Causal Machine Learning – Fall 2023

Week 2: Frequentist Inference and Influence Functions

Max H. Farrell & Sanjog Misra

Topics to cover

1. Inference based on asymptotic Normality
2. Standard errors
3. Influence functions
4. Examples in parametric models (OLS, MLE)

Example: Linear Regression

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

- Model: $Y = \alpha + \beta X + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X] = 0, \quad \mathbb{E}[\varepsilon^2 \mid X] = \sigma^2$
- Estimation $Y = \hat{\alpha} + \hat{\beta} X + e$

Standard/textbook result:

$$\sqrt{n}(\hat{\beta} - \beta) \to_d \mathcal{N}(0, V)$$

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

(Generalization to the vector case $Y = \alpha + \boldsymbol{\beta}' \boldsymbol{X} + \varepsilon$ is immediate.)

Asymptotic Normality

$$\sqrt{n}(\hat{\beta} - \beta) \to_d \mathcal{N}(0, V) \quad \approx \quad \hat{\beta} \overset{a}{\sim} \mathcal{N}(\beta, V/n)$$

We treat $\hat{\beta}$ as if it were Normally distributed

- Frequentist inference:
  - $\hat{\beta}$ is uncertain because the data is uncertain
  - $\hat{\beta}$ changes if the data changes (but the data never actually changes)
- Monte Carlo illustration

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 $\hat{\beta}$

We **actually do** see many realizations of ε, and they **vary**. We have **only one** value $\hat{\beta}$, 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)
- $\hat{\beta}$ is **one draw** from $\mathcal{N}(0, V)$ (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 $\hat{\beta}$ changes if the data changes (which it won't)?

Influence Functions

- To find V, we need to measure how $\hat{\beta}$ changes when the data changes
- View $\hat{\beta}$ as a function of the data: $\hat{\beta} := \hat{\beta}(F_n)$, with F_n the distribution of the data
  - $\hat{\beta} \to \beta$, 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 $\hat{\beta}(F_n)$ estimates

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

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

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

Need to formalize $\hat{\beta}(\mathrm{data} + 0.1)$.

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: $\hat{\alpha} = \sum_{i=1}^n y_i / n$

As a function of the distribution:

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

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
- Need to properly center and scale the statistic

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

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

Back to Regression

Derive how is $\hat{\beta}$ an average.

Maximum Likelihood

Standard MLE:

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

$$
\begin{aligned}
\hat{\theta} &= \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i, \theta) \\
\Leftrightarrow 0 &= \mathbb{E}_n \left[\ell_\theta(z_i, \hat{\theta}) \right] = \mathbb{E}_n \left[\ell_{\theta\theta}(z_i, \bar{\theta}) \right] \mathbb{E}_n \left[\ell_\theta(z_i, \theta_0) \left(\hat{\theta} - \theta_0 \right) \right]
\end{aligned}
$$

So if $\mathbb{E}_n \left[\ell_{\theta\theta}(z_i, \bar{\theta}) \right] \to_p H(\theta_0) > 0$ (ULLN), then

$$
\left(\hat{\theta} - \theta_0 \right) = \frac{1}{n} \sum_{i=1}^{n} H(\theta_0)^{-1} \ell_\theta(z_i, \theta_0)
$$

This pattern is common: inverse of something times a "residual".