**LOGIC OF INQUIRY**

**What is quantitative analysis and what does it do?**

* Quantitative analysis the study of social phenomena that uses numbers to quantify and summarize the relationships
* It allows us to operationalize research questions that test hypotheses following a scientific testing process
* scientific testing process
  + generate a theory
  + develop hypotheses
  + gather data and conduct analysis
  + make empirical generalizations
* discuss my own research with Brenden that has followed this process
  + Cross-lagged regression models with tract and year fixed effects
  + Use Arellano-Bond first-difference, generalized method moments (GMM) that produce efficient and consistent parameter for estimates when including lagged DVs
  + We also use Arellano-Bond tests for second-order autocorrelation
* pros
  + operationalize, quantify, and replicate our study
  + good at painting an overall picture of our average neighborhood
* cons
  + the coefficients summarize the average relationship, the average effect
  + limited scope about what we can say

**What are the different types of quantitative data and why is data type important?**

* There are three general type of quantitative data
  + Cross-sectional
    - Many variables at one time point
    - Holds space constant
    - Census Data
  + Time series
    - A specific variable at multiple points in time
    - Holds time constant
    - Quarterly data on housing prices
  + Pooled and Panel Data
    - Pooled Data
      * Time series of cross-sectional data
    - Panel Series: same respondent over same of time.
      * Time series of cross-sectional data that are the same observations
      * PSID
* The type of data you have impacts the types of questions you can ask, what type of regression you can run, and what kinds of things you need to consider
  + i.e., If you have time-series data you will can look at change over time, but yo might need to account for serial autocorrelation

**ORDINARY LEAST SQUARES REGRESSION (OLS)**

**What is OLS Regression and how does it work?**

* OLS is a statistical method for predicting the average relationship between two or more variables
* The point is to fit a regression line to the data, that can be interpreted as a summary of the relationship of the variables in the dataset
* Most common method is called OLS, ordinary least squares, which generates a line that fits the data, specifically my minimizing the sum of the squared errors
  + regression line: best average summary of the data
  + error term: distance between observed and predicted
    - squaring the errors acts as weighting mechanism to minimize outliers
    - errors above the line don’t cancel out those below it
* simple linear regression (bivariate regression) vs multiple regression
  + simple linear regression is only between two variables
    - summarizes the impact of an IV on a DV
    - neighborhood affordability regressed on average age of homeowner
  + multiple regression allows us to control for the influence of other variables by holding them constant
    - neighborhood affordability regressed on age, race, and income of tract
    - each coefficient summarizes the predicted effect on IV while holding the others constant

**What are the basic assumptions of regression and what is BLUE?**

* **B**est **L**inear **U**nbiased **E**stimators (BLUE)
  + **B**est means that they have the smallest possible error terms
  + **L**inear means they have a constant relationship between IV and DVs across all values of every variable
  + **U**nbiased means they would equal the true population’s effect over an unlimited number of repeated samples
* Eight basic assumptions of OLS:
* Linearity and additivity
  + Linearity:
    - That there is a constant linear relationship between DV and IVs
      * e.g., affordability a non-linear function of age,
      * if affordability changes at different rates for age, then take the square to induce normality
  + Additivity
    - There is a constant effect no matter the outcome of other variables
    - E.g., Effect of age on affordability is the same no matter someone’s income
      * Interact income with age to see get at additivity
* No measurement error
  + Assumes that the researcher is not introducing non-random error
    - E.g., gini coefficient are measured correctly, correct sample, no missing
* No specification error
  + Model includes all important variables
    - E.g., omitted variable bias
  + Relationships among variables are correctly specified
    - All variables are appropriately logged or squared (additivity)
* No multicollinearity
  + Refers to collinearity (association or dependence) of one or more variables on another
  + Mostly concerned with perfect collinearity
    - One variable or set of variables that is collinear
  + Tests
    - VIF (variance inflation factor)
    - Subset the data and regress it on other variables
  + Solutions
    - Drop or create scales
  + Not a problem because it inflates the standard errors which shrinks confidence intervals and make it more difficult to prove something significant
    - If variables are sig. in spite of multicollinearity, then def significant
* The mean of the error terms is zero
  + since it’s the line of best fit, the error terms should average to be zero
* No heteroskedasticity (homoskedasticity)
  + There is no correlation among the error terms in the model
    - Its about whether a regression model's ability to predict a DV is consistent across all values of that DV
    - Errors represent random effects for which the model cannot control, a correlation of an error term with other implies a systematic relationship that is not random
  + The assumption that there is an even spread of the error terms above and below the regression line, constant variance of the error terms
  + Possible sources
    - Meaningful variance in the dependent variable
      * I.e., Relationship between income and affordability
    - The result of some important variable being omitted
      * I.e., income as a predictor of affordability, but no account for household size
  + consequence
    - increases variance of error terms and makes estimates less efficient
  + detected
    - visualize it by plotting the residuals
    - Goldfeld-Quant that uses chunks of data
  + solution
    - perhaps model is not correctly specified
    - robust standard errors
    - use another model: GLS or WLS
    - transform DV—methodological implications
* No autocorrelation (independence)
  + The error terms are independent of all the other errors other on all other values
    - The value of one error term does not depend on another
    - y(x+1) does not depend on values of y(x)
  + Consequences
    - can underestimate the standard error term, which inflates the t score, leads to type 1 error
    - which means you could find something that is statistically significant when it is not
  + detected
    - Durbin-Watson model
    - Breusch–Godfrey
    - Lagrange Multiplier Test
  + solution
    - use another model: GLS, WLS, spatial regression, and HLM models
    - include some parameter to account for autocorrelation
* Normality
  + the error terms are normally distributed
  + Following the central limits theorem, the distribution of the error terms in the sampling population will form a bell curve or normal distribution.

