SDS 387
Linear Models

Fall 2024

Lecture 24 - Tue, Nov 21, 2024

Instructor: Prof. Ale Rinaldo

- Last time: we assume a linear model with Gaussian covariates

$$Y_i = \Phi_i^\top \beta^* + \varepsilon_i \qquad i = 1, \dots, n$$

$$\Phi_1, \dots, \Phi_n \overset{iid}{\sim} N_d(0, I_d) \quad \text{and}$$

$$\varepsilon_1, \dots, \varepsilon_n \mid \Phi_1, \dots, \Phi_n \overset{iid}{\sim} (0, \sigma^2)$$

We are interested in the exact expression for the expected excess risk of $\hat{\beta}$ OLS. If $n \geq d+2$ then this value is

$$\sigma^2 \frac{d}{n-d-1}$$

If $n = d$ or $d+1$ this value is ∞.

Next if $d > n$ we will consider the min-norm least squares estimator

$$\hat{\beta}_{MN} = \Phi^{+} Y = \Phi^{T} \underbrace{(\Phi \Phi^{T})^{-1}}_{n \times n \text{ of full rank}} Y$$

$\underset{n \times d}{\Phi}$ matrix with ith row is Φ_i^T

Last time we started analyzing the expected excess risk of $\hat{\beta}_{MN}$:

$$\mathbb{E}\left[\|\beta^* - \hat{\beta}_{MN}\|^2\right] = T_1 + T_2$$

T_1: bias term

T_2: usual "variance" term

$$= \mathbb{E}\left[\|\underbrace{(I - \Pi)}\beta^*\|^2\right] + \sigma^2\, \mathbb{E}\left[\operatorname{tr}\left((\Phi \Phi^T)^{-1}\right)\right]$$

orthogonal projection onto null(Φ)

$$= \frac{n}{d}\|\beta^*\|^2 + \sigma^2\, \underbrace{\mathbb{E}\left[\operatorname{tr}\left((\Phi \Phi^T)^{-1}\right)\right]}$$

$(\Phi \Phi^T)^{-1}$ is an inverse Wishart with scale parameter I_n and d degrees of freedom. The diagonal entries are inverse of χ^2_{d-n-1} and have expectations equal to

$$\begin{cases} \frac{1}{d-n-1} & d \geq n+2 \\ \infty & d = n \text{ or } d = n+1 \end{cases}$$

So putting together all these terms:

$$\mathbb{E}\left[\|\beta^* - \hat{\beta}_{MN}\|^2\right] = \begin{cases} \sigma^2 \dfrac{d}{n-d-1} & d \le n-2 \\ \|\beta^*\|^2 \left[\dfrac{d-n}{d}\right] + \sigma^2 \dfrac{n}{d-n-1} & d \ge n+2 \\ \infty & \text{otherwise} \end{cases}$$

see Belkin, Hsu, Xu (2020)

Letting $\gamma = \lim_n \frac{d_n}{n}$

$$= \begin{cases} \sigma^2 \dfrac{\gamma}{1-\gamma} & \gamma < 1 \\ \|\beta^*\|^2 \left(1 - \dfrac{1}{\gamma}\right) + \sigma^2 \dfrac{1}{\gamma - 1} & \gamma > 1 \\ \infty & \gamma = 1 \end{cases}$$

- Ridge regression has smaller risk profiles when it is optimally tuned $\lambda = \dfrac{\sigma^2 \gamma}{\|\beta^*\|^2}$

tuning parameter for ridge

Solution: use cross-validation. See references.

INFERENCE

$$Y_i = \Phi_i^T \beta^* + \varepsilon_i$$

$$\Phi_i \overset{iid}{\sim} P_\Phi \qquad \varepsilon_i \mid \Phi_i \sim (0, \sigma^2) \text{ indep}$$

In this task, we are interested in estimating β^*, assuming a random design and well specified model.

If the model is well specified (i.e. $\mathbb{E}[Y \mid \Phi] = \Phi^T \beta^*$ and $\mathrm{Var}(Y - \Phi^T\beta^*) = \sigma^2$) this is a simple extension of the analysis we did in the fixed-design case.

