Multiple linear regression: Least Squares Estimation of $\boldsymbol{\beta}$

- Least squares estimation for $\boldsymbol{\beta}$
- Linear independence for the columns of $\boldsymbol{X}$ is required for $\boldsymbol{\beta}$ to be uniquely specified
- Residual variance S_e^2 can be used as an estimator for σ^2

Definition 3.1

A set of vectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p\}$ is said to be linearly independent if

$$\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \cdots + \alpha_p \boldsymbol{v}_p = 0 \Longrightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$$

Otherwise it is said to be linearly dependent.

If $v_1, v_2, \ldots, v_p$ are linearly dependent, this means we can express one of the v_i's as a linear combination of the remaining v_i's.

Linear independence and X

The columns of X in

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

must be linearly independent.

Why? So that β can be uniquely identified.

If the columns of X is linearly dependent, then there exist $\underset{(r\times 1)}{\alpha} \neq 0$ such that $X\alpha = 0$. Then we can write

$$Y = X\beta + \varepsilon = X(\beta + \alpha) + \varepsilon = X\beta + \underbrace{X\alpha}_{0} + \varepsilon$$

So β is not uniquely defined in this case.

Lemma 6

If $\boldsymbol{X}_{n\times p}$ is a matrix with linearly independent columns then the symmetric, $p \times p$ matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible.
(its inverse exists)

Note: For the columns of X to be linearly independent, we need $p \leq n$.

If a matrix A is invertible, then $|A| \neq 0$.

Recall that $|A| = \sum_{i=1}^{n} \lambda_i$. ($\lambda_i$ are eigenvalues of A)

$\Rightarrow$ 0 is not an eigenvalue of A.

This implies there exists no $\alpha \neq 0$ such that $A\alpha = 0$.

$\therefore$ If X^TX is invertible, then $X^TX\alpha = 0 \iff \alpha = 0$.

Proof of Lemma 6

Show that if the columns of X are linearly independent, then $(X^T X)\alpha = 0$ if and only if $\alpha = 0$.

$$(X^T X)\alpha = 0$$

$$\alpha^T (X^T X)\alpha = 0$$

$$(X\alpha)^T (X\alpha) = 0$$

$$\|X\alpha\|^2 = 0 \qquad x^T x = \|x\|^2$$

$$\Rightarrow \quad X\alpha = 0$$

$$\Rightarrow \quad \alpha = 0$$ since the columns of X are linearly independent.

Theorem 10

If the columns of $\boldsymbol{X}$ are linearly independent columns then the least squares estimates of $\boldsymbol{\beta}$ are given uniquely by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

$Q(\beta) = \|y - X\beta\|^2$

① Show that $\frac{\partial Q(\beta)}{\partial \beta} = 0$ when $\beta = (X^T X)^{-1} X^T y$, or

② show that $Q(\beta) \geq Q(\hat{\beta})$ and equality holds only when $\beta = \hat{\beta}$.

We will use ② to prove Theorem 10.

Proof of Theorem 10

$$Q(\beta) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_r x_{ir})]^2$$

$$= (y - X\beta)^T (y - X\beta)$$

$$= \| y - X\beta \|^2$$

$$= \| \overbrace{y - X\hat\beta} + \overbrace{X\hat\beta - X\beta} \|^2$$

$\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2a^Tb$

$$= \| y - X\hat\beta \|^2 + \| X\hat\beta - X\beta \|^2 - 2\boxed{(y - X\hat\beta)^T (X\hat\beta - X\beta)}$$

$$(y - X\hat\beta)^T (X\hat\beta - X\beta)$$

$$= (y - X\hat\beta)^T X (\hat\beta - \beta)$$

$$= [y - X(X^TX)^{-1}X^Ty]^T X (\hat\beta - \beta)$$

substitute $\hat\beta = (X^TX)^{-1}X^Ty$

$$= [(I - X(X^TX)^{-1}X^T)y]^T X (\hat\beta - \beta)$$

$$= y^T \underbrace{(I - \underbrace{\boxed{X(X^TX)^{-1}X^T}}_{H})^T X}_{0} (\hat\beta - \beta)$$

H is symmetric and idempotent (i.e. $H^T = H$ and $H^2 = H$)

$$= 0$$

(We will look at the properties of H and $I - H$ in Tutorial 4.)

