Lecture 6: Research Designs and Analysis of Experimental Data

Outline:

- Types of research designs
- Internal / external validity of a study
- Analysis of experimental data
- Limitations of experimental analysis
- Applications
- Readings: A&P Chapters 1-2, I&W Lecture 1

Recall the Rubin Causal Model

$Y_i(0)$ = potential outcome for unit i if untreated

$Y_i(1)$ = potential outcome for unit i if treated

T_i = 1 if unit i treated, 0 if not

- $Y_i(1) - Y_i(0)$ = unit-level treatment effect. We seek to estimate its average (ATE)

- We observe (Y_i, T_i), where Y_i is the observed outcome:

$$Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$$

$$= E[Y_i(0)] + [Y_i(1) - Y_i(0)]T_i + [Y_i(0) - E(Y_i(0))]$$

$$= \beta_0 + \beta_{1i}T_i + u_i$$

Assume $\beta_{1i} = \beta_1$

$$= \beta_0 + \beta_1 T_i + u_i$$

Study (research) designs

- In order to estimate the causal effect of a treatment, there are 3 main types of study designs, each corresponding to a different treatment assignment mechanism:

- Randomized controlled experiments

- Observational studies

- Natural (or quasi-) experiments

Randomized controlled experiment

- In a controlled experiment, treatment assignment is random (ex: coin toss, last digit of SSN)

- Up to some random error, the random assignment balances the two groups (the treatment group T=1, and the control group T=0) with respect to all factors except the treatment
  - This makes the assumption $Cov(T_i, u_i)=0$ highly credible since T_i is randomly assigned, and thus in theory independent of u_i

- $\Rightarrow$ Causal inferences based on random controlled experiments are the most credible
  - However, in social & natural sciences, controlled experiments remain relatively rare (unpractical, expensive, etc)

Observational studies

- It is a study where the 'researcher' does not control the assignment of subjects to treatment and control
  - Another way to think about this is that the treatment (T_i) (i.e. the value of T_i) is a choice variable of the subject

- The lack of control over the treatment assignment implies that the treatment and control groups may not be comparable
  - This implies that the assumption that T_i and u_i are uncorrelated is unlikely to be satisfied

- Most data is social sciences comes from observational studies…

Natural experiment

- A natural experiment (also called quasi-experiment) is an observational study where the treatment (T_i) is partially assigned randomly by "nature"
  - Natural experiments are not controlled experiments
- Natural experiments generated by:
  - (Unanticipated) changes in policies, laws, or regulatory rules (Ex: State tax laws on tobacco, introduction of an MPA, CAA Non-Attainment designation, etc)
  - Naturally occurring (exogenous) phenomena: weather fluctuations, volcano eruptions, etc.
  - Etc
- Like in observational studies, identification of causal effects with natural experiments require important assumptions (to come)

Classic example of a natural experiment

- Suppose you are interested in measuring the causal effect of military service on civilian earnings:

- Is there a correlation between u_i and T_i (military service indicator)?
  - Yes. It depends on who applies to the military and screening process used by military administration

- Vietnam draft lottery: Between 1970-1972, the military drafted individuals according to "call numbers" (RSN) that were randomly assigned (each birth date was randomly assigned one)

- Men with RSN below a cutoff were eligible to be drafted while men with RSN above the cutoff could not
  - RSN does not perfectly determine military service: deferments (health, college), volunteering, draft-dodging

- **But** at-risk cohorts randomly assigned "incentive" to join military

- In other words, a randomly assigned factor changes *probability* an individual serves in military
  - Sounds like an RCT… but does not switch T_i from 0 to 1… instead if gives a (randomly determined) increase in the probability of having $T_i = 1$

- How can we use this?? $\Rightarrow$ Basis for instrumental variables

Internal and external validity of a study:

- Distinguishes between population/setting studied and population/setting of interest
- Population studied: population from which the sample was drawn
- Population of interest: population for which we want to apply the results of the study
- "Setting": institutional, legal, social, and economic rules/norms

