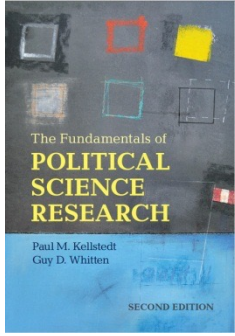
The Fundamentals of Political Science Research, 3rd Edition

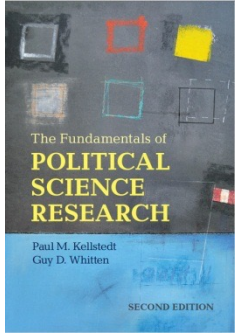
Chapter 9: Bivariate Regression Models

Chapter 9 Outline



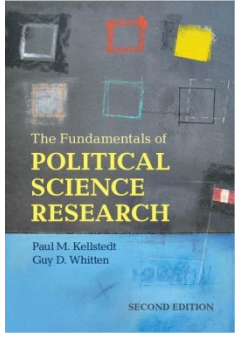
- Two-Variable
- Fitting a Line: Population ←→ Sample
- Which line fits best? Estimating the regression line
- Measuring Our Uncertainty about the OLS Regression Line and parameters
- Assumptions, More Assumptions, and Minimal Mathematical Requirements

Fitting a Line: Population ←→ Sample

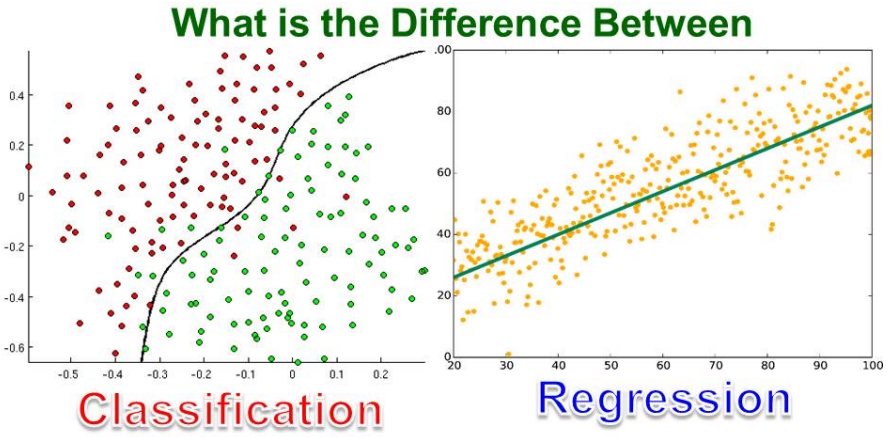


- Two-variable regression: fit the "best" line through a scatter plot of data
- This line, which is defined by its slope and y-intercept, serves as a statistical model of reality.
  - Q: Why draws a straight line?

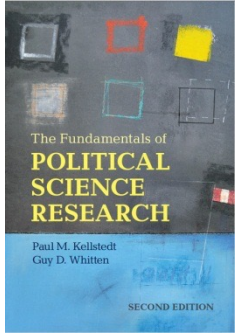
Fitting a Line: Population ←→ Sample



- Two-variable regression: fit the "best" line through a scatter plot of data
- This line, which is defined by its slope and y-intercept, serves as a statistical model of reality.
  - Q: Why draws a straight line? Because our brain can't digest complicated relationships even though they are truer in our social lives.

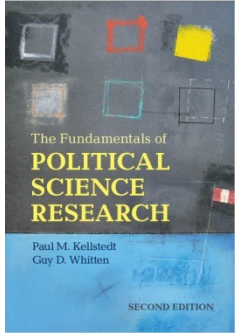


Fitting a Line: Population ←→ Sample



- Two-variable regression: fit the "best" line through a scatter plot of data
- This line, which is defined by its slope and y-intercept, serves as a statistical model of reality.
- You may remember from a geometry course the formula for a line expressed as Y=mX + b, where b is the y-intercept and m is the slope
- For a one-unit increase (run) in X, m is the corresponding amount of rise in Y (or fall in Y, if m is negative).
- Together these two elements (m and b) are described as the line's parameters.

Population regression model

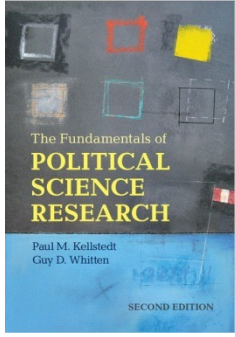


- In a two-variable regression model, we represent the y-intercept parameter by the Greek letter alpha (α) and the slope parameter by the Greek letter beta (β)
  - Y is the dependent variable and X is the independent variable.
- Our theory about the underlying population in which we are interested is expressed in the population regression model:

$$Y_i = \alpha + \beta X_i + u_i$$

- **Error term?**

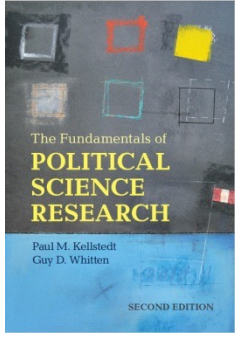
Population regression model



- In a two-variable regression model, we represent the y-intercept parameter by the Greek letter alpha (α) and the slope parameter by the Greek letter beta (β)
  - Y is the dependent variable and X is the independent variable.
- Our theory about the underlying population in which we are interested is expressed in the population regression model:

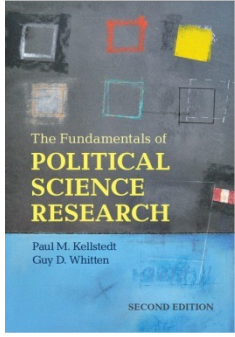
$$Y_i = \alpha + \beta X_i + u_i$$

- **Error term**: where u_i is the stochastic or random component of our dependent variable.
  - We have this term because we do not expect all of our data points to line up perfectly on a straight line.
- Thus we think about the values of our dependent variable Y_i as having a systematic component, $\alpha + \beta X_i$, and a stochastic component, u_i.



Sample regression model

- We rarely work with population data.
- To distinguish between these two, we place hats (^) over terms in the sample regression model that are estimates of terms from the unseen population regression model.
- Because they have hats, we can describe $\hat{\alpha}$ and $\hat{\beta}$ as being ``parameter estimates.''



