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 $f(\mathbf{w})$ and the X and Y as $\mathbf{w}$.

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 it 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 $f(\mathbf{w}) = \mathbf{w}^2$, which will always produce 4 when $\mathbf{w} = 2$.

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. For instance, when estimating the mean, which in this case we refer as α:

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

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:

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

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:

$$\sqrt{n}(\hat{\alpha} - \alpha) = \frac{1}{\sqrt{n}} \sum (x_i - \mu) \to N(0, \rho^2)$$

Where $(\hat{\alpha} - \alpha)$ 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. In this way, we can prove assyntotically normality.

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:

$$a_n = O(b_n)$$

refers 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.

In an example, given:

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

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

Going back to the original problem, we should aim to write:

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

For it to hold, $o_p(1)$ must go to 0. It must be smaller than $\frac{1}{\sqrt{n}}$. Otherwise we wont have the desired property as $n \to \infty$.

Therefore:

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

x_i represents observation i. Thus, the matrix $X'X$ can be written as:

$$X'X = \sum_{i=1}^{n} x_i x_i'$$

After, we use the fact the inverse of the sample average is common across all terms and we are able to express it inside the sum.

$$\bar{M}^{-1} = \left(\frac{1}{n}\sum x_j x_j'\right)^{-1}$$

Furthermore, $\bar{M} \xrightarrow{p} M$, thus $\bar{M}^{-1} - M^{-1} = o_p(1)$.

M is the probability limit of $\bar{M}$ as the sample size n approaches infinity.

We can write it as $o_p\left(\frac{1}{\sqrt{n}}\right)$ 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 in a way that "makes sense" and it provides assyntotically normality.

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 estimation:

- 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 use the CATE to get the ATE. Meaning that we get the average of the CATE. The solution is harder because you have (i) uncertainty estimating CATE, and (ii) uncertainty averaging CATE.

The issue with observational data is that there is selection bias:

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

Now, we are assuming that X captures why people select, meaning:

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

The intuition behind it is that we have "RCT" (or equivalent of RCT) for each $X = x$.

The steps are:

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

Rewriting the previous bullet points. One should estimate the regression for the treated group:

$$Y_i = \alpha_1 + X_i\beta_1 + \varepsilon_{1,i} \quad \text{for} T_i = 1$$

And a separated regression for the control group:

$$Y_i = \alpha_0 + X_i\beta_0 + \varepsilon_{0,i} \quad \text{for} T_i = 0$$

The idea of projecting everywhere means that you should predict the outcome for every unit of the sample using the treated and control regressions.

Meaning that, regardless if the observation being treated or not, you run the regression for treated and non-treated.

$$\hat{Y}_i(1) = X_i\hat{\beta}_1, \quad \hat{Y}_i(0) = X_i\hat{\beta}_0$$

From here, the individual estimated treatment effect is:

$$\hat{\tau}_i = \hat{Y}_i(1) - \hat{Y}_i(0) = X_i(\hat{\beta}_1 - \hat{\beta}_0)$$

The average treatment effect is by definition the average of the individual treatement effects:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i = \frac{1}{n}\sum_{i=1}^{n}X_i(\hat{\beta}_1 - \hat{\beta}_0)$$

Alternatively, running a saturated regression means that you run a regression that considers all data:

$$Y_i = \alpha + T_i(\alpha_1 + X_i\beta_1) + (1 - T_i)(\alpha_0 + X_i\beta_0) + \varepsilon_i$$

This can be simplied to (NOT ENTIRELY SURE ABOUT THAT PART):

$$Y_i = \alpha + \beta X_i + \tau T_i + (T_i \times X_i)\phi + \varepsilon_i$$

Again, we are making assumptions to make sure that such a method provides the ATE. That is the Conditional Independence Assumption (CIA), which means that, after conditioning on X, the potential outcomes are independent of treatment assignments.

In it strong version:

$$Y(1), Y(0) \perp T|X$$

In its weak version:

$$\mathbb{E}[Y(t)|T, X] = \mathbb{E}[Y(t)|X]$$

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 is be easier.

