Foundation of
Data Science and Analytics

Simple Linear Regression

Arun K. Timalsina

Material Adaptation : Applied Statistics and Probability for Engineers, by Montgomery and Runger. John Wiley & Sons,

11-1: Empirical Models

- Many problems in engineering and science involve exploring the relationships between two or more variables.
- **Regression analysis** is a statistical technique that is very useful for these types of problems.
- For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature.
- Regression analysis can be used to build a model to predict yield at a given temperature level.

11-1: Empirical Models

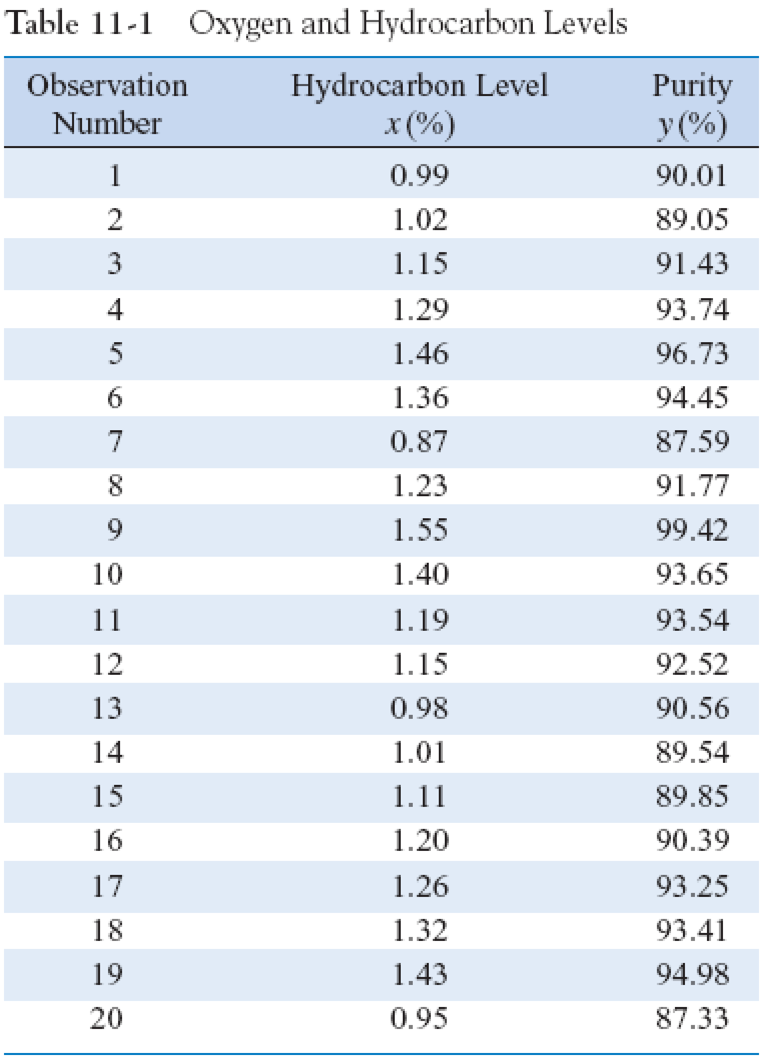


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

11-1: Empirical Models

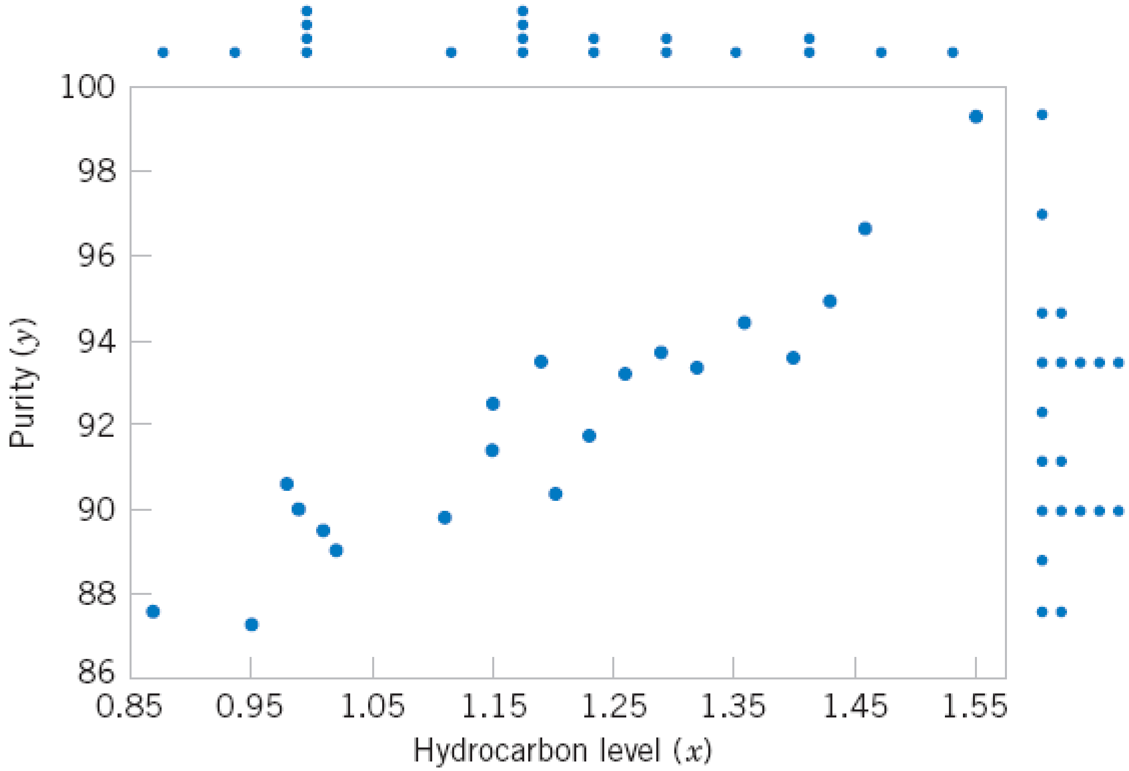


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

Figure 11-1 Scatter Diagram of oxygen purity versus hydrocarbon level from Table 11-1.

11-1: Empirical Models

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable *Y* is related to *x* by the following straight-line relationship:

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

where the slope and intercept of the line are called **regression coefficients.**
The **simple linear regression model** is given by

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

where ε is the random error term.

11-1: Empirical Models

We think of the regression model as an empirical model. Suppose that the mean and variance of ε are 0 and σ^2, respectively, then

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x$$

The variance of *Y* given *x* is

$$V(Y|x) = V(\beta_0 + \beta_1 x + \epsilon) = V(\beta_0 + \beta_1 x) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

11-1: Empirical Models

- The true regression model is a line of mean values:

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

where β_1 can be interpreted as the change in the mean of *Y* for a unit change in *x*.

- Also, the variability of Y at a particular value of *x* is determined by the error variance, σ^2.
- This implies there is a distribution of *Y*-values at each *x* and that the variance of this distribution is the same at each *x*.

11-1: Empirical Models

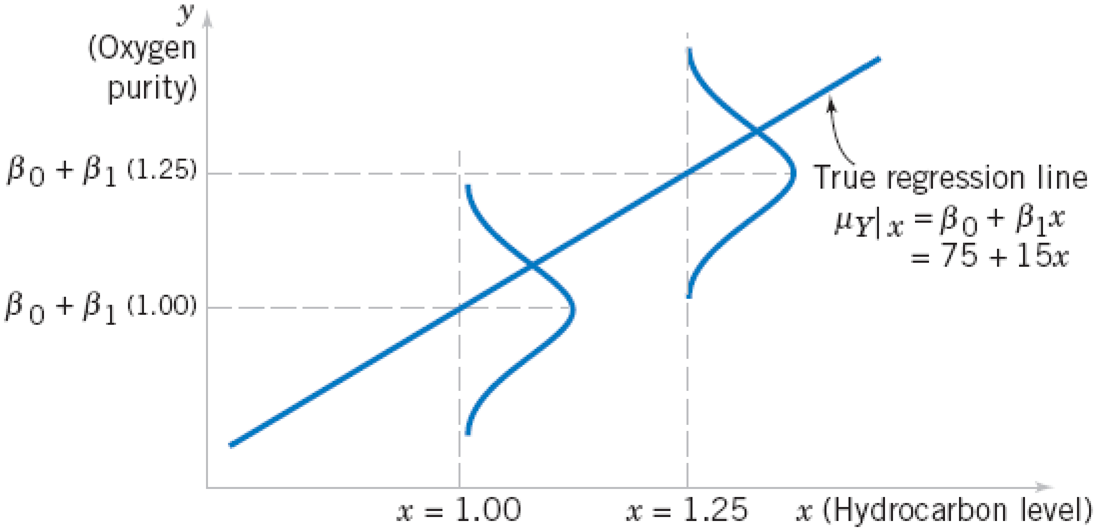


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity–hydrocarbon data.

Figure 11-2 The distribution of *Y* for a given value of *x* for the oxygen purity-hydrocarbon data.

11-2: Simple Linear Regression

- The case of **simple linear regression** considers a single **regressor** or **predictor** *x* and a **dependent** or **response variable** *Y*.
- The expected value of *Y* at each level of *x* is a random variable:

$$E(Y|x) = \beta_0 + \beta_1 x$$

- We assume that each observation, *Y*, can be described by the model

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

11-2: Simple Linear Regression

- Suppose that we have *n* pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

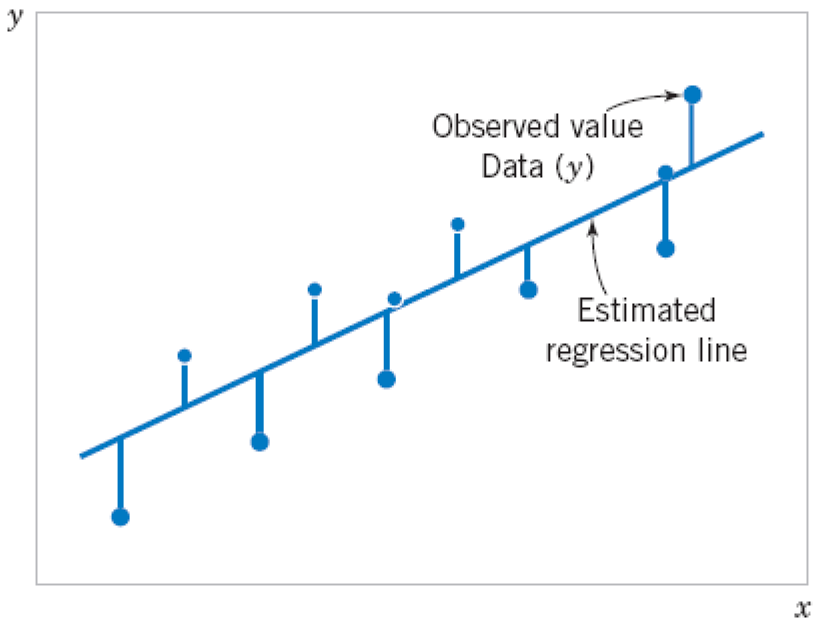


Figure 11-3 Deviations of the data from the estimated regression model.

Figure 11-3 Deviations of the data from the estimated regression model.

11-2: Simple Linear Regression

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

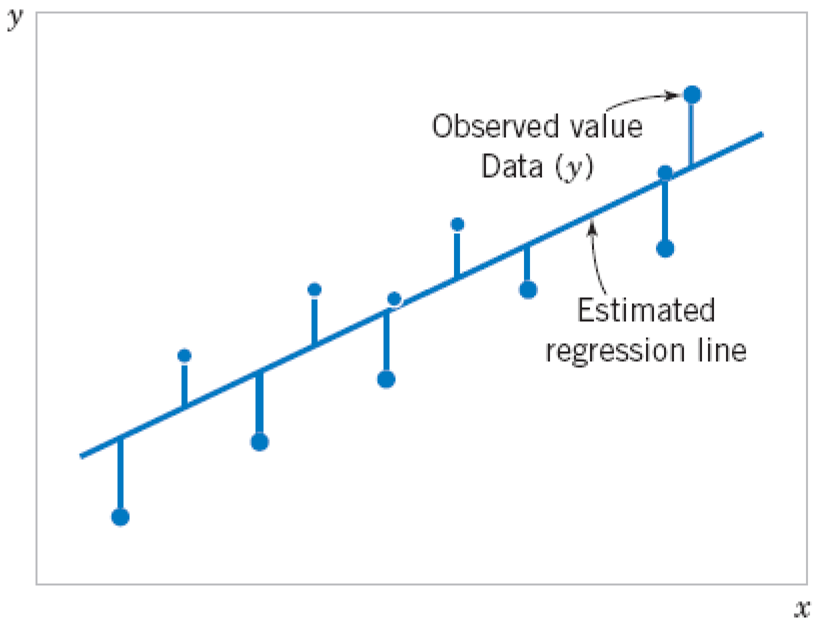


Figure 11-3 Deviations of the data from the estimated regression model.

Figure 11-3 Deviations of the data from the estimated regression model.

11-2: Simple Linear Regression

- Using Equation 11-2, the *n* observations in the sample can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

- The sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

11-2: Simple Linear Regression

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

11-2: Simple Linear Regression

Simplifying these two equations yields

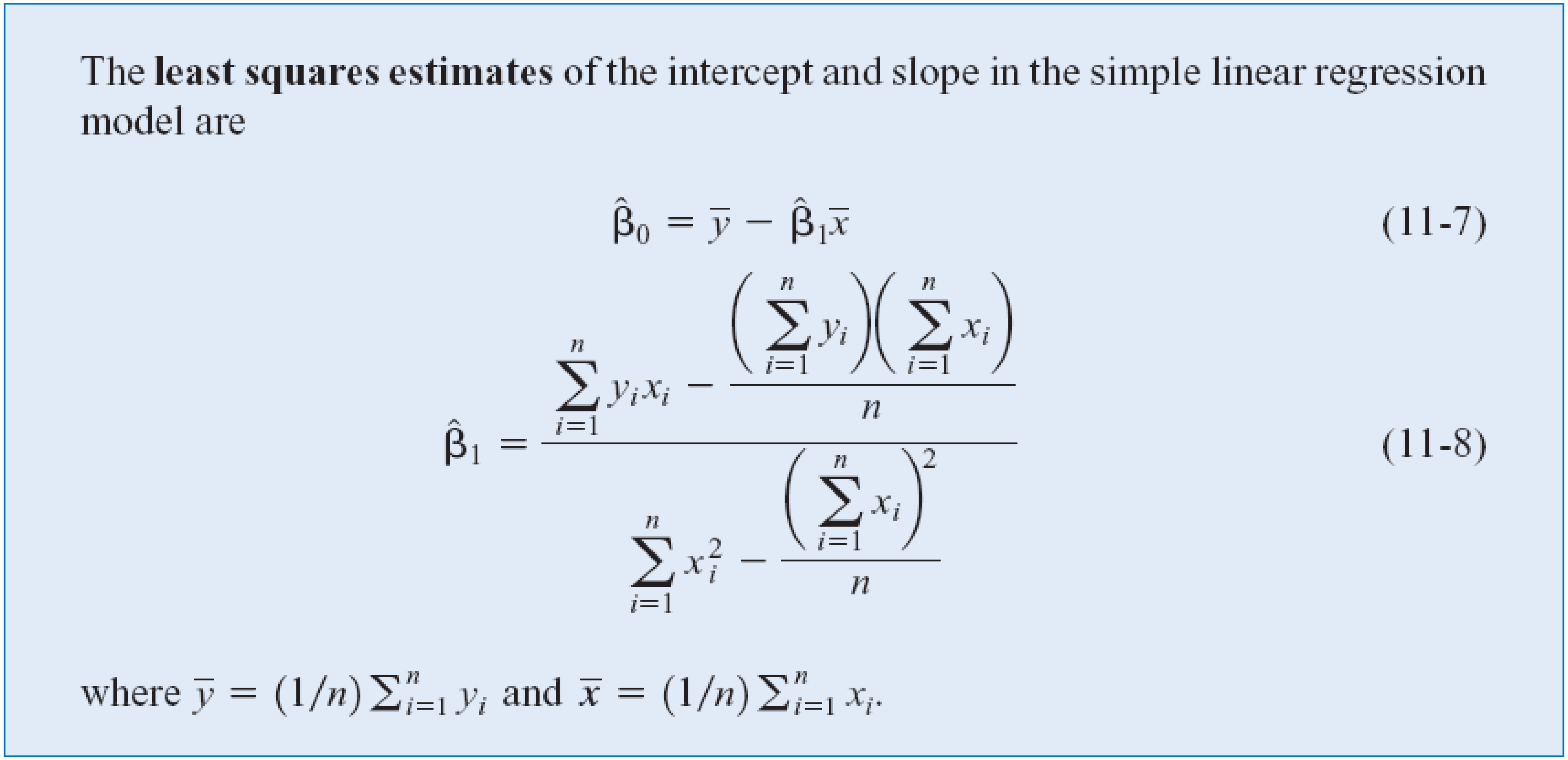
$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad (11\text{-}6)$$

Equations 11-6 are called the **least squares normal equations.** The solution to the normal equations results in the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

11-2: Simple Linear Regression

Definition



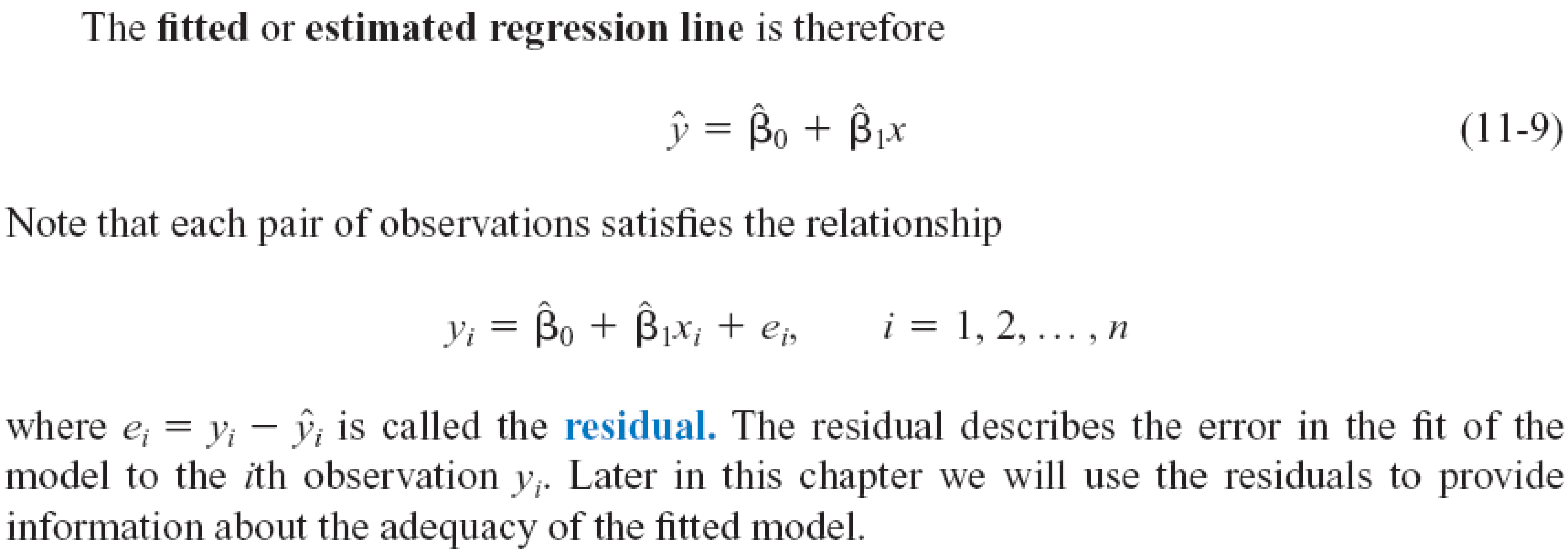
The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} y_i x_i - \frac{\left(\displaystyle\sum_{i=1}^{n} y_i\right)\left(\displaystyle\sum_{i=1}^{n} x_i\right)}{n}}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

11-2: Simple Linear Regression



The **fitted** or **estimated regression line** is therefore

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{11-9}$$

Note that each pair of observations satisfies the relationship

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \qquad i = 1, 2, \ldots, n$$

where $e_i = y_i - \hat{y}_i$ is called the **residual.** The residual describes the error in the fit of the model to the ith observation y_i. Later in this chapter we will use the residuals to provide information about the adequacy of the fitted model.