Y-hat

$$sample\ regression\ model\text{:}\ Y_i = \hat{\alpha} + \hat{\beta}X_i + \hat{u}_i$$

- In the sample regression model, α, β, and u_i get hats, but Y_i , and X_i do not.
- This is because Y_i and X_i are values for cases in the population that ended up in the sample. As such, Y_i and X_i are values are *measured* rather than estimated.
- For each X_i value, we use $\hat{\alpha}$ and $\hat{\beta}$ to calculate the predicted value of Y_i, which we call $\widehat{Y_i}$, where

$$\widehat{Y_i} = \hat{\alpha} + \hat{\beta}X_i$$

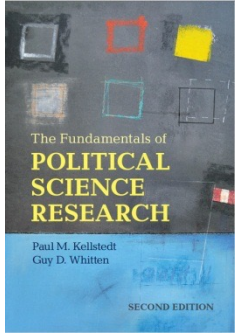
- This can also be written in terms of expectations,

$$E(Y|X_i) = \hat{\alpha} + \hat{\beta}X_i$$

- We can also write our model as

$$Y_i = \widehat{Y_i} + \hat{u}_i$$

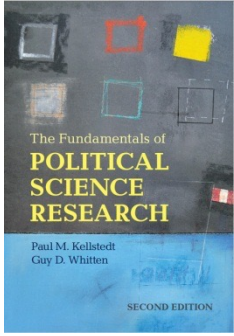
Residual



- $Y_i = \widehat{Y}_i + \hat{u}_i$ can also be rewritten in terms of $\hat{u}_i$ to get a better understanding of the estimated stochastic component:

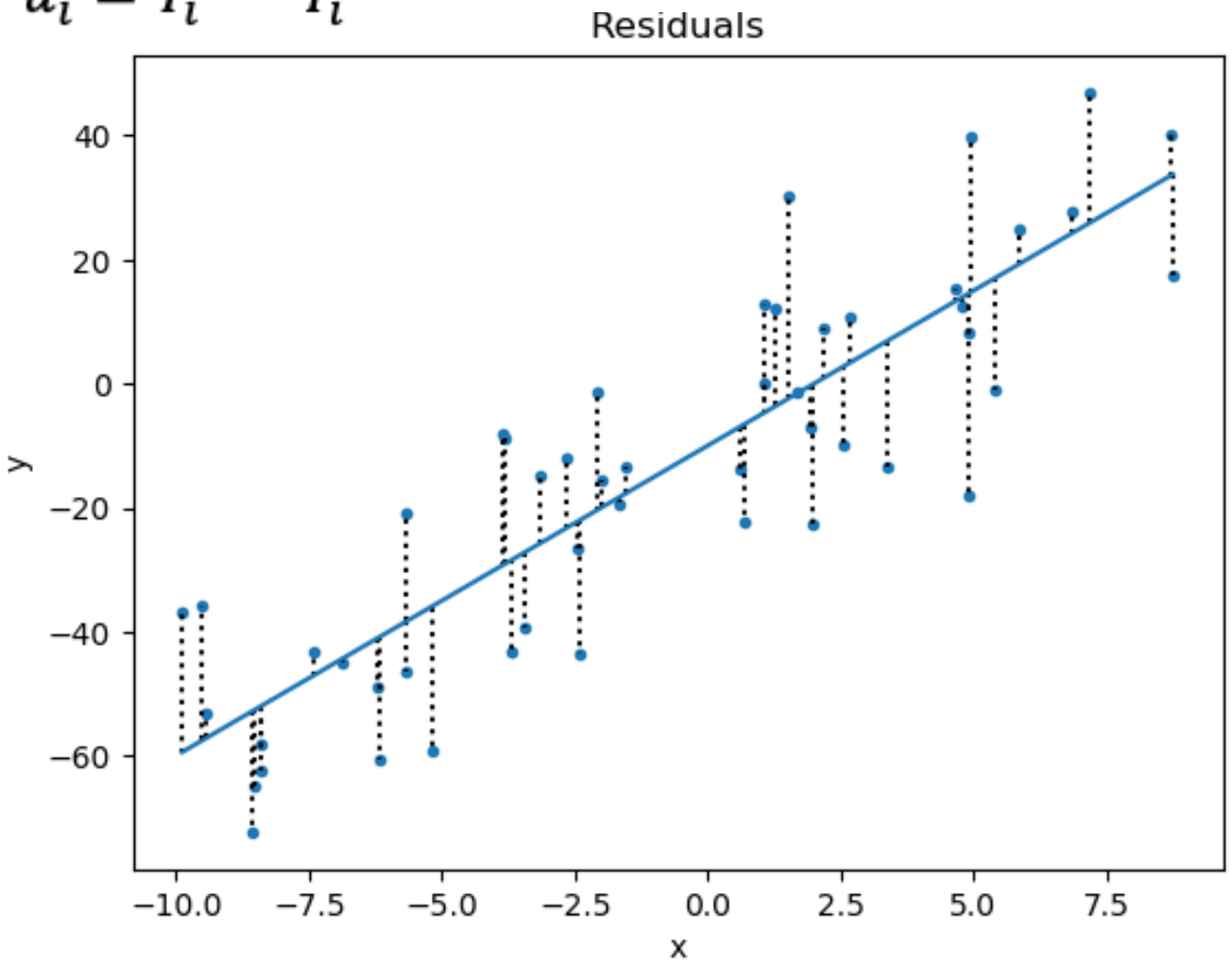
$$\hat{u}_i = Y_i - \widehat{Y}_i$$

- The difference between the actual value of the dependent variable (Y_i) and the predicted value of the dependent variable ($\widehat{Y}_i$)from our two-variable regression model.
- Residual/error term:
  - Another name for the estimated stochastic component is the residual. "Residual" is another word for "leftover," and this is appropriate, because $\hat{u}_i$ is the leftover part of Y_i after we have drawn the line defined by $\widehat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$.

