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

Dummy variables and interactions

API 202: TF Session 3

R

Nolan M. Kavanagh
February 9, 2024

Goals for today

1. **Review core concepts in regression analysis.**
2. **Review the principles of interactions in regression.**
3. **Learn how to run interacted regressions.**
4. **Practice interpreting interaction terms.**

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

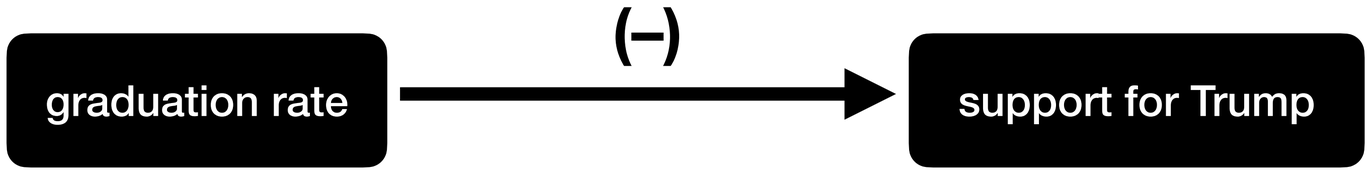
Overview of our sample data

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*

Let’s revisit the Trump story.

We’ve learned that high school graduation rates were an important predictor of Trump’s support in 2020.



In Season 4 Episode 19, we learn that Homer never technically graduated high school (even though he does later in the episode).

This screenshot is from later in the series, but would he be more or less likely to support Trump than someone who graduated high school?

We've seen this graph before.

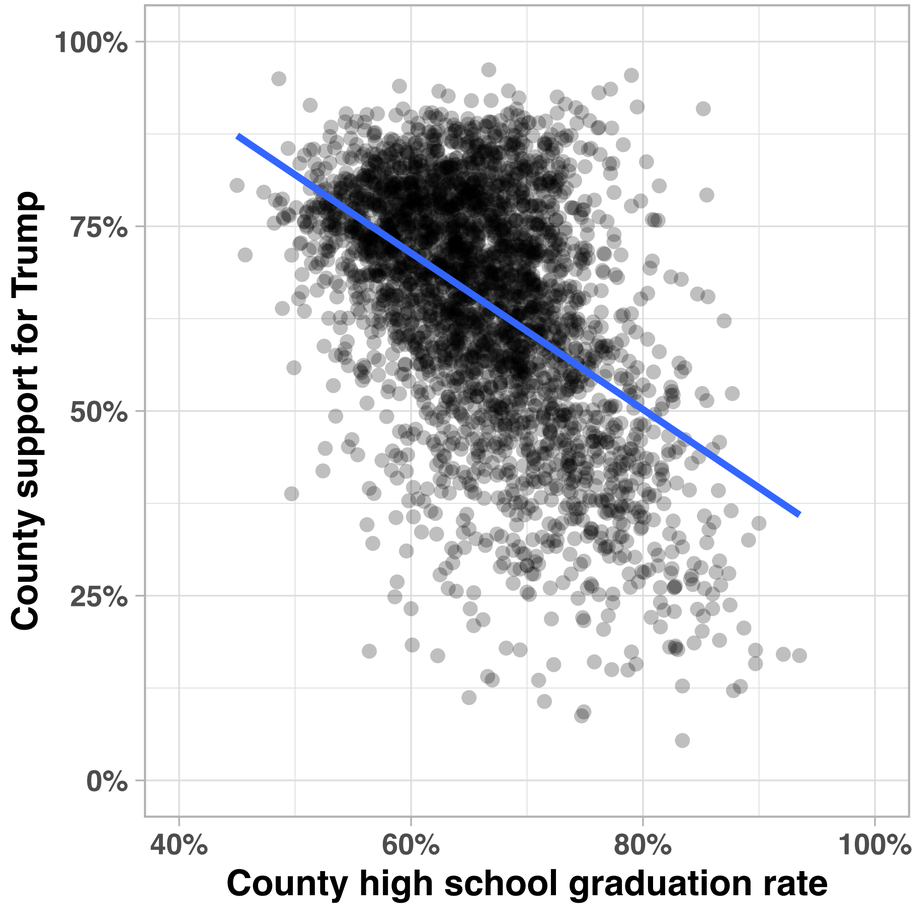
```
# Graph graduation rate and Trump support
plot_1 <- ggplot(df, aes(x=pc_hs_grad, y=pc_trump)) +

  # Add scatterplot points
  geom_point(alpha=0.25) +

  # Labels of axes
  xlab("County high school graduation rate") +
  ylab("County support for Trump") +

  # 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(40,100),
                     labels = function(x) paste0(x,"%"))
```



We've seen this regression before.

