Applied Regression Analysis

AUTHOR
Matthew Joel

For Problems 1-5, suppose that

$$Y_i = \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

Problem 1 (Derivation of slope-only OLS estimator)

We would like to find the OLS estimator of $\hat{\beta}_1$. In this context, our objective function is

$$RSS(\hat{\beta}_1) = \sum (Y_i - \hat{\beta}_1 x_i)^2.$$

Determine the formula for $\hat{\beta}_1$ that minimizes $RSS = \sum (Y_i - \hat{\beta}_1 x_i)^2$.

Solution

1. $Y_i = \beta_1 x_i + \epsilon_i$

$\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$

$$RSS(\hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 X_i)^2$$

$$\frac{d}{d\hat{\beta}_1}\left(RSS(\hat{\beta}_1)\right) = \frac{d}{d\hat{\beta}_1}\left(\sum_{i=1}^{n} Y_i - \hat{\beta}_1 X_i\right)^2$$

$$= \frac{d}{d\hat{\beta}_1} \sum_{i=1}^{n} \left(Y_i^2 - 2Y_i\hat{\beta}_1 X_i + \hat{\beta}_1^2 X_i^2\right)$$

$$0 = \sum_{i=1}^{n} \left(-2Y_i X_i + 2\hat{\beta}_1 X_i^2\right)$$

$$\sum_{i=1}^{n} Y_i X_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

Problem 2 (Verifying the minimum)

Verify that the expression from the previous problem minimizes the RSS.

Solution

2. $$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_1 X_i)$$

$$\frac{\partial^2 RSS}{\partial \hat{\beta}_1^2} = -2\sum_{i=1}^{n} X_i\left(\frac{\partial}{\partial \hat{\beta}_1}(Y_i - \hat{\beta}_1 X_i)\right)$$

$$= -2\sum_{i=1}^{n} X_i(-X_i)$$

$$= 2\sum_{i=1}^{n} X_i^2$$

Since $X_i^2 \geq 0$ for all i except $X_i = 0$,

$$\sum_{i=1}^{n} X_i^2 > 0$$

$$\therefore \frac{\partial^2 RSS}{\partial \hat{\beta}_1^2} = 2\sum_{i=1}^{n} X_i^2 > 0$$

positive second derivative confirms RSS is min @ $\hat{\beta}_1$

Problem 3 (A useful lemma)

Determine $E(Y_i \mid X = X_i)$ for arbitrary $i \in \{1, 2, \ldots, n\}$.

Solution

3. $E[Y_i \mid X = x_i]$

$$Y_i = \beta_1 X_i + \epsilon_i$$

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

$$E[Y_i \mid X = X_i] = E[\beta_1 X_i + \epsilon_i \mid X = X_i]$$

$$= E[\beta_1 X_i \mid X = X_i] + E[\epsilon_i \mid X = X_i]$$

$$= \beta_1 X_i + 0 = \beta_1 X_i$$

Problem 4 (Unbiasedness)

Determine $E(\hat{\beta}_1 \mid \mathbf{X})$.

Solution

4.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(\beta_1 X_i + \epsilon_i)}{\sum_{i=1}^{n} X_i^2}$$

$$= \frac{\sum_{i=1}^{n} X_i \beta_1 X_i + \sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}$$

$$= \beta_1 \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2} + \frac{\sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}$$

$$E[\hat{\beta}_1|X] = E\left[\beta_1 + \frac{\sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2} \middle| X\right]$$

$$= \beta_1 + E\left[\frac{\sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2} \middle| X\right]$$

$$= \beta_1 + \frac{\sum_{i=1}^{n} X_i E[\epsilon_i]}{\sum_{i=1}^{n} X_i^2}$$

when $E[\epsilon_i] = 0$:

$$E[\hat{\beta}_1|X] = \beta_1$$

$\therefore$ unbiased

Problem 5 (Variability)

Determine $\mathrm{var}(\hat{\beta}_1 \mid \mathbf{X})$.

Solution

5. $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$

$= \frac{\sum_{i=1}^{n} X_i(\beta_1 X_i + \epsilon_i)}{\sum_{i=1}^{n} X_i^2}$

$= \beta_1 + \frac{\sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}$

$Var[\hat{\beta}_1 | X] = Var\left[\frac{\sum X_i \epsilon_i}{\sum X_i^2}\right]$

$= \sum_{i=1}^{n} X_i^2 \, Var(\epsilon_i)$

$= \sigma^2 \sum_{i=1}^{n} X_i^2$

$\therefore Var[\hat{\beta}_1 | X] = \frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{\left(\sum_{i=1}^{n} X_i^2\right)^2}$

$= \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$

Some matrix-related results

For Problems 6-9, using the usual notation, assume that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and

$$\boldsymbol{\epsilon} \mid \mathbf{X} \sim \mathsf{N}(\mathbf{0}_{n\times 1}, \sigma^2 \mathbf{I}_{n\times n}).$$

Problem 6

Determine the mean of $\mathbf{H}\mathbf{y}$.

Solution

6. $y = X\beta + \epsilon$

$\epsilon | X \sim N(0_{n \times 1}, \sigma^2 I_{n \times n})$

$E[y|X] = X\beta$

$E[Hy|X] = E[H(X\beta + \epsilon) | X]$

$= HE[X\beta + \epsilon | X]$

since $E[\epsilon|X] = 0$:

$E[Hy|X] = E[Hy] = HX\beta$

Problem 7

Determine the variance of $\mathbf{Hy}$.

Solution

7. $Hy = H(X\beta + \epsilon)$

$= HX\beta + H\epsilon$

$Var[Hy] = Var[HX\beta + H\epsilon]$

$= Var[H\epsilon]$

We know $Var[\epsilon] = \sigma^2 I_{n \times n}$

$\therefore Var[H\epsilon] = H Var[\epsilon] H^T$

$= H(\sigma^2 I) H^T$

$= \sigma^2 HH^T$

Problem 8

Determine the distribution of $\mathbf{Hy}$.

Solution

8. $Hy = H(X\beta + \epsilon)$

$= HX\beta + H\epsilon$

$E[\epsilon|X] = 0$

$\therefore\ E[Hy|X] = HX\beta$

$Var[Hy|X] = Var[H\epsilon|X]$

$= H Var[\epsilon|X] H^T$

$= \sigma^2 H I_{n\times n} H^T$

$= \sigma^2 HH^T$

$\therefore\ Hy|X \sim N(HX\beta, \sigma^2 HH^T)$

Problem 9 (Math 5387 Students only)

Determine the theoretical covariance between $\hat{\mathbf{y}}$ and $\hat{\boldsymbol{\epsilon}}$, i.e., $\operatorname{cov}(\hat{\mathbf{y}}, \hat{\boldsymbol{\epsilon}})$.

Solution

9. $\hat{y} = X(X^TX)^{-1}X^Ty$

$\hat{y} = Hy$

$\hat{\epsilon} = (I-H)y$

$cov[\hat{y}, \hat{\epsilon}] = cov[H\epsilon, (I-H)\epsilon]$

since $\epsilon \sim N(0, \sigma^2 I)$:

$cov[H\epsilon, (I-H)\epsilon] = H\,cov[\epsilon](I-H)^T$

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

since H is idempotent:

$H(I-H) = H - H^2$

$= H - H = 0 = cov[\hat{y}, \hat{\epsilon}]$