Ec140 - Statistical Inference and Regression

Fernando Hoces la Guardia
07/12/2022

Housekeeping

- Problem Set #2 due this Friday at 5pm on Gradescope.
- Section merge:
  - Section 102 and 103 will be taught at the time of 102 (9:30am MW @ Evans 3)
  - Section 106 and 108 will be taught at the time of 106 (2pm TTh @ Evans 9)

Today's Lecture

- Finish Statistical Inference
  - Confidence Intervals
  - P-Hacking
- Start the Regression Journey!
  - Regression as Matching - Part I

Confidence Intervals 1/4

- Confidence Intervals flips the question of statistical significance, and asks what are the set all possible values of μ (the true population value) that are consistent with the sample mean that we observe.

- To simplify notation, let's define $\hat{\mu} = \overline{Y_1} - \overline{Y_0}$. Then the expression for the t-statistic from last class, can be written as:

$$t(\mu_0) = \frac{\hat{\mu} - \mu_0}{SE(\hat{\mu})}$$

Confidence Intervals 2/4

- For confidence intervals, we fixed the value of t at some critical value, usually the (approximate) value of $|t_{5\%}| = 2$ (2 and -2) that correspond to the 5% convention of statistical significance, and ask which values of μ could take for our data to be compatible with null hypothesis (that we would not reject):

$$
\begin{aligned}
2 \geq \frac{\hat{\mu} - \mu}{SE(\hat{\mu})} \text{ and } -2 \leq \frac{\hat{\mu} - \mu}{SE(\hat{\mu})} \\
\Leftrightarrow 2 \times SE(\hat{\mu}) \geq \hat{\mu} - \mu \text{ and } -2 \times SE(\hat{\mu}) \leq \hat{\mu} - \mu \\
\Leftrightarrow \mu \geq \hat{\mu} - 2 \times SE(\hat{\mu}) \text{ and } \mu \leq \hat{\mu} + 2 \times SE(\hat{\mu})
\end{aligned}
$$

Hence, the interval:

$$[\hat{\mu} - 2 \times SE(\hat{\mu}), \hat{\mu} + 2 \times SE(\hat{\mu}]$$

will contain the true value of μ 95% of the times.

Confidence Intervals 3/5

So we have a confidence interval for μ, *e.g.*, $[0.324, 0.588]$.

What does it mean?

Informally: The confidence interval gives us a region (interval) in which we can place some trust (confidence) for containing the parameter.

More formally: If repeatedly sample from our population and construct confidence intervals for each of these samples, $X\%$ percent of our intervals (*e.g.*, 95%) will contain the population parameter *somewhere in the interval*.

Confidence Intervals 4/5

- But this concept of 95% of the samples is pretty abstract. As we observe just one sample (the one we have in our data). In order to better grasp this concept, we will do a simulation.

