Political Science 210: Introduction to Empirical Methods

Week 5: Regression & Large-N Analysis

What can we learn from a large set of cases?

If we have a large set of *cases* (people, parties, countries) containing X, it gives us leverage to explore the X-Y relationship in some ways:

- Do X and Y *covary*?
  - We can compare across cases to see if levels or presence/absence of X is related to levels or presence/absence of Y
- Are there other *confounding* factors affecting X and Y?
  - We can “control” for variation introduced by other factors
  - But not a perfect approach: We don’t know if we’ve controlled for every situation; too many controls can lead to bad predictions and make results difficult to interpret.
- Is the relationship *generalizable* to other cases?
  - By definition, we’re studying average effects across many cases

What can we learn from a large set of cases?

But even a large set of cases can't tell us everything about X and Y:

- Does X *cause* Y?
  - Can't tell if X or Y happened first
  - Can't check for reverse causality (is Y causing X?)
  - Don't know the counterfactual or "potential outcome" for a given case: We can only compare to other cases that are different from the given case.
- What is the *causal mechanism* or the *pathway* connecting X and Y? What "story" are we telling that explains their relationship?
  - You could conduct further tests for mechanisms that fit your theory, but harder to get much within-case detail when studying across so many cases - you'd need a different approach.

Regression

Regression is an especially powerful and popular tool for measuring the relationship between concepts across a large number of cases.

Ordinary least squares (OLS) regression is a straightforward form of regression that attempts to estimate the *linear* relationship between concepts that produces the *least* amount of error when applied to a large number of cases.

In OLS, the Y variable is numeric, NOT categorical.

- There are other regression methods that allow for a categorical Y variable, but can only cover so much in one week.

Regression

The basic OLS regression formula can be written the same way that we might plot cases on a two-dimensional X-Y axis:

$$Y = a + bX$$

- **Y** is the outcome of interest (but again, we can't prove Y is "caused")
- **X** is the variable that we think can explain Y (but again, don't know causality)
- **a** is the *average* value of Y when X is zero (the y-axis "intercept")
- **b** is the amount that Y changes *on average* when X changes by one interval.

Interpreting beta

When X *changes* by one interval, what *change* will we see in Y, on average?

$$Y = a + bX$$

- Let's say we want to know if the level of wealth in a neighborhood increases voter turnout.
  - Our *cases* are neighborhoods
  - Our Y (dependent variable) is voter turnout measured by percentage of VEP who voted in the last election, which we can learn from (e.g.) polling outlets
  - Our X (independent variable) is level of wealth measured by average income (in 1000s of dollars), which we can learn from sources like US Census data.
  - Let's say we run our regression and find a beta of 0.2. What have we learned?
    - For every $1000 increase in average income, neighborhood turnout increases by 20%.

Interpreting beta

When X *changes* by one interval, what *change* will we see in Y, on average?

$$Y = a + bX$$

- Let's say we want to know if getting a bachelor's degree increases a person's level of income.
  - Our *cases* are people.
  - Our Y (dependent variable) is a person's level of income, measured in 10,000s of dollars
  - Our X (independent variable) is whether or not a person obtained a bachelor's degree
  - Let's say we run our regression and find a beta of 2.5. What have we learned?
    - A person with a bachelor's degree (X = 1) earns $25,000 more on average than a person without a bachelor's degree (X = 0).
    - The value of **a** represents how much someone without a bachelor's degree earns.

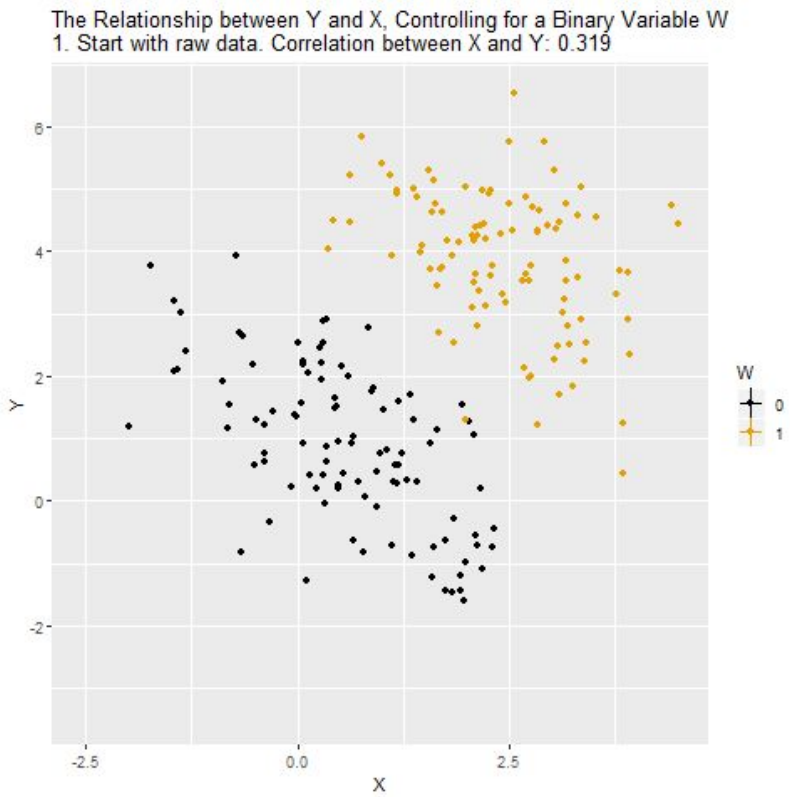
Controls

Let's say we add another independent variable, leaving us with two: X_1 and X_2.

$$Y = a + b_1X_1 + b_2X_2$$

- If X_1 and X_2 are not correlated with each other, then they each "explain" different types of changes, or variation, in Y.
- In this equation, b_1 tells us "effect" of X_1 on Y when we *remove the variation* in Y that is explained by X_2.
  - This is also called "holding constant" X_2 or "controlling for" X_2.
- When used correctly, this can allow us to either "ignore" changes in Y that are explained by X_2 and increase our confidence that we are measuring independent effects of X_1.