Internal and external validity of a study:

Internal and External Validity

Key Concept 9.1

A statistical analysis is **internally valid** if the statistical inferences about causal effects are valid for the population being studied. The analysis is **externally valid** if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

Threats to internal and external validity

- Internal validity has two components (and thus two possible categories of 'threats'):
- Estimator of the causal effect must be consistent
- Inference should have the desired significance level

- Threats to external validity:
- Differences in populations (study and target population)
- Differences in settings (study and target settings)

Threats to external validity

- Differences in population:
  - A new drug is invented and shown to be effective on lab rats. This does not mean that the drug will work on humans

- Differences in settings:
  - In the U.S., multiple studies suggest that an additional year of education increases earnings by 8-10%
  - This may not generalize to other European countries since their labor markets are more regulated, and wage-setting is more subject to centralized norms

Threats to internal validity

- **Sources of inconsistency in the regression estimate of the causal effect:**
  - Omitted variable bias
  - Functional form misspecification
  - Measurement error
  - Sample selection bias
  - Simultaneous causality
- Note: All these imply that $Cov(T_i , u_i) \neq 0$

- **Sources of incorrect inference:**
  - Heteroskedasticity (we know the solution to this)
  - Serial correlation in the regression error terms (similar solution, to come)
- Note: All these imply that $Var[u_i|T_i]$ is not constant

Randomized controlled experiment

- Random assignment ensures that treatment is independent of potential outcomes (if randomization protocol correctly implemented and complied with)

- Written as $T_i \perp (Y_i(0) , Y_i(1))$, where "$\perp$" means "statistically independent of" (i.e. uncorrelated)
  - A less-restrictive assumption would be a mean-independence assumption $E[Y_i(1) | T_i] = E[Y_i(1)]$ and $E[Y_i(0) | T_i] = E[Y_i(0)]$

- Note: For now we assume compliance with the treatment, or that any non-compliance is exogenous to the potential outcomes ("Ideal randomized controlled experiment")

Estimation of ATE in an ideal randomized controlled experiment

- When the treatment is binary, β_1 is just the difference in mean outcome (Y) in the treatment vs. control group, so we can estimate it by:

$$\overline{Y}_1 - \overline{Y}_0$$

$$\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} 1(T_i = 1) \times Y_i$$

$$\overline{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} 1(T_i = 0) \times Y_i$$

- This difference in means is sometimes called the ***differences estimator***

- (...) This difference in means is sometimes called the ***differences estimator***
  - *Later we introduce the **difference-in-difference estimator**, which will be similar, but applied to changes over time in mean outcomes across treatment and control groups*
  - *This is especially useful with quasi-experimental data where treatment and control groups may differ in their baseline characteristics*

Identification of ATE in an ideal randomized controlled experiment

- Algebra: Consider the population regression function of Y_i on T_i, and recall the interpretation of β_1 when T_i is binary:

$$Y_i = \beta_0 + \beta_1 T_i + u_i$$

$$\beta_1 = E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0]$$

$$= E[Y_i(1)T_i + Y_i(0)(1 - T_i) \mid T_i = 1] -$$

$$E[Y_i(1)T_i + Y_i(0)(1 - T_i) \mid T_i = 0]$$

$$= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]$$

$$= E[Y_i(1)] - E[Y_i(0)] = ATE$$

…Using the fact that T_i is uncorrelated with $Y_i(0)$ and $Y_i(1)$

Implication of random assignment of treatment

- A important property of random assignment of the treatment is that it should balance the pre-treatment (baseline) characteristics

- Let X_i denote the baseline or pre-treatment characteristics

- $\Rightarrow$ Random assignment of T_i implies:

$$E[X_i \mid T_i = 1] = E[X_i \mid T_i = 0]$$

- Empirically easy to test with t-test for equality of means, or a normalized difference in means
  - If not true for 95% of cases, then the randomization protocol failed. We will use a similar intuition for balancing tests in other research designs later

