Causality and Experiments in Econometrics

Charlie Gibbons
University of California, Berkeley
Economics 140

Summer 2011

Notes

Outline

Notes

Two types of studies

There are two types of studies:

- Descriptive studies
  Find patterns and relationships
  Less concerned about omitted variables
  Like interpretation of our "assumption-free" dummy variable model "Do men make more than women?"
- Causal inference
  Explain patterns and relationships
  Very concerned about omitted variables
  "Does being a woman cause a person's wage to be lower?"

Notes

Getting at causality

Recall our definition of the causal effect of treatment for observation i: $y_i(1) - y_i(0)$.

Since this difference is fundamentally unobservable, we could compare the impact of treatment on two people who are similar in every respect that is important for the outcome y_i (*e.g.*, they are similar on every observable aspect and there is no unobserved factors that influence y).

We can broaden this to look at a group of people that received treatment compared to a group that did not, where each person has a doppleganger and the unobserved factors are uncorrelated with the included ones (especially treatment itself).

If we can't find matched groups, we can model how the groups differ using observable factors. We still assume that any omitted variables are uncorrelated with observable traits.

Notes

Pre-treatment similarity

Thinking of the matched group idea, we want both groups to be identical in terms of observable and unobservable factors at the start of the study (*e.g.*, *pre-treatment*).

We want to see how they differ at the end.

Notes

Post-treatment bias

We don't necessarily want them to be identical in the middle of the study. Maybe treatment causes changes in other *mediating* variables that impact treatment. These changes should be counted as part of the causal effect. We don't want to control for these intermediate outcomes, as this leads to *post-treatment bias*.

Note the difference here with OVB: we don't want to count the impact of stuff that's just correlated with treatment *ex ante*; we do want to count the impact of intermediate outcomes that treatment causes.

Notes

Asking a good question

The first step in performing econometric analysis is to determine the **precise** question that you want to answer.
Bad: Do employers discriminate against blacks?
Better: Does being black cause a worker to have a lower wage than a white worker?

Notes

Example: Discrimination

"Does being black cause a worker to have a lower wage than a white worker?"

Does this question actually make sense?

Causality requires that a single observation has the potential to receive *and* not receive treatment.

Does it make sense to think of a person's wage when he is black and the wage *for that same person* when he is white?

Notes

Discrimination and post-treatment bias

Suppose that we can justify "being black" as a treatment.

When does treatment occur?

What pre-treatment variables can we control for? What are intermediate outcomes that we don't want to control for?

What would it mean to compare similar groups?

Notes

Randomized trials

The best approach for all quantitative scientific research is the *randomized experiment*, where a group of people is randomly divided into treatment and control groups.

This ensures that no factors are correlated with treatment (in expectation).

Notes

Experimental benchmark

After defining a question, you should think of an experiment that could answer that question if you were a dictator and amoral. This is called the *experimental benchmark*.

Example: What experiment could you do to find the impact of a college education on wages?

If we only randomize among graduating high school students, what effect are we really measuring?
The impact of a college education on wages *for people that can complete high school.*

Could randomize among all 18 year-olds instead.

Notes

What is a "precise" question?

Our question wasn't precise enough to choose the appropriate experiment—we need to reformulate it.

A precise question is one that gives enough specificity so that we can determine **the** experiment that would answer it.

Notes

Comparison to benchmark

Once we have a precise question and have thought of the experimental benchmark, we want to ask how our actual study compares to this benchmark.

How is it similar?

How is it different?

What are the implications for bias?

Notes

A discrimination study

Suppose that we want to compare individuals that are entirely similar going onto the job market, except some are black and some are white.

This doesn't fit our standard causal model:

- A person doesn't have a positive chance of being black and a chance of being white.
- Treatment is applied at birth (actually, before), so holding intermediate outcomes doesn't make sense.

We need to find a *proxy* for being black that can plausibly

- Be applied to either blacks and whites (but more likely for blacks and less likely for whites) and
- Randomized upon entering the job market.

Notes

Race-specific names

Top race-specific names for babies born in Massachusetts in 1970–1986

White girls: Emily, Anne, Jill, Allison, Sarah, Meredith, Laurie, Carrie, Kristen (3.8%)

Black girls: Aisha, Keisha, Tamika, Lakisha, Tanisha, Latoya, Kenya, Latonya, Ebony (7.1%)

White boys: Neil, Geoffrey, Brett, Brendan, Greg, Todd, Matthew, Jay, Brad (1.7%)

Black boys: Rasheed, Tremayne, Kareem, Darnell, Tyrone, Jamal, Hakim, Leroy, Jermaine (3.1%)

Notes

Reference

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*. 94(4): 991–1013.

Notes

The experiment

The authors find employers advertising openings in the *Boston Globe* and the *Chicago Tribune* and create two high-quality and two-low quality resumes that are relevant for those jobs.

Within each quality level, one is randomly chosen to have a black-sounding name and the other has a white-sounding name.

They see how many e-mail or phone responses each resume generates.

