**LOGIC OF INQUIRY**

**What Quantitative Analysis Does**. As opposed to qualitative research that gathers non-numeric data in the form of stories, first-hand accounts, observations, and interviews, quantitative research collects data in the form of numbers. Statistical analysis involves mathematical techniques to analyze numeric data, speak about the relationships among them, and test research questions. Much like qualitative research, quantitative research follows a rigorous scientific testing process: generate a theory, develop hypotheses, conduct observation and data analysis, and make empirical generalizations. This process is often repeated as we learn more about the social world.

Some of my own work has walked this process. My research colleague and I had a theory about whether supply or demand processes led gentrification in New York City neighborhoods. We hypothesized that increases in supply of gentrification preceded demand changes in neighborhood demographics and collected quantitative data on demographic change and housing prices in New York City from 2009 to 2016. We then used statistical analysis to test our theory and found that demand indeed preceded supply. While these data are not representative of the larger US population across all years, we could generalize about our population of interest, New York City during our study years.

There are both benefits and draw backs to quantitative research. While it is good at painting a general, overarching picture of social processes, it is often limited in how refined of a picture it can present. For example, our research can identify the socioeconomic mechanism—supply—that sparks gentrification, but it cannot elucidate the reasons why individuals move to gentrifying neighborhoods, how they identify them, or what kind of experience they have when they move there—this is where you need qualitative research. While one could imagine a quantitative research design that would include surveys that ask some of these questions, in general quantitative research is more concerned with the average experience or the macro level.

**Data**. There are essentially three main types of quantitative data: cross-sectional, time series, and pooled. Cross-sectional data measure many variables at a specific point in time, while time series data measure specific variables at various points in time (Kahane 2007). Cross-sectional holds space constant, while time series holds time constant. Pooled data are a combination of both, the same variables measured over time, while a panel series measures the same respondents and the same variables over time.

**ORDINARY LEAST SQUARES REGRESSION (OLS)**

**What is OLS Regression?** Linear regression is a tool for researchers to test hypotheses or theories about relationships in the social world and summarize those relationships between different social phenomena. The hypothesis, the more inequality in a neighborhood, the lower the share of housing affordable to the middle class is an example of a functional relationship that we could test with regression. It’s a statement about the effect that an independent variable—in this case level of neighborhood inequality—has on the dependent variable—the share of housing affordable to the middle class. Regression analysis provides us with coefficients that numerically summarize the general relationship between the independent and dependent variables. For instance, as the level of neighborhood inequality increases by 1 unit, the level of affordable housing decreases by 3 units. In short, regression is a statistical method for predicting the average relationship between two or more variables, with the goal of trying to summarize that relationship in ways that help us understand the average effect of a predictor on an outcome.

While simple linear regression allows us to measure the effect of one variable on another, multiple regression allows us to measure several factors at the same time by holding other variables constant. So, multiple regression would allow us to measure the effect of population change, age, or employment characteristics of each neighborhood in addition to measuring the effect of a neighborhood’s level of inequality on its housing affordability. Each coefficient in a multiple regression would summarize the predicted effect on housing affordability for each independent variable, while holding the others constant.

**How does it work?** The point of OLS regression is to fit a line to a dataset, which can then be interpreted as a summary of the relationship of the variables in the dataset. The most common method of regression is what is called Ordinary Least Squares (OLS). This method derives its name for the practice of generating a line that minimizes the sum of the squared errors. An error term is the difference in an observation between what the regression line predicts and what is observed in the data. Returning to the previous example, there is a difference between where the regression line summarizing the relationship between inequality and affordability predicts where an observation will be and where it is. The goal of OLS is to fit a line that best predicts where all those observations would be. The best method for doing this is to minimize the sum of the squared error terms (Schroeder, Sjoquist, and Stephan 1986:19). Squaring the error terms acts as a weighting mechanism in the equation that derives the regression line so that far away error terms don’t pull the line far away and so that errors below and above the line do not cancel each other out when you sum them.