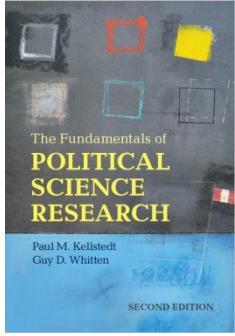


Residuals for the regression lines

$$\hat{u}_i = Y_i - \hat{Y}_i$$

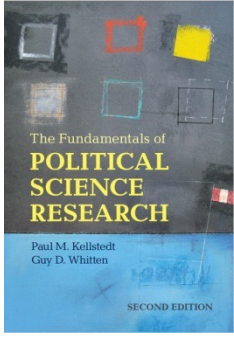


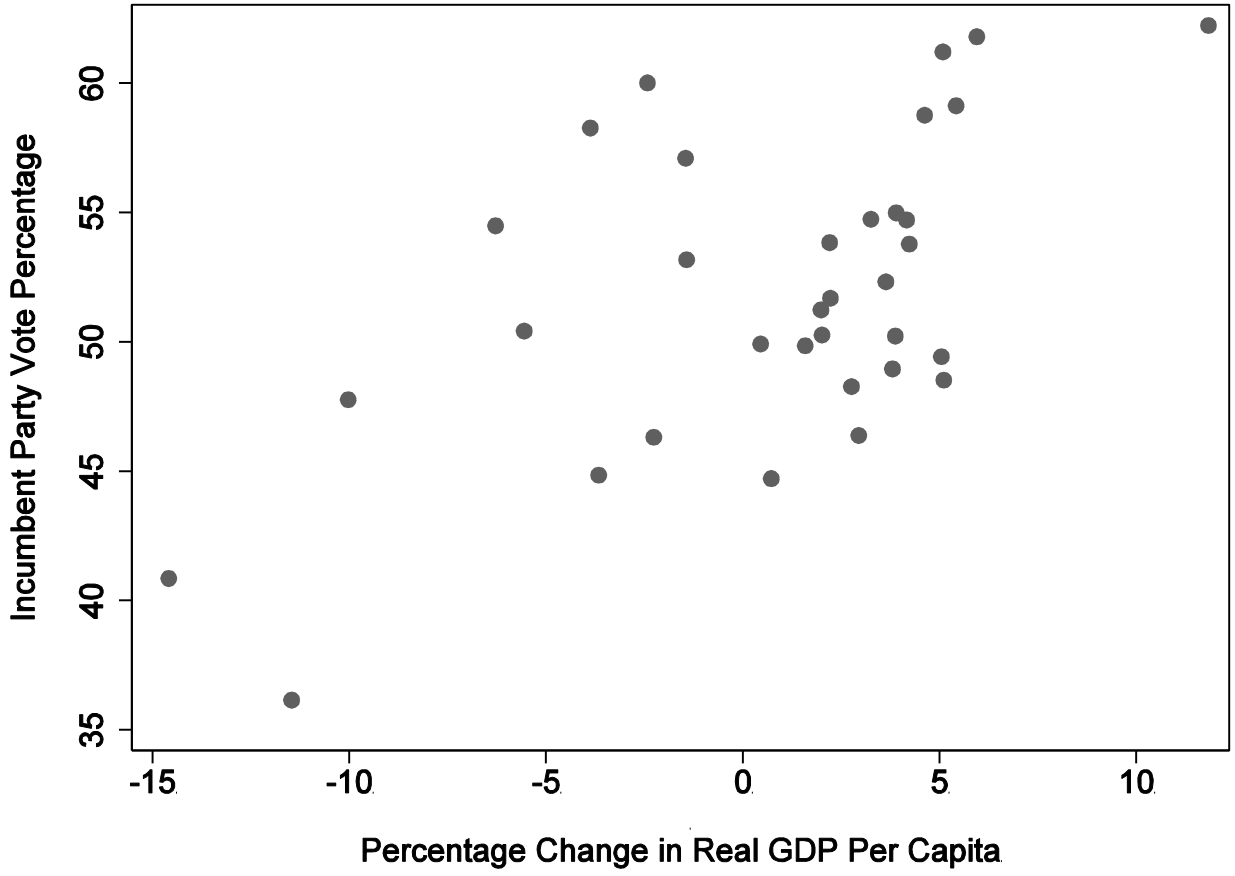
Which line fits best? Estimating the regression line



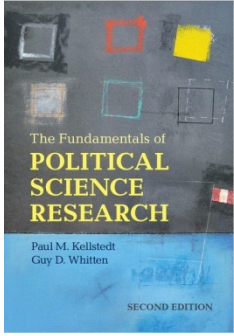
- In estimating a regression line, our task is to draw a straight line that describes the relationship between our independent variable X and our dependent variable Y.
- We clearly want to draw a line that comes as close as possible to the cases in our scatter plot of data.
- But how do we decide which line is best?

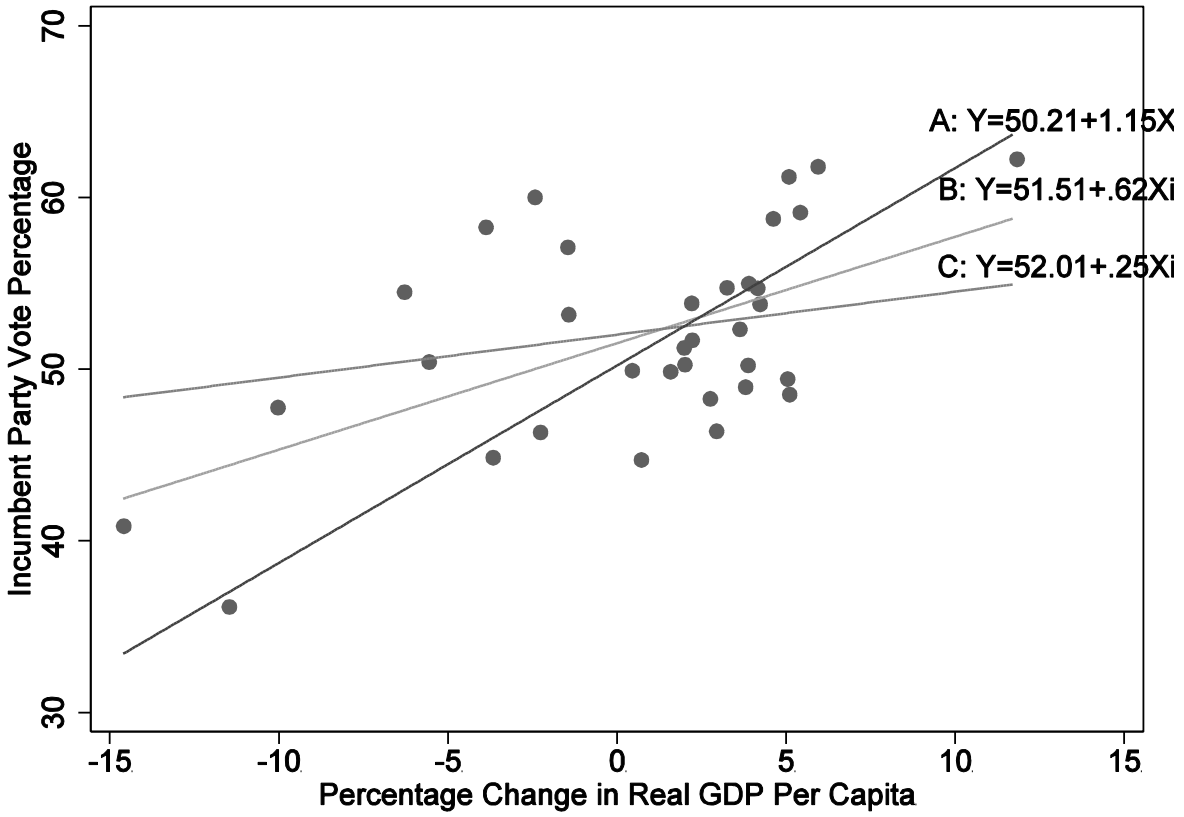
Scatter plot of change in GDP and incumbent-party vote share



