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

EXCEL

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.

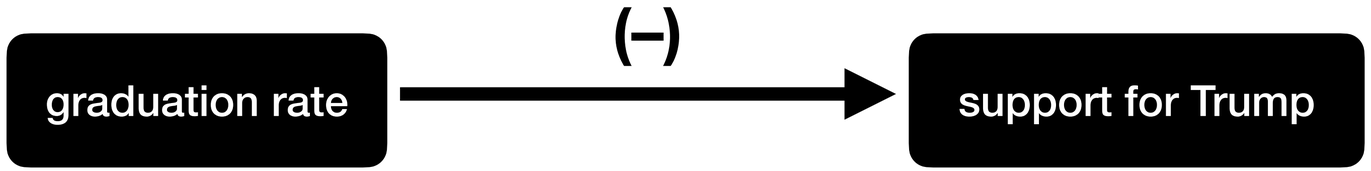
Overview of our sample data

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.

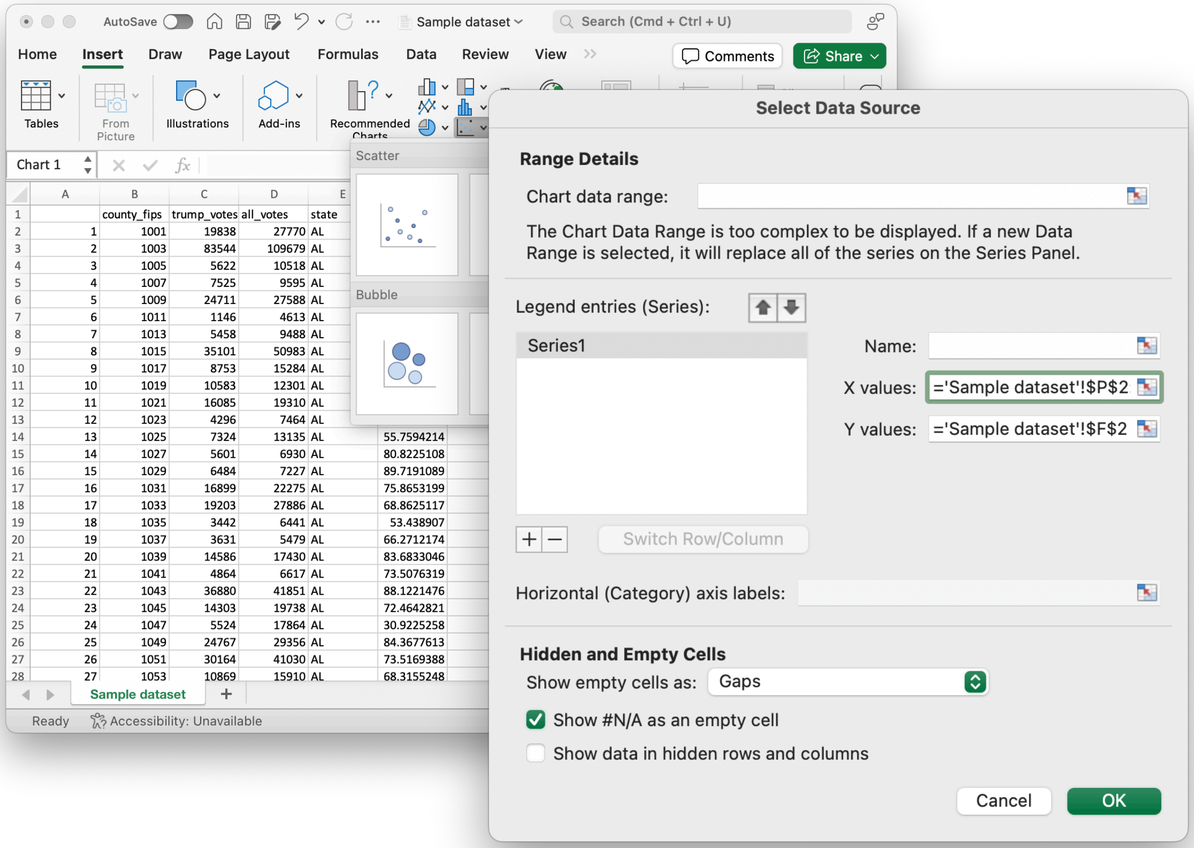
We’ve learned that high school graduation rates were an important predictor of Trump’s support in 2020.

