

# 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 2, 2024



# 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.

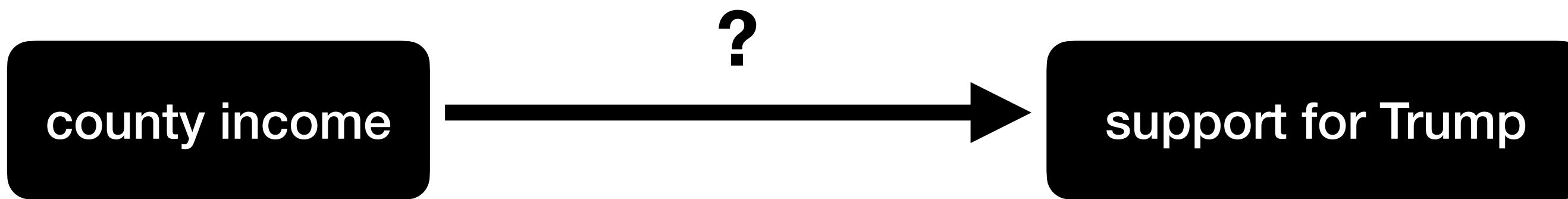
# Overview of our sample data

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

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

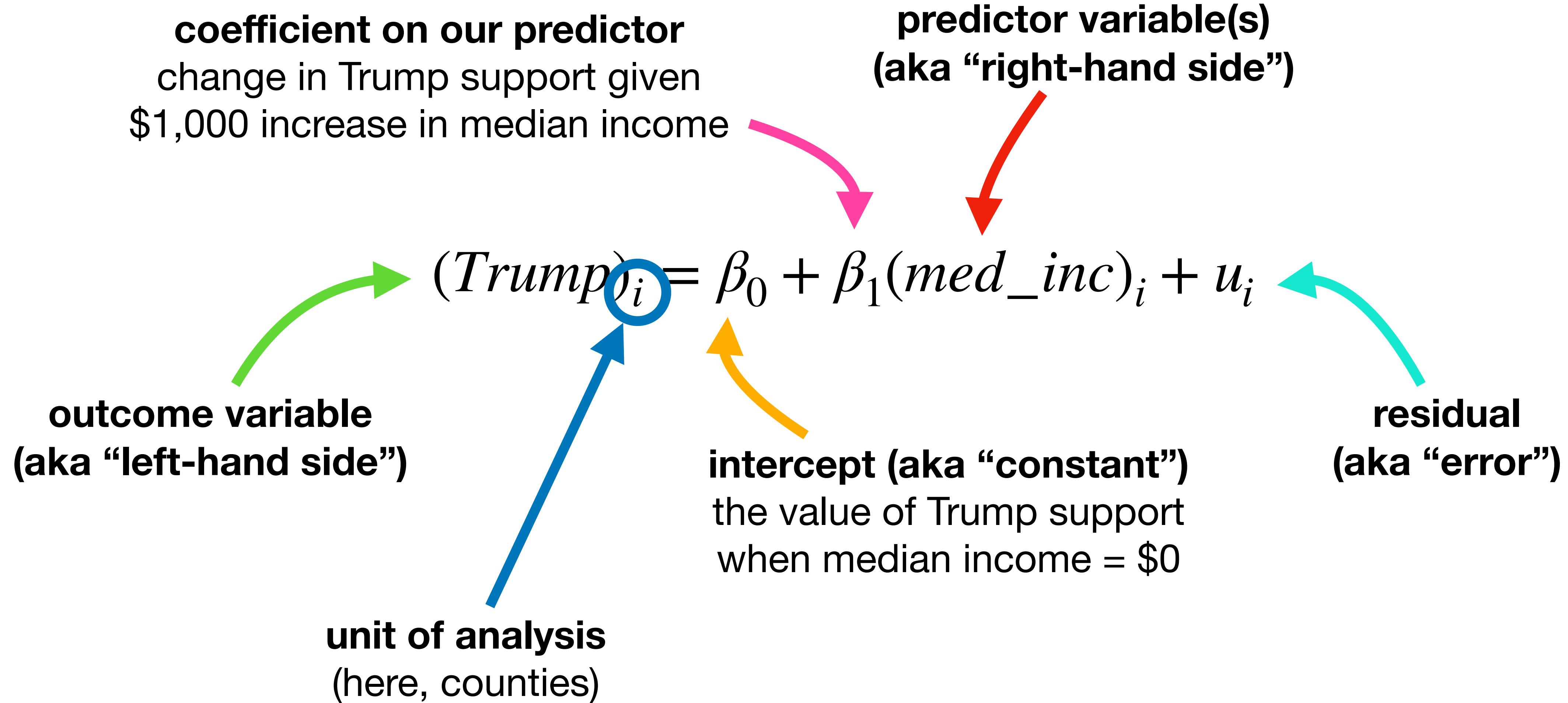
# Hmm, I have an idea!

