SDS 387
Linear Models

Fall 2025

Lecture 24 - Thu, Nov 20, 2025

Instructor: Prof. Ale Rinaldo

RANDOM DESIGN LINEAR REGRESSION

We still assume a well-specified linear regression model, but allow for the covariates (e.g. features) to be random. So now the model is

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

where Φ_i, $i=1,\ldots,n$, are iid random vectors in $\mathbb{R}^d$ from some distribution P_Φ and the errors are s.t.

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

- Note: the Φ_i's are ancillary for estimating β^* because the model is well specified
  ↳ it is ok to condition on the Φ_i's.

↳ See Buja et al. (2019) Stat Science

- So, we now observe n iid pairs

$$(Y_1, \Phi_1), \ldots, (Y_n, \Phi_n) \quad \text{in} \quad \mathbb{R} \times \mathbb{R}^d$$

↳ Remark: the first coordinate of each Φ_i is a 1, to allow for an intercept.

- Now the risk function is defined as:

$$\beta \in \mathbb{R}^d \longmapsto R(\beta) = \mathbb{E}_{\substack{Y, \Phi \\ \text{or } \varepsilon, \Phi}} \left[(Y - \Phi^T \beta)^2 \right]$$

Think of Y and Φ as (Y_{new}, Φ_{new}) a new iid draw from $P_{Y, \Phi}$

- Prop. 3.9 in Bach's book (Expression for the predictive risk).

Let $\underset{d \times d}{\Sigma_1} = \mathbb{E}\left[\Phi \Phi^T\right] \succeq 0$

Then, $\forall \beta \in \mathbb{R}^d$,

$$R(\beta) = \underbrace{(\beta - \beta^*)^T \Sigma_1 (\beta - \beta^*)}_{\|\beta - \beta^*\|^2_{\Sigma}} + \sigma^2$$

PA

$$R(\beta) = \mathbb{E}\left[\left(Y - \Phi^T\beta\right)^2\right] = \mathbb{E}\left[\left(Y - \Phi^T\beta^* + \Phi^T(\beta^* - \beta)\right)^2\right]$$

$$= \mathbb{E}\left[\left(Y - \Phi^T\beta^*\right)^2\right] + \mathbb{E}\left[\left(\Phi^T(\beta^* - \beta)\right)^2\right]$$

$$+ 2\underbrace{\mathbb{E}\left[\left(Y - \Phi^T\beta^*\right)\left(\Phi^T(\beta^* - \beta)\right)\right]}_{CV}$$

- If the model is linear, as we assume, then

$Y - \Phi^T\beta^* = \varepsilon$ is such that

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

$$\mathbb{E}[CV] = \mathbb{E}\left[\mathbb{E}[CV \mid \Phi]\right]$$

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

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

$$= 0.$$

- If is not linear, then our β^* is the projection parameter

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

projection parameter,

the coefficients of the L_2 projection of Y (or of $\mathbb{E}[Y|\Phi]$) onto the linear span of Φ

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

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

assuming $\Sigma \succ 0$ and $\mathbb{E}[Y^2] < \infty$

In this case

$\mathbb{E}[CV] = 0$ be the defining properties of L_2 projection

namely that $Y - \Phi^T\beta^*$ is uncorrelated

with any linear function of Φ.

- Regardless, $\mathbb{E}[CV] = 0$

$$\hookrightarrow R(\beta) = \mathbb{E}\left[\left(\Phi^T(\beta^* - \beta)\right)^2\right] + \mathbb{E}\left[\left(Y - \Phi^T\beta^*\right)^2\right]$$

$$= \|\beta - \beta^*\|^2_\Sigma + \sigma^2$$ if the model is well specified

$$= \|\beta - \beta^*\|^2_\Sigma + \overbrace{\mathbb{E}\left[\left(Y - \mathbb{E}[Y|\Phi]\right)^2\right]}^{\sigma^2}$$

$$+ \mathbb{E}\left[\left(\mathbb{E}[Y|\Phi] - \Phi^T\beta^*\right)^2\right]$$

η^2

non-linearity

↳ Thus $R(\beta)$ decomposes as a sum of

$\|\beta - \beta^*\|^2_{\Sigma}$ and an intrinsic, irreducible error: either σ^2 or $\sigma^2 + \eta^2$

■

↳ All we can do to minimize the risk is to minimize $\|\beta - \beta^*\|^2_{\Sigma}$ because σ^2 (or $\sigma^2 + \eta^2$) do not depend on β.

