# 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 30, 2026

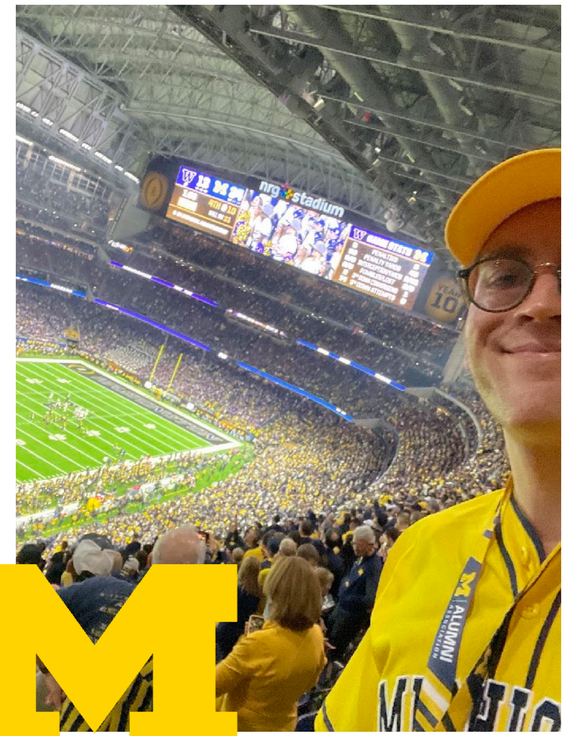
# Goals for today

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.

*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
ANIL MENON *University of M*
JUSTIN E. HEINZE *Universi*

*Why do voters in developed a focuses on economic and c that perceptions of personal this argument, we analyzed all waves that voters with worse self-reported h parties. The relationship persists even ability, as well as voters' satisfaction w influence of health on support for right self-reported economic insecurity, wh findings suggest that policies affecting political landscape.*

**INTRODUCTION**

Right-wing populist parties are sur port across the Western world Inglehart 2019). Why do voters i democracies support such parties? A grow research has identified economic insecur tural backlash as potential drivers of rec successes (Algan et al. 2017; Hochschild 20 and Norris 2016; Madde 2007; Oesch 2 2018; Smith and Hanley 2018). Accordi explanations, once-dominant socioecono perceive an erosion of their economic o or a threat to their privileged positions These threats heighten voters' perceived ity, motivating them to support parties t to restore their socioeconomic standing th multiculturalism, antiglobalism, and anti- (Inglehart and Norris 2016).

We argue that a voter's perceived health ingfully contribute to populist support v mechanism. The development of illness a often produces frustration with one's p

Nolan M. Kavanagh, Medical student, Perelman S cine, University of Pennsylvania; Lecturer, Departme ties and Oral Medicine, School of Dentistry, Universi nolan.kavanagh@perelmanschool.upenn.edu.
Anil Menon, PhD candidate, Department of Po University of Michigan, anmenon@umich.edu.
Justin E. Heinze, Assistant Professor, Departm Behavior and Health Education, School of Public Hea of Michigan, jheinze@umich.edu.

Received: June 26, 2020; revised: February 27, 2 March 23, 2021. First published online: April 26, 202

1104




My research is on the politics of health.

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


GO BLUE!

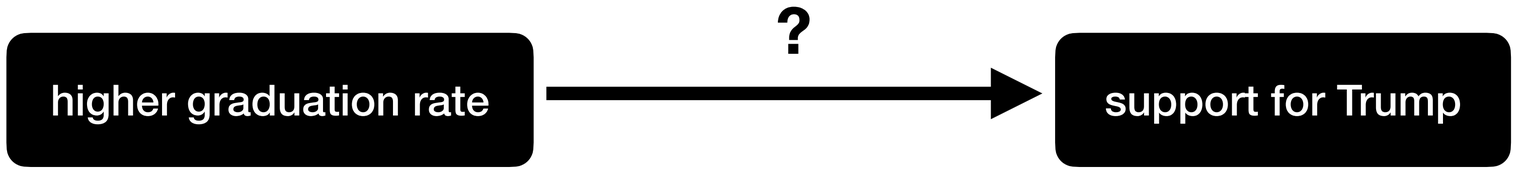
# Overview of our sample data

## 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)* |
| `median_income` | **County median income ($)** | *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* |

