Causal Machine Learning

Fernando Rocha Urbano

Autumn 2024

1 Lecture 1

We start with binary treatment and consider the following random variables in our analysis:

- Treatment: $T \in \{0, 1\}$
- Outcome: $Y(1), Y(0)$
- Covariates: X

1.a Estimands

- ITE (Individual Treatment Effect): impact of the treatment on person i. Not identified (means that you cannot observe directly).
- CATE (Conditional Average Treatment Effect): average treatment for individuals with specific realizations of observables.
- ATE (Average Treatment Effect): average treatment for all individuals.
- ATT (Average Treatment Effect on the Treated): average treatment effect on the observations treated.

1.b ATE

$$ATE = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] := \tau$$

$$\begin{aligned}\hat{\tau} &= \bar{Y}_1 - \bar{Y}_0 \\ &= \frac{1}{n_1} \sum_{treated} Y_i - \frac{1}{n_0} \sum_{control} Y_i\end{aligned}$$

Under regularity conditions this is identified (which include independence of observations). The estimate of $\hat{\tau}$ requires overlap, meaning that $\mathbb{P}[T = 1] > 0$. Otherwise, the following equation does not hold.

$$\bar{Y_1} = \frac{1}{n_1} \sum_{treated} Y_i \xrightarrow{P} \frac{\mathbb{E}[YT]}{\mathbb{P}[T = 1]}$$

For us to have a causal effect, we need the following conditions:

- Positivity/Overlap: $\mathbb{P}[T = 1] > 0$.
- SUTVA: only the treatment matters to define Y (if not, you need covariates).
- Consistency: observed outcome matches treatment "assignment".

If all conditions are met:

$$\hat{\tau} = \bar{Y_1} - \bar{Y_0} \xrightarrow{P} \mathbb{E}[Y(1)|T = 1] - \mathbb{E}[Y(0)|T = 0]$$

Reminder: $Y(k)$ is a potential outcome notation.

- $Y(1)$: outcome if individual receives the treatment (regardless if he actually received or not).
- $Y(0)$ outcome if individual does not receive the treatment (regardless if he actually received or not).

2 Lecture 2

2.a Causal Inference

With SUTVA, only your treatment matters to define the outcome. From that and other conditions, we get the average treatment effect. Furthermore, so far we have used randomization to break selection bias.

Randomization allows for the following:

$$\mathbb{E}[Y|T = 1] = \mathbb{E}[Y(1)|T = 1]$$

Nonetheless, most of the times, randomization is not possible. For instance, sick people want medical care, but not health people.

In situations where randomization is not possible, we cannot be sure that the expected value for people who took the treatment is equal to the expected value of taking the treatment.

More specifically:

- $\mathbb{E}[Y|T=1]$: expected value of taking the treatment.
- $\mathbb{E}[Y(1)|T=1]$: expected value of taking the treatment for people who took the treatment.

2.b Using Regression

We run a regression with dummy variables. If there is overlap and randomization, $\beta = \tau = ATE$.

$$Y = TY(1) + (1-T)Y(0) = \alpha + \beta T + \epsilon$$

In this case, the β_1 is exactly the difference in means:

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

Let's revisit the necessary assumptions for this model to work.

2.c Assumptions

2.c.1 Define μ_t and ϵ_t via $Y(t) = \mu$

Everyone starts from the same μ_t and has an extra ϵ

Now, we can view:

$$Y(t) = TY(1) + (1-T)Y(0)$$

$$Y(t) = \mu_0 + T(\mu_1 - \mu_0) + \epsilon_0 + (\epsilon_1 - \epsilon_0)T$$

$$Y(t) = \alpha + T\beta + \epsilon$$

What do need for that to work:

1. Rank: Variance of $T > 0$, meaning $P[T=1] > 0$ and $P[T=0] > 0$, because the variance is $p(1-p)$.
2. Orthogonality: we would like error and treatment to be uncorrelated, or preferentiabily with mean 0: $\mathbb{E}[\epsilon T] = 0$ or preferentiabily $\mathbb{E}[\epsilon|T] = 0$ (the second one implies the first one). The second one means that if you take any T, you learn nothing from the errors (ϵ), which is a great thing: this

should happen! This is loosely related to causal inference because we need a condition like this to get causal inference (the proof for this one will be in HW1).