**Was support for Trump about economic grievances?**



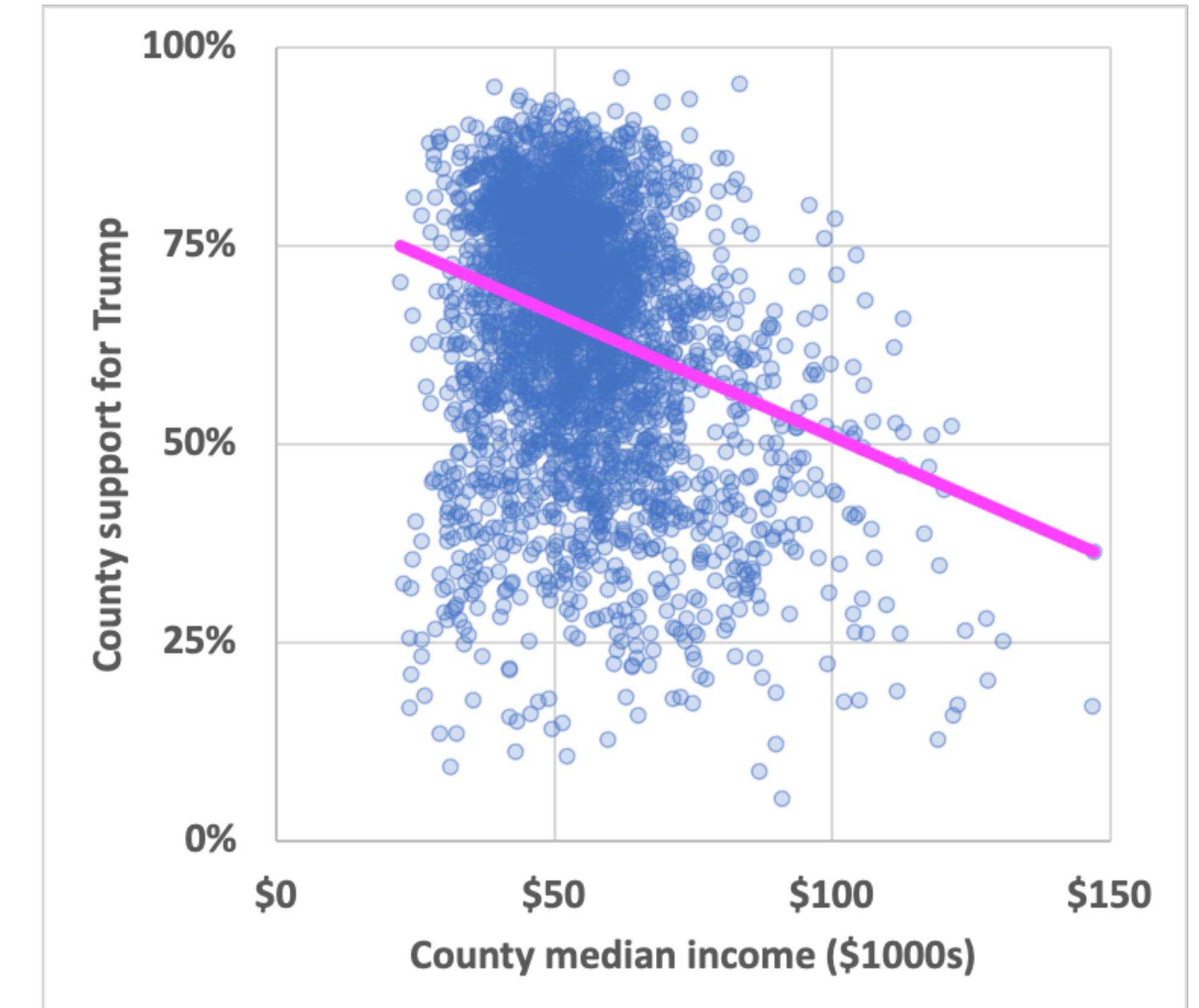
**This idea is (was?) very hot in political science  
and among pundits on MSNBC and Fox News.**

# Population regression function



# Does the graph check out?

The screenshot shows the Microsoft Excel interface with a sample dataset loaded. A scatter chart is selected, and the 'Insert' tab is active. A 'Select Data Source' dialog box is open, showing the 'Range Details' section. The X values are set to '=Sample dataset!\$P\$2:\$P\$3115' and the Y values are set to '=Sample dataset!\$F\$2:\$F\$3115'. The chart area displays a scatter plot of county median income versus county support for Trump.