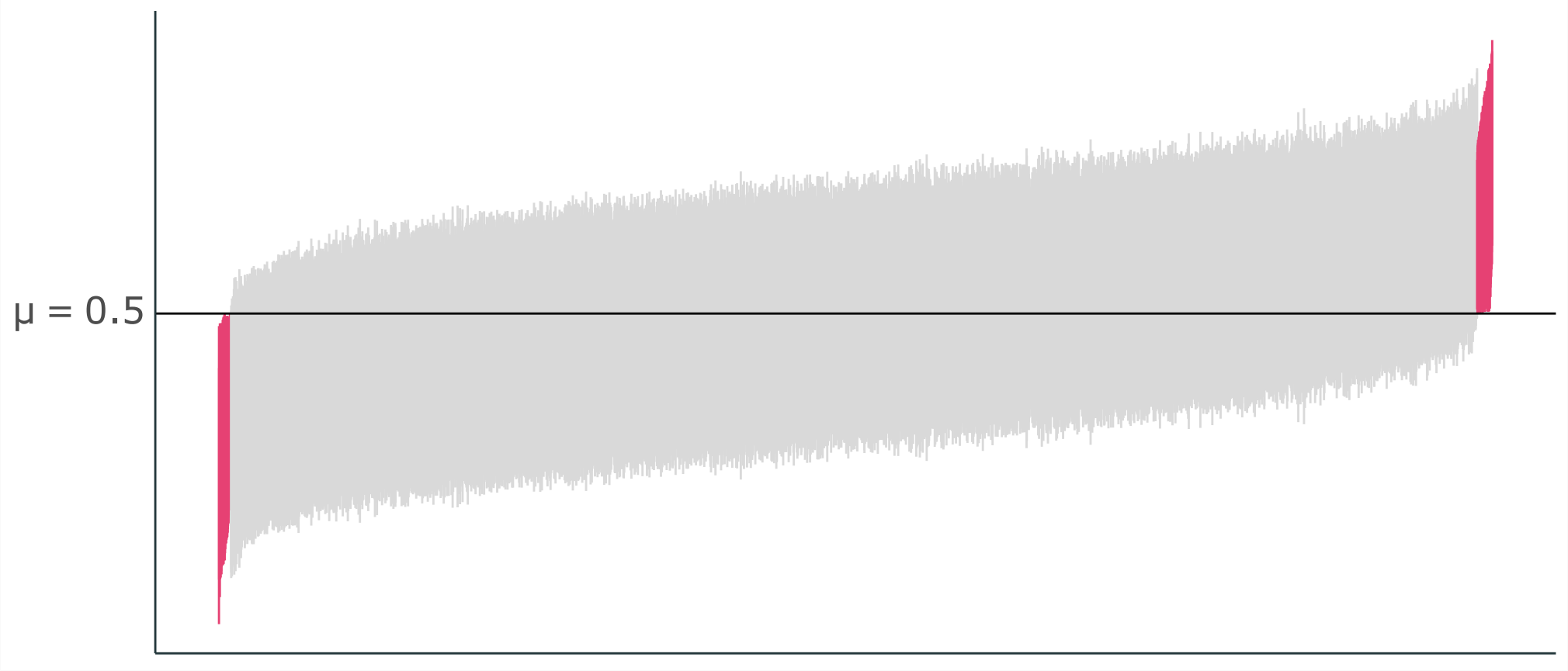
- Let's drew 10,000 samples (each of size $n = 30$) from a population and where $E(Y|D = 1) - E(Y|D = 0) = \mu = 0.5$ (we are not focusing now on whether μ is causal or not).

- One sample of $n = 30$ from the population for values of Y and D could yield an estimate: $\overline{Y_1} - \overline{Y_0} = \hat{\mu} = 0.456$ with a confidence interval $[0.324, 0.588]$.

- This is one sample (what we usually see). But in a simulation (where we know the data generating process) we can repeat this as many times as we want. So, let's draw 10,000 of these "worlds" and compute its $\hat{\mu}$ and CI for each.

Confidence Intervals 5/6

This amazing figure (from Ed Rubin's class) represents all those CI. As you can see 97.8% of 95% confidences contain the true parameter of $\mu = 0.5$.



Confidence Intervals: Warning

- After seeing so much CLT and how it applies to the t-statistic, at some point in the future you might feel the temptation to think “it must be the case that the true population parameter is normally distributed in the confidence interval, hence its more likely to be in the middle than in the corners”.

- There are two errors in that way of thinking:

