Welcome back!

Nameplates please. And technology encouraged today!

All TF materials are available at github.com/nolankav/api-202.

If you want to follow along, download the dataset here:

In R: `df <- read.csv("http://tinyurl.com/api-202-tf-3")`

In Excel: http://tinyurl.com/api-202-tf-4

Multiple regression and omitted variables

API 202: TF Session 2

EXCEL

Nolan M. Kavanagh
February 6, 2026


I'm not like a regular TF
I'm a cool TF Right, Regina?
you in
dinner
CAPSULE

Goals for today

1. **Review core concepts in bivariate analysis.**
2. **Consider an example of omitted variable bias.**
3. **Learn how to run multiple regressions.**
4. **Practice interpreting multiple regressions.**

We'll treat this session like a workshop with an interactive example.

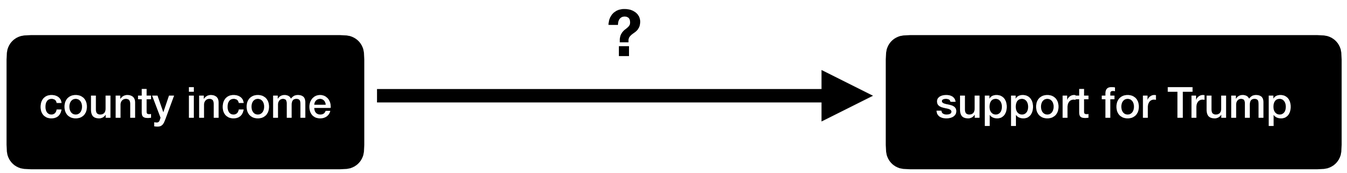
Overview of our sample data

Dataset of U.S. county-level characteristics in 2020

`state`	**State of county**	*Administrative*
`county_fips`	**County FIPS identifier**	*Administrative*
`pc_under_18`	**Percent of county under age 18**	*American Community Survey (2016–2020)*
`pc_over_65`	**Percent of county over age 65**	*American Community Survey (2016–2020)*
`pc_male`	**Percent of county that is male**	*American Community Survey (2016–2020)*
`pc_black`	**Percent of county that is Black**	*American Community Survey (2016–2020)*
`pc_latin`	**Percent of county that is Hispanic/Latino**	*American Community Survey (2016–2020)*
`pc_hs_grad`	**Percent of county that graduated high school**	*American Community Survey (2016–2020)*
`unemploy_rate`	**County unemployment rate (%)**	*American Community Survey (2016–2020)*
`med_income_000s`	**County median income ($1,000s)**	*American Community Survey (2016–2020)*
`pc_uninsured`	**Percent of county without health insurance**	*American Community Survey (2016–2020)*
`pc_trump`	**Percent of county votes for Trump in 2020**	*MIT Election Lab*

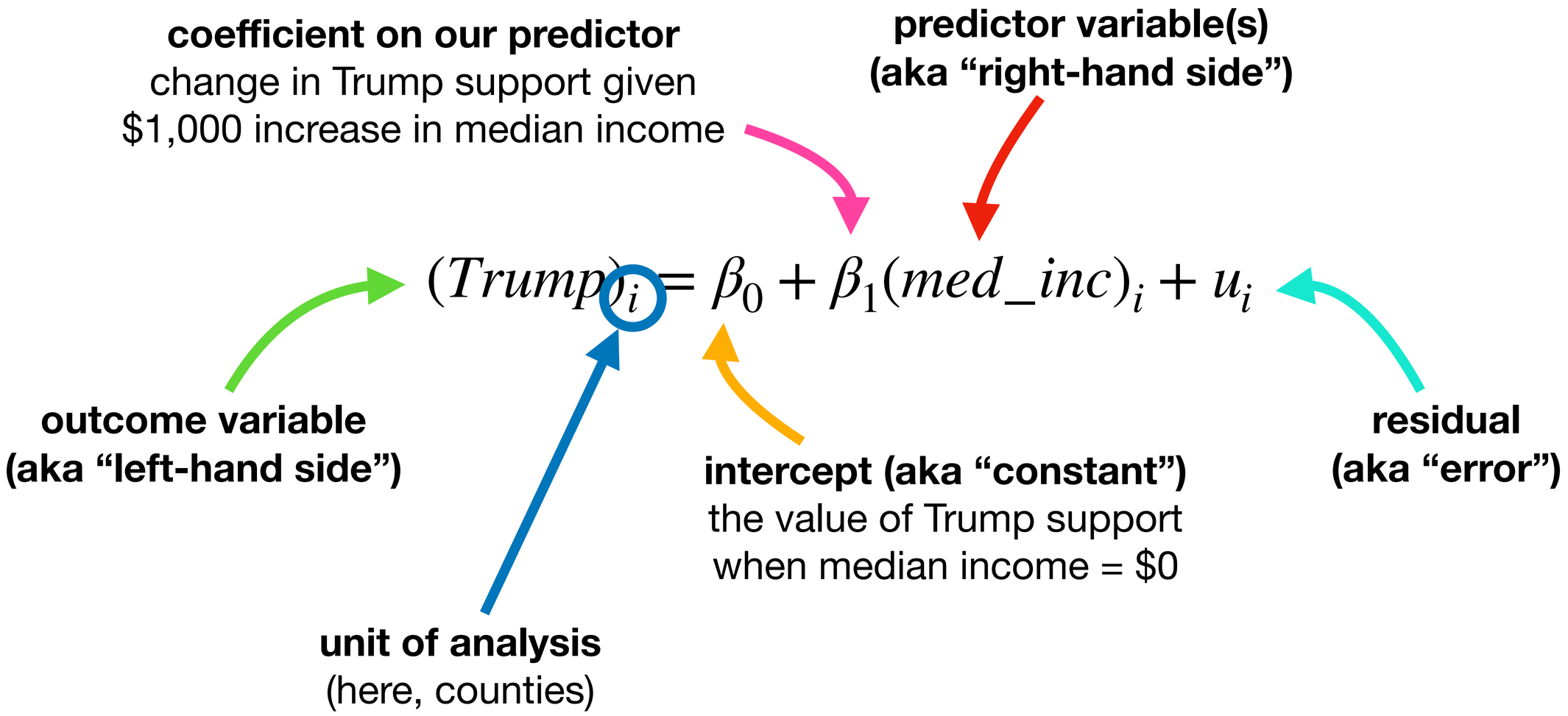
Hmm, I have an idea!