11-2: Simple Linear Regression

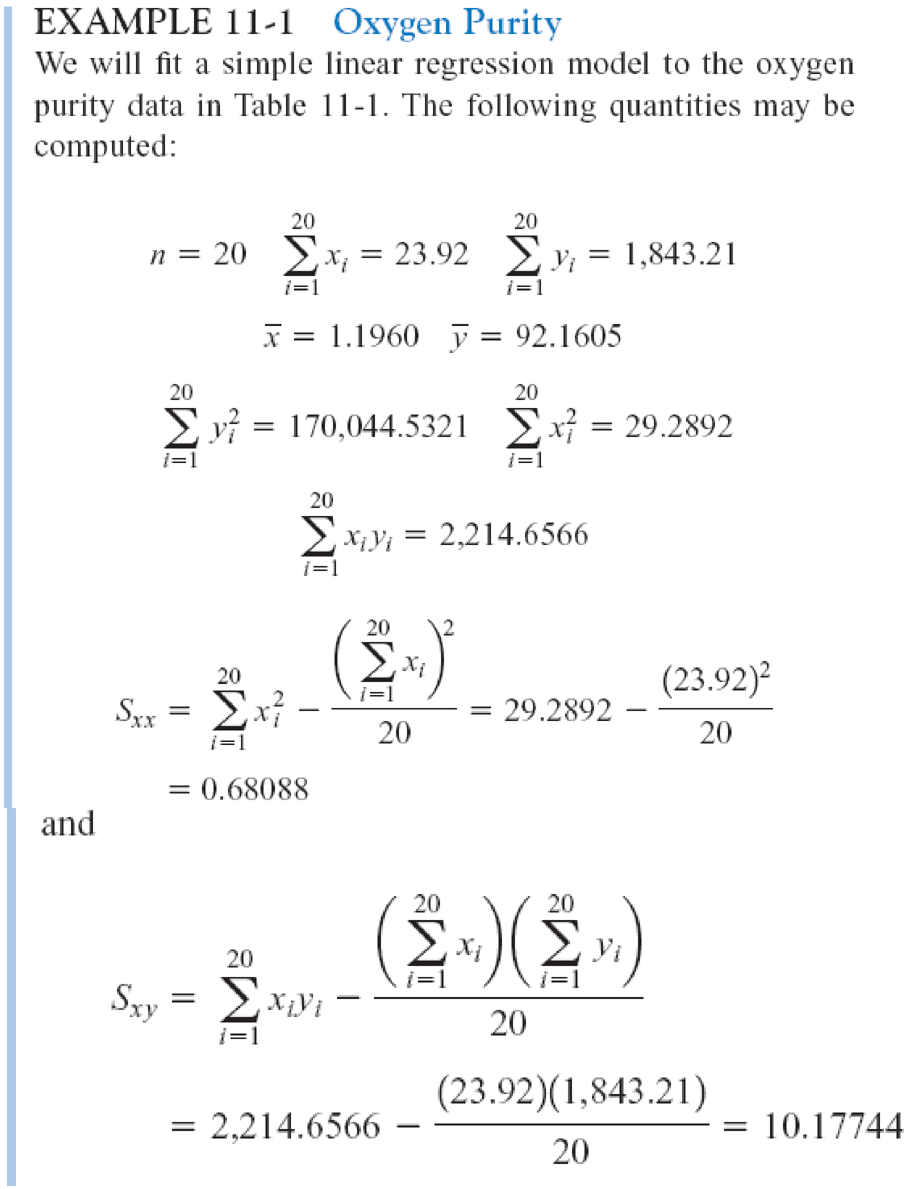
Notation

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

11-2: Simple Linear Regression

Example 11-1



EXAMPLE 11-1 Oxygen Purity

We will fit a simple linear regression model to the oxygen purity data in Table 11-1. The following quantities may be computed:

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21$$

$$\bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170{,}044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892$$

$$\sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20}$$

$$= 0.68088$$

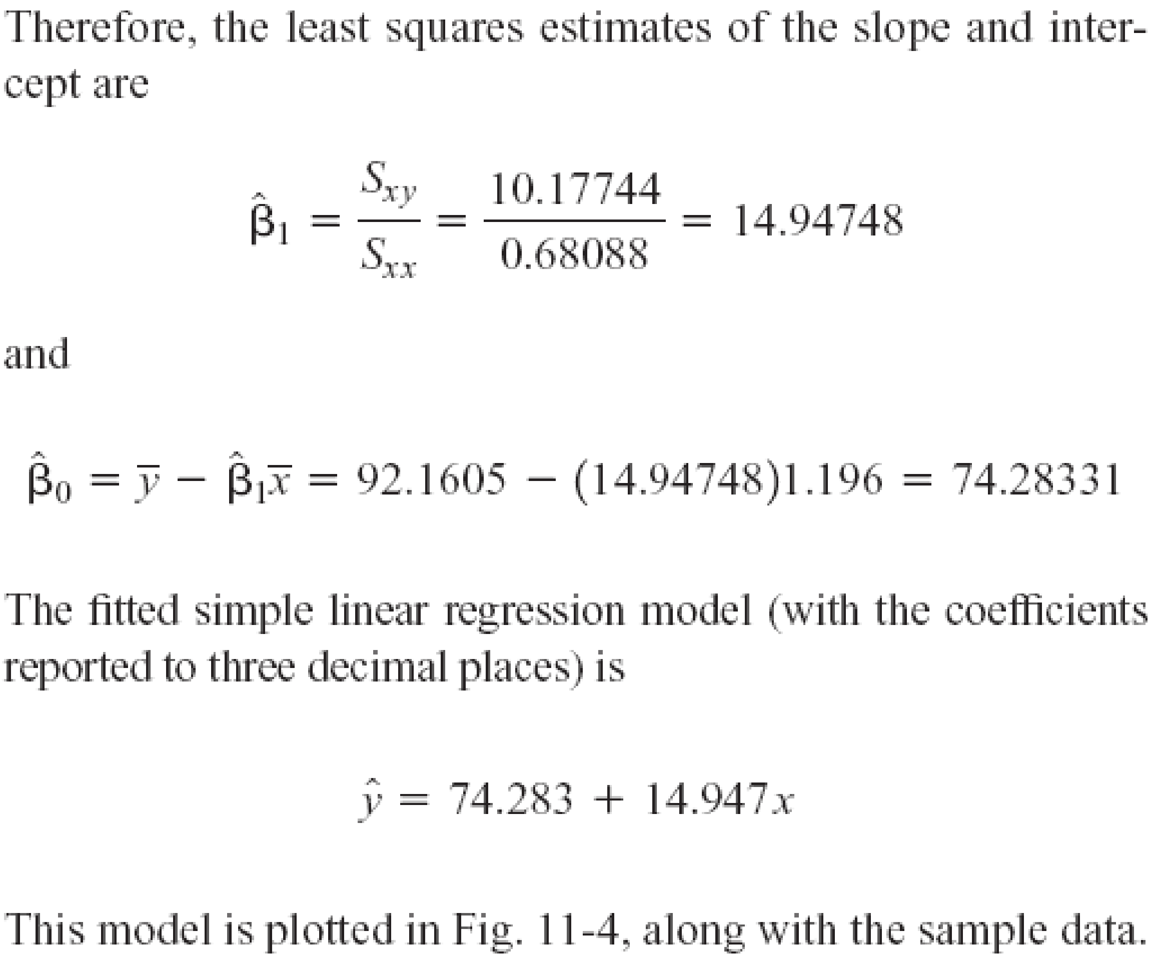
and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20}$$

$$= 2{,}214.6566 - \frac{(23.92)(1{,}843.21)}{20} = 10.17744$$

11-2: Simple Linear Regression

Example 11-1



Therefore, the least squares estimates of the slope and intercept are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

The fitted simple linear regression model (with the coefficients reported to three decimal places) is

$$\hat{y} = 74.283 + 14.947x$$

This model is plotted in Fig. 11-4, along with the sample data.

11-2: Simple Linear Regression

Example 11-1

Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model ŷ = 74.20 + 14.97x.

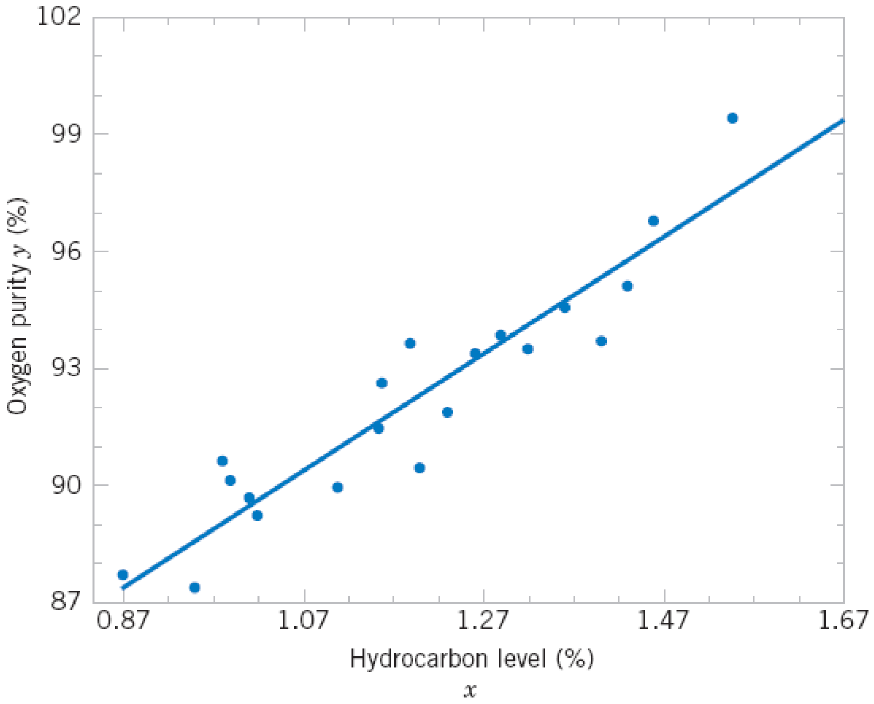
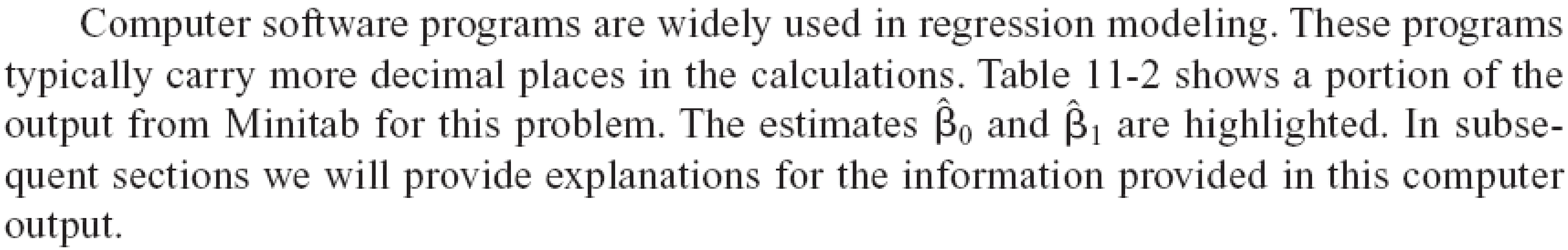


Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.

11-2: Simple Linear Regression

Example 11-1



Computer software programs are widely used in regression modeling. These programs typically carry more decimal places in the calculations. Table 11-2 shows a portion of the output from Minitab for this problem. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are highlighted. In subsequent sections we will provide explanations for the information provided in this computer output.

11-1: Empirical Models

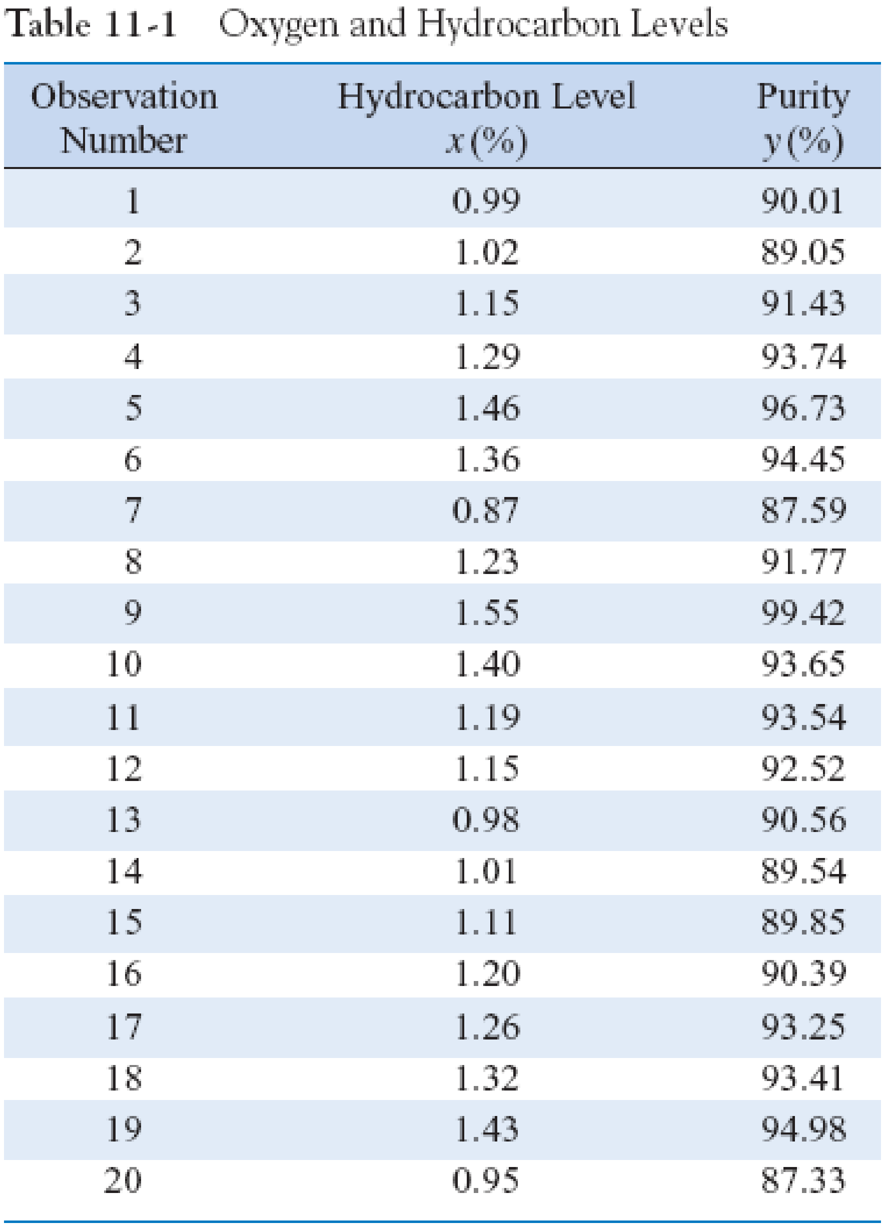


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

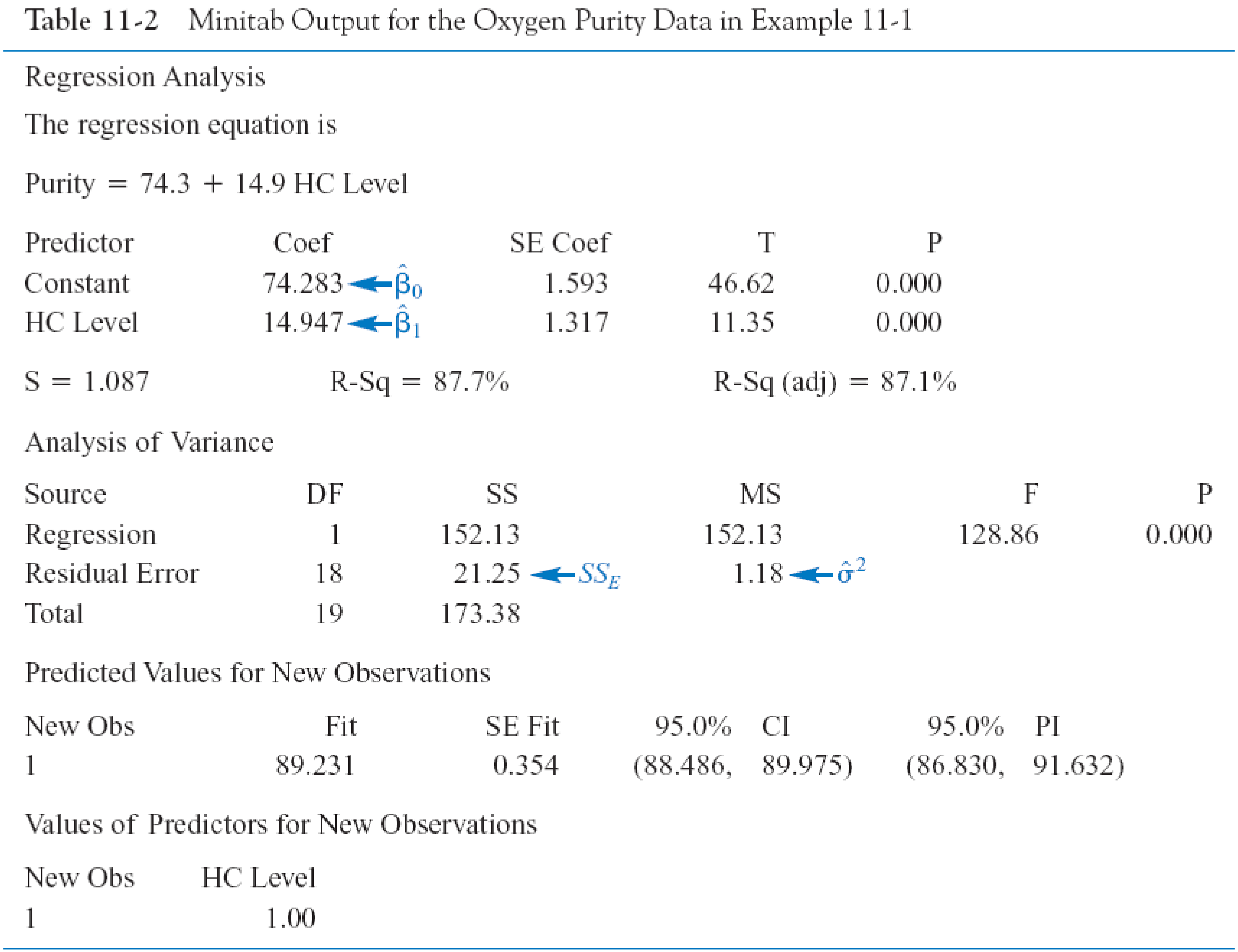


Table 11-2 Minitab Output for the Oxygen Purity Data in Example 11-1

Regression Analysis

The regression equation is

Purity = 74.3 + 14.9 HC Level

Predictor	Coef	SE Coef	T	P
Constant	74.283 ← $\hat{\beta}_0$	1.593	46.62	0.000
HC Level	14.947 ← $\hat{\beta}_1$	1.317	11.35	0.000

S = 1.087 R-Sq = 87.7% R-Sq (adj) = 87.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	152.13	152.13	128.86	0.000
Residual Error	18	21.25 ← SS_E	1.18 ← $\hat{\sigma}^2$		
Total	19	173.38			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	89.231	0.354	(88.486, 89.975)	(86.830, 91.632)

Values of Predictors for New Observations

New Obs	HC Level
1	1.00

11-2: Simple Linear Regression

Estimating σ^2

The error sum of squares is

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

It can be shown that the expected value of the error sum of squares is $E(SS_E) = (n - 2)\sigma^2$.

