Simple Linear Regression: exercises

Exercice 1 (Estimation) *By minimising*

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2.$$

Find the expression of $\hat{\beta}_1$ *and* $\hat{\beta}_2$.

Calculate the derivatives according to β_1 and β_2, we get

$$\begin{cases} \dfrac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_1} & = & -2 \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0, \\ \dfrac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_2} & = & -2 \displaystyle\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0. \end{cases}$$

The first equation yields

$$\hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

and we have the estimator for the intercept

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \tag{1}$$

where $\bar{x} = \sum_{i=1}^{n} x_i / n$. The second equation yields

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Replacing $\hat{\beta}_1$ by its expression (1) we can rewrite it as

$$\hat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}},$$

Making use of the fact that the sum $\sum (x_i - \bar{x})$ is zero, we can rewrite the expression for the gradient of the straight line as

$$\hat{\beta}_2 = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}. \tag{2}$$

To obtain this result, we assume that there are at least two points with different x axis coordinates.

Exercice 2 (Bias) *Considering the model*

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

and under the assumption $\mathcal{H}_2 : \mathbb{E}(\varepsilon_i) = 0$, *for* $i = 1, \cdots, n$ *and* $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, *evaluate the bias of* $\hat{\beta}_2$ *and* $\hat{\beta}_1$.

Replace y_i by $\beta_1 + \beta_2 x_i + \varepsilon_i$ and get

$$\begin{aligned}
\hat{\beta}_2 &= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \frac{\beta_1 \sum(x_i - \bar{x}) + \beta_2 \sum x_i(x_i - \bar{x}) + \sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2} \\
&= \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}.
\end{aligned}$$

The last term is 0 by construction.
Now $\mathbb{E}(\hat{\beta}_1) = \mathbb{E}(\bar{y}) - \bar{x}\mathbb{E}(\hat{\beta}_2) = \beta_1 + \bar{x}\beta_2 - \bar{x}\beta_2 = \beta_1$.

Exercice 3 (Variance) *Considering the model*

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

and under the assumption $\mathcal{H}_2 : \mathbb{E}(\varepsilon_i) = 0$, *for* $i = 1, \cdots, n$ *and* $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, *evaluate the variance of* $\hat{\beta}_2$ *and* $\hat{\beta}_1$.

Start with $\mathrm{V}(\hat{\beta}_2)$ since the only random variable is ε we have

$$\begin{aligned}
\mathrm{V}(\hat{\beta}_2) &= \mathrm{V}\left(\beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}\right) \\
&= \mathrm{V}\left(\frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}\right) = \frac{\mathrm{V}\left(\sum(x_i - \bar{x})\varepsilon_i\right)}{\left[\sum(x_i - \bar{x})^2\right]^2} \\
&= \frac{\sum_{i,j}(x_i - \bar{x})(x_j - \bar{x})\,\mathrm{Cov}(\varepsilon_i, \varepsilon_j)}{\left[\sum(x_i - \bar{x})^2\right]^2}.
\end{aligned}$$

Since $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$ we have

$$\mathrm{V}(\hat{\beta}_2) = \frac{\sum_i(x_i - \bar{x})^2\sigma^2}{\left[\sum_i(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}.$$

Now

$$\begin{aligned}
\mathrm{V}(\hat{\beta}_1) &= \mathrm{V}\left(\bar{y} - \hat{\beta}_2\bar{x}\right) = \mathrm{V}\left(\bar{y}\right) + V(\bar{x}\hat{\beta}_2) - 2\,\mathrm{Cov}(\bar{y}, \hat{\beta}_2\bar{x}) \\
&= \mathrm{V}\left(\frac{\sum y_i}{n}\right) + \bar{x}^2\frac{\sigma^2}{\sum(x_i - \bar{x})^2} - 2\bar{x}\,\mathrm{Cov}(\bar{y}, \hat{\beta}_2) \\
&= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum(x_i - \bar{x})^2} - 2\bar{x}\sum_i \mathrm{Cov}(\bar{y}, \hat{\beta}_2).
\end{aligned}$$

Since

$$\begin{aligned}
\mathrm{Cov}(\bar{y}, \hat{\beta}_2) &= \frac{1}{n}\,\mathrm{Cov}\left(\sum_i \left(\beta_1 + \beta_2 x_i + \varepsilon_i\right), \frac{\sum_j (x_j - \bar{x})\varepsilon_j}{\sum_j (x_j - \bar{x})^2}\right) \\
&= \frac{1}{n}\sum_i \mathrm{Cov}\left(\varepsilon_i, \frac{\sum_j (x_j - \bar{x})\varepsilon_j}{\sum_j (x_j - \bar{x})^2}\right) \\
&= \frac{1}{\sum_j (x_j - \bar{x})^2}\sum_i \frac{1}{n}\,\mathrm{Cov}\left(\varepsilon_i, \sum_j (x_j - \bar{x})\varepsilon_j\right) \\
&= \frac{\sigma^2 \frac{1}{n}\sum_i (x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} = 0.
\end{aligned}$$

We get

$$\mathrm{V}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}.$$

Exercice 4 (Covariance) *Considering the model*

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

and under the assumption $\mathcal{H}_2 : \mathbb{E}(\varepsilon_i) = 0$*, for* $i = 1, \cdots, n$ *and* $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$*, evaluate the covariance of* $\hat{\beta}_2$ *and* $\hat{\beta}_1$*.*

We have

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \mathrm{Cov}(\bar{y} - \hat{\beta}_2 \bar{x}, \hat{\beta}_2) = \mathrm{Cov}(\bar{y}, \hat{\beta}_2) - \bar{x}\,\mathrm{V}(\hat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}.$$

Exercice 5 (Sum of residuals) *Show that in a simple linear regression model the sum of the residuals is zero.*

Just write

$$\sum_i \hat{\varepsilon}_i = \sum_i (y_i - \bar{y} + \hat{\beta}_2 \bar{x} - \hat{\beta}_2 x_i) = \sum_i (y_i - \bar{y}) - \hat{\beta}_2 \sum_i (x_i - \bar{x}) = 0.$$