

Welcome!

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

What's the deal with regression?

API 202: TF Session 1

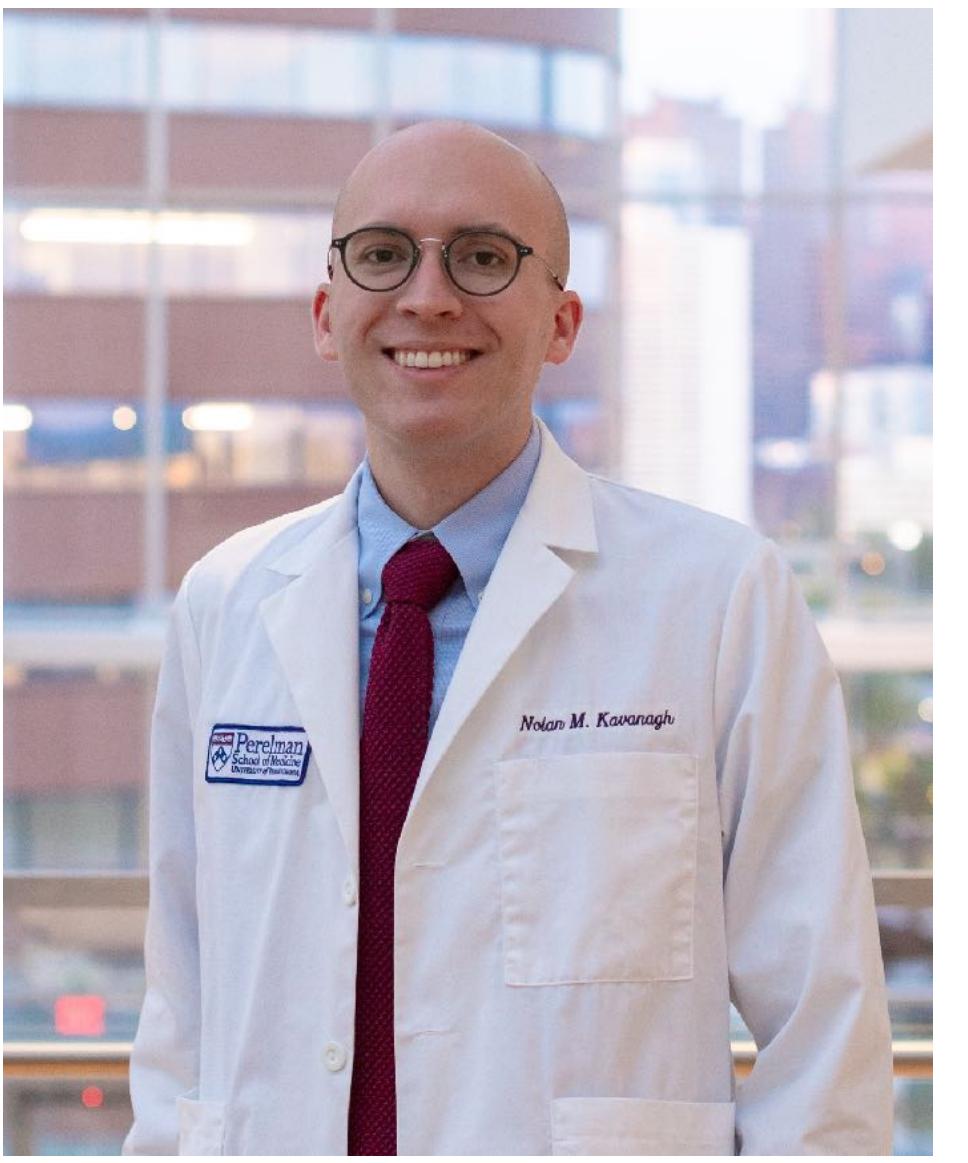
Nolan M. Kavanagh
January 31, 2025



Goals for today

- 1. Get to know each other a little better.**
- 2. Review regression notation, including the PRF vs. SRF.**
- 3. Learn how to graph bivariate relationships.**
- 4. Learn how to run bivariate regressions.**
- 5. Review how to interpret bivariate regressions.**

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



MD/PhD student in health policy.

**"I don't need backups.
I'm going to Harvard."**



Hi, I'm Nolan.



GO BLUE!

My research is on the politics of health.

American Political Science Review (2021) 115, 3, 1104–1109
doi:10.1017/S0003055421000265 © The Author(s), 2021. Published by Cambridge University Press on behalf of the American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Letter Does Health Vulnerability Predict Voting for Right-Wing Populist Parties in Europe?

NOLAN M. KAVANAGH *University of Pennsylvania*
ANIL MENON *University of Michigan*
JUSTIN E. HEINZE *University of Michigan*

Why do voters in developed democracies support right-wing populist parties? Existing research focuses on economic and cultural vulnerability, but little attention has been paid to the role of health vulnerability. We argue that health vulnerability may have a similar influence on voters. To test this argument, we analyzed all waves of the European Social Survey (2002–2020). Our findings suggest that voters with worse self-reported health were significantly more likely to vote for right-wing populist parties. The relationship persists even after accounting for measures of cultural and economic vulnerability, as well as life satisfaction, both their personal lives and their country's health system. The influence of health on support for right-wing populist parties appears to be greater than that of income and self-reported economic insecurity, while less than that of gender and attitudes about immigration. Our findings suggest that policies affecting public health could shape not only health outcomes but also the political landscape.

INTRODUCTION

Right-wing populist parties are surging in popularity across the Western world (Norris and Inglehart 2010). Why do voters in developed democracies support such parties? A growing body of research has identified economic insecurity and cultural backlash as potential drivers of recent populist successes (Algan et al. 2007; Hochschild 2016; Inglehart and Norris 2016; Kavanagh and Menon 2018; Rodrik 2018; Smith and Hanley 2018). According to these explanations, once-dominant socioeconomic groups perceive an erosion of their economic opportunities or a threat to their privileged status in society. These threats lead voters to perceive vulnerability in motivating them to support parties that promise to restore their socioeconomic standing through anti-multiculturalism, antiglobalisation, and anti-immigration (Inglehart and Norris 2016).