Was support for Trump about economic grievances?

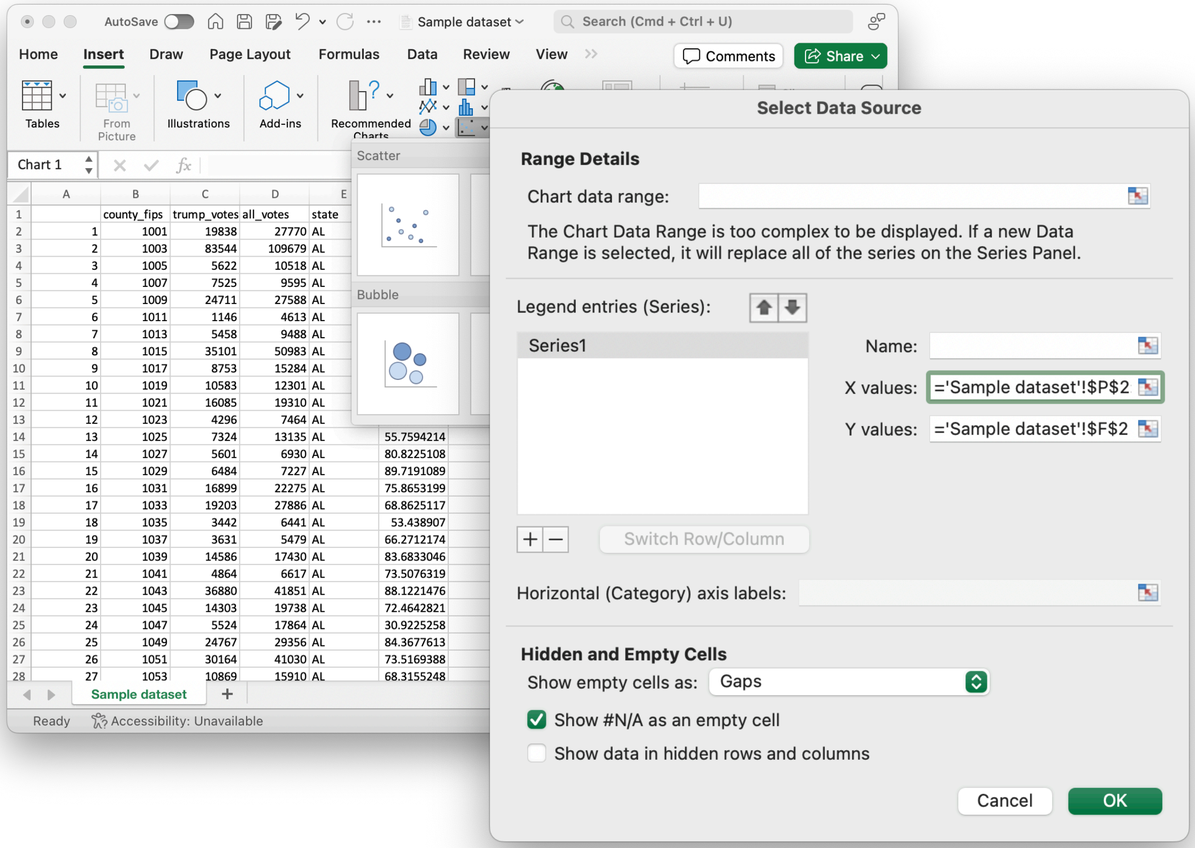


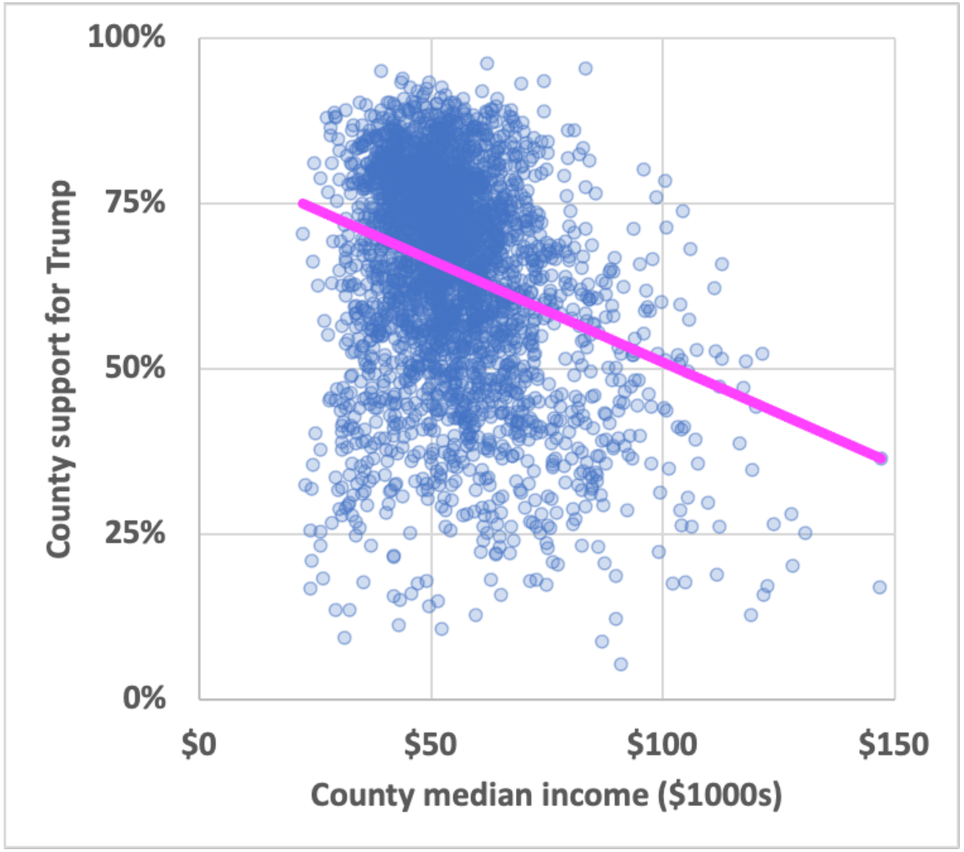
This idea is (was?) very hot in political science and among pundits on MSNBC and Fox News.