5.d.1 Formally as Maps

$\hat{\beta}_1$ and β_1 are a function of DGP (data generating process).

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:

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

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

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

How to solve that?

Influence Functions

Leave One Out Approach

Common method in the 1950s in which one would take away one single data-point and retrain the model to check how the prediction would change.

The method was quite inneficient and computational intensive.

Gradient Based Solution

The solution found was to use efficient gradient-based approximation: influence functions.

The idea started with setting weights for each individual observation when training the model:

$$L = \frac{1}{n} \sum_{i=1}^{n} w_i \ell_i(x_i, y_i)$$

The leave one out approach is the case in which $w_i = 0$ for a particular i.

The influence function, on the other hand, is a continous approximation of the method, in which the change happens in infinitesimal scale $(1 - \varepsilon)$.

Meaning tha we have a perturbation in one single datapoint.

From the previous loss, we can arrive to the fact the the loss for a individual datapoint is actually dependent on the parameters θ.

$$L(\mathbf{w}, \theta) = \frac{1}{n} \sum_{i=1}^{n} w_i \ell_i(\theta)$$

For the following steps, we define:

- $g_i(\theta) = \nabla_\theta \ell_i(\theta)$
- $h_i(\theta) = \nabla^2_\theta \ell_i(\theta)$
- $G(\theta, \mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} g_i(\theta) = \nabla_\theta L$
- $H(\theta, \mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} h_i(\theta) = \nabla^2_\theta L$
- $\hat{\theta}$ is the $\text{argmin}_\theta\ L$
- $\hat{\theta}_1$ is $\hat{\theta}$ for when $\mathbf{w} = \mathbf{1}$.
- $H_1 = h(\hat{\theta}_1, \mathbf{w} = \mathbf{1})$

Implicit Function Theorem

Given $G = \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^d$ continuous and differentiable.

For some fixed $\hat{\theta}_1 \in \mathbb{R}^d,\ \mathbf{1} \in \mathbb{R}^d, G(\hat{\theta}_1, \mathbf{1}) = \mathbf{0} \in \mathbb{R}^d$, if

$$\frac{\partial G(\hat{\theta}(w), w)}{\partial \hat{\theta}^\top} \mid_{\theta=\hat{\theta}_1, \mathbf{w}=1}$$

is invertible, then $\exists$ open $U \subset \mathbb{R}^n$ and a unique function $\hat{\theta} : U \to \mathbb{R}^d$, s.t. $\hat{\theta}(\mathbf{w}) = \hat{\theta}_1$ for $\forall\ \mathbf{w} \in U$ and $G(\hat{\theta}(\mathbf{w}, \mathbf{w})) = 0$.

Moreover, $\hat{\theta}$ is continuous and differentiable and:

$$\begin{aligned}\frac{\partial \hat{\theta}(w)}{\partial w^\top} &= -\left(\frac{\partial G(\hat{\theta}(w), w)}{\partial \hat{\theta}^\top}\right)^{-1} \frac{\partial G(\hat{\theta}(w), w)}{\partial w^\top} \\ &= -H(\hat{\theta}(w), w)^{-1} \frac{\partial G(\hat{\theta}(w), w)}{\partial w^\top}\end{aligned}$$

Evaluating the function at $w = 1$, we get:

$$\frac{\partial \hat{\theta}(w)}{\partial w^\top} = \frac{1}{2n} H_1^{-1} \left[g_1(\hat{\theta}_1), \ldots, g_n(\hat{\theta}_1)\right]$$

Thus, for an individual datapoint i:

$$\frac{\partial \hat{\theta}(w)}{\partial w^\top} = \frac{1}{2n} H_1^{-1} g_i(\hat{\theta}_1)$$

Thus, the partial of the loss function with respect to w i:

$$\frac{\partial f(\theta)}{\partial w^\top} = \frac{\partial f(\theta)}{\partial \hat{\theta}(w)} \cdot \frac{\partial \hat{\theta}(w)}{\partial w^\top}$$

Evaluated at $w = 1$.

