[Textbook Files](#)

[Full Text Online](#)

[Github](#)

[Private Github](#)

[Canvas Pages](#)

Causal Overview

1. What is the causal relationship of interest?
2. What is the ideal experiment that can be run to capture this relationship?
3. What is our strategy to use our sample data to approximate the population?
4. What are you studying, what's the sample, what are the assumptions?

Supervised learning is about using our features to make predictions that are as accurate as possible. Model revisions are done with the intent to improve the model's accuracy (RMSE, F1, etc). Causal inference is about taking the results of what has happened and then looking at variation to understand the relationships that exist among the "result" and the features that lead to the result.

What levers do we pull? What outcome can we expect from this? Why did the algo make this prediction? If we change this feature, what will happen? To evaluate in causal inference, we have to look at the results of changes that we make. What happened? What was expected to happen?

Methods

	When to use	Advantage	Disadvantage
Experiment	When we can	Gold standard	Not always feasible
Matching	Last option	Removes confounding impact of variables outside of treatment	Assumes no other unobserved differences
Panel	When multiple obs for each entity	Resolved unobserved confounds that are time invariant	Cannot deal with time variant missing variables
Difference in difference	Groups growing at similar rates before treatment	Eliminates fixed differences not related to treatment	Hard to have perfect growth rates before
Instrumental variables	When the omitted variable doesn't directly impact Y, only indirectly through one of the IV	Easy to show evidence	Can only see our sample, harder to extrapolate
Regression Discontinuity	When there is a sharp cut-off between the groups	Easy to argument / demonstrate	Only works for the local sub-group

Overview of Methods

IMPORTANT: Observational methods bear stronger assumptions, which often cannot be tested. As such, the assumptions should always be stated explicitly!

Design	When to use	Advantages	Disadvantages
Randomized Experiment	Whenever possible	Gold standard, most powerful.	Not always feasible, ethical or legal.
Matching Techniques	When all other methods are not possible (sometimes in tandem with other methods).	Resolves confounded differences between treatment and control (aside from treatment), post-hoc. Avoids overfitting, model dependence.	Assumes no unobserved differences (often implausible assumption, and it's not testable).
Panel Data Models	When there are repeated observations for each entity in your data.	Resolves unobserved confounds that are time-invariant.	Cannot deal with unobserved confounds that are time-variant.
Difference-in-Differences	If two groups are growing at similar rates before one is "treated" Baseline and follow-up data are available	Eliminates any fixed differences not related to the treatment	Can be biased if trends are not parallel before the "treatment" Ideally have 2 pre-intervention periods of data
Instrumental Variables	If you can identify an exogenous variable that strongly influences the RHS variable of interest, and does not impact the final outcome of interest (except via the RHS variable).	Resolves omitted variables, simultaneity and measurement error all at once.	Good instruments are hard to find. While relevance can be evaluated, the exclusion can not be tested empirically. Has to be argued.
Regression Discontinuity	If an intervention has a clear, sharp (arbitrary) assignment rule	Usually easy to argue and show descriptive evidence that the discontinuity is there and having an effect	Can only look at a local sub-group of the overall sample

L1 - Stats Overview

Random variable

A random variable takes a random value from some distribution

- Discrete - probability mass function - described by prob that the variable takes each of the numbers
- Continuous - probability density function - described by prob that variables takes value within interval

Moments - Expectation, Variance, Skew, Kurtosis

Linear Regression

Estimating the population with the sample that we have - min SSE

Models relationship between DV and linear combo of IV + error

- Confidence interval = range that includes the true mean with __% probability

- Type 1 error = false positive = control with significance (.001 is very low chance of type 1 error)

- R^2 = % of change in DV accounted for by our IV

  - Correct model spec is not necessarily the one with highest R^2, its the one which most closely reflects the relationship among the variables

OLS properties

- Unbiased - the exact coef value that we find probably won't be exactly the same as the "true" value, but on average our estimate of the coef should equal the true value
- Each coef estimate's distribution is normal

OLS assumptions

- Exogeneity - error is not correlated with any of our IV's. If this does not hold, OLS properties don't hold

Log Action	Change in Y	Change in X
None	__ unit	1 unit
Y ~ log(X)	__ %	1 %
log(Y) ~ X	__ % pt	1 unit
log(Y) ~ log(X)	__ %	1 %

L2 - Correlation and Causation

[Mochen Yang Notes](#)