**Basic Assumptions of OLS.** The goal of OLS regression is to produce estimators that are the **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 across all values of every variable, and **U**nbiased means they would equal the true population’s effect over an unlimited number of repeated samples. Together, OLS regression makes eight basic assumptions:

1. Linearity and additivity
2. No measurement error
3. No specification error
4. No multicollinearity
5. The mean of the error terms is zero
6. No heteroskedasticity (homoskedasticity)
7. No autocorrelation (independence)
8. Normality

**First**, OLS regression assumes there to be a constant, linear relationship between the predictor(s) and the outcome that does not vary by context and—in terms of additivity—that the effect is constant no matter the values of the other variables in the equations. In some cases there isn’t a linear relationship between a predictor and an outcome or the effect is lower for lower values while higher for higher values. Returning to my example—measuring the share of a neighborhood’s affordable housing as a function of its inequality, we would theorize that affordability could also be a function of the age of its residents. However, it could be that age has a non-linear relationship with affordability, that is, neighborhood affordability increases at different rates for different ages. Introducing a quadratic term into the model for age, by taking its square, would induce linearity in the relationship. Further in terms of additivity, OLS assumes that the effect of inequality on affordability is the same no matter the value of age or any other values of other variables included in the model. So, for example, the effect of age is the same whether incomes are high or low. One way to address this is to interact age with incomes to test the degree to which age affects income, which will give you a better sense of the unique contribution of each variable.

**Second**, OLS regression assumes that there is no measurement error. This means that the researcher is not introducing non-random measurement error into the variables. For example, OLS assumes that the Gini coefficients that measure inequality do so accurately or that age of a neighborhood’s residents were measured systematically. A **third**, related assumption of OLS is that there is no model specification error. This means two things: (1) that the regression model was correctly specified to include all the theoretically appropriate variables in the model and that no important ones were excluded, and (2) that the relationships have been correctly specified. First, if we didn’t include some measure of household income in our model of affordability as a function of inequality, we would have model specification error, which is often called omitted variable bias. While we think we are measuring the effect of inequality on affordability, some of that coefficient is therefore also measuring the effect of household income on affordability as well, although we haven’t specified it in our model. Second, if one of the variables in the model has a non-linear relationship, it has to be transformed—by introducing a quadratic term or taking the log—to make it more linear. A **fourth** assumption of OLS is that there is no multicollinearity. Multicollinearity refers to the correlation (association or dependence) of one or more variables in a regression model. While there will undoubtedly be some collinearity in a model, concerns of multicollinearity primarily refer to perfect collinearity, or high levels of correlation—as one variable varies, another varies exactly with it. However, perfect collinearity is very uncommon, and many argue that multicollinearity is not actually a problem since it inflates standard errors and shrinks confidence intervals, making it more difficult to prove a variable has a statistically significant relationship in a model. While a researcher could conclude that a significant variable is insignificant, the flipside is that if you have a variable that is significant despite multicollinearity, it is most definitely significant. While many researchers check for multicollinearity by testing for correlations between variables, VIF is often used, it is better to regress a variable suspected of multicollinearity on the others so that you do not miss how multiple variables can add to multicollinearity. Often the solution is to drop one of the variables that is multicollinear or create scales of multiple collinear variables.

The remaining assumptions pertain to the behavior of error terms. The **fifth** assumption of OLS is that the error terms are well-behaved—the average of the error terms is zero. While there are bound to be error terms above and below the regression line, they should on average cancel each other out. **Sixth**, OLS regression assumes there to be an even spread of error terms below and above the regression line, or homoskedasticity. Another way to think of this is that OLS regression assumes the error terms to have constant variance. Heteroskedasticity, non-constant variance is usually a result of an interaction of an independent variable in the model with one that is not. An example might be if we included income as a predictor of affordability but didn’t account for household size. Affordability would not only vary for households with different incomes, but also those of different sizes—as income increases it will have a different effect depending on household size. The consequence of heteroskedasticity is that it increases the variance of the estimators and biases the error terms. While the estimators are still unbiased—over a high number of samples they would still equal the average of the population—they are no longer the best in that they do not have the smallest variance. The best method to detect heteroskedasticity is to visualize the error terms and check for constant variance. There is also a statistical test that can detect heteroskedasticity, the **Goldfeld-Quandt** test using chunks of data. If there is non-constant variance that cannot be corrected, it may be that a linear model is the most appropriate model for the analysis. Perhaps a better solution would be to use a model like generalized least squares or weighted least squares regression, which account for non-constant error variance, or to transform the dependent variable, which has theoretical and methodological implications.