Population regression function



Does the graph check out?





X values: ='Sample dataset'!$P$2:$P$3115
Y values: ='Sample dataset'!$F$2:$F$3115

Does the regression check out?

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.280081455					
R Square	0.078445621					
Adjusted R Squa	0.078149492					
Standard Error	15.49839087					
Observations	3114					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	63629.79836	63629.79836	264.9032752	3.17032E-57	
Residual	3112	747502.7718	240.2001195			
Total	3113	811132.5702				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	81.9320651	1.079129799	75.92419851	0	79.81618663	84.04794358
med_inc_000s	-0.309050485	0.018988286	-16.27584945	3.17032E-57	-0.346281322	-0.271819648

Note: Trump support is measured from 0–100, so we don't have to multiply by 100 to interpret the coefficients.

Does the regression check out?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.280081455
R Square	0.078445621
Adjusted R Squa	0.078149492
Standard Error	15.49839087
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	63629.79836	63629.79836	264.9032752	3.17032E-57
Residual	3112	747502.7718	240.2001195		
Total	3113	811132.5702			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	81.9320651	1.079129799	75.92419851	0	79.81618663	84.04794358
med_inc_000s	-0.309050485	0.018988286	-16.27584945	3.17032E-57	-0.346281322	-0.271819648

Looks right to me!

Each $1,000 increase in county median income was associated with a statistically significant 0.31 percentage point (pp) decline in Trump support.

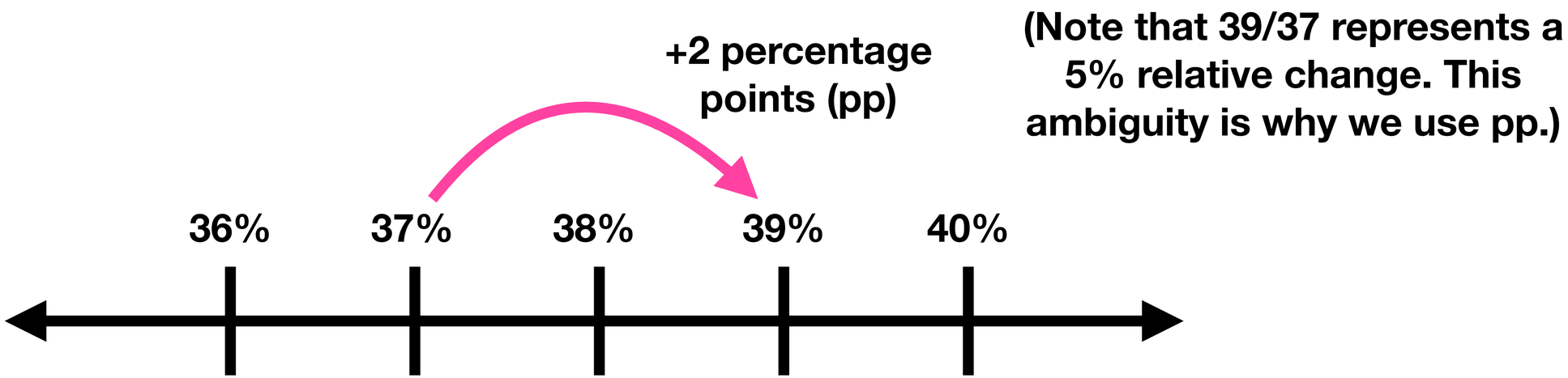
So Trump was all about economic grievances.

Case closed!

Note: Trump support is measured from 0–100, so we don't have to multiply by 100 to interpret the coefficients.

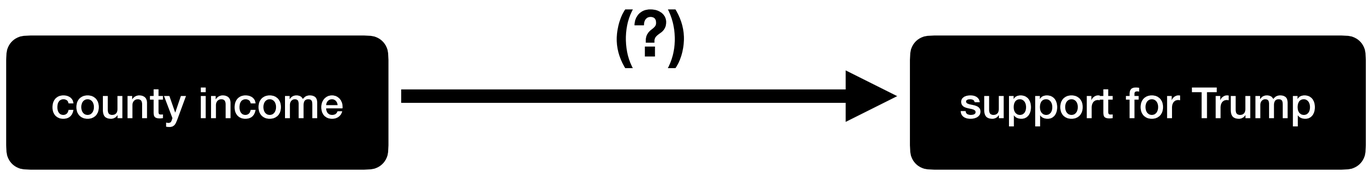
A quick detour on percentage points

- When our outcome is measured in percents (%), we describe any movement along the number line using percentage points.