If I have rank and orthogonality, the $\hat{\beta}$ converges to β.

$$\hat{\beta} \xrightarrow{P} \beta = \tau = \mathbb{E}[\tau(x)]$$

2.d Using Regression with Covariates

Running a regression with covariates is a very common practice. Even in tretment experiments, people add covariates because (i) we need to improve precision (ii) deals with heterogeneity (more important).

OLS under randomized experiments and pre-treatment X still recovers ATE and improves precision, even when it is misspecified.

2.d.1 Model

Now I say that μ_0 is also dependent on X.

$$Y = TY(1) + (1 - T)Y(0)$$

$$Y = \mu_0(X) + (\mu_1(X) - \mu_0(X))T + \epsilon_0 + (\epsilon_1 - \epsilon_0)T$$

$$Y = \alpha(X) + \beta(X)T + \epsilon$$

This shows what happen if I run a regression only dependent on $\hat{\beta}$.

2.d.2 Frisch-Waugh-Lovell Theorem (First Law of ...)

Run a regression to find out the value of a single parameter.

Given:

$$Y = \alpha + \beta_t T + \beta X + \epsilon$$

It is made in a two step procedure:

1. $T = \alpha_1 + \beta_1 X + \epsilon_1$
    - Regress the variable of interest on all other control variables and obtain the residuals.

- These residuals represent the part of the variable of interest that is orthogonal to (uncorrelated with) the control variables.

2. $Y = \alpha_2 + \beta_2 \epsilon_1 + \epsilon_2$
   - Regress the dependent variable on the residuals obtained from Step 1.
   - The coefficient (β_2) on these residuals will be the same as the coefficient on the variable of interest in the full multiple regression model.

2.d.3 Matrix format (need review)

$$\hat{\beta} = (T' M_x T)^{-1} (T M_x Y)$$

Best linear predictor of T:

$$M_x T = T - X(X'X)^{-1} XT$$

This relates to the correlation between $corr(X, T)$.

We know that:

$$\sqrt{n}(\bar{Y}_1 - \bar{Y}_0 - \tau) \to N(0, V)$$

$$\sqrt{n}(\hat{\beta} - \tau) \to N(0, \Omega)$$

We know that $\Omega \leqq V$. This happens because X is information that helps to reduce the variance of the treatment in the limit.

The magic of all that is that $\hat{\beta} = ATE$. The effect of T on Y is defined by β (if enough conditions are met).

Even when we add X, we are assuming that the relationship is linear between Y and X.

$$\beta = \mathbb{E}[\beta(X)], \text{only for sure with randomization}$$

(Simpsons Paradox will be in Homework 1)

2.d.4 Regression with Interactions: Best Practice

When we have covariates, the best practice is to run a regression wth Interactions ($\delta(X - \bar{X})$).

$$Y = \alpha + \beta T + \gamma X + \delta(X - \bar{X})T + \epsilon$$

If the functions are linear also with respect to X, this gives the conditional average treatment (CATE).

Regardless of linearity, we will still recover the average treatment effect if we have pretreatment covariates.

This one is always more efficient (variance being smaller and have them estimating the same thing - being unbiased (or assyntotically unbiased) and consistent for the same thing).

Nonetheless, this still leads to some heterogeneity.

When running such a regression, we assume linear approximation and linear relationship.

2.e X Post Treatment Problem

If we have X as post treatment, we have to deal with other problems, even when we have randomization.

Assuming randomization of treatment:

$$\begin{aligned}
\bar{Y}_{1,1} - \bar{Y}_{0,1} \xrightarrow{P} &= \mathbb{E}[Y|X=1,T=1] - \mathbb{E}[Y|X=1,T=0] \quad \text{(by LLN)} \\
&= \mathbb{E}[Y(1)|X(1)=1,T=1] - \mathbb{E}[Y(0)|X(0)=1,T=0] \quad \text{(def. of Y and X)} \\
&= \mathbb{E}[Y(1)|X(1)=1] - \mathbb{E}[Y(0)|X(0)=1] \quad \text{(because treatment is randomized)}
\end{aligned}$$

This changes if X is post treatment: which means, is the treatment causing the X.

- $X(1)$: X of people when they receive the treatmetn.
- $X(0)$: X of people when they receive the treatmetn.

