## Chapter 11 Linear REGRESSION

Bibliography: Sheldon Ross

### 11.1 INTRODUCTION

Many physical problems are concerned with determining a relationship within a set of variables. For instance, we might be interested in the relationship between the force for stretching a spring, its temperature and the distance that the spring stretches (Hooke's law). Knowledge of such a relationship would enable us to predict the output for various values of force and temperature.

In many situations, there is a single *response* variable *Y*, also called the *dependent* variable, which depends on the value of a set of *input*, also called *independent*, variables $x_1, x_2, \ldots, x_r$. The simplest type of relationship between the dependent variable *Y* and the input variables $x_1, x_2, \ldots, x_r$ is a linear relationship. That is, for some constants $\beta_0, \beta_1, \ldots, \beta_r$, the equation

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r \tag{11.1}$$

would hold. If this was the relationship between *Y* and the $x_i$, $i = 1, \ldots, r$, then it would be possible (once the $\beta_i$ were learned) to exactly predict the response for any set of input values. However, in practice, such precision is almost never attainable, and the most that one can expect is that Eq (11.1) would be valid *subject to random error*. By this we mean that the explicit relationship is

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r + e \tag{11.2}$$

where *e*, representing the *random error*, is assumed to be a random variable having mean 0. It is convenient to view the input *x* as *controlled by the data analyst* and measured with negligible error, while the response *y* is a *random variable*. That is, there is a *probability distribution* for *y* at each possible value for *x*. The mean of this distribution is:

$$E\left[Y \middle| x\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r \tag{11.3}$$

where $x = (x_1, x_2, \ldots, x_r)$ is the set of independent variables, and $E\left[Y \middle| x\right]$ is the expected response given the inputs *x*.

Eq (11.2) is called a *linear regression equation*. The quantities $\beta_0, \beta_1, \ldots, \beta_r$ are called the *regression coefficients*, and must usually be estimated from a set of data.

A regression equation containing a single independent variable — that is, one in which $r = 1$ — is called a *simple regression equation*, whereas one containing many independent variables is called a *multiple regression equation*.

Thus, a *simple* linear regression model supposes a linear relationship between the mean response and the value of a single independent variable:

$$Y = \alpha + \beta x + e \tag{11.4}$$

where *x* is the value of the independent variable, also called the input level, *Y* is the response, and *e*, representing the random error, is a random variable having mean 0.

**Example 1** Consider the following 10 data pairs $(x_i, y_i)$, $i = 1, 2, \ldots, 10$, relating *y*, the amount of spring stretch (mm) of a laboratory experiment, to *x*, the force on spring (Newtons) at which the experiment was run.

| i | $x_i$ | $y_i$ | i | $x_i$ | $y_i$ |
|---|---|---|---|---|---|
| 1 | 1.00 | 45 | 6 | 1.50 | 68 |
| 2 | 1.10 | 52 | 7 | 1.60 | 75 |
| 3 | 1.20 | 54 | 8 | 1.70 | 76 |
| 4 | 1.30 | 63 | 9 | 1.80 | 92 |
| 5 | 1.40 | 62 | 10 | 1.90 | 88 |

A plot of $y_i$ versus $x_i$ — called a *scatter diagram* — is given in Figure 11.1. This scatter diagram reflect a linear relation between *y* and *x*. It seems that a simple linear regression model would be appropriate.

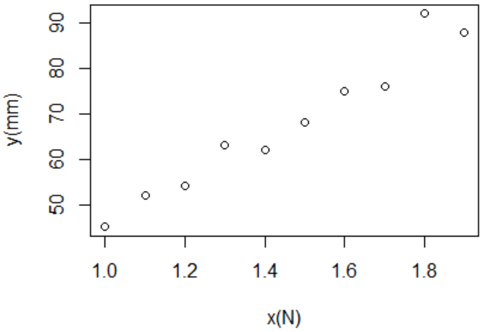

Figure 11.1 Scatter diagram

## 11.2 LEAST SQUARES ESTIMATORS OF THE REGRESSION PARAMETERS

Suppose that the responses $Y_i$ corresponding to the input values $x_i$, $i=1,2,\dots,n$ are to be observed and used to estimate α and β in a simple linear regression model. To determine estimators of α and β we reason as follows: If $\hat{\alpha}$ is the estimator of α and $\hat{\beta}$ of β, then the estimator of the response corresponding to the input variable $x_i$ would be $\hat{\alpha}+\hat{\beta}x_i$. Since the actual response is $Y_i$, the squared difference is $\left(Y_i-\hat{\alpha}-\hat{\beta}x_i\right)^2$, and so if $\hat{\alpha}$ and $\hat{\beta}$ are the estimators of α and β, then the *sum of the squared differences between the estimated responses and the actual response* values—call it $SS$ — is given by

$$SS=\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}x_i\right)^2 \qquad (11.5)$$

