$$Y = X\theta^* + \varepsilon \qquad (y_i, x_i)$$

$$\varepsilon \sim \xi \qquad \mathbb{E}[\varepsilon] = 0, \qquad Var(\varepsilon) = \sigma^2$$

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - X\theta\|^2 \qquad , X \in \mathbb{R}^{n \times (p+1)} \qquad Y \in \mathbb{R}^n$$

$$\theta, \hat{\theta} \in \mathbb{R}^{p+1}$$

$$\downarrow$$

$$(X^T X)\hat{\theta} = X^T Y \quad \text{Normal equation}$$

▷ The solution to OLS is unique iff $(X^T X)^{-1} \exists$

exer 1

$X = 1_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ X, y, What is the OLS?

$$\hat{\theta} = (X^T X)^{-1} X^T Y = (1_n^T 1_n)^{-1} 1_n^T Y = \left[(1 \cdots 1) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right]^{-1} 1_n Y =$$

$$= n^{-1} 1_n^T Y = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

exer 2

Show $\|\underbrace{Y - \hat{Y}}_{\hat{\varepsilon}}\|^2 \leq \|Y - \bar{Y} 1_n\|^2$

Let $X \in (1_n, \tilde{X})$ for $\tilde{X} \in \mathbb{R}^{n \times p}$, $X \in \mathbb{R}^{n \times (p+1)}$, $1_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$

let $\theta = (\theta_0, \theta_1, \dots, \theta_p)^T = (\theta_0, \tilde{\theta})^T$, $\theta_0 \in \mathbb{R}$, $\tilde{\theta} \in \mathbb{R}^p$

$$\min_{\theta \in \mathbb{R}^{p+1}} \| Y - X\theta \|^2 = \min_{\substack{\theta_0 \in \mathbb{R}, \\ \tilde{\theta} \in \mathbb{R}^p}} \| Y - (1_n, \tilde{X}) \begin{pmatrix} \theta_0 \\ \tilde{\theta} \end{pmatrix} \|^2$$

$$\leq \min_{\theta_0} \| Y - 1_n \theta_0 \|^2 = \| Y - \bar{Y} 1_n \|^2$$

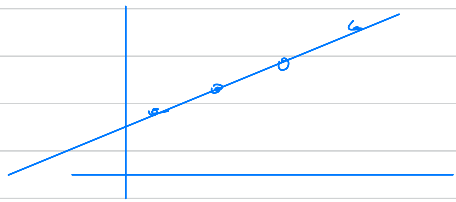
thus $\| Y - \hat{Y} \|^2 \leq \| Y - \bar{Y} 1_n \|^2$ $\square$

Proposition $0 \overset{①}{\leq} R^2 \overset{②}{\leq} 1$

Recall
$$R^2 = 1 - \frac{\| Y - \hat{Y} \|^2}{\| Y - \bar{Y} 1_n \|^2} \qquad Y = \hat{Y}$$

② for $\hat{Y} = Y \Rightarrow R^2 = 1$

① by exer2



Recall : X is random matrix
A, b deterministic matrix / vector
$\Rightarrow \mathrm{cov}(AX + b) = A\, \mathrm{Cov}(X) A^T$

Properties (we assume $\ker(X) = 0$, i.e $(X^T X)^{-1}$) $\exists$s

property 1 : $\hat{\theta}$ is unbiased θ^*, i.e $\mathbb{E}[\hat{\theta}] = \theta^*$

only random

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[(X^T X)^{-1} X^T Y] = \mathbb{E}[(X^T X)^{-1} X^T (X\theta^* + \varepsilon)] =$$

$$= (X^T X)^{-1} X^T X \theta^* + (X^T X)^{-1} X^T \underbrace{\mathbb{E}[\varepsilon]}_{0} = \theta^*$$