In the third line, we are able to remove the dependency on treatment because the treatment is randomized. The expectation of $Y(t)$ for any t does not depend on whether $T = 1$ or $T = 0$.

The $X(t)$ remains on the final line because the dependence we have a difference in expected potential outcome based solely on the value of X under treatment versus control.

Therefore, X can introduce bias if $X(1)$ and $X(0)$ represent different selection processes. We call this X pos treatment.

When this happens:

$$
\begin{aligned}
\bar{Y}_{1,1} - \bar{Y}_{0,1} \xrightarrow{P} &= \mathbb{E}[Y(1)|X(1)=1] - \mathbb{E}[Y(0)|X(0)=1] \quad \text{(Now add and subtract)} \\
&= \mathbb{E}[Y(1)|X(1)=1] - \mathbb{E}[Y(0)|X(0)=1] - \mathbb{E}[Y(0)|X(1)=1] + \mathbb{E}[Y(0)|X(1)=1] \\
&\text{(Reorganize terms)} \\
&= \mathbb{E}[Y(1)|X(1)=1] - \mathbb{E}[Y(0)|X(1)=1] - \mathbb{E}[Y(0)|X(0)=1] + \mathbb{E}[Y(0)|X(1)=1] \\
&= (\mathbb{E}[Y(1)|X(1)=1] - \mathbb{E}[Y(0)|X(1)=1]) - (\mathbb{E}[Y(0)|X(0)=1] + \mathbb{E}[Y(0)|X(1)=1]) \\
&= (\text{ATE for } X(1)=1) - (\text{Selection bias into } X=1)
\end{aligned}
$$

If the X is post treatment, by including the interactions I will be doing worst.

Let's assume that X is post treatment.

Let's assume that X is binary variable.

Doing the differences in mean by treatment:

3 Lecture 3

3.a Bad Control Example

Current situation: linear regression with covariates. The X is a dummy variable so far and pos-treatment (not orthogonal to the treatment).

Example:

$$\text{wage} = \alpha + \beta\text{gender} + \gamma' X + \epsilon$$

If other X have influence in gender, $\mathbb{E}[Y(1)|X(1)=1] - \mathbb{E}[Y(0)|X(0)=1]$ does not represent the ATE.

3.b Observational Data

Observational Data: when we cannot randomize treatment. This is often the case, since we cannot have randomize treatment.

When is a regression causal?

We will need three things:

- Overlap: plenty of people in treatment and control (depends on the data - assumption about the population, but we only see the sample).
- Selection on Observables: conditional independence (depends on the data).

  - When T is randomized, that is always true: $Y(1), Y(0) \perp T$
  - When T is not randomized (observational data), we need at least $Y(1), Y(0) \perp T|X$. Meaning that $\mathbb{E}[Y(t)|T = s, X = x] = \mathbb{E}[Y(t)|X = x]$
  - If that is not true, we have selection bias: $\mathbb{E}[Y(0)|T = 1] \neq \mathbb{E}[Y(0)|T = 0]$
- Correct specification (depends on the model - should the model be linear?)

Some assumptions are testable and some are not testable.

Correct specification can be tested: we can do the test using the data. For instance, we can test if the model should be linear or quadratic.

$$Y = \alpha + \beta X + \epsilon$$

$$Y = \alpha + \beta X + \gamma X^2 + \epsilon$$

We can either:

- Check if $\gamma \neq 0$
- Which model gives the best prediction.

Therefore, it is testable!

3.b.1 Selection on Observables

For instance, a Houseman test, RESET test (use the residual), etc...

The first two assumptions are untestable.

The bias of observable data (people select their own treatment based on expected benefits). If that is the case:

$$\mathbb{E}[Y(0)|T = 1] \neq \mathbb{E}[Y(0)|T = 0]$$

Selection on observational: T is not randomize but it is as good as if T is randomize. They are not independent of T, but they are independent of T conditional on X. Meaning $Y(1), Y(0) \perp T|X$. If we have $Y(1), Y(0) \perp T$.

We can think that:

$$Y(1), Y(0) \perp T|X \quad \rightarrow \quad \mathbb{E}[Y(t)|T = s, X = x] = \mathbb{E}[Y(t)|X = x]$$

Attention: this does not hold for the variance of the moments. Transformations on Y might make the new $Y(1), Y(0) \not\perp T|X$. For instance, if $Y_{new} = \ln(Y)$, not necessarily will hold that $Y(1), Y(0) \perp T|X$

Missing at random: means that selection on observational holds.

