

# Welcome back!

**Nameplates please. And technology encouraged today!**

**All TF materials are available at [github.com/nolankav/api-202](https://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

R

Nolan M. Kavanagh  
February 7, 2025





I'm not like a regular  
I'm a cool TF Right, Regina?



Mister Duplicity  
I hate you in  
the middle 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.

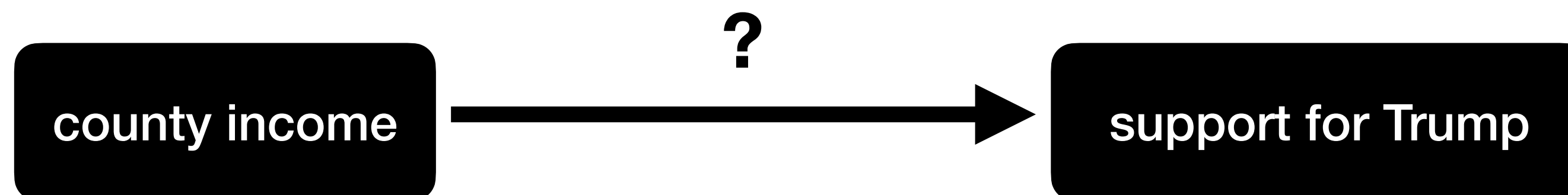
# Overview of our sample data

## 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

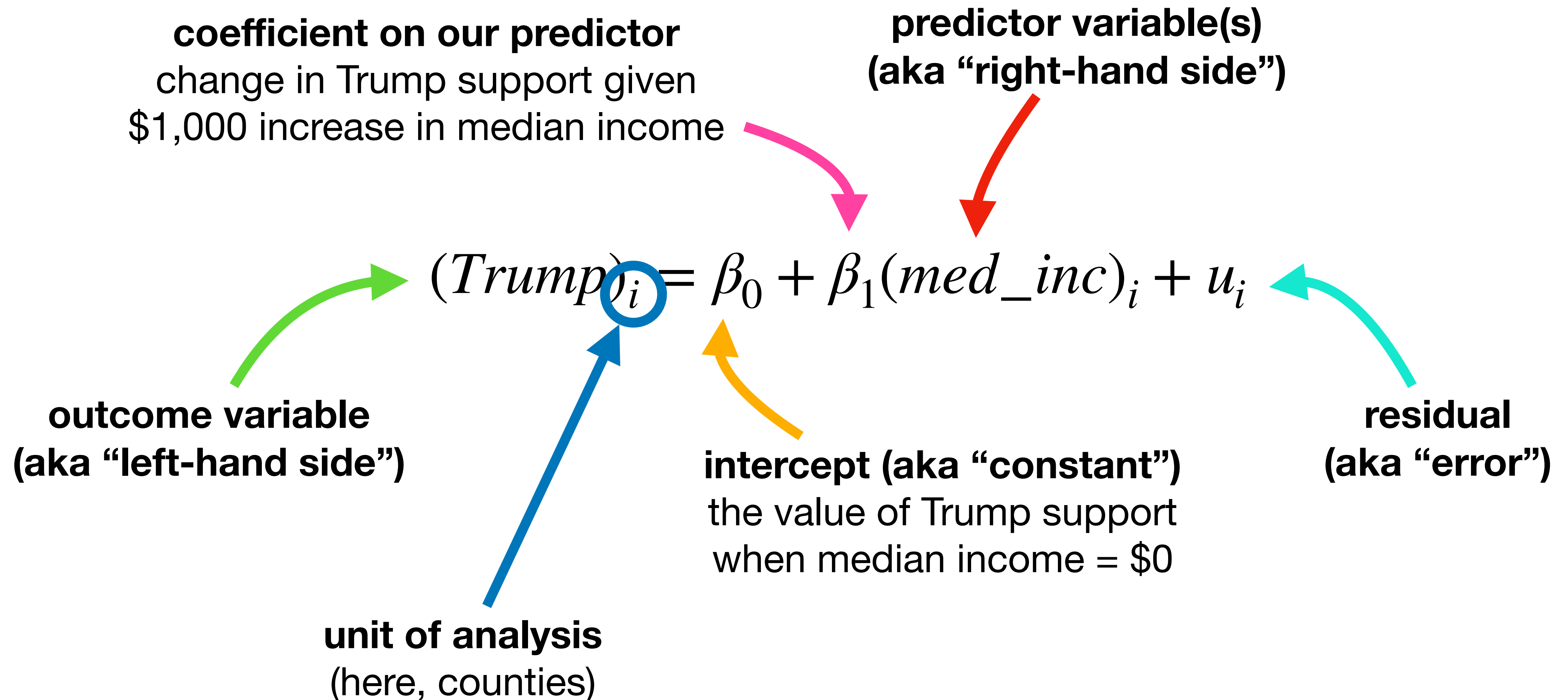
# Hmm, I have an idea!