P2: $\text{cov}(\hat\theta) = \sigma^2 (X^T X)^{-1}$

$$\text{cov}(\hat\theta) = \text{cov}\big(\underbrace{(X^T X)^{-1} X^T}_{A} Y\big) = (X^T X)^{-1} X^T \underbrace{\text{cov}(Y)}_{\sigma^2 I} \big((X^T X)^{-1} X^T\big)^T$$

$$\sigma^2 (X^T X)^{-1} \underline{X^T X} \, \underline{(X^T X)^{-1}} = \sigma^2 (X^T X)^{-1}$$

property 3 $\quad \hat\theta$ is the $\underbrace{\text{Best}}_{\text{min var}}$ $\underbrace{\text{Linear}}_{AY}$ $\underbrace{\text{Unbiased}}_{\mathbb{E}[\hat\theta] = \theta^*}$ Estimator (BLUE) $\quad \theta^*$

By linear we mean that $\hat\theta = AY$. Note that

$\hat\theta_{OLS} = \underline{(X^T X)^{-1} X^T} Y$. Thus, in OLS $A = (X^T X)^{-1} X^T$

$\mathbb{E}[AY] = \theta^*$ (its unbiased)

$\Leftrightarrow A\mathbb{E}[Y] = \theta^* \Leftrightarrow A\,\mathbb{E}[X \cdot \theta^* + \varepsilon] = \theta^*$

$\Rightarrow \underline{A} X \theta^* = \theta^* \quad \Leftrightarrow \quad AX = I$.

$\text{cov}(AY) = A\,\text{cov}(Y) A^T = \sigma^2 A A^T$

$\Rightarrow$ We choose A so that the diagonal of AA^T is min subject to those that satisfy $AX = I$

$$AA^T = \left(\left(A - (X^TX)^{-1}X^T\right) + \left((X^TX)^{-1}X^T\right)\right)\left(\left(A-(X^TX)^{-1}X^T\right) + (X^TX)^{-1}X^T\right)^T$$

$$\Rightarrow \left(A - (X^TX)^{-1}X^T\right)\left((X^TX)^{-1}X^T\right)^T = \left(A - (X^TX)^{-1}X^T\right)\left(X(X^TX)^{-1}\right) =$$

$$\Rightarrow = \underbrace{AX}_{I}(X^TX)^{-1} - (X^TX)^{-1}\underline{X^TX}(X^TX)^{-1} = 0$$

$$AA^T = \underbrace{\left(A - (X^TX)^{-1}X^T\right)}_{B}\underbrace{\left(A-(X^TX)^{-1}X^T\right)^T}_{B^T} + (X^TX)^{-1}$$

$$BB^T \qquad [BB^T]_{ii} = \sum_k B_{ik}^2 \geq 0$$

The diagonal in AA^T (i.e., the variance of the linear estimator AY) is ≥ 0 always. When is it "$=0$"?

We choose A. $AA^T = 0$: $\text{diag}(AA^T) = 0$

$\Rightarrow A = (X^TX)^{-1}X^T \rightarrow$ Note. this choice satisfies $AX = I$

$\Rightarrow A = (X^TX)^{-1}X^T$ is the BLUE $\square$

<u>exer: 3</u> Show that the predicted value $\hat{y}$ is invariant to linear changes on X

$X = [x_0, \ldots, x_p] \qquad Z = [c_0 x_0, c_1 x_1, \ldots, c_p x_p]$

How to write the transformed proj?

$D = \mathrm{diag}(c_0, c_1, \dots, c_p)$

$Z = XD$

$\hat{\theta}_x = (X^T X)^{-1} X^T Y$

$\hat{\theta}_z = (Z^T Z)^{-1} Z^T Y = ((XD)^T (XD))^{-1} (XD)^T Y =$

$= (D^T X^T X D)^{-1} D^T X^T Y = D^{-1} (X^T X)^{-1} \underbrace{D^{-1} D}_{I} \cancel{X}^T Y =$

$= D^{-1} \hat{\theta}_x$ Note! typo

$x_0 \in \mathbb{R}^{p+1} \Rightarrow$ point prediction $\hat{\theta}_x^T x_0$

let z_0 the point with linear changes $z_0 = D x_0$

point prediction for z_0?

