Notes from Jeremy

Written: 11/05/17

What is the role or purpose of quant methods in the social sciences? 🡪 Foundational part

* Repeatable studies
* Empirical support for stated hypothesis
  + Inferences from smaller population
* Compare to qualitative analysis? What does it allow us to do that qual doesn’t? Drawbacks of each. Talk about small schools study.
* Bigger sample allows us to make generalizations. Talk about small schools study.

-Sampling techniques necessary to make generalizations

-Lacks the richness and depth

Primary method of quantitative analyses (for sociologists) for examining relationships between variables: OLS.

* What is it? How does it work? How we use the results from OLS?
* OLS is a statistical method for predicting the average relationship between two or more variables. The method seeks to create a predication line that minimizes the sum of squared errors, i.e. minimizes the discrepancy between the predicted value of an outcome variable and the actual value of an outcome variable at a given level of a predictor variable(s). The regression coefficients give a numerical estimate of the strength and direction of relationships between variables. These can be expressed in a standardized or unstandardized form (standardized form is the effect of change in standard deviations of the IV on the DV)
  + Approach to understand on average relationships (bivariate of multivariate form), allows for control of other variables
  + y=
  + Talk about the slope equation, what is it? Talk about what the slope is: deviations from average values in X and Y over changes in X, so it gets interpreted as 1 unit change in X is a slope change in Y
  + We interpret coefficient estimates (come in standard or unstandardized forms)
* Regression assumptions:
* When regression assumptions are not met, then it is not considered BLUE (Best Linear Unbiased Estimate)
  + linearity – constant relationship between x and y
  + normality – error terms are normally distributed
  + homoskedascity –error terms have equal variance at each level of X
  + independence –error terms are not autocorrelated
  + multicollinearity, relationship of IVs to one another, and shared variance with the DV, no unique contribution to the regression model
* Applications of OLS
  + Goldin, Katz and Kuziemko (2006) sought to explain changes college graduation rates by gender in cohorts born in 1957-1992. From longitudinal survey data, they could see that there were changes in women’s scores of aptitude tests and course taking patterns. They use an OLS regression (but it’s weird cause the DV is dichotomous, i.e. 1 if graduated and 0 if no) to estimate how much a series of proximate determinants (i.e. high school rank, test scores, and course taking patterns) and gender and family background affect the likelihood of completing a BA for three different HS graduation cohorts (1957, 1972, 1992). They are able to identify which variables had the largest impact on reversing the college graduation gender gap.

Logistic Regression

* What is logistic regression and how does it differ form OLS? Logistic regression is used to predict categorical outcome variables. One of the assumptions of OLS regression is that the DV is unbounded, continuous and interval ratio. Categorical DVs fail to meet many of the OLS assumptions including linearity, homoscedasticity and normality. In OLS, the regression estimate can be thought of as a conditional mean of Y give a value of X, in contrast logistic regression estimates the probability that Y will take a certain value or category at a given value of X. Because an estimation of probability must be between 0 and 1, we take the logit (natural logarithm of the odds) of Y, which allows the estimation to be unbounded. Ask Hayden re logit transformation??
* The coefficients of logistic regression, the log odds, are not easy to understand, nor are the exponentiated version of these, the odds ratio. Long and Freese (2014) suggest using predicted probabilities to express coefficients.
* One will use different logistic models depending on the structure of the DV. For categorical variables with ordered categories (i.e. levels of education attained or survey scale from disagree to agree), you use ordinal logistic regression. For categorical outcomes with more than two unordered categories (i.e. married, widowed divorced or agree to disagree with an I don’t know option)
* In contrast to OLS, which uses the additive values of IVs to improve prediction of the DV, in logistic regression the analysis seeks to improve the frequency of correct versus incorrect predictions and minimize errors in prediction.
* Model fit: OLS uses R2 or a proportional reduction in error statistic to assess the substantive significance of the model. Logistic regression uses deviance to compare a saturated model to the current model to assess fit and whether a set of predictor variables help to predict outcomes. There are a number of pseudo R2s, which attempt to explain how well the model reduces deviance but should not be considered as analogous to the OLS R2 due to the heteroskedasticity of errors in logistic regression
  + Discuss applications: Bailey, Jeong & Cho, discuss why this method is appropriate for exploring the phenomenon
  + Bailey, Jeong & Cho’s (2008) study of student referral, enrollment and completion patterns in developmental education used a ordinal logistic regression to analyze national and Achieving the Dream student data. One of the major innovations of this study was to look at completion of development courses as an ordinal rather than a dichotomous variable. In the model Y=0 if a student was referred to three levels below college-ready and did not complete that level, and Y if they passed. Y=2 for students who were placed two levels below college ready and passed, etc. This approach painted a more nuanced picture of remedial course taking patterns, and allowed a student who completed one level of remediation, but not all to be seen as more successful than a student who completed no remedial courses.