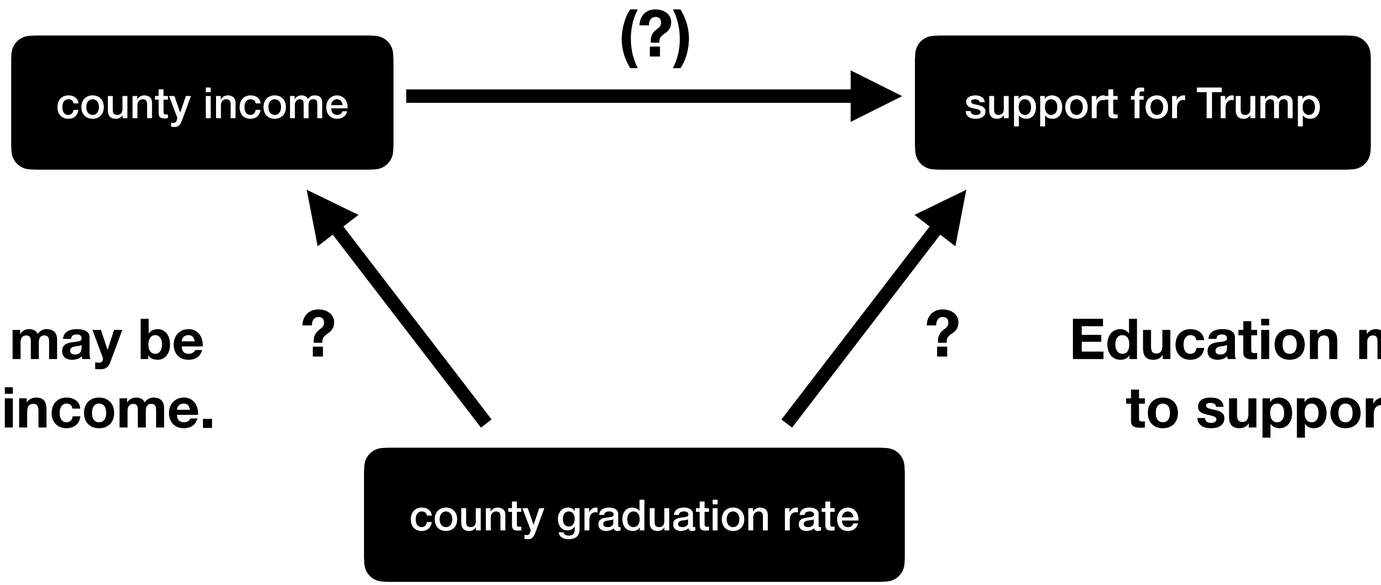


- If the outcome is measured from 0 to 100, you can interpret β_1 directly in pp. If measured 0 to 1, you must multiply by 100.

Or are we missing something?



Or are we missing something?



Graduation rates may be related to county income.

Education may be related to support for Trump.

The result? Bias in our regression.

Fine, let's add education to our analysis.

We use alpha vs. beta just to distinguish the different regressions.

Short regression $(Trump)_i = \hat{\alpha}_0 + \hat{\alpha}_1(med\_inc)_i + \hat{u}_i$

Long regression $(Trump)_i = \hat{\beta}_0 + \hat{\beta}_1(med\_inc)_i + \hat{\beta}_2(HS\_grad)_i + \hat{v}_i$

the omitted variable


You know
short and long regression functions?
Yeah, everyone in
can read
who has suffered API 202
short and long regression functions

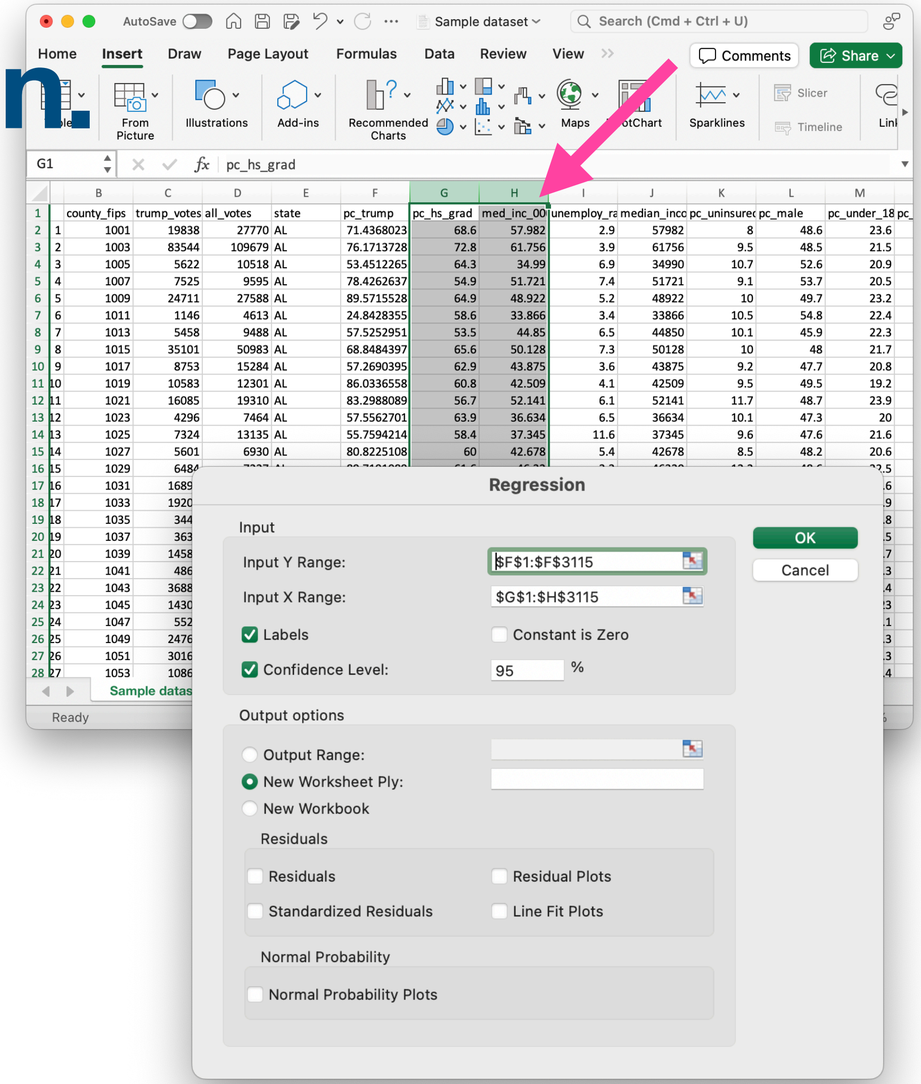
Let's run the long regression.