The *method of least squares* chooses as estimators of α and β the values of $\hat{\alpha}$ and $\hat{\beta}$ that *minimize* $SS$. To determine these estimators, we differentiate $SS$ first with respect to $\hat{\alpha}$ and then to $\hat{\beta}$ as follows:

$$\frac{\partial SS}{\partial\hat{\alpha}}=-2\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}x_i\right) \qquad (11.6)$$

$$\frac{\partial SS}{\partial\hat{\beta}}=-2\sum_{i=1}^{n}x_i\left(Y_i-\hat{\alpha}-\hat{\beta}x_i\right) \qquad (11.7)$$

Setting these partial derivatives equal to zero yields the following equations for values $\hat{\alpha}$ and $\hat{\beta}$:

$$\sum_{i=1}^{n}Y_i=n\hat{\alpha}+\hat{\beta}\sum_{i=1}^{n}x_i \qquad (11.8)$$

$$\sum_{i=1}^{n}x_iY_i=\hat{\alpha}\sum_{i=1}^{n}x_i+\hat{\beta}\sum_{i=1}^{n}x_i^2 \qquad (11.9)$$

If we let

$$\bar{Y}=\sum_{i=1}^{n}Y_i/n \quad , \qquad \bar{x}=\sum_{i=1}^{n}x_i/n$$

we can write the first equation as

$$\bar{Y}=\hat{\alpha}+\hat{\beta}\bar{x} \qquad (11.10)$$

which means that the point defined by $\left(\bar{x},\bar{Y}\right)$ is located on the *estimated regression line*.

Moreover,

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x} \tag{11.11}$$

Substituting this value of $\hat{\alpha}$ into the second equation (11.9) yields

$$\sum_{i=1}^{n} x_i Y_i = \left(\overline{Y} - \hat{\beta}\overline{x}\right) n\overline{x} + \hat{\beta}\sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{11.12}$$

**PROPOSITION 1** The least squares estimators of β and α corresponding to the data set $x_i$, $Y_i$, $i = 1, 2, \ldots, n$ are, respectively,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i - \overline{x}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{11.13}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x} \tag{11.14}$$

The straight line $\hat{\alpha} + \hat{\beta}x$ is called the *estimated regression line*.

In the next, we run R for the data from Example 1 and obtain the regression line superimposed on the scatterplot in Figure 11.2.

```
xi<-c(1.0,1.1,1.2,1.3,1.4,1.5,1.6,1.7,1.8,1.9)
> Yi<-c(45,52,54,63,62,68,75,76,92,88)
> model<-lm(Yi~xi)
> summary(model)


Call:
lm(formula = Yi ~ xi)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9091 -1.9455 -0.6273  1.4182  7.1273 

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.473      5.634  -0.794     0.45
xi            49.636      3.812  13.022 1.15e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.462 on 8 degrees of freedom
Multiple R-squared:  0.9549,   Adjusted R-squared:  0.9493
F-statistic: 169.6 on 1 and 8 DF,  p-value: 1.147e-06
```

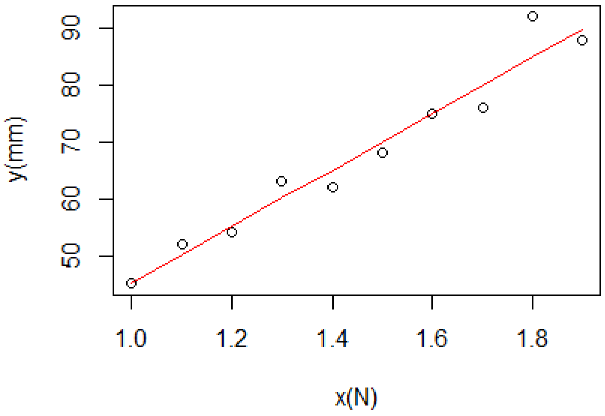

Figure 11.2 Estimated regression line $y = -4.473 + 49.636x$

## 11.3 DISTRIBUTION OF THE ESTIMATORS

To specify the distribution of the estimators $\hat{\alpha}$ and $\hat{\beta}$, it is necessary to make additional assumptions about the random errors aside from just assuming that their mean is 0. The usual approach is to assume that the *random errors are independent normal random variables* having mean 0 and variance $\sigma^2$. That is, we suppose that if $Y_i$ is the response corresponding to the input value $x_i$, then $Y_1, Y_2, \ldots, Y_n$ are independent and

$$Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right) \qquad (11.15)$$

Note that this value $\sigma^2$ is not assumed to be known but is a constant that *must be estimated from the data*.

