Week 2: Randomized Experiments

PUBL0050 Causal Inference

Julia de Romémont

Term 2 2023-24

UCL Departement of Political Science

Stop and think

- An NGO that directs development aid (cash transfers) to poor households in Country X wants to evaluate the impact of its activities.
- The NGO collects data on 10000 households from Country X, and compares the economic status of aid receivers and non-aid receivers one year after the cash was transfered.
- The analysis reveals that those who were given aid were significantly **poorer** than those who were not given aid.

Question: Should the NGO conclude that development aid makes the economic conditions of the poor worse? Why or why not?

Motivation

Does Aid Improve the Well-Being of the Extreme Poor?

A critical question in development economics is whether is it possible to reliably improve the livelihoods of the poorest households in the world by giving them aid. Banerjee et. al. study the effects of self-employment focused development aid across 6 different countries via a randomized experiment.

- Unit of analysis: 10,495 households in India, Ethiopia, Pakistan, Ghana, Honduras and Peru
- Outcomes (Y): Several outcomes, including: family assets; overall consumption; income from animals; sufficient food
- Treatment (D): 1 if the household was in the treatment group, 0 if in the control group

The Experimental Approach

> *"It is not that I believe an experiment is the **only** proper setting for discussing causality, but I do feel that an experiment is the simplest such setting."*
>
> *– Holland, 1986*

- The goal of experiments is to **eliminate** observable and unobservable confounders **by design**.
- Assumption: The world is heterogeneous and **one cannot hold everything constant** other than the treatment.
- One can, however, **randomize to render confounders ineffectual**.

This Lecture

Identification under Random Assignment

Estimation under Random Assignment

Statistical Inference under Random Assignment

Covariates and Random Assignment

Internal and External Validity

Identification under Random Assignment

Review of last week

1. The **fundamental problem of causal inference**:

$$Y_i = D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i}$$

so

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

and thus

$\tau_i = Y_{1i} - Y_{0i}$ is unobservable.

2. Instead, we focus on the **average treatment effect**:

$$\tau_{\text{ATE}} = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$$

Review of last week

3. But, the **difference-in-group means** is only unbiased sometimes!

$$E[Y_i|D=1] - E[Y_i|D=0] = \tau_{\mathsf{ATT}} + \text{selection bias}$$

4. In particular, the DIGM is unbiased when
   - $E[Y_{0i}|D=1] = E[Y_{0i}|D=0]$
     - ***i.e. there is no selection bias***
   - $\tau_{\mathsf{ATT}} = \tau_{\mathsf{ATE}}$
     - ***i.e. the ATT and ATE are the same***

Randomisation solves the selection problem

Why does random assignment allow us to make causal statements?

> *"Random assignment works not by eliminating individual differences but rather by ensuring that the mix of individuals being compared is the same"*
>
> *– Angrist and Pischke, 2015, p.16*

Intuition

With D_i assigned at random, it follows that:

- We observe Y_{1i} for a **representative sample** of the population of units, which gives an **unbiased estimate** of $E[Y_{1i}]$
- We observe Y_{0i} for a **representative sample** of the population of units, which gives an **unbiased estimate** of $E[Y_{0i}]$

From this, it follows that:

- No selection bias in *expectation*, because $E[Y_{0i}|D = 0] = E[Y_{0i}|D = 1] = E[Y_{0i}]$
- Treated units are representative of all units with respect to their potential outcomes, so $E[\tau_{\text{ATT}}] = \tau_{ATE}$

$\rightarrow$ The difference in group means, under randomisation, therefore provides an unbiased estimate of the average treatment effect

Identification under random assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

***Problem**: $\tau_{ATE} = E[Y_{1i} - Y_{0i}]$ is unobserved. But given random assignment:*

$$\begin{aligned} E[Y_i|D=1] &= E[D \cdot Y_{1i} + (1-D) \cdot Y_{0i}|D=1] \\ &= E[Y_{1i}|D=1] \\ &= E[Y_{1i}] \qquad \textit{(Random assignment)} \end{aligned}$$

$$\begin{aligned} E[Y_i|D=0] &= E[D \cdot Y_{1i} + (1-D) \cdot Y_{0i}|D=0] \\ &= E[Y_{0i}|D=0] \\ &= E[Y_{0i}] \qquad \textit{(Random assignment)} \end{aligned}$$

