

# Extending OLS: Fixed Effects, Controls, and Interactions

EH6105 – Quantitative Methods

---

Steven V. Miller

Department of Economic History and International Relations



Stockholm  
University

## Goal for Today

*Add some wrinkles to the OLS regression framework.*

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”.

## Swing States and Voter Turnout

Return to our education and turnout example: what if we're just interested in the effect of a state being a "swing state?"

- We'll follow 538's coding of "swing states": CO, FL, IA, MI, MN, NV, NH, NC, OH, PA, VA, and WI
- When  $x = 0$ , we have the  $y$ -intercept.

**Table 1:** The Effect of Being a Swing State on Voter Turnout, 2016

	Model 1
Swing State	7.371*** (1.747)
Intercept	59.087*** (0.847)
Num.Obs.	51
R2 Adj.	0.252
RMSE	5.19

+  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

## Swing States and Voter Turnout

- The estimated turnout in safe states is 59.09%
- The estimated turnout in swing states is 66.46%
- The “swing state” effect is an estimated 7.37% (s.e.: 1.75).
- $t$ -statistic:  $7.37/1.75 = 4.22$

We can rule out, with high confidence, an argument that being a “swing state” has no effect on voter turnout.

- Our findings suggest a precise positive effect.



## What About Regional Variation?

Southern states tend to have lower turnout, for any number of reasons.

- Most Southern states are safe states.
- Southern states also tend to have poorer citizens, which raise costs of voting.
- A few have larger minority populations and a gross past/recent history of votings rights restrictions.

Let's first unpack regional variation by looking at the effect of the South relative to non-Southern states on voter turnout.

**Table 2:** The Effect of Being a Southern State on Voter Turnout, 2016

	Model 1
State is in the South	-3.465+ (1.768)
Intercept	61.976*** (1.020)
Num.Obs.	51
R2 Adj.	0.054
RMSE	5.83
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

## Southern States and Voter Turnout

- The estimated turnout in non-Southern states is 61.98%
- The estimated turnout in Southern states is 58.51%
- The “South” effect is an estimated -3.46% (s.e.: 1.77).
- $t$ -statistic:  $-3.46/1.77 = -1.96$

We can rule out, with high confidence, an argument that being a Southern state has no effect on voter turnout.

- Our findings suggest a precise negative effect.
- However, don't confuse this for a large effect. The difference is an estimated 3%.
  - This amounts to about half a standard deviation change across  $y$ .

## Fixed Effects and Voter Turnout

Obviously, this last regression isn't that informative.

- It also problematically treats non-Southern states as homogenous.
- A meager  $R^2$  suggests that.

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.

**Table 3:** The Effect of State Regions on Voter Turnout, 2016

	Model 1
Northeast	6.099* (2.351)
Midwest	4.805* (2.151)
West	0.404 (2.102)
Intercept	58.512*** (1.383)
Num.Obs.	51
R2 Adj.	0.131
RMSE	5.48
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

## Region Fixed Effects and Voter Turnout

How to interpret this regression:

- All coefficients communicate the effect of that region versus the baseline category.
  - This is the South in our example.
- Estimated turnout in the South is 58.51%.
- Turnout in the Northeast is discernibly higher than the South ( $t = 2.59$ )
- Turnout in the Midwest is discernibly higher than the South ( $t = 2.23$ ).
- We cannot estimate a difference between the South and West ( $t = 0.19$ )

Notice the coefficient for the West is positive, but probability of observing it if there's no actual difference between South and West is 0.42. Kinda probable.

# 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 a college diploma.
- $x_2$ : states in the South.
- $x_3$ : state is a swing state.

Important: we do this to “control” for potential confounders.



# The Rationale

Assume you are proposing a novel argument that state-level education explains voter turnout. I might argue for omitted variable bias on these grounds:

- The “South” effect depresses state-level education and voter turnout.
- The “swing state” effect may explain state-level education (roll with it...) and increases voter turnout.

In other words, I contend your argument linking education ( $x$ ) to voter turnout ( $y$ ) is spurious to these other factors ( $z$ ).

