SDS 387
Linear Models

Fall 2025

Lecture 16 - Tue, Oct 23, 2025

Instructor: Prof. Ale Rinaldo

- HW3 is out. For the delta method question you need to invoke Lemma 2.12 in van der Vaart's book.

Let $R: \mathbb{R}^d \to \mathbb{R}$ s.t. $R(0) = 0$. Let $\{X_n\} \subseteq \mathbb{R}^d$ be a sequence of random vectors s.t. $X_n \xrightarrow{P} 0$.

Then, $\forall p > 0$

$\forall \{h_n\}$ i) if $R(h_n) = o(\|h_n\|^p)$ then $R(X_n) = o_P(\|X_n\|^p)$

ii) if $R(h_n) = O(\|h_n\|^p)$ then $R(X_n) = O_P(\|X_n\|^p)$

- Let's finish the proof of L_2 projection.

Thm 11.1 (of van der Vaart) $\hat{S}$ is the projection of T onto $\mathcal{S}$ iff i) $\hat{S} \in \mathcal{S}$ and ii) $\mathbb{E}[(T-\hat{S})S] = 0$

orthogonality $\langle T-\hat{S}, S \rangle = 0$ $\forall S \in \mathcal{S}$

1

This projection is unique (in the sense if $\hat{S}'$ is another projection then $P(\hat{S} \neq \hat{S}') = 0$)

If $\mathcal{S}$ contains the constant functions then

$$\mathbb{E}[\hat{S}] = \mathbb{E}[T] \quad \text{and} \quad \text{cov}(T - \hat{S}, S) = 0 \quad \forall S \in \mathcal{S}$$

where T and S are r.v.'s in L_2 and $\mathcal{S}$ is a vector space

Pf/ Last time we show that if orthogonality holds then $\hat{S}$ is the unique projection.

Suppose $\hat{S}$ is a projection (that is $\hat{S} \in \underset{S \in \mathcal{S}}{\text{argmin}}\, \mathbb{E}[(S-T)^2]$)

Then $\forall \alpha \in \mathbb{R}$ $\forall S \in \mathcal{S}$

$$0 \leq \mathbb{E}\Big[(T - \underbrace{\hat{S} - \alpha S}_{\in \mathcal{S}})^2\Big] - \mathbb{E}[(T - \hat{S})^2]$$

$$= \alpha^2 \, \mathbb{E}[S^2] - 2\alpha \, \mathbb{E}[(T - \hat{S}) S]$$

This is a parabola in α that has to stay above the x-axis. The zeros of this parabola are

$$\alpha = 0 \quad \text{and} \quad \alpha = 2 \frac{\mathbb{E}[(T - \hat{S}) S]}{\mathbb{E}[S^2]}$$

$\hookrightarrow$ therefore $\mathbb{E}[(T - \hat{S}) S] = 0$ and

since S is generic, the orthogonality condition is satisfied.

If S contains the constant functions then, by orthogonality
$$\mathbb{E}[(T-\hat{S})\cdot c] = 0 \quad \forall c \in \mathbb{R}$$
$$\hookrightarrow \mathbb{E}[T] = \mathbb{E}[\hat{S}]$$
∎

<u>Corollary</u> Pythagora theorem for r.v.'s:

$$\mathbb{E}[T^2] = \mathbb{E}[\hat{S}^2] + \mathbb{E}[(T-\hat{S})^2]$$

Writing $\|X\|_2^2 = \mathbb{E}[X^2]$, this gives us

$$\|T\|^2 = \|\hat{S}\|^2 + \|T-\hat{S}\|^2$$

direct sum decomposition

- The most important type of L_2 projection is the conditional expectation. Suppose Y and X are L_2 random variables and let
$$S = \{ f(X),\ f \text{ is arbitrary s.t. } \mathbb{E}[f(X)^2] < \infty \}$$
I want to find the function g s.t. $\mathbb{E}[g(X)^2] < \infty$ and $\mathbb{E}[(Y-g(X))^2] \leq \mathbb{E}[(Y-f(X))^2]$

for all $f(X) \in S$.

