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

EXCEL

What’s the deal with regression?

API 202: TF Session 1

Nolan M. Kavanagh
January 30, 2026

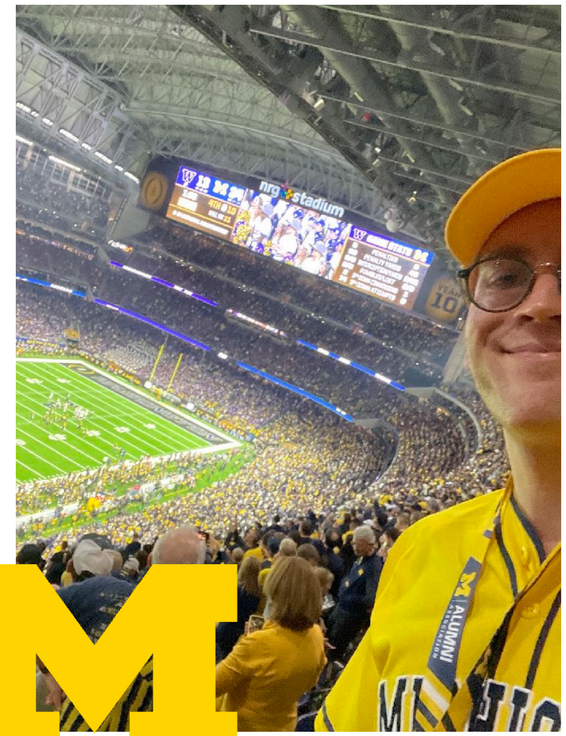
Goals for today

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!

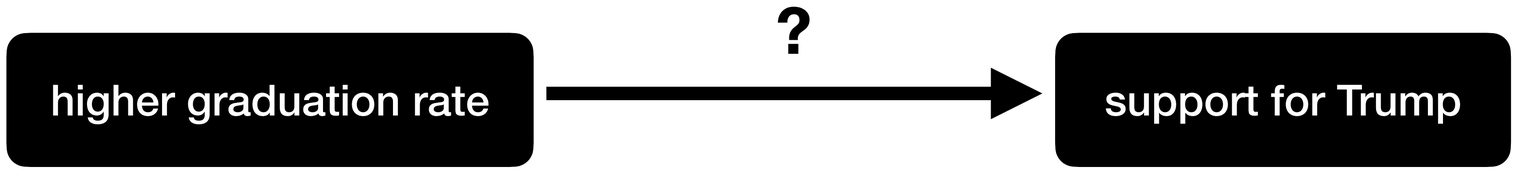
Overview of our sample data

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.

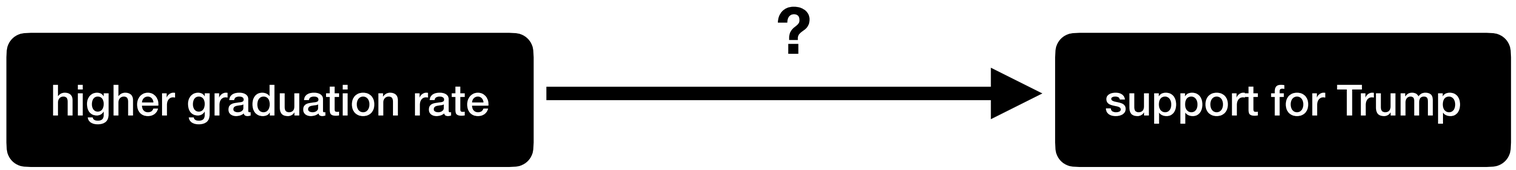
Let's say we're interested in the relationship between high school graduation and support for Trump.



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.

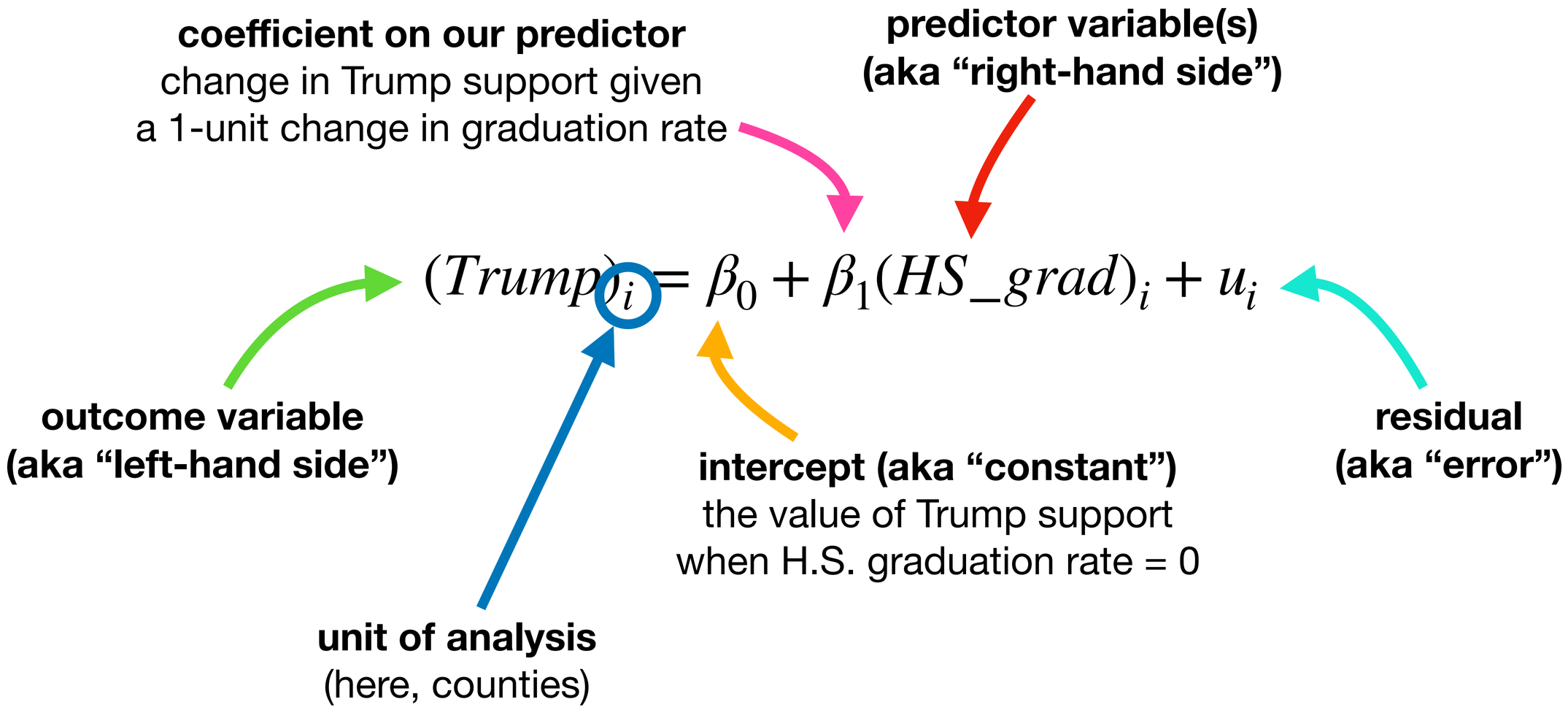


What might be the direction? Mechanism?

