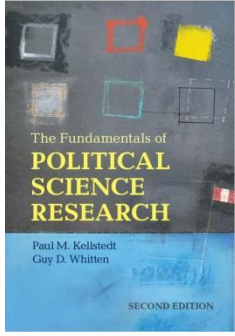
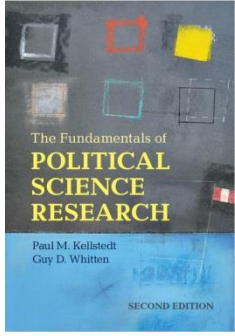
The Fundamentals of Political Science Research, 2nd Edition

Chapter 11: Multiple Regression Model Specification



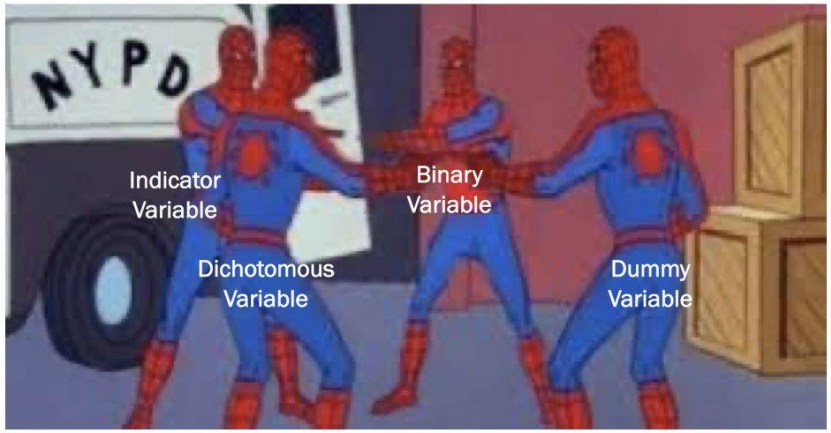
Chapter 11 Outline

- Basic data handling tips (indep. var.)
- Being Smart with Dummy Independent Variables in OLS
- Outliers and Influential Cases in OLS
- Multicollinearity

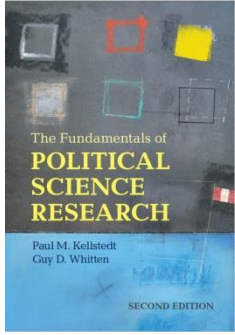


1. Dummy Variables

- When we have categorical independent variables that take on one of two possible values for all cases.
- Categorical variables like this are commonly referred to as "dummy variables."
- The most common form of dummy variable is one that takes on values of one or zero.
- These variables are also sometimes referred to as "indicator variables" when a value of one indicates the presence of a particular characteristic and a value of zero indicates the absence of that characteristic.



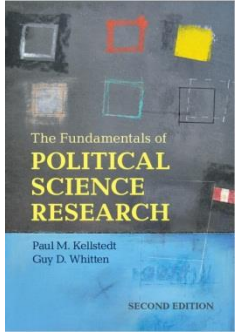
Hillary Clinton Thermometer Scores Example



- Data from 1996 NES
- Dependent variable: Hillary Clinton Thermometer Rating
- Independent variables: Income and Gender
- Each respondent's gender was coded as equaling either 1 for “male” or 2 for “female.”
- Although we could leave this gender variable as it is and run our analyses, we chose to use this variable to create two new dummy variables, “male” equaling 1 for “yes” and 0 for “no,” and “female” equaling 1 for “yes” and 0 for “no.”
- Our first inclination is to estimate an OLS model in which the specification is the following:

$$\text{Hillary Thermometer}_i = \alpha + \beta_1 \text{Income}_i + \beta_2 \text{Male}_i + \beta_3 \text{Female}_i + u_i.$$

The dummy trap



- R will report this result from the following model instead of what we asked for:

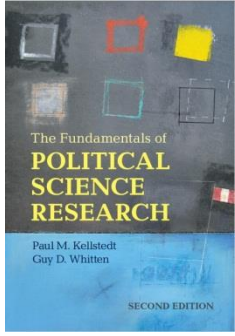
$$\text{Hillary Thermometer}_i = \alpha + \beta_1 \text{Income}_i + \beta_3 \text{Female}_i + u_i.$$

- This is the case because of the issue of "perfect multicollinearity."
- The reason that we have failed to meet this is that, for two of the independent variables in our model, Male and Female, it is the case that

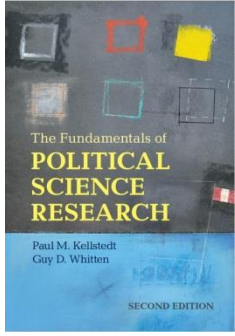
$$\text{Male}_i + \text{Female}_i = 1 \ \forall \ i.$$

- In other words, our variables Male and Female are perfectly correlated.

