the 0.05 level? How large would the absolute value of the studentized residual have to be to be significant as an outlier at the 0.05 level? Give a number.	OUTLIER NOT AN OUTLIER How large: ________________
1.6 In model A, which city has the largest absolute dffits? Name the city. What is the numerical value (with sign) of this dffits?	City: ____________ Value: ___________
1.7 Select the one letter of the one best description on the data page for what you found in 1.6. Give one letter.	Letter: ____________
1.8 Test for nonlinearity in model A using Tukey's one-degree of freedom. Give the t-statistic and the p-value. Does this test reject the linear model at the 0.05 level?	t-statistic _________ p-value: _________ REJECTS AT 0.05 DOES NOT

Name: ______________________________ ID# __________________________

PROBLEM SET #2 STATISTICS 500 FALL 2010: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

	Fill in or circle the correct answer.
2.1 The estimated coefficient for refugees in model A is 2.68 suggesting that more refugees from Eastern Germany is associated with more rapid growth of population. Test for parallelism in this slope for cities near (border =1) and far from (border = 0) the border. Give the name and value of the test statistic and the p-value. Is parallelism plausible?	Name: _________ Value: _________ P-value: _________ PLAUSIBLE NOT PLAUSIBLE
2.2 In 2.1, whether or not the parallelism is rejected, look at the *estimated* slopes of the two fitted nonparallel lines. Based on the *point estimates* of slopes, is the estimated slope near the border (border = 1) steeper upwards than the estimated slope far from the border (border = 0)?	YES NO
2.3 Plot the residuals for model A (as Y vertical) against flats (as X horizontal). Add a lowess curve to the plot. Which of the 3 shapes on the data page does the lowess plot most closely resemble? Give one Roman numeral, I, II or III. (In R, use the default settings for lowess.)	Roman numeral: ______________
2.4 Center flats at its mean and square the result. Add this centered quadratic term to model A. Test the null hypothesis that model A is correct in specifying a linear relationship between population growth and flats against the alternative that it is quadratic. Give the name and value of the test statistic and the p-value. Is a linearity plausible?	Name: _________ Value: _________ P-value: _________ PLAUSIBLE NOT PLAUSIBLE
2.5 Give the multiple squared correlation, R^2, for model A and the model in 2.4, and the estimate of the standard deviation, σ, of the true errors.	R^2 estimate of σ Model A _____________ ___________ Model in 2.4 ___________ ___________

ANSWERS

PROBLEM SET #2 STATISTICS 500 FALL 2010: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

Fit model A and use it to answer the following questions.	Fill in or circle the correct answer.
1.1 In model A, which city has the largest leverage (or hat value or h_i or Sheather's h_{ii})? (Give the name of the city.) What is the numerical value of h_i? What is the numerical value of the cut-off for judging whether h_i is large? Is it large?	City: *Datteln* h_i = *0.173* cut-off = *0.082* [✓ LARGE] NOT LARGE
1.2 From the reasons listed on the data page, write in the letter (A or B or C or D) of the one best reason for what you found in 1.1.	Letter of one best reason: *B* *Plot Y=flats versus X=rubble and find Datteln.*
1.3 Test the null hypothesis that the residuals of model A are Normal. What is the name of the test? What is the p-value? Is it plausible that the residuals are Normal?	Name:*Shapiro-Wilk* P-value: *0.0000108* PAUSIBLE [✓ NOT PLAUSIBLE]
1.4 In model A, which city has the largest absolute studentized residual? Give the name of the city and the numerical value with sign of this studentized residual.	City: *Hamm* Value: *3.088*
1.5 Is the city you identified in 1.4 a statistically significant outlier at the 0.05 level? How large would the absolute value of the studentized residual have to be to be significant as an outlier at the 0.05 level? Give a number.	OUTLIER [✓ NOT AN OUTLIER] How large: *>= 3.639* *122 tests, each 2-sided, with 116 df* *qt(1-0.025/122, 116)*
1.6 In model A, which city has the largest absolute dffits? Name the city. What is the numerical value (with sign) of this dffits?	City: *Moers* Value: *1.0763*
1.7 Select the one letter of the one best description on the data page for what you found in 1.6. Give one letter.	Letter: *a. 1.0763 is positive, so above, pulling up. Value is >1, so more than 1 standard error.*
1.8 Test for nonlinearity in model A using Tukey's one-degree of freedom. Give the t-statistic and the p-value. Does this test reject the linear model at the 0.05 level?	t-statistic *1.82* p-value: *0.072* REJECTS AT 0.05 [✓ DOES NOT] *Close, but not quite.*

Name: ______________________________ ID# __________________________

PROBLEM SET #2 STATISTICS 500 FALL 2010: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

	Fill in or circle the correct answer.
2.1 The estimated coefficient for refugees in model A is 2.68 suggesting that more refugees from Eastern Germany is associated with more rapid growth of population. Test for parallelism in this slope for cities near (border =1) and far from (border = 0) the border. Give the name and value of the test statistic and the p-value. Is parallelism plausible?	Name: *t-statistic* Value: *-1.41* P-value: *0.16* [✓ PLAUSIBLE] NOT PLAUSIBLE
2.2 In 2.1, whether or not the parallelism is rejected, look at the *estimated* slopes of the two fitted nonparallel lines. Based on the *point estimates* of slopes, is the estimated slope near the border (border = 1) steeper upwards than the estimated slope far from the border (border = 0)?	YES [✓ NO] *No, it's steeper away from the border, but from 2.1, it is not significantly different.*
2.3 Plot the residuals for model A (as Y vertical) against flats (as X horizontal). Add a lowess curve to the plot. Which of the 3 shapes on the data page does the lowess plot most closely resemble? Give one Roman numeral, I, II or III. (In R, use the default settings for lowess.)	Roman numeral: *III*
2.4 Center flats at its mean and square the result. Add this centered quadratic term to model A. Test the null hypothesis that model A is correct in specifying a linear relationship between population growth and flats against the alternative that it is quadratic. Give the name and value of the test statistic and the p-value. Is a linearity plausible?	Name: *t-statistic* Value: 3.567 P-value: *0.000527* PLAUSIBLE [✓ NOT PLAUSIBLE]
2.5 Give the multiple squared correlation, R^2, for model A and the model in 2.4, and the estimate of the standard deviation, σ, of the true errors.	R^2 estimate of σ Model A *0.230* *48.14* Model in 2.4 *0.306* *45.9*

Problem Set 2, Fall 2010 Doing the Problem Set in R

```
> modA<-lm(g3988 ~ border + rubble + flats + refugees)
> modA
 (Intercept)       border       rubble        flats     refugees
    14.4864     -32.9477      -2.1635       0.4005       2.6808
```

1.1 leverage

```
> which.max(hatvalues(modA))
22
> hatvalues(modA)[22]
0.1733585
> 2*mean(hatvalues(modA))
[1] 0.08196721
> 2*5/122
[1] 0.08196721
> gborder[22,]
    cities    g3988 dist rubble flats refugees border
22 Datteln 79.24853    0   32.7  20.4     20.1      0
```

1.2 Looking at and understanding a high leverage point

```
> summary(gborder)
> plot(rubble,flats)
> abline(v=32.7)
> abline(h=20.4)
```

1.3 Test for normality

```
> shapiro.test(modA$resid)
        Shapiro-Wilk normality test
W = 0.9319, p-value = 1.079e-05
```

1.4 Studentized residual

```
> which.max(abs(rstudent(modA)))
50
> rstudent(modA)[50]
      50
3.087756
> gborder[50,]
   cities    g3988 dist rubble flats refugees  border
50   Hamm 191.3526    0   20.3  60.3     20.5    0
```

1.5 Outlier test

```
> qt(.025/122,116)
[1] -3.63912
> qt(1-0.025/122,116)
[1] 3.63912
```

Problem Set 2, Fall 2010, continued

1.6 and 1.7 dffits

```
> which.max(abs(dffits(modA)))
81
> dffits(modA)[81]
1.076284
> gborder[81,]
   cities    g3988 dist rubble flats refugees    border
81  Moers 241.6411    0    1.6  75.7     25.3     0
```

1.8 Tukey’s one degree of freedom for nonadditivity

```
> tk<-tukey1df(modA)
> summary(lm(g3988 ~ border + rubble + flats + refugees+tk))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.3745    19.6169   0.529 0.597914
border      -39.0843    12.7927  -3.055 0.002791 **
rubble       -2.4323     0.7314  -3.326 0.001181 **
flats         0.4130     0.2644   1.562 0.120985
refugees      2.7826     0.7163   3.885 0.000171 ***
tk            0.9209     0.5064   1.819 0.071525 .
```

2.1 and 2.2 Testing parallelism

```
> brinteraction<-border*refugees
> summary(lm(g3988 ~ border + rubble + flats + refugees +
brinteraction))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.5808    21.3277   0.121 0.903896
border         18.5806    38.5088   0.483 0.630358
rubble         -2.2582     0.7233  -3.122 0.002269 **
flats           0.4285     0.2665   1.608 0.110571
refugees        3.2569     0.8256   3.945 0.000137 ***
brinteraction  -2.0877     1.4770  -1.413 0.160193
```

Slope near border estimated to be

```
> 3.2569+(-2.0877)
[1] 1.1692
```

so it is steeper (3.26) far from the border and shallower (1.17) near the border.

```
2.3 Looking for curves
> plot(flats,modA$resid)
> lines(lowess(flats,modA$resid))

2.4 Quadratic in flats
> flatsc2<-(flats-mean(flats))^2
> summary(lm(g3988 ~ border + rubble + flats + refugees+flatsc2))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.98569   18.78579   0.957 0.340352
border      -36.45579   11.92064  -3.058 0.002765 **
rubble       -2.10195    0.68979  -3.047 0.002860 **
flats         0.12088    0.26626   0.454 0.650671
refugees      2.42524    0.69118   3.509 0.000641 ***
flatsc2       0.02318    0.00650   3.567 0.000527 ***


2.5
> summary(modA)
Residual standard error: 48.14 on 117 degrees of freedom
Multiple R-squared: 0.2302,     Adjusted R-squared: 0.2038
F-statistic: 8.745 on 4 and 117 DF,  p-value: 3.271e-06

> modB<-lm(g3988 ~ border + rubble + flats + refugees+flatsc2)
> summary(modB)
Residual standard error: 45.9 on 116 degrees of freedom
Multiple R-squared: 0.3062,     Adjusted R-squared: 0.2763
F-statistic: 10.24 on 5 and 116 DF,  p-value: 3.796e-08
```

PROBLEM SET #3 STATISTICS 500 FALL 2010: DATA PAGE 1
Due in Monday 20 December 2010 at noon.
This is an exam. Do not discuss it with anyone.

The first part of this problem set again uses the data from Problems 1 and 2, from Redding and Strum (2008) The costs of remoteness: evidence from German division and reunification. *American Economic Review*, 98, 1766-1797. You can obtain the paper from the library web-page, but there is no need to do that to do the problem set.

The paper discusses the division of Germany into East and West following the Second World War. Beginning in 1949, economic activity that crossed the East/West divide was suppressed. So a West German city that was close to the East German border was geographically limited in commerce. Redding and Strum were interested in whether such cities had lower population growth than cities far from the East/West boarder.

The data for the first part are in the data.frame **gborder**. The outcome is **Y** = g3988, which is the percent growth in population from 1939 to 1988. (Germany reunified in 1990.) The variable dist is a measure of proximity to the East German border. Here, **D** = dist would be 1 if a city were on the border, it is 0 for cities 75 or more kilometers from the border, and in between it is proportional to the distance from the border, so dist=1/2 for a city 75/2 = 37.5 kilometers from the border. Redding and Strum would predict slow population growth for higher values of dist. The variables **Ru** = rubble, **F** = flats and **Re** = refugees describe disruption from World War II. Here, rubble is cubic meters of rubble per capita, flats is the number of destroyed dwellings as a percent of the 1939 stock of dwellings, and refugees is the percent of the 1961 city population that were refugees from eastern Germany. Finally, **G** = g1939 is the percent growth in the population of the city from 1919 to 1939. The actual distance to the border with East Germany is **Ad** =dist_gg_border.

```
> dim(gborder)
[1] 122  11
```

In R, you will want the **leaps package** for variable selection and the **DAAG package** for press. The first time you use these packages, you must install them at the Package menu. Every time you use these packages, including the first time, you must load them at the Packages menus.

If you are using R, the data are available on my webpage, http://www-stat.wharton.upenn.edu/~rosenbap/index.html in the objects gborder and pku. You will need to download the workspace again. You *may* need to clear your web browser's cache, so that it gets the new file, rather than using the file already on your computer. In Firefox, this would be Tools -> Clear Private Data and check cache. If you cannot find the gborder object when you download the new R workspace, you probably have not downloaded the new file and are still working with the old one.

If you are not using R, the data are available in a .txt file (notepad) at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` gborder.txt and pku.txt. The list of files here is case sensitive, upper case separate from lower case, so pku.txt is with the lower case files further down. If you cannot find the file, make sure you are looking at the lower case files.

There are three options about **turning in the exam**. (i) You can deliver it to my office 473 JMHH on Monday 20 December at noon. Any time before noon on Monday 20 December, you can (ii) place it in a sealed envelope addressed to me and leave it in my mail box in the statistics department, 4th floor JMHH, or (iii) you can leave it with Adam at the front desk in the statistics department. Make and keep a photocopy of your answer page – if something goes wrong, I can grade the photocopy. The statistics department is locked at night and on the weekend. Your **course grade** will be available from the registrar shortly after I grade the finals. I will put the **answer key** in an updated version of the bulkpack on my web page shortly after I grade the final.

PROBLEM SET #3 STATISTICS 500 FALL 2010: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

In the current analysis, we will follow the paper more closely than we did in Problem 1. They used a coded variable for proximity to the East/West German border, specifically 1 if within 75 KM of the border, 0 otherwise. In R, create the variable as follows:

```
> border<-1*(gborder$dist_gg_border<=75)
> attach(gborder)
```

Model #A

$Y = \beta_0 + \beta_1\text{border} + \beta_2 D + \beta_3 Ad + \beta_4 Ru + \beta_5 F + \beta_6 Re + \beta_7 G + \varepsilon$ with ε iid $N(0,\sigma^2)$

or $g3988 = \beta_0 + \beta_1\text{border} + \beta_2 D + \beta_3 Ad + \beta_4\text{Rubble} + \beta_5\text{Flats} + \beta_6\text{Refugees} + \beta_7 g1939 + \varepsilon$

Model #B

$Y = \gamma_0 + \gamma_1 Ru + \gamma_2 Re + \xi$ with ξ iid $N(0,\omega^2)$

Model #C

$Y = \theta_0 + \theta_1\text{border} + \theta_2 Ru + \theta_3 F + \theta_4 Re + \theta_5 G + \zeta$ with ζ iid $N(0,\upsilon^2)$

Use Y, border, D, Ad, Ru, F, Re and G to refer to specific variables.

The second data set is adapted from Sitta, A. et al. (2009) Evidence that DNA damage is associated to phenylalanine blood levels in leukocytes from phenylketonuric patients. Mutation Research, 679, 13-16. Again, you may look at the paper on the library web-page, but you do not need to do so for this problem. To simplify this problem set, the data are adapted to make a balanced design, 8 people per group, selected from the paper's unbalanced design by simple random sampling. They studied a genetic disorder, phenylketonuria (PKU), which affects the metabolism of phenylalanine (Phe). The study has three groups, a unaffected control group with 8 people, and two groups of 8 individuals with PKU. The two groups of people with PKU are distinguished by the level of Phe in their blood, highPhe is > 600 μ mol/L, lowPhe is < = 600 μ mol/L. The outcome, Y = DI, is a measure of genetic damage in certain blood cells, the comet tail assay from leukocytes. So you are to use two variables, Y = DI and group, in the object **pku** in the R workspace.

Model #D

$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$ with ε iid $N(0,\sigma^2)$ $i=1,\ldots,8$, $j=1,2,3$, with $\tau_1 + \tau_2 + \tau_3 = 0$.

In answering questions, refer to groups as "control", "lowPhe" or "highPhe".

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name: ______________________________ ID# __________________________

PROBLEM SET #3 STATISTICS 500 FALL 2010: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

Fit model A. **Use Y, border, D, Ad, Ru, F, Re and G to refer to specific variables**	Fill in or CIRCLE the correct answer
1.1 If you were to remove all variables from model A with t-statistics that were not significant in a 2-sided, 0.05 level test, which would you remove?	Give names of variables removed:
1.2 Test the null hypothesis that model B is an adequate model against the alternative that model A is better. Give the name and value of the test statistic, the degrees of freedom, the p-value. Is the null hypothesis plausible?	Name: _____________ Value: __________ Degrees of freedom: ______ P-value: ____ PLAUSIBLE NOT PLAUSIBLE
1.3 Including the empty model with no variables and model A itself, how many models can be formed from model A by deleting 0, 1, …, or 7 variables?	Number of models: _____________
1.4 Of the models in part 1.3 above, which one model has the smallest C_P statistic? List the variables included in this model.	Give names of variables in this model:
1.5 What is the numerical value of C_P for the model you identified in 1.4? If the model in 1.4 contained all of the variables with nonzero coefficients, what number would C_P be estimating? Give one number.	Value of C_P: _______________ What number would C_P be estimating? Number: ________________
1.6 Is there another model with the same number of variables as the model in 1.4 but with different variables such that the value of C_P for this other model is also consistent with this other model containing all the variables with nonzero coefficients? Circle YES or NO. If YES, then give the value of C_P and the predictor variables in this model. If NO, leave other items blank.	YES NO Value of C_P: _______________ Give names of variables in this model:
1.7 Give PRESS and C_P values for model A and C. Also, give the number of coefficients (including the constant) in these two models. If these estimates were not estimates but true values of what they estimate, which model, A or C, would predict better? CIRCLE A or C.	Model A Model C PRESS ___________ __________ C_P ___________ __________ # coeffs ___________ __________ Better Predicts A C

Name: ______________________________ ID# __________________________

PROBLEM SET #3 STATISTICS 500 FALL 2010: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

Use the pku data and Model D for these questions. Refer to groups as "control", "lowPhe" or "highPhe".	Fill in or CIRCLE the correct answer
2.1 Fit model D and test the null hypothesis that its residuals are Normal. What is the name of the test? What is the P-value? Is the null hypothesis plausible?	Name:____________ P-value: ________ PLAUSIBLE NOT PLAUSIBLE
2.2 In model D, test the null hypothesis that H_0: $\tau_1 = \tau_2 = \tau_3 = 0$. Give the name and value of the test-statistic, the degrees of freedom, the P-value, and state whether the null hypothesis is plausible.	Name:____________ Value: ________ Degrees of freedom: _____ P-value: _____ PLAUSIBLE NOT PLAUSIBLE
2.3 Test the three null hypotheses, H_{12}: $\tau_1 = \tau_2$, H_{13}: $\tau_1 = \tau_3$ and H_{23}: $\tau_2 = \tau_3$ using Tukey's method at the two-sided 0.05 level. List those hypotheses that are rejected by this method. That is, list H_{12} and/or H_{13} and/or H_{23} or write NONE.	
2.4 If model D were true and H_{12} were true but H_{13} and H_{23} were false, then the chance that Tukey's method in 2.3 will reject at least one of the hypotheses H_{12}: $\tau_1 = \tau_2$, H_{13}: $\tau_1 = \tau_3$ and H_{23}: $\tau_2 = \tau_3$ is at most 0.05 despite testing three hypotheses.	TRUE FALSE
2.5 Give two orthogonal contrasts with integer weights to test the two hypotheses that: H_C control does not differ from the average of the two PKU groups and H_{hl} that high and low Phe groups do not differ. Fill in 6 integer values.	Group control lowPhe highPhe H_C _____ ______ ______ H_{hl} _____ ______ ______

3. Use model D and the contrasts in 2.5 to fill in the following anova table.

Source	Sum of squares	Degrees of freedom	Mean Square	F-statistic
Between groups				
Contrast H_C				
Contrast H_{hl}				
Within groups				

PROBLEM SET #3 STATISTICS 500 FALL 2010: ANSWERS

Fit model A. **Use Y, border, D, Ad, Ru, F, Re and G to refer to specific variables**	Fill in or CIRCLE the correct answer Use **Ru** not **Rubble** as a variable name.
1.1 If you were to remove all variables from model A with t-statistics that were not significant in a 2-sided, 0.05 level test, which would you remove?	Give names of variables removed: *border, D, Ad, F and G.*
1.2 Test the null hypothesis that model B is an adequate model against the alternative that model A is better. Give the name and value of the test statistic, the degrees of freedom, the p-value. Is the null hypothesis plausible?	Name: *F-statistic* Value: *2.4199* Degrees of freedom: *5 and 114* P-value: *0.0399* PLAUSIBLE [✓ NOT PLAUSIBLE]
1.3 Including the empty model with no variables and model A itself, how many models can be formed from model A by deleting 0, 1, …, or 7 variables?	Number of models: $2^7=128$
1.4 Of the models in part 1.3 above, which one model has the smallest C_P statistic? List the variables included in this model.	Give names of variables in this model: *border, Ru, F, Re*
1.5 What is the numerical value of C_P for the model you identified in 1.4? If the model in 1.4 contained all of the variables with nonzero coefficients, what number would C_P be estimating? Give one number.	Value of C_P: *4.008623* What number would C_P be estimating? Number: *5*
1.6 Is there another model with the same number of variables as the model in 1.4 but with different variables such that the value of C_P for this other model is also consistent with this other model containing all the variables with nonzero coefficients? Circle YES or NO. If YES, then give the value of C_P and the predictor variables in this model. If NO, leave other items blank.	[✓ YES] NO Value of C_P: *4.733086* Give names of variables in this model: *D, Ru, F, Re*
1.7 Give PRESS and C_P values for model A and C. Also, give the number of coefficients (including the constant) in these two models. If these estimates were not estimates but true values of what they estimate, which model, A or C, would predict better? CIRCLE A or C.	Model A Model C PRESS 314,191.8 298,914.3 C_P 8.000 4.761 # coeffs 8 6 Better Predicts A [✓ C]

PROBLEM SET #3 STATISTICS 500 FALL 2010: ANSWER PAGE 2**.**

Use the pku data and Model D for these questions. Refer to groups as "control", "lowPhe" or "highPhe".	Fill in or CIRCLE the correct answer
2.1 Fit model D and test the null hypothesis that its residuals are Normal. What is the name of the test? What is the P-value? Is the null hypothesis plausible?	Name: *Shapiro-Wilk test* P-value: 0.42 [✓ PLAUSIBLE] NOT PLAUSIBLE
2.2 In model D, test the null hypothesis that H_0: $\tau_1 = \tau_2 = \tau_3 = 0$. Give the name and value of the test-statistic, the degrees of freedom, the P-value, and state whether the null hypothesis is plausible.	Name: *F-statistic* Value: *68.3* Degrees of freedom: *2 and 21* P-value: *6.4 x* 10^{-10} PLAUSIBLE [✓ NOT PLAUSIBLE]
2.3 Test the three null hypotheses, H_{12}: $\tau_1 = \tau_2$, H_{13}: $\tau_1 = \tau_3$ and H_{23}: $\tau_2 = \tau_3$ using Tukey's method at the two-sided 0.05 level. List those hypotheses that are rejected by this method. That is, list H_{12} and/or H_{13} and/or H_{23} or write NONE.	H_{12} and H_{13} and H_{23}
2.4 If model D were true and H_{12} were true but H_{13} and H_{23} were false, then the chance that Tukey's method in 2.3 will reject at least one of the hypotheses H_{12}: $\tau_1 = \tau_2$, H_{13}: $\tau_1 = \tau_3$ and H_{23}: $\tau_2 = \tau_3$ is at most 0.05 despite testing three hypotheses.	TRUE [✓ FALSE]
2.5 Give two orthogonal contrasts with integer weights to test the two hypotheses that: H_C control does not differ from the average of the two PKU groups and H_{hl} that high and low Phe groups do not differ. Fill in 6 integer values.	Group control lowPhe highPhe H_C -2 1 1 H_{hl} 0 -1 1

3. Use model D and the contrasts in 2.5 to fill in the following anova table.

Source	Sum of squares	Degrees of freedom	Mean Square	F-statistic
Between groups	11956.5	2	5978.3	68.333
Contrast H_C	9976.3	1	9976.3	114.031
Contrast H_{hl}	1980.2	1	1980.2	22.634
Within groups	1837.2	21	87.5	

Notice that 9976.3+1980.2 = 11956.5, so the sum of squares between groups has been partitioned into two parts that add to the total. This required orthogonal contrasts in a balanced design.

Most of the action is control vs Pku, much less is high vs low.

Statistics 500, Fall 2010, Problem Set 3
Doing the Problem Set in R

```
> attach(gborder)
> border<-1*(gborder$dist_gg_border<=75)
> mod<-lm(g3988~dist+border+dist_gg_border+rubble+flats+refugees+g1939)
```

1.1

```
> summary(mod)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.45402   29.64364   0.184 0.854351
dist           -17.31742   38.78915  -0.446 0.656119
border         -16.39592   24.86025  -0.660 0.510890
dist_gg_border   0.06125    0.09352   0.655 0.513842
rubble          -2.14487    0.73271  -2.927 0.004128 **
flats            0.38635    0.27158   1.423 0.157578
refugees         2.98494    0.76520   3.901 0.000163 ***
g1939           -0.23866    0.21851  -1.092 0.277036
---
Residual standard error: 48.34 on 114 degrees of freedom
Multiple R-squared: 0.2435,     Adjusted R-squared: 0.197
F-statistic: 5.242 on 7 and 114 DF,  p-value: 3.298e-05
```

1.2

```
> modLittle<-lm(g3988 ~ rubble+refugees)
> anova(modLittle,mod)
Analysis of Variance Table
Model 1: g3988 ~ rubble + refugees
Model 2: g3988~dist+border+dist_gg_border+rubble+flats+refugees+g1939
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    119 294718
2    114 266439  5     28279 2.4199 0.03993 *
```

1.3

```
> 2^7
[1] 128
```

1.4

```
> library(leaps)
> help(leaps)
> X<-cbind(dist,border,dist_gg_border,rubble,flats,refugees,g1939)
> result<-leaps(x=X,y=g3988,names=colnames(X))
> result
$which
   dist border dist_gg_border rubble flats refugees g1939
1 FALSE  FALSE          FALSE   TRUE FALSE    FALSE FALSE
1 FALSE  FALSE          FALSE  FALSE FALSE     TRUE FALSE
…
> which.min(result$Cp)
[1] 28
> result$which[28,]
dist  border dist_gg_border     rubble     flats   refugees   g1939
FALSE TRUE   FALSE               TRUE       TRUE    TRUE       FALSE
```

1.5

```
> result$Cp[28]
[1] 4.008623
> result$size[28]
[1] 5
```

It is often helpful to plot C_P:

```
> plot(result$size,result$Cp)
> abline(0,1)
```

1.6

```
> cbind(result$which,result$Cp,result$size)[result$Cp<=result$size,]
  dist border dist_gg_border rubble flats refugees g1939
4    0      1              0      1     1        1     0 4.008623 5
4    1      0              0      1     1        1     0 4.733086 5
5    0      1              0      1     1        1     1 4.761274 6
5    0      1              1      1     1        1     0 5.328695 6
5    1      0              0      1     1        1     1 5.351142 6
5    1      0              1      1     1        1     0 5.622819 6
5    1      1              0      1     1        1     0 5.748573 6
6    0      1              1      1     1        1     1 6.199318 7
6    1      1              0      1     1        1     1 6.428913 7
6    1      0              1      1     1        1     1 6.434971 7
7    1      1              1      1     1        1     1 8.000000 8
```

1.7

```
> library(DAAG)
> modC<-lm(g3988 ~ border+rubble+flats+refugees+g1939)
> press(modC)
[1] 298914.3
> press(mod)
[1] 314191.8
```

2.1

```
> mod<-aov(DI~group)
> shapiro.test(mod$residual)

        Shapiro-Wilk normality test

data:  mod$residual
W = 0.9588, p-value = 0.4151
```

2.2

```
> summary(mod)
            Df  Sum Sq Mean Sq F value    Pr(>F)
group        2 11956.6  5978.3  68.333 6.413e-10 ***
Residuals   21  1837.2    87.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2.3

```
> TukeyHSD(mod)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = DI ~ group)
$group
                  diff      lwr      upr     p adj
lowPhe-control  32.125 20.33691 43.91309 0.0000025
highPhe-control 54.375 42.58691 66.16309 0.0000000
highPhe-lowPhe  22.250 10.46191 34.03809 0.0003016
```

2.5 and 3

These are the default contrasts.

```
> contrasts(group)
        lowPhe highPhe
control      0       0
lowPhe       1       0
highPhe      0       1
```

You need to change the default contrasts.

```
> contrasts(group)[,1]<-c(-2,1,1)
> contrasts(group)[,2]<-c(0,-1,1)
> colnames(contrasts(group))<-c("Pku vs Control","High vs Low")
> contrasts(group)
        Pku vs Control High vs Low
control             -2           0
lowPhe               1          -1
highPhe              1           1
```

Now redo the model with the new contrasts and look at the model.matrix.

```
> mod<-lm(DI~group)
> model.matrix(mod)
```

Use the model matrix to create new variables.

```
> PkuVsC<-model.matrix(mod)[,2]
> HighVsLow<-model.matrix(mod)[,3]
```

Finally, do the anova. Remember you need orthogonal contrasts and a balanced design to do this.

```
> anova(lm(DI~PkuVsC+HighVsLow))
Analysis of Variance Table

Response: DI
          Df Sum Sq Mean Sq F value    Pr(>F)    
PkuVsC     1 9976.3  9976.3 114.031 6.063e-10 ***
HighVsLow  1 1980.2  1980.2  22.634 0.0001064 ***
Residuals 21 1837.2    87.5                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

PROBLEM SET #1 STATISTICS 500 FALL 2011: DATA PAGE 1

Due in class Tuesday 25 October 2011 at noon.

This is an exam. Do not discuss it with anyone.

The data are from the Joint Canada/United States Survey of Health, which was a version of the National Health Interview Survey given to both Canadians and people in the US. The data came from http://www.cdc.gov/nchs/nhis/jcush.htm , but there is no need for you to go to that web page unless you want to.

If you are using R, the data are available on my webpage, http://www-stat.wharton.upenn.edu/~rosenbap/index.html in the object `uscanada`. You will need to download the workspace again. You *may* need to clear your web browser's cache, so that it gets the new file, rather than using the file already on your computer. In Firefox, you might have to clear recent history. If you cannot find the `uscanada` object when you download the new R workspace, you probably have not downloaded the new file and are still working with the old one.

If you are not using R, the data are available in a .csv file `uscanada.csv` at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` A csv file should open in excel, so you can copy and paste it, and many programs can read csv files. The list of files here is case sensitive, upper case separate from lower case, so `uscanada.csv` is with the lower case files further down. If you cannot find the file, make sure you are looking at the lower case files

The variables are listed below. The newnames refer to the original CDC names (age is short for DJH1GAGE). In particular, PAJ1DEXP or dailyenergy is a measure of the average daily energy expended during leisure time activities by the respondent in the past three months, and it summarizes many questions about specific activities. The body mass index is a measure of obesity http://www.nhlbisupport.com/bmi/ .

```
> uscanadaLabels
        newname         name                                      label
2       country   SPJ1_TYP                                 Sample type
11          age   DHJ1GAGE                                   Age - (G)
12       female   DHJ1_SEX                                         Sex
68   cigsperday     SMJ1_6         # cigarettes per day (daily smoker)
88          bmi   HWJ1DBMI                       Body Mass Index - (D)
89       weight   HWJ1DWTK                     Weight - kilograms (D)
91       height   HWJ1DHTM                       Height - metres - (D)
93       hasdoc   HCJ1_1AA                 Has regular medical doctor
342 dailyenergy   PAJ1DEXP                   Energy expenditure - (D)
343   minutes15   PAJ1DDFR       Partic. in daily phys. act. >15 min.
347     PhysAct   PAJ1DIND               Physical activity index - (D)
353        educ   SDJ1GHED     Highest level/post-sec. educ. att. (G)
```

PROBLEM SET #1 STATISTICS 500 FALL 2011: DATA PAGE 2

This is an exam. Do not discuss it with anyone.
Due in class Tuesday 25 October 2011 at noon.

```
attach(uscanada)
```

Model #1

bmi $= \beta_0 + \beta_1$age + β_2 cigsperday + β_3 dailyenergy + ε where ε are iid $N(0,\sigma^2)$

Model #2

```
rbmi <- 1/bmi   (Reciprocal of bmi.)
```

rbmi $= \gamma_0 + \gamma_1$age + γ_2 cigsperday + γ_3 dailyenergy + η where η are iid $N(0,\omega^2)$

Model 1 has slopes β (beta), while model 2 has slopes γ (gamma), so that different things have different names. The choice of Greek letters is arbitrary.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2011: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due 25 October 2011 at noon.

1. Look at the data.	Fill in or CIRCLE the correct answer

1.a What is the smallest bmi in the data? What is the age of the person with the lowest bmi? How many cigarettes per day does this person smoke?	bmi = _____________ age = _________
	cigsperday = ____________________

1.b What is the largest bmi in the data? What is the age of the person with the largest bmi? How many cigarettes per day does this person smoke?	bmi = _____________ age = _________
	cigsperday = ____________________

1.c What is the median number of cigarettes smoked per day?	Median = _______________________

2 **Fit model 1 on the data page** and use it to answer the following questions. For the questions in part 2, assume model 1 is true.	**Fill in or CIRCLE the correct answer**
2.a Give the point estimate and 95% two-sided confidence interval for β_1 , the coefficient of age.	Estimate= _____ CI = [,]
2.b Test the hypothesis that the coefficient of cigsperday is zero, H_0: $\beta_2 = 0$, against a two-sided alternative hypothesis. What is the name of the test statistic? What is the value of the test statistic? What is the p-value? Is the null hypothesis plausible?	Name: ___________ Value: __________ p-value: __________ Circle one: Plausible Not plausible
2.c Greater daily energy expenditure is associated with a larger bmi.	TRUE FALSE
2.d Smoking more cigarettes is associated with a larger bmi.	TRUE FALSE
2.e The model has fitted about 53.75% of the variation in bmi as measured by the F statistic.	TRUE FALSE
2.f A person who smokes 1 more cigarette is estimated to have a bmi that is 2% lower.	TRUE FALSE
2.g What is the numerical value of the largest residual in the data?	Value: _______________________
2.h. What is the estimate of σ? Give the numerical value.	Value: _______________________

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2011: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due 25 October 2011 at noon.

