Lecture Topics for 36-490
Spring 2010

Following are the lecture topics we have planned for this semester, *very roughly* in the order we will present them. We have about 10 80-minute slots for lectures. In order to fit some topics in before your due dates for presentations or draft papers, we may do a couple of lectures on Wednesdays, but we will try to avoid this so we have plenty of time for individual team meetings.

- **Statistical Consulting.** The practice of statistical consulting in an academic or industry setting presents its own challenges beyond mastery of the technical analytic, skills and obtaining sufficient understanding of the problem at hand and its domain. These include communication, project management, and professional ethics. We introduced these ideas on the first day of classes, and we plan to return to them at least once more during the semester.

- **Principal Components and Factor Analysis.** Principal Components Analysis (PCA) is a *data reduction method*—it reduces a set of p variables to only $k < p$ variables, selecting linear combinations of the original variables that maximize variability among the observations. Factor Analysis (FA) is a closely related statistical model that represents each observable variable as a linear combination of a smaller set of unobservable (or "latent") variables, plus an error term (as in linear regression). We can use FA as a data reduction method like PCA, or we can use FA to derive statistical estimates of the latent variables, which may be of interest in their own right, and their influence the observed variables.

- **Mixed effects linear models.** Some of the projects this semester involve "clustering" and/or multiple observations on the same individual. In these situations there can be dependence between the observations, so it can be better to view the observations on the same individual (or from the same "cluster") as random draws in their own private experiment. This reasoning leads (roughly) to the idea of *mixed effects models*, in which some parameters are really parameters that you estimate (these are called "fixed effects"), and others are treated as draws from some distribution (these are called "random effects").

- **Missing data.** When we ran into missing data (NA's) in 36-401, our usual approach was to (1) delete the cases with missing data; and (2) try the analysis again with various filled-in values for the missing data, to see if the answers are any different from (1) (this is an example of "sensitivity analysis"). What if there is too much missing data to effectively do step (2), or what if you suspect there is useful information in the missing data (e.g., one type of person is much more likely to refuse to answer a question than another type of person; or one type of astronomical object is more likely to be outside the telescope aperture than another)? We will talk about assumptions

under which it's OK to just delete cases with missing data; and assumptions that help us decide what values to fill in (called "imputation") when it's not OK to delete.

- **Clustering.** Sometimes the purpose of data analysis is to group individuals or objects into subsets where items within each subset are somehow more closely related to each other than items from different subsets. There are a variety of ways of assessing "closeness" and different algorithms for obtaining definitions of the groups.

- **Grade of Membership and Latent Dirichlet Allocation Models.** Grade of Membership (GoM) and Latent Dirichlet Allocation (LDA) models are a family of statistical models that are closely related to clustering methods. For example if we try to cluster newspaper articles into "news reporting", "news analysis" and "opinion pieces" we might find that articles don't fit neatly into these three groups; rather many articles are a mixture of these "pure types". GoM and LDA models allow us to fully specify what we mean by "a mixture of types" and to estimate how close each article would be to each pure type.

- **Log-linear models and logistic regression.** Using modifications of linear regression to predict discrete data.

- **Poisson Process and other models for events over time.** We know the Poisson distribution as a model for the number of independent events that can happen at a fixed rate in a fixed time. What happens if the time interval is not fixed, or the rate is not fixed? There is a family of statistical models for this kind of data. A generalization of these models are the *birth and death* models that are used in queueing theory and other areas.

- **Hidden Markov models.** Sometimes the "missing data" is the true state of a system which is evolving over time, though we get some noisy and imperfect measurement of the changing state as well. Hidden Markov models (a.k.a. "state-space models") are an important way of dealing with such time series.

- **Writing and speaking.** Organizing your reports and oral presentations, and writing and speaking clearly and simply, will go a long way toward making your work easy to understand, and making it easy for us to give you high grades in the class.