2.4.3 Estimating σ^2

Note that the sums of squares are functions of the conditional random variables $Y_i = (Y|X = x_i)$. Hence, the sums of squares are random variables as well. This fact allows us to check some stochastic properties of the sums of squares, such as their expectation, variance and distribution.

Theorem 2.2. *In the full simple linear regression model we have*

$$\mathrm{E}(SS_E) = (n-2)\sigma^2$$

Proof. Proof will be given later. □

From the theorem we obtain

$$\mathrm{E}(MS_E) = \mathrm{E}\left(\frac{1}{n-2}SS_E\right) = \sigma^2$$

and so MS_E is an unbiased estimator of σ^2. It is often denoted by S^2.

Notice, that in the full model S^2 is not the sample variance. We have

$$S^2 = MS_E = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\mathrm{E}(Y_i)})^2, \quad \text{where } \widehat{\mathrm{E}(Y_i)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

It is the sample variance in the constant (null) model, where $\widehat{\mathrm{E}(Y_i)} = \widehat{\beta}_0 = \bar{Y}$ and $\nu_E = n-1$. Then

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

2.4.4 Coefficient of Determination

The coefficient of determination, denoted by R^2, is the percentage of total variation in the data explained by the fitted model, that is

$$R^2 = \frac{SS_R}{SS_T}100\% = \frac{SS_T - SS_E}{SS_T}100\% = \left(1 - \frac{SS_E}{SS_T}\right)100\%. \tag{2.12}$$

Note:

- $R^2 \in [0, 100]$.
- $R^2 = 0$ indicates that none of the variability in the response is explained by the regression model.
- $R^2 = 100$ indicates that $SS_E = 0$ and all observations fall on the fitted line exactly.

A small value of R^2 does not always imply a poor relationship between Y and X, which may, for example, follow another model.

2.4.5 Minitab Example

Example 2.4. Sparrow Wings continued

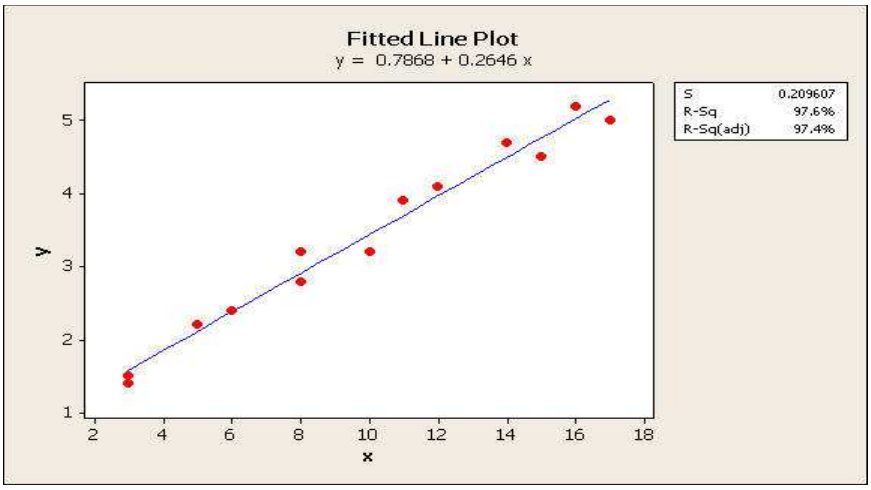


Figure 2.7: Fitted line plot for Sparrow Wings

```
The regression equation is
y = 0.787 + 0.265 x

Predictor      Coef   SE Coef      T      P
Constant     0.7868    0.1368   5.75  0.000
x           0.26463   0.01258  21.04  0.000
```

```
S = 0.209607   R-Sq = 97.6%   R-Sq(adj) = 97.4%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       1    19.446   19.446   442.60  0.000
Residual Error  11     0.483    0.044
Total           12    19.929
```

Comments:
We fitted a simple linear model of the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, 13, \quad \varepsilon_i \underset{iid}{\sim} \mathcal{N}(0,1).$$

The estimated values of the parameters are
- intercept: $\widehat{\beta}_0 \cong 0.79$
- slope: $\widehat{\beta}_1 \cong 0.26$
Both parameters are highly significant ($p < 0.001$).