$\hat{\theta}_z^T z_0 = (D^{-1} \hat{\theta}_x)^T (D x_0) = \hat{\theta}_x^T D^{-1} D x_0 = \hat{\theta}_x^T x_0$ $\square$

Cochran's lemma

Rem X is fixed Gaussian noise

$\varepsilon_i \sim N(0, \sigma^2)$. $Y = X\theta^* + \varepsilon$, X full rank

$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

slide 59 in session 1

Hat matrix. $H = X(X^T X)^{-1} X^T$

$H^2 = H = H'$

$\hat{Y} = X\hat{\theta} = \underbrace{X(X^T X)^{-1} X^T}_{H} Y = HY$

Note that $(I - H)X = 0$

(1) $\hat{\theta} \sim N(\theta^*, \sigma^2 (X^T X)^{-1})$

(2) $(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$

(3) $\hat{\theta}$, $\hat{\sigma}^2$ are independent

(4) $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$ is unbiased

(5) - Relation to T-student distn : next week

For (1)

$$\hat{\theta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\theta^* + \varepsilon) =$$

$$= (X^T X)^{-1} X^T X \theta^* + (X^T X)^{-1} X^T \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$\hat{\theta}$ is gaussian, characterized $\mathbb{E}[\hat{\theta}]$, $Var(\hat{\theta})$

(2) $V = (V_1, V_2)$

V_1 is a basis for span (X)

V is orthogonal $\in \mathbb{R}^{n \times n}$

$$\rightarrow \begin{cases} V_1^T (I-H) = 0 \\ V_2^T (I-H) = V_2^T \end{cases}$$

$H \cdot X$
$(I-H)X$

not matrix

$$(n-p-1)\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \| Y - \hat{Y} \|^2 =$$

$$\| Y - HY \|^2 = \| (I-H) Y \|^2 = \| (I-H)(X\theta^* + \varepsilon) \|^2$$

$$= \| (I-H)\varepsilon \|^2 = \| V^T (I-H) \varepsilon \|^2 =$$

$$= \| V_2^T \hat{\varepsilon} \|^2$$

let $\tilde{\varepsilon} = \underline{V_2^T \varepsilon}_{\varepsilon}$ then

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{n-p-1} \tilde{\varepsilon}_i^2 \sim \chi^2_{n-p-1}$$

$\varepsilon \sim N$

(3) Note $X^T (I - H) = 0$

$$\frac{1}{n-p-1}\hat{\sigma}^2 = \| \hat{\varepsilon} \|^2 = \| (I-H)\varepsilon \|^2$$

$$\hat{\theta} - \theta^* = (X^T X)^{-1} X^T \varepsilon$$

(4) $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$

$$\mathbb{E}\Big[\underbrace{\frac{1}{n-p-1} \frac{\hat{\sigma}^2}{\sigma^2}}_{\text{random var.}} \Big] = n-p-1$$

χ^2_{n-p-1} based on (2)

$$\frac{1}{n-p-1} \cdot \frac{\mathbb{E}[\hat{\sigma}^2]}{\sigma^2} = n-p-$$

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

MAXIMUM LIKELIHOOD ESTIMATION

let $x_i \sim \mathcal{N}(\mu, \sigma^2)$. $S = \{x_1, \ldots, x_n\}$ is a sample. Give the MLE for the params, $\hat{\mu}, \hat{\sigma}^2$.