To include multiple predictors in our regression, we need their columns to be contiguous in Excel.

1. Move `med_inc_000s` near `pc_hs_grad`. To do so, "cut" the `med_inc_000s` column, right-click on the `pc_hs_grad` column, then "Insert Cut Cells."

2. Data > Data Analysis > Regression.

3. Input the Y range as usual. Specify the X range from the upper left-most cell to the lower right-most.



For X, note that we take cell 1 from our first column of interest and cell 3115 from our second column of interest.

Y values: =$F$1:$F$3115
X values: =$G$1:$H$3115

Let’s run the long regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.483734184
R Square	0.23399876
Adjusted R Squa	0.233506313
Standard Error	14.1322285
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	189804.0159	94902.00797	475.175565	8.2974E-181
Residual	3111	621328.5542	199.7198824		
Total	3113	811132.5702			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	134.4473979	2.309471038	58.21566744	0	129.9191561	138.9756397
pc_hs_grad	-1.026998542	0.040859699	-25.13475517	4.443E-127	-1.10711325	-0.946883833
med_inc_000s	-0.029663783	0.020575403	-1.441710882	0.149484604	-0.070006528	0.010678962

Note: Trump support is measured from 0–100, so the coefficients are already in percentage points. No multiplication by 100 required here.

Let’s run the long regression.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.483734184					
R Square	0.23399876					
Adjusted R Squa	0.233506313					
Standard Error	14.1322285					
Observations	3114					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	189804.0159	94902.00797	475.175565	8.2974E-181	
Residual	3111	621328.5542	199.7198824			
Total	3113	811132.5702				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	134.4473979	2.309471038	58.21566744	0	129.9191561	138.9756397
pc_hs_grad	-1.026998542	0.040859699	-25.13475517	4.443E-127	-1.10711325	-0.946883833
med_inc_000s	-0.029663783	0.020575403	-1.441710882	0.149484604	-0.070006528	0.010678962

Note: Trump support is measured from 0–100, so the coefficients are already in percentage points. No multiplication by 100 required here.

Let’s run the long regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.483734184
R Square	0.23399876
Adjusted R Squa	0.233506313
Standard Error	14.1322285
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	189804.0159	94902.00797	475.175565	8.2974E-181
Residual	3111	621328.5542	199.7198824		
Total	3113	811132.5702			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	134.4473979	2.309471038	58.21566744	0	129.9191561	138.9756397
pc_hs_grad	-1.026998542	0.040859699	-25.13475517	4.443E-127	-1.10711325	-0.946883833
med_inc_000s	-0.029663783	0.020575403	-1.441710882	0.149484604	-0.070006528	0.010678962

Well. **.**

Controlling for high school graduation rates, each $1,000 increase in county median income was associated with a 0.03 pp decline in Trump support.

And it’s not statistically significant.

Note: Trump support is measured from 0–100, so the coefficients are already in percentage points. No multiplication by 100 required here.


She doesn't even
explain our outcome
after controls

Let’s run the long regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.483734184
R Square	0.23399876
Adjusted R Squa	0.233506313
Standard Error	14.1322285
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	189804.0159	94902.00797	475.175565	8.2974E-181
Residual	3111	621328.5542	199.7198824		
Total	3113	811132.5702			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	134.4473979	2.309471038	58.21566744	0	129.9191561	138.9756397
pc_hs_grad	-1.026998542	0.040859699	-25.13475517	4.443E-127	-1.10711325	-0.946883833
med_inc_000s	-0.029663783	0.020575403	-1.441710882	0.149484604	-0.070006528	0.010678962

Note: Trump support is measured from 0–100, so the coefficients are already in percentage points. No multiplication by 100 required here.

Let's run the long regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.483734184
R Square	0.23399876
Adjusted R Squa	0.233506313
Standard Error	14.1322285
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	189804.0159	94902.00797	475.175565	8.2974E-181
Residual	3111	621328.5542	199.7198824		
Total	3113	811132.5702			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	134.4473979	2.309471038	58.21566744	0	129.9191561	138.9756397
pc_hs_grad	-1.026998542	0.040859699	-25.13475517	4.443E-127	-1.10711325	-0.946883833
med_inc_000s	-0.029663783	0.020575403	-1.441710882	0.149484604	-0.070006528	0.010678962

Note: Trump support is measured from 0–100, so the coefficients are already in percentage points. No multiplication by 100 required here.