# Tell me a story.

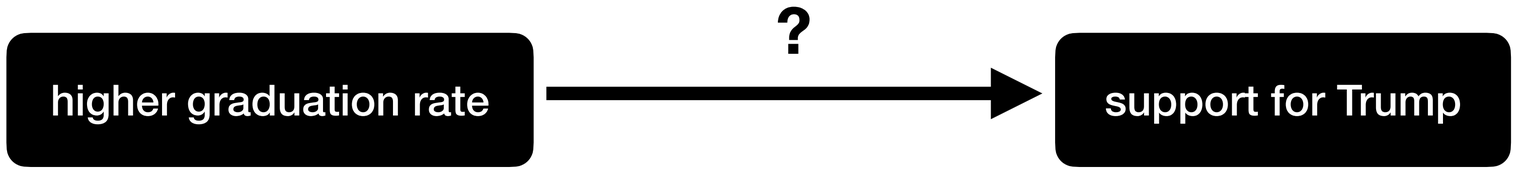
**Let's say we're interested in the relationship between high school graduation and support for Trump.**




**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.**

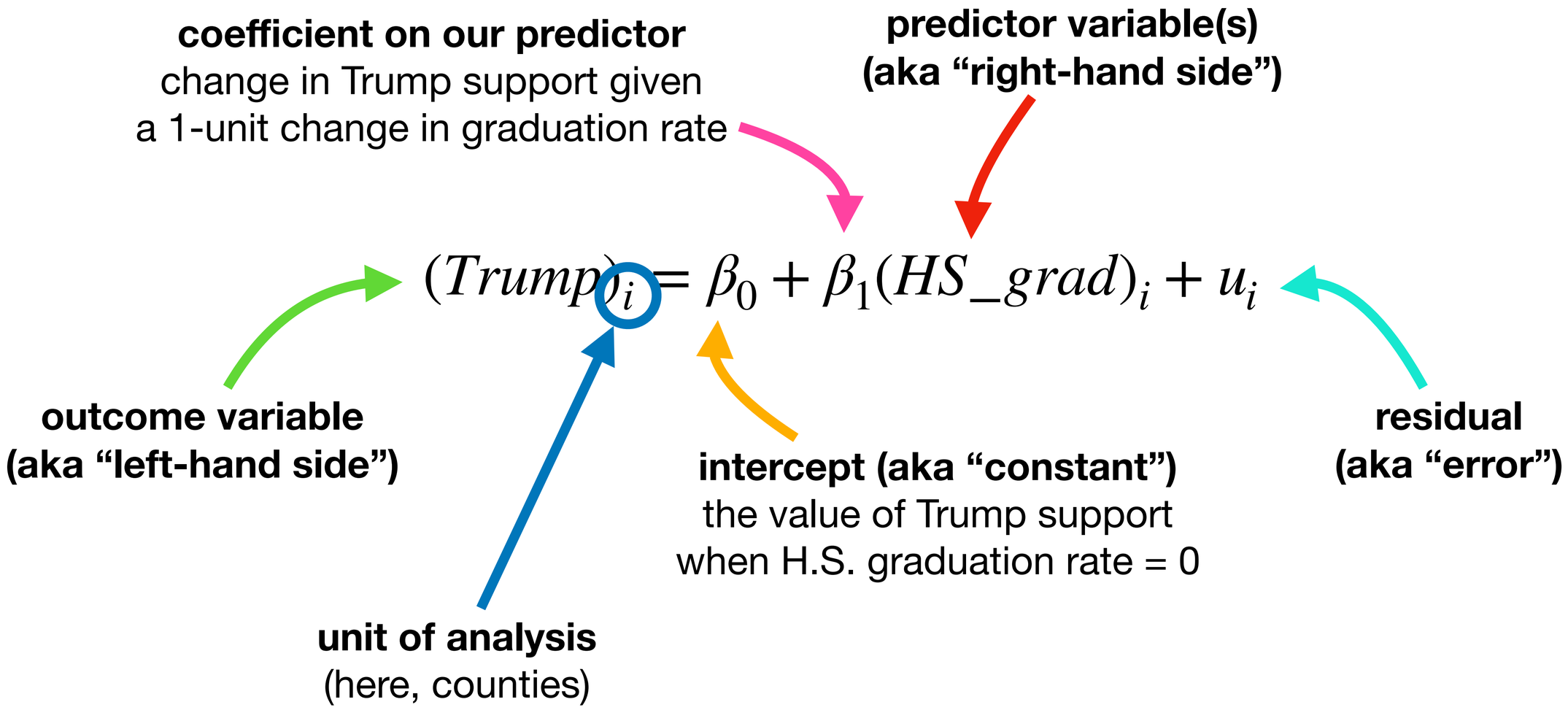



**What might be the direction? Mechanism?**

More education = liberal values = prefer multiculturalism?