We argue that voters' perceived health may similarly contribute to populist support via a similar mechanism. The development of illness and disability often produces frustration with one's physical and emotional limitations, and it prompts people to compare themselves with their peers (Banks, Gibson, and Baum 2013; Martz and Liao 2007).

This experience may increase an individual's sense of personal vulnerability regardless of their socioeconomic background. They may then blame their misfortune on the political system, and this may lead to a sense of existing local, political, and economic structures (Lacouture 2005; Nussbaum 2018). If true, individuals who suffer poorer health and more disability would be desirous of changing the political status quo. This desire to change the political system may draw them toward parties that campaign for a fundamental restructuring of a "biased and broken" system.

As such, health-related vulnerability may contribute to the rise of right-wing populism. Indeed, some research has associated declining population health with right-wing populist voting. U.S. counties that experienced the greatest rise in mortality over recent decades, especially among whites, were most likely to vote for President Donald Trump in the 2016 election (Kopstein et al. 2016; Bilal, Knapp, and Cooper 2018; Bor 2017). Similar associations have been shown for rates of chronic opioid use (Goodwin et al. 2018) and other markers of poor health (Wasylyshyn, et al., 2018; Kopstein et al. 2017). In the U.K., voters that experienced greater rises in "deaths of despair" due to suicide or drug overdose in the previous decade were more likely to vote for Brexit (Koltai et al. 2019). However, the relationship between poor health and right-wing populist voting remains to be investigated after the COVID-19 virus, appropriate controls for economic and cultural vulnerability.

Understanding how poor health influences right-wing populist support could have important implications for

Nolan M. Kavanagh Medical student, Perleman School of Medicine, University of Pennsylvania; Lecturer, Department of Periodontics and Oral Medicine, School of Dentistry, University of Michigan; nolan.m.kavanagh@perlemanmedicine.upenn.edu

Justin E. Heinze Assistant Professor, Department of Political Science, University of Michigan; jheinze@umich.edu

Received: June 26, 2020; revised: February 27, 2021; accepted: March 23, 2021. First published online: April 26, 2021.

1104

My go-to karaoke song is "Since U Been Gone."



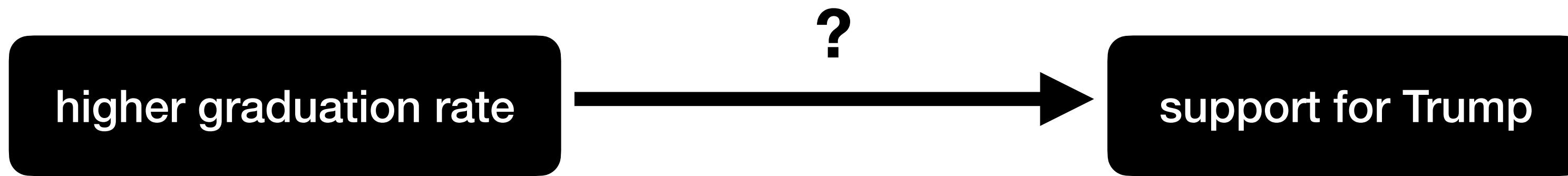
Overview of our sample data

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

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

Tell me a story.

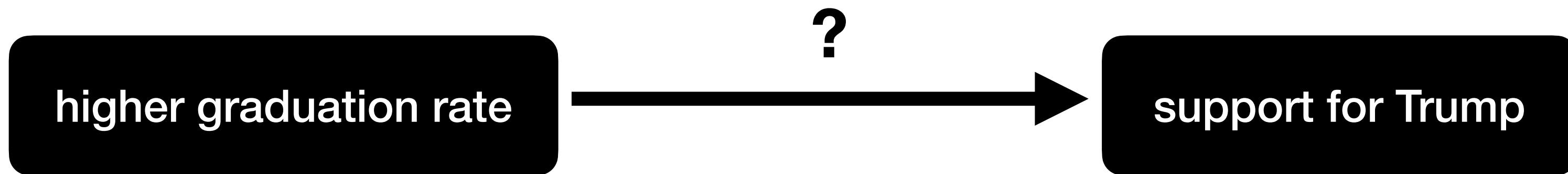
Let's say we're interested in the relationship between high school graduation and support for Trump.



What might be the direction? Mechanism?

Tell me a story.

Let's say we're interested in the relationship between high school graduation and support for Trump.