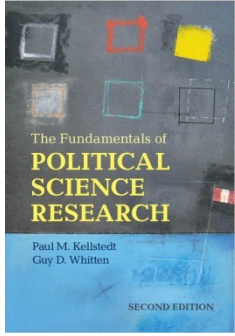


Three possible lines





Which line is “best?”



- One possibility is to add together the absolute value of the residuals for each line:

$$\sum_{i=1}^{n} |\hat{u}_i|$$

- Another possibility is to add together the squared value of each the residuals for each line:

$$\sum_{i=1}^{n} \hat{u}_i^{\,2}$$

- With either choice, we want to choose the line that has the smallest total value.

Measures of total residuals for three different lines

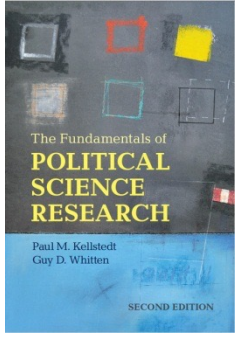


Table 8.1: Measures of total residuals for three different lines

Line	Parametric formula	$\sum_{i=1}^{n} \lvert\hat{u}_i\rvert$	$\sum_{i=1}^{n} \hat{u}_i^2$
A	$Y = 50.21 + 1.15X_i$	149.91	1086.95
B	$Y = 51.51 + 0.62X_i$	137.60	785.56
C	$Y = 52.01 + 0.25X_i$	146.50	926.16

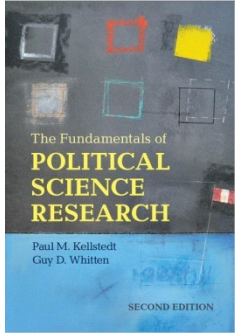
- B does a better job of fitting the data than lines A and C
- Although the absolute-value calculation is just as valid as the squared residual calculation, statisticians have tended to prefer the latter (both methods identify the same line as being ``best'').

- Thus we draw a line that "minimizes the sum of the *squared* residuals"

$$\sum_{i=1}^{n} \hat{u}_i^{\,2}$$

- This technique for estimating the parameters of a regression model is known as ``ordinary least-squares'' (OLS) regression.

OLS parameter estimates

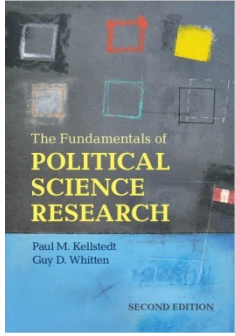


- For a two-variable OLS regression, the formulae for the parameter estimates of the line are:

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \\ \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X}.\end{aligned}$$

- How do we derive these estimands?

OLS parameter estimates

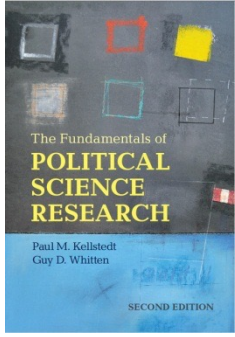


- For a two-variable OLS regression, the formulae for the parameter estimates of the line are:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \\ \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X}. \end{aligned}$$

- How do we derive these estimands? Ans: Calculus

OLS parameter estimates

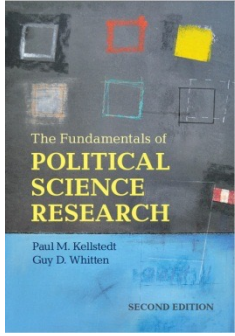


- For a two-variable OLS regression, the formulae for the parameter estimates of the line are:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

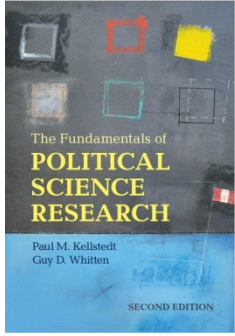
- How do we derive these estimands?
- Numerator: If we examine the formula for β, we can see that the numerator is the same as the numerator for calculating the covariance between X and Y.
  - Thus the logic of how each case contributes to this formula is the same. → shows the direction of relationship
- Denominator: The denominator in the formula for $\hat{\beta}$ is the sum of squared deviations of the X_i values from the mean value of X ($\bar{X}$).
  - Thus, for a given covariance between X and Y, the more (less) spread out X is, the less (more) steep the estimated slope of the regression line.

Measuring Our Uncertainty about the OLS Regression Line



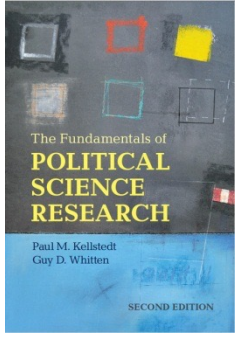
- With an OLS regression model, we have several different ways in which to measure our uncertainty.
- There are two uncertainties: which two?

Measuring Our Uncertainty about the OLS Regression Line



- With an OLS regression model, we have several different ways in which to measure our uncertainty.
- There are two uncertainties: which two?
- We discuss these measures 1) in terms of the overall fit between X and Y first and 2) then discuss the uncertainty about individual parameters.
- Our uncertainty about individual parameters is used in the testing of our hypotheses.

