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 $P_{Col(X)} = X(X^T X)^{-1} X^T$

(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

(2) Residual

Mean $\quad E(\hat{R} \mid X) = E(Y - \hat{Y} \mid X) \qquad \underset{n\times 1}{\hat{Y}} = P \underset{n\times 1}{Y}$

$= (I-P)\,E(Y \mid X)$

$= (I-P)\,X\beta = 0 \qquad (I-P)X = 0$

Covariance $\quad cov(\underset{n\times 1}{\hat{R}} \mid X) = cov(Y - \hat{Y} \mid X)$

$= E[(Y-\hat{Y})(Y-\hat{Y})^T \mid X]$

$= E\{(I-P)\,Y\,\{(I-P)Y\}^T \mid X\}$

$= (I-P)\,E(YY^T \mid X)\,(I-P)^T \quad ①$

By $Y = X\beta + \epsilon$, $E(YY^T \mid X) = E\{(X\beta+\epsilon)(X\beta+\epsilon)^T \mid X\}$

$= E\{X\beta\beta^T X^T + \epsilon(X\beta)^T + X\beta\epsilon + \epsilon\epsilon^T \mid X\}$

$= X\beta\beta^T X^T + \sigma^2 I$

$① = (I-P)\{X\beta\beta^T X^T + E(\epsilon \mid X)(X\beta)^T + X\beta E(\epsilon \mid X) + E(\epsilon\epsilon^T \mid X)\}(I-P)^T$

$= (I-P)\,\sigma^2 I\,(I-P)^T$

$= \sigma^2 (I-P)$

By $\quad (I-P)X = 0$

$E(\epsilon \mid X) = 0$

$E(\epsilon\epsilon^T \mid X) = \sigma^2 I$

$cov(\hat{R} \mid X) = \sigma^2 (I-P)$

Residuals sum of squares (RSS)

$$RSS = \| Y - \hat{Y} \|^2 \qquad \hat{R} = Y - \hat{Y}$$

$$= \| (I-P) Y \|^2 \qquad Y - \hat{Y} = (I-P) Y$$

$$= \{(I-P)Y\}^T (I-P) Y$$

$$= Y^T (I-P)^T (I-P) Y \qquad (I-P)^T(I-P) = I-P$$

$$= Y^T (I-P) Y$$

This is a quadratic form in Y

mean RSS $\quad E(RSS \mid X)$

$$= E\{ Y^T (I-P) Y \mid X \} \qquad (E1)$$

$$= E[tr\{ (I-P) YY^T \} \mid X] \qquad (E2)$$

$$= \sigma^2 tr(I-P) \qquad (E3)$$

$$= \sigma^2 (n-p) \qquad (\text{By property on } P)$$

(Sep. 21st Notes)

$$\Rightarrow E\left(\frac{RSS}{n-p}\right) = \sigma^2$$

$\Rightarrow \frac{RSS}{n-p}$ is an unbiased estimator of σ^2

(E1) $\Rightarrow$ (E2)

$$\underset{1\times n}{\overset{\text{random}}{Y^T}} \ \underset{n\times n}{\overset{\text{fixed}}{(I-P)}} \ \underset{n\times 1}{\overset{\text{random}}{Y}} \qquad \text{scalar}$$

$$= tr\{ Y^T (I-P) Y \}$$

$$= tr\{ \underbrace{(I-P)}_{\text{fixed}} \underbrace{YY^T}_{\text{random}} \}$$

(E2) $\Rightarrow$ (E3)

$$(E2) = tr\{(I-P)\, \underline{E(YY^T|X)}\}$$

$$= tr\{(I-P)(\underline{X\beta\beta^T X^T} + \sigma^2 I)\} \quad \text{By } (I-P)X = 0$$

$$= tr\{(I-P)\sigma^2 I\}$$

Remark : ① $X \in \mathbb{R}^{n\times p}$ $\qquad \sigma^2 = E\left(\frac{RSS}{n-p}\right)$

② p covariates and 1 intercept $\quad X \in \mathbb{R}^{n\times(p+1)} \qquad \sigma^2 = E\left(\frac{RSS}{n-(p+1)}\right)$

Discussions on the SLR

$p=2 \quad$ 1 intercept + 1 covariate

$$Y = \alpha + \beta X + \epsilon \qquad (\alpha, \beta \in \mathbb{R})$$

(1) OLS estimates: unbiased $\qquad E(\hat{\alpha}|X) = \alpha$

$$E(\hat{\beta}|X) = \beta$$

Follows from MLR conclusion.