The ANOVA table also shows the significance of the regression (slope), that is the null hypothesis

$$H_0 : \beta_1 = 0$$

versus the alternative

$$H_1 : \beta_1 \neq 0$$

can be rejected on the significance level $\alpha < 0.001$ ($p \cong 0.000$).

The tests require the assumptions of the normality and of constant variance of random errors. It should be checked whether the assumptions are approximately met. If not, the tests may not be valid.

The value of R^2 is very high, i.e., $R^2 = 97.6$. It means that the fitted model explains the variability in the observed responses very well.

The graph shows that the observations lie along the fitted line and there are no strange points which are far from the line or which could strongly affect the slope.

Final conclusions:
We can conclude that the data indicate that the length of sparrows' wings depends linearly on their age (within the range 3 - 18 days). The mean increase in the wing's length per day is estimated as $\widehat{\beta}_1 \cong 0.26$ cm.

However, it might be wrong to predict the length or its increase per day outside the range of the observed time. We would expect that the growth slows down in time and so the relationship becomes non-linear. □

2.5 Residuals

2.5.1 Crude Residuals

In Section 2.4.1 we defined the residuals as

$$e_i = Y_i - \widehat{Y}_i.$$

These are often called *crude residuals*. We have

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

We also have seen that

$$\sum_{i=1}^{n} e_i = 0.$$

Now the question is what is the expectation and the variance of crude residuals?

The mean of the ith residual is

$$\begin{aligned} E[e_i] &= E[Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i] \\ &= E[Y_i] - E[\widehat{\beta}_0] - x_i E[\widehat{\beta}_1] \\ &= \beta_0 + \beta_1 x_i - \beta_0 - \beta_1 x_i \\ &= 0. \end{aligned}$$

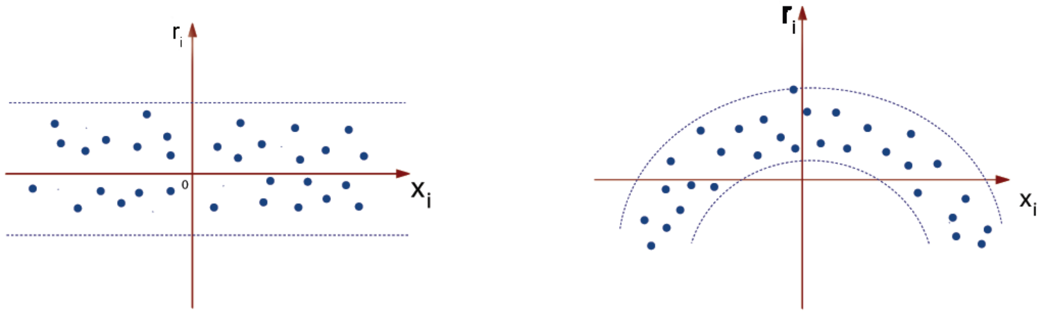
The variance is given by

$$\operatorname{var}[e_i] = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right] = \sigma^2 (1 - h_{ii}),$$

which can be shown by writing e_i as a linear combination of the Y_i's. Note that it depends on i, that is the variance of e_i is not constant, unlike that of ε_i. Similarly it can be shown that the covariance of two residuals e_i and e_j is

$$\operatorname{cov}[e_i, e_j] = -\sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right] = -\sigma^2 h_{ij}.$$

We know that $\operatorname{var}[\varepsilon_i] = \sigma^2$ and $\operatorname{cov}[\varepsilon_i, \varepsilon_j] = 0$. So the crude residuals e_i do not quite mimic the properties of ε_i.



(a) (b)

Figure 2.8: (a) No problem apparent (b) Clear non-linearity

2.5.2 Standardized/Studentized Residuals

To standardize a random variable we subtract its mean and divide by its standard error. Hence, to standardize residuals we calculate

$$d_i = \frac{e_i - \mathrm{E}(e_i)}{\sqrt{\operatorname{var} e_i}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}.$$

Then

$$d_i \sim \mathcal{N}(0,1).$$

They are not independent, though for large samples the correlation should be small.