**Seventh**, a related assumption is that there is no autocorrelation. This means that the error terms are independent of all the other errors other on all other values. If errors represent random effects that for which the model cannot control, a correlation of an error term with other implies a systematic relationship that is not random. The presence of autocorrelation increases the variance of the estimators and biases the error terms. This is most common in time-series data, usually called serial autocorrelation, when the present value is affected in some way from a previous value, but it is also common with spatial data—called spatial autocorrelation—when values are correlated to those in relation to distance. In our example, the inequality within a neighborhood or the share of its affordable housing over time is most likely predicated on previous values of inequality or affordability. Again, generalized least squares or weighted least squares regression are some models that can account for this type of systematic error correlation. The **eighth** and final assumption of OLS regression is that the error terms are normally distributed. Following the central limits theorem, even if the error terms are not normally distributed in the population, the distribution of the error terms in the sampling population will form a bell curve or normal distribution.

**Applications of OLS.** One interesting application of OLS regression to housing issues is the work of Faber and Ellen (2016). They sought to explain differences in home equity trends by race from 2003 to 2009. Using the American Housing Survey (AHS), a nationally representative longitudinal study of housing units conducted by the US Census every two years, they construct their dependent variable, home equity, as a measurement of the difference between a respondent’s report of home value and the outstanding principal on associated mortgages. Their explanatory variables are dummies for race, with non-Hispanic white as the reference group. They control for housing market appreciation by region, housing unit characteristics like home value at the beginning of the study period and the decade of construction, as well as household characteristics like education and logged income. They find that over the study period, all four groups saw increases in home equity, but that they followed different trajectories. Most significantly, they found that blacks and Hispanics gained less home equity than whites and were more likely to end the period underwater.

Some of my recent work with Paul Attewell uses OLS regression to test the earnings benefits for individual who attend some college but receive no degree as opposed to those who only received high school degrees. Using 1-year estimates of IPUMs micro data from 2013 to 2017, this research tries to understand the effect on earnings of going to some college but not earning a degree. Earnings is a continuous variable, so it is appropriate for OLS. However, it is right skewed. Transformation makes sense not only because of its skew, but because we theorize a non-constant relationship with the independent variable. We therefore take the log and regress it on our explanatory variable, a binary variable where 0 equal those who received only a high school degree, and where 1 equals those who went to some college but did not receive a degree. We control for age as a categorical variable, state, year, and marital status. To control for selection, we run regression separately by sexrace, so for example, a separate regression for White males or Hispanic females. Preliminary findings suggest that on average there are significant effects of some college, no degree on earning for all sex race groups.

**LOGISTIC REGRESSION**

**What is Logistic Regression?** Logistic regression is used when the outcome variable is a categorical instead of a continuous variable, because having a categorical outcome variable violates three principal assumptions of OLS regression: linearity, independence, and homoskedasticity. A categorical outcome variable violates the assumption of linearity because rather than expressing the outcome as a continuous distribution, it is expressed as a probability that is bounded between 0 and 1. OLS regression predicts outcomes beyond those bounds, which means that the predicted outcome is no longer a linear function of X where there exists a constant relationship between X and Y at different values. While linear regression parameters are still unbiased (meaning they resemble the population parameters over a repeated sample), standard errors and confidence intervals would be invalid. Since the outcome is a bounded function of the mean value of X, the outcome depends on X and is not constant as the mean changes from 0 to 1. This means that the error terms are not independent of the predictors—there is autocorrelation. With autocorrelation, comes heteroskedasticity, non-constant error term variance, because the error terms vary with each other as well as with the predictors. To correct for this and satisfy assumptions of regression, the model can be linearized by taking the logit function of the outcome variable—taking the log of the odds. The logit function is a sigmoidal curve that is bounded by 0 and 1, is nearly linear in the middle, but is curved at either end as X approaches either very small or very large numbers.

**Data Structure.** The structure of the dependent variable will influence what type of logistic regression can be used. Binary logistic regression is used for dichotomous dependent variables (own a home or do not), ordinal logistic regression is used for ordered categorical outcome variables (levels of education or survey scales), and multinomial logistic regression is used when the outcome in unordered (i.e., different types of tenure status).

**Interpreting Logistic Regression.** In contrast to OLS that estimates the conditional mean as a function of Y of a given a value of X, logistic regression estimates the probability that Y will take a certain value or category at a given value of X. “The predicted value of the dependent variable can be interpreted as the predicted probability that a case falls into the higher of the two categories on the dependent variable, given its value on the independent value” (Menard 2002:7). The resulting coefficient in logistic regression models is the log odds, which are difficult to interpret. Two alternatives can be used, (1) the log odds are exponentiated to become odds ratios, which express the odds of the outcome occurring as a function of X, or after converting the log odds to the odds ratios, they (2) can be interpreted as a percent change in odds (Demaris 1995:960). A third option suggested by Long and Freese (2014) is to use the margins command to obtain predicted probabilities for coefficients instead.