The least squares estimator $\hat{\beta}$ of β can be expressed as

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{x} \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad \Rightarrow \quad \hat{\beta} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right) Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad (11.16)$$

It is a linear combination of the independent normal random variables $Y_i$, $i = 1, 2, \ldots, n$ and so is itself *normally distributed*.

Using Eq (11.16), the *mean* and *variance* of $\hat{\beta}$ are computed as follows:

$$E\left[\hat{\beta}\right] = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)E\left[Y_i\right]}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(\alpha + \beta x_i\right)}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$E\left[\hat{\beta}\right] = \frac{\alpha\sum_{i=1}^{n}\left(x_i - \overline{x}\right) + \beta\sum_{i=1}^{n} x_i\left(x_i - \overline{x}\right)}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$E\left[\hat{\beta}\right] = \beta\frac{\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \beta \qquad (11.17)$$

since $\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0$

Thus $E\left[\hat{\beta}\right] = \beta$ and so $\hat{\beta}$ is an *unbiased* estimator of β. (11.18)

$$Var\left(\hat{\beta}\right) = \frac{Var\left(\sum_{i=1}^{n}\left(x_i - \overline{x}\right)Y_i\right)}{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)^2}$$

$$Var\left(\hat{\beta}\right) = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 Var\left(Y_i\right)}{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)^2} \qquad \text{by independence}$$

$$Var\left(\hat{\beta}\right) = \frac{\sigma^2\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)^2} \qquad (11.19)$$

Since $\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}\left(x_i^2-2\overline{x}x_i+\overline{x}^2\right)=\sum_{i=1}^{n}x_i^2-2n\overline{x}^2+n\overline{x}^2=\sum_{i=1}^{n}x_i^2-n\overline{x}^2$

$$Var\left(\hat{\beta}\right)=\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2} \qquad (11.20)$$

The estimator $\hat{\alpha}$ is

$$\hat{\alpha}=\overline{Y}-\hat{\beta}\overline{x}=\sum_{i=1}^{n}\frac{Y_i}{n}-\hat{\beta}\overline{x} \qquad (11.21)$$

and shows that $\hat{\alpha}$ can also be expressed as a linear combination of the independent normal random variables $Y_i$, $i=1,2,\ldots,n$ and is thus also *normally distributed*. Its mean is obtained from

$$E[\hat{\alpha}]=\sum_{i=1}^{n}\frac{E[Y_i]}{n}-\overline{x}E\left[\hat{\beta}\right]$$

$$E[\hat{\alpha}]=\sum_{i=1}^{n}\frac{(\alpha+\beta x_i)}{n}-\overline{x}\beta=\alpha+\beta\overline{x}-\overline{x}\beta=\alpha \qquad (11.22)$$

Thus $E[\hat{\alpha}]=\alpha$ and $\hat{\alpha}$ is also an *unbiased* estimator. (11.23)

The variance of $\hat{\alpha}$ is computed by expressing $\hat{\alpha}$ as a linear combination of the $Y_i$.

$$Var\left(\hat{\alpha}\right)=Var\left(\sum_{i=1}^{n}\frac{Y_i}{n}-\hat{\beta}\overline{x}\right)=\frac{1}{n^2}\sum_{i=1}^{n}Var\left(Y_i\right)+\overline{x}^2Var\left(\hat{\beta}\right)$$

$$Var\left(\hat{\alpha}\right)=\frac{1}{n^2}n\sigma^2+\overline{x}^2\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}=\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}\right)$$

$$Var\left(\hat{\alpha}\right)=\sigma^2\left(\frac{\sum_{i=1}^{n}x_i^2-n\overline{x}^2+n\overline{x}^2}{n\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)}\right)$$

$$Var\left(\hat{\alpha}\right)=\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)} \qquad (11.24)$$