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}] = \underbrace{E[Y_i|D=1] - E[Y_i|D=0]}_{\textit{Difference in Means}}$$

Identification under random assignment

Implication

The difference in group means (DIGM) ***is*** an **unbiased estimator** of τ_{ATE} when D_i **is assigned at random.**

Estimation under Random Assignment

Notation reminder

Greek letters

- Letters like β, α, μ, or τ are *the truth* (parameters) in the population
- Letters with extra marks like $\hat{\beta}$, $\hat{\alpha}$, $\hat{\mu}$, or $\hat{\tau}$ are our *estimate* of the truth based on our sample

Latin letters

- Letters like X or Y represent *actual data* (variables) in our sample
- Letters with extra marks like $\bar{X}$ or $\bar{Y}$ are *statistics* from our sample

Notation reminder

Data → Estimator → Estimate → Parameter

Data	Y
Estimator	$\bar{Y} = \frac{\sum_i^N Y_i}{N}$
Estimate	$\widehat{\mu}$
Parameter	μ

$$Y \rightarrow \frac{\sum_i^N Y_i}{N} \rightarrow \widehat{\mu} \xrightarrow{hopefully} \mu$$

Analogy principle

The **analogy principle** tells us to estimate a characteristic in the population using the same characteristic in the sample.

- For example, following the analogy principle, we estimate the *population* ***mean*** $\mu_Y = E[Y]$ using the *sample* ***mean*** $\bar{Y}$:

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

- To estimate the *population* ***variance*** $\sigma_Y^2 = Var[Y] = E[(Y - E[Y])^2]$, use the *sample* ***variance***:

$$\hat{\sigma}_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

Properties of estimators

To assess an estimator, we assume that it has a **sampling distribution** – i.e. the distribution of estimates produced under repeated sampling.

- **Unbiasedness**: Is the estimator's sampling distribution centered on the true parameter value?
- **Consistency**: As the sample size grows to infinity, does the estimator's sampling distribution converge to the true value?

Intuition: we want an estimator that, *on average*, gets the right answer, and where the estimation error decreases *as the sample size increases*.

Properties of estimators

What is being repeatedly sampled?

- **Statistical inference** → repeated sampling of units from a population
- **Causal inference** → repeatedly allocating units in the sample to different treatment assignments

Estimation under random assignment

Consider a randomized trial with N individuals in the sample.

Estimand

$\tau_{ATE} = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$

Estimator

By the analogy principle we use $\hat{\tau}_{ATE} = \bar{Y}_1 - \bar{Y}_0$*, where*

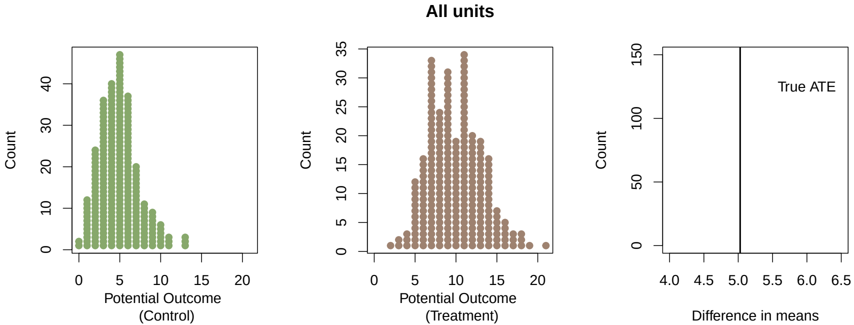
$$\bar{Y}_1 = \frac{\sum Y_i \cdot D_i}{\sum D_i} = \frac{1}{N_1} \sum_{D_i=1} Y_i$$

$$\bar{Y}_0 = \frac{\sum Y_i \cdot (1 - D_i)}{\sum (1 - D_i)} = \frac{1}{N_0} \sum_{D_i=0} Y_i$$