Goodness-of-Fit: Root Mean-Squared Error (RMSE)



- Uncertainty about the "model": goodness-of-fit
- Measures of the overall fit between a regression model and the dependent variable are called "goodness-of-fit measures."
- One of the most intuitive of these measures is root mean-squared error (RMSE):

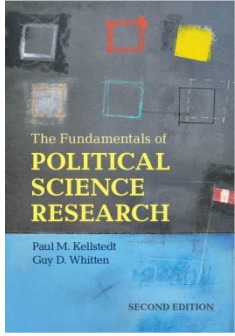
$$\text{root MSE} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n}}.$$

- The squaring and then taking the square root of the quantities in this formula are done to adjust for the fact that some of our residuals will be positive and some will be negative.

$$\hat{u}_i = Y_i - \widehat{Y}_i$$

- RMSE is particularly useful in evaluating prediction accuracy
  - E.g. Our model is off by 4.95 points in predicting the DV (percentage of the incumbent party's vote share)
  - You need a comparison set (compared to what?)

Goodness-of-Fit: R-Squared Statistic



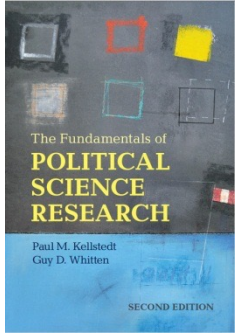
- Another popular indicator of the model's goodness-of-fit is the R-squared statistic (typically written as R^2).
- The R^2 statistic ranges between zero and one, indicating the proportion of the variation in the dependent variable that is accounted for by the model.
- The formula for total variation in Y (areas a and b in the Venn diagram figure, also known as the total sum of squares (TSS), is

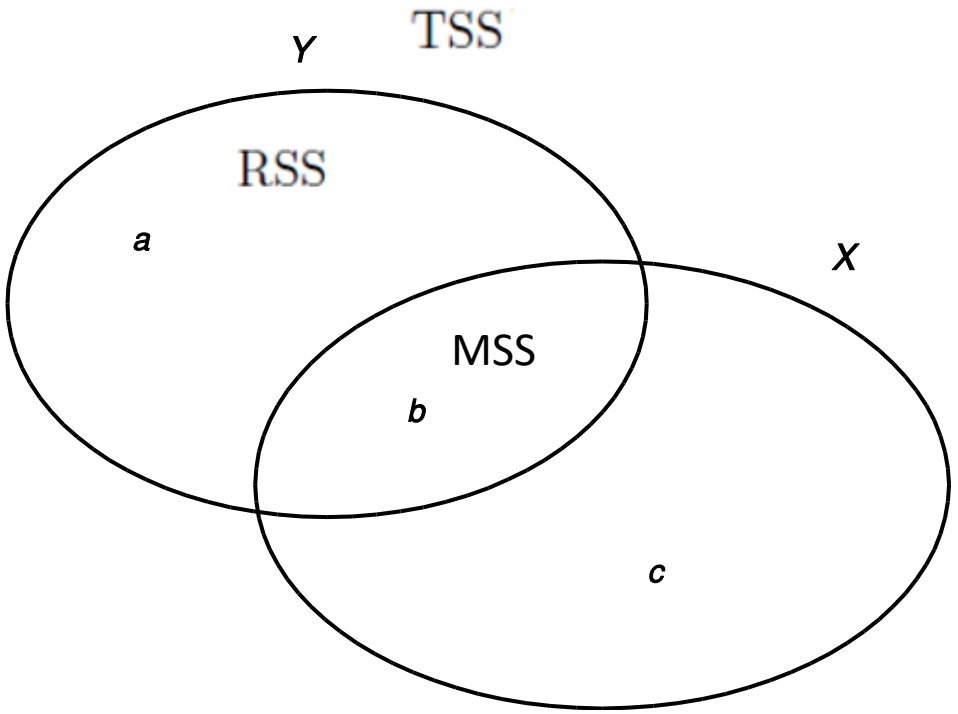
$$\text{TSS} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

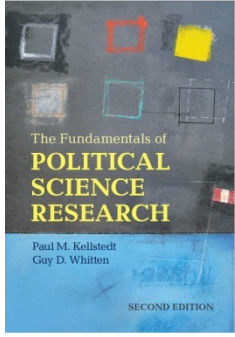
- The formula for the residual variation in Y, area a, that is not accounted for by X, called the residual sum of squares (RSS), is

$$\text{RSS} = \sum_{i=1}^{n} \hat{u}_i^2.$$

Venn diagram of variance and covariance for X and Y







Goodness-of-Fit: R-Squared Statistic

- Once we have TSS and RSS two quantities, we can calculate the R^2 statistic as

$$\frac{\text{TSS}}{\text{TSS}} - \frac{\text{RSS}}{\text{TSS}} \qquad R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

- The formula for the other part of TSS that is not the RSS, called the model sum of squares (MSS), is

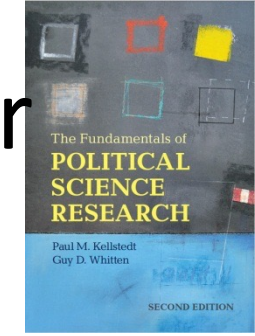
$$\text{MSS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

- This can also be used to calculate R^2 as

$$R^2 = \frac{\text{MSS}}{\text{TSS}}.$$

- E.g. Our model explains 33% of the variation in our DV. The larger, the better
- Again, you need a comparison set

Uncertainty about the "Parameter Estimates"

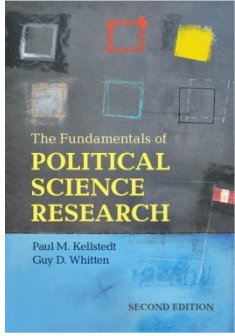


- In Chapter 7 we discussed how we use the normal distribution (supported by the central limit theorem) to estimate confidence intervals for the unseen population mean from sample data.
- We will use the same concept to estimate uncertainty in parameters
- The formulae for estimating confidence intervals are

$$\hat{\beta} \pm [t \times \text{se}(\hat{\beta})],$$
$$\hat{\alpha} \pm [t \times \text{se}(\hat{\alpha})],$$

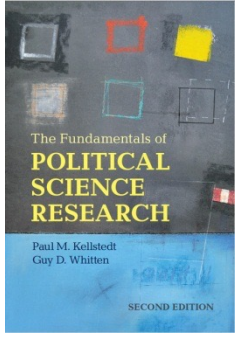
- where the value for t is determined from the t-table such as the one provided in Appendix B.