Visualizing control

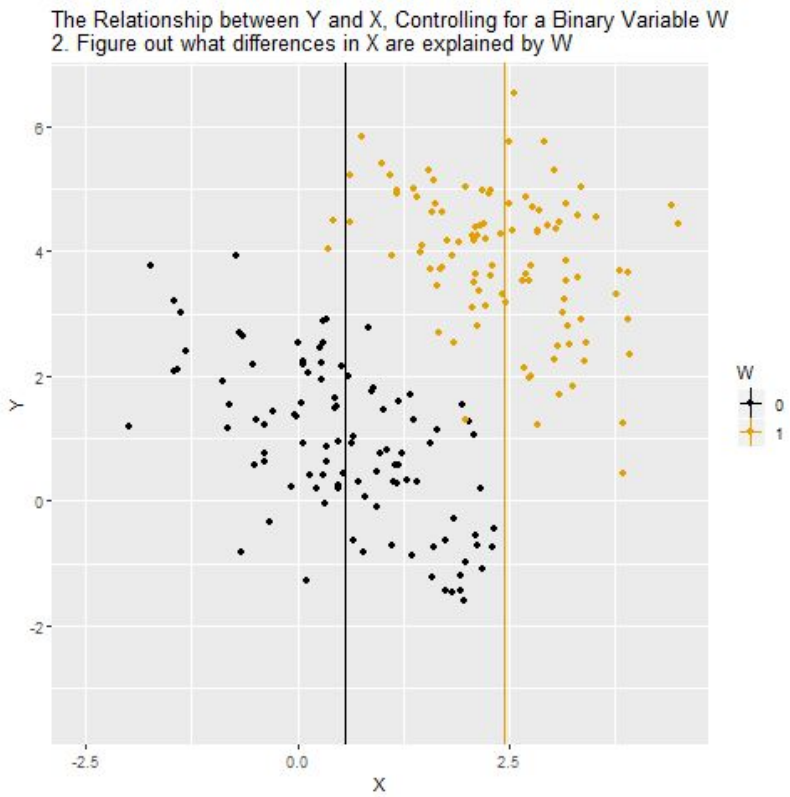


Say we have a set of cases with a positive (but fairly weak) correlation between measures of X and measures of Y.

But a third variable, W, is a categorical variable with values of either 0 (yellow) or 1 (black). We think it might explain some of the variation we see, independent of the relationship between X and Y.