Proof of Theorem 10

$$Q(\beta) = \|y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\beta\|^2$$

$$= Q(\hat{\beta}) + \underbrace{\|X\hat{\beta} - X\beta\|^2}_{\geqslant 0}$$

$$\geqslant Q(\hat{\beta})$$

Equality holds only if $\|X\hat{\beta} - X\beta\|^2 = 0$

$$X\hat{\beta} - X\beta = 0$$

$$\Rightarrow \quad \beta = \hat{\beta}$$

Hence, $Q(\beta)$ is uniquely minimised at $\beta = \hat{\beta}$.

Example 3.3

Check that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ agrees with the expressions given for the case of simple linear regression.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \qquad y = X\beta + \varepsilon$$

① $$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

If $X = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$,

then $X^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

② $$(X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

③ $$X^T y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

Example 3.3

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{array}{c} \begin{bmatrix} \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right) \\ -\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) + n\left(\sum_{i=1}^n x_i y_i\right) \end{bmatrix} \end{array} \begin{array}{c} \hat{\beta}_0 \\ \hat{\beta}_1 \end{array}$$

After further algebraic manipulations, we can show that the above will give the same expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ in Theorem 7. Please try this as an exercise.

Estimation of σ^2

The residual variance is

$$\boxed{S_e^2 = \frac{1}{n-p} \left\| \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right\|^2,}$$

where $\underline{p = r + 1}$, i.e., the number of β's.

$$S_e^2 = \frac{1}{n-p} \| Y - X(X^T X)^{-1} X^T Y \|^2$$

$$= \frac{1}{n-p} \| Y - HY \|^2$$

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

Example 3.4

Consider the data in Example 3.1 again:

x	-1	0	2	-2	5	6	8	11	12	-3
y	-5	-4	2	-7	6	9	13	21	20	-9

a) Use the matrix approach to fit a least-squares line to these data points.

b) Compute s_e^2 using the matrix approach.

a) $Y = \begin{bmatrix} -5 \\ -4 \\ \vdots \\ -9 \end{bmatrix}$, $X = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & -3 \end{bmatrix}$, $\hat{\beta} = (X^T X)^{-1} X^T Y$

Example 3.4 Solution

x	-1	0	2	-2	5	6	8	11	12	-3
y	-5	-4	2	-7	6	9	13	21	20	-9

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ -1 & 0 & \cdots & -3 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & -3 \end{bmatrix} = \begin{bmatrix} 10 & 38 \\ 38 & 408 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{10(408) - 38^2} \begin{bmatrix} 408 & -38 \\ -38 & 10 \end{bmatrix} = \frac{1}{2636} \begin{bmatrix} 408 & -38 \\ -38 & 10 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ -1 & 0 & \cdots & -3 \end{bmatrix} \begin{bmatrix} -5 \\ -4 \\ \vdots \\ -9 \end{bmatrix} = \begin{bmatrix} 46 \\ 709 \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} (X^T Y) = \frac{1}{2636} \begin{bmatrix} 408 & -38 \\ -38 & 10 \end{bmatrix} \begin{bmatrix} 46 \\ 709 \end{bmatrix} \approx \begin{bmatrix} -3.101 \\ 2.0266 \end{bmatrix}$$

Example 3.4 Solution

b) $$S_e^2 = \frac{1}{n-p} \| (I-H)Y \|^2$$

$$= \frac{1}{n-p} \left[(I-H)Y \right]^T (I-H)Y$$

$$= \frac{1}{n-p} Y^T \underbrace{(I-H)^T (I-H)}_{I-H} Y \quad \text{(as } I-H \text{ is symmetric and idempotent)}$$

$$= \frac{1}{n-p} Y^T (I-H) Y$$

$$= \frac{1}{10-2} \left(\frac{1}{2636} \right) (20598)$$

$$\approx 0.97$$

$$H = X(X^TX)^{-1}X^T = \frac{1}{2636} \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & -3 \end{bmatrix} \begin{bmatrix} 408 & -38 \\ -38 & 10 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ -1 & 0 & \cdots & -3 \end{bmatrix}$$

$$= \frac{1}{2636} \begin{bmatrix} 494 & 446 & \cdots & 590 \\ 446 & 408 & \cdots & 522 \\ \vdots & \vdots & & \vdots \\ 590 & 522 & \cdots & 726 \end{bmatrix}$$