11-2: Simple Linear Regression

Estimating σ^2

An **unbiased estimator** of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \qquad (11\text{-}13)$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \qquad (11\text{-}14)$$

11-3: Properties of the Least Squares Estimators

- Slope Properties

$$E(\hat{\beta}_1) = \beta_1 \qquad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

- Intercept Properties

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right]$$

11-1: Empirical Models

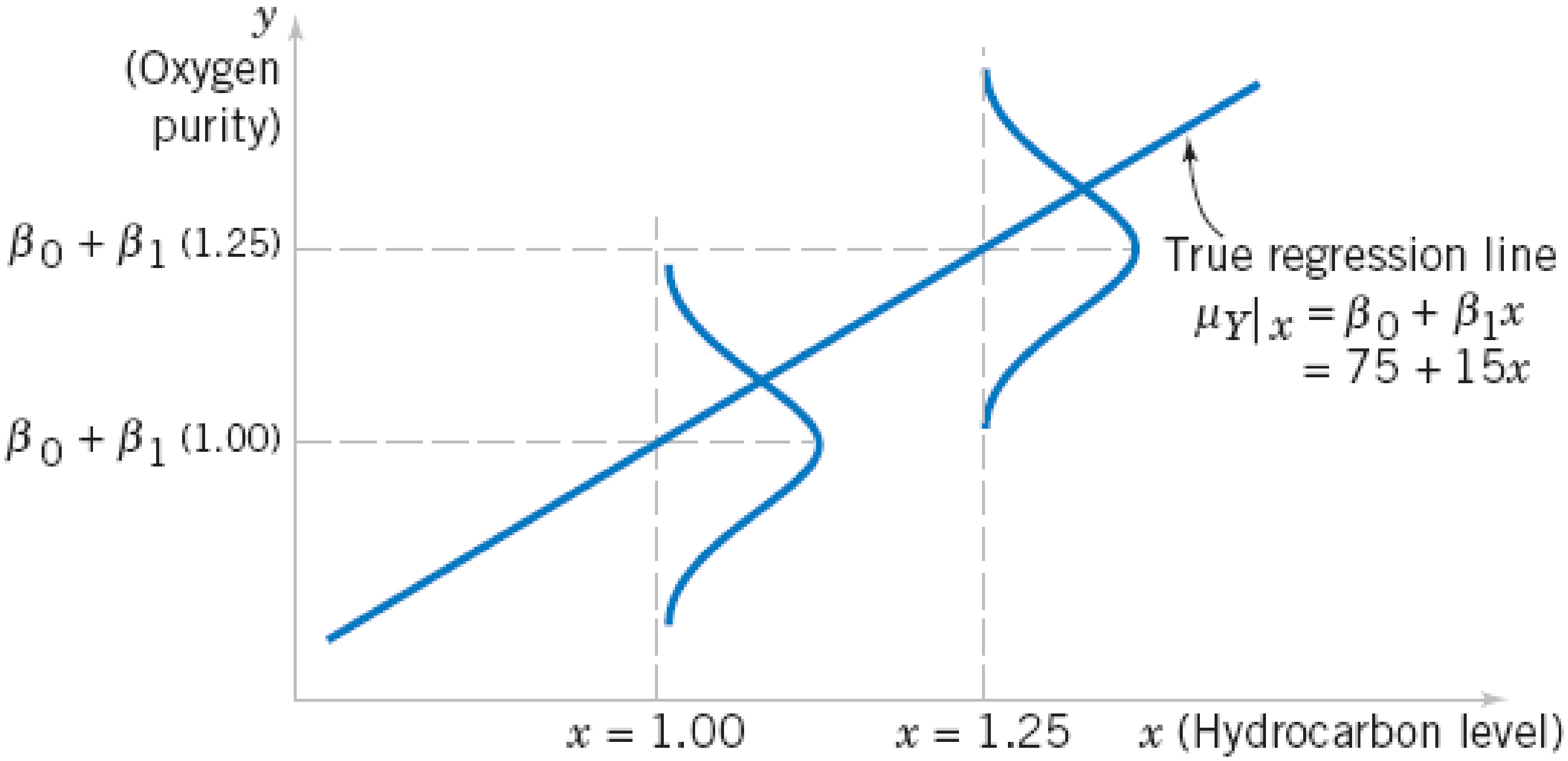


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity–hydrocarbon data.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

Suppose we wish to test

$$H_0\colon \beta_1 = \beta_{1,0}$$
$$H_1\colon \beta_1 \neq \beta_{1,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

The test statistic could also be written as:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

We would reject the null hypothesis if

$$|t_0| > t_{\alpha/2,n-2}$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

Suppose we wish to test

$$H_0\colon \beta_0 = \beta_{0,0}$$
$$H_1\colon \beta_0 \neq \beta_{0,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

We would reject the null hypothesis if

$$|t_0| > t_{\alpha/2,n-2}$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

An important special case of the hypotheses of Equation 11-18 is

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**. *Failure* to reject H_0 is equivalent to concluding that there is no linear relationship between *x* and *Y*.

11-4: Hypothesis Tests in Simple Linear Regression

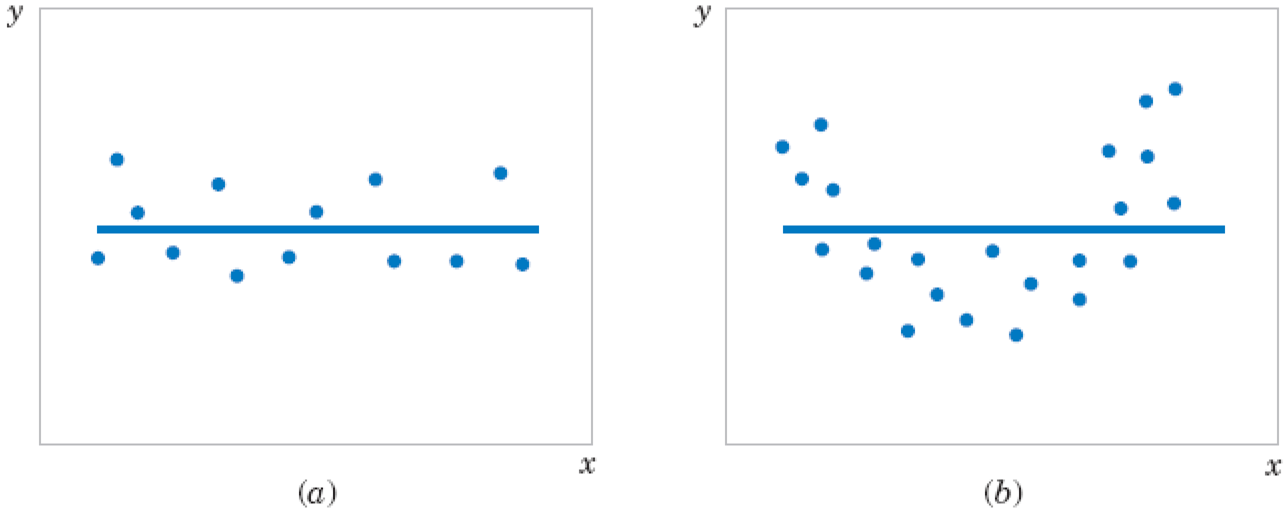


Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

11-4: Hypothesis Tests in Simple Linear Regression

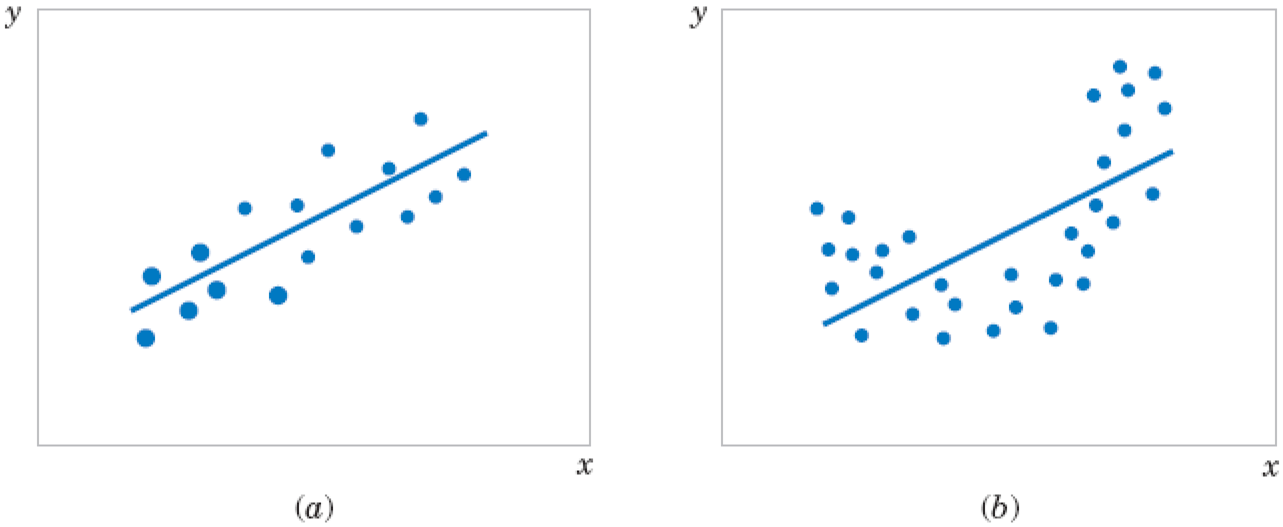
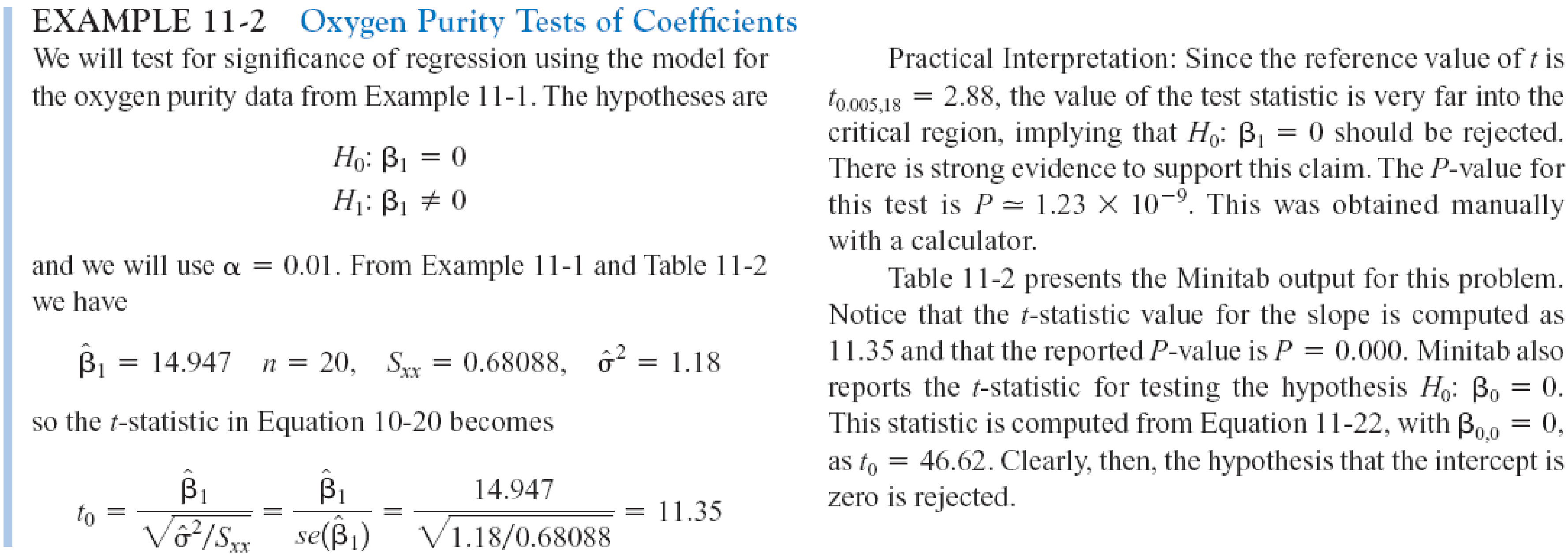


Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

11-4: Hypothesis Tests in Simple Linear Regression

Example 11-2



EXAMPLE 11-2 Oxygen Purity Tests of Coefficients

We will test for significance of regression using the model for the oxygen purity data from Example 11-1. The hypotheses are

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 \neq 0$$

and we will use $\alpha = 0.01$. From Example 11-1 and Table 11-2 we have

$$\hat{\beta}_1 = 14.947 \quad n = 20, \quad S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$$

so the t-statistic in Equation 10-20 becomes

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{14.947}{\sqrt{1.18/0.68088}} = 11.35$$

Practical Interpretation: Since the reference value of t is $t_{0.005,18} = 2.88$, the value of the test statistic is very far into the critical region, implying that $H_0\colon \beta_1 = 0$ should be rejected. There is strong evidence to support this claim. The P-value for this test is $P \simeq 1.23 \times 10^{-9}$. This was obtained manually with a calculator.

Table 11-2 presents the Minitab output for this problem. Notice that the t-statistic value for the slope is computed as 11.35 and that the reported P-value is $P = 0.000$. Minitab also reports the t-statistic for testing the hypothesis $H_0\colon \beta_0 = 0$. This statistic is computed from Equation 11-22, with $\beta_{0,0} = 0$, as $t_0 = 46.62$. Clearly, then, the hypothesis that the intercept is zero is rejected.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The analysis of variance identity is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (11\text{-}24)$$

Symbolically,

$$SS_T = SS_R + SS_E \qquad (11\text{-}25)$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

If the null hypothesis, H_0: $\beta_1 = 0$ is true, the statistic

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} \qquad (11\text{-}26)$$

follows the $F_{1,n\text{-}2}$ distribution and we would reject if $f_0 > f_{\alpha,1,n\text{-}2}$.

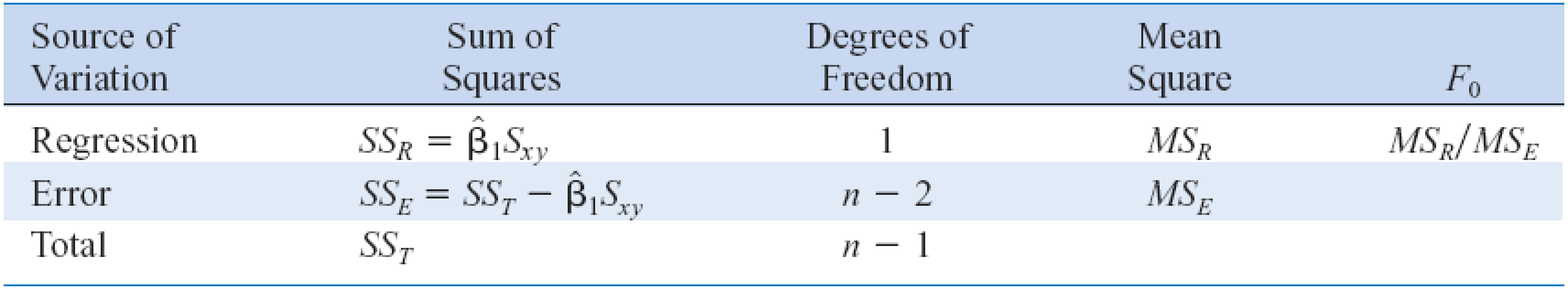
11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The quantities, MS_R and MS_E are called **mean squares**.

Analysis of variance table:

Table 11-3 Analysis of Variance for Testing Significance of Regression

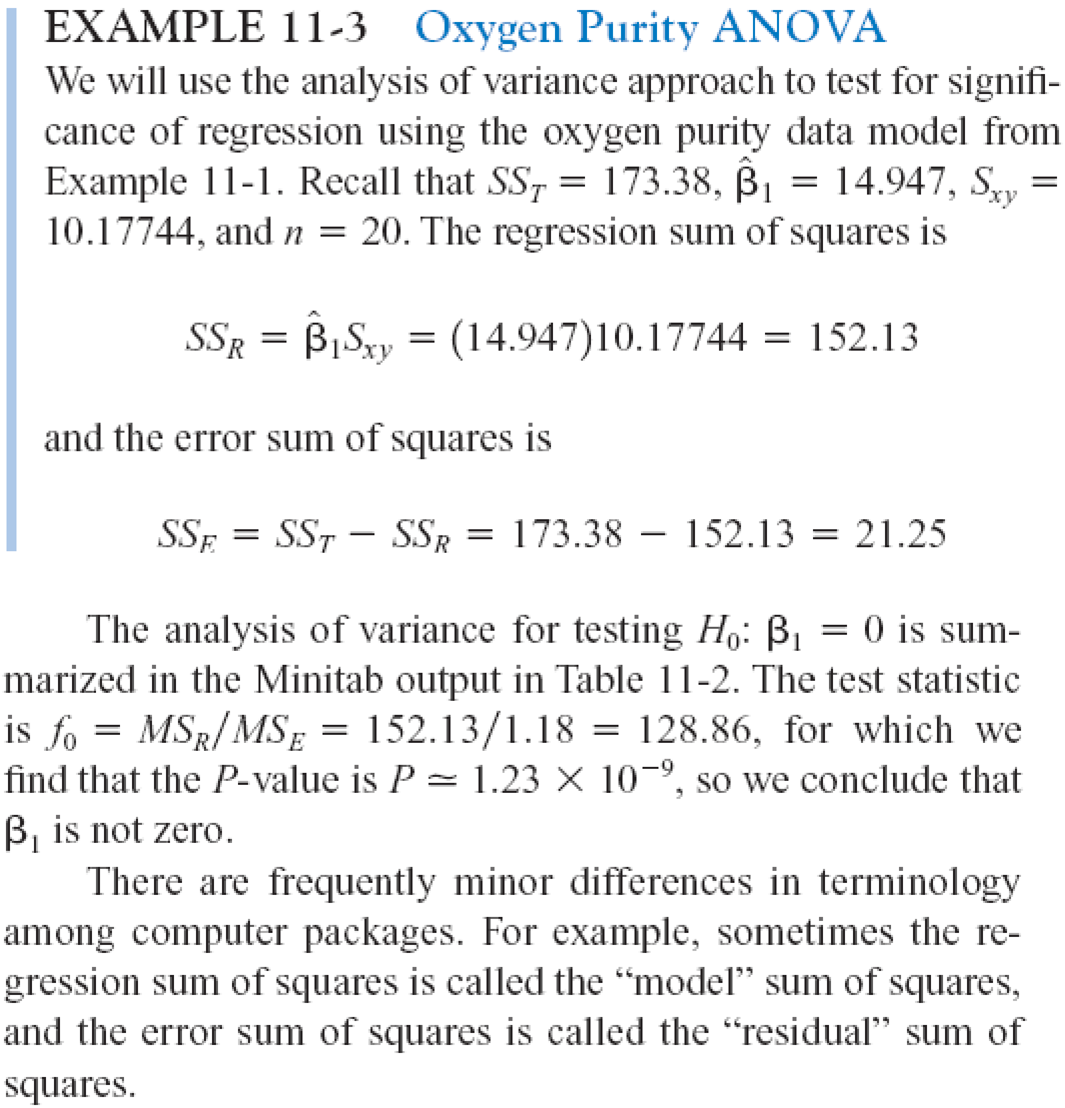


Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R / MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n - 2$	MS_E	
Total	SS_T	$n - 1$		

Note that $MS_E = \hat{\sigma}^2$.

