SDS 387
Linear Models

Fall 2025

Lecture 25 - Tue, Dec 2, 2025

Instructor: Prof. Ale Rinaldo

STATISTICAL INFERENCE IN ASSUMPTION-LEAN SETTINGS

- Assume just that the pair $(Y, \Phi) \sim P_{Y,\Phi}$ on $\mathbb{R} \times \mathbb{R}^d$ have each 2 moments. $\mathbb{E}[Y^2] < \infty$ and $\Sigma = \mathbb{E}[\Phi \Phi^T]$ is invertible

- Then we saw that the projection parameter

$$\beta^* = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \ \mathbb{E}\left[(Y - \Phi^T \beta)^2\right]$$

$$= \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \ \mathbb{E}\left[\left(\mathbb{E}[Y \mid \Phi] - \Phi^T \beta\right)^2\right]$$

$$= \Sigma^{-1} \underbrace{\mathbb{E}[Y \cdot \Phi]}_{\Gamma \in \mathbb{R}^d}$$

is well defined!

- In particular β^* satisfies the normal equations

$$\Sigma \beta^* = \Gamma$$

- Using the theory of L_2 projections, this implies that

$$\mathbb{E}\left[\left(Y - \Phi^T\beta^*\right)\Phi^T a\right] = \mathbb{E}\left[\left(\mathbb{E}[Y|\Phi] - \Phi^T\beta^*\right)\Phi^T a\right] = 0$$

$$\forall a \in \mathbb{R}^d$$

- So based on this

$$Y = \Phi^T\beta^* + \underbrace{\left(\mathbb{E}[Y|\Phi] - \Phi^T\beta^*\right)}_{\eta} + \underbrace{\left(Y - \overbrace{\mathbb{E}[Y|\Phi]}^{\text{regression function}}\right)}_{\varepsilon}$$

non-linearity
(= 0 if the model is well specified)

intrinsic variability

$$= \Phi^T\beta^* + \underset{\eta+\varepsilon}{\underset{\downarrow}{\delta}}$$

Importantly ~) $\mathbb{E}[\delta^2] = \mathbb{E}[\eta^2] + \underbrace{\mathbb{E}[\varepsilon^2]}$

Variance term because $\mathbb{E}[\varepsilon] = 0$ by law of iterated expectation

iii) η is orthogonal (in an L_2 sense) to the linear span of Φ:

$$\mathbb{E}[\eta \cdot \Phi(j)] = 0$$
$$j = 1, \dots, d$$

ε is orthogonal to all r.v.'s of the form $f(\Phi)$ any f s.t. $\mathrm{Var}[f(\Phi)] < \infty$

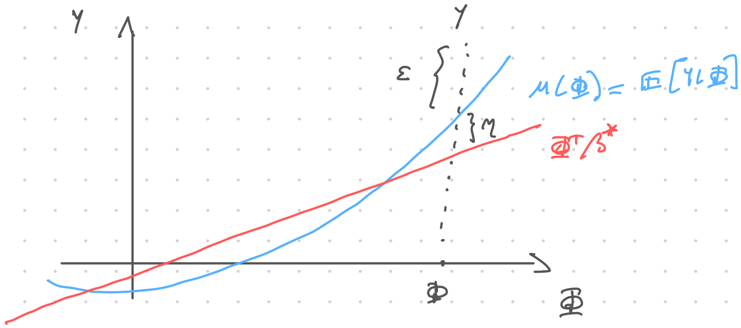
↳ $\mathbb{E}[\varepsilon \cdot \eta] = 0$

- When the model is not well-specified the distribution of Φ has be taken into account because β^* depends on it. This is of course no longer the case when the model is well-specified (i.e. $\mathbb{E}[Y | \Phi] = \Phi^T \beta^*$)

  ↳ For a general discussion see Proposition 4.1 in Buja et al. (2019)

- Take home message: when the model is not linear we are facing an extra source of variability, namely the non-linearity!

  ↳ See Figure 1 in Buja et al. (2019)



- <u>Remark</u>: the issue is not just an increase in variance, but also the fact that the variance of ε and η given Φ depends on Φ → variance not constant.
  conditional

- Assume n iid pairs (Y_i, Φ_i) $i = 1, \dots, n$ from the unknown data generating distribution $P_{Y,\Phi}$.