However, we do not know σ^2. If we replace σ^2 by S^2 we get the so called *studentized residuals* (in Minitab they are called standardized residuals),

$$r_i = \frac{e_i}{\sqrt{S^2(1-h_{ii})}}.$$

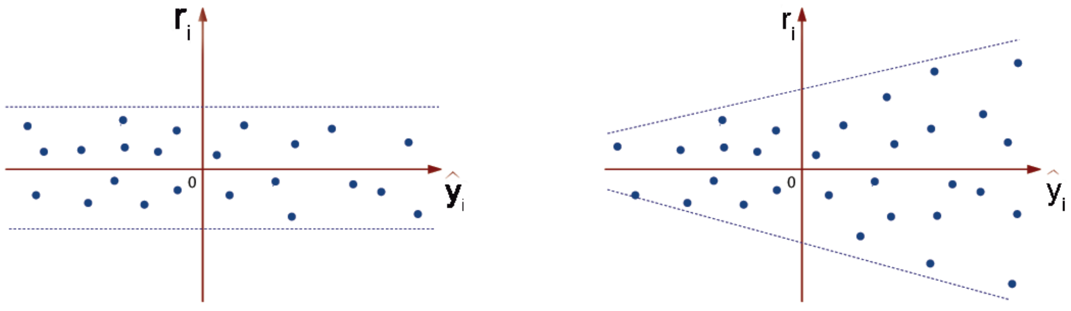
For large samples they will approximate the standard d_i.

2.5.3 Residual plots

Shapes of various residual plots can show whether the model assumptions are approximately met.

To check linearity, we plot r_i against x_i, as it is shown in Figure 2.8.

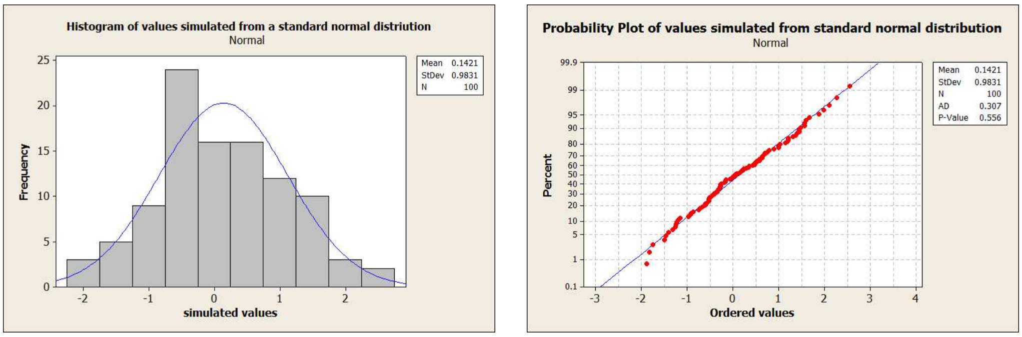
To check the assumption of constant variance (homoscedasticity), we plot r_i against the fitted values $\widehat{y}_i$, as it is shown in Figure 2.9. This plot can also indicate whether the assumption of model linearity is approximately satisfied.



(a) (b)

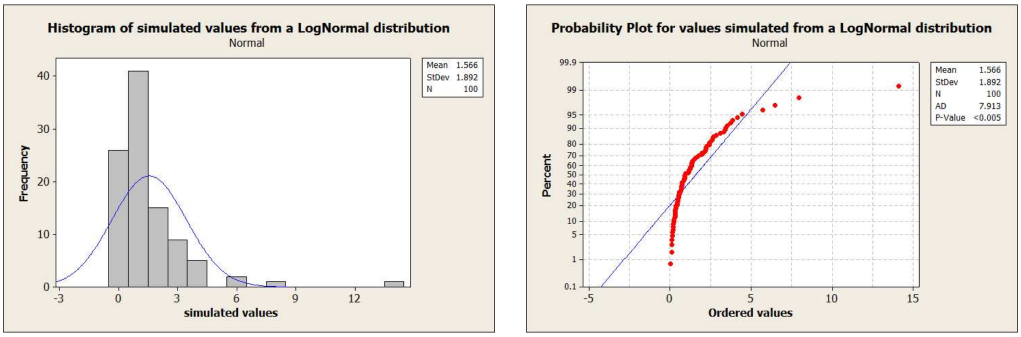
Figure 2.9: (a) No problem apparent (b) Variance increases as the mean response increases

To check whether the distribution of the residuals follows a symmetric shape of the normal distribution we can draw so called *Normal Probability Plot*. It plots each value of ordered residuals vs. the percentage of values in the sample that are less than or equal to it, along a fitted distribution line. The scales are transformed so that the fitted distribution forms a straight line. A plot that departs substantially from linearity suggests that the error distribution is not normal as shown in plots 2.10 - 2.13.