11-4: Hypothesis Tests in Simple Linear Regression

Example 11-3



EXAMPLE 11-3 Oxygen Purity ANOVA

We will use the analysis of variance approach to test for significance of regression using the oxygen purity data model from Example 11-1. Recall that $SS_T = 173.38$, $\hat{\beta}_1 = 14.947$, $S_{xy} = 10.17744$, and $n = 20$. The regression sum of squares is

$$SS_R = \hat{\beta}_1 S_{xy} = (14.947)10.17744 = 152.13$$

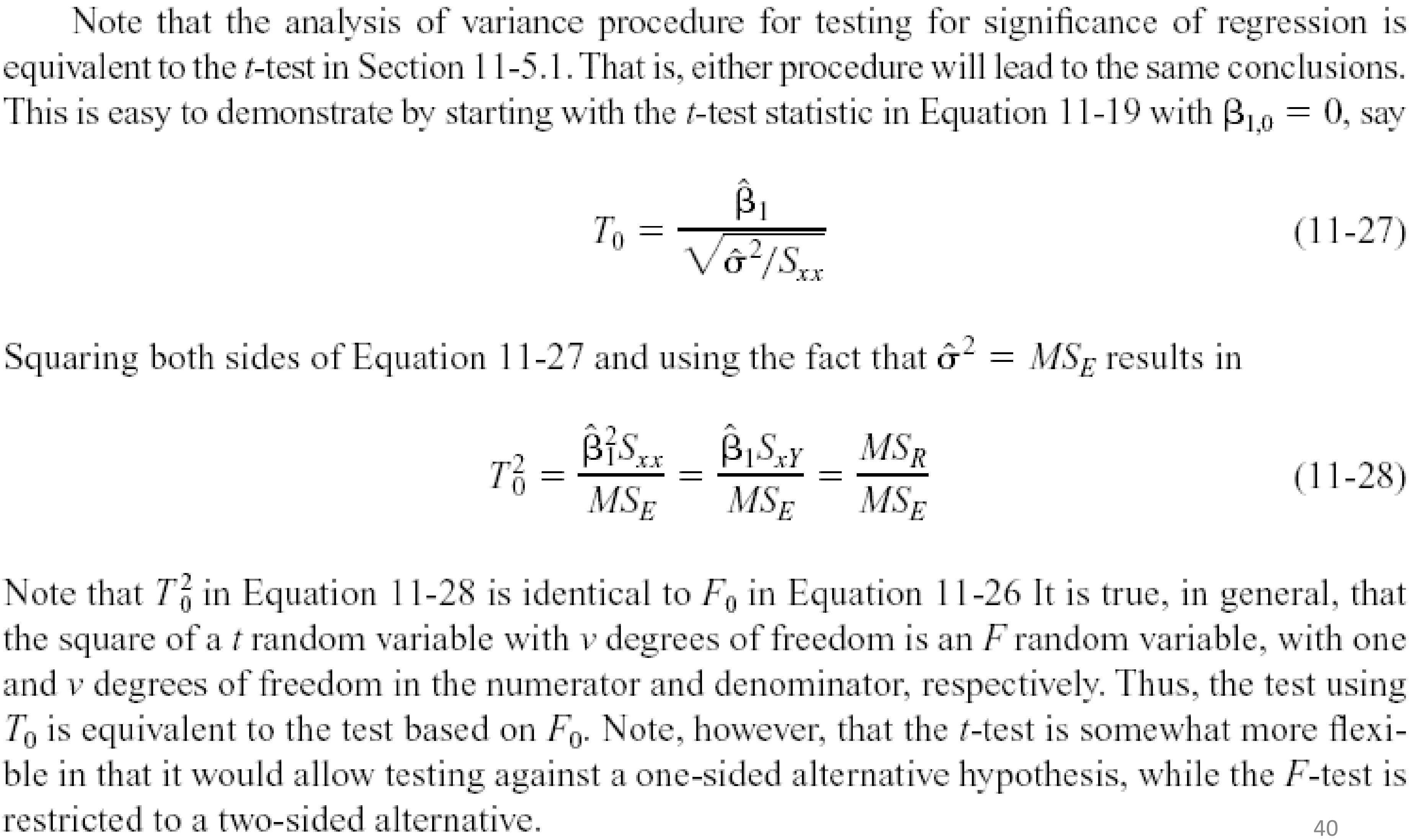
and the error sum of squares is

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

The analysis of variance for testing H_0: $\beta_1 = 0$ is summarized in the Minitab output in Table 11-2. The test statistic is $f_0 = MS_R/MS_E = 152.13/1.18 = 128.86$, for which we find that the P-value is $P \simeq 1.23 \times 10^{-9}$, so we conclude that β_1 is not zero.

There are frequently minor differences in terminology among computer packages. For example, sometimes the regression sum of squares is called the "model" sum of squares, and the error sum of squares is called the "residual" sum of squares.

11-4: Hypothesis Tests in Simple Linear Regression



Note that the analysis of variance procedure for testing for significance of regression is equivalent to the t-test in Section 11-5.1. That is, either procedure will lead to the same conclusions. This is easy to demonstrate by starting with the t-test statistic in Equation 11-19 with $\beta_{1,0} = 0$, say

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \qquad (11\text{-}27)$$

Squaring both sides of Equation 11-27 and using the fact that $\hat{\sigma}^2 = MS_E$ results in

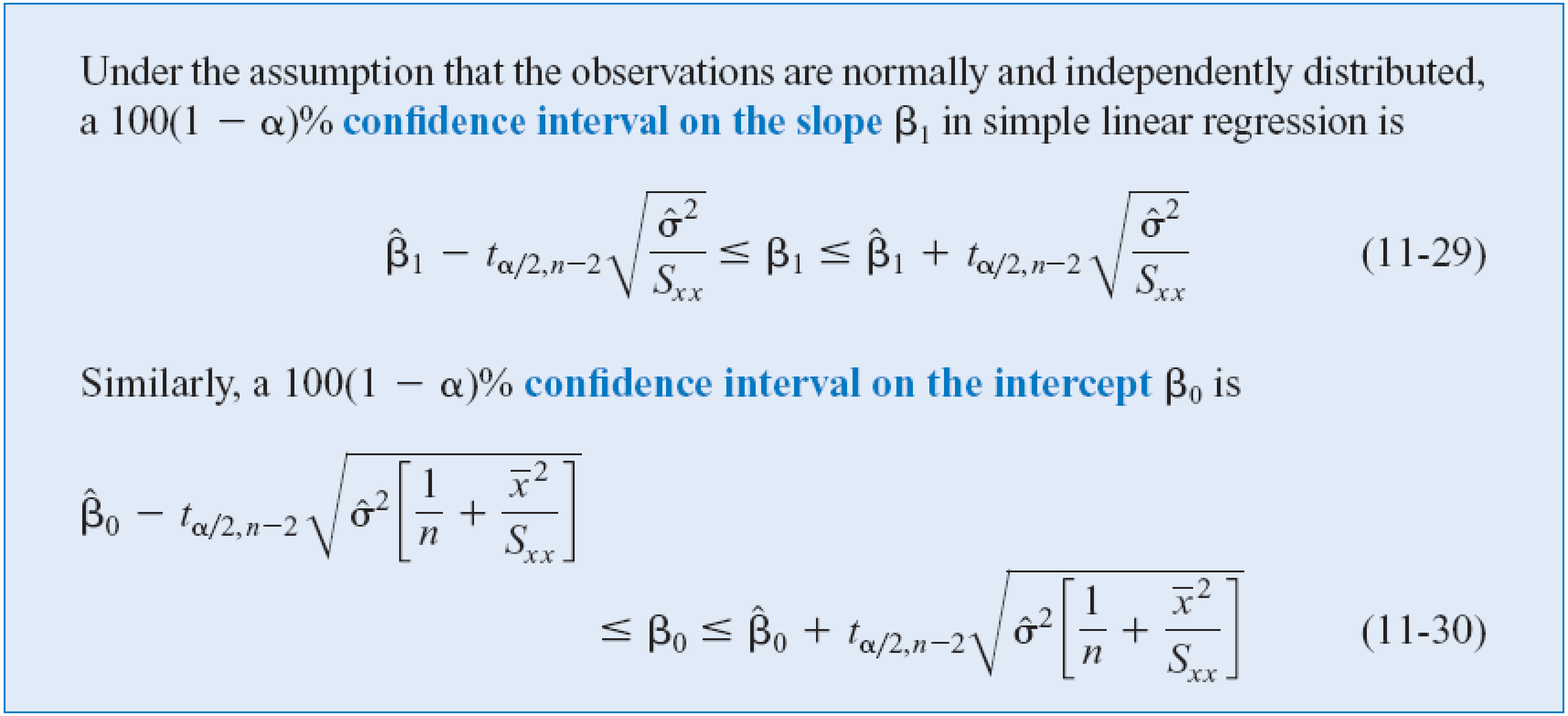
$$T_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_E} = \frac{\hat{\beta}_1 S_{xY}}{MS_E} = \frac{MS_R}{MS_E} \qquad (11\text{-}28)$$

Note that T_0^2 in Equation 11-28 is identical to F_0 in Equation 11-26 It is true, in general, that the square of a t random variable with v degrees of freedom is an F random variable, with one and v degrees of freedom in the numerator and denominator, respectively. Thus, the test using T_0 is equivalent to the test based on F_0. Note, however, that the t-test is somewhat more flexible in that it would allow testing against a one-sided alternative hypothesis, while the F-test is restricted to a two-sided alternative.

11-5: Confidence Intervals

11-5.1 Confidence Intervals on the Slope and Intercept

Definition



Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ **confidence interval on the slope** β_1 in simple linear regression is

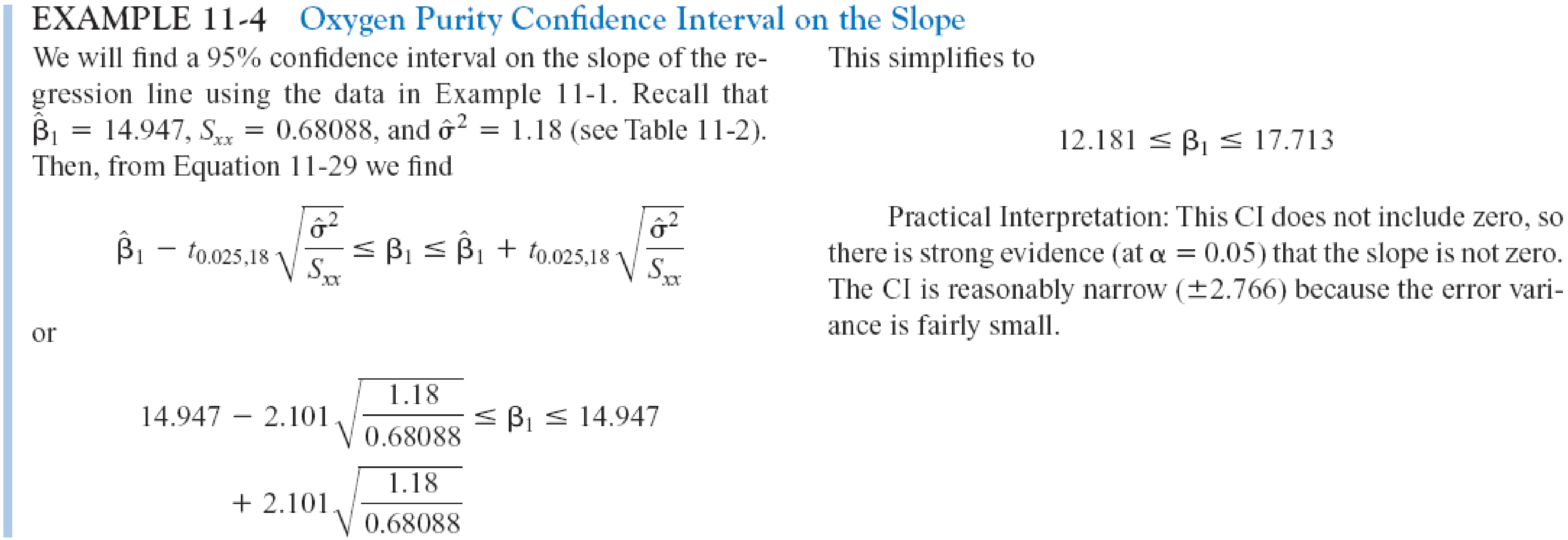
$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad (11\text{-}29)$$

Similarly, a $100(1-\alpha)\%$ **confidence interval on the intercept** β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]} \qquad (11\text{-}30)$$

11-5: Confidence Intervals

Example 11-4



EXAMPLE 11-4 Oxygen Purity Confidence Interval on the Slope

We will find a 95% confidence interval on the slope of the regression line using the data in Example 11-1. Recall that $\hat{\beta}_1 = 14.947$, $S_{xx} = 0.68088$, and $\hat{\sigma}^2 = 1.18$ (see Table 11-2). Then, from Equation 11-29 we find

$$\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

or

$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \leq \beta_1 \leq 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

This simplifies to

$$12.181 \leq \beta_1 \leq 17.713$$

Practical Interpretation: This CI does not include zero, so there is strong evidence (at $\alpha = 0.05$) that the slope is not zero. The CI is reasonably narrow (±2.766) because the error variance is fairly small.

11-5: Confidence Intervals

11-5.2 Confidence Interval on the Mean Response

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Definition

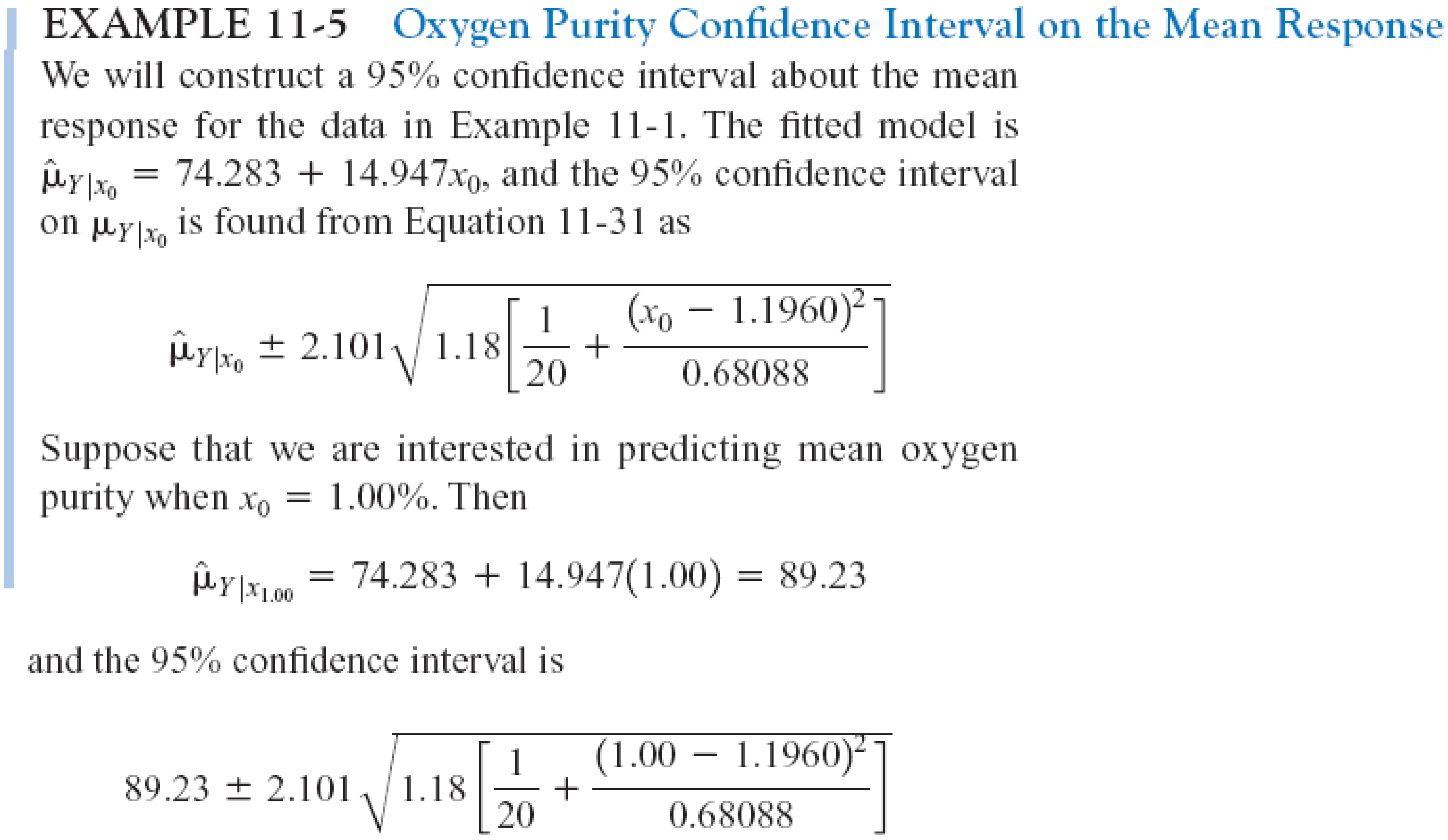
A $100(1-\alpha)\%$ **confidence interval about the mean response** at the value of $x = x_0$, say $\mu_{Y|x_0}$, is given by

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$
$$\leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \qquad (11\text{-}31)$$

where $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is computed from the fitted regression model.

11-5: Confidence Intervals

Example 11-5



EXAMPLE 11-5 Oxygen Purity Confidence Interval on the Mean Response

We will construct a 95% confidence interval about the mean response for the data in Example 11-1. The fitted model is $\hat{\mu}_{Y|x_0} = 74.283 + 14.947x_0$, and the 95% confidence interval on $\mu_{Y|x_0}$ is found from Equation 11-31 as

$$\hat{\mu}_{Y|x_0} \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(x_0 - 1.1960)^2}{0.68088}\right]}$$

Suppose that we are interested in predicting mean oxygen purity when $x_0 = 1.00\%$. Then

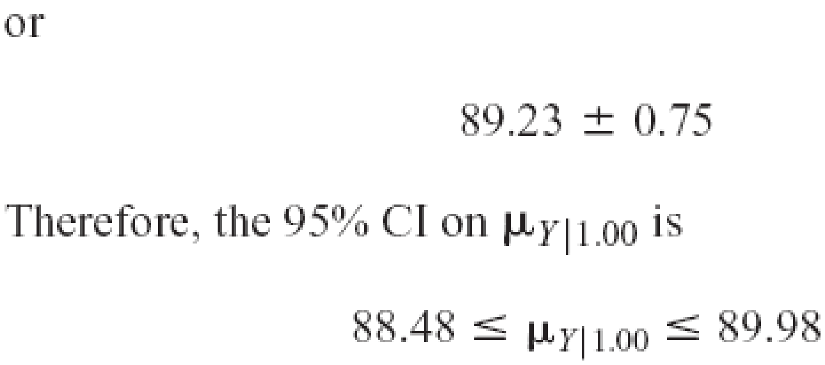
$$\hat{\mu}_{Y|x_{1.00}} = 74.283 + 14.947(1.00) = 89.23$$

and the 95% confidence interval is

$$89.23 \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

11-5: Confidence Intervals

Example 11-5

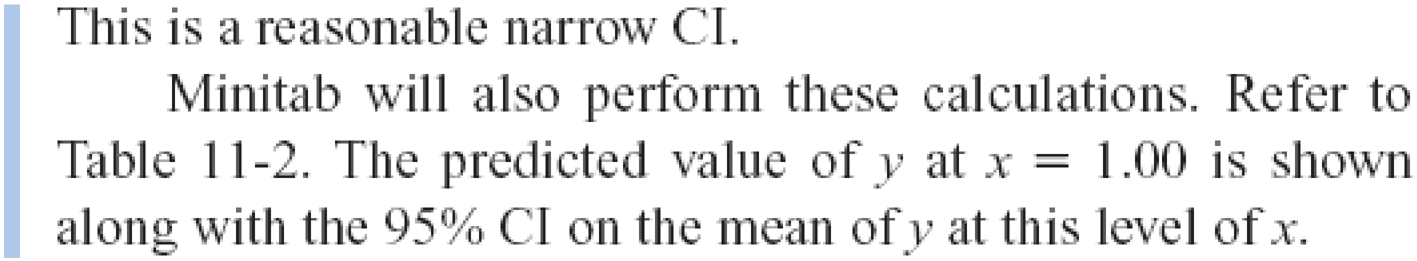


or

$$89.23 \pm 0.75$$

Therefore, the 95% CI on $\mu_{Y|1.00}$ is

$$88.48 \leq \mu_{Y|1.00} \leq 89.98$$

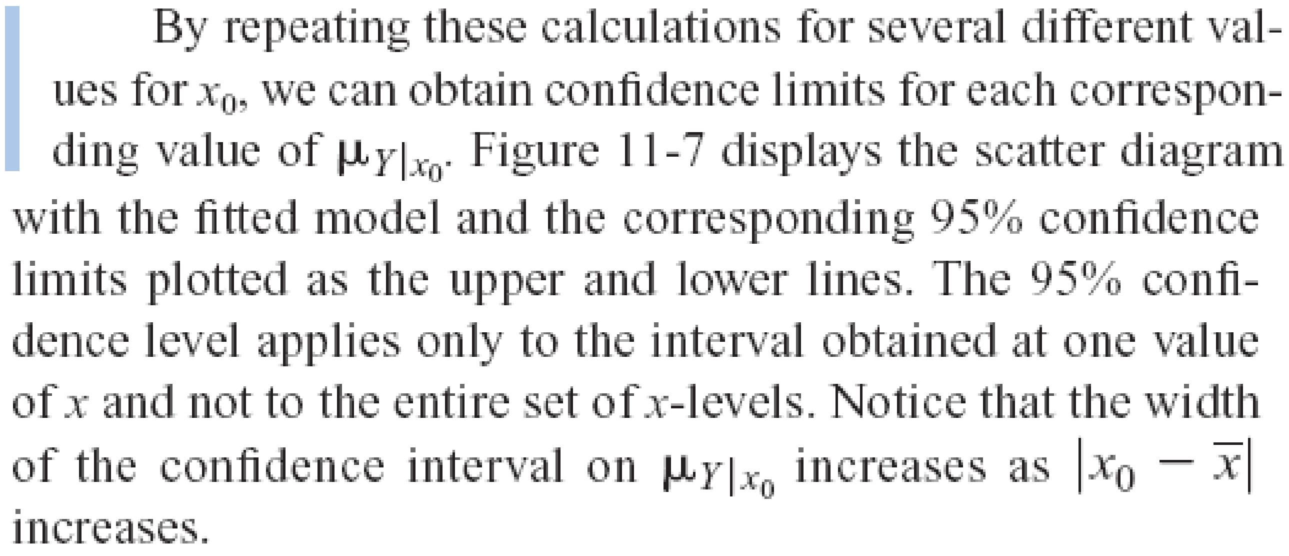


This is a reasonable narrow CI.