3.b.2 Example of Selection on Observational

Regress wages on education with covariates.

The problem is that people select into different levels of education because of the covariates. That is a problem because there is a causal dependence of features. But, if I am able to put every covariate that explains education, it is as good as having a randomized experiment. We only need to put all the covariates that affect the treatment (education) and the target variable (wage).

$$ATE = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y(1)|X] - \mathbb{E}[Y(0)|X]]$$

$\mathbb{E}[Y(1)|X]$ is identified.

3.b.3 Overlap

Overlap is necessary because we need to observe X in both states because, otherwise it cannot be orthogonal conditional on X.

Because of selection on observables:

$$\begin{aligned}\mathbb{E}[YT|X] &= \mathbb{E}[Y(1)T|X] \\ &= \frac{\mathbb{E}[Y(1)|X]\mathbb{E}[T|X]}{\mathbb{E}[T|X]}\end{aligned}$$

The previous require overlap in order for $\mathbb{E}[YT|X]$ to exist (this is still working with non parametric models).

3.c Correct Specification

For a linear model to hold:

$$\mathbb{E}[Y|T = 1, X] = \alpha_1 + \beta_1 X$$

and:

$$\mathbb{E}[Y|T = 0, X] = \alpha_0 + \beta_0 X$$

If the intercepts and slopes are different:

$$\mathbb{E}[Y(1) - Y(0)|X = x] = \tau(x) = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)X$$

CATE is the $\tau(x)$ at a specific X.

The specification of the linear model has to be true for that to work.

When T is randomize, we do not even need a linear specification to be true in order to get ATE (because $T \perp X$).

This can also be viewed in the OLS estimator:

$$\hat{Y}_i = \hat{\alpha} + \hat{\tau}T_i + \hat{\gamma}_1' X_i + \hat{\gamma}_2'(X_i - \bar{X})T_i$$

If the model is correctly specified:

$$Y_i = \alpha + \tau T_i + \gamma_1' X_i + \gamma_2'(X_i - \bar{X})T_i$$

$$Y(t) = \mu(t, X) + \epsilon_t, \quad \text{with} \quad \mu(t, X) = \alpha + \gamma_1' X + T(\tau + \gamma_2'(X - \mathbb{E}[X]))$$

In this scenario:

- $\hat{\tau} \xrightarrow{P}$ ATE
- CATEs: $\tau(x) = \tau + \gamma_2' x$

3.d Heterogeneous Effects

Our problem is that different people have different treatment effects.

$$CATE = \mathbb{E}[Y(1) - Y(0)|X = x] = \tau(x)$$

The more X we have, the more fine grained it becomes.

In the limit, we have that for person i:

$$Y_i = \alpha_i + \beta_i T_i$$

We can think of that as:

$$\alpha_i = Y_i(0)$$

$$\beta_i = Y_i(1) - Y_i(0) = ITE$$

For that to hold:

$$cov(\alpha_i, T|X) = 0$$

$$cov(\beta_i, T|X) = 0$$

We can't really estimate this well, but in the end, that is my ultimate goal. This would allow me to know the treatment effect for each individual person.

In practice, we are able to know the treatment effect for groups of individuals.

Question for later: can I measure how much the CATE represents the actual individual treatment effect?

For the ITE to hold when we do not have X:

$$cov(\alpha_i, T|X) = 0$$

$$cov(\beta_i, T|X) = 0$$

If this is true we can do the following:

$$\begin{aligned}
\mathbb{E}[Y|T, X] &= \mathbb{E}[\alpha_i + \beta_i T|T, X] \\
&= \mathbb{E}[\alpha_i|T, X] + \mathbb{E}[\beta_i T|T, X] \\
&= \mathbb{E}[\alpha_i, X] + \mathbb{E}[\beta_i T|T, X] \\
&= \mathbb{E}[\alpha_i, X] + T\mathbb{E}[\beta_i|T, X] \\
&= \alpha + T\beta(X) \quad \text{(by Chamberlain)}
\end{aligned}$$

Discussion Section 1

Key Properties of Estimators

Unbiased

If its expected value equals to the true parameter value.

$$\mathbb{E}[\hat{\theta}] = \theta$$

The parameter is, on average, correct. Unbiasedness means that in repeated samples, the estimator will be centered around the true value of the parameter. This is a finite sample property.

