Causal Machine Learning – Autumn Quarter 2024–2025

Slides Set #2: Frequentist Inference and Influence Functions

Max H. Farrell

(version 1.1, compiled October 10, 2024)

Topics to cover

1. Inference based on asymptotic Normality
2. Standard errors
3. Estimators as maps from data sets to $\mathbb{R}$
4. Influence functions
5. Examples in parametric models (OLS, MLE)
6. Two step estimation

Example: Linear Regression

Fit a linear model (for simplicity, only one variable)

- Model: $Y = \beta_0 + \beta_1 X + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X] = 0, \quad \mathbb{E}[\varepsilon^2 \mid X] = \sigma^2$
- Data: $\{y_i, x_i\},\ i = 1, \ldots, n$
- Estimation $y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + e_i$

Standard/textbook result.

$$\sqrt{n}(\widehat{\beta} - \beta) \rightarrow_d \mathcal{N}(0, V)$$

1. What does this mean?
2. Where does this come from?
3. What is V?

What Variance?

We talk about "variance" a lot.

- σ^2 is the variance of ε (or $\mathbb{V}[Y \mid X]$).
- V is the (asymptotic) variance of $\widehat{\beta}$

We **actually do** see many realizations of ε, and they **vary**. We have **only one** value $\widehat{\beta}$, so how does it "vary"? How does V quantify the precision of the estimator?

- ε is **one draw** from $(0, \sigma^2)$ (We see n of these)
- $\widehat{\beta}$ is **one draw** from $\mathcal{N}(\beta, V)$ (Wee see **1** of these)
- Example: You flip n coins. Each of the n tosses $\{0, 1\}$ is Bernoulli distributed, but the mean is Normally distributed.

Both σ^2 and V measure how much each draw bounces around

- Standard errors are just estimates of this, since we don't know V.
- How much will $\widehat{\beta}$ changes if the data changes (which it won't)?

Frequentist Inference

Functions/Maps

- $q = f(w) = w^2$ is a function that maps $\mathbb{R} \mapsto \mathbb{R}$
- Different $w \in \mathbb{R}$ yield different $q \in \mathbb{R}$
- $f(w)$ is a function, but after you know w, q is fixed.

Random Variable

- Every RV is a map from a **sample space** to the **real line**. E.g. $X : \mathcal{S} \to \mathbb{R}$
- Different $\omega \in \mathcal{S}$ yield different realizations $x_i \in \mathbb{R}$
- X is random, but after you know ω, x_i is fixed
- Column of data is a set of n realizations of this map: $x_i = X(\omega),\ \omega \in \mathcal{S}$

Frequentist Inference

Estimators are Random Variables

- $\widehat{\beta} := \widehat{\beta}(\omega) \to \mathbb{R}$
- The sample space $\mathcal{S} = \{\text{all possible data sets}\}$
  - Each point $\omega \in \mathcal{S}$ is a data set of size n
  - Write F for the poulation distribution of the random variables (Y, X)
  - F_n is the empirical distribution of the data set
  - We will write $\widehat{\beta} = \widehat{\beta}(\omega) = \widehat{\beta}(F_n)$
- Different $\omega = F_n \in \mathcal{S}$ yield different realizations $\widehat{\beta} \in \mathbb{R}$
- $\widehat{\beta}$ is random, but after you know F_n, $\widehat{\beta}$ is fixed
- $\widehat{\beta}$ changes if the data changes (but the data never actually changes)
- $\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(0, V) \quad \approx \quad \widehat{\beta} \overset{a}{\sim} \mathcal{N}(\beta, V/n)$
- Monte Carlo illustration

Influence Functions

- To find V, we need to measure how $\widehat{\beta}$ changes when the data changes
- View $\widehat{\beta}$ as a function of the data: $\widehat{\beta} := \widehat{\beta}(F_n)$, with F_n the distribution of the data
  - $\widehat{\beta} \to \beta$, which is also a function of the population "data": $\beta(F)$
  - If F_n are draws from F, then $\beta(F)$ is defined as what $\widehat{\beta}(F_n)$ estimates

Just like any other function, we can ask what happens to the output if the input changes a little.