Traditional OLS hypothesis testing



- Slope estimate: Although we can test hypotheses about either the slope or the intercept parameter, we are usually more concerned with tests about the slope parameter.
- We build a null hypothesis and an alternative hypothesis
- $\hat{B} = 0$: In particular, we are usually concerned with testing the hypothesis that the population slope parameter is equal to zero → a flat line
  - A flat line means no covariation between X and Y
- The logic of this hypothesis test corresponds closely with the logic of the bivariate hypothesis tests introduced in Chapter 7.
  - We observe a sample slope parameter, which is an estimate of the population slope.
  - Then, from the value of this parameter estimate, the confidence interval around it, and the size of our sample, we evaluate how likely it is that we observe this sample slope if the true but unobserved population slope is equal to zero.
  - If the answer is "very likely," then we conclude that the population slope is equal to zero.

Two-Tailed Hypothesis Tests



- The most common form of statistical hypothesis tests about the parameters from an OLS regression model is a two-tailed hypothesis test that the slope parameter is equal to zero.
- It is expressed as

$$H_0: \beta = 0, \quad \text{~relationship}$$
$$H_1: \beta \neq 0, \quad \text{relationship}$$

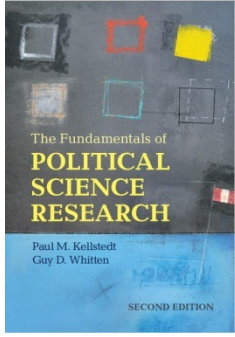
- where H_0 is the null hypothesis and H_1 is the alternative hypothesis.
- Note that these two rival hypotheses are expressed in terms of the "slope parameter" from the population regression model.
- To test which of these two hypotheses is supported, we calculate a t-ratio in which β is set equal to the value specified in the null hypothesis (in this case zero because $H_0: \beta=0$), which we represent as β^*:

$$t_{n-k} = \frac{\hat{\beta} - \beta^*}{\text{se}(\hat{\beta})}.$$

Cross Tab

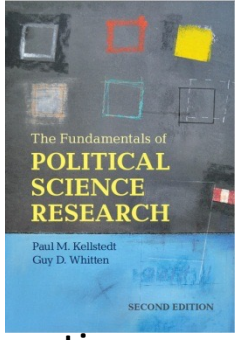
$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{se}(\bar{Y}_1 - \bar{Y}_2)}$$

The Relationship between "Confidence Intervals" and "Two-Tailed Hypothesis Tests"



- These two methods for making inferences are mathematically related to each other.
- We can tell this because they each rely on the t-table.
- The relationship between the two is such that, if the 95% confidence interval does not include a particular value (e.g., 0), then the null hypothesis that the population parameter equals that value (a two-tailed hypothesis test) will have a p-value smaller than .05.

One-Tailed Hypothesis Tests



- Most political science hypotheses are that a parameter is either positive or negative and not just that the parameter is different from zero. This is what we call a ``directional hypothesis.''
- When our theory leads to a directional hypothesis, it is expressed as

  - H_0: B = 0
  - H_1: B > 0 or B < 0

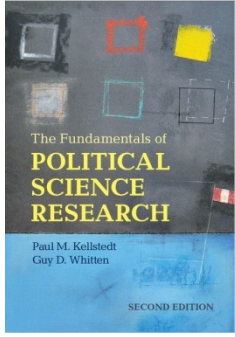
- where H_0 is the null hypothesis and H_1 is the alternative hypothesis.

- As was the case with the two-tailed test, these two rival hypotheses are expressed in terms of the slope parameter from the population regression model.

- To test which of these two hypotheses is supported, we calculate a t-ratio where β is set equal to the value specified in the null hypothesis, which we represent as β^*

- <u>The One-tailed test is easier to achieve statistical significance than the two-tailed test because the rejection region is wider</u>, but political scientists most of the time still use two-tailed even though their hypothesis is directional

OLS Assumptions about the Population Stochastic Component u_i

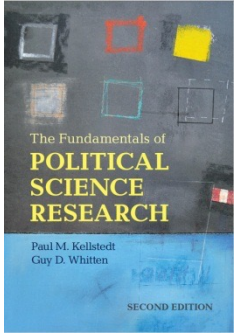


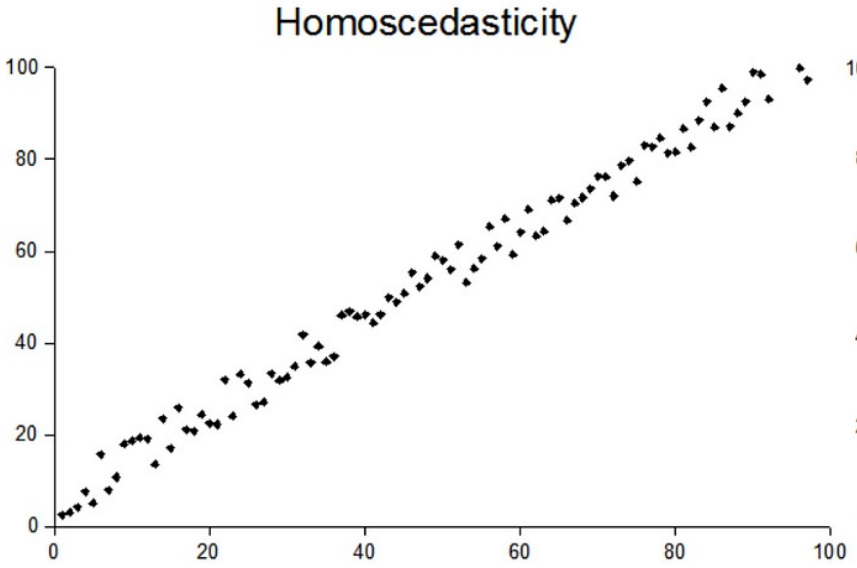
- The most important assumptions about the population stochastic component u_i are about its distribution. These can be summarized as

