Chapter 2
Simple Linear Regression

2.1 Introduction and Least Squares Estimates

Regression analysis is a method for investigating the functional relationship among variables. In this chapter we consider problems involving modeling the relationship between two variables. These problems are commonly referred to as simple linear regression or straight-line regression. In later chapters we shall consider problems involving modeling the relationship between three or more variables.

In particular we next consider problems involving modeling the relationship between two variables as a straight line, that is, when Y is modeled as a linear function of X.

Example: A regression model for the timing of production runs
We shall consider the following example taken from Foster, Stine and Waterman (1997, pages 191–199) throughout this chapter. The original data are in the form of the time taken (in minutes) for a production run, Y, and the number of items produced, X, for 20 randomly selected orders as supervised by three managers. At this stage we shall only consider the data for one of the managers (see Table 2.1 and Figure 2.1). We wish to develop an equation to model the relationship between Y, the run time, and X, the run size.

A scatter plot of the data like that given in Figure 2.1 should **ALWAYS** be drawn to obtain an idea of the sort of relationship that exists between two variables (e.g., linear, quadratic, exponential, etc.).

2.1.1 Simple Linear Regression Models

When data are collected in pairs the standard notation used to designate this is:

$$(x_1, y_1),(x_2, y_2), \ldots ,(x_n, y_n)$$

where x_1 denotes the first value of the so-called X-variable and y_1 denotes the first value of the so-called Y-variable. The X-variable is called the **explanatory** or **predictor variable**, while the Y-variable is called the **response variable** or the **dependent variable**. The X-variable often has a different status to the Y-variable in that:

S.J. Sheather, *A Modern Approach to Regression with R*,
DOI: 10.1007/978-0-387-09608-7_2, © Springer Science+Business Media LLC 2009

Table 2.1 Production data (production.txt)

Case	Run time	Run size	Case	Run time	Run size
1	195	175	11	220	337
2	215	189	12	168	58
3	243	344	13	207	146
4	162	88	14	225	277
5	185	114	15	169	123
6	231	338	16	215	227
7	234	271	17	147	63
8	166	173	18	230	337
9	253	284	19	208	146
10	196	277	20	172	68

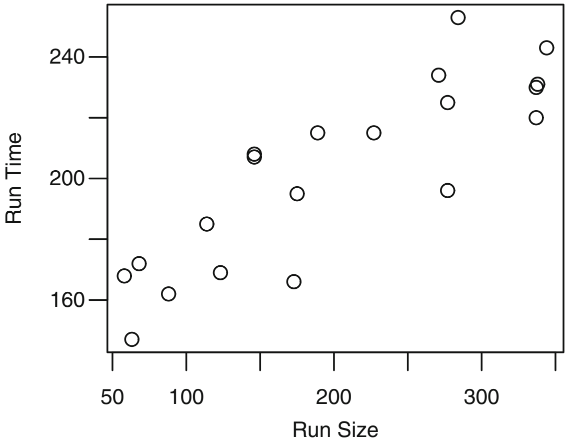

Figure 2.1 A scatter plot of the production data

- It can be thought of as a potential predictor of the Y-variable
- Its value can sometimes be chosen by the person undertaking the study

Simple linear regression is typically used to model the relationship between two variables Y and X so that given a specific value of X, that is, $X = x$, we can predict the value of Y. Mathematically, the regression of a random variable Y on a random variable X is

$$\mathrm{E}(Y|X=x),$$

the expected value of Y when X takes the specific value x. For example, if X = Day of the week and Y = Sales at a given company, then the regression of Y on X represents the mean (or average) sales on a given day.

The regression of Y on X is linear if

$$\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x \tag{2.1}$$

where the unknown parameters β_0 and β_1 determine the intercept and the slope of a specific straight line, respectively. Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent realizations of the random variable Y that are observed at the values $x_1, x_2, \ldots, x_n$ of a random variable X. If the regression of Y on X is linear, then for $i = 1, 2, \ldots, n$

$$Y_i = \mathrm{E}(Y \mid X = x) + e_i = \beta_0 + \beta_1 x + e_i$$

where e_i is the random error in Y_i and is such that $\mathrm{E}(e|X) = 0$.

The random error term is there since there will almost certainly be some variation in Y due strictly to random phenomenon that cannot be predicted or explained. In other words, all unexplained variation is called **random error**. Thus, the random error term does not depend on x, nor does it contain any information about Y (otherwise it would be a systematic error).

We shall begin by assuming that

$$\mathrm{Var}(Y \mid X = x) = \sigma^2. \tag{2.2}$$

In Chapter 4 we shall see how this last assumption can be relaxed.

Estimating the population slope and intercept

Suppose for example that X = height and Y = weight of a randomly selected individual from some population, then for a straight line regression model the mean weight of individuals of a given height would be a linear function of that height. In practice, we usually have a sample of data instead of the whole population. The slope β_1 and intercept β_0 are unknown, since these are the values for the whole population. Thus, we wish to use the given data to estimate the slope and the intercept. This can be achieved by finding the equation of the line which "best" fits our data, that is, choose b_0 and b_1 such that $\hat{y}_i = b_0 + b_1 x_i$ is as "close" as possible to y_i. Here the notation $\hat{y}_i$ is used to denote the value of the line of best fit in order to distinguish it from the observed values of y, that is, y_i. We shall refer to $\hat{y}_i$ as the ith **predicted value** or the **fitted value** of y_i.

Residuals

In practice, we wish to minimize the difference between the actual value of y (y_i) and the predicted value of y ($\hat{y}_i$). This difference is called the residual, $\hat{e}_i$, that is,

$$\hat{e}_i = y_i - \hat{y}_i.$$

Figure 2.2 shows a hypothetical situation based on six data points. Marked on this plot is a **line of best fit**, $\hat{y}_i$ along with the residuals.

Least squares line of best fit

A very popular method of choosing b_0 and b_1 is called the method of least squares. As the name suggests b_0 and b_1 are chosen to minimize the sum of squared residuals (or residual sum of squares [RSS]),

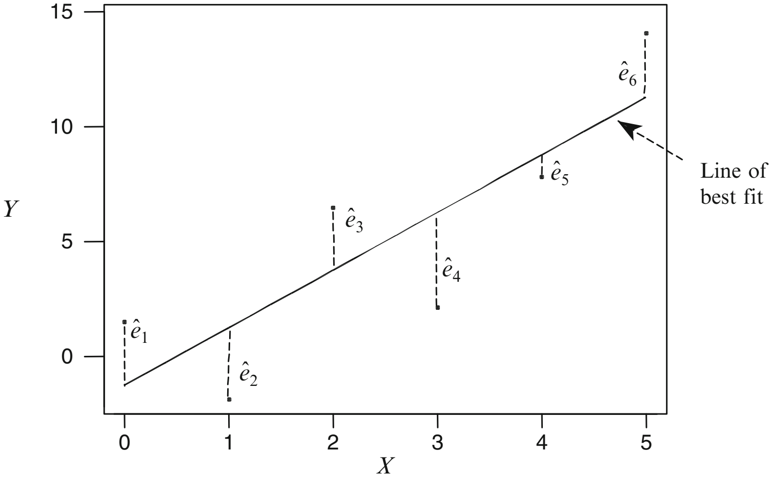

Figure 2.2 A scatter plot of data with a line of best fit and the residuals identified

$$\text{RSS} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

For RSS to be a minimum with respect to b_0 and b_1 we require

$$\frac{\partial \text{RSS}}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

and

$$\frac{\partial \text{RSS}}{\partial b_1} = -2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