More education = more income = prefer lower taxes?

# Population regression function



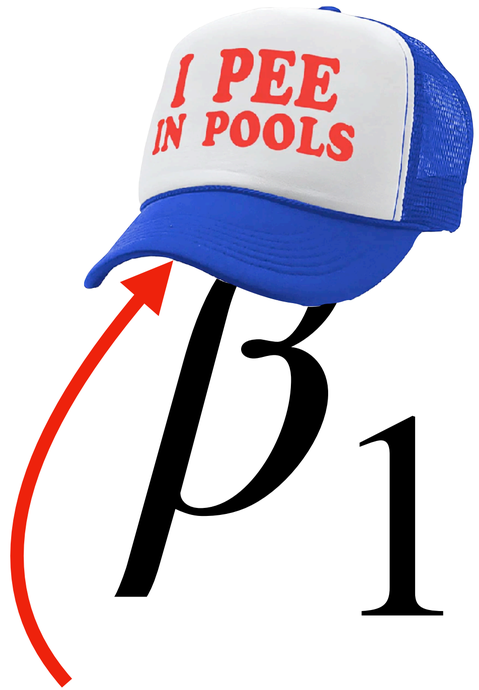
## Population regression function

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

**We add "hats" to signify estimated values in our sample.**

## Sample regression function

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

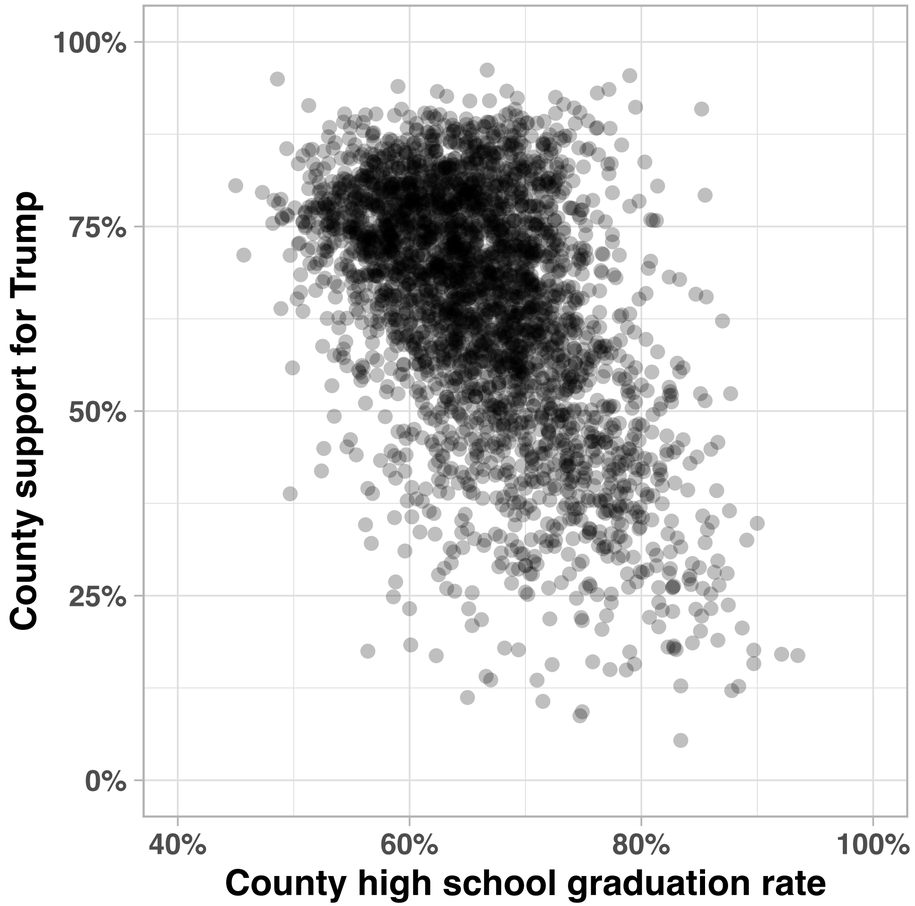



**Only a <u>specific</u> sample would ever wear this hat.**


- That's the hat I gave her, she's wearing it as a
She's a menace to society.
regression estimate

