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-1")`

In Excel: http://tinyurl.com/api-202-tf-2

Multiple regression and omitted variables

API 202: TF Session 2

R

Nolan M. Kavanagh
February 2, 2024

Goals for today

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

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

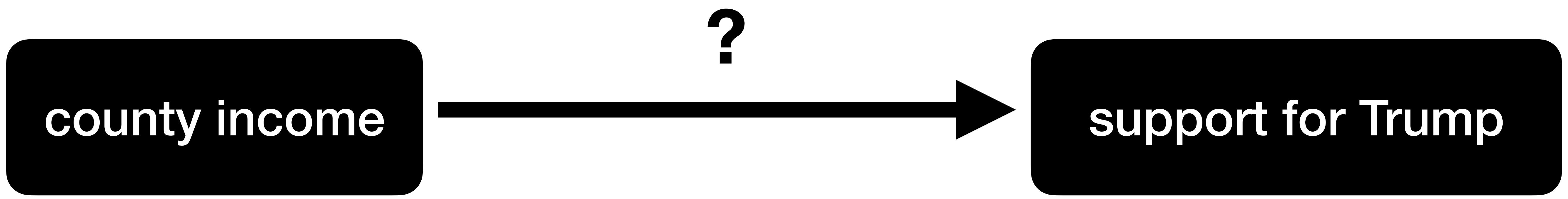
Overview of our sample data

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

coefficient on our predictor
change in Trump support given
$1,000 increase in median income

predictor variable(s)
(aka “right-hand side”)

$$(Trump)_i = \beta_0 + \beta_1 (med\_inc)_i + u_i$$

outcome variable
(aka “left-hand side”)

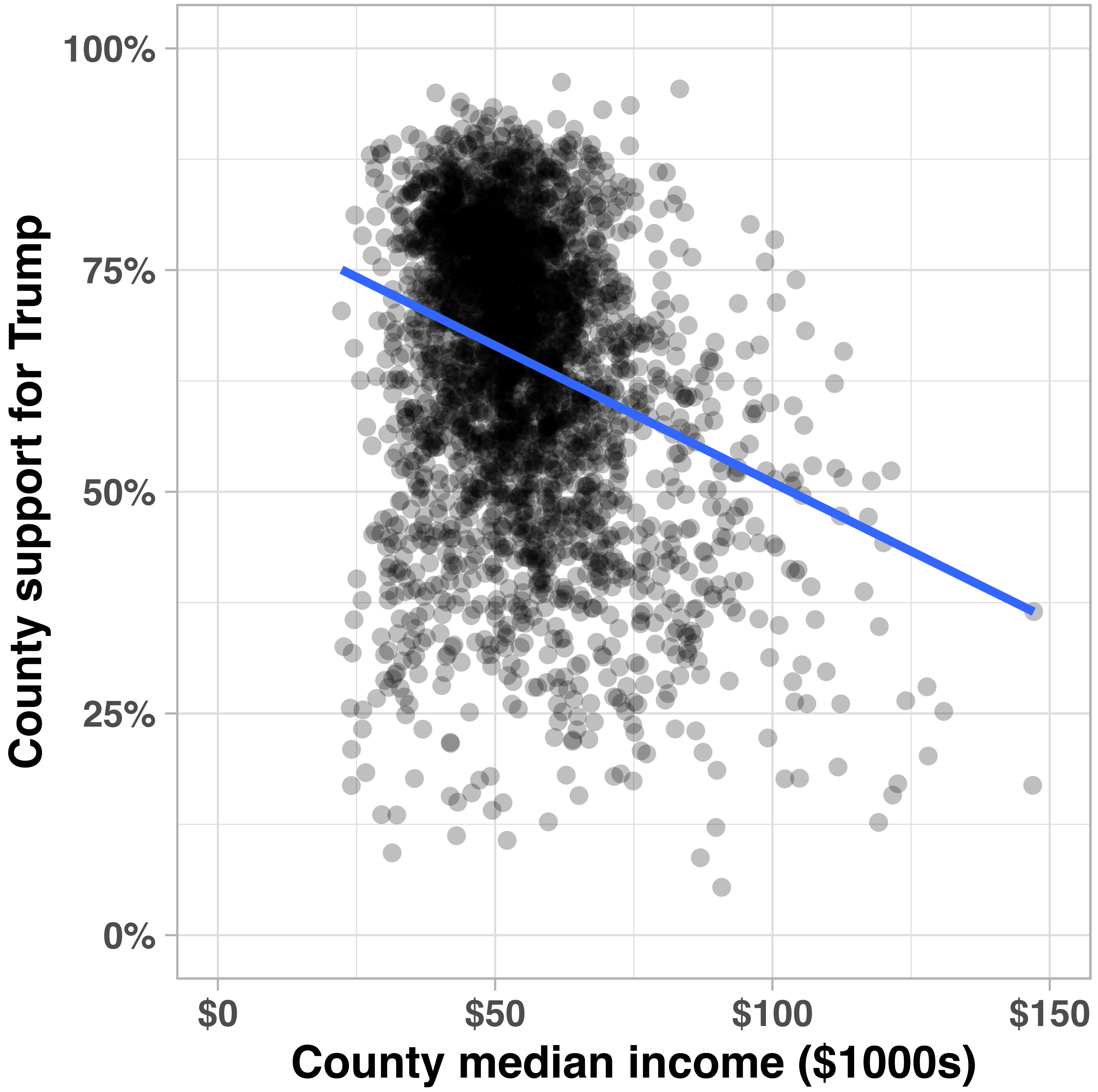
unit of analysis
(here, counties)