More education = liberal values = prefer multiculturalism?

More education = more income = prefer lower taxes?

Population regression function



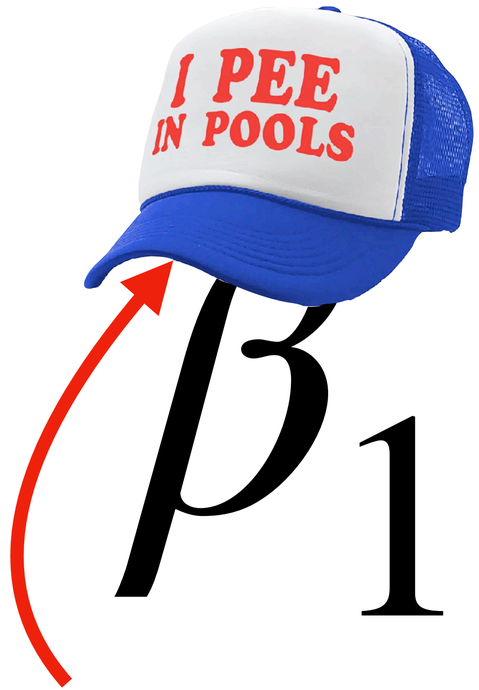
Population regression function

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

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

Sample regression function

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

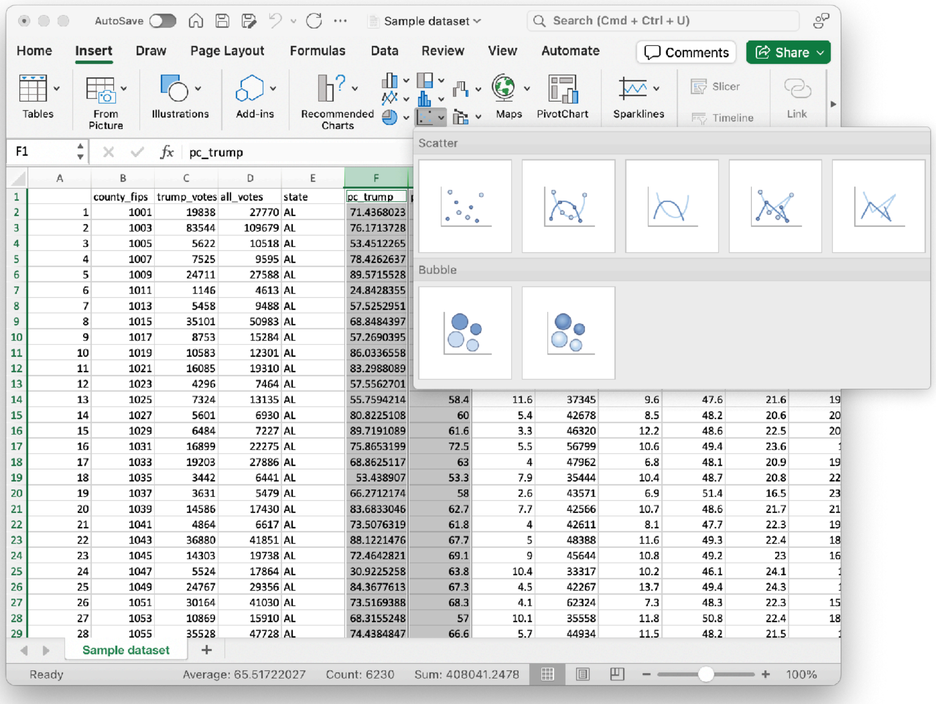


Only a <u>specific</u> sample would ever wear this hat.

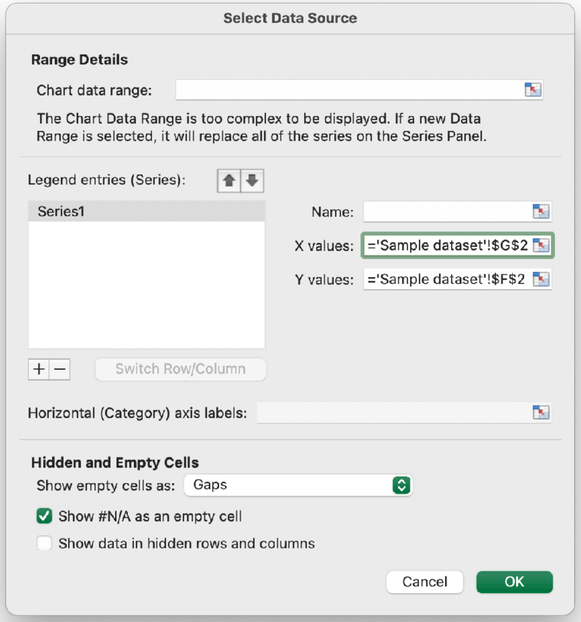

- That's the hat I gave her, she's wearing it as a
She's a menace to society.
regression estimate

Let’s graph our data.

1. Insert a scatterplot.



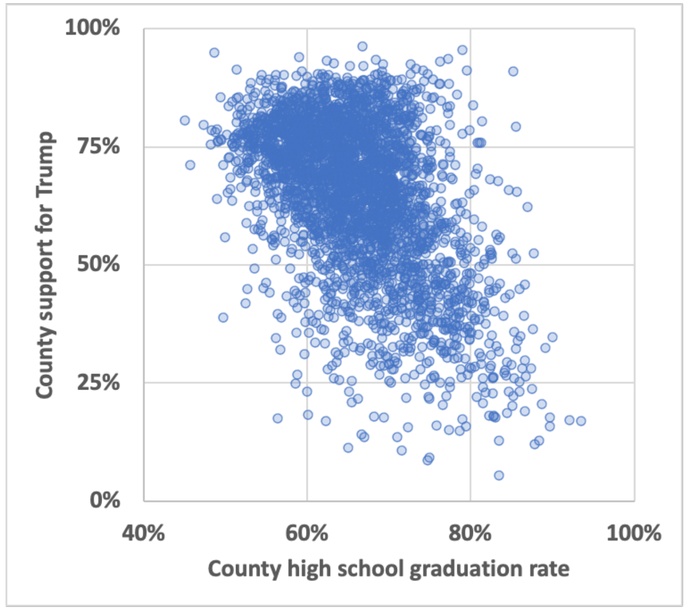
2. Delete the X & Y ranges and replace with correct columns.



