MA026IU PROBABILITY, STATISTIC AND RANDOM PROCESS

Part 2D

Simple Linear Regression and Correlation

Chapter 11 Title Slide

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.

Simple Linear Regression

- The **simple linear regression** considers a single **regressor** or **predictor** x and a **dependent** or **response variable** Y.

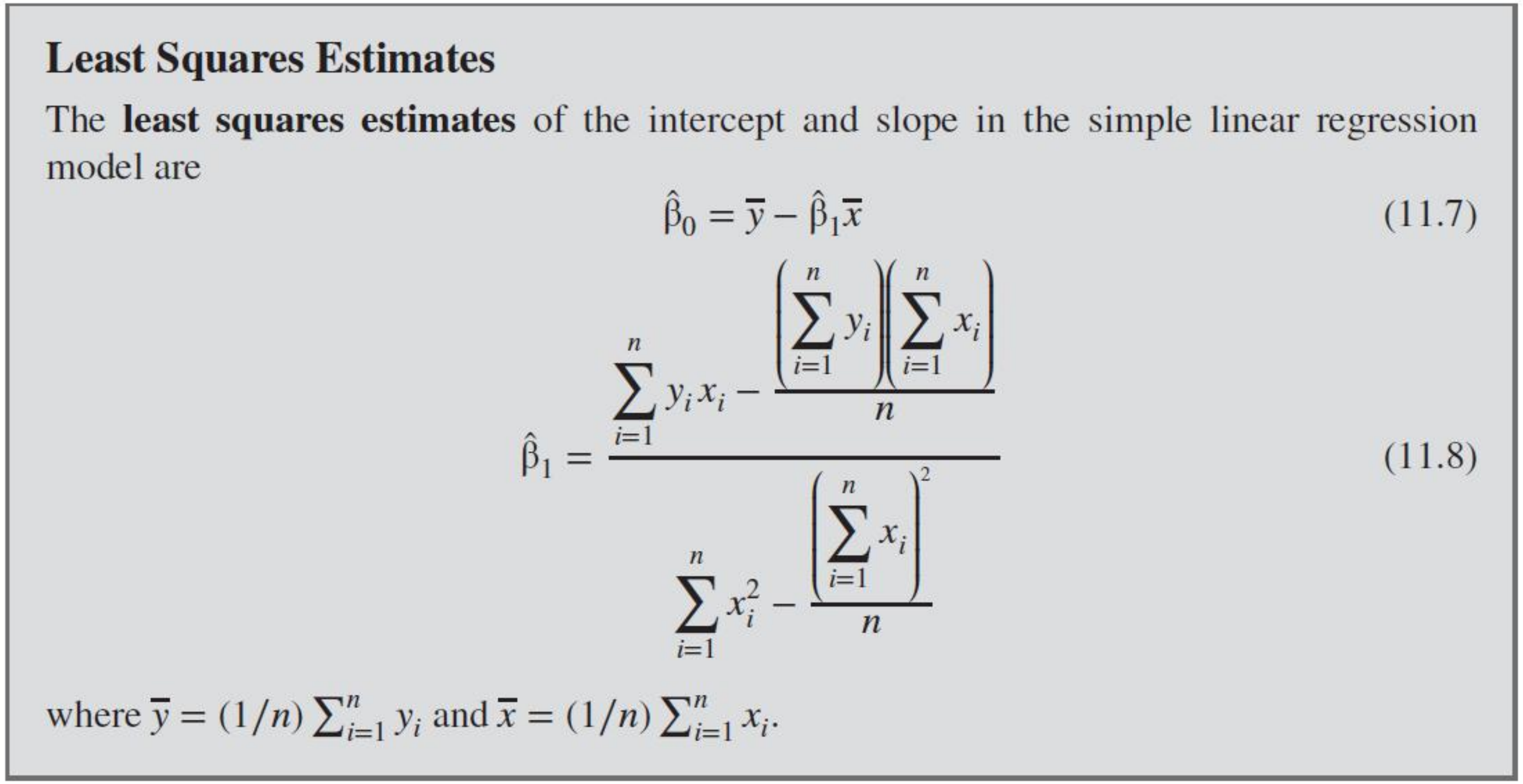
- The expected value of Y at each level of x is a random variable:

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

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

$$Y \; = \beta_0 + \beta_1 x + \varepsilon$$

Least Squares Estimates



Least Squares Estimates

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

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{11.7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \dfrac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11.8}$$

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

Simple Linear Regression

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

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

- Note that each pair of observations satisfies the relationship

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + e_i$$

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 i^{th} observation y_i.