Unbiasedness is a useful property that we generally want our estimators to have, but sometimes we are willing to trade off some bias for a large reduction in variance of the estimator.

If an estimator is unbiased but not consistent, the variance does not go to zero.

Consistency

An estimator $\hat{\theta}_n$ of parameter θ is consistent if it converges in probability to the true parameter θ as sample size goes to infinity. Consistency is assyntotically.

Recall the definition of convergence in probability. For any $\epsilon > 0$:

$$\lim_{n \to \infty} \mathbb{P}[|\hat{\theta} - \theta| > \epsilon]$$

This is the LLN. For it, you need (i) finite variance, (ii) independence.

Consistent does not imply unbiased (and vice-versa).

- $\hat{\theta}_n = X_1$: unbiased but not consistent.
- $\hat{\theta}_n = \frac{1}{n-2} \sum_{i=1}^{n} X_i$: consistent but not unbiased.
- $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$: consistent and unbiased.
- $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - 5)$: not consistent and not unbiased.
- OLS estimator: only consistent and unbiased if $\mathbb{E}[u|x] = 0$ (among other regularities).

Asymptotic Unbiased

Asymptotic Unbiased if its bias disappears as the sample size grows:

$$\lim_{n \to \infty} \mathbb{E}[\hat{\theta}_n] = \theta$$

Efficiency

Having the minimum variance among a class of estimators.

Frisch-Waugh-Lovell

FWL states that in a regression model with outcome y and multiple regressors X_1 and X_2 (each a matrix of covariates).

Our regression model is:

$$y = X_1\beta_1 + X_2\beta_2 + u$$

We are interested in using data ($\{y_i, X_{1,i}, X_{2,i}\}$) to estimate β_1.

We can obtain the coefficients of X_1 by:

1. regressing X_1 on X_2: obtain $M_{X_2}X_1$
2. regressing y on X_2: obtain $M_{X_2}y$
3. running a regression with the residuals of the two previous regressions: $M_{X_2}y = \beta_1 M_{X_2}X_1$

Items (1) and (2) are often called residualizing or partialing out the effect of X_2.

This is an often-used result because, among other reasons, it helps by taking advantage of dimensionality reduction.

We call M_X the residual matrix or annihilator matrix that projects the orthogonal complement of X:

$$M_X = I - X(X'X)^{-1}X$$

Formal statement

Identification

Refers to the ability to learn the true value of a parameter from the data. Intuitively, it answers: if we had infinite data, could we calculate the value of this parameter uniquely?

Without identification, parameter estimates are meaninless because the data cannot uniquely reveal the true parameter value.

3.d.1 Point Identification

A parameter is point identified if there is exactly one value of θ that is consistent with the observed data.

Partial identification is becoming a more important topic.

4 Lecture 4

4.a Confidence Interval

If I do sampling a infinity amount of times, I expect that the sample populational estimate will be inside the CI $X\%$ of the time, where X is the confidence level of my CI.

4.b Parameter Variance

The treatment effect will have variance that will be probably difference from the $\hat{\beta}$ variance. The variance of $\hat{\beta}$ is how good my estimate is. The variance of the treatment effect is how much the treatment effect changes.

The treatment effect might have a super weird distribution. On the other hand, the estimate of the average treatment might be fairly precise. The opposite can also happen.

4.c Changes in Estimate

$$\hat{\beta} = [0, 1](X'X)^{-1}(X'Y)$$

We can view $\hat{\beta}$ as a function that maps rows and columns into a vector in $\mathbb{R}^p$.

We can think about $\hat{\beta}$ as a $f(w)$ and the X and Y as w.

With this perception, we could even take the derivative of $\hat{\beta}$:

$$\frac{\partial \hat{\beta}}{\partial X, Y}$$

Nonetheless, X, Y are too many numbers. Taking a derivative with respect to so many numbers would be impractical and inconclusive.

Therefore, we think of X, Y as a dataset and we should do the derivative with respect to the dataset.

A particular dataset maps to a $\hat{\beta}$. Any other dataset maps to another $\hat{\beta}$.