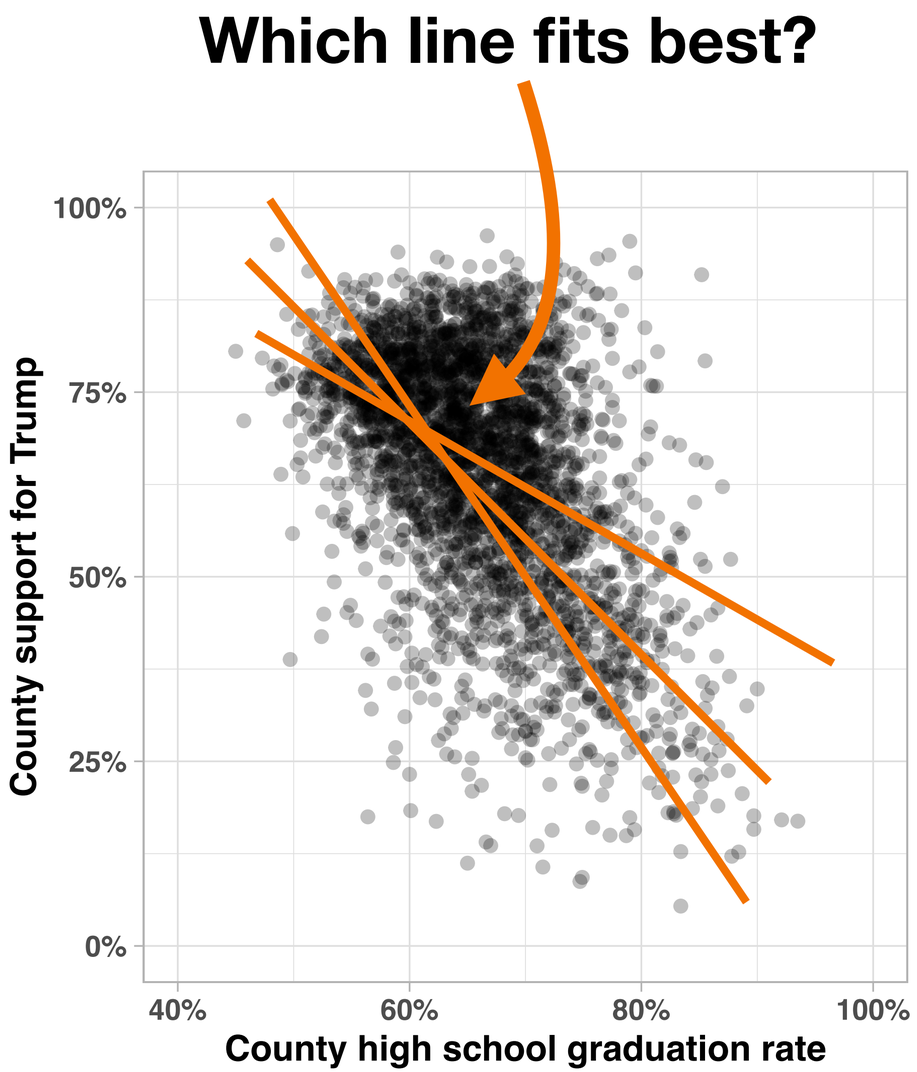
# Let's graph our data.

```
# 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,"%"))
```