**How are categorical variables included in OLS?**

* Only can include continuous variables, but can include dummy variables, 0 and 1
  + Must exclude one category which will serve as a reference
    - Significance of variables can change based on reference category
    - Should choose a meaningful reference category
  + Interpret the effect as compared to the reference category

**What is the point of an interaction affect?**

* Interactions are tests for conditional relationships
  + It’s the effect of X1 on Y that depends on different levels of X2
* When you run an interaction, you would get three coefficients
  + Two coefficients that summarize the effect of each on the outcome, and the third that summarizes the effect of the interaction
  + i.e., affordability regressed on income and age, with interaction of age

**What are some applications or examples from your own work or from what you’ve read?**

* **Faber and Ellen:** where they used OLS regression to explain difference in home equity trends by race from 2003 to 2009
  + American Housing Survey, longitudinal, nationally representative sample of housing units
  + OLS appropriate
    - DV: home equity: difference between reported home value and outstanding principal
    - EV: race with white as reference
    - Controls
      * Household demographics
      * baseline equity
      * change in housing prices
      * housing unit characteristics
  + All saw increases in home equity, but blacks and Hispanics saw less as compared to white
* **Paul Attewell:** Earnings benefit for those with some college, no degree compared to HS only
  + IPUMS 1 -year samples from 2013 to 2017
  + OLS appropriate
    - DV: log of earnings, because there is a non-linear relationship
    - EV: binary HS and SC, no degree
    - Controls: age, martial status, year and break the models down by sexrace

**LOGISTIC REGRESSION**

**What is Logistic Regression and how does it differ from OLS regression?**

* When the dependent variable is a categorical instead of continuous variable
  + Binary, order, and multinomial
    - Homeowner, levels of education, various tenure statuses
* Violation of three assumptions of regression
  + Linearity
    - The categorical nature of outcome bounds its distribution between 0 -1
      * Probability that Y is a one
      * bounded between 0 and 1
    - OLS predicts outcomes beyond those bounds,
      * outcomes are no longer linear, constant function of X
    - Still unbiased, but standard errors and confidence intervals are invalid
  + Independence
    - The outcome is a bounded function of the mean of X that therefore depends on the value of X
    - No longer independent
  + Homoskedasticity
    - The error terms vary with each other, especially with values near the middle
* Logistic regression uses a logit function to transform the DV, taking the natural log of the odds that Y equals one of the categories
  + The logit function is a sigmoidal curve that is nearly linear in the middle but is bounded at both ends as X approaches high or low values
  + For this reason and because it is still unbiased, some use OLS regression with categorical DVs

**What are the difficulties of interpreting logistic regression?**

* Logistic regression estimates the probability that Y will be a 1
  + It can be interpreted as the odds that a case falls into the higher of the two categories of the dependent variable
* The resulting coefficient is in the logit scale as logged odds
  + Coefficients can be transformed:
    - odds ratios: exponentiated, subtract 1, and multiply 100
    - margin commands to obtain predicted probabilities