Exercise : ✓ $\begin{cases} \hat{\beta} = \dfrac{\widehat{cov}(X,Y)}{\widehat{var}(X)} \\ \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \end{cases} \Rightarrow \begin{cases} E(\hat{\alpha}|X) = \alpha \\ E(\hat{\beta}|X) = \beta \end{cases}$

OLS estimate covariance $\mathrm{cov}\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = \underbrace{\sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}}_{2\times 2}$

$$\underset{n\times 2}{\mathbf{X}} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Exercise 1. Plug-in $\mathbf{X}$ into $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

2. $\hat{\beta}, \hat{\alpha}$ formula (after lecture)

Hint: $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \dfrac{1}{ad-bc}\begin{pmatrix} d & -b \\ -b & a \end{pmatrix}$

$$\mathrm{var}(\hat{\beta}\mid x) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\mathrm{var}(\hat{\alpha}\mid x) = \frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sigma^2 \times \frac{\bar{x}^2 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \checkmark$$

$$= \sigma^2\left\{\underset{\checkmark}{\frac{1}{n}} + \underbrace{\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}_{\geq 0}\right\}$$

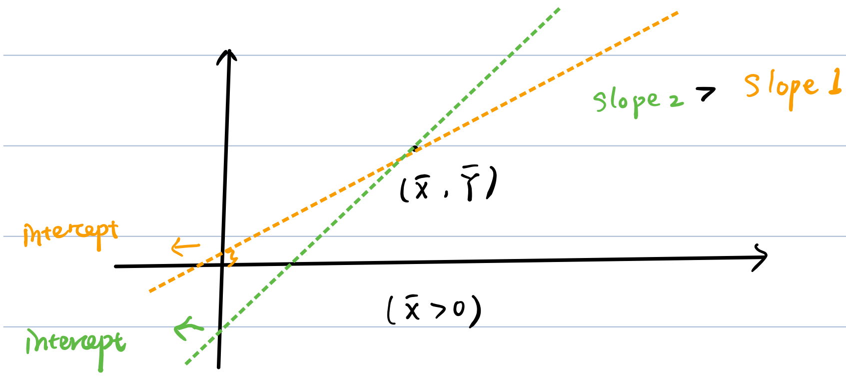
equality holds if $\bar{x}=0$

$\mathrm{var}(\hat{\alpha}\mid x)$ is minimized if $\bar{x}=0$.

$$\mathrm{cov}(\hat{\alpha},\hat{\beta}\mid x) = -\sigma^2\frac{\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

If $\bar{x}=0$, $\hat{\alpha}$ and $\hat{\beta}$ are uncorrelated.

If $\bar{x}>0$, covariance is negative.



OLS fit always pass $(\bar{X}, \bar{Y})$ data center.

(2) Residuals

$$\hat{R}_i = Y_i - \hat{Y}_i$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$= Y_i - \hat{\alpha} - \hat{\beta} X_i$$

$$= Y_i - \bar{Y} - \hat{\beta}(X_i - \bar{X})$$

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

$$\bar{Y} = \alpha + \beta \bar{X} + \bar{\epsilon}$$

$$\bar{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i$$

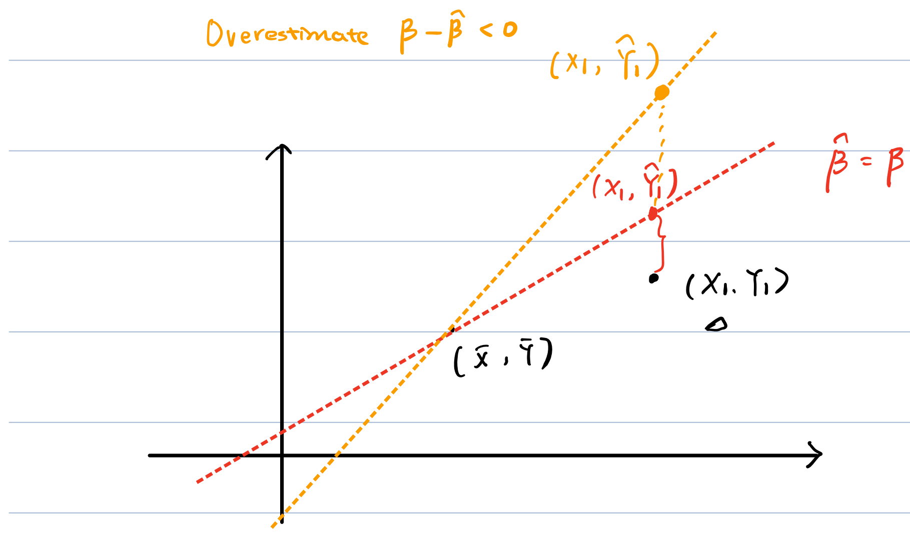
$$= (\cancel{\alpha} + \beta X_i + \epsilon_i) - (\cancel{\alpha} + \beta \bar{X} + \bar{\epsilon}) - \hat{\beta}(X_i - \bar{X})$$