intercept (aka “constant”)
the value of Trump support
when median income = $0

residual
(aka “error”)

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?

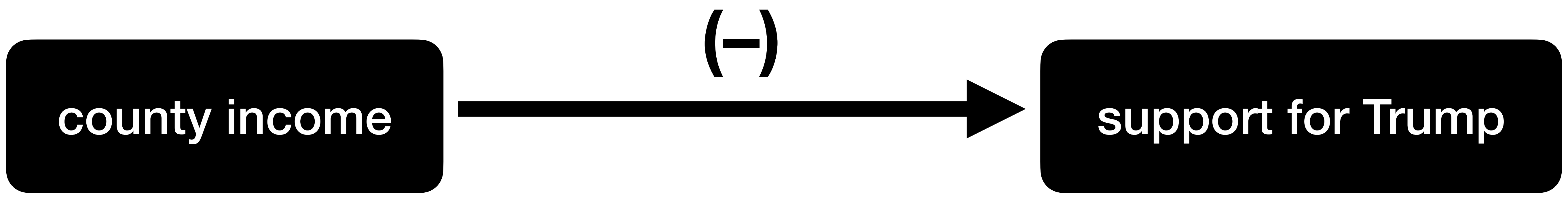
```
# Estimate regression
reg_1 <- lm(pc_trump ~ med_inc_000s, data=df)
summary(reg_1)

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!

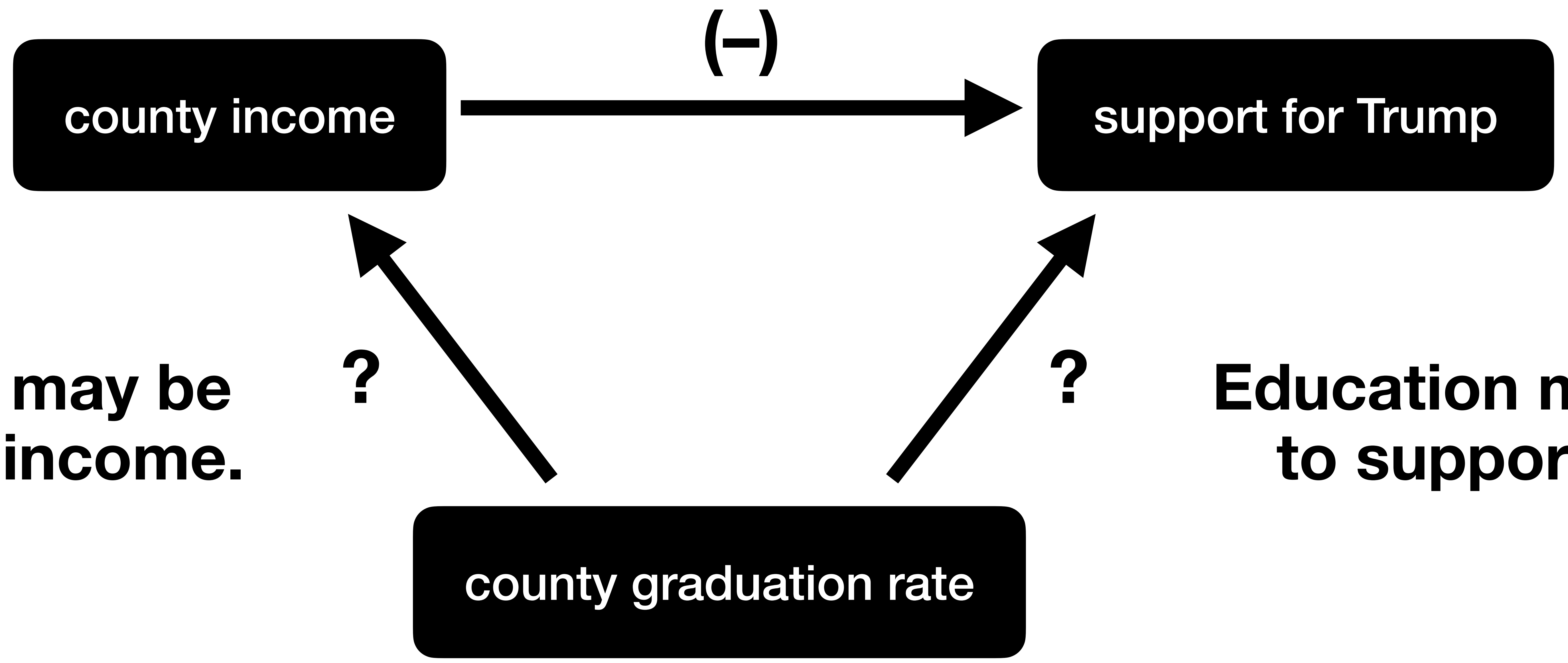
So Trump was all about economic grievances.

Case closed!

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

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

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)
```

