Estimating β

- some versions of linear model:
  - observed data $(y_i, x_{i1}, \ldots, x_{im})$, $i=1, \ldots, n$. Regard y_i as a realization of a random variable Y_i.

functional form

$$Y_i = \sum_{j=0}^{p-1} \beta_j \cdot g_j(x_{i1}, \ldots, x_{im}) + \epsilon_i, \quad E(\epsilon_i) = 0, \quad Var(\epsilon_i) = \sigma^2, \quad Cor(\epsilon_{i_1}, \epsilon_{i_2}) = 0, \text{ for } i_1 \neq i_2$$

mean, variance, r.v. | not r.v. | deterministic component | random component

matrix form

- Y: random, X_D: fixed; $Y = X_M \beta + \varepsilon$, $E(\varepsilon) = \mathbf{0}$, $Var(\varepsilon) = \sigma^2 I$ (data matrix; e.g. data from DOE; model matrix; variance-covariance matrix; $\varepsilon \equiv Y - X_M\beta$ (LM); $Cov(\varepsilon)$)
- different
- Y: random, X_D: fixed; Y is a random vector with $E(Y) = X_M \beta$, $Var(Y) = \sigma^2 I$ ← easier to be extended to GLM (e.g. binomial y_x, Poisson y_x, ...)
- Y: random, X_D: random; $E(Y|X_D) = X_M \beta$, $Var(Y|X_D) = \sigma^2 I$

sampling model (future lecture)

(X_D, Y): joint distribution → $Y|X_D = X_M\beta + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$

→ $Y|X_D$: conditional dist. ← Linear model is built on this

Notes:

1. At this stage, no specific distribution assumption imposed on ε, only assume they (1) have *zero mean*, (2) have *constant variance*, and (3) are *mutually uncorrelated*. (i.e., ε_i's can have any distributional forms (non-parametric approach))
2. parameters in the model: β and σ^2

- a geometric view – an example

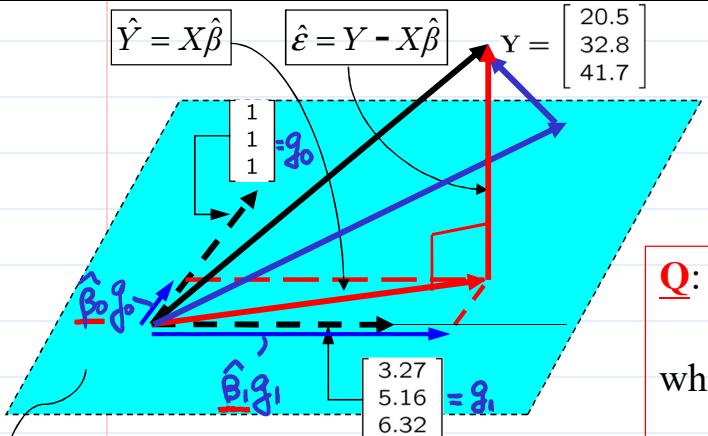
model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i=1, 2, 3$ ⇒ $Y = X\beta + \varepsilon$ (sum of columns in X multiplied by β)

data matrix (y, x) | Y | model matrix X

$$\begin{bmatrix} 20.5 \\ 32.8 \\ 41.7 \end{bmatrix} = \begin{bmatrix} 1 & 3.27 \\ 1 & 5.16 \\ 1 & 6.32 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \Rightarrow \begin{bmatrix} 20.5 \\ 32.8 \\ 41.7 \end{bmatrix} = \beta_0 \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{=g_0} + \beta_1 \underbrace{\begin{bmatrix} 3.27 \\ 5.16 \\ 6.32 \end{bmatrix}}_{=g_1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

$g_0, g_1 \in \mathbb{R}^3$

# of observations = 3, # of parameters in β = 2 ⇒ $Y \in \mathbb{R}^3$, $\beta \in \mathbb{R}^2$



span{X} = Ω: the space spanned by $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 3.27 \\ 5.16 \\ 6.32 \end{bmatrix}$,

i.e., $\left\{ \beta_0 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} 3.27 \\ 5.16 \\ 6.32 \end{bmatrix} \middle| \beta_0, \beta_1 \in \mathbb{R} \right\}$.

its dimension = 2 But, $\Omega \subset \mathbb{R}^3$

estimation of β (i.e., finding $\hat{\beta}$) is equivalent to finding a vector (i.e., $X\hat{\beta}$) on the 2-dim ***model space*** $\Omega = span\{X\}$

規律 | 隨機 | $\hat{Y}$ | $Y - \hat{Y} = Y - X\hat{\beta}$ as "small" as possible

Q: $Y = X\beta + \varepsilon = X\hat{\beta} + \hat{\epsilon}$,

what $\hat{\beta}$ would best "separate" $X\hat{\beta}$ from $\hat{\epsilon}$?