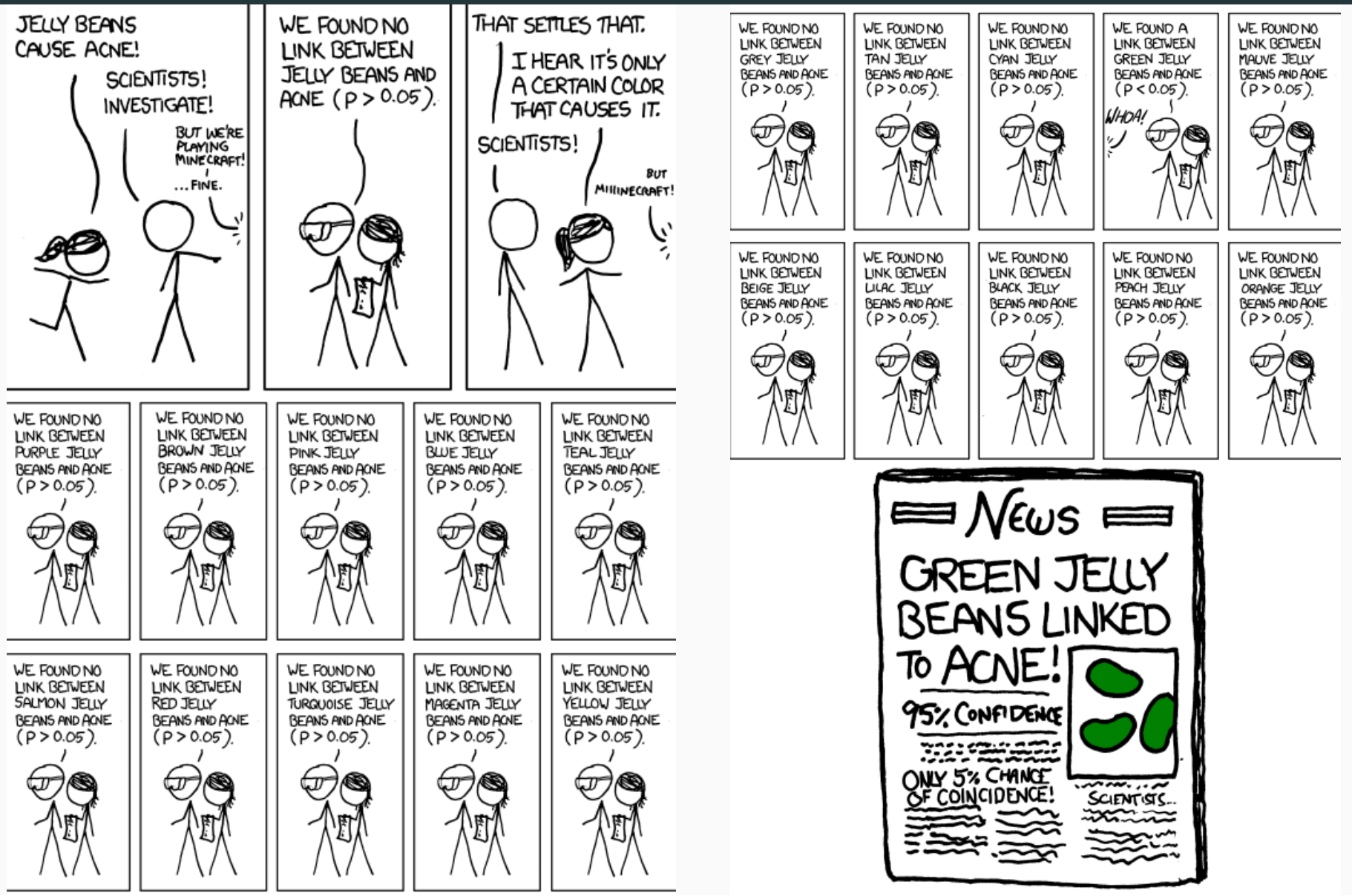
    1. The true parameter does not have a distribution (remember it's a fix quantity)
    2. Even if you were to focus on something like the “distribution of likely truths” (whatever that might be), we do not know if it is the sum of i.i.d RVs, hence nothing tells us that the CLT applies here.

Absent any additional information, our best guess is that the true parameter is uniformly distributed in this range.