3. **Fit model 1 on the data page** and use it to answer the following questions. The questions in part 3 ask whether model 1 is a reasonable model for these data.	**Fill in or CIRCLE the correct answer**
3a. Based on a boxplot of residuals, very large negative residuals occur more often than very large positive residuals.	TRUE FALSE
3b. Because the normal plot exhibits an inverted S-shaped curve, the residuals appear to be skewed to the left rather than Normal.	TRUE FALSE
3c. Because the plot of residuals against fitted values exhibits an inverted U shape, it is clear that the relationship is nonlinear.	TRUE FALSE
3d. In the plot of residuals against fitted values, there is one extremely large fitted bmi far away from other points for a very old person who smokes 10 cigerattes per day and who has dailyenergy = 3.1.	TRUE FALSE

4. **Fit model 2 on the data page** and use it to answer the following questions.	**Fill in or CIRCLE the correct answer**
4a. The Normal quantile plot of the residuals from model 2 looks straighter than the Normal quantile plot of residuals from model 1.	TRUE FALSE
4b. In model 2, test the hypothesis that the coefficients of age and cigarettes per day are simultaneously both zero, H_0: $\gamma_1=\gamma_2=0$. What is the name of the test statistic? What is the numerical value of the test statistic. What are its degrees of freedom (DF)? What is the p-value? Is the null hypothesis plausible?	Name: ___________ Value: __________ p-value: __________ DF: ____________ Circle one: Plausible Not plausible
4c. Consider age = 25, dailyenergy =1 for a nonsmoker, cigsperday = 0. Give the point estimate and 95% two-sided confidence interval for $\gamma_0 + \gamma_1 25 + \gamma_2 0 + \gamma_3 1$ assuming model 2 is true. Then take reciprocals to express this in terms of bmi.	Estimate and interval for $\gamma_0 + \gamma_1 25 + \gamma_2 0 + \gamma_3 1$ Estimate: _________ CI: _____________ Take reciprocals of the estimate and CI 1/ Estimate =______ 1/CI=[,]

Answers

PROBLEM SET #1 STATISTICS 500 FALL 2011: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone.

2. Look at the data.	Fill in or CIRCLE the correct answer

1.a What is the smallest bmi in the data? What is the age of the person with the lowest bmi? How many cigarettes per day does this person smoke?	bmi = 13.6 age = 23
	cigsperday = 0

1.b What is the largest bmi in the data? What is the age of the person with the largest bmi? How many cigarettes per day does this person smoke?	bmi = 82.5 age = 40
	cigsperday = 0

1.c What is the median number of cigarettes smoked per day?	Median = 0

2 **Fit model 1 on the data page** and use it to answer the following questions. For the questions in part 2, assume model 1 is true.	**Fill in or CIRCLE the correct answer**
2.a Give the point estimate and 95% two-sided confidence interval for β_1 , the coefficient of age.	Estimate= 0.0205 CI = [0.0138, 0.0272]
2.b Test the hypothesis that the coefficient of cigsperday is zero, H_0: $\beta_2 = 0$, against a two-sided alternative hypothesis. What is the name of the test statistic? What is the value of the test statistic? What is the p-value? Is the null hypothesis plausible?	Name: t-test Value: -2.98 p-value: 0.00287 Circle one: Plausible [✓ Not plausible]
2.c Greater daily energy expenditure is associated with a larger bmi.	TRUE [✓ FALSE]
2.d Smoking more cigarettes is associated with a larger bmi.	TRUE [✓ FALSE]
2.e The model has fitted about 53.75% of the variation in bmi as measured by the F statistic.	TRUE [✓ FALSE] Not what F is. Use R^2.
2.f A person who smokes 1 more cigarette is estimated to have a bmi that is 2% lower.	TRUE [✓ FALSE] Units of BMI, not percent.
2.g What is the numerical value of the largest residual in the data?	Value: 56.153
2.h. What is the estimate of σ? Give the numerical value.	Value: 5.219

Answers

PROBLEM SET #1 STATISTICS 500 FALL 2011: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

3. **Fit model 1 on the data page** and use it to answer the following questions. The questions in part 3 ask whether model 1 is a reasonable model for these data.	**Fill in or CIRCLE the correct answer**
3a. Based on a boxplot of residuals, very large negative residuals occur more often than very large positive residuals.	TRUE [✓ FALSE] Large positive residuals are common (large bmi).
3b. Because the normal plot exhibits an inverted S-shaped curve, the residuals appear to be skewed to the left rather than Normal.	TRUE [✓ FALSE] The curve is not inverted S-shaped, and the residuals are skewed right, not left. Not Normal.
3c. Because the plot of residuals against fitted values exhibits an inverted U shape, it is clear that the relationship is nonlinear.	TRUE [✓ FALSE] Not U-shaped.
3d. In the plot of residuals against fitted values, there is one extremely large fitted bmi far away from other points for a very old person who smokes 10 cigerattes per day and who has dailyenergy = 3.1.	TRUE [✓ FALSE] Actually, there is a very low fitted bmi …

4. **Fit model 2 on the data page** and use it to answer the following questions.	**Fill in or CIRCLE the correct answer**
4a. The Normal quantile plot of the residuals from model 2 looks straighter than the Normal quantile plot of residuals from model 1.	[✓ TRUE] FALSE Much, much straighter, but still not quite Normal.
4b. In model 2, test the hypothesis that the coefficients of age and cigarettes per day are simultaneously both zero, H_0: $\gamma_1=\gamma_2=0$. What is the name of the test statistic? What is the numerical value of the test statistic. What are its degrees of freedom (DF)? What is the p-value? Is the null hypothesis plausible? (10 points)	Name: F-test Value: 40.111 p-value: 2.2×10^{-16} DF: 2 & 8028 Circle one: Plausible [✓ Not plausible]
4c. Consider age = 25, dailyenergy =1 for a nonsmoker, cigsperday = 0. Give the point estimate and 95% two-sided confidence interval for $\gamma_0 + \gamma_1 25 + \gamma_2 0 + \gamma_3 1$ assuming model 2 is true. Then take reciprocals to express this in terms of bmi. (10 points)	Estimate and interval for $\gamma_0 + \gamma_1 25 + \gamma_2 0 + \gamma_3 1$ Estimate: 0.0400 CI: [0.0398, 0.0403] Take reciprocals of the estimate and CI 1/ Estimate = 24.967 1/CI=[24.79, 25.15]

Problem 1, Fall 2011, Statistics 500
Doing the Problem Set in R

```
> attach(uscanada)

Question 1.
> which.min(bmi)
[1] 1612
> uscanada[1612,]
     age cigsperday dailyenergy  bmi height weight
1701  23          0         0.8 13.6  1.803   44.1
> which.max(bmi)
[1] 6316
> uscanada[6316,]
     age cigsperday dailyenergy  bmi height weight
6812  40          0           1 82.5  1.651    225

> summary(cigsperday)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   3.127   0.000  60.000

Question 2.
> mod<-lm(bmi~age+cigsperday+dailyenergy)
> mod
Coefficients:
(Intercept)          age   cigsperday  dailyenergy
   25.77529      0.02049     -0.02228     -0.24727

> confint(mod)
                  2.5 %       97.5 %
(Intercept) 25.39913509 26.151436946
age          0.01382004  0.027152447
cigsperday  -0.03692123 -0.007632065
dailyenergy -0.29741617 -0.197127872

> summary(mod)
Residuals:
    Min      1Q  Median      3Q     Max
-13.096  -3.511  -0.778   2.582  56.153

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.775286   0.191888 134.324  < 2e-16 ***
age          0.020486   0.003401   6.024 1.77e-09 ***
cigsperday  -0.022277   0.007471  -2.982  0.00287 **
dailyenergy -0.247272   0.025580  -9.666  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.219 on 8028 degrees of freedom
Multiple R-squared: 0.01969,    Adjusted R-squared: 0.01932
F-statistic: 53.75 on 3 and 8028 DF,  p-value: < 2.2e-16
```

Problem 1, Fall 2011, Statistics 500, continued

Question 3.

```
> res<-mod$residual
> boxplot(res)
> qqnorm(res)
> qqline(res)
> shapiro.test(sample(res,5000))
```

(For some reason, R won’t do the test with more than 5000 observations.)

```
        Shapiro-Wilk normality test
data:  sample(res, 5000)
W = 0.9309, p-value < 2.2e-16

> fit<-mod$fitted
> plot(fit,res)
> lines(lowess(fit,res),col="red")
```

Question 4.

```
> rbmi<-1/bmi
> cbind(bmi,rbmi)[1:4,]
      bmi       rbmi
[1,] 21.1 0.04739336
[2,] 22.4 0.04464286
[3,] 20.4 0.04901961
[4,] 28.4 0.03521127
> modfull<-lm(rbmi~age+cigsperday+dailyenergy)
> qqnorm(modfull$residual)
> modreduced<-lm(rbmi~dailyenergy)
> anova(modreduced,modfull)
Analysis of Variance Table
Model 1: rbmi ~ dailyenergy
Model 2: rbmi ~ age + cigsperday + dailyenergy
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   8030 0.41101
2   8028 0.40694  2 0.0040665 40.111 < 2.2e-16 ***
---
> predict(modfull,data.frame(age=25,cigsperday=0,dailyenergy=1),
interval="confidence")
         fit        lwr        upr
1 0.04005285 0.03976237 0.04034334
> 1/predict(modfull,data.frame(age=25,cigsperday=0,dailyenergy=1),
interval="confidence")
       fit      lwr      upr
1 24.96701 25.14941 24.78724
```

PROBLEM SET #2 STATISTICS 500 FALL 2011: DATA PAGE 1

This is an exam. Do not discuss it with anyone.

As in problem 1, the data are from the Joint Canada/United States Survey of Health, which was a version of the National Health Interview Survey given to both Canadians and people in the US. The data came from http://www.cdc.gov/nchs/nhis/jcush.htm , but there is no need for you to go to that web page unless you want to.

If you are using R, the data are available on my webpage, http://www-stat.wharton.upenn.edu/~rosenbap/index.html in the object `uscanada`. You will need to download the workspace again. You *may* need to clear your web browser's cache, so that it gets the new file, rather than using the file already on your computer. In Firefox, you might have to clear recent history. If you cannot find the `uscanada` object when you download the new R workspace, you probably have not downloaded the new file and are still working with the old one.

If you are not using R, the data are available in a .csv file `uscanada.csv` at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` A csv file should open in excel, so you can copy and paste it, and many programs can read csv files. The list of files here is case sensitive, upper case separate from lower case, so `uscanada.csv` is with the lower case files further down. If you cannot find the file, make sure you are looking at the lower case files

The variables are listed below. The newnames refer to the original CDC names (age is short for DJH1GAGE). In particular, PAJ1DEXP or dailyenergy is a measure of the average daily energy expended during leisure time activities by the respondent in the past three months, and it summarizes many questions about specific activities. The body mass index is a measure of obesity http://www.nhlbisupport.com/bmi/ .

```
> uscanadaLabels
        newname         name                                      label
2       country   SPJ1_TYP                                 Sample type
11          age   DHJ1GAGE                                   Age - (G)
12       female   DHJ1_SEX                                         Sex
68   cigsperday     SMJ1_6         # cigarettes per day (daily smoker)
88          bmi   HWJ1DBMI                       Body Mass Index - (D)
89       weight   HWJ1DWTK                     Weight - kilograms (D)
91       height   HWJ1DHTM                       Height - metres - (D)
93       hasdoc   HCJ1_1AA                 Has regular medical doctor
342 dailyenergy   PAJ1DEXP                   Energy expenditure - (D)
343   minutes15   PAJ1DDFR       Partic. in daily phys. act. >15 min.
347     PhysAct   PAJ1DIND               Physical activity index - (D)
353        educ   SDJ1GHED     Highest level/post-sec. educ. att. (G)
```

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

PROBLEM SET #2 STATISTICS 500 FALL 2011: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due in class Tuesday 22 November 2011 at noon.

Model #1

bmi $= \beta_0 + \beta_1$age + β_2 cigsperday + β_3 dailyenergy + ε where ε are iid $N(0,\sigma^2)$

Model #2

```
rbmi <- 1/bmi   (Reciprocal of bmi.)
```

rbmi $= \gamma_0 + \gamma_1$age + γ_2 cigsperday + γ_3 dailyenergy + η where η are iid $N(0,\omega^2)$

Remark #1: In the first problem set, you fit models 1 and 2 and discovered that rbmi = 1/bmi in model #2 had residuals whose Normal plot looked much straighter (hence less wildly non-Normal) than the residuals from model #1 using bmi. However, the reciprocal of bmi seems hard to interpret. The NHLBI recommends a bmi between 18.5 and 24.9, and the midpoint of that range is about 22. So let us call 22 the "recommended bmi". Consider the variable fr = (22-bmi)/bmi.

Model #3

```
> fr <-(22-bmi)/bmi
```

fr $= \lambda_0 + \lambda_1$age + λ_2 cigsperday + λ_3 dailyenergy + ι where ι are iid $N(0,\delta^2)$

Plot fr versus rbmi. Calculate the residuals from models 2 and 3 and plot them against each other. Do not turn in the plots – just look at them as an aid to answering the questions. Question 1.6 asks for the **best interpretation** of fr from the list below

A. If your fr is 0.20, your bmi is 20% higher than the recommended bmi of 22.
B. If your fr is -0.20, your bmi is 20% higher than the recommended bmi of 22.
C. If your fr is 0.20, your weight must fall by 20% to reach the recommended bmi of 22.
D. If your fr is -0.20, your weight must fall by 20% to reach the recommended bmi of 22.

Model #4 Fit model 4 below, and look at the Normal plot of the residuals, plot residuals vs fitted with a lowess curve in col= "red". Do not turn in plots – look at them.

$$\text{fr} = \theta_0 + \theta_1\text{age} + \theta_2 \text{ female} + \theta_3 \text{ hasdoc} + \theta_4 \text{ dailyenergy} + \theta_5 \text{ country} + \xi$$
$$\text{where } \xi \text{ are iid } N(0,\Delta^2)$$

Question 2 asks what model 4 predicts for two people who are "otherwise the same," so for example, a male and a female who are otherwise the same have the same age, both have or both do not have a regular doctor, same daily energy expenditure and live the in the same country. "Otherwise the same" means the same in terms of x's in the model except those specifically mentioned.

Model #5 Fit model 5 below and plot its fitted values (y) against age (x).
Define centered age as age2 = (age-mean(age))^2. Plot age2 versus age.

fr $= \kappa_0 + \kappa_1$age + κ_2 female + κ_3 hasdoc + κ_4 dailyenergy+ κ_5 country + κ_6 age2 + ζ where ζ are iid $N(0,\phi^2)$

Name: _____________________________ ID# _________________________

PROBLEM SET #2 STATISTICS 500 FALL 2011: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due 22 November 2011 at noon.

1. **Read Remark 1 on the data page.**	Fill in the correct answer
1.1 What is the correlation between 1/bmi and fr = (22-bmi)/bmi. Give the numerical value of the ususal (i.e. Pearson) correlation.	Correlation = ____________________
1.2 What is the correlation between the residuals of models 2 and 3 on the data page? Give the numerical value.	Correlation = ____________________
1.3 For the four bmi's listed, give the numerical values of fr = (22-bmi)/bmi. Two digits beyond the decimal are sufficient, so .333333 is ok as .33.	bmi 20 30 35 44 fr ___ ____ ____ ____
1.4 To achieve the recommended bmi of 22, what percentage (0-100%) would a person with a bmi of 44 have to lose? Give one number between 0 and 100%.	______________ %

1.5 If X is a random variable with finite nonzero variance and a and b are two constants with b>0, what is the correlation between X and a+bX? Give a number. If you don't know the number, run an experiment. Write fr as a+bX where X=1/bmi by giving the value of a and b that make this true.	Correlation = ____________________
	a = __________
	b = __________

1.6 Using the list of best interpretations of fr on the data page, select the one best interpretation. Give one letter A-D.	One letter: _________________

2. Fit model #4 on the data page and use it for the following questions. Read about "otherwise the same" on the data page	Circle the correct answer
2.1 For a male and female who are otherwise the same, model 4 predicts the female needs to lose a larger fraction of her weight to achieve a bmi of 22.	TRUE FALSE
2.2 For a person aged 70 and another aged 25 who are otherwise the same, the model predicts the 25-year old needs to lose a larger fraction of his/her weight to achieve the recommended bmi of 22.	TRUE FALSE
2.3 The constant term θ_0 in model #4 is the fractional weight loss recommended for the average person in the data set.	TRUE FALSE

Name: ______________________________ ID# __________________________

PROBLEM SET #2 STATISTICS 500 FALL 2011: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due 22 November 2011 at noon.

3. Use models 4 and 5 on the data page to answer the following questions.	Fill in or circle the correct answer
3.1 Use Tukey's 1 degree-of-freedom for non-additivity to test the null hypothesis that model #4 is correct against the alternative hypothesis that some curvature is present. Give the value of the t-test statistic, the P-value, and state whether the null hypothesis is plausible.	t-value: __________ p-value: _________ PLAUSIBLE NOT PLAUSIBLE
3.2 What is the numerical value of the correlation between age and age^2? What is the numerical value of the correlation between age and centered age^2, namely age2 = (age-mean(age))^2.	With age^2: _______________ With age2: _______________
3.3 Fit model 5 and test the null hypothesis that the relationship between fr and age is linear as in model 4 versus the alternative that it is not linear but rather needs a quadratic term as in model 5. Give the name and value of the test statistic, the P-value, and state whether the null hypothesis is plausible.	Name: _________ Value: __________ P-value: _________ PLAUSIBLE NOT PLAUSIBLE
3.4 If the point estimate of κ_6, the coefficient of age2, were actually the true value of κ_6, then the model would predict that a 20 year old and an 80 old would both need to lose more weight than a 55 year old who is otherwise the same to reach the recommended bmi of 22.	TRUE FALSE
3.5 In model 5, which individual has the largest absolute studentized residual (rstudent())? Give the row number. What is the numerical value of the studentized residual? Is it true that this individual has a bmi of 82.5?	Which row?________ Value: ______ TRUE FALSE
3.6 Test at the 0.05 level the null hypothesis that model 5 has no outliers. What is the value of statistic? What are the degrees of freedom (DF)? Does the test reject the null hypothesis of no outlier, thereby finding at least one outlier?	Value: _________ DF: ______________ Circle one Rejects Does not reject Finds outlier Not an outlier

PROBLEM SET #2 STATISTICS 500 FALL 2011: ANSWER PAGE 1, Answers

This is an exam. Do not discuss it with anyone. Due 22 November 2011 at noon.

2. **Read Remark 1 on the data page.**	Fill in the correct answer (7 points each)
1.1 What is the correlation between 1/bmi and fr = (22-bmi)/bmi. Give the numerical value of the ususal (i.e. Pearson) correlation.	Correlation = 1
1.2 What is the correlation between the residuals of models 2 and 3 on the data page? Give the numerical value.	Correlation = 1
1.3 For the four bmi's listed, give the numerical values of fr = (22-bmi)/bmi. Two digits beyond the decimal are sufficient, so .333333 is ok as .33.	bmi 20 30 35 44 fr 0.1 -0.27 -0.37 -0.50
1.4 To achieve the recommended bmi of 22, what percentage (0-100%) would a person with a bmi of 44 have to lose? Give one number between 0 and 100%.	50 %

1.5 If X is a random variable with finite nonzero variance and a and b are two constants with b>0, what is the correlation between X and a+bX? Give a number. If you don't know the number, run an experiment. Write fr as a+bX where X=1/bmi by giving the value of a and b that make this true.	Correlation = 1
	a = -1
	b = 22

1.6 Using the list of best interpretations of fr on the data page, select the one best interpretation. Give one letter A-D.	One letter: D (22-27.5)/27.5 = -0.2 27.5*(1-.2) = 22

2. Fit model #4 on the data page and use it for the following questions. Read about "otherwise the same" on the data page	Circle the correct answer 5 points each
2.1 For a male and female who are otherwise the same, model 4 predicts the female needs to lose a larger fraction of her weight to achieve a bmi of 22.	TRUE [✓ FALSE]
2.2 For a person aged 70 and another aged 25 who are otherwise the same, the model predicts the 25-year old needs to lose a larger fraction of his/her weight to achieve the recommended bmi of 22.	TRUE [✓ FALSE]
2.3 The constant term θ_0 in model #4 is the fractional weight loss recommended for the average person in the data set.	TRUE [✓ FALSE] A person has predicted value θ_0 if all of their x's were 0, which often makes no sense.

PROBLEM SET #2 STATISTICS 500 FALL 2011: ANSWER PAGE 2, Answers

This is an exam. Do not discuss it with anyone. Due 22 November 2011 at noon.

3. Use models 4 and 5 on the data page to answer the following questions.	Fill in or circle the correct answer 7 points each
3.1 Use Tukey's 1 degree-of-freedom for non-additivity to test the null hypothesis that model #4 is correct against the alternative hypothesis that some curvature is present. Give the value of the t-test statistic, the P-value, and state whether the null hypothesis is plausible.	t-value: -5.27 p-value: 1.4 x 10^{-7} PLAUSIBLE [✓ NOT PLAUSIBLE]
3.2 What is the numerical value of the correlation between age and age^2? What is the numerical value of the correlation between age and centered age^2, namely age2 = (age-mean(age))^2.	With age^2: 0.984 With age2: 0.255
3.3 Fit model 5 and test the null hypothesis that the relationship between fr and age is linear as in model 4 versus the alternative that it is not linear but rather needs a quadratic term as in model 5. Give the name and value of the test statistic, the P-value, and state whether the null hypothesis is plausible.	Name: t-test Value: 14.33 F-test ok, F = t^2 with 1 df in numerator. P-value: 10^{-16} PLAUSIBLE [✓ NOT PLAUSIBLE]
3.4 If the point estimate of κ_6, the coefficient of age2, were actually the true value of κ_6, then the model would predict that a 20 year old and an 80 old would both need to lose more weight than a 55 year old who is otherwise the same to reach the recommended bmi of 22.	TRUE [✓ FALSE]
3.5 In model 5, which individual has the largest absolute studentized residual (rstudent())? Give the row number. What is the numerical value of the studentized residual? Is it true that this individual has a bmi of 82.5?	Which row? 1612 Value: 4.38 TRUE [✓ FALSE]
3.6 Test at the 0.05 level the null hypothesis that model 5 has no outliers. What is the value of statistic? What are the degrees of freedom (DF)? Does the test reject the null hypothesis of no outlier, thereby finding at least one outlier?	Value: 4.38 DF: 8024 1 df lost for outlier dummy variable. Circle one Rejects Finds outlier [✓ Does not reject Not an outlier]

Problem Set 2, Fall 2011, Statistics 500

Doing the Problem Set in R

Question 1.

```
> fr<-(22-bmi)/bmi
> cor(1/bmi,fr)
[1] 1
> rbmi<-1/bmi
> mod2<-lm(rbmi~cigsperday+age+dailyenergy)
> mod3<-lm(fr~cigsperday+age+dailyenergy)
> cor(mod2$resid,mod3$resid)
[1] 1
> bmilist
[1] 20 30 35 44
> bmilist[1]<-22
> (22-bmilist)/bmilist
[1]  0.0000000 -0.2666667 -0.3714286 -0.5000000
> round((22-bmilist)/bmilist,2)
[1]  0.00 -0.27 -0.37 -0.50
```

Question 2.

```
> mod4<-lm(fr~age+female+hasdoc+dailyenergy+country)
> qqnorm(mod4$resid)
> plot(mod4$fit,mod4$resid)
> lines(lowess(mod4$fit,mod4$resid),col="red")
> summary(mod4)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.1292268  0.0065706 -19.667  < 2e-16 ***
age                  -0.0007601  0.0001027  -7.401 1.49e-13 ***
female                0.0606567  0.0034861  17.400  < 2e-16 ***
hasdocNo regular doc  0.0225307  0.0047035   4.790 1.70e-06 ***
dailyenergy           0.0070861  0.0007577   9.353  < 2e-16 ***
countryUS            -0.0258247  0.0034975  -7.384 1.69e-13 ***
---
```

Question 3.

```
> summary(lm(fr ~ age + female + hasdoc + dailyenergy +
country+tukey1df(mod4)))
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.1340317  0.0066228 -20.238  < 2e-16 ***
age                  -0.0007641  0.0001025  -7.452 1.02e-13 ***
female                0.0617226  0.0034862  17.705  < 2e-16 ***
hasdocNo regular doc  0.0216226  0.0046989   4.602 4.26e-06 ***
dailyenergy           0.0056699  0.0008027   7.063 1.76e-12 ***
countryUS            -0.0254848  0.0034923  -7.297 3.21e-13 ***
tukey1df(mod4)       -1.2319893  0.2338398  -5.269 1.41e-07 ***
---

> age2<-(age-mean(age))^2
```

Problem Set 2, Fall 2011, Statistics 500, continued.

```
> plot(age,age2)
> cor(age,age2)
[1] 0.2550511
> cor(age,age^2)
[1] 0.9843236
```

```
> mod5<-lm(fr~age+female+hasdoc+dailyenergy+country+age2)
> summary(mod5)
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.304e-01  6.489e-03 -20.091  < 2e-16 ***
age                  -1.160e-03  1.052e-04 -11.028  < 2e-16 ***
female                5.760e-02  3.449e-03  16.700  < 2e-16 ***
hasdocNo regular doc  1.783e-02  4.656e-03   3.829 0.000129 ***
dailyenergy           6.649e-03  7.488e-04   8.880  < 2e-16 ***
countryUS            -2.501e-02  3.454e-03  -7.241 4.87e-13 ***
age2                  7.704e-05  5.374e-06  14.335  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.1515 on 8025 degrees of freedom
Multiple R-squared: 0.08394,    Adjusted R-squared: 0.08326
F-statistic: 122.6 on 6 and 8025 DF,  p-value: < 2.2e-16
> plot(age,mod5$fit)
> which.max(abs(rstudent(mod5)))
1612
> length(age)
[1] 8032
> indiv1612<-rep(0,8032)
> indiv1612[1612]<-1
> 44.1*2.2
[1] 97.02
> 1.803*39
[1] 70.317
> summary(lm(fr~age+female+hasdoc+dailyenergy+country+age2+indiv1612))
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.310e-01  6.483e-03 -20.202  < 2e-16 ***
age                  -1.150e-03  1.051e-04 -10.941  < 2e-16 ***
female                5.748e-02  3.445e-03  16.682  < 2e-16 ***
hasdocNo regular doc  1.801e-02  4.651e-03   3.872 0.000109 ***
dailyenergy           6.678e-03  7.480e-04   8.929  < 2e-16 ***
countryUS            -2.482e-02  3.451e-03  -7.191 6.99e-13 ***
age2                  7.666e-05  5.369e-06  14.278  < 2e-16 ***
indiv1612             6.633e-01  1.514e-01   4.382 1.19e-05 ***
> 0.05/8032
[1] 6.2251e-06
> qt(1-0.025/8032,8024)
[1] 4.521613
> max(abs(rstudent(mod5)))
[1] 4.381769
```

PROBLEM SET #3 STATISTICS 500 FALL 2011: DATA PAGE 1
Due Thursday 15 December 2011 at noon.
This is an exam. Do not discuss it with anyone.

Two data sets are used. The first is the same as in problems 1 and 2, the Joint Canada/United States Survey of Health. The second data set is from Allison, Truett and Cicchetti, Domenic V. (1976), Sleep in Mammals: Ecological and Constitutional Correlates, *Science*, 194: 732-734. The paper is available from JSTOR on the library web page, but there is no need to read it unless you are interested in doing so.

If you are using R, the data are available on my webpage, http://www-stat.wharton.upenn.edu/~rosenbap/index.html in the objects `uscanada` and `sleepST500`. You will need to download the workspace again. You *may* need to clear your web browser's cache, so that it gets the new file, rather than using the file already on your computer. If you cannot find the `sleepST500`, then you probably have not downloaded the new file and are still working with the old one.

If you are not using R, the data are available in a .csv files `uscanada.csv` and `sleepST500.csv` at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` A csv file should open in excel, so you can copy and paste it, and many programs can read csv files. Please note that there are several files with similar names, so make sure you have the correct files. The list of files here is case sensitive, upper case separate from lower case, so `uscanada.csv` and `sleepST500.csv` are with the lower case files further down. If you cannot find the file, make sure you are looking at the lower case files.

In `sleepST500`, look at two variables, y_{ij} = `totalsleep`, which is total hours per day of sleep, and `sleepdanger`, which forms three groups of 16 mammals each based on the danger they face when asleep. The bat and the jaguar are in the group in least danger when asleep, while the guinea pig is in most danger. Before doing anything else, you should plot the data, `boxplot(totalsleep~ sleepdanger)`. The model for the sleep data, **Model 1**, is

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where ε_{ij} are iid $N(0,\sigma^2)$, $i=1,2,3$, $j=1,2,\ldots,16$, $\alpha_1+\alpha_2+\alpha_3=0$

where i=1 for least, i=2 for middle, i=3 for most danger. The overall null hypothesis, H_0: $\alpha_1=\alpha_2=\alpha_3=0$, has three subhypotheses, H_{12}: $\alpha_1=\alpha_2$, H_{13}: $\alpha_1=\alpha_3$, and H_{23}: $\alpha_2=\alpha_3$, and you should refer to these hypothesis as H_{12}, etc. on the answer page.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

PROBLEM SET #3 STATISTICS 500 FALL 2011: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due Thursday 15 December 2011 at noon.

The variables for the US-Canada data are listed below. The newnames refer to the original CDC names (age is short for DJH1GAGE). In particular, PAJ1DEXP or dailyenergy is a measure of the average daily energy expended during leisure time activities by the respondent in the past three months, and it summarizes many questions about specific activities. The body mass index is a measure of obesity http://www.nhlbisupport.com/bmi/ .

```
> uscanadaLabels
        newname        name                                    label
2       country   SPJ1_TYP                               Sample type
11          age   DHJ1GAGE                               Age - (G)
12       female   DHJ1_SEX                                     Sex
68   cigsperday     SMJ1_6        # cigarettes per day (daily smoker)
88          bmi   HWJ1DBMI                   Body Mass Index - (D)
89       weight   HWJ1DWTK                 Weight - kilograms (D)
91       height   HWJ1DHTM                   Height - metres - (D)
93       hasdoc   HCJ1_1AA                Has regular medical doctor
342 dailyenergy   PAJ1DEXP                 Energy expenditure - (D)
343   minutes15   PAJ1DDFR       Partic. in daily phys. act. >15 min.
347     PhysAct   PAJ1DIND               Physical activity index - (D)
353        educ   SDJ1GHED     Highest level/post-sec. educ. att. (G)
```

Model #2

```
> fr <-(22-bmi)/bmi
```

$$fr = \gamma_0 + \gamma_1 age + \gamma_2\, female + \gamma_3\, cigsperday + \gamma_4\, dailyenergy + \eta$$

where η are iid $N(0,\omega^2)$

The problem will ask you to consider all submodels of model 2, including model 2 itself, the model with $fr = \gamma_0$ with no predictors, and all the models that use a subset of the variables age, female, cigsperday and dailyenergy.

There are three options about **turning in the exam**. (i) You can deliver it to my office 473 JMHH on Thursday 15 December at noon. Any time before noon on 15 December, you can (ii) place it in a sealed envelope addressed to me and leave it in my mail box in the statistics department, 4th floor JMHH, or (iii) you can leave it with Adam at the front desk in the statistics department. **Make and keep a photocopy of your answer page** – if something goes wrong, I can grade the photocopy. The statistics department is locked at night and on the weekend. Your **course grade** will be available from the registrar shortly after I grade the finals. I will put the **answer key** in an updated version of the bulkpack on my web page shortly after I grade the final. I no longer distribute course materials by US or campus mail.

Last name: _______________ **First name**:_____________ ID# __________________

PROBLEM SET #3 STATISTICS 500 FALL 2011: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due 15 December 2011 at noon.

1 Use the USCanada data and model 2 on the data page to answer the following questions. Assume model 2 is true.	Fill in or Circle the Correct Answer
1.1 How many submodels does model 2 have? Include model 2 itself and the empty model in your count.	How many: ____________________
1.2 Of the submodels of model 2, which one model is estimated to have the smallest expected total squared error of prediction as judged by C_P? List the names of the variables in this model, the value of C_P, and the size of the model.	List variables: C_P = ___________ size = ___________
1.3 Of the submodels of model 2, use C_P to identify all the models whose C_P value estimates that the model contains all variables with nonzero coefficients. **Of these models**, pick the **one model** with the smallest size. List the names of the variables in this model, the value of C_P, and the size of the model.	List variables: C_P = ___________ size = ___________
1.4 Give the values of PRESS for model 2 and for the submodel with exactly one predictor namely age.	PRESS for Model 2: ________________ PRESS with Age alone: _____________

2 Use the USCanada data and model 2 on the data page to answer the following questions.	Fill in or Circle the Correct Answer
2.1 In model 2, how many observational have large hatvalues h_i (or leverages) by the standard we discussed in class? Give one number.	How many? ___________________
2.2 The one person with the largest hatvalue is a 21-year old female nonsmoker with an unusually high value for dailyenergy. Give the largest hat value.	Circle one: TRUE FALSE Largest hatvalue = __________________
2.3 The person with the largest absolute dffits, |dffits|, shifts his/her own fitted value by more than half of its standard error. What is the signed value of dffits for the person with the largest |dffits|?	Circle one: TRUE FALSE Value of dffits = _______________

Last name: _______________ **First name**:_____________ ID# __________________

PROBLEM SET #3 STATISTICS 500 FALL 2011: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due 15 December 2011 at noon.

3. Use sleepST500 for question 3. Assume the model for the sleep data, model 1, on the data page is true in question 3 and use it in answering the following questions.	Fill in or **circle** the correct answer.
3.1 Test the null hypothesis that H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$. What is the name of the test statistic? What is the value of the test statistic? What are the degrees of freedom? What is the p-value? Is the null hypothesis plausible?	Name: _____________ Value: ________ DF: _____ , ______ P-value: _________ Circle one: PLAUSIBLE NOT PLAUSIBLE
3.2 Use Tukey's method of multiple comparisons to create three simultaneous 95% confidence intervals for the differences $\alpha_i - \alpha_k$. Give the numerical values of the endpoints of the confidence intervals expressed in hours per day.	Middle-Most: [,] Least-Most: [,] Least-Middle: [,]
Each of the three confidence intervals in 3.2 covers its parameter $\alpha_i - \alpha_k$ in 95% of experiments, but because of the Bonferroni inequality, all three confidence intervals cover their respective parameter in only 1-0.05x3 = 85% of experiments.	Circle one: TRUE FALSE
Suppose H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ is false. Under this supposition **circle all of the true statements**, if any. (That is, circle 0, 1, 2, 3 or 4 statements.) Refer to the data page for notation, eg H_{12}.	If H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ is false, then: 1) It is possible that H_{12} is true 2) It is possible that H_{12} and H_{13} are both true 3) It is possible that H_{12} and H_{13} and H_{23} are all false 4) At most one of H_{12} and H_{13} and H_{23} is true.
Under the model for the sleep data, what is the smallest hatvalue (i.e. leverage)? What is the largest hatvalue? Give two numbers. How many hatvalues are large as judged by the standard we discussed in class?	Smallest = _________ Largest = _______ How many? _____________
Holm's procedure rejects H_{13}: $\alpha_1 = \alpha_3$ (ie least=most) but accepts the two other subhypotheses H_{12}: $\alpha_1 = \alpha_2$ and H_{23}: $\alpha_2 = \alpha_3$	Circle one: TRUE FALSE

PROBLEM SET #3 STATISTICS 500 FALL 2011: ANSWERS

This is an exam. Do not discuss it with anyone.

