SDS 387
Linear Models

Fall 2025

Lecture 19 - Tue, Nov 4, 2025

Instructor: Prof. Ale Rinaldo

OLS ESTIMATOR IN FIXED-DESIGN SETTING

Today we will assume

$$Y_i = \underset{\sim}{\Phi}_i^\top \beta^* + \varepsilon_i \qquad \text{where } \varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} (0, \sigma^2)$$

$i = 1, \ldots, n$

and $\Phi_1, \ldots, \Phi_n$ are fixed (deterministic) vectors in $\mathbb{R}^d$

$\downarrow$

2 assumptions

i) linearity

$$\mathbb{E}[Y_i] = \underset{\sim}{\Phi}_i^\top \beta^*$$

(Aside, if the $\underset{\sim}{\Phi}_i$'s were random linearity means $\mathbb{E}[Y_i \mid \underset{\sim}{\Phi}_i] = \underset{\sim}{\Phi}_i^\top \beta^*$)

ii) fixed covariates

- The fixed covariates assumption is unrealistic but you can assume that we are conditioning on the $\underset{\sim}{\Phi}_i$'s.

- When the model is linear and the covariates are random (i.e. $\mathbb{E}[Y_i \mid \Phi_i] = \Phi_i^\top \beta^*$) the distribution of the covariates (hence the covariates themselves) is ancillary. So it is natural to condition on the Φ_i's. This is only true if linearity holds (otherwise we know that the projection parameter $\beta^* = \mathbb{E}[\Phi_1\Phi_1^\top]^{-1}\,\mathbb{E}[Y_1\Phi_1]$ depends on the distribution of the covariates).

See Buja et al. (2019)

- Remark: If $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ then the likelihood of the data $Y_1, \dots, Y_n$ is

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{-\frac{\| \underset{n\times 1}{Y} - \underset{n\times d}{\Phi}\beta^* \|^2}{2\sigma^2}\right\}$$

and the OLS $\hat{\beta}$ is the MLE of β^*.

- Now, for any $\beta \in \mathbb{R}^d$, the risk of β is

vector in $\mathbb{R}^n$ of errors $\varepsilon = \begin{bmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{bmatrix}$

$$R(\beta) = \mathbb{E}_Y\left[\frac{\|Y - \Phi\beta\|^2}{n}\right] = \mathbb{E}_\varepsilon\left[\frac{\|\Phi(\beta^* - \beta) + \varepsilon\|^2}{n}\right]$$

$$= (\beta^* - \beta)^\top \underbrace{\frac{\Phi^\top\Phi}{n}}_{= \frac{1}{n}\sum_{i=1}^n \Phi_i \Phi_i^\top} (\beta^* - \beta) + \mathbb{E}_\varepsilon\left[\frac{\|\varepsilon\|^2}{n}\right]$$

$$= (\beta^* - \beta)^\top \hat{\Sigma} (\beta^* - \beta) + \sigma^2$$

$$= \|(\beta^* - \beta)\|^2_{\hat{\Sigma}} + \sigma^2$$

$$\downarrow \qquad\qquad \downarrow R(\beta^*)$$

The quantity $R(\beta) - R(\beta^*) \geq 0$ is the excess risk

- Remark: You can think of $R(\beta)$ as $\mathbb{E}_{Y|\Phi}\left[\frac{\|Y - \Phi\beta\|^2}{n} \,\middle|\, \Phi\right]$ if Φ were random (assuming linearity).

- So let's analyze $R(\hat{\beta}) - R(\beta^*)$, the excess risk of the OLS $\hat{\beta}$. Now this excess risk is a random variable! Let's compute its expectation:

- Remark: think of $R(\hat{\beta})$ as $\mathbb{E}_{Y_{new}}\left[\frac{\|Y_{new} - \Phi\hat{\beta}\|^2}{n}\right]$ where $Y_{new} \in \mathbb{R}^n$ is a new draw of data independent of the observations.

- $$\mathbb{E}[R(\hat{\beta})] - R(\beta^*) = \mathbb{E}_{\hat{\beta}}\Big[\underbrace{(\beta^* - \hat{\beta})^T \hat{\Sigma} (\beta^* - \hat{\beta})}\Big]$$

$$\|\beta^* - \hat{\beta}\|^2_{\hat{\Sigma}} = \|\beta^* - \mathbb{E}[\hat{\beta}] + \mathbb{E}[\hat{\beta}] - \hat{\beta}\|^2_{\hat{\Sigma}} = \dots$$

$= $ add / subtract $\mathbb{E}[\hat{\beta}]$

Exercise!

$$= \underbrace{\mathbb{E}\left[\|\hat{\beta} - \mathbb{E}[\hat{\beta}]\|^2_{\hat{\Sigma}}\right]}_{\text{variance term for } \hat{\beta}} + \underbrace{\|\beta^* - \mathbb{E}[\hat{\beta}]\|^2_{\hat{\Sigma}}}_{\text{bias term}}$$

$$\mathbb{E}\left[(\hat{\beta} - \mathbb{E}[\hat{\beta}])^T \hat{\Sigma} (\hat{\beta} - \mathbb{E}[\hat{\beta}])\right] = \mathbb{E}\left[\mathrm{tr}\left(\hat{\Sigma}(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])^T\right)\right]$$

$$= \mathrm{tr}\left(\hat{\Sigma}\, \underbrace{\mathbb{E}\left[(\hat{\beta}-\mathbb{E}[\hat{\beta}])(\hat{\beta}-\mathbb{E}[\hat{\beta}])^T\right]}_{\mathrm{Var}[\hat{\beta}]}\right)$$