Avoiding the dummy trap



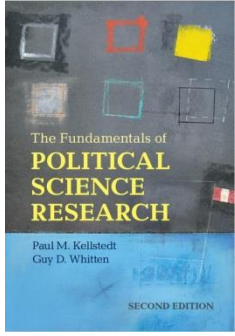
- To avoid the dummy-variable trap, we have to omit one of our dummy variables.
  - If we have a categorical variable with two categories only (male/female), turn them into one variable
  - If we have a categorical variable with more than two categories (Protestant, Catholic, Jewish, non-religious), turn them into multiple binary variables



Two models of the effects of gender and income on Hillary Clinton Thermometer scores

Independent variable	Model 1	Model 2
Male	–	-8.08^{***}
		(1.50)
Female	8.08^{***}	–
	(1.50)	
Income	-0.84^{***}	-0.84^{***}
	(0.12)	(0.12)
Intercept	61.18^{***}	69.26^{***}
	(2.22)	(1.92)
n	1542	1542
R^2	.06	.06

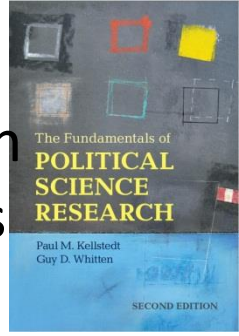
Notes: The dependent variable in both models is the respondent's thermometer score for Hillary Clinton. Standard errors in parentheses. Two-sided t-tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.



Using Dummy Variables to Test Hypotheses about a Categorical Independent Variable with "More Than Two Values"

Value	Category	Frequency	Percent
0	Protestant	683	39.85
1	Catholic	346	20.19
2	Jewish	22	1.28
3	Other	153	8.93
4	None	510	29.75

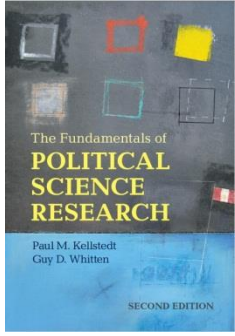
- When we have a categorical variable with more than two categories and we want to include it in an OLS model, things get more complicated.
- The best strategy for modeling the effects of such an independent variable is to include a dummy variable for all values of that independent variable *except one*.
  - In R fastDummies::**dummy_cols**(fastDummies_example)



The same model of religion and income on Hillary Clinton Thermometer scores with different reference categories

