

# Elements of statistical learning: Chapter 3

July 21, 2020

## 1 Linear models

- Sampling properties of  $\hat{\beta}$

# LS estimator

Let  $\mathbf{X}$  be an  $N \times (p + 1)$  matrix of explanatory variables and  $\mathbf{y}$  an  $N \times 1$  vector of outputs. Then we know the LS estimator  $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

[see lecture slides "ESL1" for recap and proof].

## The "hat" matrix

As such for the fitted linear model

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} \\ &= \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_H \mathbf{y} \\ &= H\mathbf{y}\end{aligned}$$

where  $H$  is commonly referred to as the hat matrix.

# $H$ the projection matrix

Let us denote the column vectors of  $\mathbf{X}$  by  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_p$  with  $\mathbf{x}_0 \equiv 1$ .

- These vectors span a subspace of  $\mathbb{R}^N$ , also referred to as a column vector of  $\mathbf{X}$ .
- We minimize  $RSS(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$  by choosing  $\hat{\beta}$ , so that the residual vector  $\mathbf{y} - \hat{\mathbf{y}}$  is orthogonal to this subspace.
- the hat matrix  $\mathbf{H}$  computes the orthogonal projection, and hence it is also known as the projection matrix.

## Assumptions

- 1 Observations  $y_i$  are uncorrelated have constant variance  $\sigma^2$
- 2  $x_i$  are fixed (i.e. non-stochastic)

$$\begin{aligned}\text{var}(\hat{\beta}) &= \text{var} [(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta + (\mathbf{X}'\mathbf{X})^{-1}\epsilon] \\&= \text{var} [(\mathbf{X}'\mathbf{X})^{-1}\epsilon] \\&= \mathbb{E} \{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \epsilon [(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \epsilon]'\} \\&= \mathbb{E} \{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \epsilon \epsilon' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \} \\&= \mathbb{E} \{ (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{X}) \epsilon \epsilon' (\mathbf{X}'\mathbf{X})^{-1} \}\end{aligned}$$

Note that  $\epsilon$  is the error term and has zero mean and also remember that  $\mathbf{X}$  is fixed, and thus

$$\mathbb{E}[aZ] = a\mathbb{E}[Z]$$

where  $Z$  is a random variable and  $a$  is a constant. Therefore,

$$\begin{aligned} \text{var}(\hat{\beta}) &= \mathbb{E} \{ \epsilon \epsilon' (\mathbf{X}' \mathbf{X})^{-1} \} \\ &= (\mathbf{X}' \mathbf{X})^{-1} E \{ \epsilon \epsilon' \} \\ &= (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 \end{aligned}$$

where  $\sigma^2$  can be calculated by

$$\sigma^2 = \frac{1}{N - p - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

thus, assuming the errors are further Gaussian

$$\hat{\beta} \sim N(\beta, (\mathbf{X}' \mathbf{X})^{-1} \sigma^2)$$