Session 9: Randomized controlled trials

MGT 581 | Introduction to econometrics

Michaël Aklin

PASU Lab | EPFL

Last time...

- Threats to inference
- Directed acyclic graphs

Today:

- Randomized controlled trial
- Design
- Statistical power

Readings:

- Stock and Watson (2011), ch 13.1, 13.2
- Angrist and Pischke (2008) ch 2

Experimental methods

- **DAGs** show us what we should **condition** on (control/adjust for)
- Problem: we often don't know what to condition on! How can you build a DAG with high levels of trust?
- Solution: use research designs that by-pass the need to know the DAG
- Idea: can you **control** or **know** the origin of variation in **treatment assignment**
- **Experimental methods**: treatment is randomized
- **Quasi-experimental methods**: treatment is *as if* randomized

Randomized controlled trials

- RCTs: type of **randomization-based inference**
- Backstory: Ronald Fisher's *tea* test (Fisher 1935)
- Imagine having to make 4 choices $\{$Tea, Milk$\}$, with $Pr(\text{Tea}) = Pr(\text{Milk}) = 0.5$.
- Null: people don't know difference (= pick at random)
- What is the probability of making 4 correct choices in a row by luck?
- $p = 0.5^4 = 0.065$
- That's a p value! Pr of getting such data given the null

Benefits of randomization

$$
\begin{aligned}
\text{Naïve difference} &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
&= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] \\
&+ E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
&= E[Y_i(1) - Y_i(0)|D_i = 1] \\
&+ E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
&\equiv \text{ATT} + \text{Selection (baseline) bias}
\end{aligned}
$$

Question: how to get rid of selection bias?

- Randomize D_i: makes D independent from $Y(0)$, $Y(1)$.
- We write: $D \perp Y(0), Y(1)$. We call it the **ignorability assumption** (Rosenbaum and Rubin 1983)
- What $\cdot \perp \cdot$ means: $E[Y(0)|D] = E[Y(0)]$ and $E[Y(1)|D] = E[Y(1)]$
- Then we can simplify:

$$\begin{aligned}
\text{Naïve difference} &= E[Y_i(1) - Y_i(0)|D_i = 1] \\
&\quad + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
&= E[Y_i(1) - Y_i(0)|D_i = 1] \\
&= E[Y_i(1) - Y_i(0)] \\
&\equiv \text{Average treatment effect}
\end{aligned}$$

- Can be estimated with OLS or with $\overline{Y(1)} - \overline{Y(0)}$:

$$\widehat{\text{ATE}} = \frac{1}{N_1} \sum_{D=1} Y_i - \frac{1}{N_0} \sum_{D=0} Y_i$$

- RCT: removes backdoor relations between D and Y
- Is also clear from Frisch-Waugh: $\hat{\beta}$ converges to zero by construction in:

$$D_i = \alpha + \beta \mathbf{X}_i + \mu_i$$

- No need to **know** the data-generating process to remove confounding
- ATE can be estimated with a two-sample t-test (!)...
- ... but you might as well estimate a linear model with pre-treatment confounders (why?)

RCT in practice

1. Define population of interest, $\mathcal{U}$, and unit of analysis i.
2. Define treatment of interest, D.
   - Ideally: well-defined, not bundled.
   - Can be **direct** treatment or **indirect** (encouragement)
3. Define outcome(s) of interest, Y.
4. Build hypothesis, connecting $D \rightarrow Y$, with H_0, H_1.
5. Identify how you will measure D, Y (survey? administrative data? direct observation?)