Recall, the density is

$$p(x_i, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma^2} \right)$$

1) give the likelihood of the sample $x_1 \ldots x_n$

2) give the log-likelihood

3) derivate w.r.t. μ and σ^2

4) solve the equation in (3)

$S = \{x_1, \ldots, x_n\} \qquad x_i \sim \mathcal{N}(\_, \_)$

1) $\ell(\mu, \sigma^2; S) = \prod_{i=1}^{n} p(x_i; \mu, \sigma^2) =$

$$= \prod_{i=1}^{n} (2\pi\sigma)^{-\frac{1}{2}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma)^{-\frac{n}{2}} \exp\left(-\sum \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

2) $\mathcal{L}(\mu, \sigma^2, S) = -\frac{n}{2} \lg(2\pi) - \frac{n}{2} \lg \sigma^2 - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$

3) $\underset{\mu, \sigma^2}{\operatorname{argmax}} \mathcal{L}(\mu, \sigma^2; S) \iff \nabla \mathcal{L}(\mu, \sigma^2, S) = 0$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2} \sum (x_i - \hat{\mu})^2 = 0$$

$$\sum (x_i - \hat{\mu}) = 0 \longrightarrow \boxed{\hat{\mu} = \frac{1}{n} \sum x_i = \bar{x}}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \frac{-n}{2\hat{\sigma}^2} - \left(\frac{1}{2} \sum (x_i - \hat{\mu})^2\right) \underbrace{\frac{d}{d\hat{\sigma}^2}\left(\frac{1}{\sigma^2}\right)}_{\frac{1}{(\hat{\sigma}^2)^2}} = 0$$

$$\frac{1}{\hat{\sigma}^2}\left(\frac{1}{\hat{\sigma}^2} \sum (x_i - \mu_i)^2 - n\right) = 0$$

$$\boxed{\frac{1}{n} \sum (x_i - \mu_i)^2 = \hat{\sigma}^2}$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^{2})$$

Back to regression: $y_i = \theta_0^* + \theta_1^* x_i + \varepsilon_i$

We observe $\{(x_i, y_i)\}_{i=1}^n$ $\qquad \mathbb{E}[y_i] = \theta_0^* + \theta_1^* x$

We want to estimate $\theta_0^*, \theta_1^*, \sigma^2$

1) $$L(\theta_0, \theta_1, \sigma^2; S) = \prod_{i=1}^{n} p((x_i, y_i), \theta_0, \theta_1, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y_i - \overbrace{[\theta_0 + \theta_1 x]}{})^2}{2\sigma^2} \right)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \right)$$

2) log likelihood $\mathcal{L}(\theta_0, \theta_1, \sigma^2) = \lg L(\dots)$

$$= -\frac{n}{2} \lg(2\pi) - \frac{n}{2} \lg \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

3) $\underset{\theta_0\, \theta_1\, \sigma^2}{\arg\max}\ \mathcal{L}(\theta_0, \theta_1\ \sigma^2_\varepsilon, S)$

Find the partial derivatives w.r.t $\theta_0, \theta_1, \sigma^2$

$$\frac{\partial \mathcal{L}}{\partial \theta_0} = \frac{1}{\sigma^2} \sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i) = 0$$

$$\boxed{\hat{\theta}_0^{MLE} = \bar{y} + \hat{\theta}_1 \bar{x}}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = \frac{1}{\sigma^2} \sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i) x_i = 0$$

$$\boxed{\hat{\theta}_1^{MLE} = \frac{cov(x, y)}{var(\cancel{x})}}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \frac{-n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \cdot \sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i)^2 = 0$$

$$\frac{+n}{2\hat{\sigma}^2} = \frac{1}{2(\hat{\sigma}^2)^2} \cdot \sum (y_i - \hat{\theta}_0 - \theta_1 x_i)^2$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i)^2$$

$$\hat{\sigma}^2_{OLS} = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \qquad X \in \mathbb{R}^{n \times (p+1)}$$

Simple regression (i.e. 1D) $X, y \in \mathbb{R}^n$

in this case $p = 1$ $\Rightarrow$

$$\hat{\sigma}^2_{OLS} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i)^2$$

$\hat{\sigma}^2_{MLE}$ is biased

but $\hat{\theta}_0$, $\hat{\theta}_1$ they are the same

$\varepsilon_i \sim \sigma^2$