Introduction

Applied Regression and Other Multivariable Methods
Chapters 1 and 2

What is Statistics?

- Science of learning from a set of information (data)
- Attempts to discern underlying features in the data
- Encompasses the following three tasks
  - Collecting the information

    Commonly classified into study variables

    How to measure each study variable / # of replicates?

    Experimental vs observational study?
  - Describing the information

    Graphical and numerical summaries of Stat 501
  - Interpreting/generalizing the information

    Interest in assessing relationships among variables

    Interest in formal testing of hypotheses

    Will focus on regression/ANOVA analytic techniques

Classification of Variables

- Choice of analysis often depends on classification
- Variables classified as
  - **Nominal/Qualitative -** variable classified into one of several groups or categories

    Ordinal - Letter grade, State of disease

    Non-ordinal - Eye color, Blood type
  - **Interval/Quantitative -** variable takes on numerical values for which arithmetic operations make sense

    Continuous - Height, Serum creatinine level

    Discrete - Number of defects, Manatee count

    Choice based on presence of a gap between values

    Ratio-scale variable has scale with true zero
- Can convert ordinal to discrete quantitative

  Example: State of Disease : 0 - absent, 1 - present

Classification of Variables

- Variables also classified based on orientation
  - **Response or dependent** variable

    To be described in terms of other study variables

    May vary study objective to study objective
  - **Predictor or independent** variables

    Study variables of interest experimentally

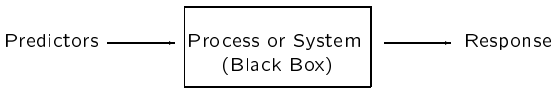
    Used to describe response variable
  - **Nuisance or control** variables

    Variables that may affect response

    Not of interest experimentally

    Sometimes called covariates or confounders

Diagram of a Research Problem



- Have process quantified by specific response
- Uncertainty in system response to fixed predictors
- Uncertainty due to nuisance factors/inherent noise
- Interested in studying process
  - Which predictors are correlated with process response?
  - Which predictors affect process response?
  - What level of predictors for specific response?
  - What predictor combination results in low uncertainty?
- Will use linear model to describe process

Observational vs Experimental Study

An **observational study** is the easiest to run but most difficult to draw definitive conclusions (cause/effect). There is no control of the predictor variables.

An **experimental study** combines strategies of running an experiment with statistical tools for decision making. Have control of the predictor variables.

- Develop empirical **regression model** to explain process
- Plan experiment to obtain definitive conclusions from model

Features of experiment to consider

- Statement of problem
- What response will be used?
- What change in response is important?
- What predictors to study?
- How many observations to be taken?
- What are resources? costs?
- Which inputs are most important?
- Are there nuisance variables?
- Block on control variables?
- What is experimental unit?

Examples of Experiments

Exp 1 To study the effects of pesticides on a bird's calcium content, an experimenter randomly (and equally) allocated sixty-five chicks to five diets (a control and four with a different pesticide included). After a month each chick's calcium content (mg) in one cm length of bone was measured.

Exp 2 A psychologist is interested in studying the mean and variance in IQs of 1st grade children from the low income areas of several major cities. Six grade schools were randomly chosen (from the low income areas) and from each of these schools, five 1st grade children were randomly chosen and had their IQs measured.

Exp 3 Brewer's malt is produced from germinating barley. The following is an experiment to determine the best conditions to germinate the barley. A total of thirty lots of barley seeds (100 seeds per lot) were equally and randomly assigned to ten germination conditions. The conditions are combinations of the week after harvest (1, 3, 6, 9, or 12 weeks) and the amount of water used in the process (4 ml or 8 ml).