In conclusion, we have a tendency to think of $\hat{\beta}$ as a estimate, but we should think about as a function of the dataset. Given a dataset, it will always produce the same result because it is a deterministic function. In a dummy way, it works in the same way as $f(w) = w^2$, which will always produce 4 when $w = 2$.

The difference here is that I want to check a frequency histogram of $\hat{\beta}$ in the output $\mathbb{R}^p$. I expect a normal distribution as a map from multiple datasets to $\hat{\beta}$.

4.c.1 When I talk about datasets, how far apart are they?

The previous idea is also valid for more simple parameteres, like $\hat{\mu}$.
We can represent the dataset by its empirical distribution function.

$$\alpha(F_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \quad \text{where } \alpha \text{ is a function}$$

What if I had F_{n-1}: same dataset, but without one row?

$$\alpha(F_{n-1}) = \frac{1}{n-1}\sum_{i=1}^{n-1} X_i = \bar{X}$$

$$\begin{aligned}\alpha(F_n) - \alpha(F_{n-1}) &= \frac{X_n}{n} + \left(\frac{1}{n} - \frac{1}{n-1}\right)\sum_{i=1}^{n} X_i \\ &= \frac{X_n}{n} - \frac{1}{n}\alpha(F_{n-1})\end{aligned}$$

Therefore, the "derivative" of it:

$$\frac{f(w+\Delta) - f(w)}{w+\Delta-w} = \frac{\frac{X_n}{n} - \frac{1}{n}\alpha(F_{n-1})}{\frac{1}{n}}$$

Now, for F_∞:

$$\frac{\partial \alpha(F_\infty)}{\partial i} = X_i - \mu$$

Which can also be written as:

$$\sqrt{n}(\bar{X} - \mu) = \frac{1}{n}$$

5 Lecture 5

Following the example of the previous lecture, we call α the sample mean:

$$\alpha(F) = \int X \, dF = \mathbb{E}[X]$$

How can we do the derivative of $\alpha(F)$?
The methods to do the derivative is:

- Small change in the dataset (maybe change or add one observation)
- Change the distribution:
  - $F_\varepsilon = (1-\varepsilon)F + \varepsilon G$: we can use the idea of drawing from the right distribution $(1-\varepsilon)$ of the times and from a corrupted distribution every ε times. We want to check that to know how much my parameter changes:

$$\begin{aligned}\alpha(F_\varepsilon) - \alpha(F) &= \int X\, dF_\varepsilon - \int X\, dF \\ &= (1-\varepsilon)\int X\, dF + \varepsilon \int X\, dG - \int X\, dF \\ &= \varepsilon\alpha(G) - \varepsilon\alpha(F) \qquad\qquad = \varepsilon\end{aligned}$$

  - influence functions comes from how much an observation change the parameter.
- Take the derivative with respect to ε:

$$\begin{aligned}\frac{\partial \alpha(F_\varepsilon)}{\partial \varepsilon} &= \frac{\partial}{\partial \varepsilon}\int X\left[(1-\varepsilon)dF + \varepsilon dG\right] \\ &= \frac{\partial}{\partial \varepsilon}\int X(1-\varepsilon)dF(x)dx + \int X\varepsilon dG(x)dx \\ &= \varepsilon\alpha(G) - \varepsilon\alpha(F) \quad \text{(same result)}\end{aligned}$$

5.a What is an influence function?

In statistics, an influence function is a tool used in robust statistics to measure the sensitivity of a statistical estimator to small changes or contaminations in the data. If assesses how an infinitesimal amount of contamination at any point in the data space affects the estimator, providing insights into the robustness of the estimator against outliers.

In a formal distribution:

$$\mathrm{IF}(x, T, F) = \lim_{\varepsilon \to 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}$$

Where $F_\varepsilon = (1-\varepsilon)F + \varepsilon G$

5.b Central Limit Theorem

Asymptotic Normality: according to central limit theorem, under certain conditions, the sum or average of large number of independent random variables will be approximately normally distributed.

Example:

$$\sqrt{n}(\bar{x} - \mu) = \frac{1}{\sqrt{n}} \sum (x_i - \mu)$$

Where $(x_i - \mu)$ can be viewed as $(\alpha(G) - \alpha(F))$

The central limit theorem always work with a:

- w that has expected value of 0. Example: $(\bar{x} - \mu)$.
- w with variance bigger than 0 and smaller than ∞.