I recommend excluding column headers from the X and Y values (i.e. start at cell 2).

Shortcut for selecting cells: Click on cell 2, then press SHIFT + CTL/⌘ + DOWN.

3. Chart Design > Add Chart Element, like axis labels.

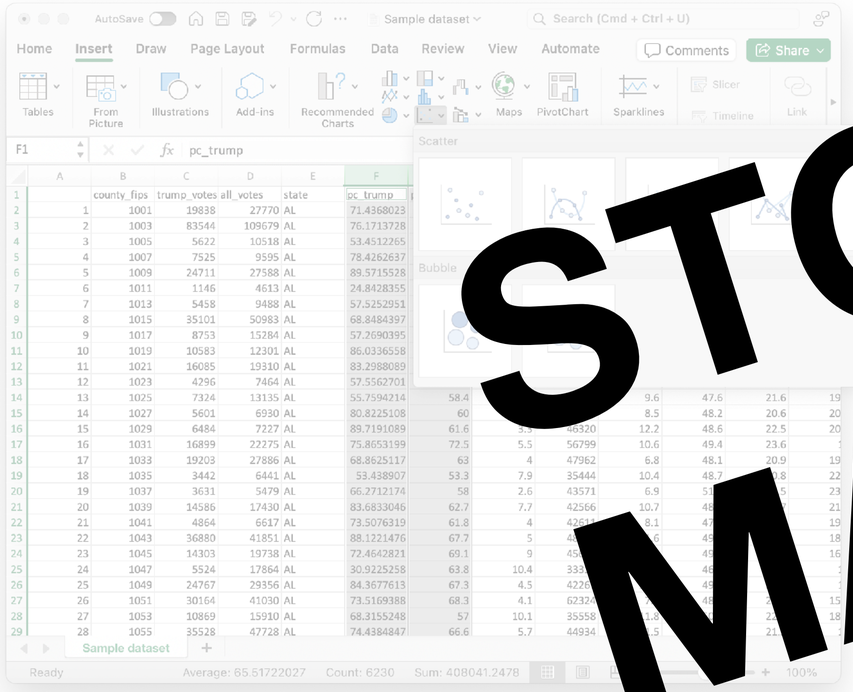


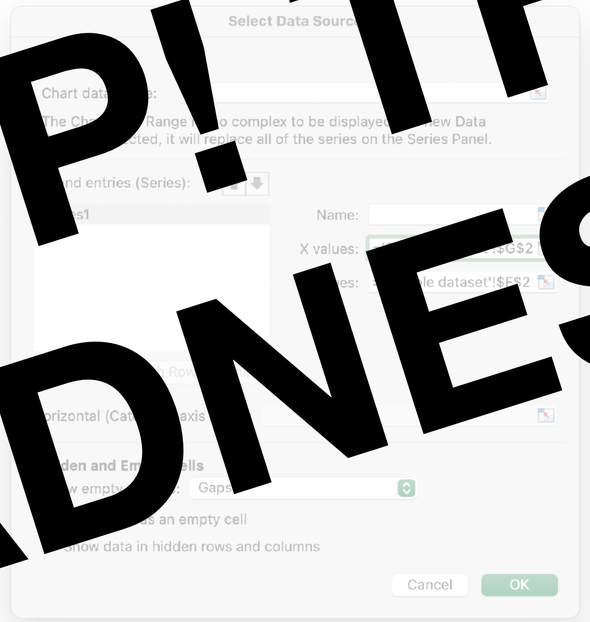
To modify an existing element, double-click on it and a formatting window will pop up on the right.

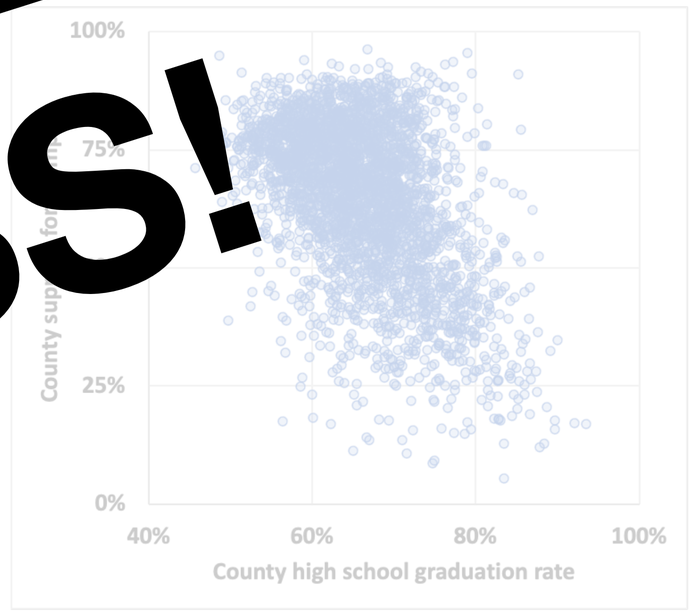
Let’s graph our data.

1. Insert a scatterplot.

2. Delete the X & Y ranges and replace with correct columns. Chart Design > Add Chart Element, like axis labels.







STOP! THE! MADNESS!

I recommend excluding column headers from the X and Y values (i.e. start at cell 2).

Shortcut for selecting cells: Click on cell 2, then press SHIFT + CTL/⌘ + DOWN.

To modify an existing element, double-click on it and a formatting window will pop up on the right.

