36-401 Class Notes for 8/29/00

Howard Seltman

Course Overview

- Important principles
  - Start and end data analyses with the subject matter, not the math or statistics.
  - Clarify the objects (cases, subjects, patients, plots, lots, etc.) being measured, the conditions of measurement, the identity of the outcome (a.k.a. response or dependent) variables, the identity of the explanatory (a.k.a. independent or predictor) variables, the units of measurement, the variable types (categorical vs. ordinal vs. continuous), the (marginal) distributions of the variables, and the relationships (e.g. correlation) between the explanatory variables.
  - Perform exploratory graphics.
  - Clearly explain what you did and what your conclusions are. If you understand what is going on, but can't communicate it, there is little benefit to the analysis, and little job security for you!
  - A (good) picture is worth a thousand words. Well annotated graphs go a long way towards communicating about your analyses.
- Statistical models
  - A statistical model is a simplified description of a part of the world that we are interested in studying and understanding. It generally has two parts: the "structural" part of the model describes the relationship between explanatory variables and, e.g. the mean of the outcome variables, while the "measurement" part of the model describes the statistical nature of the deviations from the mean.
  - "All models are wrong, but some are useful." George E.P. Box
- Where does regression fit among statistical techniques?

	ANOVA	**Regression**	**Contingency Tables**
Outcome variable			
Explanatory variables			
Null hypotheses			
Measurement model			

Categorical/continuous:
Non-normal:
Multiple outcomes:
Time to failure with censoring:

Non-linearizable:
Correlated observations:
Other goals, e.g. classification:

- Overview of linear regression

  If we have measured a continuous response (outcome) variable and one or more continuous and/or categorical explanatory variables on appropriate scales, then the regression model states that the average outcome at any fixed combination of explanatory variables can be expressed as a linear combination of the explanatory variables and a set of unknown constants called "coefficients". The variation about the mean, for randomly chosen cases with a fixed combination of explanatory variables is Gaussian.

  E.g. Units are randomly selected corn fields in N.J. Y=crop yield in bushels per acre, W=rainfall+irrigation in inches, F=fertilizer added in pounds N/acre.

  One model is

$$Y = \beta_0 + \beta_1 W + \beta_2 F + \beta_{12} WF + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

  Estimates of $\beta_0, \beta_1, \beta_2$, and β_{12} can be obtained by maximum likelihood or the method of least squares. Standard errors come from studying the distributional properties of the estimators. Predictions with confidence intervals can then be calculated. P-values can be used to test $\beta_i = 0$.

  Important assumptions are that the covariates are fixed (not random), the units are independent of each other, the structural model is formulated correctly, the errors follow a normal distribution, and the variance of the errors is constant across all values of the covariates.

Breakout Session

Hartford, Connecticut BioBlitz
Smithsonian Magazine, April 2000

1. A BioBlitz is an event in which dozens of scientists fan out across some unlikely habitat, hell-bent on recording every species they can find, dead or alive, in a 24-hour period. The scientists launching themselves into Hartford's 695-acre Keney Park at 3 p.m. that Friday were armed with spotting scopes, sweep nets, pit traps, scalpels and fish stunners. They were prepared to dance like butterflies, sing like chickadees or do almost anything else this scavenger hunt required. One collector was seeking someone to donate bait for his dung beetle trap. Another was erecting what looked like a device for extracting sunbeams from cucumbers.

   It consisted of a big galvanized-steel funnel mounted on a tripod, and it was actually meant to extract animals from soil samples. With a coffee can, a slight, white-bearded man named Carl Rettenmeyer cut out a disk of grassy soil and leaf litter "about the size of a quarter-pounder, with lettuce." This sample would sit overnight on a screen in mid-funnel, with an electric light overhead to make things uncomfortable, and a killing jar below to collect whatever came creeping out.

   By Saturday afternoon, the "quarter-pounder" of lawn alone would yield 23 separate species, and 89 individual animals, including mites, thrips and an awful lot of springtails, "little bitty things smaller than the head of a pin." The springtails run around with their tails tucked under their abdomens, spring-loaded so that, in case of danger, they can execute a "we-outta-here" backflip, six inches into the air. It is a skill two-legged city-dwellers can only envy.

   "The area of the sample is 11.044639 square inches," Rettenmeyer said. "Now multiply that by..." Then he came up with his best estimate: that there are 50 million individuals in just the top inch of soil in an ordinary acre here, and 35 billion animals in Keney Park - nearly six for every man, woman and child on earth.

   What's wrong with this article excerpt?

2. Chemical analysis

   Solutions were created by adding 0, 25, 50, 75 and 100 μg of the antibiotic adacillin to 1 mL of serum from patients not taking the drug. The samples were exposed for a fixed time to a reagent that reacts with adacillin to produce a colored product. The light absorption was measured on duplicate samples at each concentration with the following results: absorbance= 0.20, 0.22, 0.31, 0.31, 0.42, 0.44, 0.49, 0.50, 0.59, 0.61, respectively.

   Two patients on the drug had absorbances of 0.35 and 0.92. What concentrations of drug do the patients have? Comment on the reliability of your results.

3. "Long-term effects of sulphur dioxide on crops: an analysis of dose-response relations", T.M. Roberts, Phil. Trans. T. Soc. Lond., B305, 299-316, 1984.

   Towards what audience is the Abstract aimed? What are some good and bad points about the writing?

   What are the units measured, outcome variable, and explanatory variables?

What does the table communicate? Is it done well?

What do the graphs communicate? Comment on the clarity of the graphical presentations.

What are the good and bad points of the Conclusions section?