Simple Linear Regression

Notation

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

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

Example 1 | Oxygen Purity

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

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

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

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

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

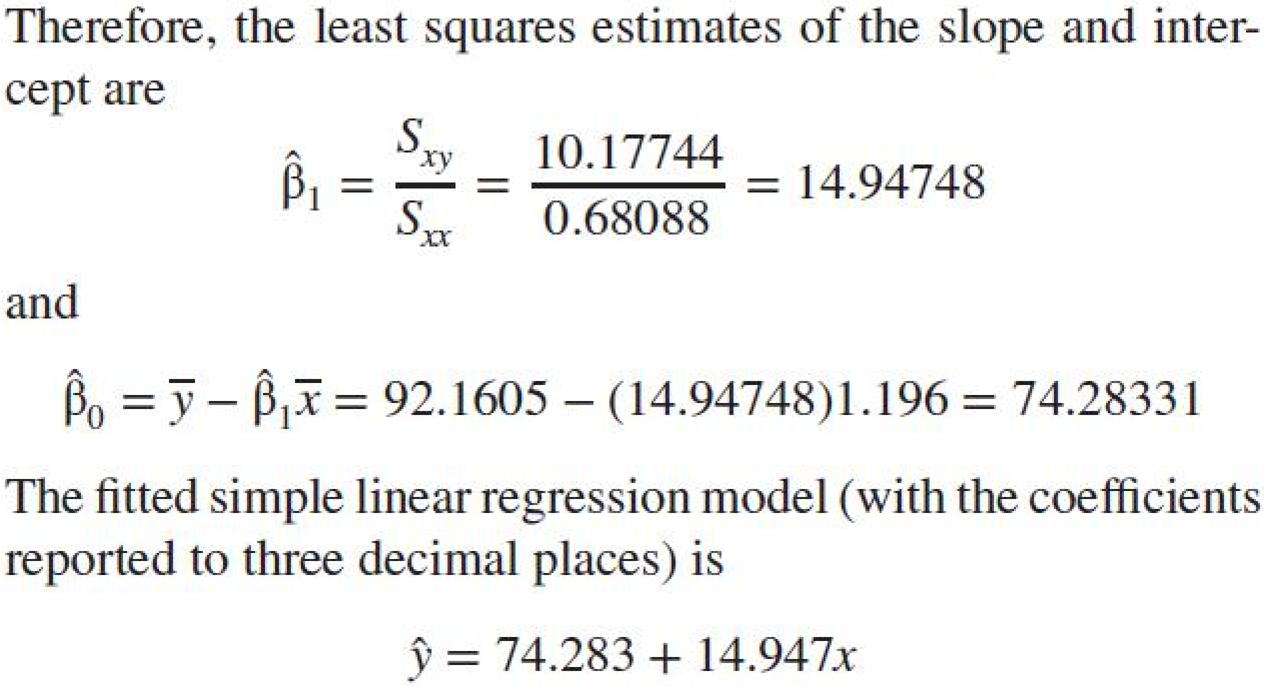
$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20} = 0.68088$$

and

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

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

Example 1 | Oxygen Purity (ctd)



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

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

Simple Linear Regression

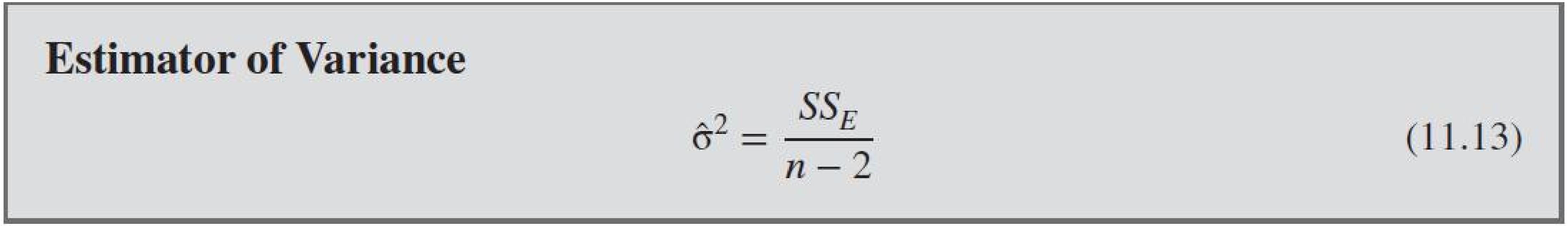
Estimating $\boldsymbol{\sigma^2}$

The error sum of squares is

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

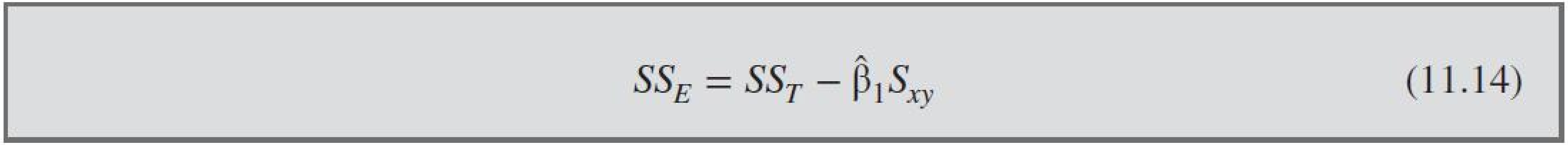
It can be shown that the expected value of the error sum of squares is $E(SSE) = (n - 2)\sigma^2$.

