2.2 Least Squares Estimation

Estimation is a method of finding values of the unknown model parameters for a given data set so that the model fits the data in a “best” way. There are various estimation methods, depending on how do we define “best”. In this section we consider the ***Method of Least Squares Estimation*** (***LS*** or ***LSE***).

The ***LS estimators*** of the model parameters β_0 and β_1 minimize the sum of squares of errors denoted by $S(\beta_0, \beta_2)$. That is, the estimators minimize

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2. \tag{2.2}$$

The “best” here means the smallest value of $S(\beta_0, \beta_1)$. S is a function of the parameters and so to find its minimum we differentiate it with respect to β_0 and β_1, then equate the derivatives to zero. We have

$$\begin{cases} \frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)] \\ \frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)]x_i \end{cases} \tag{2.3}$$

When compared to zero we obtain so called *normal equations*:

$$\begin{cases} \sum_{i=1}^{n}(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n}(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)x_i = \sum_{i=1}^{n} x_i Y_i \end{cases} \tag{2.4}$$

This set of equations can be written as

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i \end{cases} \tag{2.5}$$

The solutions to these equations are

$$\begin{aligned} \widehat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \bar{Y} - \widehat{\beta}_1 \bar{x} \end{aligned} \tag{2.6}$$

and, from the second normal equation

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^n x_i Y_i - \frac{1}{n}(\sum_{i=1}^n x_i)(\sum_{i=1}^n Y_i)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{S_{xY}}{S_{xx}},\end{aligned} \tag{2.7}$$

where

$$S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}), \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

To check that $S(\beta_0, \beta_1)$ attains a minimum at $(\widehat{\beta}_0, \widehat{\beta}_1)$ we calculate second derivatives and evaluate the determinant

$$\begin{vmatrix} \frac{\partial^2 S}{\partial \beta_0^2} & \frac{\partial S}{\partial \beta_0 \beta_1} \\ \\ \frac{\partial S}{\partial \beta_1 \beta_0} & \frac{\partial^2 S}{\partial \beta_1^2} \end{vmatrix} = \begin{vmatrix} 2n & 2\sum_{i=1}^n x_i \\ \\ 2\sum_{i=1}^n x_i & 2\sum_{i=1}^n x_i^2 \end{vmatrix} = 4n \sum_{i=1}^n (x_i - \bar{x})^2 > 0$$

for all β_0, β_1 (it does not depend on the parameters).

Also, $\frac{\partial^2 S}{\partial \beta_0^2} > 0$ (and $\frac{\partial^2 S}{\partial \beta_1^2} > 0$) for all β_0, β_1. This means that the function $S(\beta_0, \beta_1)$ attains a minimum at $(\widehat{\beta}_0, \widehat{\beta}_1)$ given by (2.6) and (2.7).

Remark 2.1. Note that the estimators depend on Y. They are functions of Y which is a random variable and so the estimators of the model parameters are random variables too. When we calculate the values of the estimators for a given data set, i.e. for ***observed*** values of Y at given values of X, we obtain so called ***estimates*** of the parameters. We may obtain different estimates of β_0 and β_1 calculated for different data sets fitted by the same kind of model. □

Example 2.3. (Wing's length cont.)
For the given data in Example 2.1 we obtain

$$\sum_{i=1}^{13} y_i = 44.1, \quad \sum_{i=1}^{13} x_i = 124.$$
$$\sum_{i=1}^{13} x_i y_i = 488.3, \quad \sum_{i=1}^{13} x_i^2 = 1418.$$

MINITAB

Stat → Basic Statistics → Store Descriptive Statistics...

Hence, the estimates of the model parameters are

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}\\ &= \frac{488.3 - \frac{1}{13} \times 124 \times 44.1}{1418 - \frac{1}{13} \times 124^2}\\ &= 0.28761\end{aligned}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = \frac{1}{13} \times 44.1 - 0.28761 \times \frac{1}{13} \times 124 = 0.649$$

and the estimated (fitted) linear model is

$$\widehat{y}_i = 0.649 + 0.288 x_i.$$

From this fitted model we may calculate values of the wing's length of sparrows for any age within the used age interval. For example, we may estimate the wing's length of sparrows of age 7 days (the missing value). It is

$$\widehat{y}_i = 0.649 + 0.288 \times 7 = 2.664 \text{ cm}.$$

MINITAB

Calc → Calculator...