1 Use the USCanada data and model 2 on the data page to answer the following questions. Assume model 2 is true.	Fill in or Circle the Correct Answer (7 points each)
1.1 How many submodels does model 2 have? Include model 2 itself and the empty model in your count.	How many: 2^4 = 16
1.2 Of the submodels of model 2, which one model is estimated to have the smallest expected total squared error of prediction as judged by C_P? List the names of the variables in this model, the value of C_P, and the size of the model.	List variables: Age, female, cigsperday, dailyenergy C_P = 5 size = 5
1.3 Of the submodels of model 2, use C_P to identify all the models whose C_P value estimates that the model contains all variables with nonzero coefficients. **Of these models**, pick the **one model** with the smallest size. List the names of the variables in this model, the value of C_P, and the size of the model.	List variables: Age, female, cigsperday, dailyenergy C_P = 5 size = 5
1.4 Give the values of PRESS for model 2 and for the submodel with exactly one predictor namely age.	PRESS for Model 2: 190.30 PRESS with Age alone: 198.73

2 Use the USCanada data and model 2 on the data page to answer the following questions.	Fill in or Circle the Correct Answer (8 points each)
2.1 In model 2, how many observational have large hatvalues h_i (or leverages) by the standard we discussed in class? Give one number.	How many? 547
2.2 The one person with the largest hatvalue is a 21-year old female nonsmoker with an unusually high value for dailyenergy. Give the largest hat value.	Circle one: [✓ TRUE] [✓ FALSE] (True and false both accepted – the variable female was not defined on the data page.) Largest hatvalue = 0.01958
2.3 The person with the largest absolute dffits, |dffits|, shifts his/her own fitted value by more than half of its standard error. What is the signed value of dffits for the person with the largest |dffits|?	Circle one: TRUE [✓ FALSE] Value of dffits = -0.1609

PROBLEM SET #3 STATISTICS 500 FALL 2011: ANSWERS

This is an exam. Do not discuss it with anyone.

3. Use sleepST500 for question 3. Assume the model for the sleep data, model 1, on the data page is true in question 3 and use it in answering the following questions.	Fill in or **circle** the correct answer. (8 points each)
3.1 Test the null hypothesis that H_0: $\alpha_1= \alpha_2= \alpha_3=0$. What is the name of the test statistic? What is the value of the test statistic? What are the degrees of freedom? What is the p-value? Is the null hypothesis plausible?	Name: F-test Value: 11.78 DF: 2 , 45 P-value: 7.7×10^{-5} Circle one: PLAUSIBLE [✓ NOT PLAUSIBLE]
3.2 Use Tukey's method of multiple comparisons to create three simultaneous 95% confidence intervals for the differences $\alpha_i- \alpha_k$. Give the numerical values of the endpoints of the confidence intervals expressed in hours per day.	Middle-Most: [1.364, 7.798] Least-Most: [2.996, 9.429] Least-Middle: [-1.586, 4.848]
Each of the three confidence intervals in 3.2 covers its parameter $\alpha_i- \alpha_k$ in 95% of experiments, but because of the Bonferroni inequality, all three confidence intervals cover their respective parameter in only 1-0.05x3 = 85% of experiments.	Circle one: TRUE [✓ FALSE]
Suppose H_0: $\alpha_1= \alpha_2= \alpha_3=0$ is false. Under this supposition **circle all of the true statements**, if any. (That is, circle 0, 1, 2, 3 or 4 statements.) Refer to the data page for notation, eg H_{12}.	If H_0: $\alpha_1= \alpha_2= \alpha_3=0$ is false, then: [✓ 5) It is possible that H_{12} is true] 6) It is possible that H_{12} and H_{13} are both true [✓ 7) It is possible that H_{12} and H_{13} and H_{23} are all false] [✓ 8) At most one of H_{12} and H_{13} and H_{23} is true.]
Under the model for the sleep data, what is the smallest hatvalue (i.e. leverage)? What is the largest hatvalue? Give two numbers. How many hatvalues are large as judged by the standard we discussed in class?	Smallest = 0.0625 Largest = 0.0625 How many? 0
Holm's procedure rejects H_{13}: $\alpha_1= \alpha_3$ (ie least=most) but accepts the two other subhypotheses H_{12}: $\alpha_1= \alpha_2$ and H_{23}: $\alpha_2= \alpha_3$	Circle one: TRUE [✓ FALSE]

Problem 3, Fall 2011: Doing the Problem Set in R

Question 1.

```
> attach(uscanada)
> y<-(22-bmi)/bmi
> library(leaps)
> x<-uscanada[,c(2,3,5,9)]
> x[1,]
  age female cigsperday dailyenergy
1  44      1          0         1.3
> leaps(x=x,y=y,names=colnames(x))
$which
    age female cigsperday dailyenergy
1 FALSE   TRUE      FALSE       FALSE
1  TRUE  FALSE      FALSE       FALSE
1 FALSE  FALSE      FALSE        TRUE
1 FALSE  FALSE       TRUE       FALSE
2 FALSE   TRUE      FALSE        TRUE
2  TRUE   TRUE      FALSE       FALSE
2 FALSE   TRUE       TRUE       FALSE
2  TRUE  FALSE      FALSE        TRUE
2  TRUE  FALSE       TRUE       FALSE
2 FALSE  FALSE       TRUE        TRUE
3  TRUE   TRUE      FALSE        TRUE
3 FALSE   TRUE       TRUE        TRUE
3  TRUE   TRUE       TRUE       FALSE
3  TRUE  FALSE       TRUE        TRUE
4  TRUE   TRUE       TRUE        TRUE
$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5
$Cp
 [1] 222.86448 360.45091 373.19887 447.93843  99.87721 114.15371
205.02159 301.70463 355.51457 361.02689  22.59063  75.88034 102.04606
294.07859   5.00000

> fr<-(22-bmi)/bmi
> mod2<-lm(fr~age+female+cigsperday+dailyenergy)
> PRESS(mod2)
$PRESS
[1] 190.2983

> PRESS(lm(fr~age))
$PRESS
[1] 198.7274
```

Question 2:

```
> boxplot(hatvalues(mod2))
> sum(hatvalues(mod2)>=2*mean(hatvalues(mod2)))
[1] 547
> which.max(hatvalues(mod2))
6743
> uscanada[6743,]
     country age female cigsperday       dailyenergy  bmi
7285      US  21      0   0                  30.8     23.4
```

Problem 3, Fall 2011: Doing it in R, continued

```
> summary(hatvalues(mod2))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0002423 0.0003537 0.0004649 0.0006225 0.0006564 0.0195800 
> 0.0195800/0.0006225
[1] 31.45382

> summary(dffits(mod2))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
-0.1609000 -0.0142100  0.0002489  0.0002899  0.0143400  0.1417000 
> which.min(dffits(mod2))
5464
```

Question 3:

```
> attach(sleepST500)
> boxplot(totalsleep~sleepdanger)
> summary(aov(totalsleep~sleepdanger))
            Df Sum Sq Mean Sq F value    Pr(>F)    
sleepdanger  2 331.97 165.984  11.777 7.702e-05 ***
Residuals   45 634.24  14.094                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 
> TukeyHSD(aov(totalsleep~sleepdanger))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = totalsleep ~ sleepdanger)

$sleepdanger
                diff       lwr      upr     p adj
Middle-Most  4.58125  1.364335 7.798165 0.0034411
Least-Most   6.21250  2.995585 9.429415 0.0000774
Least-Middle 1.63125 -1.585665 4.848165 0.4425727

> summary(hatvalues(aov(totalsleep~sleepdanger)))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0625  0.0625  0.0625  0.0625  0.0625  0.0625 

> pairwise.t.test(totalsleep,sleepdanger)

        Pairwise comparisons using t tests with pooled SD 

data:  totalsleep and sleepdanger 

       Most   Middle
Middle 0.0024 -     
Least  8e-05  0.2255

P value adjustment method: holm 
```

PROBLEM SET #1 STATISTICS 500 FALL 2012: DATA PAGE 1

Due in class at noon.

This is an exam. Do not discuss it with anyone.

The data are from NHANES, the 2009-2010 National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm). The data are in a data.frame called "fish" with 5000 adults and 43 variables in the course workspace – you must download it again. A csv file, `fish.csv`, is available for those not using R:
`http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/`
SEQN is the NHANES id number. This is a portion of NHANES 2009-2010.
age in years
female = 1 for female, 0 for male
povertyr is income expressed as a ratio of the poverty level (INDFMPIR), so 2 means twice the poverty level. Capped at 5.
education is 1-5 and is described in **educationf**. (DMDEDUC2)
mercury is the mercury level in the blood, (LBXTHG, mercury total ug/L)
cadmium is the cadmium level in the blood (LBXBCD - Blood cadmium ug/L)
lead is lead level in the blood (LBXBPB - Blood lead ug/dL)
The rest of the data frame describes consumption of fish or shellfish over the prior 30 days. **tfish** is total number of servings of fish in the past 30 days, **tshell** is total number of servings of shell fish, breaded is total number of servings of breaded fish (part of tfish), etc. Mr. 51696 is 54, earns a little more than poverty despite being a college graduate, ate 24 servings of fish consisting of 12 servings of tuna and 12 of sardines. Because his mercury level is high, his mercindx is low.

```
> fish[1:2,]
   SEQN age female femalef povertyr education       educationf mercury
1 51696  54      0    male     1.39         5 College Graduate    4.60
2 51796  62      1  female     5.00         5 College Graduate    0.85
  cadmium lead tfish tshell breaded tuna bass catfish cod flatfish
1    0.25 2.01    24      0       0   12    0       0   0        0
2    0.37 0.93    11      6       0    3    0       0   4        0
  haddock mackerel perch pike pollack porgy salmon sardines seabass
1       0        0     0    0       0     0      0       12       0
2       0        0     0    0       0     0      2        0       0
  shark swordfish trout walleye otherfish unknownfish clams crabs
1     0         0     0       0         0           0     0     0
2     0         2     0       0         0           0     0     2
  crayfish lobsters mussels oysters scallops shrimp othershellfish
1        0        0       0       0        0      0              0
2        0        0       0       0        0      4              0
  unknownshellfish
1                0
2                0
> dim(fish)
[1] 5019   43
```

If a question says "A and B and C", true-or-false, then it is true if A and B and C are each true, and it is false if A is true, B is true, but C is false. "North Carolina is north of South Carolina and the moon is made of green cheese" is false. "A is true because of B" is false if A is true, B is true, but A is not true because of B. "A", true-or-false, is false if A is too crazy mean anything sufficiently coherent that it could be true.

PROBLEM SET #1 STATISTICS 500 FALL 201: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

```
> attach(fish)
```

Model #1

mercury $= \beta_0 + \beta_1$ age + β_2 povertyr + β_3 education + β_4 tfish+ β_5 tshell + ε where ε are iid $N(0,\sigma^2)$

Model #2

Define a new variable, lmercury

```
> lmercury<-log(mercury)
```

lmercury $= \gamma_0 + \gamma_1$ age + γ_2 povertyr + γ_3 education + γ_4 tfish+ γ_5 tshell + η where η are iid $N(0,\omega^2)$

Model 1 has slopes β (beta), while model 2 has slopes γ (gamma), so that different things have different names. The choice of Greek letters is arbitrary.

It is often useful to put two plots next to each other on the same page so you can see the same point in both plots. If you type

```
> par(mfrow=c(1,2))
```

then the next two plots will appear on the same page, the first on the left, the second on the right. For question 2, try doing this with a boxplot of the residuals on the left and a quantile-quantile plot of the residuals on the right. The command sets a graphics parameter (that's the 'par'), and it says that there should be 1 row of graphs with 2 columns, filling in the first row first. By setting graph parameters, you can control many aspects of a graph. The free document R for Beginners by Paradis (http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf) contains lots of useful information about graph parameters (see page 43).

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam. **You must sign** the statement on the answer page saying you did not discuss the exam. A perfect exam paper without a signature receives **no credit**.
Due noon in class Thursday 26 Oct.

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2012: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon in class Thursday 26 Oct.

"This exam is my own work. I have not discussed it with anyone."

Your signature: ______________________________

Question (Part 1) (6 points each)	Fill in or CIRCLE the correct answer.
1.1 Plot y=mercury against x=tfish. The four people with the highest levels of mercury all ate more than 20 servings of fish in the previous month.	TRUE FALSE
1.2 Add a lowess smooth to the plot in 1.1. (Use color to see clearly.) The curve tilts upwards, suggesting higher levels of mercury in the blood of people who ate more servings of fish in the previous month.	TRUE FALSE
1.3 A boxplot of mercury levels suggests the distribution is symmetric about its median and free of extreme observations.	TRUE FALSE
1.4 The one person with the highest level of mercury ate two servings of 'otherfish' and one serving of 'scallops' in the previous month.	TRUE FALSE

Fit model 1 from the data page. Use it to answer the questions in part 2 below	Fill in or CIRCLE the correct answer. (Part 2) (7 points each)
2.1 A quantile-quantile plot of residuals from model 1 confirms that the errors in model 1 are correctly modeled as Normally distributed with mean zero and constant variance.	TRUE FALSE
2.2 The Shapiro-Wilk test is a test of the null hypothesis that a group of independent observations is not Normally distributed. Therefore, a small P-value from this test confirms that the observations are Normal.	TRUE FALSE
2.3 Do the Shapiro-Wilk test on the residuals from model 1. What is the P-value? Is it plausible that the residuals are Normally distributed with constant variance?	P-value: ________________ PLAUSIBLE NOT PLAUSIBLE
2.4 Although there are indications that the residuals are not Normal, this is entirely due to a single outlier identified in question 1.4.	TRUE FALSE

Fit model 2 from the data page (Part 3) (6pts)	Fill in or CIRCLE the correct answer.
3. The quantile-quantile plot and Shapiro-Wilk test of of residuals from model 2 confirm that model 2 has Normal errors.	TRUE FALSE

Name: ______________________________ ID# _________________________

PROBLEM SET #1 STATISTICS 500 FALL 201: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon in class Thursday 26 Oct.

Fit model 2 from the data page. For the purpose of answering questions in part 4 below, assume that model 2 is true.	Part 4 Fill in or CIRCLE the correct answer. (7 points each)
4.1 In model 2, test the null hypothesis H_0: $\gamma_1=\gamma_2=\gamma_3=\gamma_4=\gamma_5=0$. What is the name of the test statistic? What is the numerical value of the test statistic? What are the degrees of freedom (DF)? What is the P-value? Is the null hypothesis plausible?	Name:__________ Value: __________ DF = (____, ____) P-value: _________ PLAUSIBLE NOT PLAUSIBLE
4.2 In model 2, test the null hypothesis that the coefficient of education is zero, H_0: $\gamma_3=0$. What is the name of the test statistic? What is the numerical value of the test statistic? What are the degrees of freedom (DF)? What is the P-value? Is the null hypothesis plausible?	Name:__________ Value: __________ DF = ______ P-value: ___________ PLAUSIBLE NOT PLAUSIBLE
4.3 Using the answer to 4.2 and `boxplot(lmercury~educationf)` it is safe to say that professors emit mercury during lectures.	TRUE FALSE CANNOT BE DETERMINED FROM NHANES
4.4 Test the null hypothesis H_0: $\gamma_4=\gamma_5=0$ that neither tfish nor tshell has a nonzero coefficient. What is the name of the test statistic? What is the numerical value of the test statistic? What are the degrees of freedom (DF)? What is the P-value? Is the null hypothesis plausible?	Name:__________ Value: __________ DF = (____, ____) P-value: _________ PLAUSIBLE NOT PLAUSIBLE

Fit model 2 and use it for part 5 below	Fill in or CIRCLE the correct answer. (7 points each)
5.1 For model 2, plot residuals against fitted values, adding a lowess smooth (best in color). The lowess smooth shows no distinctive pattern relevant to regression.	TRUE FALSE Value: ______________
5.2 For model 2, plot residuals against tfish, adding a lowess smooth (best in color). The lowess smooth shows no distinctive pattern relevant to regression.	TRUE FALSE

Problem set 1, Fall 2012, Statistics 500, Answer Page.

Question (Part 1) (6 points each)	Fill in or CIRCLE the correct answer.
1.1 Plot y=mercury against x=tfish. The four people with the highest levels of mercury all ate more than 20 servings of fish in the previous month.	TRUE [✓ FALSE]
1.2 Add a lowess smooth to the plot in 1.1. (Use color to see clearly.) The curve tilts upwards, suggesting higher levels of mercury in the blood of people who ate more servings of fish in the previous month.	[✓ TRUE] FALSE
1.3 A boxplot of mercury levels suggests the distribution is symmetric about its median and free of extreme observations.	TRUE [✓ FALSE]
1.4 The one person with the highest level of mercury ate two servings of 'otherfish' and one serving of 'scallops' in the previous month.	[✓ TRUE] FALSE

Fit model 1 from the data page. Use it to answer the questions in part 2 below	Fill in or CIRCLE the correct answer. (Part 2) (7 points each)
2.1 A quantile-quantile plot of residuals from model 1 confirms that the errors in model 1 are correctly modeled as Normally distributed with mean zero and constant variance.	TRUE [✓ FALSE]
2.2 The Shapiro-Wilk test is a test of the null hypothesis that a group of independent observations is not Normally distributed. Therefore, a small P-value from this test confirms that the observations are Normal.	TRUE [✓ FALSE]
2.3 Do the Shapiro-Wilk test on the residuals from model 1. What is the P-value? Is it plausible that the residuals are Normally distributed with constant variance?	P-value: $< 2.2x10^{-16}$ PLAUSIBLE [✓ NOT PLAUSIBLE]
2.4 Although there are indications that the residuals are not Normal, this is entirely due to a single outlier identified in question 1.4.	TRUE [✓ FALSE]

Fit model 2 from the data page (Part 3) (6pts)	Fill in or CIRCLE the correct answer.
3. The quantile-quantile plot and Shapiro-Wilk test of of residuals from model 2 confirm that model 2 has Normal errors.	TRUE [✓ FALSE]

Problem set 1, Fall 2012, Statistics 500, Answer Page, 2.

Fit model 2 from the data page. For the purpose of answering questions in part 4	Part 4 Fill in or CIRCLE the correct answer.

below, assume that model 2 is true.	(7 points each)
4.1 In model 2, test the null hypothesis H_0: $\gamma_1=\gamma_2=\gamma_3=\gamma_4=\gamma_5=0$. What is the name of the test statistic? What is the numerical value of the test statistic? What are the degrees of freedom (DF)? What is the P-value? Is the null hypothesis plausible?	Name: F-test Value: 364.2 DF = (5, 4994) P-value: $< 2.2\text{x}10^{-16}$ PLAUSIBLE [✓ NOT PLAUSIBLE]
4.2 In model 2, test the null hypothesis that the coefficient of education is zero, H_0: $\gamma_3=0$. What is the name of the test statistic? What is the numerical value of the test statistic? What are the degrees of freedom (DF)? What is the P-value? Is the null hypothesis plausible?	Name: t-test Value: 4.92 DF = 4994 P-value: $8.79 \text{ x } 10^{-7}$ PLAUSIBLE [✓ NOT PLAUSIBLE]
4.3 Using the answer to 4.2 and `boxplot(lmercury~educationf)` it is safe to say that professors emit mercury during lectures.	TRUE FALSE [✓ CANNOT BE DETERMINED FROM NHANES]
4.4 Test the null hypothesis H_0: $\gamma_4=\gamma_5=0$ that neither tfish nor tshell has a nonzero coefficient. What is the name of the test statistic? What is the numerical value of the test statistic? What are the degrees of freedom (DF)? What is the P-value? Is the null hypothesis plausible?	Name: F-test Value: 603.35 DF = (2, 4994) P-value: $< 2.2\text{x}10^{-16}$ PLAUSIBLE [✓ NOT PLAUSIBLE]

Fit model 2 and use it for part 5 below	Fill in or CIRCLE the correct answer. (7 points each)
5.1 For model 2, plot residuals against fitted values, adding a lowess smooth (best in color). The lowess smooth shows no distinctive pattern relevant to regression.	TRUE [✓ FALSE]
5.2 For model 2, plot residuals against tfish, adding a lowess smooth (best in color). The lowess smooth shows no distinctive pattern relevant to regression.	TRUE [✓ FALSE]

Problem Set 1 Fall 2012 Statistics 500 Answers: Doing the Problem Set in R

```
> attach(fish)
1.
> plot(tfish,mercury)
> lines(lowess(tfish,mercury),col="red")
> boxplot(mercury)
> which.max(mercury)
[1] 1184
> fish[1184,]
SEQN  age female femalef povertyr education, etc
54251  48      0    male     0.98         5, etc

2.
> mod1<-lm(mercury ~ age+povertyr+education+tfish+tshell)
> plot(mod1$fit,mod1$resid)
> lines(lowess(mod1$fit,mod1$resid),col="red")
> qqnorm(mod1$resid)
> qqline(mod1$resid)
> boxplot(mod1$resid)
> shapiro.test(mod1$resid)
        Shapiro-Wilk normality test
W = 0.4581, p-value < 2.2e-16
The null hypothesis is Normality.
3.
> lmercury<-log(mercury)
> mod2<-lm(lmercury ~ age +
povertyr+education+tfish+tshell)
> shapiro.test(mod2$resid)
        Shapiro-Wilk normality test
W = 0.9884, p-value < 2.2e-16
> qqnorm(mod2$resid)
> qqline(mod2$resid)
Very far from Normal in many ways.
4.1-4.3
> summary(mod2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.911659   0.047897 -19.034  < 2e-16 ***
age          0.003154   0.000672   4.694 2.75e-06 ***
povertyr     0.087227   0.008071  10.807  < 2e-16 ***
education    0.050681   0.010294   4.923 8.79e-07 ***
tfish        0.074570   0.002815  26.487  < 2e-16 ***
tshell       0.048100   0.003657  13.154  < 2e-16 ***
Residual standard error: 0.8142 on 4994 degrees of freedom
Multiple R-squared: 0.2672,     Adjusted R-squared: 0.2665
F-statistic: 364.2 on 5 and 4994 DF,  p-value: < 2.2e-16
```

Problem Set 1 Fall 2012 Statistics 500 Answers: Doing the Problem Set in R, continued

```
4.4
> modr<-lm(lmercury ~ age + povertyr + education)
> anova(modr,mod2)
Analysis of Variance Table
Model 1: lmercury ~ age + povertyr + education
Model 2: lmercury ~ age + povertyr + education + tfish +
tshell
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   4996 4110.2
2   4994 3310.3  2    799.88 603.35 < 2.2e-16 ***

5.
> plot(mod2$fit,mod2$resid)
> lines(lowess(mod2$fit,mod2$resid),col="red")
> plot(tfish,mod2$resid)
> lines(lowess(tfish,mod2$resid),col="red")
The curves are inverted U’s suggesting curvature that
belongs in the model, not in the residuals.
```

PROBLEM SET #2 STATISTICS 500 FALL 2012: DATA PAGE 1

Due in class at noon on Tuesday 4 December 2012.
This is an exam. Do not discuss it with anyone.

The data are again from NHANES, the 2009-2010 National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm). The data are in a data.frame called "fish" with 5000 adults and 43 variables in the course workspace – you must download it again. A csv file, `fish.csv`, is available for those not using R:
`http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/`
SEQN is the NHANES id number. This is a portion of NHANES 2009-2010.
age in years
female = 1 for female, 0 for male
povertyr is income expressed as a ratio of the poverty level (INDFMPIR), so 2 means twice the poverty level. Capped at 5.
education is 1-5 and is described in **educationf**. (DMDEDUC2)
mercury is the mercury level in the blood, (LBXTHG, mercury total ug/L)
cadmium is the cadmium level in the blood (LBXBCD - Blood cadmium ug/L)
lead is lead level in the blood (LBXBPB - Blood lead ug/dL)
The rest of the data frame describes consumption of fish or shellfish over the prior 30 days. **tfish** is total number of servings of fish in the past 30 days, **tshell** is total number of servings of shell fish, breaded is total number of servings of breaded fish (part of tfish), etc. Ms. 52964 is 80, earns more than 5 times the poverty level, is a college graduate, ate 4 servings of fish, 4 servings of shellfish, including tuna, cod, haddock, salmon, clams and shrimp.

```
> fish[1:2,]
      SEQN age female femalef povertyr education       educationf
580  52964  80       1  female        5         5 College Graduate
1092 57154  60       0    male        5         5 College Graduate
     mercury cadmium lead tfish tshell breaded tuna bass catfish
580     1.23    0.56 2.39     4      4       0    1    0       0
1092    2.00    0.33 2.03     5      4       0    1    0       2
     cod flatfish haddock mackerel perch pike pollack porgy
580    1        0       1        0     0    0       0     0
1092   0        0       0        0     0    0       0     0
     salmon sardines seabass shark swordfish trout walleye
580       1        0       0     0         0     0       0
1092      2        0       0     0         0     0       0
     otherfish unknownfish clams crabs crayfish lobsters mussels
580          0           0     2     0        0        0       0
1092         0           0     0     0        0        0       0
     oysters scallops shrimp othershellfish unknownshellfish
580        0        0      2              0                0
1092       0        0      2              2                0
> dim(fish)
[1] 5000   43
```

If a question says "A and B and C", true-or-false, then it is true if A and B and C are each true, and it is false if A is true, B is true, but C is false. "North Carolina is north of South Carolina and the moon is made of green cheese" is false. "A is true because of B" is false if A is true, B is true, but A is not true because of B. "A", true-or-false, is false if A is too crazy mean anything sufficiently coherent that it could be true.

PROBLEM SET #2 STATISTICS 500 FALL 201: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

```
> attach(fish)
```

Model #1

mercury $= \gamma_0 + \gamma_1$ age + γ_2 povertyr + γ_3 education + γ_4 female + γ_5 tfish + γ_6 tshell
+ γ_6 swordfish + η where η are iid $N(0,\omega^2)$

Model #2

Define a new variable, lmercury

```
> l2mercury<-log2(mercury)     (That l is an L not a 1=one)
```

l2mercury $= \beta_0 + \beta_1$ age + β_2 povertyr + β_3 education + β_4 female + β_5 tfish + β_6 tshell
+ β_7 swordfish + ε where ε are iid $N(0,\sigma^2)$

Model 1 has slopes γ (gamma), while model 2 has slopes β (beta), so that different things have different names. The choice of Greek letters is arbitrary.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam. Due noon in class Tuesday 4 December 2012.

Last Name: ___________________ **First** Name: _____________ ID# ____________

PROBLEM SET #2 STATISTICS 500 FALL 2012: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon in class

Part 1: **Fit Model 1** from the data page and use it to answer the questions in Part 1.	Fill in or CIRCLE the correct answer.
1.1 The P-value from the Shapiro-Wilk test of Normality applied to the residuals from model 1 is very small, less than 0.0001.	TRUE FALSE
1.2 The Box-Cox method (in the MASS package) applied to model 1 suggests that a transformation of mercury by something between the log and the reciprocal would improve the fit of a Normal theory linear model.	TRUE FALSE

Part 2: **Fit model 2** from the data page. For Part 2, assume that model 2 is true.	Fill in or CIRCLE the correct answer.
2.1 Model 2 uses base 2 logs, that is $\log_2$. If $\log_2(A) - \log_2(B) = 2$ then A = 2B.	TRUE FALSE
2.2 What is the 95% confidence interval (CI) for β_4, the coefficient of female?	CI : [________, __________]
2.3 Based on the confidence interval in 2.2, model 2 says that if a man and a woman were the same in terms of other variables in the model (age, povertryr, etc), then one should predict the woman will have half the mercury in her blood that a man will have. (Base this on the confidence interval in 2.2).	TRUE FALSE
2.4 If the coefficient of swordfish, β_7, in model 2 were 0 (that is, if H_0: $\beta_7=0$ were true), then a serving of swordfish would be associated with the same amount of mercury as a serving of a typical serving of fish in tfish.	TRUE FALSE
2.5 In model 2, the hypothesis in 2.4, H_0: $\beta_7=0$, is judged plausible by the appropriate test.	TRUE FALSE

Part 3: Model 2 assumes that tfish has the same slope, namely β_5, for men and women.	Fill in or CIRCLE the correct answer.
3.1 For model 2, test the null hypothesis H_0 that tfish has the same slope for men and women against the alternative that they have different slopes. What is the **name** of the test statistic? What is the **value** of the test statistic? What is the **P-value**? Is H_0 **plausible**?	Name:_________ Value:______ P-value: ____________ CIRCLE ONE Plausible Not Plausible

Last Name: ________________ **First** Name: ____________ ID# ___________

PROBLEM SET #2 STATISTICS 500 FALL 201: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon in class

Part 4: Model 2 fits the $\log_2$(mercury) as linear in tfish. Fit a model similar to model 2 in all ways except that it is quadratic in the one variable tfish alone. (I.e., no crossproducts with other variables. Remember to center.	CIRCLE the correct answer.
4.1 Based on the appropriate statistical test, the linear model 2 seems accurate and the quadratic in tfish is not needed.	TRUE FALSE
4.2 If the point estimate of the coefficient of the quadratic-in-tfish (ie, the single estimated coefficient, not its confidence interval) were taken as the true coefficient, then it would be adding a bit of a U-shape bend (as opposed to a bit of an inverted U shape).	U-shape Inverted U-shape

Part 5: Use the studentized residuals from model 2 to answer part 5.	CIRCLE the correct answer.
5.1 Which person has the largest ABSOLUTE studentized residual? Give the SEQN number. What are the values of mercury, tfish, tshell and swordfish for this person? What is the value (with sign +/-) for this studentized residual?	SEQN: _______ mercury:_____ tfish:____ tshell:____ studentized residual: _________
5.2 Under model 2, test the null hypothesis that the person you identified in 5.1 is not an outlier against the alternative that he is an outlier. What is the value of the t-statistic for this person? What are the degrees of freedom (df)? How small must the two-sided P-value from the t-distribution be for the most extreme observation to be an outlier? Is this person an outlier?	t-statistic:______ df:_______ P-value cutpoint: _________ Outlier Not an Outlier

Part 6: Base your answers on model 2.	CIRCLE the correct answer.
6.1 The person with the largest leverage (ie hatvalue) ate 32 servings of fish during the month including 8 servings of swordfish (True/False). Give the value of this largest leverage.	TRUE FALSE Value of leverage: _________
6.2 Which person has the largest <u>absolute</u> DFFITS? Give the SEQN and the value of DFFITS for this individual with its sign (+/-). This person does not stand out in a boxplot of the 5000 DFFITS (true/false)	SEQN:_______ DFFITS:______ TRUE FALSE

PROBLEM SET #2 STATISTICS 500 FALL 2012: ANSWER PAGE 1

Answers

Part 1: **Fit Model 1** from the data page and use it to answer the questions in Part 1.	Fill in or CIRCLE the correct answer.
1.1 The P-value from the Shapiro-Wilk test of Normality applied to the residuals from model 1 is very small, less than 0.0001.	[✓ TRUE] FALSE
1.2 The Box-Cox method (in the MASS package) applied to model 1 suggests that a transformation of mercury by something between the log and the reciprocal would improve the fit of a Normal theory linear model.	[✓ TRUE] FALSE

Part 2: **Fit model 2** from the data page. For Part 2, assume that model 2 is true.	Fill in or CIRCLE the correct answer
2.1 Model 2 uses base 2 logs, that is $\log_2$. If $\log_2(A) - \log_2(B) = 2$ then A = 2B.	TRUE [✓ FALSE]
2.2 What is the 95% confidence interval (CI) for β_4, the coefficient of female?	CI : [-0.161 , -0.031]
2.3 Based on the confidence interval in 2.2, model 2 says that if a man and a woman were the same in terms of other variables in the model (age, povertryr, etc), then one should predict the woman will have half the mercury in her blood that a man will have. (Base this on the confidence interval in 2.2).	TRUE [✓ FALSE]
2.4 If the coefficient of swordfish, β_7, in model 2 were 0 (that is, if H_0: $\beta_7=0$ were true), then a serving of swordfish would be associated with the same amount of mercury as a serving of a typical serving of fish in tfish.	[✓ TRUE] FALSE
2.5 In model 2, the hypothesis in 2.4, H_0: $\beta_7=0$, is judged plausible by the appropriate test.	TRUE [✓ FALSE]

Part 3: Model 2 assumes that tfish has the same slope, namely β_5, for men and women.	Fill in or CIRCLE the correct answer.
3.1 For model 2, test the null hypothesis H_0 that tfish has the same slope for men and women against the alternative that they have different slopes. What is the **name** of the test statistic? What is the **value** of the test statistic? What is the **P-value**? Is H_0 **plausible**?	Name: t-test Value: -1.621 P-value: 0.105 CIRCLE ONE [✓ Plausible] Not Plausible

PROBLEM SET #2 STATISTICS 500 FALL 201: ANSWER PAGE 2

Answers

Part 4: Model 2 fits the log$_2$(mercury) as linear in tfish. Fit a model similar to model 2 in all ways except that it is quadratic in the one variable tfish alone. (I.e., no crossproducts with other variables. Remember to center.	CIRCLE the correct answer.
4.1 Based on the appropriate statistical test, the linear model 2 seems accurate and the quadratic in tfish is not needed.	TRUE [✓ FALSE]
4.2 If the point estimate of the coefficient of the quadratic in tfish (ie, the single estimated coefficient, not its confidence interval) were taken as the true coefficient, then it would be adding a bit of a U-shape bend (as opposed to a bit of an inverted U shape).	U-shape [✓ Inverted U-shape]

Part 5: Use the studentized residuals from model 2 to answer part 5.	CIRCLE the correct answer.
5.1 Which person has the largest ABSOLUTE studentized residual? Give the SEQN number. What are the values of mercury, tfish, tshell and swordfish for this person? What is the value (with sign +/-) for this studentized residual?	SEQN: 54251 mercury: 85.7 tfish: 2 tshell: 1 studentized residual: 5.748
5.2 5.2 Under model 2, test the null hypothesis that the person you identified in 5.1 is not an outlier against the alternative that he is an outlier. What is the value of the t-statistic for this person? What are the degrees of freedom (df)? How small must the two-sided P-value from the t-distribution be for the most extreme observation to be an outlier? Is this person an outlier?	t-statistic: 5.748 df: 4991 P-value cutpoint: 0.00001==.05/5000 [✓ Outlier] Not an Outlier

Part 6: Base your answers on model 2.	CIRCLE the correct answer.
6.1 The person with the largest leverage (ie hatvalue) ate 32 servings of fish during the month including 8 servings of swordfish (True/False). Give the value of this largest leverage.	[✓ TRUE] FALSE Value: 0.2685
6.2 Which person has the largest <u>absolute</u> DFFITS? Give the SEQN and the value of DFFITS for this individual with its sign (+/-). This person does <u>not</u> stand out as extreme in a boxplot of the 5000 DFFITS (true/false)	SEQN: 52288 DFFITS: -3.008 TRUE [✓ FALSE]

Problem 2, Fall 2012 Answers
Doing the Problem in R

```
> attach(fish)
```

Part 1.

```
> md1<-lm(mercury~age+povertyr+education+female+tfish+tshell+swordfish)
```

1.1

```
> shapiro.test(md1$resid)
        Shapiro-Wilk normality test
data:  md1$resid
W = 0.4574, p-value < 2.2e-16
```

This is strong evidence that the residuals are not Normal.