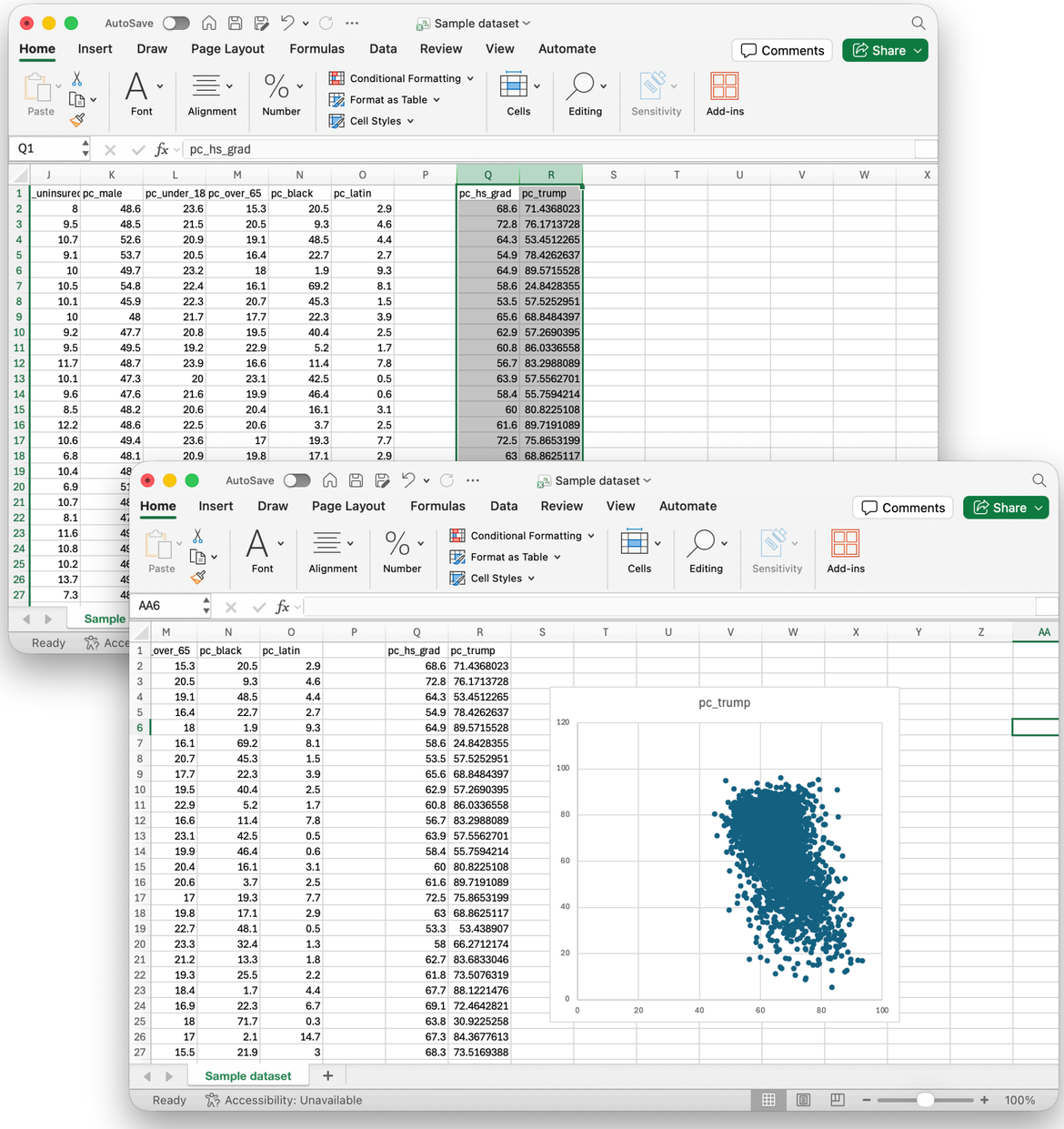
Look, there's a quick and dirty way.

1. Cut and paste the columns next to each another: X first, then Y.

2. Highlight both columns and select "Insert Scatterplot." Voilà.

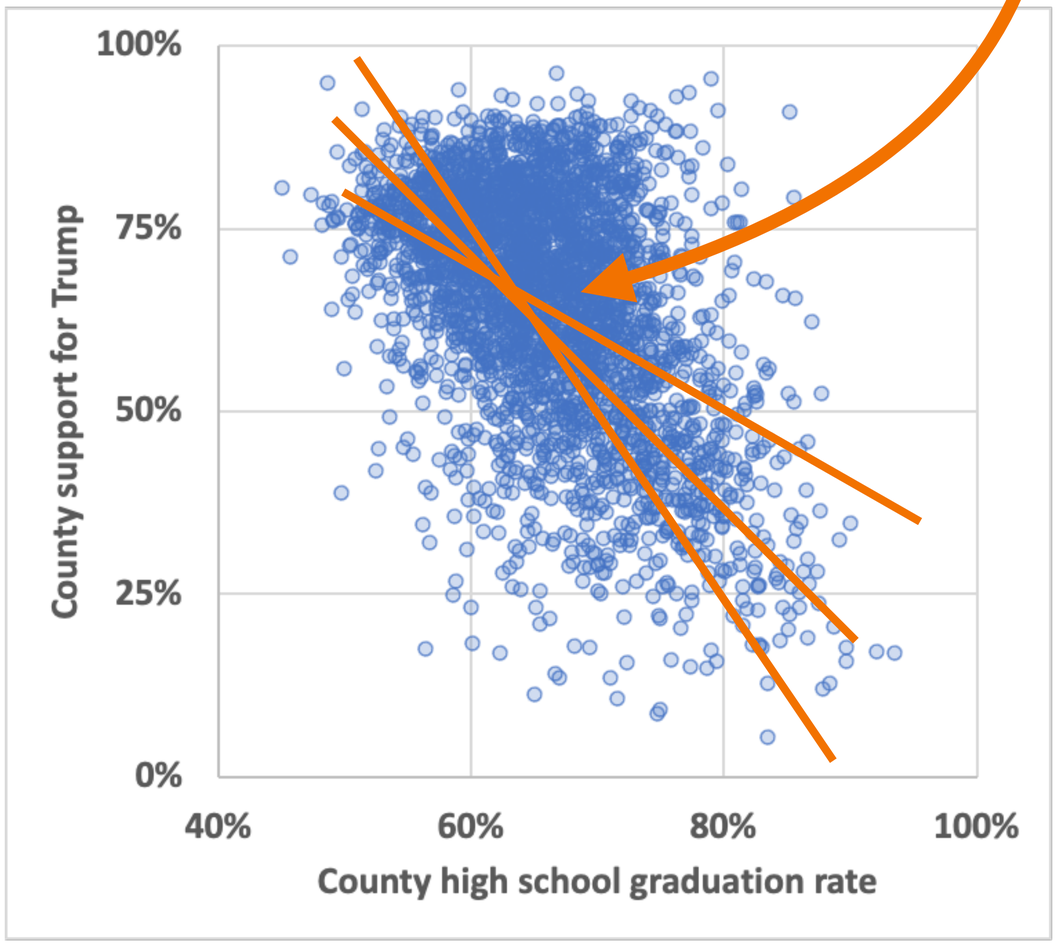
But this method isn't foolproof.