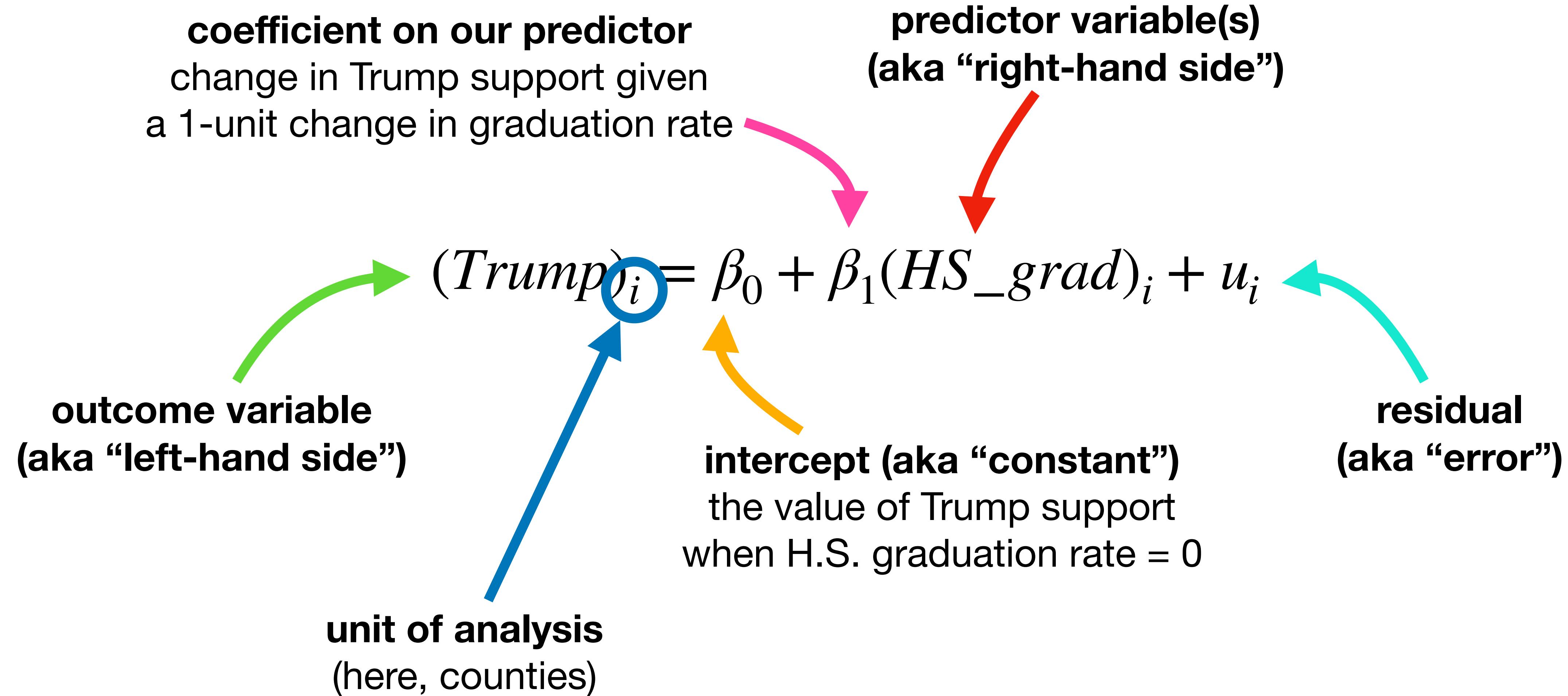


What might be the direction? Mechanism?

More education = liberal values = prefer multiculturalism?

More education = more income = prefer lower taxes?

Population regression function



Population regression function

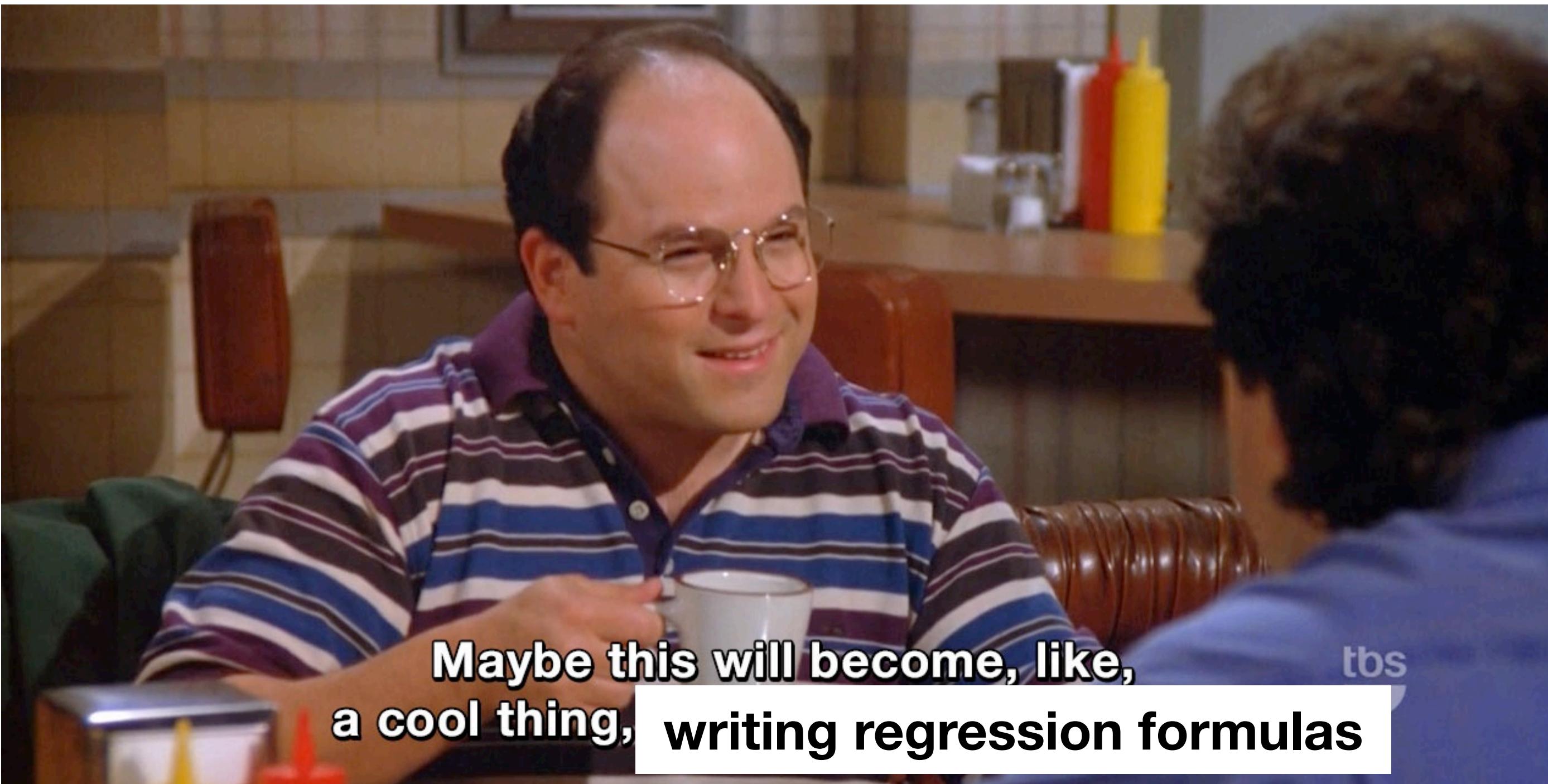
$$(Trump)_i = \beta_0 + \beta_1(HS_grad)_i + u_i$$

Sample regression function

$$(Trump)_i = \hat{\beta}_0 + \hat{\beta}_1(HS_grad)_i + \hat{u}_i$$

We add “hats” to signify estimated values in our sample.