(Q: What's its meaning?) | if very close

⇒ find $\hat{\beta}$ s.t. $X\hat{\beta}$ is "close" to Y ← $\hat{Y} \approx Y$

- Q: is it reasonable? | as similar as possible
- Q: what's the measurement of closeness?

- a geometric view – a general description
  - n = # of observations; p = # of parameters in $\boldsymbol{\beta}$
  - $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^p$
  - assume $p \le n$
  - assume $\boldsymbol{X}$ is of full rank, rank($\boldsymbol{X}$)=p ($\Leftrightarrow$ $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular $\Leftrightarrow$ the p columns of $\boldsymbol{X}$ are linearly independent $\Leftrightarrow$ the model space Ω spanned by the columns of $\boldsymbol{X}$ is of dimension p)

model matrix $X_{n\times p}$ = [g_0 g_1 ⋯ g_{p-1}]; # of columns = p = dim(Ω)

Note: Data sets with $p > n$ do exist

$g_0, g_1, \cdots, g_{p-1} \in \mathbb{R}^n$

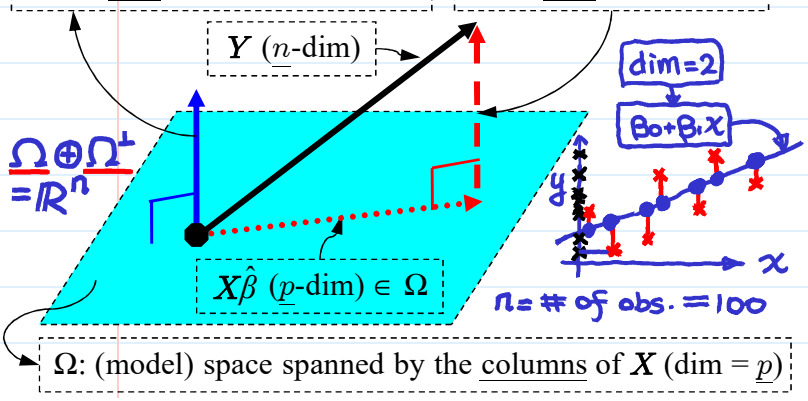
X^TX: a $p \times p$ matrix

$$\Omega \subset \mathbb{R}^n,\ \Omega^\perp \subset \mathbb{R}^n \qquad \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = \beta_0 \underbrace{\begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix}}_{=g_0} + \beta_1 \underbrace{\begin{bmatrix} g_{11} \\ g_{21} \\ \cdots \\ g_{n1} \end{bmatrix}}_{=g_1} + \cdots + \beta_{p-1} \underbrace{\begin{bmatrix} g_{1\,p-1} \\ g_{2\,p-1} \\ \cdots \\ g_{n\,p-1} \end{bmatrix}}_{g_{p-1}=} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{bmatrix}$$

$\Omega^\perp$: residual space (orthogonal complement of Ω, dim = $n-p$)

$\hat{\boldsymbol{\varepsilon}}$ ($(n-p)$-dim) $\in \Omega^\perp$

$\mu_i = E(Y_i)$, $i = 1, \cdots, n$; $(\mu_1, \cdots, \mu_n) \in \mathbb{R}^n$



Ω: (model) space spanned by the columns of $\boldsymbol{X}$ (dim = p)

$Y \in \mathbb{R}^n$, $X\beta \in \mathbb{R}^p$, $\varepsilon \in \mathbb{R}^n$

data	=	systematic structure	+	random error
n-dim	=	p-dim	+	$(n-p)$-dim
Y		$X\hat{\beta}$		$\hat{\varepsilon}$

Use a simpler structure (p-dim) to describe the complex (n-dim) Y

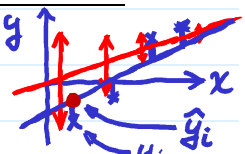
$\Rightarrow$ if successful (i.e, $Y \approx X\hat{\beta}$), main structure of Y is p-dim

規律

- find $\hat{\beta}$ such that $X\hat{\beta}$ is as "close" as possible to Y, where closeness is measured by Euclidean distance (**Q**: when, i.e. under what assumption, is this a reasonable measurement of closeness?)