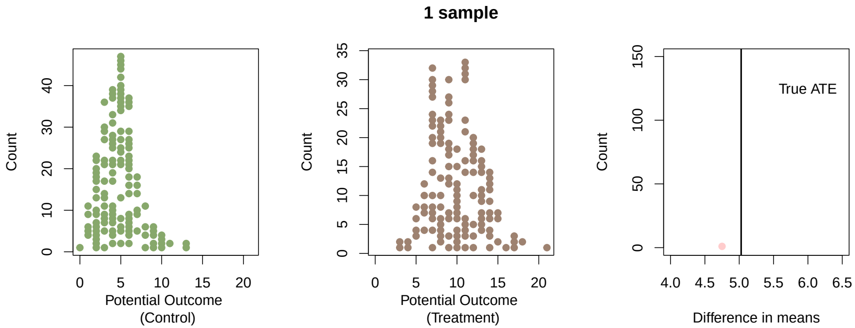
with $N_1 = \sum_i D_i$ *and* $N_0 = N - N_1$*.*

Is $\hat{\tau}_{\text{ATE}}$ an unbiased estimator of τ_{ATE}?

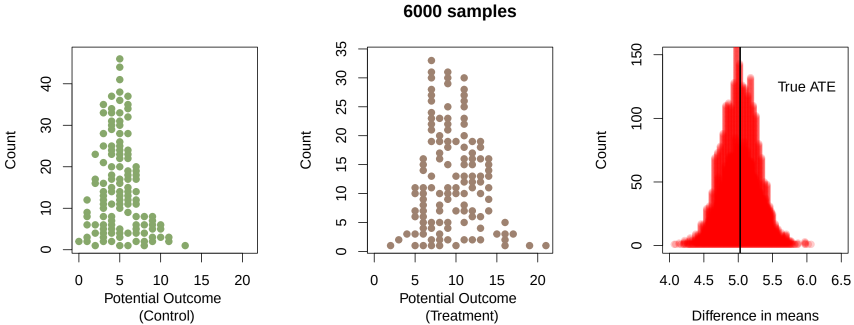
Unbiasedness under random assignment



Unbiasedness under random assignment



Unbiasedness under random assignment



Implications

- Randomization of the treatment makes the difference in group means an **unbiased** estimator of the true ATE
- **On average**, you can expect a randomized experiment to get the right answer
- This does not guarantee that the answer you get from any particular randomization will be exactly correct!

SATE and PATE

- The difference in group means is an unbiased estimator for the **S**ample **A**verage **T**reatment **Effect** (or, SATE).
- In intro stats classes, we would be concerned with making inferences about a **population parameter** (i.e. $E[Y_i]$) from a **sample statistic** (i.e $\bar{Y}_i$)
- Is the difference in group means is an unbiased estimator for the **P**opulation **A**verage **T**reatment **Effect** (or, PATE)?
- Answer: Only when the sample is drawn at random from the population! In the social sciences, this is very rare.
  - We will generally focus on the SATE
  - $\hat{\tau}_{\text{SATE}}$ will be *internally valid*, but may not be *externally valid*
  - More on this later

What is the (S)ATE in the Aid experiment?

In this study, families assigned to the treatment received:

1. A one-time transfer of a productive asset (normally an animal)
2. A regular transfer of food or cash for a few months
3. Technical skills training
4. Access to a savings account
5. Some health education and some basic health services

We will focus on the following outcomes:

1. `consumption` - Monthly household consumption ($, PPP)
2. `assets` - Index of family asset values (continuous)
3. `food` - Whether household has enough food every day (binary)
4. `livestock` - Monthly income from livestock ($, PPP)

What is the (S)ATE in the Aid experiment?

We can calculate the difference-in-means for each outcome in R:

```
consumption_ate <- mean(aid$consumption[aid$treatment==1]) -
                      mean(aid$consumption[aid$treatment==0])

asset_ate <- mean(aid$assets[aid$treatment==1]) -
                  mean(aid$assets[aid$treatment==0])

food_ate <- mean(aid$food[aid$treatment==1]) -
                      mean(aid$food[aid$treatment==0])

livestock_ate <- mean(aid$livestock[aid$treatment==1]) -
                      mean(aid$livestock[aid$treatment==0])
```

What is the (S)ATE in the Aid experiment?

Outcome	ATE
Consumption	2.19
Assets	0.27
Food	0.03
Livestock	17.65

→ In all cases, the positive ATE indicates that the treatment improves outcomes.

How large are these effects?

- **Substantive answer:** It depends on the scale on which the outcome variable is measured!
  - Consumption → treated households increase their monthly consumption by about \$2, compared to a control average of \$78
  - Livestock → treated households increase their income from livestock by about \$18, compared to a control group average of \$72
  - **Key point**: it is important to interpret the substantive magnitude of your estimated treatment effects!
- **Statistical answer:** How well do you remember your previous courses?