Example

- RCT on effect of new website layout on consumer engagement ("A/B test")
- Treatment: updated website
  - Note: 'bundled' treatment (not just about one change)
- Outcomes?
  - Probability of buying a good on the website
  - Probability of returning on the website within 7 days
  - Probability of signing up for newsletter
- H_0: new website generates no difference in sales volumnes ($H_0 : \beta = 0; H_1 : \beta \neq 0$)
- Measurement: website will be randomized for each visitor; each visitor connected to a sale (or not)

5. Conduct power analysis

- Idea: find out how many observations you need to have a reasonable chance of detecting an effect
- Or: what kind of effect can you detect for a given sample size
- Useful to design a good RCT.
- If sample too small: too hard to detect an effect, null result won't be informative. Study is **underpowered**.
- If sample too large: too expensive.
- Starting point: **minimum detectable effect** (MDE): smallest effect that, if true, has an x% probability of generating a statistically significant effect for a given α (Bloom 1995)

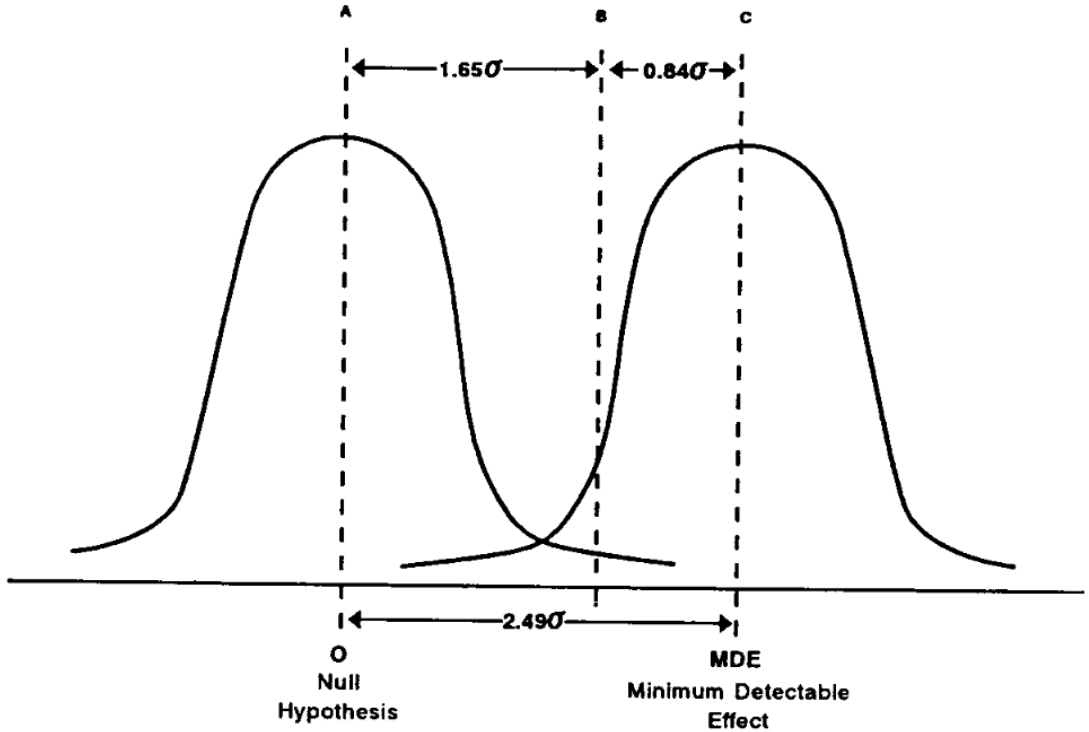

Figure 1: An Illustration of the Relationship Between the Minimum Detectable Effect and the Standard Error of an Impact Estimator

Figure 1: MDE for one-tailed test ($\alpha = 0.05$) with 80% power. You want 80% of the mass of right curve to be above B. Source: Bloom (1995)

- Power of the test: κ (often $\kappa = 0.8$)
- Type I error: α (often $\alpha = 0.05$)
- Recall: $\hat{\beta}/se(\hat{\beta}) = t^*$
- We build on this to identify the Minimum detectable effect (MDE) $\equiv |\beta_{\alpha,\kappa,se(\hat{\beta})}| = se(\hat{\beta})(t_{\alpha/2} + t_{1-\kappa})$
- We know the critical t values in large sample
  - For $\alpha = 0.05$, 2-tailed test: $t_{\alpha/2} = t_{0.025} = 1.96$
  - For 80% power: $t_{1-\kappa} = 0.84$
  - Thus: MDE $= 2.8se(\hat{\beta})$