Estimator of Variance



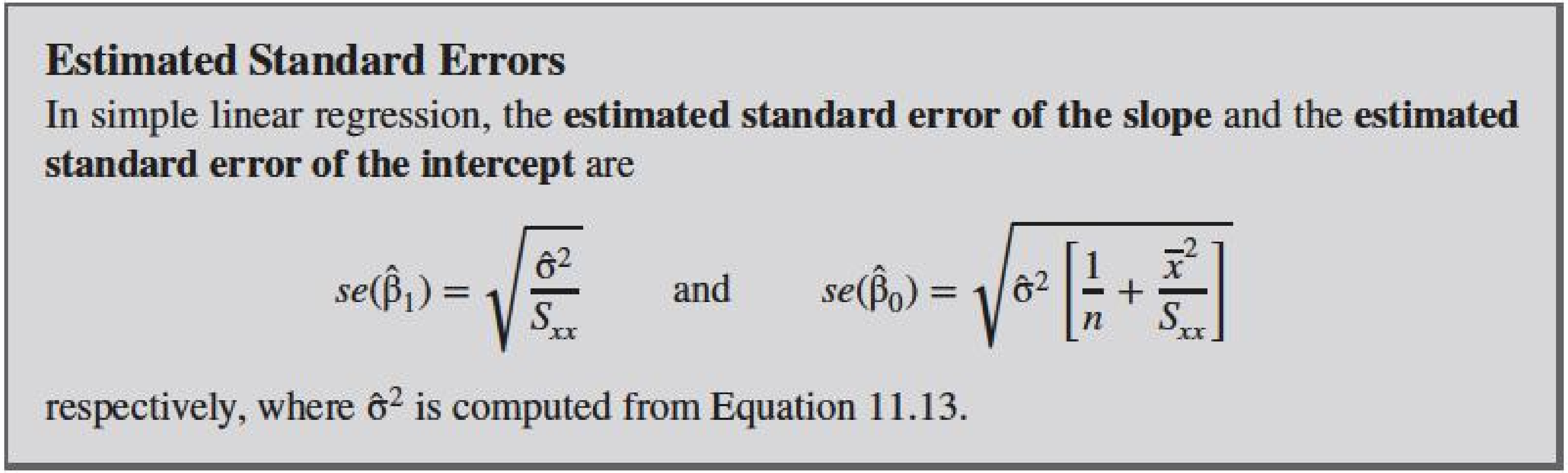
Estimator of Variance

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \tag{11.13}$$



$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \tag{11.14}$$

Properties of the Least Squares Estimators



Estimated Standard Errors

In simple linear regression, the **estimated standard error of the slope** and the **estimated standard error of the intercept** are

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad \text{and} \qquad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

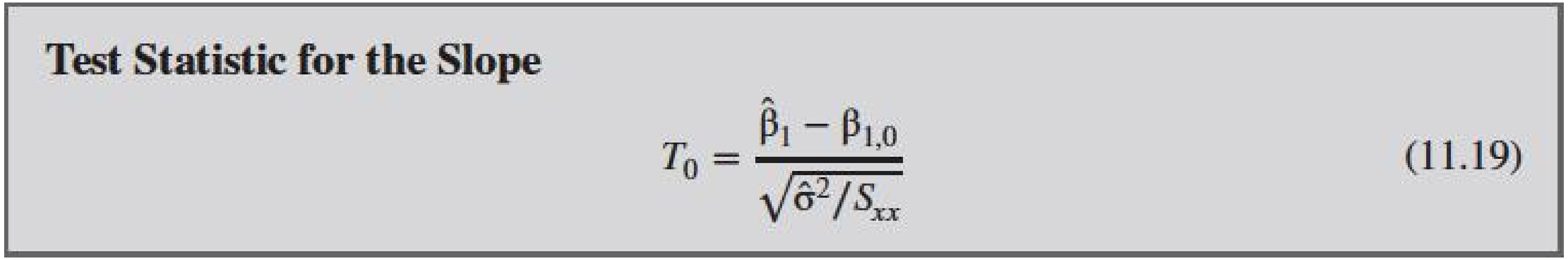
respectively, where $\hat{\sigma}^2$ is computed from Equation 11.13.