**HLM**

* What is it for?

HLM is used for nested data, which is very common in the social sciences. For example, observations nested in individuals (i.e. longitudinal data 🡪 how do little children develop vocabulary?), metaanalysis uses HLM to look at outcomes across studies; in this approach, outcomes are nested within studies, and HLM is useful for looking at educational settings as this data is often nested i.e. individual student outcomes, nested in students, nested in classrooms, within schools, etc.

Good things that HLM does:

* Improved estimation of individual effects through pooling data. For example, a group of researchers in the 80s wanted to model a separate admissions equations for minority students to business school, but there were so few minority students in business schools that OLS estimations for individual schools would be poorly estimated.
* Cross-level effects: HLM allows the testing of hypotheses of how variables at one level of the hierarchy affects variables at another level. For example, how does a school’s average level of SES affect the relationship between SES and academic performance for individual students?
* Partitioning variance-covariance components: For example decomposing student growth trajectories into within and between school components.
* Violates independence, why

Data appropriate for HLM analyses have a complicated error structure that violate the independence and homoscedasticity assumptions of OLS. Errors terms in HLM tend to be clustered or autocorrelated due to the nested structure of the data, subjects are exposed to similar treatments i.e. have the same teacher, or are in the same school. Or, the case of longitudinal data, the observations are taken for the same person, thus outcomes at each point are likely to be related to outcomes at the previous point. Error terms in OLS also have unequal variances.

* Structure of HLM

First level variables would be student level variables predicting individual student outcomes i.e. gender, race, SES. Second level variables (depending on what the second level is) could be classroom level (i.e. teacher tenure, class size, curriculum design) or school level variables (i.e. school type charter or religious, geographical location, finding)

* 1st level is the random effects model (demographic characteristics), 2nd level are mean effects. Look at variation in non-repeated measures and repeated measures and see if there is any interaction
* If you can identify variation across second level units then HLM is appropriate
* 1st level: random coefficients model, 2nd level means as outcomes model, coefficients as estimates model which looks at cross level interactions
* Purpose of each level of the model

Missing data

The important feature of missing data in terms of how consequential it may be for results is the degree of randomness of the missing items.

* Identify different types of missing dates
  + Missing Completely at Random means that the missingness is not related to observed or unobserved characteristics of the sample. There is no systematic observable difference between observations with full data and those with missing data. This is ignorable.
  + Missing at Random means that missingness is independent of unobserved characteristics, conditional upon observed data.???? It is more common for data to be Missing at Random.
  + Missing Not at Random means that the missingness is conditional upon observed or unobserved characteristics of the sample.
* Identify different methods for addressing missing data, imputation
  + Common procedures for addressing missing data include data weighting, imputation and “other corrective schemes.”
* Impact of missing data and imputation may have on study results
  + Missing data can make it difficult to claim that findings are generalizable, i.e. complicate the ability to make inferences based on the data.
  + Using different imputation schemes can lead to different outcomes from analyses, i.e. different imputation schemes can create variation in a dataset
* Heckman Correction procedure
* Read Jeremy’s paper with Elaine Ecklund
  + This paper attempts to analyze why certain populations may not respond to survey questions. A main finding was that the group of people targeted for the survey matters i.e. there are specific reasons that a group of people may avoid answering certain survey questions.

Causal Inference

* Propensity score matching
* RD studies
* Counterfactual