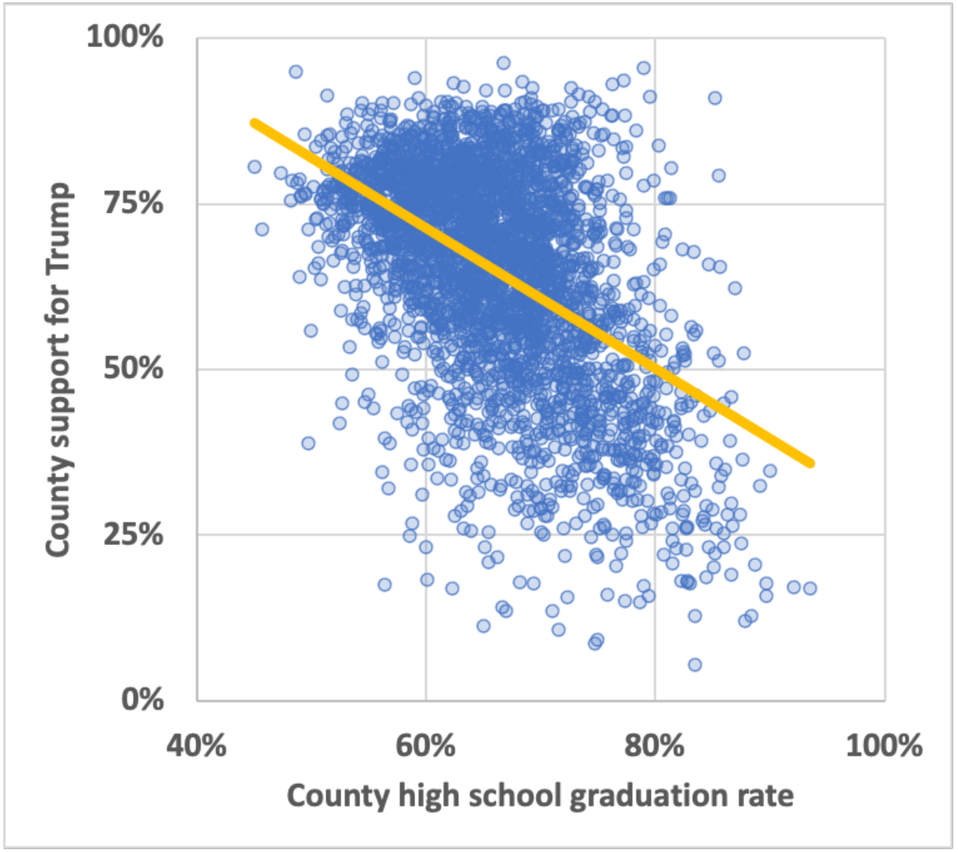


In Season 4 Episode 19, we learn that Homer never technically graduated high school (even though he does later in the episode).

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.





X values: ='Sample dataset'!$G$2:$G$3115
Y values: ='Sample dataset'!$F$2:$F$3115

We've seen this regression before.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.4832049							
R Square	0.23348698							
Adjusted R S	0.23324067							
Standard Err	14.1346772							
Observations	3114							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	189388.892	189388.892	947.944069	6.07E-182			
Residual	3112	621743.678	199.7891					
Total	3113	811132.57						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	134.921115	2.28637329	59.0109742	0	130.438162	139.404068	130.438162	139.404068
pc_hs_grad	-1.0588226	0.03438998	-30.7887	6.07E-182	-1.126252	-0.9913933	-1.126252	-0.9913933

We've seen this regression before.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.4832049
R Square	0.23348698
Adjusted R S	0.23324067
Standard Err	14.1346772
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	189388.892	189388.892	947.944069	6.07E-182
Residual	3112	621743.678	199.7891		
Total	3113	811132.57			

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	134.921115	2.28637329	59.0109742	0	130.438162	139.404068	130.438162	139.404068
pc_hs_grad	-1.0588226	0.03438998	-30.7887	6.07E-182	-1.126252	-0.9913933	-1.126252	-0.9913933