The log likelihood is the criterion for selecting parameters in the logistic regression model (Menard 2002:20). Logistic regression takes the log of the maximum likelihood (ML) function and calculates the maximum likelihood estimates for which the values of the parameters have the greatest likelihood of generating the observed sample of data. This process is much like trying to find the summit a hill, blindfolded, and only able to test the slope of the hill by sticking a foot out in one direction at a time. After a while, you’ll find the summit form where you started. ML is an iterative process that adds more variables one at a time to the model and determines whether there is an increasing ability of predicting a one over a zero. Unlike OLS where you can account for 100% of all variance in the errors, ML is an iterative process that increases with the model’s ability to correctly predict the 1’s over the 0’s. As such, maximum likelihood starts with some arbitrary values and compares the likelihood of other values across the distribution to determine which estimators have the highest likelihood. It essentially fits a normal distribution on the data until it arrives at the estimators that have the highest likelihood of being the population parameters. In contrast to OLS that uses the additive values of predictors to improve the prediction of the outcome, logistic regression looks to improve the frequency of correct versus incorrect predictions and minimize errors in prediction. While the pseudo R2 is often used to compare models by explaining how they reduce the deviance in likelihood, they are not exactly analogous to OLS’s R2 because of the heteroskedasticity of errors in logistic regression. The best approach to identifying which model is better is to conduct a 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.

**Applications of Logistic Regression.** Faber and Ellen (2016) use the same covariates they used in predicting home equity to now predict whether or not a household is underwater—they owe more than the house is worth. Moving from a continuous dependent variable requires they use a logit function to fit their estimates. However, it also means that OLS regression assumptions of linearity, independence, and homoskedasticity are violated. No longer a linear relationship because the outcome is whether or not a household is underwater, the error terms now also vary with each other as well as with the predictors since the outcomes depend on the values of the predictors. Therefore, using logistic regression models, Faber and Ellen can predict the likelihood of a household to be underwater. They find that white homeowners that were able to hold onto their homes throughout the study period were less likely to be underwater than their Latino or black counterparts.

One interesting application of logistic regression was to study the likelihood of displacement in the gentrification literature. Freeman and Braconi (2004) ingeniously construct a dataset from the New York City Housing and Vacancy Survey to measure the displacement pressures on disadvantaged households, which they defined as households below the poverty line and those headed by non-college graduates. Within gentrifying neighborhoods, they found that poor households were 19 percent less likely to move out of gentrifying neighborhoods than poor households in non-gentrifying neighborhoods, while non-college heads of households were 15 percent less likely than their counterparts in other neighborhoods in the city (Freeman and Braconi 2004). Logistic regression allowed them to quantify the likelihood of displacement of residents in gentrifying neighborhoods, an important empirical question for which qualitative research methods did not allow. Timberlake and Johns-Wolfe (2017) use multinomial logistic regression to examine the impact of neighborhood ethnoracial composition on the likelihood that neighborhoods that could gentrify do gentrify over time. Their outcome variable has five categories, whether a neighborhood (1) did not gentrify, (2) gentrified white, (3) black, (4) Hispanic, or (5) mixed at the end of the study period. They found that the percentage of Black residents in 1980 was negatively associated with gentrified white and positively associated with gentrified black neighborhoods, and that percent Hispanic in 1980 was positively associated with gentrified Hispanic neighborhoods. The strength of this study was that it allowed to test for different pathways of gentrification, contesting the idea that all neighborhood gentrification involved minority communities becoming white. Their study gets at the nuance of how neighborhoods gentrify and indicates that neighborhoods can and do gentrifying along racial lines and that whites are less likely to gentrify predominantly black neighborhoods.

**DIFFERENT TYPES OF MODELS TO TAKE SPACE SERIOUSLY**

**Space**. With the advent of new tools of spatial analysis, Logan (2012) urges us to take space and spatial relations more seriously in urban sociology. He argues four main points. First, he asserts that analyses that use “place-level data become spatial when they introduce the relative locations of places as a major consideration”(Logan 2012:521). With a wider prevalence of a spatial data and methods for analyzing spatial relations and context like multilevel modeling and spatial regression, we are increasing required to incorporate spatial thinking into our work. Second, he argues that maps are powerful tools that reinforce findings on the extent of variation of in a place characteristic and demonstrate that variation has a spatial pattern. Good map making is a balance of directing the viewer to specific spatial patterns while also stimulating their interpretive imagination. Third, spatial analysis deals with terms like proximity, connection, exposure, access, all of which involve measuring distance (Logan 2012:521). However, distance should be treated as a complex sociological concept, not a fixed measure (Logan 2012:521). Finally, spatial dependence is the phenomenon that similar things tend to be clustered together. It has many potential sources, and “its interpretation therefore depends heavily on theory” (Logan 2012).

