SDS 387
Linear Models

Fall 2024

Lecture 22 - Tue, Nov 14, 2024

Instructor: Prof. Ale Rinaldo

- HW 4 is now due Thu, Nov 21, by midnight and project report due on Monday Nov 18, by midnight.
- Last time: we finally finished discussing OLS properties under fixed-design, well-specified model.
- We are now dropping both assumptions. The lack of linearity and randomness of the design matrix create extra-complications in particular an increase in variability.
- Let's assume for now that $\underset{n \times d}{\Phi}$ is random but the model is well-specified. This means that our observations are

$$(Y_1, \Phi_1), \ldots, (Y_n, \Phi_n) \overset{iid}{\sim} P_{Y,\Phi} \text{ on } \mathbb{R} \times \mathbb{R}^d$$

$$Y_i = \Phi_i^T \beta^* + \varepsilon_i \quad \text{where } \varepsilon_1, \ldots, \varepsilon_n \mid \Phi_1, \ldots, \Phi_n \overset{iid}{\sim} (0, \sigma^2)$$

Now the definition of the risk has to be modified:

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

$\downarrow$ expectation wrt joint distribution of Y and Φ

Prop 3.9 in Bach's book

$$R(\beta) = \|\beta - \beta^*\|_\Sigma^2 + \sigma^2$$

$\rightarrow (\beta-\beta^*)^T \Sigma (\beta-\beta^*)$

where $\sigma^2 = R(\beta^*) = \inf_\beta R(\beta)$ and $\Sigma = \mathbb{E}[\Phi\Phi^T]$.

Pf/ For any $\beta \in \mathbb{R}^d$

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

$$= \mathbb{E}\left[\left(Y - \Phi^T\beta^*\right)^2\right] + \mathbb{E}\left[\left(\Phi^T(\beta^*-\beta)\right)^2\right] + \underbrace{2\,\mathbb{E}\left[\left(Y-\Phi^T\beta^*\right)\left(\Phi^T(\beta^*-\beta)\right)\right]}_{=0}$$

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

$$= \sigma^2 + \underbrace{(\beta^*-\beta)^T\,\mathbb{E}[\Phi\Phi^T]\,(\beta^*-\beta)}_{\|\beta^*-\beta\|_\Sigma^2}$$

$\blacksquare$

Because σ^2 is intrinsic noise quantity, we will focus on the excess risk:

$$R(\beta) - \sigma^2 = \|\beta^* - \beta\|_\Sigma^2$$

- So, now assume we observe $(Y_1, \Phi_1), \dots, (Y_n, \Phi_n)$

and compute the OLS $\hat{\beta} = \hat{\Sigma}^{-1} \frac{\sum_{i=1}^{n} y_i \cdot \Phi_i}{n}$

where $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \Phi_i \Phi_i^T$ which we assume to be invertible with probability 1. A sufficient condition for this is $\underset{d \times d}{\Sigma}$ is of full rank and $n \geq d$

$\Downarrow$

so the distribution of the Φ_i's does not concentrate on any affine linear subspace

Prop. 3.10 The expected excess risk of $\hat{\beta}$ (OLS estimator) is:

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

Pf/ Notation let $\hat{\Phi}$ be $n \times d$ matrix with rows given by $\Phi_1^T, \ldots, \Phi_n^T$. This is the random design matrix. So in particular

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

Also let Y and ε be n-dimensional vectors of responses and errors, so

because the model is linear

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

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

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

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

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

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

- Condition on the covariates
- Standard trick when $\hat{\Phi}$ is random, OK to do this when model is well-specified

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

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

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

potential issue : $\hat{\Sigma}$ may not be well conditioned !

↳ we will talk about this next week

- What if the model is not well specified? That is what if $\underset{\in \mathbb{R}}{\mathbb{E}}[Y | \underset{\in \mathbb{R}^d}{\Phi}] \neq \Phi^T \beta^*$ (the regression function is not linear).

Then we can define as our parameter

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

→ "best" (in L2) linear predictor of Y

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

= "best" linear approximation to the regression function in L_2

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

this unique as long as $\underset{d \times d}{\Sigma} = \mathbb{E}[\Phi \Phi^T]$ is invertible and $\mathbb{E}[Y^2] < \infty$.

- β^* is sometimes called the <u>projection</u> <u>parameter</u>
- More than one joint distribution of (Y, Φ) can have the same projection parameter!
- β^* is the vector of coefficients of the L_2 projection of Y onto the linear span of Φ (ie the set of r.v.'s of the form $\{\Phi^T\beta,\ \beta \in \mathbb{R}^d\}$)

  measure of linear association between Y and the vector Φ.

- Then, one can show that, in this situation the risk

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

$$\underbrace{\mathbb{E}\left[\left(Y - \mathbb{E}[Y \mid \Phi]\right)^2\right]}_{\sigma^2 \text{ intrinsic/unavoidable variance}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y \mid \Phi] - \Phi^T\beta^*\right)^2\right]}_{\text{non-linearity which is } = 0 \text{ when model is linear}} + \underbrace{\|\beta^* - \beta\|^2_{\Sigma}}_{\text{familiar term}}$$

Recall $\mathbb{E}[Y \mid \Phi] = \underset{f}{\text{argmin}}\; \mathbb{E}\left[(Y - f(\Phi))^2\right]$