For an estimator T, the influence function $\text{IF}(x, T, F)$ measures the effect of a small contamination at point x on T.

The variance of this function over the distribution F gives the asymptotic variance of T:

$$\sigma^2 = \int [IF(x; T, F)]^2 \, dF(x)$$

Asymptotic variance is the variance of an estimator's sampling distribution as the sample size approaches infinity. It quantifies the estimator's variability in large samples and is used to describe its limiting normal distribution for inference.

For asymptotic variance we can apply CLT:

$$\sqrt{n}(T_n - T(F)) \xrightarrow{d} N(0, \sigma^2)$$

5.c Influence Function for OLS

Now, lets look at the solution of the OLS. What is the influence function of the OLS estimator?

The solution for $\hat{\beta}$

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

Our goal right now is to write $\hat{\beta}$ as assyntotically normal (follow the CLT):

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(X_i, Y_i)$$

That must be a function $\phi(X_i, Y_i)$ that has mean 0 and finite variance that allows for us to be able to that.

My way to prove assyntotically normality is to write it in this format.

We cannot get there exactly, but we can get close enough in a way that the difference is o_1. This is the same idea as we have in computer science, meaning that:

$$a_n = O(b_n)$$

Which translates to:

$$\lim_{n\to\infty} \frac{a_n}{b_n} < \infty$$

Now, little o means:

$$a_n = o(b_n)$$

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 0$$

When using little o in statistics, we often mean that such a thing hold with high probability, which is written as o_p.

Another example:

Given:

$$X_i \sim N(\mu, v)$$

$$\bar{X} - \mu = o_p(1)$$

Going back to the original problem:

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(X_i, Y_i) + o_p(1)$$

For it to hold, $o_p(1)$ must go to 0. It must be smaller than $\frac{1}{\sqrt{n}}$. Otherwise it will not work as $n \to \infty$.

Therefore:

$$
\begin{aligned}
\hat{\beta} - \beta &= (X'X)^{-1}(X'Y) - \beta \\
&= (X'X)^{-1}\left[X'(Y - X\beta)\right] \quad \text{(divide by } n\text{)} \\
&= \left(\frac{1}{n}\sum x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum x_i \varepsilon_i\right) \\
&= \frac{1}{n}\sum\left(\frac{1}{n}\sum x_j x_i'\right)^{-1} x_i \varepsilon_i \\
&= \frac{1}{n}\sum M^{-1} x_i \varepsilon_i + \frac{1}{n}\sum(\bar{M}^{-1} - M^{-1}) x_i \varepsilon_i \\
&= \frac{1}{n}\sum_i (M^{-1} x_i \varepsilon_i) + o_p(\frac{1}{\sqrt{n}})
\end{aligned}
$$

We can write it as $o_p(\frac{1}{\sqrt{n}}$ because $\sqrt{n}(\bar{M}^{-1} - M^{-1})\frac{1}{n}\sum(x_i\varepsilon_i)$:

$$
(\bar{M}^{-1} - M^{-1}) = \frac{1}{n}\sum x_i x_i') - \mathbb{E}[xx]^{-1} \to 0
$$

$$
\frac{1}{n}\sum(x_i\varepsilon_i) = O_p(1)
$$

(it just needed to be smaller than ∞, given that the other term is 0)

In simpler terms: get the estimator and to be centered around the true parameter and make it look like something that makes sense.

As we can see that for this to hold, M^{-1} must exist. Therefore, the identification assumptions for OLS must exist.

5.d Derive ATE in a non Randomized Experiment

Two step process:

- estimate $\alpha(x)$ and $\beta(x)$
- Use these to estimate τ

Lets assume for a moment that the CATE are linear functions. If we run a regression in the treatment or control group, I can recover the CATE.

$$
\text{CATE} = \beta(x) = \tau(x) = x'\beta_1 - x'\beta_0
$$

We need to estimate the estimators and then get the average of these estimators.

It becomes harder because the mean is done with estimated parameters.

The steps are:

- Run a regression in treatment and control groups separaately, then project everywhere (or run a saturated model).
- Then:
$$\hat{\tau} = \mathbb{E}[\hat{Y}(1)] - \mathbb{E}[\hat{Y}(0)] = \frac{1}{n}\sum_{i=1}^{n} x_i\hat{\beta}_1 - \sum_{i=1}^{n} x_i\hat{\beta}_0$$

In our $\hat{\tau}$ is now a composition of two functions. We want to figure out how the map that generates $\hat{\tau}$ changes with changes in dataset.

