# Mid-Term Review

**March 3, 2014**

## 1. Explanations and Evidence

a. Correlation and Causality
b. Selection bias
c. Confounding: endogeneity/exogeneity
    i. Confounding factors: it is difficult to decouple what is driving what
d. Natural Experiments
    i. Ex: Israeli school policy allowed for no confounders
        1. Y-variable: test performance
        2. X-variable: class size
e. Randomize and Intervene
    i. Multiple explanations for correlative data
    ii. What is wrong with simple observation?
        1. Ex: Smaller classes have higher performance rates—difficult to test

## 2. Exploring Multivariate Data

a. Basic plots/Summaries
    i. Contingency table (titanic example)
    ii. Boxplots/histograms/dot plots
    iii. Scatter plots
    iv. Lattice plots (GPA:SAT example)
b. Group wise Models
    i. Group means
    ii. Coefficients/parameters of the model
    iii. Fitted/model values
    iv. Residuals – Actual Value = Fitted Value + Residual
    v. Taking the "x-ness" out of y
    vi. Regression: least squares (know least squares equation)
    vii. Nonlinear transformations: logs, power laws, polynomial fits: adding x^2, x^3, etc.
c. Have y and want to adjust for x, simply take residuals
d. Reducing Uncertainty
    i. Standard deviation is the average error
    ii. Adding information will always reduce your error
    iii. Squished data—take log of both sides
        1. Power law is the reverse
            a. Don't forget extra step of undoing log by exponentiating!

        iv. When fitting polynomials remember it is a tradeoff between fit and simplicity

3. **Predictable and Unpredictable Variation**
    a. Coverage intervals
    b. Standard deviation = “average error”
    c. Remember R^2 = PV/TV
        i. Closer to 1 means more predictive variation
        ii. Closer to 0 means bad fit
    d. Naïve prediction intervals (Level 2 prediction incorporate some magnitude of error, better than plug-in)
        i. Says that future error will be like my past error
        ii. Naïve prediction intervals do not take in to account the unpredictability of the estimates or the predictable variation

4. **Quantifying Uncertainty (parameter/prediction)**
    a. Definition of a sampling distribution
    b. Standard error (standard deviation of sampling distribution)
    c. Confidence intervals
        i. Informal/intuitive
        ii. Formal/mathematical version
    d. Frequentist Coverage Property (Truth in advertising: what you see is what you get)

Q: How do you estimate these things? (since really they can’t exist)

A: Bootstrapping/ Normal Linear Regression Model

    e. Bootstrapping
        i. Taking repeated samples of my sample with replacement, omission, and ties
        ii. Bootstrapped confidence interval to estimate parameter uncertainty
    f. Normal Linear Regression Model
        i. Know equation structure and assumption about data generating process which says that residuals are drawn forma normal distribution---the residual is an aggregation of nudges (or other forces we have left out of the model)
        ii. Be able to use assumptions and *Read Output
    g. Cross-Validation
        i. Another resampling based method
        ii. Split data set into 2 sets arbitrarily—make one set training and one set testing
            1. Can be used to determine how well we estimated
            2. It is necessary to do multiple cross validation splits
                a. This helps us estimate general error of model

5. **Grouping Variables in Regression**
    a. Having both quantitative and qualitative predictors
        i. Dummy variables (baseline offset format)
        ii. Interaction terms (change slope by group)
            1. Dummy variable*quantitative variable
            2. Slope: rate of change of y as x changes
6. **Multiple Regression**
    a. Partial slope (holds other variables constant or statistically adjusts)
    b. Statistical adjustment
    c. Criteria for model choice
    d. Structure of multiple regression equation
    e. How do I know what I need to add to a regression model?
        i. Look at R^2 for precision of predictions
        ii. Look at practical affect size

7. **Hypothesis Testing**
    a. Neyman-Pearson Test (6 steps)
            1. Formulate a null hypothesis
            2. Choose a discrepancy measure (ex: t statistic)
            3. Compute (or simulate) a sampling distribution
            4. Choose R, your rejection region
            5. Calculate alpha = size of rejection region as a fraction (e.g. 0.05)
            6. Look at actual value of t for your data set and determine if t falls into rejection region or not
        ii. Permutation test:  shuffling the cards (shuffled cards become null hypothesis)