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

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
```

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. Shit.

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
predict 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
```

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 adjusted association is statistically significant at the 5% level.

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.)

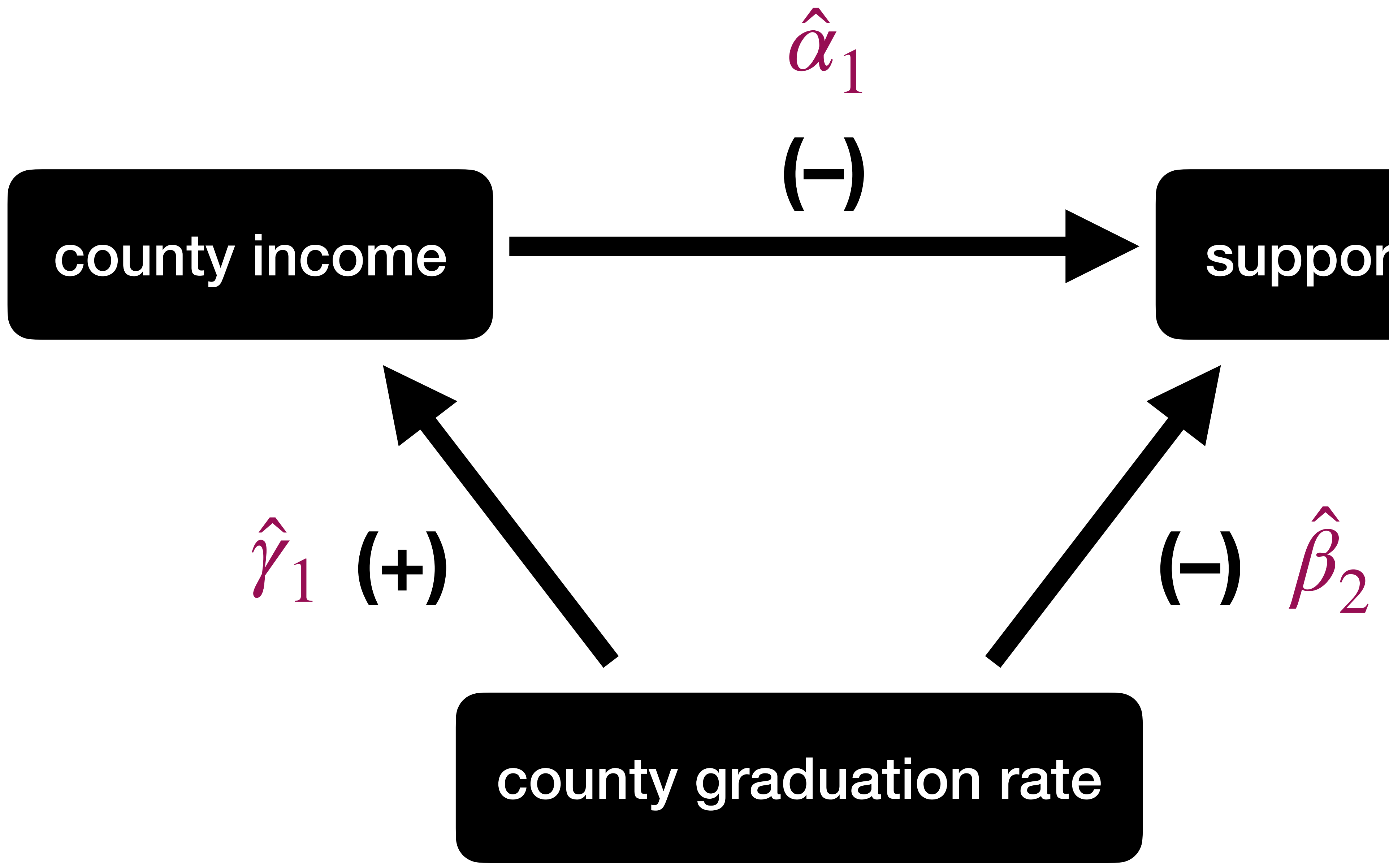