  We saw that we can estimate β^* using OLS

$$\hat{\beta} = \hat{\Sigma}^{-1} \hat{\Gamma} \quad \text{where}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \Phi_i \Phi_i^T \qquad \hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} Y_i \cdot \Phi_i$$

Then $\mathbb{E}[\hat{\beta}] \neq \beta^*$ (it is if the model is well-specified)

$$\mathrm{Var}[\hat{\beta}] = \mathbb{E}\left[\mathrm{Var}[\hat{\beta} \mid \Phi_1, \dots, \Phi_n]\right] + \underbrace{\mathrm{Var}\left[\mathbb{E}[\hat{\beta} \mid \Phi_1, \dots, \Phi_n]\right]}_{= 0 \text{ when the model is well-specified}}$$

- $\hat{\beta}$ is nonetheless a consistent estimator of β^*

$$\hat{\beta} \xrightarrow{P} \beta^*$$

Pf/ Recall that $\hat{\beta} = \hat{\Sigma}^{-1} \cdot \hat{\Gamma}$ (assume that $\hat{\Sigma}$ is invertible!)

So WLLN $\hat{\Sigma} \xrightarrow{P} \Sigma = \mathbb{E}[\Phi \Phi^T]$

$\hat{\Gamma} \xrightarrow{P} \Gamma = \mathbb{E}[Y \cdot \Phi]$

By CMT $\hat{\Sigma}^{-1} \xrightarrow{P} \Sigma^{-1}$ so by Slutsky's

$$\hat{\beta} \xrightarrow{P} \Sigma^{-1} \Gamma = \beta^* \quad \square$$

<u>Remark</u>: this is much more complicated in high-dim settings where $d = d(n)$.

- <u>CLT for $\hat{\beta}$.</u>

- Recall: if Φ is random and the model is linear

$$\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N_d(0, \sigma^2 \Sigma^{-1})$$

$\hookrightarrow \Sigma = \mathbb{E}[\Phi \Phi^T]$

- To establish a CLT for $\hat{\beta}$ is assumption lean settings, let's consider these quantities:

$$\psi_i = \Sigma^{-1} \Phi_i (Y_i - \Phi_i^T \beta^*) \qquad \text{not computable!}$$

$i = 1, \ldots, n$

Then
$$\frac{1}{n}\sum_{i=1}^{n}\psi_i = \Sigma^{-1}\left(\hat{\Gamma} - \hat{\Sigma}\beta^*\right)$$

Next,
$$\hat{\Sigma}\left(\hat{\beta}-\beta^*\right) = \hat{\Gamma} - \hat{\Sigma}\beta^*$$

$\hookrightarrow$
$$\Sigma^{-1}\hat{\Sigma}\sqrt{n}\left(\hat{\beta}-\beta^*\right) = \sqrt{n}\,\Sigma^{-1}\left(\hat{\Gamma} - \hat{\Sigma}\beta^*\right)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i$$

- Now, $\mathbb{E}[\psi_i] = 0$ by the normal equations

$$\mathrm{Var}[\psi_i] = \Sigma^{-1} V \Sigma^{-1}$$ where

by iid condition

$$V = \mathrm{Var}\left[\Phi_i\left(Y_i - \Phi_i^T\beta^*\right)\right]$$

the sandwich variance

- Remark if the model is well specified and $Y_i - \Phi_i^T\beta^* = \varepsilon_i \sim N(0, \sigma^2)$ then

$$\mathrm{Var}[\psi_i] = \sigma^2\Sigma^{-1}$$

- So by multivariate CLT

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i \xrightarrow{d} N_d\left(0, \Sigma^{-1}V\Sigma^{-1}\right)$$

$\hookrightarrow$
$$\Sigma^{-1}\hat{\Sigma}\sqrt{n}\left(\hat{\beta}-\beta^*\right) \xrightarrow{d} N_d\left(0, \Sigma^{-1}V\Sigma^{-1}\right)$$

But $$\Sigma^{-1}\hat{\Sigma}\sqrt{n}(\hat{\beta}-\beta^*) - \sqrt{n}(\hat{\beta}-\beta^*) = o_p(1)$$

because $\left(\Sigma^{-1}\hat{\Sigma} - I_d\right) \xrightarrow{p} 0$ by CMT

$$\hookrightarrow \boxed{\sqrt{n}(\hat{\beta}-\beta^*) \xrightarrow{d} N_d(0, \Sigma^{-1} V \Sigma^{-1})}$$

- Issue: we need to have a consistent estimator of the sandwich covariance. A natural estimator is the plug-in estimator:

$$\hat{\Sigma}^{-1}\hat{V}\hat{\Sigma}^{-1}$$

where $$\hat{V} = \frac{1}{n}\sum_{i=1}^{n} \Phi_i \Phi_i^T (Y_i - \Phi_i^T \hat{\beta})^2$$