The quantities $Y_i - \hat{\alpha} - \hat{\beta} x_i$, $i = 1, 2, \ldots, n$, which represent the differences between the actual responses (that is, the $Y_i$) and their least squares estimators (that is, $\hat{\alpha} + \hat{\beta} x_i$) are called the *residuals*. The *sum of squares of the residuals*

$$SS_R = \sum_{i=1}^{n} \left( Y_i - \hat{\alpha} - \hat{\beta} x_i \right)^2 \qquad (11.25)$$

can be utilized *to estimate the unknown error variance* $\sigma^2$. Indeed, it can be shown that

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2} \qquad (11.26)$$

That is, $SS_R / \sigma^2$ has a chi-square distribution with $n-2$ degrees of freedom, which implies that

$$E\left[ \frac{SS_R}{\sigma^2} \right] = n - 2 \qquad (11.27)$$

or

$$E\left[ \frac{SS_R}{n-2} \right] = \sigma^2 \qquad (11.28)$$

Thus $SS_R / (n-2)$ is an *unbiased estimator of* $\sigma^2$. In addition, it can be shown that $SS_R$ is independent of the pair $\hat{\alpha}$ and $\hat{\beta}$.

**REMARKS**

A plausibility argument as to why $SS_R / \sigma^2$ might have a chi-square distribution with $n-2$ degrees of freedom and be independent of $\hat{\alpha}$ and $\hat{\beta}$ runs as follows. Because the $Y_i$ are independent normal random variables, it follows that $\dfrac{Y_i - E[Y_i]}{\sqrt{Var(Y_i)}}$, $i = 1, 2, \ldots, n$ are independent standard normals and so

$$\sum_{i=1}^{n} \frac{\left( Y_i - E[Y_i] \right)^2}{Var(Y_i)} = \sum_{i=1}^{n} \frac{\left( Y_i - \alpha - \beta x_i \right)^2}{\sigma^2} \sim \chi^2_n \qquad (11.29)$$

Now if we substitute the estimators $\hat{\alpha}$ and $\hat{\beta}$ for α and β, then 2 degrees of freedom are lost, and so $SS_R / \sigma^2$ has a chi-square distribution with $n-2$ degrees of freedom.

The fact that $SS_R$ is independent of $\hat{\alpha}$ and $\hat{\beta}$ is quite similar to the fundamental result (theorem) that in normal sampling $\bar{X}$ and $S^2$ are independent. Indeed this latter result states that if $Y_1, Y_2, \ldots, Y_n$ is a normal sample with population mean μ and variance $\sigma^2$, then if in the sum of squares $\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2}$, which has a chi-square distribution with *n* degrees of freedom, one substitutes the estimator $\bar{Y}$ for μ to obtain the new sum of squares $\sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{\sigma^2}$ , then this quantity [equal to $(n-1)S^2/\sigma^2$] will be independent of $\bar{Y}$ and will have a chi-square distribution with $n-1$ degrees of freedom. Since $SS_R/\sigma^2$ is obtained by substituting the estimators $\hat{\alpha}$ and $\hat{\beta}$ for α and β in the sum of squares $\sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2}$, it is not unreasonable to expect that this quantity might be independent of $\hat{\alpha}$ and $\hat{\beta}$.

**Notation** If we let

$$S_{xY} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \qquad (11.30)$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad (11.31)$$

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \qquad (11.32)$$

then the least squares estimators can be expressed as

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{x}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad \Rightarrow \quad \hat{\beta} = \frac{S_{xY}}{S_{xx}} \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} \qquad (11.33)$$

**PROPOSITION** Suppose that the responses $Y_i$, $i = 1, 2, \ldots, n$ are independent normal random variables with means $\alpha + \beta x_i$ and common variance $\sigma^2$. The least squares estimators of β and α

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} \qquad (11.34)$$

are distributed as follows:

$$\hat{\alpha} \sim N\left(\alpha,\ \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}\right) \tag{11.35}$$

$$\hat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right) \tag{11.36}$$

In addition, if

$$SS_R = \sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 \tag{11.37}$$

is the sum of squares of the residuals, then

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2} \tag{11.38}$$

Also, $SS_R$ can be computed from

$$SS_R = \frac{S_{xx}S_{YY} - \left(S_{xY}\right)^2}{S_{xx}} \tag{11.39}$$

## 11.4 STATISTICAL INFERENCES ABOUT THE REGRESSION PARAMETERS

Using Proposition, it is a simple matter to devise hypothesis tests and confidence intervals for the regression parameters.