Hypothesis Tests in Simple Linear Regression

Use of *t*-Tests

- The appropriate hypotheses are

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



Test Statistic for the Slope

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \qquad (11.19)$$

Use of *t*-Tests

We could also write it as

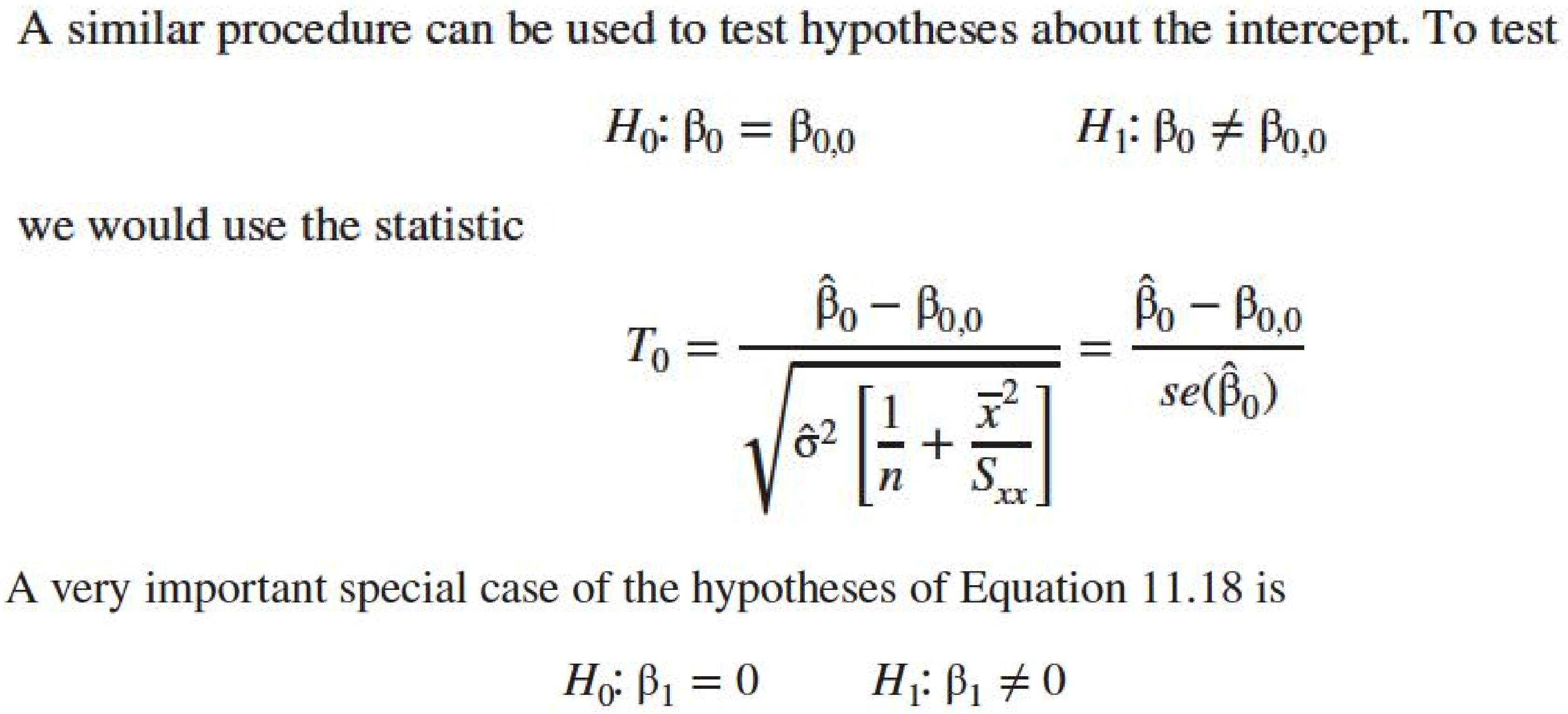
$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

We would reject

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

Sec 11.4.1 Use of *t*-Tests

Use of *t*-Tests



A similar procedure can be used to test hypotheses about the intercept. To test

$$H_0\text{: } \beta_0 = \beta_{0,0} \qquad H_1\text{: } \beta_0 \neq \beta_{0,0}$$

we would use the statistic

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

A very important special case of the hypotheses of Equation 11.18 is

$$H_0\text{: } \beta_1 = 0 \qquad H_1\text{: } \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**

Example 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

$$H_0\text{: } \beta_1 = 0$$

$$H_1\text{: } \beta_1 \neq 0$$

and we will use $\alpha = 0.01$. From Example 11.1 and Table 11.2 we have

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

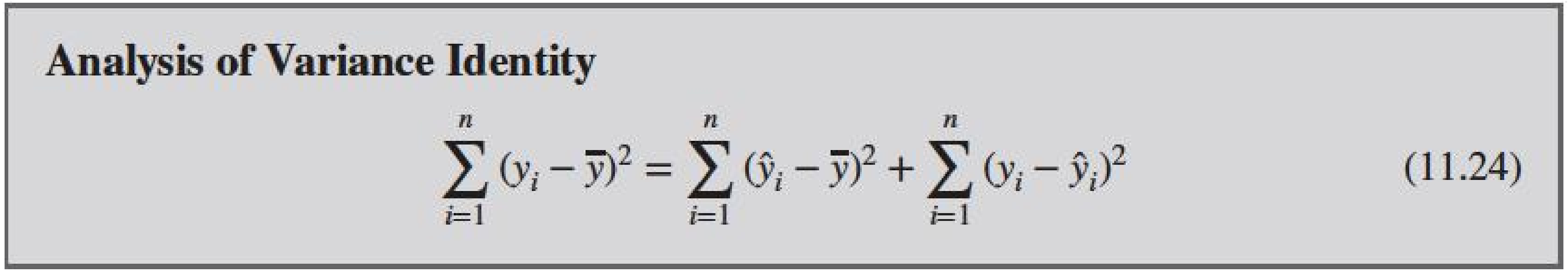
so the *t*-statistic in Equation 10.20 becomes

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

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: $\beta_1 = 0$ 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.

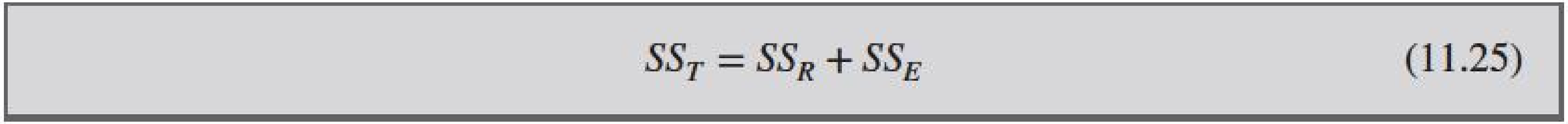
Analysis of Variance Approach to Test Significance of Regression

A method called **analysis of variance** can be used to test for significance of regression, The procedure partitions the total variability in the response variable into meaningful components as the basis for the test. The **analysis of variance identity** is as follows