- That’s why you “control.” Not to soak up variation but to test for effect of potential confounders.

**Table 4:** Simple Models of Voter Turnout, 2016

	Swing State Model	South Model	Full Model
Swing State	7.371*** (1.747)		7.008*** (1.546)
State is in the South		-3.465+ (1.768)	-1.940 (1.415)
% College Diploma			0.384** (0.111)
Intercept	59.087*** (0.847)	61.976*** (1.020)	48.479*** (3.468)
Num.Obs.	51	51	51
R2 Adj.	0.252	0.054	0.420
RMSE	5.19	5.83	4.47

+  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

## Multiple Regression

- Estimated turnout for 1) a state not in the South that's 2) not a swing state and in which 3) no one graduated from college: 48.48%
  - This seems reasonable, but recall the minimum on this variable is WV (19.2%).
  - This parameter is effectively useless.
- The partial regression coefficient for % college diploma: 0.38 ( $t = 3.47$ ).
- The partial regression coefficient for the South is insignificant.
  - Conceivable explanation: education levels have a more precise effect and muddy the estimated negative effect of the South.
- The estimated effect of being a “swing state” is to increase voter turnout by an estimated 7.01% ( $t = 4.53$ )

# Interactive 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 **interactive 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 use 2012 ANES data to evaluate whether there's something to this.

## Our Data

**IVs:** Party ID, political knowledge, interaction between both

- Party ID: (0 = Dem, 1 = Independent, 2 = GOP)
- Political knowledge: does respondent know who Speaker of the House is?



Only 40% of respondents ( $n= 5,914$ ) in the 2012 ANES data knew who the Speaker of the House was.

**DV:** latent pro-choice score via graded response model of favor/oppose/neither abortion if:

- non-fatal health risk to woman
- fatal health risk to woman
- woman pregnant via incest
- woman pregnant via rape
- birth defect cases
- financial hardship cases
- woman wants to select child gender
- it's woman's choice.

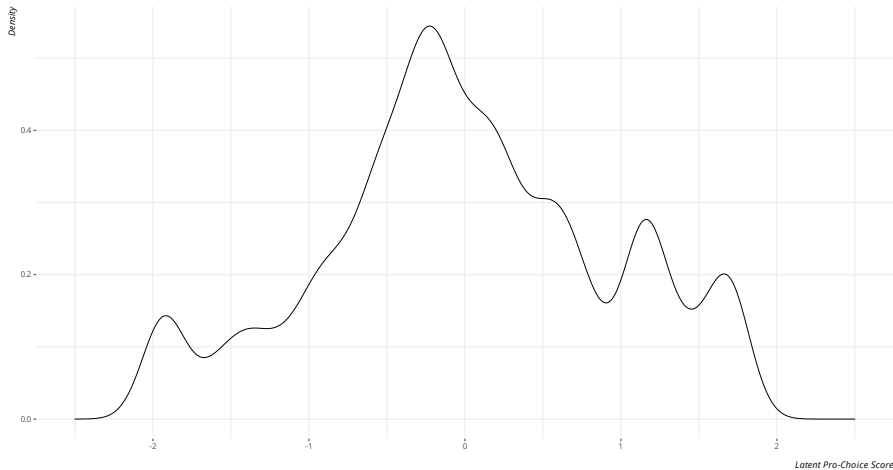
Emerging estimate has a mean of zero and standard deviation of one.

- Higher values = more “pro-choice.”



## Density Plot of Latent Pro-Choice Score (ANES, 2012)

The data were generated from a graded response model to have an approximate mean of 0 and standard deviation of 1.



Data: ANES (2012). Data available as `anes_prochoice` in `stevedata`. [Github: svmler/stevedata](#)

## Interactive 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 = doesn't know Speaker, 1 = knows Speaker).
- $x_1 * x_2$  = product of the two variables.

## A Caution About Constituent Terms

*Be careful with interpreting regression coefficients for constituent terms of an interaction.*

- The regression coefficient for party ID is effect of party ID when political knowledge = 0.
- The political knowledge coefficient is effect of knowledge when party ID variable = 0 (i.e. among Democrats).

