Extending the Linear Model: Fixed Effects, Controls, and Interactions

EH6127 – 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.

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 applied social science.

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

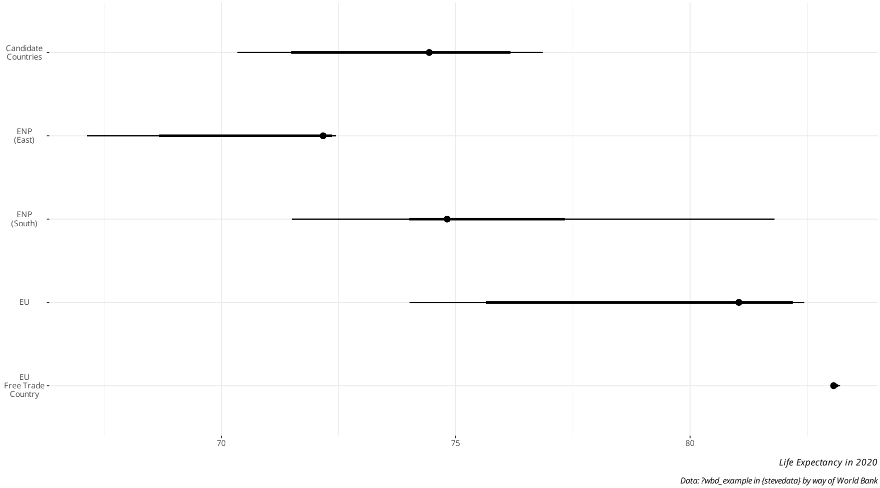
Eurostat Categories and Life Expectancy

Return to our life expectancy example: what if we're just interested in categorical difference, by Eurostat category?

- Categories: EU (e.g. Sweden), EFTA (e.g. Norway), UK (i.e. those guys), EUCC (e.g. BiH), PC (i.e. Kosovo, Georgia), ENP-E (i.e. AM, BY, AZ), ENP-S (e.g. Algeria), OEC (i.e. Russia)
- Let's make this somewhat honest and drop the UK and Russia and combine the PC and EUCC countries.

The Distribution of Life Expectancy in 2020, by Eurostat Category

The largest categorical differences seem to focus on the ENP-E countries as well as the free trade countries.



Eurostat Categories and Life Expectancy

Let's look at two things here:

1. A comparison of the ENP-E to the rest of the data.
2. A full comparison among categories.

Table 1: The Correlates of Life Expectancy for Eurostat Category States, 2020

	Model 1
ENP (East)	-7.597**
	(2.315)
Intercept	78.097***
	(0.585)
Num.Obs.	47
R2 Adj.	0.175

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

Life Expectancy and the ENP-East

- The estimated life expectancy in other Eurostat category states is 78.1
- The estimated life expectancy in ENP-East states is 70.5
- The "ENP-East effect" is an estimated -7.6 (s.e.: 2.31).
- t-statistic: -7.6/2.31 = -3.28

We can rule out, with high confidence, an argument that being an ENP-East state has no effect on life expectancy.

- Our findings suggest a precise negative effect.

What About Other Variation?

Obviously, this last regression isn't that informative.

- The baseline category is quite heterogeneous.
- It's impressive to pick up 17% of the variation with alone, though.

We can specify other categories 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 2: The Correlates of Life Expectancy for Eurostat Category States, 2020

	Model 1	**Model 2**
ENP (East)	-7.597**	-9.107***
	(2.315)	(1.793)
ENP (South)		-4.019**
		(1.192)
EU Free Trade		3.485+
		(1.793)
Candidate Country		-5.589***
		(1.192)
Intercept	78.097***	79.606***
	(0.585)	(0.587)
Num.Obs.	47	47
R2 Adj.	0.175	0.528

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

Categorical Fixed Effects and Life Expectancy

How to interpret this regression:

- All coefficients communicate the effect of that category versus the baseline category.
    - I forced this to be the EU for ease of comparison, but default is whatever comes first.
    - Just be mindful: *everything* is benchmarked to the baseline.
- Estimated life expectancy in the EU is 79.61.
- Life expectancy in the candidate countries is discernibly lower than the EU (t = -4.69).
- Life expectancy in the FTA countries is discernibly higher than the EU (t = 1.94).
- Life expectancy in the ENP-East is discernibly lower than the EU (t = -5.08).
- Life expectancy in the ENP-South is discernibly lower than the EU (t = -5.08).

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 what we did last time with human capital, but do more. Let:

- x_1: human capital score [0:1]
- x_2: real GDP per capita (2015 USD)
- x_3: categorical fixed effects

Important: we do this to "control" for potential confounders.