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

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've seen this regression before.

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

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
```

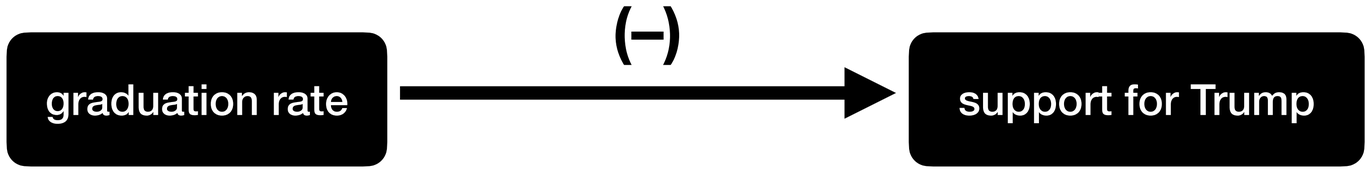
For each 1 percentage point (pp) increase in a county's high school graduation rate, the estimated support for Trump decreases by 1.1 pp.

The association is statistically significant.

Note: These variables are scaled 0–100, not 0–1.

Let’s revisit the Trump story.

We’ve learned that high school graduation rates were an important predictor of Trump’s support in 2020.



But is this true for every community?

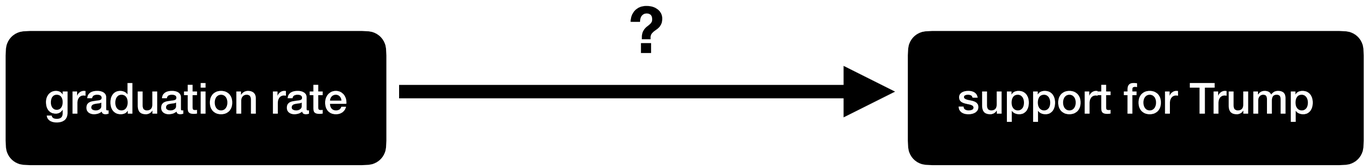
In Season 4 Episode 19, we learn that Homer never technically graduated high school (even though he does later in the episode).

This screenshot is from later in the series, but would he be more or less likely to support Trump than someone who graduated high school?

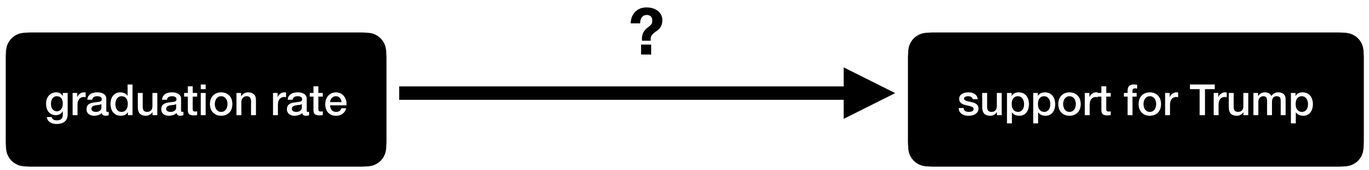
Let’s consider minoritized communities.

Is the association different for majority-Black or Latin counties, compared to majority-white counties?

Majority-Black or Latin counties



Other counties



OK, let's make a dummy variable.

It should equal "1" for counties that are majority-Black or Latin.

Meanwhile, it should equal "0" for all other counties.