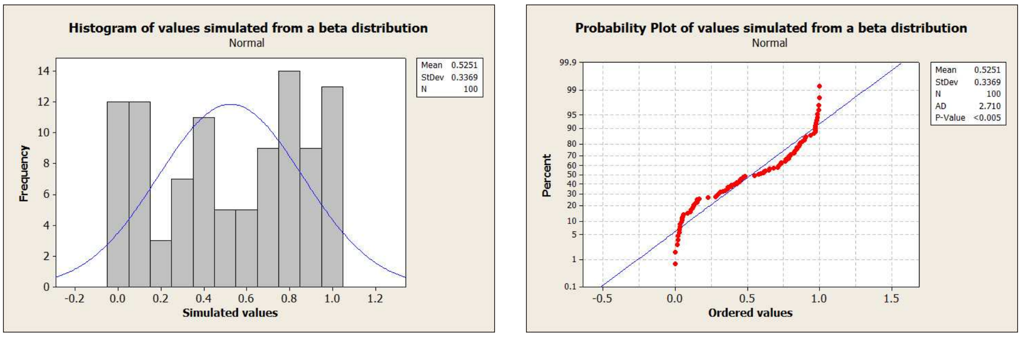
(a) (b)

Figure 2.10: (a) Histogram of data simulated from standard normal distribution, (b) Normal Probability Plot, no problem apparent.



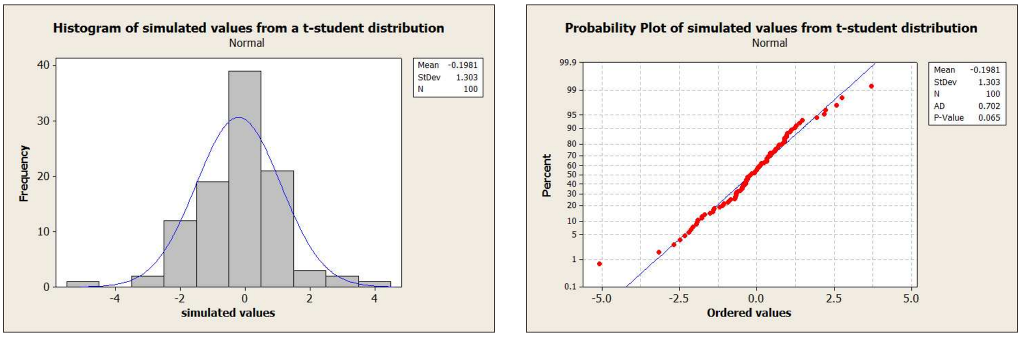
(a) (b)

Figure 2.11: (a) Histogram of data simulated from a Log-normal distribution, (b) Normal Probability Plot indicates skewness of the distribution.



(a) (b)

Figure 2.12: (a) Histogram of data simulated from a Beta distribution, (b) Normal Probability Plot indicates light tails.



(a) (b)

Figure 2.13: (a) Histogram of data simulated from a Student t-distribution, (b) Normal Probability Plot indicates heavy tails.

2.6 Inference about the regression parameters

Example 2.5. Overheads.
A company builds custom electronic instruments and computer components. All jobs are manufactured to customer specifications. The firm wants to be able to estimate its overhead cost. As part of a preliminary investigation, the firm decides to focus on a particular department and investigates the relationship between total departmental overhead cost (Y) and total direct labor hours (X). The data for the most recent 16 months are plotted in Figure 2.14.

Two objectives of this investigation are

1. to summarize for management the relationship between total departmental overhead and total direct labor hours.
2. to estimate the expected and to predict the actual total departmental overhead from the total direct labor hours.