1.2

```
> library(MASS)
> boxcox(md1)
```

The plausible values of λ are between 0 and -1, closer to 0. Remember 0 is the log, while -1 is the reciprocal. Actually, -1/10 is somewhat better than 0 or -1 according to boxcox.

Part 2.

2.1 Right idea, but wrong value:

```
> log2(8)-log2(2)
[1] 2
> 8/2
[1] 4
> log2(12)-log2(3)
[1] 2
> 12/3
[1] 4
```

2.2

```
> md2<-
lm(l2mercury~age+povertyr+education+female+tfish+tshell+swordfish)
> confint(md2)
                  2.5 %       97.5 %
(Intercept) -1.377588658 -1.097776489
age          0.002494721  0.006278769
…
female      -0.161245428 -0.030922609
…
swordfish    0.349718792  0.651199942
```

2.3 Taking antilogs, the interval of multipliers does not include ½.

```
> 2^(-0.161245428)
[1] 0.8942528
> 2^(-0.030922609)
[1] 0.9787942

> summary(md2)

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.2376826  0.0713647 -17.343  < 2e-16 ***
…
swordfish    0.5004594  0.0768912   6.509 8.33e-11 ***
```

Problem 2, Fall 2012 Answers

Part 3

```
> tfishfemale<-tfish*female
> summary(lm(l2mercury ~ age + povertyr + education +
female +
+     tfish + tshell + swordfish + tfishfemale))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.2572879  0.0723705 -17.373  < 2e-16 ***
age          0.0043904  0.0009649   4.550 5.49e-06 ***
…
tfishfemale -0.0121300  0.0074821  -1.621    0.105
```

Part 4

```
> tfish2<-(tfish-mean(tfish))^2
> summary(lm(l2mercury ~ age + povertyr + education + female +
+      tfish + tshell + swordfish + tfish2))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.1999318  0.0702395 -17.083  < 2e-16 ***
age          0.0033022  0.0009527   3.466 0.000533 ***
…
tfish2      -0.0039278  0.0003004 -13.074  < 2e-16 ***
```

Part 5

```
> which.max(abs(rstudent(md2)))
1184
1184
> fish[1184,]
> rstudent(md2)[1184]
5.74752
> 2*5000*pt(-5.74752,4990)
[1] 4.797004e-05
```

Part 6

6.1

```
> which.max(hatvalues(md2))
1010
> fish[1010,]
> max(hatvalues(md2))
[1] 0.2685011
> mean(hatvalues(md2))*2
[1] 0.0032
```

6.2

```
> which.max(abs(dffits(md2)))
1010
> dffits(md2)[1010]
     1010
-3.007509  Wow!  So much for model 2!
> boxplot(dffits(md2))
```

PROBLEM SET #3 STATISTICS 500 FALL 2012: DATA PAGE 1

Due at noon Wednesday December 19, 2012, at my office, 473 JMHH.

This is an exam. Do not discuss it with anyone.

The first data are again from NHANES, the 2009-2010 National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm). The data are in a data.frame called "fish" with 5000 adults and 43 variables in the course workspace – you must download it again. A csv file, `fish.csv`, is available for those not using R:
http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/
SEQN is the NHANES id number. This is a portion of NHANES 2009-2010.
age in years **female** = 1 for female, 0 for male
mercury is the mercury level in the blood, (LBXTHG, mercury total ug/L)
The rest of the data frame describes consumption of fish or shellfish over the prior 30 days. **tfish** is total number of servings of fish in the past 30 days, **tshell** is total number of servings of shell fish, breaded is total number of servings of breaded fish (part of tfish), etc. Ms. 52964 is 80, earns more than 5 times the poverty level, is a college graduate, ate 4 servings of fish, 4 servings of shellfish, including tuna, cod, haddock, salmon, clams and shrimp.

```
> dim(fish)
[1] 5000   43
```

If a question says "A and B and C", true-or-false, then it is true if A and B and C are each true, and it is false if A is true, B is true, but C is false. "North Carolina is north of South Carolina and the moon is made of green cheese" is false. "A is true because of B" is false if A is true, B is true, but A is not true because of B. "A", true-or-false, is false if A is too crazy mean anything sufficiently coherent that it could be true.

FishMod: For the fish data, let y = log2(mercury) and consider a linear model for y using the following predictors: "age" "female" "tfish" "breaded" "tuna" "cod" "salmon" "sardines" "shark" "swordfish". So there are 10 predictors. Assume y is linear in the predictors with independent errors have mean zero, constant variance, and a Normal distribution.

The **ceramic data** is in the object `ceramic` in the course R workspace. It is also available `ceramic.csv` at the web page above. It is taken from an experiment done at the Ceramics Division, Materials Science and Engineering Lab, NIST. They describe it as follows: "The original data set was part of a high performance ceramics experiment with the goal of characterizing the effect of grinding parameters on sintered reaction-bonded silicon nitride. " There are 32 observations. The outcome, Y, is the strength of the ceramic material. We will focus on two factors, grit = wheel grit (140/170 or 80/100) and direction (longitudinal or transverse). The variable GD combines grit and direction into one nominal variable with four levels. Do a one-way anova with 4 groups. As a linear model (the **CeramicModel**), assume Y=strength is independently and Normally distributed with constant variance and a mean that depends upon grit and direction (GD).

IMPORTANT: When asked to give the name of a group, use the short form in gd, such as 140:L.

PROBLEM SET #3 STATISTICS 500 FALL 201: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

```
> ceramic
   speed rate grit direction batch strength order          GD    gd
1     -1   -1   -1        -1    -1   680.45    17 140/170:long 140:L
2      1   -1   -1        -1    -1   722.48    30 140/170:long 140:L
3     -1    1   -1        -1    -1   702.14    14 140/170:long 140:L
4      1    1   -1        -1    -1   666.93     8 140/170:long 140:L
5     -1   -1    1        -1    -1   703.67    32  80/100:long  80:L
6      1   -1    1        -1    -1   642.14    20  80/100:long  80:L
7     -1    1    1        -1    -1   692.98    26  80/100:long  80:L
8      1    1    1        -1    -1   669.26    24  80/100:long  80:L
9     -1   -1   -1         1    -1   491.58    10 140/170:tran 140:T
10     1   -1   -1         1    -1   475.52    16 140/170:tran 140:T
11    -1    1   -1         1    -1   478.76    27 140/170:tran 140:T
12     1    1   -1         1    -1   568.23    18 140/170:tran 140:T
13    -1   -1    1         1    -1   444.72     3  80/100:tran  80:T
14     1   -1    1         1    -1   410.37    19  80/100:tran  80:T
15    -1    1    1         1    -1   428.51    31  80/100:tran  80:T
16     1    1    1         1    -1   491.47    15  80/100:tran  80:T
17    -1   -1   -1        -1     1   607.34    12 140/170:long 140:L
18     1   -1   -1        -1     1   620.80     1 140/170:long 140:L
19    -1    1   -1        -1     1   610.55     4 140/170:long 140:L
20     1    1   -1        -1     1   638.04    23 140/170:long 140:L
21    -1   -1    1        -1     1   585.19     2  80/100:long  80:L
22     1   -1    1        -1     1   586.17    28  80/100:long  80:L
23    -1    1    1        -1     1   601.67    11  80/100:long  80:L
24     1    1    1        -1     1   608.31     9  80/100:long  80:L
25    -1   -1   -1         1     1   442.90    25 140/170:tran 140:T
26     1   -1   -1         1     1   434.41    21 140/170:tran 140:T
27    -1    1   -1         1     1   417.66     6 140/170:tran 140:T
28     1    1   -1         1     1   510.84     7 140/170:tran 140:T
29    -1   -1    1         1     1   392.11     5  80/100:tran  80:T
30     1   -1    1         1     1   343.22    13  80/100:tran  80:T
31    -1    1    1         1     1   385.52    22  80/100:tran  80:T
32     1    1    1         1     1   446.73    29  80/100:tran  80:T
```

Remark on question 3.6: The degrees of freedom between the four levels of GD may be partitioned into a main effect of G, a main effect of D, and an interaction or cross-product of G and D. You can do this with regression (grit, direction) or with contrasts.

Follow instructions. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. **Make and keep a photocopy of your answer page**. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

The exam is due **at noon Wednesday December 19, 2012, at my office, 473 JMHH**. You may turn in the exam early by placing it in an envelope addressed to me and leaving it in my mail box in statistics, 4th floor, Huntsman Hall. You may give it to Adam at the front desk in statistics if you prefer. Make and keep a photocopy of your answer page. The answer key will be posted in the revised bulk pack on-line.

Last Name: __________________ **First** Name: _____________ ID# ____________

PROBLEM SET #3 STATISTICS 500 FALL 2012: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due Wed, December 19, 2012, noon.

Use FishMod to answer the questions in part 1.	Fill in/CIRCLE the Correct Answer
1.1 Consider all of the models that can be formed from FishMod by using subsets of variables, including the model with 10 predictors and the model with no predictors. How many models are there?	Number of models = __________
1.2 For the models in 1.1, what is the smallest value of C_P? How many predictor variables are in this model? How many regression slope (beta) parameters are in this model, including the constant as one parameter?	Smallest C_P: ___________ predictors: _____ parameters:____
1.3 For the 10 predictors in FishMod, list the predictors that are NOT in the model in 1.2 with the smallest C_P.	List of predictor names:
1.4 Whenever any model in any problem is the submodel with the smallest C_P then by virtue of having the smallest C_P this model is estimated to have all predictors with nonzero coefficients.	TRUE FALSE
1.5 The model identified in 1.2 clearly does NOT have all of the predictors with nonzero coefficients based on comparing the value of C_P and the number of parameters in this model.	TRUE FALSE
1.6 Of the models mentioned in 1.1, the model identified in 1.2 (smallest C_P) also is the model with the largest R^2.	TRUE FALSE
1.7 Comparing C_P values to the number of parameters in the model, there is a model with 4 predictors that is estimated to have all of the predictors with nonzero coefficients, but the model in 1.2 is estimated to predict y more accurately.	TRUE FALSE
1.8 Using the model with all 10 predictors, which one of the 10 predictors has the largest variance inflation factor (vif)? What is the value of this one vif? What is the R^2 of this variable with the other 9 predictors in the 10 predictor model?	Variable name: _____________ vif: __________ R^2: ____________
1.9 In the model in 1.2 (lowest C_P), a serving of breaded fish is estimated to be worse than a serving of fish unspecified by the variables in the model, breaded fish being associated with extra mercury.	TRUE FALSE

Last Name: ___________________ **First** Name: _____________ ID# ____________

PROBLEM SET #2 STATISTICS 500 FALL 201: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone.

2. View the ceramic data in terms of a one-way analysis of variance with four groups defined by GD. Fill in the following analysis of variance table.

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic
Between Groups				
Within Groups (Residual)				XXXXXXXX XXXXXXXX

3. Use the ceramic data and the CeramicModel to answer the questions in part 3.	Fill in or CIRCLE the correct answer
3.1 Use the anova table in part 2 to test the null hypothesis that the four groups do not differ. What is the P-value for the F-statistic? Is it plausible that the four groups do not differ?	P-value: ________________ PLAUSIBLE NOT PLAUSIBLE
3.2 Four treatment groups defined by GD may be compared in pairs, group 1 to group 2, group 1 to group 3, etc. How many distinct comparisons are there of two groups? (Group 1 with 2 is the same comparison as group 2 with group 1).	Number of comparisons: _____________
3.3 Use Tukey's method to perform all of the comparisons in 3.2 at an experiment-wise error rate of 5%. List ALL comparisons that are NOT significant. (If none, write none.) One possible comparison is "140:L vs 80:T".	
3.4 Use Holm's method to perform all of the comparisons in 3.2 at an experiment-wise error rate of 5%. List ALL comparisons that are NOT significant. (If none, write none.) One possible comparison is "140:L vs 80:T".	
3.5 To say that the experiment-wise error rate is strongly controlled at 5% is to say that, no matter which groups truly differ, the chance of falsely declaring at least one pair of groups different is at most 5%.	TRUE FALSE
3.6 See remark on the data page. Test the null hypothesis H_0 that there is no interaction between grit and direction.	H_0 is: P-value:_______ Plausible Not Plausible

PROBLEM SET #3 STATISTICS 500 FALL 2012: DATA PAGE 1

Due in class at noon .

This is an exam. Do not discuss it with anyone.

The first data are again from NHANES, the 2009-2010 National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm). The data are in a data.frame called “fish” with 5000 adults and 43 variables in the course workspace – you must download it again. A csv file, `fish.csv`, is available for those not using R: `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/`
SEQN is the NHANES id number. This is a portion of NHANES 2009-2010.
age in years **female** = 1 for female, 0 for male
mercury is the mercury level in the blood, (LBXTHG, mercury total ug/L)
The rest of the data frame describes consumption of fish or shellfish over the prior 30 days. **tfish** is total number of servings of fish in the past 30 days, **tshell** is total number of servings of shell fish, breaded is total number of servings of breaded fish (part of tfish), etc. Ms. 52964 is 80, earns more than 5 times the poverty level, is a college graduate, ate 4 servings of fish, 4 servings of shellfish, including tuna, cod, haddock, salmon, clams and shrimp.

```
> dim(fish)
[1] 5000   43
```

If a question says “A and B and C”, true-or-false, then it is true if A and B and C are each true, and it is false if A is true, B is true, but C is false. “North Carolina is north of South Carolina and the moon is made of green cheese” is false. “A is true because of B” is false if A is true, B is true, but A is not true because of B. “A”, true-or-false, is false if A is too crazy mean anything sufficiently coherent that it could be true.

FishMod: For the fish data, let y = log2(mercury) and consider a linear model for y using the following predictors: "age" "female" "tfish" "breaded" "tuna" "cod" "salmon" "sardines" "shark" "swordfish". So there are 10 predictors. Assume y is linear in the predictors with independent errors have mean zero, constant variance, and a Normal distribution.

The **ceramic data** is in the object `ceramic` in the course R workspace. It is taken from an experiment done at the Ceramics Division, Materials Science and Engineering Lab, NIST. They describe it as follows: “The original data set was part of a high performance ceramics experiment with the goal of characterizing the effect of grinding parameters on sintered reaction-bonded silicon nitride. “ There are 32 observations. The outcome, Y, is the strength of the ceramic material. We will focus on two factors, grit = wheel grit (140/170 or 80/100) and direction (longitudinal or transverse). The variable GD combines grit and direction into one nominal variable with four levels. As a linear model (the **CeramicModel**), assume Y=strength is independently and Normally distributed with constant variance and a mean that depends upon grit and direction (GD).

IMPORTANT: When asked to give the name of a group, use the short form in gd, such as 140:L.

PROBLEM SET #3 STATISTICS 500 FALL 2012: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

```
> ceramic
   speed rate grit direction batch strength order           GD    gd
1     -1   -1   -1        -1    -1   680.45    17 140/170:long 140:L
2      1   -1   -1        -1    -1   722.48    30 140/170:long 140:L
3     -1    1   -1        -1    -1   702.14    14 140/170:long 140:L
4      1    1   -1        -1    -1   666.93     8 140/170:long 140:L
5     -1   -1    1        -1    -1   703.67    32  80/100:long  80:L
6      1   -1    1        -1    -1   642.14    20  80/100:long  80:L
7     -1    1    1        -1    -1   692.98    26  80/100:long  80:L
8      1    1    1        -1    -1   669.26    24  80/100:long  80:L
9     -1   -1   -1         1    -1   491.58    10 140/170:tran 140:T
10     1   -1   -1         1    -1   475.52    16 140/170:tran 140:T
11    -1    1   -1         1    -1   478.76    27 140/170:tran 140:T
12     1    1   -1         1    -1   568.23    18 140/170:tran 140:T
13    -1   -1    1         1    -1   444.72     3  80/100:tran  80:T
14     1   -1    1         1    -1   410.37    19  80/100:tran  80:T
15    -1    1    1         1    -1   428.51    31  80/100:tran  80:T
16     1    1    1         1    -1   491.47    15  80/100:tran  80:T
17    -1   -1   -1        -1     1   607.34    12 140/170:long 140:L
18     1   -1   -1        -1     1   620.80     1 140/170:long 140:L
19    -1    1   -1        -1     1   610.55     4 140/170:long 140:L
20     1    1   -1        -1     1   638.04    23 140/170:long 140:L
21    -1   -1    1        -1     1   585.19     2  80/100:long  80:L
22     1   -1    1        -1     1   586.17    28  80/100:long  80:L
23    -1    1    1        -1     1   601.67    11  80/100:long  80:L
24     1    1    1        -1     1   608.31     9  80/100:long  80:L
25    -1   -1   -1         1     1   442.90    25 140/170:tran 140:T
26     1   -1   -1         1     1   434.41    21 140/170:tran 140:T
27    -1    1   -1         1     1   417.66     6 140/170:tran 140:T
28     1    1   -1         1     1   510.84     7 140/170:tran 140:T
29    -1   -1    1         1     1   392.11     5  80/100:tran  80:T
30     1   -1    1         1     1   343.22    13  80/100:tran  80:T
31    -1    1    1         1     1   385.52    22  80/100:tran  80:T
32     1    1    1         1     1   446.73    29  80/100:tran  80:T
```

Remark on question 3.6: The degrees of freedom between the four levels of GD may be partitioned into a main effect of G, a main effect of D, and an interaction or cross-product of G and D. You can do this with regression (grit, direction) or with contrasts.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Write your name and id number on **both sides** of the answer page. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

PROBLEM SET #3 STATISTICS 500 FALL 2012: ANSWER PAGE 1, answers

This is an exam. Do not discuss it with anyone.

Use FishMod to answer the questions in part 1.	Fill in/CIRCLE the Correct Answer
1.1 Consider all of the models that can be formed from FishMod by using subsets of variables, including the model with 10 predictors and the model with no predictors. How many models are there?	Number of models = 2^{10} =1024
1.2 For the models in 1.1, what is the smallest value of C_P? How many predictor variables are in this model? How many regression slope (beta) parameters are in this model, including the constant as one parameter?	Smallest C_P: 8.51 predictors: 8 parameters: 9
1.3 For the 10 predictors in FishMod, list the predictors that are NOT in the model in 1.2 with the smallest C_P.	List of predictor names: Tuna Cod
1.4 Whenever any model in any problem is the submodel with the smallest C_P then by virtue of having the smallest C_P this model is estimated to have all predictors with nonzero coefficients.	TRUE [✓ FALSE]
1.5 The model identified in 1.2 clearly does NOT have all of the predictors with nonzero coefficients based on comparing the value of C_P and the number of parameters in this model.	TRUE [✓ FALSE]
1.6 Of the models mentioned in 1.1, the model identified in 1.2 (smallest C_P) also is the model with the largest R^2.	TRUE [✓ FALSE]
1.7 Comparing C_P values to the number of parameters in the model, there is a model with 4 predictors that is estimated to have all of the predictors with nonzero coefficients, but the model in 1.2 is estimated to predict y more accurately.	TRUE [✓ FALSE]
1.8 Using the model with all 10 predictors, which one of the 10 predictors has the largest variance inflation factor (vif)? What is the value of this one vif? What is the R^2 of this variable with the other 9 predictors in the 10 predictor model?	Variable name: tfish vif: 3.03 R^2: 0.67
1.9 In the model in 1.2 (lowest C_P), a serving of breaded fish is estimated to be worse than a serving of fish unspecified by the variables in the model, breaded fish being associated with extra mercury.	TRUE [✓ FALSE]

PROBLEM SET #2 STATISTICS 500 FALL 2012: ANSWER PAGE 2, answers

This is an exam. Do not discuss it with anyone.

2. View the ceramic data in terms of a one-way analysis of variance with four groups defined by GD. Fill in the following analysis of variance table.

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic
Between Groups	330955	3	110318	51.6
Within Groups (Residual)	59845	28	2137	XXXXXXXX XXXXXXXX

3. Use the ceramic data/CeramicModel to answer the questions in part 3.	Fill in or CIRCLE the correct answer
3.1 Use the anova table in part 2 to test the null hypothesis that the four groups do not differ. What is the P-value for the F-statistic? Is it plausible that the four groups do not differ?	P-value: 1.56×10^{-11} PLAUSIBLE [✓ NOT PLAUSIBLE]
3.2 Four treatment groups defined by GD may be compared in pairs, group 1 to group 2, group 1 to group 3, etc. How many distinct comparisons are there of two groups? (Group 1 with 2 is the same comparison as group 2 with group 1).	Number of comparisons: 4x3/2 = 6
3.3 Use Tukey's method to perform all of the comparisons in 3.2 at an experiment-wise error rate of 5%. List ALL comparisons that are NOT significant. (If none, write none.) One possible comparison is "140:L vs 80:T".	80:T vs 140:T 80:L vs 140:L
3.4 Use Holm's method to perform all of the comparisons in 3.2 at an experiment-wise error rate of 5%. List ALL comparisons that are NOT significant. (If none, write none.) One possible comparison is "140:L vs 80:T".	80:L vs 140:L
3.5 To say that the experiment-wise error rate is strongly controlled at 5% is to say that, no matter which groups truly differ, the chance of falsely declaring at least one pair of groups different is at most 5%.	[✓ TRUE] FALSE
3.6 See remark on the data page. Test the null hypothesis H_0 that there is no interaction between grit and direction.	P-value: 0.234 H_0 is: [✓ Plausible] Not Plausible

Statistics 500, Problem Set 3, Fall 2012

Doing the Problem Set in R

Problem 1: fish

```
> X<-fish[,c(2,3,11,13,14,17,25,26,28,29)]
> y<-log2(fish$mercury)
> rfish<-leaps(x=X,y=y,names=colnames(X))
> which.min(rfish$Cp)
[1] 71
> rfish$size[71]
[1] 9
> rfish$Cp[71]
[1] 8.514598
> rfish$which[71,]
      age    female     tfish   breaded      tuna       cod    salmon
     TRUE      TRUE      TRUE      TRUE     FALSE     FALSE      TRUE
 sardines     shark swordfish
     TRUE      TRUE      TRUE
```

1.8 Variance inflation factor

```
> library(DAAG)
> md1<-
lm(y~age+female+tfish+breaded+tuna+cod+salmon+sardines+shark+swordfish)
> vif(md1)
      age    female     tfish   breaded      tuna
   1.0199    1.0092    3.0295    1.0911    1.7010
      cod    salmon  sardines     shark swordfish
   1.1187    1.6189    1.1391    1.0096    1.0402
> 1-(1/vif(md1))
        age      female       tfish     breaded
0.019511717 0.009116132 0.669912527 0.083493722
       tuna         cod      salmon    sardines
0.412110523 0.106105301 0.382296621 0.122113950
      shark   swordfish
0.009508716 0.038646414
```

Problem 2 and 3.1

```
> anova(lm(strength~GD))
Analysis of Variance Table

Response: strength
          Df Sum Sq Mean Sq F value    Pr(>F)
GD         3 330955  110318  51.615 1.565e-11 ***
Residuals 28  59845    2137
```

Statistics 500, Problem Set 3, Fall 2012
Doing the Problem Set in R, continued

3.3

```
> TukeyHSD(aov(strength~GD))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = strength ~ GD)
$GD
                                diff        lwr         upr     p adj
140/170:tran-140/170:long -178.60375 -241.71667 -115.490825 0.0000001
80/100:long-140/170:long   -19.91750  -83.03042   43.195425 0.8243531
80/100:tran-140/170:long  -238.26000 -301.37292 -175.147075 0.0000000
80/100:long-140/170:tran   158.68625   95.57333  221.799175 0.0000011
80/100:tran-140/170:tran   -59.65625 -122.76917    3.456675 0.0690593
80/100:tran-80/100:long   -218.34250 -281.45542 -155.229575 0.0000000
```

Looking at the results for Tukey’s test, one might think there is no effect of grit. If it’s long vs long or tran vs tran, the pair comparison is not significant.

3.4

```
> pairwise.t.test(strength,gd)
     Pairwise comparisons using t tests with pooled SD
data:  strength and gd
      140:L   140:T   80:L
140:T 8.2e-08 -       -
80:L  0.396   5.5e-07 -
80:T  2.9e-10 0.031   1.7e-09
P value adjustment method: holm
```

Notice that Holm’s method found an additional difference, an effect of grit at direction tran. So grit matters after all. The Bonferroni method fails to find a grit effect, agreeing with Tukey’s method.

```
> pairwise.t.test(strength,gd,p.adjust.method="b")
      Pairwise comparisons using t tests with pooled SD
data:  strength and gd

      140:L   140:T   80:L
140:T 1.2e-07 -       -
80:L  1.000   1.1e-06 -
80:T  2.9e-10 0.092   2.0e-09
P value adjustment method: bonferroni
```

3.5

```
> summary(aov(strength~grit*direction))
               Df Sum Sq Mean Sq  F value    Pr(>F)
grit            1  12664   12664   5.9251   0.02156 *
direction       1 315133  315133 147.4420 1.127e-12 ***
grit:direction  1   3158    3158   1.4777   0.23429
Residuals      28  59845    2137
```

You could also do this with t-tests in a regression model with interactions.

PROBLEM SET #1 STATISTICS 500 FALL 2013: DATA PAGE 1
Due in class at noon on Tuesday, October 22, 2013.
This is an exam. Do not discuss it with anyone.

The data are from NHANES, the 1999-2000 National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm). The data describe people over age 60 who took a cognitive test; see below. The data are in a data.frame called "cogscore" with 1108 adults and 16 variables in the course workspace – you must download it again. A csv file, `cogscore.csv`, is available for those not using R:
`http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/`
The list is case-sensitivity, so `cogscore.csv` is at the end, with the lower case files. The file is simplified in several ways; in particular, missing data have been removed.

cogsc cognitive test score (age>=60) (CFDRIGHT)
age in years
bmi body mass index
educ is 1-5 and is described in **educf**. (DMDEDUC2)
alcohol Alcohol consumed in grams (DRXTALCO)
caffeine Caffeine consumed in miligrams (DRXTCAFF)
lead is lead level in the blood (LBXBPB - Blood lead ug/dL)
cadmium is the cadmium level in the blood (LBXBCD - Blood cadmium ug/L)
smoke100life =1 if smoked 100 cigarettes in life, 0 otherwise (SMQ020)
smokenow Smoke now, explained in smokenowf (SMQ040)
cigsperday Cigarettes per day for people who smoke everyday (SMD070)
female = 1 for female, 0 for male
vigorous =1 if any vigorous physical activity in past 30 days, 0 otherwise (PAD200)
moderate =1 if any moderate physical activity in past 30 days, 0 otherwise (PAD320)

The cogscore is described by NHANES as follows:
"This section contains the results of a version of the WAIS III (Wechsler Adult Intelligence Scale, Third Edition) Digit Symbol – Coding module conducted during the household interview. The subtest was administered under a licensing agreement with The Psychological Corporation. In this coding exercise, participants copy symbols that are paired with numbers. Using the key provided at the top of the exercise form, the participant draws the symbol under the corresponding number. The score is the number of correct symbols drawn within a period of 120 seconds. One point is given for each correctly drawn symbol completed within the time limit. The maximum score is 133. Sample items are provided for initial practice. Participants who are unable to complete any of the sample items do not continue with the remainder of the exercise."

Starbucks web page indicates that a grande Pike's roast brewed coffee has 330 mg of caffeine.

The cogscore is described at http://en.wikipedia.org/wiki/Digit_symbol_substitution_test

PROBLEM SET #1 STATISTICS 500 FALL 2013: DATA PAGE 2

This is an exam. Do not discuss it with anyone. Due Tuesday, October 22, 2013.

Before you do anything else, plot the data in various ways. For example:

```
> attach(cogscore)
> boxplot(cogsc~educf)
> plot(age,cogsc)
> lines(lowess(age,cogsc))
> plot(lead,cogsc)
> lines(lowess(lead,cogsc))
```

Model #1

cogsc $= \beta_0 + \beta_1$ age + β_2 bmi + β_3 educ + β_4 alcohol + β_5 cigsperday + ε where ε are iid $N(0,\sigma^2)$

Model #2

cogsc $= \gamma_0 + \gamma_1$ age + γ_2 educ + γ_3 lead + γ_4 caffeine + η where η are iid $N(0,\omega^2)$

Model 1 has slopes β (beta), while model 2 has slopes γ (gamma), so that different things have different names. The choice of Greek letters is arbitrary.

It is often useful to put two plots next to each other on the same page so you can see the same point in both plots. If you type

```
> par(mfrow=c(1,2))
```

then the next two plots will appear on the same page, the first on the left, the second on the right. For example, you can compare a boxplot and a Normal quantile plot in this way. The command sets a graphics parameter (that's the 'par'), and it says that there should be 1 row of graphs with 2 columns, filling in the first row first. By setting graph parameters, you can control many aspects of a graph. The free document R for Beginners by Paradis (http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf) contains lots of useful information about graph parameters (see page 43).

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name: _____________________________ ID# _________________________

PROBLEM SET #1 STATISTICS 500 FALL 2013: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon in class Tuesday 22 Oct.

Question (Part 1)	CIRCLE the correct answer.
1.1 In the plot of cogsc (as y) against age (60-85 as x), the lowess curve increases suggesting higher scores at older ages.	TRUE FALSE
1.2 Looking at alcohol consumption by the levels of education, the sample upper (3rd) quartile of alcohol consumption is highest among individuals with a college degree.	TRUE FALSE
1.3 The sample median cogsc increases over the five levels of education.	TRUE FALSE
1.4 Using Pearson (i.e. usual) correlation, it is correct to say that an individual who is a standard deviation above average on education is expected to be more than half a standard deviation above average on cogsc.	TRUE FALSE

Fit model 1 from the data page. Use it to answer the questions in part 2 below	Fill in or CIRCLE the correct answer.
2.1 Test the null hypothesis that H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ in model 1. Give the name of the test, the numerical value of the test statistics, the P-value, and indicate whether the null hypothesis is plausible.	Name:_____________ Value: ______ P-value: _______________ Circle one: PLAUSIBLE NOT PLAUSIBLE
2.2 Test the null hypothesis that H_0: $\beta_2=0$ in model 1, that is, the coefficient of BMI. Give the name of the test, the numerical value of the test statistics, the P-value, and indicate whether the null hypothesis is plausible.	Name:_____________ Value: ______ P-value: _______________ Circle one: PLAUSIBLE NOT PLAUSIBLE
2.3 If four variables in a regression model have individual P-values above 0.10, then that indicates it is reasonable to drop all four variables from the regression.	TRUE FALSE
2.4 Give the 95% confidence interval for β_3 (the coefficient of educ) in Model #1.	[________, _________]
2.5 Do the Shapiro-Wilk test on the residuals from model 1. What is the P-value? Is it plausible that the residuals are Normally distributed with constant variance?	P-value: _______________ Circle one: PLAUSIBLE NOT PLAUSIBLE

Name: _____________________________ ID# _________________________

PROBLEM SET #1 STATISTICS 500 FALL 2013: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon in class Tuesday 22 Oct.

Fit model 1 from the data page. Use it to answer question 3.	Fill in or CIRCLE the correct answer.
3.1 In model 1, test the null hypothesis H_0: $\beta_4=\beta_5=0$, that both alcohol and cigsperday have zero coefficients. What is the name of the test statistic? What is the numerical value of the test statistic? What are the degrees of freedom (DF)? What is the P-value? Is the null hypothesis plausible? (16 points)	Name:__________ Value: __________ DF = (____, ____) P-value: _________ Circle one: PLAUSIBLE NOT PLAUSIBLE
3.2 In a Normal linear model with five variables (like model 1), if H_0: $\beta_4=0$ is not rejected at the 0.05 level and H_0: $\beta_5=0$ is not rejected at the 0.05 level, then H_0: $\beta_4=\beta_5=0$ will not be rejected at the 0.05 level.	TRUE FALSE
3.3 Use the added variable plot (Sheather section 6.1.3) to consider adding caffeine to model #1. Although the slope in this plot is tilted up, there is one person who looks unusual in terms of caffeine consumption, a 69 year-old nonsmoking man, who consumed much more caffeine than most people.	TRUE FALSE

Fit model 2 and use it for part 4 below	Fill in or CIRCLE the correct answer.
4.1 Using a test of a general linear hypothesis (a partial F test), we can reject model 1 in favor of model 2 at the 0.05 level.	TRUE FALSE
4.2 If one boxplots the residuals from model 2 by the five levels of education, the median residual for the middle HS/Ged group is slightly positive, while the medians for <9 grade and college grad are slightly negative.	TRUE FALSE
4.3 The Normal quantile plot (and Shaprio-Wilk test) suggest that the residuals from model 2 are not Normal, with an upper tail that is too short for the Normal and a lower tail that is too long for the Normal – too many people have exceptionally low cognitive score residuals.	TRUE FALSE

Questions are 6 points each, except as noted.

PROBLEM SET #1 STATISTICS 500 FALL 2013: ANSWERS

This is an exam. Do not discuss it with anyone.

Question (Part 1) (6 points each)	CIRCLE the correct answer.
1.1 In the plot of cogsc (as y) against age (60-85 as x), the lowess curve increases suggesting higher scores at older ages.	TRUE [✓ FALSE] The plot tilts down, not up.
1.2 Looking at alcohol consumption by the levels of education, the sample upper (3rd) quartile of alcohol consumption is highest among individuals with a college degree.	[✓ TRUE] FALSE
1.3 The sample median cogsc increases over the five levels of education.	[✓ TRUE] FALSE
1.4 Using Pearson (i.e. usual) correlation, it is correct to say that an individual who is a standard deviation above average on education is expected to be more than half a standard deviation above average on cogsc.	[✓ TRUE] FALSE

Fit model 1 from the data page. Use it to answer the questions in part 2 below	Fill in or CIRCLE the correct answer.
2.1 Test the null hypothesis that H_0: $\beta_1= \beta_2= \beta_3= \beta_4= \beta_5=0$ in model 1. Give the name of the test, the numerical value of the test statistics, the P-value, and indicate whether the null hypothesis is plausible.	Name: F-test Value: 114.1 P-value: 2.2 x 10^{-16} Circle one: PLAUSIBLE [✓ NOT PLAUSIBLE]
2.2 Test the null hypothesis that H_0: $\beta_2=0$ in model 1, that is, the coefficient of BMI. Give the name of the test, the numerical value of the test statistics, the P-value, and indicate whether the null hypothesis is plausible.	Name: t-test Value: -0.806 P-value: 0.42 Circle one: [✓ PLAUSIBLE] NOT PLAUSIBLE
2.3 If four variables in a regression model have individual P-values above 0.10, then that indicates it is reasonable to drop all four variables from the regression.	TRUE [✓ FALSE] The t-test asks about one variable assuming you keep all the others.
2.4 Give the 95% confidence interval for b_3 (the coefficient of educ) in Model #1.	[7.10 , 8.47]
2.5 Do the Shapiro-Wilk test on the residuals from model 1. What is the P-value? Is it plausible that the residuals are Normally distributed with constant variance?	P-value: 3.2 x 10^{-6} Circle one: PLAUSIBLE [✓ NOT PLAUSIBLE]

PROBLEM SET #1 STATISTICS 500 FALL 201: ANSWERS

This is an exam. Do not discuss it with anyone.