- We need to show that

$$\hat{\Sigma}^{-1}\hat{V}\hat{\Sigma}^{-1} \xrightarrow{p} \Sigma^{-1} V \Sigma^{-1}$$

We already know that $\hat{\Sigma}^{-1} \xrightarrow{p} \Sigma^{-1}$. We need to show that $\hat{V} \xrightarrow{p} V$

- We will first define

$$\tilde{V} = \frac{1}{n}\sum_{i=1}^{n} \Phi_i \Phi_i^T (Y_i - \Phi_i^T \beta^*)^2 \quad \text{(not computable)}$$

By WLLN $\tilde{V} \xrightarrow{p} V$

So all we need to to is to prove that

$$\hat{V} - \tilde{V} \xrightarrow{p} 0$$

We have

$$\hat{V} - \tilde{V} = \frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{\Phi}_i \underset{\sim}{\Phi}_i^T \left[(\underset{\sim}{\Phi}_i^T \hat{\beta})^2 - (\underset{\sim}{\Phi}_i^T \beta^*)^2 + 2 Y_i \underset{\sim}{\Phi}_i^T (\beta^* - \hat{\beta}) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{\Phi}_i \underset{\sim}{\Phi}_i^T \Big[\underbrace{\left(\underset{\sim}{\Phi}_i^T (\hat{\beta} - \beta^*)\right)^2 + 2 \left(Y_i - \underset{\sim}{\Phi}_i^T \beta^*\right) \underset{\sim}{\Phi}_i^T (\beta^* - \hat{\beta})} \Big]$$

Next,

$$\| \hat{V} - \tilde{V} \|_{op} \le \frac{1}{n} \sum_{i=1}^{n} \| \underset{\sim}{\Phi}_i \|^2 \left[\quad \cdot \quad \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \| \underset{\sim}{\Phi}_i \|^2 \left(\underset{\sim}{\Phi}_i^T (\hat{\beta} - \beta^*)\right)^2 + \frac{2}{n} \sum_{i=1}^{n} \| \underset{\sim}{\Phi}_i \| \, | Y_i - \underset{\sim}{\Phi}_i^T \beta^* | \, \| \underset{\sim}{\Phi}_i \| \, | \underset{\sim}{\Phi}_i^T (\beta^* - \hat{\beta}) |$$

by Cauchy Schwartz

$$\le \underbrace{\frac{1}{n} \sum_{i=1}^{n} \| \underset{\sim}{\Phi}_i \|^2 \left(\underset{\sim}{\Phi}_i^T (\hat{\beta} - \beta^*)\right)^2}_{A} + 2 \sqrt{\underbrace{\frac{1}{n} \sum_{i=1}^{n} \| \underset{\sim}{\Phi}_i \|^2 (Y_i - \underset{\sim}{\Phi}_i^T \beta^*)^2}_{B}} \sqrt{\underbrace{\frac{1}{n} \sum_{i=1}^{n} \| \underset{\sim}{\Phi}_i \|^2 \left[\underset{\sim}{\Phi}_i^T (\beta^* - \hat{\beta}) \right]}_{A}}$$

$$= A + 2 \sqrt{A} \sqrt{B}$$

Next by Cauchy Schwartz

$$A \le \underbrace{\left[\frac{1}{n} \sum_{i=1}^{n} \| \underset{\sim}{\Phi}_i \|^4 \right]}_{\xrightarrow{p} \mathbb{E}[\| \underset{\sim}{\Phi}_i \|^4]} \underbrace{\| \hat{\beta} - \beta^* \|}_{\xrightarrow{p} 0} \xrightarrow{p} 0$$

$\xrightarrow{p} \mathbb{E}[\| \underset{\sim}{\Phi}_i \|^4]$ which we assume to be finite

As for B:

$$B \xrightarrow{p} tr\left(Var\left(\Phi_1 (Y_1 - \Phi_1^T \beta^*) \right)\right) = tr(V)$$

also finite

By Slutsky's

$$A + 2\sqrt{A}\sqrt{B} \xrightarrow{p} 0$$

↳ For large n:

$$\sqrt{n}(\hat{\beta} - \beta^*) \approx N_d(0, \hat{\Sigma}^{-1}\hat{V}\hat{\Sigma}^{-1})$$

- Remark: To carry out this program in high-dim settings is highly non-trivial.