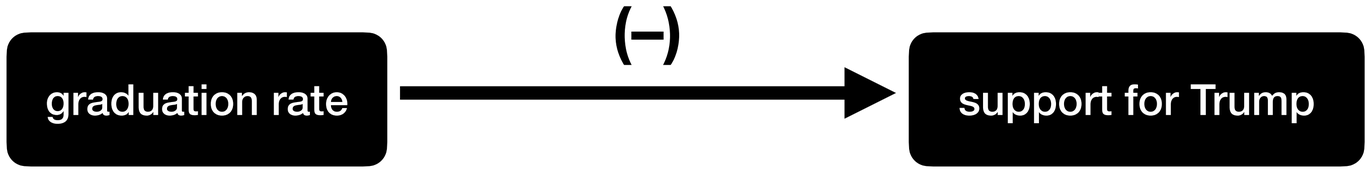
For each 1 percentage point (pp) increase in a county's high school graduation rate, the estimated support for Trump decreases by 1.1 pp.

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?

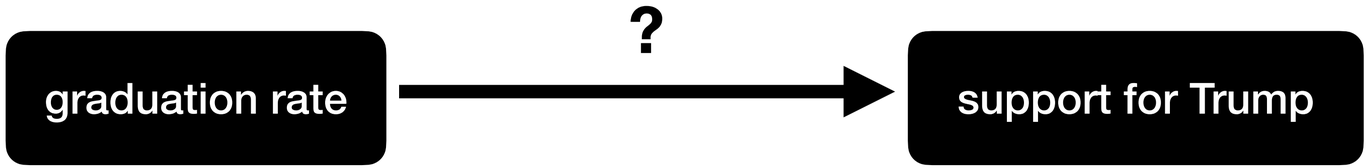
In Season 4 Episode 19, we learn that Homer never technically graduated high school (even though he does later in the episode).

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?

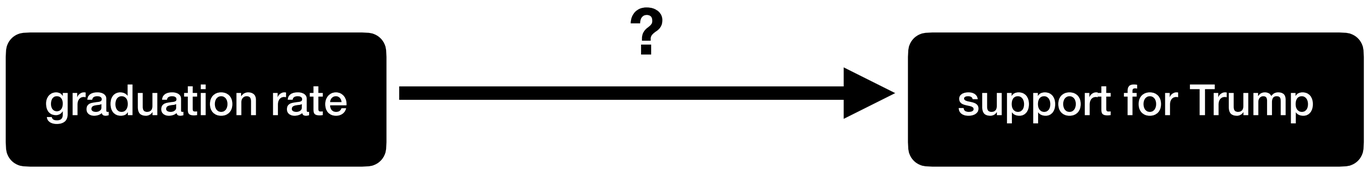
Let’s consider minoritized communities.

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.

N	O	P	Q	R
pc_black	pc_latin	med_inc_000s	majority	
20.5	2.9	57.982	=IF(OR(N2>=50,O2>=50),1,0)	
9.3	4.6	61.756	0	
48.5	4.4	34.99	0	
22.7	2.7	51.721	0	
1.9	9.3	48.922	0	
69.2	8.1	33.866	1	
45.3	1.5	44.85	0	
22.3	3.9	50.128	0	

=IF(OR(N2>=50, O2>=50), 1, 0)

OR([can be this], [or can be this])

IF([criterion], [value if true], [value if false])

OK, let's make an interacted variable.

This is easier.

Just multiply our graduation rate and dummy columns together.

G	Q	R
pc_hs_grad	majority	majority_hs_grad
68.6	0	=Q2*G2
72.8	0	0
64.3	0	0
54.9	0	0
64.9	0	0
58.6	1	58.6
53.5	0	0
65.6	0	0
62.9	0	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%.