Fit model 1 from the data page. Use it to answer question 3.	Fill in or CIRCLE the correct answer.
3.1 In model 1, test the null hypothesis H_0: $\beta_4 = \beta_5 = 0$, that both alcohol and cigsperday have zero coefficients. What is the name of the test statistic? What is the numerical value of the test statistic? What are the degrees of freedom (DF)? What is the P-value? Is the null hypothesis plausible? (16 points)	Name: F-test Value: 1.036 DF = (2, 1102) P-value: 0.355 Circle one: [✓ PLAUSIBLE] NOT PLAUSIBLE
3.2 In a Normal linear model with five variables (like model 1), if H_0: $\beta_4=0$ is not rejected at the 0.05 level and H_0: $\beta_5=0$ is not rejected at the 0.05 level, then H_0: $\beta_4 = \beta_5 = 0$ will not be rejected at the 0.05 level.	TRUE [✓ FALSE] This is the same issue as in question 2.3. Because predictors are often correlated, you may not need x_4 if you keep x_5, and you may not need x_5 if you keep x_4, but you may need to keep one of them.
3.3 Use the added variable plot (Sheather section 6.1.3) to consider adding caffeine to model #1. Although the slope in this plot is tilted up, there is one person who looks unusual in terms of caffeine consumption, a 69 year-old nonsmoking man, who consumed much more caffeine than most people.	[✓ TRUE] FALSE 3067/330 = 9.3 grande Starbucks coffees

Fit model 2 and use it for part 4 below	Fill in or CIRCLE the correct answer.
4.1 Using a test of a general linear hypothesis (a partial F test), we can reject model 1 in favor of model 2 at the 0.05 level.	TRUE [✓ FALSE] A general linear hypothesis compares nested models, one simpler than the other. These two models are not nested.
4.2 If one boxplots the residuals from model 2 by the five levels of education, the median residual for the middle HS/Ged group is slightly positive, while the medians for <9 grade and college grad are slightly negative.	[✓ TRUE] FALSE This is true and it suggests that the relationship between education and test score is somewhat curved, not a line.
4.3 The Normal quantile plot (and Shaprio-Wilk test) suggest that the residuals from model 2 are not Normal, with an upper tail that is too short for the Normal and a lower tail that is too long for the Normal – too many people have exceptionally low cognitive score residuals.	TRUE [✓ FALSE] The residuals are not Normal, but it is the upper tail that is thicker than Normal. Too many very high scores, not too many very low ones.

Questions are 6 points each, except as noted.

Doing the Problem Set in R

Problem 1, Fall 2013, Stat 500

```
Question 1.
1.1
> plot(age,cogsc)
> lines(lowess(age,cogsc))
(Goes down, not up.)
1.2
> boxplot(alcohol~educf)
1.3
> boxplot(cogsc~educf)
1.4
> cor(cogsc,educ)
[1] 0.543796
0.543796 > ½
Question 2.
> mod1<-lm(cogsc~age+bmi+educ+alcohol+cigsperday)
> summary(mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 61.24402    5.91401  10.356   <2e-16 ***
age         -0.55571    0.06468  -8.591   <2e-16 ***
bmi         -0.07426    0.09217  -0.806    0.421
educ         7.78226    0.34835  22.341   <2e-16 ***
alcohol     -0.02200    0.02946  -0.747    0.455
cigsperday  -0.09343    0.08265  -1.131    0.258
Residual standard error: 15.48 on 1102 degrees of freedom
Multiple R-squared:  0.341,     Adjusted R-squared:  0.338
F-statistic: 114.1 on 5 and 1102 DF,  p-value: < 2.2e-16
2.4
> confint(mod1)
                  2.5 %      97.5 %
(Intercept) 49.64001788 72.84801685
age         -0.68263302 -0.42879445
bmi         -0.25510106  0.10658373
educ         7.09876314  8.46575698
alcohol     -0.07980479  0.03579602
cigsperday  -0.25559409  0.06872682
2.5
> shapiro.test(mod1$residual)
        Shapiro-Wilk normality test
data:  mod1$residual
W = 0.9911, p-value = 3.225e-06
Question 3:
3.1 General linear hypothesis
```

```
> modred<-lm(cogsc~age+bmi+educ)
> anova(modred,mod1)
Analysis of Variance Table
Model 1: cogsc ~ age + bmi + educ
Model 2: cogsc ~ age + bmi + educ + alcohol + cigsperday
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1104 264593
2   1102 264096  2    496.78 1.0365  0.355
```

3.3 Added variable plot (use addedvarplot or do it yourself)

```
> addedvarplot(mod1,caffeine)
> modc<-lm(caffeine~age+bmi+educ+alcohol+cigsperday)
> plot(modc$resid,mod1$resid)
> which.max(modc$resid)
1002
> cogscore[1002,]
     cogsc age   bmi educ educf alcohol caffeine lead
1282    18  69 20.41    1  <9th       0   3066.8    3
     cadmium smoke100life smokenow cigsperday female
1282     1.2            1        3          0      0
     smokenowf vigorous moderate
1282        No        0        0
> summary(caffeine)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.82  102.70  167.90  251.60 3067.00
```

Question 4

4.1 The general linear hypothesis (or partial F-test) compares two models, one nested within the other. The reduced or simpler model must contain all the variables in the full or more complex model. Models 1 and 2 are not related in this way, so they cannot be compared by the partial F-test.

4.2

```
> boxplot(mod2$resid~educf)
> abline(h=0)
```

4.3

```
> par(mfrow=c(1,2))
> boxplot(mod2$resid)
> qqnorm(mod2$resid)
> qqline(mod2$resid)
> shapiro.test(mod2$resid)
        Shapiro-Wilk normality test
W = 0.9904, p-value = 1.198e-06
```

PROBLEM SET #2 STATISTICS 500 FALL 2013: DATA PAGE 1
Due in class at noon on Tuesday 26 November 2013 noon in class.
This is an exam. Do not discuss it with anyone.

The data are from the Wisconsin Longitudinal Study (http://www.ssc.wisc.edu/wlsresearch/) and they simplify a study by Springer et al. (2007) relating adult anger to abuse by parents as a child. The Study interviewed adults and asked about their childhoods. The data are in the object abuse500 in the R workspace for the course: http://www-stat.wharton.upenn.edu/~rosenbap/index.html. You will need to download it again. If you cannot find abuse500, it probably means you need to clear your browser's cache and download again. If you are not using R, then there is a csv file called abuse500.csv at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` The list is case-sensitivity, so `cogscore.csv` is at the end, with the lower case files. The file is simplified in several ways; in particular, missing data have been removed.

The anger score is based on structured questions due to Speilberger (1996). Higher scores mean more anger. Abuse is based on two questions: "During the first 16 years of your life, how much did your father/mother slap, shove or throw things at you?" The study by Springer et al. (2007) does not address some important practical questions, such as the accuracy of adult statements about childhood events, but these practical questions are not part of the current problem set.

id id number from the Wisconsin Longitudinal Study
anger Score on Speilberger (1996) anger scale. (nua34rec)
Fabuse 1 means Father was abusive – (nw036rer)
Mabuse 1 means Father was abusive – (nw037rer)
female 1 means respondent is female, 0 means male
age age of respondent (sa029re)
siblings respondent's number of siblings (sibstt)
Feducation years of education of respondent's father (edfa57q)
Meducation years of education of respondent's mother (edmo57q)

When asked to identify a person in the data set, please use the row number (or rownames(abuse500), 1, 2, …, 2841. Please do not use id.

Part 4 has three questions that ask you to expand model 2. When answering question 4.1, you add some things. When answering question 4.2 you add other things. Do not fit a model with both the things for 4.1 and the things for 4.2 – do these two questions separately.

PROBLEM SET #2 STATISTICS 500 FALL 2013: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due in class at noon on 26 November 2013 noon in class.

Before you do anything else, plot the data in various ways. For example:

```
> attach(abuse500)
> boxplot(anger~factor(Fabuse):factor(Mabuse))
> plot(age,anger)
> lines(lowess(age,anger),col=”red”)
```

Model #1

anger $= \beta_0 + \beta_1$ Fabuse + β_2 Mabuse + β_3 female + β_4 age + β_5 siblings
+ β_6 Feducation + + β_5 Meducation + ε where ε are iid $N(0,\sigma^2)$

Calculate the log-base-2 of anger+1. (You cannot take the log of 0.)

```
> L2anger<-log2(anger+1)
```

Model #2

L2anger $= \gamma_0 + \gamma_1$ Fabuse + γ_2 Mabuse + γ_3 female + γ_4 age + γ_5 siblings
+ γ_6 Feducation + γ_5 Meducation + η where η are iid $N(0,\omega^2)$

Model 1 has slopes β (beta), while model 2 has slopes γ (gamma), so that different things have different names. The choice of Greek letters is arbitrary.

Spielberger, C. D. (1996), State-Trait Anger Expression Inventory Professional Manual. Odessa, FL: Psychological Assessment Resources.

Springer, K. W., Sheridan, J., Kuo, D., and Carnes, M. (2007), "Long term phsyical and mental health consequences of childhood physical abuse," Child Abuse and Neglect, 31, 517-530.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name (last, first): ______________________________ ID# ___________________

PROBLEM SET #2 STATISTICS 500 FALL 2013: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon in class 26 November.

Part 1: Use model 1 for part 1.	Fill in or CIRCLE the correct answer
Looking at the Normal quantile plot, the residuals from model 1 do not look like observations from a Normal distribution: they are asymmetric with a long right tail.	1.1 TRUE FALSE
The largest absolute studentized residual from model 1 is about 7.405 for an individual with anger score of 70.	1.2 TRUE FALSE
Test the null hypothesis that there are no outliers in model 1. Do a 2-sided test for each residual adjusting using the Bonferroni inequality to obtain a familywise error rate of 0.05. In the t-distribution, what are the degrees of freedom? What is the critical value, that is, the absolute t-value just significant as an outlier? This critical value is a property of the t-distribution, and depends upon the data only through the sample size and the number of predictors in the model. By this standard, is their at least one outlier?	1.3 Degrees of freedom:________ Critical t-value: ___________ CIRCLE ONE YES OUTLIER(S) NO OUTLIERS
Using the method in question 1.3, how many individuals would be judged outliers? Of these outliers, if any, how many are above/below the fitted regression?	1.4 #outliers: ______________ #outliers above: ______ #below:______
In light of what you have discovered in questions 1.1-1.4, it is reasonable to think of the assumption of Normal errors in in model 1 as almost true apart from perhaps a few (<=3) outliers that should be removed before working further with model 1.	1.5 TRUE FALSE

Part 2: Use models 1 and 2 from the data page for the questions in part 2.	Fill in or CIRCLE the correct answer **Identify individuals by row of abuse500**
Repeat the method in questions 1.3 and 1.4 for model 2, stating the number of outliers detected, above and below the regression plane.	2.1 #outliers: ______________ #outliers above: ______ #below:______
Which individual has the largest leverage or hat-value for model 1? For model 2?	2.2 Model 1:_______ Model 2:________
The person identified in 2.2 for model 2 has large leverage because he is angry and his father had an unusual Feducation	2.3 TRUE FALSE

Name (last, first): _____________________________ ID# _____________________

PROBLEM SET #2 STATISTICS 500 FALL 2013: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon in class 26 November.

Part 3: Use model 2 from the data page for the questions in part 3.	Fill in or CIRCLE the correct answer **Identify individuals by row of abuse500**
In model 2, which observation has the largest absolute dffits? In model 2, which observation has the largest Cook's distance?	3.1 Which for |dffits|: _______________ Which for Cook's distance:___________
The individual identified in 3.1 with the largest absolute dffits is below the regression plane pulling it down by half of the standard deviation of his/her fitted Y.	3.2 TRUE FALSE
Looking at dfbetas, the individual identified in 3.1 with the largest absolute dffits is having his biggest impact on the estimate coefficient of "siblings" pulling that coefficient upwards.	3.3 TRUE FALSE

Part 4: Part 4 asks you to start with model 2 and add things to it. So Y should be L2anger, not anger.	Fill in or CIRCLE the correct answer Read about part 4 on the data page!
One might imagine that abuse by both father and mother might be associated with much greater anger than the sum of the two parts. In model 2, test the null hypothesis H_0 that there is no interaction between Fabuse and Mabuse. Give the t-statistic and P-value testing H_0, and say whether H_0 is plausible. How many people had both Fabuse=1 and Mabuse=1? (8 points)	4.1 t-statistic:__________ P-value:__________ Circle one: PLAUSIBLE NOT PLAUSIBLE How many:_________
In model 2, test the null hypothesis H_0 that the relationship between anger and age is linear against the alternative hypothesis that it is quadratic. Give the t-statistic and P-value testing H_0, and say whether H_0 is plausible. (8 points)	4.2 t-statistic:__________ P-value:__________ Circle one: PLAUSIBLE NOT PLAUSIBLE
Use Tukey's one degree of freedom test, tukey1df, to test the null hypothesis H_0 that the transformation to logs in model 2 is ok and further transformation is not needed. Give the t-statistic and P-value testing H_0, and say whether H_0 is plausible.	4.3 t-statistic:__________ P-value:__________ Circle one: PLAUSIBLE NOT PLAUSIBLE

Questions are 7 points each except as noted (4.1 and 4.2)

PROBLEM SET #2 STATISTICS 500 FALL 2013: ANSWER PAGE 1

Part 1: Use model 1 for part 1.	Fill in or CIRCLE the correct answer
Looking at the Normal quantile plot, the residuals from model 1 do not look like observations from a Normal distribution: they are asymmetric with a long right tail.	1.1 [✓ TRUE] FALSE
The largest absolute studentized residual from model 1 is about 7.405 for an individual with anger score of 70.	1.2 [✓ TRUE] FALSE
Test the null hypothesis that there are no outliers in model 1. Do a 2-sided test for each residual adjusting using the Bonferroni inequality to obtain a familywise error rate of 0.05. In the t-distribution, what are the degrees of freedom? What is the critical value, that is, the absolute t-value just significant as an outlier? This critical value is a property of the t-distribution, and depends upon the data only through the sample size and the number of predictors in the model. By this standard, is their at least one outlier?	1.3 Degrees of freedom: 2832 = 2833-1 Critical t-value: 4.30 CIRCLE ONE [✓ YES OUTLIER(S)] NO OUTLIERS
Using the method in question 1.3, how many individuals would be judged outliers? Of these outliers, if any, how many are above/below the fitted regression?	1.4 #outliers: 19 #outliers above: 19 #below: 0
In light of what you have discovered in questions 1.1-1.4, it is reasonable to think of the assumption of Normal errors in in model 1 as almost true apart from perhaps a few (<=3) outliers that should be removed before working further with model 1.	1.5 TRUE [✓ FALSE]

Part 2: Use models 1 and 2 from the data page for the questions in part 2.	Fill in or CIRCLE the correct answer **Identify individuals by row of abuse500**
Repeat the method in questions 1.3 and 1.4 for model 2, stating the number of outliers detected, above and below the regression plane.	2.1 #outliers: 0 #outliers above: 0 #below: 0
Which individual has the largest **leverage or hat-value** for model 1? For model 2?	2.2 Same person because x did not change Model 1: 824 Model 2: 824
The person identified in 2.2 for model 2 has large leverage because he is angry and his father had an unusual Feducation	2.3 Leverage depends on x, not y. TRUE [✓ FALSE]

PROBLEM SET #2 STATISTICS 500 FALL 2013: ANSWER PAGE 2

Part 3: Use model 2 from the data page for the questions in part 3.	Fill in or CIRCLE the correct answer **Identify individuals by row of abuse500**
In model 2, which observation has the largest absolute dffits? In model 2, which observation has the largest Cook's distance?	3.1 Which for |dffits|: 824 Which for Cook's distance: 824 Two very similar measures.
The individual identified in 3.1 with the largest absolute dffits is below the regression plane pulling it down by half of the standard deviation of his/her fitted Y.	3.2 TRUE [✓ FALSE] He is about a quarter standard deviation above, not half standard deviation below.
Looking at dfbetas, the individual identified in 3.1 with the largest absolute dffits is having his biggest impact on the estimated coefficient of "siblings" pulling that coefficient upwards.	3.3 A fairly angry guy with lots of siblings [✓ TRUE] FALSE

Part 4: Part 4 asks you to start with model 2 and add things to it. So Y should be L2anger, not anger.	Fill in or CIRCLE the correct answer Read about part 4 on the data page!
One might imagine that abuse by both father and mother might be associated with much greater anger than the sum of the two parts. In model 2, test the null hypothesis H_0 that there is no interaction between Fabuse and Mabuse. Give the t-statistic and P-value testing H_0, and say whether H_0 is plausible. How many people had both Fabuse=1 and Mabuse=1? (8 points)	4.1 t-statistic: -0.052 P-value: 0.96 Circle one: [✓ PLAUSIBLE] NOT PLAUSIBLE How many: 93
In model 2, test the null hypothesis H_0 that the relationship between anger and age is linear against the alternative hypothesis that it is quadratic. Give the t-statistic and P-value testing H_0, and say whether H_0 is plausible. (8 points)	4.2 t-statistic: 1.565 P-value: 0.1177 Circle one: [✓ PLAUSIBLE] NOT PLAUSIBLE
Use Tukey's one degree of freedom test, tukey1df, to test the null hypothesis H_0 that the transformation to logs in model 2 is ok and further transformation is not needed. Give the t-statistic and P-value testing H_0, and say whether H_0 is plausible.	4.3 t-statistic: 1.219 P-value: 0.223 Circle one: [✓ PLAUSIBLE] NOT PLAUSIBLE

Doing the Problem Set in R: Problem 2, Fall 2013, Stat 500

```
Part 1.
> mod<-lm(anger ~ Fabuse + Mabuse + female + age + siblings +
     Feducation + Meducation)
1.1
> qqnorm(mod$resid)
> hist(mod$resid)
> boxplot(mod$resid)
1.2
> which.max(abs(rstudent(mod)))
2839
2839
> rstudent(mod)[2839]
    2839
7.404952
> abuse500[2839,]
         id anger Fabuse Mabuse female age siblings Feducation Meducation
2839 933941    70      0      0      0  48        6         14         18
1.3
> dim(abuse500)
[1] 2841    9
> mod$df
[1] 2833
> qt(.025/2841,2832)
[1] -4.30073
1.4
> sum(rstudent(mod)>=4.30073)
[1] 19
> sum(rstudent(mod)<=-4.30073)
[1] 0
2.1
> l2anger<-log2(anger+1)
> modl<-lm(l2anger~Fabuse+Mabuse+female+age+siblings+Feducation+Meducation)
> qqnorm(rstudent(modl))
> sum(rstudent(modl)>=4.30073)
[1] 0
> sum(rstudent(modl)<=-4.30073)
[1] 0
2.2
> which.max(hatvalues(modl))
824
824
> which.max(hatvalues(mod))
824
824
Of course, they are the same, because taking logs changed y, not x, and
leverages or hatvalues are determined by the x’s.
> abuse500[824,]
      X     id anger Fabuse Mabuse female age siblings Feducation Meducation
824 824 909671    14      0      0      0  62       26         12          7
Twenty-six siblings!  Leverage depends on x, not y.
```

Doing the Problem Set in R, continued: Problem 2, Fall 2013, Stat 500

3.1

```
> which.max(abs(dffits(modl)))
824
824
```

We’ve seen him before.

```
> which.max(cooks.distance(modl))
824
824
```

dffits and Cook’s distance are very similar!

3.2

```
> max(dffits(modl))
[1] 0.249942
> min(dffits(modl))
[1] -0.2265019
> round(dfbetas(modl)[824,],2)
Fabuse      Mabus   female     age  siblings  Feducation  Meducation
 -0.02    0.00   -0.02      0.03        0.24        0.06       -0.01
```

4.1

```
> table(Fabuse,Mabuse)
      Mabuse
Fabuse    0    1
     0 2490  111
     1  147   93
> int<-Fabuse*Mabuse
> sum(int)
[1] 93
> modli<-
lm(l2anger~Fabuse+Mabuse+female+age+siblings+Feducation+Meducation+int)
> summary(modli)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.334622   0.244131  21.851   <2e-16 ***
Fabuse       0.250927   0.119624   2.098   0.0360 *
…
int         -0.011979   0.231484  -0.052   0.9587
```

4.2 Remember to center before squaring.

```
> age2c<-(age-mean(age))^2
> modlq<-
lm(l2anger~Fabuse+Mabuse+female+age+siblings+Feducation+Meducation+age2c)
> summary(modlq)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.3206203  0.2441803  21.790   <2e-16 ***
Fabuse       0.2579148  0.1028093   2.509   0.0122 *
...
Meducation  -0.0045446  0.0103519  -0.439   0.6607
age2c        0.0005629  0.0003597   1.565   0.1177
```

4.3

```
> t1df<-tukey1df(modl)
> modlt<-
lm(l2anger~Fabuse+Mabuse+female+age+siblings+Feducation+Meducation+t1df)
> summary(modlt)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.328745   0.244104  21.830  < 2e-16 ***
Fabuse       0.247670   0.102621   2.413  0.01587 *
...
Meducation  -0.004600   0.010359  -0.444  0.65704
t1df         0.586381   0.481214   1.219  0.22312
```

PROBLEM SET #3 STATISTICS 500 FALL 2013: DATA PAGE 1

Due at noon on Wednesday, 18 December 2013, my office 473 JMHH.

This is an exam. Do not discuss it with anyone.

The cogscore data are the same as in the first problem set, but the IBS data will need to be downloaded. The cogscore data from NHANES describe people over age 60 who took a cognitive test; see below. The data are in a data.frame in the course workspace called “cogscore”. A csv file, `cogscore.csv`, is available for those not using R:
http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/
Do not use the csv file if you are using R.

cogsc cognitive test score (age>=60) (CFDRIGHT)
age in years
bmi body mass index
educ is 1-5 and is described in **educf**. (DMDEDUC2)
alcohol Alcohol consumed in grams (DRXTALCO)
caffeine Caffeine consumed in miligrams (DRXTCAFF)
lead is lead level in the blood (LBXBPB - Blood lead ug/dL)
cadmium is the cadmium level in the blood (LBXBCD - Blood cadmium ug/L)
smoke100life =1 if smoked 100 cigarettes in life, 0 otherwise (SMQ020)
smokenow Smoke now, explained in smokenowf (SMQ040)
cigsperday Cigarettes per day for people who smoke everyday (SMD070)
female = 1 for female, 0 for male
vigorous =1 if any vigorous physical activity in past 30 days, 0 otherwise (PAD200)
moderate =1 if any moderate physical activity in past 30 days, 0 otherwise (PAD320)
The cogscore is a version of “WAIS III (Wechsler Adult Intelligence Scale, Third Edition) Digit Symbol – Coding module conducted during the household interview.”

You will be doing a cross-validated variable selection. Typically, this would be a random half of the data, but you will NOT use a random half. So that everyone uses the same half, you will build the model using the first 554 rows of data (x1, y2) and validate using the remaining 554 rows (x2, y2). Do this as follows:

```
> dim(cogscore)
[1] 1108   16
> 1108/2
[1] 554
> x<-cogscore[,c(2,3,4,6,12,13,15,16)]
> head(x)
  age   bmi educ alcohol cigsperday female vigorous moderate
3  62 36.94    3     0.0          0      0        0        0
> x1<-x[1:554,]
> y1<- cogscore$cogsc[1:554]
> x2<-x[555:1108,]
> y2<- cogscore$cogsc[555:1108]
```

Important note for parts 1&2: For part 1 questions, use leaps (best subsets regression) to predict y1 from x1. For part 2 questions, use lm (regression) to relate y2 to x2 using the variables selected in part 1. Note that the predictors are: `age bmi educ alcohol cigsperday female vigorous moderate`

PROBLEM SET #3 STATISTICS 500 FALL 2013: DATA PAGE 2

Due at noon on Wednesday, 18 December 2013, my office 473 JMHH.

The second data set is the object IBS in the course workspace or IBS.csv at http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/
You will need to download the course workspace again. There is a csv file IBS.csv for those not using R; see web page on data page 1. Data are from a clinical trial of a drug intended to reduce the abdominal pain associated with irritable bowel syndrome (IBS). The data are from a paper by E. Biesheuvel andand L. Hothorn (2002) in the *Biometrical Journal* 44, 101-116, but there is no need to consult the paper unless you want to. I removed a few people to make a balanced design, so use IBS.

There are 70 people in each of five dose groups, one of which is a placebo (zero dose). The doses were blinded. The dose groups are 0, 1, 2, 3, 4. The outcome is iapain, a measure of improvement in abdominal pain from baseline, with larger numbers indicating greater improvement (good) and negative numbers indicating that a person got worse, not better, under the drug or placebo (bad). The dose is given as a number and as a factor (dose, dosef). There is an indicator of gender (1=female, 0=male). R handles numbers and factors differently, so you should think about when to use dose vs dosef.

```
  dose   dosef female     iapain
1    0 Placebo      1 1.26153846
> dim(IBS)
[1] 350   4
```

There are five groups, i=0,1,2,3,4 and seventy people in group i, j=1,2,…,70. The model for the IBS data is y_{ij}=iapain score for the jth person in group I, with
$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where the errors ε_{ij} are independent, identically distributed (iid) with Normal distributions having mean 0 and variance σ^2. **IMPORTANT**: When referring to pairs of groups in **questions 3.2 and 3.5**, refer to them by number, so 0-1 is placebo vs verylow dose. Remember, the groups are numbered 0, 1, 2, 3, 4, NOT 1, 2, 3, 4, 5. Please do not get the questions wrong by numbering the groups 1 to 5. **Question 3.7** asks you to make 4 **orthogonal** contrasts, each contrast comparing one dose group to the average of all higher dose groups, e.g., low versus the average of medium and high, asking whether low dose is not as effective as a higher dose. Use **integer contrast weights** as in (4, -6, 2, 1, -1), but use appropriate weights with orthogonal contrasts. **Follow instructions**. **Write your name**, last name first, on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. **Do not turn in additional pages**. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer**. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam. The exam is due **at noon Wednesday December 18, 2012, at my office, 473 JMHH**. You may turn in the exam early by placing it in an envelope addressed to me and leaving it in my mail box in statistics, 4th floor, JMHH. If you prefer, give it to Noel at the front desk in statistics. Make and keep a photocopy of your answer page. The answer key will be posted in the revised bulk pack on-line.

Name (Last, First): _________________________ ID# __________________________

PROBLEM SET #3 STATISTICS 500 FALL 2013: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon Wed., 18 December.

(Part 1). Read the important note on the data page. **Use y1 and x1 to answer part 1**. Define them as on the data page.	All questions in part 1 are based on y1 and x1 CIRCLE the correct answer.
1.1 Which model has the smallest C_P value using y1 and x1? List all of the predictor (x) variables in this model.	
1.2 What is the numerical value of C_P in the model you selected in 1.1? What is the "size" of this model, that is, the value we plot C_P against? What is R^2 for this model?	C_P = _________ Size = __________ R^2 = _________
1.3 When C_P is greater than the size of the model (see 1.2), then this fact is a sign that the model contains unneeded variables.	TRUE FALSE
1.4 A small C_P is a sign that the model omits variables needed for prediction.	TRUE FALSE
1.5 Among models fitted by leaps that have C_P less than or equal to the size of the model, the variable vigorous is in all of these models but the variable educ is only some of these models.	TRUE FALSE
1.6 Using all 8 predictors in x1 to fit y1 in a regression using lm, which one of the 8 predictors has the largest VIF = variance inflation factor? What is the numerical value of VIF? What is the R^2 of this predictor with the other 7 predictors?	Variable name: ___________ VIF: __________ R^2 :__________

(Part 2). Read the important note on the data page. **Use y2 and x2 to answer part 2**. Define them as on the data page. Do not use leaps for part 2; use only lm.	All questions in part 2 are based on y2 and x2 Fill in or CIRCLE the correct answer.
2.1 Fit a regression predicting cogsc using just the variables identified in question 1.1 and using only the data in y2 and x2. What is the F-statistic for this regression? What are its df = degrees of freedom? What is the P-value?	F-statistic: _______ df= (_____, _____) P-value: ___________
2.2 The multiple squared correlation R^2 for the regression in 2.1 is lower than for the regression using these same variables in y1 and x1 (in 1.2).	TRUE FALSE
2.3 The model in 2.1 has a lower value of PRESS = predicted residual sum of squares than the model with 8 predictors using y2, x2.	TRUE FALSE

Name (Last, First): _________________________ ID# __________________________

PROBLEM SET #3 STATISTICS 500 FALL 2013: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon Wed., 18 December.

<table>
<tr><td>(Part 3). Part 3 uses the IBS data. As noted on the data page, dose and dosef (numeric vs factor) behave differently in R.</td><td>Answer part 3 assuming the model for the IBS data given on the data page.
Fill in or CIRCLE the correct answer.</td></tr>
<tr><td>3.1 Test the null hypothesis H_0 that the improvement in abdominal pain iapain is the same in the five groups. What is the name of the test statistic? What is its value? What is the P-value? Is the null hypothesis plausible?</td><td>Name:______ Value:____ P-value:____
CIRCLE ONE
PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>3.2 Use Tukey's method to compare all pairs of two of the five groups. Controlling the family-wise error rate at 5%, list all pairs of groups that differ significantly. If none, write NONE. See important note on the data page.</td><td>Use groups numbers, eg 0-1, as in the important note on the data page.</td></tr>
<tr><td>3.3 Weak control of the family wise-error rate at 5% means that if H_0 in 3.1 is true, the chance of finding at least one significant difference is at most 5%. (5 points)</td><td>TRUE FALSE</td></tr>
<tr><td>3.4 Strong control of the family wise-error rate at 5% means that even if every one of the null hypotheses tested is false, the chance of finding at least one significant difference is at most 5%. (5 points)</td><td>TRUE FALSE</td></tr>
<tr><td>3.5 Repeat question 3.2, but use Holm's method instead (using the pooled estimate of the standard deviation, the default in R).</td><td></td></tr>
<tr><td>3.6 Holm's method controls the family-wise error rate in the strong sense, but Tukey's method does not.</td><td>TRUE FALSE</td></tr>
<tr><td>3.7 Make 4 <b>orthogonal integer</b> contrast weights as described on the <b>data page</b> to compare individual doses to the average of higher doses. Each of the four rows is one contrast. The columns are the five groups.</td><td><table><tr><td>Placebo</td><td>Verylow</td><td>Low</td><td>Med</td><td>High</td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr></table></td></tr>
<tr><td>3.8 Use the contrasts in 3.7 to fill</td><td>in the anova table below.</td></tr>
</table>

Source	**Sum of squares**	**Degrees of freedom**	**Mean Square**	**F**	**p-value**
Between groups					
Placebo vs higher dose					
Verylow vs higher dose					
Low vs higher dose					
Medium vs High					
Within groups (error)				**xxxx**	**xxxxxxxx**

Problem Set 3, Fall 2013, Statistics 500

DOING THE PROBLEM SET IN R

Part 1.

```
> dim(cogscore)
[1] 1108   16
> 1108/2
[1] 554
> x<-cogscore[,c(2,3,4,6,12,13,15,16)]
> head(x)
  age   bmi educ alcohol cigsperday female vigorous moderate
3  62 36.94    3     0.0          0      0        0        0
> x1<-x[1:554,]
> y1<- cogscore$cogsc[1:554]
> x2<-x[555:1108,]
> y2<- cogscore$cogsc[555:1108]
> library(leaps)
> mod<-leaps(x=x1,y=y1,names=colnames(x1))
> which.min(mod$Cp)
[1] 29
> mod$which[29,]
       age   bmi  educ  alcohol cigsperday  female vigorous   moderate
      TRU  FALSE  TRUE  FALSE   FALSE       TRUE       TRUE      FALSE
> mod$Cp[29]
[1] 2.503731
> mod$which[mod$Cp<=mod$size,]
   age   bmi educ alcohol cigsperday female vigorous moderate
3 TRUE FALSE TRUE   FALSE      FALSE   TRUE    FALSE    FALSE
4 TRUE FALSE TRUE   FALSE      FALSE   TRUE     TRUE    FALSE
...
> attach(x1)
> library(DAAG)
> summary(lm(y1~age+educ+female+vigorous))
... Multiple R-squared:  0.3534
> vif(lm(y1~age+bmi+educ+alcohol+cigsperday+female+vigorous+moderate))
age    bmi     educ    alcohol cigsperday     female   vigorous
moderate
1.1250 1.0926 1.1321     1.0509     1.0786     1.0598     1.0948
1.1594
> 1-1/1.1595
[1] 0.1375593
> summary(lm(educ~age+bmi+alcohol+cigsperday+female+vigorous+moderate))
...
Multiple R-squared:  0.1167
```

Part 2

```
> detach(x1)
> attach(x2)
> summary(lm(y2~age+educ+female+vigorous))
… Multiple R-squared:  0.3609
F-statistic: 77.52 on 4 and 549 DF,  p-value: < 2.2e-16
>
press(lm(y2~age+bmi+educ+alcohol+cigsperday+female+vigorous+moderate))
[1] 126586.5
> press(lm(y2~age+educ+female+vigorous))
[1] 126547.9
```

Part 3. IBS data. 3.1

```
> summary(aov(iapain~dosef))
             Df Sum Sq Mean Sq F value Pr(>F)
dosef        4    7.28  1.8193   3.179 0.0138 *
Residuals   345 197.42  0.5722
> TukeyHSD(mod)
  Tukey multiple comparisons of mean 95% family-wise confidence level
                       diff         lwr       upr     p adj
VeryLow-Placebo  0.24157129 -0.10906624 0.5922088 0.3250271
Low-Placebo      0.33210369 -0.01853383 0.6827412 0.0730586
Medium-Placebo   0.40359625  0.05295872 0.7542338 0.0148874
High-Placebo     0.36255280  0.01191528 0.7131903 0.0386954
Low-VeryLow      0.09053240 -0.26010512 0.4411699 0.9545933
…
High-Medium     -0.04104344 -0.39168096 0.3095941 0.9976952
```

3.5

```
> pairwise.t.test(iapain,dosef)
        Pairwise comparisons using t tests with pooled SD
        Placebo VeryLow Low   Medium
VeryLow 0.418   -       -     -
Low     0.078   1.000   -     -
Medium  0.017   1.000   1.000 -
High    0.044   1.000   1.000 1.000
P value adjustment method: holm
```

3.7

```
> dosef2<-dosef
> contrasts(dosef2)<-cbind(placebo=c(-4,1,1,1,1),
   verylow=c(0,-3,1,1,1),low=c(0,0,-2,1,1),medium=c(0,0,0,-1,1))
> summary(aov(iapain~dosef2))
             Df Sum Sq Mean Sq F value Pr(>F)
dosef2        4   7.28  1.8193   3.179 0.0138 *
Residuals   345 197.42  0.5722
```

3.8

```
> mm<-model.matrix(aov(iapain~dosef2))
> mm<-as.data.frame(mm)
> attach(mm)
> mod2<-lm(iapain~dosef2placebo+dosef2verylow+dosef2low+dosef2medium)
> anova(mod2)
Analysis of Variance Table
               Df  Sum Sq Mean Sq F value   Pr(>F)
dosef2placebo   1   6.283  6.2829 10.9798 0.001019 **
dosef2verylow   1   0.814  0.8139  1.4224 0.233829
dosef2low       1   0.121  0.1212  0.2119 0.645591
dosef2medium    1   0.059  0.0590  0.1030 0.748411
Residuals     345 197.418  0.5722
```

PROBLEM SET #3 STATISTICS 500 FALL 2013: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon Wed., 18 December.