Meanwhile, each 1 pp increase in a county's high school graduation rate was associated with 1.0 pp less support for Trump, controlling for county median income.

This association is statistically significant at the 5% level.

Let’s run the long regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.483734184
R Square	0.23399876
Adjusted R Squa	0.233506313
Standard Error	14.1322285
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	189804.0159	94902.00797	475.175565	8.2974E-181
Residual	3111	621328.5542	199.7198824		
Total	3113	811132.5702			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	134.4473979	2.309471038	58.21566744	0	129.9191561	138.9756397
pc_hs_grad	-1.026998542	0.040859699	-25.13475517	4.443E-127	-1.10711325	-0.946883833
med_inc_000s	-0.029663783	0.020575403	-1.441710882	0.149484604	-0.070006528	0.010678962

Note: Trump support is measured from 0–100, so the coefficients are already in percentage points. No multiplication by 100 required here.

Let’s run the long regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.483734184
R Square	0.23399876
Adjusted R Squa	0.233506313
Standard Error	14.1322285
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	189804.0159	94902.00797	475.175565	8.2974E-181
Residual	3111	621328.5542	199.7198824		
Total	3113	811132.5702			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	134.4473979	2.309471038	58.21566744	0	129.9191561	138.9756397
pc_hs_grad	-1.026998542	0.040859699	-25.13475517	4.443E-127	-1.10711325	-0.946883833
med_inc_000s	-0.029663783	0.020575403	-1.441710882	0.149484604	-0.070006528	0.010678962

When county median income AND high school graduation rates are set to 0, the expected support for Trump is 134%.

(Obviously, this isn’t a meaningful value.)

Note: Trump support is measured from 0–100, so the coefficients are already in percentage points. No multiplication by 100 required here.

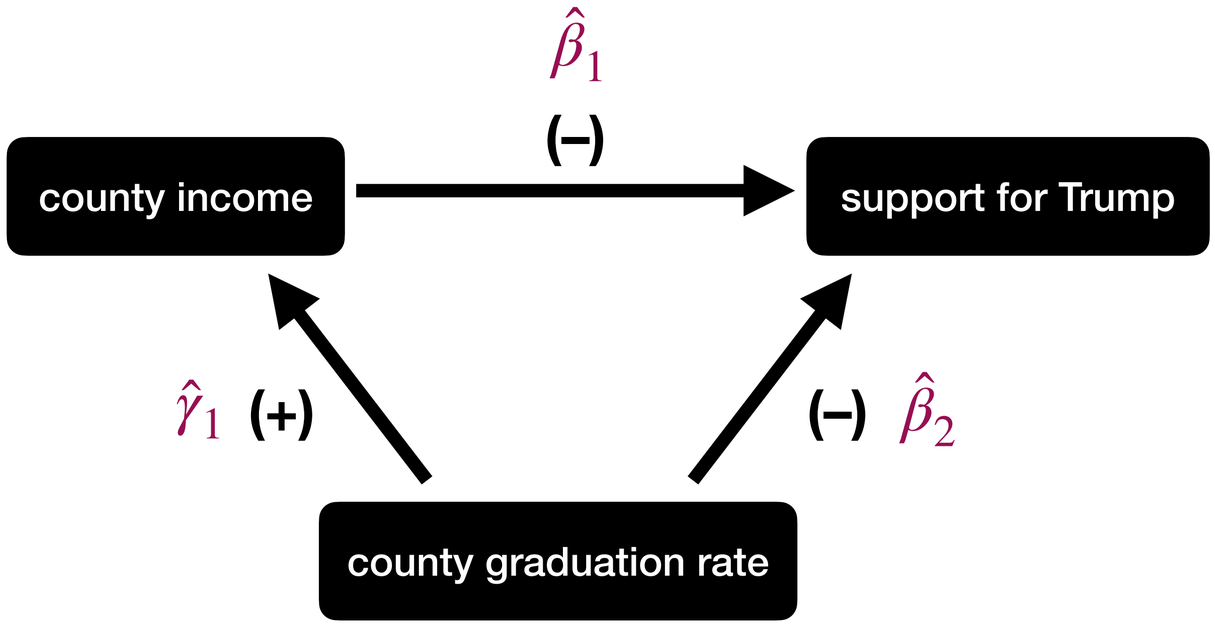
Womp.

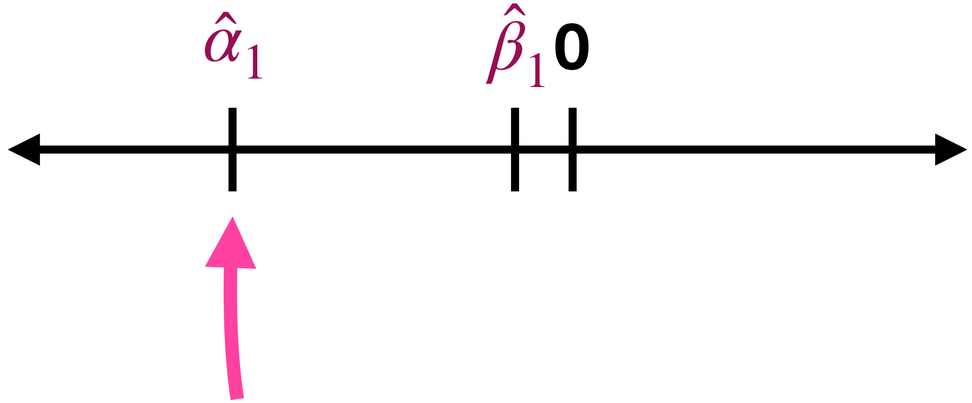
		Model 1	Model 2	
Intercept	$\hat{\alpha}_0$	**81.93** (-1.08) P<0.001	**134.45** (-2.31) P<0.001	$\hat{\beta}_0$
County median income ($1000s)	$\hat{\alpha}_1$	**-0.31** (0.02) P<0.001	**-0.03** (0.02) P=0.149	$\hat{\beta}_1$
County graduation rate			**-1.03** (0.04) P<0.001	$\hat{\beta}_2$
Num.Obs.		3114	3114	
R2		0.078	0.234	
R2 Adj.		0.078	0.234	

