

SHEET - An Introduction to Statistical Learning
Chapter 3 - Linear Regression

4 December 2023

1 Linear Regression

1.1 Courses' Demonstrations

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i th value of X_i . Then $e_i = y_i - \hat{y}_i$ represents the i th residual - this is the difference between the i th observed response value and the i th response value that is predicted by our linear model. We define the residual sum of squares (RSS) as

$$\begin{aligned} \text{RSS} &= e_1^2 + e_2^2 + \cdots + e_n^2 \\ &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \end{aligned}$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned} \tag{3.4}$$

where \bar{y} is the sample mean, defined as

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Demonstration : $\hat{\beta}_0$ and $\hat{\beta}_1$

We search RSS as

$$f(\hat{\beta}_0, \hat{\beta}_1) = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We get the minimum where

$$\frac{\partial f}{\partial \hat{\beta}_k} = 0 \text{ with } k \in \{0, 1\}$$

We have

$$\bar{y} = \sum_{i=1}^n y_i \text{ and } \bar{x} = \sum_{i=1}^n x_i$$

Then

$$\begin{cases} \frac{\partial f}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial f}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

$$\begin{aligned}
& \begin{cases} \sum_{i=1}^n y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n x_i y_i = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \end{cases} \\
& \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \sum_{i=1}^n x_i y_i - n\hat{\beta}_0 \bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \end{cases} \\
& \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \sum_{i=1}^n x_i y_i - n(\bar{y} - \hat{\beta}_1 \bar{x})\bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \end{cases} \\
& \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ (\sum_{i=1}^n x_i y_i - n\bar{y}\bar{x}) - \hat{\beta}_1 (\sum_{i=1}^n x_i^2 - n\bar{x}^2) = 0 \end{cases}
\end{aligned}$$

We have

$$\begin{cases} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i y_i - \bar{y}\bar{x}) \\ \sum_{i=1}^n (x_i^2 - \bar{x}^2) = \sum_{i=1}^n (x_i^2 - \bar{x}^2) \end{cases}$$

Finally

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{cases}$$

By developping we can have

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (y_i)(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

Demonstration : $\hat{\beta}_1$

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{\bar{x} \sum_{i=1}^n (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{\bar{x}n(\bar{y} - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

In the same way

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_1 x_i + \beta_0 + \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

We assume that the True relationship between X and Y takes the form

$$Y = f(X) + \varepsilon$$

for some unknown function f, where ε is a mean-zero random error term. If f is to be approximated by a linear function then we can write the relationship as

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (3.5)$$

How accurate is the sample mean $\hat{\mu}$ as an estimate of μ ? In general, we answer this question by computing the standard error of $\hat{\mu}$, written as the standard error $SE(\hat{\mu})$. A reasonable estimate is $\hat{\mu} = \bar{y}$. We have the well-known formula :

$$Var(\hat{\mu}) = SE(\hat{\mu}^2) = \frac{\sigma^2}{n} \quad (3.7)$$

where σ is the standard deviation of each of the realizations y_i of Y.

Demonstration : $Var(\hat{\mu})$

We have

$$\hat{\mu} = \bar{y}$$

$$Var(X + Y) = Var(X) + Var(Y)$$

$$Var(aX) = a^2 Var(X) \text{ with } a = \text{cste}$$

$$Var(y_i) = \sigma^2$$

Then

$$Var(\hat{\mu}) = Var(\bar{y}) = Var\left(\frac{1}{n} \sum_{i=1}^n y_i\right)$$

$$= \frac{1}{n^2} Var\left(\sum_{i=1}^n y_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(y_i)$$

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

$$\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n} =$$

In a similar vein, we can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1 . To compute the standard errors associated with β_0 and β_1 , we use the following formulas :

$$SE(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad (3.8)$$

$$SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.8)$$

where $\sigma^2 = Var(\varepsilon)$

Preliminaries :
We have

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i y_i \quad \text{with} \quad w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \sum_{i=1}^n w_i &= \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{n(\bar{x} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0 \\ \sum_{i=1}^n w_i x_i &= \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x} + \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n ((x_i - \bar{x}) \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\bar{x} \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 1 + 0 = 1 \\ \sum_{i=1}^n w_i^2 &= \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ y_i &= \beta_1 x_i + \beta_0 + \varepsilon_i \end{aligned}$$

Démonstration : $E[\hat{\beta}_1]$

$$\begin{aligned} \mathbf{E}[\hat{\beta}_1] &= E \left[\sum_{i=1}^n w_i y_i \right] = E \left[\sum_{i=1}^n w_i (\beta_1 x_i + \beta_0 + \varepsilon_i) \right] \\ &= E[\beta_1 \sum_{i=1}^n w_i x_i] + E[\beta_0 \sum_{i=1}^n w_i] + \sum_{i=1}^n E[w_i \varepsilon_i] \\ &= E[\beta_1 * 1] + E[\beta_0 * 0] + \sum_{i=1}^n E[\varepsilon_i] E[w_i] \\ \mathbf{E}[\hat{\beta}_1] &= \beta_1 \end{aligned}$$