(Part 1). Read note on the data page. **Use y1 and x1 to answer part 1**. Defined on the data page.	All questions in part 1 are based on y1 and x1 CIRCLE the correct answer.
1.1 Which model has the smallest C_P value using y1 and x1? List all of the predictor (x) variables in this model.	age, educ, female, vigorous
1.2 What is the numerical value of C_P in the model you selected in 1.1? What is the "size" of this model, that is, the value we plot C_P against? What is R^2 for this model?	C_P = 2.50 Size = 5 R^2 = 0.3534
1.3 When C_P is greater than the size of the model (see 1.2), then this fact is a sign that the model contains unneeded variables.	TRUE [✓ FALSE]
1.4 A small C_P is a sign that the model omits variables needed for prediction.	TRUE [✓ FALSE]
1.5 Among models fitted by leaps that have C_P less than or equal to the size of the model, the variable vigorous is in all of these models but the variable educ is only some of these models.	TRUE [✓ FALSE]
1.6 Using all 8 predictors in x1 to fit y1 in a regression using lm, which one of the 8 predictors has the largest VIF = variance inflation factor? What is the numerical value of VIF? What is the R^2 of this predictor with the other 7 predictors?	Variable name: moderate VIF: 1.159 R^2 : 0.138 These predictors are not highly correlated.

(Part 2). Read note on the data page. **Use y2 and x2 to answer part 2**. Define them as on the data page. Do not use leaps for part 2; use only lm.	All questions in part 2 are based on y2 and x2 Fill in or CIRCLE the correct answer.
2.1 Fit a regression predicting cogsc using just the variables identified in question 1.1 and using only the data in y2 and x2. What is the F-statistic for this regression? What are its df = degrees of freedom? What is the P-value?	F-statistic: 77.52 df= (4, 549) P-value: 2.2×10^{-16}
2.2 The multiple squared correlation R^2 for the regression in 2.1 is lower than for the regression using these same variables in y1 and x1 (in 1.2).	TRUE [✓ FALSE]
2.3 The model in 2.1 has a lower value of PRESS = predicted residual sum of squares than the model with 8 predictors using y2 and x2.	[✓ TRUE] FALSE

Name (Last, First): ______________________ ID# ______________________

PROBLEM SET #3 STATISTICS 500 FALL 2013: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon Wed., 18 December.

(Part 3). Part 3 uses the IBS data. As noted on the data page, dose and dosef (numeric vs factor) behave differently in R.	Answer part 3 assuming the model for the IBS data given on the data page. Fill in or CIRCLE the correct answer.

<table>
<tr><td>3.1 Test the null hypothesis H_0 that the improvement in abdominal pain iapain is the same in the four groups. What is the name of the test statistic? What is its value? What is the P-value? Is the null hypothesis plausible?</td><td>Name: F-test Value: 3.179 P-value: 0.0138
CIRCLE ONE
PLAUSIBLE [✓ NOT PLAUSIBLE]</td></tr>
<tr><td>3.2 Use Tukey's method to compare all pairs of two of the five groups. Controlling the family-wise error rate at 5%, list all pairs of groups that differ significantly. If none, write NONE. See important note on the data page.</td><td>Use groups numbers, eg 0-1, as in the important note on the data page.
0-3 0-4</td></tr>
<tr><td>3.3 Weak control of the family wise-error rate at 5% means that if H_0 in 3.1 is true, the chance of finding at least one significant difference is at most 5%.</td><td>[✓ TRUE] FALSE</td></tr>
<tr><td>3.4 Strong control of the family wise-error rate at 5% means that even if every one of the null hypotheses tested is false, the chance of finding at least one significant difference is at most 5%.</td><td>If a null hypothesis is false, you want a high probability of rejecting it, not 5%
TRUE [✓ FALSE]</td></tr>
<tr><td>3.5 Repeat question 3.2, but use Holm's method instead (using the pooled estimate of the standard deviation, the default in R).</td><td>0-3 0-4</td></tr>
<tr><td>3.6 Holm's method controls the family-wise error rate in the strong sense, but Tukey's method does not.</td><td>Both control it in the strong sense.
TRUE [✓ FALSE]</td></tr>
<tr><td>3.7 Make 4 <b>orthogonal integer</b> contrast weights as described on the <b>data page</b> to compare individual doses to the average of higher doses. Each of the four rows is one contrast. The columns are the five groups.</td><td><table>
<tr><th>Placebo</th><th>Verylow</th><th>Low</th><th>Med</th><th>High</th></tr>
<tr><td>-4</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>0</td><td>-3</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>0</td><td>0</td><td>-2</td><td>1</td><td>1</td></tr>
<tr><td>0</td><td>0</td><td>0</td><td>-1</td><td>1</td></tr>
</table></td></tr>
<tr><td>3.8 Use the contrasts in 3.7 to fill in the anova</td><td>table below.</td></tr>
</table>

Source	**Sum of squares**	**Degrees of freedom**	**Mean Square**	**F**	**p-value**
Between groups	7.28	4	1.8193	3.18	0.0138
Placebo vs higher dose	6.283	1	6.283	10.98	0.001019
Verylow vs higher dose	0.814	1	0.814	1.42	0.23
Low vs higher dose	0.121	1	0.121	0.21	0.65
Medium vs High	0.059	1	0.059	0.10	0.75
Within groups (error)	197.418	345	0.5722	xxxx	xxxxxxxx

PROBLEM SET #1 STATISTICS 500 FALL 2014: DATA PAGE 1

Due in class at noon on Tuesday, October 21, 2014.

This is an exam. Do not discuss it with anyone.

The data are from NHANES, the 2009-2010 National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm). There is no need to visit the webpage unless you want to. The file adultcal describes calories consumed on the first interview day for individuals 20 years old or older. Calories are deduced from a food interview.

The file is simplified in several ways; in particular, missing data have been removed.

SEQN nhanes sequence number or id
age – age in year 0-19
female – 1 if female, 0 if male
ed and edf record education. Type table(adultcal$edf) for categories.
income – ratio of family income to the poverty level, capped at 5 times.
married – 1 if married or living with partner, 0 otherwise
bmi – body mass index
waist – waist circumference in cm 1 cm = 0.393701 inches
calories – calories consumed on first interview day

```
> head(adultcal)
    SEQN age female ed                       edf income married   bmi waist calories
1  51624  34      0  3      High School Grad/GED   1.36       1 32.22 100.4     1844
5  51628  60      1  3      High School Grad/GED   0.69       0 42.39 118.2     1913
6  51629  26      0  2                9-11 grade   1.01       1 32.61 103.7     3123
7  51630  49      1  4 Some college or AA degree   1.91       1 30.57 107.8     1345
10 51633  80      0  4 Some college or AA degree   1.27       1 26.04  91.1     1565
12 51635  80      0  2                9-11 grade   1.69       0 27.62 113.7     1479
```

```
> dim(adultcal)
[1] 5000   10
```

The data are in the object `adultcal` in the course workspace at http://www-stat.wharton.upenn.edu/~rosenbap/ . You will have to download the workspace again to have the current version with `adultcal`. If you download the workspace and `adultcal` is not there, it probably means that you web browser remembers the last time you downloaded the file and thinks (incorrectly) that you do not need to download it again – in this case, clear the browser's memory and try again. There is a csv file `adultcal.csv` with the data at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` if you wish to use software other than R. The csv file should open in excel and other packages.

PROBLEM SET #1 STATISTICS 500 FALL 2014: DATA PAGE 2

This is an exam. Do not discuss it with anyone. Due Tuesday, October 21, 2014

Before you do anything else, plot the data in various ways. For example:

```
> attach(adultcal)
> boxplot(calories)
> plot(age,calories)
> lines(lowess(age,calories),col="red",lwd=2)
> boxplot(calories~female)
> boxplot(calories~female:married)
 etc
```

DO NOT TURN IN THE PLOTS.

Model #1

calories $= \beta_0 + \beta_1$ age + β_2 female + β_3 ed + β_4 income + ε where ε are iid $N(0,\sigma^2)$

Model #2

calories $= \gamma_0 + \gamma_1$ age + γ_2 female + γ_3 income + η where η are iid $N(0,\omega^2)$

Model #3

calories $= \lambda_0 + \lambda_1$ age + λ_2 female + ζ where ζ are iid $N(0,\kappa^2)$

Model 1 has slopes β (beta), while model 2 has slopes γ (gamma), so that different things have different names. The choice of Greek letters is arbitrary. The same is true for model 3.

It is often useful to put two plots next to each other on the same page so you can see the same point in both plots. If you type

```
> par(mfrow=c(1,2))
```

then the next two plots will appear on the same page, the first on the left, the second on the right. For example, you can compare a boxplot and a Normal quantile plot in this way. The command sets a graphics parameter (that's the 'par'), and it says that there should be 1 row of graphs with 2 columns, filling in the first row first. By setting graph parameters, you can control many aspects of a graph. The free document R for Beginners by Paradis (http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf) contains lots of useful information about graph parameters (see page 43).

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name: _____________________________ ID# _________________________

PROBLEM SET #1 STATISTICS 500 FALL 2014: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due **Tuesday, October 21 noon**

Question (Part 1)	CIRCLE the correct answer
1.1 There is one person who consumed more than 10,000 calories.	TRUE FALSE
1.2 The lower quartile of calories for males is above the median for females.	TRUE FALSE
1.3 The median waist size for females is more than 37 inches = 93.98 cm.	TRUE FALSE
1.4 Of the five education categories, the lowest median of calories is for the <9th grade category.	TRUE FALSE

Fit model 1 from the data page.	Fill in or CIRCLE the correct answer.
2.1 Test the null hypothesis that the coefficient of income in model 1 is zero, $H_0:\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the two-sided P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name:________ Value:__________ P-value: ____________________ Circle one: Plausible Not Plausible
2.2 Test the null hypothesis that all four coefficients are zero, $H_0:\beta_1=\beta_2=\beta_3=\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name:________ Value:__________ P-value: ____________________ Circle one: Plausible Not Plausible
2.3 Two people have the same gender, the same education, and the same income, but one is 30 years old and the other is 40 years old. Using just the least squares estimate of the coefficient β_1 of age, the model would guess that the 40 year-old consumes 500 calories less than the 30 year old.	TRUE FALSE
2.4 Give the 95% confidence interval for the coefficient β_2 of female. If a man and a woman had the same age, education and income, the model would predict higher calories consumed for the woman.	95% CI: [,] TRUE FALSE

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2014: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due **Tuesday, October 21 noon**

Fit models 2 and 3 from the data page.	Fill in or CIRCLE the correct answer.
3.1 Assuming model 2 is true, test the null hypothesis that the coefficient of income in model 1 is zero, H_0:γ_3=0. What is the name of the test? What is the numerical value of the test statistic? What is the two-sided P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name:________ Value:__________ P-value: ____________________ Circle one: Plausible Not Plausible
3.2 Assuming model 1 is true, test the null hypothesis that model 3 is also true, that is, test H_0:β_3=β_4=0. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name:________ Value:__________ P-value: ____________________ Circle one: Plausible Not Plausible

Use the fit of model 1 to answer questions in part 4.	Fill in or CIRCLE the correct answer.
4.1 The Normal quantile plot of the residuals from model 1 gives the appearance of residuals that are Normally distributed.	TRUE FALSE
4.2 Test Normality of the residuals using the Shapiro-Wilk test. What is the P-value?	P-value: ____________________
4.3 The Normal plot of residuals suggests negative skewness, a long left-hand tail, with too many people consuming far fewer calories than the model predicts. This impression of negative skewness is reinforced by a boxplot of the residuals.	TRUE FALSE
4.4 Plot residuals as y against fitted values as x. Plot the absolute value of residuals as y against fitted values as x. Add a lowess curve (in red, so you can see it) in the second plot. The assumption of constant variance is clearly violated here, with larger absolute residuals being more common at low fitted calories (say 1500) than at higher fitted calories (say 2500), so the variance looks larger when the fitted values are smaller.	TRUE FALSE

PROBLEM SET #1 STATISTICS 500 FALL 2014: ANSWER PAGE 1: ANSWERS

This is an exam. Do not discuss it with anyone. Due noon in class

7 points each, except as noted

Question (Part 1)	CIRCLE the correct answer
1.1 There is one person who consumed more than 10,000 calories.	[✓ TRUE] FALSE
1.2 The lower quartile of calories for males is above the median for females.	[✓ TRUE] FALSE
1.3 The median waist size for females is more than 37 inches = 93.98 cm.	[✓ TRUE] FALSE
1.4 Of the five education categories, the lowest median of calories is the <9th grade category.	[✓ TRUE] FALSE

Fit model 1 from the data page.	Fill in or CIRCLE the correct answer.
2.1 Test the null hypothesis that the coefficient of income in model 1 is zero, $H_0:\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the two-sided P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name: t-test Value: 1.469 P-value: 0.1420 Circle one: [✓ Plausible] Not Plausible
2.2 Test the null hypothesis that all four coefficients are zero, $H_0:\beta_1=\beta_2=\beta_3=\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name: F-test Value: 258.5 P-value: <2.2 x 10^{-16} Circle one: Plausible [✓ Not Plausible]
2.3 Two people have the same gender, the same education, and the same income, but one is 30 years old and the other is 40 years old. Using just the least squares estimate of the coefficient β_1 of age, the model would guess that the 40 year-old consumes 500 calories less than the 30 year old.	TRUE [✓ FALSE] -12.1585 x 10 is not -500
2.4 Give the 95% confidence interval for the coefficient β_2 of female. If a man and a woman had the same age, education and income, the model would predict higher calories consumed for the woman.	95% CI: [-735.8, -637.6] Women consume less, maybe -686.7 TRUE [✓ FALSE]

PROBLEM SET #1 STATISTICS 500 FALL 2014: ANSWER PAGE 2: ANSWERS

This is an exam. Do not discuss it with anyone. Due in class on

Fit models 2 and 3 from the data page.	Fill in or CIRCLE the correct answer.
3.1 Assuming model 2 is true, test the null hypothesis that the coefficient of income in model 1 is zero, H_0:γ_3=0. What is the name of the test? What is the numerical value of the test statistic? What is the two-sided P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name: t-test Value: 2.53 P-value: 0.0115 Circle one: Plausible [✓ Not Plausible]
3.2 Assuming model 1 is true, test the null hypothesis that model 3 is also true, that is, test H_0:β_3=β_4=0. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible using the conventional 0.05 level standard? (9 points)	Name: F-test Value: 4.658 P-value: 0.009526 Circle one: Plausible [✓ Not Plausible]

Use the fit of model 1 to answer questions in part 4.	Fill in or CIRCLE the correct answer.
4.1 The Normal quantile plot of the residuals from model 1 gives the appearance of residuals that are Normally distributed.	TRUE [✓ FALSE] Remember, we expect a straight line for Normal data, and this is curved.
4.2 Test Normality of the residuals using the Shapiro-Wilk test. What is the P-value?	P-value: 2.2×10^{-16}
4.3 The Normal plot of residuals suggests negative skewness, a long left-hand tail, with too many people consuming far fewer calories than the model predicts. This impression of negative skewness is reinforced by a boxplot of the residuals.	TRUE [✓ FALSE] It is definitely skewed, but skewed right, not left. Sometimes people consume a lot more calories than the model expects
4.4 Plot residuals as y against fitted values as x. Plot the absolute value of residuals as y against fitted values as x. Add a lowess curve (in red, so you can see it) in the second plot. The assumption of constant variance is clearly violated here, with larger absolute residuals being more common at low fitted calories (say 1500) than at higher fitted calories (say 2500), so the variance looks larger when the fitted values are smaller.	TRUE [✓ FALSE] The assumption of constant variance looks wrong here, but calories are more unstable when fitted calories are higher.

Problem Set 1, Fall 2014
DOING THE PROBLEM SET IN R

```
1.1
> max(calories)
[1] 10463
1.2
> summary(calories[female==1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     70    1281    1691    1777    2150    5814 
> summary(calories[female==0])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    300    1735    2275    2450    2964   10460 
1.3
> tapply(waist*.393701,female,summary)
$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  25.08   35.91   39.49   40.01   43.66   66.42 
$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  23.50   33.62   37.64   38.24   42.24   64.96 
1.4
> tapply(calories,edf,summary)
$`<9th grade`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    279    1252    1734    1891    2336    6892 
$`9-11 grade`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    305    1415    1891    2087    2581    8077 
$`High School Grad/GED`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    164    1475    1994    2174    2643    9315 
$`Some college or AA degree`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    274    1464    1948    2136    2597   10460 
$`BA degree+`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     70    1513    1986    2110    2533    7301 
2
> md<-lm(calories~age+female+ed+income)
> summary(md)
Call:
lm(formula = calories ~ age + female + ed + income)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2958.7537    53.0998  55.721   <2e-16 ***
age          -12.1585     0.7219 -16.842   <2e-16 ***
female      -686.7017    25.0221 -27.444   <2e-16 ***
ed            18.8074    11.0156   1.707   0.0878 .  
income        12.8205     8.7301   1.469   0.1420    
Residual standard error: 880.8 on 4995 degrees of freedom
Multiple R-squared:  0.1715,    Adjusted R-squared:  0.1708 
F-statistic: 258.5 on 4 and 4995 DF,  p-value: < 2.2e-16
```

```
> confint(md)
                   2.5 %     97.5 %
(Intercept) 2854.654814 3062.85258
age          -13.573750  -10.74320
female      -735.756123 -637.64735
ed            -2.788033   40.40274
income        -4.294376   29.93546
3.1
> md2<-lm(calories~age+female+income)
> summary(md2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3013.9077    42.1511   71.50   <2e-16 ***
age          -12.3652     0.7118  -17.37   <2e-16 ***
female      -684.8067    25.0023  -27.39   <2e-16 ***
income        19.6402     7.7643    2.53   0.0115 *
3.2
> md3<-lm(calories~age+female)
> anova(md3,md)
Analysis of Variance Table
Model 1: calories ~ age + female
Model 2: calories ~ age + female + ed + income
  Res.Df        RSS Df Sum of Sq      F   Pr(>F)
1   4997 3882746217
2   4995 3875518001  2   7228216 4.6581 0.009526 **
4.
> qqnorm(md$resid)
> shapiro.test(md$resid)
        Shapiro-Wilk normality test
data:  md$resid
W = 0.9276, p-value < 2.2e-16
> boxplot(md$resid)
> plot(md$fit,md$resid)
> plot(md$fit,abs(md$resid)
> lines(lowess(md$fit,abs(md$resid)),col="red")
```

PROBLEM SET #2 STATISTICS 500 FALL 2014: DATA PAGE 1

Due in class at noon in class on Tuesday November 25, 2014.

This is an exam. Do not discuss it with anyone.

The data are from a paper by Card, Chetty and Weber (CCW) (2007) Cash-on-hand and competing models of intertemporal behavior: new evidence from the labor market. *Quarterly Journal of Economics*, 1511-1560. In their Table 1, their estimation sample had 650,922 people. The data sample for this problem set is simpler and smaller. It began as a random sample of 5000 individuals. Then people with missing data on key variables were removed, leaving 3923 people. The study is in various ways more complex than the data for the problem set, sometimes in interesting ways, but this is a problem set about regression, not a comprehensive study.

In Austria, a person who is unemployed may receive various unemployment benefits depending upon past employment. A person who has 3 or more years of job tenure receives a severance payment equivalent to two-months of pretax salary – i.e., two months wages. There is no severance payment for a person with less than 3 years of job tenure. The variables in the object ccwSt500 are as follows.

Variable name here	Variable name in CCW	Meaning
sevpay	sevpay	Severence pay, 1=yes, 0 =no
age	age	Age
female	female	1=female, 0=male
highed	high_ed	post-compulsory schooling, 1=yes, 0=no
married	married	1=married, 0=not
austrian	austrian	1=Austrian, 0=other
bluecollar	bluecollar	1=lost blue collar job, 0=lost other job
priorwage	ann_wage	Wage before job loss, euros/year
tenure	tenure_mths	Tenure (ie time) in previous job in months
uduration	uduration_mths	Duration of nonemployment in months
wrk12	wrk12<-pmax(12-uduration,0)	Months worked in the 12 months following unemployment
nextwage	ne_wage0	Monthly wage in next job (sometimes missing)
wage12	wage12<-ne_wage0*wrk12 wage12[wrk12==0]<-0	(Approximately) total wages in the 12 months after job loss.
id	1:3923	1 to 3923

The original paper had various motivations. One was to ask whether severance pay caused people to stay out of work longer. Another was to ask whether staying out longer was, in a sense, a good thing, because it gave people the chance to find a better job. You will do various regression with the data, but if you want to take a serious look at the paper's questions, you might want to look at the paper itself.

PROBLEM SET #2 STATISTICS 500 FALL 2014: DATA PAGE 2

This is an exam. Do not discuss it with anyone. Due Tuesday November 25, 2014.

```
> dim(ccwSt500)
[1] 3923   13
> colnames(ccwSt500)
 [1] "id"          "sevpay"     "female"     "highed"
"married"    "austrian"   "bluecollar" "priorwage"
"tenure"     "uduration"  "wrk12"      "nextwage"
"wage12"
```

The data are in the object `ccwSt500` in the course workspace at http://www-stat.wharton.upenn.edu/~rosenbap/ . You will have to download the workspace again to have the current version with `ccwSt500`. If you download the workspace and `ccwSt500` is not there, it probably means that you web browser remembers the last time you downloaded the file and thinks (incorrectly) that you do not need to download it again – in this case, clear the browser's memory and try again. There is a csv file `ccwSt500.csv` with the data at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` if you wish to use software other than R. The csv file should open in excel and other packages.

The variable wage12 is (approximately) what the person earned in the year after losing the job: it is the number of months work in the next 12 months times the salary at the new job. It is not perfect – the person might have gotten a raise in the middle of the year – but let's ignore its imperfections. Define a new variable wageloss as the difference between the annual wage prior to job loss minus wage12

```
> attach(ccwSt500)
> wageloss<-priorwage-wage12
```

so this what the individual would have earned at the salary of the job just lost minus what the individual earned in the year after job loss. A positive number is **bad** for the individual: it means his/her wage income went down quite a bit, by not working and perhaps by working for less. A negative number is **good** for the individual – despite the job loss, he/she earned more. How many people had a negative wage loss?

Model #1

$$\text{wageloss} = \beta_0 + \beta_1 \text{sevpay} + \beta_2 \text{age} + \beta_3 \text{female} + \beta_4 \text{highed} + \beta_5 \text{married} + \beta_6 \text{austrian} + \beta_7 \text{bluecollar} + \varepsilon \quad \text{where} \quad \varepsilon \text{ are iid } N(0,\sigma^2)$$

After fitting model 1, you should plot the residuals in the usual ways, even if questions do not ask you to do this.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Last name: ______________ **First** name:_____________ ID# __________________

PROBLEM SET #2 STATISTICS 500 FALL 2014: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due **Tuesday November 25, 2014.**

Fit model 1 on the data page. Use it for the questions in part 1.	Fill in or circle the correct answer.
1.1 Give the estimate and the 95% confidence interval for the β_1 the coefficient of sevpay.	Estimate:_______ CI: [,]
1.2 For two people who look the same in terms of all predictors in model 1 except sevpay, the model predicts 1081.78 euro less wage loss for the person who received the severance payment.	TRUE FALSE
1.3 Model 1 assumes that the relationship between wageloss and age is parallel for sevpay=1 and sevpay=0. In model 1, test this assumption. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), the two-sided p-value and state whether parallelism is plausible.	Name: __________ Value:_____________ DF: _____________ P-value:__________ PLAUSIBLE NOT PLAUSIBLE
1.4 Model 1 assumes that the relationship between wage loss and age is linear, not quadratic. In model 1, test the assumption that the relationship is linear against the alternative that it is quadratic in age. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom, the two-sided p-value and state whether a linear relationship with age is plausible.	Name: __________ Value:_____________ DF: _____________ P-value:__________ PLAUSIBLE NOT PLAUSIBLE
1.5 Use Tukey's method to test the null hypothesis that no transformation of wageloss is needed against the alternative that a power transformation would be helpful. Give the name, value, DF, two-sided p-value and state whether no transformation is needed.	Name: __________ Value:_____________ DF: _____________ P-value:__________ The null hypothesis that no transformation is needed is: PLAUSIBLE NOT PLAUSIBLE
1.6 $(y^p-1)/p$ is very close to the base-10 log of y, that is $\log_{10}(y)$, for p very near 0.	TRUE FALSE
1.7 In model 1, test the hypothesis that there is no interaction between female and married. Give the name, value, DF, two-sided p-value and state whether the hypothesis of no interaction is plausible.	Name: __________ Value:_____________ DF: _____________ P-value:__________ PLAUSIBLE NOT PLAUSIBLE

Last name: _______________ **First** name:_____________ ID# ___________________

PROBLEM SET #2 STATISTICS 500 FALL 2014: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due **Tuesday November 25, 2014.**

	Fill in or circle the correct answer.

2.1 Which observation in model 1 has the largest leverage (i.e. hatvalue)? Give the id number in the last column of ccwSt500. What is the numerical value of this person's leverage? Is this leverage large by the rule judging the size of the leverages?	id = __________
	leverage = _____________
	Circle one
	LARGE NOT LARGE

2.2 The individual identified in 2.1 has the leverage he/she does because the wage12 is so much lower than the priorwage.	TRUE FALSE
2.3 Which observation in model 1 has the largest absolute studentized residual (rstudent)? Give the id number in the last column of ccwSt500. What is the numerical value of this person's studentized residual (with its sign + or –)? This individual went from a low priorwage to a much higher wage12 (True or false)?	id = __________ studentized residual = _____________ Circle one TRUE FALSE
2.4 Is the person in 2.3 an outlier at the 0.05 level? What absolute value of the studentized residual would just barely reject a person as an outlier at the 0.05 level in model 1? What are the degrees of freedom used in computing this cut-off value?	Circle one Outlier: YES NO Value: ________________ Degrees of freedom: __________
2.5 Testing the null hypothesis of no outliers at the 0.05 level using the Bonferroni inequality with studentized residuals means we expect only one out of every 20 people in a regression to be wrongly judged an outlier.	Circle one TRUE FALSE
2.6 Which person had the largest absolute dffits? Give the id#. What is the value of this person's dffits with its sign +/-. This individual moved his/her fitted value up by 3.1 times its standard error (T/F)	id = __________dffits = _____________ Circle one TRUE FALSE
2.7 Do a Normal quantile plot of residuals from model 1 and add a qqline. The person identified in 2.6 is recognizably off the line, but even without this person, the residuals look longer-tailed than Normal.	Circle one TRUE FALSE

PROBLEM SET #2 STATISTICS 500 FALL 2014: ANSWER PAGE 1: **ANSWERS**

Fit model 1 on the data page. Use it for the questions in part 1.	Fill in or circle the correct answer.
1.1 Give the estimate and the 95% confidence interval for the β_1 the coefficient of sevpay.	Estimate: 1081.78 CI: [510.2, 1653.4]
1.2 For two people who look the same in terms of all predictors in model 1 except sevpay, the model predicts 1081.78 euro less wage loss for the person who received the severance payment.	TRUE [✓ FALSE] Associated with more, not less, wage loss.
1.3 Model 1 assumes that the relationship between wageloss and age is parallel for sevpay=1 and sevpay=0. In model 1, test this assumption. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), the two-sided p-value and state whether parallelism is plausible.	Name: t-statistic Value: 1.316 DF: 3914 P-value: 0.188 [✓ PLAUSIBLE] NOT PLAUSIBLE
1.4 Model 1 assumes that the relationship between wage loss and age is linear, not quadratic. In model 1, test the assumption that the relationship is linear against the alternative that it is quadratic in age. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom, the two-sided p-value and state whether a linear relationship with age is plausible.	Name: t-statistic Value: -2.225 DF: 3914 P-value: 0.026 PLAUSIBLE [✓ NOT PLAUSIBLE]
1.5 Use Tukey's method to test the null hypothesis that no transformation of wageloss is needed against the alternative that a power transformation would be helpful. Give the name, value, DF, two-sided p-value and state whether no transformation is needed.	Name: t-statistic Value: 6.14 DF: 3914 P-value: 9.31×10^{-10} The null hypothesis that no transformation is needed is: PLAUSIBLE [✓ NOT PLAUSIBLE]
1.6 $(y^p-1)/p$ is very close to the base-10 log of y, that is $\log_{10}(y)$, for p very near 0. *Base 10 and base e are different!*	TRUE [✓ FALSE] log10(3)= 0.4771213, log(3)=1.098612 ((3^0.001)-1)/0.001=1.099216
1.7 In model 1, test the hypothesis that there is no interaction between female and married. Give the name, value, DF, two-sided p-value and state whether the hypothesis of no interaction is plausible.	Name: t-statistic Value: 2.419 DF: 3914 P-value: 0.0156 PLAUSIBLE [✓ NOT PLAUSIBLE]

PROBLEM SET #2 STATISTICS 500 FALL 2014: ANSWER PAGE 2: ANSWERS

	Fill in or circle the correct answer.

2.1 Which observation in model 1 has the largest leverage (i.e. hatvalue)? Give the id number in the last column of ccwSt500. What is the numerical value of this person's leverage? Is this leverage large by the rule judging the size of the leverages?	id = 659
	leverage = 0.00557
	Circle one
	[✓ LARGE] NOT LARGE

2.2 The individual identified in 2.1 has the leverage he/she does because the wage12 is so much lower than the priorwage.	TRUE [✓ FALSE] Leverage is about x, not about y.
2.3 Which observation in model 1 has the largest absolute studentized residual (rstudent)? Give the id number in the last column of ccwSt500. What is the numerical value of this person's studentized residual (with its sign + or –)? This individual went from a low priorwage to a much higher wage12 (True or false)?	id = 393 studentized residual = -5.209 Circle one [✓ TRUE] FALSE
2.4 Is the person in 2.3 an outlier at the 0.05 level? What absolute value of the studentized residual would just barely reject a person as an outlier at the 0.05 level in model 1? What are the degrees of freedom used in computing this cut-off value?	Circle one Outlier: [✓ YES] NO Value: 4.370027 Degrees of freedom: 3914
2.5 Testing the null hypothesis of no outliers at the 0.05 level using the Bonferroni inequality with studentized residuals means we expect only one out of every 20 people in a regression to be wrongly judged an outlier.	Circle one TRUE [✓ FALSE] That would mean 3923*0.05 = 196 false outliers in the current data set!
2.6 Which person had the largest absolute dffits? Give the id#. What is the value of this person's dffits with its sign +/-. This individual moved his/her fitted value up by 3.1 times its standard error (T/F)	id = 393 dffits = -0.31 Circle one TRUE [✓ FALSE]
2.7 Do a Normal quantile plot of residuals from model 1 and add a qqline. The person identified in 2.6 is recognizably off the line, but even without this person, the residuals look longer-tailed than Normal.	Circle one [✓ TRUE] FALSE These data would be more appropriate for m-estimation or some other robust or nonparametric method.

Statistics 500 Fall 2014 Problem Set 2

Doing the Problem Set in R

```
> attach(ccwSt500)
> wageloss<-priorwage-wage12
Question 1.1
> mod<-
lm(wageloss~sevpay+age+female+highed+married+austrian+bluecollar)
> summary(mod)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3499.90     753.14   4.647 3.48e-06 ***
sevpay       1081.78     291.56   3.710 0.000210 ***
age           130.41      16.58   7.865 4.73e-15 ***
…
bluecollar  -2376.52     266.18  -8.928  < 2e-16 ***
---
> confint(mod)
                  2.5 %     97.5 %
(Intercept)  2023.31434  4976.4899
sevpay        510.15462  1653.4031
age            97.90418   162.9198
…
Question 1.3
> isevpayage<-sevpay*age
>summary(lm(wageloss~sevpay+age+female+highed+married+austrian+
    bluecollar+isevpayage))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3819.27     791.23   4.827 1.44e-06 ***
sevpay       -460.82    1208.27  -0.381 0.702935
age           120.34      18.26   6.590 4.97e-11 ***
…
isevpayage     47.79      36.33   1.316 0.188395
```

```
Question 1.4
> age2<-(age-mean(age))^2
> summary(lm(wageloss~sevpay+age+female+highed+married+austrian+
     bluecollar+age2))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3069.583    777.216   3.949 7.97e-05 ***
sevpay       1057.242    291.622   3.625 0.000292 ***
…
age2           -4.279      1.923  -2.225 0.026156 *
Question 1.5
> summary(lm(wageloss~sevpay+age+female+highed+married+austrian+
    bluecollar+tukey1df(mod)))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2856.3795   756.9425   3.774 0.000163 ***
sevpay          948.5399   291.0163   3.259 0.001126 **
…
tukey1df(mod)     1.8795     0.3063   6.136 9.31e-10 ***

Question 1.7
> fm<-female*married
> summary(lm(wageloss~sevpay+age+female+highed+married+austrian+
    bluecollar+fm))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3694.32     756.95   4.881 1.10e-06 ***
sevpay       1091.00     291.40   3.744 0.000184 ***
…
fm           1162.50     480.51   2.419 0.015596 *
Question 2.1
> which.max(hatvalues(mod))
659
659
> ccwSt500[659,]
    sevpay age female highed married austrian bluecollar priorwage
tenure uduration    wrk12 nextwage   wage12  id
834      0  49      1      0       1        0          1  19698.12
23.22581 0.6451613 11.35484 772.9453 8776.669 659
> hatvalues(mod)[659]
        659
0.005571402
Question 2.3
> rstudent(mod)[393]
      393
-5.209135
> ccwSt500[393,]
    sevpay age female highed married austrian bluecollar priorwage
tenure uduration    wrk12 nextwage   wage12  id
492      1  44      0      0       0        1          0  1348.568
51.48387 0.4516129 11.54839 2427.689 28035.89 393
Question 2.4
> dim(ccwSt500)
[1] 3923   14
> qt(0.025/3923,3914)
[1] -4.370027
Question 2.5
> which.max(abs(dffits(mod)))
393
393
> dffits(mod)[393]
       393
-0.3108647
Question 2.7
> qqnorm(mod$resid)
> which.min(mod$resid)
393
393
> qqline(mod$resid)
```

PROBLEM SET #3 STATISTICS 500 FALL 2014: DATA PAGE 1

Due at noon Thursday, December 18, in my office, 473 JMHH.