Rearranging terms in these last two equations gives

$$\sum_{i=1}^{n} y_i = b_0 n + b_1 \sum_{i=1}^{n} x_i$$

and

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 .$$

These last two equations are called the **normal equations**. Solving these equations for b_0 and b_1 gives the so-called **least squares estimates** of the intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.3}$$

and the slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{SXY}{SXX} \ . \tag{2.4}$$

Regression Output from R

The least squares estimates for the production data were calculated using R, giving the following results:

```
Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   149.74770   8.32815    17.98   6.00e-13 ***
RunSize         0.25924   0.03714     6.98   1.61e-06 ***
---
Signif. codes:0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-Squared: 0.7302,      Adjusted R-squared: 0.7152
F-statistic: 48.72 on 1 and 18 DF, p-value: 1.615e-06
```

The least squares line of best fit for the production data

Figure 2.3 shows a scatter plot of the production data with the least squares line of best fit. The equation of the least squares line of best fit is

$$y = 149.7 + 0.26x.$$

Let us look at the results that we have obtained from the line of best fit in Figure 2.3. The intercept in Figure 2.3 is 149.7, which is where the line of best fit crosses the run time axis. The slope of the line in Figure 2.3 is 0.26. Thus, we say that each additional unit to be produced is predicted to add 0.26 minutes to the run time. The intercept in the model has the following interpretation: for any production run, the average set up time is 149.7 minutes.

Estimating the variance of the random error term

Consider the linear regression model with constant variance given by (2.1) and (2.2). In this case,

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, 2, \ldots, n)$$

where the random error e_i has mean 0 and variance σ^2. We wish to estimate $\sigma^2 = \text{Var}(e)$. Notice that

$$e_i = Y_i - (\beta_0 + \beta_1 x_i) = Y_i - \text{unknown regression line at } x_i.$$

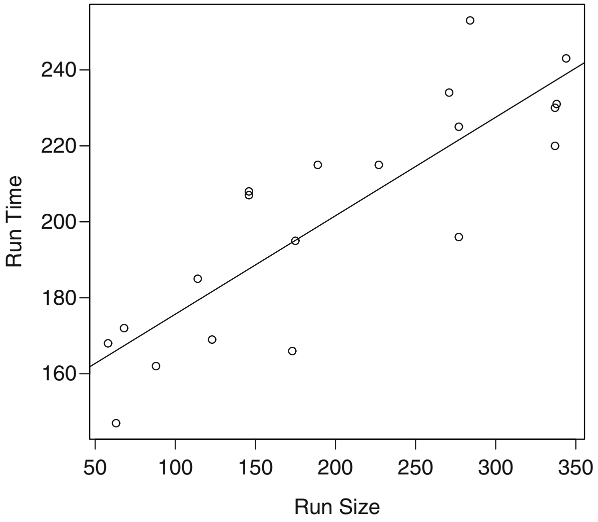

Figure 2.3 A plot of the production data with the least squares line of best fit

Since β_0 and β_1 are unknown all we can do is estimate these errors by replacing β_0 and β_1 by their respective least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ giving the residuals

$$\hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = Y_i - \text{estimated regression line at } x_i.$$

These residuals can be used to estimate σ^2. In fact it can be shown that

$$S^2 = \frac{\text{RSS}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2$$

is an unbiased estimate of σ^2.

Two points to note are:

1. $\bar{\hat{e}} = 0$ (since $\sum \hat{e}_i = 0$ as the least squares estimates minimize $\text{RSS} = \sum \hat{e}_i^2$)
2. The divisor in S^2 is $n-2$ since we have estimated two parameters, namely β_0 and β_1.

2.2 Inferences About the Slope and the Intercept

In this section, we shall develop methods for finding confidence intervals and for performing hypothesis tests about the slope and the intercept of the regression line.

2.2.1 Assumptions Necessary in Order to Make Inferences About the Regression Model

Throughout this section we shall make the following assumptions:

1. Y is related to x by the simple linear regression model
   $Y_i = \beta_0 + \beta_1 x_i + e_i \ (i = 1,\ldots,n)$, i.e., $\mathrm{E}(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$
2. The errors $e_1, e_2, \ldots, e_n$ are independent of each other
3. The errors $e_1, e_2, \ldots, e_n$ have a common variance σ^2
4. The errors are normally distributed with a mean of 0 and variance σ^2, that is, $e \mid X \sim N(0, \sigma^2)$

Methods for checking these four assumptions will be considered in Chapter 3. In addition, since the regression model is conditional on X we can assume that the values of the predictor variable, x_1, x_2, …, x_n are known fixed constants.

2.2.2 Inferences About the Slope of the Regression Line

Recall from (2.4) that the least squares estimate of β_1 is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{SXY}{SXX}$$

Since, $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$ we find that

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} (x_i - \overline{x}) y_i - \overline{y} \sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} (x_i - \overline{x}) y_i$$

Thus, we can rewrite $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \text{ where } c_i = \frac{x_i - \overline{x}}{SXX} \tag{2.5}$$

We shall see that this version of $\hat{\beta}_1$ will be used whenever we study its theoretical properties.

Under the above assumptions, we shall show in Section 2.7 that

$$\mathrm{E}(\hat{\beta}_1 \mid X) = \beta_1 \tag{2.6}$$

$$\mathrm{Var}(\hat{\beta}_1 \mid X) = \frac{\sigma^2}{SXX} \tag{2.7}$$

$$\hat{\beta}_1 \mid X \sim N\left(\beta_1, \frac{\sigma^2}{SXX}\right) \tag{2.8}$$

Note that in (2.7) the variance of the least squares slope estimate decreases as *SXX* increases (i.e., as the variability in the *X*'s increases). This is an important fact to note if the experimenter has control over the choice of the values of the *X* variable.

Standardizing (2.8) gives

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma \big/ \sqrt{SXX}} \sim N(0,1)$$

If σ were known then we could use a *Z* to test hypotheses and find confidence intervals for β_1. When σ is unknown (as is usually the case) replacing σ by *S*, the standard deviation of the residuals results in

$$T = \frac{\hat{\beta}_1 - \beta_1}{S \big/ \sqrt{SXX}} = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)}$$

where $\mathrm{se}(\hat{\beta}_1) = S \big/ \sqrt{SXX}$ is the estimated standard error (se) of $\hat{\beta}_1$, which is given directly by R. In the production example the *X*-variable is *RunSize* and so $\mathrm{se}(\hat{\beta}_1) = 0.03714$.