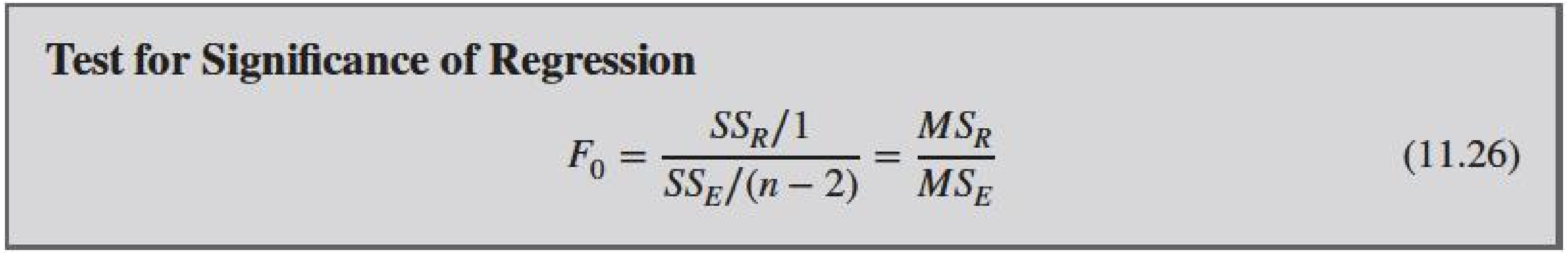
Analysis of Variance Identity

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad (11.24)$$



$$SS_T=SS_R+SS_E \qquad (11.25)$$

Analysis of Variance Approach to Test Significance of Regression



Test for Significance of Regression

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} \qquad (11.26)$$

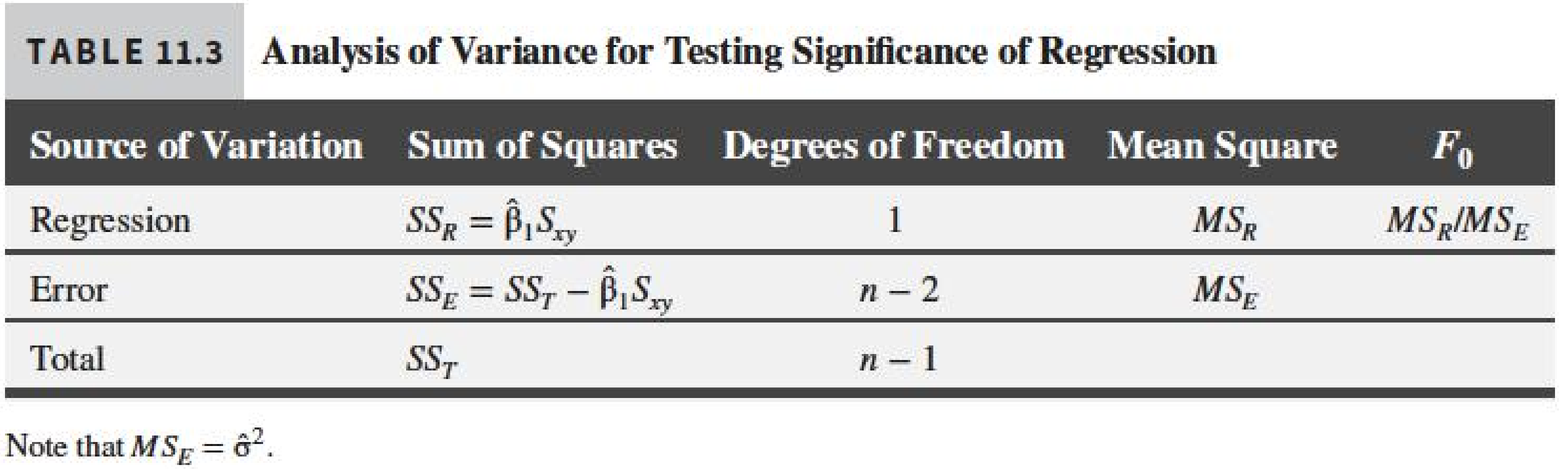


TABLE 11.3 Analysis of Variance for Testing Significance of Regression

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

Example 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 $SS_T = 173.38$, $\widehat{\beta}_1 = 14.947$, $S_{xy} = 10.17744$, and *n* = 20. The regression sum of squares is