```
# Generate dummy variable
# Majority-Black or Latin counties
df <- df %>% mutate(
  majority = case_when(
    pc_black >= 50 | pc_latin >= 50 ~ 1,
    TRUE ~ 0
  ))
```

If a county meets either criterion, assign it a value of 1.

For all other counties, put 0.

Let’s consider our regression function.

Let's consider our regression function.

interaction between our two predictors

$$(Trump)_i = \beta_0 + \beta_1(HS\_grad)_i + \beta_2(majority)_i + \beta_3(HS\_grad * majority)_i + u_i$$

high school graduation rate
measured in percent (0–100)

dummy for majority-Black/Latin
1 = majority-Black/Latin county
0 = all other counties

Let’s consider our regression function.

Other counties

$$(Trump)_i = \beta_0 + \beta_1(HS\_grad)_i + \beta_2(majority)_i + \beta_3(HS\_grad * majority)_i + u_i$$

here, majority = 0

Majority-Black or Latin counties

here, majority = 1

$$(Trump)_i = \beta_0 + \beta_1(HS\_grad)_i + \beta_2(majority)_i + \beta_3(HS\_grad * majority)_i + u_i$$

Let's consider our regression function.

Other counties

here, majority = 0

$$(Trump)_i = \beta_0 + \beta_1(HS\_grad)_i + \beta_2(majority)_i + \beta_3(HS\_grad * majority)_i + u_i$$

these terms go to 0

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

Majority-Black or Latin counties

here, majority = 1

these terms are just 1

$$(Trump)_i = \beta_0 + \beta_1(HS\_grad)_i + \beta_2(majority) + \beta_3(HS\_grad * majority)_i + u_i$$

$$(Trump)_i = \beta_0 + \beta_1(HS\_grad)_i + \beta_2 + \beta_3(HS\_grad)_i + u_i$$

$$(Trump)_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)(HS\_grad)_i + u_i$$

rearrange our terms