Controlling for individual characteristics

- As we just saw, In a ideal randomized control experiment, you can estimate the average treatment effect by a simple regression of Y on the treatment indicator variable T:

$$Y_i = \beta_0 + \beta_1 T_i + u_i$$

- Typically you also observe additional subject pre-treatment characteristics, $X_{1i},\ldots,X_{ri}$

- Differences estimator with additional regressors:

$$Y_i = \beta_0 + \beta_1 T_i + \gamma_1 X_{1i} + \ldots + \gamma_r X_{ri} + u_i$$

Why controlling for individual characteristics?

- ***1. Efficiency***: more precise estimator of β_1 (inclusion of the X's always leads to smaller standard errors for $\hat{\beta}_1$)

- ***2. Check for randomization***: If *T* is randomly assigned, then the OLS estimators of β_1 with and without the *X*'s should be similar – if they are not, this suggests that *T* was not properly randomly assigned (a problem with the experiment implementation)

Advantages of ideal random controlled experiments

- Identifying assumption is transparent and easy to interpret
- Identifying assumption can be tested on pre-treatment characteristics (randomization test)
- Statistical analysis simple
- Results easy to present and explain

Potential problems with experiments in practice

- **A. Threats to Internal Validity**
  - 1. Failure to randomize
  - 2. Failure to follow treatment protocol
  - 3. Experimental effects

- **B. Threats to External Validity**
  - Non-representative sample
  - Non-representative treatment
  - General equilibrium effects (scale effects)

- **C. Not every important question can be answered with an actual controlled experiment**
  - i.e. effect of air pollution on human health

Threats to internal validity in experiments

- ***1. Failure to randomize*** (or imperfect randomization)
- Ex: openings in job training program are filled with higher proportions with early applicants; latecomers more likely to be controls

- Randomization device fails (or is incorrectly used)

- $\Rightarrow$ May result is correlation between T and u (biased estimates of the average treatment effect parameter β_1)

- $\Rightarrow$ Testable (to some extent) with randomization test on sample averages of X by treatment status

Threats to internal validity (ctd)

- **2. *Failure to follow experimental protocol* (or non-*compliance)***
- **$\Rightarrow$ *realized T_i not equal to initial T/C assignment*)**
- (i) Some controls get the treatment; Some treated "don't take" treatment; "No-shows", etc
- (ii) Attrition (some subjects drop out)

- $\Rightarrow$ May result is correlation between T and u (biased estimates of the average treatment effect parameter β_1)

- Sometimes there is a solution to the non-compliance problem

Identification of ATT with partial compliance

- Consider the differences estimator:

- $Y_i = \beta_0 + \beta_1 T_i + u_i$

- Suppose there is partial compliance: some of the treated don't "take" the treatment; some of the controls receive the treatment; No-shows

- $\Rightarrow T$ (actual treatment status) differs from initial treatment assignment

- If T is correlated with u, and OLS is biased

Identification of ATT with partial compliance (ctd)

- Suppose initial assignment to control/treament groups is random (and denote the assignment variable by Z)

- Then (1) corr(Z,T) $\neq 0$ and (2) corr(Z,u) $= 0$

- Then β_1 can be estimated by Two-Stage Least Squares, with instrumental variable where the instrument for the endogenous variable T is Z = initial assignment
  - We cover TSLS and instrumental variables in the coming lecture

- This can be extended to include X's (included exogenous variables)

Threats to internal validity (ctd)

- ***3. Experimental effects***
- (i) Experimenter bias (conscious or subconscious): The experimenter (think of a clinical drug trial) may exert "extra effort" or "extra care," with treatment individuals
  - Drug companies have lots of $$$ at stake in clinical trials
  - Double-blind protocol as a result...
- (ii) Subject behavior might be affected by being in an experiment (*Hawthorne* effect)
  - Also addressed by double-blind protocol
- Both may lead the OLS estimator (the differences estimator) to be biased