**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?

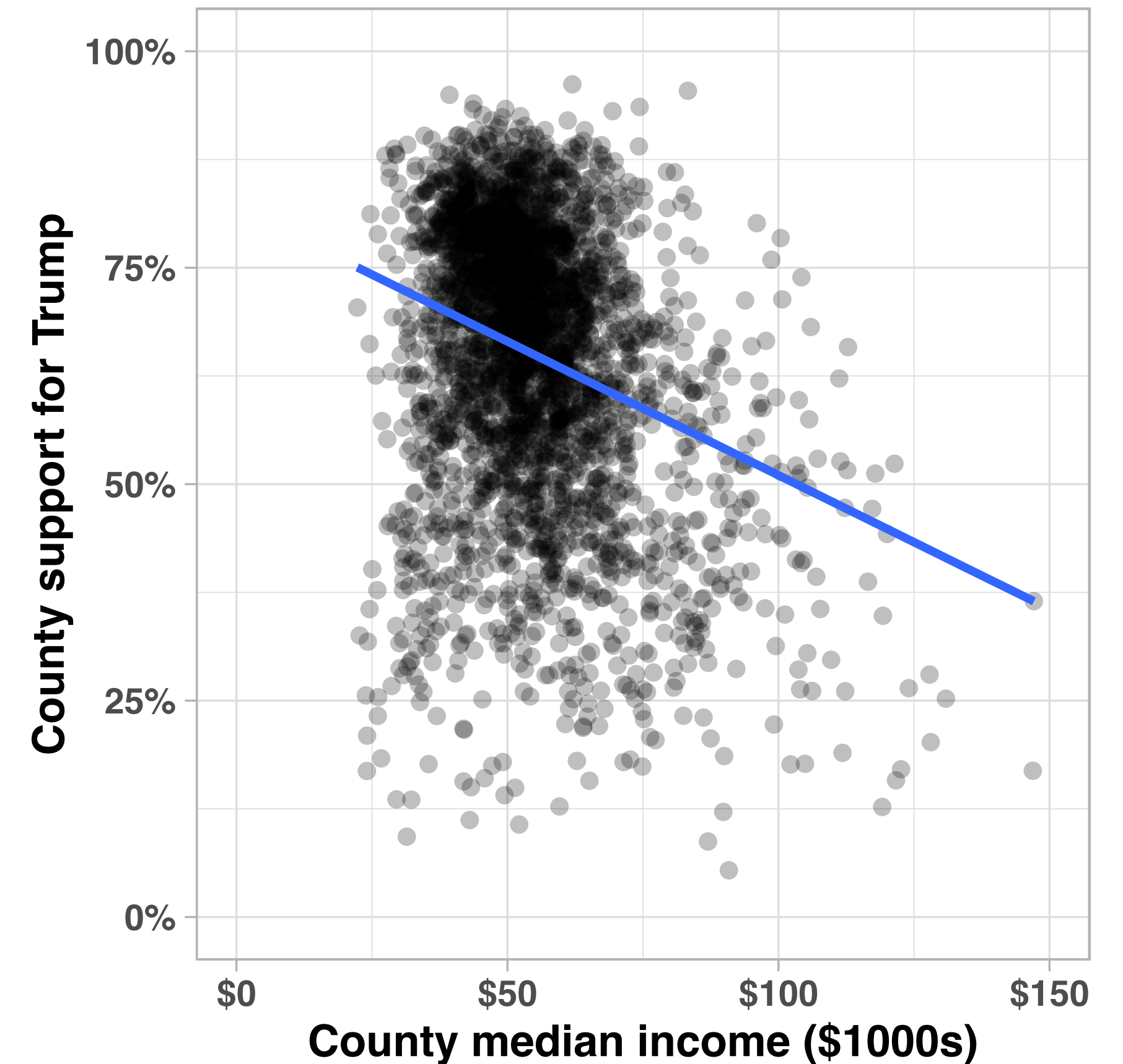
```
# Graph median income and Trump support
plot_1 <- ggplot(df, aes(x=med_inc_000s, y=pc_trump)) +

# Add scatterplot points
geom_point(alpha=0.25) +

# Labels of axes
xlab("County median income (000s)") +
ylab("County support for Trump") +

# Add best fit line
geom_smooth(method="lm", se=F, formula = y~x) +

# Cosmetic changes
theme_light() + theme(text = element_text(face="bold")) +
scale_y_continuous(limits=c(0,100),
                   labels = function(x) paste0(x,"%")) +
scale_x_continuous(limits=c(0,150),
                   labels = scales::dollar_format())
```





# Does the regression check out?

```
# Estimate regression
reg_1 <- lm(pc_trump ~ med_inc_000s, data=df)
summary(reg_1)
```

Call:  
lm(formula = pc\_trump ~ med\_inc\_000s, data = df)

Residuals:

Min	1Q	Median	3Q	Max
-62.940	-8.985	3.256	11.042	39.239

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	81.93207	1.07913	75.92	<2e-16 ***
med_inc_000s	-0.30905	0.01899	-16.28	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.5 on 3112 degrees of freedom

Multiple R-squared: 0.07845, Adjusted R-squared: 0.07815

F-statistic: 264.9 on 1 and 3112 DF, p-value: < 2.2e-16

Looks right to me!

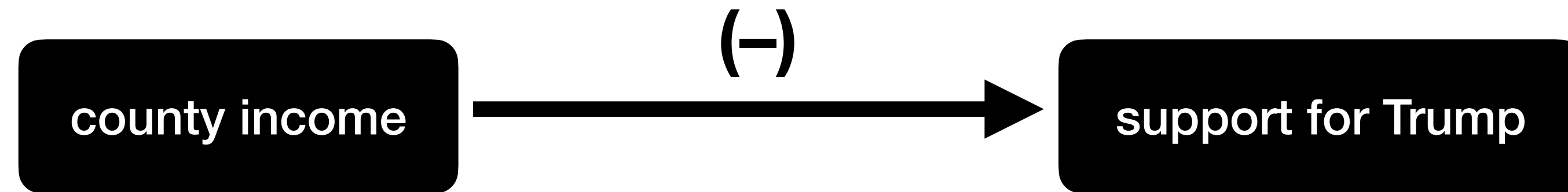


So Trump was all about  
economic grievances.

Case closed!