Short regression $(Trump)_i = \hat{\alpha}_0 + \hat{\alpha}_1(med\_inc)_i + \hat{u}_i$

Long regression $(Trump)_i = \hat{\beta}_0 + \hat{\beta}_1(med\_inc)_i + \hat{\beta}_2(HS\_grad)_i + \hat{v}_i$

Clearly, we were missing something.





Relative to the true β_1 (–), our estimate of α_1 was even more negative.

Bias formula $\alpha_1 - \beta_1 = \beta_2 * \gamma_1 = (-)(+) = (-)$


omitted variable
Nolan
'd be like, "Why didn't you
control for me
?"
And I'd be like, "Why are you so obsessed with me?"

Bias: sign or size?

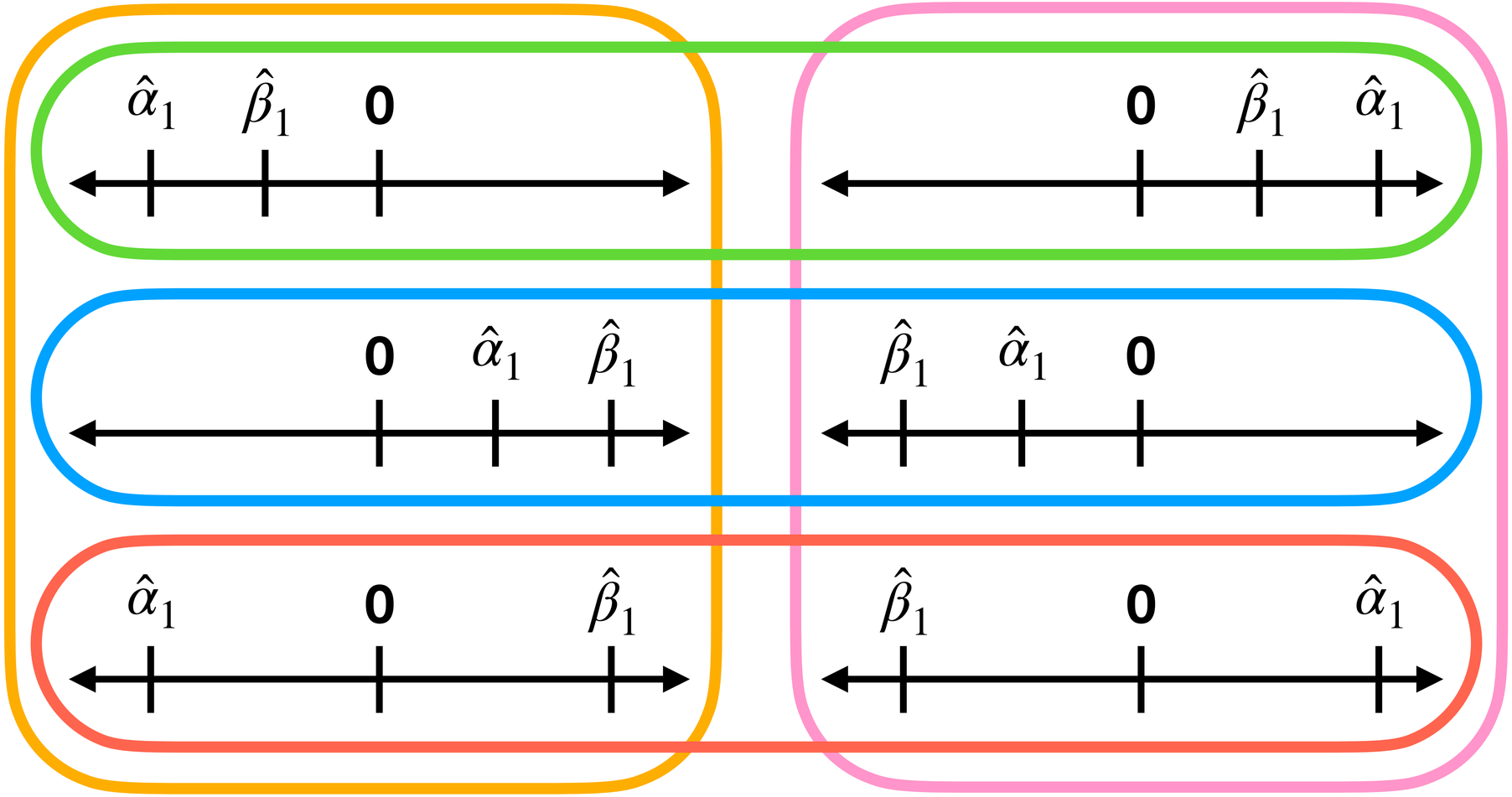
Overstatement

i.e. α_1 is farther from 0

Understatement

i.e. α_1 is closer to 0

Sign flip!