It can be shown that under the above assumptions that *T* has a *t*-distribution with $n - 2$ degrees of freedom, that is

$$T = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2}$$

Notice that the degrees of freedom satisfies the following formula

degrees of freedom = sample size – number of mean parameters estimated.

In this case we are estimating two such parameters, namely, β_0 and β_1.

For **testing the hypothesis** $H_0 : \beta_1 = \beta_1^0$ the test statistic is

$$T = \frac{\hat{\beta}_1 - \beta_1^0}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

R provides the value of *T* and the *p*-value associated with testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$ (i.e., for the choice $\beta_1^0 = 0$). In the production example the *X*-variable is *RunSize* and $T = 6.98$, which results in a *p*-value less than 0.0001.

A $100(1-\alpha)\%$ **confidence interval** for β_1, the slope of the regression line, is given by

$$(\hat{\beta}_1 - t(\alpha/2, n-2)\,\text{se}(\hat{\beta}_1), \hat{\beta}_1 + t(\alpha/2, n-2)\,\text{se}(\hat{\beta}_1))$$

where $t(\alpha/2, \text{n} - 2)$ is the 100(1– $\alpha/2$)th quantile of the t-distribution with $n - 2$ degrees of freedom.

In the production example the X-variable is *RunSize* and $\hat{\beta}_1 = 0.25924, \text{se}(\hat{\beta}_1) =$ 0.03714, $t\,(0.025, 20–2 = 18) = 2.1009$. Thus a 95% confidence interval for β_1 is given by

$$(0.25924 \pm 2.1009 \times 0.03714) = (0.25924 \pm 0.07803) = (0.181, 0.337)$$

2.2.3 Inferences About the Intercept of the Regression Line

Recall from (2.3) that the least squares estimate of β_0 is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Under the assumptions given previously we shall show in Section 2.7 that

$$\text{E}(\hat{\beta}_0 \mid X) = \beta_0 \tag{2.9}$$

$$\text{Var}(\hat{\beta}_0 \mid X) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right) \tag{2.10}$$

$$\hat{\beta}_0 \mid X \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)\right) \tag{2.11}$$

Standardizing (2.11) gives

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{{}^1\!/_n + {}^{\bar{x}^2}\!/_{SXX}}} \sim N(0,1)$$

If σ were known then we could use Z to test hypotheses and find confidence intervals for β_0. When σ is unknown (as is usually the case) replacing σ by S results in

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{{}^1\!/_n + {}^{\bar{x}^2}\!/_{SXX}}} = \frac{\hat{\beta}_0 - \beta_0}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}$$

where $\text{se}(\hat{\beta}_0) = S\sqrt{{}^1\!/_n + {}^{\bar{x}^2}\!/_{SXX}}$ is the estimated standard error of $\hat{\beta}_0$, which is given directly by R. In the production example the intercept is called *Intercept* and so $\text{se}(\hat{\beta}_0) = 8.32815$.

For **testing the hypothesis** $H_0 : \beta_0 = \beta_0^0$ the test statistic is

$$T = \frac{\hat{\beta}_0 - \beta_0^0}{\operatorname{se}(\hat{\beta}_0)} \sim t_{n-2} \text{when } H_0 \text{ is true.}$$

R provides the value of T and the p-value associated with testing $H_0 : \beta_0 = 0$ against $H_A : \beta_0 \neq 0$. In the production example the intercept is called *Intercept* and T = 17.98 which results in a p-value < 0.0001.

A 100(1 $-\alpha$)% **confidence interval** for β_0, the intercept of the regression line, is given by

$$(\hat{\beta}_0 - t(\alpha/2, n-2)\operatorname{se}(\hat{\beta}_0), \hat{\beta}_0 + t(\alpha/2, n-2)\operatorname{se}(\hat{\beta}_0))$$

where $t(\alpha/2, n-2)$ is the 100(1–α/2)th quantile of the t-distribution with n – 2 degrees of freedom.

In the production example,

$$\hat{\beta}_0 = 149.7477, \operatorname{se}(\hat{\beta}_0) = 8.32815, t(0.025, 20-2=18) = 2.1009.$$

Thus a 95% confidence interval for β_0 is given by

$$(149.7477 \pm 2.1009 \times 8.32815) = (149.748 \pm 17.497) = (132.3, 167.2)$$

Regression Output from R: 95% confidence intervals

```
                 2.5%    97.5%
(Intercept) 132.251  167.244
RunSize       0.181    0.337
```

2.3 Confidence Intervals for the Population Regression Line

In this section we consider the problem of finding a confidence interval for the unknown population regression line at a given value of X, which we shall denote by x^*. First, recall from (2.1) that the population regression line at $X = x^*$ is given by

$$\mathrm{E}(Y \mid X = x^*) = \beta_0 + \beta_1 x^*$$

An estimator of this unknown quantity is the value of the estimated regression equation at $X = x^*$, namely,

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Under the assumptions stated previously, it can be shown that

$$\mathrm{E}(\hat{y}^*) = \mathrm{E}(\hat{y} \mid X = x^*) = \beta_0 + \beta_1 x^* \quad (2.12)$$

$$\operatorname{Var}(\hat{y}^*) = \operatorname{Var}(\hat{y} \mid X = x^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX} \right) \tag{2.13}$$

$$\hat{y}^* = \hat{y} \mid X = x^* \sim N\left(\beta_0 + \beta_1 x^*, \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX} \right) \right) \tag{2.14}$$

Standardizing (2.14) gives

$$Z = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{\sigma \sqrt{(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX})}} \sim N(0,1)$$

Replacing σ by S results in

$$T = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{S \sqrt{(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX})}} \sim t_{n-2}$$

A 100(1 – α)% **confidence interval** for $\mathrm{E}(Y \mid X = x^*) = \beta_0 + \beta_1 x^*$, the population regression line at $X = x^*$, is given by

$$\begin{aligned} &\hat{y}^* \pm t(\alpha/2, n-2) S \sqrt{(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX})} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t(\alpha/2, n-2) S \sqrt{(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX})} \end{aligned}$$

where $t(\alpha/2, n-2)$ is the 100(1–α/2)th quantile of the *t*-distribution with n – 2 degrees of freedom.

2.4 Prediction Intervals for the Actual Value of *Y*

In this section we consider the problem of finding a prediction interval for the actual value of *Y* at *x**, a given value of *X*.

Important Notes:

1. $\mathrm{E}(Y \mid X = x^*)$, the expected value or average value of *Y* for a given value *x** of *X*, is what one would expect *Y* to be in the long run when *X* = *x**. $\mathrm{E}(Y \mid X = x^*)$ is therefore a fixed but unknown quantity whereas *Y* can take a number of values when *X* = *x**.

2. $\mathrm{E}(Y \mid X = x^*)$, the value of the regression line at $X = x^*$, is entirely different from Y^*, a single value of Y when $X = x^*$. In particular, Y^* need not lie on the population regression line.
3. A **confidence interval** is always reported for a **parameter** (e.g., $\mathrm{E}(Y|X = x^*) = \beta_0 + \beta_1 x^*$) and a ***prediction interval*** is reported for the value of a ***random variable*** (e.g., Y^*).