https://nickchk.com/causalgraphs.html

Visualizing control

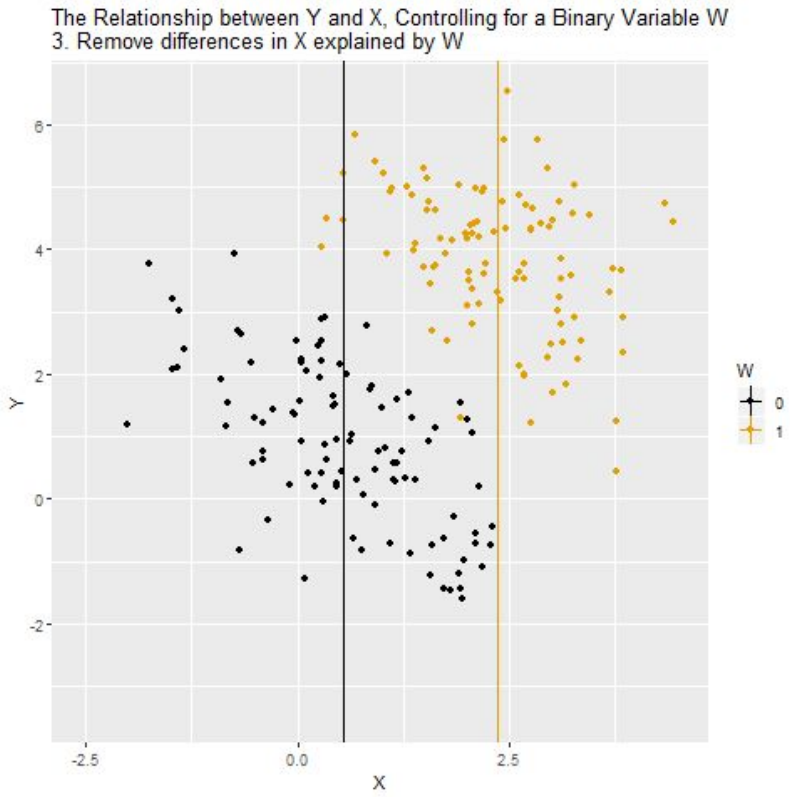


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

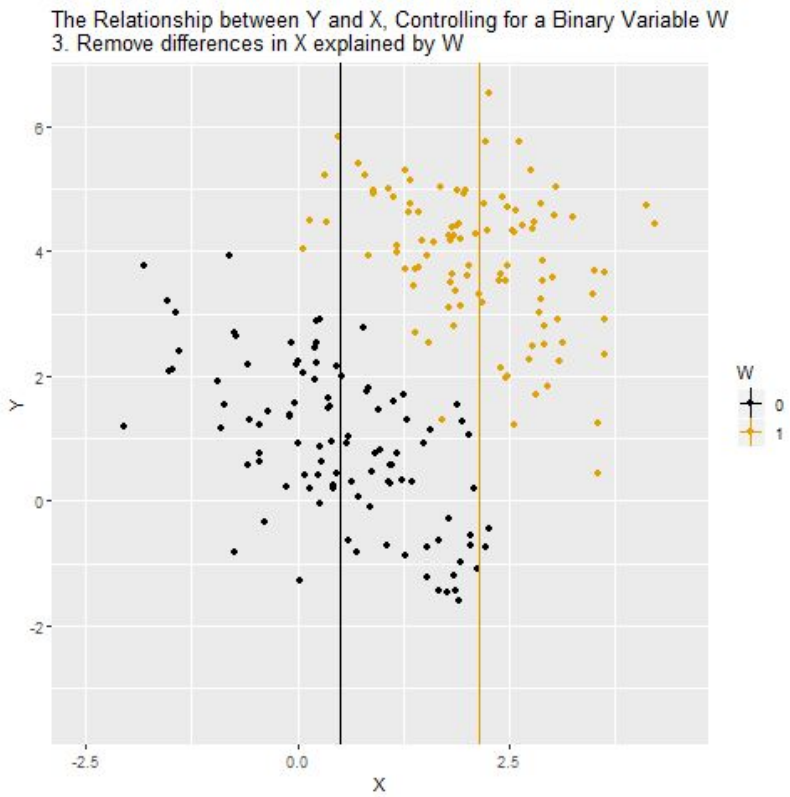


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

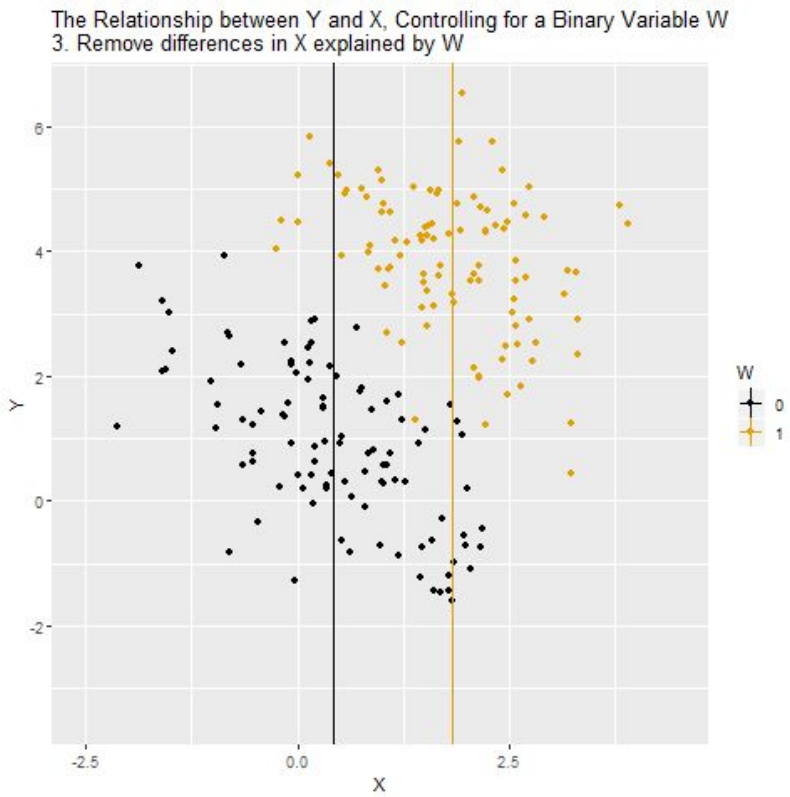


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

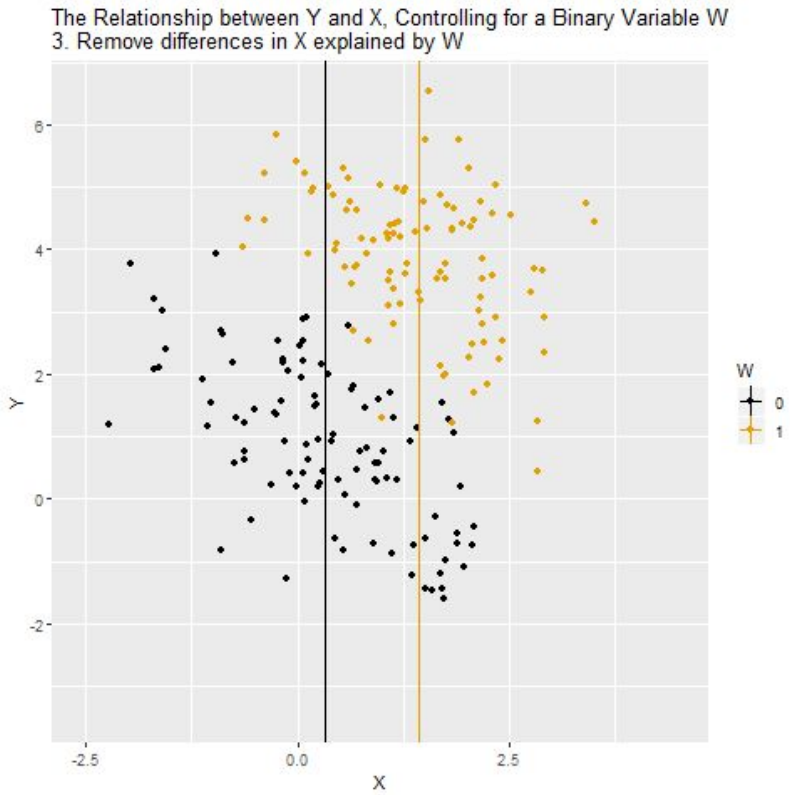


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

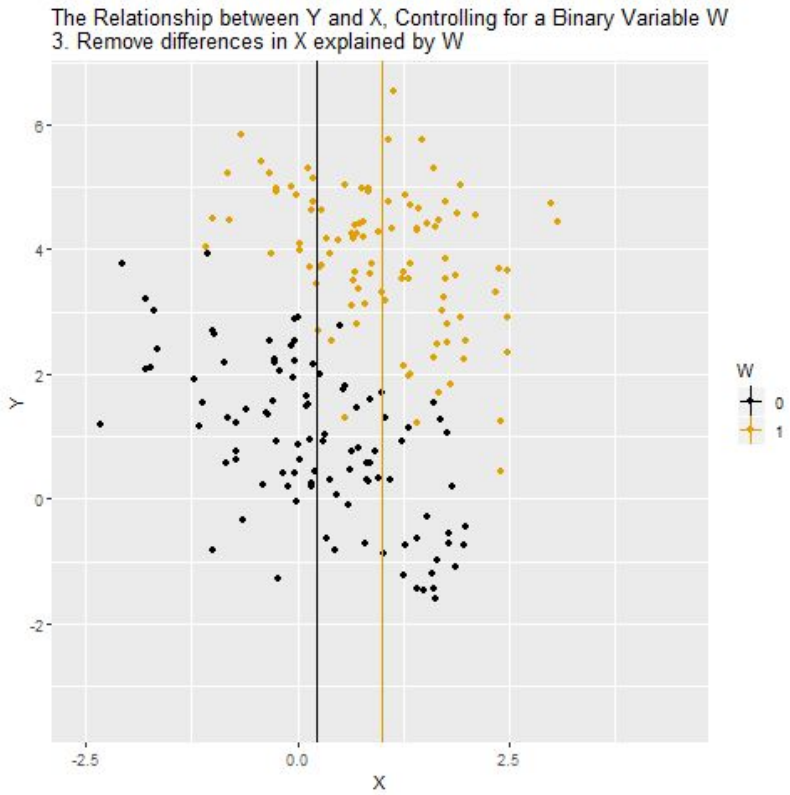


