REGRESSION & BOOTSTRAP

- BOOTSTRAP OF POINTS
- BOOTSTRAP OF RESIDUALS

FIRST: ON THE MODELS OF REGRESSION

DATA: $(y_i, c_i)_{i=1}^{n}$

y_i IS AN OUTCOME OF Y_i WHERE

$$Y_i = c_i^T \beta + \varepsilon_i = \sum_{j=1}^{p} c_{ij} \beta_j + \varepsilon_i \; ; \; i=1,\dots,n$$

$c_i^T = (c_{i1}, c_{i2}, \dots, c_{ip})$ IS KNOWN, A COVARIATE WITH INFORMATION ABOUT Y_i.

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \rightarrow \text{PARAMETERS TO BE ESTIMATED}$$

ε_i I.I.D. R.V.'S : $\varepsilon_i \in N(0, \sigma^2)$

MULTIPLE LINEAR REGRESSION

$$Y_i = c_i^T \beta + \varepsilon_i \qquad i = 1, \dots, n$$

$$c_i^T = (1, x_{i1}, x_{i2}, \dots, x_{ik}) \qquad 1 \times (k+1)$$

$$\beta = \begin{pmatrix} a \\ b_1 \\ \vdots \\ b_k \end{pmatrix} \qquad p = k+1$$

$$p \ll n$$

(p is much smaller than n)

TASK: ESTIMATE β FROM

$$(y_i, c_i)_{i=1}^{n}$$

CRITERION $$Q(\beta) = \sum_{i=1}^{n} (y_i - c_i^T \beta)^2$$

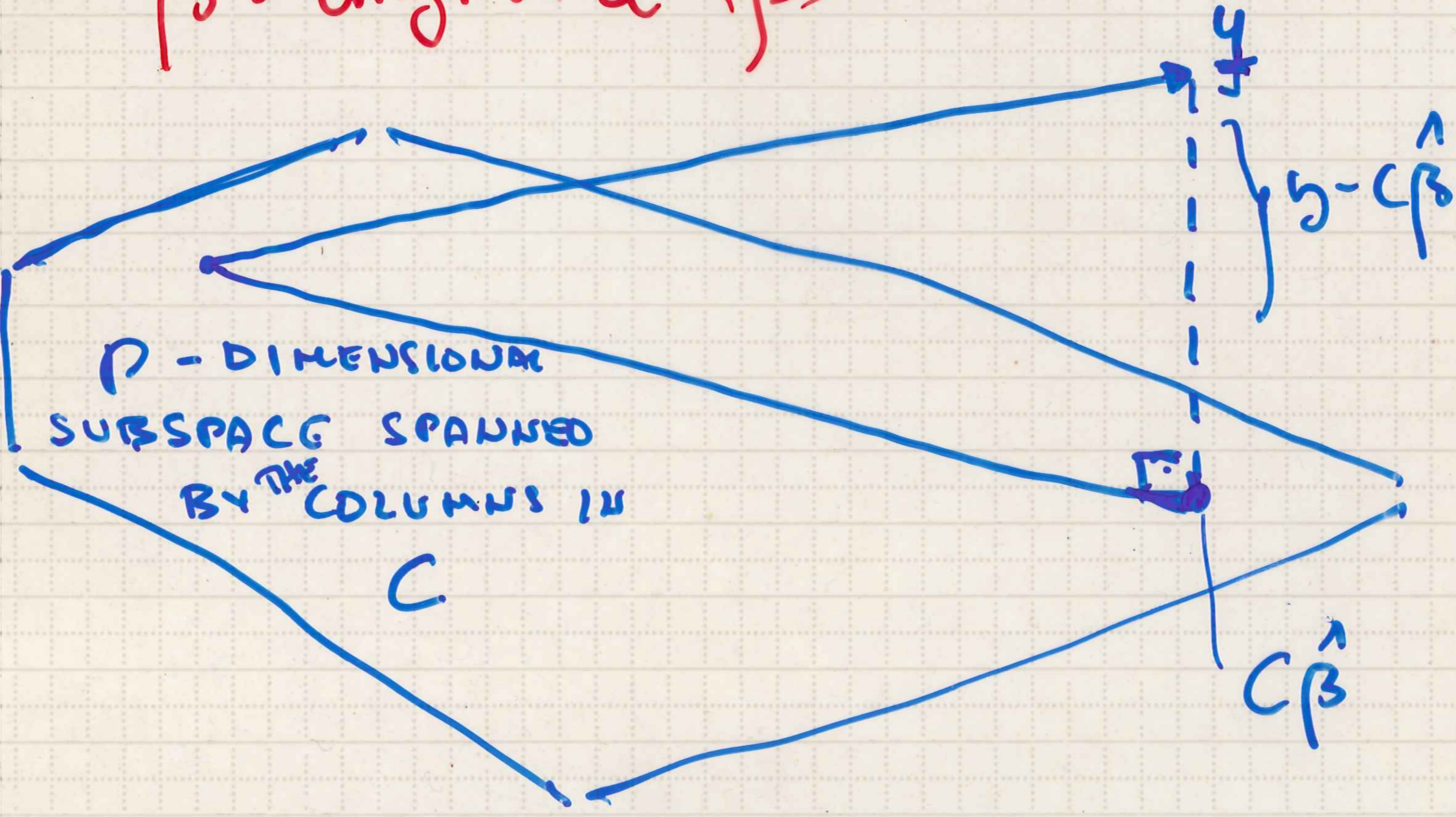
IS TO BE MINIMIZED.