We base our prediction of Y when $X = x^*$ (that is of Y^*) on

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

The error in our prediction is

$$Y^* - \hat{y}^* = \beta_0 + \beta_1 x^* + e^* - \hat{y}^* = \mathrm{E}(Y \mid X = x^*) - \hat{y}^* + e^*$$

that is, the deviation between $\mathrm{E}(Y \mid X = x^*)$ and $\hat{y}^*$ plus the random fluctuation e^* (which represents the deviation of Y^* from $\mathrm{E}(Y|X = x^*)$). Thus the variability in the error for predicting a single value of Y will exceed the variability for estimating the expected value of Y (because of the random error e^*).

It can be shown that under the previously stated assumptions that

$$\mathrm{E}(Y^* - \hat{y}^*) = \mathrm{E}(Y - \hat{y} \mid X = x^*) = 0 \tag{2.15}$$

$$\mathrm{Var}(Y^* - \hat{y}^*) = \mathrm{Var}(Y - \hat{y} \mid X = x^*) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right] \tag{2.16}$$

$$Y^* - \hat{y}^* \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right]\right) \tag{2.17}$$

Standardizing (2.17) and replacing σ by S gives

$$T = \frac{Y^* - \hat{y}^*}{S\sqrt{(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX})}} \sim t_{n-2}$$

A $100(1-\alpha)\%$ **prediction interval** for Y^*, the value of Y at $X = x^*$, is given by

$$\begin{aligned} &\hat{y}^* \pm t(\alpha/2, n-2) S\sqrt{(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX})} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t(\alpha/2, n-2) S\sqrt{(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX})} \end{aligned}$$

where $t(\alpha/2,n{-}2)$ is the $100(1{-}\alpha/2)$th quantile of the *t*-distribution with $n - 2$ degrees of freedom.

Regression Output from R

Ninety-five percent confidence intervals for the population regression line (i.e., the average *RunTime)* at *RunSize* = 50, 100, 150, 200, 250, 300, 350 are:

```
       fit      lwr      upr
1 162.7099 148.6204 176.7994
2 175.6720 164.6568 186.6872
3 188.6342 179.9969 197.2714
4 201.5963 193.9600 209.2326
5 214.5585 206.0455 223.0714
6 227.5206 216.7006 238.3407
7 240.4828 226.6220 254.3435
```

Ninety-five percent prediction intervals for the actual value of *Y* (i.e., the actual *RunTime)* at at *RunSize* = 50, 100, 150, 200, 250, 300, 350 are:

```
       fit      lwr      upr
1 162.7099 125.7720 199.6478
2 175.6720 139.7940 211.5500
3 188.6342 153.4135 223.8548
4 201.5963 166.6076 236.5850
5 214.5585 179.3681 249.7489
6 227.5206 191.7021 263.3392
7 240.4828 203.6315 277.3340
```

Notice that each prediction interval is considerably wider than the corresponding confidence interval, as is expected.

2.5 Analysis of Variance

There is a linear association between *Y* and *x* if

$$Y = \beta_0 + \beta_1 x + e$$

and $\beta_1 \neq 0$. If we knew that $\beta_1 \neq 0$ then we would predict *Y* by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

On the other hand, if we knew that $\beta_1 = 0$ then we predict *Y* by

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

To test whether there is a linear association between *Y* and *X* we have to test

$$H_0 : \beta_1 = 0 \text{ against } H_A : \beta_1 \neq 0 .$$

We can perform this test using the following *t*-statistic

$$T = \frac{\hat{\beta}_1 - 0}{\operatorname{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

We next look at a different test statistic which can be used when there is more than one predictor variable, that is, in multiple regression. First, we introduce some terminology.

Define the total corrected sum of squares of the *Y*'s by

$$\text{SST} = SYY = \sum_i^n (y_i - \overline{y})^2$$

Recall that the residual sum of squares is given by

$$\text{RSS} = \sum_i^n (y_i - \hat{y}_i)^2$$

Define the regression sum of squares (i.e., sum of squares explained by the regression model) by

$$\text{SSreg} = \sum_i^n (\hat{y}_i - \overline{y})^2$$

It is clear that SSreg is close to zero if for each *i*, $\hat{y}_i$ is close to $\bar{y}$ while SSreg is large if $\hat{y}_i$ differs from $\bar{y}$ for most values of *x*.

We next look at the hypothetical situation in Figure 2.4 with just a single data point (x_i, y_i) shown along with the least squares regression line and the mean of *y* based on all *n* data points. It is apparent from Figure 2.4 that $y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$.

Further, it can be shown that

SST	= SSreg	+ RSS
Total sample	= Variability explained by	+ Unexplained (or error)
variability	the model	variability

See exercise 6 in Section 2.7 for details.

If

$$Y = \beta_0 + \beta_1 x + e \text{ and } \beta_1 \neq 0$$

then RSS should be "small" and SSreg should be "close" to SST. But how small is "small" and how close is "close"?

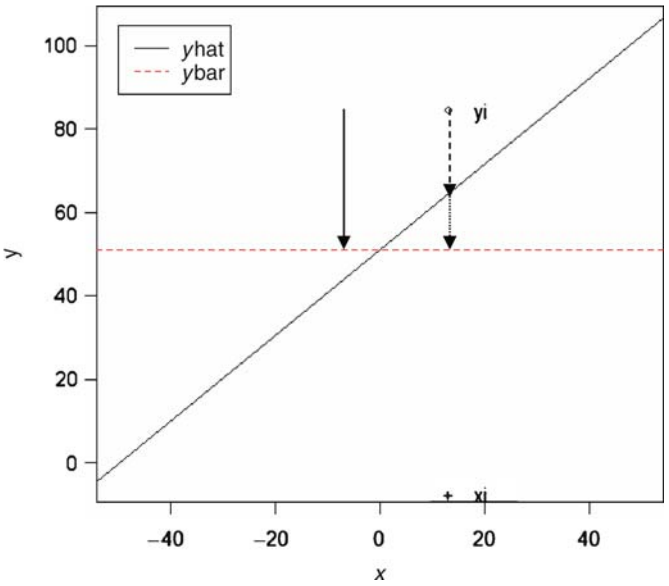

Figure 2.4 Graphical depiction that $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

To test

$$H_0 : \beta_1 = 0 \text{ against } H_A : \beta_1 \neq 0$$

we can use the test statistic

$$F = \frac{\text{SSreg}/1}{\text{RSS}/(n-2)}$$

since RSS has (n – 2) degrees of freedom and SSreg has 1 degree of freedom.

Under the assumption that $e_1, e_2, \ldots, e_n$ are independent and normally distributed with mean 0 and variance σ^2, it can be shown that F has an F distribution with 1 and n – 2 degrees of freedom when H_0 is true, that is,

$$F = \frac{\text{SSreg}/1}{\text{RSS}/(n-2)} \sim F_{1,n-2} \text{ when } H_0 \text{ is true}$$

Form of test: reject H_0 at level α if $F > F_{\alpha,1,n-2}$ (which can be obtained from table of the F distribution). However, all statistical packages report the corresponding p-value.