**How is logistic regression estimated?**

* Uses the maximum likelihood function to calculate the values that have the greatest likelihood of generating the observed values and therefore minimize error
  + Unlike OLS where you could theoretically explain all the variance in the errors, ML is about increasing the ability to correctly predict 1’s over 0’s
* ML is an iterative process that adds more variables one at a time, increasing the model’s ability to predict 1’s over 0’s
  + Start with some arbitrary values and compares the likelihood of other values across the distribution, essentially fitting a normal distribution until it arrives at the estimators with the highest likelihood of being the population parameters.
* How determine the best model?
  + likelihood ratio test that compares the model with the likelihood of the data under the full model against the likelihood of the data under a model with fewer predictors
  + pseudo R2 is not appropriate because of the heteroskedastcity of errors

**What are some applications or examples from your own work or from what you’ve read?**

* Faber and Ellen, from earlier
  + Data: American Housing Survey 2003-2009
  + Question: second half, same covariates to predict likelihood of being underwater
  + DV: continuous -----> binary
    - 0, 1: likelihood of being underwater
    - no longer linear relationship, violation of 3 assumptions
  + Findings: found white homeowners who held onto home were more less likely to be underwater than their Latino and black counterparts
* Freeman and Braconi
  + Data: Housing and Vacancy Survey from the 1990s
  + Question: measure the displacement pressures on disadvantaged households in gentrifying neighborhoods
  + DV: likelihood of moving out of a neighborhood
    - Disadvantaged households: no college degree, below poverty line
  + EV: live in a gentrifying neighborhood
  + Controls: demographic controls, housing unit characteristics, neighborhood rating
  + Finding
    - poor households were 19% less likely to move
    - non-college households were 15% less likely to move
* Timberlake and Johns-Wolfe
  + Data: Decennial census data in NYC and Chicago
  + Question: how does ethnoracial composition impact gentrification
  + DV: multinomial logistic regression with 5 categories
    - did not gentrify
    - gentrify white
    - gentrify black
    - gentrify Hispanic
    - gentrify mixed
  + EV: percent shite, black, Latino
  + controls: population characteristics, housing characteristics, proximity to CBD
  + Finding: Lots
    - % Black in 1980 was negatively associated with gentrified white and positively associated with gentrified black neighborhoods
    - % Hispanic in 1980 was positively associated with gentrified Hispanic
  + Strength: study was able to test the likelihood of gentrifying along pathways, really getting at the nuance of gentrification along racial lines

**HIERARCHICAL LINER MODELS (HLM)**

**What is HLM and what are the four basic assumptions?**

* Hierarchical linear models (HLM) aka Multilevel Models
* Specific type of regression that accounts for violations of assumption caused by nested data
  + Developed by educational researchers to measure the effect of context on individuals
    - Effect of a classroom environment on a student
  + Can be translated to spatial contexts
    - Effect of neighborhoods on individuals, city on neighborhoods
* Violations
  + Linear models cannot account for complicated error structure of nested data
    - Error terms are clustered by contextual unit because observations are exposed to similar treatments within groups
  + HLM relaxes independence assumptions and allows for correlated error structure
* Four basic assumptions
  1. Homoscedasticity
     + No correlation between clusters
  2. Independence
     + Errors at L1 are independent from errors at L2
  3. Normal distribution of errors
     + L1 errors are normally distributed
  4. Normal distribution
     + L2 intercepts and slopes are normally distributed
* Pros
  1. Allows for cross-level interactions
     + See how variables at one level affect the variables at another
  2. HLM is about partitioning variance-covariance components to see how much variance in the DV is explained by context
  3. Leverages pooled sample to improve estimation of individual effects
     + Assess admissions of minority students to business school
     + OLS would not work bc too few students
     + HLM uses maximum likelihood

**What type of estimation is used in HLM?**

* HLM uses Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML)
* They produce the same estimates for fixed effects,
* REML takes degrees of freedom into consideration, and thereby produces less biased random effects than full ML.