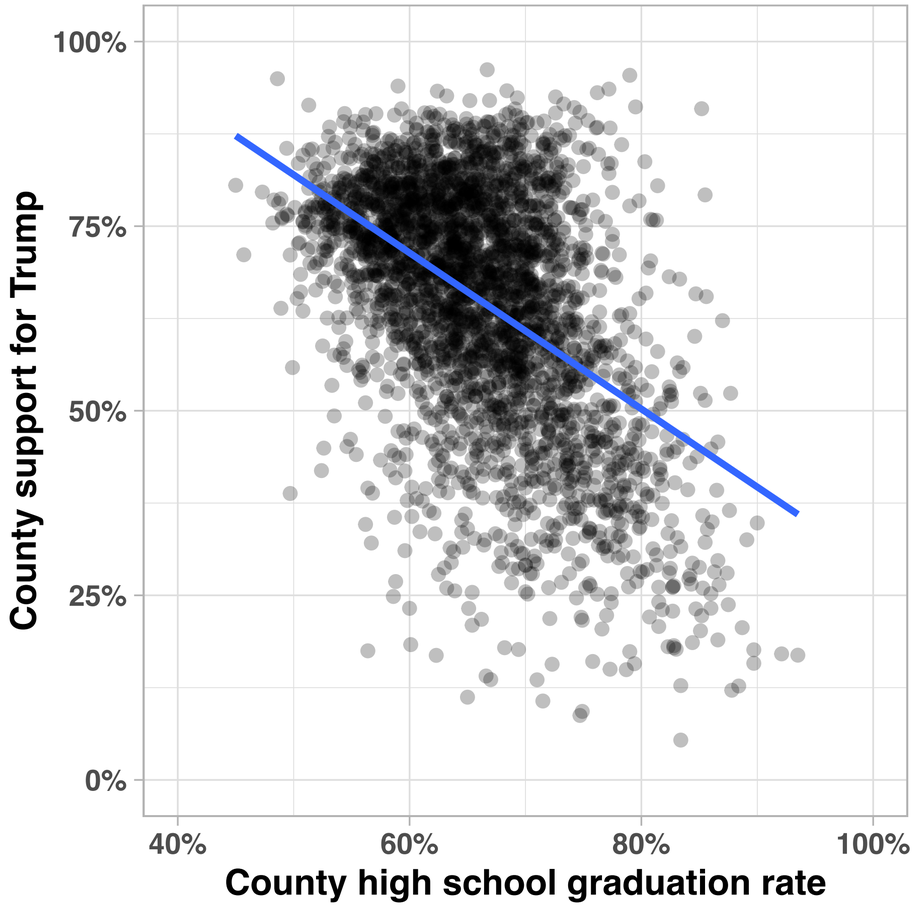
# Let's graph our data.

```
# 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,"%"))
```



# 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)
```




Maybe this will become, like,
a cool thing,
making sick graphs
tbs
Then maybe baldness will catch on.
tbs

# Alright, let's run a regression.

Save it as object "reg_1"