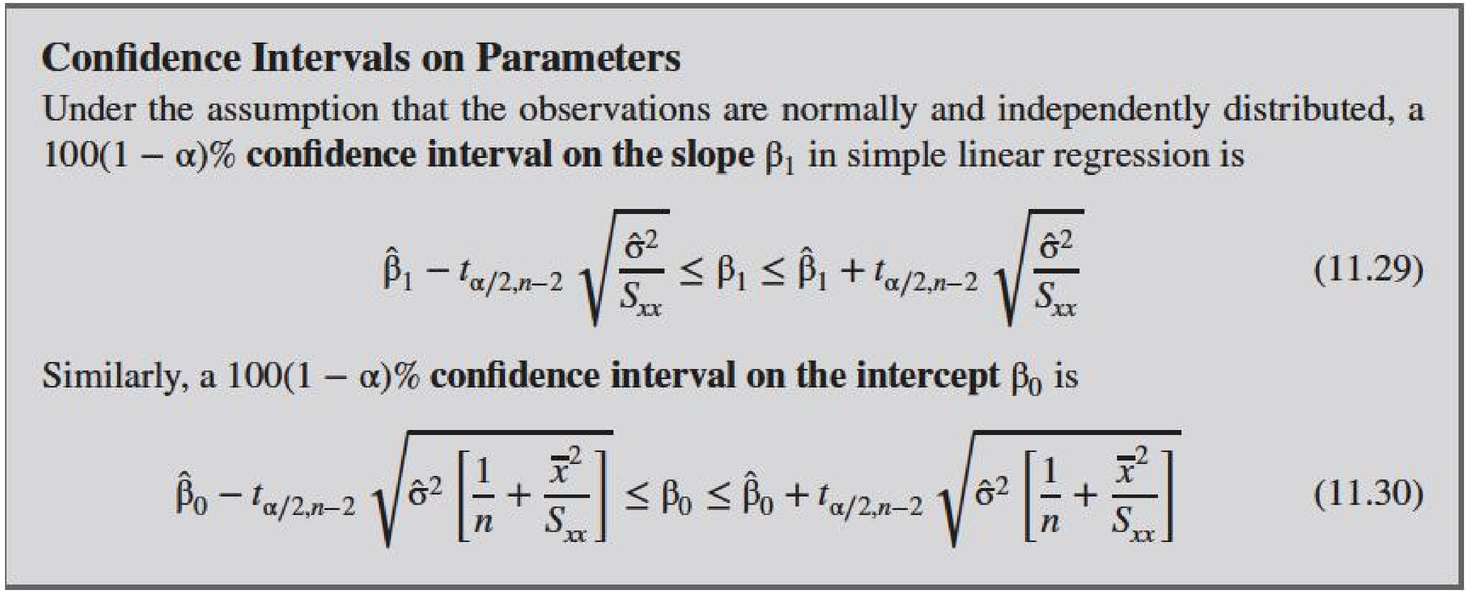
$$SS_R = \hat{\beta}_1 S_{xy} = (14.947)10.17744 = 152.13$$

and the error sum of squares is $SS_E = SST - SSR = 173.38 - 152.13 = 21.25$

- 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 $f_0 = MSR/MSE = 152.13/1.18 = 128.86$, for which we find that the *P*-value is $P \cong 1.23 \times 10^{-9}$, 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.

Confidence Intervals

Confidence Intervals on the Slope and Intercept



Confidence Intervals on Parameters

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

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad (11.29)$$

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

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \qquad (11.30)$$

Example 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 $\hat{\beta}_1 = 14.947, S_{xx} = 0.68088$, Then, from Equation 11.29 we find