Much of my work on housing affordability, inequality, and segregation deals directly with spatial patterns within and across neighborhoods and cities. As I analyze these spatial relations, Logan reminds me that there are multiple levels of spatial analysis and that social phenomena tend to cluster at various levels and influenced by spatial patterns that have different sources at different levels. As such, regression analysis and the more advanced tools of hierarchical linear modeling and spatial regression are useful tools of quantitative analysis that allow me to systematically analyze spatial patterns and their effects.

**HIERARCHICAL LINER MODELS (HLM)**

Hierarchical Linear Models have grown more common with the increasing ease of computation and need for models that theoretically account for nested data structure. Nested data are common in the social sciences, for example, observations nested within individuals overtime, students within classrooms, and census tracts within counties within states. When linear methods are used with these data, they are unable to account for their complicated error structure that violates assumptions of independence and homoskedasticity. Error terms in nested data are clustered or autocorrelated because observations are exposed to similar treatments within groups—census tracts are from the same county—or outcomes at prior points are related to those at later points—as with individuals over time. HLM relaxes independence assumptions and allows for correlated error structure.

There are four basic assumptions of HLM:

1. That residuals are independent between clusters
2. There is independence of errors at cluster level from individual level
3. Level 1 residuals are normally distributed and have constant variances
4. Level 2 intercept and slope have multivariate normal distribution with a constant covariance matrix (Finch, Bolin, and Kelley 2014:36–37)

**First**, HLM assumes that random intercept and slope(s) are independent of each other across clusters. This means, for example, that there is no autocorrelation between clusters of level 2 variables and that clusters are independent of each other. **Second**, HLM assumes that errors at the cluster level are independent from errors at the individual level. This assumption deals with correlation of errors between cluster and individual. **Third**, HLM assumes that level 1 residuals are normally distributed. This is a similar assumption to what we see in OLS about homoskedasticity and normality. **Fourth**, HLM assumes that level 2 intercepts and slopes are normally distributed.

There are three key benefits of using HLM is that it leverages a pooled sample to improve estimation of individual effects. For example, a group of researchers in the 1980’s 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. Additionally, HLM allows for cross-level interactions that allow researchers to account for how variables at one level of the hierarchy affects variables at another. For example, how does a school’s average level of SES affect the relationship between SES and academic performance for individual students? Third, HLM allows researchers to partition variance-covariance components that point to how much of the variance in the dependent variable is accounted for by conext, and how much is reduced when new variables are introduced.

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

**Building an HLM model:** The first class of models can be thought of as unconstrained or null models. These models have no level 1 or level 2 predictors and let the dependent variable vary by context. This serves to verify variation across higher level contexts, which justifies the use of HLM. In so doing, researchers can calculate the interclass correlation (ICC) coefficient, the proportion of variance in the outcome explained between groups and is calculated by dividing the between group variance by the total variance. Depending on theoretical considerations and research questions, researchers can use either *random intercept models* or *random intercept and slopes models*. Random intercept models introduce level 1 and level 2 predictors, where level 1 variables vary randomly by context or means serve as predictors. These would answer questions about the average level 2 effects on level 1 predictors. Random intercept *and* slopes models let intercepts and slopes of level 1 variables vary across Level 2 units, model Level 1 intercepts and slopes using Level 2 predictors, and use cross level interaction. These class of HLM can be thought of as explanatory models that help researchers better account for variances in the dependent variable by interacting higher level variables or by allowing slopes to vary.

**An Application of HLM.** Emily Molina uses a multilevel logistic model that combines property level data on foreclosures and home sales with demographic data in census tracts to estimate (1) odds of a sale to an investor, (2) odds of a sale to a corporate investor, and (3) odds of property being flipped within study period. Using this method circumvents issues of correlated outcomes among properties within same census tract and produces more robust and accurate estimates of neighborhood characteristics on investor properties.