Statistical Inference under Random Assignment

Uncertainty and random assignment

- The sampling distribution shows that any given estimate $\hat{\tau}_{\text{ATE}}$ may be some distance from the true τ_{ATE}
- This goes to show that there still is uncertainty: we do not observe all units and we do not observe all potential outcomes
- How can we characterise this uncertainty?
  - **Hypothesis testing**: Can we convince a skeptic that the treatment effect is not equal to 0?
  - **Confidence intervals**: What are the range of values that likely bracket the true average treatment effect?

Standard error and t-statistic

The **standard error**[1] estimates the amount of sampling variability in the ATE estimator.

$$\widehat{\mathsf{SE}}_{\mathsf{ATE}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}$$

where

- N_1 and N_0 are the number of treatment and control observations
- σ_1^2 and σ_0^2 are the variances of the treatment and control groups

We can use the standard error to construct a **t-statistic** for the difference in means:

$$t_{\mathsf{ATE}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\widehat{\mathsf{SE}}_{\mathsf{ATE}}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}}$$

[1] The standard deviation of the sampling distribution.

Null hypothesis and critical values

The test of differences in means with large N relies on the t-test that we know from an intro stats class.

- Null and alternative hypotheses:

$$H_0 : E[Y_1] = E[Y_0], \quad H_1 : E[Y_1] \neq E[Y_0]$$

- Under the null hypothesis, the t-statistic is distributed according to a standard normal distribution

$$t \xrightarrow{d} N(0,1)$$

- We reject the null hypothesis H_0 against the alternative H_1 at the 5% significance level if
  - if $|t| > 1.96$ or, equivalently,
  - if the p-value < 0.05.
- 95% confidence interval for the $\tau_{\text{ATE}} : (\bar{Y}_1 - \bar{Y}_0) \pm 1.96 * \widehat{\text{SE}}_{\text{ATE}}$

P-values and confidence intervals

P-value

The probability of observing a test-statistic as large or larger (in absolute terms) than the one we observe, under the assumption that the null hypothesis is true.

Confidence interval

If we repeated the experiment many times, the confidence intervals we construct would include the true ATE in 95% of replications. In other words, the confidence interval has a .95 probability of *bracketing* (or: containing) the true ATE.

T-test for the difference in means (example)

Outcome	ATE
Consumption	2.19
Assets	0.27
Food	2.60
Livestock	17.65

1. Are these treatment effects significantly different from zero?
2. What are the plausible values for the true ATEs, given our data?

T-test in R (by hand)

$$t_{\text{ATE}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\widehat{\text{SE}}_{\text{ATE}}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} = 3.9$$

```r
y_bar_1 <- mean(aid$livestock[aid$treatment == 1])
y_bar_0 <- mean(aid$livestock[aid$treatment == 0])
```

```r
n_1 <- sum(aid$treatment == 1)
n_0 <- sum(aid$treatment == 0)
```

```r
sigma_1 <- var(aid$livestock[aid$treatment == 1])
sigma_0 <- var(aid$livestock[aid$treatment == 0])
```

```r
st_err <- sqrt(sigma_1/n_1 + sigma_0/n_0)
```

```r
t_statistic <- (y_bar_1 - y_bar_0)/st_err
t_statistic
```

```
## [1] 3.903147
```

Can we reject the null hypothesis of no effect at the 95% conf. level? **Yes!**

Confidence Interval in R

$$(\bar{Y}_1 - \bar{Y}_0) \pm 1.96 * \widehat{\mathsf{SE}}_{\mathsf{ATE}}$$

```
lower_conf_int <- (y_bar_1 - y_bar_0) - 1.96 * st_err
upper_conf_int <- (y_bar_1 - y_bar_0) + 1.96 * st_err

lower_conf_int
```

```
## [1] 8.788918
```

```
upper_conf_int
```

```
## [1] 26.51921
```

T-test in R

```r
t.test(x = aid$livestock[aid$treatment == 1],
       y = aid$livestock[aid$treatment == 0])
```

```
...
## t = 3.9031, df = 7743.3, p-value = 9.576e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.787695 26.520433
## sample estimates:
## mean of x mean of y
##  89.78622  72.13215
...
```