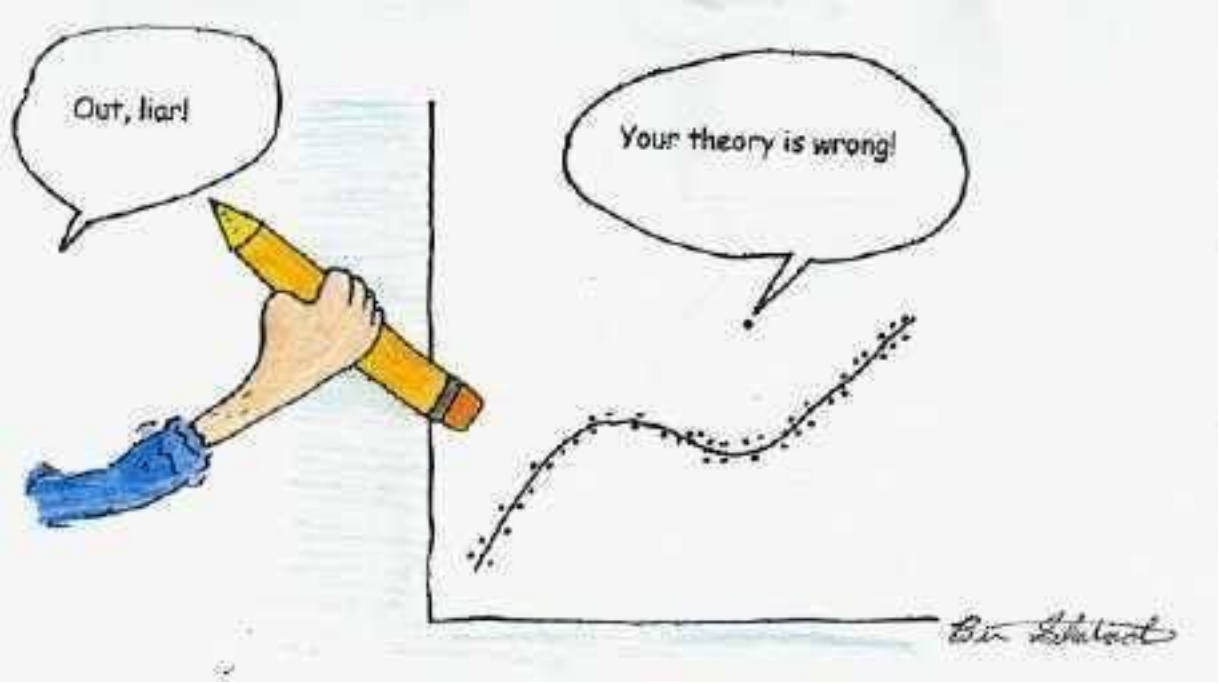
Independent variable	Model 1	Model 2	Model 3	Model 4	Model 5
Income	-0.97^{***}	-0.97^{***}	-0.97^{***}	-0.97^{***}	-0.97^{***}
	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)
Protestant	-4.24^{*}	-6.66^{*}	-24.82^{***}	-6.30^{**}	
	(1.77)	(2.68)	(6.70)	(2.02)	
Catholic	2.07	-0.35	-18.51^{**}		6.30^{**}
	(2.12)	(2.93)	(6.80)		(2.02)
Jewish	20.58^{**}	18.16^{**}		18.51^{**}	24.82^{***}
	(6.73)	(7.02)		(6.80)	(6.70)
Other	2.42		-18.16^{**}	0.35	6.66^{*}
	(2.75)		(7.02)	(2.93)	(2.68)
None		-2.42	-20.58^{**}	-2.07	4.24^{*}
		(2.75)	(6.73)	(2.12)	(1.77)
Intercept	68.40^{***}	70.83^{***}	88.98^{***}	70.47^{***}	64.17^{***}
	(2.19)	(2.88)	(6.83)	(2.53)	(2.10)
n	1542	1542	1542	1542	1542
R^2	.06	.06	.06	.06	.06

Notes: The dependent variable in both models is the respondent's thermometer score for Hillary Clinton. Standard errors in parentheses.
Two-sided t-tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

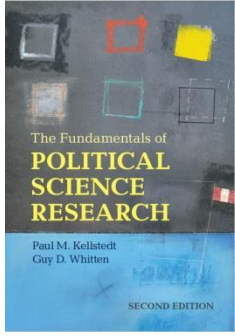


2. Outliers and Influential Cases in OLS

- Having outliers can draw suspicions in our findings



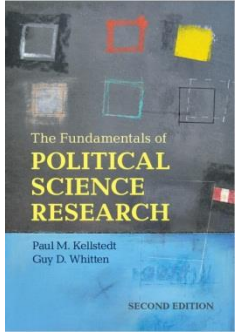
Outliers and Influential Cases in OLS



- In the regression setting, individual cases can be outliers in several different ways:
  - Leverage: They can have unusual independent variable values. This is known as a case having large "leverage."
  - Residual: They can have large residual values (usually we look at squared residuals to identify outliers of this variety).
  - They can have both large leverage and large residual values.
- The relationship among these different concepts of outliers for a single case in OLS is often summarized as separate contributions to "influence" in the following formula:

$$\text{influence}_i = \text{leverage}_i \times \text{residual}_i.$$

Identifying Influential Cases

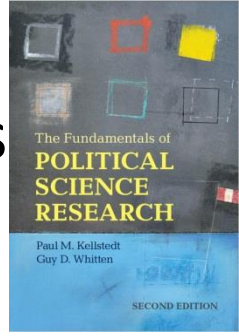


- One of the most famous cases of outliers/influential cases in political data comes from the 2000 U.S. presidential election in Florida.
- In an attempt to measure the extent to which ballot irregularities may have influenced election results, a variety of models were estimated in which the raw vote numbers for candidates across different counties were the dependent variables of interest.
- As an example of such a model, we will work with the following:

$$\text{Buchanan}_i = \alpha + \beta \text{Gore}_i + u_i.$$

- In this model the cases are individual counties in Florida,
  - The dependent variable (Buchanan_i) is the number of votes in each Florida county for the independent candidate Patrick Buchanan
  - The independent variable is the number of votes in each Florida county for the Democratic Party's nominee Al Gore (Gore_i).