```
# Estimate regression
# LHS = support for Trump
# RHS = high school graduation 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
```

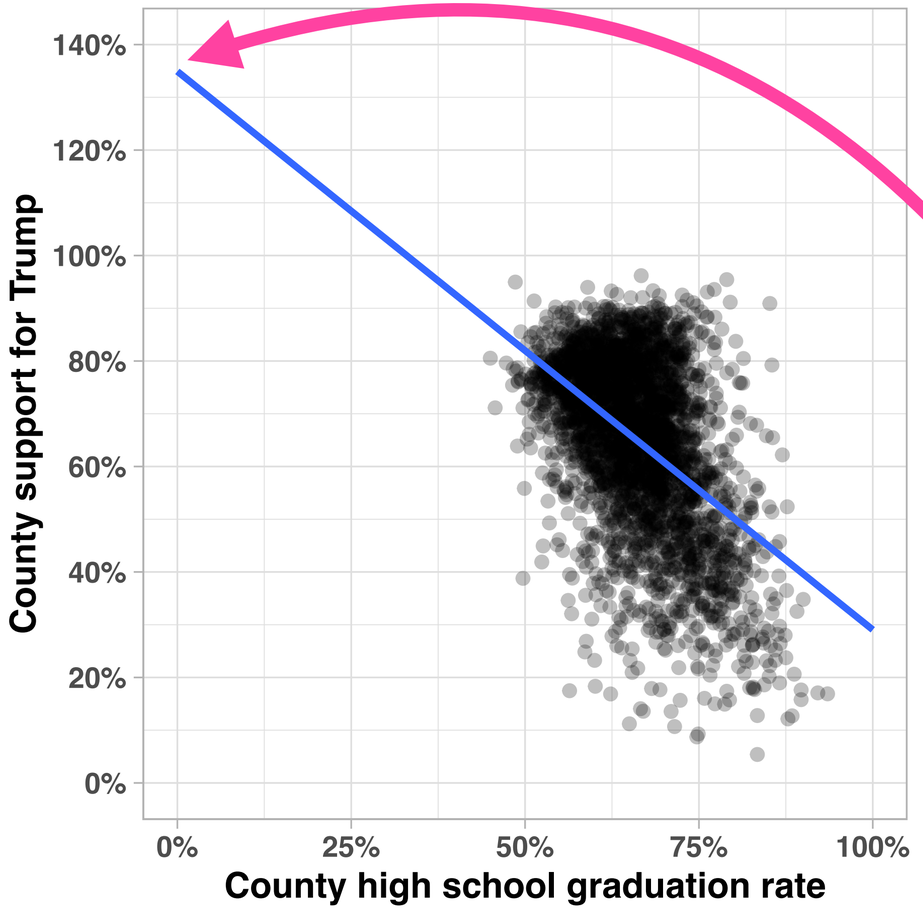
Note: Both X and Y are measured from 0–100.

**We'll start with the intercept.**

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

*(Is a graduation rate of 0% meaningful?)*

# Alright, let's interpret a regression.




**We'll start with the intercept.**

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

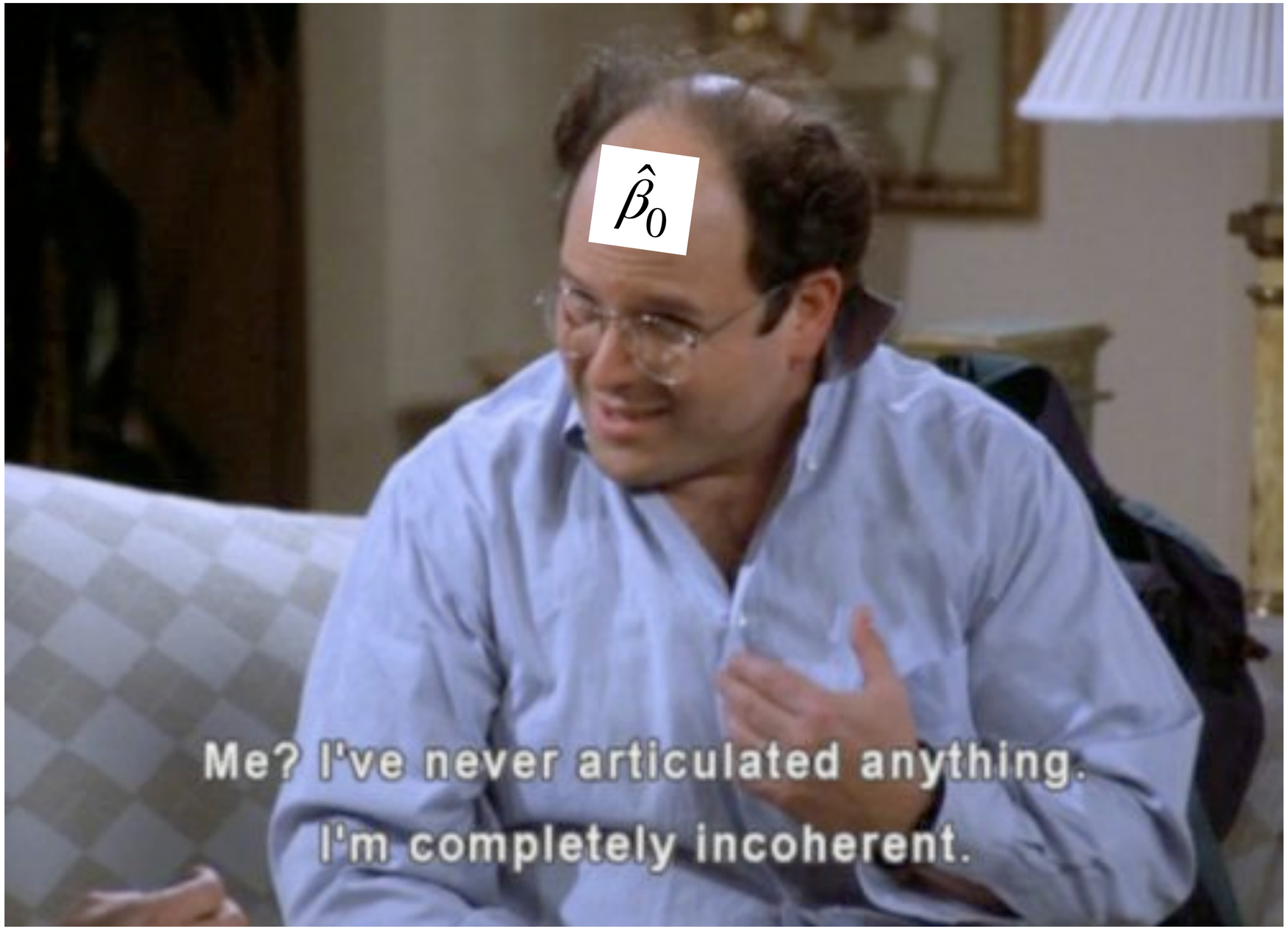
When 0% of a county has graduated high school, the predicted support for Trump is 134.9%.

*(Is a graduation rate of 0% meaningful?)*

The standard error is 2.3 pp. This gives us a 95% C.I. of 134.9 ± 1.96*2.3 = [130.4% 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.**