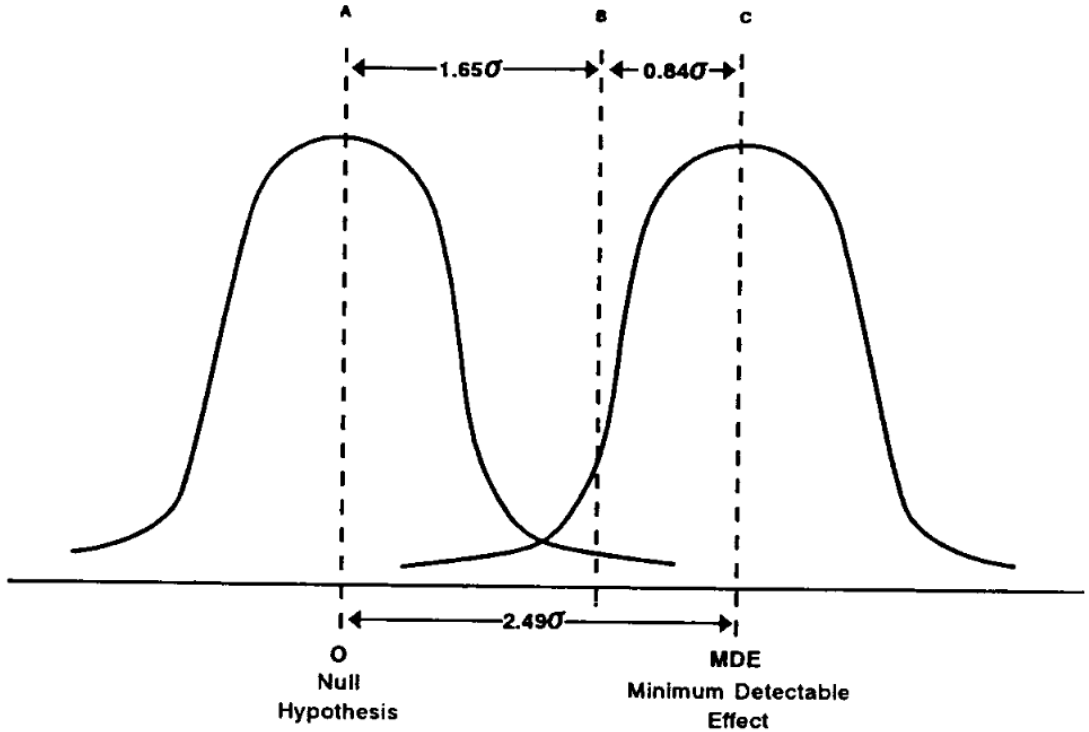

Figure 1: An Illustration of the Relationship Between the Minimum Detectable Effect and the Standard Error of an Impact Estimator

Figure 2: MDE for one-tailed test ($\alpha = 0.05$) with 80% power. You want 80% of the mass of right curve to be above B. Source: Bloom (1995)

- We can now connect this to the sample size needed!
- We will need to estimate the standard error of $\hat{\beta}$
- General expression:

$$\begin{aligned} se(\hat{\beta}) &= se(\overline{Y(1)} - \overline{Y(0)}) \\ &= \sqrt{\frac{\sigma_{y1}^2}{N_1} + \frac{\sigma_{y0}^2}{N_0}} \end{aligned}$$

- In RCT with probabilty of treatment p and sample size N:

$$se(\hat{\beta}) = \sqrt{\frac{1}{N}\left(\frac{\sigma_{y1}^2}{p} + \frac{\sigma_{y0}^2}{1-p}\right)}$$

- Recall that MDE $= se(\hat{\beta})(t_{\alpha/2} + t_{1-\kappa})$
- Plug in estimate for $se(\hat{\beta})$ and solve for N:

$$N = \frac{(t_{\alpha/2} + t_{1-\kappa})^2 \left(\frac{\sigma_{y1}^2}{p} + \frac{\sigma_{y0}^2}{1-p} \right)}{\text{MDE}^2}$$

- Implications...
- N decreases in the size of the MDE
- N increases in σ_{y1}, σ_{y0}
- N decreases when α increases, κ decreases
- (N also decreases using within-unit variation - more later)