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

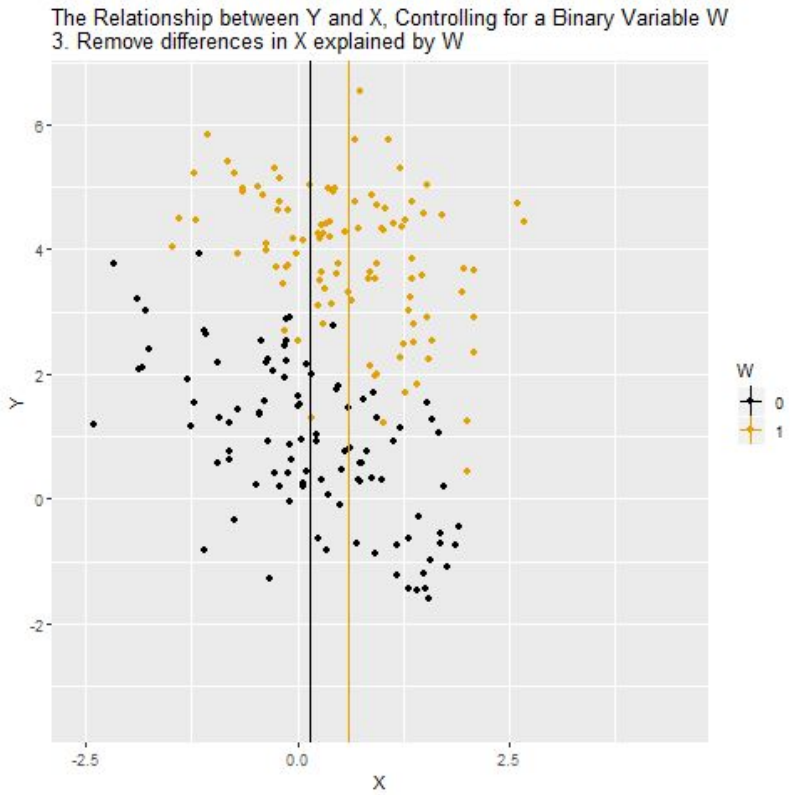


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

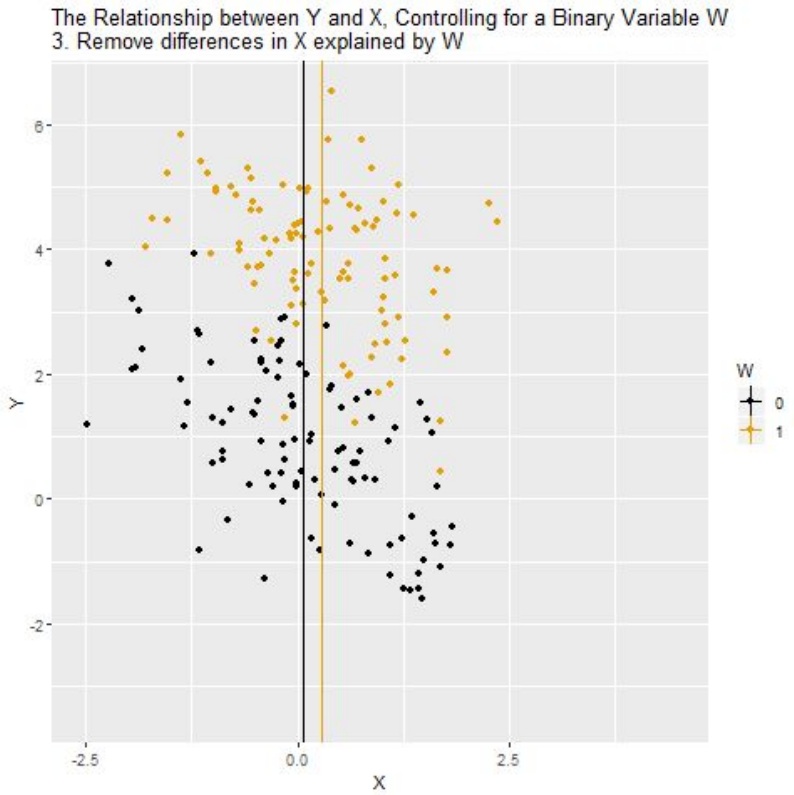


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

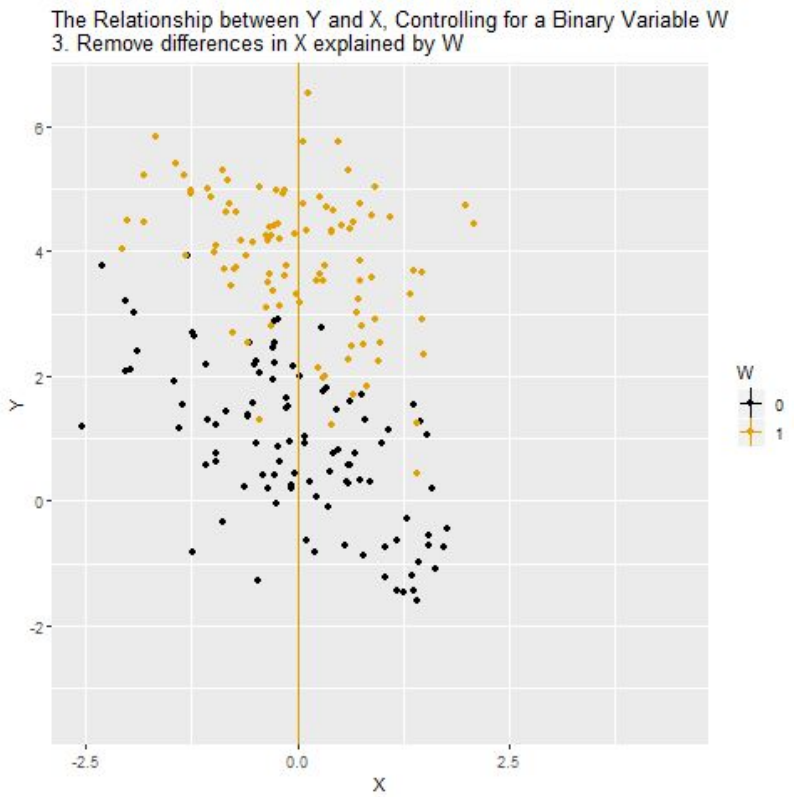


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

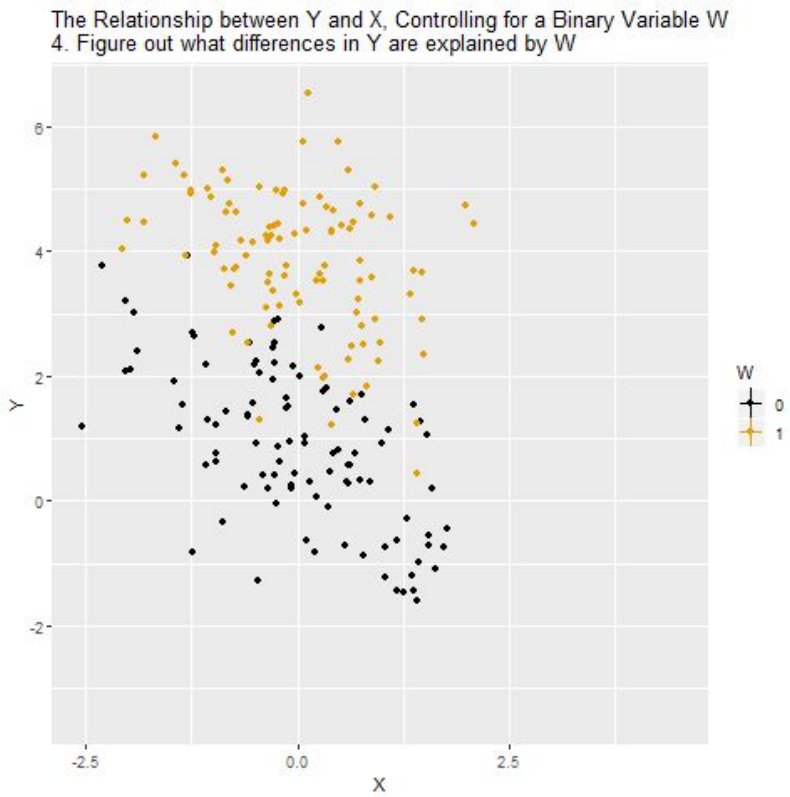


So to “control” for W, we subtract the averages of W that explain X and Y.

