Extending OLS: Fixed Effects and Controls

POSC 3410 – Quantitative Methods in Political Science

Steven V. Miller

Department of Political Science

Goal for Today

Add some wrinkles to the OLS regression framework.

Introduction

By this point, I think you could be doing your own research.

- You know what variables are.
- You know how to describe them.
- You know how to propose an explanation for variations in them.
- You know how to set up a research design to test an argument.
- You even know how to interpret a regression coefficient!

Limitations in Bivariate Regression

However, simple bivariate OLS is never enough.

- Variables of interest in political science are rarely interval.
- Bivariate regression does not control for confounders.

This lecture will deal with those topics accordingly.

Dummy Variables as Predictors

Dummy variables are everywhere in political science.

- They play an important role in "fixed effects" regression.
- Sometimes we're just interested in the effect of "one thing".

The South and Voter Turnout

Return to our education and turnout example: what if we're just interested in the effect of the South?

- We'll code states in the South as 1.
- When $x = 0$, we have the y-intercept.

The South and Voter Turnout

Table 8-3 Voter Turnout in Southern and Nonsouthern States

$\hat{y}$	=	$\hat{a}$	+	$\hat{b}$	(x)
Estimated turnout	=	Estimated turnout when South is 0	+	Mean change in turnout	(South)
Estimated turnout	=	46.34	+	–9.44	(South)
Standard error of $\hat{b}$				1.98	
t-statistic				–4.78	
P-value				.000	
Adjusted *R*-square = .31					

Sources: Turnout is the percentage of the voting eligible population in the 2006 elections, calculated by Michael McDonald, Department of Public and International Affairs, George Mason University, Fairfax, Va., and made available through his Web site: http://elections.gmu.edu/voter_turnout.htm. The independent variable is based on the definition of census regions: www.census.gov/geo/www/us_regdiv.pdf.

The South and Voter Turnout

- The estimate turnout in non-southern states is 46.34%
- The estimated turnout in the South is 36.9%.
- The "South effect" is -9.44 (se: 1.98).
- t-statistic: $-9.44/1.98 = -4.78$

Fixed Effects and Voter Turnout

Obviously, this regression isn't that informative.

- It also problematically treats non-Southern states as homogenous.

We can specify other regions as "fixed effects".

- These treat predictors as a series of dummy variables for each value of x.
- One predictor (or group) is left out as "baseline category".
  - Otherwise, we'd have no y-intercept.

Fixed Effects and Voter Turnout

Table 8-4 Estimating Voter Turnout in Four Regions

Estimated turnout	=	$\hat{a}$	+	$\hat{b}_1$ (Northeast)	+	$\hat{b}_2$ (West)	+	$\hat{b}_3$ (South)
		48.73	+	–2.69	+	–4.36	+	–11.82
Standard error of $\hat{b}$				2.85		2.58		2.47
t-statistic				–.95		–1.69		–4.79
P-value				.35		.10		.00
Adjusted *R*-square = .32								

Sources: Turnout is the percentage of the voting eligible population in the 2006 elections, calculated by Michael McDonald, Department of Public and International Affairs, George Mason University, Fairfax, Va., and made available through his Web site: http://elections.gmu.edu/voter_turnout.htm. Dummy variables for region are based on the definition of census regions: www.census.gov/geo/www/us_regdiv.pdf.

Fixed Effects and Voter Turnout

How to interpret Table 8-4:

- All coefficients communicate the effect of that region versus the baseline category.
  - This is the Midwest in our example.
- Estimated turnout in the Midwest is 48.73%.
- Northeast turnout doesn't different much from the Midwest ($t =$ -.95).
- Turnout may be lower in the West than the Midwest ($t =$ -1.69).
- The effect of the South is to depress turnout when compared to Midwest states.

Multiple Regression

Your previous example is basically an applied **multiple regression**.

- However, it lacks control variables.

Multiple regression produces **partial regression coefficients**.