Let’s consider our regression function.

Other counties

here, majority = 0

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

Majority-Black or Latin counties

here, majority = 1

$$(Trump)_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)(HS\_grad)_i + u_i$$

Let's consider our regression function.

Other counties

here, majority = 0

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

β_2 gives us the difference in intercepts
i.e. the difference in expected Trump support for majority-Black/Latin vs. other counties with graduation rates of 0%.

Majority-Black or Latin counties

here, majority = 1

$$(Trump)_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)(HS\_grad)_i + u_i$$

Let's consider our regression function.

Other counties

here, majority = 0

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

β_2 gives us the difference in intercepts
i.e. the difference in expected Trump support for majority-Black/Latin vs. other counties with graduation rates of 0%.

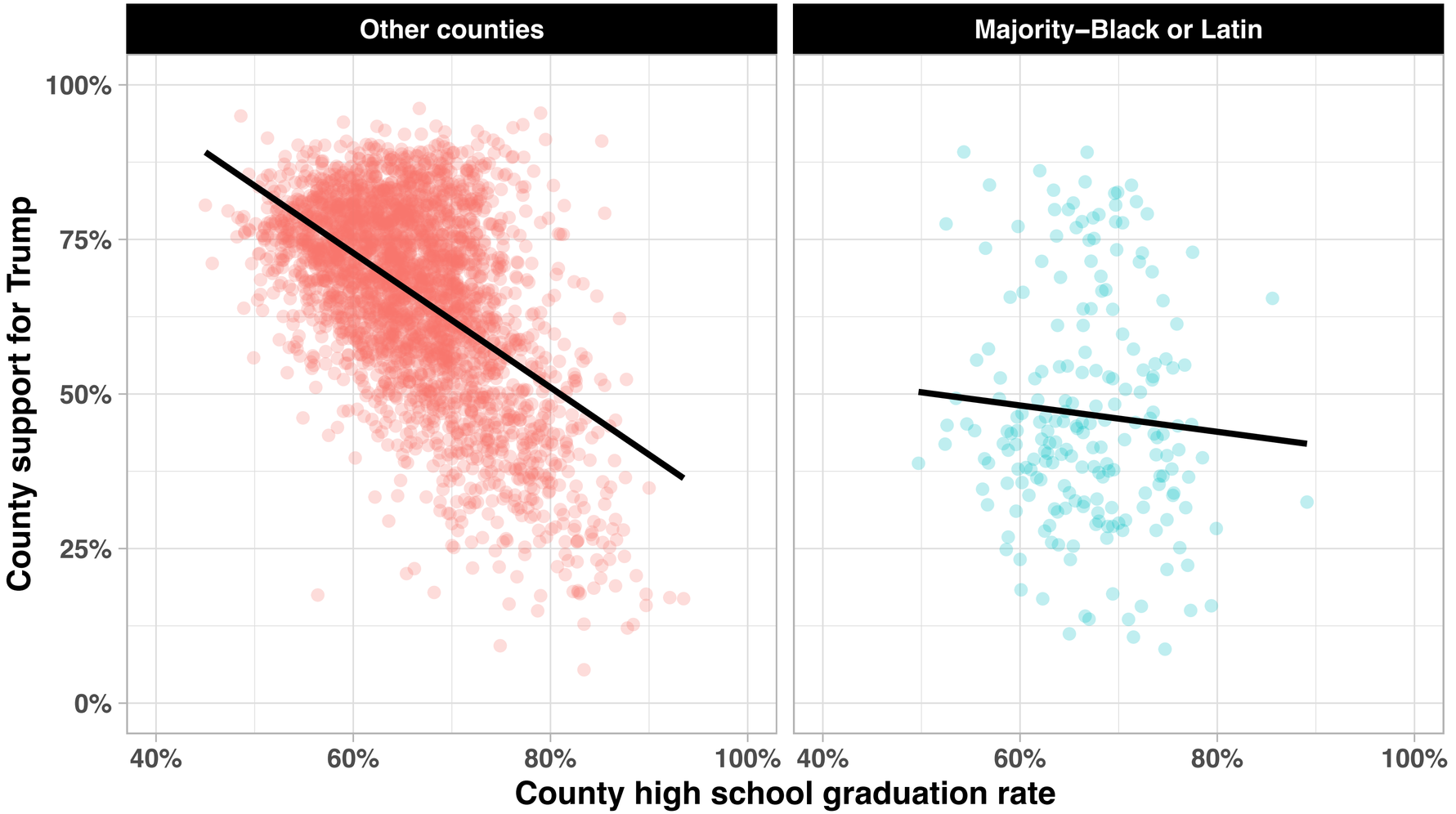
β_3 gives us the difference in slopes
i.e. the difference in the association between graduation rates and Trump support for majority-Black/Latin vs. other counties.

Majority-Black or Latin counties

here, majority = 1

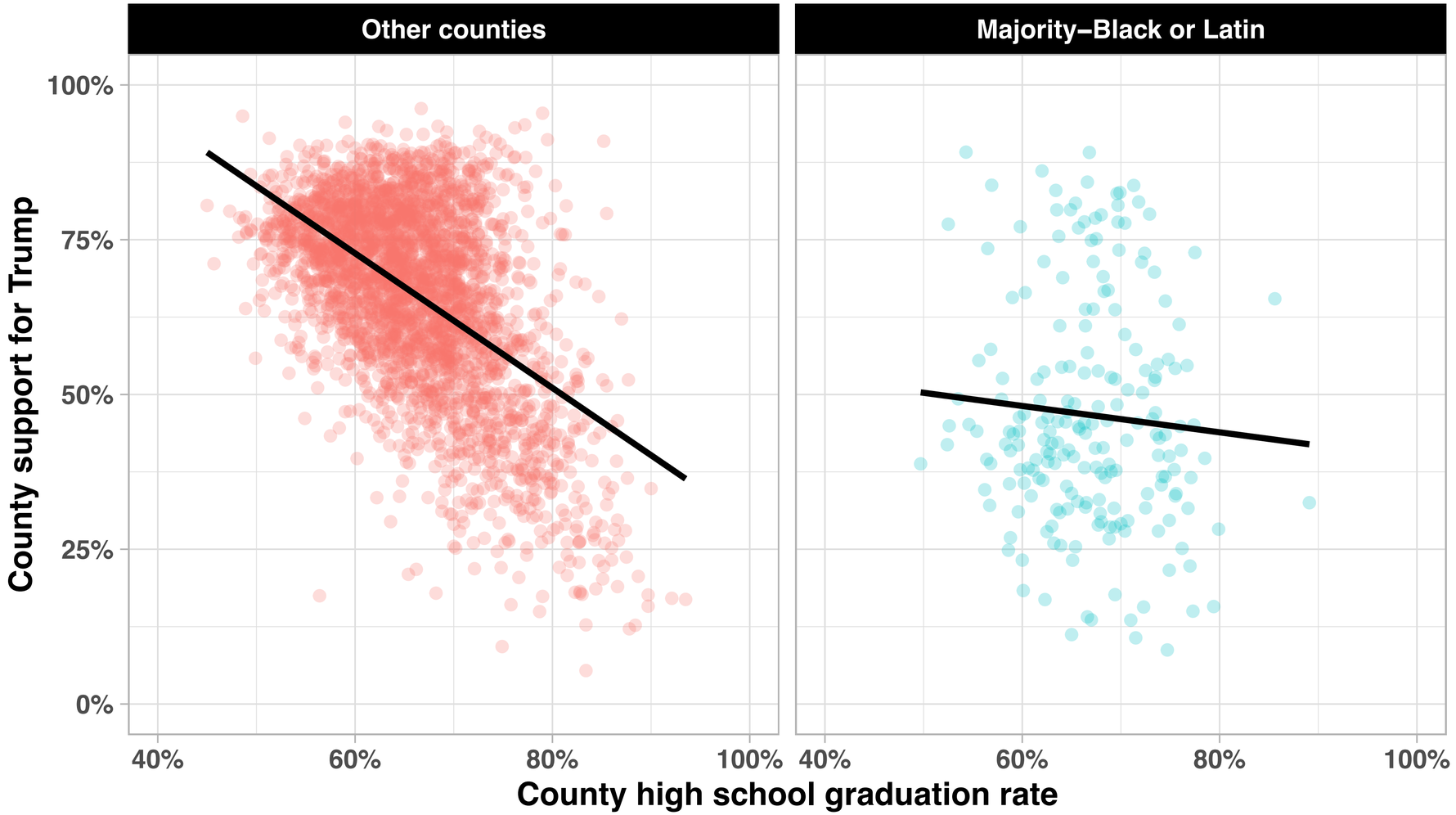
$$(Trump)_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)(HS\_grad)_i + u_i$$

Show me the graph already!



See the online script file for the code to make these graphs.

Show me the graph already!



Clearly, these are different relationships!

See the online script file for the code to make these graphs.

Show me the regression already!