- Assume throughout that $\frac{\Phi^T \Phi}{n} = \hat{\Sigma}_n$ is invertible with prob. 1. Then

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n \Phi_i \Phi_i^T$$

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

To see this, write

$$\hat{\beta} = \beta^* + \hat{\Sigma}^{-1} \frac{\Phi^T \varepsilon}{n}$$

and notice that

i) $\hat{\Sigma} \xrightarrow{P} \Sigma = \mathbb{E}[\Phi_i \Phi_i^T]$ by WLLN

so $\hat{\Sigma}^{-1} \xrightarrow{P} \Sigma^{-1}$ by CMT

ii) $\frac{\Phi^T \varepsilon}{n} \xrightarrow{P} \mathbb{E}[\Phi_i \cdot \varepsilon_i] =$

by WLLN $\quad \mathbb{E}_{\Phi_i}\Big[\Phi_i \underbrace{\mathbb{E}[\varepsilon_i \mid \Phi_i]}_{= 0}\Big] = 0$

$$\text{(}\sim\text{)} \quad \hat{\Sigma}^{-1} \frac{\Phi^T \varepsilon}{n} \xrightarrow{P} \Sigma^{-1} \cdot 0 = 0$$

by Slutsky

Also

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

This is because

$$\sqrt{n}\left(\hat{\beta} - \beta^*\right) = \hat{\Sigma}^{-1} \frac{\Phi^T \varepsilon}{\sqrt{n}}$$

$$\xrightarrow{P} \Sigma^{-1}$$

Next $$\frac{\Phi^T \varepsilon}{\sqrt{n}} = \sqrt{n} \sum_{i=1}^{n} \frac{\Phi_i \cdot \varepsilon_i}{n}$$

↳ normalized average of iid random vectors

We saw that $\forall i$ $\mathbb{E}\left[\Phi_i \cdot \varepsilon_i\right] = 0$. Also

$$\mathrm{Var}\left[\Phi_i \cdot \varepsilon_i\right] = \mathbb{E}\left[\varepsilon_i^2 \Phi_i \Phi_i^T\right]$$

$$= \mathbb{E}_{\Phi_i}\Big[\underbrace{\mathbb{E}\left[\varepsilon_i^2 \mid \Phi_i\right]}_{\sigma^2} \Phi_i \Phi_i^T\Big]$$

$$= \sigma^2 \mathbb{E}\left[\Phi \Phi_i^T\right] = \sigma^2 \Sigma$$

So by CLT

$$\sqrt{n}\,\frac{\Phi^T \varepsilon}{n} \xrightarrow{d} N_d(0, \sigma^2 \Sigma)$$

Finally by Slutsky's theorem:

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

$$= N_d(0, \sigma^2 \Sigma^{-1})$$

INFERENCE IN ASSUMPTION-LEAN SETTINGS (OR DISTRIBUTION-FREE)

Assumption-lean means
1) $\mathbb{E}[Y \mid \Phi] \neq \Phi^T \beta^*$
2) random design.

More generally, it means we are not willing to make assumptions beyond minimal ones that are needed for the model to be well-defined (e.g. moment assumptions)

So, assume $(\Phi, Y) \sim P_{\Phi, Y} \in \mathbb{R}^d \times \mathbb{R}$

covariates/ features/ regressors → Φ; Y → response

Let's only assume that Φ and Y have 2 moments

$\mathbb{E}[Y^2] < \infty$ and $\mathbb{E}[\Phi \Phi^T] = \underset{d \times d}{\Sigma}$ full rank

Then, trivially:

$$Y = \mathbb{E}[Y \mid \Phi] + \underbrace{Y - \mathbb{E}[Y \mid \Phi]}_{\varepsilon}$$

$\mathbb{E}[Y \mid \Phi]$: regression function $x \in \mathbb{R}^d \mapsto \mathbb{E}[Y \mid \Phi = x]$

$\varepsilon \longrightarrow$ by construction $\mathbb{E}[\varepsilon \mid \Phi] = 0$ so $\mathbb{E}[\varepsilon] = 0$

- In general $\mathrm{Var}[\varepsilon \mid \Phi]$ may depend on Φ

- We saw that

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

projection parameter

$$= \Sigma^{-1} \mathbb{E}[\Phi \cdot Y] \qquad \text{unique!}$$

In particular β^* satisfies the "normal equations":

$$\Sigma \beta^* = \mathbb{E}[\Phi \cdot Y]$$

This implies that $\forall a \in \mathbb{R}^d$

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

With this in mind, we have the following decomposition:

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

$$= \Phi^T\beta^* + \eta + \varepsilon$$

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

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

Also:

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

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

$$\Phi = \begin{bmatrix} \Phi(1) \\ \vdots \\ \Phi(d) \end{bmatrix}$$

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

In particular

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

$\hookrightarrow$ function of Φ