Statistics 474/974 Criminology 474/974
Modern Regression for the Social, Behavioral and Biological Sciences

Section 1 (Stat 474/974) Monday and Wednesday, 9:00 to 10:30
VANP (active learning format)
Section 2 (Crim 474/974) Tuesday and Thursday, 9:00 to 10:30
VANP 113 (active learning format)

Professor Berk, 564 McNeil Hall, berkr@sas.upenn.edu

October 20, 2019

Conventional regression analysis within the generalized linear model is usually premised on a causal model responsible for the data, typically of a parametric form. This approach has dominated the social, biomedical, and environmental sciences since the 1970s. Because it rests on so many untestable assumptions, the causal modeling formulation was never universally accepted by statisticians and computer scientists. Over the years, there have been increasingly skeptical assessments from social scientists and econometricians as well. Ed Leamer's early paper "Let's Take the Con Out of Econometrics" is still a good read and on target.

Over the past two decades a broader perspective has been developed that seems more appropriate for the way data are actually analyzed. This broader perspective has the following features.

1. The data generation process is independent realizations from a joint probability distribution of predictors and one or more response variables.
2. Causal mechanisms can be introduced, but are formally unnecessary.
3. The model can be parametric, semiparametric, or nonparametric.
4. The modeling process can be highly inductive.
5. Unbiased estimates are no longer the holy grail. Extensions to some forms of machine learning can follow naturally (and will in this course).

Statistics 474/974 and Criminology 474/974 adopt this modern perspective. They emphasize intuitive understanding and applications. Proofs are introduced only as absolutely necessary. The target audience is juniors and seniors in the

social, behavioral, and biological sciences or graduate students from those disciplines. Prerequisites are at a minimum Statistics 111 or 101 and 112 or 102 (or the equivalent, not AP classes). Good control of linear regression is absolutely essential as background. Better still is good control over the generalized linear model. A background in linear algebra is helpful but not essential.

Grading will be based on up to four short research papers in which a serious data analysis is required. These analyses will be undertaken using the statistical programming language R. R is free, runs on all major platforms and can be downloaded from www.r-project.org/. Free documentation can also be downloaded from that site. A working knowledge of R is assumed. Most of the more advanced statistical procedures covered in the course are not available in conventional statistical packages such as SPSS, STATA or JMP. Some of the procedures can be found in libraries for Python (e.g. sklearn), but the coverage is spotty. It will prove handy to implement R from within RStudio. The single user "Desktop" version is also free: https://rstudio.com/.

The text is *Statistical Learning from a Regression Perspective,* third edition (Richard Berk, Springer Series in Statistics, 2019). **Don't get the first or second edition**. They are badly dated. If the third edition of the text is unavailable at the beginning of the semester, free drafts of the first chapter or two will be provided electronically.

Office hours are by appointment. It is usually possible to find a mutually convenient time within 24 hours of a request. When it is practical, e-mail will usually lead to a faster turnaround than waiting for a face-to-face meeting.

An important heads up. This is a course in data analysis, or what some now call "analytics" or "data science." Within the limits of what can be done in a classroom setting, the intent is to provide hands-on experience with real world data analysis problems. In the real world, there is rarely any task that closely approximates a midterm or final exam, let alone a standardized test. Yet, most students come to this course fine-tuned for this form of evaluation, and some find data analysis very challenging. Moreover, prowess in mathematical statistics or computer science certainly can help, but some students experience the transition from formal proofs to the analysis of real data quite daunting. By and large, Penn students do well in this course, but this is not a conventional course either in statistical content or the means by which each student's work is evaluated. The task is to make sense of data.

A final heads up. Both sections will meet an active learning classroom. The content is the same in both sections and "counts" the same toward various requirements. Working in groups is encouraged so that students can learn from one another. Students will be expected to carefully read the text **outside of class** and then do hands-on applications in class. There will be very little lecturing. This format has worked extremely well for some students but not well for others. The active learning format means that students have to **really** read the text outside of class and then be on task when class meets. Students who take control of their education thrive under these arrangements. Passive learners may struggle.