*Inferences Concerning* β

An important hypothesis to consider regarding the simple linear regression model

$$Y = \alpha + \beta x + e \tag{11.40}$$

is the hypothesis that $\beta = 0$. Its importance derives from the fact that it is equivalent to stating that the mean response does not depend on the input variable. To test

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0 \tag{11.41}$$

note that, from the last Proposition, $\hat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)$, so

$$\frac{\hat{\beta}-\beta}{\sqrt{\sigma^2 / S_{xx}}} = \sqrt{S_{xx}}\,\frac{\hat{\beta}-\beta}{\sigma} \sim N\left(0,1\right) \qquad (11.42)$$

and is independent of

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2} \qquad (11.43)$$

From the definition of a *t*-random variable $T_n = \frac{Z}{\sqrt{\chi^2_n / n}}$ it follows that

$$\frac{\sqrt{S_{xx}}\,\frac{\hat{\beta}-\beta}{\sigma}}{\sqrt{\frac{SS_R}{\sigma^2\left(n-2\right)}}} = \sqrt{\frac{\left(n-2\right)S_{xx}}{SS_R}}\left(\hat{\beta}-\beta\right) \sim t_{n-2} \qquad (11.44)$$

That is, $\sqrt{\left(n-2\right)S_{xx} / SS_R}\left(\hat{\beta}-\beta\right)$ has a *t*-distribution with $n-2$ degrees of freedom. Therefore, if $H_0$ is true (and so $\beta = 0$), then

$$\sqrt{\frac{\left(n-2\right)S_{xx}}{SS_R}}\hat{\beta} \sim t_{n-2} \qquad (11.45)$$

which gives rise to the following test of $H_0$.

Hypothesis Test of $H_0 : \beta = 0$

A significance level γ test of $H_0$ is to

Reject $H_0$ if $\sqrt{\frac{\left(n-2\right)S_{xx}}{SS_R}}\left|\hat{\beta}\right| > t_{\gamma/2,n-2}$ (11.46)

Accept $H_0$ otherwise

This test can be performed by first computing the value of the test statistic $\sqrt{\left(n-2\right)S_{xx} / SS_R}\left|\hat{\beta}\right|$ - call its value $v$ – and then rejecting $H_0$ if the desired significance level is at least as large as

$$\begin{aligned} p-value &= P\left(\left|T_{n-2}\right| > v\right) \\ &= 2P\left(T_{n-2} > v\right) \end{aligned} \qquad (11.47)$$

where $T_{n-2}$ is a *t*-random variable with $n-2$ degrees of freedom.

**Example 2** An individual claims that the fuel consumption of his automobile does not depend on how fast the car is driven. To test the plausibility of this hypothesis, the car was tested at various speeds between 45 and 70 miles per hour. The miles per gallon attained at each of these speeds was determined, with the following data resulting:

| Speed | Miles per Gallon |
|---|---|
| 45 | 24.2 |
| 50 | 25.0 |
| 55 | 23.3 |
| 60 | 22.0 |
| 65 | 21.5 |
| 70 | 20.6 |
| 75 | 19.8 |

Do these data refute the claim that the mileage per gallon of gas is unaffected by the speed at which the car is being driven?

Suppose that a simple linear regression model

$$Y = \alpha + \beta x + e$$

relates *Y*, the miles per gallon of the car, to *x*, the speed at which it is being driven. The claim being made is that the regression coefficient β is equal to 0. To see if the data are strong enough to refute this claim, we need to see if it leads to a rejection of the null hypothesis when testing

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0$$

To compute the value of the test statistic, we first compute the values of $S_{xx}$, $S_{YY}$, and $S_{xY}$. A hand calculation yields that

$$S_{xx} = 700 \qquad S_{YY} = 21.757 \qquad S_{xY} = -119$$

$$SS_R = \frac{S_{xx}S_{YY} - \left(S_{xY}\right)^2}{S_{xx}} = \frac{700 \times 21.757 - 119^2}{700} = 1.527$$

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} = -\frac{119}{700} = -0.17$$

the value of the *test statistic* is

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\left|\hat{\beta}\right| = \sqrt{\frac{5 \times 700}{1.527}} \times 0.17 = 8.139$$