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

# Does the regression check out?

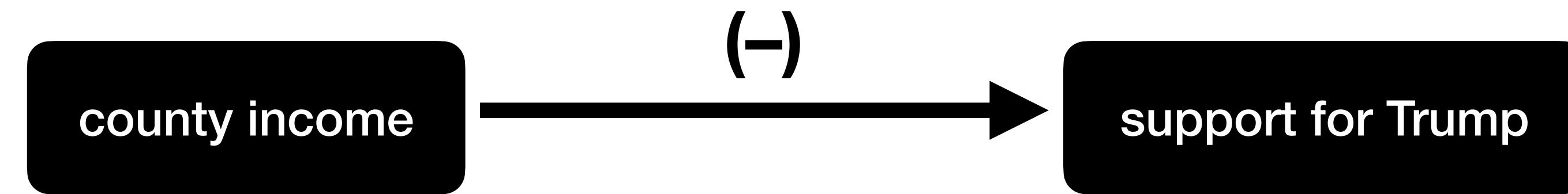
SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.280081455					
R Square	0.078445621					
Adjusted R Squa	0.078149492					
Standard Error	15.49839087					
Observations	3114					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	63629.79836	63629.79836	264.9032752	3.17032E-57	
Residual	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

Looks right to me!

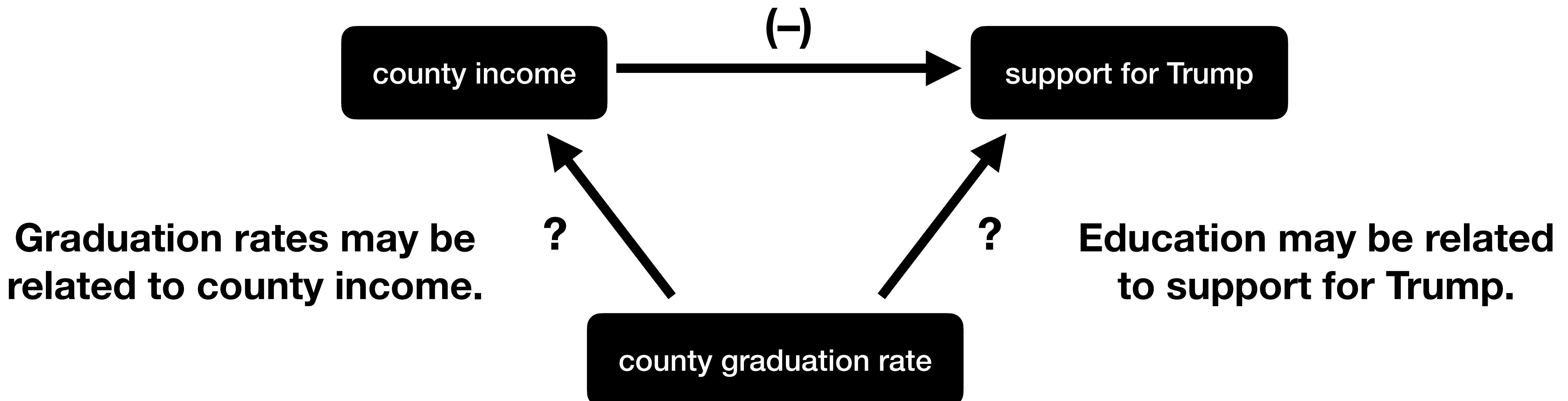
So Trump was all about economic grievances.

Case closed!

# Or are we missing something?



# Or are we missing something?



**The result? Bias in our regression.**

# Fine, let's add education to our analysis.

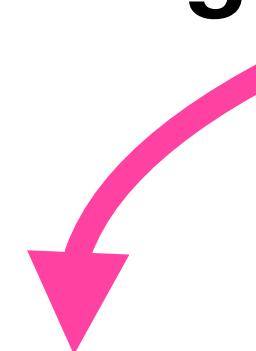
We use alpha vs. beta just to distinguish the different regressions.