Where $\frac{\partial f(\theta)}{\partial \hat{\theta}(w)} = g(\hat{\theta}_1)$

Thus, for a testing sample:

$$\frac{\partial \ell_{\text{test}}(\theta)}{\partial w^\top} = -\frac{1}{2n} g_{\text{test}}(\hat{\theta}_1) H_1^{-1} g_i(\hat{\theta}_1)$$

It can be viewed as an inner product between g_{test} and $g_i(\hat{\theta}_1)$ multiplied by a "kernel", which is H_1^{-1}.

6 Lecture 6

Start with the regression for the treated group:

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

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

6.a Mean of Treated Group 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.b 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.

7 Lecture 7

7.a Doubly Robust Estimation

Doubly Robust Esimation: the estimator is consistent if only one of the steps is consistent. Therefore, it is consistent even if one of the functions is incorrect.

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\mathbb{E}[Y(1)|X] + \frac{\tau(Y - \mu_1(x))}{p(x)}\right]$$

Here we have the propensity score $p(x)$: and we did not plug in a extreme $p(x)$ (meaning not 0 and ∞), the $\mathbb{E}[Y(1)]$ is consistent because the second term tends towards to 0.

The same happen if we misspecified $\mu_1(x)$ and specify correctly $p(x)$: we still expect the estimator to be consistent.

In a linear estimator, we have:

$$\mathbb{E}[Y(1)|X] = \mu_1(x) = \frac{1}{n}\sum_{i=1}^{n} x_i\hat{\beta}$$

$$\frac{t_i}{p(x_i)}\left[y_i - \hat{\mu}_1(x_i)\right] = \hat{M}_1^{-1}\hat{\varepsilon}_i$$

7.a.1 General Case

Show the two step estimation for a general case.

7.b Nonparametric and ML

In this class, ML is a substitute for Nonparametric regression.

We are going to try to learn those functions with more flexible specifications. We are also going to try to learn the specification from the data.

We now have the same data, but different assumptions. We will now assume other stuff about the function and that the estimators are well behaved in a certain way.

We can start using simple nonparametric inference and infer the same for modern ML methods.

We can show that Deep Learning are piecewise linear regressions.

7.b.1 Crash Course in Nonparametric

When we use nonparametric estimation, we are often assume "smoothness", which we define that the function has some number of well behaved derivatives.

If $x_1 \approx x_2$ (are close enough to each other), we assume $f(x_1) \approx f(x_2)$.

7.b.2 Taylor

Taylor approximation tells an approximation for $f(x)$ in any x based on the derivatives and an initial point x_0.

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \dots$$

The first approximation is:

$$
\begin{aligned}
f(x) &\approx f(x_0) + f'(x_0)(x - x_0) \\
&\approx [f(x_0) + f'(x_0)(x_0)] + f'(x_0)x \\
&\approx \alpha + \beta x
\end{aligned}
$$

In which polinomial should I stop?

The bigger polinomial, the bigger the variance, the smaller my bias.

$$
\begin{aligned}
\hat{f}(x) &= |f(x) - \bar{f}(x)| + |\bar{f}(x) - \hat{f}(x)| \\
&= \text{Bias} + \text{Variance}
\end{aligned}
$$

For a very large gap between x and x_0, a linear function can only approximate so much. For a large gap, you can expect the approximation to be poor, given that the linear function can only approximate so much. We define this as bias.

Instead of using a linear approximation for the entire dataset, we can also cut the dataset in parts and estimate the line for each part of X. This will limit my bias, given that now we have the gap between x and x_0 "half" of the size (or at least smaller).

$$
\mathbb{F}_{\text{2L}} \begin{cases} (a_1 + b_1 X)\mathbb{I}\{x < M_d\} \\ (a_2 + b_2 X)\mathbb{I}\{x \geq M_d\} \end{cases}
$$

My variance, on the other hand, increased.

The variance:

$$
\mathbb{F}_{\text{L}} = \left(\frac{k}{1}\right)^2 + \frac{1}{n}
$$

$$
\mathbb{F}_{\text{2L}} = \left(\frac{k}{2}\right)^2 + \frac{2}{n}
$$

The variance tells you how much your estimate changes when the data changes.