First, we subtract the average value of of W that best explains X…

https://nickchk.com/causalgraphs.html

Visualizing control

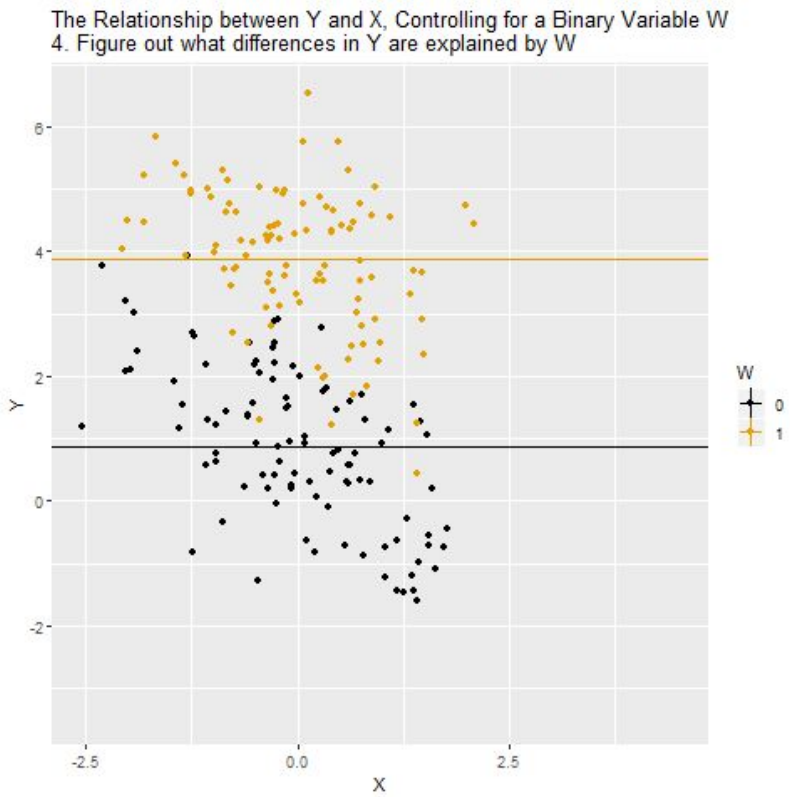


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

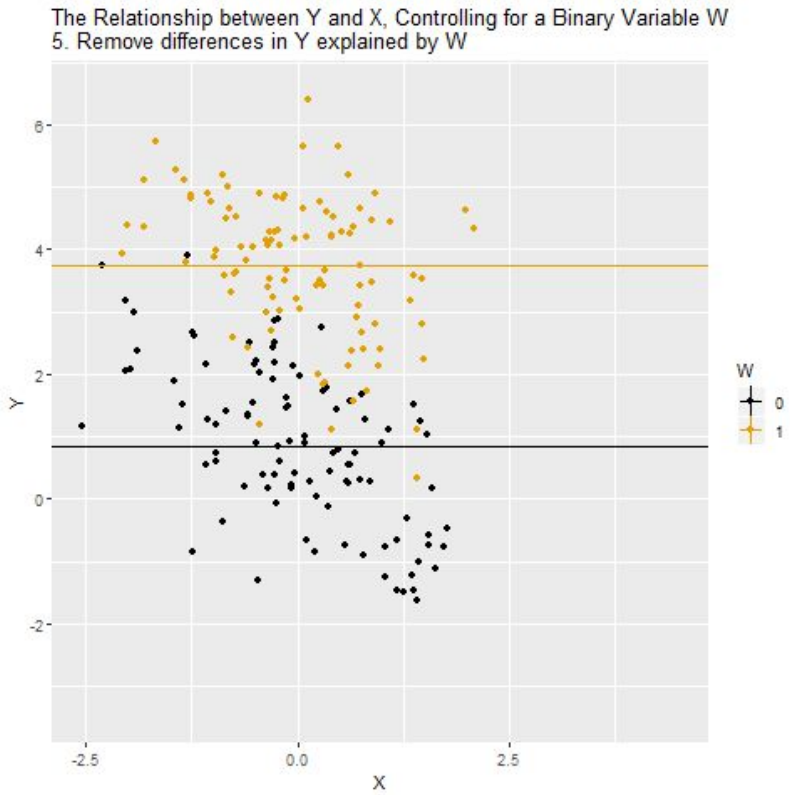


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

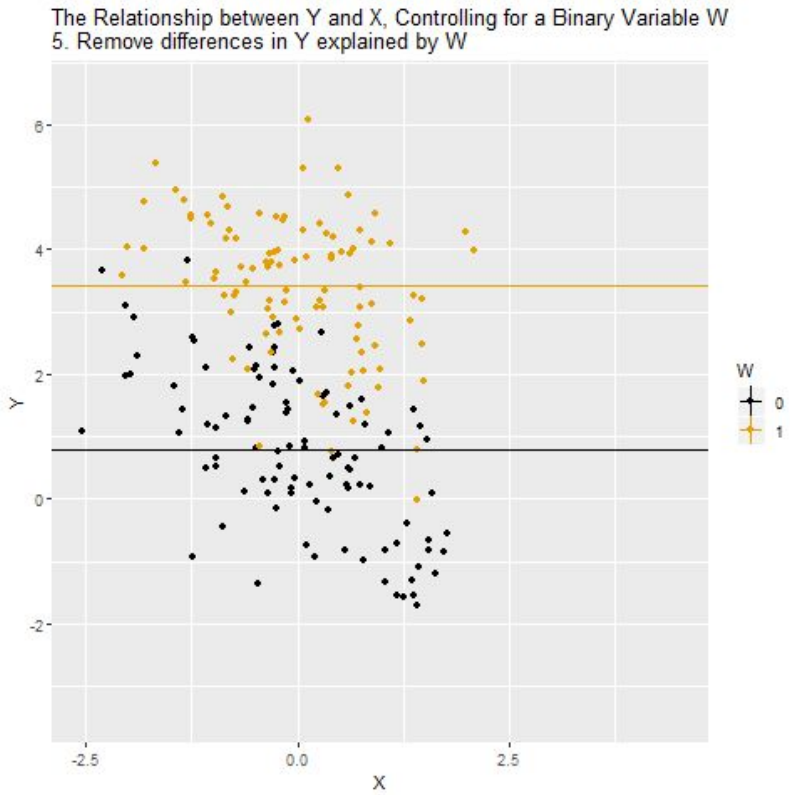


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

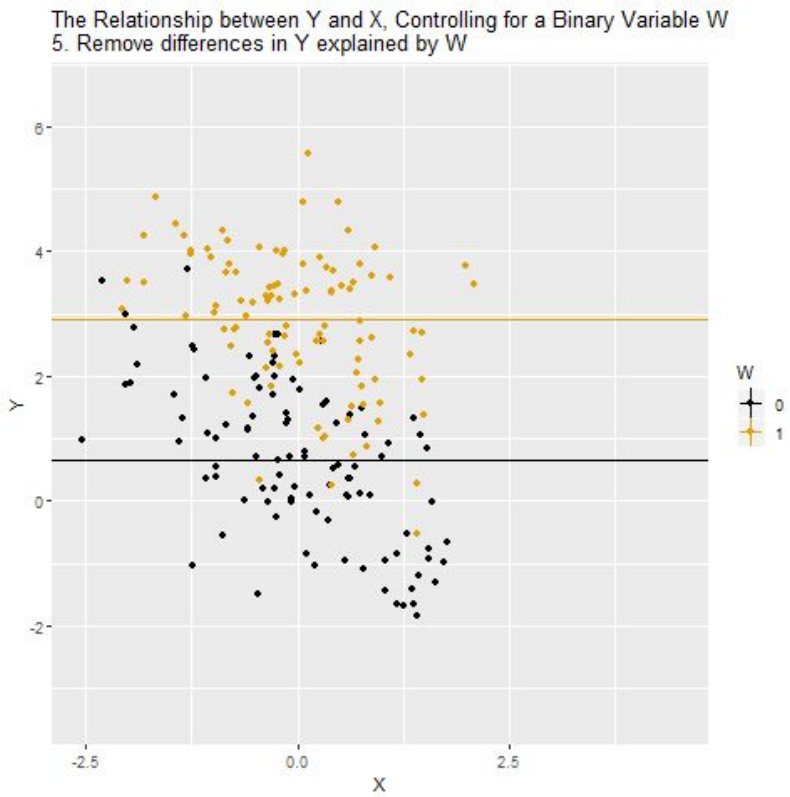


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

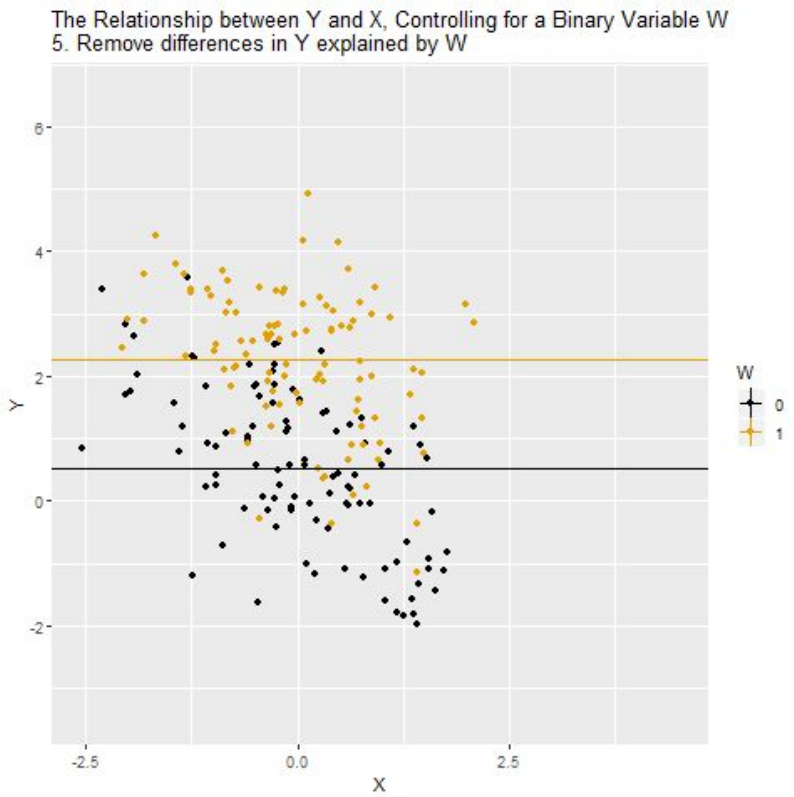


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

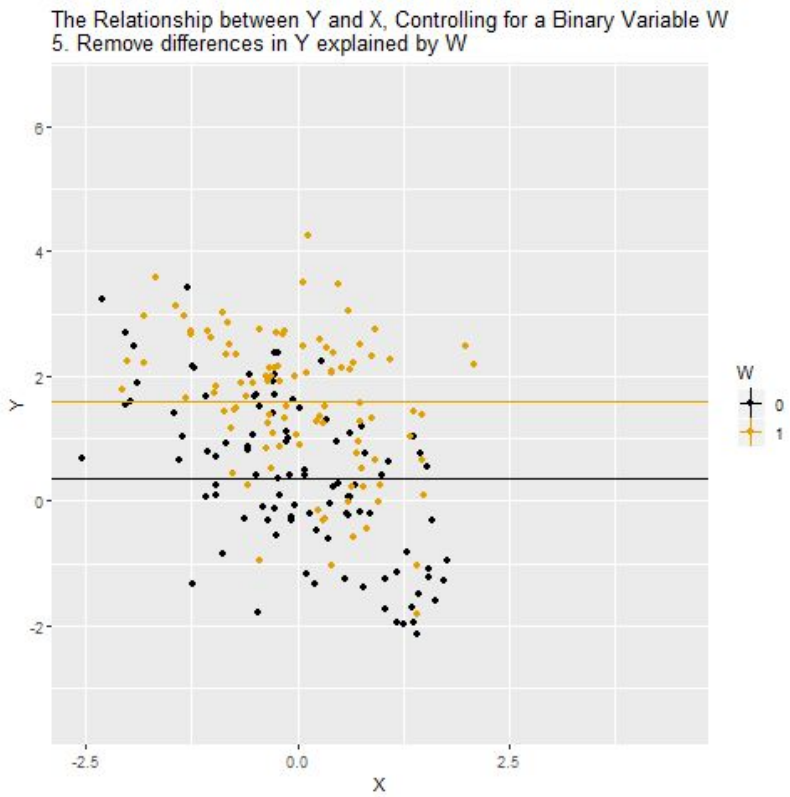


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

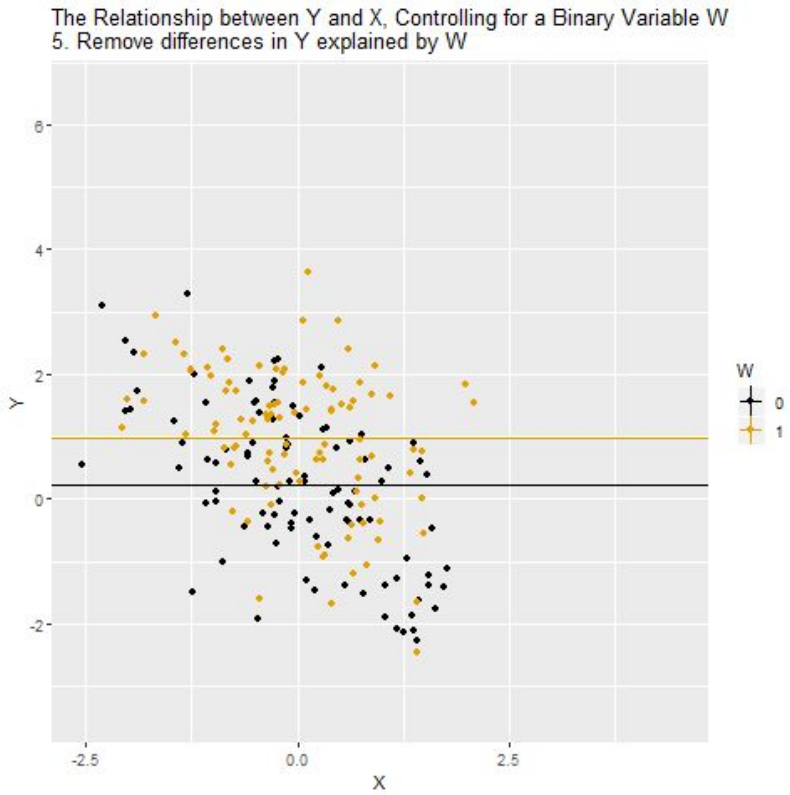


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

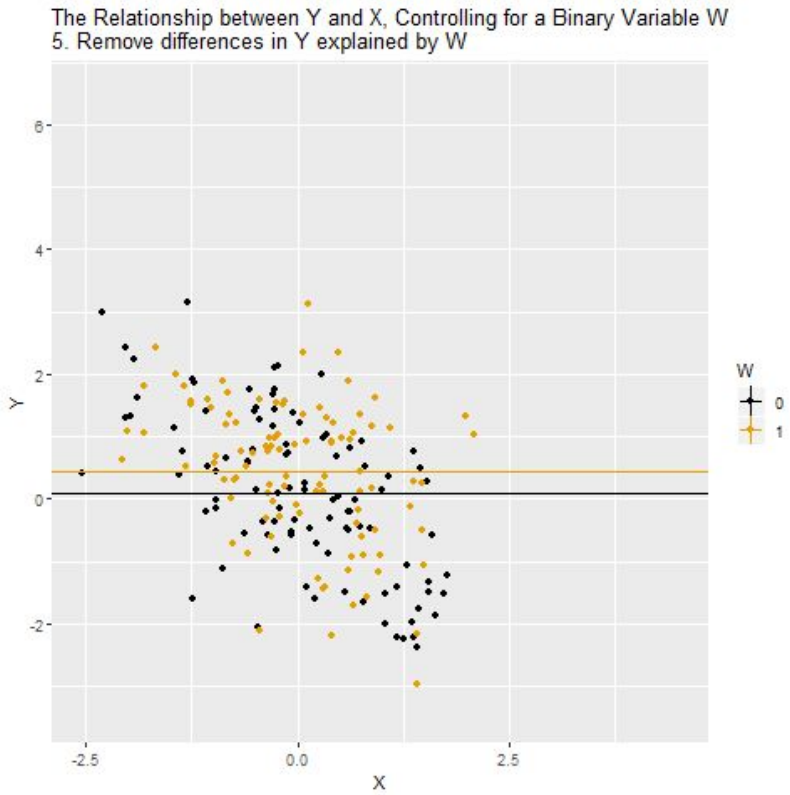


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

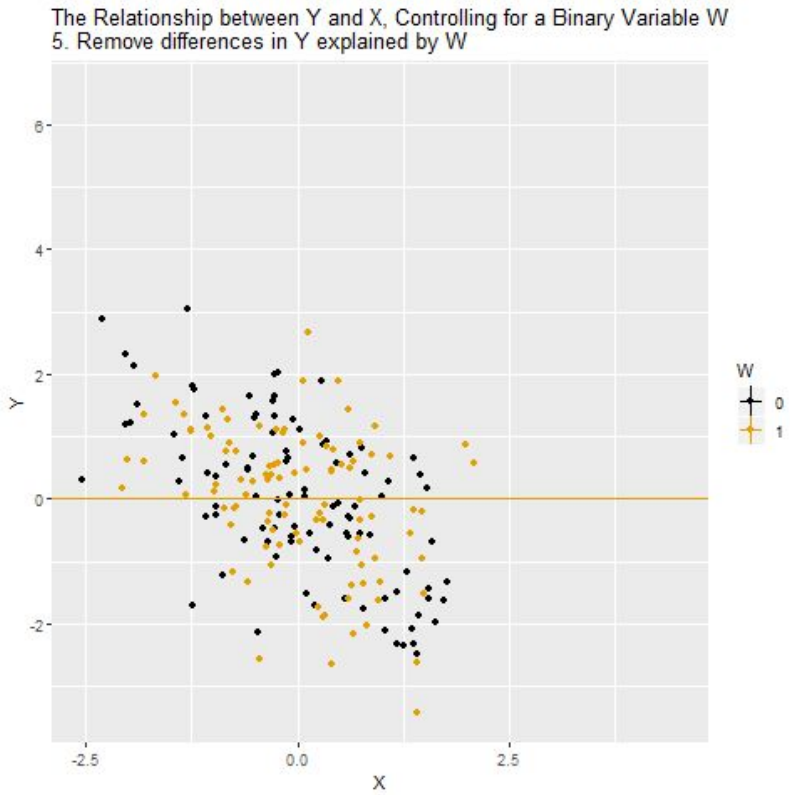


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control

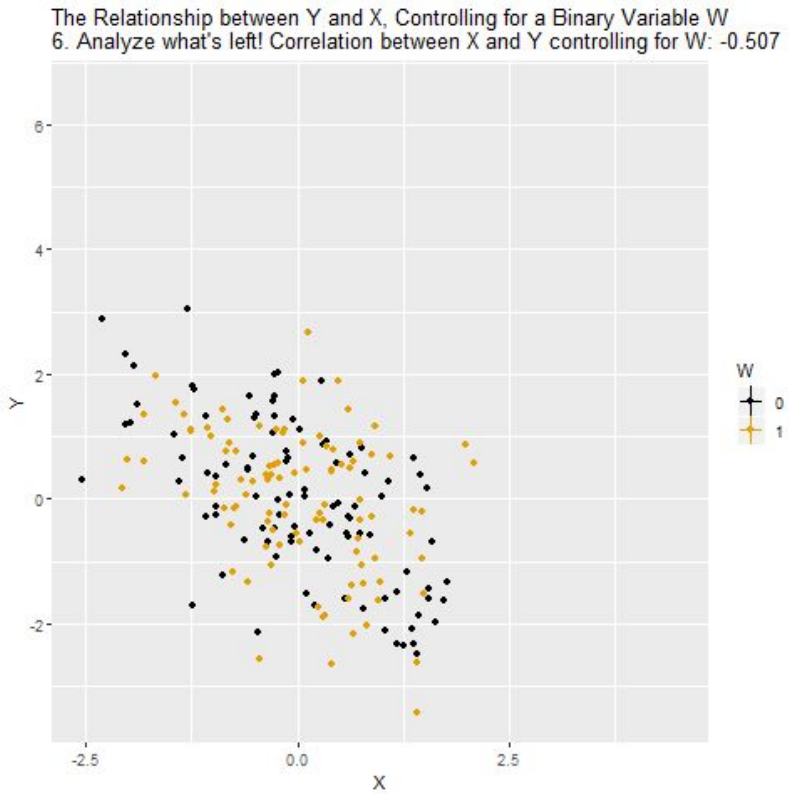


So to “control” for W, we subtract the averages of W that explain X and Y.

…Then, we subtract the average value of W that explains Y.

https://nickchk.com/causalgraphs.html

Visualizing control



After controlling for W, the correlation between X and Y is now stronger and it's *in the opposite direction* (that is, negative instead of positive) from when we weren't controlling for W!

https://nickchk.com/causalgraphs.html

Statistical significance

Statistical significance does *not* mean X is important for explaining Y.