Maybe this will become, like,
a cool thing, writing regression formulas

tbs

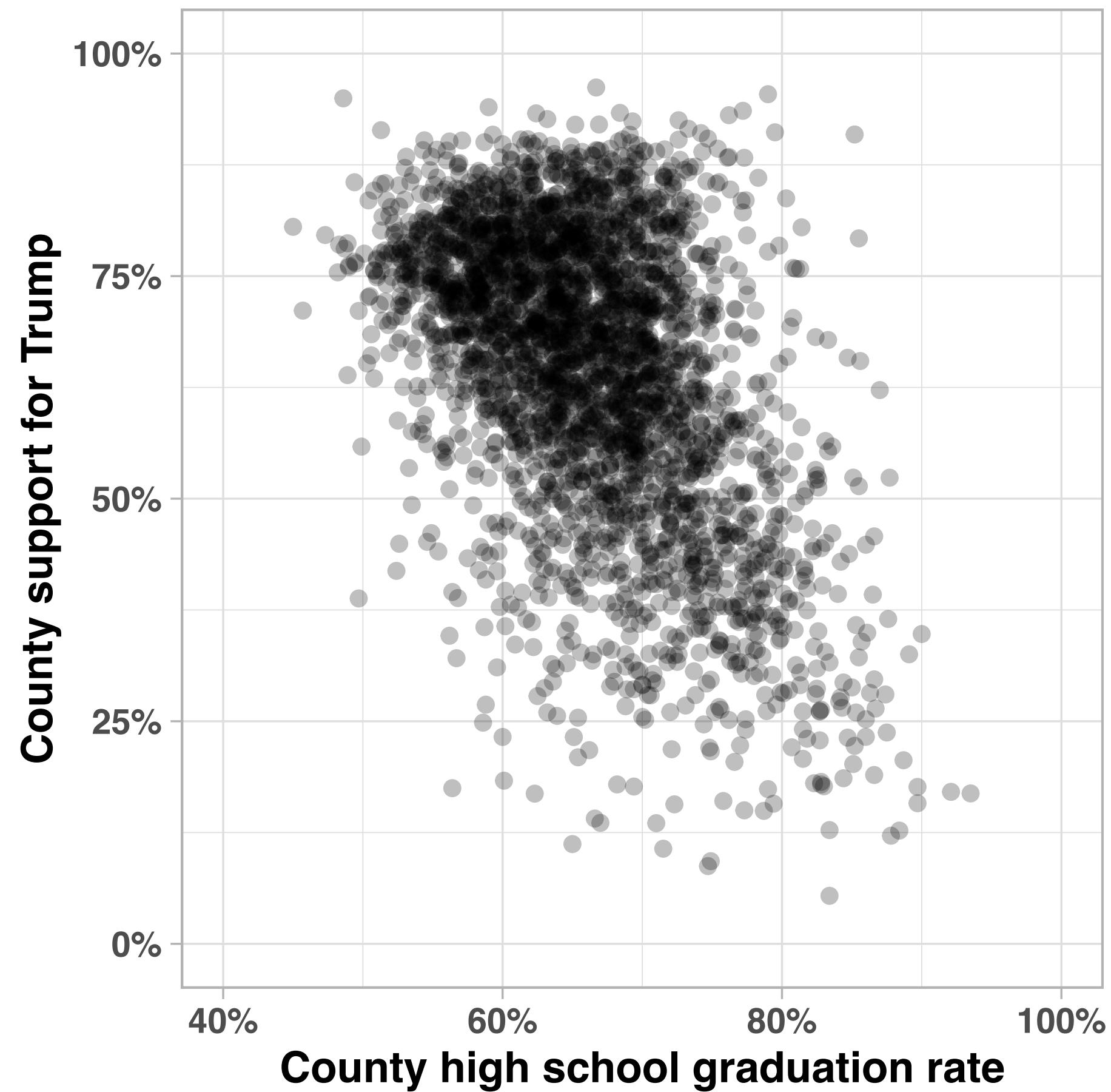


Then maybe baldness will catch on.

tbs

Let's graph our data.