**Discuss how to construct an HLM model.**

* First class of models is call the unconstrained or null models
  + No L1 or L2 predictors
  + Just DV that varies by context
    - Essentially the mean varies by each context
  + can get the interclass correlation coefficient (ICC)
    - gives you the proportion of variance of DV explained by the context
    - between group variance / group variance
* random intercept
  + L1 and L2 predictors
  + L1 varies randomly by context (random intercept) : predictor
  + Average L2 effects on L1 predictors
* random intercept and slopes models
  + let’s the intercepts and slopes of L1 variable vary across L2 units
    - creates an individual regression line where each case has its own slope instead of taking on L2 context’s slope
  + model L1 intercepts and slopes using L2 predictors
* cross-level interactions
  + interaction a L2 with an L1
  + what’s the effect of L1 X on Y that depends on different levels of L2 X

**What are the benefits of using HLM and when can OLS with standard clustered errors be used instead?**

* OLS model with clustered standard errors can account for the complicated error structure associated with HLM models
* Pros
  + It’s easier to use and understand than HLM models
  + Why use difficult method when there is a simplier method?
  + Can get the same results with basic modeling techniques
* Cons
  + OLS is not designed to model nested error structure
  + Can’t let slopes vary by context
  + Can’t run cross-level interaction
* Run them both and compare results
  + Usually aren’t that different for simple regression models

**What are some applications or examples from your own work or from what you’ve read?**

* Emily Molina use a multilevel logistic model
  + Data: census data and data on foreclosures and home sales in LA county
    - Foreclosure 2008-2009
    - Census from 2000
    - Sale price data from 2005-2007
  + Levels:
    - L1: Foreclosures
    - L2: census tract
  + Question: what has happened to foreclosed properties after the financial crisis and how have neighborhood characteristics impacted those trajectories
  + DV: three variables
    - Odds of being sold to investor
    - Odds of being sold to corporate investor
    - Odds of being flipped within study period
  + Finding: two major findings about geography and race
    - Inner-city and inner-ring suburban tracts experienced more corporate investment and speculation while exurban benefited from owner-occupied investment
    - Corporate investor targeted predominantly white neighborhoods while smaller investors targeted non-white communities, tho everyone was likely to flip in communities of color
* Work with Paul
  + Data: census data and data on foreclosures and home sales in LA county
    - 1-year IPUMs data from 2013-2017,
  + Levels:
    - L1: Statepumas
    - L2: individuals
  + Question: what are the spillover effects on HS earnings for those living in neighborhoods with a higher share of people are who have any SC?
  + DV:
    - Log of earnings for those with only a HS degree
  + Controls
    - Age, marital status, race, sex, foreign born, % any SC, cost of living controls
  + Finding:
    - Significant findings of a spillover effect, but very small
    - For every 1% increase in % any SC, 0.001 % increase in log of HS earnings

**SPATIAL REGRESSION**

**What is spatial regression and why should we use it?**

* Best illustration of the importance of space is through example
  + John Snow example of how things cluster in space
    - Cholera outbreak in Soho, London in 1850s
    - Plotted individual deaths
    - Saw spatial clustering around broad street water pump
    - Further investigation revealed it was contaminated
  + Flu example
    - Likelihood of catching the flu
    - Individual immune system, personal characteristics
* Spatial Regression is a technique that takes spatial structure and interaction seriously
  + Specifically, it is a regression technique that corrects for spatial dependence, spatial autocorrelation, clustering in space
  + Spatial dependence
    - statistically significant spatial clustering
    - measure of similarity or correlation with nearby observations
  + Problem
    - violates OLS assumptions of independence and homoskedasticity
    - spatial dependence creates first-order spatial autocorrelation that increases the standard errors
    - increases the likelihood of Type 1 error—rejecting the null hypothesis that there is a significant difference when there is not.
  + Detect
    - Run an OLS regression and check the clustering of the standard errors
    - If there is clustering and you suspect some spatial component, it is likely that there is spatial autocorrelation
    - Calculate a global Moran’s I statistic
      * Define what counts as a neighbor
        + Queen, rook, bishop
      * Calculate deviations from the mean and take the cross products
        + positive Moran’s I: spatial clustering
        + negative Moran’s I: dispersion
        + zero: randomness
      * Essentially the correlation coefficient, the slope of the regression line for the spatial regression
  + Solution
    - In same way that serial autocorrelation is corrected, by modeling the spatial correlation in the regression equation in some way
    - There are two ways to do this
      * Spatial lag
      * Spatial error