□

Remark 2.2. Two special cases of the simple linear model are

- no-intercept model

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

  which implies that $\mathrm{E}(Y|X = 0) = 0$, and

- constant model

  $$Y_i = \beta_0 + \varepsilon_i,$$

  which implies that the response variable Y does not depend on the explanatory variable X.

□

2.3 Properties of the Estimators

Definition 2.1. *If $\widehat{\theta}$ is an estimator of θ and $E[\widehat{\theta}] = \theta$, then we say $\widehat{\theta}$ is* unbiased *for θ.*

Note that in this definition $\widehat{\theta}$ is a random variable. We must distinguish between $\widehat{\theta}$ when it is an estimate and when it is an estimator. As a function of the random variables Y_i it is an estimator. Its value obtained for a given data set (observed y_i) is an estimate.

The parameter estimator $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i, \quad \text{where} \quad c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{S_{xx}}. \tag{2.8}$$

We have assumed that $Y_1, Y_2, \ldots, Y_n$ are normally distributed and hence using the result that a linear combination of normal random variables is also a normal random variable, $\widehat{\beta}_1$ is also normally distributed. We now derive the mean and variance of $\widehat{\beta}_1$ using the representation (2.8).

$$\begin{aligned} E[\widehat{\beta}_1] &= E[\sum_{i=1}^{n} c_i Y_i] \\ &= \sum_{i=1}^{n} c_i E[Y_i] \\ &= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \end{aligned}$$

but $\sum c_i = 0$ and $\sum c_i x_i = 1$ as $\sum (x_i - \bar{x})x_i = S_{xx}$, so $E[\widehat{\beta}_1] = \beta_1$. Thus $\widehat{\beta}_1$ is unbiased for β_1. Also

$$\begin{aligned} \text{var}[\widehat{\beta}_1] &= \text{var}\left[\sum_{i=1}^{n} c_i Y_i\right] \\ &= \sum_{i=1}^{n} c_i^2 \,\text{var}[Y_i] \quad \text{since the } Y\text{'s are independent} \\ &= \sum_{i=1}^{n} \sigma^2 (x_i - \bar{x})^2/[S_{xx}]^2 \\ &= \sigma^2/S_{xx}. \end{aligned}$$

Hence,

$$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Similarly it can be shown that

$$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right).$$

2.4 Assessing the Model

2.4.1 Analysis of Variance Table

Parameter estimates obtained for the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

can be used to estimate the mean response corresponding to each variable Y_i, that is,

$$\widehat{\mathrm{E}(Y_i)} = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \ \ i = 1, \ldots, n.$$

These, for a given data set (x_i, y_i), are called ***fitted values*** and are denoted by $\widehat{y}_i$. They are points on the ***fitted regression line*** corresponding to the values of x_i. The observed values y_i usually do not fall exactly on the line and so are not equal to the fitted values $\widehat{y}_i$, as it is shown in Figure 2.5.

The ***residuals*** (also called crude residuals) are defined as

$$e_i := Y_i - \widehat{Y}_i, \quad i = 1, \ldots, n, \tag{2.9}$$

These are estimators of the random errors ε_i.

Thus

$$\begin{aligned} e_i &= Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \\ &= Y_i - \bar{Y} - \widehat{\beta}_1(x_i - \bar{x}). \end{aligned}$$