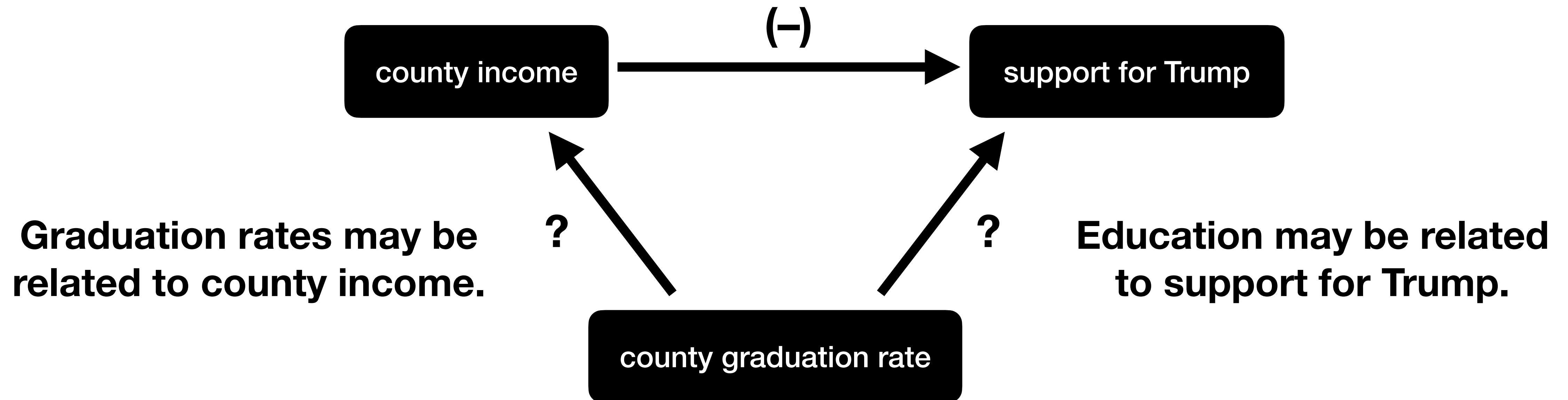
**Note:** Trump support is measured from 0–100, so we don't have to multiply by 100 to interpret the coefficients.

# 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.

```
# Estimate long regression  
reg_2 <- lm(pc_trump ~ med_inc_000s + pc_hs_grad, data=df)  
summary(reg_2)
```


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

**To include multiple predictors in our regression, we just add them to the right-hand side with a “+”.**

# Let's run the long regression.

```
# Estimate long regression
reg_2 <- lm(pc_trump ~ med_inc_000s + pc_hs_grad, data=df)
summary(reg_2)
```

Call:

```
lm(formula = pc_trump ~ med_inc_000s + pc_hs_grad, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-57.641	-8.134	0.859	9.436	45.269

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	134.44740	2.30947	58.216	<2e-16 ***
med_inc_000s	-0.02966	0.02058	-1.442	0.149
pc_hs_grad	-1.02700	0.04086	-25.135	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.13 on 3111 degrees of freedom  
Multiple R-squared: 0.234, Adjusted R-squared: 0.2335  
F-statistic: 475.2 on 2 and 3111 DF, p-value: < 2.2e-16

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



# Let's run the long regression.

```
# Estimate long regression
reg_2 <- lm(pc_trump ~ med_inc_000s + pc_hs_grad, data=df)
summary(reg_2)
```

```
Call:
lm(formula = pc_trump ~ med_inc_000s + pc_hs_grad, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-57.641	-8.134	0.859	9.436	45.269

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	134.44740	2.30947	58.216	<2e-16 ***
med inc 000s	-0.02966	0.02058	-1.442	0.149
pc_hs_grad	-1.02700	0.04086	-25.135	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.13 on 3111 degrees of freedom  
Multiple R-squared: 0.234, Adjusted R-squared: 0.2335  
F-statistic: 475.2 on 2 and 3111 DF, p-value: < 2.2e-16

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

# Let's run the long regression.

```
# Estimate long regression
reg_2 <- lm(pc_trump ~ med_inc_000s + pc_hs_grad, data=df)
summary(reg_2)
```

Call:

```
lm(formula = pc_trump ~ med_inc_000s + pc_hs_grad, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-57.641	-8.134	0.859	9.436	45.269

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	134.44740	2.30947	58.216	<2e-16 ***
med inc 000s	-0.02966	0.02058	-1.442	0.149
pc_hs_grad	-1.02700	0.04086	-25.135	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.13 on 3111 degrees of freedom  
Multiple R-squared: 0.234, Adjusted R-squared: 0.2335  
F-statistic: 475.2 on 2 and 3111 DF, p-value: < 2.2e-16

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.

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





**She doesn't even**

**explain our outcome  
after controls**



# Let's run the long regression.

```
# Estimate long regression
```

```
reg_2 <- lm(pc_trump ~ med_inc_000s + pc_hs_grad, data=df)
summary(reg_2)
```

Call:

```
lm(formula = pc_trump ~ med_inc_000s + pc_hs_grad, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-57.641	-8.134	0.859	9.436	45.269

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	134.44740	2.30947	58.216	<2e-16 ***
med inc 000s	-0.02966	0.02058	-1.442	0.149
pc hs grad	-1.02700	0.04086	-25.135	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.13 on 3111 degrees of freedom

Multiple R-squared: 0.234, Adjusted R-squared: 0.2335

F-statistic: 475.2 on 2 and 3111 DF, p-value: < 2.2e-16

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



# Let's run the long regression.

```
# Estimate long regression
reg_2 <- lm(pc_trump ~ med_inc_000s + pc_hs_grad, data=df)
summary(reg_2)
```

Call:

```
lm(formula = pc_trump ~ med_inc_000s + pc_hs_grad, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-57.641	-8.134	0.859	9.436	45.269

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	134.44740	2.30947	58.216	<2e-16 ***
med inc 000s	-0.02966	0.02058	-1.442	0.149
pc hs grad	-1.02700	0.04086	-25.135	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.13 on 3111 degrees of freedom  
Multiple R-squared: 0.234, Adjusted R-squared: 0.2335  
F-statistic: 475.2 on 2 and 3111 DF, p-value: < 2.2e-16

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.

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

# Let's run the long regression.

