Monday, March 03, 2014

Scribing – Midterm 1 Review
By: Sarah Qin

Section 1: Explanations and Evidence

- Big picture was thinking critically, going beyond “correlation does not mean causation”
- Selection bias
- Confounding: Endogeneity/Exogeneity
  - Confounder is systematically associated with both predictor and response
    - ie baseline intelligence
  - Exogenous means variation from outside the system, no inward pointing arrow
    - ie coin flip for treatment or placebo
    - ie Israeli classes of 13 and 20 are exogenous variable as well, luck
- Natural experiments/real experiments
  - Examples of natural experiments: Do smaller class sizes lead to improved test scores?
  - Question that must occur to you: What’s wrong with mere observation?
    - Generates confounders: best students or worst students in smaller classes
  - Need experiments, but they are expensive and hard to run
  - Set the stage for ability to do statistical adjustment instead of natural adjustment
  - What is driver of action and result of action?
- Randomize and intervene- good observational evidence has both characteristics

Section 2: Exploring Multivariate Data

- Basic plots/summaries: Allow you to visualize variation in data
  - Contingency table—Titanic example
  - Boxplots/Dot plots –More than one group
  - Histograms
  - Scatterplots
  - Lattice plots—Taking a scatter plot and stratifying them by group; multivariate reasoning
- Group wise models:
  - Computing group means
  - Coefficients/parameters of the model (what you’re trying to estimate in data)
  - Fitted/model values
  - Residuals
    - Actual value = Fitted value + Residual
  - Taking the “x”-ness out of “y”: take y and strip a systematic association with x, residual is what is left over, what is point lowest from the line? Statistical adjustment
  - Regression: Know the least squares regression equation

        - Plug-in predictions: Start at the x where you want to predict, go vertically up to the line, and go horizontally to the y axis. Read off your prediction.
        - Slope: how fast y changes with x
    - Nonlinear transformations: logs, power logs
    - Polynomial fits: adding x^2, x^3, etc.
    - Examples: Tooth length and vitamin C graph has more between group variation than within group variation
- Standard deviation is simply the average error
    - Standard error of residual relates to plug-in predictions
- Adding information will always reduce uncertainty, question is: how much?
- Body weight vs brain weight example: data is squished, so take log of both sides to spread things out and get a nice linear relationship
- Simple models are more likely to generalize better to future cases
    - In example, red model may generalize future best. Blue model appears to fit well with data until you look into future months, when it just drops to zero. Green model looks like it will go downwards.

Section 3: Predictable and Unpredictable Variation

- In statistics, we care about both. Unpredictable variation shows limitations of model
- Coverage intervals – confidence interval of sampling distribution
- Standard deviation = “average error” in forecasting
- Naïve prediction intervals—Naïve because it doesn’t take into account the uncertainty of the parameters of the model, ignores the fact that slope and intercept estimates may be off because of sampling variability
- Prediction level 1 (plug-in prediction) vs Prediction level 2 (incorporates magnitude of error, not just a guess)
- R squared
    - TV = PV + UV
    - Ratio of predictable variation to total variation
    - Variation of fitted values/ original variation of data
    - If close to 1, fit is good; if close to 0, fit is poor

Section 4: Quantifying Uncertainty Part 1 (Parameter/Prediction Uncertainty)

- Sampling distribution unifies all concepts in this section
- Standard error is the standard deviation of a sampling distribution
- Confidence intervals
    - Informal/intuitive “chop” way: range of plausible values you are willing to back
    - Formal/Mathematical version: defined by frequentive coverage property (Truth in advertising property, what you see is what you get)
    - Both are idealized, thought experiments. You will never see a real sampling standard error or confidence interval.

- How to deal with real data set without “gold standard”:
    - Bootstrapping- don’t have population to do repeated samples from, take repeated samples with replacement, reproduce/replicate getting samples from real population
        - Bootstrapped standard errors
        - Bootstrapped confidence intervals- “chop” method
    - Normal Linear Regression Model—Aggregation of random up-down nudges can be modeled with a normal (Gaussian) distribution
- Cross-validation method: another resampling based method, split data set into two
        - One half is training, other half is testing/validation (make predictions)
        - Can’t just do one split, have to do multiple times and average over a hundred times to give notion of variability
        - Know the purpose of this method and how it differs from bootstrapping

Section 5: Grouping Variables in Regression

- Dummy variables (baseline/offset form) change the intercept group by group
- Interaction terms change the slope group by group
    - Take more than one predictor and multiply them together
    - ie volume is interaction of 3 variables (length, width, height) in fishing example
        - dummy variable * quantitative variable
    - Slope: rate of change in y as x changes

Section 6: Multiple regressions

- Partial slope
- Statistical adjustment
- Criteria for model choice

Section 7: Hypothesis Testing

- Neyman-Pearson test (6 steps)
- Permutation test: shuffling the cards analogy
- Precision of predictions
- Precision of estimates
- Substantive effect size
- How much does x affect y?