and

$$\sum e_i = 0.$$

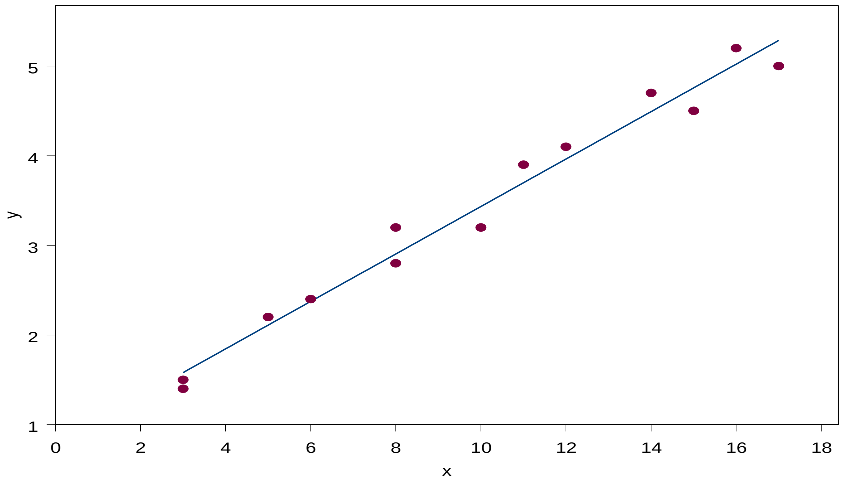

Figure 2.5: Observations and fitted line for the Sparrow wing's length data.

Also note that the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ minimize the function $S(\beta_0, \beta_1)$. The minimum is called the ***Residual Sum of Squares*** and is denoted by SS_E, that is,

$$SS_E = \sum_{i=1}^{n}[Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2. \tag{2.10}$$

Consider the constant model

$$Y_i = \beta_0 + \varepsilon_i.$$

For this model $\widehat{\beta}_0 = \bar{Y}$ and we have

$$\widehat{Y}_i = \bar{Y}, \quad e_i = Y_i - \widehat{Y}_i = Y_i - \bar{Y}$$

and

$$SS_E = SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

It is called the ***Total Sum of Squares*** and is denoted by SS_T. For a constant model $SS_E = SS_T$. When the model is non constant, i.e. there is a significant slope, the difference $Y_i - \bar{Y}$ can be split into two components: one due to the regression model fit and one due to the residuals, that is

$$Y_i - \bar{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \bar{Y}).$$

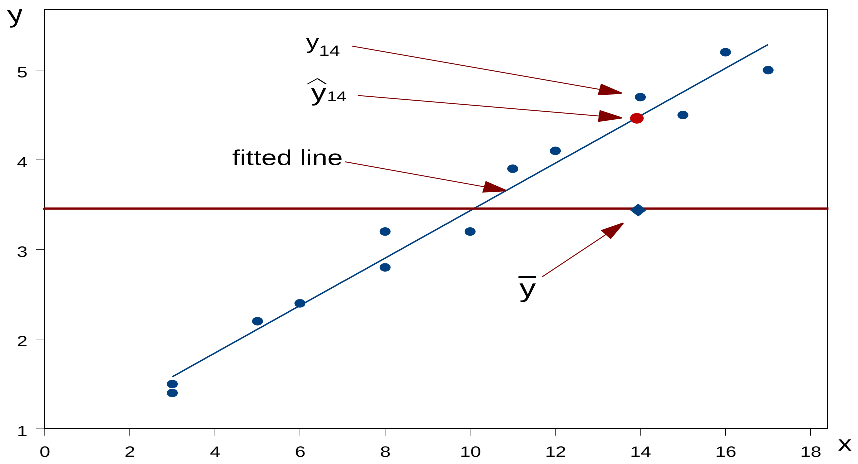

Figure 2.6: Observations, fitted line and the mean for a constant model.

For a given data set it could be represented as in Figure 2.6.

The following theorem gives such an identity for the respective sums of squares.

Theorem 2.1. *Analysis of Variance Identity.*
In the simple linear regression model the total sum of squares is a sum of the regression sum of squares and the residual sum of squares, that is

$$SS_T = SS_R + SS_E, \tag{2.11}$$

where

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ SS_R &= \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2 \\ SS_E &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \end{aligned}$$

Proof.