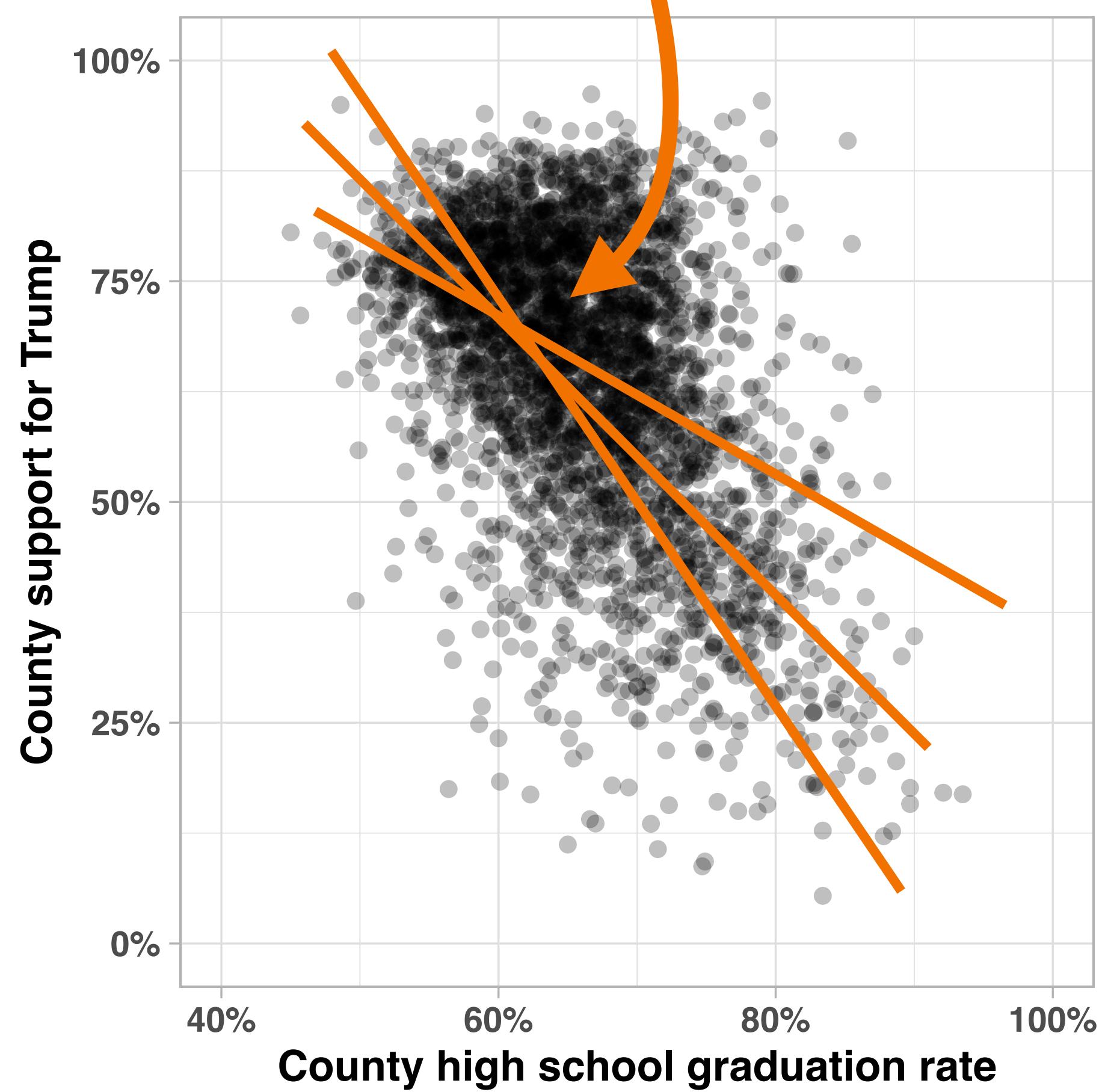
```
# Graph high school graduation and Trump support
plot_1 <- ggplot(df, aes(x=pc_hs_grad, y=pc_trump)) +
  # Add scatterplot points
  geom_point(alpha=0.25) +
  # Labels of axes
  xlab("County high school graduation rate") +
  ylab("County support for Trump") +
  # Cosmetic changes
  theme_light() + theme(text = element_text(face="bold")) +
  scale_y_continuous(limits=c(0,100),
                     labels = function(x) paste0(x,"%")) +
  scale_x_continuous(limits=c(40,100),
                     labels = function(x) paste0(x,"%"))
```



Let's graph our data.

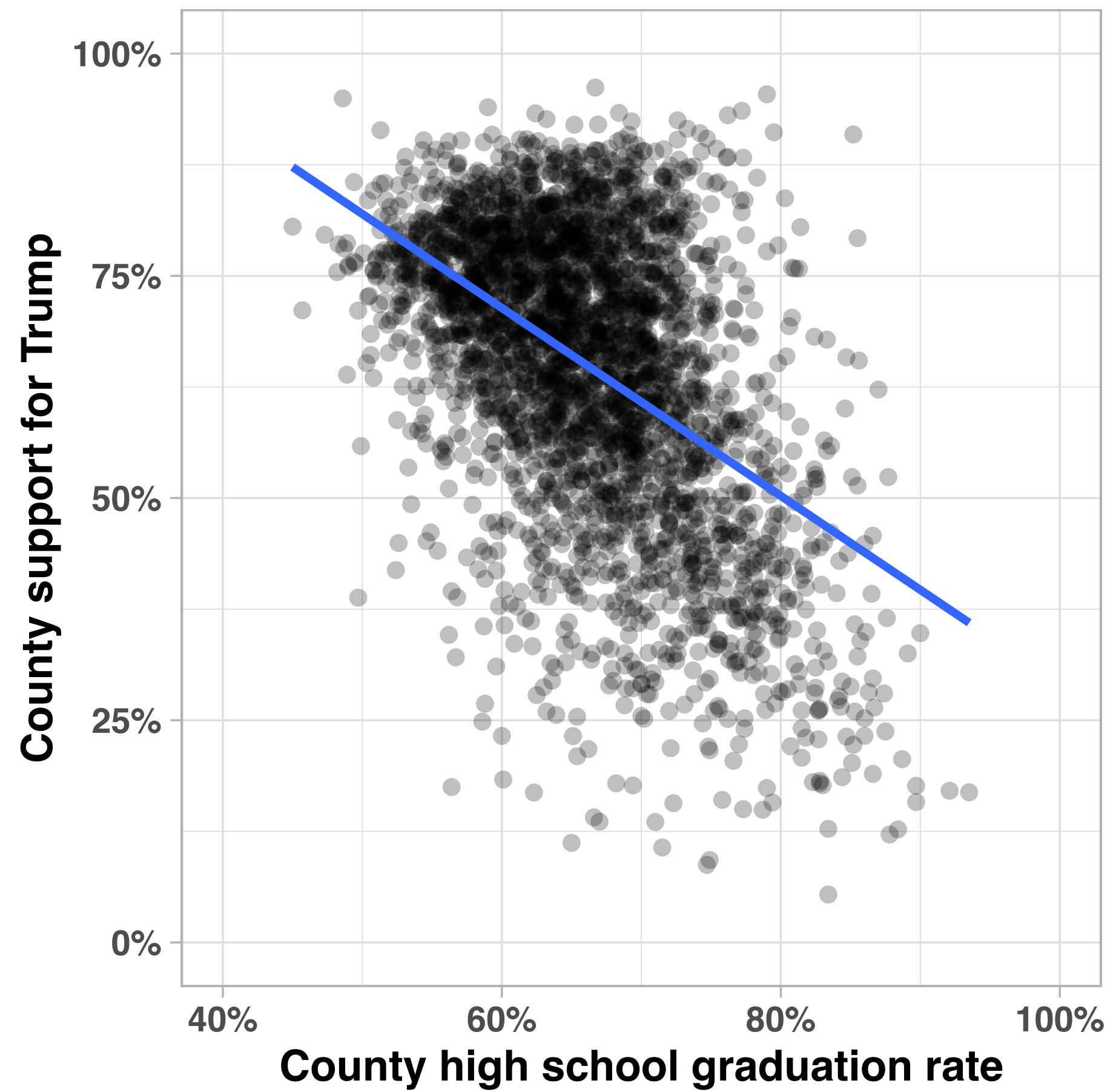
```
# Graph high school graduation and Trump support
plot_1 <- ggplot(df, aes(x=pc_hs_grad, y=pc_trump)) +  
  
  # Add scatterplot points
  geom_point(alpha=0.25) +  
  
  # Labels of axes
  xlab("County high school graduation rate") +
  ylab("County support for Trump") +  
  
  # Cosmetic changes
  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(40,100),
                     labels = function(x) paste0(x,"%"))
```

Which line fits best?