Instead, it means that we are confident that the relationship between X and Y isn't just due to random chance based on the sample we drew from the population.

So if a relationship between X and Y is significant, we will re-sample cases from the population and continue to find a relationship a certain percentage of the time.

- $p < 0.05$: We will continue to find an effect of X on Y 95% of the time
- $p < 0.01$: We will continue to find an effect of X on Y 99% of the time
- $p < 0.001$: We will continue to find an effect of X on Y 99.9% of the time.

Again, greater statistical significance does *not* mean X has a larger impact on Y. It's about how confident we are that there is an actual relationship between X and Y in the "real world," *assuming our sample is randomly selected from the population of interest*.

Other (better?) measures of error

Standard error

- Usually in parenthesis next to each beta coefficient (0.27)
- Reflects the degree of variation in your estimated average “effect”.
- Larger standard error means cases vary more widely in their X-Y relationship
    - Large standard errors means we’re less sure we’ll find the same average effect if we conduct the study again.
    - If our standard errors are large enough, we might just be measuring random noise

Confidence interval

- Calculated based on standard error
    - For a 95% confidence interval, multiply the standard error by 1.96 (or round to 2) and add to the coefficient for the upper bound, then subtract from the coefficient for the lower bound.
    - This tells us that when we select new cases to use in our model, our average estimates will fall within that upper and lower bound 95% of the time.

