2.3 Projection and geometric view

[Proj 1] Definition of projection & least squares fit

[Proj 2] (1) Properties of the projection map $\Rightarrow$ induces a matrix

(1) Uniqueness (2) Linearity (3) Idempotent

(4) Map $\Rightarrow$ Matrix: specific form $P_{Col(X)} = X(X^TX)^{-1}X^T$

(5) Relationship with OLS (6) $I - P$ is also projection

(2) Properties of the projection matrix in OLS

(Use the properties to prove OLS conclusions)

3. Statistical properties for OLS

[Stat 1] From only moment structure $E(Y \mid X) = X^T\beta$

$\Rightarrow$ Additive model $Y = X^T\beta + \epsilon$ (Assumptions on ϵ)

[Stat 2] Properties: Mean and variance of $\hat{\beta}$ ✓

and residuals ✓ (RSS) ✓

✓ General formula in MLR

✓ Special forms in SLR and interpretations

[Stat 3] Optimality $\Rightarrow$ Gauss-Markov Theorem

Optimality of OLS $\Rightarrow$ Gauss-Markov Theorem

Why OLS estimates $\hat{\beta}$ not other estimates?

Nice properties: Unbiasedness $E(\hat{\beta} \mid X) = \beta$

$E(\frac{RSS}{n-p} \mid X) = \sigma^2$

$\triangle$ Goal: For a reasonable class of estimates of β, OLS $\hat{\beta}_j$ is an unbiased estimate of β_j with the smallest variance.

$\triangle$ $\underset{}{\beta_j} = \underset{1\times p}{e_j^T} \underset{p\times 1}{\beta}$ $\quad e_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \rightarrow$ jth position indicator

$e_j^T \beta$ unbiased small variance

We can generalize this to $c^T \underset{p\times 1}{\beta}$ for any given $c \in \mathbb{R}^p$.

$\triangle$ Gauss-Markov Theorem

When columns of $\mathbf{X}$ are linearly independent ($\hat{\beta} = (X^TX)^{-1}X^TY$)

among the class of linear unbiased estimates of $\underset{1\times 1}{c^T\beta}$,

$c^T \hat{\beta}$ is the unique estimate with the minimum variance.

$\triangle$ We say $c^T\hat{\beta}$ is the best linear unbiased estimate of $c^T\beta$.

(BLUE)

Linear unbiased estimate: any estimate in the form

$m^T Y$ ($Y \in \mathbb{R}^n$, $m \in \mathbb{R}^n$) $E(m^T Y \mid \mathbf{X}) = \beta$

Proof: Note that $\underset{p\times n}{\hat{\beta}} = \underset{n\times 1}{B Y}$ $\quad B = (X^T X)^{-1} X^T$

For any linear unbiased estimator $\underbrace{MY}_{p\times 1} \Rightarrow E(MY|X) = \underset{p\times 1}{\beta}$

Goal: $\underline{\mathrm{var}(c^T M Y \mid X)} \geq \mathrm{var}(c^T \hat{\beta} \mid X) = \underline{\mathrm{var}(c^T B Y} \mid X)$

$$\mathrm{var}(\underbrace{c^T M Y} \mid X) = \mathrm{var}\{c^T(\underbrace{M - B}_{\text{term 1}} + \underbrace{B}_{\text{term 2}})Y \mid X\}$$

$$= (1) + (2) + (3) = (1) + (2) \geq (2) = \mathrm{var}(c^T B Y \mid X)$$

(The equality holds $\Leftrightarrow$ $(1) = 0$ $\Leftrightarrow$ $c^T(M-B)Y = 0$ $\Leftrightarrow$ $c^T M Y = c^T B Y = c^T \hat{\beta}$)

$$(1) = \mathrm{var}\{c^T(M-B)Y \mid X\} \quad \underbrace{\geq 0}$$

$$(2) = \mathrm{var}\{c^T B Y \mid X\} \quad \underbrace{\geq 0}$$

$$(3) = 2\,\mathrm{cov}\{c^T(M-B)Y,\ c^T B Y \mid X\}$$ (Will prove $(3) = 0$)

$$= 2c^T(M-B)\,\mathrm{cov}(Y, Y \mid X) \times B^T c$$ (By $\mathrm{cov}(AY, BY) = A\,\mathrm{cov}(Y) B^T$)

$$= 2c^T(M-B) \times \sigma^2 I_n \times B^T c$$ (By (3.1))