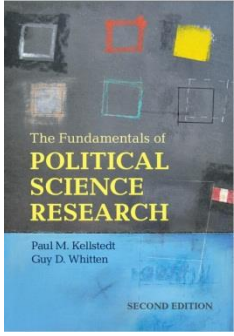
Votes for Gore and Buchanan in Florida counties in the 2000 U.S. presidential election

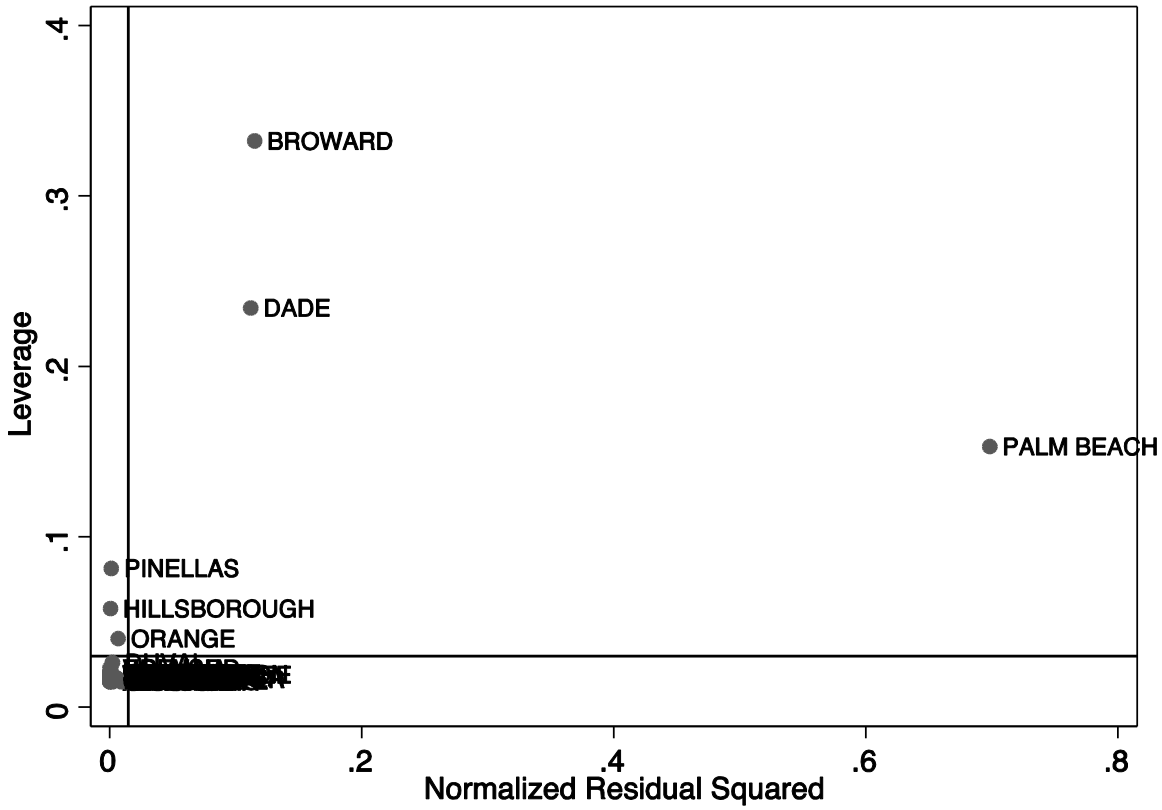


Independent variable	Parameter estimate
Votes for Gore	0.004***
	(0.0005)
Intercept	80.63*
	(46.4)
n	67
R^2	.48

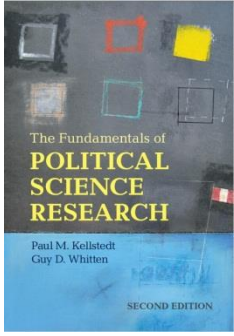
Notes: The dependent variable is the number of votes for Patrick Buchanan. Standard errors in parentheses.
Two-sided t-tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