The usual way of setting out this test is to use an Analysis of variance table

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Regression	1	SSreg	SSreg/1	$F = \dfrac{\text{SSreg}/1}{\text{RSS}/(n-2)}$
Residual	$n - 2$	RSS	RSS/$(n - 2)$	
Total	$n - 1$	SST		

Notes:

1. It can be shown that in the case of simple linear regression $T = \dfrac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$ and $F = \dfrac{\text{SSreg}/1}{\text{RSS}/(n-2)} \sim F_{1,n-2}$ are related via $F = T^2$
2. R^2, the coefficient of determination of the regression line, is defined as the proportion of the total sample variability in the Y's explained by the regression model, that is,

$$R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$$

The reason this quantity is called R^2 is that it is equal to the square of the correlation between Y and X. It is arguably one of the most commonly misused statistics.

Regression Output from R

```
Analysis of Variance Table
Response: RunTime
           Df   Sum Sq   Mean Sq   F value     Pr(>F)
RunSize     1  12868.4   12868.4    48.717  1.615e-06   ***
Residuals  18   4754.6     264.1
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Notice that the observed F-value of 48.717 is just the square of the observed t-value 6.98 which can be found between Figures 2.2 and 2.3. We shall see in Chapter 5 that Analysis of Variance overcomes the problems associated with multiple t-tests which occur when there are many predictor variables.

2.6 Dummy Variable Regression

So far we have only considered situations in which the predictor or X-variable is quantitative (i.e., takes numerical values). We next consider so-called **dummy variable regression**, which is used in its simplest form when a predictor is categorical with two values (e.g., gender) rather than quantitative. The resulting regression models allow us to test for the difference between the means of two groups. We shall see in a later topic that the concept of a dummy variable can be extended to include problems involving more than two groups.

Using dummy variable regression to compare new and old methods

We shall consider the following example throughout this section. It is taken from Foster, Stine and Waterman (1997, pages 142–148). In this example, we consider a large food processing center that needs to be able to switch from one type of package to another quickly to react to changes in order patterns. Consultants have developed a new method for changing the production line and used it to produce a sample of 48 change-over times (in minutes). Also available is an independent sample of 72 change-over times (in minutes) for the existing method. These two sets of times can be found on book web site in the file called changeover_times.txt. The first three and the last three rows of the data from this file are reproduced below in Table 2.2. Plots of the data appear in Figure 2.5.

We wish to develop an equation to model the relationship between Y, the change-over time and X, the dummy variable corresponding to New and hence test whether the mean change-over time is reduced using the new method.

We consider the simple linear regression model

$$Y = \beta_0 + \beta_1 x + e$$

where Y = change-over time and x is the dummy variable (i.e., $x = 1$ if the time corresponds to the new change-over method and 0 if it corresponds to the existing method).

Regression Output from R