```
# Estimate interacted regression
reg_2 <- lm(pc_trump ~ pc_hs_grad + majority + pc_hs_grad*majority, data=df)
summary(reg_2)

Call:
lm(formula = pc_trump ~ pc_hs_grad + majority + pc_hs_grad *
    majority, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-59.233  -8.557  -0.072   8.636  45.507

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          138.04241    2.20351  62.647  < 2e-16 ***
pc_hs_grad            -1.08732    0.03317 -32.785  < 2e-16 ***
majority             -77.12618    9.81818  -7.855 5.44e-15 ***
pc_hs_grad:majority    0.87450    0.14638   5.974 2.58e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.27 on 3110 degrees of freedom
Multiple R-squared:  0.3246,           Adjusted R-squared:  0.3239
F-statistic: 498.2 on 3 and 3110 DF,  p-value: < 2.2e-16
```

Show me the regression already!

```
# Estimate interacted regression
reg_2 <- lm(pc_trump ~ pc_hs_grad + majority + pc_hs_grad*majority, data=df)
summary(reg_2)

Call:
lm(formula = pc_trump ~ pc_hs_grad + majority + pc_hs_grad *
    majority, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-59.233  -8.557  -0.072   8.636  45.507

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        138.04241    2.20351  62.647  < 2e-16 ***
pc_hs_grad          -1.08732    0.03317 -32.785  < 2e-16 ***
majority           -77.12618    9.81818  -7.855 5.44e-15 ***
pc_hs_grad:majority  0.87450    0.14638   5.974 2.58e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.27 on 3110 degrees of freedom
Multiple R-squared:  0.3246,            Adjusted R-squared:  0.3239
F-statistic: 498.2 on 3 and 3110 DF,  p-value: < 2.2e-16
```

The expected support for Trump in an "other" county with a high school graduation rate of 0% is 138%.

It's significantly different from 0.

It's also not especially meaningful.

Show me the regression already!

```
# Estimate interacted regression
reg_2 <- lm(pc_trump ~ pc_hs_grad + majority + pc_hs_grad*majority, data=df)
summary(reg_2)

Call:
lm(formula = pc_trump ~ pc_hs_grad + majority + pc_hs_grad *
    majority, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-59.233  -8.557  -0.072   8.636  45.507

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          138.04241    2.20351  62.647  < 2e-16 ***
pc_hs_grad            -1.08732    0.03317 -32.785  < 2e-16 ***
majority             -77.12618    9.81818  -7.855 5.44e-15 ***
pc_hs_grad:majority    0.87450    0.14638   5.974 2.58e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.27 on 3110 degrees of freedom
Multiple R-squared:  0.3246,            Adjusted R-squared:  0.3239
F-statistic: 498.2 on 3 and 3110 DF,  p-value: < 2.2e-16
```

For "other" counties, a 1 pp increase in the graduation rate is associated with 1.1 pp lower support for Trump.

It's significantly different from 0.

Show me the regression already!