Since $t_{0.005,5} = 4.032$ it follows that the hypothesis $\beta = 0$ is rejected at the 1 percent level of significance. Thus, the claim that the mileage does not depend on the speed at which the car is driven is rejected; there is strong evidence that increased speeds lead to decreased mileages.

```
  speed<-c(45,50,55,60,65,70,75)
> miles<-c(24.2,25.0,23.3,22.0,21.5,20.6,19.8)
> sxx<-sum((speed-mean(speed))^2)
> sxx
[1] 700
> syy<-sum((miles-mean(miles))^2)
> syy
[1] 21.75714
> sxy<-sum((speed-mean(speed))*(miles-mean(miles)))
> sxy
[1] -119
> ssr<-(sxx*syy-sxy^2)/sxx
> ssr
[1] 1.527143
> bet<-sxy/sxx
> bet
[1] -0.17
> v<-abs(bet)*sqrt((7-2)*sxx/ssr)
> v
[1] 8.138476
> qt(0.005,5,lower.tail = FALSE)
[1] 4.032143
```

A *confidence interval* estimator for β is easily obtained from Equation (11.44):

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\left(\hat{\beta}-\beta\right) \sim t_{n-2} \tag{11.48}$$

Indeed, it follows from this Equation that for any $a$, $0 < a < 1$,

$$P\left(-t_{a/2,n-2} < \sqrt{\frac{(n-2)S_{xx}}{SS_R}}\left(\hat{\beta}-\beta\right) < t_{a/2,n-2}\right) = 1-a \tag{11.49}$$

or, equivalently,

$$P\left(-\sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2} < \left(\hat{\beta}-\beta\right) < \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right) = 1-a$$

$$P\left(\hat{\beta}-\sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2} < \beta < \hat{\beta}+\sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right) = 1-a \tag{11.50}$$

*Confidence Interval for β*

A $100(1-a)$ percent confidence interval estimator of β is

$$\left(\hat{\beta}-\sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\ ,\ \hat{\beta}+\sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right) \qquad (11.51)$$

**REMARK**

The result that

$$\frac{\hat{\beta}-\beta}{\sqrt{\sigma^2/S_{xx}}}\sim N(0,1) \qquad (11.52)$$

cannot be immediately applied to make inferences about β since it involves the unknown parameter $\sigma^2$. Instead, we replace $\sigma^2$ by its estimator $SS_R/(n-2)$, which has the effect of changing the distribution of the statistic from the standard normal to the *t*-distribution with $n-2$ degrees of freedom.

**Example 3** Derive a 95 percent confidence interval estimate of β in the previous Example 2.

Since $t_{0.025,5}=2.571$, it follows from the computations of this example that the 95 percent confidence interval is

$$\left(\hat{\beta}-\sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\ ,\ \hat{\beta}+\sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right)$$

$$\left(-0.17-2.571\sqrt{\frac{1.527}{5\times700}},\ -0.17+2.571\sqrt{\frac{1.527}{5\times700}}\right)=(-0.224,\ -0.116)$$

*Inferences Concerning α*

The determination of confidence intervals and hypothesis tests for α is accomplished in exactly the same manner as was done for β. Specifically, the last Proposition can be used to show that

$$\frac{\hat{\alpha}-\alpha}{\sqrt{\dfrac{\sigma^2\sum_{i=1}^{n}x_i^2}{nS_{xx}}}}\sim N(0,1) \qquad (11.53)$$

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2} \tag{11.54}$$

$$\frac{\dfrac{\hat{\alpha}-\alpha}{\sqrt{\dfrac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}}}{\sqrt{\dfrac{SS_R}{\sigma^2(n-2)}}} = \sqrt{\frac{n(n-2)S_{xx}}{SS_R \sum_{i=1}^{n} x_i^2}}\left(\hat{\alpha}-\alpha\right) \sim t_{n-2} \tag{11.55}$$

From this Eq, $\forall a \in (0,1)$,

$$P\left(-t_{a/2,n-2} < \sqrt{\frac{n(n-2)S_{xx}}{SS_R \sum_{i=1}^{n} x_i^2}}\left(\hat{\alpha}-\alpha\right) < t_{a/2,n-2}\right) = 1-a \tag{11.56}$$

$$P\left(-t_{a/2,n-2}\sqrt{\frac{SS_R \sum_{i=1}^{n} x_i^2}{n(n-2)S_{xx}}} < \hat{\alpha}-\alpha < t_{a/2,n-2}\sqrt{\frac{SS_R \sum_{i=1}^{n} x_i^2}{n(n-2)S_{xx}}}\right) = 1-a$$

$$P\left(\hat{\alpha}-t_{a/2,n-2}\sqrt{\frac{SS_R \sum_{i=1}^{n} x_i^2}{n(n-2)S_{xx}}} < \alpha < \hat{\alpha}+t_{a/2,n-2}\sqrt{\frac{SS_R \sum_{i=1}^{n} x_i^2}{n(n-2)S_{xx}}}\right) = 1-a \tag{11.57}$$