Minitab will also perform these calculations. Refer to Table 11-2. The predicted value of y at $x = 1.00$ is shown along with the 95% CI on the mean of y at this level of x.

11-5: Confidence Intervals

Example 11-5



By repeating these calculations for several different values for x_0, we can obtain confidence limits for each corresponding value of $\mu_{Y|x_0}$. Figure 11-7 displays the scatter diagram with the fitted model and the corresponding 95% confidence limits plotted as the upper and lower lines. The 95% confidence level applies only to the interval obtained at one value of x and not to the entire set of x-levels. Notice that the width of the confidence interval on $\mu_{Y|x_0}$ increases as $|x_0 - \overline{x}|$ increases.

11-5: Confidence Intervals

Figure 11-7

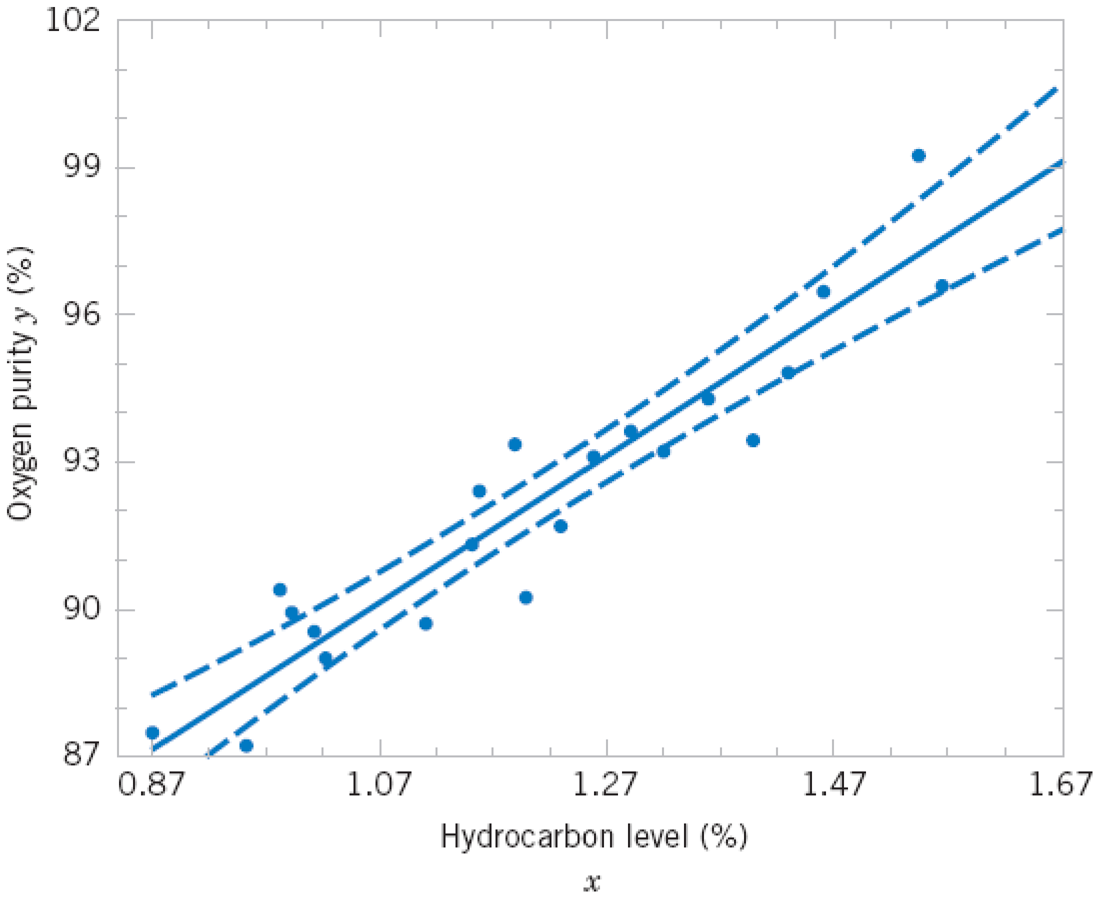


Figure 11-7 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line and 95 percent confidence limits on $\mu_{Y|x0}$.

11-6: Prediction of New Observations

If x_0 is the value of the regressor variable of interest,

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

is the point estimator of the new or future value of the response, Y_0.

11-6: Prediction of New Observations

Definition

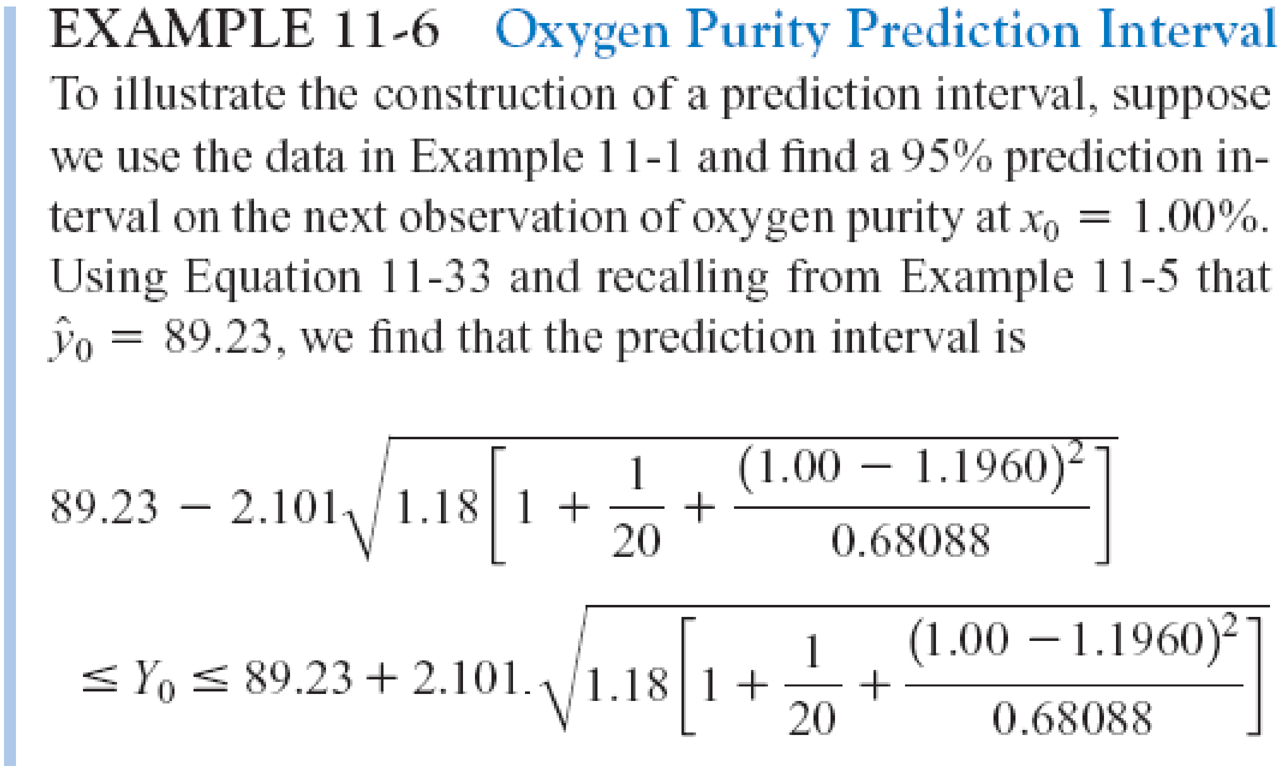
A $100(1-\alpha)\%$ **prediction interval on a future observation** Y_0 at the value x_0 is given by

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \qquad (11\text{-}33)$$

The value $\hat{y}_0$ is computed from the regression model $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

11-6: Prediction of New Observations

Example 11-6



EXAMPLE 11-6 Oxygen Purity Prediction Interval

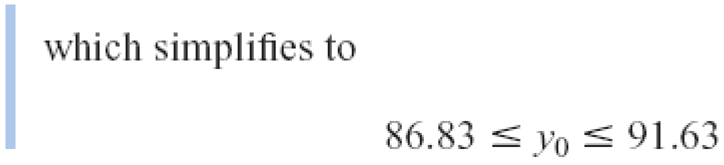
To illustrate the construction of a prediction interval, suppose we use the data in Example 11-1 and find a 95% prediction interval on the next observation of oxygen purity at $x_0 = 1.00\%$. Using Equation 11-33 and recalling from Example 11-5 that $\hat{y}_0 = 89.23$, we find that the prediction interval is

$$89.23 - 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

$$\leq Y_0 \leq 89.23 + 2.101.\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

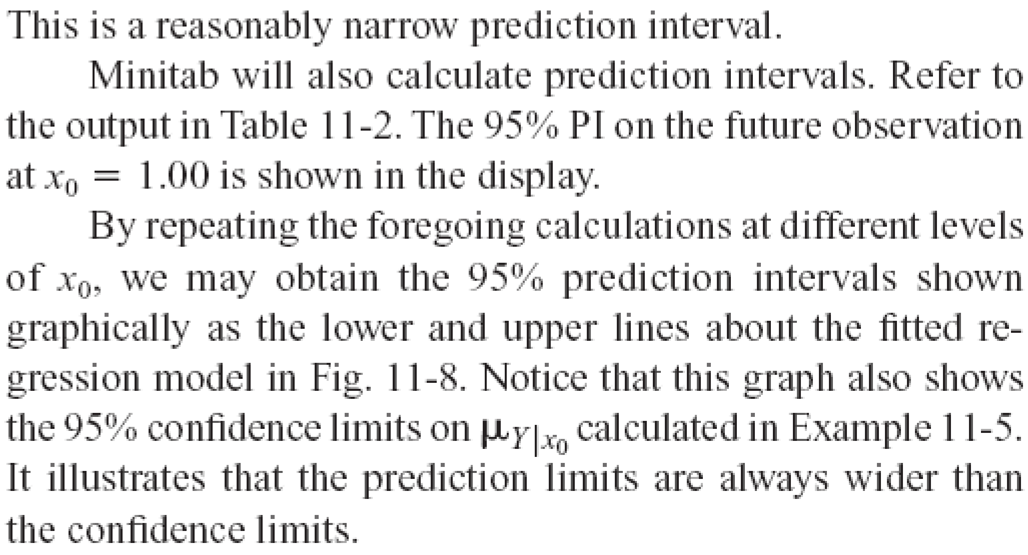
11-6: Prediction of New Observations

Example 11-6



which simplifies to

$$86.83 \leq y_0 \leq 91.63$$



This is a reasonably narrow prediction interval.

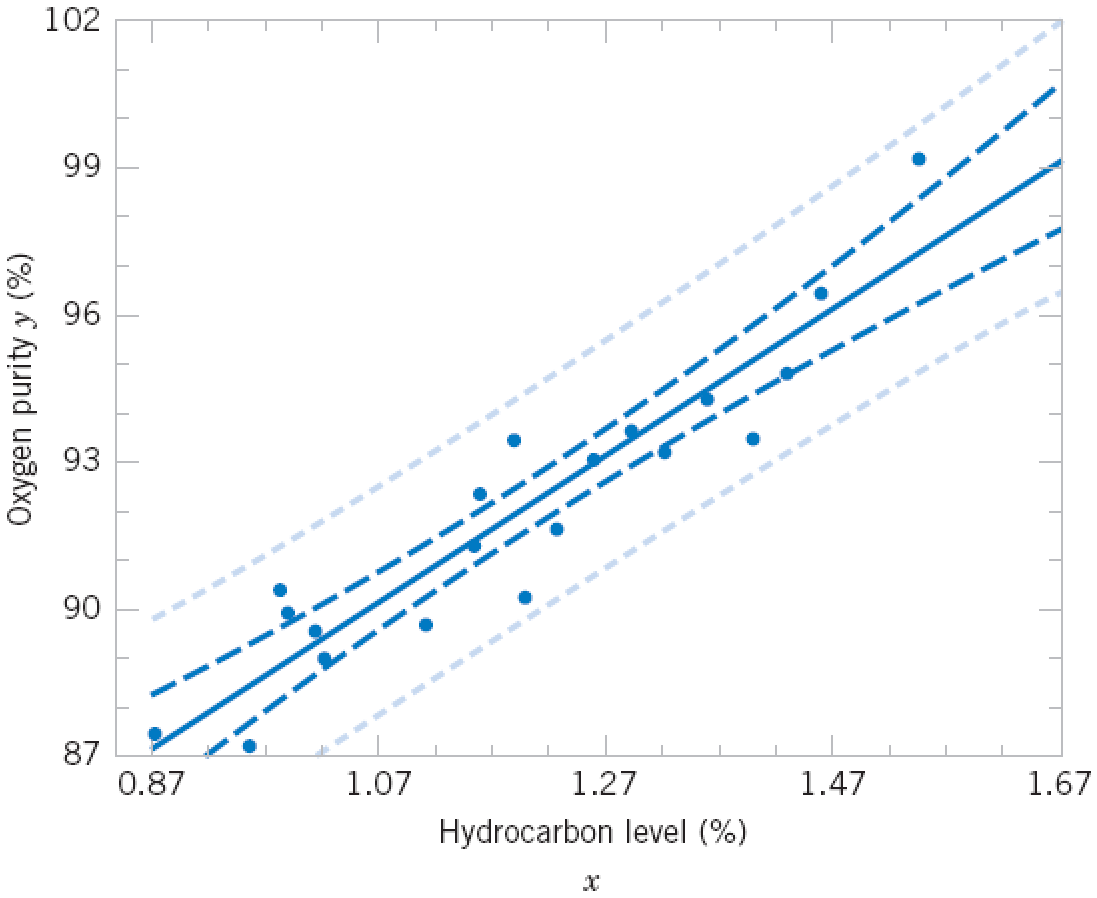
Minitab will also calculate prediction intervals. Refer to the output in Table 11-2. The 95% PI on the future observation at $x_0 = 1.00$ is shown in the display.

By repeating the foregoing calculations at different levels of x_0, we may obtain the 95% prediction intervals shown graphically as the lower and upper lines about the fitted regression model in Fig. 11-8. Notice that this graph also shows the 95% confidence limits on $\mu_{Y|x_0}$ calculated in Example 11-5. It illustrates that the prediction limits are always wider than the confidence limits.

11-6: Prediction of New Observations

Figure 11-8

Figure 11-8 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line, 95% prediction limits (outer lines) , and 95% confidence limits on $\mu_{Y|x0}$.



11-7: Adequacy of the Regression Model

- Fitting a regression model requires several **assumptions.**
  1. Errors are uncorrelated random variables with mean zero;
  2. Errors have constant variance; and,
  3. Errors be normally distributed.
- The analyst should always consider the validity of these assumptions to be doubtful and conduct analyses to examine the adequacy of the model

11-7: Adequacy of the Regression Model

11-7.1 Residual Analysis

- The **residuals** from a regression model are $e_i = y_i - \hat{y}_i$, where y_i is an actual observation and $\hat{y}_i$ is the corresponding fitted value from the regression model.
- Analysis of the residuals is frequently helpful in checking the assumption that the errors are approximately normally distributed with constant variance, and in determining whether additional terms in the model would be useful.

11-7: Adequacy of the Regression Model

11-7.1 Residual Analysis

Figure 11-9 Patterns for residual plots. (a) satisfactory, (b) funnel, (c) double bow, (d) nonlinear.
[Adapted from Montgomery, Peck, and Vining (2006).]

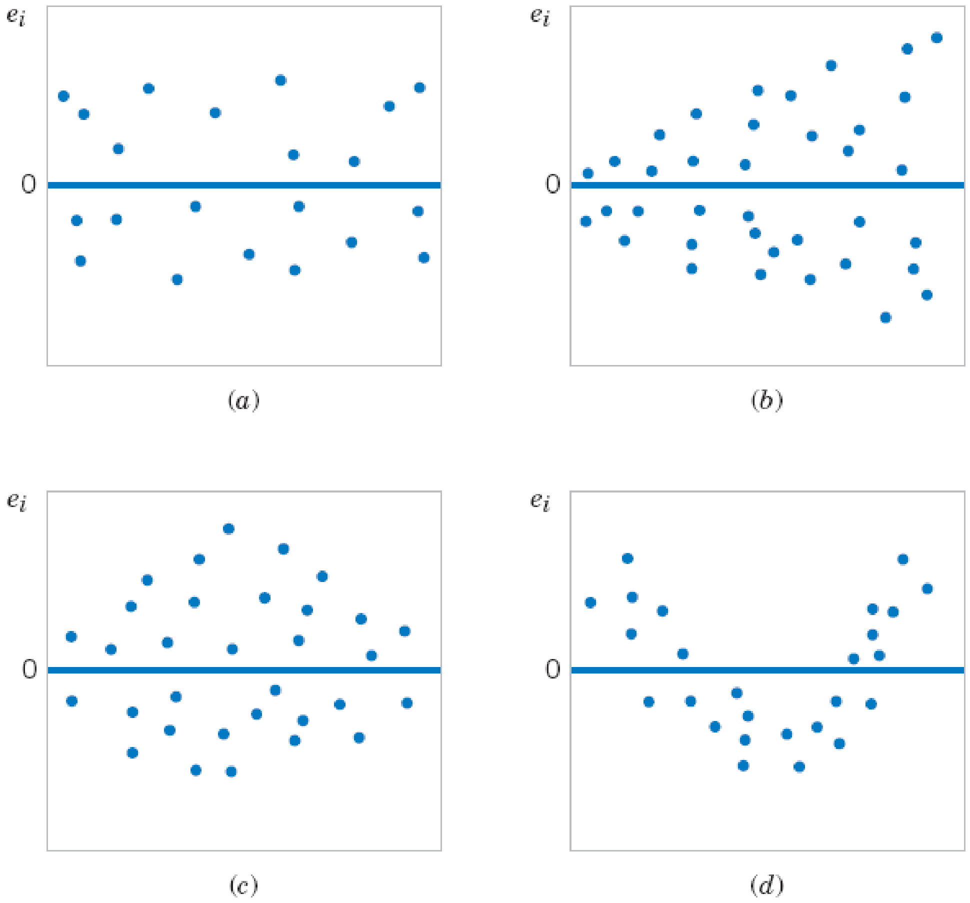
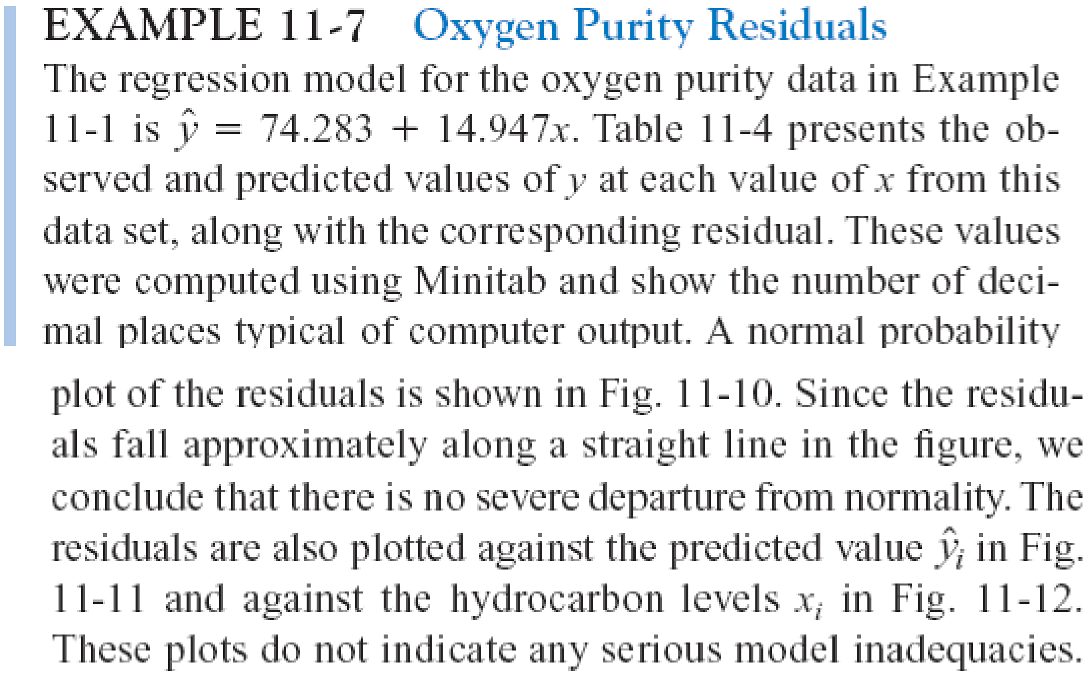


Figure 11-9 Patterns for residual plots. (a) Satisfactory, (b) Funnel, (c) Double bow, (d) Nonlinear. [Adapted from Montgomery, Peck, and Vining (2006).]