- Let's say a state wants to improve turnout in an upcoming election.
- Election officials place television ads on major broadcast stations informing viewers about the election.
- However, levels of TV viewership is not the same across the state's towns: Some towns are in more rural areas where signal reception is bad, and other towns have younger populations where residents are more likely to be watching programs streaming on their laptop instead of broadcast stations.
- Officials also already know that towns with higher income levels vote at higher rates.
- After the election, officials test whether their ad had any effect through multivariate regression. The dependent variable is a town's level of turnout, and the dependent variables are (1) the percentage of people in a town who make more than the national median income and (2) the percentage of people who watch broadcast programs.

	Beta	**Std. error**	**95% CI**	**p**
% who make more than nat'l median income	3.2	(1.05)	[0.1 , 6.3]	<0.05*
% who watch broadcast TV	0.9	(0.05)	[0.8, 1.0]	<0.001***
Intercept	1.3	(0.75)	[-0.2, 2.8]	<0.1

1. What is the "effect" of a town's level of income on voter turnout? What's the "effect" of TV viewership?
2. Based on these results, what can we say has the biggest impact on voter turnout?
3. How certain are we that we would get the same results if the same referendum were held again?

Inspired by a well-known study by Gilens and Page (2014), researchers want to find out which group has the most influence on American politics: "Average" American citizens, economic elites, or organized interest groups. To find out, they collect bills that received roll-call votes in Congress between 1980 and 2010 to use as cases. Using this data, they run four separate OLS regression models, with the outcome variable in each model representing the number of legislative votes in favor of each bill.