We have "double changes" in this estimation: data changes and coefficients changes.

If we had known the influence function in advance, figuring out the distribution would be easier.

5.d.1 Formally as Maps

$\hat{\beta}_1$ and β_1 are a function of DGP.

We want to prove that the first part is consistent. For the treatment group:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i\hat{\beta}_1$$

Where:

$$\begin{aligned}\mu_i &= \mathbb{E}[Y(1)] \\ &= \mathbb{E}[X]'\beta_1\end{aligned}$$

Therefore:

$$\hat{\mu}_1 = (\frac{1}{n}\sum_{i=1}^{n} x_i)\hat{\beta}_1 \to_p \mathbb{E}[X]\beta_1 = \mu_1$$

Now, considering that this part is done, we can do the CLT for τ:

$$\sqrt{n}(\hat{\tau}-\tau) = \sqrt{n}(\frac{1}{n}\sum_i x_i - \mathbb{E}[X])\hat{\beta}_1 \quad \text{(does not work - goes to } \infty\text{)}$$
$$= \left(\frac{1}{n}\right)\sqrt{n}(\hat{\beta}_1-\beta_1) + \sqrt{n}\left(\frac{1}{n}\sum x_i\right)\beta_1 \quad \text{(also does not work - goes to } \infty\text{)}$$

How to solve that?

6 Lecture 6

Start with the regression for the treated group:

$$\sqrt{n}(\mathbb{E}[\hat{Y(1)}]) - \mathbb{E}[Y(1)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(x_i,y_i,t_i) + o_p(1)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i\hat{\beta}_1 - \mathbb{E}[x\beta_1]) \quad (\beta_1 \text{ is only for the treated group})$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i\beta_1 - \mathbb{E}[x\beta_1]) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i\hat{\beta}_1 - x\beta_1)$$
(from now, only changes in the second term)
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i\beta_1 - \mathbb{E}[x\beta_1]) + \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\sqrt{n}(\hat{\beta}_1-\beta_1)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i\beta_1 - \mathbb{E}[x\beta_1]) + \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\frac{n_1}{n}\right)^{-\frac{1}{2}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}M_1^{-1}t_ix_i\varepsilon_i + o_p(1)\right]$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i\beta_1 - \mathbb{E}[x\beta_1]) + (\mathbb{E}[X]\times o_p(1))\left(\mathbb{P}[T=1]\right)^{-\frac{1}{2}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}M_1^{-1}t_ix_i\varepsilon_i + o_p(1)\right]$$

In this case, we see that second term contains $o_p(1)$.

CONFERIR $\mathbb{E}[X]\times o_p(1)$: tem $o_p(1)$?

Now, lets talk about the mean of the treated Y: μ_1

We defined that the conditional mean of Y is a linear function:

$$\begin{aligned}\mu_1 &= \mathbb{E}[Y(1)] = \mu_1(F)\\ &= \mathbb{E}[E(Y(1)|X)]\\ &= \int (x\beta_1)\,dF\end{aligned}$$

From here, we can get how much the μ changes with changes in F:

$$\frac{\partial \mu_1(F, \beta_1(F))}{\partial F} = \frac{\partial \mu_1}{\partial F} + \frac{\partial \mu_1}{\partial \beta_1} \frac{\partial \beta_1}{\partial F}$$

We can see that the average of the treated is dependent on β_1 and F.

6.a The Problem of the Variance with 2 Step Estimation

We aim to get θ_*:

$$\theta_* = \mathbb{E}[Y(1)] = \mathbb{E}[X\beta_1]$$

$$\gamma_* = \beta_1$$

$$\hat{\gamma} = (X'X)^{-1}X'Y$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i \hat{\beta}_i$$

Therefore, we have 3 steps for what we call now θ_A (A because is the first method to achieve such result)

- Get $\hat{\gamma}$ and $\hat{M}$
- Get $\hat{\theta}_A$
- Estimate the distribution:

$$\hat{\theta}_A \sim N(\theta_A, V_1 + V_2)$$

Due to the two step estimation, we have $V_1 + V_2$, which is a problematic estimate. Thus, we will use method B, which will make the first part more difficult, but facilitate the estimation of the variance.