What happens to $f(w) = w^2$ when w changes a little?

- $f(2) = 4, \;\; f(2 + 0.1) = 4.41$
- $f'(w) = 2w$

Need to formalize $\widehat{\beta}(\text{data} + 0.1)$. Need to find the derivative of $\widehat{\beta}$ with respect to the data set.

Really Simple Example: Sample Mean

Forget about X, assume we only have Y

- Model: $Y = \alpha + \nu, \quad \mathbb{E}[\nu] = 0, \quad \mathbb{E}[\nu^2] = \rho^2$
- Estimation: $\widehat{\alpha} = \sum_{i=1}^{n} y_i / n$

As a function of the distribution:

- $\widehat{\alpha} = \widehat{\alpha}(F_n) = \int y dF_n(y) = \mathbb{E}_n[Y] = \frac{1}{n} \sum_{i=1}^{n} y_i$
- $\alpha = \alpha(F) = \int y dF(y) = \mathbb{E}[Y]$

How to think about the data changing?

1. **Influence** of one data point on the statistic $\alpha(F)$
2. Perturbation of the data
3. Explicit derivative

Really Simple Example: Sample Mean

Both the **influence function** and the **CLT** capture how the statistic changes when the data changes.

Now we connect the two.

- The CLT applies to **averages**, and the influence function is **exactly** what you are averaging
  - For large n, F_n is close to F, so we examine $\alpha(F_n) - \alpha(F)$
- Need to properly center and scale the statistic

$$\sqrt{n}\,(\widehat{\alpha} - \alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{(y_i - \mathbb{E}[Y])}_{\substack{\text{influence}\\\text{function}}} \rightarrow_d \mathcal{N}(0, \rho^2)$$

- The asymptotic variance is just the variance of the influence function!
- Standard errors are just estimates of this variance

Back to Regression

Instead of differentiation, this time let's just derive forward until we get something in the right format:

$$\widehat{\beta} - \beta = \frac{1}{n}\sum_{i=1}^{n} M^{-1} x_i \varepsilon_i + o_P(1/\sqrt{n})$$

Note the format: Inverse times residual

- Inverse is also key for identification and is second derivative

Maximum Likelihood

Standard MLE:

- Data z_i, Parameter θ, Negative log likelihood $\ell(z, \theta)$
- $\theta_0 = \arg\min_\theta \mathbb{E}[\ell(Z, \theta)]$

$$\widehat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i, \theta)$$

$$\Leftrightarrow 0 = \mathbb{E}_n \left[\ell_\theta(z_i, \widehat{\theta}) \right] = \mathbb{E}_n \left[\ell_\theta(z_i, \theta_0) + \mathbb{E}_n \left[\ell_{\theta\theta}(z_i, \bar{\theta}) \right] \left(\widehat{\theta} - \theta_0 \right) \right]$$

So if $\mathbb{E}_n \left[\ell_{\theta\theta}(z_i, \bar{\theta}) \right] \to_p \Lambda(\theta_0) > 0$, then (Need **U**LLN, see Newey & McFadden)

$$\left(\widehat{\theta} - \theta_0 \right) = \frac{1}{n} \sum_{i=1}^{n} -\Lambda(\theta_0)^{-1} \ell_\theta(z_i, \theta_0) + o_P(1/\sqrt{n})$$

Note the format: Inverse times residual

- Inverse is also key for identification and is second derivative

Back to treatment effects

Just need to identify the CATE

- If we have $\tau(x)$, then we can average to get $\tau = \mathbb{E}[\tau(X)]$
- If we have $\mathbb{E}[Y(1) \mid X]$, we can average to get $\mathbb{E}[Y(1)]$

Two strategies

- Imputation: $\mathbb{E}[Y \mid T = 1, X = x]$
- Inverse weighting: $\mathbb{E}[YT \mid X = x] / \mathbb{E}[T \mid X = x]$