Threats to causal inference

1. Sample selection bias - sample doesn't represent the population we care about, can't generalize
2. Endogeneity - the error is related to one of our variables
   1. Omitted variable bias - we haven't accounted for everything, "harm" becomes greater as data increases - to be an omitted variable, variable needs to be correlated with both X and Y
      - Multico doesn't bias the coef, just increases the standard error of our estimate and makes the variable appear to be insignifcant. We prefer to err on the side of committing multico instead of committing omitted variable bias.
   2. Simultaneity bias - X and Y cause each other, there's something else "behind" them
   3. Measurement error - could be random or systematic - if random error is related to one of the variables, it doesn't just "go away". Our model will not actually reflect reality.

Requirements for Causality

1. Correlation - the "prereq" for causal inference, but is not enough on its own
2. Temporal precendence - X has to lead to Y
3. Free from the threats above (ie free from endogeneity)

L3 - Experiments

[Mochen Yang Notes](#)

[Code](#)

Randomized Experiments

The biggest concern in experiments is that was have a **sample selection bias**. We get rid of this by using random assignment. If our subjects are not actually randomly assigned though, we cannot make a claim that our treatment "caused" something.

Objective is to get rid of any kind of confounding explanations for why Y is changing. We are looking to control for all variables to isolate the impact of the variable of interest in our response.

Experiments should be used when we have comparable subjects, we can randomize among them, and we have an outcome of interest that we can measure reliably. We can check how good our randomization was by comparing the *other* features between the groups and seeing if there are statistical differences.

- Simple - completely random
- Block - sort subjects into groups, and then randomize

Concerns

- Lack of data - unreliable coef estimates, especially if we are looking to pick up on small differences
- Subject interference - treatment spillover, someone becoming aware of the experiment, social desirability bias, etc
- Lack of valid control group

More than one treatment

Example - red or green + square or circle

Running an experiment with 2 potential treatment will provide us with 4 possibilities. We can look at 1 treatment alone, the other treatment alone, the interaction, and the control group. If the coef for one of the treatment alone is positive **and** the interaction coef is positive, it means that the addition of the other treatment had an **amplifying** impact.

L4 - Matching Methods

Reference textbook section 3.3.1

Issues with controlling for factors

- Maybe we don't actually have all of the factors (low stat power)
- Maybe the factors don't impact DV linearly (model dependency)
- Might have "sub-groups" with different impacts

The objective of matching is to create an "experiment" from an existing dataset when there wasn't actually an experiment being run. We are looking to find pairs that are as similar as possible on everything **except** for the treatment that they receive so that we can simulate what a random experiment would have been.

We only want to use the points that have a match, and then throw away the other observations. Matching is good for removing confounded differences between the treatment vs control group, but it also assumes that there is no unobserved differences which is often implausible.

We should verify our results by using multiple techniques. We should also try different hyperparameters for each model to make sure that our findings are robust.

Propensity Score Matching

We are **not** matching on characteristics, but on prob of being in the positive group.

- We first run a logic or probit model to see whether a point will get the treatment (positive group)
- We then set a specific threshold for what we consider to be "similar" for a pair. The pair uses the "actual" results for possible pairs, and then uses the "expected" results from our model to compare.
- We then find the difference between possible pairs, and only keep the pairs where the difference is below this threshold. We can use either *with replacement* or *without replacement* if we are okay with re-using an observation.

Use when we have plenty of variables, used as a way to simulate randomization

Issues - (1) Still possibility of OVB, (2) might not be any good matches, (3) logit/probit are limited

Coarsened Exact Matching

Instead of matching on propensity score or matching exactly on variable differences, we match on the comparison of **discretized** versions of the variables.

- Keep discrete and categorical, convert continuous into bins.
- We must have **exact** matches on everything besides the treatment in order to create the pair.

L5 - Panel Data

Cross section: one observation per entity

Panel data: multiple observations and multiple entities (over time or other dimensions)

Example: Observing how a company performs over different store locations

Motivation: What if we are missing a confounding variable?

We have a **time invariant** missing variable when the missing variable is stable over time for each of the entities. We are able to make the impact of this missing variable disappear by using the techniques below. (Note that the techniques below only apply when we have OLS linear regression.)

Fixed Effect

We can use this technique even if the omitted variable is related to any of our IV. We turn the missing variable into a value instead. This adds a new intercept for each entity.