Démonstration : $E[\hat{\beta}_0]$

$$\begin{aligned}
 \mathbf{E}[\hat{\beta}_0] &= E[\bar{y} - \hat{\beta}_1 \bar{x}] = E\left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_1 x_i)\right] = \frac{1}{n} \sum_{i=1}^n E[y_i - \hat{\beta}_1 x_i] \\
 &= \frac{1}{n} \sum_{i=1}^n E[(\beta_1 x_i + \beta_0 + \varepsilon_i) - \hat{\beta}_1 x_i] = \frac{1}{n} \sum_{i=1}^n (E[(\beta_1)E[x_i] + E[\beta_0] + E[\varepsilon_i]] - E[\hat{\beta}_1]E[x_i]) \\
 &= \frac{1}{n} \sum_{i=1}^n (\beta_1 x_i + \beta_0 + 0 - \beta_1 x_i) = \frac{1}{n} \sum_{i=1}^n \beta_0 = \beta_0
 \end{aligned}$$

Démonstration : $Var(\hat{\beta}_1)$

We have

$$\begin{aligned}
 \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i y_i \quad \text{with} \quad w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
 Var(\hat{\beta}_1) &= Var\left(\sum_{i=1}^n w_i y_i\right) = \sum_{i=1}^n Var(w_i y_i) = \sum_{i=1}^n w_i^2 Var(y_i) = \sigma^2 \sum_{i=1}^n w_i^2 \\
 Var(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
 \end{aligned}$$

Démonstration : $Var(\hat{\beta}_0)$

We have

$$\begin{aligned}
 \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
 Cov(X, cste) &= E[(X - E[X])(cste - E[cste])] = 0 \\
 Var(\bar{y}) &= \frac{\sigma^2}{n}
 \end{aligned}$$

Then

$$\begin{aligned}
 Var(\hat{\beta}_0) &= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1) \\
 &= \frac{\sigma^2}{n} + \bar{x}^2 Var(\hat{\beta}_1) \quad \text{because } \bar{y} = cste \\
 &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
 Var(\hat{\beta}_0) &= \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \sigma^2
 \end{aligned}$$

The estimate of σ is known as the residual standard error, and is given by the formula residual standard error

$$\text{RSE} = \frac{\sqrt{RSS}}{n-2}$$

Preliminaries :

$$\begin{aligned}\beta_1 - \hat{\beta}_1 &= -\frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n (y_i - \hat{y}_i) = n\bar{y} - \sum_{i=1}^n \hat{y}_i \\ &= n(\hat{\beta}_1 \bar{x} + \hat{\beta}_0) - \sum_{i=1}^n (\hat{\beta}_1 x_i + \hat{\beta}_0) = 0 \\ \frac{\partial f}{\partial \hat{\beta}_1} &= -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \sum_{i=1}^n \hat{\varepsilon}_i x_i &= \sum_{i=1}^n (y_i - \hat{y}_i) x_i = \sum_{i=1}^n (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0) x_i = 0 \\ \sum_{i=1}^n \hat{\varepsilon}_i \hat{y}_i &= \sum_{i=1}^n \hat{\varepsilon}_i (\hat{\beta}_1 x_i + \hat{\beta}_0) = \hat{\beta}_1 \sum_{i=1}^n \hat{\varepsilon}_i x_i + \hat{\beta}_0 \sum_{i=1}^n \hat{\varepsilon}_i = 0\end{aligned}$$

Demonstration : $\sum_{i=1}^n \varepsilon_i^2$

We have

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n \varepsilon_i (\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^n ((\hat{\beta}_1 x_i + \hat{\beta}_0) - (\hat{\beta}_1 \bar{x} + \hat{\beta}_0))^2 \\
&= \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2
\end{aligned}$$

$$E[X^2] = Var(X) + (E[X])^2$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

Then

$$\begin{aligned}
\sum_{i=1}^n \varepsilon_i^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^n y_i^2 - n\bar{y}^2 - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\
E\left[\sum_{i=1}^n \varepsilon_i^2\right] &= \sum_{i=1}^n E[y_i^2] - nE[\bar{y}^2] - E[\hat{\beta}_1^2] S_{xx} \\
&= \sum_{i=1}^n (\sigma^2 + (\beta_1 x_i + \beta_0)^2) - n\left(\frac{\sigma^2}{n} + (\beta_1 \bar{x} + \beta_0)^2\right) - \left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) S_{xx} \\
&= n\sigma^2 - \sigma^2 - \sigma^2 + \sum_{i=1}^n (\beta_1 x_i + \beta_0)^2 - n(\beta_1 \bar{x} + \beta_0)^2 - \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\
&= \sigma^2(n-2) \\
\text{RSE}^2 &= \frac{\text{RSS}}{n-2}
\end{aligned}$$

Recall that in the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta_1 = 0$. In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = \dots = \beta_p = 0$. As in the simple linear regression setting, we use a hypothesis test to answer this question. We test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

versus the alternative

$$H_a : \text{at least one } \beta_j \text{ is non-zero}$$

This hypothesis test is performed by computing the F-statistic,

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)} \quad (3.23)$$

where, as with simple linear regression, $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. If the linear model assumptions are correct, one can show that

$$E[RSS/(n - p - 1)] = \sigma^2$$

and that, provided H_0 is true,

$$E[(TSS - RSS)/p] = \sigma^2.$$

Demonstration : if H_0 True then $E[RSS/(n - p - 1)] = \sigma^2$
We have