```
# Estimate interacted regression
reg_2 <- lm(pc_trump ~ pc_hs_grad + majority + pc_hs_grad*majority, data=df)
summary(reg_2)

Call:
lm(formula = pc_trump ~ pc_hs_grad + majority + pc_hs_grad *
    majority, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-59.233  -8.557  -0.072   8.636  45.507

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          138.04241    2.20351  62.647  < 2e-16 ***
pc_hs_grad            -1.08732    0.03317 -32.785  < 2e-16 ***
majority             -77.12618    9.81818  -7.855 5.44e-15 ***
pc_hs_grad:majority    0.87450    0.14638   5.974 2.58e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.27 on 3110 degrees of freedom
Multiple R-squared:  0.3246,            Adjusted R-squared:  0.3239
F-statistic: 498.2 on 3 and 3110 DF,  p-value: < 2.2e-16
```

A majority-Black or Latin county with 0% graduation has, on average, 77 pp less support for Trump than an "other" county with 0% graduation.

It's significantly different from 0.

Doing the math: 138 – 77 = 61% support.

Show me the regression already!

```
# Estimate interacted regression
reg_2 <- lm(pc_trump ~ pc_hs_grad + majority + pc_hs_grad*majority, data=df)
summary(reg_2)

Call:
lm(formula = pc_trump ~ pc_hs_grad + majority + pc_hs_grad *
    majority, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-59.233  -8.557  -0.072   8.636  45.507