- **Hypothesis test** → we can be *very confident* that the true effect of the treatment *on livestock income* is not zero.
- **Confidence intervals** → the effect of the aid programme on livestock income was between \$8.8 and \$26.5

T-test in R

```r
t.test(x = aid$consumption[aid$treatment == 1],
       y = aid$consumption[aid$treatment == 0])
```

```
...
## t = 1.5751, df = 7542.9, p-value = 0.1153
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5346328  4.9065788
## sample estimates:
## mean of x mean of y
##  80.06653  77.88056
...
```

- **Hypothesis test** → we *cannot be confident* that the true effect of the treatment *on consumption* is not zero.
- **Confidence intervals** → the effect of the aid programme on consumption was between -$0.5 and $4

p-Values for all outcomes

Outcome	ATE	p-value	Upper CI	Lower CI
Consumption	2.19	0.12	4.96	-0.59
Assets	0.27	<0.00	0.31	0.22
Food	2.60	<0.00	0.04	0.01
Livestock	17.65	<0.00	26.77	8.53

1. We can reject the null hypothesis of no-effect for Assets, Food, and Livestock, but not for Consumption
2. The confidence intervals provide plausible values for the effect sizes in the population

Regression as a difference-in-means estimator

Recall the linear regression model:

$$E[Y_i] = \alpha + \beta D_i$$

When D is binary, we can express the regression coefficients as

$$\begin{aligned} E[Y_i|D=0] &= \alpha \\ E[Y_i|D=1] &= \alpha + \beta \end{aligned}$$

By randomization:

$$\begin{aligned} E[Y_i|D=0] &= E[Y_{0i}] = \alpha \\ E[Y_i|D=1] &= E[Y_{1i}] = \alpha + \beta \end{aligned}$$

And so:

$$\begin{aligned} E[Y_{1i}] - E[Y_{0i}] &= E[Y_i|D=1] - E[Y_i|D=0] \\ &= (\alpha + \beta) - (\alpha) \\ &= \beta \end{aligned}$$

Regression as a difference-in-means estimator

Implication: When D is binary, the linear regression coefficients provide:

1. $\alpha =$ Estimated average potential outcome under control
2. $\alpha + \beta =$ Estimated average potential outcome under treatment
3. $\beta =$ Estimated average treatment effect

Regression in R

```r
summary(lm(livestock ~ treatment, data = aid))
```

```
...
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   72.132      2.815  25.622  < 2e-16 ***
## treatment     17.654      4.653   3.794 0.000149 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 216.4 on 9320 degrees of freedom
## Multiple R-squared:  0.001542,   Adjusted R-squared:  0.001435
## F-statistic:  14.4 on 1 and 9320 DF,  p-value: 0.000149
...
```

Covariates and Random Assignment

Uses of covariates in experiments

In contrast to observational studies, in experiments we don't need to "control" for other factors in order to get an unbiased estimate of τ_{ATE}. **Why?**

Because, in expectation, there is no selection bias.

Three uses of covariates in experimental settings:

1. Randomization ("balance") checks
2. Uncertainty reduction
3. Heterogeneous treatment effects

Balance checks

- Randomization ensures that all pre-treatment covariates (observable, unobservable) are balanced in expectation (across randomizations).
- But recall our definition of selection bias:

$$E[Y_{0i}|D=1] \neq E[Y_{0i}|D=0]$$

  - We can never test for selection bias directly. **Why?**
  - Because $E[Y_{0i}|D=1]$ cannot be observed.

Balance tests are often used to check for differences in **observable** pre-treatment covariates between the treatment and control groups:

- covariate-by-covariate comparison of means (e.g. via t-tests)
- multivariate regression of treatment status (DV) on all covariates

Balance checks

Basic idea:

- Compare the difference in **covariate** means between treatment and control groups
- If there is no difference, on average, this suggests randomization has been successful
- If there are differences, on average, this suggests randomization has been unsuccessful (or you have been unlucky!)

