MA50260 Statistical Modelling

Lecture 3: Linear Regression Estimation

Ilaria Bussoli

February 13, 2024

Content of today's lecture

Recall the linear regression model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, \ldots, n.$$

What are the assumption on $\epsilon_1, \ldots, \epsilon_n$?

Content of today's lecture

Recall the linear regression model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, \ldots, n.$$

What are the assumption on $\epsilon_1, \ldots, \epsilon_n$?

- Mutual independence,
- $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, for $i = 1, \ldots, n$.

Content of today's lecture

Recall the linear regression model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, \ldots, n.$$

What are the assumption on $\epsilon_1, \ldots, \epsilon_n$?

- Mutual independence,
- $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, for $i = 1, \ldots, n$.

Today, we consider the estimation of the:

- Regression coefficients $\underline{\beta} = (\beta_1, \ldots \beta_p)$.
- Residual variance σ^2.
- Residuals $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$.

Estimation of $\underline{\beta}$ (I)

We estimate $\underline{\beta}$ by considering the sum of squares

$$S\left(\underline{\beta}\right) = \sum_{i=1}^{n} (y_i - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2 \, .$$

We aim to derive the **least square estimate** $\hat{\underline{\beta}}$ minimizing $S(\underline{\beta})$.

Estimation of $\underline{\beta}$ (I)

We estimate $\underline{\beta}$ by considering the sum of squares

$$S\left(\underline{\beta}\right) = \sum_{i=1}^{n} \left(y_i - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p}\right)^2 .$$

We aim to derive the **least square estimate** $\hat{\underline{\beta}}$ minimizing $S(\underline{\beta})$.

To motivate this approach, note that

$$\epsilon_i = Y_i - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p}.$$

Thus, our approach is equivalent to minimizing the sum of squared observed residuals.

Estimation of $\underline{\beta}$ (II)

To find $\hat{\underline{\beta}}$ we must solve the equation,

$$-2\mathbf{X}^T\left(\mathbf{y}-\mathbf{X}\hat{\underline{\beta}}\right)=\mathbf{0}.$$

Estimation of $\underline{\beta}$ (II)

To find $\hat{\underline{\beta}}$ we must solve the equation,

$$-2\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\hat{\underline{\beta}}\right) = \mathbf{0}.$$

The solution to this equation is

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

which we term the **least square estimate**.

Estimation of $\underline{\beta}$ (II)

To find $\underline{\hat{\beta}}$ we must solve the equation,

$$-2\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}\right) = \mathbf{0}.$$

The solution to this equation is

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

which we term the **least square estimate**.

Remarks:

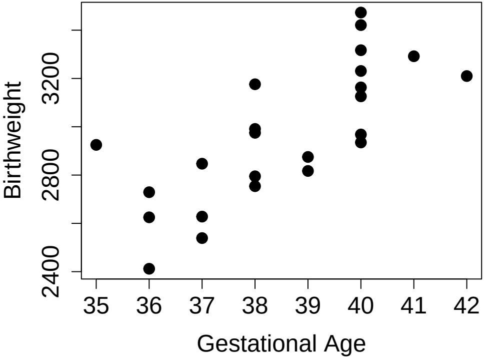
- The design matrix $\mathbf{X}$ must have linearly independent columns;
- We must check the second-order condition to verify that $\underline{\hat{\beta}}$ is a minimum.

Example 1: Birth Weight (I)

Last week, we considered the simple linear regression model

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where Y_i is the birth weight and x_i is the gestational age.



Example 1: Birth Weight (II)

We can now calculate the lest square estimates for β_1 and β_2.