Then $g(X) = \mathbb{E}[Y \mid X]$. This is true because $X \mapsto \mathbb{E}[Y \mid X]$ satisfies the orthogonality condition

$$\mathbb{E}\left[\left(Y - \mathbb{E}[Y \mid X]\right) f(X)\right] = \mathbb{E}\left[Y f(X)\right] - \mathbb{E}\left[\mathbb{E}[Y \mid X] f(X)\right]$$

$\Downarrow$

any $f(\cdot)$ s.t. $\mathbb{E}[f^2(X)] < \infty$ $\quad = 0 \quad$ by law of iterated expectation

$$\left(\mathbb{E}\left[\left(\mathbb{E}[Y \mid X]\right)^2\right] \leq \mathbb{E}\left[\mathbb{E}[Y^2 \mid X]\right] = \mathbb{E}[Y^2] < \infty\right)$$

by conditional Jensen

- Remark: Conditional expectation is well-defined even without a second moment. In this case the defining condition is

$$\mathbb{E}\left[Y \, 1\{X \in A\}\right] = \mathbb{E}\left[\mathbb{E}[Y \mid X] \, 1\{X \in A\}\right]$$

all set A.

LINEAR REGRESSION

- <u>General regression settings</u>:

response or dependent variable Y univariate r.v. of interest

- covariates
- independent variable
- features
- explanatory variables

X random vector in $\mathbb{R}^d$

- Our goal is to "model" or "learn" Y using X

- Assuming Y and X have finite second moments ($\mathbb{E}[Y^2] < \infty$ and

$$\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right]$$

$$= \mathbb{E}[XX^T] - \mathbb{E}[X](\mathbb{E}[X])^T$$

exist) this can be cast as the problem of finding or modeling a function $f : \mathbb{R}^d \to \mathbb{R}$ s.t.

$$\mathbb{E}\left[(Y - f(X))^2\right]$$ is "small".

- Hierarchy of modeling assumption

Agnostic or model free approach:

$$Y = \mathbb{E}[Y|X] + (\underbrace{Y - \mathbb{E}[Y|X]}_{\varepsilon \text{ error}})$$

$\downarrow$

best approximation of Y using X

The function $x \in \mathbb{R}^d \longmapsto \mathbb{E}[Y|X=x]$ is the regression function. So letting $f^{OPT}(X) = \mathbb{E}[Y|X]$

$$Y = f^{OPT}(X) + \varepsilon$$

signal error

Properties: i) $\mathbb{E}[\varepsilon] = \mathbb{E}[(Y - \mathbb{E}[Y|X])] = 0$

$\downarrow$ natural desirable property of ε

ii) $\mathbb{E}[f^{OPT}(X)\,\varepsilon] = \mathrm{Cov}[f^{OPT}(X)\,\varepsilon] = 0$

by orthogonality

$\hookrightarrow$ Remark This does not mean $\varepsilon \perp\!\!\!\perp f^{OPT}(X)$

- Signal + independent noise assumption non-parametric regression

$$Y = f(X) + \varepsilon \quad \text{where} \quad \mathbb{E}[\varepsilon] = 0$$
$$\varepsilon \perp\!\!\!\perp X$$

where f belongs a (large) class of functions satisfying certain properties. Typically these functions are assumed to be "smooth"

- Parametric regression

$$Y = f(X) + \varepsilon \qquad \mathbb{E}[\varepsilon] = 0$$
$$X \perp\!\!\!\perp \varepsilon = 0$$

where f is a function belonging to a parametric class, i.e. each f in this class is parametrized by a parameter $\theta \in \Theta \subseteq \mathbb{R}^m$

so regression function if f_{θ^*} some $\theta^* \in \Theta$

- Linear regression:

$$Y = f_{\theta}^*(X) + \varepsilon \qquad \mathbb{E}[\varepsilon] = 0 \quad X \perp\!\!\!\perp \varepsilon$$

where $f_{\theta^*}(X) = X^T \theta^*$, with

θ^* unknown but in a set $\Theta \subseteq \mathbb{R}^d$

<u>Remark</u>: to allow for an intercept term, we assume that the first coordinate of X is a constant, say 1. In this case

$$f_{\theta^*}(X) = \theta_1^* + \sum_{j=2}^{d} \theta_j^* X_j$$

- We usually need to have an intercept in your model, to make sense of ANOVA decomposition and R^2 statistic.

- <u>Mis-specified modeling</u>: You are in an agnostic setting

  Fit a parametric model and carry out inference on some well-defined projection parameter.

- The fixed-X setting: the covariates are deterministic

- 2 tasks: 1) Prediction: we want to predict a new instance of response variable, say Y_{new}, using a new instance of the covariates X_{new}.

We observe X_{new} and would like to predict Y_{new}

2) statistical inference on θ^* the "true" parameter indexing the regression function, or the projection parameter if the model is mis-specified