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          138.04241    2.20351  62.647  < 2e-16 ***
pc_hs_grad            -1.08732    0.03317 -32.785  < 2e-16 ***
majority             -77.12618    9.81818  -7.855 5.44e-15 ***
pc_hs_grad:majority    0.87450    0.14638   5.974 2.58e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.27 on 3110 degrees of freedom
Multiple R-squared:  0.3246,            Adjusted R-squared:  0.3239
F-statistic: 498.2 on 3 and 3110 DF,  p-value: < 2.2e-16
```

The association between graduation rates and Trump support is 0.87 pp more positive for majority-Black or Latin counties, compared to "other counties."

It's significantly different from 0.

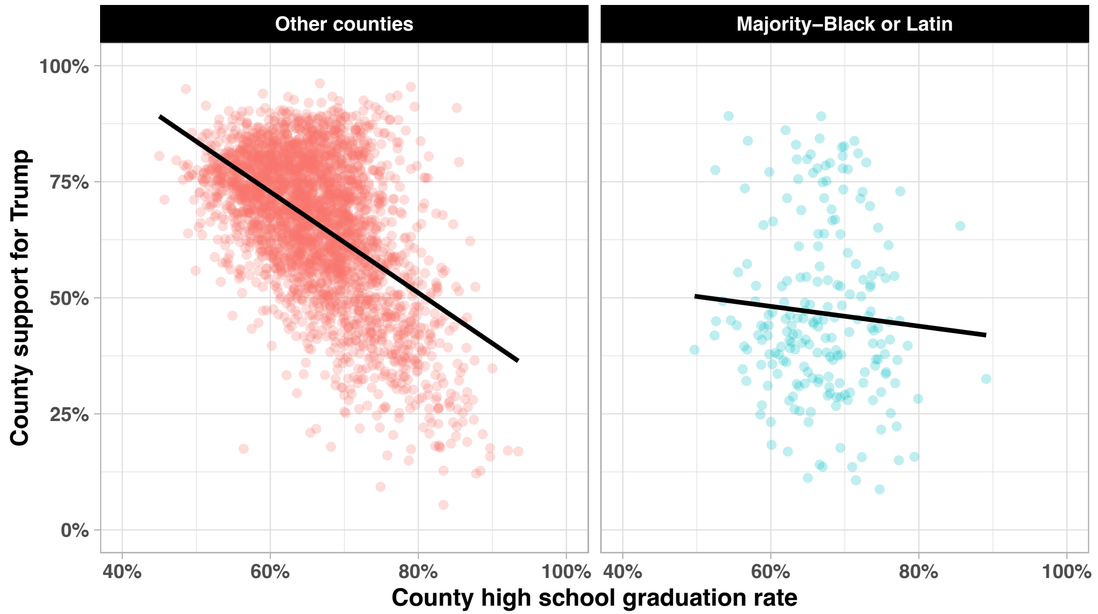
Doing the math: –1.09 + 0.87 = –0.22, which is a much flatter slope.

Putting it all together

	Model 1	Model 2
Intercept	**134.92*** **	**138.04*** **
	(2.29)	(2.20)
County graduation rate	**-1.06*** **	**-1.09*** **
	(0.03)	(0.03)
Majority-Black or Latin county		**-77.13*** **
		(9.82)
Grad. rate * Majority-Black or Latin		**0.87*** **
		(0.15)
Num.Obs.	3114	3114
R2	0.233	0.325
R2 Adj.	0.233	0.325

*p<0.05, **p<0.01, ***p<0.001

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



$$(Trump)_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)(HS\_grad)_i + u_i$$

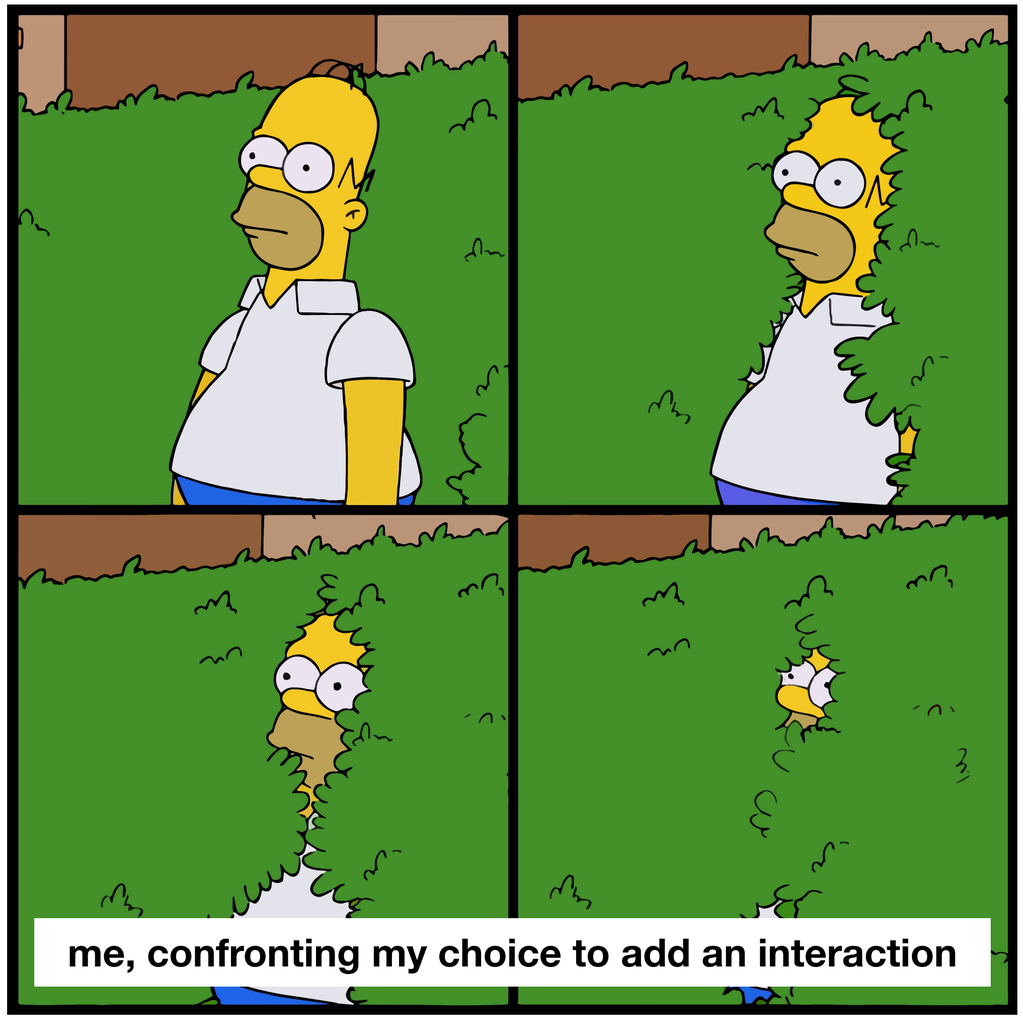
$\beta_0 + \beta_2 = 138 - 77 = 61$

$\beta_1 + \beta_3 = -1.09 + 0.87 = -0.22$

What did we learn?

Interactions are tough to interpret.

We don’t always need them.
But they can add richness to our models.



However, watch out for statistical power. We need much larger samples to estimate interactions than main effects.