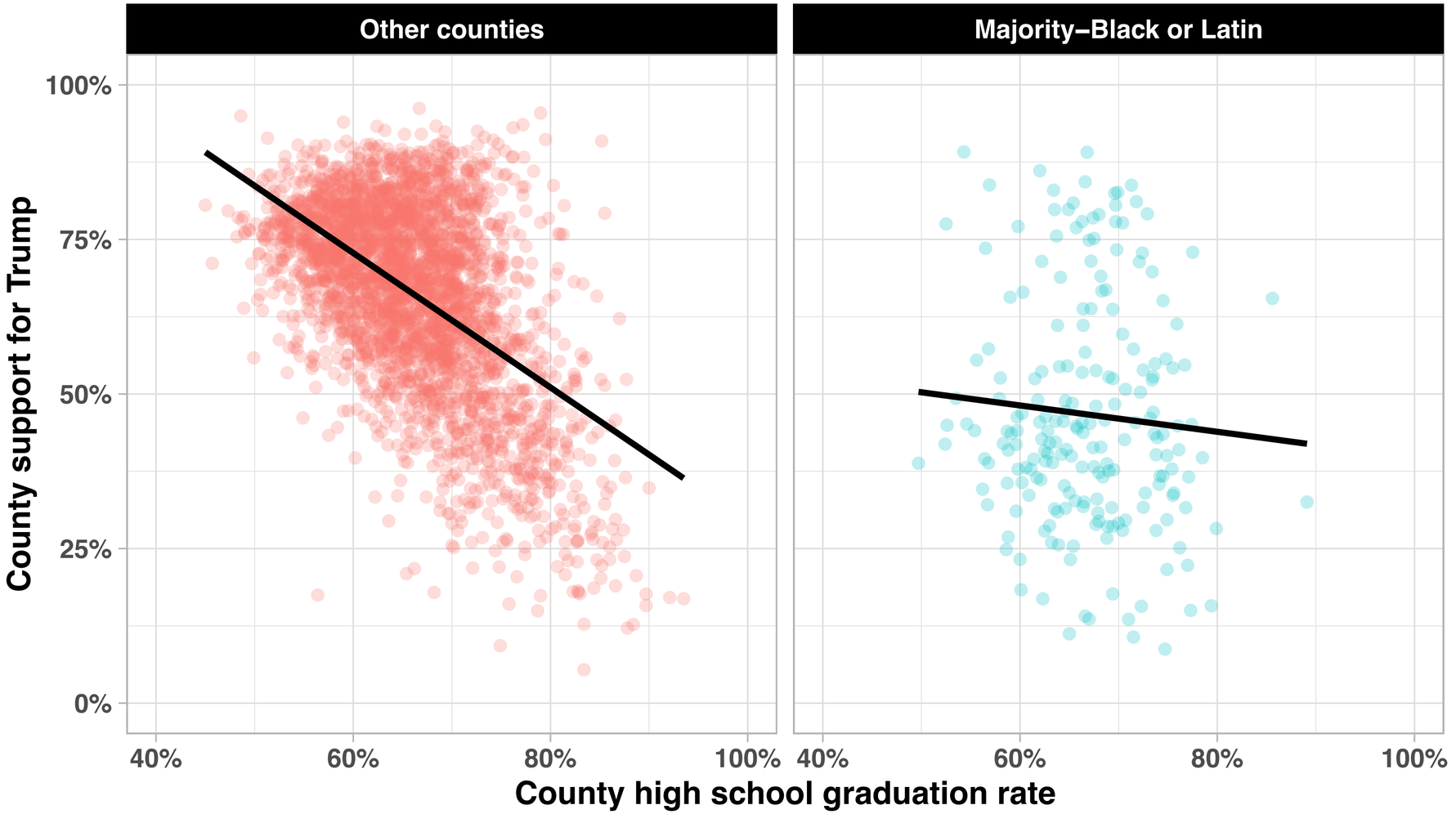
β_3 gives us the difference in slopes
i.e. the difference in the association between graduation rates and Trump support for majority-Black/Latin vs. other counties.