11-7: Adequacy of the Regression Model

Example 11-7



EXAMPLE 11-7 Oxygen Purity Residuals

The regression model for the oxygen purity data in Example 11-1 is $\hat{y} = 74.283 + 14.947x$. Table 11-4 presents the observed and predicted values of y at each value of x from this data set, along with the corresponding residual. These values were computed using Minitab and show the number of decimal places typical of computer output. A normal probability plot of the residuals is shown in Fig. 11-10. Since the residuals fall approximately along a straight line in the figure, we conclude that there is no severe departure from normality. The residuals are also plotted against the predicted value $\hat{y}_i$ in Fig. 11-11 and against the hydrocarbon levels x_i in Fig. 11-12. These plots do not indicate any serious model inadequacies.

11-7: Adequacy of the Regression Model

Example 11-7

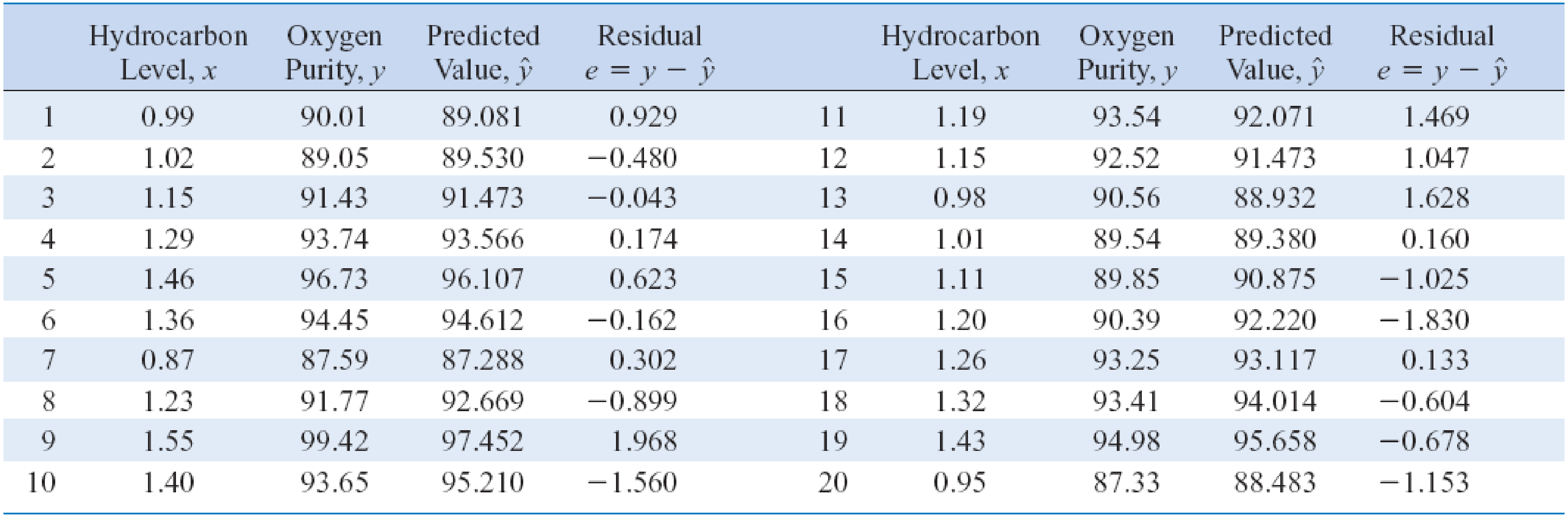


Table 11-4 Oxygen Purity Data from Example 11-1, Predicted Values, and Residuals

	Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, $\hat{y}$	Residual $e = y - \hat{y}$		Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, $\hat{y}$	Residual $e = y - \hat{y}$
1	0.99	90.01	89.081	0.929	11	1.19	93.54	92.071	1.469
2	1.02	89.05	89.530	−0.480	12	1.15	92.52	91.473	1.047
3	1.15	91.43	91.473	−0.043	13	0.98	90.56	88.932	1.628
4	1.29	93.74	93.566	0.174	14	1.01	89.54	89.380	0.160
5	1.46	96.73	96.107	0.623	15	1.11	89.85	90.875	−1.025
6	1.36	94.45	94.612	−0.162	16	1.20	90.39	92.220	−1.830
7	0.87	87.59	87.288	0.302	17	1.26	93.25	93.117	0.133
8	1.23	91.77	92.669	−0.899	18	1.32	93.41	94.014	−0.604
9	1.55	99.42	97.452	1.968	19	1.43	94.98	95.658	−0.678
10	1.40	93.65	95.210	−1.560	20	0.95	87.33	88.483	−1.153

11-7: Adequacy of the Regression Model

Example 11-7

Figure 11-10 Normal probability plot of residuals, Example 11-7.

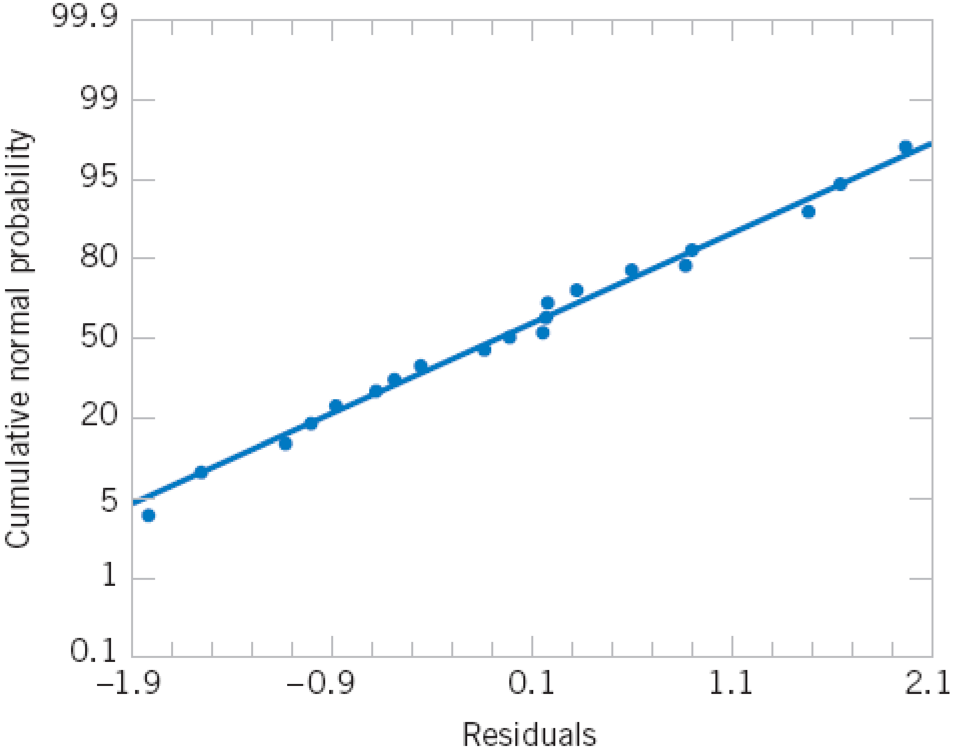


Figure 11-10 Normal probability plot of residuals, Example 11-7.

11-7: Adequacy of the Regression Model

Example 11-7

Figure 11-11 Plot of residuals versus predicted oxygen purity, $\hat{y}$, Example 11-7.

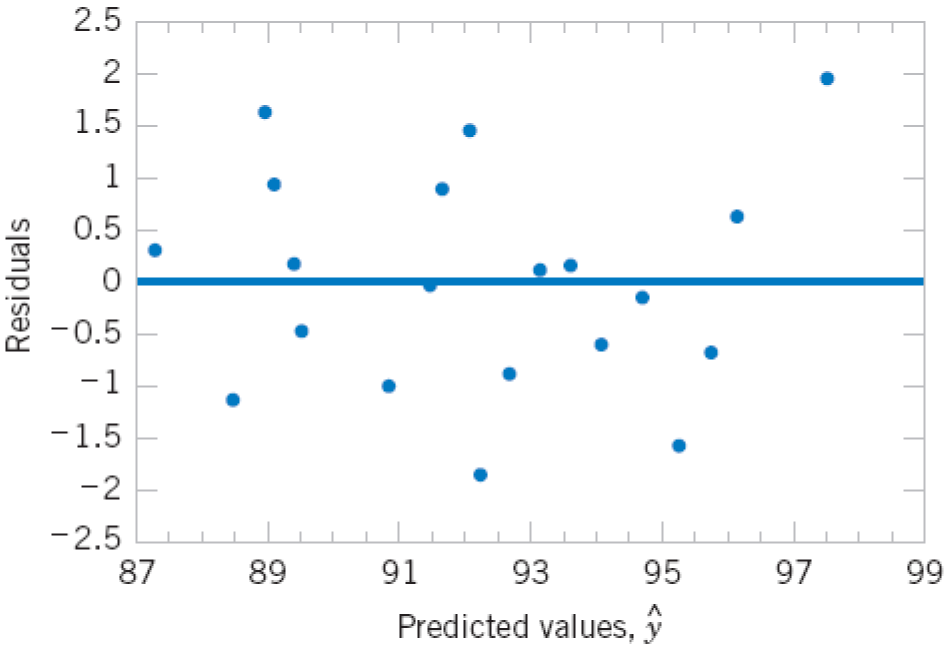


Figure 11-11 Plot of residuals versus predicted oxygen purity $\hat{y}$, Example 11-7.

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \leq R^2 \leq 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

11-7: Adequacy of the Regression Model

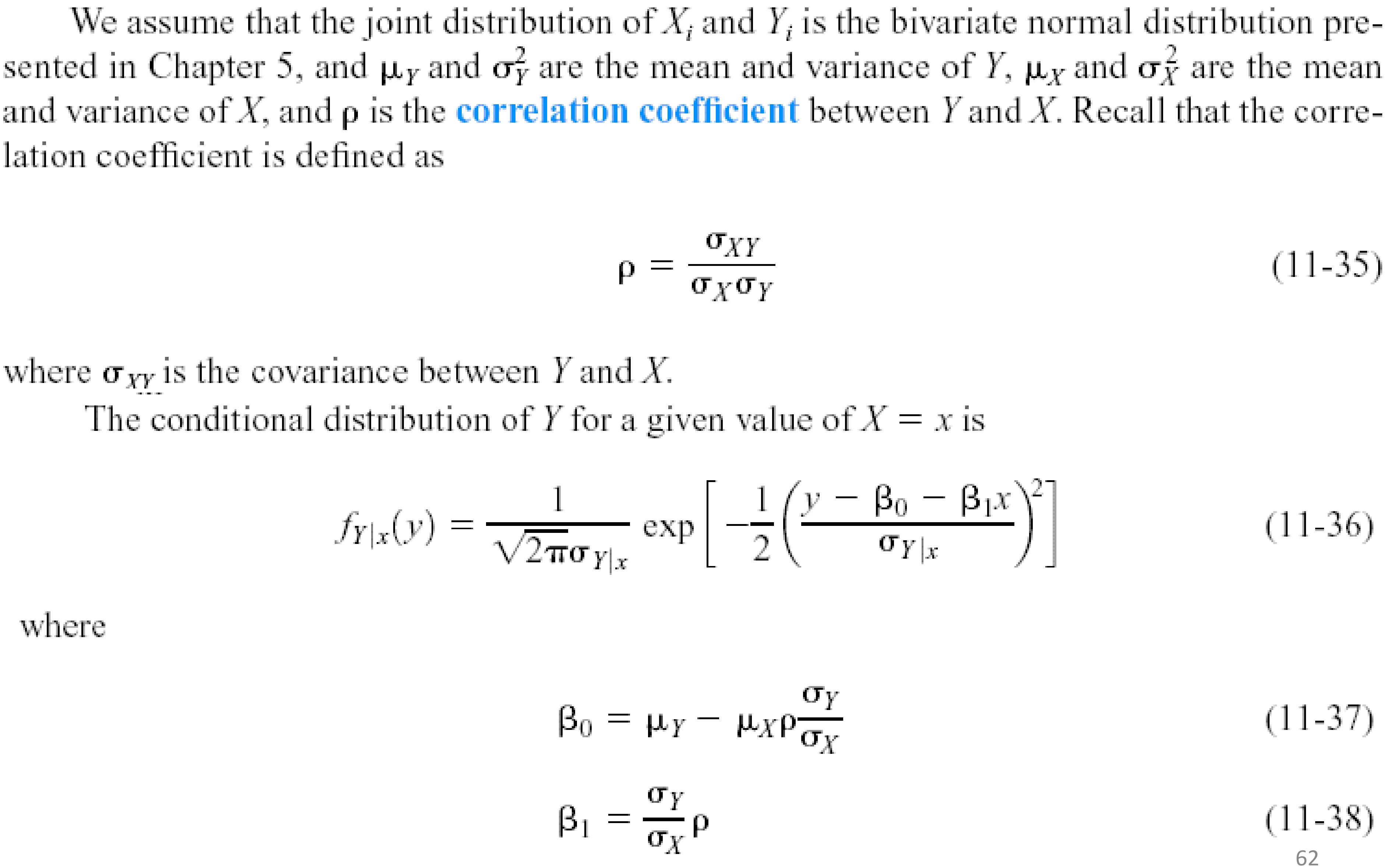
11-7.2 Coefficient of Determination (R^2)

- For the oxygen purity regression model,

$$\begin{aligned} R^2 &= SS_R/SS_T \\ &= 152.13/173.38 \\ &= 0.877 \end{aligned}$$

- Thus, the model accounts for 87.7% of the variability in the data.

11-8: Correlation



We assume that the joint distribution of X_i and Y_i is the bivariate normal distribution presented in Chapter 5, and μ_Y and σ_Y^2 are the mean and variance of Y, μ_X and σ_X^2 are the mean and variance of X, and ρ is the **correlation coefficient** between Y and X. Recall that the correlation coefficient is defined as

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{11-35}$$

where σ_{XY} is the covariance between Y and X.

The conditional distribution of Y for a given value of $X = x$ is

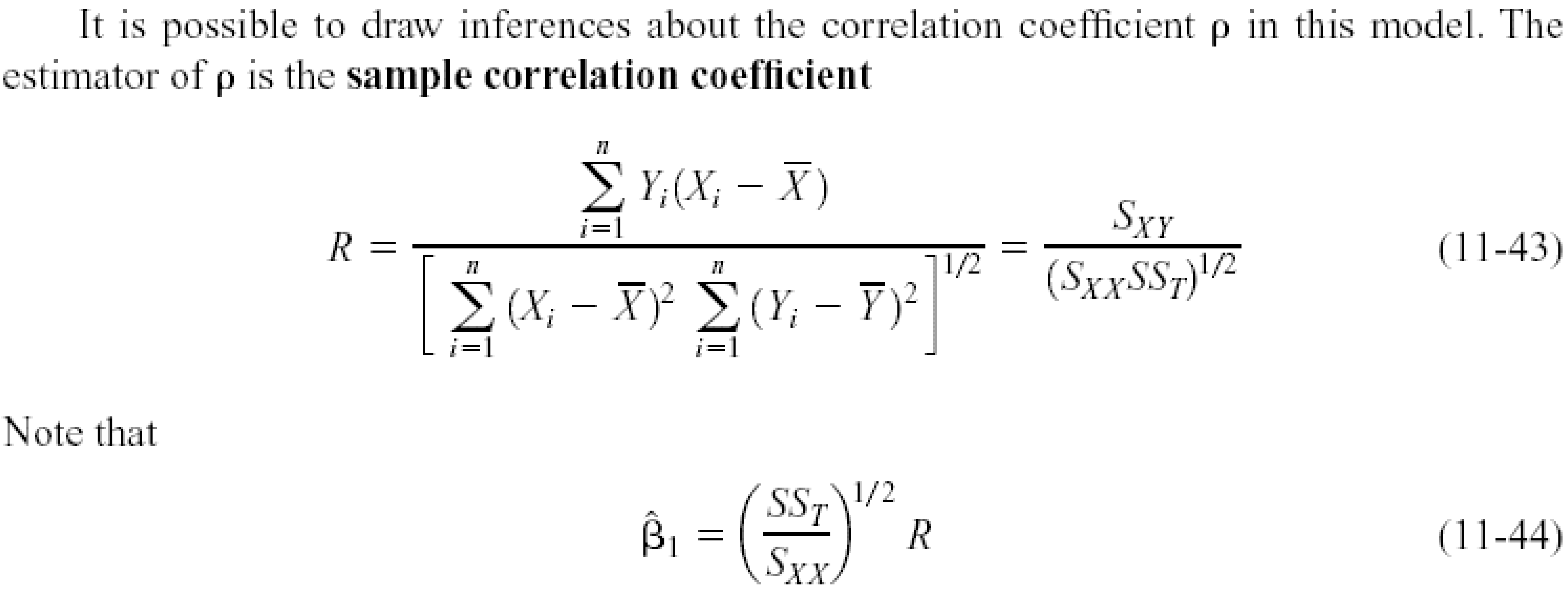
$$f_{Y|x}(y) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp\left[-\frac{1}{2}\left(\frac{y - \beta_0 - \beta_1 x}{\sigma_{Y|x}}\right)^2\right] \tag{11-36}$$

where

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X} \tag{11-37}$$

$$\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho \tag{11-38}$$

11-8: Correlation



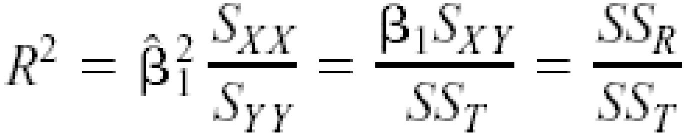
It is possible to draw inferences about the correlation coefficient ρ in this model. The estimator of ρ is the **sample correlation coefficient**

$$R = \frac{\sum_{i=1}^{n} Y_i(X_i - \overline{X})}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]^{1/2}} = \frac{S_{XY}}{(S_{XX}SS_T)^{1/2}} \qquad (11\text{-}43)$$

Note that

$$\hat{\beta}_1 = \left(\frac{SS_T}{S_{XX}}\right)^{1/2} R \qquad (11\text{-}44)$$

We may also write:



$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T}$$

11-8: Correlation

It is often useful to test the hypotheses

$$H_0\colon \rho = 0$$
$$H_1\colon \rho \neq 0$$

The appropriate test statistic for these hypotheses is

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \qquad (11\text{-}46)$$

Reject H_0 if $|t_0| > t_{\alpha/2,n-2}$.