Next, let's evaluate the variance and bias of $\hat{\beta}$:

<u>bias</u>: $\mathbb{E}[\hat{\beta}] = (\Phi^T\Phi)^{-1}\Phi^T \underbrace{\mathbb{E}[Y]}_{\Phi\beta^*} = \beta^*$ because $\hat{\Sigma}$ is invertible

$\downarrow$

no bias

$\downarrow$

This derivation is valid also when Φ is random:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\big[\underbrace{\mathbb{E}[\hat{\beta}\,|\,\Phi]}_{\beta^*}\big] = \beta^*$$

$\downarrow$

Linearity here is crucial.

<u>variance</u>

$$\mathrm{Var}[\hat{\beta}] = \mathrm{Var}\left[(\Phi^T\Phi)^{-1}\Phi^T Y\right]$$

$$= (\Phi^T\Phi)^{-1}\Phi^T \underbrace{\mathrm{Var}[Y]}_{\sigma^2 I_n} \Phi(\Phi^T\Phi)^{-1}$$

$\mathrm{Var}[AY] = A\,\mathrm{Var}[Y]\,A^T$

$$= \sigma^2(\Phi^T\Phi)^{-1}$$

- So we can now plug-in these expressions and get

$$\mathbb{E}[R(\hat{\beta})] - R(\beta^*) = \sigma^2\frac{d}{n}$$

Pf/ We only need to evaluate

$$\mathbb{E}\left[\| \hat\beta - \underbrace{\mathbb{E}[\hat\beta]}_{\beta^*} \|^2_{\hat\Sigma} \right] = \mathbb{E}\left[\| \cancel{\beta^*} + (\Phi^T\Phi)^{-1}\Phi^T\varepsilon - \cancel{\beta^*} \|^2_{\hat\Sigma} \right]$$

$(\Phi^T\Phi)^{-1}\Phi^T(\Phi\beta^* + \varepsilon)$

$$= \mathbb{E}\left[\| \hat\Sigma^{-1} \frac{\Phi^T\varepsilon}{n} \|^2_{\hat\Sigma} \right]$$

$$= \mathbb{E}\left[\varepsilon^T \frac{\Phi}{n} \hat\Sigma^{-1} \cancel{\hat\Sigma} \cancel{\hat\Sigma^{-1}} \frac{\Phi^T}{n}\varepsilon \right]$$

$$= \frac{1}{n} \mathbb{E}\left[\varepsilon^T \underbrace{\Phi(\Phi^T\Phi)^{-1}\Phi^T}_{H \text{ the hat matrix}} \varepsilon \right]$$

$$= \frac{1}{n} \mathbb{E}\left[\varepsilon^T H \varepsilon \right]$$

$$= \frac{1}{n} \mathbb{E}\left[\mathrm{tr}(H \varepsilon\varepsilon^T) \right]$$

$$= \frac{1}{n} \mathrm{tr}\left(H \underbrace{\mathbb{E}[\varepsilon\varepsilon^T]}_{\sigma^2 I_n} \right)$$

$$= \frac{\sigma^2}{n} \mathrm{tr}(H)$$

$$= \sigma^2 \frac{d}{n}$$

■

other proof:

$$\mathbb{E}\left[\| \hat\beta - \underbrace{\beta^*}_{\mathbb{E}[\hat\beta]} \|^2_{\hat\Sigma} \right] = \mathbb{E}\left[\mathrm{tr}\left(\hat\Sigma (\hat\beta - \beta^*)(\hat\beta - \beta^*)^T \right) \right]$$

$$= \mathrm{tr}\left(\hat\Sigma \, \mathrm{Var}[\hat\beta] \right)$$

$$= \mathrm{tr}\left(\hat\Sigma \frac{\sigma^2}{n} \hat\Sigma^{-1} \right)$$

$$= \frac{\sigma^2}{n} \mathrm{tr}(I_d) = \sigma^2 \frac{d}{n}$$

■

Remarks:

i) The $\sigma^2 \frac{d}{n}$ bound for the excess risk of $\hat{\beta}$ (OLS) is optimal, in a minimax sense

ii) More refined analysis will give you high prob. bounds for excess risk.

iii) Recall that this result tells that

$$\mathbb{E}\left[R(\hat{\beta})\right] = \mathbb{E}_{Y_{new}, Y}\left[\frac{\| Y_{new} - \Phi\hat{\beta}\|^2}{n}\right]$$

$$= \sigma^2\left(1 + \frac{d}{n}\right)$$

$\downarrow$

this is called the out-of-sample risk

What if we used the in-sample expected risk?

$$\mathbb{E}\left[\hat{R}(\hat{\beta})\right] = \mathbb{E}_{Y}\left[\frac{\| Y - \Phi\hat{\beta}\|^2}{n}\right] = \sigma^2\left(1 - \frac{d}{n}\right)$$

HW

$\downarrow$

Wrong measure of risk! The risk is at least $R(\beta^*) = \sigma^2$ while $\mathbb{E}\left[\hat{R}(\hat{\beta})\right] < \sigma^2$