**Short regression**

$$(Trump)_i = \alpha_0 + \alpha_1(med\_inc)_i + u_i$$

**Long regression**

$$(Trump)_i = \beta_0 + \beta_1(med\_inc)_i + \beta_2(HS\_grad)_i + v_i$$



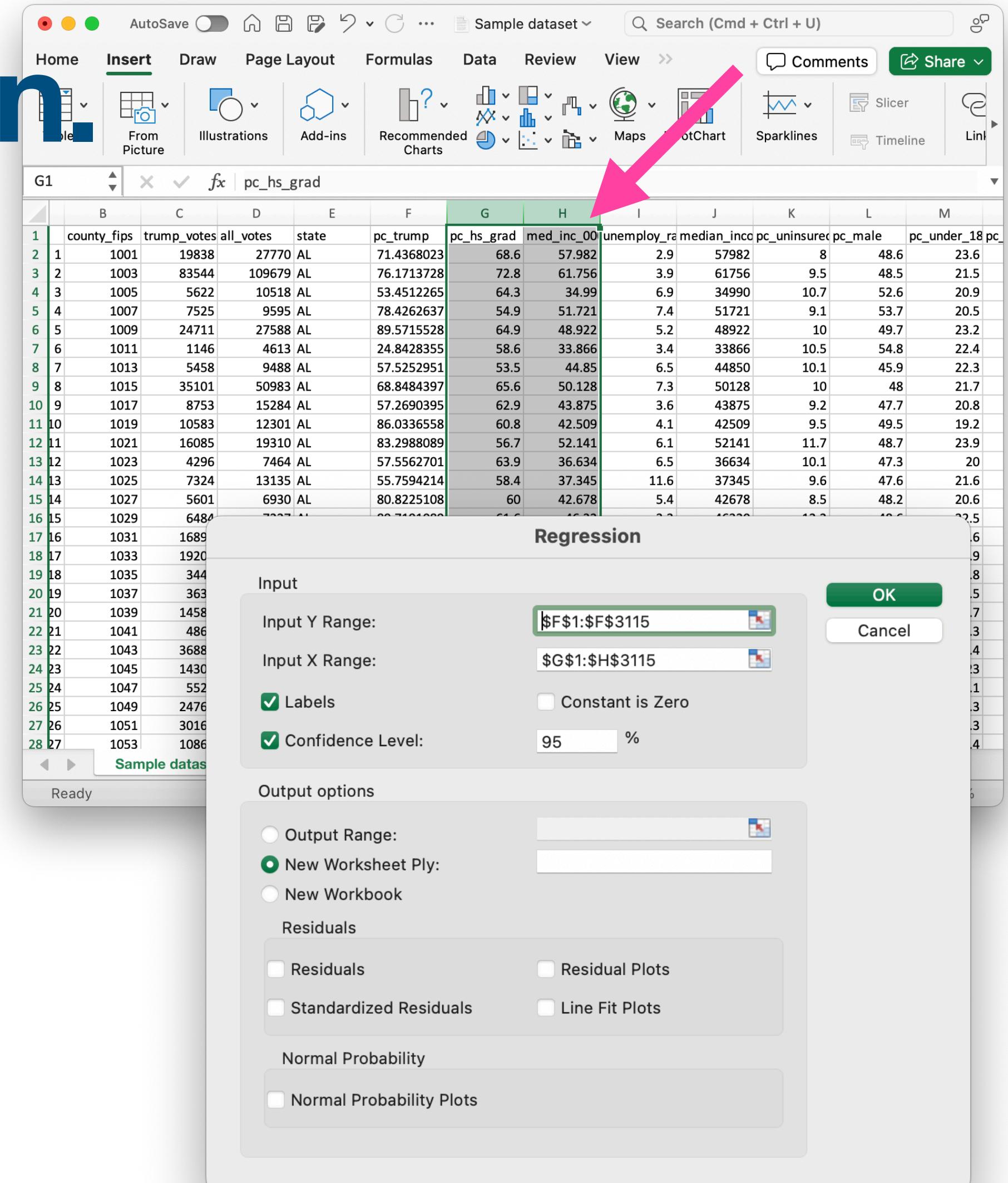
the omitted variable

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



X values: =`$F$1:$F$3115`  
Y values: =`$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

# 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

# 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 is associated with a 0.03 pp decline in Trump support.

And it's not statistically significant.