11-8: Correlation

The test procedure for the hypothesis

$$H_0\colon \rho = \rho_0$$
$$H_1\colon \rho \neq \rho_0$$

where $\rho_0 \neq 0$ is somewhat more complicated. In this case, the appropriate test statistic is

$$Z_0 = (\operatorname{arctanh} R - \operatorname{arctanh} \rho_0)(n - 3)^{1/2} \qquad (11\text{-}49)$$

Reject H_0 if $|z_0| > z_{\alpha/2}$.

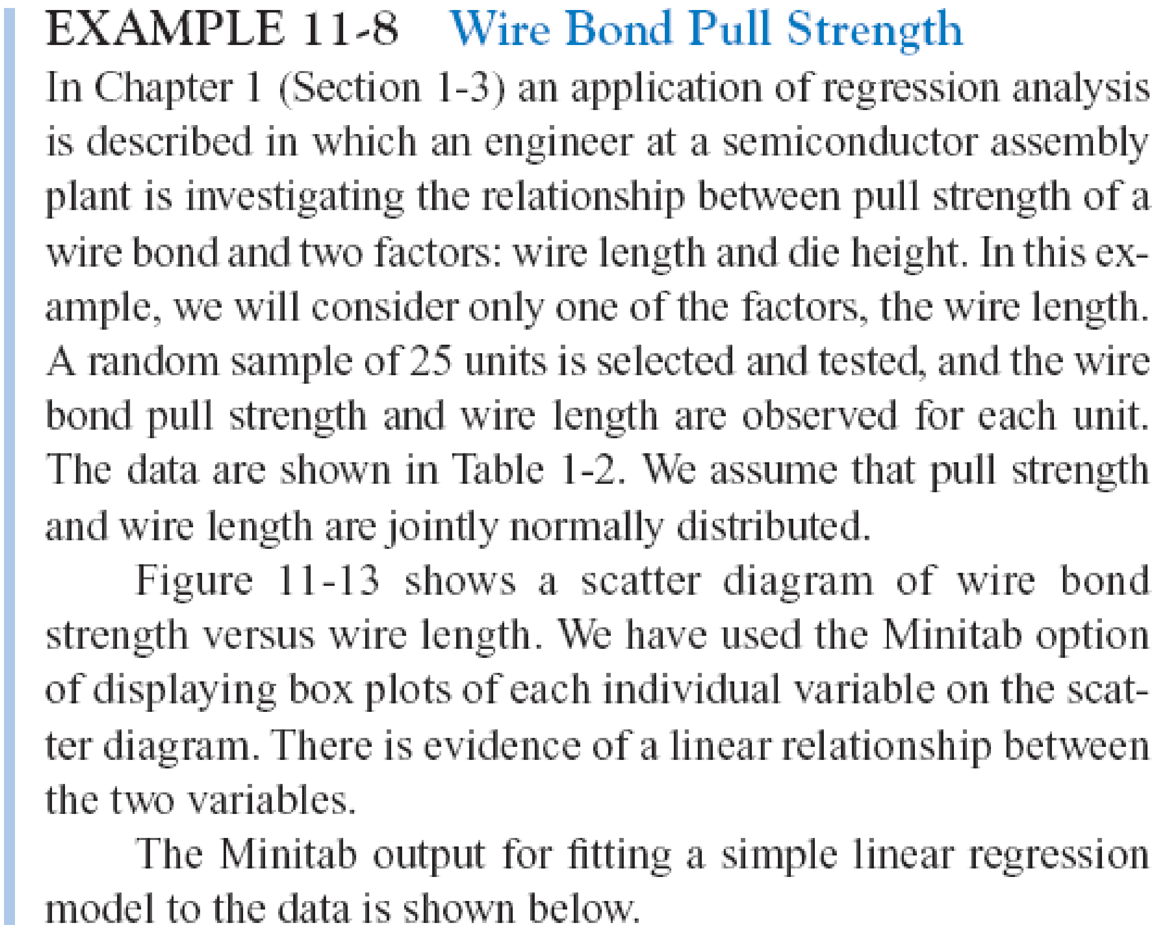
11-8: Correlation

The approximate 100(1- α)% confidence interval is

$$\tanh\left(\operatorname{arctanh} r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(\operatorname{arctanh} r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \qquad (11\text{-}50)$$

11-8: Correlation

Example 11-8



EXAMPLE 11-8 Wire Bond Pull Strength

In Chapter 1 (Section 1-3) an application of regression analysis is described in which an engineer at a semiconductor assembly plant is investigating the relationship between pull strength of a wire bond and two factors: wire length and die height. In this example, we will consider only one of the factors, the wire length. A random sample of 25 units is selected and tested, and the wire bond pull strength and wire length are observed for each unit. The data are shown in Table 1-2. We assume that pull strength and wire length are jointly normally distributed.

Figure 11-13 shows a scatter diagram of wire bond strength versus wire length. We have used the Minitab option of displaying box plots of each individual variable on the scatter diagram. There is evidence of a linear relationship between the two variables.

The Minitab output for fitting a simple linear regression model to the data is shown below.

11-8: Correlation

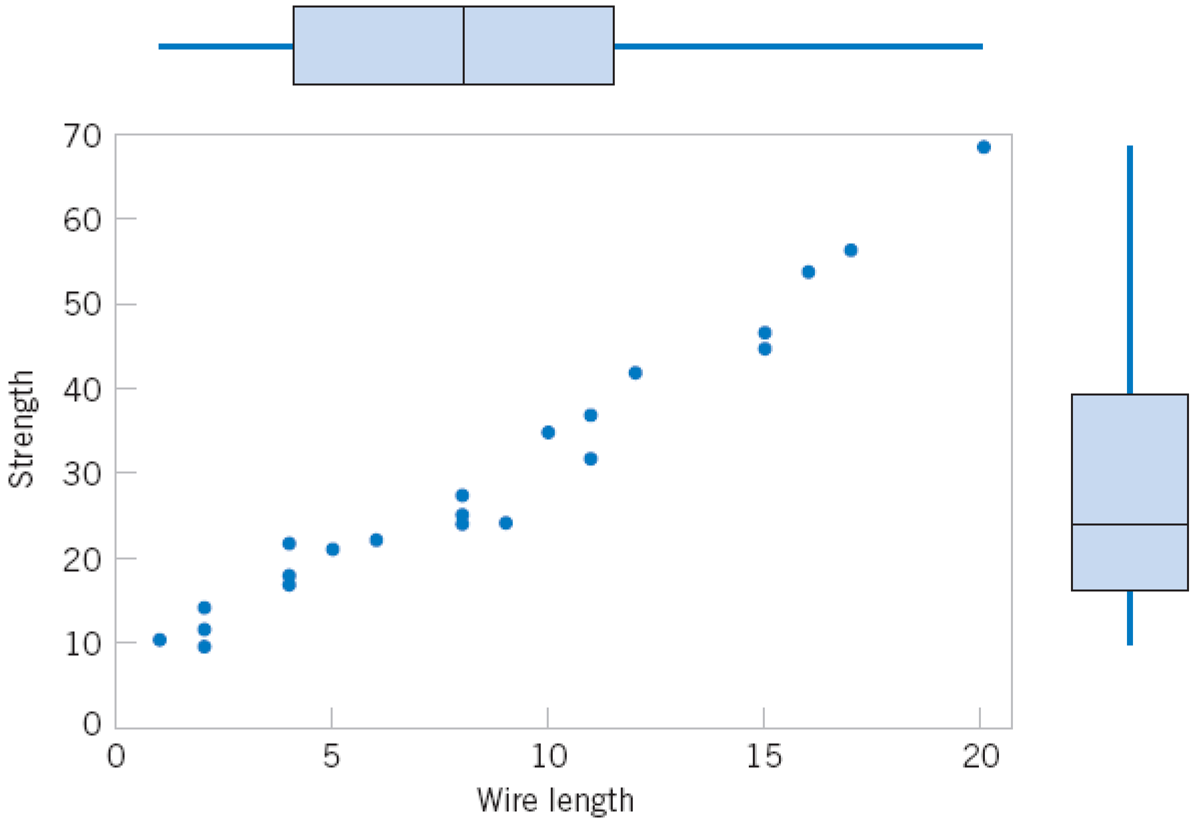
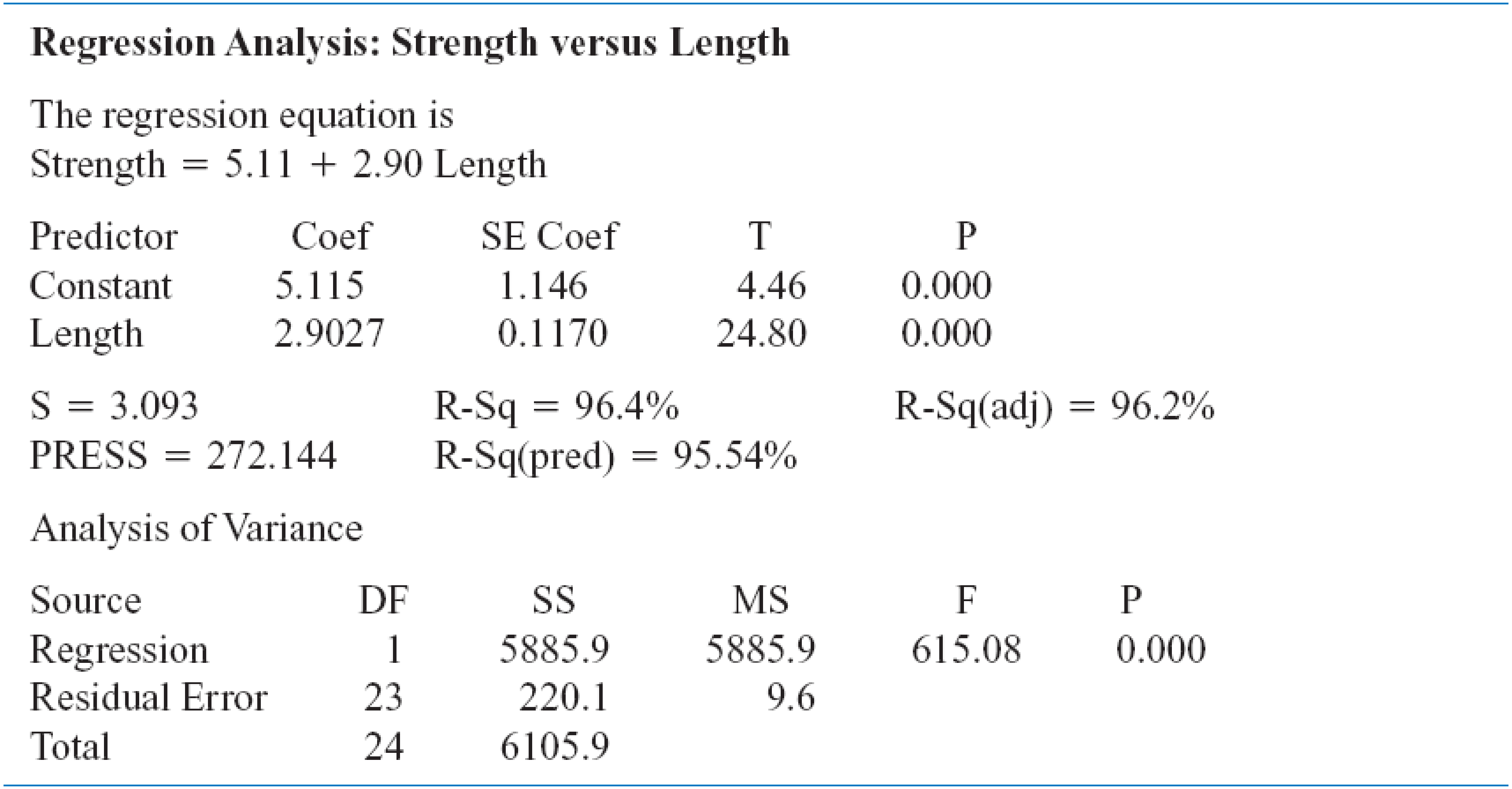


Figure 11-13 Scatter plot of wire bond strength versus wire length, Example 11-8.

Figure 11-13 Scatter plot of wire bond strength versus wire length, Example 11-8.

11-8: Correlation

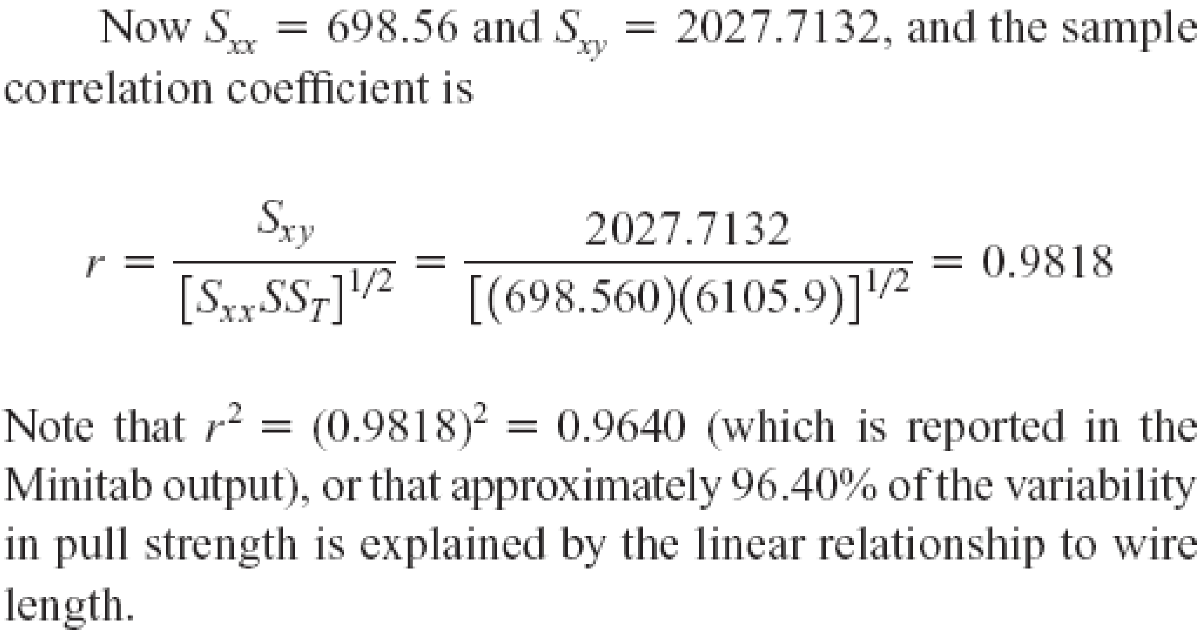
Minitab Output for Example 11-8



```
Regression Analysis: Strength versus Length

The regression equation is
Strength = 5.11 + 2.90 Length

Predictor      Coef     SE Coef       T        P
Constant      5.115      1.146      4.46    0.000
Length       2.9027     0.1170     24.80    0.000

S = 3.093               R-Sq = 96.4%             R-Sq(adj) = 96.2%
PRESS = 272.144         R-Sq(pred) = 95.54%

Analysis of Variance

Source            DF       SS        MS        F        P
Regression         1    5885.9    5885.9    615.08    0.000
Residual Error    23     220.1       9.6
Total             24    6105.9
```

11-8: Correlation

Example 11-8 (continued)



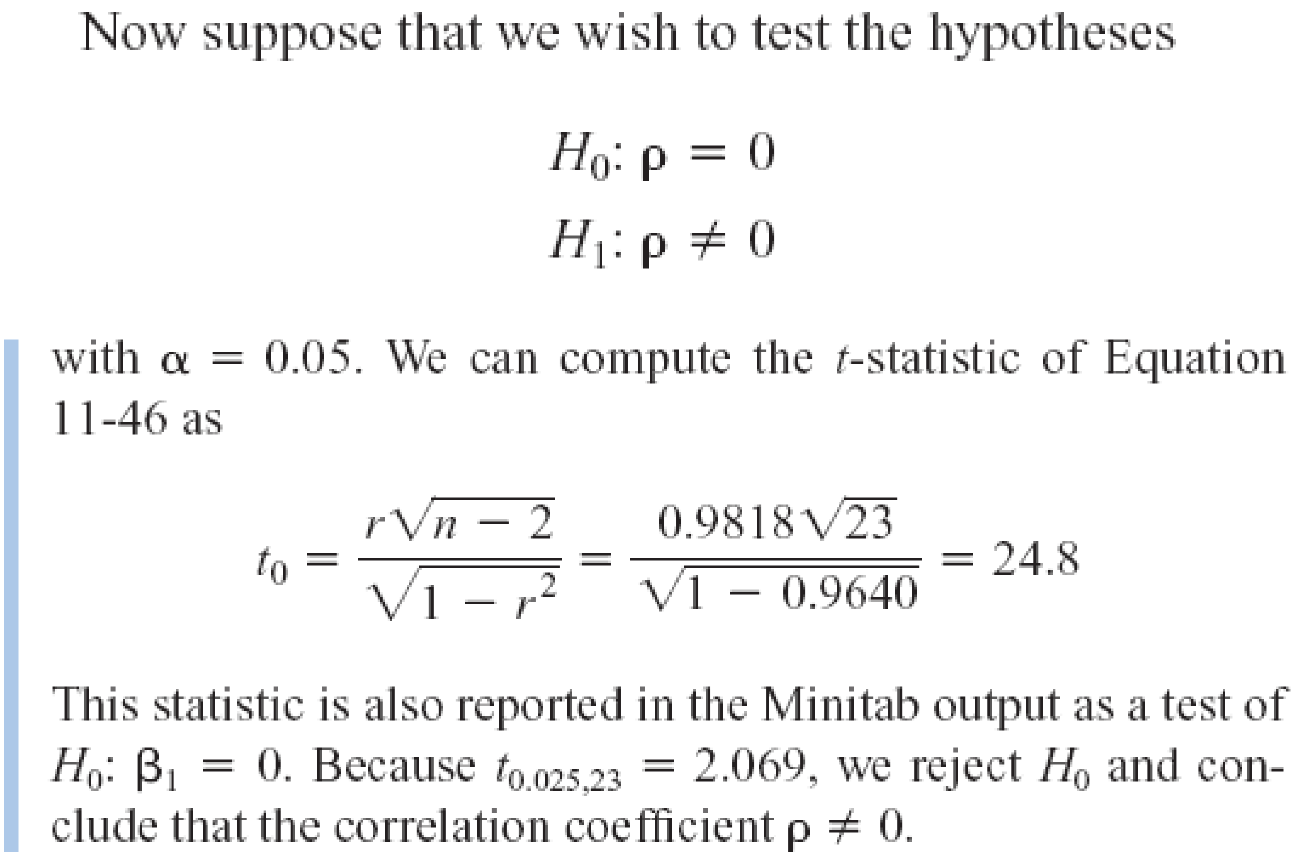
Now $S_{xx} = 698.56$ and $S_{xy} = 2027.7132$, and the sample correlation coefficient is

$$r = \frac{S_{xy}}{[S_{xx}SS_T]^{1/2}} = \frac{2027.7132}{[(698.560)(6105.9)]^{1/2}} = 0.9818$$

Note that $r^2 = (0.9818)^2 = 0.9640$ (which is reported in the Minitab output), or that approximately 96.40% of the variability in pull strength is explained by the linear relationship to wire length.

