SDS 387
Linear Models

Fall 2024

Lecture 15 - Thu, Oct 17, 2024

Instructor: Prof. Ale Rinaldo

- Reminder: no class on Tue, Oct 22

- HW3, Q2: it was rewritten and simplified. In the solutions, you will find the following result.

If $\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow{d} X$ and $g''(\mu) \neq 0$

then
$$n\left[g(\bar{X}_n) - g(\mu)\right] \xrightarrow{d} \frac{g''(\mu)}{2} X^2$$

Thus, if $\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$

$$n\,(\bar{X}_n^2 - \mu^2) \xrightarrow{d} \sigma^2 \chi_1^2(\mu^2)$$

which is well defined $\forall \mu \in \mathbb{R}$

- Uses Lemma 2.12 in van der Vaart

Let $R: \mathbb{R}^d \to \mathbb{R}$ s.t. $R(0) = 0$. Let $\{x_n\} \subset \mathbb{R}^d$ s.t. $X_n \xrightarrow{P} 0$. Then, $\forall p > 0$,

i) if $R(h) = o(\|h\|^p)$ then $R(X_n) = o_P(\|X_n\|^p)$

ii) if $R(h) = O(\|h\|^p)$ then $R(X_n) = O_p(\|X_n\|^p)$

- Last time: linear regression modeling

$$\mathbb{E}[Y \mid X = x] = x^T \beta \qquad \text{some } \beta \in \mathbb{R}^d$$

regression function

univariate response variable

d-dimensional vector of covariates

<u>Remarks</u> i) linearity here refers to β. We would call this model:

$$\mathbb{E}[Y \mid X = x] = \phi(x)^T \beta \qquad \beta \in \mathbb{R}^{d'}$$

also linear, where $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a <u>feature vector</u>. Example

$$\mathbb{E}[Y \mid X = x] = \alpha_0 + \alpha_1 x + \alpha_2 x^2$$

is a linear model (in $(\alpha_0, \alpha_1, \alpha_2)$).

ii) Typically we include an intercept term in the regression function:

$$\mathbb{E}[Y \mid X = x] = \beta_0 + x^T \beta$$

This is important for ANOVA testing and for a correct interpretation of R^2 coefficient.

We will always include the intercept, thus we will not write this explicitly. You can think of X or $\Phi(X)$ as a vector whose first coordinate is non-random and equal to 1

- 2 inferential tasks:
  1) statistical inference about β
  2) prediction

<u>statistical inference</u>

- If the model is well-specified (i.e. $\mathbb{E}[Y \mid X = x] = \beta^T x$ then β is clearly the parameter of interest.

- What if $\mathbb{E}[Y \mid X = x]$ is not linear? In this mis-specified setting we need to first identify the target parameter. This can be defined by looking at the best linear approximation to $\mathbb{E}[Y \mid X]$:

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

This is well-defined and unique provided that Y and X have 2 moment; in particular $\Sigma = \mathbb{E}[X X^T]$ needs to be invertible. Also, β^* is also equal to

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

<u>Thm</u> If Σ is invertible and Y has 2 moments

$$\beta^* = \Sigma^{-1} \mathbb{E}[Y \cdot X]$$

Pf/ β^* is the minimizer of

$$\mathbb{E}[(X^T\beta)^2] - 2\mathbb{E}[\mathbb{E}[Y|X] \cdot X^T\beta]$$

over all $\beta \in \mathbb{R}^d$

Because of moment assumptions we can take the derivative wrt to β inside the expectation and obtain the first order optimality condition

$$\mathbb{E}[\cancel{2} XX^T\beta] - \cancel{2}\, \mathbb{E}[\mathbb{E}[Y|X] \cdot X] = 0$$

Solution is $\mathbb{E}[XX^T]^{-1} \mathbb{E}[Y \cdot X]$

By convexity this is unique! ∎

<u>Remark</u>: What is β^*? It it is the vector of coefficients of the L_2 projection of Y onto the linear span of X (the vector space of all linear functions of X)!