Womp.

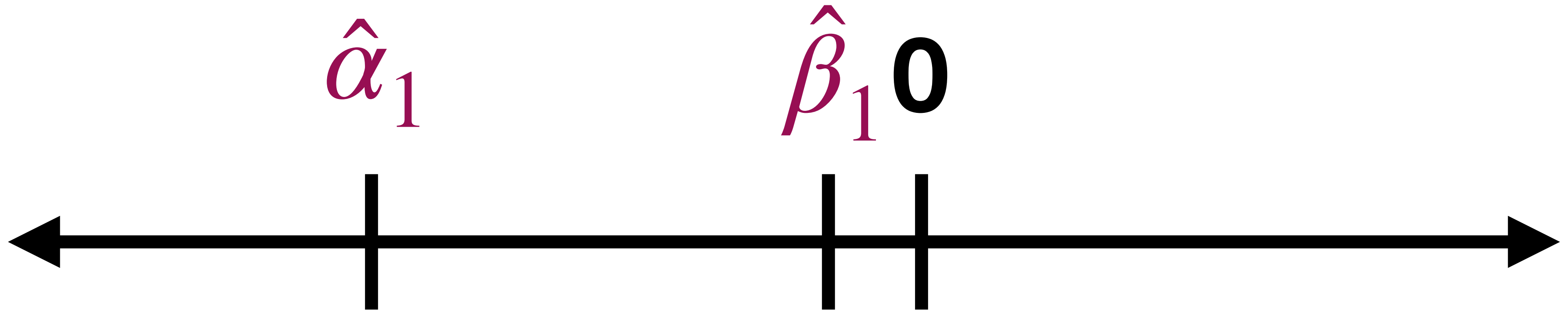
		Model 1	Model 2	
Intercept		**81.93**	**134.45**	
		(-1.08)	(-2.31)	
		P<0.001	P<0.001	
County median income ($1000s)	$\hat{\alpha}_1$	**-0.31**	**-0.03**	$\hat{\beta}_1$
		(0.02)	(0.02)	
		P<0.001	P=0.149	
County graduation rate			**-1.03**	$\hat{\beta}_2$
			(0.04)	
			P<0.001	
Num.Obs.		3114	3114	
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 $(Trump)_i = \beta_0 + \beta_1(med\_inc)_i + \beta_2(HS\_grad)_i + u_i$

Clearly, we were missing something.

