LINEAR REGRESSION

GIVEN A PAIR OF RV $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, in linear regression analysis we are interested to relate Y with the observations X using a REGRESSION FUNCTION

$$m: \mathbb{R}^d \to \mathbb{R}$$

$$m(x) = \mathbb{E}(Y \mid X = x)$$

Why do we choose this form? Given a generic function $f: \mathbb{R}^d \to \mathbb{R}$, in general

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

LIE (LAW ITERATED EXPECTATIONS)

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

$$= \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] \mathbb{P}(A_i)$$

$$= \mathbb{E}\left[\underbrace{(f(X) - m(X))}_{\text{NUMERO, SCALARE}} \cdot \mathbb{E}\left[(m(X) - Y) \mid X\right]\right] =$$

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

$$= m(X) \qquad = \emptyset$$

$$\mathbb{E}\left[(f(X) - Y)^2\right] = \mathbb{E}\left[(f(X) - m(X) + m(X) - Y)^2\right] =$$

$$\mathbb{E}\left[(f(X) - m(X))^2 + 2\left((f(X) - m(X))(m(X) - Y)\right) + (m(X) - Y)^2\right]$$

$$= \mathbb{E}\left[(f(X) - m(X))^2\right] + \underbrace{2\mathbb{E}\left[(f(X) - m(X))(m(X) - Y)\right]}_{= \emptyset} + \mathbb{E}\left[(m(X) - Y)^2\right]$$

$$= \mathbb{E}\left[(f(X) - m(X))^2\right] + \mathbb{E}\left[(m(X) - Y)^2\right]$$

This quantity is minimized when $f(X) = m(X)$ because the quantity $\mathbb{E}\left[(m(X) - Y)^2\right]$ is unavoidable, is always present

So $m(X)$ satisfies the following $\mathbb{E}\left[(m(X) - Y)^2\right] = \min_{f: \mathbb{R}^d \to \mathbb{R}} \mathbb{E}\left[(f(X) - Y)^2\right]$ and we can say that the OPTIMAL APPROXIMATION (wrt the MEAN SQUARE) of Y by a function f is the regression function $m(X)$

X, Y are RVs so they are UNKNOWN $\xrightarrow{\text{BUT}}$ we can estimate $m(X)$ using <u>data</u>

LINEAR REGRESSION MODEL

$\mathcal{D} = \{(\bar{x}_1, y_1)\ (\bar{x}_2, y_2) \dots (\bar{x}_n, y_n)\}$ $(\bar{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ and they are <u>i.i.d.</u>

$$y_i = m(x_i) + \varepsilon_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_d x_{id} + \varepsilon_i$$

$$i = 1, \dots, n$$

! n is the dimension of the data set

d is the dimension of the vector $\rightarrow$ i.e. d coordinates

$\bar{x}_i = (x_{i1}, \dots, x_{id})^d$

θ_0 is the intercept , ε_i is the noise such that $\mathbb{E}[\varepsilon_i] = 0$

$\mathrm{Var}[\varepsilon_i] = \sigma^2$

$$\bar{x}_i \longrightarrow \bar{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{id} \end{pmatrix} \in \mathbb{R}^{d+1} \qquad \bar{\theta} = \begin{pmatrix} \theta_0 \\ \\ \theta_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

$$y_i = (1, x_{i1}, x_{i2}, \dots, x_{id}) \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_d \end{pmatrix} + \varepsilon_i = \bar{x}^T \cdot \bar{\theta} + \varepsilon_i \qquad (i = 1, \dots, n)$$

Since the previous is valid for each point of our dataset . We can do

$$\bar{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^n \quad , \quad \bar{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix} \in \mathbb{R}^n$$

$$X = \begin{pmatrix} - & \bar{x}_1^T & - \\ & \vdots & \\ - & \bar{x}_m^T & - \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & & x_{1d} \\ & & & \\ 1 & x_{m1} & & x_{md} \end{pmatrix}$$

$[m \times (d+1)]$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & & x_{1d} \\ & & & \\ 1 & x_{m1} & & x_{md} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \\ \theta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \\ \varepsilon_m \end{pmatrix} = \bar{y} = X\bar{\theta} + \bar{\varepsilon}$$

$$y_i = \sum_{j=0}^{d} X_{ij}\, \theta_j + \varepsilon_i$$

ε_i : Residuals

LINEAR REGRESSION MODEL

→ LEAST SQUARES ⇒ METHOD FOR ESTIMATING THE REGRESSION MODEL

MINIMIZING THE RESIDUALS SUM OF SQUARES (RSS)

$$RSS(\bar{\theta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} x_{ij}\theta_j \right)^2$$

In general, $\bar{u}^T\bar{u} = \sum_{i=1}^{K} u_i^2 = \|u\|_2^2 = \langle u, u \rangle$

$$RSS(\bar{\theta}) = (\bar{y} - X\bar{\theta})^T (\bar{y} - X\bar{\theta}) = (\bar{y}^T - \bar{\theta}^T X^T)(\bar{y} - X\bar{\theta})$$

$$= \bar{y}^T\bar{y} - \bar{y}^T X\bar{\theta} - \bar{\theta}^T X^T \bar{y} + \bar{\theta}^T X^T X \bar{\theta} \qquad (y^T X\bar{\theta})^T$$

$$= \bar{y}^T\bar{y} - 2\bar{\theta}^T X^T \bar{y} + \bar{\theta}^T X^T X \bar{\theta}$$

$n = 4$, $d = 2$ $\quad \frac{1}{2}2\begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = 8 \qquad 3\ 3\ 4\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ $\quad 6 + 6 + 8 = 20$

$\begin{bmatrix} 2 & 2 & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = 8 \qquad \begin{bmatrix} 8 & 6 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = 20$

$$\min_{\bar{\theta}} RSS(\bar{\theta}) \Rightarrow \left.\frac{\partial RSS(\bar{\theta})}{\partial \bar{\theta}}\right|_{\theta=\hat{\theta}} = \emptyset - 2X^T\bar{y} + 2X^T X\hat{\theta} = \emptyset$$

$\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_d)^T$ is the estimator the minimize RSS

$$\hat{\beta} = \arg\min_{\beta} RSS(\beta) = (X^T X)^{-1} X^T \bar{y}$$

EXAMPLE $\quad n = 3, \ d = 1$

$$\{(1, 3.1), (3, 10.5), (4, 17.6)\}$$

$$\bar{x} = \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix} \qquad \bar{y} = \begin{pmatrix} 3.1 \\ 10.5 \\ 17.6 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$$

SINGULAR MATRIX → NOT INVERTIBLE BECAUSE OF THE DETERMINANT

determinant = 0

→ if $X^T X$ is not invertible ?

REGULARIZATION

- RIDGE → it introduce a penalty term depending on the square (L^2-norm) of the vector of the regression coefficient

$$\hat{\theta} = \arg\min_{\theta} \left[\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{d} x_{ij} \theta_j \right)^2 + \lambda \sum_{j=1}^{d} \theta_j^2 \right]$$

$$RSS = \left(\bar{y} - X\bar{\theta} \right)^T \left(\bar{y} - X\bar{\theta} \right) + \lambda \bar{\theta}^T \bar{\theta}$$

$$\hat{\theta}^{RIDGE} = \left(X^T X + \lambda I \right)^{-1} X^T \bar{y}$$

- LASSO → PENALITATION TERM IS $\lambda \sum_{j=1}^{d} |\theta_i|$ (L^1-NORM)

This regularization has not CLOSED FORM

we need something to compute it