Excel will only "intuit" the desired scatterplot *if* the columns are ordered correctly and you promise it your firstborn child.



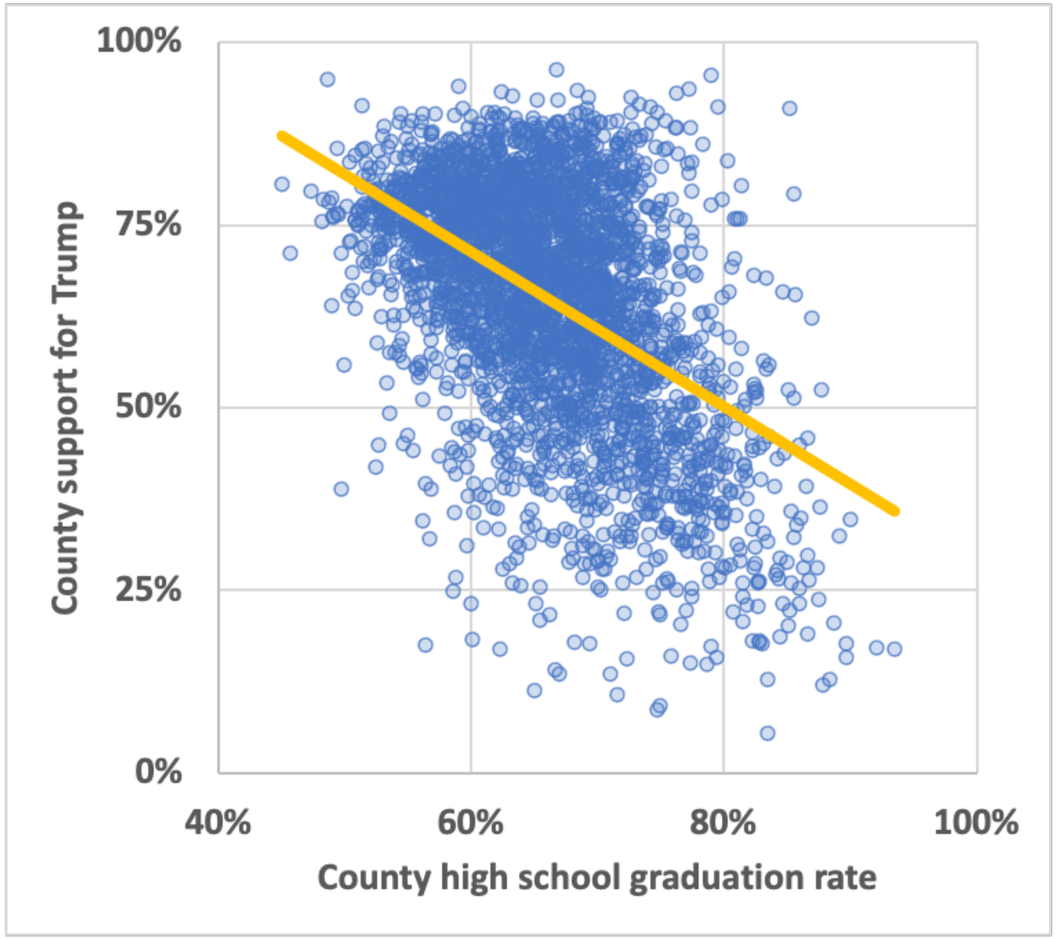
Let’s graph our data.

Which line fits best?



Let’s graph our data.

We can add a linear model under Add Chart Element > Trendline > Linear.

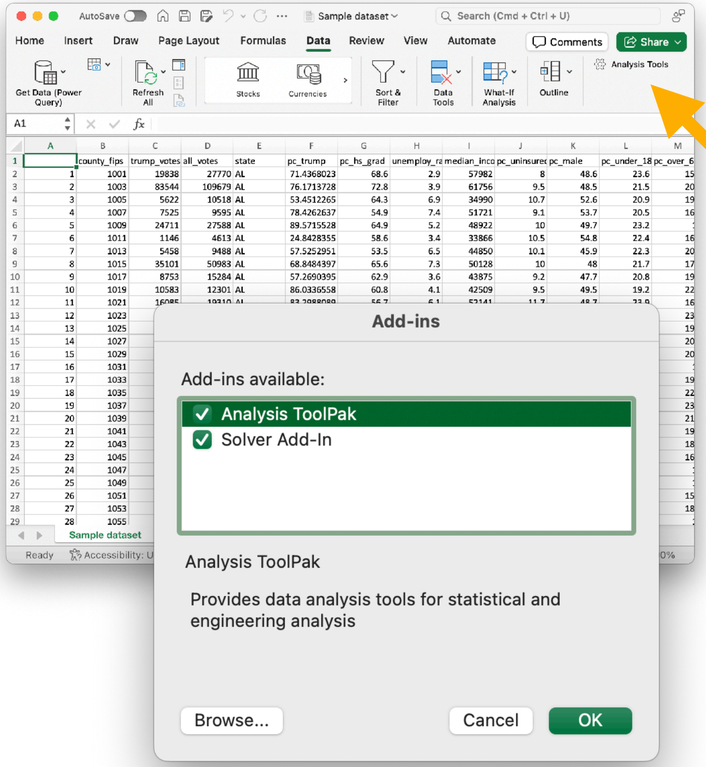


…but we still want to run the regression.