```
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept) 17.8611       0.8905     20.058      <2e-16    ***
New         -3.1736       1.4080     -2.254      0.0260    *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.556 on 118 degrees of freedom
Multiple R-Squared: 0.04128, Adjusted R-squared: 0.03315
F-statistic: 5.081 on 1 and 118 DF, p-value: 0.02604
```

We can test whether there is significant reduction in the change-over time for the new method by testing the significance of the dummy variable, that is, we wish to test whether the coefficient of x is zero or less than zero, that is:

$$H_0 : \beta_1 = 0 \text{ against } H_A : \beta_1 < 0$$

We use the one-sided "<" alternative since we are interested in whether the new method has lead to a reduction in mean change-over time. The test statistic is

$$T = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

Table 2.2 Change-over time data (changeover_times.txt)

Method	Y, Change-over time	X, New
Existing	19	0
Existing	24	0
Existing	39	0
.	.	.
New	14	1
New	40	1
New	35	1

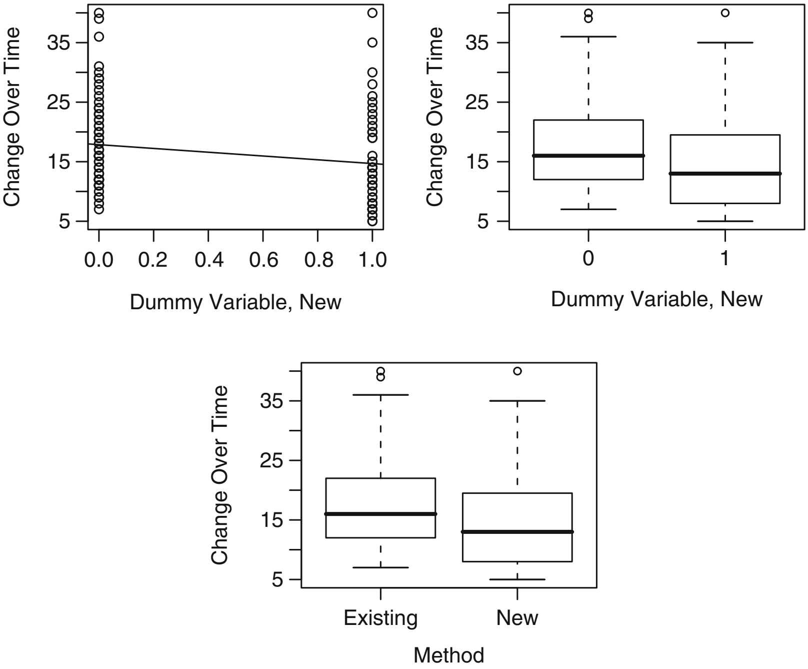

Figure 2.5 A scatter plot and box plots of the change-over time data

In this case,

$$T = -2.254.$$

(This result can be found in the output in the column headed 't value'). The associated p-value is given by

$$p - value = P(T < -2.254 \text{ when } H_0 \text{ is true}) = \frac{0.026}{2} = 0.013$$

as the two-sided $p\text{-}value = P(T \neq -2.254 \text{ when } H_0 \text{ is true}) = 0.026.$

This means that there is significant evidence of a reduction in the mean change-over time for the new method.

Next consider the group consisting of those times associated with the **new change-over method**. For this group, the dummy variable, *x* is equal to 1. Thus, we can estimate the mean change-over time for the new method as:

$$17.8611 + (-3.1736) \times 1 = 14.6875 = 14.7 \text{ minutes}$$

Next consider the group consisting of those times associated with the **existing change-over method**. For this group, the dummy variable, *x* is equal to 0. Thus, we can estimate the mean change-over time for the new method as:

$$17.8611 + (-3.1736) \times 0 = 17.8611 = 17.9 \text{ minutes}$$

The new change-over method produces a reduction in the mean change-over time of 3.2 min from 17.9 to 14.7 minutes (Notice that the reduction in the mean change-over time for the new method is just the coefficient of the dummy variable.) This reduction is **statistically significant**.

A 95% confidence interval for the reduction in mean change-over time due to the new method is given by

$$(\hat{\beta}_1 - t(\alpha/2, n-2)\mathrm{se}(\hat{\beta}_1), \hat{\beta}_1 + t(\alpha/2, n-2)\mathrm{se}(\hat{\beta}_1))$$

where $t(\alpha/2, n-2)$ is the 100(1$-\alpha/2$)th quantile of the *t*-distribution with $n-2$ degrees of freedom. In this example the *X*-variable is the dummy variable *New* and $\hat{\beta}_1 = -3.1736, \mathrm{se}(\hat{\beta}_1) = 1.4080, t(0.025, 120-2 = 118) = 1.9803.$ Thus a 95% confidence interval for β_1 (in minutes) is given by

$$(-3.1736 \pm 1.9803 \times 1.4080) = (-3.1736 \pm 2.7883) = (-5.96, -0.39).$$

Finally, the company should adopt the new method if a reduction of time of this size is of **practical significance**.

2.7 Derivations of Results

In this section, we shall derive some results given earlier about the least squares estimates of the slope and the intercept as well as results about confidence intervals and prediction intervals.

Throughout this section we shall make the following assumptions:

1. *Y* is related to *x* by the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i \ (i = 1, \ldots, n), i.e., \mathrm{E}(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$
2. The errors $e_1, e_2, \ldots, e_n$ are independent of each other
3. The errors $e_1, e_2, \ldots, e_n$ have a common variance σ^2
4. The errors are normally distributed with a mean of 0 and variance σ^2(especially when the sample size is small), that is, $e \mid X \sim N(0, \sigma^2)$

In addition, since the regression model is conditional on X we can assume that the values of the predictor variable, $x_1, x_2, \ldots, x_n$ are known fixed constants.

2.7.1 Inferences about the Slope of the Regression Line

Recall from (2.5) that the least squares estimate of β_1 is given by

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \text{ where } c_i = \frac{x_i - \bar{x}}{SXX}.$$

Under the above assumptions we shall derive (2.6), (2.7) and (2.8).

To derive (2.6) let's consider

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1 \mid X) &= \mathrm{E}\left[\sum_{i=1}^{n} c_i y_i \mid X = x_i\right] \\
&= \sum_{i=1}^{n} c_i E\left[y_i \mid X = x_i\right] \\
&= \sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 x_i\right) \\
&= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \\
&= \beta_0 \sum_{i=1}^{n} \left\{\frac{x_i - \bar{x}}{SXX}\right\} + \beta_1 \sum_{i=1}^{n} \left\{\frac{x_i - \bar{x}}{SXX}\right\} x_i \\
&= \beta_1
\end{aligned}$$

since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = SXX.$

To derive (2.7) let's consider

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1 \mid X) &= \mathrm{Var}\left[\sum_{i=1}^{n} c_i y_i \mid X = x_i\right] \\
&= \sum_{i=1}^{n} c_i^2 \mathrm{Var}(y_i \mid X = x_i) \\
&= \sigma^2 \sum_{i=1}^{n} c_i^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left\{\frac{x_i - \bar{x}}{SXX}\right\}^2 \\
&= \frac{\sigma^2}{SXX}
\end{aligned}$$

Finally we derive (2.8). Under assumption (4), the errors $e_i|X$ are normally distributed. Since $y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1,2,...,n)$, $Y_i|X$ is normally distributed. Since $\hat{\beta}_1 \mid X$ is a linear combination of the y_i's, $\hat{\beta}_1 \mid X$ is normally distributed.

2.7.2 Inferences about the Intercept of the Regression Line

Recall from (2.3) that the least squares estimate of β_0 is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Under the assumptions given previously we shall derive (2.9), (2.10) and (2.11). To derive (2.9) we shall use the fact that

$$\mathrm{E}(\hat{\beta}_0 \mid X) = \mathrm{E}(\bar{y} \mid X) - \mathrm{E}(\hat{\beta}_1 \mid X)\bar{x}$$

The first piece of the last equation is

$$\begin{aligned}\mathrm{E}(\bar{y} \mid X) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(y_i \mid X = x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\beta_0 + \beta_1 x_i + e_i) \\ &= \beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \beta_0 + \beta_1 \bar{x}\end{aligned}$$

The second piece of that equation is

$$\mathrm{E}(\hat{\beta}_1 \mid X)\bar{x} = \beta_1 \bar{x}.$$

Thus,

$$\mathrm{E}(\hat{\beta}_0 \mid X) = \mathrm{E}(\bar{y} \mid X) - \mathrm{E}(\hat{\beta}_1 \mid X)\bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$$

To derive (2.10) let's consider

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0 \mid X) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1 \bar{x} \mid X) \\ &= \mathrm{Var}(\bar{y} \mid X) + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1 \mid X) - 2\bar{x}\mathrm{Cov}(\bar{y}, \hat{\beta}_1 \mid X)\end{aligned}$$

The first term is given by

$$\mathrm{Var}(\bar{y} \mid X) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} y_i \mid X = x_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

From (2.7),

$$\mathrm{Var}(\hat{\beta}_1 \mid X) = \frac{\sigma^2}{SXX}$$

Finally,

$$\mathrm{Cov}(\bar{y}, \hat{\beta}_1 \mid X) = \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} c_i y_i\right) = \frac{1}{n}\sum_{i=1}^{n} c_i \mathrm{Cov}(y_i, y_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0$$

So,