Notes

Table 1
Mean Call-Back Rates By Racial Soundingness of Names [a]

	Call-Back Rate for White Names	*Call-Back Rate for African American Names*	*Ratio*	*Difference (p-value)*
Sample:				
All sent resumes	**10.06%** [2445]	**6.70%** [2445]	**1.50**	**3.35%** (.0000)
Chicago	**8.61%** [1359]	**5.81%** [1359]	**1.48**	**2.80%** (.0024)
Boston	**11.88%** [1086]	**7.83%** [1086]	**1.52**	**4.05%** (.0008)
Females	**10.33%** [1868]	**6.87%** [1893]	**1.50**	**3.46%** (.0001)
Females in administrative jobs	**10.93%** [1363]	**6.81%** [1364]	**1.60**	**4.12%** (.0001)
Females in sales jobs	**8.71%** [505]	**6.99%** [529]	**1.25**	**1.72%** (.1520)
Males	**9.19%** [577]	**6.16%** [552]	**1.49**	**3.03%** (.0283)

Notes

Table 2
Distribution of Call-Backs By Employment Ad [a]

Equal Treatment:	*No Call-back*	*1W+1B*	*2W+2B*
87.37%	82.56%	3.46%	1.35%
[1162]	[1098]	[46]	[18]
Whites Favored (WF):	*1W+0B*	*2W+0B*	*2W+1B*
8.87%	5.93%	1.50%	1.43%
[118]	[79]	[20]	[19]
African Americans Favored (BF):	*1B+0W*	*2B+0W*	*2B+1W*
3.76%	2.78%	0.45%	0.53%
[50]	[37]	[6]	[7]
Ho: WF=BF			
p=.0000			

Notes

Table 4
Average Call-Back Rates
By Racial Soundingness of Names and Resume Quality [a]

Panel A: Subjective Measure of Quality

	Low	High	Ratio	Difference (p-value)
White Names	**8.80%**	**11.31%**	**1.29**	**2.51%**
	[1216]	[1229]		(.0391)
African American Names	**6.41%**	**6.99%**	**1.09**	**0.58%**
	[1216]	[1229]		(.5644)

Notes

Table 5
Effect of Resume Characteristics on Likelihood of Call-Back [a]

Dependent Variable: Call-Back Dummy

Sample:	**All Resumes**	**White Names**	**African American Names**
Years of experience (*10)	.07	.13	.02
	(.03)	(.04)	(.03)
Years of experience^2 (*100)	-.02	-.04	-.00
	(.01)	(.02)	(.01)
Volunteering? (Y=1)	-.01	-.01	.00
	(.01)	(.02)	(.01)
Military experience? (Y=1)	-.00	.01	-.01
	(.02)	(.02)	(.02)
Email? (Y=1)	.02	.03	.00
	(.01)	(.01)	(.01)
Employment holes? (Y=1)	.02	.03	.01
	(.01)	(.02)	(.01)
Work in school? (Y=1)	.01	.02	-.00
	(.01)	(.01)	(.01)
Honors? (Y=1)	.05	.07	.02
	(.02)	(.03)	(.02)
Computer skills? (Y=1)	-.02	-.03	-.00
	(.01)	(.02)	(.01)
Special skills? (Y=1)	.05	.07	.04
	(.01)	(.02)	(.01)
Ho: Resume characteristics effects are all zero (p-value)	55.73 (.0000)	59.83 (.0000)	20.78 (.0227)
Standard deviation of predicted call-back	.047	.064	.034
Sample size	4890	2445	2445

Notes

Re. Table 5

Why does the previous table give the standard deviation of the predicted call-back?

It gives a sense of how big the coefficients are.

Notes

Interpretation

What question does this study answer?
What is the causal effect of a black-sounding name on interview call-back rates relative to white-sounding names?

Is this what we care about?
No, we really care about

- All blacks relative to all whites and
- Labor market outcomes (employment, salaries).

Notes

Alternative story

Rather than interpret these findings as evidence for discrimination against blacks, what is another interpretation?

Probability that the mother of kids with white-sounding names graduated from high school: 92%

Probability that the mother of kids with black-sounding names graduated from high school: 63%

Employers may be discriminating against people from low socioeconomic status backgrounds.

Notes

Unbiased estimates

The great feature of randomized experiments is that we don't need any other predictors in the model to get unbiased results—treatment is uncorrelated with everything in the error term.

But should we bring the observable characteristics in the error term into the model?

If we add more predictors, how do we expected our estimate of the causal effect to change?
Not at all, since it's already unbiased and uncorrelated with the added characteristics.

Can we interpret the coefficients on the added predictors?
No, the predictors might be correlated with the error term and the coefficients may be biased.

Notes

Reducing standard errors

Recall that the standard error is

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_k\right) = \frac{\sigma^2}{(1 - R_k^2)(N-1)\widehat{\mathrm{Var}}(x_k)}.$$

- $(N-1)\mathrm{Var}(x_k)$ doesn't change when we add more predictors.
- R_k^2 measures how well we can predict treatment using the other predictors. Since these factors should be uncorrelated, this should be near 0 no matter what we add.
- If we add important predictors, we can lower our $\hat{\sigma}^2$.

Hence, our standard errors fall when we add more relevant predictors.

Notes

Notes