least square criterion: $\sum_{i=1}^n [y_i - (X\hat{\beta})_i]^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 \cdot 1 = \|\hat{\varepsilon}\|^2$ (y_i, $\hat{y}_i$)

$\mathrm{Var}(\varepsilon) = \sigma^2 I$ – constant variance – uncorrelated



- $X\hat{\beta}$ is the orthogonal projection of Y onto Ω, (the space spanned by the columns of X), i.e.,

surrogate of Xβ (規律)

$$X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{n\times n}\underbrace{Y}_{n\times p} = \underbrace{H}_{n\times n}\underbrace{Y}_{n\times p}$$

an $n \times n$ matrix

IF $\Omega_1 = \mathrm{span}\{X_1\} = \mathrm{span}\{X_2\} = \Omega_2$, $X_1(X_1^TX_1)^{-1}X_1^T = H_1 = H_2 = X_2(X_2^TX_2)^{-1}X_2^T$

where $X(X^TX)^{-1}X^T$ is the *orthogonal projection matrix* of Ω (called *hat matrix* H).

- $\hat{\varepsilon}$: *residuals* = difference between Y and $X\hat{\beta}$ (i.e, $Y - X\hat{\beta}$)

surrogate of ε (error); OPM of $\Omega^\perp$: $(I-H)Y =$

$\mathrm{cov}(X\hat{\beta}, \hat{\varepsilon}) = 0_{n\times n}$

- $X\hat{\beta} \perp \hat{\varepsilon}$, $\perp$: geometrically orthogonal $\Rightarrow (X\hat{\beta})^T\hat{\varepsilon} = 0$ (r.v.'s)

different properties

(**Note**: actually, the two random vectors are *uncorrelated* $\Rightarrow$ when normal distribution assumption is imposed on $\boldsymbol{\varepsilon}$, they become *independent*)

- $n-p$: degree of freedom of $\hat{\varepsilon}$ (**Q**: what is a geometric interpretation for the degree of freedom?) (**Q**: what is the constraint on $\hat{\varepsilon}$?)

$X^T\hat{\varepsilon} = 0_{p\times 1}$

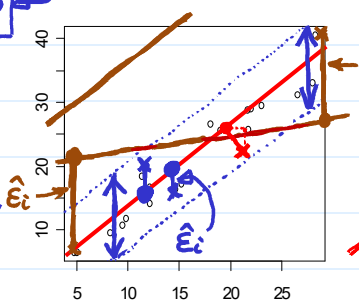
$\dim(\varepsilon) = n$ cf. $\hat{\varepsilon} \in \Omega^\perp \subset \mathbb{R}^n$, $\dim(\Omega^\perp) = n-p$

If $\mathbb{1} \in X$ or $\mathbb{1} \in \Omega \Rightarrow \mathbb{1}^T\hat{\varepsilon} = 0$ ($\mathbb{1}$: vector of 1; $\sum_{i=1}^n \hat{\varepsilon}_i$)

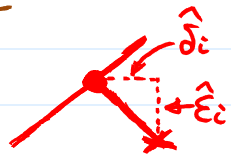
- $\hat{\beta}$ is the ordinary least square estimator

- (ordinary) least square estimator — distribution form not specified
  - assume $\boldsymbol{\varepsilon}$ are (i) uncorrelated (ii) equal variance ($Var(\boldsymbol{\varepsilon}) = \sigma^2 I$)
  - define the best $\hat{\boldsymbol{\beta}}$ as that minimizes *sum of squared error*: $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$ — length² of $\hat{\boldsymbol{\varepsilon}}$; $\hat{\epsilon}_i = y_i - (X\hat{\beta})_i$
    (Q: why? — reasonable criterion?)
    - easy to calculate
    - variation due to random error only appears on y-axis
    - variation minimization ⇒ $\hat{Y}$ close to Y
    - same scale in y_i's

Why not minimize $\sum_{i=1}^{n}|\epsilon_i|$? or minimize $\max_{1\le i\le n}|\epsilon_i|$?