Covariates in the Aid example

The aid data includes a number of covariates:

1. `assets_baseline` - assets of the family before the experiment
2. `food_baseline` - food security of the family before the experiment
3. `consumption_baseline` - family's total monthly consumption before the experiment
4. `country` - country in which the family lives

Balance checks (via t-test)

```r
t.test(aid$assets_baseline[aid$treatment == 1],
aid$assets_baseline[aid$treatment == 0])$p.value
```

```
## [1] 0.1326054
```

```r
t.test(aid$food_baseline[aid$treatment == 1],
aid$food_baseline[aid$treatment == 0])$p.value
```

```
## [1] 0.8617334
```

```r
t.test(aid$consumption_baseline[aid$treatment == 1],
aid$consumption_baseline[aid$treatment == 0])$p.value
```

```
## [1] 0.3174575
```

→ Large p-values indicate there are no systematic differences between treatment and control groups, which suggests randomization was successful.

Balance checks (via multiple regression)

```r
summary(lm(treatment ~ consumption_baseline + food_baseline +
                       assets_baseline, data = aid))
```

```
...
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.614e-01  7.067e-03  51.138   <2e-16 ***
## consumption_baseline  6.810e-05  7.452e-05   0.914    0.361
## food_baseline        -2.260e-04  5.021e-03  -0.045    0.964
## assets_baseline       7.193e-03  4.706e-03   1.529    0.126
...
```

→ None of the covariates is a strong predictor of the treatment, which again suggests that the treatment and control groups do not differ systematically from each other, on average.

Covariates for uncertainty reduction

- Controlling for other variables in experimental regression can generate more precise treatment estimates
- If you have a variable X_i that predicts Y_i then controlling for X_i will reduce the residual variance
  - (So long as X_i is not correlated with D_i, which it will not be in an experiment!)
- **Implication**: Controlling for X_i may lead to smaller standard errors for D_i
- See MM p. 95 – 97 for more on this

Uncertainty reduction in the Aid experiment

```r
simple_model <- lm( livestock ~  treatment, data = aid)

multiple_model <- lm(livestock ~  treatment + consumption_baseline +
                                  assets_baseline + food_baseline +
                                  country,
                     data = aid)
```

Uncertainty reduction in the Aid experiment

Comparing simple and multiple regression

	livestock	
Treatment	17.7 (4.7)	18.2 (4.1)
Consump. baseline		−0.01 (0.03)
Asset baseline		19.5 (1.8)
Food baseline		4.1 (1.9)
Ghana		−66.0 (7.3)
Honduras		−62.2 (7.3)
India		−40.6 (9.0)
Packistan		−65.4 (8.8)
Peru		212.6 (7.7)
Intercept	72.1 (2.8)	71.1 (6.7)
Observations	9,322	9,322
R^2	0.002	0.3

Heterogeneous treatment effects

In many settings, we will be interested not only in average treatment effects (ATE), but in **conditional** average treatment effects:

Definition: Conditional Average Treatment Effects

$$\tau_{\mathsf{CATE}_X} = E[Y_{1i} - Y_{0i} | X_i = x]$$

- Many social science theories generate conditional predictions
- Those implementing policies may care about effects on particular subgroups

Heterogeneous treatment effects

One way to estimate treatment-effect heterogeneity is to include **interactions** between the treatment variable D and a covariate X.

Interaction effects

An interaction effect exists if the effect of the treatment D, on the outcome Y, differs for units with different values of a covariate X.

We can build this into our regression models by including the **product** of the treatment indicator and an explanatory variable.

Interaction effects in the Aid experiment

Is the effect of aid greater in Ethiopia than in other countries?

$$Y_i = \alpha + \beta_{\text{Treatment}} D_i + \beta_{\text{Ethiopia}} \text{Ethiopia}_i + \beta_{\text{treatment}\cdot\text{Ethiopia}} (D_i \cdot \text{Ethiopia}_i)$$

```r
ethiopia_model <- lm(livestock ~ treatment * ethiopia, data = aid)
summary(ethiopia_model)
```

```
...
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          75.935      2.924  25.966  < 2e-16 ***
## treatment             6.681      4.928   1.356    0.175
## ethiopia            -48.956     10.493  -4.666 3.12e-06 ***
## treatment:ethiopia  103.214     15.139   6.818 9.82e-12 ***
...
```

Interaction effects in the Aid experiment

```
...
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             75.935      2.924  25.966  < 2e-16 ***
## treatment                6.681      4.928   1.356    0.175
## ethiopia               -48.956     10.493  -4.666 3.12e-06 ***
## treatment:ethiopia     103.214     15.139   6.818 9.82e-12 ***
...
```

What is the ATE for households in Ethiopia?