$$= 2\sigma^2 c^T (M-B) B^T c$$

$$= 2\sigma^2 c^T (M-B) X (X^T X)^{-1} c$$ (By $B = (X^T X)^{-1} X^T$)

$$= 2\sigma^2 c^T (\underbrace{MX - BX}_{=0})(X^T X)^{-1} c$$ (By (3.2))

$$= 0$$

(3.1) $Y = X\beta + \underset{n\times 1}{\epsilon}$ conditioning on X, $\underset{n\times 1}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$

$$\mathrm{cov}(Y, Y \mid X) = \mathrm{cov}(\epsilon, \epsilon \mid X) = \sigma^2 \times I_n$$ (Moment assumptions on ϵ, Notes Sep 26)

(13.7) By unbiasedness $E(BY \mid X) = B\,E(Y \mid X) = BX\beta = \beta$

$= E(MY \mid X) = M\,E(Y \mid X) = MX\beta$

$\Rightarrow$ For <u>any</u> value of β, $\quad BX\beta = MX\beta$

$(BX - MX)\beta = 0$

$\Rightarrow$ $BX - MX = 0$

Δ The above proof shows $\operatorname{Var}(c^T MY \mid X) - \operatorname{var}(c^T \hat{\beta} \mid X) \geqslant 0$

$\Rightarrow$ $c^T \left\{ \operatorname{cov}(MY \mid X) - \operatorname{cov}(\hat{\beta} \mid X) \right\} c \geqslant 0$ for any c

$\Rightarrow$ $\operatorname{cov}(MY \mid X) - \operatorname{cov}(\hat{\beta} \mid X)$ is positive semi-definite.

Optimality: $\underset{1\times 1}{\operatorname{Var}(C^T \hat{\beta})}$ minimum

$\underset{p\times p}{\operatorname{cov}(\hat{\beta})}$ "minimum"

Exercise. If columns of X are linearly dependent,

there is no matrix C such that CY is

an unbiased estimator of β.

Proof: Suppose $E(CY \mid X) = \beta$

$\Rightarrow$ $C\,E(Y \mid X) = CX\beta = \beta$ for all values of β

$\Rightarrow$ $CX = I$

$\Rightarrow$ Since X columns are linearly dependent, there exists

$a \neq 0 \in \mathbb{R}^p$ such that $Xa = 0$.

$$\Rightarrow \quad c X a = I \times a$$
$$\downarrow \qquad \downarrow$$
$$0 = a \quad (\text{Contradicts with } a \neq 0)$$

[Gauss - Markov 2] (X may not have full rank.)

Let $\underset{n\times 1}{\theta} = \underset{n\times p}{X}\underset{p\times 1}{\beta}$ and $\hat{\theta} = P_{Col(X)}(Y)$,

where recall $P_{Col(X)}(Y) = \underset{\eta \in Col(X)}{\arg\min} \|Y - \eta\|^2$.

(Notes Sep 21)

The projection of Y onto the column space of X.

$c^T\hat{\theta}$ is the BLUE for $c^T\theta$ give any $c \in \mathbb{R}^n$.

Proof:

Let X_1 be an $n\times r$ matrix $(r \leq p)$ with linearly independent columns and $col(\underset{n\times r}{X_1}) = col(\underset{n\times p}{X})$.

The $P_{Col(X)}(Y) = P_{Col(X_1)}(Y)$ and X_1 is of full rank.

By notes **[Proj 3]** in Sep 21, $\underset{n\times n}{\boxed{P_1}} = \underset{n\times r}{X_1}\underbrace{(X_1^TX_1)^{-1}}_{r\times r}\underbrace{X_1^T}_{r\times n}$

such that (P1) $P_{Col(X_1)}(Y) = P_1 \cdot Y \Rightarrow \hat{\theta} = P_1 Y$

(P2) $P_1^T = P_1$ and $P_1^TP_1 = P_1^2 = P_1$

(P3) $P_{Col(X_1)}(X) = X \Rightarrow \underset{n\times n}{P_1} \times \underset{n\times p}{X} = \underset{n\times p}{X}$

(P4) $\underset{n\times n}{P_1}\underset{n\times 1}{\theta} = P_1 X\beta = X\beta = \theta$ (by (P3))

$$\Rightarrow P_1\theta = \theta$$

Suppose $d^T Y$ is a linear unbiased estimator for $c^T\theta$.