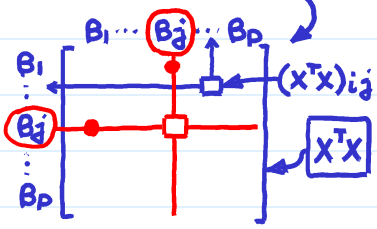
$E(\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}) = (n-p)\sigma^2$ under OLS



$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (Y - X\beta)^T(Y - X\beta) = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta \quad (*)$$

($Y^T - \beta^TX^T =$; β^TX^TY: $1\times p$, $p\times 1$)

⇒ a second-order polynomial of $\boldsymbol{\beta}$

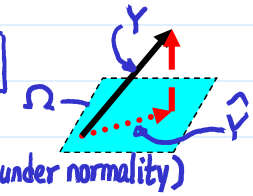


- One method of finding the minimizer is to differentiate (*) w.r.t. $\boldsymbol{\beta}$ and set the derivatives equal to zero

$$\Rightarrow \frac{\partial}{\partial\beta}\epsilon^T\epsilon = -2X^TY + 2X^TX\beta = \mathbf{0}$$

- By calculus, $\hat{\boldsymbol{\beta}}$ is the solution of $X^TX\beta = X^TY$ ⇐ called *normal equation* — sufficient statistics of β (under normality)

(X^T: $p\times n$ matrix ($p<n$); $X^T\hat{Y} = X^TY \Leftarrow \hat{Y}$)



- assume X^TX is non-singular (Q: when would it be singular?) — check LNp.3-3 condition for X to be of full rank

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^TY \Rightarrow X\hat{\beta} = X(X^TX)^{-1}X^TY \equiv HY$$

(a linear function of y_i's; unique solution)

- $H_{n\times n} = X(X^TX)^{-1}X^T$ is called *hat matrix* in statistics. It is the orthogonal projection matrix onto Ω, the space spanned by the columns of X
  - $H^2 = H$ ⇒ *idempotent*
  - $H^T = H$ ⇒ *symmetric*

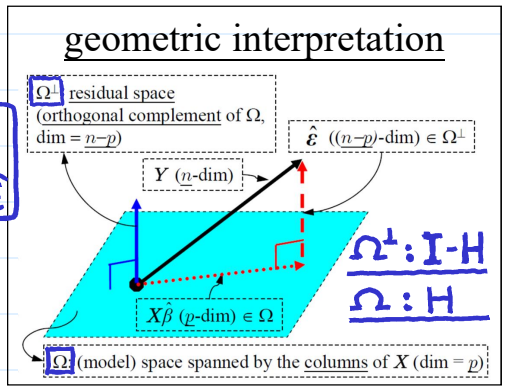
  (a matrix with the 2 properties must be an orthogonal projection matrix)

$H^2 = [X(X^TX)^{-1}X^T][X(X^TX)^{-1}X^T] = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$

(exercise)
  - $(I-H)^2 = (I-H)$ and $(I-H)^T = (I-H)$ — $(I-H)Y = Y - HY = Y - X\hat{\beta} = \hat{\varepsilon}$
  - $H(I-H) = (I-H)H = \mathbf{0}$ — their columns orthogonal; $n\times n$ matrix

(n)
  - the eigenvalues of H (or $I-H$) are either 0 or 1; # of 1's = p (or $n-p$); # of 0's = $n-p$ (or p)
  - $Hx = x$ if and only if x lies in Ω
  - $Hx = \mathbf{0}$ if and only if x lies in $\Omega^\perp$, the orthogonal complement of Ω

geometric interpretation



- *predicted values*: $\hat{Y} = X\hat{\beta} = HY$ ↔ c.f. $X\beta$ ← unknown
- *residuals*: $\hat{\boldsymbol{\varepsilon}} = Y - X\hat{\beta} = Y - \hat{Y} = (I-H)Y$ ↔ c.f. ε ← unknown; $Y = X\beta + \varepsilon$; $\varepsilon^T\varepsilon$ (c.f.)
- *residual sum of squares* (*RSS*): $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = [Y^T(I-H)^T][(I-H)Y] = Y^T(I-H)Y = \varepsilon^T(I-H)\varepsilon$ (c.f.)

($= \sum_{i=1}^{n}\hat{\varepsilon}_i^2$ = length² of $\hat{\varepsilon}$)

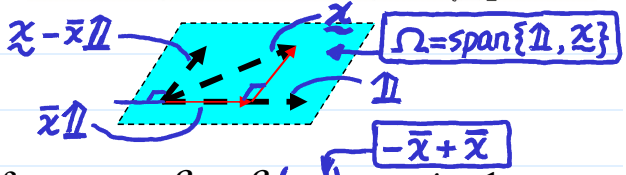
- examples of calculating ordinary least square estimator $\hat{\beta}$
  - example 1 (one-sample problem). functional form: $y_i = \mu + \varepsilon_i$, $i = 1, ..., n$.