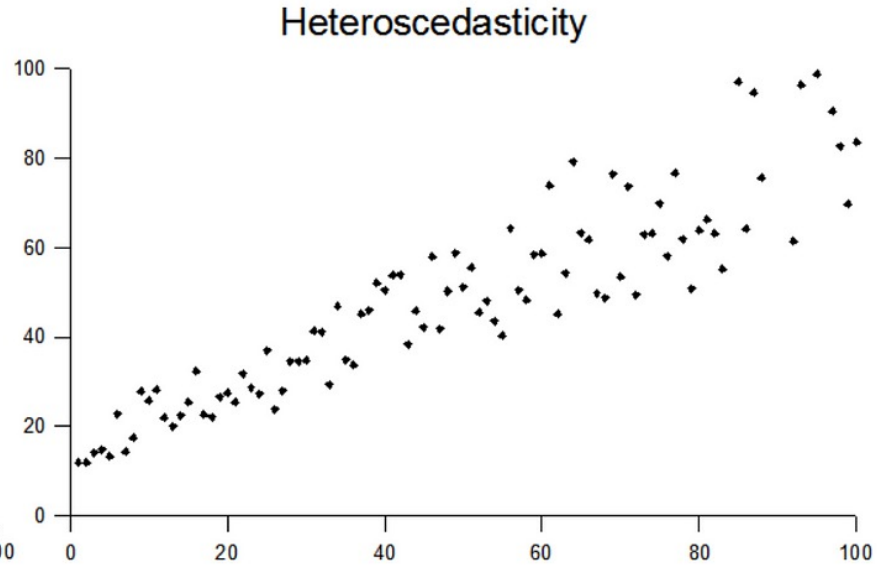
$$u_i \backsim N(0, \sigma^2)$$

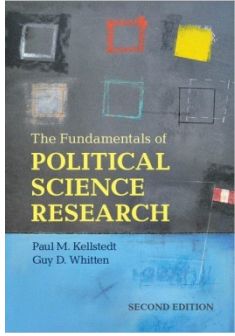
- which means that we assume that u_i is distributed normally ($\backsim$ N) with the mean equal to zero and the variance equal to σ^2.
- This compact mathematical statement contains three of the five assumptions that we make about the population stochastic component any time we estimate a regression model:

1. u_i Is normally distributed→ so we can use t-statistics (central limit theorem and the 68-95-99 rule)
   - Often violated
2. E(u_i)=0: (zero mean error) zero bias for "in average" prediction;
   - no omitted variables → often violated
3. u_i has variance σ^2: Homoscedasticity; each unit has the same variance (E.g. violated if some elections are harder; clusters)
   - Often violated
4. cov(u_i, u_j)=0 $\forall\ i \neq j$: no spatial and temporal autocorrelation
   - Often ignored
5. X values are measured without error
   - We assume any variability from our regression line is due to the random component μ and not to measurement problems in X.
   - E.g. Real GDP per capita


The Fundamentals of
POLITICAL
SCIENCE
RESEARCH
Paul M. Kellstedt
Guy D. Whitten
SECOND EDITION