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

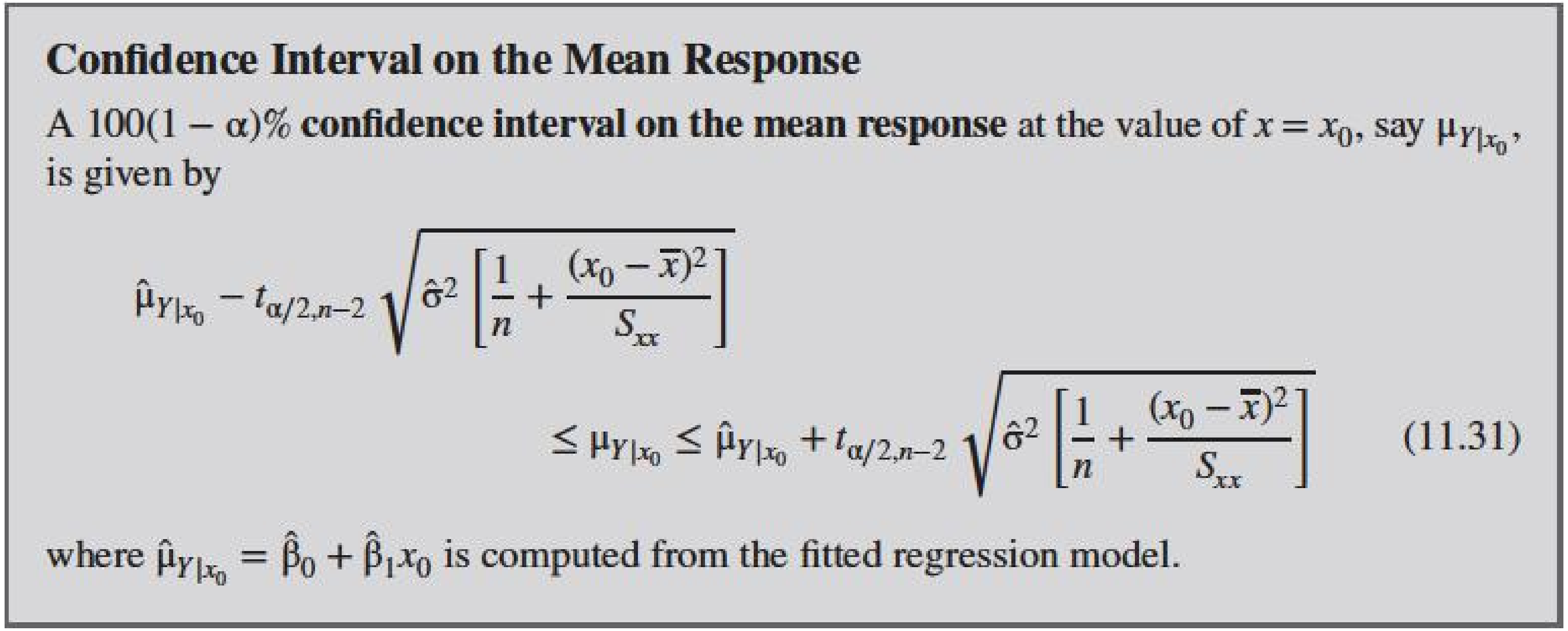
or

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

This simplifies to $12.181 \le \beta_1 \le 17.713$

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

Confidence Interval on the Mean Response



Confidence Interval on the Mean Response

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

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

$$\le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]} \qquad (11.31)$$

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

Sec 11.5.2 Confidence Intervals on the Mean Response

Example 5 | Oxygen Purity Confidence Interval on the Mean Response

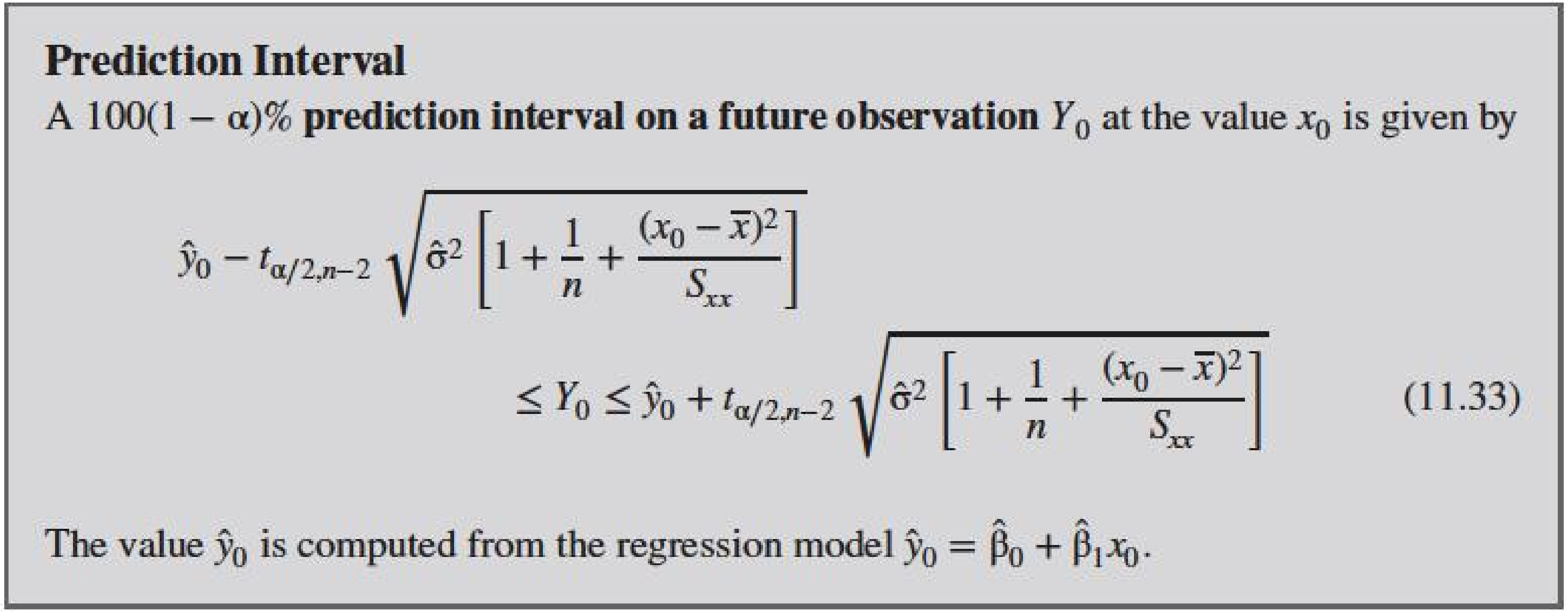
- We will construct a 95% confidence interval about the mean response for the data in Example 11.1. The fitted model is $\hat{\mu}_{Y|x_{1.00}} = 74.283 + 14.947(1.00) = 89.23$ and the 95% confidence interval on $\mu_{Y|x_0}$ is found from Equation 11.31 as $\hat{\mu}_{Y|x_0} = 74.283 + 14.947\,x_0$
- Suppose that we are interested in predicting mean oxygen purity when x_0 = 1.00%. Then

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

and the 95% confidence interval is $89.23 \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$ or 89.23 ± 0.75

- Therefore, the 95% CI on $\mu_{Y|1.00}$ is $88.48 \le \mu_{Y|1.00} \le 89.98$
- This is a reasonable, narrow CI.

Prediction of New Observations



Prediction Interval

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

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]}$$

$$\le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]} \qquad (11.33)$$

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

Example 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 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

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

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

which simplifies to

$$86.83 \leq y_0 \leq 91.63$$

This is a reasonably narrow prediction interval.

Adequacy of the Regression Model

- Fitting a regression model requires making several **assumptions**
- Estimating the model parameters requires assuming that the errors are uncorrelated random variables with mean zero and constant variance
- Tests of hypotheses and interval estimation require that the 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 that has been tentatively entertained

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
- As an approximate check of normality, the experimenter can construct a frequency histogram of the residuals or a **normal probability plot of residuals**

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.