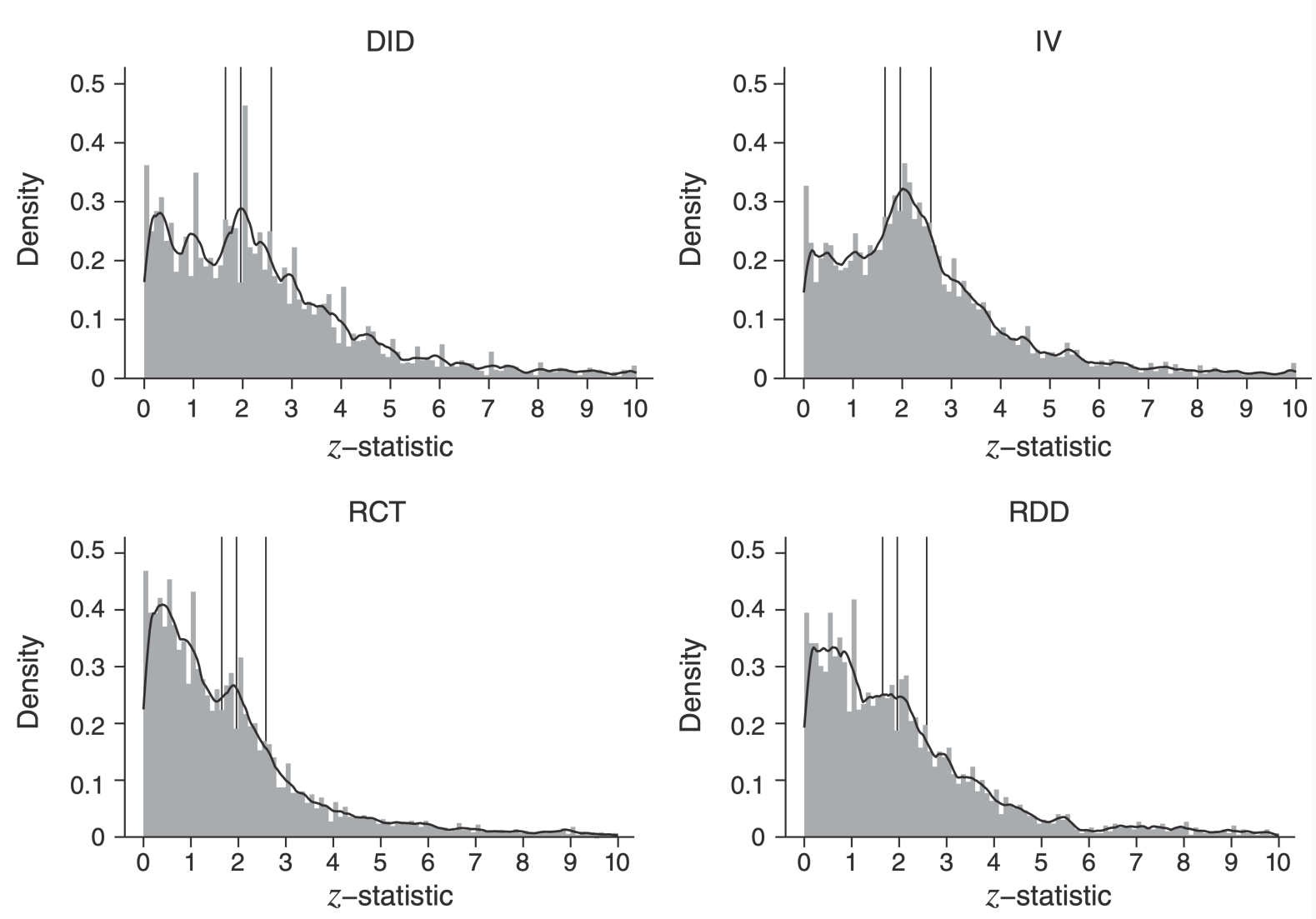
P-Hacking

- **Definition**: flexibility in data analysis allows portrayal of *almost anything* as below an arbitrary p-value threshold.
- Statistical significance loses its meaning.
- Also called *specification-searching*, *fishing*, *researcher degrees of freedom*, or *data-mining*.
- Not something only evil people do. It's subconscious, or simply built into how we have practice statistical inference (until very recently).

P-Hacking in One (Fictional) Picture: XQCD



P-Hacking in One (Real) Picture: Economic Papers



(Brodeur et. al., 2020)

P-hacking Solutions

- Registrations
- Pre-Analysis Plans
- Computational Reproducibility

(Check out bitss.org if you want to learn more about this!)

Regression: The Next 4 (to 6) Lectures Ahead

- Regression as Matching on Groups (Part I today). Ch2 of MM up to page 68 (not included).
- Regression as Conditional Expectation and Line Fitting. Ch2 of MM, Appendix + others.
- Multiple Regression and Omitted Variable Bias. Ch2 of MM pages 68-79.
- Regression Inference, Binary Variables and Logarithms. Ch2 of MM, Appendix + others.