Majority-Black or Latin counties

here, majority = 1

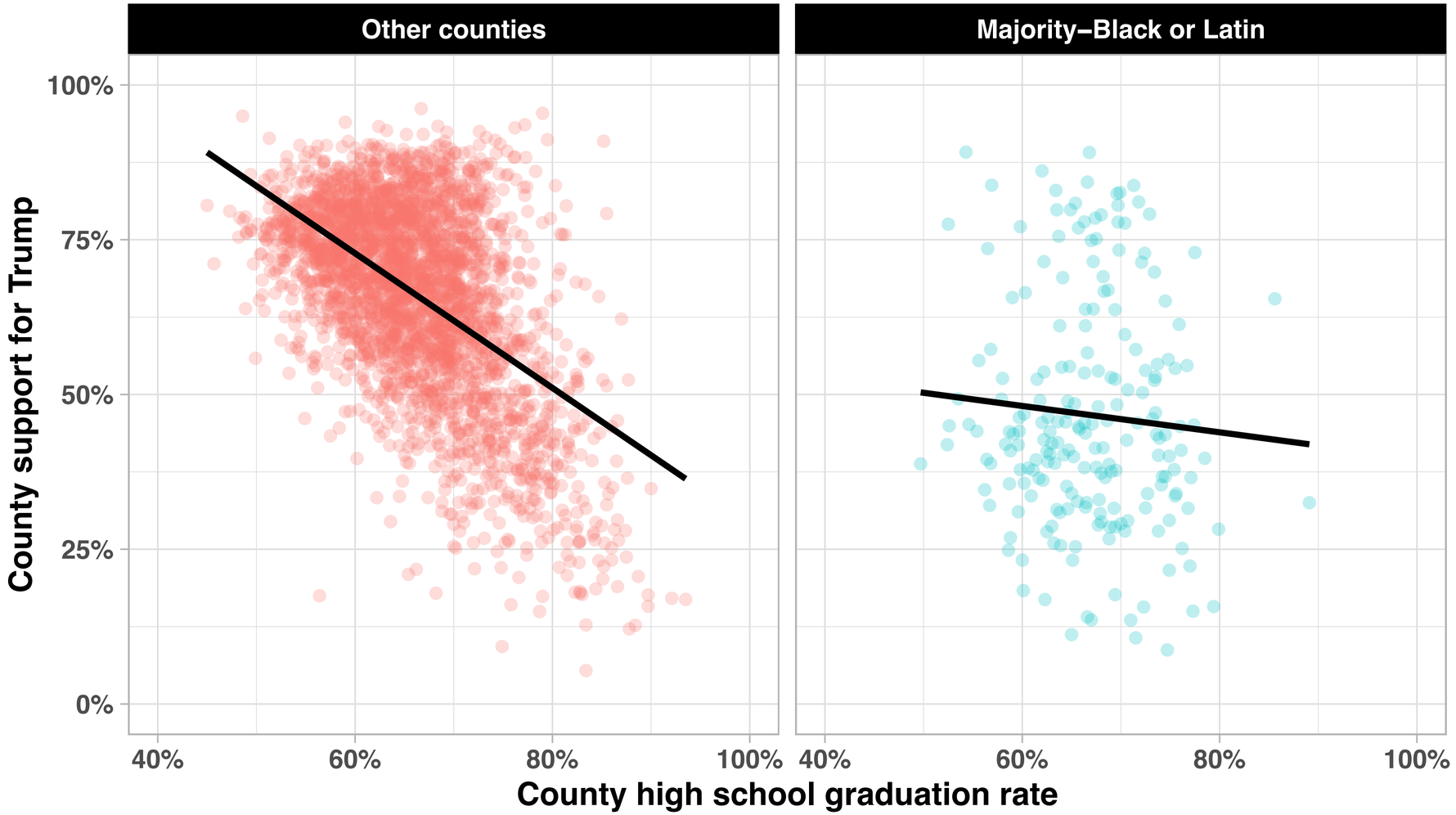
$$(Trump)_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)(HS\_grad)_i + u_i$$

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!

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.5697146
R Square	0.32457472
Adjusted R S	0.32392319
Standard Err	13.2725503
Observations	3114

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	263273.13	87757.71	498.168797	2.495E-264
Residual	3110	547859.44	176.160592		
Total	3113	811132.57			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	138.042414	2.20351028	62.6465942	0	133.721932	142.362897
pc_hs_grad	-1.0873224	0.03316569	-32.784555	9.642E-203	-1.1523512	-1.0222935
majority	-77.126184	9.8181767	-7.8554487	5.4412E-15	-96.376948	-57.875419
majority_hs_	0.87450095	0.14638393	5.97402289	2.5768E-09	0.58748201	1.16151988

Show me the regression already!