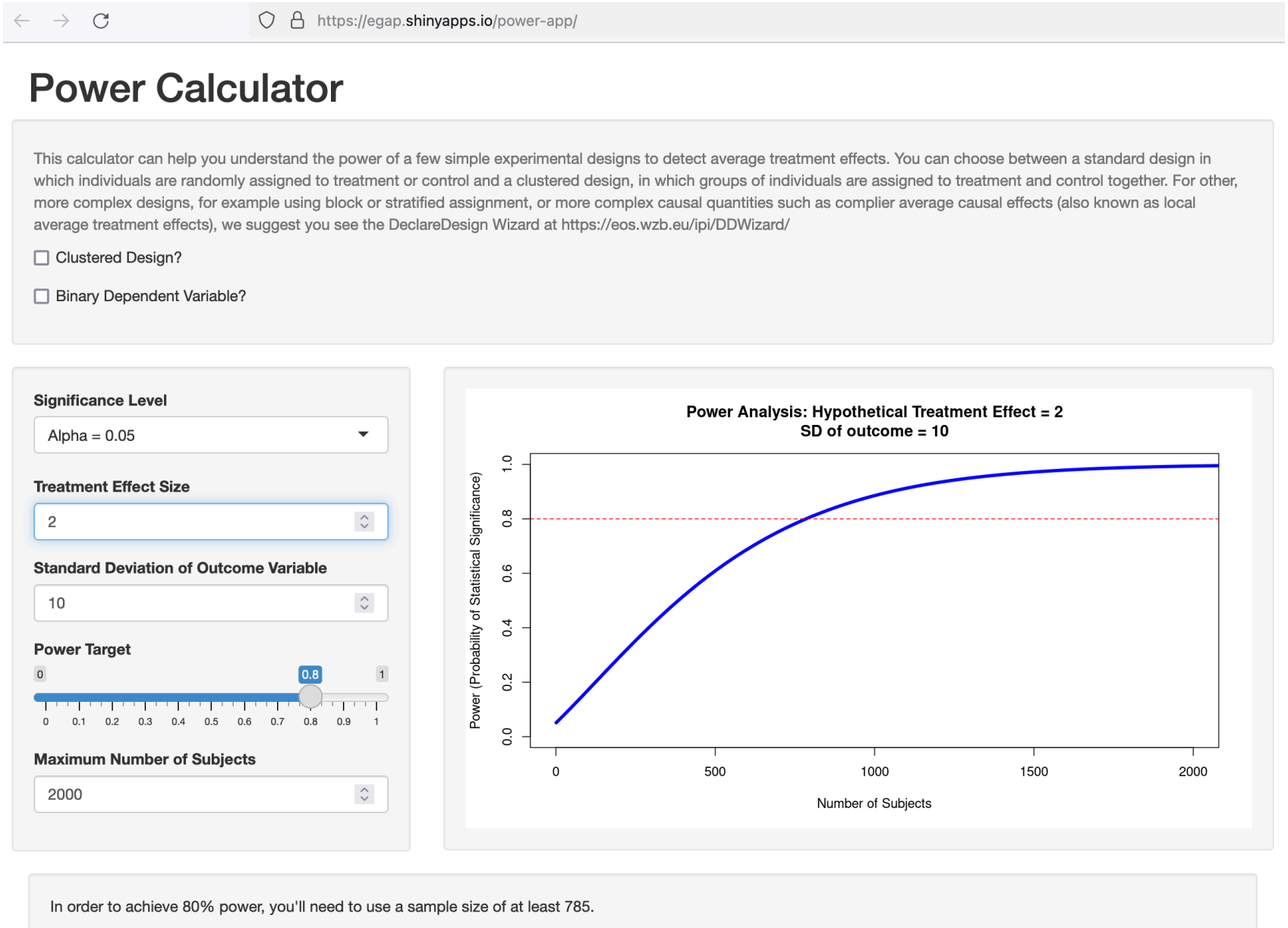

Figure 3: EGAP power calculator

6. Use a random number generators to treat units in your sample

- Generally: $Pr(\text{Treatment}) = 0.5$ maximizes power
- Typically: Bernoulli trials
- Each unit will then either be treated (Bernoulli $= 1$) or not