$$E(d^T Y \mid X) = d^T E(Y \mid X) = d^T X\beta = d^T\theta$$

$$= c^T\theta \quad (\text{By unbiasedness})$$

$$\Rightarrow \quad d^T\theta - c^T\theta = 0$$

By (P4) above, $P_1\theta = \theta \Rightarrow (d-c)^T P_1 \theta = 0$ for any values of θ

$$\Rightarrow \underset{1\times n}{(d-c)^T} \underset{n\times n}{P_1} = \underset{1\times n}{0}$$

$$\Rightarrow \underline{d^T P_1 = c^T P_1} \Rightarrow P_1^T d = P_1^T c \Rightarrow P_1 d = P_1 c$$

Goal: $\operatorname{var}(d^T Y \mid X) \geq \operatorname{var}(c^T\hat\theta \mid X)$

$$\operatorname{var}(d^T Y \mid X) = \operatorname{var}\{(d^T Y - c^T\hat\theta) + c^T\hat\theta \mid X\}$$

$$= (1) + (2) + (3) \geq (2) \qquad (\text{Equality holds} \Leftrightarrow d^T Y = c^T\hat\theta)$$

$$(1) = \operatorname{var}\{d^T Y - c^T\hat\theta \mid X\} \geq 0$$

$$(2) = \operatorname{var}\{c^T\hat\theta \mid X\} \geq 0$$

$$(3) = 2\operatorname{cov}(d^T Y - c^T\hat\theta,\ c^T\hat\theta \mid X) \qquad (\text{By } \hat\theta = \underset{n\times n}{P_1}\underset{n\times 1}{Y})$$

$$= 2\operatorname{cov}(d^T Y - c^T P_1 Y,\ c^T P_1 Y \mid X)$$

$$= 2(d^T - c^T P_1)\operatorname{cov}(Y, Y \mid X) P_1^T c \qquad (\text{By } \operatorname{cov}(AY, BY) = A\operatorname{cov}(Y,Y)B^T)$$

$$= 2(d^T - c^T P_1)\sigma^2 I P_1 c \qquad (\text{By } \operatorname{cov}(Y,Y|X) = \sigma^2 I_n \text{ and (P2) } P_1^T = P_1)$$

$$= 2\sigma^2 (d^T - \underline{c^T P_1}) \underline{P_1 c}$$

$$= 2\sigma^2 (d^T - \underline{d^T P_1}) \underline{P_1 d}$$

$$= 2\sigma^2 (d^T P_1 d - d^T P_1^2 d) \qquad (\text{By (P2)}, P_1^2 = P_1)$$

$$= 2\sigma^2 (d^T P_1 d - d^T P_1 d) = 0$$

△ If columns of X are linearly dependent

$$\begin{cases} \theta = X\beta & \text{always has linear unbiased estimator} \\ \beta & \text{does NOT} \end{cases}$$

△ Estimable Function

The parametric function $\underset{1\times p}{a^T}\underset{p\times 1}{\beta}$ is said to be estimable if it has a linear unbiased estimator $b^T Y$ for β.

[See, e.g. Qualification B 21 . Q3(a)]

Conclusion: $a^T\beta$ is estimable if and only if $a \in C(X^T)$.

Proof: There exists $b^T Y$ such that

$$E(b^T Y \mid X) = b^T E(Y \mid X) = b^T X\beta = a^T\beta$$

for any values of β.

$$\Leftrightarrow \quad a^T = b^T X \quad \text{or} \quad a = \underline{X^T b}$$

(a is a linear combination of colums of X^T)

$\in Col(X^T)$

[Stat 4] Regression inference with Gaussian errors
Normal

[4.1] Multivariate normal distribution

(1.1) Standard multivariate normal distribution

Random vector $\mathbf{z} = (z_1 \cdots z_p)^T$ follows a p-dimensional standard normal if

(1) each $z_j \sim N(0,1)$ independently.