11-8: Correlation

Example 11-8 (continued)



Now suppose that we wish to test the hypotheses

$$H_0\colon \rho = 0$$
$$H_1\colon \rho \neq 0$$

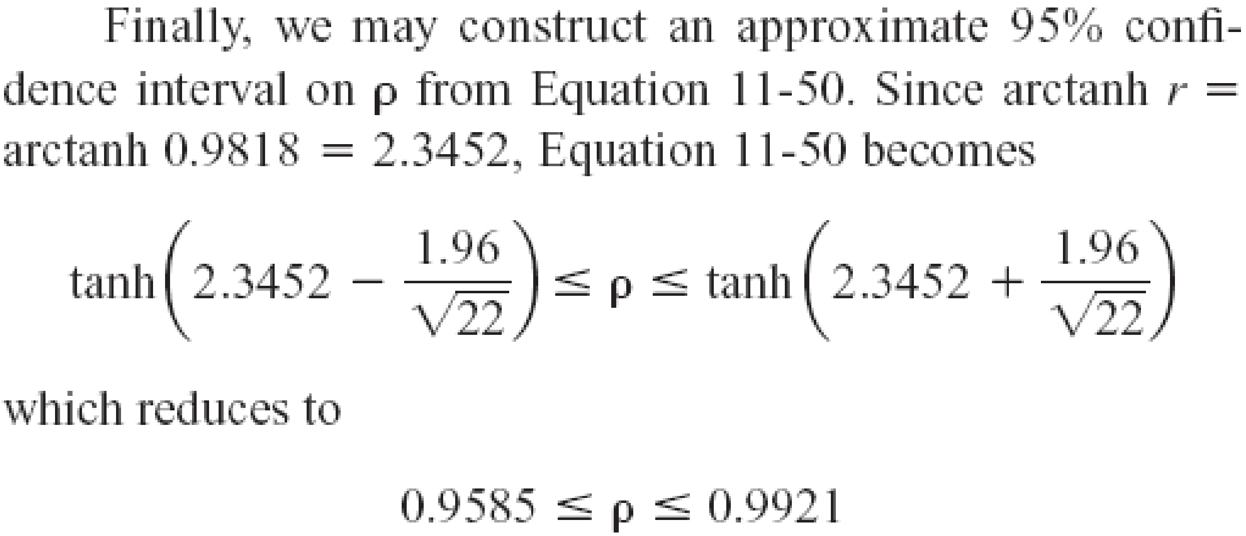
with $\alpha = 0.05$. We can compute the t-statistic of Equation 11-46 as

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9818\sqrt{23}}{\sqrt{1-0.9640}} = 24.8$$

This statistic is also reported in the Minitab output as a test of $H_0\colon \beta_1 = 0$. Because $t_{0.025,23} = 2.069$, we reject H_0 and conclude that the correlation coefficient $\rho \neq 0$.

11-8: Correlation

Example 11-8 (continued)



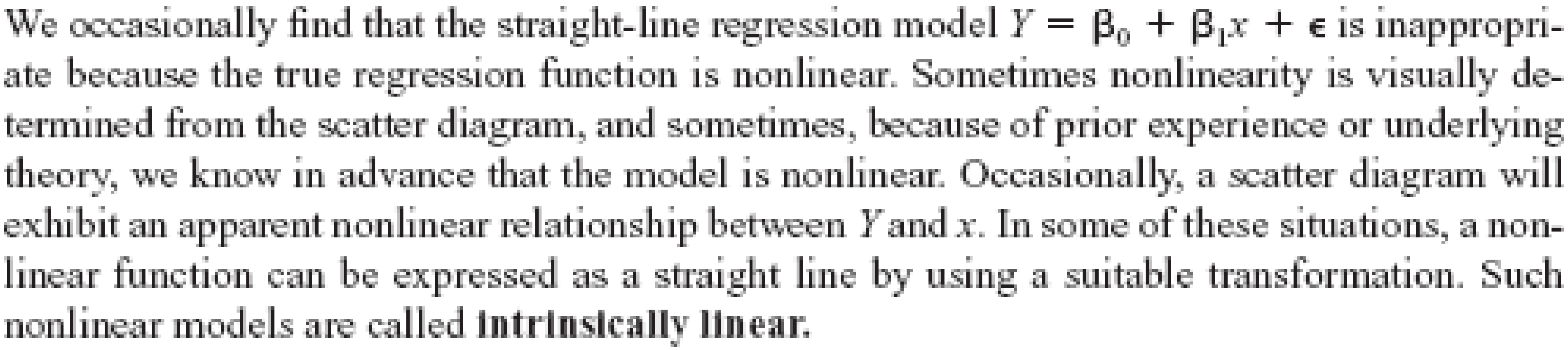
Finally, we may construct an approximate 95% confidence interval on ρ from Equation 11-50. Since arctanh r = arctanh 0.9818 = 2.3452, Equation 11-50 becomes

$$\tanh\left(2.3452 - \frac{1.96}{\sqrt{22}}\right) \le \rho \le \tanh\left(2.3452 + \frac{1.96}{\sqrt{22}}\right)$$

which reduces to

$$0.9585 \le \rho \le 0.9921$$

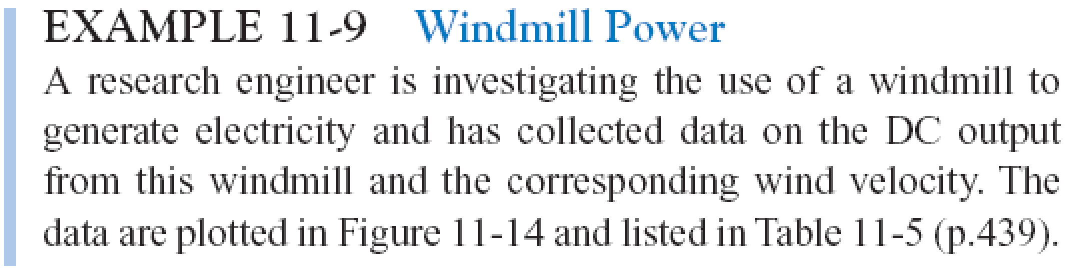
11-9: Transformation and Logistic Regression



We occasionally find that the straight-line regression model $Y = \beta_0 + \beta_1 x + \epsilon$ is inappropriate because the true regression function is nonlinear. Sometimes nonlinearity is visually determined from the scatter diagram, and sometimes, because of prior experience or underlying theory, we know in advance that the model is nonlinear. Occasionally, a scatter diagram will exhibit an apparent nonlinear relationship between Y and x. In some of these situations, a nonlinear function can be expressed as a straight line by using a suitable transformation. Such nonlinear models are called **intrinsically linear.**

11-9: Transformation and Logistic Regression

Example 11-9



EXAMPLE 11-9 Windmill Power

A research engineer is investigating the use of a windmill to generate electricity and has collected data on the DC output from this windmill and the corresponding wind velocity. The data are plotted in Figure 11-14 and listed in Table 11-5 (p.439).

Table 11-5 Observed Values y_i and Regressor Variable x_i for Example 11-9.

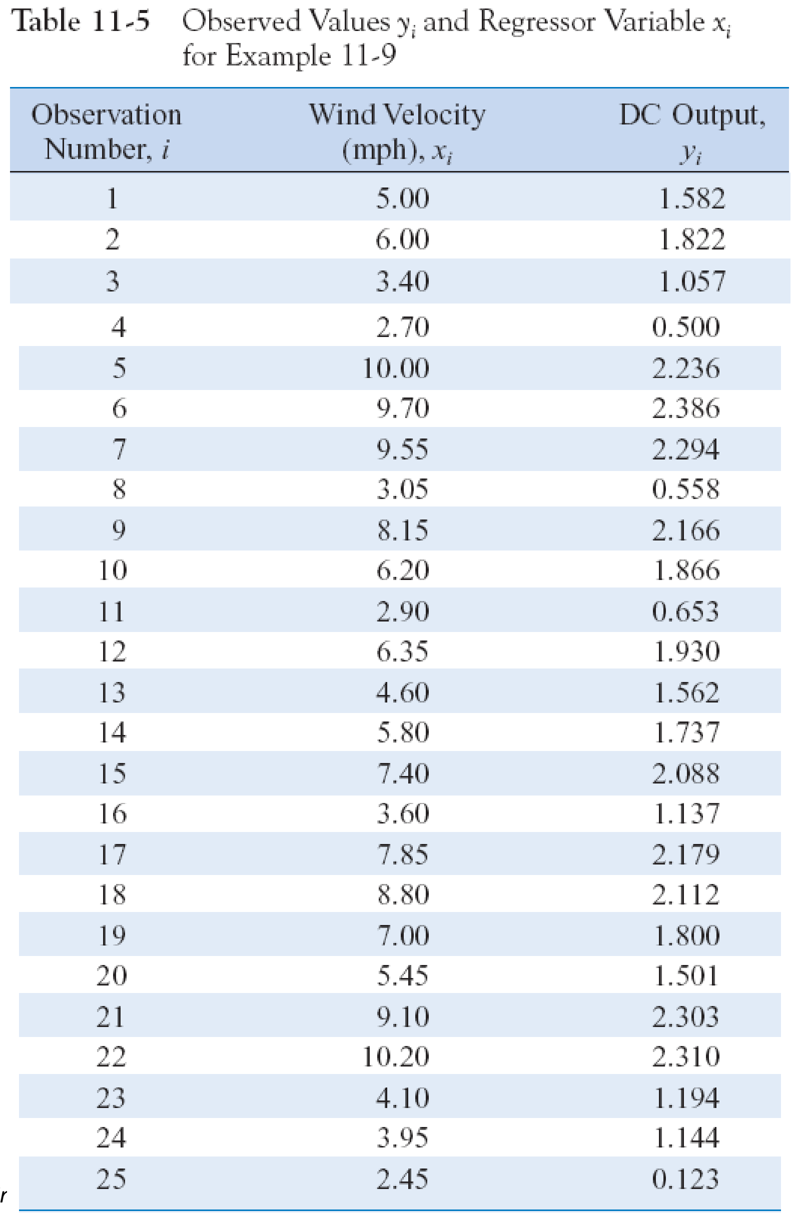


Table 11-5 Observed Values y_i and Regressor Variable x_i for Example 11-9

Observation Number, i	Wind Velocity (mph), x_i	DC Output, y_i
1	5.00	1.582
2	6.00	1.822
3	3.40	1.057
4	2.70	0.500
5	10.00	2.236
6	9.70	2.386
7	9.55	2.294
8	3.05	0.558
9	8.15	2.166
10	6.20	1.866
11	2.90	0.653
12	6.35	1.930
13	4.60	1.562
14	5.80	1.737
15	7.40	2.088
16	3.60	1.137
17	7.85	2.179
18	8.80	2.112
19	7.00	1.800
20	5.45	1.501
21	9.10	2.303
22	10.20	2.310
23	4.10	1.194
24	3.95	1.144
25	2.45	0.123

11-9: Transformation and Logistic Regression

Example 11-9 (Continued)

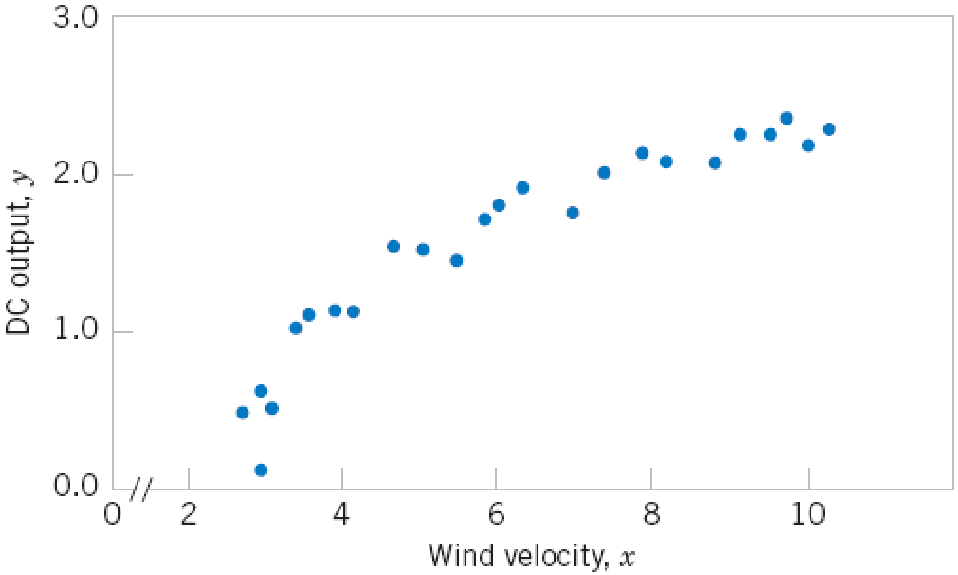


Figure 11-14 Plot of DC output y versus wind velocity x for the windmill data.

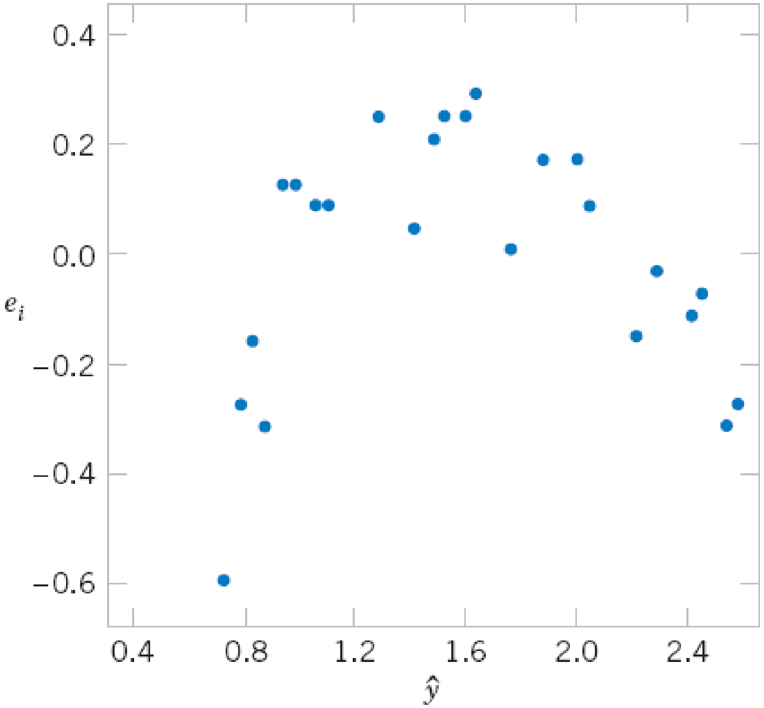


Figure 11-15 Plot of residuals e_i versus fitted values $\hat{y}_i$ for the windmill data.

11-9: Transformation and Logistic Regression

Example 11-9 (Continued)

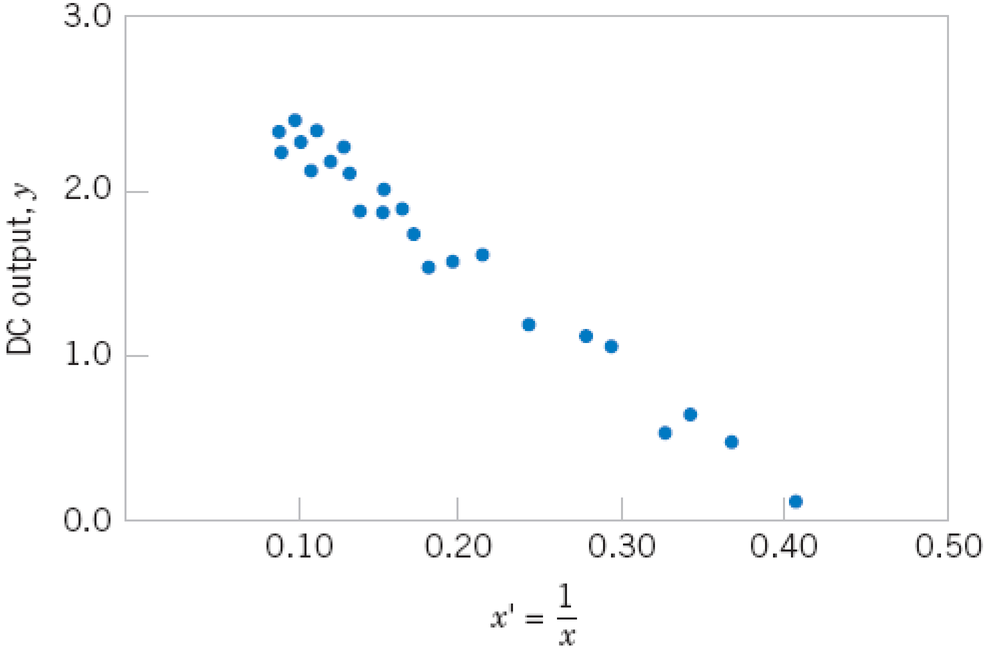


Figure 11-16 Plot of DC output versus $x' = 1/x$ for the windmill data.

11-9: Transformation and Logistic Regression

Example 11-9 (Continued)

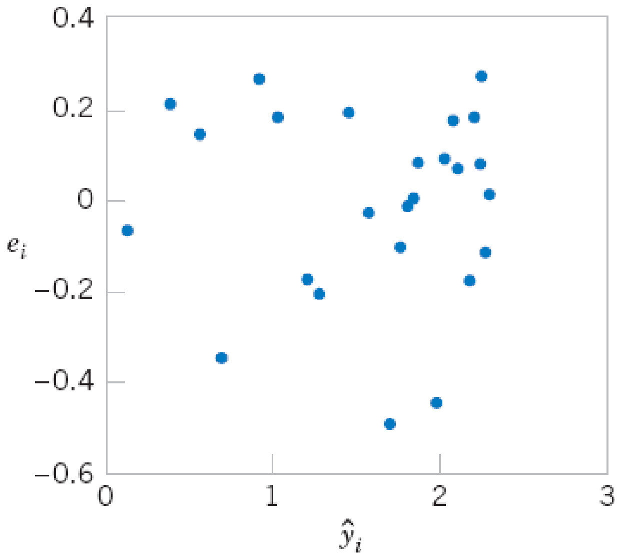


Figure 11-17 Plot of residuals versus fitted values $\hat{y}_i$ for the transformed model for the windmill data.

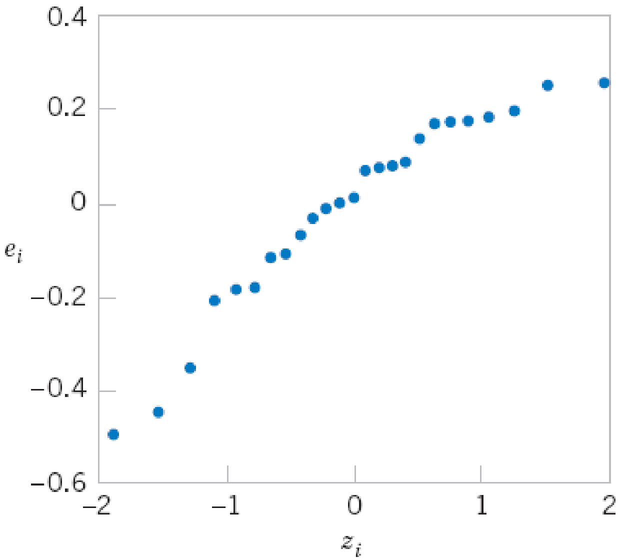


Figure 11-18 Normal probability plot of the residuals for the transformed model for the windmill data.

A plot of the residuals from the transformed model versus $\hat{y}$ is shown in Figure 11-17. This plot does not reveal any serious problem with inequality of variance. The normal probability plot, shown in Figure 11-18, gives a mild indication that the errors come from a distribution with heavier tails than the normal (notice the slight upward and downward curve at the extremes). This normal probability plot has the z-score value plotted on the horizontal axis. Since there is no strong signal of model inadequacy, we conclude that the transformed model is satisfactory.

Important Terms & Concepts of Chapter 11

Analysis of variance test in regression

Confidence interval on mean response

Correlation coefficient

Empirical model

Confidence intervals on model parameters

Intrinsically linear model

Least squares estimation of regression model parameters

Logistics regression

Model adequacy checking

Odds ratio

Prediction interval on a future observation

Regression analysis

Residual plots

Residuals

Scatter diagram

Simple linear regression model standard error

Statistical test on model parameters

Transformations