$$Q(\beta) = \| y - C\beta \|^2$$

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad C = \begin{pmatrix} c_1^T \\ \vdots \\ c_n^T \end{pmatrix} \quad n \times p \text{ MATRIX}$$

$$\| \underline{x} \| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\hat{\beta} = \text{Argmin } Q(\beta)$$



$$\underline{Y} = C\beta + \underline{\varepsilon}$$

$\hat{\beta}$ SATISFIES

$$C^T C \hat{\beta} = C^T y$$

IF $(C^T C)^{-1}$ EXISTS

$$\hat{\beta} = (C^T C)^{-1} C^T y$$

IT CAN BE SHOWN THAT

1) $E_F(\hat{\beta}) = \beta$ $\quad \beta = T(F)$

2) $\mathrm{Cov}(\hat{\beta}) = \sigma^2 (C^T C)^{-1}$

SIMPLE LINEAR REGRESSION

$$Y_i = a + b x_i + \varepsilon_i$$

$$\beta = \begin{pmatrix} a \\ b \end{pmatrix} \qquad c_i = (1, x_i)^T = \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

DECENTERING

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

$$c_i^T = (1, x_i - \bar{x}) \qquad \beta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

$$\cancel{\alpha = a + \beta \bar{x}}$$

$$\alpha = a - \beta \bar{x}$$

$$C = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$$

$$C^T C = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{pmatrix}$$

$$(C^T C)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{1}^{n} (x_i - \bar{x})^2} \end{pmatrix}$$

$$\boxed{Cov(\beta) = \sigma^2 (C^T C)^{-1}}$$

1) BOOTSTRAP OF POINTS

$(y_i, c_i)_{i=1}^{n}$ SAMPLES OF A $p \times 1$ DIMENSIONAL DISTRIBUTION

$\hat{F}$: empirical dist'n

BOOTSTRAP SAMPLES $(Y_i^*, C_i^*)_{i=1}^{n}$

$$P\left((Y_j^*, C_j^*) = (y_i, c_i) \right) = \frac{1}{n} \quad j=1,\dots,n$$

THEN WE HAVE

$$C^* = \begin{pmatrix} c_1^* \\ \vdots \\ c_n^* \end{pmatrix}$$ $n \times p$ -MATRIX

$$\hat{\beta}^* = (C^{*T} C^*)^{-1} C^{*T} Y^*$$

2) BOOTSTRAP OF RESIDUALS

MODEL: $y_i = c_i \beta + \varepsilon_i \qquad i=1, \dots, n$

STEP ONE: ESTIMATE β USING $(c_i, y_i)_{i=1}^{n}$

$\rightarrow \hat{\beta}$

COMPUTE THE RESIDUALS

$$\hat{\varepsilon}_i = y_i - c_i \hat{\beta} \qquad i=1, \dots, n$$

STEP TWO: USE BOOTSTRAP OF $\hat{\varepsilon}_i$

TO GET $\hat{\varepsilon}^* = \begin{pmatrix} \hat{\varepsilon}_1^* \\ \vdots \\ \hat{\varepsilon}_n^* \end{pmatrix}$

AND THEN

$$y_i^* = c_i \hat{\beta} + \hat{\varepsilon}_i^*$$

$$P\left(\hat{\varepsilon}_j^* = \hat{\varepsilon}_j\right) = \frac{1}{n} \qquad j=1, \dots, n$$

STEP THREE:

$$\hat{\beta}^{*} = (C^T C)^{-1} C^T y^{*}$$

REPEAT : STEP ONE – STEP THREE

B TIMES, THEN WE HAVE

$$\hat{\beta}^{*1}, \hat{\beta}^{*2}, \ldots, \hat{\beta}^{*B}$$

+ APPROXIMATE THE DISTN OF $\hat{\beta}^{*}$ ∎