$$\text{ATE}_{\text{Ethiopia}} = \beta_{\text{treatment}} + \beta_{\text{treatment}\cdot\text{Ethiopia}} = 6.68 + 103.21 = 109.9$$

What is the ATE for households **not** in Ethiopia?

$$\text{ATE}_{\text{Not Ethiopia}} = \beta_{\text{treatment}} = 6.68$$

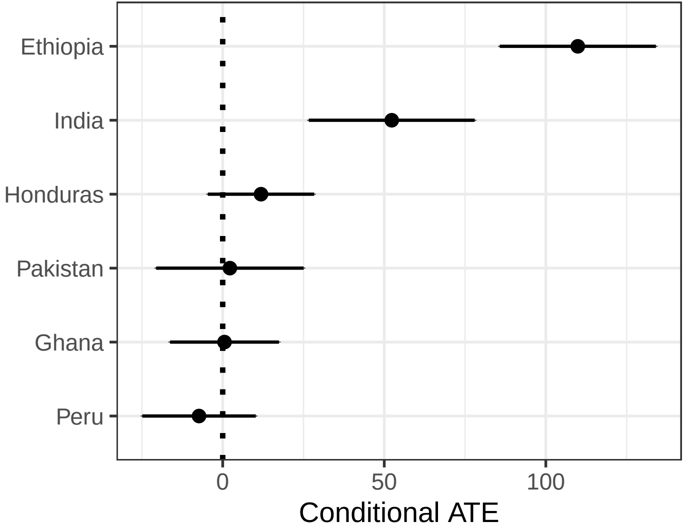
Interaction effects in the Aid experiment

More generally, we can model the ATE for each country, separately:

```r
interaction_model <- lm(livestock ~ treatment * country, data = aid)
summary(interaction_model)
```

```
...
## (Intercept)                 26.979      8.681   3.108  0.00189 **
## treatment                  109.896     12.332   8.912  < 2e-16 ***
## countryGhana               -17.677      9.721  -1.818  0.06902 .
## countryHonduras            -16.433      9.922  -1.656  0.09771 .
## countryIndia               -13.853     12.872  -1.076  0.28187
## countryPakistan            -13.455     12.046  -1.117  0.26406
## countryPeru                265.335     10.136  26.178  < 2e-16 ***
## treatment:countryGhana    -109.342     15.034  -7.273 3.80e-13 ***
## treatment:countryHonduras  -98.039     14.920  -6.571 5.26e-11 ***
## treatment:countryIndia     -57.581     17.994  -3.200  0.00138 **
## treatment:countryPakistan -107.669     16.966  -6.346 2.31e-10 ***
## treatment:countryPeru     -117.226     15.245  -7.689 1.63e-14 ***
...
```

Interaction effects in the Aid experiments



Internal and External Validity

Threats to internal and external validity

- **Internal validity**
  - Are the causal assumptions satisfied in this study? (i.e. does $E[Y_{0i}|D=1] = E[Y_{0i}|D=0]$?)
  - Can we estimate a credible treatment effect for our particular sample?
  - Fails when there are differences between treated and controls (other than the treatment itself) that affect the outcome and that we cannot control for
- **External validity**
  - Can the conclusions we draw be generalised beyond our study?
  - Can we extrapolate our estimates to other populations?
  - Fails when outside the evaluation environment the treatment has a different effect

Threats to internal validity

- Failure of randomization
  - Many heuristics people would use to randomize are not really random
  - E.g. sorting by letter of alphabet of first letter of surname
- Non-compliance with experimental protocol
  - People do not always do what they are told!
  - e.g. Treated individuals refusing treatment
  - e.g. Control individuals taking treatment
  - More on this in week 7
- Attrition
  - Subjects dropping out of the study after randomization
  - If attrition is related to treatment → bias from differences between those who remain in and those who leave

(More) threats to internal validity

- Small samples
  - Do not cause bias, but does lead to imprecision
- Hawthorne effect
  - Subjects behaving differently because they know they are being studied
  - e.g. workers' productivity increasing because they are being watched, not because of the treatment