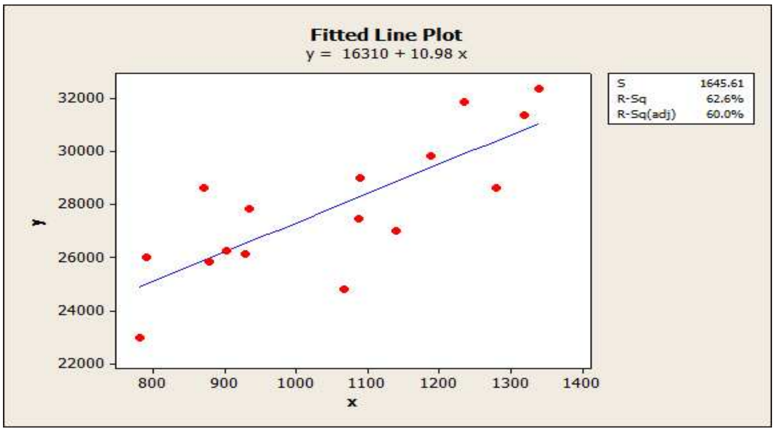


Figure 2.14: Plot of overheads data

```
The regression equation is
Ovhd = 16310 + 11.0 Labor

Predictor     Coef     SE Coef      T       P
Constant      16310    2421       6.74    0.000
Labor         10.982   2.268      4.84    0.000

S = 1645.61   R-Sq = 62.6%   R-Sq(adj) = 60.0%
```

```
Analysis of Variance
Source          DF        SS           MS      F      P
Regression       1     63517077   63517077  23.46  0.000
Residual Error  14     37912232    2708017
Total           15    101429309

Unusual Observations
Obs   Labor   Ovhd    Fit    SE Fit  Residual  St Resid
  6   1067    24817  28028    413      -3211     -2.02R

R denotes an observation with a large standardized residual.
```

Comments:

- The model fit is $\widehat{y}_i = 16310 + 11x_i$. There is a significant relationship between the overheads and the labor hours ($p < 0.001$ in ANOVA).
- The increase of labor hours by 1 will increase the mean overheads by about £11 ($\widehat{\beta}_1 = 11.0$).
- There is rather large variability in the data; the percentage of total variation explained by the model is rather small ($R^2 = 62.6$).

The model allows us to estimate the total overhead cost as a function of labour hours, but as we noticed, there is large variability in the data. In such a case, the point estimates may not be very reliable. Anyway, point estimates should always be accompanied by their standard errors. Then we can also find confidence intervals (CI) for the unknown model parameters, or test their non-significance. □

Note that for the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where } \varepsilon_i \underset{iid}{\sim} N(0, \sigma^2), \tag{2.13}$$

we obtained the following LSE of the parameters β_0 and β_1:

$$\begin{aligned} \widehat{\beta}_0 &= \bar{Y} - \widehat{\beta}_1 \bar{x} \\ \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

We now derive results which allow us to make inference about the regression parameters and predictions.

2.6.1 Inference about β_1

We proved the following result in Section 2.3.

Theorem 2.3. *In the full simple linear model (SLM) the distribution of the LSE of* β_1, $\widehat{\beta}_1$, *is normal with the expectation* $\mathrm{E}(\widehat{\beta}_1) = \beta_1$ *and the variance* $\mathrm{var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, *that is*

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right). \tag{2.14}$$

□

Remark 2.3. For large samples, where there is no assumption of normality of Y_i, the sampling distribution of $\widehat{\beta}_1$ is approximately normal. □

Theorem 2.3 allows us to derive a confidence interval (CI) for β_1 and a test of non-significance for β_1. After standarisation of $\widehat{\beta}_1$ we obtain

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0, 1).$$

However, the error variance is usually not known and it is replaced by its estimator. Then the normal distribution changes to a Student t-distribution. The explanation is following.

Lemma 2.1. *If* $Z \sim N(0, 1)$ *and* $U \sim \chi^2_\nu$, *and* Z *and* U *are independent, then*

$$\frac{Z}{\sqrt{U/\nu}} \sim t_\nu.$$

□

Here we have,

$$Z = \frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0, 1).$$

We will see later that

$$U = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

and S^2 and $\widehat{\beta}_1$ are independent. It follows that

$$T = \frac{\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2(n-2)}}} = \frac{\widehat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t_{n-2}. \tag{2.15}$$