…unless X=0 is meaningful!

# Alright, let's interpret a regression.

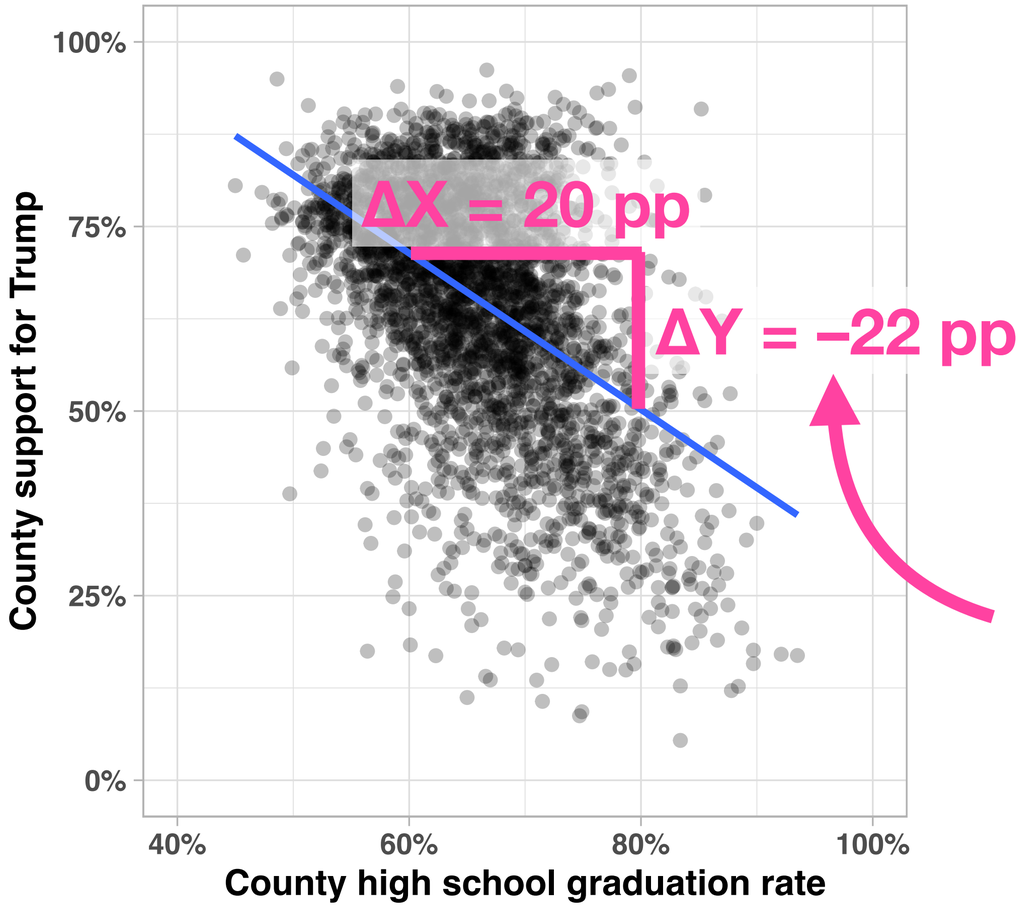
```
Call:
lm(formula = pc_trump ~ pc_hs_grad, data = df)

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
```

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

**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.

# Alright, let's interpret a regression.




**That looks about right!**

**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.

# 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.

**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 ± 1.96*0.03 = [–1.13 pp to –0.99 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.**


$\hat{\beta}_1$
You're strange and beautiful
and sensitive.

# 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?**