This is an exam. Do not discuss it with anyone.

Data set `antineoplastic` is adapted from an article by Kopjar and Garaj-Vrhovac in Mutagenesis (2001) vol 16, #1, pp 71-78, concerned with the possibility that working with antineoplastic drugs (i.e., cancer chemotherapies) might damage the DNA of the nurses and doctors. DNA damage is measured by the tail moment of the comet assay. In the comet assay, genetic material from a cell is placed in an electric field, which pulls DNA in one direction. DNA is a large and hard to move, but if some of the DNA is broken it moves more. The assay creates the appearance of a comet with a tail, intact DNA being in the comet's head, broken DNA in the tail. Large values of the tail moment y_{ij} are taken to mean greater DNA breakage, not a good thing. There are 3 groups of different people, i=1,2,3, each of size 20, j=1,…,20: (i) gloves = workers wearing gloves, (ii) hood = workers wearing gloves in a safety cabinet under a laminar hood in a that vents fumes upwards, (iii) control = individuals with no exposure to antineoplastic drugs. You should do `boxplot(tailmoment~group)` and think about what you see.

Model 1 says $y_{ij} = \mu + \gamma_i + \varepsilon_{ij}$ where ε_{ij} are iid $N(0,\sigma^2)$ and $\gamma_1 + \gamma_2 + \gamma_3 = 0$ **Hypothesis set A** has three hypotheses A = $\{H_{12}: \gamma_1 = \gamma_2, H_{13}: \gamma_1 = \gamma_3, H_{23}: \gamma_2 = \gamma_3\}$. Hypothesis set B is similar to hypothesis set A, but it is for the situation with 4 groups rather than the 3 groups in the antineoplastic data. **Hypothesis set B** has six hypotheses B = $\{H_{12}: \gamma_1 = \gamma_2, H_{13}: \gamma_1 = \gamma_3, H_{23}: \gamma_2 = \gamma_3, H_{14}: \gamma_1 = \gamma_4, H_{24}: \gamma_2 = \gamma_4, H_{34}: \gamma_3 = \gamma_4\}$. When a question asks about hypothesis set B, it is not a question about the antineoplastic data, but about some analogous situation with 4 groups instead of three.

Question 2.5 asks you to build two orthogonal contrasts, `exp.control` and `glove.hood`. It is an easy question, but if you mess it up, then you mess up several more questions.

The two contrasts you built in question 2.5 constitute 2 new variables. **Model 2** uses these two variables plus two more from the data set, namely `age` and `smoker`, which is 1 for a smoker, 0 for a nonsmoker. **Model 2** is:

tailmoment = $\beta_0 + \beta_1$ `exp.control` + β_2 `glove.hood` + β_3 `age` + β_4 `smoker` + ε_{ij} where ε_{ij} are iid $N(0,\sigma^2)$.

Model 2 has 2^4 submodels formed by deleting predictors, including the model with all 4 predictors and the model with no predictors (just the constant).

Model 3 has

tailmoment = $\beta_0 + \beta_1$ `exp.control` + β_2 `glove.hood` + ε_{ij} where ε_{ij} are iid $N(0,\sigma^2)$, so it is like model 2, but without age and smoker.

Question 2.1 mentions the experimentwise error rate and the familywise error rate. These are two names for the same thing.

PROBLEM SET #3 STATISTICS 500 FALL 2014: DATA PAGE 2
This is an exam. Do not discuss it with anyone.
Due at noon Thursday, December 18, in my office, 473 JMHH.

The data are in the object `antineoplastic` in the course workspace at http://www-stat.wharton.upenn.edu/~rosenbap/ . You will have to download the workspace again to have the current version with `antineoplastic`. If you download the workspace and `antineoplastic` is not there, it probably means that you web browser remembers the last time you downloaded the file and thinks (incorrectly) that you do not need to download it again – in this case, clear the browser's memory and try again. There is a csv file `antineoplastic.csv` with the data at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` if you wish to use software other than R. The csv file should open in excel and other packages.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

The exam is due Thursday, Dec 18, 2014 at noon. You may turn in the exam early by placing it in an envelope addressed to me and leaving it in my mail box in statistics, 4th floor, JMHH. If you prefer, give it to Noel at the front desk in statistics. **Make and keep a photocopy of your answer page**. The answer key will be posted in the revised bulk pack on-line. You can compare your photocopy to the on-line answer page.

HAVE A GREAT HOLIDAY!

Last name: _______________ **First** name:_____________ ID# ___________________

PROBLEM SET #3 STATISTICS 500 FALL 2014: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon Thursday, December 18.

Use the antineoplastic data and model 1 to answer the following questions.	Fill in or CIRCLE the correct answer
1.1 Assuming model 1 is true, test the null hypothesis H_0: $\gamma_1 = \gamma_2 = \gamma_3 = 0$. Give the name of the test, the value of the test statistics, the degrees of freedom (DF), the P-value and indicate whether the null hypothesis is plausible.	Name:_________ Value: ___________ DF:____________ P-value:__________ PLAUSIBLE NOT PLAUSIBLE
1.2 Assuming model 1 is true, use the Tukey method to build three simultaneous 95% confidence intervals for $\gamma_1 - \gamma_2$, $\gamma_1 - \gamma_3$, and $\gamma_2 - \gamma_3$, but REPORT HERE ONLY the interval for $\gamma_1 - \gamma_2 = \gamma_{glove} - \gamma_{hood}$, being careful to get the sign (+/-) correct for glove-minus-hood. The interval suggests that people working with a hood and gloves had lower tail moments than people working with just gloves. (T or F)	95% simultaneous interval: $\gamma_1 - \gamma_2 = \gamma_{glove} - \gamma_{hood}$ [,] TRUE FALSE
1.3 If model 1 were true, the three intervals in 1.2 would all cover their three parameters in at least 95% of experiments.	TRUE FALSE
1.4 Under model 1 for the antineoplastic data, if you used the t-test to compare two group means, but you used the three group pooled estimate of σ^2, then what would be the degrees of freedom (DF3) for the test? What would the degrees of freedom (DF2) be if you used just the data from the two groups being compared to estimate σ^2? What is the two-sided 95% critical value for |t| for a **single** t-test with the corresponding DF?	3 group estimate DF3: _____________ 2 group estimate DF2: _____________ 3 group critical |t|:________________ 2 group critical |t|:________________ (This question asks about doing **one** t-test with DF3 or DF2 degrees of freedom. It is **NOT** about testing **several** hypotheses.)
1.5 If you test k true null hypotheses, and there is probability λ that you reject each one, then you expect to reject λk of these true null hypotheses.	TRUE FALSE
1.6 Use Holm's method and the two-sided pairwise t-test with pooled estimate of σ^2 to test the three null hypothesis in hypothesis set A. Give the Holm-adjusted p-values.	Give 3 adjusted p-values H_{12}: $\gamma_1 = \gamma_2$ or $\gamma_{glove} = \gamma_{hood}$:___________ H_{13}: $\gamma_1 = \gamma_3$, or $\gamma_{glove} = \gamma_{control}$: _________ H_{23}: $\gamma_2 = \gamma_3$}, or $\gamma_{hood} = \gamma_{control}$: _________

Due **Last** name: _______________ **First** name:_____________ ID#______________

PROBLEM SET #3 STATISTICS 500 FALL 2014: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon Thursday Dec 18.

Use model 1 and the antineoplastic data to answer the following questions.	Fill in or CIRCLE the correct answer
2.1 If pairwise t-tests adjusted by the Bonferroni inequality reject a particular hypothesis with an experimentwise or familywise error rate of 0.05, then Holm's method rejects that hypothesis also.	TRUE FALSE
2.2 Hypothesis set A could contain exactly 1 true hypothesis or exactly 2 true hypotheses or exactly 3 true hypotheses, which is why the R function defaults to Holm's method.	TRUE FALSE
2.3 Hypothesis set B could contain exactly 1 true and 5 false null hypotheses.	TRUE FALSE
2.4 In hypothesis set A for model 1, the contrasts for Hypotheses H_{12}: $\gamma_1 = \gamma_2$ and H_{23}: $\gamma_2 = \gamma_3$ are two orthogonal contrasts.	TRUE FALSE
2.5 Give two orthogonal contrasts with integer weights for (a) exposed to drugs versus control, (b) glove only vs gloves plus hood. Fill in 6 integers as contrast weights for 2 contrasts.	a ______ _______ ________ b ______ _______ ________ Glove Hood Control

3. Use model 1 and the contrasts in question 2.5 to fill in the following anova table.

	Sum of squares	DF	Mean Square	F
Between Groups				
Exposed versus Control				
Gloves versus Hood				
With Groups Residual				

Use models 2 and 3 for part 4.	Fill in or CIRCLE the correct answer
4.1 Which of the 16 submodels of model 2 has the smallest C_P? List the variables in this model, its size (1+#vars), the value of C_P.	Variable in this model (**list names**): Size= _________ C_P= _______________
4.2 What is the PRESS value for model 2 and for model 3?	Model 2 Model 3 PRESS= ___________ ____________
4.3 How many observations have large leverages or hatvalues in model 2? In model 3? Give two counts.	Model 2:________ Model 3:___________

4.4 Give the variance inflation factors for models 2 and 3.

Put in VIFs	exp.control	glove.hood	age	smoker
Model 2				
Model 3			XXXXXXXX	XXXXXXXX

Answers

PROBLEM SET #3 STATISTICS 500 FALL 2014: ANSWER PAGE 1

6 points each, except #3 which is 10 points

Use the antineoplastic data and model 1 to answer the following questions.	Fill in or CIRCLE the correct answer
1.1 Assuming model 1 is true, test the null hypothesis H_0: $\gamma_1 = \gamma_2 = \gamma_3 = 0$. Give the name of the test, the value of the test statistics, the degrees of freedom (DF), the P-value and indicate whether the null hypothesis is plausible.	Name: F-test Value: 56.62 DF: 2 and 57 P-value: 2.86×10^{-14} PLAUSIBLE [✓ NOT PLAUSIBLE]
1.2 Assuming model 1 is true, use the Tukey method to build three simultaneous 95% confidence intervals for $\gamma_1 - \gamma_2$, $\gamma_1 - \gamma_3$, and $\gamma_2 - \gamma_3$, but REPORT HERE ONLY the interval for $\gamma_1 - \gamma_2 = \gamma_{glove} - \gamma_{hood}$, being careful to get the sign (+/-) correct for glove-minus-hood. The interval suggests that people working with a hood and gloves had lower tail moments than people working with just gloves. (T or F)	95% simultaneous interval: $\gamma_1 - \gamma_2 = \gamma_{glove} - \gamma_{hood}$ [.969 , 3.481] [✓ TRUE] FALSE
1.3 If model 1 were true, the three intervals in 1.2 would all cover their three parameters in at least 95% of experiments.	[✓ TRUE] FALSE
1.4 Under model 1 for the antineoplastic data, if you used the t-test to compare two group means, but you used the three group pooled estimate of σ^2, then what would be the degrees of freedom (DF3) for the test? What would the degrees of freedom (DF2) be if you used just the data from the two groups being compared to estimate σ^2? What is the two-sided 95% critical value for |t| for a **single** t-test with the corresponding DF?	3 group estimate DF3: 57 = 60-3 2 group estimate DF2: 38 = 40-2 3 group critical |t|: 2.002 2 group critical |t|: 2.024 Once you have 38 DF for error, going to 57 does not add much. And you have to assume the three groups have the same σ^2 to get the extra DF. The 2-group test doesn't assume anything about the third group.
1.5 If you test k true null hypotheses, and there is probability λ that you reject each one, then you expect to reject λk of these true null hypotheses.	[✓ TRUE] FALSE
1.6 Use Holm's method and the two-sided pairwise t-test with pooled estimate of σ^2 to test the three null hypothesis in hypothesis set A. Give the Holm-adjusted p-values.	Give 3 adjusted p-values H_{12}: $\gamma_1 = \gamma_2$ or $\gamma_{glove} = \gamma_{hood}$: 7.7×10^{-5} H_{13}: $\gamma_1 = \gamma_3$, or $\gamma_{glove} = \gamma_{control}$: 1.4×10^{-14} H_{23}: $\gamma_2 = \gamma_3$}, or $\gamma_{hood} = \gamma_{control}$: 8.8×10^{-8}

Answers

PROBLEM SET #3 STATISTICS 500 FALL 2014: ANSWER PAGE 2

Use model 1 and the antineoplastic data to answer the following questions.	Fill in or CIRCLE the correct answer
2.1 If pairwise t-tests adjusted by the Bonferroni inequality reject a particular hypothesis with an experimentwise or familywise error rate of 0.05, then Holm's method rejects that hypothesis also.	[✓ TRUE] FALSE
2.2 Hypothesis set A could contain exactly 1 true hypothesis or exactly 2 true hypotheses or exactly 3 true hypotheses, which is why the R function defaults to Holm's method.	TRUE [✓ FALSE]
2.3 Hypothesis set B could contain exactly 1 true and 5 false null hypotheses.	[✓ TRUE] FALSE
2.4 In hypothesis set A for model 1, the contrasts for Hypotheses H_{12}: $\gamma_1 = \gamma_2$ and H_{23}: $\gamma_2 = \gamma_3$ are two orthogonal contrasts.	TRUE [✓ FALSE]
2.5 Give two orthogonal contrasts with integer weights for (a) exposed to drugs versus control, (b) glove only vs gloves plus hood. Fill in 6 integers as contrast weights for 2 contrasts.	a 1 1 -2 b 1 -1 0 Glove Hood Control

3. Use model 1 and the contrasts in question 2.5 to fill in the following anova table.

	Sum of squares	DF	Mean Square	F
Between Groups	308.6	2	154.3	56.6
Exposed versus Control	259.1	1	259.1	95.1
Gloves versus Hood	49.5	1	49.5	18.2
With Groups Residual	155.3	57	2.72	

Use models 2 and 3 for part 4.	Fill in or CIRCLE the correct answer
4.1 Which of the 16 submodels of model 2 has the smallest C_P? List the variables in this model, its size (1+#vars), the value of C_P.	Variable in this model (**list names**): exp.control and glove.hood Size= 3 C_P= 3.447
4.2 What is the PRESS value for model 2 and for model 3?	Model 2 Model 3 PRESS= 177.8 172.1
4.3 How many observations have large leverages or hatvalues in model 2? In model 3? Give two counts.	Model 2: 1 Model 3: 0 Without age and smoker, the design is balanced, constant leverage.

4.4 Give the variance inflation factors for models 2 and 3.

Put in VIFs	exp.control	glove.hood	age	smoker
Model 2	1.458	1.057	1.170	1.432
Model 3	1.000	1.000	XXXXXXXX	XXXXXXXX

DOING THE PROBLEM SET IN R

Problem Set 3, Fall 2014, Statistics 500, Answers

```
1.1
> summary(aov(tailmoment~group))
            Df Sum Sq Mean Sq F value   Pr(>F)    
group        2  308.6  154.29   56.62 2.86e-14 ***
Residuals   57  155.3    2.72                     
1.2
> TukeyHSD(aov(tailmoment~group))
Tukey multiple comparisons of means 95% family-wise confidence level
Fit: aov(formula = tailmoment ~ group)
$group
                 diff       lwr        upr    p adj
hood-glove    -2.2250 -3.481165 -0.9688347 2.24e-04
control-glove -5.5205 -6.776665 -4.2643347 0.00e+00
control-hood  -3.2955 -4.551665 -2.0393347 1.00e-07
1.4
> qt(.025,57)
[1] -2.002465
> qt(.025,38)
[1] -2.024394
1.5 (Holm method is the default)
> pairwise.t.test(tailmoment,group)
        Pairwise comparisons using t tests with pooled SD 
        glove   hood   
hood    7.7e-05 -      
control 1.4e-14 8.8e-08
P value adjustment method: holm 
3.
> exp.control<-c(1,1,-2)
> glove.hood<-c(1,-1,0)
> contrasts(group)<-cbind(exp.control,glove.hood)
> contrasts(group)
        exp.control glove.hood
glove             1          1
hood              1         -1
control          -2          0
> mm<-model.matrix(aov(tailmoment~group))
> head(mm)
  (Intercept) groupexp.control groupglove.hood
1           1                1               1
2           1                1              -1
3           1               -2               0
> is.data.frame(mm)
[1] FALSE
```

Problem Set 3, Fall 2014, Statistics 500, Answers continued

```
> mm<-as.data.frame(mm)
> attach(mm)
> anova(lm(tailmoment~groupexp.control+groupglove.hood))
                 Df  Sum Sq Mean Sq F value    Pr(>F)    
groupexp.control  1 259.073 259.073  95.076 9.408e-14 ***
groupglove.hood   1  49.506  49.506  18.168 7.678e-05 ***
Residuals        57 155.320   2.725                      
```

Because contrasts are orthogonal, order does not matter:

```
> anova(lm(tailmoment~groupglove.hood+groupexp.control))
                 Df  Sum Sq Mean Sq F value    Pr(>F)    
groupglove.hood   1  49.506  49.506  18.168 7.678e-05 ***
groupexp.control  1 259.073 259.073  95.076 9.408e-14 ***
Residuals        57 155.320   2.725                      
> library(leaps)
> x<-cbind(groupexp.control,groupglove.hood,age,smoke)
> mod<-leaps(x=x,y=tailmoment,names=colnames(x))
> cbind(mod$which,mod$size,mod$Cp)
  groupexp.control groupglove.hood age smoke              
1                0               0   1     0 2 114.906221
2                1               1   0     0 3   3.446797
2                1               0   1     0 3  17.478437
2                1               0   0     1 3  20.867206
2                0               1   1     0 3  96.002119
3                1               1   1     0 4   4.204977
3                1               1   0     1 4   4.556732
4                1               1   1     1 5   5.000000
> modfull<-
lm(tailmoment~groupexp.control+groupglove.hood+age+smoke)
> library(DAAG)
> press(modfull)
[1] 177.8312
> press(lm(tailmoment~groupexp.control+groupglove.hood))
[1] 172.0993
> vif(modfull)
groupexp.control  groupglove.hood              age
smoke 
          1.4584           1.0568           1.1699
1.4316 
> vif(lm(tailmoment~groupexp.control+groupglove.hood))
groupexp.control  groupglove.hood 
               1                1 
> summary(hatvalues(modfull))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.05007 0.07058 0.08064 0.08333 0.09565 0.17330 
> 2*0.08333
[1] 0.16666
> summary(hatvalues(lm(tailmoment~groupexp.control+groupglove.hood)))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.05    0.05    0.05    0.05    0.05    0.05 
```

PROBLEM SET #1 STATISTICS 500 FALL 2015: DATA PAGE 1

Due in class at noon in class on Tuesday 13 October 2015

This is an exam. Do not discuss it with anyone.

The data are selected for illustration from a study by Tager et al (1979), *American Journal of Epidemiology*, 110, 15-26, but have been substantially simplified. The data concern fev = forced expiratory volume of children measured in liters, which is an index of pulmonary function and is the volume of air expelled after 1 second of constant effort. Other variables are age of the child in years, height of the child in inches, female = 1 for female, 0 for male, and smoker = 1 if current smoker, 0 if not a current smoker. Each row of data is a different child.

The data are in an object called `rfev2` in the course workspace at http://www-stat.wharton.upenn.edu/~rosenbap/ If you are not using R, then the link data.csv on the same page will give you the data as a csv-file that many programs can read, including excel.

```
> head(rfev2)
    id   fev age height female smoker
1  301 1.708   9   57.0      1      0
2  451 1.724   8   67.5      1      0
3  501 1.720   7   54.5      1      0
4  642 1.558   9   53.0      0      0
5  901 1.895   9   57.0      0      0
6 1701 2.336   8   61.0      1      0

> dim(rfev2)
[1] 654   6
```

Model 1: fev = $\beta_0 + \beta_1$ age + β_2 height + ε where ε is iid $N(0, \sigma^2)$

Model 2: fev = $\gamma_0 + \gamma_1$ age + γ_2 height + γ_3 female + ζ where ζ is iid $N(0, \omega^2)$

Model 1 has betas and model 2 has gammas so that different things have different symbols. It does not matter which Greek letter we use.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Last name:______________________ First Name: _________________

ID#:____________________

Statistics 500, Problem 1, Fall 2015, p1. **This problem set is an exam. Do not discuss it with anyone**.

Fit model 1 from the data page and use it to answer the questions in part 1.	Fill in or circle the correct answer
1.1 In model 1, what is the numerical value of the estimate of σ?	Estimate of σ:
1.2 Under model 1, give the 95% two-sided confidence interval for the coefficient of age, β_1.	95% CI: [,]
1.3 In model 1, test the null hypothesis that the coefficient of height, β_2, is zero, H_0: $\beta_2=0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.	Name:_________ Value:__________ DF:___________ P-value:_____________ Circle one: H_0 is PLAUSIBLE NOT PLAUSIBLE
1.4 In model 1, test the null hypothesis that both slopes are zero, H_0: $\beta_1=\beta_2=0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.	Name:_________ Value:__________ DF:___________ P-value:_____________ Circle one: H_0 is PLAUSIBLE NOT PLAUSIBLE
1.5 In model 1, the correlation between fev and fitted fev (i.e., yhat) is 0.7664.	Circle one: TRUE FALSE

Use the residuals from model 1 on the data page to answer questions in part 2.	Fill in or circle the correct answer
2.1 Do a normal plot, a boxplot, and a Shapiro-Wilk test using the residuals from model 1. Circle the **letters** of true statements, perhaps more than one, perhaps none. E.g., you might circle a. if you thought a was true and the rest false.	Circle the LETTER for each true statement a. The residuals look Normal. b. The residuals have a short right tail and a long left tail compared to the Normal. c. The residuals have a long right tail and a long left tail compared to the Normal. d. The Shapiro-Wilk tests accepts Normality as a plausible distribution for the residuals.
2.2 Plot residuals as y against fitted values as x. Add the lowess smooth. Plot residuals as y against height as x. Add the	Circle the LETTER for each true statement a. The lowess smooth of the plot

lowess smooth. Use round() to round the fitted values to integers, forming 4 groups, 1, 2, 3, and 4; then, boxplot the residuals in these four groups. Circle the **letters** of true statements, perhaps more than one, perhaps none.	against **height** clearly exhibits an *inverted* (i.e., upside down) U-shape, indicating nonlinearity. b. The plots and boxplots clearly indicate that the assumption of constant variance is correct.
2.3 The child with the largest absolute residual is 15 years old, 69 inches tall, with the largest fev in the data set.	Circle one: TRUE FALSE

Last name:______________________ First Name: __________________
ID#:____________________

Statistics 500, Problem 1, Fall 2015, p2. **This problem set is an exam. Do not discuss it with anyone**.

<table>
<tr><td>Fit model 2 and use it for questions in part 3.</td><td>Fill in or circle the correct answer</td></tr>
<tr><td>3.1 In model 2, female children are estimated to have a higher fev than male children of the same height and age.</td><td>Circle one
TRUE FALSE</td></tr>
<tr><td>3.2 In model 2, test the null hypothesis that the coefficient of age and the coefficient of female are both zero, H_0: $\gamma_1 = \gamma_3 = 0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.</td><td>Name:__________ Value:___________
DF:____________ P-value:______________
Circle one:
H_0 is PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>3.3 For the hypothesis tested in question 3.2, fill in the following ANOVA = analysis of variance table. In the table, SS means sum-of-squares, DF means degrees-of-freedom, and MS means mean-square.</td><td><table>
<tr><td></td><td>SS</td><td>DF</td><td>MS</td></tr>
<tr><td>Full Model</td><td></td><td></td><td></td></tr>
<tr><td>Reduced Model</td><td></td><td></td><td></td></tr>
<tr><td>Added variables</td><td></td><td></td><td></td></tr>
<tr><td>Residual</td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td>3.4 Using model 2, give a 2-sided 95% confidence interval for location of the regression surface (ie, the fitted fev value) for a female, 5 foot 3 inches tall, age 15.</td><td>95% Interval [,]</td></tr>
<tr><td>3.5 Using model 2, give a 2-sided 95% interval for the fev for a new child who is a female, 5 foot 3 inches tall, age 15. This new child would be in addition to the 654 children in the data set, but is imagined to follow the same model 2.</td><td>95% Interval [,]</td></tr>
</table>

3.6 If the number of children in the data set, currently 654, where to increase and increase but always follow the same model 2, then the interval in question 3.4 would shrink to a point, but the interval in question 3.5 would not.	Circle one TRUE FALSE

Part 4 asks hypothetical questions.	Circle the correct answer
4.1 In some data set, it could happen that a test of H_0: $\gamma_1 = \gamma_3 = 0$, as in part 3.2, could reject at the 0.05 level, but two separate test of H_0: $\gamma_1 = 0$ and H_0: $\gamma_3 = 0$ in the same model could both accept at the two-sided 0.05 level.	Circle one TRUE FALSE
4.2 The actual study from which these data were selected measured each child at several ages to study growth in fev. Imagine having 654 children, each measured at 3 ages, and fitting model 2 to these 654x3 = 1962 measurements. A problem with model 2 is that it is not very plausible that 3 fev's from one child at 3 ages will be independent.	Circle one TRUE FALSE

Statistics 500, Fall 2105, Problem 1, Answers (6 points each except as noted)

Fit model 1 from the data page and use it to answer the questions in part 1.	Fill in or circle the correct answer
1.1 In model 1, what is the numerical value of the estimate of σ?	Estimate of σ: 0.4197
1.2 Under model 1, give the 95% two-sided confidence interval for the coefficient of age, β_1.	95% CI: [0.0364, 0.0722]
1.3 In model 1, test the null hypothesis that the coefficient of height, β_2, is zero, H_0: $\beta_2=0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.	t-DF are DF for σ^2, here 651. Name: t-statistic Value: 23.263 DF: 651 P-value: $< 2 \times 10^{-16}$ Circle one: H_0 is PLAUSIBLE [✓ NOT] PLAUSIBLE
1.4 In model 1, test the null hypothesis that both slopes are zero, H_0: $\beta_1=\beta_2=0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and	Name: F-statistic Value: 1068 DF: 2 and 651 P-value: $< 2.2 \times 10^{-16}$ Circle one: H_0 is PLAUSIBLE [✓ NOT]

state whether the null hypothesis is plausible.	PLAUSIBLE
1.5 In model 1, the correlation between fev and fitted fev (i.e., yhat) is 0.7664.	Circle one: TRUE [✓ FALSE]

Use the residuals from model 1 on the data page to answer questions in part 2.	Fill in or circle the correct answer
2.1 Do a normal plot, a boxplot, and a Shapiro-Wilk test using the residuals from model 1. Circle the **letters** of true statements, perhaps more than one, perhaps none. E.g., you might circle a. if you thought a was true and the rest false.	Circle the LETTER for each true statement e. The residuals look Normal. f. The residuals have a short right tail and a long left tail compared to the Normal. ○ g. The residuals have a long right tail and a long left tail compared to the Normal. h. The Shapiro-Wilk tests accepts Normality as a plausible distribution for the residuals.
2.2 Plot residuals as y against fitted values as x. Add the lowess smooth. Plot residuals as y against height as x. Add the lowess smooth. Use round() to round the fitted values to integers, forming 4 groups, 1, 2, 3, and 4; then, boxplot the residuals in these four groups. Circle the **letters** of true statements, perhaps more than one, perhaps none.	Circle the LETTER for each true statement c. The lowess smooth of the plot against **height** clearly exhibits an inverted (i.e., upside down) U-shape, indicating nonlinearity. d. The plots and boxplots clearly indicate that the assumption of constant variance is correct.
2.3 The child with the largest absolute residual is 15 years old, 69 inches tall, with the largest fev in the data set.	Circle one: [✓ TRUE] FALSE

<table>
<tr><th>Fit model 2 and use it in part 3.</th><th>Fill in or circle the correct answer</th></tr>
<tr><td>3.1 In model 2, female children are estimated to have a higher fev than male children of the same height and age.</td><td>Circle one
TRUE [✓ FALSE]</td></tr>
<tr><td>3.2 In model 2, test the null hypothesis that the coefficient of age and the coefficient of female are both zero, H_0: $\gamma_1 = \gamma_3 = 0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible. (8 points)</td><td>An F has (numerator,denominator) DF-2 numbers.
Name: F-statistic Value: 30.213
DF: 2 and 650 P-value: 2.84 x 10^{-13}
Circle one:
H_0 is PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>3.3 For the hypothesis tested in question 3.2, fill in the following ANOVA = analysis of variance table. In the table, SS means sum-of-squares, DF means degrees-of-freedom, and MS means mean-square. (8 points)</td><td>
<table>
<tr><th></th><th>SS</th><th>DF</th><th>MS</th></tr>
<tr><td>Full Model</td><td>380.27</td><td>3</td><td>126.76</td></tr>
<tr><td>Reduced Model</td><td>369.99</td><td>1</td><td>369.99</td></tr>
<tr><td>Added variables</td><td>10.28</td><td>2</td><td>5.14</td></tr>
<tr><td>Residual</td><td>110.65</td><td>650</td><td>0.170</td></tr>
</table>
</td></tr>
<tr><td>3.4 Using model 2, give a 2-sided 95% confidence interval for location of the regression surface (ie, the fitted fev value) for a female, 5 foot 3 inches tall, age 15.</td><td>95% Interval [2.97, 3.14]</td></tr>
<tr><td>3.5 Using model 2, give a 2-sided 95% interval for the fev for a new child who is a female, 5 foot 3 inches tall, age 15. This new child would be in addition to the 654 children in the data set, but is imagined to follow the same model 2.</td><td>95% Interval [2.24, 3.87]</td></tr>
<tr><td>3.6 If the number of children in the data set, currently 654, where to increase and increase but always follow the same model 2, then the interval in question 3.4 would shrink to a point, but the interval in question 3.5 would not.</td><td>With more and more data, we learn where line is, but not where a new child is.
Circle one
[✓ TRUE] FALSE</td></tr>
</table>

Part 4 asks hypothetical questions.	Circle the correct answer
4.1 In some data set, it could happen that a test of H_0: $\gamma_1 = \gamma_3 = 0$, as in part 3.2, could reject at the 0.05 level, but two separate test of H_0:	This might happen if x_1 and x_3 are highly correlated. [✓ TRUE] FALSE

$\gamma_1 = 0$ and H_0: $\gamma_3 = 0$ in the same model could both accept at the two-sided 0.05 level.	
4.2 The actual study from which these data were selected measured each child at several ages to study growth in fev. Imagine having 654 children, each measured at 3 ages, and fitting model 2 to these 654x3 = 1962 measurements. A problem with model 2 is that it is not very plausible that 3 fev's from one child at 3 ages will be independent.	Circle one [✓ TRUE] FALSE Multiple measures on each child are called "repeated measures" or "panel data" or "longitudinal data," and typically cannot be viewed as independent observations, violating an assumption of regression.

Statistics 500, Problem 1, Fall 2015: Doing the Problem Set in R.

1.

```
> md1<-lm(fev~age+height)
> summary(md1)
lm(formula = fev ~ age + height)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.610466   0.224271 -20.558  < 2e-16 ***
age          0.054281   0.009106   5.961 4.11e-09 ***
height       0.109712   0.004716  23.263  < 2e-16 ***
Residual standard error: 0.4197 on 651 degrees of freedom
Multiple R-squared: 0.7664,     Adjusted R-squared: 0.7657
F-statistic:  1068 on 2 and 651 DF,  p-value: < 2.2e-16
> confint(md1)
                  2.5 %      97.5 %
(Intercept) -5.05084726 -4.17008507
age          0.03639976  0.07216159
height       0.10045104  0.11897263
> cor(md1$fitted,fev)
[1] 0.8754474
> cor(md1$fitted,fev)^2
[1] 0.7664081   This is R^2, not R.
2.
> boxplot(md1$resid)
> qqnorm(md1$resid)
> qqline(md1$resid)
> shapiro.test(md1$resid)
        Shapiro-Wilk normality test
data:  md1$resid
W = 0.9865, p-value = 9.794e-06
> plot(md1$fit,md1$resid)
> lines(lowess(md1$fit,md1$resid),col="red")
> plot(height,md1$resid)
> lines(lowess(height,md1$resid),col="red")
> boxplot(md1$resid~round(md1$fit))
> which.max(abs(md1$resid))
624
> rfev2[624,]
       id   fev age height female smoker
624 25941 5.793  15     69       0      0
> summary(fev)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.791   1.981   2.548   2.637   3.118   5.793
```

```
3.1
> md2<-lm(fev~age+height+female)
> summary(md2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.287448   0.230253 -18.621  < 2e-16 ***
age          0.061364   0.009069   6.766 2.96e-11 ***
height       0.104560   0.004756  21.986  < 2e-16 ***
female      -0.161112   0.033125  -4.864 1.45e-06 ***

Residual standard error: 0.4126 on 650 degrees of freedom
Multiple R-squared: 0.7746,     Adjusted R-squared: 0.7736
F-statistic: 744.6 on 3 and 650 DF,  p-value: < 2.2e-16
3.2-3.3
> mdh<-lm(fev~height)
> anova(mdh,md2)
Analysis of Variance Table
Model 1: fev ~ height
Model 2: fev ~ age + height + female
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    652 120.93
2    650 110.65  2    10.286 30.213 2.84e-13 ***
3.4
> predict(md2,data.frame(age=15,height=63,female=1),interval="confidence")
       fit      lwr      upr
1 3.059155 2.973354 3.144956
3.5
> predict(md2,data.frame(age=15,height=63,female=1),interval="prediction")
       fit      lwr      upr
1 3.059155 2.244461 3.873849
```

PROBLEM SET #2 STATISTICS 500 FALL 2015: DATA PAGE 1

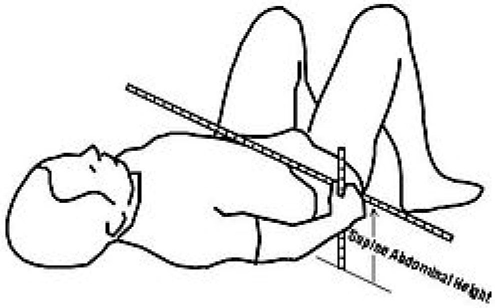
Due in class at noon in class on Tuesday 24 November 2015

This is an exam. Do not discuss it with anyone.