Let's graph our data.

```
# Graph high school graduation and Trump support
plot_2 <- ggplot(df, aes(x=pc_hs_grad, y=pc_trump)) +  
  
  # Add scatterplot points
  geom_point(alpha=0.25) +  
  
  # Labels of axes
  xlab("County high school graduation rate") +
  ylab("County support for Trump") +  
  
  # 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(40,100),
                     labels = function(x) paste0(x,"%")) +  
  
  # Add line of best fit
  geom_smooth(method="lm", se=F, formula = y~x)
```



Alright, let's run a regression.

Save it as
object “reg_1”

```
# Estimate regression
# LHS = support for Trump
# RHS = unemployment rate
reg_1 <- lm(pc_trump ~ pc_hs_grad, data=df)
summary(reg_1)
```

Get regression output

$$(Trump)_i = \hat{\beta}_0 + \hat{\beta}_1(HS_grad)_i + \hat{u}_i$$

Alright, let's interpret a regression.

```
Call:  
lm(formula = pc_trump ~ pc_hs_grad, data = df)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-57.719 -8.173  0.833  9.423  46.200  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 134.92112   2.28637   59.01 <2e-16 ***  
pc_hs_grad   -1.05882   0.03439  -30.79 <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 14.13 on 3112 degrees of freedom  
Multiple R-squared:  0.2335,          Adjusted R-squared:  0.2332  
F-statistic: 947.9 on 1 and 3112 DF,  p-value: < 2.2e-16
```

Note: Both X and Y are measured from 0–100.

Alright, let's interpret a regression.

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-57.719	-8.173	0.833	9.423	46.200

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	134.92112	2.28637	59.01	<2e-16 ***							
pc_hs_grad	-1.05882	0.03439	-30.79	<2e-16 ***							

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

Residual standard error: 14.13 on 3112 degrees of freedom

Multiple R-squared: 0.2335, Adjusted R-squared: 0.2332

F-statistic: 947.9 on 1 and 3112 DF, p-value: < 2.2e-16

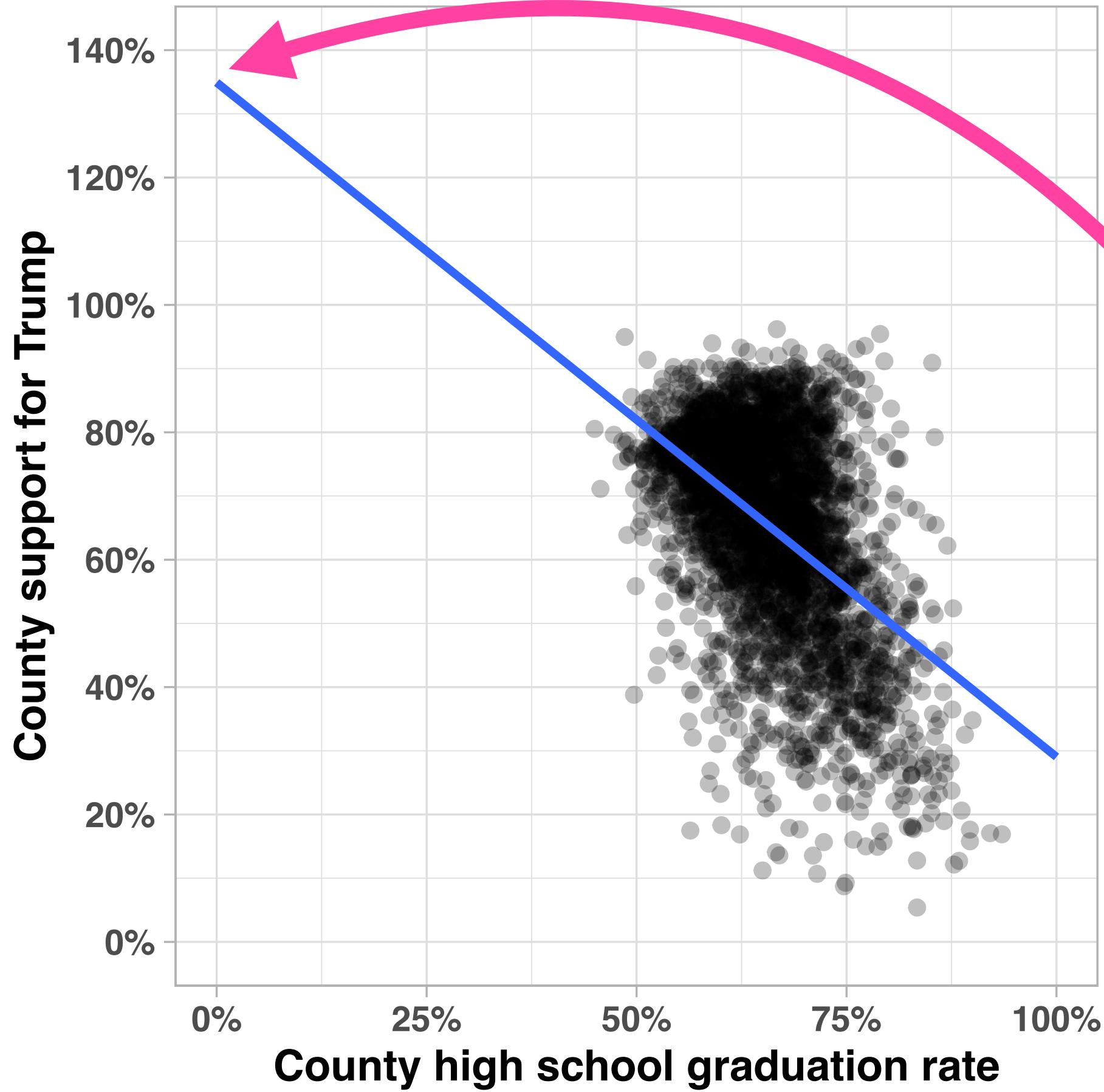
We'll start with the intercept.