$$\begin{aligned}
SS_T &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}[(Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \bar{Y})]^2 \\
&= \sum_{i=1}^{n}[(Y_i - \widehat{Y}_i)^2 + (\widehat{Y}_i - \bar{Y})^2 + 2(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \bar{Y})] \\
&= SS_E + SS_R + 2A,
\end{aligned}$$

where

$$\begin{aligned}
A &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \bar{Y}) \\
&= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)\widehat{Y}_i - \bar{Y}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i) \\
&= \sum_{i=1}^{n} e_i\widehat{Y}_i - \bar{Y}\underbrace{\sum_{i=1}^{n} e_i}_{=0} \\
&= \sum_{i=1}^{n} e_i(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \\
&= \widehat{\beta}_0\underbrace{\sum_{i=1}^{n} e_i}_{=0} + \widehat{\beta}_1\underbrace{\sum_{i=1}^{n} e_i x_i}_{=0}.
\end{aligned}$$

Hence $A = 0$. □

For a given data set the model fit (regression) sum of squares, SS_R, represents the variability in the observations y_i accounted for by the fitted model, the residual sum of squares, SS_E, represents the variability in y_i accounted for by the differences between the observations and the fitted values.

The Analysis of Variance (ANOVA) Table shows the sources of variation, the sums of squares and the statistic, based on the sums of squares, for testing the significance of regression slope.

ANOVA table

Source of variation	d.f.	SS	MS	VR
Regression	$\nu_R = 1$	SS_R	$MS_R = \frac{SS_R}{\nu_R}$	$\frac{MS_R}{MS_E}$
Residual	$\nu_E = n - 2$	SS_E	$MS_E = \frac{SS_E}{\nu_E}$	
Total	$\nu_T = n - 1$	SS_T		

The “d.f.” is short for “degrees of freedom”.

What are degrees of freedom?

For an intuitive explanation consider the observations $y_1, y_2, \ldots, y_n$ and assume that their sum is fixed, say equal to a, that is

$$y_1 + y_2 + \ldots + y_n = a.$$

For a fixed value of the sum a there are $n-1$ arbitrary y-values but one y-value is determined by the difference of a and the $n-1$ arbitrary y values. This one value is not free, it depends on the other y-values an on a. We say, that there are $n-1$ independent (free to vary) pieces of information and one piece is taken up by a.

Estimates of parameters can be based upon different amounts of information. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom. This is why in order to calculate

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

we have $n-1$ free to vary pieces of information from the collected data, that is we have $n-1$ degrees of freedom. The one degree of freedom is taken up by $\bar{y}$. Similarly, for

$$SS_E = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

we have two degrees of freedom taken up: one by $\widehat{\beta}_0$ and one by $\widehat{\beta}_1$ (both depend on $y_1, y_2, \ldots, y_n$). Hence, there are $n-2$ independent pieces of information to calculate SS_E.

Finally, as $SS_R = SS_T - SS_E$ we can calculate the d.f. for SS_R as a difference between d.f. for SS_T and for SS_E, that is $\nu_R = (n-1) - (n-2) = 1$.

In the ANOVA table there are also included so called ***Mean Squares (MS)***, which can be thought of as measures of average variation.

The last column of the table contains the ***Variance Ratio (VR)***

$$\frac{MS_R}{MS_E}.$$

It measures the variation explained by the model fit relative to the variation due to residuals.

2.4.2 F test

The mean squares are function of random variables Y_i and so is their ratio. We denote it by F. We will see later, that if $\beta_1 = 0$, then

$$F = \frac{MS_R}{MS_E} \sim \mathcal{F}_{1,n-2}.$$

Thus, to test the null hypothesis

$$H_0 : \beta_1 = 0$$

versus the alternative

$$H_1 : \beta_1 \neq 0,$$

we use the variance ratio F as the test statistic. Under H_0 the ratio has $\mathcal{F}$ distribution with 1 and $n-2$ degrees of freedom.

We reject H_0 at a significance level α if

$$F_{cal} > \mathcal{F}_{\alpha;1,n-2},$$

where F_{cal} denotes the value of the variance ratio F calculated for a given data set and $\mathcal{F}_{\alpha;1,n-2}$ is such that

$$P(F > \mathcal{F}_{\alpha;1,n-2}) = \alpha.$$

There is no evidence to reject H_0 if $F_{cal} < \mathcal{F}_{\alpha;1,n-2}$.

Rejecting H_0 means that the slope $\beta_1 \neq 0$ and the full regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

is better then the constant model

$$Y_i = \beta_0 + \varepsilon_i.$$