The 100(1 − a) percent confidence interval for α is the interval

$$\left(\hat{\alpha}-t_{a/2,n-2}\sqrt{\frac{SS_R \sum_{i=1}^{n} x_i^2}{n(n-2)S_{xx}}},\ \ \hat{\alpha}+t_{a/2,n-2}\sqrt{\frac{SS_R \sum_{i=1}^{n} x_i^2}{n(n-2)S_{xx}}}\right) \tag{11.58}$$

*Summary of Distributional Results*

Model: $Y = \alpha + \beta x + e$ , $e \sim N\left(0,\sigma^2\right)$

Data: $\left(x_i, Y_i\right)$, $i = 1,2,\ldots,n$

| Inferences About | Use the Distributional Result |
|---|---|
| $\beta$<br>$\alpha$ | $\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\left(\hat{\beta}-\beta\right)\sim t_{n-2}$<br>$\sqrt{\frac{n(n-2)S_{xx}}{SS_R\sum_{i=1}^{n}x_i^2}}\left(\hat{\alpha}-\alpha\right)\sim t_{n-2}$ |

## 11.5 THE COEFFICIENT OF DETERMINATION AND THE SAMPLE CORRELATION COEFFICIENT

Suppose we wanted to measure the amount of variation in the set of response values $Y_1, Y_2, \ldots, Y_n$ corresponding to the set of input values $x_1, x_2, \ldots, x_n$. A standard measure in statistics of the amount of variation in a set of values $Y_1, Y_2, \ldots, Y_n$ is given by the quantity

$$S_{YY} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \qquad (11.59)$$

For instance, if all the $Y_i$ are equal — and thus are all equal to $\bar{Y}$ — then $S_{YY}$ would equal 0.

The variation in the values of the $Y_i$ arises from two factors. First, because the input values $x_i$ are different, the response variables $Y_i$ all have different mean values, which will result in some variation in their values. Second, the variation also arises from the fact that even when the differences in the input values are taken into account, each of the response variables $Y_i$ has variance $\sigma^2$ and thus will not exactly equal the predicted value at its input $x_i$.

Let us consider now the question as to how much of the variation in the values of the response variables is due to the different input values, and how much is due to the inherent variance of the responses even when the input values are taken into account. To answer this question, note that the quantity

$$SS_R = \sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 \qquad (11.60)$$

measures the remaining amount of variation in the response values after the different input values have been taken into account.

Thus,

$$S_{YY} - SS_R \qquad (11.61)$$

represents the amount of variation in the response variables that is explained by the different input values (or *by the model*), and so the quantity $R^2$ defined by

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}} \qquad (11.62)$$

represents the proportion of the *variation in the response variables that is explained by the different input values* (or, by the model). $R^2$ is called the *coefficient of determination*.

The coefficient of determination $R^2$ will have a value between 0 and 1. A value of $R^2$ near 1 indicates that most of the variation of the response data is explained by the different input values (or by the model), whereas a value of $R^2$ near 0 indicates that little of the variation is explained by the different input values.

The value of $R^2$ is often used as an indicator of how well the regression model fits the data, with a value near 1 indicating a good fit, and one near 0 indicating a poor fit. In other words, if the regression model is able to explain most of the variation in the response data, then it is considered to fit the data well.

Recall that we defined the sample correlation coefficient *r* of the set of data pairs $(x_i, Y_i)$, $i = 1, 2, \ldots, n$, by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad (11.63)$$

It was noted that *r* provided a measure of the degree to which high values of *x* are paired with high values of *Y* and low values of *x* with low values of *Y*. A value of *r* near +1 indicated that large *x* values were strongly associated with large *Y* values and small *x* values were strongly associated with small *Y* values, whereas a value near −1 indicated that large *x* values were strongly associated with small *Y* values and small *x* values with large *Y* values.

In the notation of this chapter,

$$r = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}} \qquad (11.64)$$

Upon using identity

$$SS_R = \frac{S_{xx}S_{YY} - \left(S_{xY}\right)^2}{S_{xx}} \tag{11.65}$$

We see that

$$r^2 = \frac{S_{xY}^2}{S_{xx}S_{YY}} = \frac{S_{xx}S_{YY} - SS_R S_{xx}}{S_{xx}S_{YY}} = 1 - \frac{SS_R}{S_{YY}} = R^2 \tag{11.66}$$

That is,

$$|r| = \sqrt{R^2} \tag{11.67}$$

and so, except for its sign indicating whether it is positive or negative, the sample correlation coefficient is equal to the square root of the coefficient of determination. The sign of *r* is the same as that of $\hat{\beta}$.