(2) the density of $\mathbf{z}$ is

$$p(\mathbf{z}) = p(z_1 \cdots z_p) = \prod_{j=1}^{p} p(z_j)$$
$$= \prod_{j=1}^{p} (2\pi)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} z_j^2\right)$$
$$= (2\pi)^{-\frac{p}{2}} \exp\left(-\sum_{j=1}^{p} \tfrac{1}{2} z_j^2\right)$$
$$= (2\pi)^{-\frac{p}{2}} \exp\left(-\tfrac{1}{2} \mathbf{z}^T \mathbf{z}\right)$$

We write $\mathbf{z} \sim N_p(\mathbf{0}_p, I_p)$ $\quad 0_p = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $\quad I_p = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ & \ddots & & \\ 0 & \cdots & & 1 \end{pmatrix}_{p \times p}$

$E(\mathbf{z}) = \mathbf{0}_p \qquad \mathrm{cov}(\mathbf{z}) = I_p$

(1.2) Multivariate normal distribution with mean μ and covariance Σ

(Consider positive definite Σ in the following)

Let $\boxed{\underset{p\times 1}{\mathbf{M}} = \underset{p\times p}{\Sigma^{\frac{1}{2}}} \underset{p\times 1}{\mathbf{z}} + \underset{p\times 1}{\mu}}$ $\qquad \mathbf{z} \sim N_p(0, I)$

We have $E(M) = \Sigma^{\frac{1}{2}} E(\mathbf{z}) + \mu = 0 + \mu = \mu$

$$\mathrm{cov}(M) = \mathrm{cov}(\Sigma^{\frac{1}{2}} \mathbf{z}) = \Sigma^{\frac{1}{2}} \mathrm{cov}(\mathbf{z}) (\Sigma^{\frac{1}{2}})^T = \Sigma^{\frac{1}{2}} I_p \Sigma^{\frac{1}{2}} = \underset{p\times p}{\Sigma}$$

$$\Rightarrow \quad \mathbf{z} = \Sigma^{-\frac{1}{2}} (\mathbf{M} - \boldsymbol{\mu}) \quad \circledast$$

By the change of variables in calculus

Jacobian matrix $\left(\frac{\partial z_i}{\partial M_j}\right)_{n \times n} = \Sigma^{-\frac{1}{2}}$

$$1 = \int_{\mathbb{R}^p} p(\mathbf{z})\, d\mathbf{z} = \int_{\mathbb{R}^p} (2\pi)^{-\frac{p}{2}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right) d\mathbf{z}$$

$$= \int_{\mathbb{R}^p} (2\pi)^{-\frac{p}{2}} \exp\left[-\frac{1}{2}\left\{\underline{\Sigma^{-\frac{1}{2}}(\mathbf{M}-\boldsymbol{\mu})}\right\}^T \left\{\Sigma^{-\frac{1}{2}}(\mathbf{M}-\boldsymbol{\mu})\right\}\right] \det\left[\left(\frac{\partial z_i}{\partial M_j}\right)_{n\times n}\right] d\mathbf{M}$$

$$= \int_{\mathbb{R}^n} (2\pi)^{-\frac{p}{2}} \exp\left[-\frac{1}{2}(\mathbf{M}-\boldsymbol{\mu})^T \underbrace{\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}}_{\Sigma^{-1}} (\mathbf{M}-\boldsymbol{\mu})\right] \det\left[\Sigma^{-\frac{1}{2}}\right] d\mathbf{M}$$

$$\Rightarrow \quad p(\mathbf{M}) = (2\pi)^{-\frac{p}{2}} \det\left[\Sigma^{-\frac{1}{2}}\right] \exp\left\{-\frac{1}{2}(\mathbf{M}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{M}-\boldsymbol{\mu})\right\}$$

$$= (2\pi)^{-\frac{p}{2}} \left\{\det(\Sigma)\right\}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{M}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{M}-\boldsymbol{\mu})\right\}$$

(By property of determinant)

is the density of random vector $\mathbf{M}$.

$\Rightarrow$ Random vector $\mathbf{M}$ with density $p(\mathbf{M})$ is called multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ

Moment generating function of $\mathbf{M}$

(1) By $\mathbf{M} = \Sigma^{\frac{1}{2}}\mathbf{z} + \boldsymbol{\mu}$ where $\mathbf{z} \sim N_p(0, I_p)$

the MGF of $\mathbf{M}$ is defined as

$$g_{\mathbf{M}}(t) = E\left[\exp(t^T \mathbf{M})\right] = E\left[\exp\left(t^T(\Sigma^{\frac{1}{2}}\mathbf{z} + \boldsymbol{\mu})\right)\right]$$

Exercise $\underset{1\times 1}{Z} \sim N(0,1)$ Derive $E[\exp(sZ)]$ $(s \in \mathbb{R})$