**Table 5:** A Simple Interaction Between Partisanship and Political Knowledge on Pro-Choice Attitudes (ANES, 2012)

	<b>Model 1</b>
Partisanship (D to R)	-0.237*** (0.020)
Political Knowledge	0.414*** (0.036)
Partisanship*Political Knowledge	-0.184*** (0.031)
Intercept	0.099*** (0.022)
Num.Obs.	5196
R2 Adj.	0.091
RMSE	0.88
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

## Interactive Effects

How to interpret Table 5:

- Our estimate of pro-choice scores is 0.099 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 -0.184 from  $\hat{y}$ .
- Political knowledge leads to higher pro-choice scores *among Democrats*.

## Interactive Effects

Here's what this does for Democrats:

- $\hat{y}$  for low-knowledge Democrats: 0.099.
- $\hat{y}$  for high-knowledge Democrats: 0.513.

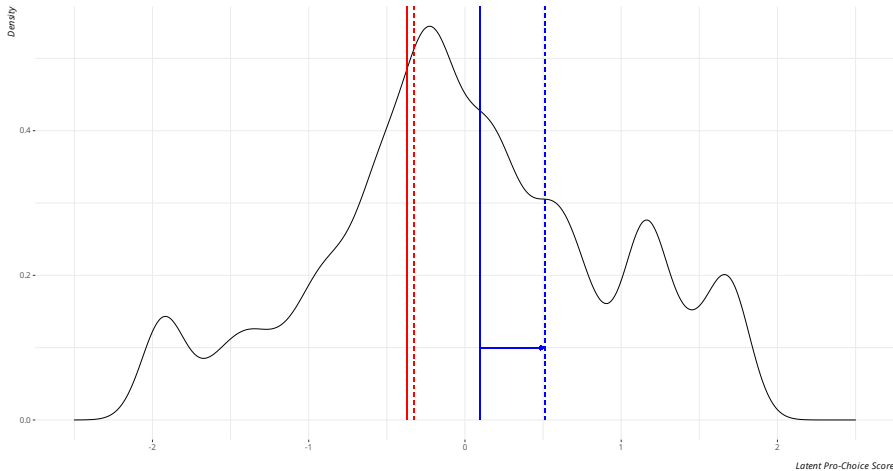
What this does for Republicans is arguably more interesting.

- $\hat{y}$  for low-knowledge Republicans: -0.374.
- $\hat{y}$  for high-knowledge Republicans: -0.328.

You see a huge effect of political knowledge on Democrats and, perhaps, no large (or even discernible) effect on Republicans.

## Density Plot of Latent Pro-Choice Score With Emphasized Interactive Effects

Notice the effect of political knowledge on pro-choice attitudes is larger among the Democrats than the Republicans.



Data: ANES (2012). Data available as `anes_prochoice` in `stevedata`. Github: [svmler/stevedata](https://github.com/svmler/stevedata). Party IDs are intuitively color-coded. Solid lines = low knowledge. Dashed lines = high knowledge.

# Conclusion

- Moving from bivariate OLS to multiple regression isn't really a big to-do.
  - It just means there are more parameters on the right-hand side of the equation.
  - What comes back are "partial" associations or regression coefficients.
  - This is where "ceteris paribus" language emerges.
- "Fixed effects" as you may encounter them = categorical dummy variables.
  - Something has to be a baseline, and that's what you're comparing against.
- Interactions = two (or more) things get multiplied together.
  - Constituent terms of  $x_1$  ( $x_2$ ): effect of  $x_1$  ( $x_2$ ) when  $x_2$  ( $x_1$ ) is 0.
  - Be mindful an "insignificant" interactive term may hide something.
  - Both things really have to have a 0 for the regression coefficients to communicate something.



# Table of Contents

Introduction

Extending OLS

- Dummy Variables as Predictors

- Fixed Effects in Regression

- Multiple Regression

- Interactive Effects

Conclusion