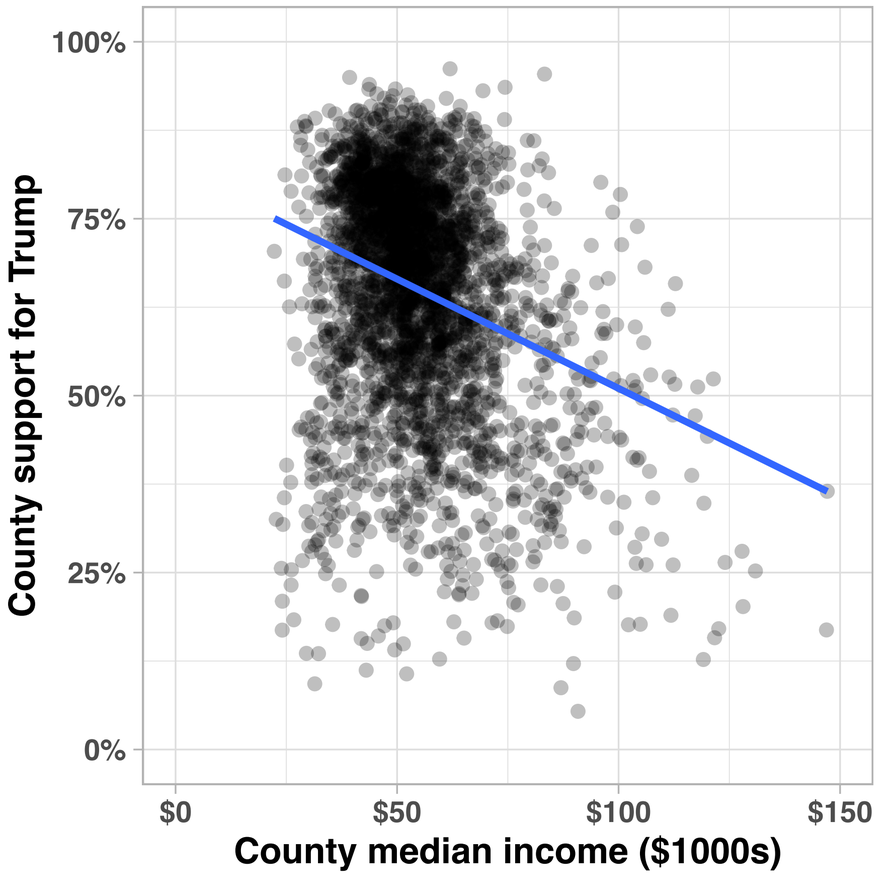
Negative bias (–)

i.e. α_1 is to the left of β_1

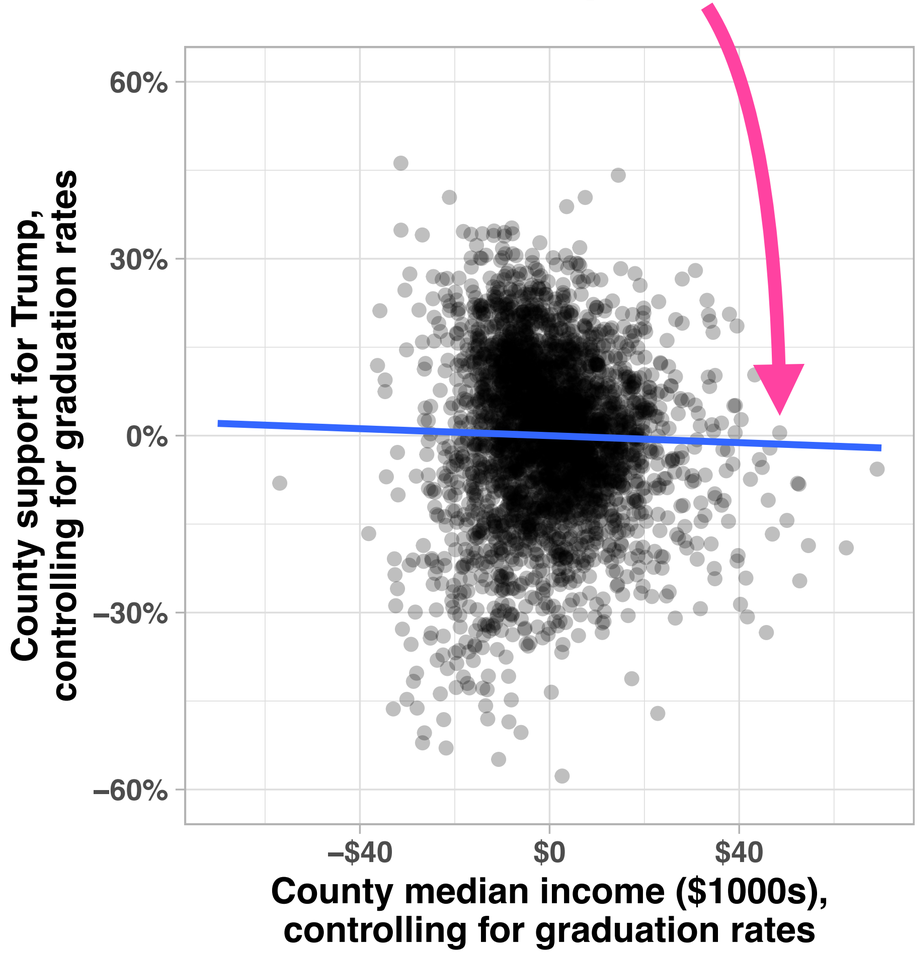
Positive bias (+)

i.e. α_1 is to the right of β_1

What happens to our graph when we control for education?



This is the same slope as Model 2.



P.S. The code to do this optional exercise is in the Github, but we won't be reviewing it in class.

OK, what did we learn?

Omitted variables can mess up our regressions.

Think carefully about what might be missing.

Is our new model causal? Or are we missing something *else*?

How many omitted variables can there be?