Threats to external validity

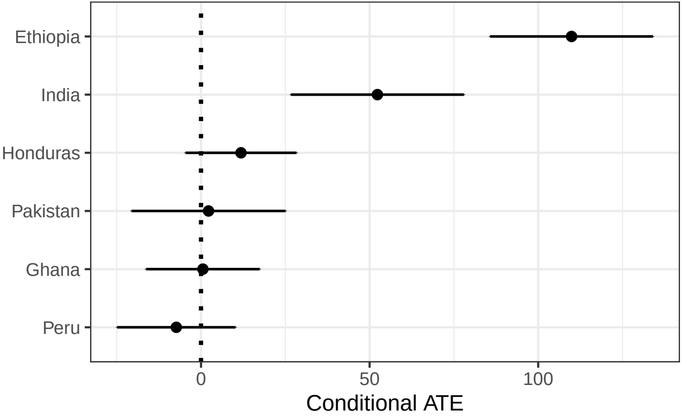
- Non-representative sample
  - Experiments are often conducted on convenience samples
  - Experiment participants may differ from population of interest
  - e.g. Students; MTurk workers; volunteers
- Non-representative treatment
  - The treatment differs in actual implementations
  - e.g. survey experiment about the effects of media priming
- General equilibrium effects
  - Scaling up an experiment may change the environment such that the treatment no longer has the same effect
  - e.g. Increasing government-provided training may decrease employer-provided training

Which matters most?

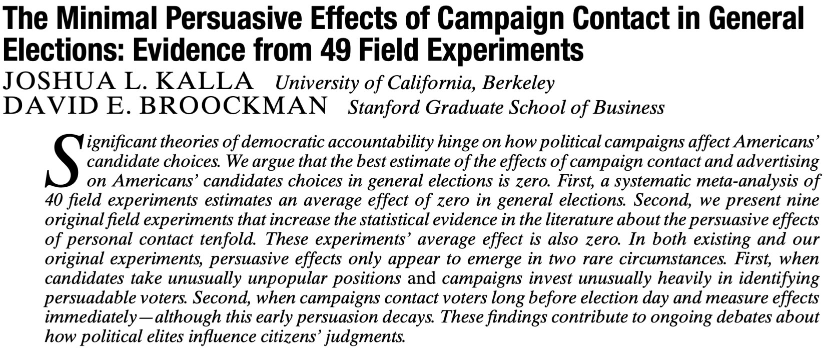
> *"One common view is that internal validity comes first. If you do not know the effects of the treatment on the units in your study, you are not well-positioned to infer the effects on units you did not study who live in circumstances you did not study."*
>
> *–Rosenbaum, 2010, p. 56*

- Randomization ensures internal validity
- External validity can be addressed by comparing many different internally valid studies
- Many of the external validity criticisms of experiments can also be levelled at observational studies

Running multiple experiments for generalisation



Multiple experiments for generalisation



The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments

JOSHUA L. KALLA *University of California, Berkeley*
DAVID E. BROOCKMAN *Stanford Graduate School of Business*

Significant theories of democratic accountability hinge on how political campaigns affect Americans' candidate choices. We argue that the best estimate of the effects of campaign contact and advertising on Americans' candidates choices in general elections is zero. First, a systematic meta-analysis of 40 field experiments estimates an average effect of zero in general elections. Second, we present nine original field experiments that increase the statistical evidence in the literature about the persuasive effects of personal contact tenfold. These experiments' average effect is also zero. In both existing and our original experiments, persuasive effects only appear to emerge in two rare circumstances. First, when candidates take unusually unpopular positions and *campaigns invest unusually heavily in identifying persuadable voters. Second, when campaigns contact voters long before election day and measure effects immediately—although this early persuasion decays. These findings contribute to ongoing debates about how political elites influence citizens' judgments.*

Types of experiments and validity

- **Field experiment**
  - Natural setting, real intervention, real outcomes
- **Survey experiment**
  - Survey setting (normally online), ‘weak’ interventions, survey outcomes
- **Lab experiment**
  - Lab setting (normally online), contrived interventions and outcomes (money often used as an incentive)

Which of these is best for establishing internally valid estimates?
Which of these is best for establishing externally valid estimates?

Conclusion

- Random assignment solves the identification problem in causal inference by balancing treatment and control groups with respect to **observed and unobserved** confounders
- For this reason, randomized experiments are often thought of as the **gold standard** for causal inference
- Analysing experiments is very straightforward: **t-tests and linear regression** are both suitable choices for estimation
- There is potentially a trade off between **external and internal validity** in experimental research