Applications:

- Project STAR (Student-Teacher Achievement Ratio)
  - 4-year study, $12 million
- Applies to children in grades K-3
- Upon entering the school system, a student was randomly assigned to one of three groups:
  - regular class (22 – 25 students)
  - regular class + aide
  - small class (13 – 17 students)
- *Y* = Combined score on a math test and Stanford Achievement Test reading component

Randomization test and mean outcome by treatment status

TABLE I

COMPARISON OF MEAN CHARACTERISTICS OF TREATMENTS AND CONTROLS: UNADJUSTED DATA

A. Students who entered STAR in kindergarten[b]

Variable	Small	Regular	Regular/Aide	Joint *P*-Value[a]
1. Free lunch[c]	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition rate[d]	.49	.52	.53	.02
5. Class size in kindergarten	15.1	22.4	22.8	.00
6. Percentile score in kindergarten	54.7	49.9	50.0	.00

Note: rows 1-3 are pre-treatment characteristics.

Regression analysis of experimental STAR data

- The "differences" regression model:
- $Y_i = \beta_0 + \beta_1 SmallClass_i + \beta_2 RegAide_i + u_i$

- *Note: since treatment can take 3 values, we include 2 treatment indicators:*
- $SmallClass_i$ = 1 if in a small class
- $RegAide_i$ = 1 if in regular class with aide

- Additional regressors (*X*'s)
  - teacher experience
  - free lunch eligibility (indicator or low SES)
  - gender, race

TABLE 13.1 Project STAR: Differences Estimates of Effect on Standardized Test Scores of Class Size Treatment Group

	Grade			
Regressor	**K**	**1**	**2**	**3**
Small class	13.90** (2.45)	29.78** (2.83)	19.39** (2.71)	15.59** (2.40)
Regular size with aide	0.31 (2.27)	11.96** (2.65)	3.48 (2.54)	−0.29 (2.27)
Intercept	918.04** (1.63)	1039.39** (1.78)	1157.81** (1.82)	1228.51** (1.68)
Number of observations	5786	6379	6049	5967

The regressions were estimated using the Project STAR Public Access Data Set described in Appendix 13.1. The dependent variable is the student's combined score on the math and reading portions of the Stanford Achievement Test. Standard errors are given in parentheses under the coefficients. **The individual coefficient is statistically significant at the 1% significance level using a two-sided test.

Summary: the STAR experiment

- Main findings:
- The effects are statistically significant but small quantitatively (small class size effect = 1-3% of average score)

- Effect is sustained but not cumulative or increasing $\Rightarrow$ biggest effect at the youngest grades

- Threats to internal validity:
- Partial compliance/attrition
  - Can use TSLS with Z = initial assignment
  - As it turns out, TSLS and OLS estimates are similar (Krueger (1999)), so this bias seems not to be large

Random manipulations of racial perceptions

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

By Marianne Bertrand and Sendhil Mullainathan*

Random manipulations of racial perceptions

- Bertrand&Mullainathan "Are Emily and Greg More Employable Than Lakisha and Jamal", American Economic Review, 2004
- Create fake CVs and send replies to job ads in Boston and Chicago newspapers
- Allocate names at random to CVs – either 'black-sounding' names or 'white-sounding' names
- Outcome variable is call-back rate on job applications
- Interpretation – not direct measure of racial discrimination, just effect of having a 'black-sounding' name – may have other implications
  - But name uncorrelated by construction with other material on CV

TABLE A1—FIRST NAMES USED IN EXPERIMENT

White female Name	L(W)/L(B)	Perception White
Allison	∞	0.926
Anne	∞	0.962
Carrie	∞	0.923
Emily	∞	0.925
Jill	∞	0.889
Laurie	∞	0.963
Kristen	∞	0.963
Meredith	∞	0.926
Sarah	∞	0.852
Fraction of all births:		
3.8 percent		