The Rationale

Assume you are proposing a novel argument that human capital explains life expectancy. I might argue for omitted variable bias on these grounds:

- You've misspecified "capital"; it's more material than "human".
- You've missed that some regions "are just different".

In other words, I contend your argument linking human capital (x) to life expectancy (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 3: The Correlates of Life Expectancy for Eurostat Category States, 2020

	Model 1	Model 2	Model 3
Human Capital			26.617***
			(5.815)
Real GDP per Capita			0.000**
			(0.000)
ENP (East)	-7.597**	-9.107***	-4.369**
	(2.315)	(1.793)	(1.360)
ENP (South)		-4.019**	2.258+
		(1.192)	(1.236)
EU Free Trade		3.485+	0.453
		(1.793)	(1.359)
Candidate Country		-5.589***	-0.537
		(1.192)	(1.056)
Intercept	78.097***	79.606***	58.071***
	(0.585)	(0.587)	(4.068)
Num.Obs.	47	47	47
R2 Adj.	0.175	0.528	0.790

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

Multiple Regression

Estimated life expectancy for EU state with no human capital and no money: 58.07

- This parameter is effectively useless, given how you modeled the data.
- (It's not a problem, though there are advanced tools available to make use of this).

Other takeaways:

- Partial [min-max] effect of human capital: 26.62
- Partial effect of GDP per capita is positive and significant.
  - *Don't* read much into coefficient size, only direction and significance.
  - e.g. a 45,000 USD increase in GDP per capita increases life expectancy by an estimated 2.8 years.
- Partialing out human capital and real GDP per capita, only the ENP differences remain.

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 example: we want to measure political trust in Sweden by ideology.

- *But*, a la Converse (1954) and Zaller (1992), political opinions are filtered through the politically aware.
- i.e. there's no standalone effect of ideology independent from political engagement.

Let's use 2019/2020 SOM data to evaluate whether there's something to this.

Our Data

IVs: ideology, political interest

- Ideology: (0 = “clearly to the left”, 4 = “clearly to the right”)
- Political interest: (0 = “not at all interested” or “not particularly interested”, 1 = “very/rather interested”)

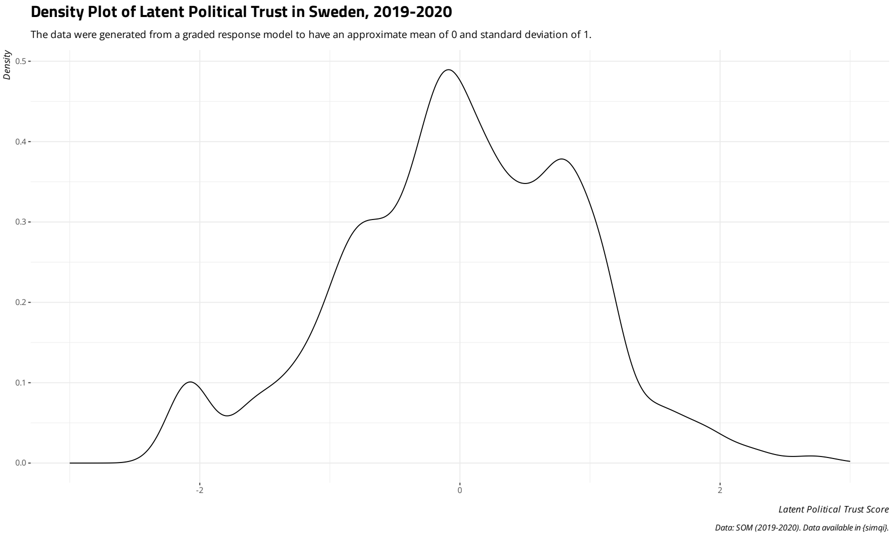
Our Data

DV: latent political trust based on various items. Including:

- government
- parliament
- political parties
- Swedish politicians

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

- Higher values = more political trust



Density Plot of Latent Political Trust in Sweden, 2019-2020
The data were generated from a graded response model to have an approximate mean of 0 and standard deviation of 1.
Density
0.5
0.4
0.3
0.2
0.1
0.0
-2
0
2
Latent Political Trust Score
Data: SOM (2019-2020). Data available in {simqi}.

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 political trust score.
- x_1 = ideology (0 = “clearly to the left”).
- x_2 = political interest (0 = “not at all/not particularly interested”).
- $x_1 * x_2$ = 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 ideology is effect of increasing ideology when the interest variable = 0 (i.e. low/no-interest).
- The political interest variable is effect of interest when ideology = 0 (i.e. among the furthest Left).