When 0% of a county has graduated high school, support for Trump is a predicted 134.9%.

(Is a graduation rate of 0% meaningful?)

Note: Both X and Y are measured from 0–100.

Alright, let's interpret a regression.



We'll start with the intercept.

When 0% of a county has graduated high school, support for Trump is a predicted 134.9%.

(Is a graduation rate of 0% meaningful?)

That looks
about right!

...but here, it's not meaningful. We don't have any counties near 0% graduation rates, and we can't ever have 135% support!

Alright, let's interpret a regression.

```
Call:  
lm(formula = pc_trump ~ pc_hs_grad, data = df)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-57.719 -8.173  0.833  9.423  46.200  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 134.92112   2.28637   59.01 <2e-16 ***  
pc_hs_grad   -1.05882   0.03439  -30.79 <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 14.13 on 3112 degrees of freedom  
Multiple R-squared:  0.2335,          Adjusted R-squared:  0.2332  
F-statistic: 947.9 on 1 and 3112 DF,  p-value: < 2.2e-16
```

Note: Both X and Y are measured from 0–100.

We'll start with the intercept.

When 0% of a county has graduated high school, support for Trump is a predicted 134.9%.

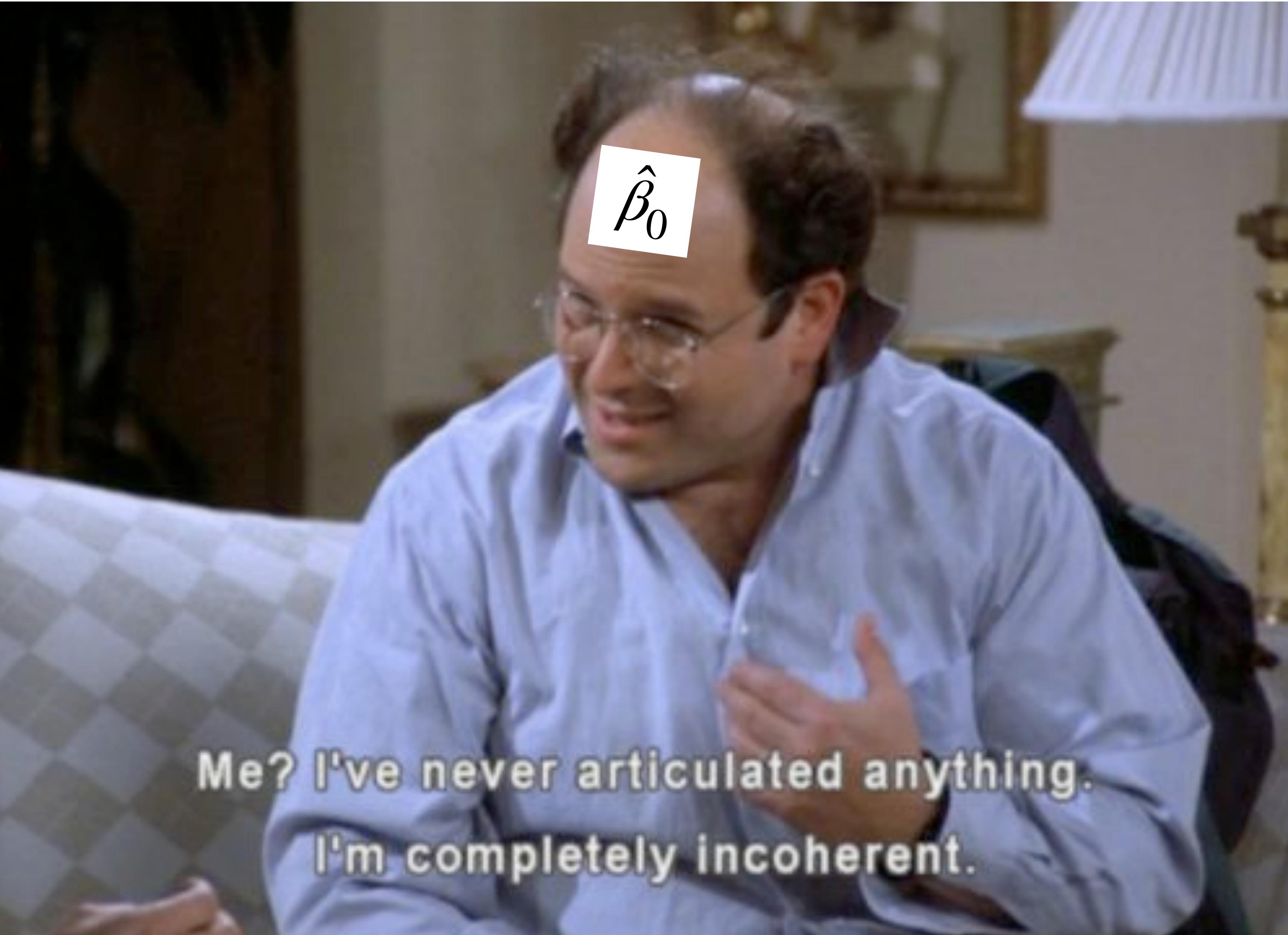
(Is a graduation rate of 0% meaningful?)

The standard error is 2.3 pp. This gives us a 95% C.I. of $134.9 \pm 1.96 \cdot 2.3 = [130.4\% \text{ to } 139.4\%]$.

The t-statistic is 59.0. The p-value is <0.05.

Thus, we can conclude that the intercept is significantly different from 0.

$$\hat{\beta}_0$$



Me? I've never articulated anything.
I'm completely incoherent.

...unless $X=0$ is meaningful!

Alright, let's interpret a regression.

