STATISTICS 474/974
MODERN REGRESSION FOR THE SOCIAL, BEHAVIORAL AND BIOLOGICAL SCIENCES

PROFESSOR BERK, 564 MCNEIL HALL, BERKR@WHARTON.UPENN.EDU

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 widely popular among statisticians and computer scientists. (See, for example, work by statisticians such as Rubin, Holland, Rosenbaum, Freedman, and Sacks). But over the years there have been increasingly skeptical assessments from social scientists and econometricians as well (e.g., Leamer, Duncan, Berk, Winship, Manski, Heckman, Imbens, Angrist). Ed Leamer's early paper "Let's Take the Con Out of Econometrics" is still a good read and on target.

The good news is that 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 a realization from a joint 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) A form of penalized least squares or penalized maximum likelihood can be used for estimation so that one arrives at a regularized result trading off bias against variance. Unbiased estimates are no longer the holy grail. Extensions to some forms of machine learning can follow naturally.

Statistics 474 examines this modern perspective. The emphasis will be on intuitive understanding and applications. Proofs will be 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). A background in linear algebra is helpful but not essential.

Grading will be based on 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. Most of the more advanced statistical procedures covered in the course are not available in conventional statistical packages such as SPSS, STATA or JMP.

The text is *Statistical Learning from a Regression Perspective* (Richard Berk, Springer Series in Statistics, 2008). The plan to work through the text at a pace of about 4 lectures for each chapter. But the pace can be accelerated or decelerated as needed.

Office hours are by appointment. It is usually possible to find a 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.