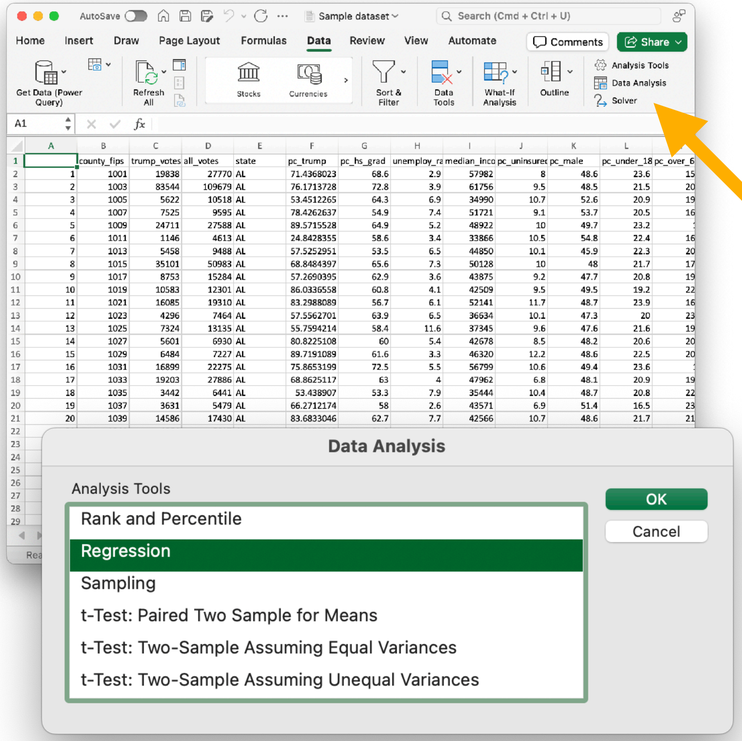

Maybe this will become, like,
a cool thing,
making sick graphs
tbs
Then maybe baldness will catch on.
tbs

Alright, let's run a regression.

1. Data > Analysis Tools. Add the Analysis ToolPak.

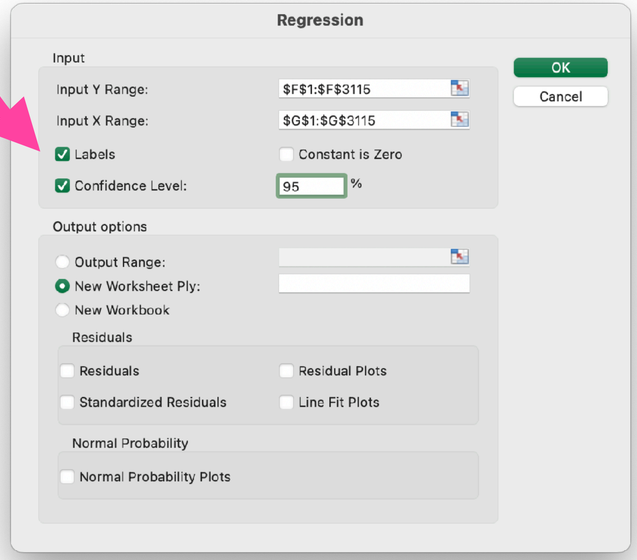


2. Click Data Analysis. Then, select Regression.



3. Modify X and Y ranges. Add labels and 95% C.I.s.

Check "Labels" if the first cell is a variable name.



<u>Other tidbits</u>

- Constant should <u>not</u> be 0.
- Can ask for residuals if you plan to use them for another analysis.
- For ranges, have to give first and last cell in the range: "$F$1:$F$3115". Can't just name the column.

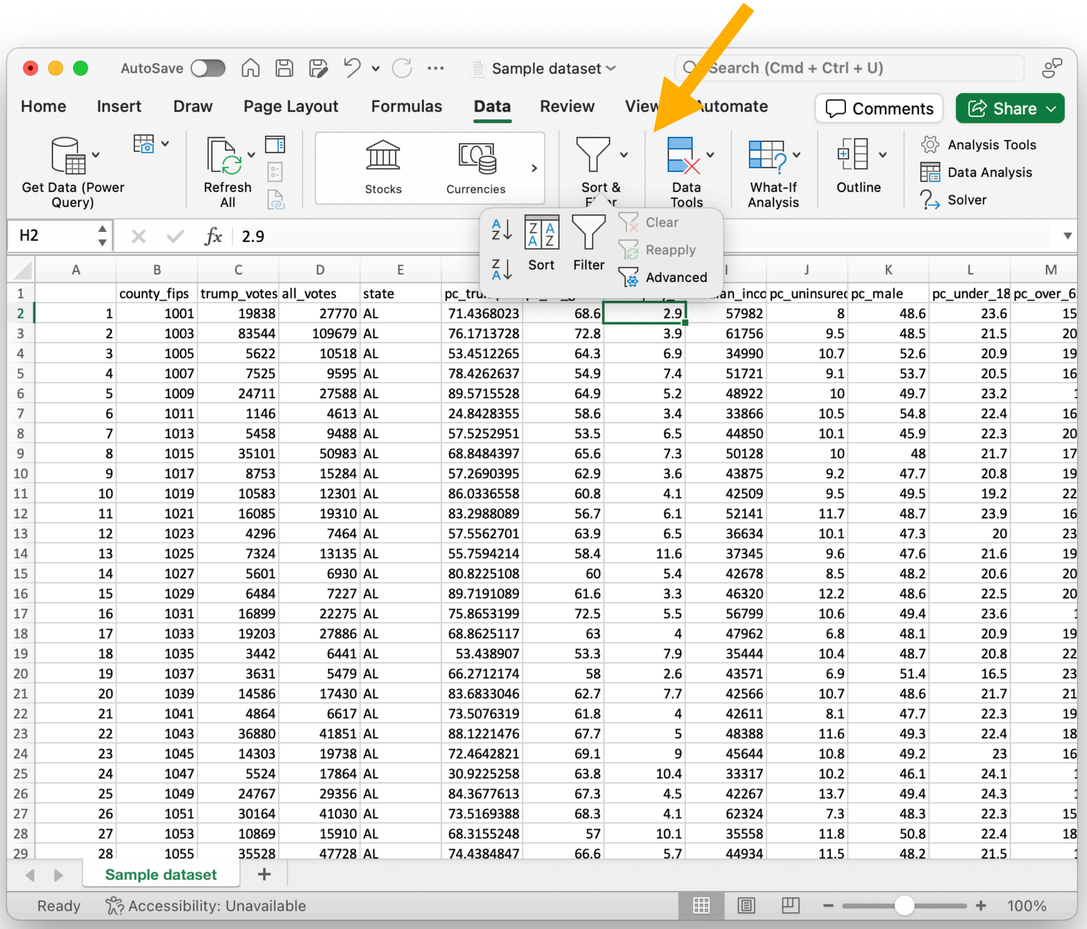
Alright, let's run a regression.

If your X or Y columns have missing data (e.g. blank cells or "NAs"), you'll get an error and the regression won't run.

We have a few options for dealing with it:

- "Sort" the column(s) with missing data so that the NAs go to the bottom. Then, specify a range that excludes the missing rows.

- Use "Filter" and uncheck the NAs. Then, copy the remaining rows into a new sheet and run the regression on these rows.