The bias tells me how far my selected model is distant from the data generating process. On the other hand, the variance tells me how much the estimation of my data given the model is distant from the data generating process. This implies that, by estimating the parameters many times, how big is the probability that I am "quite far" from the actual value.

8 Lecture 8

8.a Piecewise Regression

Divide the regression in different parts (bins) and have the polinomial of size p.

$$\mathbb{F}_n = \left\{ \sum_{j=1}^{J} \mathbb{I}\{x \in b_j\}(p(x)'\beta_j), \quad b \in \mathbb{R}^{ph} \right\}$$

As either p or j converge to infinity, we can approximate perfectly any function.

Now, let's say we are working in a specific bin: now we are able to localize.

If I sum close enough everything becomes a line.

$$f(x) = f(x_0) - f'(x_0)x_0 + f'(x_0)x = \beta_i + \beta_i^2$$

As long as the $f''(x)$ is finite. Meaning:

$$\sup_x |f''(x)| < C$$

How far apart can x and x_0 be? They cannot be more far apart than the size of the bin.

So far, we have been saying that the bins are the all the same size. We could also:

- Have the same number of observations in each bin.
- Look at the data and check for more obvious pattern: nonetheless, most time you won't be able to do that.

The smoothness that you use has to always be smaller to the smoothness that exists.

If you are fitting local quadratic, we have to assume J^{-3}.

If you try to set 5-degree polinomial but the function is only twice derivable. This would be a problem and you would only be able to approximate to the 2nd degree.

$$\begin{aligned}
\hat{f}(x) - f(x) &= \hat{f}(x) - \bar{f}(x) + \bar{f}(x) - f(x) \\
&= p(x)(B'B)^{-1}(B'Y) - p(x)\bar{\beta} + O(J^{-p-1}) \\
&= p(x)(B'B)^{-1}(B'[Y-G]) + p(x)'(B'B)^{-1}B'[G - B\bar{\beta}] \\
&\quad + O(J^{-p-1})
\end{aligned}$$

We are able to say that, in the limit (I THINK), this is equal to zero:

$$p(x)'(B'B)^{-1}B'[G - B\bar{\beta}]$$

Looking specifically at:

$$\begin{aligned}\hat{f}(x) - f(x) &= p(x)(B'B)^{-1}(B'[Y - G]) + O(J^{-p-1}) \\ &= p(x)'(B'B)^{-1}(B'[Y - G])\end{aligned}$$

Where we know that:

$$(B'B) = \frac{1}{n_j}\sum_{i=1}^{n} p(x_i)p(x_i)\mathbb{I}\{x_i \in b_j\}$$

This tells us that j must goes slower to infinity than n, otherwise some bins will be empty:

$$n_j \approx \frac{n}{J} \to \infty$$

This function follows the central limit theorem:

$$\sqrt{\frac{n}{J}}(\hat{f}(x) - f(x)) = \sqrt{\frac{n}{J}}p(x)(B'B)^{-1}(B'[Y - G]) + \sqrt{\frac{n}{J}}O(J^{-p-1})$$

Knowing that J is going to infinity, but not too fast, we expect that:

$$\sqrt{\frac{n}{J}}(\hat{f}(x) - f(x)) = N(0, V) + \sqrt{\frac{n}{J}}O(J^{-p-1})$$

Where:

$$G = \begin{bmatrix}\mathbb{E}[Y|X_1] \\ \mathbb{E}[Y|X_2] \\ \vdots \\ \mathbb{E}[Y|X_n]\end{bmatrix} = \begin{bmatrix}f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n)\end{bmatrix}$$

$\bar{\beta}$ is the best slope for that bin.

8.a.1 Conclusion

We want to set in a way that the benefit of making J larger (smaller bias) is equal to the detriment of making it larger (bigger variance).

We are optimizing γ and J.

For inference, we would have a different optimal J than for point estimate. The problem now is that we have a weird term when dealing with inference:

$$\sqrt{\frac{n}{J}} J^{-p-1}$$

The trade-off between point estimate and inference is not clear...

