SDS 387
Linear Models

Fall 2024

Lecture 25 - Tue, Dec 3, 2024

Instructor: Prof. Ale Rinaldo

- Last time: Assumption - lean inference: i) linear model is mis-specified; ii) covariates are random
  ↳ see Statistical Science paper Models as approximations, part I

- White (1980) Consequences and Detection of mis-specified non-linear regression Models, JASA 76, 374, 419 - 433

- $(\underset{d\times 1}{\Phi}, Y) \sim P_{\Phi, Y}$ on $\mathbb{R}^{d+1}$ but no assumptions on the regression function $x \in \mathbb{R}^d \longmapsto \mathbb{E}[Y \mid \Phi = x]$ is made. We only assume 2nd moments for Y and Φ.

- We can always write

$$Y = \underbrace{\mathbb{E}[Y \mid \Phi]}_{\substack{\text{signal} \\ \text{regression function}}} + \underbrace{Y - \mathbb{E}[Y \mid \Phi]}_{\substack{\varepsilon \\ \text{"noise", error}}}$$

where $\mathbb{E}[\varepsilon \mid \Phi] = 0$

$\mathbb{E}[\varepsilon] = 0$

- We saw (and you should do it as an exercise) that, even if the model is not linear, the <u>projection parameter</u>

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

$$= \Sigma^{-1} \cdot \Gamma$$

where $\Sigma = \mathbb{E}[\Phi \Phi^T]$ and $\underset{d \times 1}{\Gamma} = \mathbb{E}[\Phi \cdot Y]$

assuming that Σ is invertible (and assuming $\mathbb{E}[Y^2] < \infty$)

Furthermore β^* satisfies

$$\Sigma \beta^* = \Gamma \qquad \text{normal equations}$$

- β^* is the focus of inference : vector of coefficients of the "best" approximation of Y or $\mathbb{E}[Y \mid \Phi]$ by linear functions of Φ.

measure of multivariate linear association btw Y and Φ

- Last time we saw a fundamental decomposition :

$$Y = \Phi^T \beta^* + \underbrace{\left(\mathbb{E}[Y \mid \Phi] - \Phi^T \beta^*\right)}_{\substack{\text{non-linearity} \\ \eta}} + \underbrace{\left(Y - \mathbb{E}[Y \mid \Phi]\right)}_{\substack{\text{error} \\ \varepsilon}}$$

So $\qquad Y = \Phi^T \beta^* + \underbrace{\delta}_{\eta + \varepsilon}$

<u>Remark</u> : $\qquad \mathbb{E}[\delta^2] = \mathbb{E}[\eta^2] + \mathbb{E}[\varepsilon^2]$

i) η is orthogonal to the linear span Φ $\;\rightarrow \{a^T\Phi,\ a \in \mathbb{R}^d\}$

$$\left[\quad \mathbb{E}[\eta \cdot \underbrace{\Phi(j)}_{j^{th} \text{ coordinate of } \Phi}] = 0 \quad \forall j \quad \right]$$

ii) ε is orthogonal to all r.v.'s of the form $f(\Phi)$ where $f: \mathbb{R}^d \to \mathbb{R}$ s.t. $\mathbb{E}[f^2(\Phi)] < \infty$

$\hookrightarrow$ as a result, $\quad \mathbb{E}[\eta \cdot \varepsilon] = 0$

- Now the distribution of Φ has to be taken into account, because β^* depends on it.

if $\qquad Y = \underbrace{\Phi^T \beta^*}_{\text{some } \beta^*} + \underbrace{\varepsilon}_{\text{mean zero}}$

then $\quad \beta^*$ does not depend on the distribution of Φ

- Nonlinearity + random covariates $\rightarrow$ extra uncertainty

- Assume n iid observations from $P_{\Phi, Y}$:

$$(\Phi_1, Y_1), \dots, (\Phi_n, Y_n) \overset{iid}{\sim} P_{\Phi, Y}$$

$\hookrightarrow$ unknown

Let $\underset{n \times (d+1)}{\Phi} = \begin{bmatrix} 1 & \Phi_1^T \\ 1 & \Phi_2^T \\ & \vdots \\ 1 & \Phi_n^T \end{bmatrix}$ be the random design matrix

and consider the OLS estimator:

$$\boxed{\hat{\beta} = \hat{\Sigma}^{-1} \cdot \hat{\Gamma}}$$

where $\hat{\Sigma} = \dfrac{\Phi^T \Phi}{n} = \dfrac{1}{n} \sum_{i=1}^{n} \Phi_i \Phi_i^T$

$\downarrow$

plug-in estimator for β^*

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

<u>Remark</u>: $\mathbb{E}[\hat{\beta}] \neq \beta^*$

$$\mathrm{Var}[\hat{\beta}] = \mathbb{E}[\mathrm{Var}[\hat{\beta} \mid \Phi]] + \mathrm{Var}[\mathbb{E}[\hat{\beta} \mid \Phi]]$$

$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$

if $\mathbb{E}[Y \mid \Phi] = \Phi \beta^*$ fixed

then $\mathbb{E}[\hat{\beta} \mid \Phi] = \beta^*$

so $\mathrm{Var}[\mathbb{E}[\hat{\beta} \mid \Phi]] = 0$

- <u>Consistency</u>:

$$\hat{\beta} \xrightarrow{P} \beta^* \quad \text{as} \quad n \to \infty \quad (\text{keeping } d \text{ fixed!})$$

Pf/ $\hat{\beta} = \hat{\Sigma}^{-1} \cdot \hat{\Gamma}$

Now $\hat{\Sigma} \xrightarrow{P} \Sigma$ by WLLN and $\hat{\Sigma}^{-1} \xrightarrow{P} \Sigma^{-1}$ by CMT

Next $\hat{\Gamma} \xrightarrow{P} \Gamma$ by WLLN

So $\hat{\beta} = \hat{\Sigma}^{-1} \hat{\Gamma} \xrightarrow{P} \Sigma^{-1} \Gamma = \beta^*$ by Slutsky's theorem. ∎

<u>Remark</u>: When d grows with n, it is still not known how to eliminate the bias $\mathbb{E}[\hat{\beta}] - \beta^*$ efficiently.

- CLT for $\hat{\beta}$

To establish a CLT for $\hat{\beta}$, define

$$\psi_i = \Sigma^{-1} \Phi_i (Y_i - \Phi_i^T \beta^*) \in \mathbb{R}^d$$

$$i = 1, \ldots, n$$

Then:

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

Next,

$$\hat{\Sigma} (\hat{\beta} - \beta^*) = \hat{\Gamma} - \hat{\Sigma} \beta^*$$

Therefore:

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

$$= \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\psi_i$$

So we can focus on RHS. The ψ_i's are iid with

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

↳ exercise

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

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

asymptotic sandwich variance

If model is well specified and $y_i - \Phi_i^T\beta^* = \varepsilon_i \sim (0, \sigma^2)$

$\mathrm{Var}[\psi_i] = \sigma^2\Sigma^{-1}$ the usual asymptotic variance!

↳ exercise

By LF CLT:

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

Next $\Sigma^{-1}\hat{\Sigma} \xrightarrow{p} I_d$ so by Slutsky's theorem

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

↓

different asymptotic variance term in the well-specified case