$$\mathrm{Var}(\hat{\beta}_0 \mid X) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$$

Result (2.11) follows from the fact that under assumption (4), $Y_i \mid X$ (and hence ȳ) are normally distributed as is $\hat{\beta}_1 \mid X$.

2.7.3 Confidence Intervals for the Population Regression Line

Recall that the population regression line at $X = x^*$ is given by

$$\mathrm{E}(Y \mid X = x^*) = \beta_0 + \beta_1 x^*$$

An estimator the population regression line at $X = x^*$ (i.e., $\mathrm{E}(Y \mid X = x^*) = \beta_0 + \beta_1 x^*$) is the value of the estimated regression equation at $X = x^*$, namely,

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Under the assumptions stated previously, we shall derive (2.12), (2.13) and (2.14). First, notice that (2.12) follows from the following earlier established results $\mathrm{E}(\hat{\beta}_0 \mid X = x^*) = \beta_0$ and $\mathrm{E}(\hat{\beta}_1 \mid X = x^*) = \beta_1$.

Next, consider (2.13)

$$\begin{aligned}
&\mathrm{Var}(\hat{y} \mid X = x^*) \\
&= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x \mid X = x^*) \\
&= \mathrm{Var}(\hat{\beta}_0 \mid X = x^*) + x^{*2}\,\mathrm{Var}(\hat{\beta}_1 \mid X = x^*) + 2x^*\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 \mid X = x^*)
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 \mid X = x^*) &= \mathrm{Cov}(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1 \mid X = x^*) \\
&= \mathrm{Cov}(\bar{y}, \hat{\beta}_1 \mid X = x^*) - \bar{x}\,\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_1) \\
&= 0 - \bar{x}\,\mathrm{Var}(\hat{\beta}_1) \\
&= \frac{-\bar{x}\sigma^2}{SXX}
\end{aligned}$$

So that,

$$\begin{aligned}\operatorname{Var}(\hat{y} \mid X = x^*) &= \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{SXX}\right)+x^{*2}\frac{\sigma^2}{SXX}-\frac{2x^*\bar{x}\sigma^2}{SXX}\\ &= \sigma^2\left(\frac{1}{n}+\frac{(x^*-\bar{x})^2}{SXX}\right)\end{aligned}$$

Result (2.14) follows from the fact that under assumption (4), $\hat{\beta}_0 \mid X$ is normally distributed as is $\hat{\beta}_1 \mid X$.

2.7.4 Prediction Intervals for the Actual Value of Y

We base our prediction of Y when $X = x^*$ (that is of Y^*) on

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

The error in our prediction is

$$Y^* - \hat{y}^* = \beta_0 + \beta_1 x^* + e^* - \hat{y}^* = \mathrm{E}(Y \mid X = x^*) - \hat{y}^* + e^*$$

that is, the deviation between $\mathrm{E}(Y \mid X = x^*)$ and $\hat{y}^*$ plus the random fluctuation e^* (which represents the deviation of Y^* from $\mathrm{E}(Y \mid X = x^*)$).

Under the assumptions stated previously, we shall derive (2.15), (2.16) and (2.17). First, we consider (2.15)

$$\begin{aligned}\mathrm{E}(Y^* - \hat{y}^*) &= \mathrm{E}(Y - \hat{y} \mid X = x^*)\\ &= \mathrm{E}(Y \mid X = x^*) - \mathrm{E}(\hat{\beta}_0 + \hat{\beta}_1 x \mid X = x^*)\\ &= 0\end{aligned}$$

In considering (2.16), notice that $\hat{y}$ is independent of Y^*, a future value of Y. Thus,

$$\begin{aligned}\operatorname{Var}(Y^* - \hat{y}^*) &= \operatorname{Var}(Y - \hat{y} \mid X = x^*)\\ &= \operatorname{Var}(Y \mid X = x^*) + \operatorname{Var}(\hat{y} \mid X = x^*) - 2\operatorname{Cov}(Y, \hat{y} \mid X = x^*)\\ &= \sigma^2 + \sigma^2\left[\frac{1}{n}+\frac{(x^*-\bar{x})^2}{SXX}\right] - 0\\ &= \sigma^2\left[1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{SXX}\right]\end{aligned}$$

Finally, (2.17) follows since both ŷ and Y^* are normally distributed.

2.8 Exercises

1. The web site www.playbill.com provides weekly reports on the box office ticket sales for plays on Broadway in New York. We shall consider the data for the week October 11–17, 2004 (referred to below as the current week). The data are in the form of the gross box office results for the current week and the gross box office results for the previous week (i.e., October 3–10, 2004). The data, plotted in Figure 2.6, are available on the book web site in the file playbill.csv.

   Fit the following model to the data: $Y = \beta_0 + \beta_1 x + e$ where Y is the gross box office results for the current week (in \$) and x is the gross box office results for the previous week (in \$). Complete the following tasks:

   (a) Find a 95% confidence interval for the slope of the regression model, β_1. Is 1 a plausible value for β_1? Give a reason to support your answer.
   (b) Test the null hypothesis $H_0 : \beta_0 = 10000$ against a two-sided alternative. Interpret your result.
   (c) Use the fitted regression model to estimate the gross box office results for the current week (in \$) for a production with \$400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office

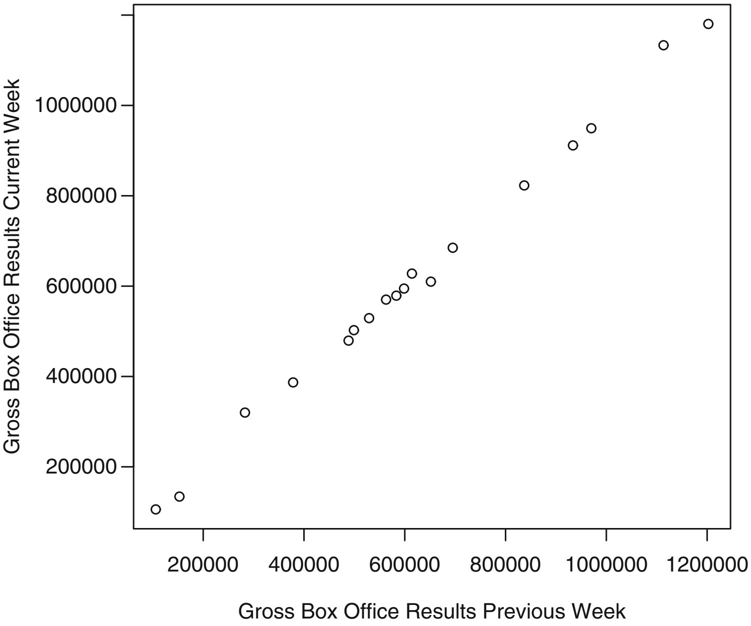

Figure 2.6 Scatter plot of gross box office results from Broadway

results for the current week (in \$) for a production with \$400,000 in gross box office the previous week. Is \$450,000 a feasible value for the gross box office results in the current week, for a production with \$400,000 in gross box office the previous week? Give a reason to support your answer.

(d) Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this week's gross box office results. Comment on the appropriateness of this rule.

2. A story by James R. Hagerty entitled *With Buyers Sidelined, Home Prices Slide* published in the Thursday October 25, 2007 edition of the *Wall Street Journal* contained data on so-called fundamental housing indicators in major real estate markets across the US. The author argues that… *prices are generally falling and overdue loan payments are piling up*. Thus, we shall consider data presented in the article on

   Y = Percentage change in average price from July 2006 to July 2007 (based on the S&P/Case-Shiller national housing index); and
   x = Percentage of mortgage loans 30 days or more overdue in latest quarter (based on data from Equifax and Moody's).

   The data are available on the book web site in the file indicators.txt. Fit the following model to the data: $Y = \beta_0 + \beta_1 x + e$. Complete the following tasks:

   (a) Find a 95% confidence interval for the slope of the regression model, β_1. On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.
   (b) Use the fitted regression model to estimate $\mathrm{E}(Y|X{=}4)$. Find a 95% confidence interval for $\mathrm{E}(Y|X{=}4)$. Is 0% a feasible value for $\mathrm{E}(Y|X{=}4)$? Give a reason to support your answer.

3. The manager of the purchasing department of a large company would like to develop a regression model to predict the average amount of time it takes to process a given number of invoices. Over a 30-day period, data are collected on the number of invoices processed and the total time taken (in hours). The data are available on the book web site in the file invoices.txt. The following model was fit to the data: $Y = \beta_0 + \beta_1 x + e$ where Y is the processing time and x is the number of invoices. A plot of the data and the fitted model can be found in Figure 2.7. Utilizing the output from the fit of this model provided below, complete the following tasks.

   (a) Find a 95% confidence interval for the start-up time, i.e., β_0.
   (b) Suppose that a best practice benchmark for the average processing time for an additional invoice is 0.01 hours (or 0.6 minutes). Test the null hypothesis $H_0 : \beta_1 = 0.01$ against a two-sided alternative. Interpret your result.
   (c) Find a point estimate and a 95% prediction interval for the time taken to process 130 invoices.

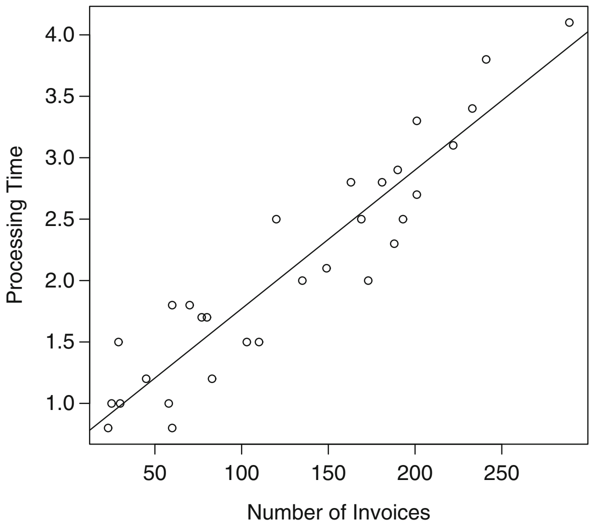

Figure 2.7 Scatter plot of the invoice data

Regression output from R for the invoice data