**SPATIAL REGRESSION**

Spatial regression is an attempt to take individual interactions and structure seriously. For example, while voter turnout may be more dependent on the individual characteristics of a specific voter, space is central to being able to predict risk of infection of the flu. An individual’s risk of catching the flu may partially depend on their individual health or the strength of their immune system, however, their risk of infection greatly increases with their spatial proximity to others that are infected. The salience of spatial analysis was made clear in John Snow’s famous analysis of the cholera outbreak in England in the 1850’s. By plotting the location of cholera deaths on a map, Snow recognized a spatial clustering around the Broad Street well, which had become infected from because of its proximity to a burial site from the cholera epidemic. His analysis shows how spatial analysis can help researchers better understand the relationship of individuals to space, and specifically in the cholera case, spatial correlation to casual analysis.

In the cholera analysis, the data take a geolocated point form, but regional data attributed to geographical areas—often called areal or lattice data—are perhaps just as common. When these values cluster in space, they create first-order spatial autocorrelation, increasing the standard errors and creating the likelihood of Type 1 error—rejecting the null hypothesis that there is a significant difference when there is not. Therefore, the goal of spatial regression is to correct for spatial autocorrelation, much in the same way that we correct for serial autocorrelation, by introducing into the regression equation a parameter that accounts for spatial correlation.

**Moran’s I.** Perhaps the most important step is to determine if there is spatial autocorrelation by calculating the Moran’s I statistic. This is done by measuring the cross-product of nearby spatial units. Moran’s I is calculated as follows:

1. Calculate the mean and variance for each spatial unit
2. Subtract the mean from nearby[[1]](#footnote-1) units to create deviations from the mean for each
3. Deviations of nearby units are multiplied to create a cross-product and then summed
4. Deviations above the mean indicate positive spatial clustering, while deviation below the mean indicate negative spatial clustering. Negative and positive values will sum closer to zero and indicate no spatial clustering.

The index results will allow the researcher to either reject or accept the null hypothesis that there is spatial autocorrelation. Moran’s I is a global statistic, that is, it is a single statistics computed for the entire dataset across all spatial units. It is a “global correlation of the values of an observation with those of its neighbor” (Ward and Gleditsch 2008:23). It is essentially the slope of the regression line in spatial regression. Anselin’s Local Moran’s I can be calculated in a similar way for each individual spatial unit, allowing for more granularity in understanding the local spatial relationships. The global versus local is a theme that will come into play with geographically weighted regression (GWR).

**Spatial Lag.** Spatial autocorrelation can be thought of as a nuisance or as substance, that is, there are research questions in which we the spatial relationship is what we are trying to better understand—substance. Take for example, the spatial distribution of democracy across the globe. We could run an OLS regression predicting a country’s polity score (score of how democratic they are) by their GDP as does Ward and Gleditsch (2008). However, when you analyze the standard errors of the model, you would see that they are clustered. More digging would reveal they are clustered by space, that is, spatial proximity to a democratic country influence’s the level of democracy in another country—democracy is not evenly distribute across space. Such an analysis requires that researchers correct for that unequal distribution.

Ordinary regression that does not consider spatial relationships will overestimate the standard errors and likely lead to Type 1 Error and is not appropriate for spatial analysis. This is because spatial autocorrelation violates the assumption of regression that the error terms are independent of each other. To correct for this, you can run a “spatially lagged dependent model,” which allows researchers to correct spatial dependence by incorporating a parameter of correlation of the dependent variable on the right side of the equation. This entails specifying a weight matrix, a rook, bishop, or queen criterion, which depends on a theoretical understanding of how space structures the relationship. Moreover, choice of a spatial weighting criterion can greatly affect the results. After incorporating the spatial lag variable, you can compare the standard errors of the spatially lagged model with those of the non-spatial model using Moran’s I to verify there was a reduction in the spatial clustering.

Two final thoughts. One important difference between spatial regression and fixed effects models that include dummies for spatial characteristics is that spatial regression gives you one statistic that is adjusted for changes in any of the spatial units. Fixed effects models are not as dynamic, that is, any change in a country’s level of democracy will not be reflected in the statistic of any other country’s fixed effect. In the democracy example, the coefficient of spatial lag is the effect of the level of democracy among connected countries on a country’s level of democracy. Finally, when using spatial lag models, researchers are attempting to better understand the spatial mechanism that structures the relationships in their data. In the democracy example, space plays a role in shaping how democratic a country is while in the cholera example space determines an individual risk of becoming infected.