The observed response vector and the design matrix are

$$\mathbf{y} = \begin{bmatrix} 2968 \\ 2795 \\ 3163 \\ 2925 \\ \vdots \\ 2875 \\ 3231 \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & 40 \\ 1 & 38 \\ 1 & 40 \\ 1 & 35 \\ \vdots & \\ 1 & 39 \\ 1 & 40 \end{bmatrix}.$$

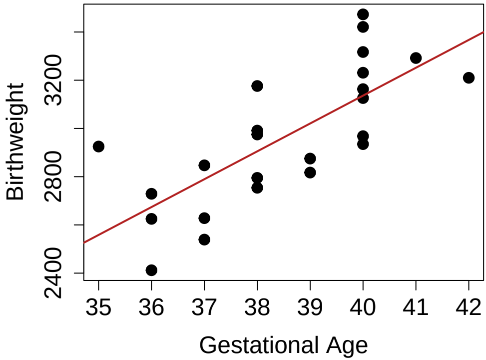
We then calculate

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \underline{\hat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} -1485 \\ 115.5 \end{pmatrix}.$$

Example 1: Birth Weight (III)

Thus, our model estimate is

$$\mathbb{E}(Y_i) = \mu_i = -1485 + 115.5x_i, \qquad i = 1, \ldots, n.$$



In practice, we usually use the `lm` function in R to derive the least square estimate $\rightarrow$ MA50258 Applied Statistics.

Example 2: Gas Consumption (I)

We study the impact of outside temperature on gas consumption. Information on whether insulation was installed is also provided.

Consider the model

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,1} x_{i,2}, \qquad i = 1, \ldots, n,$$

where

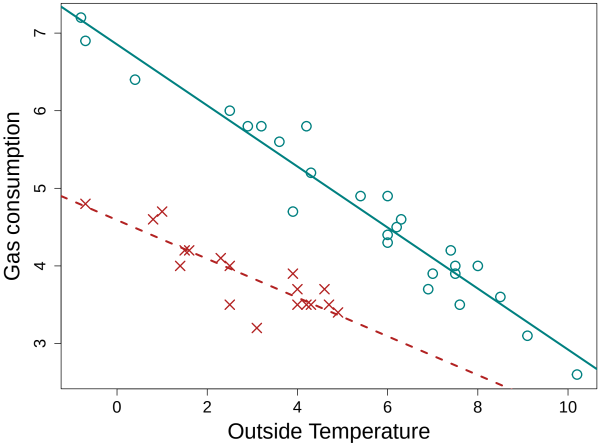
- $x_{i,1}$ is the outside temperature;
- $x_{i,2} = 1$ if cavity wall insulation was installed, and $x_{i,2} = 0$ otherwise.

The least square estimate is

$$\underline{\hat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (6.85,\ -0.393,\ -2.26,\ 0.144)^T.$$

Example 2: Gas Consumption (II)

Estimated models before ($\circ$) and after ($\times$) cavity wall insulation.



Predicted Values

Given the least square estimate $\underline{\hat{\beta}}$, we derive the **predicted value** as

$$\hat{\mu}_i = \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}, \qquad i = 1, \ldots, n.$$

The value $\hat{\mu}_i$ is our estimate for $\mathbb{E}(Y_i)$, conditional on $x_{i,1}, \ldots, x_{i,p}$.

Predicted Values

Given the least square estimate $\underline{\hat{\beta}}$, we derive the **predicted value** as

$$\hat{\mu}_i = \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}, \qquad i = 1, \ldots, n.$$

The value $\hat{\mu}_i$ is our estimate for $\mathbb{E}(Y_i)$, conditional on $x_{i,1}, \ldots, x_{i,p}$.

We can also obtain predicted values for unobserved combinations of explanatory variables.

For instance, our predicted expected birth weight for a child born at 34 weeks gestational age is

$$\hat{\mu} = -1485 + 115.5 \times 34 = 2442 \text{ grams}.$$

However, care should be taken regarding extrapolation.

Estimation of σ^2

We estimate the residual variance based on the **estimated residuals**

$$\hat{\epsilon}_i = y_i - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_p x_{i,p}, \qquad i = 1, \ldots, n.$$

The estimate of the residual variance is then

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

For the birth weight example with $n = 24$ observations, we

- Derive $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$,
- Calculate
$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \hat{\epsilon}_i^2 \approx 37094.$$