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.5697146					
R Square	0.32457472					
Adjusted R S	0.32392319					
Standard Err	13.2725503					
Observations	3114					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	263273.13	87757.71	498.168797	2.495E-264	
Residual	3110	547859.44	176.160592			
Total	3113	811132.57				
	Coefficients	*tandard Erro*	*t Stat*	*P-val*	*Lower 95%*	*Upper 95%*
Intercept	138.042414	2.20351028	62.6465942	0	133.721932	142.362897
pc_hs_grad	-1.0873224	0.03316569	-32.784555	9.642E-203	-1.1523512	-1.0222935
majority	-77.126184	9.8181767	-7.8554487	5.4412E-15	-96.376948	-57.875419
majority_hs_	0.87450095	0.14638393	5.97402289	2.5768E-09	0.58748201	1.16151988

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!

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.5697146					
R Square	0.32457472					
Adjusted R S	0.32392319					
Standard Err	13.2725503					
Observations	3114					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	263273.13	87757.71	498.168797	2.495E-264	
Residual	3110	547859.44	176.160592			
Total	3113	811132.57				
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lowe 95%*	*Upper 95%*
Intercept	138.042414	2.20351028	62.6465942	0	133.721932	142.362897
pc_hs_grad	-1.0873224	0.03316569	-32.784555	9.642E-203	-1.1523512	-1.0222935
majority	-77.126184	9.8181767	-7.8554487	5.4412E-15	-96.376948	-57.875419
majority_hs_	0.87450095	0.14638393	5.97402289	2.5768E-09	0.58748201	1.16151988

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!

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.5697146					
R Square	0.32457472					
Adjusted R S	0.32392319					
Standard Err	13.2725503					
Observations	3114					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	263273.13	87757.71	498.168797	2.495E-264	
Residual	3110	547859.44	176.160592			
Total	3113	811132.57				
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	138.042414	2.20351028	62.6465942	0	133.721932	142.362897
pc_hs_grad	-1.0873224	0.03316569	-32.784555	9.642E-03	-1.1523512	-1.0222935
majority	-77.126184	9.8181767	-7.8554487	5.4412E-15	-96.376948	-57.875419
majority_hs_	0.87450095	0.14638393	5.97402289	2.5768E-09	0.58748201	1.16151988

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!

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.5697146					
R Square	0.32457472					
Adjusted R S	0.32392319					
Standard Err	13.2725503					
Observations	3114					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	263273.13	87757.71	498.168797	2.495E-264	
Residual	3110	547859.44	176.160592			
Total	3113	811132.57				
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	138.042414	2.20351028	62.6465942	0	133.721922	142.362897
pc_hs_grad	-1.0873224	0.03316569	-32.784555	9.642E-203	-1.1523512	-1.0222935
majority	-77.126184	9.8181767	-7.8554487	5.4[illegible]E-15	-96.376948	-57.875419
majority_hs_	0.87450095	0.14638393	5.97402289	2.5768E-09	0.58748201	1.16151988

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.

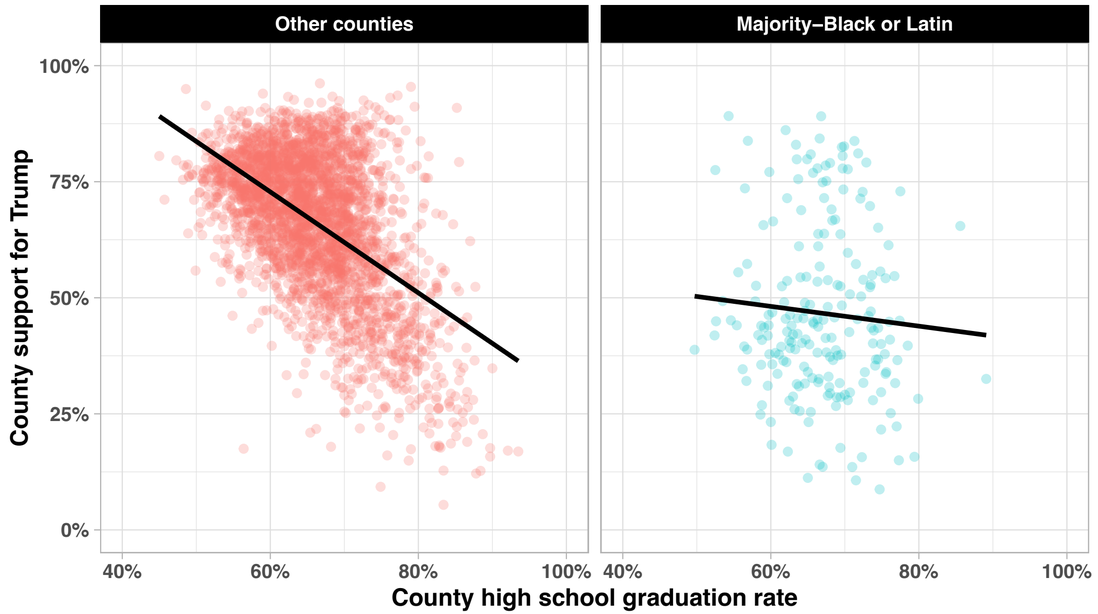
Doing the math: –1.09 + 0.87 = –0.22, which is a much flatter slope.