Regression: The Next 4 (to 6) Lectures Ahead

- **Regression as Matching on Groups (Part I today). Ch2 of MM up to page 68 (not included).**
- Regression as Conditional Expectation and Line Fitting. Ch2 of MM, Appendix + others.
- Multiple Regression and Omitted Variable Bias. Ch2 of MM pages 68-79.
- Regression Inference, Binary Variables and Logarithms. Ch2 of MM, Appendix + others.

Before we Begin: A Comment on the Tone of MM

- MM is a huge contribution to econometrics by making it more accessible and concrete. It can help to make economics more diverse by clearly presenting the value of these topics without the barrier of a strong background in math.
- However, I fear that some of its tone is still highly elitist and more likely to appeal to men than women and underrepresented groups.
- That tone is very noticeable in a series of videos produced to supplement the book, but it can also be found in the text of the book. I will try to flag those instances and propose alternative interpretations.
- I didn't question this tone 10 years ago!
- Let's try to focus on the great parts of the book, and be open to identify and discuss some of its limitations (in a sense, is a good exercise to detect BS, even among great teachers).

Regression as Matching on Groups

What to do if we Cannot Run an Experiment?

- Forget about unobservables for a minute and assume that we there is selection bias only in observables (e.g. age, income, others).
- One way to approach this would be to look at the differences within each group (e.g. ages 40-65 with incomes 40-80k) and interpret those differences as the result of an RCT within that group (or cell). This is what regression does.
- Regression is the second **research design tool** we review in this course.
- Regression alone is rarely used to justify causality. Because it's hard to believe that there is no selection on unobservables.

Individual/Policy Choice Issue: Private or Public College? 1/2

- We explore the concept of regression starting from our second real-life policy (and personal) decision based on causal evidence.
- Average yearly tuition for a private four-year college in US (2012): $29,000
- Average yearly tuition for a public four-year college in US (2012): $9,000
- Is it worth spending (or subsidizing) this $80,000 difference (20k x 4) so you (or more students overall) can go to elite private colleges?
- One dimension to assess this question is the causal effect of college on earnings. And this will be the center of this example, but first we need to talk about other possible dimensions (different from earnings).

Individual/Policy Choice Issue: Private or Public College? 2/2

- MM suggests that private ed might be better than public ed in many ways: “smaller classes, better facilities, more distinguished professors, smarter students”).
- Can you identify which part of that statement is true, and which BS? (here is a tip)
- Can you suggest some ways in which a public education is better than a private (here is another tip)?
- Now let’s go back and focus on the earnings dimension.

Simple Difference in Groups for “Private” Treatment 1/3

- First, let’s define the treatment for this setting as having attended a private four-year college, and the control as having attended a public four-year college.

- Now, you are told that a simple difference in groups shows that student from private institutions earn between 14% and 21% more than students from public universities:

$$\mathbb{E}(\textbf{Difference in group means}) =$$
$$\mathbb{E}(Y_i|D_i=1) - \mathbb{E}(Y_i|D_i=0) = \kappa + \underbrace{\mathbb{E}(Y_{i0}|D_i=1) - \mathbb{E}(Y_{i0}|D_i=0)}_{\textbf{Selection bias}}$$

- How should we read the terms $\mathbb{E}(Y_{i0}|D_i=1), \mathbb{E}(Y_{i0}|D_i=0)$ in this case?

Simple Difference in Groups for “Private” Treatment 2/3

- They are the expected earnings for treatment and control in the counterfactual world where they did not receive a private college, and did receive a public college education.
- MM suggests some reasons why these two could be different: elite private students tend to have higher GPAs, SATs, more motivation, plus other skills and talents, than elite public college students.
- Can you identify which part of that statement is true, and which BS? Can you think of additional variables (in addition to “motivation”, “smarts”, and “skills and talents”) that could also contribute to selection bias?

Simple Difference in Groups for “Private” Treatment 3/3

- How about receiving intense tutoring? Attending an elite high school? connections? Or having parents with knowledge of the system?

- All of the above play a similar role in selection bias, but unlike the MM interpretation, they do not suggest that private students are inherently better than public students (I am not suggesting that the latter should replace the former, only complement).