7. Measure pre-treatment data

- Check balance of covariates between treatment and control group (plot)
- If imbalance: possibly re-randomize, control for imbalanced vars Morgan and Rubin (2012)

8. Implement treatment and measure outcomes
9. Estimate average treatment effect (w, w/o adjustments for pre-treatment variables)

Misc #1: treatment

- Direct vs. indirect treatments: can you ensure **compliance** with treatment?
- If not: at best you can **encourage** treatment take-up
- Example: you tell people to read an article about something sad to see how it affects their consumption of chocolate
  - The treatment is something like Mood $\in \{\text{Happy, Sad}\}$
  - But you cannot force people to be sad. You can only encourage them to be so...
- Need to account for non-compliers (people who will not be sad despite being encouraged to be so). With non-compliers:
  - **Local average treatment effect** (LATE): effect of treatment *among compliers*
  - **Intent to treat** (ITT): effect of treatment averaging those who took the treatment and those who didn't

Figure 4: LATE and ITT

LATE: effect of solar panels on kerosene spending among households that were encouraged to buy them and did so

ITT: effect of encouragement to buy solar panels on kerosene spending (regardless of whether they did buy it)

- ITT can be estimated easily as:

$$\widehat{\text{ITT}} = \frac{1}{N_1} \sum_{i \in Z=1} (Y_i | Z = 1) - \frac{1}{N_0} \sum_{i \in Z=0} (Y_i | Z = 0)$$

- Also obtainable by estimating (where $\hat{\beta}$ is the estimate of ITT):

$$Y_i = \alpha + \beta Z_i + \varepsilon_i$$

- Interesting quantity from a policy perspective: overall effect accounting for the fact that some won't comply
- LATE is more complicated. Typically estimated with instrumental variables (more soon)

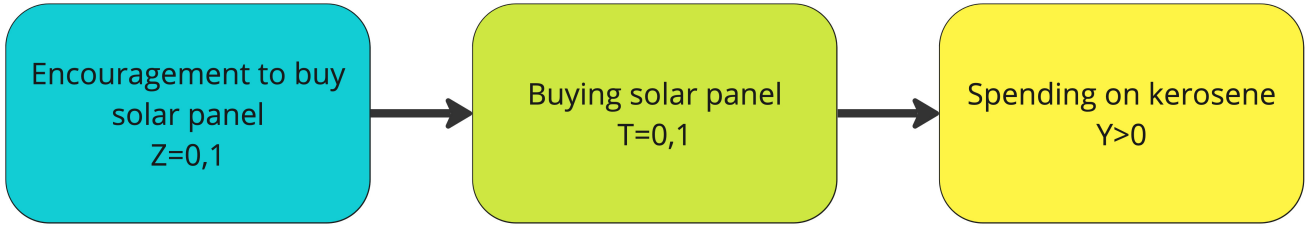

Table 2. Effect of MGP solar microgrids on household kerosene spending. Effects are shown for spending in the private market (A), the PDS (B), and overall (C). The SEs are clustered by habitation and are shown in parentheses. All dependent variables are measured in rupees per month. n = 3825; number of households, 1281. *p < 0.10, **p < 0.05, ***p < 0.01.

	ITT (OLS)		LATE (IV)	
	(1)	(2)	(3)	(4)
(A) Kerosene spending on private market				
Treatment	−14.01*** (5.28)	−14.49** (6.91)	−47.49** (19.83)	−49.36** (24.62)
Household FE		Yes		Yes
Wave FE	Yes	Yes	Yes	Yes
Pretreatment mean for control group = 72				
First-stage estimate			0.29	0.29
First-stage F statistic			10.15	10.01

Figure 5: Source: Aklin et al. (2017)

Misc #2: treatment assignment

- At what **level can you assign treatment**?
- Generally: individual unit
  - Treatment and outcomes are measured at the same level
- Sometimes: treatment is jointly given to entire groups
- Example: you want to study the effect of a new technology on farming output. Hard to give this technology to a single farmer in a village.
- Implication: randomization happens at the aggregate level
  - Eg: you treat all farmers in a given village
- Has consequence for standard errors...