	A	B	C	D	E	J	K	L	M
1		county_fips	trump_votes	all_votes	state	pc_uninsured	pc_male	pc_under_18	pc_over_6
2	1	1001	19838	27770	AL	8	48.6	23.6	15
3	2	1003	83544	109679	AL	9.5	48.5	21.5	20
4	3	1005	5622	10518	AL	10.7	52.6	20.9	19
5	4	1007	7525	9595	AL	9.1	53.7	20.5	16
6	5	1009	24711	27588	AL	10	49.7	23.2	
7	6	1011	1146	4613	AL	10.5	54.8	22.4	16
8	7	1013	5458	9488	AL	10.1	45.9	22.3	20
9	8	1015	35101	50983	AL	10	48	21.7	17
10	9	1017	8753	15284	AL	9.2	47.7	20.8	19
11	10	1019	10583	12301	AL	9.5	49.5	19.2	22
12	11	1021	16085	19310	AL	11.7	48.7	23.9	16
13	12	1023	4296	7464	AL	10.1	47.3	20	23
14	13	1025	7324	13135	AL	9.6	47.6	21.6	19
15	14	1027	5601	6930	AL	8.5	48.2	20.6	20
16	15	1029	6484	7227	AL	12.2	48.6	22.5	20
17	16	1031	16899	22275	AL	10.6	49.4	23.6	
18	17	1033	19203	27886	AL	6.8	48.1	20.9	19
19	18	1035	3442	6441	AL	10.4	48.7	20.8	22
20	19	1037	3631	5479	AL	6.9	51.4	16.5	23
21	20	1039	14586	17430	AL	10.7	48.6	21.7	21
22	21	1041	4864	6617	AL	8.1	47.7	22.3	19
23	22	1043	36880	41851	AL	11.6	49.3	22.4	18
24	23	1045	14303	19738	AL	10.8	49.2	23	16
25	24	1047	5524	17864	AL	10.2	46.1	24.1	
26	25	1049	24767	29356	AL	13.7	49.4	24.3	
27	26	1051	30164	41030	AL	7.3	48.3	22.3	15
28	27	1053	10869	15910	AL	11.8	50.8	22.4	18
29	28	1055	35528	47728	AL	11.5	48.2	21.5	

Alright, let's interpret a regression.

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

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

Alright, let's interpret a regression.

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

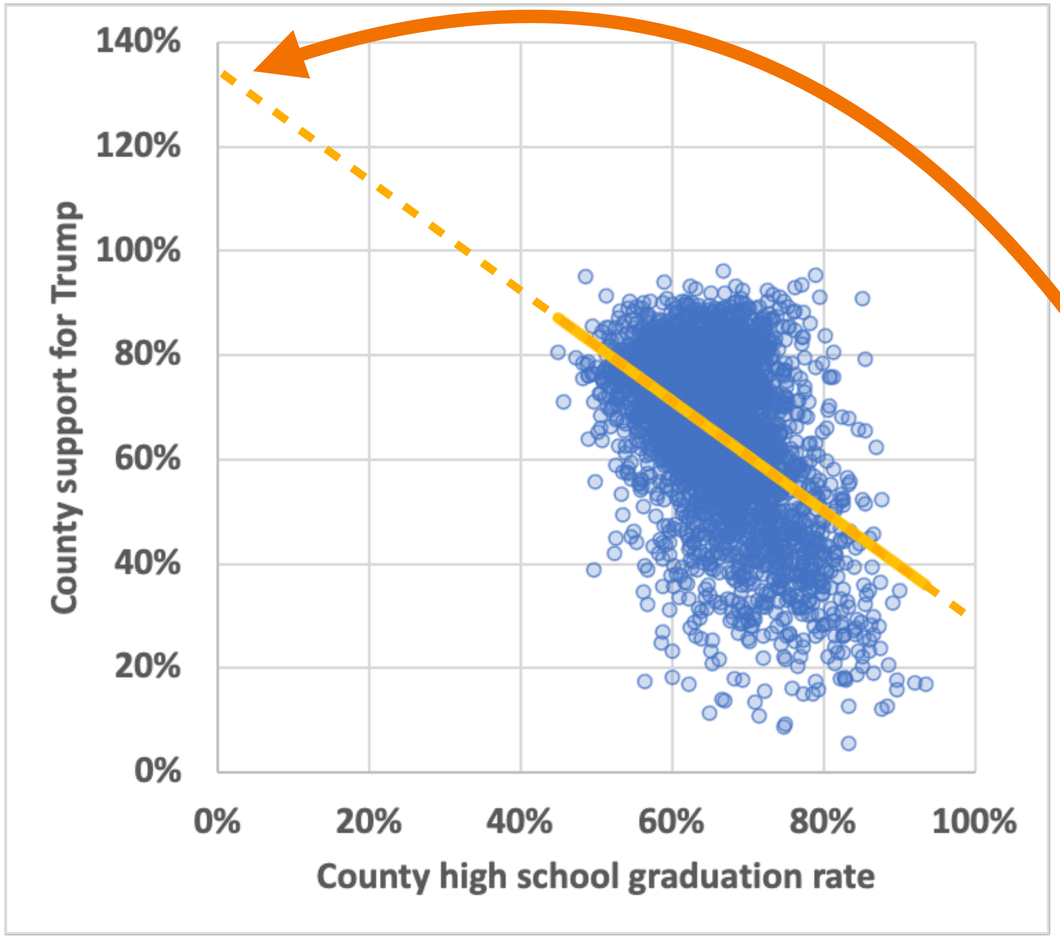
We'll start with the intercept.

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

(Is a graduation rate of 0% meaningful?)

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, 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.

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

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

We'll start with the intercept.

