Stat 414 – Day 1
Review of General Linear Model

Review: Least Squares Regression

(a) What is the statistical model for simple linear regression?

(b) What does “least squares” mean?

(c) What are the assumptions necessary for valid inference (confidence intervals and p-values) about the parameters of the model?

(d) What issues can arise if the assumptions in (c) are not met?

(e) How does one decide whether these assumptions are met?

(f) What can be done if the assumptions in (c) are not met?

The basic regression model is a special case of the **General Linear Model** (which includes categorical predictors = ANOVA, multivariate regression, ANCOVA), which is a special case of the **Generalized Linear Model** (typically what is meant by "GLM"): distribution function, link function, linear predictor (e.g., logistic regression).

Example 1: For each scenario below, identify the response and the explanatory variables and then consider each of the LINE assumptions in the context of the study, commenting on possible problems with the assumptions.
(a) Francis Galton suspected that a son's height could be predicted using the father's height. He collected observations on heights of fathers and their firstborn sons.

(b) Is the time spent studying predictive of success on an exam? The time spent studying for an exam, in hours, and success, measured as Pass or Fail, are recorded for randomly selected students.

(c) A researcher suspects that loud music can affect how quickly drivers react. She randomly selects drivers to drive the same stretch of road with varying levels of music volume. Stopping distances for each driver are measured along with the decibel level of the music on their car radio.

(d) Do wealthy families tend to have fewer children compared to lower income families? Annual income and family size are recorded for a random sample of families.

(e) The yield of wheat per acre for the month of July is thought to be related to the rainfall. A researcher randomly selects acres of wheat and records the rainfall and bushels of wheat per acre.

(f) Investigators collected the weight, sex, and amount of exercise for a random sample of college students.

(g) Medical researchers investigated the outcome of a particular surgery for patients with comparable stages of disease but different ages. The ten hospitals in the study had at least two surgeons performing the surgery of interest. Patients were randomly selected for each surgeon at each hospital. The surgery outcome was recorded on a scale of one to ten.

Example 2: The **Election2004Data.xlsx** datafile contains demographic variables for the year 2000 for 3,115 U.S. counties across 49 states (no Alaska), as well as the county-level election results in 2004 (proportion voting for Bush).

(a) Copy and paste columns F-O into R using the Rscript from the Lecture Notes page.

(b) What are the observational units in this study?

(c) Examine the histogram of the percentage of voters who voted for George W. Bush. Is this distribution approximately normal? Is it required to be? Does anything surprise you about this distribution? What about the other distributions?

(d) Is there evidence that the Democratic Party was able to tap into the youth vote? Look at the scatterplot of PctBush vs. Pct18.24. What do you notice? What do you suggest next?

(e) Is there evidence that income and voting for the Republican candidate are positively associated? Do you want to log transform income? Are there any downsides to log transforming?

(f) One of the issues that year was “traditional family values.” As a proxy, we have a variable, PctFamily, defined as the percentage of residents living in a household with husband, wife, and at least one child under age 18. What do you observe?

(g) Find the least squares line for predicting PctBush from PctFamily and interpret the slope and intercept coefficients in context.

(h) What do you learn from the residual plots?

(i) Fit the multiple regression model without median family income (Income) and using Income as the response. Does Income appear useful to add to this model? In what form?

(j) Fit the multiple regression model with median family income (model 3), using only the complete cases. Do the validity conditions appear to be met? Multiple R-squared?

(k) Examine the residuals versus each explanatory variable. What do you suggest next?

(l) Add $PctBlack^2$ to the model. Any issues? (*Hint*: Multicollinearity? Why?)

(m) Does *centering* the variables first help? How so?

(n) Investigate the three suggested interactions.

(o) What is the best way to test “can I remove all three interactions from the model”?

(p) What would a significant interaction imply in this model?