- Recall our variance-covariance matrix:

$$VC = \begin{bmatrix} E[u_1^2] & E[u_1 u_2] & ... & E[u_1 u_n] \\ E[u_2 u_1] & E[u_2^2] & ... & E[u_2 u_n] \\ ... & ... & ... & ... \\ E[u_n u_1] & ... & ... & E[u_n^2] \end{bmatrix}$$

- Two assumptions can simplify this: **homoskedasticity** and **independence** across units

$$VC = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & ... & \sigma^2 \end{bmatrix}$$

- We also said: homoskedasticity is unlikely. But what about independence?

- If I treat a group (eg a village): then it may well be that there is correlation of residuals across units *within* this group (village)
- Eg: if the technology is particularly effective in villages with characteristic W, then they will all have abnormally high residuals
- Thus: correlation *across* units *within* groups. Unlikely that $a, b... = 0$.

$$VC = \begin{bmatrix} \sigma^2 & a & ... & b \\ a & \sigma^2 & ... & c \\ ... & ... & ... & ... \\ b & c & ... & \sigma^2 \end{bmatrix}$$

- Note: logic can be expanded to observational data.
  - Spatial correlation (high u in Germany $\rightarrow$ high u in France)
  - Temporal correlation (high u at $t = 1 \rightarrow$ high u at $t = 2$)

- Solution: allow residuals to be correlated within groups. Gives rise to **clustered standard errors**
- Block-diagonal matrix. Here with two groups, G_a and G_b:

$$VC = \begin{bmatrix} \sigma^2 & a & a & 0 & 0 & 0 \\ a & \sigma^2 & a & 0 & 0 & 0 \\ a & a & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & b & b \\ 0 & 0 & 0 & b & \sigma^2 & b \\ 0 & 0 & 0 & b & b & \sigma^2 \end{bmatrix}$$

- Can be estimated via conventional tools; see Liang and Zeger (1986), Abadie et al. (2023), De Chaisemartin and Ramirez-Cuellar (2024)

Limitations of experimental methods

- **Cost**: intervention, sample size, data collection, etc.
- **Feasibility**. Eg: how can you randomize the religious beliefs? Can you randomize a government policy?
- **Ethics**
  - Can you randomize access to a life-saving drug?
  - Audit experiments

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

By MARIANNE BERTRAND AND SENDHIL MULLAINATHAN*

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (*JEL* J71, J64).

Figure 6: Source: Bertrand and Mullainathan (2004)

Conclusion

- RCTs: powerful designs to identify causal effects
- But not always feasible. Thus: other approaches needed
- Can we mimic experiments with **observational** (non-experimental) data?
- Idea behind **quasi-experiments**: we might be able to identify situations that are as-if experiments even though the analyst doesn't control treatment assignment
- Range of quasi-experiments: instrumental variables, difference-in-difference, regression discontinuity, etc.

Questions?

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *Quarterly Journal of Economics* 138 (1): 1–35.

Aklin, Michaël, Patrick Bayer, S. P. Harish, and Johannes Urpelainen. 2017. "Does Basic Energy Access Generate Socioeconomic Benefits? A Field Experiment with Off-Grid Solar Power in India." *Science Advances* 3 (5): e1602153.

Angrist, Joshua, and Steffan J. Pischke. 2008. *Mostly Harmless Econometics*. Princeton: Princeton University Press.

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.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19 (5): 547–56.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–232.

De Chaisemartin, Clément, and Jaime Ramirez-Cuellar. 2024. "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" *American Economic Journal: Applied Economics* 16 (1): 193–212.

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh: Oliver; Boyd.

Liang, Kung-Yee, and Scott L Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.

Morgan, Kari Lock, and Donald B Rubin. 2012. "Rerandomization to Improve Covariate Balance in Experiments." *Annals of Statistics* 40 (2): 1263–82.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Stock, James H., and Mark W. Watson. 2011. *Introduction to Econometrics, 3rd Edition*. Pearson.