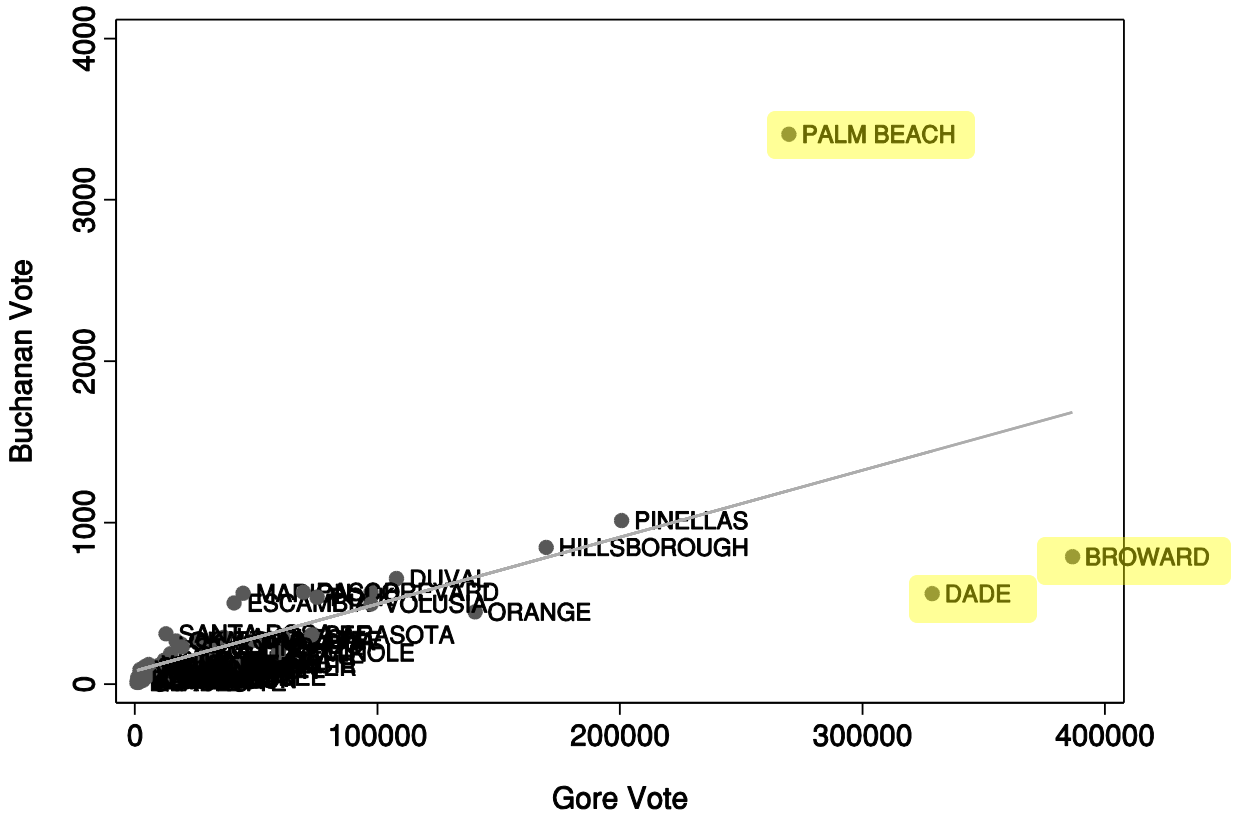
Leverage-versus-residual plot

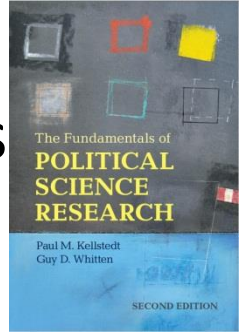




OLS line with scatter plot for Florida 2000







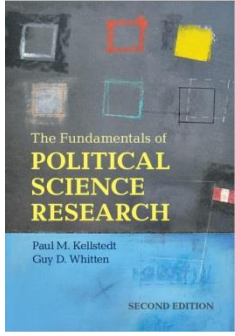
Votes for Gore and Buchanan in Florida counties in the 2000 U.S. presidential election