**Spatial Error.** While spatial dependence in spatial lag models is substantive and researchers are concerned with the spatial mechanisms that structure relations across space, spatial dependence in spatial error models is considered a technical nuisance and researchers not substantively interested in how space structures relations. This occurs when spatial dependence enters a model through the errors and not through any systematic component of the model. Spatial error becomes something that a researcher wants to control for but is not interested in substantively. This means that rather than yi affecting yj directly, the errors of these spatial units are correlated, thus violating the basic assumption of regression. One example of such a case where you might not be interested in the substance of spatial dependence would be a hedonic housing price model that is interested in determining correct housing prices.

To correct for spatial correlation of the error terms, requires separating the error in the regression from the spatially correlated component and including that component in the regression. Much like spatial lag models that include the lagged dependent variable, the spatial error model includes the correlation of the error terms. It is important to note the substantive differences in what each model is doing, however. Spatial lag models simultaneously incorporate the feedback of between observations yi, yj, and yk, and so on, into the regression model. This means that the differences in the independent variables affect outcomes on the observations. Spatial error models only incorporate dependence that enters the model through error terms, implying that observations are really only related due to unmeasured factors that are for some unknown reason correlate across the distances among the observations. The question of which model to use is difficult to answer because there are few good tests for this, since the models are so different. However, Anselin (1988) determines that the Lagrange Multiplier test, which uses the maximum likelihood function, has “good power” and shows robustness” in determining which method to use. That said however, Ward (2008) argues that theory should take precedence in deciding which model to use. Recent scholarship has also advocated that Spatial Durbin Error Models be used, which incorporate both lagged dependent variables as well as spatial dependent disturbances.

**Geographically Weighted Regression (GWR).** Advances in computational power have allowed for social scientists to be more precise in their ability to understand spatial dependence. Spatial lag and spatial error models are really important for understanding the role of spatial dependence across space. However, the only provide one statistic that can be thought of as a global average. The best illustration of this is with the difference between Moran’s I and Local Moran’s I. Moran’s I is a global average of the correlation across all the smaller spatial units, while the local Moran’s I is a measure of spatial correlation for each spatial unit as compared to its neighbors. The local statistic is calculated on the pairs of points or areas which neighbor the spatial unit, while the global statistic is the average across all possible points or areas. Spatial lag and spatial error produce global regression coefficients, while geographically weighted regression produces local regression coefficients.

Obviously, GWR is computationally cumbersome, but provides researchers with better fitting regression lines for each spatial unit. Spatial residuals are lower and much less spatially dependent. It develops a local regression line that incorporates a specific kernel type, bandwidth, distance, and number of spatial units considered neighbors to model varying spatial relationships at the local level where coefficients are allowed to vary for each spatial unit. Categorical and nominal data should be avoided as well as dummy variables. Adaptive weighting schemes that adapt to the distance between points can be used to adjust for varying proximity, for example, in urban and suburban areas across cities.

**Sample size?**

**MISSING DATA**

Missing data can dramatically impact a study’s findings in two ways. First, missing data can dramatically affect the findings themselves. One could imagine how different results could be if all of the missing data were included. Second, missing data can make it difficult to claim that the findings are generalizable. Specifically, it contains the ability of researchers to make common inferences based on the data because there are so many missing observations. Missingness can occur for a few reasons. First, it can occur because of non-response. A survey respondent can refuse or forget to answer a question on a survey. This is especially true on some type of surveys where specific types of questions are less likely to be answered, for example, questions about income, sexual preference, or religious beliefs. Missingness can be due to attrition, that is, when a respondent drops out of a survey. This is very common in longitudinal studies, where subjects are lost over time. Finally, concerning administrative data, missingness can be the result of data unavailability or censorship. Perhaps data was censored to protect a respondent’s identity, because there were too few cases, or because the data is simply not available for whatever reasons.

**Types.** There are three main types of missing data. First, there is data that is classified as **Missing Completely at Random** (MCAR). This means that the missingness is not related to observed or unobserved characteristics of the sample and that there is no systematic observable difference between observations with full data and those with missing data. For example, MCAR would be violated if the average age of those who reported their income was younger than those who did not. This would be easy to test by simply dividing the sample into two—those who reported and those who did not—and testing for a difference in age. This type of missingness is ignorable.

A second type of missingness is called **Missing at Random (MAR),** which means that missingness is independent of unobserved characteristics, conditional upon observed data. For example, MAR would be satisfied if the probability of missing data on income were dependent upon marital status, but within each category of marital status, the probability of missing income was unrelated. There is no way to detect for whether MAR is satisfied, however, because data is not available to test whether a specific marital status group is less or more likely to report missing income. As long as the missingness can be fully accounted for by the variables in the data, this problem can be addressed. The final type of missingness is **Missing Not at Random** (MNAR), which means that it is not MCAR or MAR. An example of this might be missing data of males in filling out a survey depression because of their depression.