The above gives additional meaning to the sample correlation coefficient. For instance, if a data set has its sample correlation coefficient *r* equal to 0.9, then this implies that a simple linear regression model for these data explains 81 percent (since $R^2 = 0.9^2 = 0.81$) of the variation in the response values. That is, 81 percent of the variation in the response values is explained by the different input values (or by the model).

## 11.6 ANALYSIS OF RESIDUALS: ASSESSING THE MODEL

The initial step for ascertaining whether or not the simple linear regression model

$$Y = \alpha + \beta x + e, \quad e \sim N\left(0, \sigma^2\right) \tag{11.68}$$

is appropriate in a given situation is to investigate the scatter diagram. Indeed, this is often sufficient to convince one that the regression model is or is not correct. Moreover, the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ should be computed and the residual $Y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right), \quad i = 1, 2, \ldots, n$ analyzed. The analysis begins by *normalizing*, or *standardizing*, the residuals by dividing them by $\sqrt{SS_R / (n-2)}$, the estimate of the standard deviation of the $Y_i$. The resulting quantities

$$\frac{Y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right)}{\sqrt{SS_R / (n-2)}}, \; i = 1, 2, \ldots, n \tag{11.69}$$

are called the *standardized residuals*.

When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables, and thus should be randomly distributed about 0 with about 95 percent of their values being between −2 and +2 (since $P(-1.96 < Z < 1.96) = 0.95$). In addition, a plot of the standardized residuals should not indicate any distinct pattern. Indeed, any indication of a distinct pattern should make one suspicious about the validity of the assumed simple linear regression model.

Figure 11.3 presents three different scatter diagrams and their associated standardized residuals. The first of these, as indicated both by its scatter diagram and the random nature of its standardized residuals, appears to fit the straight-line model quite well. The second residual plot shows a discernible pattern, in that the residuals appear to be first decreasing and then increasing as the input level increases. This often means that higher order (than just linear) terms are needed to describe the relationship between the input and response. This is also indicated by the scatter diagram. The third standardized residual plot also shows a pattern, in that the absolute value of the residuals appear to be increasing, as the input level increases. This indicates that the variance of the response is not constant but increases with the input level.

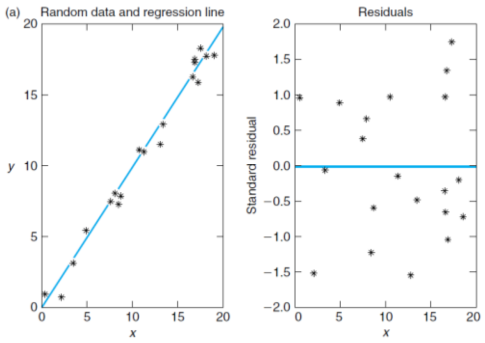

Figure 11.3a.

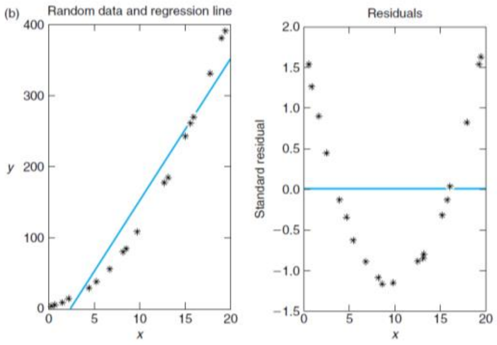

Figure 11.3b

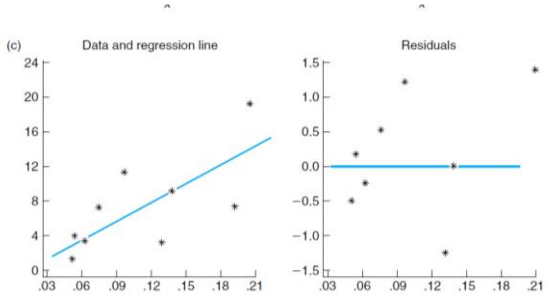
(c)
Data and regression line
24
20
16
12
8
4
0
.03
.06
.09
.12
.15
.18
.21
Residuals
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
.03
.06
.09
.12
.15
.18
.21

Figure 11. 3c