**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

# 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

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

# 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.)

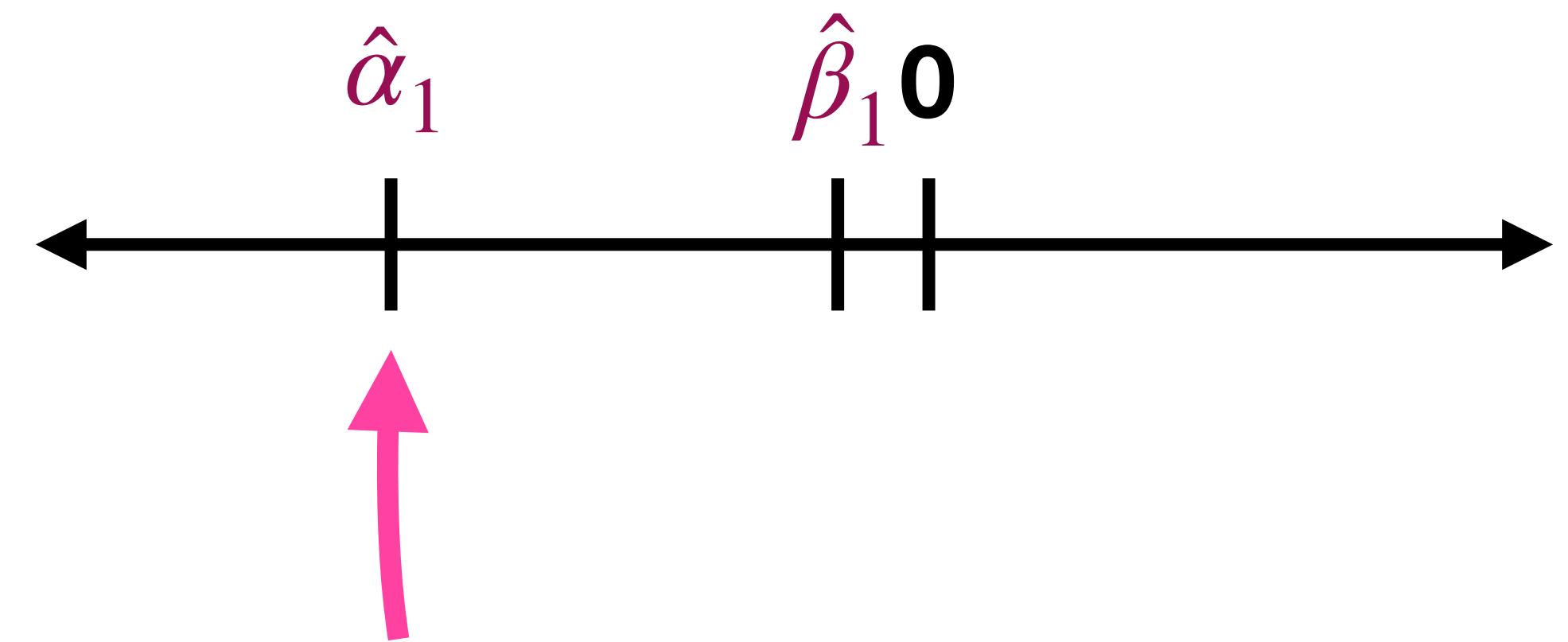
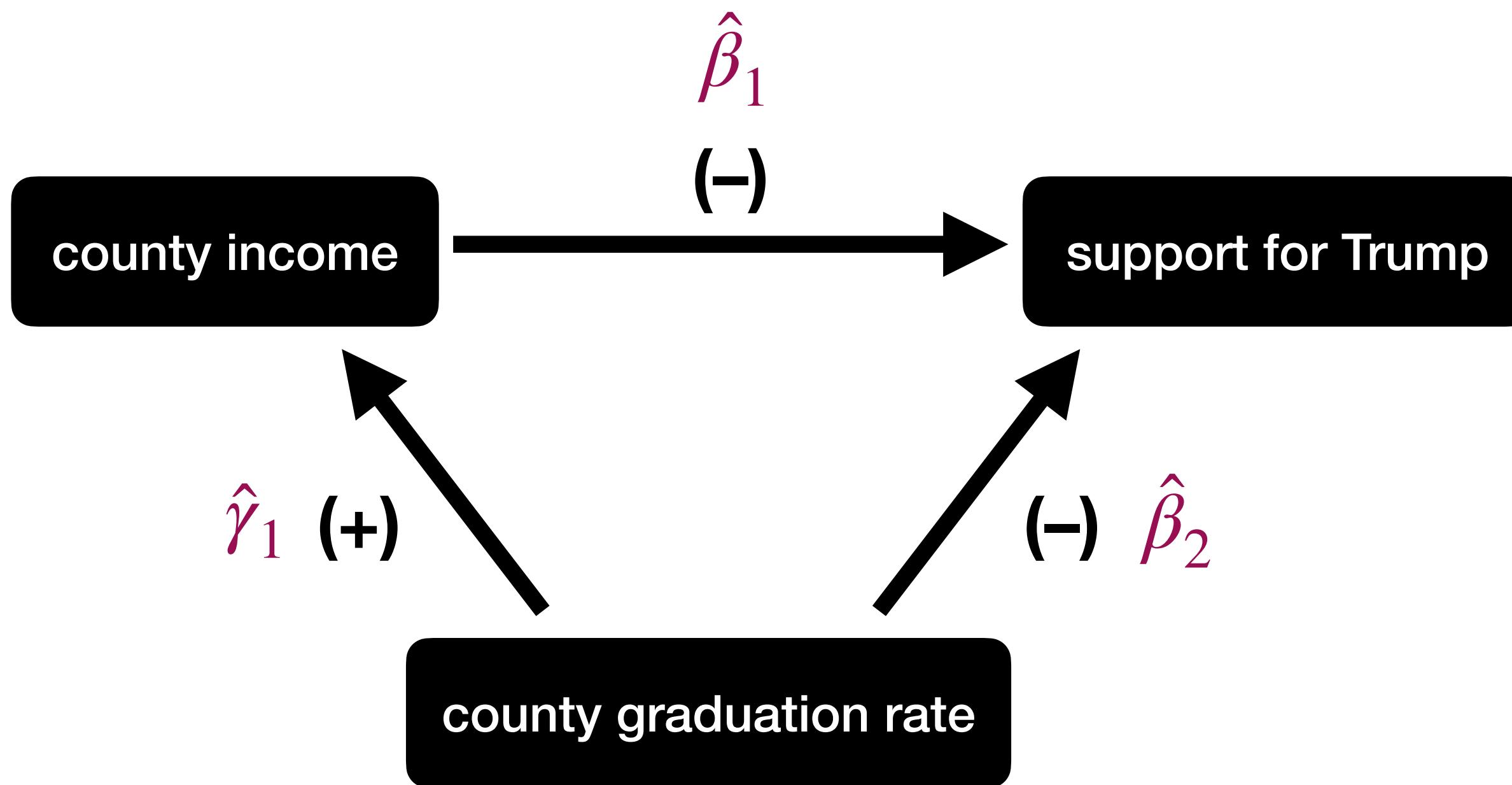
# Womp.

