Statistics 722, Spring 2017
Predictive Analytics for Business

Professor Richard Waterman, waterman@wharton.upenn.edu. 443 Huntsman Hall

TA: Sameer Deshpande, dsameer@wharton.upenn.edu. 434 Huntsman Hall.

Classes meet: Mon/Wed 9:00am – 10:30am and 10:30am – 12:00pm, Room 351 SH-DH.

Overview

This seven-week course introduces students to statistical techniques that extend the ideas of regression introduced in Statistics 102 and Statistics 613. These extensions include methods for automatically building regression models from large collections of predictors, modeling categorical responses for classification, and important tools from data mining (trees, forests and nets). The course focuses on the practical use of modern methodologies that are often associated with data mining and machine learning. The treatment of these methods in Statistics 722 adds both breadth (e.g., logistic regression) and depth (e.g., model selection) to the topics covered in Statistics 102/613.

The course begins with regression, but from the point of view of predictive modeling using automatic tools. This topic allows some review of the foundations of regression, but the emphasis lies in new areas. The name of the game in predictive modeling is to be able to predict the behavior of new data. If, for example, I can show a bank how to predict who will default on a loan better than their existing system, the bank can make a lot more profit. If I help a company identify those in the market most interested in its products, then it can construct a much more focused product launch.

Rather than think of regression as a tool for modeling how one or two carefully chosen measurements are related to a response, we'll consider how to search for predictors hidden in large databases. What's large? Research models now handle 100,000 predictors. We won't get quite so far, but we'll work with some large datasets with many predictors.

Moving beyond ordinary regression, the course turns to methodologies that extend the regression model itself. Some of these methodologies expand the form of the predictive model. Generalized linear models, in particular logistic regression, allow one to apply the ideas of regression to the analysis of categorical responses. These models are commonly used to study buyer behavior, identifying how various factors affect the purchase decision. Classification trees (CART) models present a more radical departure from regression. These models organize the data into subsets that behave similarly. Random forests are ensembles of trees and are one of the most popular data mining tools available. Neural networks have re-asserted themselves in the last few years and provide a very powerful alternative to classical methods, especially for models with extremely complicated non-linear structure.

Audience

This course presumes that students are familiar with the inferential methods covered in Statistics 102/613 (including hypothesis tests, confidence intervals, p-values, use and interpretation of least squares regression). The course also presumes familiarity with the software program JMP or at the least, a willingness to learn it.

Materials

Lecture notes and topical papers available via Canvas.

Software

The software used in the course is JMP Pro, version **13**. This software is available on the Wharton Network and for purchase from the Computer Connection at the Bookstore. Here's the current information I have about the purchase process (which will be updated if the process changes):

"Students will have to physically go to the Book store to get a copy. When they purchase it from the store it they will then have to wait for the Software Licensing team to email them back with the proper license file."

Books

As for Stat 613, Stine and Foster, Statistics for Business (SfB), Third Edition. ISBN 978-0134497167.

JMP manuals, available from the JMP help menu. SAS Institute Inc. 2017. *JMP® 13 Fitting Linear Models*, *Second Edition*. Cary, NC: SAS Institute Inc. and *JMP® 13 Predictive and Specialized Modeling*, *Second Edition*. Cary, NC: SAS Institute Inc.

Optional: An Introduction to Statistical Learning (available for download). Gareth James, et. al. 2013. ISBN 978-1-4614-7137-0. This book is more mathematically focused than we will be in the course, but it is still a useful resource to have on hand. There is a version available at: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf

Grading

Homeworks: There will be four homeworks in the Quarter. These consist of hands-on data analyses and are intended to maintain the pace of the course and develop familiarity with the relevant methods.

Final Exam: Closed-book, in-class, multiple-choice format.

Course grade: 50% homeworks (10, 10, 10 and 20), 50% final exam.

Assignment due dates

A1: 1/24/2018

A2: 2/5/2018

A3: 2/14/2018

A4: 3/4/2018

Annotated Course Outline and Readings

I. Review of multiple regression

Lecture 1.

Coincidences happen in life, particularly as instant news brings us events from all around the world. Was that an accident, or does it imply that something deeper is going on? Coming to the wrong conclusions can be expensive or dangerous. Coincidences happen in modeling as well, with equally troubling consequences. In this class we will take the opportunity to quickly review multiple regression and the associated hypothesis testing framework, setting up some of the issues such as *multiplicity* that we will explore later in the course.

Review multiple regression, SfB, Chapters 23, 24, and 25.

JMP software for multiple regression, FLM. Chapters 2, 3 and 4, **Fitting Standard Least Squares Models**.

II. Regression modeling

Lecture 2. Stepwise regression

Regression gets pretty hard when you have a lot of predictors to consider. The careful, thoughtful process motivated by substance works with small data sets, but can miss quite a few things when you start to run out of time. The modeling process seems quite iterative, so why not let a computer take over the routine? We will introduce some new model selection criteria, including AIC and BIC and discuss their goals and environments where one might be preferred over the other.

Review stepwise regression, SfB, Statistics in Action, Automated modeling, p. 767.