The best estimator is not equal to the best t-statistics.

The linear regression OLS does not have this bias-variance trade-off.

9 Lecture 9

The convergence rate of these more complicated models is slower.

We still have that:

$$\sqrt{\frac{n}{J}} \left(\hat{f}(x) - f(x) \right) \to N(0, \Omega)$$

Thus, we can have the the CI.

If we understand the distribution of $f(x)$, we can understand the CI for any x.

1. We do not know the variance:
$$\sqrt{n} \left(\hat{\beta} - \beta \right) \to N(0, V)$$
This would allow us to get the CI once we have the variance.
2. Estimate the variance:
$$\hat{V} \to V$$
3. Estimate with the sample variance:
$$N(0, \hat{V})$$

The strong approximation does the same, but in a uniform way in x:

$$\sup_x \left| \mathbb{P} \left[\frac{\sqrt{\frac{n}{J}} (\hat{f}(x) - f(x))}{V(x)^{\frac{1}{2}}} < z \right] - \mathbb{P} \left[N(x) < z \right] \right| \to 0$$

With that, we are able to check the CATE.

We do some coding:

- We use the `binsreg` to do polinomial bin regression.
- `cb` gives a legitimate band: you have to take the confidence interval of bins when the number of bins goes to infinity. This is different from just connecting some random confidence intervals.
- How can we test if the $f(x)$ is linear? We can check if we can fit a line within the region of the 95% confidence interval. We can test whatever kind of function we want.
- We can think of a "lower function" as the lower bound and the "upper function" as the upper bound.
- We can decide to define the number of bins and the polinomial degree. Furthermore, we can also make that the function fitted in each of the bins is connected to the function in the bins next to it.
- In real life, we do not expect the CI of the function to be super smooth. We expect spikes.

When can the CI not contain the point estimate?

The J optimal for defining the function is not the same as the optimal J to define the confidence interval.

$$||\hat{f} - f|| = \frac{J}{n} + J^{-2(p+1)}$$

In our example (to show a bad result), we used a bin with polinomial degree 0 and the function polinomial degree for the function.

If you want to make inference in piecewise constant, you should use linear. If you make inference in piecewise linear, you should use quadratic. The overall rule, is that the inference should be done with a bigger polinomial degree.

9.a Trees

Regression trees are piecewise constant, but we pick the pieces based on the data.

We split the data by some optimal way of spliting: grid-search.

The grid-search optimizes one split at a time. We look at the split in the data that maximizes the R^2.

Discussion 4

9.b Hypothesis Testing

For some unknown parameter μ which is a function of data generating distribution P, we are concerned with testing the bull hypothesis that μ falls within some set S.

$$H_0 : \mu(P) \in S$$

$$H_a : \mu(P) \notin S$$

- Type I Error: rejecting H_0 when it is true.
- Type II Error: not rejecting H_0 when it is false.

COPIAR RESTANTE DO SLIDE.

Influence Functions

Measure the sensitivity of an estimate to infinitesimal perturbations in the data distribution.

$$F_\varepsilon = (1 - \varepsilon)F + \varepsilon\delta_{x_i}$$

The influence function $IF(x_i, T, F)$ at point x_i is defined as:

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

10 Lecture 10

10.a Trees

- Trees are easy to do and interpret.
- They fit nonlinearities and selection automatically.
- Because trees are always piecewise flat.
- The binary variables are easy to check.

Trees are generally not used, but they are nice as a building block.

How to stop a tree? There are several ways:

- A minimum number of observations per leaf.
- A maximum depth.
- Prune than tree: grow it really deep and than decide which pieces to cut off.

10.b Random Forest

Forests do this over and over and average the result.

We use samples with replacement. we select a portion of the observations and also a portion of the features (not select every feature). For within each sample, an observation can come more than once.

The average result is:

$$\hat{f}(x)_{\mathrm{RF}} = \frac{1}{n_T} \sum_{t=1}^{T} \hat{f}_t(x)$$

Where T is the number of trees and the $\hat{f}(x)_{\mathrm{RF}}$ is the estimate function of the data.