Independent variable	Model 1	Model 2	Model 3	Model 4	Model 5
Gore	0.004***	0.003***	0.003***	0.005***	0.005***
	(0.0005)	(0.0002)	(0.0002)	(0.0003)	(0.0003)
Palm Beach dummy		2606.3***		2095.5***	
		(150.4)		(110.6)	
Broward dummy				−1066.0***	
				(131.5)	
Dade dummy				−1025.6***	
				(120.6)	
Intercept	80.6*	110.8***	110.8***	59.0***	59.0***
	(46.4)	(19.7)	(19.7)	(13.8)	(13.8)
n	67	67	66	67	64
R^2	.48	.91	.63	.96	.82

Notes: The dependent variable is the number of votes for Patrick Buchanan.
Standard errors in parentheses.
Two-sided t-tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

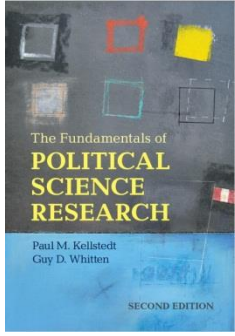
Report + address the outliers

How to deal with outliers



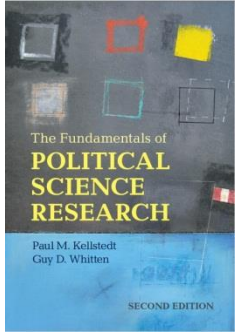
- 1. measurement errors
  - Need to remove them
- 2. Unusual in nature
  - Need theory explanations to keep it.
  - E.g. Military expenditures and the US

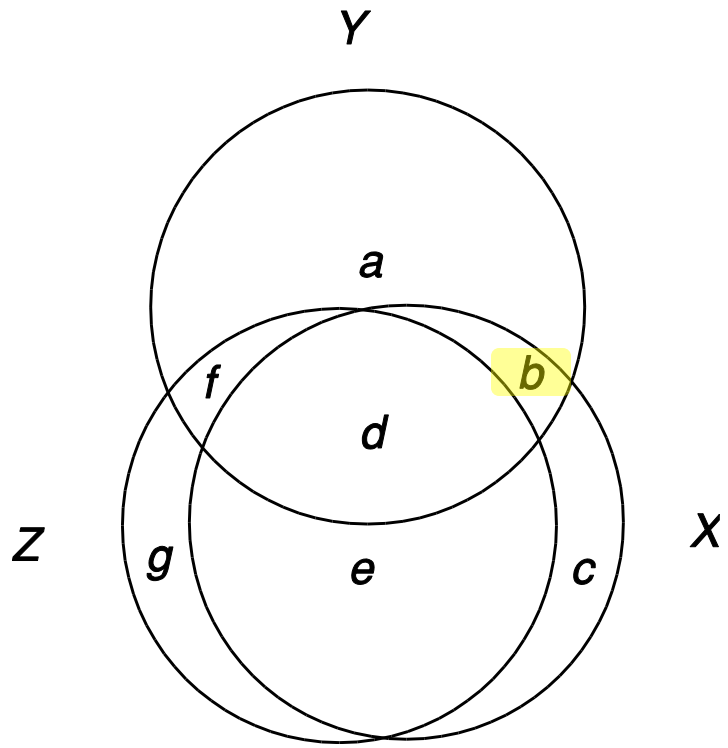
Multicollinearity



- Perfect multicollinearity: it occurs when one independent variable is an "exact linear function" of one or more other independent variables in a model.
  - Female and male
  - In practice, perfect multicollinearity is usually the result of a small number of cases relative to the number of parameters we are estimating, limited independent variable values, or model misspecification.
- High multicollinearity: when people refer to multicollinearity, they almost always mean "high multicollinearity." From here on, when we refer to "multicollinearity," we will mean "high, but less-than-perfect, multicollinearity"
- Multicollinearity is induced by (1) a small number of degrees of freedom (2) and/or high correlation between independent variables.
  - Civil war onset ~ Country GDP growth + Income per capital