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

**Short regression**  $(Trump)_i = \alpha_0 + \alpha_1(med\_inc)_i + u_i$

**Long regression**  $(Trump)_i = \beta_0 + \beta_1(med\_inc)_i + \beta_2(HS\_grad)_i + v_i$

# Clearly, we were missing something.



Relative to the true  $\beta_1$  (-), our estimate of  $\alpha_1$  was even more negative.

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

# Bias: sign or size?

**Overstatement**

i.e.  $\hat{\alpha}_1$  is farther from 0

**Understatement**

i.e.  $\hat{\alpha}_1$  is closer to 0

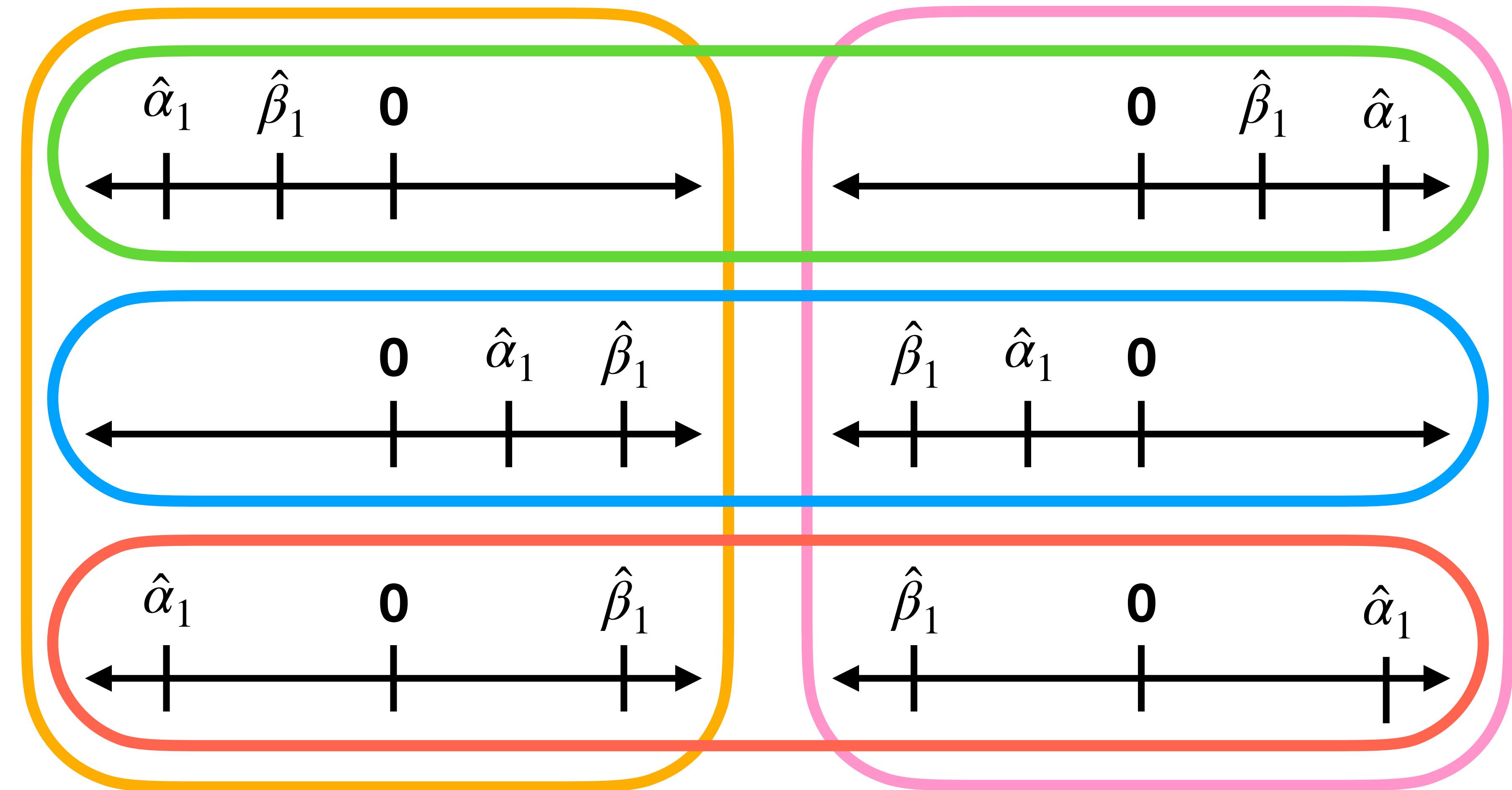
**Sign flip!**

**Negative bias (-)**

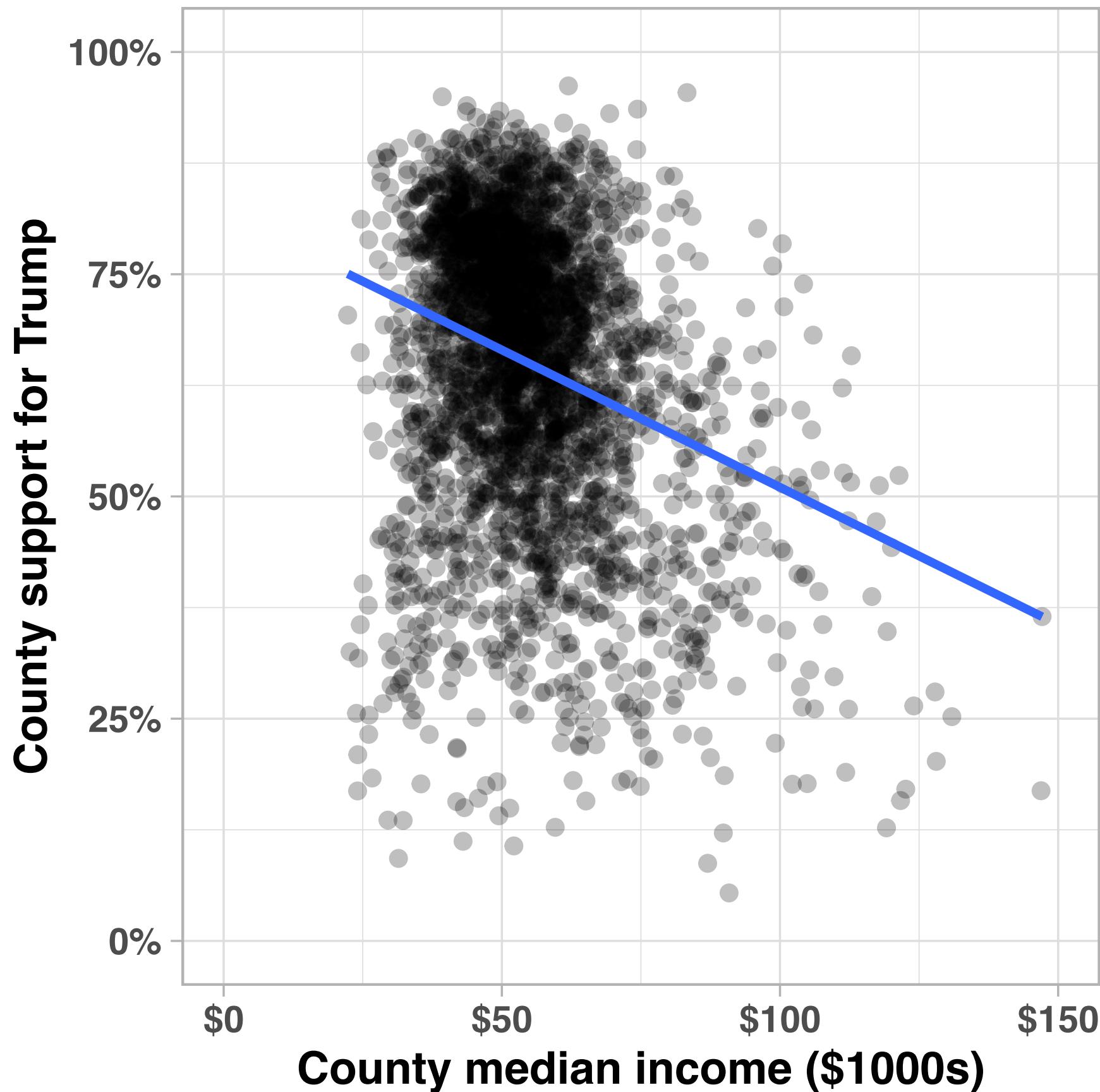
i.e.  $\hat{\alpha}_1$  is to the left of  $\beta_1$