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

Residuals:

Min	1Q	Median	3Q	Max
-57.719	-8.173	0.833	9.423	46.200

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	134.92112	2.28637	59.01	<2e-16 ***
pc hs grad	-1.05882	0.03439	-30.79	<2e-16 ***

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

Residual standard error: 14.13 on 3112 degrees of freedom

Multiple R-squared: 0.2335, Adjusted R-squared: 0.2332

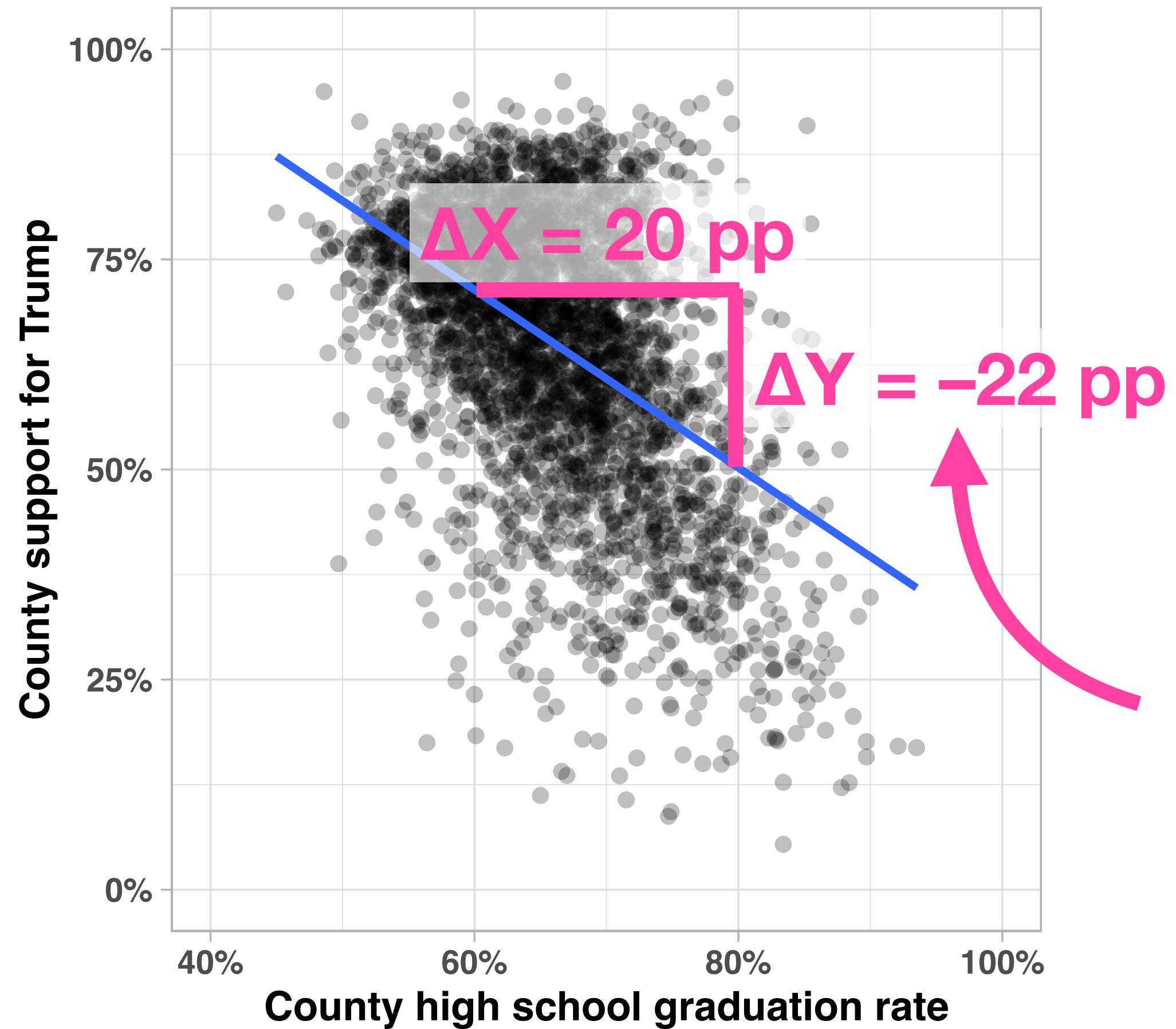
F-statistic: 947.9 on 1 and 3112 DF, p-value: < 2.2e-16

Now for the coefficient on pc_hs_grad.

For each 1 percentage point (pp) increase in a county's high school graduation rate, there is an associated 1.1 pp decrease in support for Trump, on average.

Note: Both X and Y are measured from 0–100.

Alright, let's interpret a regression.



Now for the coefficient on `pc_hs_grad`.

For each 1 percentage point (pp) increase in a county's high school graduation rate, there is an associated 1.1 pp decrease in support for Trump, on average.

That looks
about right!

Alright, let's interpret a regression.

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

Residuals:

Min	1Q	Median	3Q	Max
-57.719	-8.173	0.833	9.423	46.200

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	134.92112	2.28637	59.01	<2e-16 ***
pc hs grad	-1.05882	0.03439	-30.79	<2e-16 ***

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

Residual standard error: 14.13 on 3112 degrees of freedom

Multiple R-squared: 0.2335, Adjusted R-squared: 0.2332

F-statistic: 947.9 on 1 and 3112 DF, p-value: < 2.2e-16

Now for the coefficient on pc_hs_grad.

For each 1 percentage point (pp) increase in a county's high school graduation rate, there is an associated 1.1 pp decrease in support for Trump, on average.

The standard error is 0.03 pp. This gives us a 95% C.I. of $-1.06 \pm 1.96 \cdot 0.03 = [-1.13 \text{ pp to } -0.99 \text{ pp}]$.

The t-statistic is -30.8. The p-value is <0.05.

Thus, the coefficient is significantly different from 0. There's a negative association between graduation rates and Trump support.

Note: Both X and Y are measured from 0–100.

That's all good. But is it causal?

We'll spend lots of time in API 202
asking this very question.

What problems with a causal
interpretation come to mind?

What influences are we missing?