- To identify this causal effect, one proposal is to use data from applications and choices between elite private and elite public colleges. The **key underlying assumption** is that at some point luck (or lack thereof) starts playing a role in the final assignment of the treatment.

Intuition of Controlling For Observables

- Assume that all that matters is SAT. If we compare two individuals with the same SAT: Harvey and Uma both with 1400, but Harvey choose private and the Uma choose public, then the comparison would hold other things equal (by assumption).

- Now relax that assumption: we know that women make, on average, less than men, what if the difference we observe between Harvey and Uma is caused by gender (discrimination or something else) and not by type of school?

- Repeat thought experiment, but now for individuals with the same SAT and gender. This is the logic of regression. We match on characteristics, also called a matching estimator, where we hold fixed, or control for, a set of characteristics.

Real Life Example: Regression and Causal Effects of Private College

- Dale and Krueger (2002) analyze data from college applications, admissions and final choice for individuals that apply

- The key idea of the paper is that instead of measuring all characteristics where treatment and control will differ, they argue that they have a measure that closely summarizes all those unobserved characteristics: college application and college decisions.

- Supposedly application information is a good proxy for motivation, and acceptance is a good proxy of capacity. In my view, this could have been a good argument 20 years ago, but not today (Harvard's Legacy+Athlete bonus, college admissions scandal, additional evidence). For the purpose of the example let's assume that these are good proxies for all other things.

Intuition Behind Control Strategy

TABLE 2.1
The college matching matrix

Applicant group	Student	Private			Public			1996 earnings
		Ivy	Leafy	Smart	All State	Tall State	Altered State	
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Intuition Behind Control Strategy: Notes 1/2

- Grouped by application and admission decision at the university level.
- Within a group there can be variation in final decisions.
- Within group variation for group A is negative (-5k). Group B has a positive difference (30k). There are many combinations of such university-application-decisions-groups.
- Group C and D have all private and all public respectively, so nothing to learn here in terms of private-public diffs (all treatment or all control).

TABLE 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Intuition Behind Control Strategy: Notes 2/2

- Simple average is a good estimate of causal effects (given out assumptions): $12,500, also another good estimate is the weighted average: 9,000. Giving more weight to more data makes more efficient use of information, leading to a more precise estimate.
- Comparing within groups we can argue that we are holding Y_0 (potential earnings if no treatment) constant.
- Simple group difference would estimate 19.5K (all) or 20K (just A and B) diff.
- Selection bias emerges when comparing across, instead of within, groups. Group A was much wealthier (107K) than group B (45K), and also had more students in private schools.

TABLE 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Ready to Understand Regressions! 1/3

- Think of regression as an automated matcher: regression estimates are weighted averages of multiple matched comparisons (similar to groups A and B before).

- Regression ingredients:

  - Dependent variable, or outcome variable. In our example: earnings in 20 years after graduation.
  - Treatment variable, in our case, a binary variable indicating 1 for private and 0 for public.
  - A set of control variables, in our example variables that identify sets of schools to which students apply and were admitted too.
  - Observations: C&D are excluded from our sample because they do not provide information regarding the relevant comparison we want to make.

Ready to Understand Regressions! 2/3

Regression equation:

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i$$

- The difference between A and P is conceptual, not formal. The research design justifies the role each variable plays. In our case, P plays a primary role, while A is secondary (we don't care much if it's actually measuring a causal relationship).
- Intercept/constant, α
- Causal effect of treatment β, and
- The effect of being a group A student, γ. (not relevant to us)
- The residual, e_i, defined as the difference between observed (Y_i) and fitted values $(\hat{Y}_i)$.

Ready to Understand Regressions! 3/3

- What regression does (more detail on this next lecture): chooses α, β and γ, to minimize the sum of squared residuals. Executing this minimization is often called "Estimating" or "Running" a regression. We will explore a little of theory, and how to run regressions in a little. But first, let's focus on the result of running a regression.

- Simple toy example (from table 2.1): β of 10,000 shows that the regression estimate is somewhere in between the simple group comparison (12.5k) and weighted group comparison (9K).

Acknowledgments

- Ed Rubin's Undergraduate Econometrics II
- XQCD
- BITSS
- ScPoEconometrics
- XQCD
- MM
- Matt Hollian