```
# Estimate long regression
reg_2 <- lm(pc_trump ~ med_inc_000s + pc_hs_grad, data=df)
summary(reg_2)
```

Call:

```
lm(formula = pc_trump ~ med_inc_000s + pc_hs_grad, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-57.641	-8.134	0.859	9.436	45.269

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	134.44740	2.30947	58.216	<2e-16 ***
med_inc_000s	-0.02966	0.02058	-1.442	0.149
pc_hs_grad	-1.02700	0.04086	-25.135	<2e-16 ***
---				

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.13 on 3111 degrees of freedom

Multiple R-squared: 0.234, Adjusted R-squared: 0.2335

F-statistic: 475.2 on 2 and 3111 DF, p-value: < 2.2e-16

**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.

```
# Estimate long regression
reg_2 <- lm(pc_trump ~ med_inc_000s + pc_hs_grad, data=df)
summary(reg_2)
```

Call:

```
lm(formula = pc_trump ~ med_inc_000s + pc_hs_grad, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-57.641	-8.134	0.859	9.436	45.269

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	134.44740	2.30947	58.216	<2e-16 ***
med_inc_000s	-0.02966	0.02058	-1.442	0.149
pc_hs_grad	-1.02700	0.04086	-25.135	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.13 on 3111 degrees of freedom  
Multiple R-squared: 0.234, Adjusted R-squared: 0.2335  
F-statistic: 475.2 on 2 and 3111 DF, p-value: < 2.2e-16

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.



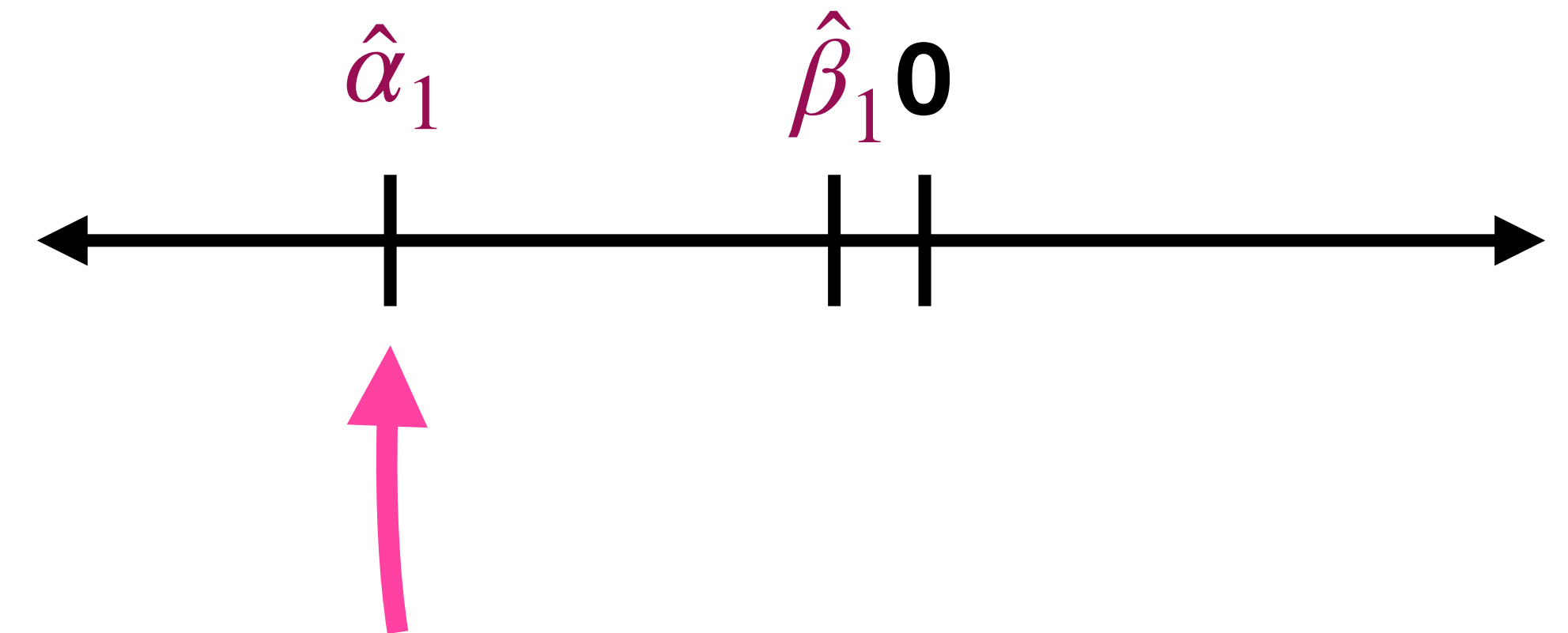
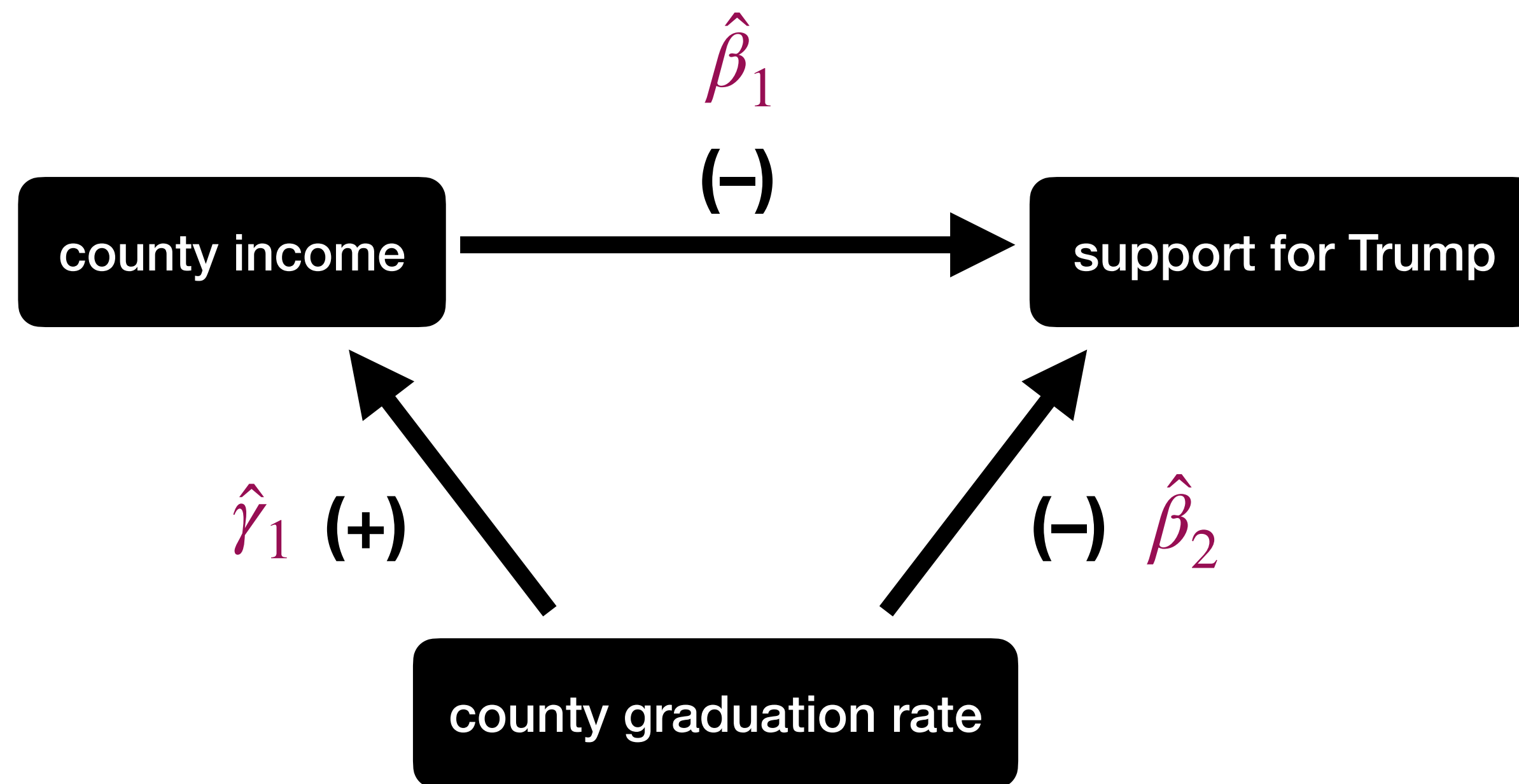