- Model 1 measures the preferences of "average" Americans through the % of median-income citizens who support each bill, according to survey data:
  - $Y_{\text{leg. votes}} = a + bX_{\text{avg.support}}$
- Model 2 measures the preferences of economic elites through the percentage of upper-income citizens (top 10%) who support each bill:
  - $Y_{\text{leg. votes}} = a + bX_{\text{wealthy support}}$
- Model 3 measures the preferences of interest groups through a set of organized lobbying groups that have openly supported each bill:
  - $Y_{\text{leg. votes}} = a + bX_{\text{group support}}$
- Model 4 predicts the outcome on all three independent variables:
  - $Y_{\text{leg. votes}} = a + bX_{\text{avg.support}} + bX_{\text{wealthy support}} + bX_{\text{group support}}$

Table 3
Policy outcomes and the policy preferences of average citizens, economic elites, and interest groups

	Model 1	Model 2	Model 3	Model 4
Preferences of average citizens	.64 (.08)***	—	—	.03 (.08)
Preferences of economic elites	—	.81 (.08)***	—	.76 (.08)***
Alignment of interest groups	—	—	.59 (.09)***	.56 (.09)***
R-sq	.031	.049	.028	.074

***p<.001

Study Table 3 above and consider the following questions:

- What did the researchers find in each of the first three models?
- What happened in Model 4? What does it suggest about the factors impacting policymaking?
- Is the design convincing? Is there any way you might change it?