**What is spatial lag and spatial error?**

* Spatial dependence can be thought about in two ways
  + Substance
  + Nuisance
* Spatial lag = substance
  + Definition: Spatial distribution of democracy across globe
    - OLS to predict polity (democracy) score base on their GDP
    - Tests for spatial dependence confirm clustered by space
    - Need to correct for spatial dependence
    - Tells you something theoretically, not just something to control away
  + Solution
    - Incorporate a spatially lagged variable to account for spatial dependence
    - Moran’s I: the correlation of DV across space
  + Checks
    - Can compare errors in OLS to Spatial model
  + Pros
    - Democracy: better than fixed effect models by country because it is dynamic
      * changes in one unit will be incorporated into spatial lag in spatial model but not in fixed effects models
    - forces researcher to think spatially and incorporate space into a model and to better understand spatial mechanisms that affect our data
* Spatial Error= nuisance
  + Definition:
    - When there is no substantive interpretation of the interaction of Y1 and Y2
      * Spatial correlation enters through the error terms
      * Errors are spatially correlated
    - Technical nuisance, statistical problem to adjust for
  + Example
    - Dyadic trade flows between countries
    - DV: volume of trade between countries
    - Trades are not independent because they share the same sender
    - Flow from one will be correlated with reverse flow to another
      * Error terms would be linked
      * Spatial error model that incorporates spatial weights
  + Solution
    - Similar to spatial lag, include weighting structure that accounts for correlated error terms
  + Implications
    - observations are related due to unmeasured and factors that for some unknown reason correlate across the distances among the observations

**What is the difference between the two and how do you know which to use?**

* Some: Lagrange Multiplier Test
* Others: use theoretical understanding of the analysis, often go with spatial lag
* New methods: Spatial Durbin Error Model that account for both

**What is GWR and when should it be used?**

* Difference between global and local
  + Spatial regression only produces one statistic for global analysis
  + Local Anselin’s Moran’s I provide local statistics
  + Does not get at variation
* GWR: Produces local regression coefficients for each spaital unit
  + Uses a specific kernel type, specific bandwidth, and specific distance matrix to calculate local regression coeffeicents
  + Adaptive weightin for when observations are more or less spread out
* Pros
  + Regression coefficients for each spatial unit
  + Allows for understanding of local dynamics and variation
  + Provides better fitting regression lines for local units
* Cons
  + Sub-setting the data can be dangerous
  + Never use without global model
  + Avoid cateogorical and nominal data

**MISSING DATA**

**What are the common problems with missing data?**

* Two big problems with missing data
  + Can dramatically affect the findings themselves
  + Difficult to claim that findings are generalizable
* Sources of missing data
  + Non-response
  + Attrition
  + Censorships, suppression, not available

**What are the three main types of missing data?**

* 3 main Types of missing data
  + Missing Completely at Random (MCAR)
    - missingness is not related to observed or unobserved characteristics of the sample
    - no systematic observable difference between observations with full data and those with missing data
    - example: income by age
      * compare income of those missing data with those not missing and see if age is the same
    - generally ignorable
  + Missing at Random (MAR) [aka Missing Conditionally at Random]
    - Missingness is independent of unobservable characteristics
    - Conditional on observed data
    - Can account for it in dataset
    - Example: income by marital status
      * Dependent upon marital status
      * Probability of missing data for each category of marital is unrelated
    - No way to test
    - Generally ignorable
  + Missing Not at Random (MNAR)
    - Not MCAR or MAR
      * Depressed males didn’t fill out survey on depression because they were depressed

**What are some strategies to address these problems?**

* three common strategies
  + listwise deletion (casewise)
    - delete all the observations in the sample/analysis that have missing data
      * great for MCAR, even better for MAR
      * looks like stratified sample with MAR data
    - default of programs like Stata
    - pairwise deletion
      * uses all possible data
      * biases standard errors and tests statistics
  + maximum likelihood
    - does not impute or add responses for missing values
    - circumvents missing data problems
      * estimating a parameter value for each observation that is the most likely to have resulted in the observed data
  + multiple imputation
    - imputes missing data by estimating them multiple times uses the data that you have
    - includes an error measure that upwardly adjusts standard errors
    - averages across all iterations to produce consistent, efficient, and normal estimates

**How can we better think about missing data as it relates to our work?**

* Porter and Eckland (2012) analyze the reasons why certain populations may or may not respond to survey questions.
* Two major findings
  + First, survey target group matters as individuals may not respond for different reasons
  + Second, why people don’t respond matters and may not be for the reasons we generally suspect
    - Religion among Academic Scientists (RAAS)
    - Thought they wouldn’t respond because of their religiosity
      * Similar thinking as to why rich people don’t report their incomes
    - Finding
      * the survey was not exhaustive of religious views
      * changes the natures of the survey and the research question
    - instead of systematically filling in missing data, we need to understand the social significance of why it might be missing