```
Call:
lm(formula = Time ~ Invoices)

Coefficients:
             Estimate  Std. Error  t value    Pr(>|t|)
(Intercept) 0.6417099   0.1222707    5.248    1.41e-05  ***
Invoices    0.0112916   0.0008184   13.797    5.17e-14  ***
---
Residual standard error: 0.3298 on 28 degrees of freedom
Multiple R-Squared: 0.8718, Adjusted R-squared: 0.8672
F-statistic: 190.4 on 1 and 28 DF, p-value: 5.175e-14

mean(Time)
2.1
median(Time)
2
mean(Invoices)
130.0
median(Invoices)
127.5
```

4. Straight-line regression through the origin:
   In this question we shall make the following assumptions:

   (1) Y is related to x by the simple linear regression model $Y_i = \beta x_i + e_i$ $(i = 1, 2, \ldots, n)$, i.e., $\mathrm{E}(Y \mid X = x_i) = \beta x_i$

(2) The errors $e_1, e_2,..., e_n$ are independent of each other

(3) The errors $e_1, e_2,..., e_n$ have a common variance σ^2

(4) The errors are normally distributed with a mean of 0 and variance σ^2 (especially when the sample size is small), i.e., $e \mid X \sim N(0, \sigma^2)$

In addition, since the regression model is conditional on X we can assume that the values of the predictor variable, $x_1, x_2, \ldots, x_n$ are known fixed constants.

(a) Show that the least squares estimate of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(b) Under the above assumptions show that

(i) $\mathrm{E}(\hat{\beta} \mid X) = \beta$

(ii) $\mathrm{Var}(\hat{\beta} \mid X) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

(iii) $\hat{\beta} \mid X \sim N(\beta, \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2})$

5. Two alternative straight line regression models have been proposed for Y. In the first model, Y is a linear function of x_1, while in the second model Y is a linear function of x_2. The plot in the first column of Figure 2.8 is that of Y against x_1, while the plot in the second column below is that of Y against x_2. These plots also show the least squares regression lines. In the following statements RSS stands for residual sum of squares, while SSreg stands for regression sum of squares. Which one of the following statements is true?

(a) RSS for model 1 is greater than RSS for model 2, while SSreg for model 1 is greater than SSreg for model 2.

(b) RSS for model 1 is less than RSS for model 2, while SSreg for model 1 is less than SSreg for model 2.

(c) RSS for model 1 is greater than RSS for model 2, while SSreg for model 1 is less than SSreg for model 2.

(d) RSS for model 1 is less than RSS for model 2, while SSreg for model 1 is greater than SSreg for model 2.

Give a detailed reason to support your choice.

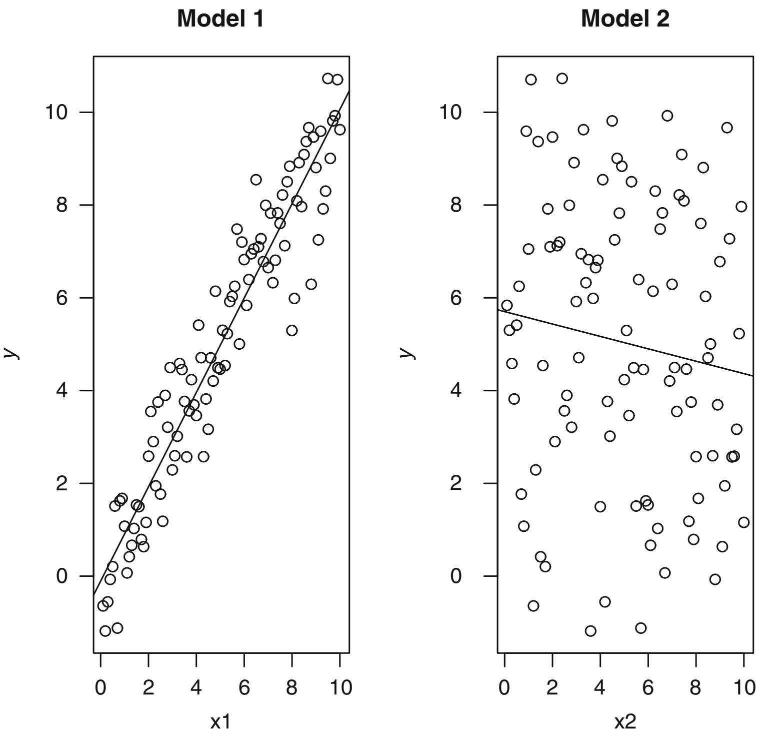

Figure 2.8 Scatter plots and least squares lines

6. In this problem we will show that SST=SSreg+RSS . To do this we will show that $\sum_{i=1}^{n}(y_i - \hat{y}_i)\,(\hat{y}_i - \overline{y}) = 0.$

   (a) Show that $(y_i - \hat{y}_i) = (y_i - \overline{y}) - \hat{\beta}_1(x_i - \overline{x})$.

   (b) Show that $(\hat{y}_i - \overline{y}) = \hat{\beta}_1(x_i - \overline{x})$.

   (c) Utilizing the fact that $\hat{\beta}_1 = \dfrac{SXY}{SXX}$, show that $\sum_{i=1}^{n}(y_i - \hat{y}_i)\,(\hat{y}_i - \overline{y}) = 0.$

7. A statistics professor has been involved in a collaborative research project with two entomologists. The statistics part of the project involves fitting regression models to large data sets. Together they have written and submitted a manuscript to an entomology journal. The manuscript contains a number of scatter plots with each showing an estimated regression line (based on a valid model) and

associated individual 95% confidence intervals for the regression function at each x value, as well as the observed data. A referee has asked the following question:

> I don't understand how 95% of the observations fall outside the 95% CI as depicted in the figures.

Briefly explain how it is entirely possible that 95% of the observations fall outside the 95% CI as depicted in the figures.