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

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



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

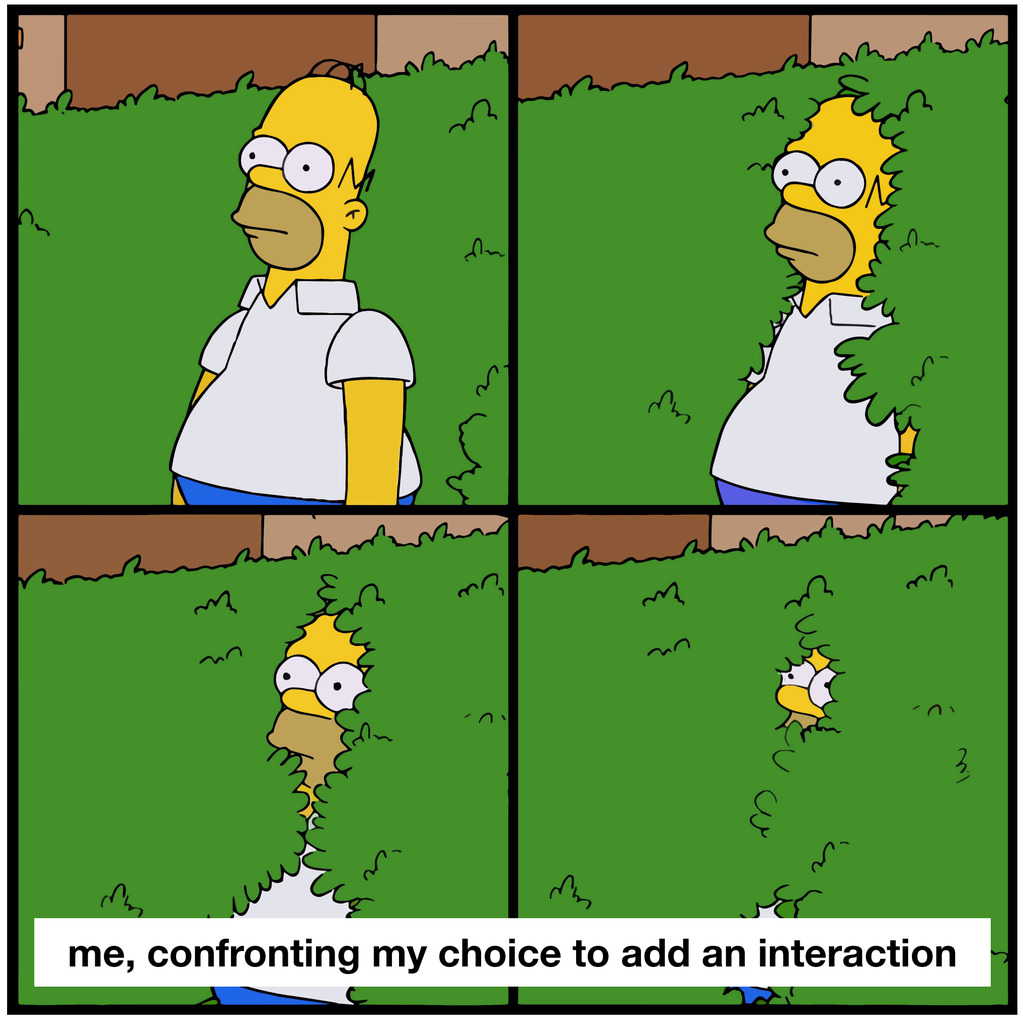
$\beta_0 + \beta_2 = 138 - 77 = 61$

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

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.