**Methods to Address.** There are several common methods for addressing ignorable missing data in the literature; however, only a few of them provide acceptable results. The most robust is listwise (casewise) deletion. This is when any cases or observations that are missing data are dropped or excluded from the analysis. This is the default process in Stata, for example. However, listwise deletion will yield biased samples for MAR, unless the missingness is only related to anther independent variable instead of the dependent variable. While pairwise deletion gives the researcher more data by using cases with data despite missing on other observations, it is not recommended it produces biased standard errors and test statistics. Similarly, a dummy variable adjustment procedure that regresses the dependent variable on a dummy for missingness, a variable with values for when data is missing, along with all other controls, produces biased estimates of coefficients.

Two other methods that are improvements over listwise deletion are multiple imputation and maximum likelihood. Maximum likelihood is a good alternative for when data are MAR. This method does not impute any data, but rather analyzes the data in full and creates maximum likelihood estimates. Maximum likelihood estimates of a parameter value for each observation that is the most likely to have resulted in the observed data. This is essentially a way around the missing data problem. The second method is multiple imputation, which imputes values for missing data by estimating them multiple times. This procedure includes error measures that will make values slightly different every time but will adjust standard errors upward. It then averages across estimates to produce consistent, asymptotically efficient, and asymptotically normal estimates.

Multiple imputation and maximum likelihood can be adjusted to correct for missing data that is non-ignorable. Another procedure is Heckman’s classic model for selection bias is another way to correct for missing non-ignorable data. Essentially, this procedure models the selection process along with the equation using a likelihood function.

**Other thoughts.** Porter and Eckland (2012) analyze the reasons why certain populations may or may not respond to survey questions. They had two important findings that should change how we think about missingness. First, they found that the group of people targeted for the survey matters: there are specific reasons that a group of people might avoid answering certain survey questions. For example, the reasons that people did not respond to the Religion among Academic Scientists (RAAS) survey was not because of their religiosity, but rathe because of the inability of the survey to capture their capture their views. Therefore, and this is their second major finding, missingness needs to be investigated so as to better understand the mechanisms driving missingness. It was not the assumed unwillingness of respondents to talk about their beliefs, but was the inability of the survey to be exhaustive, and this can dramatically change the study’s findings.

**References**

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Dordrecht.

Demaris, Alfred. 1995. “A Tutorial in Logistic Regression.” *Journal of Marriage and Family* 57(4):956–68.

Faber, Jacob W. and Ingrid Gould Ellen. 2016. “Race and the Housing Cycle: Differences in Home Equity Trends Among Long-Term Homeowners.” *Housing Policy Debate* 26(3):456–73.

Finch, W. Holmes, Jocelyn Bolin, and Ken Kelley. 2014. *Multilevel Modeling Using R*. Boca Raton: Taylor & Francis Group.

Kahane, Leo H. 2007. “Regression Basics.”

Logan, John R. 2012. “Making a Place for Space: Spatial Thinking in Social Science.” *Annual Review of Sociology* 38(1):507–24.

Long, J. Scott and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. Third edit. College Station, Texas : Stata Press Publication, StataCorp LP.

Menard, Scott W. 2002. *Applied Logistic Regression Analysis*. 2nd ed.. Thousand Oaks, Calif.: Thousand Oaks, Calif. : Sage Publications.

Porter, Jeremy and Elaine Ecklund. 2012. “Missing Data in Sociological Research: An Overview of Recent Trends and an Illustration for Controversial Questions, Active Nonrespondents, and Targeted Samples.” *The American Sociologist* 43(4):448–68.

Schroeder, Larry D., David L. Sjoquist, and Paula E. Stephan. 1986. *Understanding Regression Analysis*. Newbury Park: United States of America, California, Newbury Park: SAGE Publications, Inc.

Timberlake, Jeffrey M. and Elaina Johns-Wolfe. 2017. “Neighborhood Ethnoracial Composition and Gentrification in Chicago and New York, 1980 to 2010.” *Urban Affairs Review* 53(2):236–72.

Ward, Michael Don and Kristian Skrede Gleditsch. 2008. *Spatial Regression Models*. Thousand Oaks: Thousand Oaks : Sage Publications.

1. Nearby units will vary depending on specification of spatial weight matrices. [↑](#footnote-ref-1)