African-American female Name	L(B)/L(W)	Perception Black
Aisha	209	0.97
Ebony	∞	0.9
Keisha	116	0.93
Kenya	∞	0.967
Lakisha	∞	0.967
Latonya	∞	1
Latoya	∞	1
Tamika	284	1
Tanisha	∞	1
Fraction of all births:		
7.1 percent		

White male Name	L(W)/L(B)	Perception White
Brad	∞	1
Brendan	∞	0.667
Geoffrey	∞	0.731
Greg	∞	1
Brett	∞	0.923
Jay	∞	0.926
Matthew	∞	0.888
Neil	∞	0.654
Todd	∞	0.926
Fraction of all births:		
1.7 percent		

African-American male Name	L(B)/L(W)	Perception Black
Darnell	∞	0.967
Hakim		0.933
Jamal	257	0.967
Jermaine	90.5	1
Kareem	∞	0.967
Leroy	44.5	0.933
Rasheed	∞	0.931
Tremayne	∞	0.897
Tyrone	62.5	0.900
Fraction of all births:		
3.1 percent		

TABLE 3—RESUME CHARACTERISTICS: SUMMARY STATISTI

Sample:	All resumes	White names	African-American
Characteristic:			
College degree (Y = 1)	0.72 (0.45)	0.72 (0.45)	0.72 (0.45)
Years of experience	7.84 (5.04)	7.86 (5.07)	7.83 (5.01)
Volunteering experience? (Y = 1)	0.41 (0.49)	0.41 (0.49)	0.41 (0.49)
Military experience? (Y = 1)	0.10 (0.30)	0.09 (0.29)	0.10 (0.30)
E-mail address? (Y = 1)	0.48 (0.50)	0.48 (0.50)	0.48 (0.50)
Employment holes? (Y = 1)	0.45 (0.50)	0.45 (0.50)	0.45 (0.50)
Work in school? (Y = 1)	0.56 (0.50)	0.56 (0.50)	0.56 (0.50)
Honors? (Y = 1)	0.05 (0.22)	0.05 (0.23)	0.05 (0.22)
Computer skills? (Y = 1)	0.82 (0.38)	0.81 (0.39)	0.83 (0.37)
Special skills? (Y = 1)	0.33 (0.47)	0.33 (0.47)	0.33 (0.47)
Fraction high school dropouts in applicant's zip code	0.19 (0.08)	0.19 (0.08)	0.19 (0.08)
Fraction college or more in applicant's zip code	0.21 (0.17)	0.21 (0.17)	0.21 (0.17)
Fraction Whites in applicant's zip code	0.54 (0.33)	0.55 (0.33)	0.54 (0.33)
Fraction African-Americans in applicant's zip code	0.31 (0.33)	0.31 (0.33)	0.31 (0.33)
Log(median per capital income) in applicant's zip code	9.55 (0.56)	9.55 (0.56)	9.55 (0.55)
Sample size	4,870	2,435	2,435

Shows that information contained in CVs is the same (on average) across white-sounding name and black-sounding name applicants

Implies that difference in call-back rates is not due to differences in CV

2018

Main results

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65	6.45	1.50	3.20
	[2,435]	[2,435]		(0.0000)
Chicago	8.06	5.40	1.49	2.66
	[1,352]	[1,352]		(0.0057)
Boston	11.63	7.76	1.50	4.05
	[1,083]	[1,083]		(0.0023)
Females	9.89	6.63	1.49	3.26
	[1,860]	[1,886]		(0.0003)
Females in administrative jobs	10.46	6.55	1.60	3.91
	[1,358]	[1,359]		(0.0003)
Females in sales jobs	8.37	6.83	1.22	1.54
	[502]	[527]		(0.3523)
Males	8.87	5.83	1.52	3.04
	[575]	[549]		(0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the p-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