Multiple Regression

Let's return to our state voter turnout example. Let:

- x_1: % of citizens in state having at least a HS diploma.
- x_2: states in the South.

Multiple Regression

Table 8-5 Regression Estimates for Education Level and South, for Dependent Variable, Voter Turnout

Estimated turnout	=	$\hat{a}$	+	$\hat{b}_1$(Education)	+	$\hat{b}_2$(South)
		3.70	+	.74	+	–7.57
Standard error of $\hat{b}$				.23		1.90
t-statistic				3.24		–4.00
P-value				.00		.00
Adjusted R-square = .42						

Sources: Turnout is the percentage of the voting eligible population in the 2006 elections, calculated by Michael McDonald, Department of Public and International Affairs, George Mason University, Fairfax, Va., and made available through his Web site: http://elections.gmu.edu/voter_turnout.htm. Percentage high school or higher is calculated from U.S. Census Bureau data, www.census.gov/compendia/smadb/TableA-22 .pdf. The dummy variable for southern region is based on the definition of census regions: www.census.gov/geo/www/us_regdiv.pdf.

Multiple Regression

- Estimate turnout for state not in the South in which no one graduated from high school: 3.7%.
- The partial regression coefficient for education: .74 ($t = 3.24$).
- The partial regression coefficient for the South: -7.57 ($t =$ -4.00).

Interaction Effects

Multiple regression is linear and additive.

- However, some effects (say: x_1) may depend on the value of some other variable (say: x_2).

In regression, we call this an **interaction effect**.

A Real World Example

Consider this argument from Zaller (1992):

- Democrats are weakly more pro-choice than Republicans.
- However, the difference is very stark among the politically aware.

Let's first start by visualizing this.

Controlled Comparison Table

Table 8-6 Mean Pro-choice Scores by Partisanship, Controlling for Political Knowledge

	Political knowledge		
Partisanship	*Low*	*High*	Total
Democrat	4.3	5.7	4.8
	(184)	(111)	(295)
Independent	3.7	4.6	4.1
	(205)	(142)	(347)
Republican	2.9	2.8	2.8
	(144)	(156)	(300)
Total	3.7	4.2	3.9
	(533)	(409)	(942)

Source: 2004 American National Election Study.

Interaction Effects

Our regression formula would look like this:

$$\hat{y} = \hat{a} + \hat{b_1}(x_1) + \hat{b_2}(x_2) + \hat{b_3}(x_1 * x_2)$$

where:

- $\hat{y}$ = estimated pro-choice scale score.
- x_1 = partisanship (0 = Dems, 1 = Ind., 2 = GOP).
- x_2 = political knowledge (0 = low, 1 = high).
- $x_1 * x_2$ = product of the two variables.

Interaction Effects

Table 8-7 Modeling Interaction: Partisanship, Political Knowledge, and Abortion Opinions

Variable	Coefficient	Standard error	*t*-statistic	Significance
Constant	4.33			
Partisanship	–.70	.15	–4.67	.000
High political knowledge	1.50	.29	5.17	.000
Party identification * High political knowledge	–.76	.22	–3.45	.001
Adjusted R-square = .11				

Source: 2004 American National Election Study.

Interaction Effects

How to interpret Table 8-7:

- Our estimate of pro-choice scores is 4.33 for low-knowledge Democrats.
- $\hat{b_1}$, $\hat{b_2}$, and $\hat{b_3}$ are all statistically significant.
- When x_1 and $x_2 = 1$, subtract -.76 from $\hat{y}$.
- Political knowledge leads to higher pro-choice scores *among Democrats*.

What this does for Republicans is kind of interesting.

- $\hat{y}$ for low-knowledge Republicans: 2.93.
- $\hat{y}$ for high-knowledge Republicans: 2.91.

Conclusion

This chapter is the culmination of everything discussed previously.

- It's basically what quantitative political science is.

Regrettably, we can only use OLS for interval-level dependent variables.

- We rarely have that.
- Next, we'll discuss what to do with non-normal responses.

Table of Contents