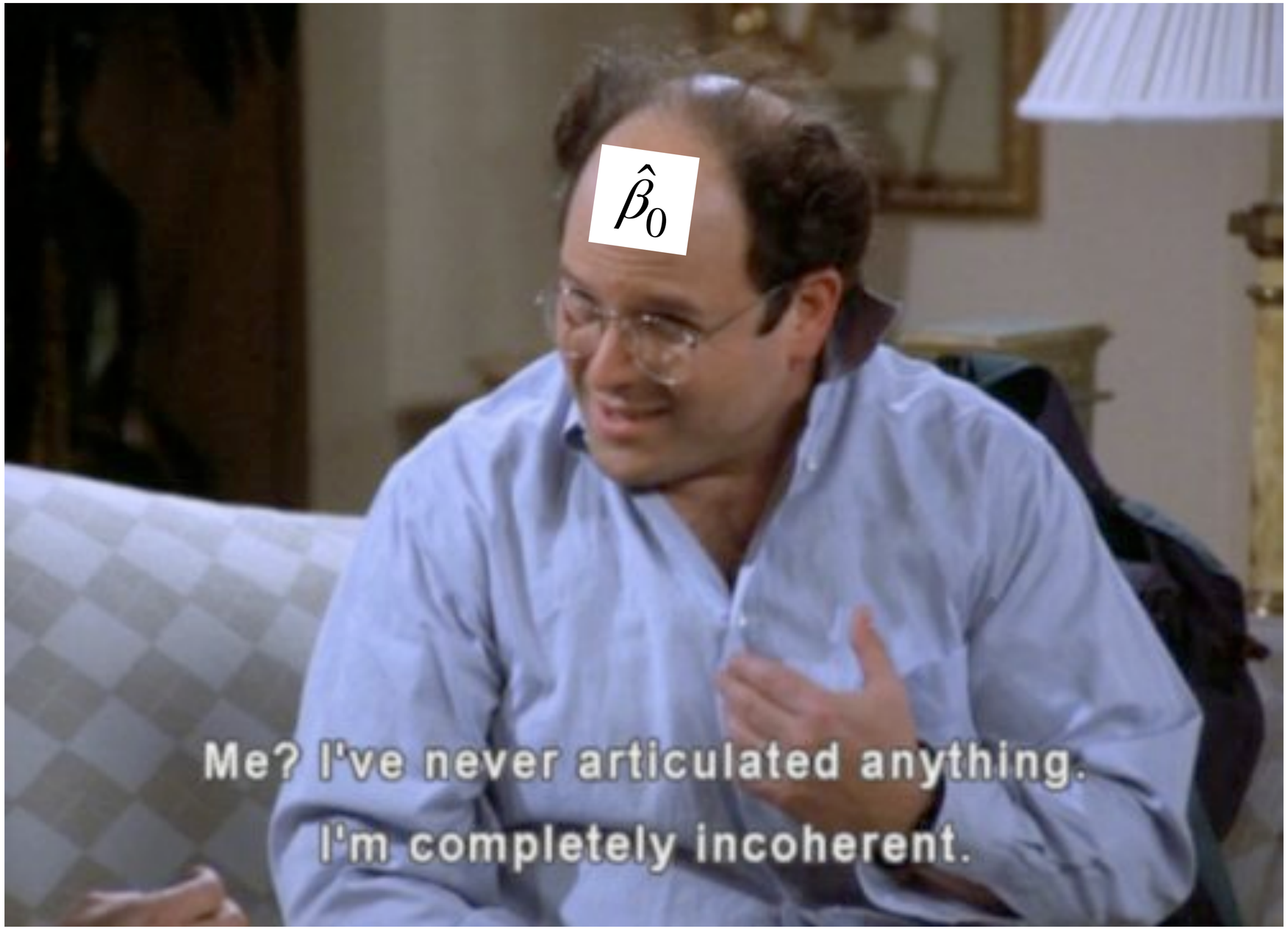
When 0% of a county has graduated high school, the predicted support for Trump is 134.9%.

(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.

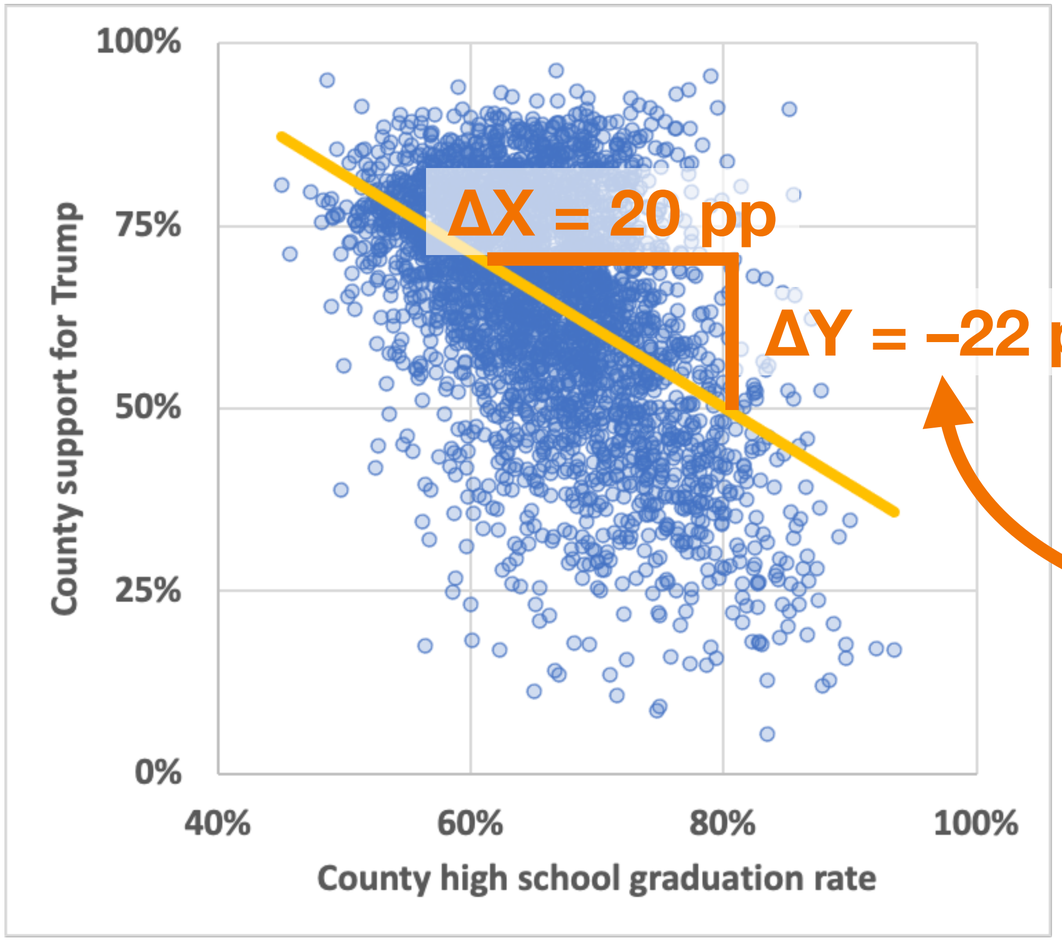
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.4832049							
R Square	0.23348698							
Adjusted R S	0.23324067							
Standard Err	14.1346772							
Observations	3114							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	189388.892	189388.892	947.944069	6.07E-182			
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

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.



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.

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

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?