# Womp.

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

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

# Clearly, we were missing something.



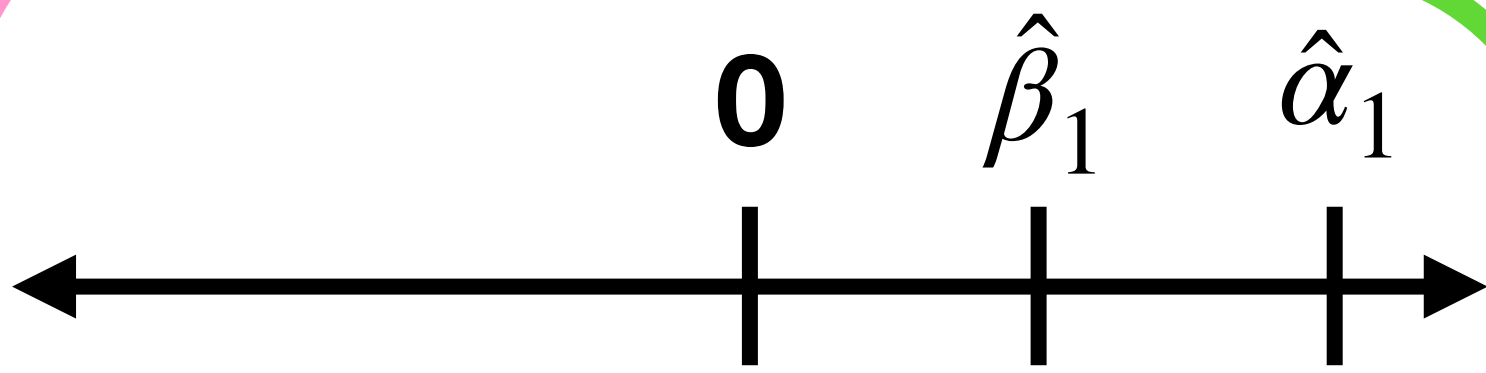
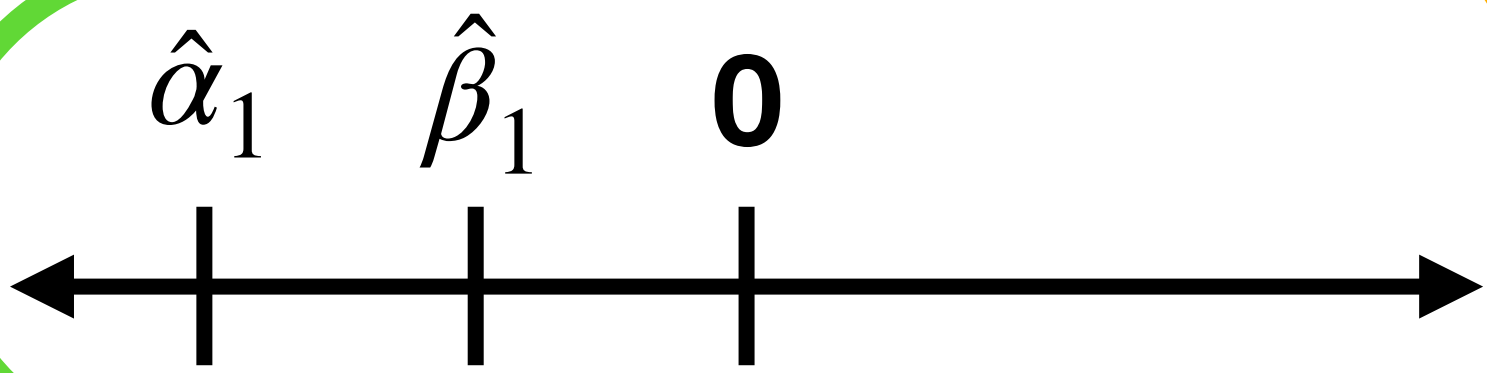
Relative to the true  $\beta_1$   $(-)$ , our estimate of  $\alpha_1$  was even more negative.

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

# Bias: sign or size?

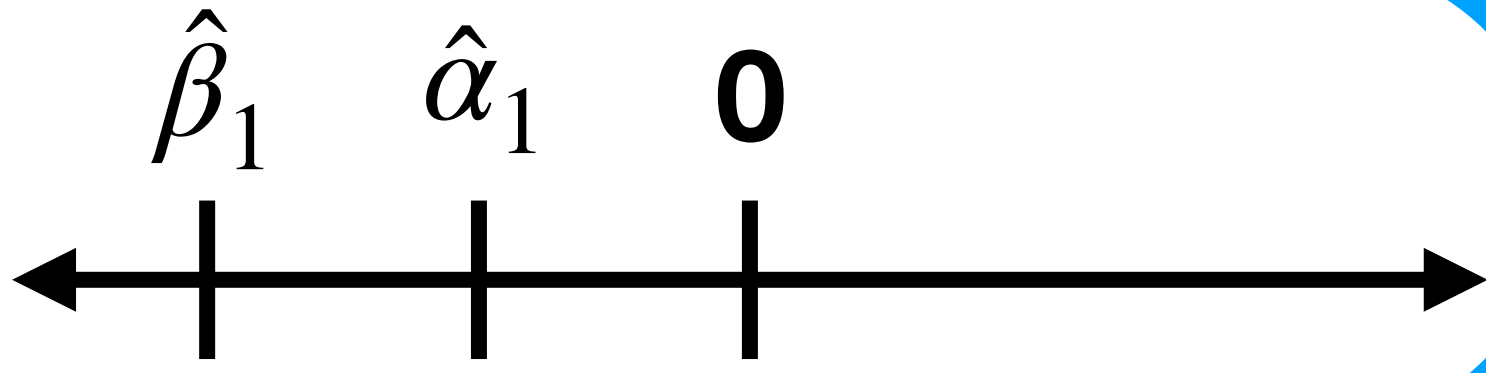