We cannot measure the effect that this missing variable has, all we can do is try to get rid of the impact on the other variables. We also cannot handle if the variable is not holding stable over time.

Method 1: De-mean data (within)

1. Looks at each entity
2. Finds the average of each column
3. Adjusts the column by taking the difference of the value and the average

Method 2: Add dummy

This provides the exact same results as method 1. We are looking to see the impact of "being this person" on the outcome.

1. Creates a new dummy variable for every single entity in the data

   100 people would mean 99 new variables

Method 3: First differencing

1. Looks at each entity
2. Finds the lag of each column
3. Adjusts the column by taking the difference of the value and the lag

Random Effect

We can use this when we believe the omitted variable is not related to any of our IV.

We turn the missing variable into a random variable instead. We then add the variable for each entity onto the error term. We must assume that the missing variable is **not** correlated to any of the X. This approach will provide us with more precise regression estimates, but will provide incorrect coefficients is the assumption does not hold.

Hausman

We should use the fixed effect model if the hypothesis is rejected, but we can use random effects if hypothesis is accepted.

L6 - Difference in Difference

Here we are looking to address the same issue as panel data, but we want to be able to account for missing variables that **change over time**.

We need to have some people who **are** treated which leads to the change over time, and some who are not. We compare the average change in the control vs the average change in the test. Both groups need to have an omitted underlying factor that is leading to them changing over time.

Example: What effect does price change have on demand? We have an app on Android and Apple.

We could give one platform the treatment and the other as the control. We then look at the change that occurs in the control group, and then factor that in before computing the change in the test. The remaining change is the impact that we care about. We have to assume that the missing underlying factor (D1) that led to the difference in Apple has the same effect on Android.
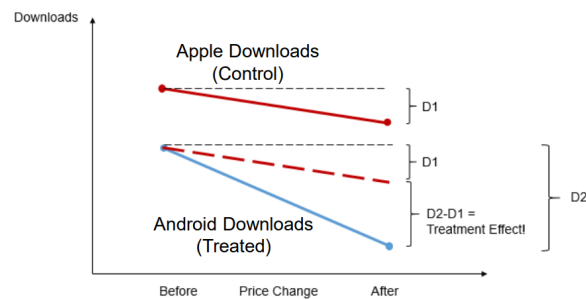
Assumptions

- Parallel trend - control & test group need to be changing in the same way before out treatment
- No interference - treated subjects can't be influencing our control subjects

Model Specification and Estimation

After - binary, 1 if after the treatment

Treat - binary, 1 if member of the treatment group

L7 - Instrumental Variables

We are looking to find a way to isolate the effect of the IV we care about and "exclude" the omitted variable it is correlated with. We do this by adding in an instrumental variable that is correlated with the IV, but is not correlated to the DV. For an instrumental variable to be meaningful, it has to hold for the whole population. Our "new" set of IV will then be free from OVB.

Requirements

- Relevance - Instrument has to be correlated with our IV
- Exclusion (more important) - Omitted can't be correlated with error

Evaluation

- First stage F-stat should be at least 10 for strong instrumental variable (conceptually similar to R^2)
- Hausman Test - tests if OLS or IV regression is more efficient
- Sargan Test - tests whether any of our IV is unnecessary, doesn't say which one

L8 - Regression Disc

Example: If an offered unemployment is too good, it may de-incentivize someone to look for work. In an ideal experiment, we have a cutoff of 50-years-old - older people receive better benefits. We then compare the people around this threshold, ages 46-54 (local area). The smaller the local area we choose, the greater stat significance. However, harder to extrapolate.

$$Y = \beta 0 + \beta 1 * Z + \beta 2 * (X - Xc) + error$$

$$Number\ of\ days = \beta 0 + \beta 1 * Dummy\ of\ Treatment + \beta 2 * (Age - Age\ Cutoff) + error$$

Even if we decide to include an interaction term of our Z * X, we only care about the interpretation of Z (the dummy of the cutoff). The interaction term is providing us with the change of the degree of slope.

Requirements

- Needs to be arbitrarily assigned, subjects cannot be placing themselves
    - Self selection: If the people on one side are different from the other side in their characteristics, we have an issue. We can use matching to filter out unhelpful people.
- DV must be continuous and smooth function of IV, especially around cutoff
    -

Regression disc doesn't really work for time series, especially if the effect ramps up over time.

We can also run "fuzzy" disc, which means crossing the threshold makes it more likely you'll get the treatment.