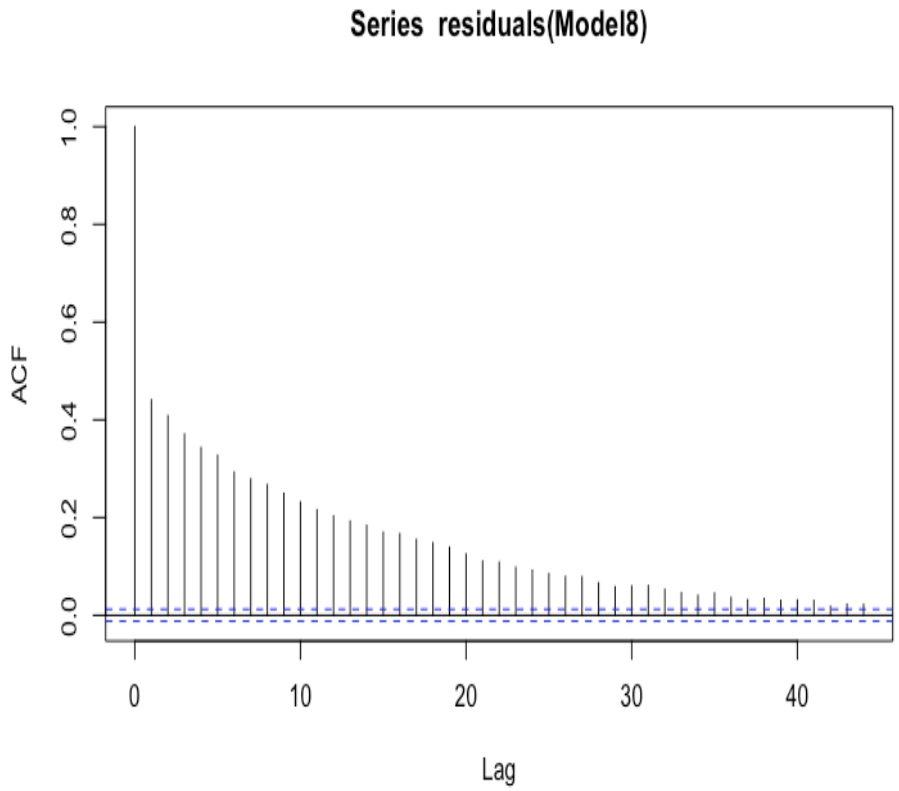

Homoscedasticity
100
80
60
40
20
0
0
20
40
60
80
100


Heteroscedasticity
100
80
60
40
20
0
0
20
40
60
80
100

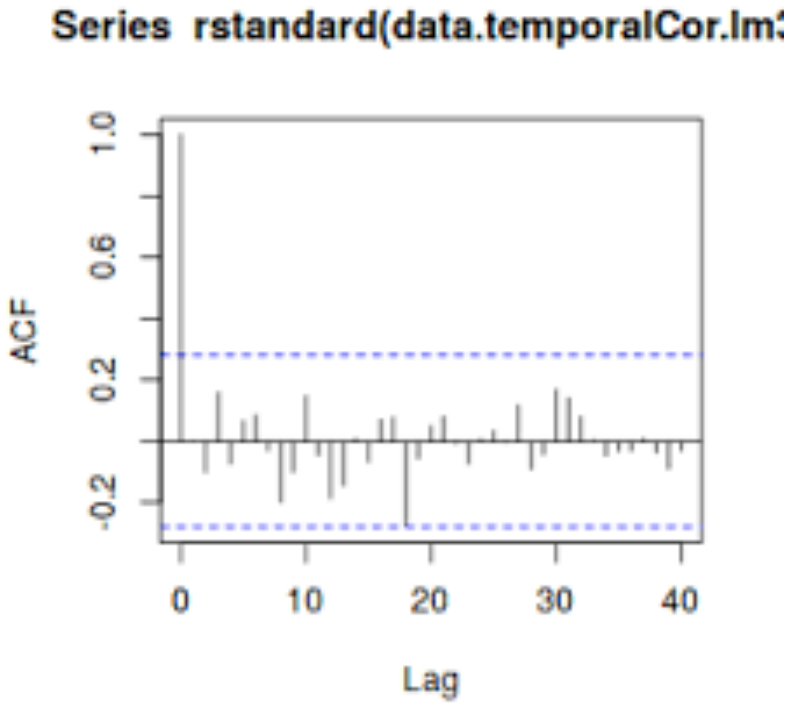


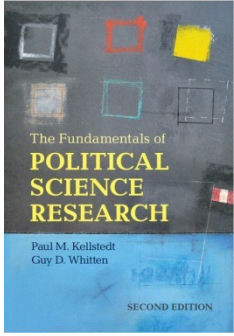
Temporal Autocorrelation

Bad

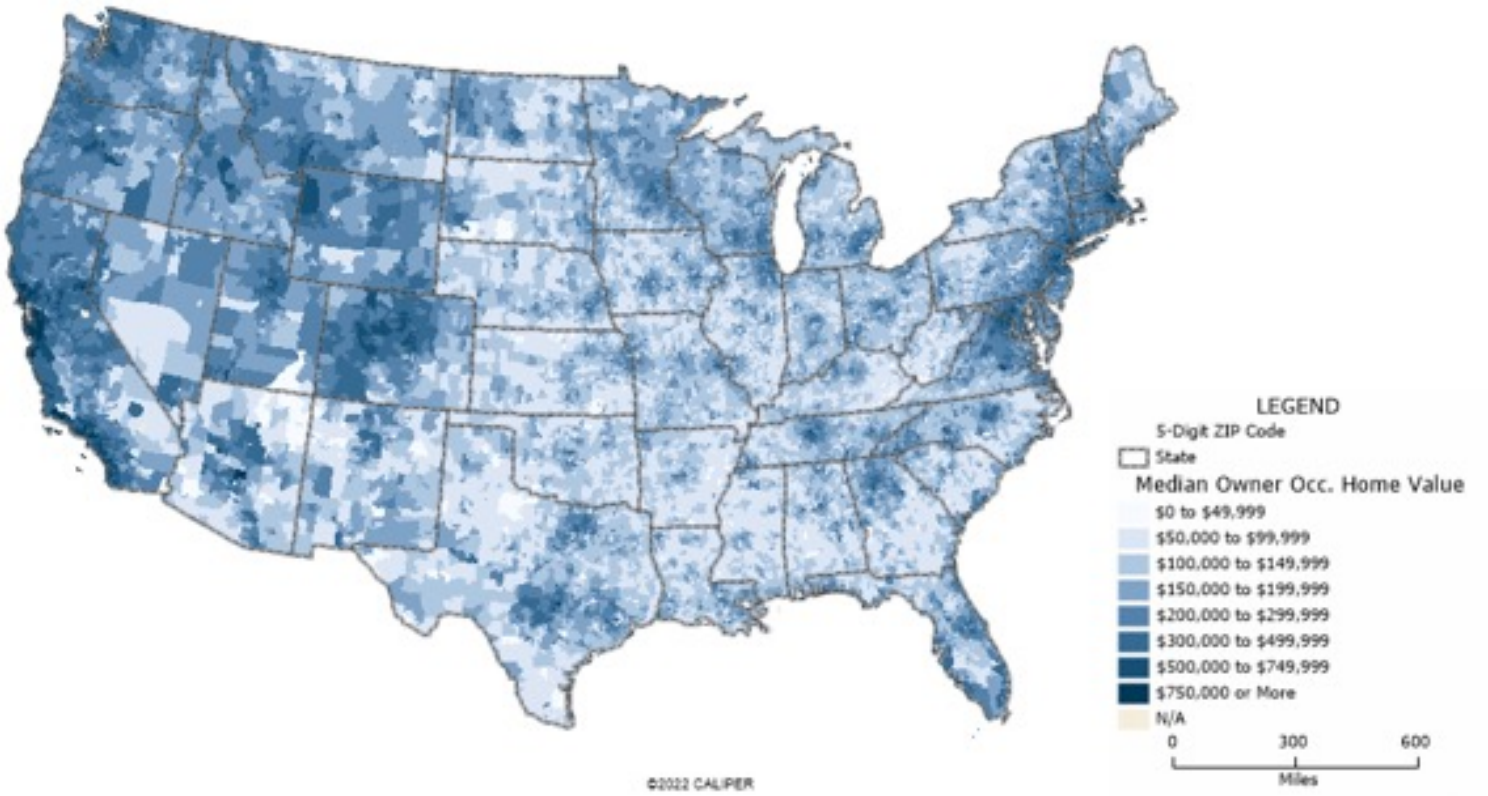


Good

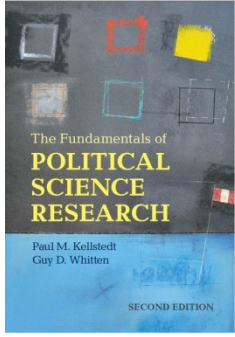




Spatial Autocorrelation

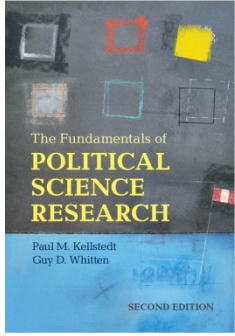


Assumptions about Our Model Specification



- We have the correct model specification
  - No Causal Variables Left Out; No Non-causal Variables Included
  - Parametric linearity: monotonicity
    - One unit increase in the change of GDP leads to 10% increase in the probability of conflict (across all x values)
    - Is it true in reality?

Minimal Mathematical Requirements



- For a two-variable regression model, we have two minimal requirements that must be met by our sample data before we can estimate our parameters.
- We will add to these requirements when we expand to multiple regression models.
  - *X* Must Vary
  - $n > k$