- β^* is vector measuring linear association between Y and X

Prediction

In prediction (the main objective of ML models), we want to predict a new response, say Y^{new}, using X^{new}. Our goal is the to minimize the prediction error, i.e. to solve the problem

$$\min_{\beta \in \mathbb{R}^d} \underbrace{\mathbb{E}\left[\left(Y^{new} - (X^{new})^T \beta\right)^2\right]}_{\text{prediction MSE}}$$

Of course the solution is β^*. Suppose we instead use a different vector $\beta \in \mathbb{R}^d$. How large is the error that we make by using the wrong β?

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

$$= \mathbb{E}\left[(Y - X^T\beta^*)^2\right] + \mathbb{E}\left[(X^T(\beta^* - \beta))^2\right]$$
$$+ 2 \underbrace{\mathbb{E}\left[(Y - X^T\beta^*) X^T(\beta^* - \beta)\right]}_{= 0 \text{ by orthogonality of } L_2 \text{ projection}}$$

$$= \underbrace{\mathbb{E}\left[(Y - X^T\beta^*)^2\right]}_{\text{systematic error}} + \underbrace{(\beta^* - \beta)^T \Sigma (\beta^* - \beta)}_{\|\beta^* - \beta\|_\Sigma^2}$$

If $\mathbb{E}[Y \mid X] = X^T \beta^*$ then the systematic error is usually written as σ^2, the variance of Y

(E.g. assuming: $Y = X^T\beta + \varepsilon$ where $\varepsilon \sim (0, \sigma^2)$ $\varepsilon \perp\!\!\!\perp X$)

and $\|\beta^* - \beta\|_{\Sigma}^2$ is a measure of how well we are estimating the true regression function

DATA

- Suppose we observe a sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ of n points of iid realizations from the joint distribution of Y and X.

- I will write $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$ and

$$\underset{n \times d}{X} = [X_1 \; \ldots \; X_n]^T \quad \text{or} \quad \underset{n \times d}{\Phi} = [\varphi(X_1) \; \ldots \; \varphi(X_n)]^T$$

- To estimate β^* we will minimize the empirical MSE or predictive risk:

$$\hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \Phi(X_i)^T \beta)^2$$

expectation wrt empirical measure

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

$$= \frac{1}{n} \|Y - \Phi\beta\|^2$$

ordinary least squares

- The OLS estimator of β^* is the minimizer of $\hat{R}(\beta)$ over all $\beta \in \mathbb{R}^d$.

- Thm Assume that Φ is of full column rank ($\text{rank}(\Phi) = d \leq n$)

Then

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\text{argmin}}\, \hat{R}(\beta) = (\Phi^T \Phi)^{-1} \Phi^T Y$$

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

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

Φ_i i^{th} row of Φ

Notice that $\hat{\beta} = \left(\mathbb{E}_n[\Phi(X)\Phi^T(X)]\right)^{-1} \mathbb{E}_n[Y \Phi(X)]$

where $\mathbb{E}_n[\cdot]$ expectation wrt empirical measure

- $\hat{\beta}$ is the plug-in estimator for β^*

Pf/ $\beta \to \hat{R}(\beta)$ is strictly convex because Φ is of full column rank

[strict convexity follows because the Hessian of $\hat{R}(\beta)$ is $\underbrace{\frac{\Phi^T \Phi}{n}}_{d \times d}$ which is pd by assumption]

So the minimizer is found by first order optimality condition : $\nabla \hat{R}(\beta) = 0$ solve for β

This yields:

$$\nabla \hat{R}(\beta) = -\frac{2}{n} \Phi^T (Y - \Phi\beta) = 0$$

↳ Normal equations

Solution is

$$\hat{\beta} = \underbrace{(\Phi^T \Phi)^{-1}}_{\text{invertible by assumption}} \Phi^T Y$$

$\blacksquare$