Experimental approach in ecology

Kelp Beds and Sea Otters: An Experimental Approach
Author(s): David O. Duggins
Source: *Ecology*, Vol. 61, No. 3 (Jun., 1980), pp. 447-453
Published by: Ecological Society of America
Stable URL: http://www.jstor.org/stable/1937405
Accessed: 21-04-2015 18:18 UTC

My attempt at explaining the background…

- Sea otters eat sea urchins… Without sea otters, the sea urchin population overgrows and this leads to overgrazing of kelp
- Paper wants to measure the “causal effect” of sea otters on kelp density (I think..)
- Observational studies (called comparative studies) have shown this by comparing populations of sea urchins and kelp density across sites with and without sea otters
- Duggins argues that an experimental approach is needed to understand the mechanisms at play

Experimental approach

- "examination of the process is accomplished by performing *in situ* experimental manipulations of single species and then measuring the relevant parameters as the system responds to that manipulation"

- "in essence we are mimicking (in a small scale and in a controlled manner) the influence of sea otters"

- ⇒ To simulate predation upon urchins, Duggins removes urchins from experimental plots (N=3) using adjacent plots with urchins as controls (N=2)

- ⇒ Measures of kelp density were collected at these sites every 2 months over a 2 year period
  - Surrounding areas were also cleared of urchins (p.448). Why?

Results

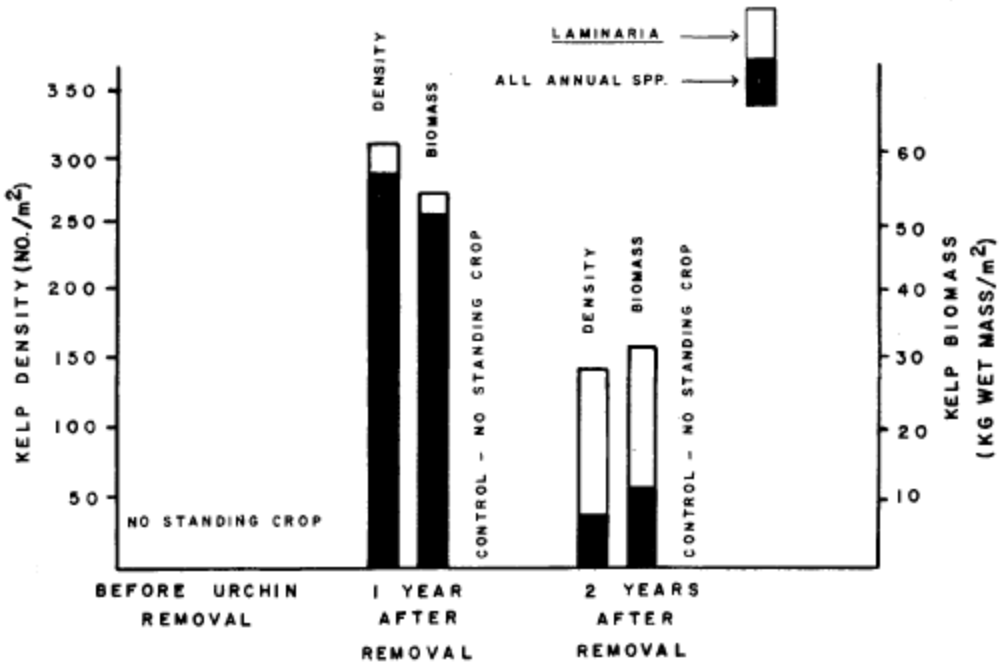

FIG. 1. Mean kelp densities and biomasses in urchin removal experiments ($N = 3$) and controls ($N = 2$). Values shown from samples taken in July 1976, 1977, and 1978. Shaded portions represent contribution by annuals, unshaded represents *Laminaria*.

- 'Treatment' leads to increase density of kelp in experimental areas. Type of kelp switch to *Laminaria* in year 2