What is happening when I take an average? More specifically, what is the average for a particular $X = x$?

We take the average of the estimates for that particular point. For any point, we are just averaging.

Nonetheless, for the entire $\hat{f}(x)_{\mathrm{RF}}$ we will see that some points are more important than others because they appear more. Furthermore, in bins that have more poits, the weight of a single bin is less important.

Therefore, in the end, the $\hat{f}(x)_{\mathrm{RF}}$ will be a weighted average of the points.

$$\hat{f}(x)_{\mathrm{RF}} = \frac{1}{n} \sum_{i=1}^{T} w_i Y_i$$

The function of weights is more discountinous in situations where we have less trees and less discountinous when we have more trees.

A single tree with sampling would have the most discountinous weights.

Remember that we are looking for local expectation:

$$\mathbb{E}[Y|X]$$

In the end, random forest requires us to check many possibilities of hyperparameters, making the result dependent on choices.

10.c Deep Neural Networks

Did not write anything

11 Lecture 11

We reviewed:

- Neural Networks.
- Activation functions, specially ReLU.
- Conceptuall NN are just fitting very complicated piecewise functions.
- At the very end node, we should not take an activation function (unless we actually need one - for instance, logistic regression would use LOGIT transformation and counting problems would use Poasson).

The results from NN:

$$\left|\hat{f}_{\text{DNN}} - f(x)\right| = O_p(\text{FINISH WRITING FROM THE SLIDE})$$

The result is quite similar to the results of OLS.

The result is mainly a function of the number of parameters.

The magic of machine learning methods is that they are able to approximate and find out stuff about the data without you specifically telling it.

Machine Learning by itself figures out a low dimensional space to make predictions.

How does a computer takes derivatives?

1. Numerical derivatives: just evaluate the function in x and $x + h$.
2. Symbolic derivatives: given a set of rules, we make the derivatives (similar to what we do).
3. Automatic derivative: the computer without knowning derivatives, does the derivative with the data. Thus, I am not going to know the derivative, but I will approximate the derivative.

11.a Automatic derivative

Given a complicated function:

$$f(x) = \log(x) \exp\left(x^{\left(-\cos(x)^2\right)}\right)$$

That means that we have:

- $\log(x) \to b$
- $\cos(x) \to -x^2 \to x^x \to \exp(x) \to a$
- $a \times b \to f(x)$

Here, we start by taking the derivative from the end to the start.

Meaning the derivative from $f(x) \to a$ and $f(x) \to b$, from $a \to \exp(x)$, etc...

For that, we need tensors.

REVIEW TENSORS

Discussion Section 5

REVIEW DISCUSSION

12 Lecture 12

Number of parameters in NN: there are two regimes.

One regime thinks about the variance-bias tradeoff. The more parameters we have, the higher the variance.

The other regime is overparamterization.

12.a Basis Functions

We can approximate $f(x)$ with different kind of functions. For instance, polinomials. With polinomial, we can transform our feature in more features by using polinomial, piecewise, etc... we can transform a low dimensional linear problem into a high dimensional one.

13 Lecture 13

To test machine learning, we often divide the data in:

- training, validation, testing
- training, validation

The testing data will serve as a hold out sample: we can only use the testing sample to validate my final model.

13.a Trees

Trees are fine for hypothesis generation, but not fine for hypothesis testing.

13.b Forest

Trying to interpret the feature relationship:

- Partial dependence plot: shows the partial dependence, meaning how the prediction change with respect to a particular feature.
- Variable dependence plot: shows the importance of the features (how much is that variable is being used as a splitting variable and how much of a difference this makes). In our example, `age` is the most "important". That does not mean that being older/younger causes the target variable. So far, it is only non-causal relationships.

Statistically significance and hability to predict well are different things.

Partial dependence plot can be used for NN. Variable dependence plot cannot be used in NN.

13.c Semiparametric Inference & Double Machine Learning

How to do statistical inference in ML?