The data are from the 2013-2014 NHANES. The data are in an object nhanes1314st500 in the course workspace. It is also available from my web page as data.csv

```
> dim(nhanes14st500)
[1] 4576   13
> head(nhanes14st500)
    seqn lbxgh female age systolic diastolic pulse weight height  bmi waist  sad
y
1  73557  13.9      0  69      122        72    86   78.3  171.3 26.7 100.0 20.5  1.23443
06
2  73558   9.1      0  54      156        62    74   89.5  176.8 28.6 107.6 24.2  0.74297
55
```

The variables are as follows. (i) seqn is the NHANES id number, (ii) lbxgh is a measure of glycohemoglobin, recorded as a percent, where high values, above 6.5, are indicative of diabetes, (iii) female = 1 for female, 0 for male, (iv) age is in years, (v) systolic and diastolic record blood pressure, (vi) pulse is pulse, (vii) weight is recorded in kilograms, (viii) height, waist circumference and sad are recorded in centimeters. sad is sagittal abdominal diameter, the height indicated below – it is thought to be an improvement on “bmi” = body mass index as a measure of health risk.



The final variable, y, is just a transformation of lbxgh obtained as follows. For your work, you can use y as instructed without considering yjPower. The yjPower function is extends the Box-Cox transformation to permit both positive and negative values and is described in Yeo and Johnson (2000); however, there is no need to consult this paper unless you want to.

```
> library(car)
> help("yjPower")
> y<-yjPower(lbxgh-6.5,0,jacobian.adjusted = TRUE)
```

I suggest doing the following plots to understand how y relates to lbxgh. Notice that y>0 if and only if lbxgh>6.5, so both are values consistent with diabetes.

```
> par(mfrow=c(1,2))
> boxplot(lbxgh)
> boxplot(y)
> plot(lbxgh,y)
> abline(h=0)
> abline(v=6.5)
```

Yeo, In-Kwon and Johnson, Richard (2000) A new family of power transformations to improve normality or symmetry.*Biometrika*, 87, 954-959.

PROBLEM SET #2 STATISTICS 500 FALL 2015: DATA PAGE 2

Due in class at noon in class on Tuesday 24 November 2015

This is an exam. Do not discuss it with anyone.

You do not have to look at the following web pages unless you want to. The first describes lbxgh and suggests 6.5% as distinguishing diabetes. The second provides general information about NHANES.
http://www.niddk.nih.gov/health-information/health-topics/diagnostic-tests/a1c-test-diabetes/Pages/index.aspx

http://www.cdc.gov/nchs/nhanes.htm

You will be comparing sad and bmi as predictors of lbxgh, adjusting for age and gender.

Model 1: lbxgh = $\beta_0 + \beta_1$ female + β_2 age + β_3 bmi + β_4 sad + ε where ε is iid $N(0, \sigma^2)$

Model 2: y = $\gamma_0 + \gamma_1$ female + γ_2 age + γ_3 bmi + γ_4 sad + ζ where ζ is iid $N(0, \omega^2)$
Model 1 has betas and model 2 has gammas so that different things have different symbols. It does not matter which Greek letter we use.

Question 1 asks you to compare plots of studentized residuals for models 1 and 2. The best way to compare the plots is to use `par(mfrow=c(1,2))` to put two plots right next to each other. When thinking about Normal plots, it is often helpful to add the qqline, and to do the Shapiro-Wilk test, so do that here.

Note!: If a test can be done either as a t-test or as an F-test with equivalent results, do it as a t-test.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. **This is an exam**. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Last name:_____________________ First Name: _________________
ID#:___________________

Statistics 500, Problem 2, Fall 2015, p1. **This problem set is an exam. Do not discuss it with anyone**.

1. Fit models 1 and 2 from the data page. In question 1, always use the studentized residuals from `rstudent(.)`	**CIRCLE**/FILL IN THE CORRECT ANSWER
1.1 Compare the Normal plots of studentized residuals from models 1 and 2, and do the Shapiro-Wilk test on the studentized residuals. Is this true? “Neither set of studentized residuals looks Normal, but the Normal plot is closer to a line for model 2 than for model 1.”	TRUE FALSE
1.2 The Normal plot of studentized residuals from model 1 shows them to be skewed right compared to the Normal distribution.	TRUE FALSE
1.3 The Normal plot of studentized residuals from model 2 shows the residuals to have shorter tails than the Normal distribution.	TRUE FALSE
1.4 Compare the boxplots of studentized residuals from models 1 and 2. Is this true? “The studentized residuals from model 1 look more nearly symmetric about their median than the residuals from model 2.”	TRUE FALSE
1.5 If you were to test for outliers in model 1 at the two-sided 0.05 level using the Bonferonni adjustment, what is the critical value that an absolute studentized residual needs to exceed to be declared an outlier? What are the degrees of freedom? How many outliers are there?	Model 1. Critical value: ____________ Degrees of freedom: __________ Number of outliers: ___________
1.6 If you were to test for outliers in model 2 at the two-sided 0.05 level using the Bonferonni adjustment, what is the critical value that an absolute studentized residual needs to exceed to be declared an outlier? What are the degrees of freedom? How many outliers are there?	Model 2. Critical value: ____________ Degrees of freedom: __________ Number of outliers: ___________

2. Use model 2 in part 2.	
2.1 In model 2, which observation has the largest leverage? Give the row #. What is the numerical value of the leverage?	Row #__________ Leverage value:____________

2.2 The person identified in 2.1 has large leverage because, despite having a high bmi and sad, the person does not have lbxgh>6.5 or y>0.	TRUE FALSE
2.3 How many individuals have large leverage by our standard rule of thumb? Give one number.	How many?_______________
2.4 If you added 1 to the y for the person identified in question 2.1 and reran the regression with this new y, the predicted value or yhat for this person would increase by more than 0.5.	TRUE FALSE

Last name:______________________ First Name: __________________
ID#:____________________

Statistics 500, Problem 2, Fall 2015, p2. **This problem set is an exam. Do not discuss it with anyone**.

3. Question 3 asks about model 2.	**CIRCLE**/FILL IN THE CORRECT ANSWER
3.1 Which observation has the largest absolute value of dffits? Give a row number. What is the value (with its sign) of the dffits?	Row #____________ dffits:_____________
3.2 The individual in 3.1 has a large dffits because she has the lowest lbxgh and y in the data set and the highest sad.	TRUE FALSE
3.2 Using dfbetas, the individual in 3.1 is seen to move the coefficient of sad by half its standard error.	TRUE FALSE

:

4. Question 4 asks about model 2 and its extensions.	**CIRCLE**/FILL IN THE CORRECT ANSWER
4.1 In model 2, test for interaction between sad and female. What is the name of the test statistic? What is the numerical value of the test statistic? What is P-value? Is it plausible that there is no interaction between sad and female?	Name:_____________ Value: _______________ P-value:______ Circle one PLAUSIBLE NOT PLAUSIBLE
4.2 In model 2, use a **centered** quadratic term in sad to test whether the relationship with sad is linear. What is the name of the test statistic? What is the numerical value of the test statistic? What is P-value? Is it plausible that relationship between y and sad is linear?	Name:_____________ Value: _______________ P-value:______ Circle one PLAUSIBLE NOT PLAUSIBLE
4.3 Use Tukey's one-degree of freedom method to test the null hypothesis that no transformation of y is needed in model 2. What is the P-value? Is the null hypothesis plausible	P-value:______ Circle one PLAUSIBLE NOT PLAUSIBLE
4.4 The original question is whether sad is better than bmi at predicting diabetes. In the fit of model 2, it is clear that sad is not needed as predictor of y if you have bmi, age and female.	TRUE FALSE

4.5 Give the squared multiple correlation, R^2, for models 1 and 2.	Model 1 R^2:________ Model 2 R^2:________
4.6 If you plot residuals of model 2 (as y) against fitted values from model 2 (as x) and add a lowess smooth, you see a dramatic inverted-U shape.	TRUE FALSE

Answers

Statistics 500, Problem 2, Fall 2015, p1. **This problem set is an exam. Do not discuss it with anyone**.

1. Fit models 1 and 2 from the data page. In question 1, always use the studentized residuals from `rstudent(.) (5 points each)`	**CIRCLE**/FILL IN THE CORRECT ANSWER
1.1 Compare the Normal plots of studentized residuals from models 1 and 2, and do the Shapiro-Wilk test on the studentized residuals. Is this true? "Neither set of studentized residuals looks Normal, but the Normal plot is closer to a line for model 2 than for model 1."	[✓ TRUE] FALSE
1.2 The Normal plot of studentized residuals from model 1 shows them to be skewed right compared to the Normal distribution.	[✓ TRUE] FALSE
1.3 The Normal plot of studentized residuals from model 2 shows the residuals to have shorter tails than the Normal distribution.	TRUE [✓ FALSE]
1.4 Compare the boxplots of studentized residuals from models 1 and 2. Is this true? "The studentized residuals from model 1 look more nearly symmetric about their median than the residuals from model 2."	TRUE [✓ FALSE]
1.5 If you were to test for outliers in model 1 at the two-sided 0.05 level using the Bonferonni adjustment, what is the critical value that an absolute studentized residual needs to exceed to be declared an outlier? What are the degrees of freedom? How many outliers are there?	Model 1. Critical value: 4.402874 Degrees of freedom: 4570 Number of outliers: 64
1.6 If you were to test for outliers in model 2 at the two-sided 0.05 level using the Bonferonni adjustment, what is the critical value that an absolute studentized residual needs to exceed to be declared an outlier? What are the degrees of freedom? How many outliers are there?	Model 2. Critical value: 4.402874 Degrees of freedom: 4570 Number of outliers: 4

2. Use model 2 in part 2. (5 points each)	
2.1 In model 2, which observation has the largest leverage? Give the row #. What is the numerical value of the leverage?	Row #3972 Leverage value: 0.0158

2.2 The person identified in 2.1 has large leverage because, despite having a high bmi and sad, the person does not have lbxgh>6.5 or y>0.	TRUE [✓ FALSE]
2.3 How many individuals have large leverage by our standard rule of thumb? Give one number.	How many? 182
2.4 If you added 1 to the y for the person identified in question 2.1 and reran the regression with this new y, the predicted value or yhat for this person would increase by more than 0.5.	TRUE [✓ FALSE]

Answers

Statistics 500, Problem 2, Fall 2015, p2. **This problem set is an exam. Do not discuss it with anyone**.

3. Question 3 asks about model 2. (5 points each)	**CIRCLE**/FILL IN THE CORRECT ANSWER
3.1 Which observation has the largest absolute value of dffits? Give a row number. What is the value (with its sign) of the dffits?	Row #1785 dffits: -0.243
3.2 The individual in 3.1 has a large dffits because she has the lowest lbxgh and y in the data set and the highest sad.	TRUE [✓ FALSE]
3.3 Using dfbetas, the individual in 3.1 is seen to move the coefficient of sad by half its standard error.	TRUE [✓ FALSE]

:

4. Question 4 asks about model 2 and its extensions. (6 points each, except 5 for 4.5)	**CIRCLE**/FILL IN THE CORRECT ANSWER
4.1 In model 2, test for interaction between sad and female. What is the name of the test statistic? What is the numerical value of the test statistic? What is P-value? Is it plausible that there is no interaction between sad and female?	Name: t-test Value: 0.199 P-value: 0.84 Circle one [✓ PLAUSIBLE] NOT PLAUSIBLE
4.2 In model 2, use a **centered** quadratic term in sad to test whether the relationship with sad is linear. What is the name of the test statistic? What is the numerical value of the test statistic? What is P-value? Is it plausible that relationship between y and sad is linear?	Name: t-test Value: 2.562 P-value: 0.0104 Circle one PLAUSIBLE [✓ NOT PLAUSIBLE]
4.3 Use Tukey's one-degree of freedom method to test the null hypothesis that no transformation of y is needed in model 2. What is the P-value? Is the null hypothesis plausible	P-value: 0.997 Circle one [✓ PLAUSIBLE] NOT PLAUSIBLE
4.4 The original question is whether sad is better than bmi at predicting diabetes. In the fit of model 2, it is clear that sad is not needed as predictor of y if you have bmi, age and female.	TRUE [✓ FALSE]
4.5 Give the squared multiple correlation, R^2, for models 1 and 2. (5 points)	Model 1 R^2: 0.1444 Model 2 R^2: 0.2668

4.6 If you plot residuals of model 2 (as y) against fitted values from model 2 (as x) and add a lowess smooth, you see a dramatic inverted-U shape.	TRUE [✓ FALSE]

Fall 2015, Problem Set 2, Doing the Problem Set in R

Part 1.

```
> m1<-lm(lbxgh ~ age + female + sad + bmi)
> m2<-lm(y ~ age + female + sad + bmi)
> par(mfrow=c(1,2))
> boxplot(m1$residuals)
> boxplot(m2$residuals)
> qqnorm(rstudent(m1))
> qqline(rstudent(m1))
> qqnorm(rstudent(m2))
> qqline(rstudent(m2))
> shapiro.test(rstudent(m1))
        Shapiro-Wilk normality test
data:  rstudent(m1)
W = 0.63903, p-value < 2.2e-16
> shapiro.test(rstudent(m2))
        Shapiro-Wilk normality test
data:  rstudent(m2)
W = 0.97161, p-value < 2.2e-16
> boxplot(rstudent(m1))
> boxplot(rstudent(m2))
> qt(.025/4576,4570)
[1] -4.402874
> help("outlierTest")
> library(car)
> outlierTest(m1,n.max=2000)
      rstudent unadjusted p-value Bonferonni p
3292 11.826358         8.2646e-32   3.7819e-28
4313  9.537447         2.3109e-21   1.0575e-17
…
3967  4.414978         1.0335e-05   4.7294e-02
> length(outlierTest(m1,n.max=2000)$rstudent)
[1] 64
> sum(abs(rstudent(m1))>=4.402874)
[1] 64
> outlierTest(m2,n.max=2000)
      rstudent unadjusted p-value Bonferonni p
3362 -6.614237         4.1638e-11   1.9054e-07
1785 -5.760300         8.9470e-09   4.0942e-05
3988 -4.615171         4.0353e-06   1.8466e-02
898  -4.468362         8.0726e-06   3.6940e-02
> sum(abs(rstudent(m2))>=4.402874)
[1] 4
```

Part 2

```
> which.max(hatvalues(m2))
3972
3972
> nhanes14st500[3972,]
> summary(sad)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.10   19.30   22.20   22.62   25.40   40.00
> summary(bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.10   24.10   27.70   28.75   32.10   67.50
> max(hatvalues(m2))
```

```
[1] 0.0158098
> mean(hatvalues(m2))
[1] 0.001092657
> 2*mean(hatvalues(m2))
[1] 0.002185315
> 2*(5/4576)
[1] 0.002185315
> sum(hatvalues(m2)>=0.002185315)
[1] 182
```

Part 3

```
> which.max(abs(dffits(m2)))
1785
> nhanes14st500[1785,]
> dffits(m2)[1785]
-0.2430739
> dfbetas(m2)[1785,]
(Intercept)         age      female         sad         bmi
 0.10847980  0.06293416 -0.06076407  0.01112735 -0.08829110
```

Part 4

4.1

```
> fs<-female*sad
> m2a<-lm(y ~ age + female + sad + bmi + fs)
> summary(m2a)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.4558271  0.0660460 -37.184   <2e-16 ***
age          0.0139349  0.0005342  26.084   <2e-16 ***
female      -0.0092342  0.0872408  -0.106    0.916    
sad          0.0476955  0.0050535   9.438   <2e-16 ***
bmi         -0.0046524  0.0032582  -1.428    0.153    
fs           0.0007604  0.0038272   0.199    0.843    
```

4.2

```
> s2<-(sad-mean(sad))^2
> m2b<-lm(y ~ age + female + sad + bmi + s2)
> summary(m2b)
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.4394789  0.0481298 -50.685   <2e-16 ***
age          0.0140531  0.0005355  26.245   <2e-16 ***
female       0.0056009  0.0180846   0.310   0.7568    
sad          0.0471909  0.0049120   9.607   <2e-16 ***
bmi         -0.0055487  0.0032213  -1.723   0.0850 .  
s2           0.0008271  0.0003229   2.562   0.0104 *  
```

4.3

```
> summary(lm(y ~ age + female + sad + bmi+tukey1df(m2)))
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.4650399  0.0473411 -52.070   <2e-16 ***
age           0.0139385  0.0005348  26.065   <2e-16 ***
female        0.0077219  0.0180860   0.427    0.669    
sad           0.0479333  0.0049454   9.693   <2e-16 ***
bmi          -0.0045282  0.0032177  -1.407    0.159    
tukey1df(m2) -0.0004076  0.1090079  -0.004    0.997    
> plot(m2$fitted.values,m2$residuals)
> lines(lowess(m2$fitted.values,m2$residuals),col="red")
```

PROBLEM SET #3 STATISTICS 500 FALL 2015: DATA PAGE 1

Due in class at noon, Thursday, 17 Decmeber 2015

This is an exam. Do not discuss it with anyone.

For part 1, the data are the same as for Problem Set 2. These data are from the 2013-2014 NHANES. The data are in an object nhanes1314st500 in the course workspace.

```
> dim(nhanes14st500)
[1] 4576   13
> head(nhanes14st500)
    seqn lbxgh female age systolic diastolic pulse weight height  bmi waist  sad
y
1  73557  13.9      0  69      122        72    86   78.3  171.3 26.7 100.0 20.5  1.23443
06
2  73558   9.1      0  54      156        62    74   89.5  176.8 28.6 107.6 24.2  0.74297
55
```

The variables are as follows. (i) seqn is the NHANES id number, (ii) lbxgh is a measure of glycohemoglobin, recorded as a percent, where high values, above 6.5, are indicative of diabetes, (iii) female = 1 for female, 0 for male, (iv) age is in years, (v) systolic and diastolic record blood pressure, (vi) pulse is pulse, (vii) weight is recorded in kilograms, (viii) height, waist circumference and sad are recorded in centimeters. sad is sagittal abdominal diameter, the height indicated below – it is thought to be an improvement on "bmi" = body mass index as a measure of health risk.

The final variable, y, is just a transformation of lbxgh as in Problem Set 2.

Create the following x matrix.

```
> x<-nhanes14st500[,c(3,4,8,9,10,11,12)]
> head(x)
   female age weight height  bmi waist  sad
1       0  69   78.3  171.3 26.7 100.0 20.5
2       0  54   89.5  176.8 28.6 107.6 24.2
> dim(x)
[1] 4576    7
```

So you will be considering 7 predictors, female, age, weight, height, bmi, waist and sad. As in the second problem set, you will predict y, the transformed lbxgh, contained in `nhanes14st500` as its last column. You will use leaps. You should plot C_P against the size of the model and think about the plot before doing part 1.

Several questions say "C_P suggests xyz." Remember C_P is an estimate of a population quantity J_P, so C_P has some sampling error and is not equal to J_P. "C_P suggests xyz" means "if we ignore the sampling error in the estimate, pretending that J_P=C_P then xyz would be true." Example: "C_P suggests the moon is made of green cheese" means "if the true value of J_P were equal to the observed value of C_P, then the moon would be made of green cheese."

PROBLEM SET #3 STATISTICS 500 FALL 2015: DATA PAGE 2
Due in class at noon on Thursday 17 December 2015
This is an exam. Do not discuss it with anyone.

The second data set for part 2 is from NHANES 2011-2012. The data are in an object `smkdustfume` in the course workspace and are available on my web page as `data.csv`. It has 868 people in four groups of size 217, 868=4x217, recorded in the variables dfsmkf and dfsmki. Groups `dustfume` and `both` were exposed at work for at least 10 years to dusts or fumes. Groups `smoker` and `both` were daily smokers. Groups `neither` and `dustfumes` smoked fewer than 100 cigarettes in their lives. The variable `dustfumesy` is the total years of exposure to dusts and fumes, whereas `mineraly`, `organic, efumesy` and `ofumesy` record years of exposure to mineral dust, organic dust, exhaust fumes and other fumes. Other variables are `SEQN = NHANES id, age, female=1, fev1, fvc, and fvratio = fev1/fvc`, where fvratio is the Tiffeneau-Pinelli index, a measure of chronic obstructive lung disease. Lower values of `fvratio` indicate poor lung function.

You should plot the data in `boxplot(fvratio~dfsmkf)`. You should see how big the groups are in `table(dfsmkf)`. Question 2.4 asks for group names: the names are in dfsmkf.

The model for part 2, **MODEL A**, is $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where for i = 1, 2, 3, 4 groups defined by dfmskf and j = 1, 2, …, 217 people in group i, where $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$ and ε_{ij} are iid $N(0,\sigma^2)$, and y_{ij} is the fvratio for the jth person in group i.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. **This is an exam**. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

The exam is due in my office, 473 JMHH, Thursday, Dec 17, 2014 at noon. You may turn in the exam early by placing it in an envelope addressed to me and leaving it in my mail box in statistics, 4th floor, JMHH. If you prefer, give it to Noelle at the front desk in statistics. **Make and keep a photocopy of your answer page**. The answer key will be posted in the revised bulk pack on-line. You can compare your photocopy to the on-line answer page.

Have a great holiday!

Statistics 500, Problem 3, Fall 2015, p1. **This problem set is an exam. Do not discuss it with anyone**.

Use leaps and the 7 predictors in x for the nhanes14st500 data to predict y.	Fill in or CIRCLE the correct answer.
1.1 With 7 predictors, how many possible regressions can be formed as subsets of the 7 predictors? In your count, include the regression with 7 predictors and the regression with no predictors.	Number: __________________
1.2 Among regressions with just one of the seven predictors, which single predictor yields the smallest C_P? What is the "size" of this model? What is the value of C_P?	Which predictor? (variable name):____________ Size=_______ C_P = ______________
1.3 A model with fewer predictors may or may not produce better predictions than a model with more predictors. Does the value of C_P in question 1.2 suggest it produces better predictions than the best two-predictor model (as judged by C_P)?	CIRCLE ONE YES NO
1.4 Of the 7 models that use 6 of the 7 predictors, which one is worst in the opinion of C_P? Name the variable left out of this worst model. What is the "size" of this model? What is the value of C_P?	Which predictor is left out:____________ Size=_______ C_P = ______________
1.5 The value of C_P for the model in question 1.4 suggests that the total squared error of predictions from the model in 1.4 are more than 60 times larger than the total squared error of the predictions from the model with all 7 variables.	CIRCLE ONE TRUE FALSE
1.6 Give the value of C_P for the best 5 predictor model and for the best 6 predictor model. List predictors in the best 6 predictor model that are not in the best 5 predictor model. List predictors, if any, in the best 5 predictor model that are not in the best 6 predictor model; if none, write "none". "5" is short for the best regression with 5 predictors, and "6" is short for the best regression with 6 predictors.	5 predictor 6 predictor C_P= _______________ C_P= _______________ Names of predictors in: 5 but not in 6: _____________________________ 6 but not in 5: _____________________________
1.7 The 6 predictor model in question 1.6	CIRCLE ONE

is estimated to make worse predictions than the 5 predictor model in question 1.6.	TRUE FALSE
1.8 At the 0.05 level, Spjotvoll's method rejects as inadequate every model that excludes female, and every model that excludes sad, but it rejects neither of the two models in question 1.6.	CIRCLE ONE TRUE FALSE
1.9 Give the variance inflation factors for age and weight in the model with all 7 predictors.	Age vif = _________ Weight vif = ____________
1.10 Give the value of press for the 7 predictor model and the (bad) 6 predictor model in question 1.4.	7-predictor press:________ bad-6-press:_______

Due December 17, Thursday, noon. **This problem set is an exam. Do not discuss it with anyone**.

Use the smkdustfume data and MODEL A to answer questions in part 2.	Fill in or CIRCLE the correct answer
2.1 Do a one-way analysis of variance to test the null hypothesis that the four groups do not differ in their fvratios. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible?	Name:___________ Value: ________________ P-value:____________ PLAUSIBLE NOT PLAUSIBLE
2.2 Use Holm's method to compare the four groups in pairs controlling the family-wise error rate despite doing many t-tests. How many tests are done when comparing 4 groups in all possible pairs? List the pairs of groups that do **not** differ significantly at the 0.05 level by Holm's method, listing pairs of groups **by name**, so if abc does not differ from xyz, write (abc, xyz). If none, write NONE. Do not use group #s, **use group names**.	How many tests?_______________ Which pairs of groups do **not** differ? List (name1, name2) or write NONE.
2.3 Holm's method controls the familywise error rate in the weak sense but not in the strong sense, but the Bonferroni method controls in the strong sense.	TRUE FALSE
2.4 If you used the Bonferroni method, not the Holm method, the P-value comparing (both,smoker) would be 3 times as large as with Holm's method.	TRUE FALSE

2.5 In the table below, create 3 orthogonal contrasts, one representing the main effect of smoking, one representing the main effect of dust/fumes and one representing their interaction. Use integer contrast weights, not fractions or decimals.

Group	Neither	Smoker	Dustfumes	Both
Contrast:Smoking Main effect				
Contrast:Dust/Fumes Main effect				
Contrast:Interaction				

2.6 Use the contrasts in 2.5 to fill in the following detailed anova table. DF= degrees of freedom.

	Sum of squares	DF	Mean Square	F	P-value
Between Groups					

Smoking Main					
Dust/Fume Main					
Interaction					
Residual within groups				XXXXXX XXXXXX	XXXXXX XXXXXX

Statistics 500, Problem 3, Fall 2015, **ANSWERS**

Use leaps and the 7 predictors in x for the nhanes14st500 data to predict y.	Fill in or CIRCLE the correct answer. (6 points each)
1.1 With 7 predictors, how many possible regressions can be formed as subsets of the 7 predictors? In your count, include the regression with 7 predictors and the regression with no predictors.	Number: $2^7 = 128$
1.2 Among regressions with just one of the seven predictors, which single predictor yields the smallest C_P? What is the "size" of this model? What is the value of C_P?	Which predictor? (variable name): age Size=2 $C_P = 529.284$
1.3 A model with fewer predictors may or may not produce better predictions that a model with more predictors. Does the value of C_P in question 1.2 suggest it produces better predictions than the best two-predictor model (as judged by C_P)?	CIRCLE ONE YES [✓ NO]
1.4 Of the 7 models that use 6 of the 7 predictors, which one is worst in the opinion of C_P? Name the variable left out of this worst model. What is the "size" of this model? What is the value of C_P?	Which predictor is left out: age Size=7 $C_P = 503.73$
1.5 The value of C_P for the model in question 1.4 suggests that the total squared error of predictions from the model in 1.4 are more than 60 times larger than the total squared error of the predictions from the model with all 7 variables.	CIRCLE ONE [✓ TRUE] FALSE
1.6 Give the value of C_P for the best 5 predictor model and for the best 6 predictor model. List predictors in the best 6 predictor model that are not in the best 5 predictor model. List predictors, if any, in the best 5 predictor model that are not in the best 6 predictor model; if none, write "none". "5" is short for the best regression with 5 predictors, and "6" is short for the best regression with 6 predictors.	5 predictor 6 predictor C_P= 6.38 C_P= 6.07 Names of predictors in: 5 but not in 6: none 6 but not in 5: weight
1.7 The 6 predictor model in question 1.6 is estimated to make worse predictions than the 5 predictor model in question 1.6.	CIRCLE ONE TRUE [✓ FALSE]
1.8 At the 0.05 level, Spjotvoll's method rejects as inadequate every model that	CIRCLE ONE

excludes female, and every model that excludes sad, but it rejects neither of the two models in question 1.6.	[✓ TRUE] FALSE
1.9 Give the variance inflation factors for age and weight in the model with all 7 predictors.	Age vif = 1.455 Weight vif = 89.232
1.10 Give the value of press for the 7 predictor model and the (bad) 6 predictor model in question 1.4.	7-predictor: 1399 bad-6- press: 1551

Use the smkdustfume data and MODEL A to answer questions in part 2.	Fill in or CIRCLE the correct answer 2.1-4 are 6 points, 2.5 and 2.6 are 8 points
2.1 Do a one-way analysis of variance to test the null hypothesis that the four groups do not differ in their fvratios. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible?	Name: F-test Value: 22.74 P-value: 3.56 x 10^{-14} PLAUSIBLE [✓ NOT PLAUSIBLE]
2.2 Use Holm's method to compare the four groups in pairs controlling the family-wise error rate despite doing many t-tests. How many tests are done when comparing 4 groups in all possible pairs? List the pairs of groups that do **not** differ significantly at the 0.05 level by Holm's method, listing pairs of groups **by name**, so if abc does not differ from xyz, write (abs, xyz). If none, write NONE. Do not use group #s, **use group names**.	How many tests? 6 Which pairs of groups do **not** differ? List (name1, name2) or write NONE. (dustfumes, neither)
2.3 Holm's method controls the familywise error rate in the weak sense but not in the strong sense, but the Bonferroni method controls in the strong sense	TRUE [✓ FALSE]
2.4 If you used the Bonferroni method, not the Holm method, the P-value comparing (both,smoker) would be 3 times as large as with Holm's method.	[✓ TRUE] FALSE

2.5 In the table below, create 3 orthogonal contrasts, one representing the main effect of smoking, one representing the main effect of dust/fumes and one representing their interaction. Use integer contrast weights, not fractions or decimals.

Group	Neither	Smoker	Dustfumes	Both
Contrast:Smoking Main effect	-1	1	-1	1
Contrast:Dust/Fumes Main effect	-1	-1	1	1
Contrast:Interaction	1	-1	-1	1

2.6 Use the contrasts in 2.5 to fill in the following detailed anova table. DF= degrees of freedom.

	Sum of squares	DF	Mean Square	F	P-value
Between Groups	0.4665	3	0.1555	22.74	3.56 x 10^{-14}
Smoking Main	0.4126	1	0.4126	60.34	2.25 x 10^{-14}
Dust/Fume Main	0.0316	1	0.0316	4.63	0.03177

Interaction	0.0222	1	0.0222	3.25	0.07184
Residual within groups	5.9083	864	0.0068	XXXXXX XXXXXX	XXXXXX XXXXXX

PROBLEM SET #3 STATISTICS 500 FALL 2015 ANSWERS
DOING THE PROBLEM SET IN R

1.1

```
> 2^7
[1] 12
> attach(nhanes14st500)
> x<-nhanes14st500[,c(3,4,8,9,10,11,12)]
> nhl<-leaps(x=x,y=y,names=colnames(x))
> plot(nhleaps$size,nhleaps$Cp)
> abline(0,1)
> plot(nhl$size,nhl$Cp)
> abline(0,1)
> cbind(nhl$which,nhl$size,nhl$Cp)[35:55,]
  female age weight height bmi waist sad
4      0   1      0      0   1     1   1 5  46.273615
4      1   1      0      0   0     1   1 5  46.286152
4      1   1      0      0   1     0   1 5  51.533812
5      1   1      0      1   1     0   1 6   6.375960
5      1   1      1      1   0     0   1 6   9.553443
5      1   1      0      1   0     1   1 6  12.083396
…
5      1   1      0      0   1     1   1 6  48.271690
6      1   1      1      1   1     0   1 7   6.066912
6      1   1      0      1   1     1   1 7   8.375293
6      1   1      1      1   0     1   1 7  11.390767
6      1   1      1      0   1     1   1 7  19.194041
6      0   1      1      1   1     1   1 7  21.556761
6      1   1      1      1   1     1   0 7  90.768968
6      1   0      1      1   1     1   1 7 503.726589
7      1   1      1      1   1     1   1 8   8.000000
```

```
> nhsp<-spjotvoll(x,y)
> nhsp[!nhsp$inadequate,]
    p     Cp    Fp  pval adjusted.pval inadequate female age weight height bmi waist sad
99  6  6.376 1.188 0.305         0.305      FALSE      1   1      0      1   1     0   1
100 6  9.553 2.777 0.062         0.062      FALSE      1   1      1      1   0     0   1
120 7  6.067 0.067 0.796         0.796      FALSE      1   1      1      1   1     0   1
121 7  8.375 2.375 0.123         0.305      FALSE      1   1      0      1   1     1   1
122 7 11.391 5.391 0.020         0.062      FALSE      1   1      1      1   0     1   1
127 8  8.000    NA 1.000         1.000      FALSE      1   1      1      1   1     1   1
```

```
> library(DAAG)
> m<-lm(y~female+age+weight+height+bmi+waist+sad)
> vif(m)
 female     age  weight  height     bmi   waist     sad
 1.9853  1.4553 89.2320 20.8960 70.8950 14.7730 11.1800
> m<-lm(y~female+age+weight+height+bmi+waist+sad)
> press(m)
[1] 1399.262
> m2<-lm(y~female+weight+height+bmi+waist+sad)
> press(m2)
[1] 1551.129
```

PROBLEM SET #3 STATISTICS 500 FALL 2015 ANSWERS
DOING THE PROBLEM SET IN R

```
2.
> attach(smkdustfume)
> boxplot(fvratio~dfsmkf)
> table(dfsmkf)
2.1
> anova(lm(fvratio~dfsmkf))
Analysis of Variance Table
Response: fvratio
           Df Sum Sq  Mean Sq F value    Pr(>F)
dfsmkf      3 0.4665 0.155499  22.739 3.559e-14 ***
Residuals 864 5.9083 0.006838

> pairwise.t.test(fvratio,dfsmkf)
        Pairwise comparisons using t tests with pooled SD
          both    dustfumes neither
dustfumes 1.2e-10 -         -
neither   2.8e-11 0.80542   -
smoker    0.01060 0.00023   0.00011
P value adjustment method: holm

> pairwise.t.test(fvratio,dfsmkf,p.adj="bonf")
        Pairwise comparisons using t tests with pooled SD
          both    dustfumes neither
dustfumes 1.5e-10 -         -
neither   2.8e-11 1.00000   -
smoker    0.03180 0.00046   0.00016
P value adjustment method: Bonferroni
0.03180 = 3*0.01060

> contrasts(dfsmkf)
> smk<-c(1,-1,-1,1)
> dufu<-c(1,1,-1,-1)
> interact<-smk*dufu
> contrasts(dfsmkf)<-cbind(smk,dufu,interact)
> contrasts(dfsmkf)
          smk dufu interact
both        1    1        1
dustfumes  -1    1       -1
neither    -1   -1        1
smoker      1   -1       -1
> m<-model.matrix(lm(fvratio~dfsmkf))
> head(m)
> cor(m[,2:4])
> m<-as.data.frame(m)
> summary(lm(fvratio~m$dfsmkfsmk+m$dfsmkfdufu+m$dfsmkfinteract))
> anova(lm(fvratio~m$dfsmkfsmk+m$dfsmkfdufu+m$dfsmkfinteract))
Analysis of Variance Table
Response: fvratio
                  Df Sum Sq Mean Sq F value    Pr(>F)
m$dfsmkfsmk        1 0.4126 0.41265 60.3442 2.254e-14 ***
m$dfsmkfdufu       1 0.0316 0.03163  4.6260   0.03177 *
m$dfsmkfinteract   1 0.0222 0.02221  3.2483   0.07184 .
Residuals        864 5.9083 0.00684
```