Table 4: A Simple Interaction Between Ideology and Political Interest on Political Trust (SOM, 2019-2020)

	Model 1
Political Interest	0.374***
	(0.075)
Ideology (L to R)	-0.172***
	(0.028)
Political Interest*Ideology	-0.071*
	(0.032)
Intercept	0.204**
	(0.066)
R2 Adj.	0.104
Num.Obs.	2841

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

Interactive Effects

How to interpret this table:

- Our estimate of political trust is 0.204 for the no-interest maximally Left
- $\hat{b_1}$, $\hat{b_2}$, and $\hat{b_3}$ are all statistically significant.
- When x_1 and $x_2 = 1$, subtract -0.071 from $\hat{y}$.

Interactive Effects

Here's what this does for the maximally Left:

- $\hat{y}$ for low/no-interest Left: 0.204.
- $\hat{y}$ for high-interest Left: 0.578.

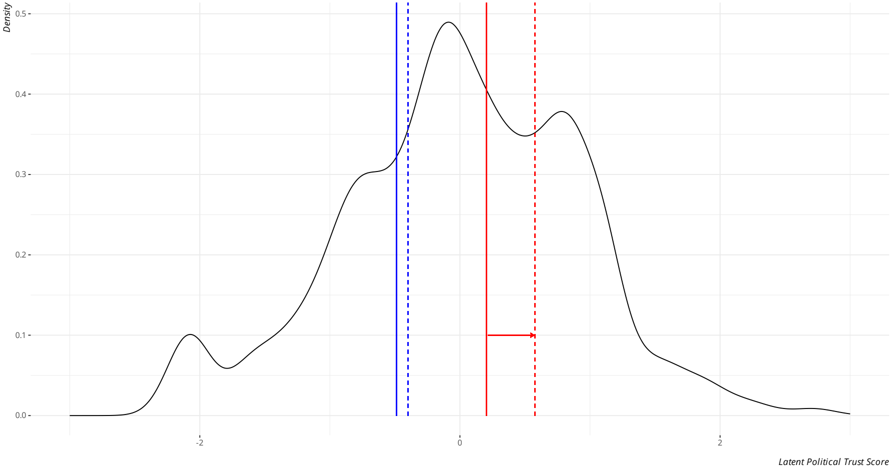
What this does for the maximally Right.

- $\hat{y}$ for low/no-interest Right: -0.486.
- $\hat{y}$ for high-interest Right: -0.398.

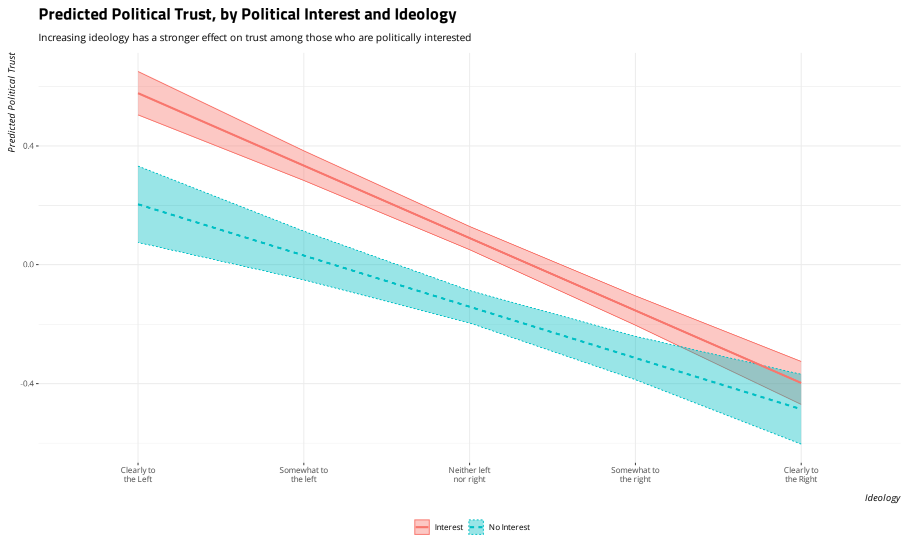
You see a huge effect of political interest on the Left, but a much smaller one on the right.

Density Plot of Latent Political Trust in Sweden, 2019-2020

Notice the effect of political interest is much stronger for the left than right.



Data: SOM (2019-2020). Data available in {simqi}.



Predicted Political Trust, by Political Interest and Ideology
Increasing ideology has a stronger effect on trust among those who are politically interested
Predicted Political Trust
0.4
0.0
-0.4
Clearly to the Left
Somewhat to the left
Neither left nor right
Somewhat to the right
Clearly to the Right
Ideology
Interest
No Interest

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 x1 (x2): effect of x1 (x2) when x2 (x1) 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