The excess risk now is

$$R(\beta) - \begin{cases} \sigma^2 & \text{linear model} \\ \sigma^2 + \eta^2 & \text{mis-specified model.} \end{cases}$$

• Assume we have data (n iid pairs $(y_i, \Phi_i)_{i=1,\dots,n}$)

then we can compute the OLS estimator

$$\hat{\beta} = \hat{\Sigma}^{-1} \frac{\sum_{i=1}^{n} y_i \Phi_i}{n}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \Phi_i \Phi_i^T = \hat{\mathbb{E}}_n [\Phi \Phi^T]$$

• The excess risk of $\hat{\beta}$ is $R(\hat{\beta}) = \|\hat{\beta} - \beta^*\|^2_{\Sigma}$

random variable

· **Prop 3.10** The expected excess risk of $\hat{\beta}$ is:

$$\mathbb{E}[R(\hat{\beta})] = \frac{\sigma^2}{n} \mathbb{E}\left[\operatorname{tr}(\Sigma \hat{\Sigma}^{-1})\right]$$

Remark $\mathbb{E}\left[\operatorname{tr}(\Sigma \hat{\Sigma}^{-1})\right] \geq d$ because

on the cone of PD matrices the map

$$A \mapsto \operatorname{tr}(A^{-1}) \text{ is convex}$$

$$\hookrightarrow \mathbb{E}\left[\operatorname{tr}(\Sigma \hat{\Sigma}^{-1})\right] \geq \operatorname{tr}\left(\mathbb{E}[\hat{\Sigma}]\Sigma^{-1}\right)$$

see page 64 of Bach's book

$$= d$$

Pf/ Write Φ for the $n \times d$ matrix with rows $\Phi_1^T, \dots, \Phi_n^T$.

So $\hat{\Sigma} = \frac{1}{n} \Phi^T \Phi$

Similarly let $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n$

So

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

Remark We are assuming throughout that $\hat{\Sigma}$ is invertible!

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

expected excess risk of $\hat{\beta}$

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

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

$$= \mathbb{E}_{\Phi, \varepsilon}\left[\operatorname{tr}\left(\Sigma \hat{\Sigma}^{-1} \frac{\Phi^T \varepsilon}{n} \varepsilon^T \frac{\Phi}{n} \hat{\Sigma}^{-1} \right)\right]$$

$$= \mathbb{E}_{\Phi}\left[\mathbb{E}_{\varepsilon | \Phi}\left[\quad " \quad \right]\right]$$

$$= \mathbb{E}_{\Phi}\left[\frac{1}{n} \operatorname{tr}\left(\Sigma \hat{\Sigma}^{-1} \frac{\Phi^T}{n} \underbrace{\mathbb{E}_{\varepsilon | \Phi}\left[\varepsilon \varepsilon^T | \Phi\right]}_{= \sigma^2 I_n} \Phi \hat{\Sigma}^{-1} \right)\right]$$

$$= \frac{\sigma^2}{n} \mathbb{E}_{\Phi}\left[\operatorname{tr}\left(\Sigma \hat{\Sigma}^{-1} \underbrace{\frac{\Phi^T \Phi}{n}}_{\cancel{\hat{\Sigma}}} \cancel{\hat{\Sigma}^{-1}} \right)\right]$$

$$= \frac{\sigma^2}{n} \mathbb{E}_{\Phi}\left[\operatorname{tr}\left(\Sigma \hat{\Sigma}^{-1} \right)\right]$$

$\boxtimes$

- Theorem 1 by Mourtada (2022) (AOS 50(4), 2157–2178)

i) Assume that $d > n$ or the distribution of the Φ_i's is degenerate (supported on a affine subspace of $\mathbb{R}^d$)

$\hat{\Sigma}$ is not invertible $\leftarrow$

Then the minimax risk is infinity !

iii) if $n \geq d$ and $\hat{\Sigma}$ is invertible then

the minimax risk :

$$\inf_{\tilde{\beta}} \sup_{\beta \in \mathbb{R}^d} \mathbb{E}_{\beta} \left[R(\tilde{\beta}) \right] = \frac{\sigma^2}{n} \mathbb{E} \left[\operatorname{tr}\left(\Sigma \hat{\Sigma}^{-1} \right) \right]$$

$\tilde{\beta}$ ↓ estimator

$\mathbb{E}_{\beta}$ ↓ expectation wrt Y, Φ where $Y = \Phi\beta + \varepsilon$

↓

OLS is minimax optimal !!