$$Y = X\beta + \epsilon \quad \text{matrix form}$$

$$Y = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 \\ \cdots \\ 1 \end{bmatrix} \equiv 1, \ \beta = [\mu].$$

$$X^T X = 1^T 1 = n$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = (1/n)\ 1^T Y = (1/n)\sum_{i=1}^n y_i = \bar{y}.$$



  - example 2 (simple regression). functional form: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, ..., n$.

$$Y = X\beta + \epsilon \quad \text{matrix form} \quad \Rightarrow \ y_i = (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + \epsilon_i \equiv \beta_0' + \beta_1(x_i - \bar{x}) + \epsilon_i.$$

$$Y = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 - \bar{x} \\ \cdots & \cdots \\ 1 & x_n - \bar{x} \end{bmatrix} = [1 \ \ x \ \ x - \bar{x}1], \beta = \begin{bmatrix} \beta_0' \\ \beta_1 \end{bmatrix}.$$

$$X^T X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 - \bar{x} & \cdots & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} 1 & x_1 - \bar{x} \\ \cdots & \cdots \\ 1 & x_n - \bar{x} \end{bmatrix} = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix}.$$

$\hat{\beta}_0 = \hat{\beta}_0' - \hat{\beta}_1\bar{x}$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum(x_i - \bar{x})^2} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 - \bar{x} & \cdots & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum(x_i - \bar{x})^2} \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum[(x_i - \bar{x})y_i] \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{\sum[(x_i - \bar{x})y_i]}{\sum(x_i - \bar{x})^2} \end{bmatrix} \begin{matrix} \leftarrow \hat{\beta}_0' \\ \leftarrow \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \hat{\rho}\frac{\hat{\sigma}_y}{\hat{\sigma}_x} \end{matrix}$$

regression effect

$$y_i = \left(\bar{y} - \hat{\rho}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}\bar{x}\right) + \hat{\rho}\frac{\hat{\sigma}_y}{\hat{\sigma}_x} x_i + \epsilon_i$$

$$\left(\frac{y_i - \bar{y}}{\hat{\sigma}_y}\right) = \hat{\rho}\left(\frac{x_i - \bar{x}}{\hat{\sigma}_x}\right) + \left(\frac{\epsilon_i}{\hat{\sigma}_y}\right)$$

- mean and covariance matrix of OLS estimator $\hat{\beta}$

$\hat{\beta} = (X^T X)^{-1} X^T Y$ is a $p \times 1$ vector of random variables, so

  - mean: $E(\hat{\beta}) = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X\beta = \beta$ (i.e., *unbiased*)

  - $Cov(\hat{\beta}) = (X^T X)^{-1} X^T \sigma^2 I\, X (X^T X)^{-1} = (X^T X)^{-1}\sigma^2$ ($\Rightarrow$ irrelevant to Y and $\beta \Rightarrow$ **Note**: if we can control X, can decide the var-cov matrix *before* observing Y)

Since $\hat{\beta}$ is a random vectors, $(X^T X)^{-1}\sigma^2$ is a variance-covariance matrix.

  - $se(\hat{\beta}_i) = \sqrt{(X^T X)^{-1}_{ii}}\,\hat{\sigma}$

  - how to calculate the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$? (Q: Is this correlation influenced by σ?) (Q: What's the difference between this correlation and the correlation between predictors g_i and g_j? **Ans**: the former is calculated from $(X^T X)^{-1}$ while the latter from $X^T X$)

$$\frac{(X^T X)^{-1}_{ij}\sigma^2}{\sqrt{(X^T X)^{-1}_{ii}\sigma^2}\sqrt{(X^T X)^{-1}_{jj}\sigma^2}}$$

$$X = [\,|\,\cdots\,|\,\cdots\,|\,] = [\,|\,\cdots\,|\,g_i\,|\,\cdots\,|\,g_j\,|\,]$$

assume $\bar{g}_i = 0$

If $X^T X = \begin{bmatrix} \square & 0 \\ 0 & \square \end{bmatrix}$, $\Rightarrow (X^T X)^{-1} = \begin{bmatrix} \square & 0 \\ 0 & \square \end{bmatrix}$

check LNp.3-7, Ex2.

$g_i \perp g_j \overset{?}{\Rightarrow} cor(\hat{\beta}_i, \hat{\beta}_j) = 0$

In general cases, $(X^T X)_{ij} = 0 \not\Rightarrow (X^T X)^{-1}_{ij} = 0$

❖ **Reading:** Faraway (2005, 1st ed.), 2.3, 2.4, 2.5 LNp.3-7

❖ **Further reading**: D&S, 4.4, 5.1, 5.2, 20.1, 20.2