SDS 387
Linear Models

Fall 2025

Lecture 17 - Tue, Oct 28, 2025

Instructor: Prof. Ale Rinaldo

- Last time: regression modeling. Task of relating a response variable Y to a vector of covariates $X \in \mathbb{R}^d$.

  Regression function $x \in \mathbb{R}^d \longmapsto \mathbb{E}[Y \mid X = x]$

  L_2 projection of Y on the vector space of functions of X

  target in regression

- We will be focussing on linear regression, i.e. we will approximate the regression function with a linear function

  $x \longmapsto x^T \beta$ some $\beta \in \mathbb{R}^d$

  recall the first coordinate of the vector x is a 1 to allow for an intercept

- 2 tasks in regression modeling

hypothesis testing
confidence sets

1) statistical inference: carry out inference on a target parameter. If the model is linear

i.e. $\mathbb{E}[Y \mid X=x] = x^T\beta^*$ some β^*

then our target is β^*.

What if the model is mis-specified (i.e. the regression function is not necessarily linear)

In this case there still exists a natural target parameter - the parameter β^* that gives us the "best" linear approximation to regression function

This is called the projection parameter. It is well defined provided that

$\mathbb{E}[Y^2] < \infty$ and $\mathbb{E}[XX^T]$ exists (is invertible)

It is equal to

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

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

$$= \left(\underbrace{\mathbb{E}[XX^T]}_{d \times d}\right)^{-1} \underbrace{\mathbb{E}[Y \cdot X]}_{d \times 1}$$

Pf/ β^* is the minimizer of

$$\beta \in \mathbb{R}^d \longmapsto \mathbb{E}[(X^T\beta)^2] - 2\,\mathbb{E}\left[\mathbb{E}[Y|X] \cdot (X^T\beta)\right]$$

The gradient at β is

$$\mathbb{E}[2\, XX^T\beta] - 2\,\mathbb{E}\left[\mathbb{E}[Y|X] \cdot X\right]$$

Because this is a strictly convex function it is enough to set gradient $= 0$ and solve for β. The solution is

$$\beta^* = (\mathbb{E}[XX^T])^{-1}\, \underbrace{\mathbb{E}\left[\mathbb{E}[Y|X] \cdot X\right]}_{= \mathbb{E}[Y \cdot X]}$$

$\blacksquare$

- Interpretation of β^*: vector of coefficient of the L_2 projection of Y onto the linear span of X (vector space of all linear combinations of X).

↓ it is a parameter expressing linear association btw Y and X

2) <u>prediction</u> : We want to predict a new observation of the response Y_{new} (which we do not observe) based on a new observation of the covariates X_{new}, which we do observe.

Formally, we want to minimize the prediction error:

$$\mathbb{E}\left[\left(Y_{new} - X_{new}^T \beta\right)^2\right]$$

over all $\beta \in \mathbb{R}^d$. The minimizer here is of course the projection parameter.

We want to quantify the prediction risk (let me write Y for Y_{new}, X for X_{new})

any $\beta \in \mathbb{R}^d$

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

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

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

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

$$= \mathbb{E}\left[\left(Y - X^T\beta^*\right)^2\right] + \underbrace{\|\beta^* - \beta\|_\Sigma^2 \quad \text{where } \Sigma = \mathbb{E}[XX^T]}_{\substack{\text{component of the} \\ \text{prediction risk that} \\ \text{depends on } \beta}}$$

Next

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

by orthogonality

$$= \underbrace{\mathbb{E}\left[\left(Y - \mathbb{E}[Y|X]\right)^2\right]}_{\substack{\sigma^2 \\ \text{unavoidable} \\ \text{variance}}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y|X] - X^T\beta^*\right)^2\right]}_{\substack{\text{non-linearity} \\ \text{term } \eta^2}}$$

So the prediction risk is

$$\mathbb{E}\left[\left(Y_{new} - X_{new}^T\beta\right)^2\right] = \underbrace{\|\beta - \beta^*\|_\Sigma^2}_{\substack{\text{estimation} \\ \text{error}}} + \underbrace{\sigma^2}_{\substack{\text{irreducible} \\ \text{variance}}} + \eta^2 \rightarrow \text{non-linearity}$$

$\downarrow$

by minimizing estimation error we minimize the prediction error.

- Suppose we observe a sample of n iid pairs

$$(Y_1, X_1), \ldots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$$

from some unknown joint distribution of (Y, X).

We will concatenate these observations:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$$

$$\underset{n \times d}{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$$ or following Bach's book

$$\underset{n \times d}{\Phi} = \begin{bmatrix} \varphi(X_1)^T \\ \vdots \\ \varphi(X_n)^T \end{bmatrix} = \begin{bmatrix} \Phi_1^T \\ \vdots \\ \Phi_n^T \end{bmatrix}$$ $\varphi(\cdot)$ a feature function

↓ feature matrix

↳ each $\Phi_i \in \mathbb{R}^d$

- We need to estimate β^* (either projection parameter or the actual parameter in a linear model)

We do this by minimizing the empirical MSE

$$\beta \in \mathbb{R}^d \mapsto \hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^n (Y_i - \Phi_i^T \beta)^2$$

$$= \hat{\mathbb{E}}_n \left[(Y_i - \Phi_i^T \beta)^2 \right]$$ ← expected value wrt the empirical measure: probability measure that puts $\frac{1}{n}$-mass on each observation

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

- The minimizer, $\hat{\beta}$, of $\hat{R}(\beta)$ over all $\beta \in \mathbb{R}^d$ is called the $\underbrace{\text{Ordinary Least Squares}}_{\text{OLS}}$ estimator

Assume that $\underset{n \times d}{\underline{\Phi}}$ is of full column rank

$$\text{rank}(\Phi) = d \leq n.$$

Then $$\hat{\beta} = \text{argmin } \hat{R}(\beta) = (\underbrace{\Phi^T \Phi}_{\substack{\text{well-defined} \\ \text{by rank assumption}}})^{-1} \Phi^T Y$$

$$= \left(\hat{\mathbb{E}}_n[\Phi_i \Phi_i^T]\right)^{-1} \hat{\mathbb{E}}_n[Y_i \cdot \Phi_i]$$

same expression for the projection parameter, except we are now using the empirical measure

↳ Plug-in estimator of β^*