Observational Data - Binary Treatment

- Recall the selection bias problem: $\mathbb{E}[Y(0) \mid T=1] \neq \mathbb{E}[Y(0) \mid T=0]$
- Randomization made this go away
- Key idea with observational data: X captures why people select
  $\Rightarrow \mathbb{E}[Y(0) \mid T = 1, X = x] = \mathbb{E}[Y(0) \mid X = x]$
- Intuition: need an RCT for each $X = x$
- CIA, unconfoundedness, missing at random, ...
  - Strong version: $Y(1), Y(0) \perp\!\!\!\perp T \mid X$
  - Weak version: $\mathbb{E}[Y(t) \mid T, X] = \mathbb{E}[Y(t) \mid X]$
- Also still need overlap, consistency, SUTVA

Two step estimation

- Our goal is to estimate $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ **and** provide inference
- $Y = \alpha(X) + \beta(X)T + \varepsilon$ is w.l.o.g.
- $\tau = \mathbb{E}[\beta(X)]$
- In an RCT you recover the average of heterogeneous effects:

$$\hat{Y} = \widehat{\alpha} + \widehat{\beta}T \quad \longrightarrow \quad \widehat{\beta} \to_p \mathbb{E}[\beta(X)]$$

- But in general this fails
  - Need to account for heterogeneity, but we also want to exploit it
  - Need to get the CATE correct
- Two step estimation:
  1. Estimate $\alpha(x)$ and $\beta(x)$
  2. Use these to estimate $\tau = \mathbb{E}[\beta(X)]$ and do inference

Example: Linear models

- Assume a correctly specified linear (or other parametric) model:

$$\mu_t(x) = \mathbb{E}[Y(t) \mid X = x] = x'\beta_t$$

- $\text{CATE} = \beta(x) = \tau(x) = x'\beta_1 - x'\beta_0$
- Run a regression in treatment and control groups separately, then project everywhere (or run a saturated model).
- Then $\widehat{\tau} = \widehat{\mathbb{E}[Y(1)]} - \widehat{\mathbb{E}[Y(0)]} = \frac{1}{n}\sum_{i=1}^{n} x_i\widehat{\beta}_1 - \frac{1}{n}\sum_{i=1}^{n} x_i\widehat{\beta}_0.$

Big questions for today:

- How do we do inference for τ even though we estimate β_t first
- How can we change our approach to make this easier/better?
- Where do influence functions fit in?

Intuition for the problem

When we estimate $\mathbb{E}[Y(1)]$ there are two sources of uncertainty:

1. Usual frequentist parameter uncertainty: when the data changes the numbers change

   If we knew β_1 or $\widehat{\beta}_1$ was fixed, we'd have a standard CLT:

$$\sqrt{n}\left(\widehat{\mathbb{E}[Y(1)]} - \mathbb{E}[Y(1)]\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ x_i \widehat{\beta}_1 - \mathbb{E}[Y(1)] \right\} \rightarrow_d \mathcal{N}(0, \Sigma),$$

   Data changes $\rightarrow$ x_i changes $\rightarrow$ $x_i \widehat{\beta}_1$ changes $\rightarrow$ $\widehat{\mathbb{E}[Y(1)]}$ changes

2. Model uncertainty – when the data changes the function(al) $\widehat{\beta}_1(F_n)$ changes

$$\begin{array}{ccccc} & \nearrow & x_i \text{ changes} & \searrow & \\ \text{data changes} & & & & x_i\widehat{\beta}_1 \text{ changes twice} \\ & \searrow & \widehat{\beta}_1 \text{ changes} & \nearrow & \end{array}$$

Formally as Maps

- $\widehat{\beta}_1$ and β_1 are functions of the DGP
  - In fact the **same** function: $\widehat{\beta}_1(F_n) \to_p \beta_1(F) = \beta_1$
- Averaging is a function of the DGP
  - Remember $\frac{1}{n}\sum_{i=1}^{n} y_i$ versus $\mathbb{E}[Y]$
- So $\widehat{\mathbb{E}[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n} x_i\widehat{\beta}_1$ is function of the data twice

$$\widehat{\mathbb{E}[Y(1)]} = \widehat{\mathbb{E}[Y(1)]}(F_n) = \widehat{\mathbb{E}[Y(1)]}(F_n, \widehat{\beta}_1(F_n))$$
$$\widehat{\mathbb{E}[Y(1)]} \to_p \mathbb{E}[Y(1)] = \mathbb{E}_F[X\beta(F)]$$