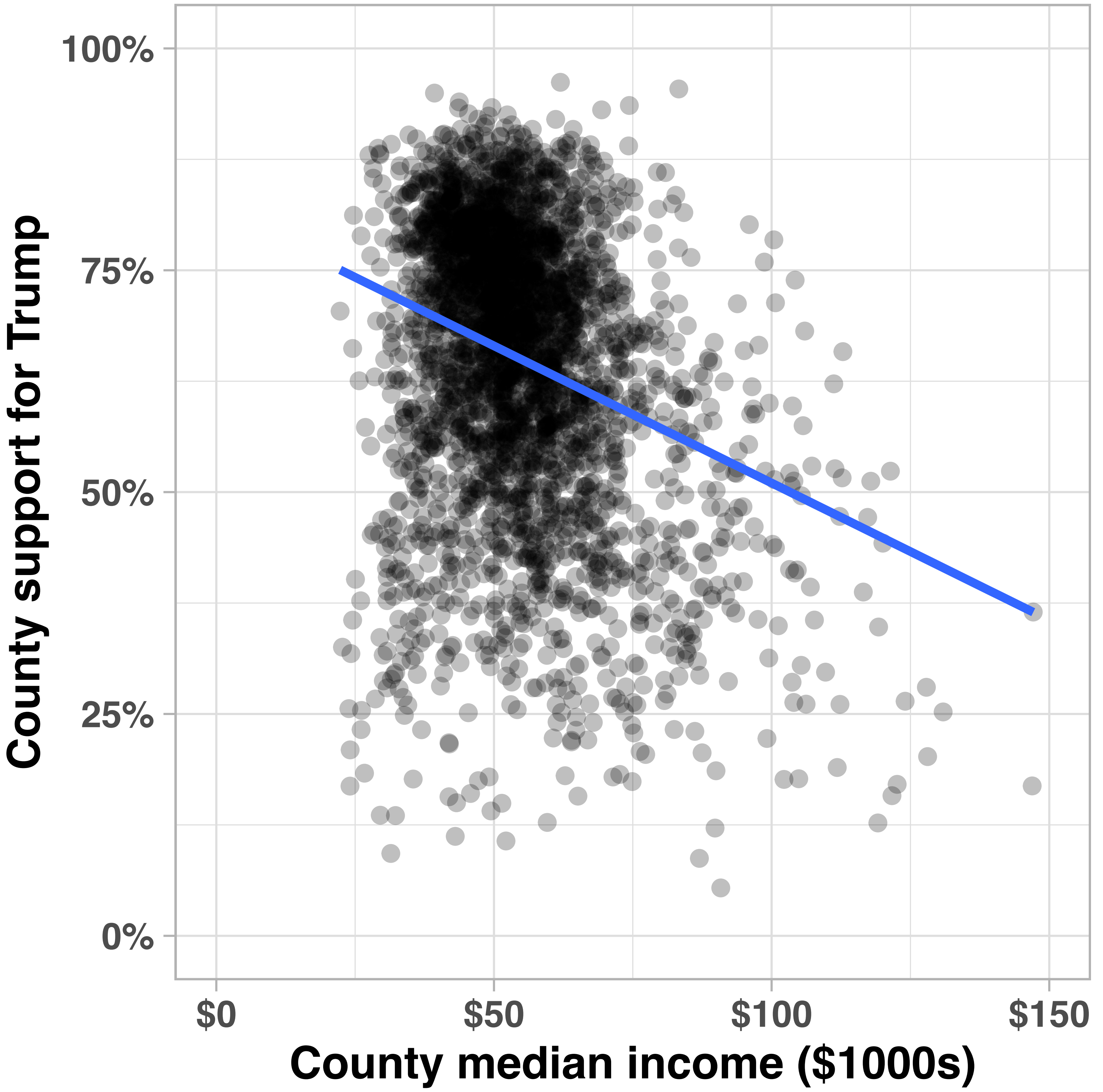




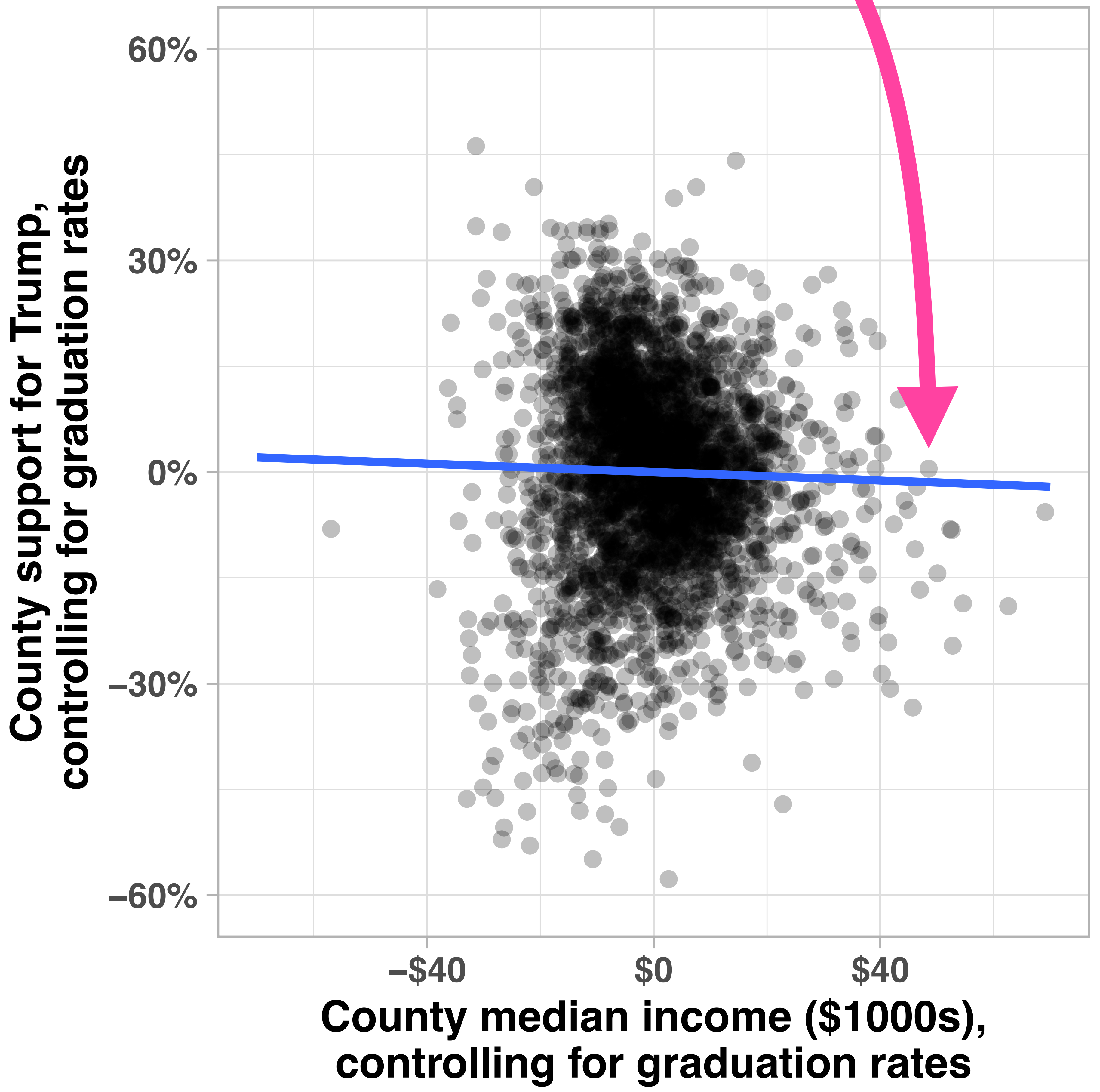
Relative to the true β_1 (–), our estimate of α_1 is even more negative.

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

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 script file, 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*?