JMP software: FLM, Chapter 5. **Stepwise regression models.**

Lecture 3. Resampling techniques: the bootstrap and cross-validation

One of the key activities associated with all statistical analyses is the identification of a measure of uncertainty, typically achieved through the calculation of a standard error and associated confidence interval. Unfortunately, for many statistical features of interest there aren't simple expressions for these standard errors and not all statistics follow a normal distribution (especially in small samples). The Bootstrap is a technique that allows for the creation of standard errors even in difficult problems, but it does so in a non-formulaic way. It does this by *resampling* the original data set, and uses this computationally intensive approach to tackle the problem.

The idea of using the same dataset many times is at the heart of cross-validation too. Cross-validation is very useful for model selection. In particular, stepwise regression, like any greedy routine, can be fooled if you give it too many choices. It's the "EverReady bunny": the model grows bigger and bigger. But, its predictions get worse and worse. The solution is figuring out how to unplug it before it goes too far.

Saving some data for later can be useful. But, do you really have enough to save just to see how well your model is doing? Often, setting aside data is the only way to know the predictive quality of your model. That's what cross-validation is all about: build a model on one part of the data, and then test it on another.

JMP Software: FLM, Chapter 5. **Stepwise regression models.**

Lecture 4. More on cross-validation

We will continue out discussion of cross-validation and see that when models are chosen by cross-validation it can help to have a third portion of the dataset withheld, in order to obtain a realistic estimate of out-ot-sample model performance.

We will also see a cross-validated R^2 used as a stopping criteria for step-wise regression.

III. Regression models for classification

Lecture 5. Classification using regression: discriminant analysis

If you use a two-level categorical variable as the response in regression (i.e., Y is a 0/1 dummy variable), then you're doing discriminant analysis. The connection gets a little weaker when you have more than two groups, but the modeling works much the same. You can do it all with good old regression, but it's a lot easier if the software takes on the burden for you.

JMP software: FLM, Chapter 10. Performing Logistic Regression on Nominal and Ordinal Responses

Lecture 6. Logistic regression and maximum likelihood

If you try to use least squares regression to classify cases (such as distinguishing buyers from browsers, honest borrowers from cheats), the predictions can be rather silly and the assumptions of discriminant analysis can be too unrealistic. Calibration can help, but it's often better to start with an approach that must give a sensible answer. That's logistic regression. This first class about logistic regression sticks to the case with one predictor.

JMP software: FLM, Chapter 11. **Logistic Regression Models**

Lecture 7. Multiple logistic regression

Logistic regression can use more than one predictor, just like regular regression. But once we allow several predictors, how are we supposed to decide which? If we remember that logistic regression is basically like a linear regression, we can exploit ideas from stepwise regression.

Lecture 8. Actionable summaries for classification techniques

How should you describe and present the results of a logistic regression? Creating quintiles and deciles in terms of the predicted probability of response, can be very helpful. We will also discuss the Area under Curve (AuC) and the Lift Chart as ways of quantifying the value of a predictive model for a categorical response. Though we will first see them in the context of logistic regression, the AuC and Lift Chart apply equally as well to the other classification techniques we will discuss later in the course.

JMP software: FLM, Chapter 11. **Logistic Regression Models**

IV. Classification and regression trees (a.k.a., CART)

Lecture 9. Introduction to tree-based models

A regression model is an equation. A CART model is a tree. How's a tree work? How does it compare to using an equation? How do you pick the "best" tree? These are the topics for today, sticking to problems with a single predictor.

JMP software: SM, Chapter 5. **Partition Models**

Lecture 10. Classification trees

Now for some real tree-based models which explore the effects of many possible predictors.

If you like CART modeling and want more details than are in the class notes, then have a look at the book Classification and Regression Trees by Breiman, Friedman, and Olshen. It's a lucid classic. The text is not required, but you can find it from Amazon and elsewhere.

JMP software: SM, Chapter 5. **Partition Models**

Lecture 11. Regression trees. Bonferroni, calibration and missing data

You can use trees to model a continuous response, and the results may be better than a regression. Bonferroni is a technique that is used to manage the false positive rate in multiple-testing and is the first-line of defense in the battle against multiplicity. Calibration ensures that when you create a predictive model, nothing is “left on the table”. And missing data happens all the time, and we will present an approach, that is reasonable in the predictive modeling context.

JMP software: SM, Chapter 5. **Partition Models**

V. Machine learning algorithms

Lecture 12. Random forests and neural networks

This last class will introduce two of the most popular data mining tools available today. These are the *random forest* and *neural network*. Random forests are ensembles of trees and can provide a very reliable prediction tool, albeit one that is hard to interpret. Neural networks are essentially massive non-linear regression models. They can be very useful when there is a complex and non-linear structure to the underlying data generation mechanism. Both of these methodologies involve multiple “tuning parameters”, and we will discuss ways to identify optimal sets of these parameters.

JMP software: SM, Chapter 6. **Bootstrap forest**, Chapter 4 **Neural networks**

Lecture 13. Review