$$= (\beta - \hat{\beta})(X_i - \bar{X}) + \epsilon_i - \bar{\epsilon} \quad ☆$$

Interpretation

The residuals $\hat{R}_i$ are not only influenced by errors ϵ but also $(\beta - \hat{\beta})(X_i - \bar{X})$, i.e. how well we recover the true slope β.

(If $\hat{\beta} = \beta$, view $\epsilon_i - \bar{\epsilon}$ as baseline errors.)



$$\left.\begin{aligned} \hat{R}_{1,\hat{\beta}=\beta} &= Y_1 - \hat{Y}_{1,\hat{\beta}=\beta} \\ \hat{R}_{2,\hat{\beta}>\beta} &= Y_1 - \hat{Y}_{1,\hat{\beta}>\beta} \end{aligned}\right\} \Rightarrow \hat{R}_{2,\hat{\beta}>\beta} \text{ is more negative than } \hat{R}_{1,\hat{\beta}=\beta}$$

1. Overestimate β: $\beta - \hat{\beta} < 0$ as $\hat{\beta} \to +\infty$

For i with $X_i - \bar{X} > 0$ (right of mean) $\hat{R}_i \downarrow -\infty$

$X_i - \bar{X} < 0$ (left of mean) $\hat{R}_i \uparrow +\infty$

2. Underestimate β $\beta - \hat{\beta} > 0$

For i with $X_i - \bar{X} > 0$, $\hat{R}_i \uparrow +\infty$

$X_i - \bar{X} < 0$, $\hat{R}_i \downarrow -\infty$

Covariance of residuals: $\underset{n\times 1}{\mathrm{cov}(\hat{R} \mid X)} = \sigma^2 (I - P)$

$P = X(X^T X)^{-1} X^T$

① Plug in $\underset{n\times 2}{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$ ✓

② Direct calculate through $\hat{R}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$

Exercise: $\operatorname{var}(\hat{R}_i \mid X) = \sigma^2 \times (I - P)_{(i,i)} = \sigma^2 \times (1 - P_{ii})$

$$P = X(X^T X)^{-1} X^T = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \underbrace{\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{n \times 2} \underbrace{\begin{pmatrix} \frac{\sum_{i=1}^{n} x_i^2}{n} - x_1 \bar{x} & \frac{\sum_{i=1}^{n} x_i^2}{n} - x_2 \bar{x} & \cdots & \frac{\sum_{i=1}^{n} x_i^2}{n} - x_n \bar{x} \\ -\bar{x} + x_1 & \cdots & \cdots & -\bar{x} + x \end{pmatrix}}_{2 \times n}$$

i-th diagonal $\quad P_{ii} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - 2x_i \bar{x} + x_i^2 \right)$

$$\operatorname{var}(\hat{R}_i \mid X) = \sigma^2 \times (1 - P_{ii})$$

$$= \sigma^2 \times \left(1 - \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \left(\frac{1}{n} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) + (x_j - \bar{x})^2 \right\} \right)$$

$$= \sigma^2 \times \left(1 - \frac{1}{n} - \underbrace{\frac{(x_j - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}_{\geq 0} \right)$$

$$\leq \sigma^2 \times \left(1 - \frac{1}{n}\right) < \sigma^2 = \operatorname{var}(\epsilon_i \mid X) \qquad n \geq 2$$

$\hat{R}_i$ has less variability than ϵ_i

Optimality of OLS $\Rightarrow$ Gauss–Markov Theorem

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

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

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

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

△ $\underset{}{\beta_j} = \underset{1\times p}{e_j^T} \underset{p\times 1}{\beta}$ $\quad e_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \to$ 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 $c \in \mathbb{R}^p$.

△ Gauss-Markov Theorem

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

among the class of linear unbiased estimates of $c^T\beta$,

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

△ 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

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