In parametric models:

$$\begin{aligned}\mathbb{E}[\hat{Y(1)}] &= \frac{1}{n}\sum_i x_i'\hat{\beta}_i \\ &= \frac{1}{n}\sum_i x_i'\beta_i + \left(\frac{1}{n}\sum_i x_i'\right)\left(\hat{\beta}_1 - \beta_1\right)\end{aligned}$$

For piecewise linear regression

We are assuming that it is a smooth function and we are approximating with J pieces.

$$\begin{aligned}\sqrt{n}\left(\mathbb{E}[\hat{Y(1)}] - \mathbb{E}[Y(1)]\right) &= \frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_i^n p_J(x)'\hat{\gamma} - \mathbb{E}[Y(1)]\right) \\ &= \frac{1}{\sqrt{n}}\sum_i^r(\mu_i^*(x_i) - \mathbb{E}[Y(1)]) \\ &\quad + \frac{1}{\sqrt{r}}\sum_i^r(p_J(x_i)'\gamma_n^* - \mu_i^*(x_i)) \\ &\quad + \frac{1}{\sqrt{n}}\sum_i p_J(x_i)'(\hat{\gamma} - \gamma_n^*)\end{aligned}$$

The $p_J(x_i)'$ the best estimator is this class of functions. It is the best possible piecewise polinomial. If you have infinite J, the approximation becomes the function.

In it:

13.c.1 Bias term (second term)

$$\sum_i^r(p_J(x_i)'\gamma_n^* - \mu_i^*(x_i))$$

Each if the individual i is of size $J^{-(p+1)}$.

If J is big enough, I can ignore this term. If we do not have big enough J, bias will always exist. Meaning that, if your function is quadratic and we do a linear approximation, we will always have bias...

13.c.2 Variance term (third term)

$$\frac{1}{\sqrt{n}}\sum_i p_J(x_i)'(\hat{\gamma} - \gamma_n^*)$$

If the J is bigger, the number of observations in each bin is smaller. Thus, this term is bigger and has a lower convergence rate when J is bigger.

Thus, in order for the bias to go to 0, we have the problem with the variance...

Otherwise, there will always be an error in the:

$$\sqrt{n}\left(\mathbb{E}[\hat{Y(1)}] - \mathbb{E}[Y(1)]\right)$$

Discussion 6

MLE does not have any universal finite sample guarantees, but it does have some nice asymptotic properties.

14 Lecture 14

14.a Double Machine Learning

Average of Potential Outcome:

$$\mu(F) = \mathbb{E}[Y(1)] = \mathbb{E}[\mathbb{E}[Y(1) \mid X]]$$

We view it as:

$$\mathbb{E}[\mathbb{E}[Y(1) \mid X]] = \Theta(F, X)$$

It maps to a function space of X.

$$\mu(F_\varepsilon) = \int \Theta(F_\varepsilon, X) f_\varepsilon(X)\, dX$$

$$\frac{\partial \mu(F_\varepsilon)}{\partial \varepsilon} \mid_{\varepsilon \to 0} = \int \Theta(F, X) \frac{f_\varepsilon(X)}{\partial \varepsilon} \mid_{\varepsilon = 0} \; dX + \int \frac{\partial \Theta(F_\varepsilon, X)}{\partial \varepsilon} f(X)\, dX$$

Looking specific at the second term, we know that according to the first order condition:

$$\frac{\partial \Theta_\varepsilon}{\partial \varepsilon} : \mathbb{E}[T(Y - \Theta(X))|X] = 0$$

This always holds for every distribution.

This is true for the distribution in the perfurbation:

$$\frac{\partial \Theta_\varepsilon}{\partial \varepsilon} : \mathbb{E}_\varepsilon[T(Y - \Theta_\varepsilon(X))|X] = 0$$

This last term can be viewed as:

$$\mathbb{E}_\varepsilon[T(Y - \Theta_\varepsilon(X))|X] = \int t(y - \Theta_\varepsilon(x)) f_\varepsilon(y, t \mid x)\, dydt = 0$$

A BUNCH OF STEPS...