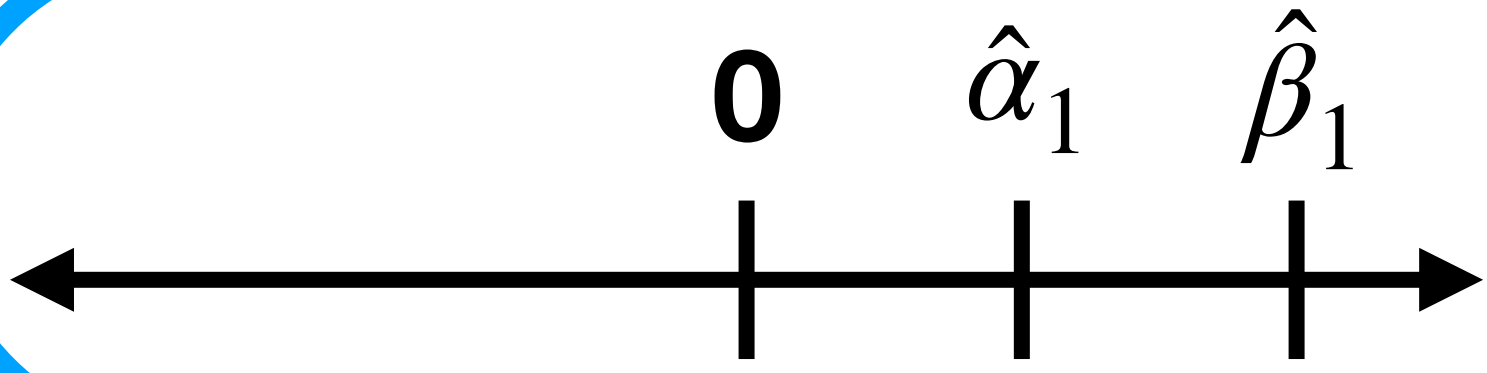
**Overstatement**

i.e.  $\alpha_1$  is farther from 0

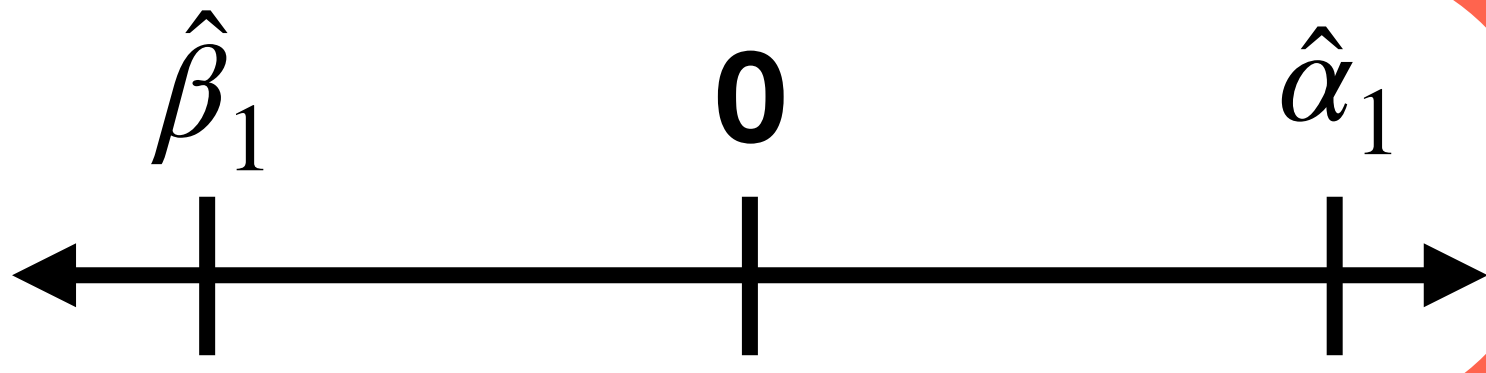
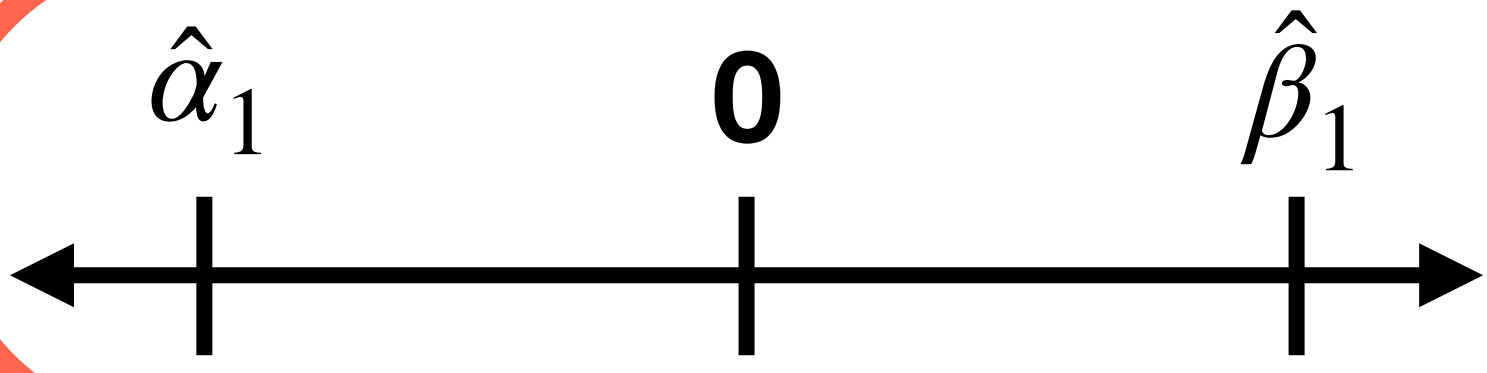


**Understatement**

i.e.  $\alpha_1$  is closer to 0



**Sign flip!**



**Negative bias (-)**

i.e.  $\alpha_1$  is to the left of  $\beta_1$

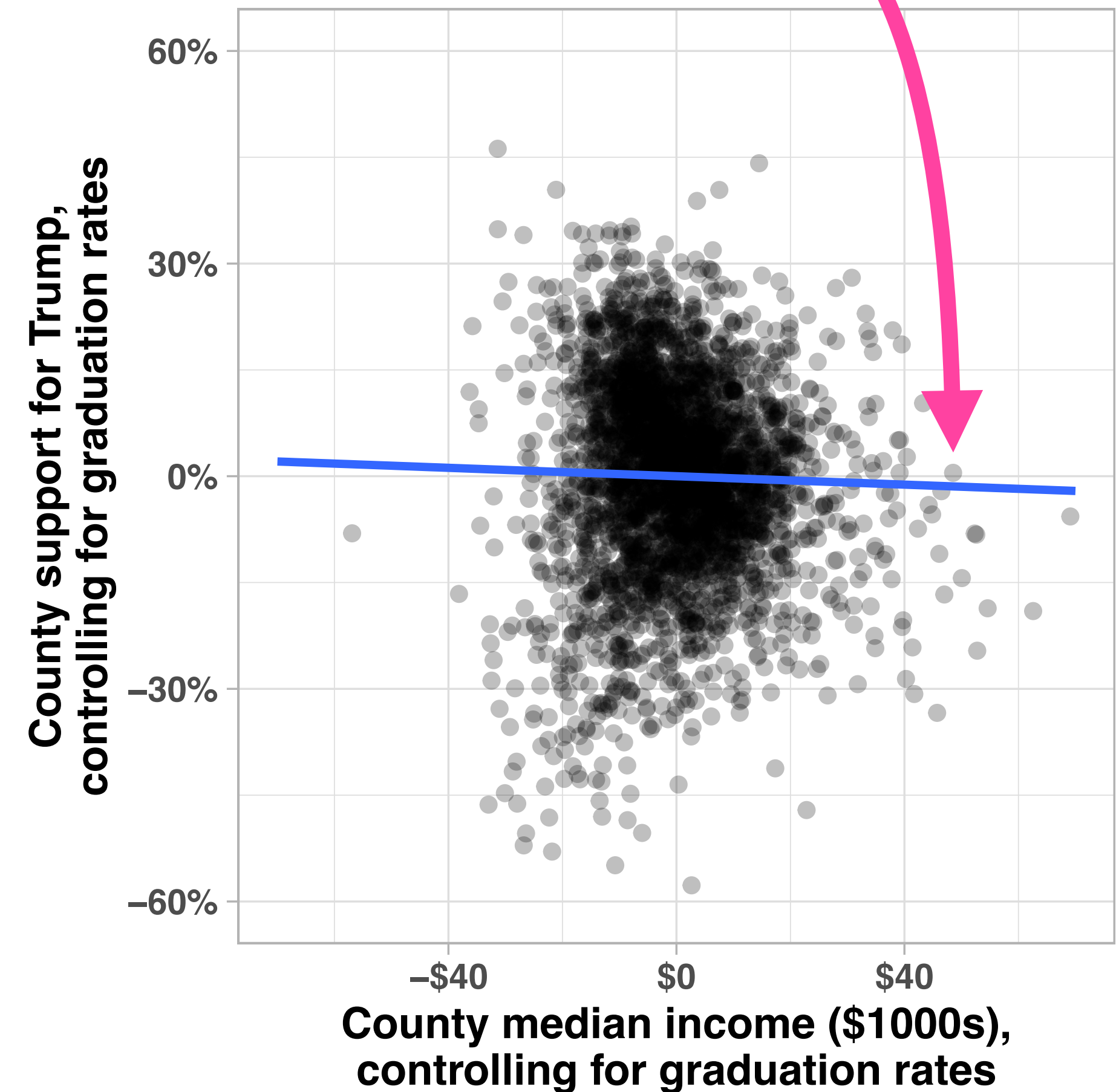
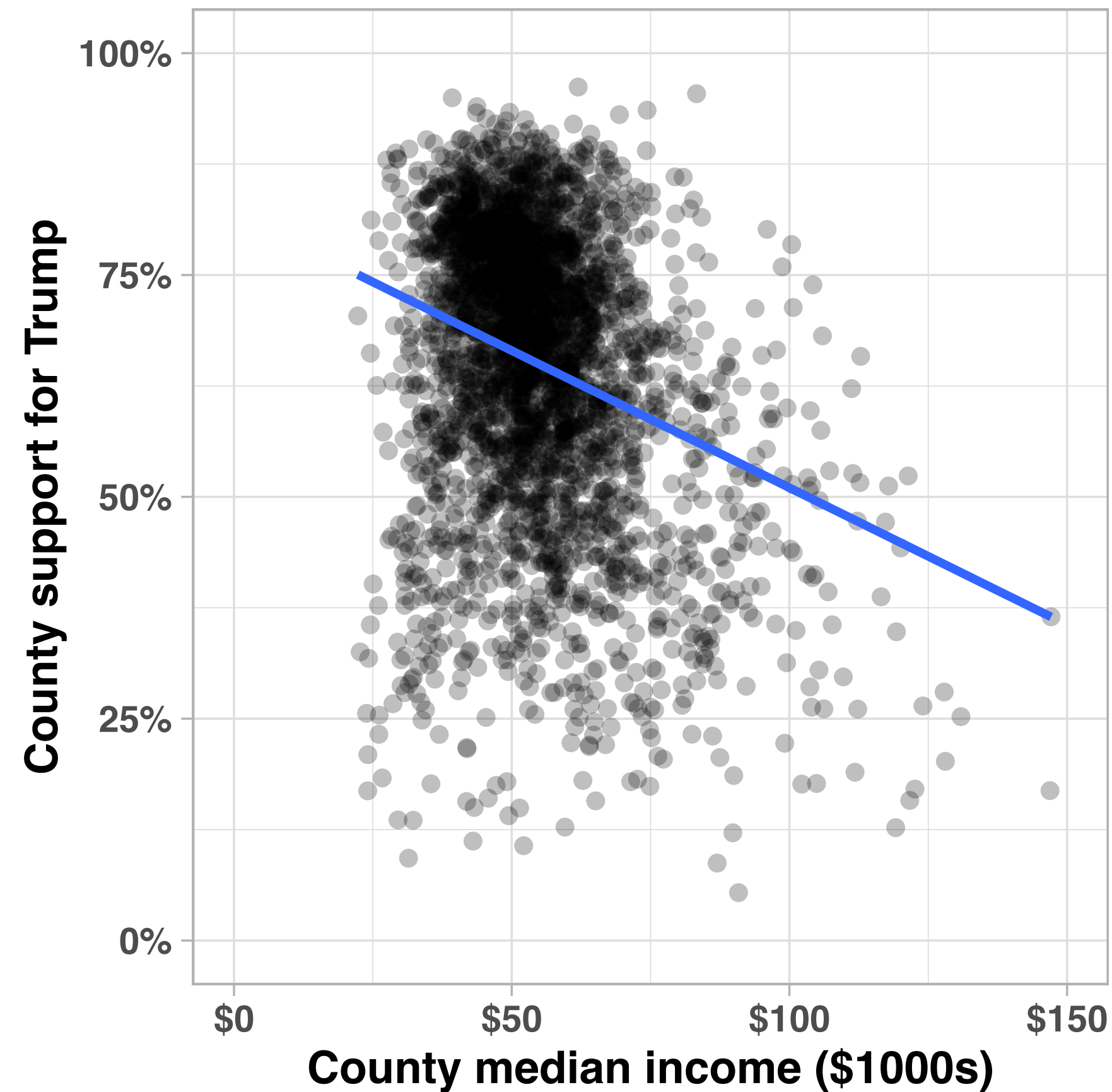
**Positive bias (+)**

i.e.  $\alpha_1$  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*?**