- Derive IF for $\widehat{\mathbb{E}[Y(1)]} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i\widehat{\beta}_1 \ldots$

Key idea: use the IF for estimation

- Can we find a **different** function of the data that still estimates $\mathbb{E}[Y(1)] = \mathbb{E}_F[X\beta(F)]$, but without this two step estimation problem?
- Yes! We use the influence function

$$\widetilde{\mathbb{E}[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(z_i) = \frac{1}{n}\sum_{i=1}^{n} \left\{ x_i'\widehat{\beta}_1 + \mathbb{E}_n[x_i']\hat{M}_1^{-1} t_i x_i \hat{\varepsilon}_i \right\}$$

- Some Monte Carlos

Doubly robust estimation

Similar idea, but from a different angle

- We already saw two ways to identify $\mathbb{E}[Y(1)]$

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\mathbb{E}\left[Y \mid T = 1, X\right]\right] = \mathbb{E}\left[\frac{TY}{p(X)}\right]$$

- So we can use one or the other estimator:

$$\widehat{\mathbb{E}[Y(1)]}_{\text{IMP}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(x_i) \qquad \widehat{\mathbb{E}[Y(1)]}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{t_i y_i}{\hat{p}(x_i)}$$

- Each relies on a first step estimator: $\widehat{\mu}_1(x) = \widehat{\mathbb{E}}\left[Y \mid T = 1, X = x\right]$ and $\hat{p}(x_i) = \widehat{\mathbb{P}}[T = 1 \mid X = x]$
  - First step has to be right

Doubly robust estimation

Basic idea of doubly robust estimation:

- Two chances to get the right answer
- Cost: do **both** first step estimators
- Benefit: ATE is right if **either** first step is right

$$\widehat{\mathbb{E}[Y(1)]}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(x_i) + \frac{t_i\left(y_i - \widehat{\mu}_1(x_i)\right)}{\hat{p}(x_i)}$$

- What if only $\widehat{\mu}_1(x_i)$ is right? What if only $\hat{p}(x_i)$ is right?
- What if both are "close"?
- Consistent? Bias limit/asymptotic/finite?
- Related to IF?

General Case

Two step M/Z estimation

1. $\theta_0 = \arg\min_\theta \mathbb{E}[\ell(Z,\theta)] = \arg\operatorname{zero}_\theta \mathbb{E}[\ell_\theta(Z,\theta)]$

$$\Rightarrow \left(\widehat{\theta} - \theta_0\right) = \frac{1}{n}\sum_{i=1}^{n} -\Lambda^{-1}\ell_\theta(z_i,\theta_0) + o_P(1/\sqrt{n})$$

2. $\mu_0 = \arg\min_m \mathbb{E}[g(Z,m,\theta_0)] = \arg\operatorname{zero}_m \mathbb{E}[g_\mu(Z,m,\theta_0)]$

   $\widehat{\mu} = \arg\min_m \mathbb{E}_n[g(z_i,m,\widehat{\theta})] = \arg\operatorname{zero}_m \mathbb{E}_n[g_\mu(z_i,m,\widehat{\theta})]$

$$\begin{aligned}\Rightarrow (\widehat{\mu} - \mu_0) &= \frac{1}{n}\sum_{i=1}^{n} -\Omega^{-1} g_\mu(Z,\mu_0,\widehat{\theta}) + o_P(1/\sqrt{n}) \\ &= \frac{1}{n}\sum_{i=1}^{n} -\Omega^{-1}\left\{g_\mu(Z,\mu_0,\theta_0) + \mathbb{E}[g_{\mu,\theta}(Z,\mu_0,\theta_0)]\Lambda^{-1}\ell_\theta(z_i,\theta_0)\right\} + o_P(1/\sqrt{n})\end{aligned}$$

Familiar format

- Inverse $\times$ residual
- plug+in + gradient $\times$ correction