**Positive bias (+)**

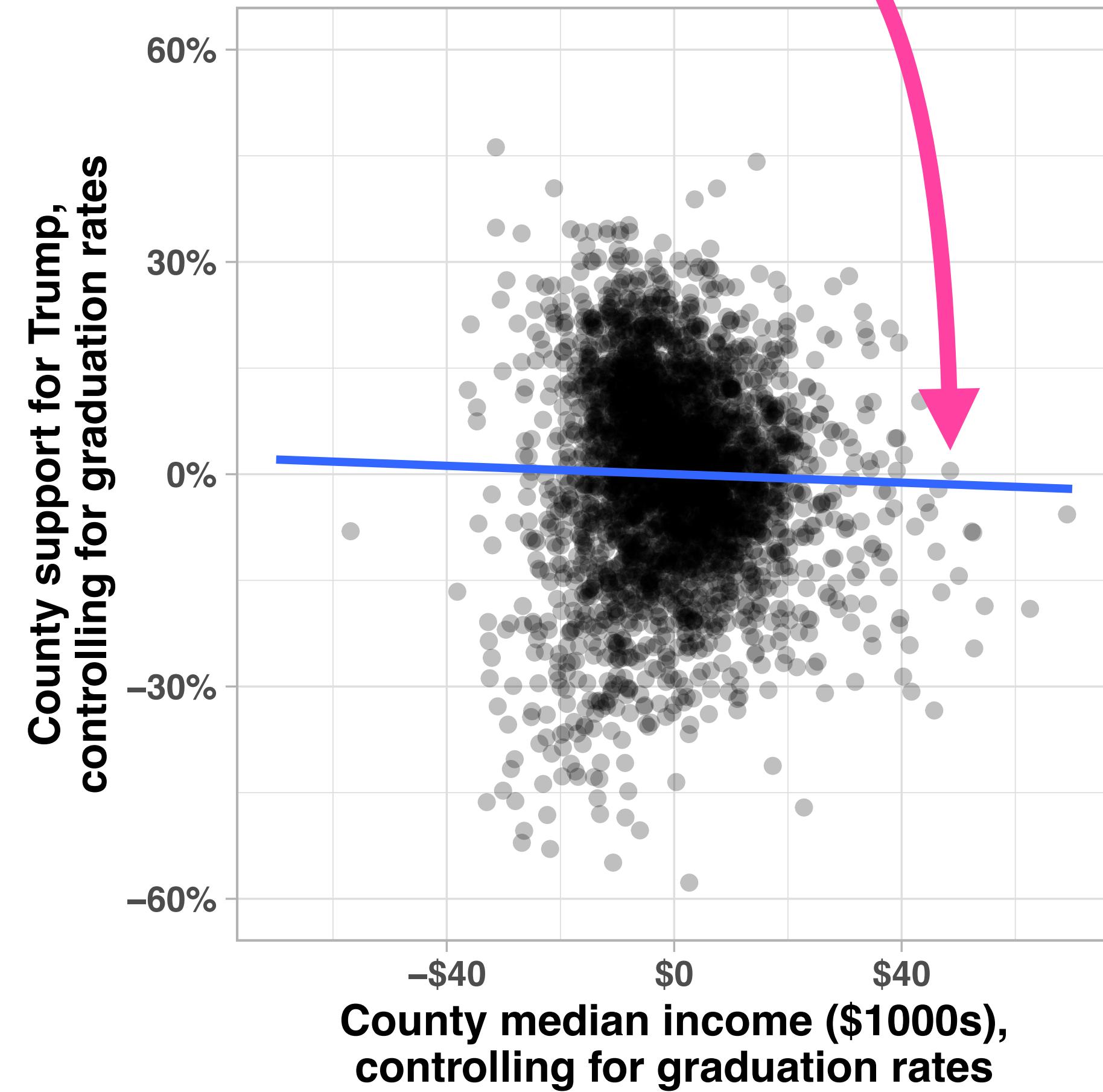
i.e.  $\hat{\alpha}_1$  is to the right of  $\beta_1$



# 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?**