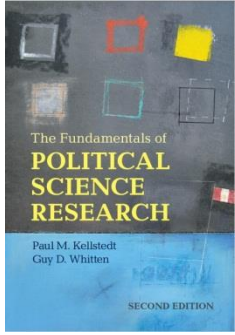
Venn diagram with multicollinearity





Result insignificant when having the multicollinearity issue

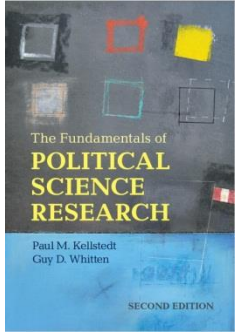
Detecting Multicollinearity



- It is important to know when you have multicollinearity.
  - If you have a high ${R^2}$ statistic, but none (or very few) of our parameter estimates is statistically significant, you should be suspicious of multicollinearity.
  - When you add and remove independent variables from our model, the parameter estimates for other independent variables (and especially their standard errors) change substantially.
- A more formal way to diagnose multicollinearity is to calculate the "variance inflation factor" (VIF) for each of our independent variables.
  - This calculation is based on an auxiliary regression model in which one independent variable, which we will call X_j, is the dependent variable and all of the other independent variables are independent variables.

$$\text{VIF}_j = \frac{1}{(1 - R_j^2)}.$$

Multicollinearity: A Real-World Example

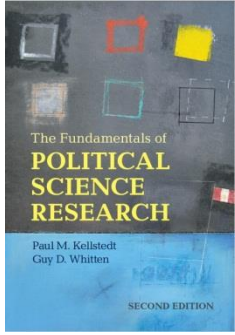


- We estimate a model of the thermometer scores for U.S. voters for George W. Bush in 2004. Our model specification is the following:

$$\begin{aligned}\text{Bush Thermometer}_i &= \alpha + \beta_1 \text{Income}_i + \beta_2 \text{Ideology}_i + \beta_3 \text{Education}_i \\ &\quad + \beta_4 \text{Party ID}_i + u_i.\end{aligned}$$

- Although we have distinct theories about the causal impact of each independent variable on peoples' feelings toward Bush, the table on the next slide indicates that some of these independent variables are substantially correlated with each other.

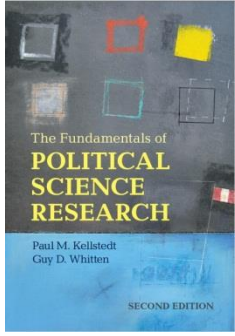
Pairwise correlations between independent variables



	Bush Therm.	Income	Ideology	Education	Party ID
Bush Therm.	1.00				
Income	0.09***	1.00			
Ideology	0.56***	0.13***	1.00		
Education	−0.07***	0.44***	−0.06*	1.00	
Party ID	0.69***	0.15***	0.60***	0.06*	1.00

Notes: Cell entries are correlation coefficients. Two-sided t-tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

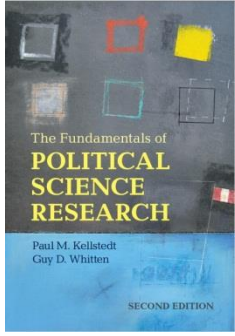
Model results from random draws of increasing size from the 2004 NES



Independent variable	Model 1	Model 2	Model 3
Income	0.77	0.72	0.11
	(0.90)	(0.51)	(0.15)
	{1.63}	{1.16}	{1.24}
Ideology	7.02	4.57*	4.26***
	(5.53)	(2.22)	(0.67)
	{3.50}	{1.78}	{1.58}
Education	−6.29	−2.50	−1.88***
	(3.32)	(1.83)	(0.55)
	{1.42}	{1.23}	{1.22}
Party ID	6.83	8.44***	10.00***
	(3.98)	(1.58)	(0.46)
	{3.05}	{1.70}	{1.56}
Intercept	21.92	12.03	13.73***
	(23.45)	(13.03)	(3.56)
n	20	74	821
R^2	.71	.56	.57

Notes: The dependent variable is the the respondent's thermometer score for George W. Bush. Standard errors in parentheses; VIF statistics in braces.
Two-sided t-tests: *** indicates $p < .01$; ** indicates $p < .05$; * indicates $p < .10$.

Multicollinearity: What Should I Do?



- The reason why multicollinearity is "vexing" is that there is no magical statistical cure for it.
- What is the best thing to do when you have multicollinearity?
  - Easy (in theory): *Collect more data*. But data are expensive to collect. If we had more data, we would use them and we wouldn't have hit this problem in the first place.
  - So, if you do not have an easy way increase your sample size, then multicollinearity ends up being something that you just have to live with.
- Robustness: It is important to know that you have multicollinearity and to present your multicollinearity by reporting the results of VIF statistics or what happens to your model when you add and drop the "guilty" variables.