Example 11.7b | Oxygen Purity Residuals

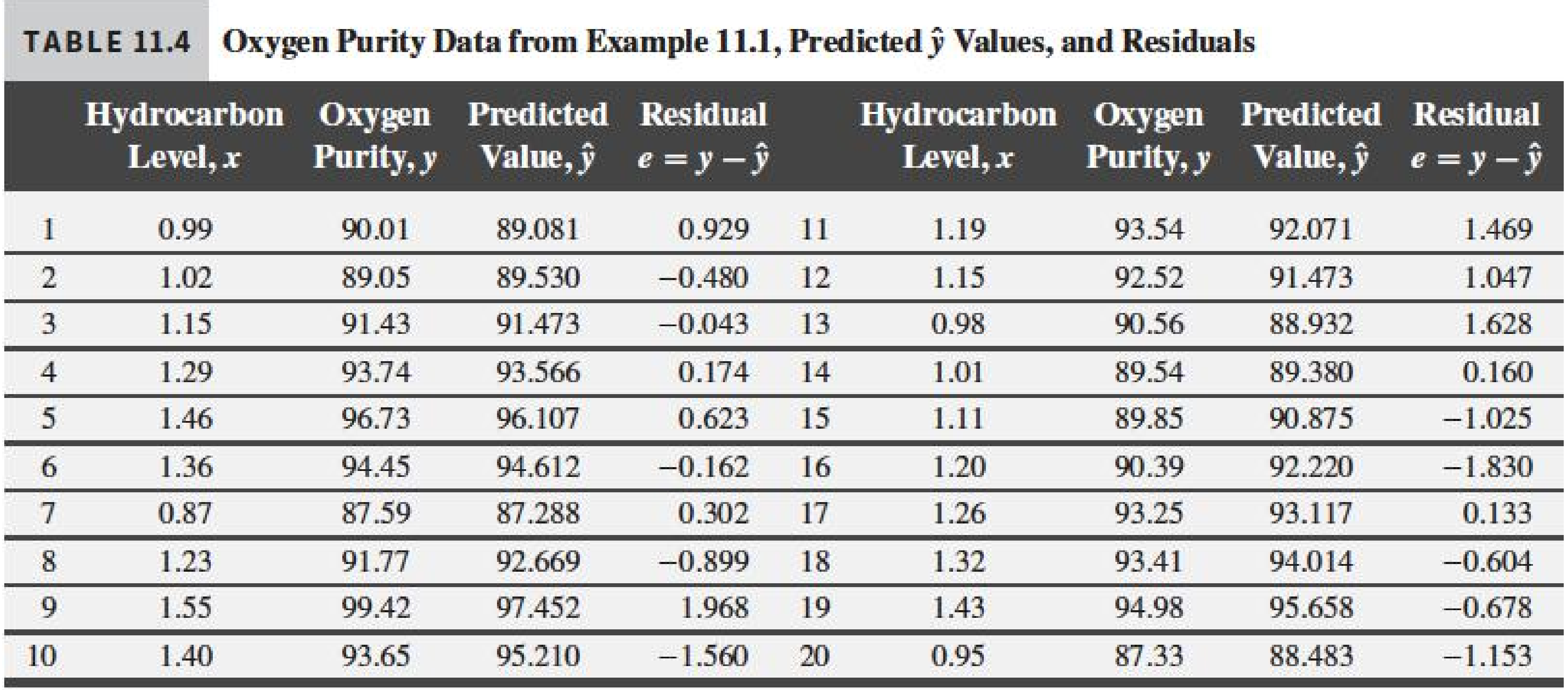


TABLE 11.4 Oxygen Purity Data from Example 11.1, Predicted $\hat{y}$ Values, and Residuals

	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

Sec 11.7.1 Residual Analysis

Example 11.7c | Oxygen Purity Residuals

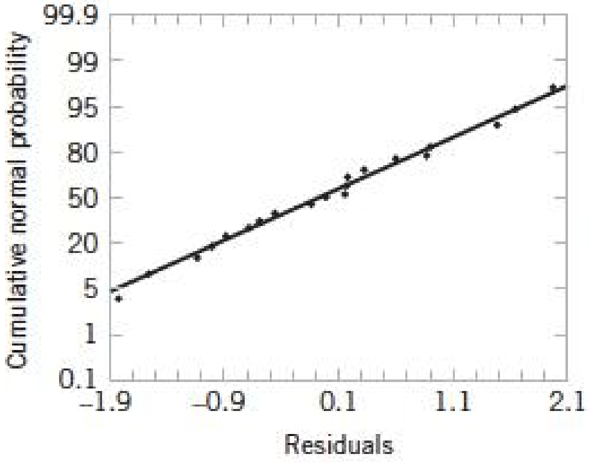


FIGURE 11.10

Normal probability plot of residuals.

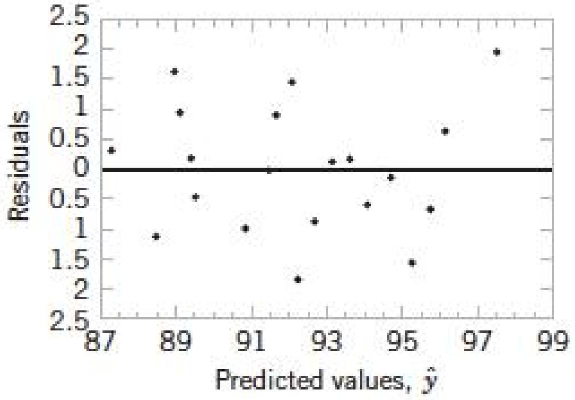


FIGURE 11.11

Plot of residuals versus predicted oxygen purity $\hat{y}$.

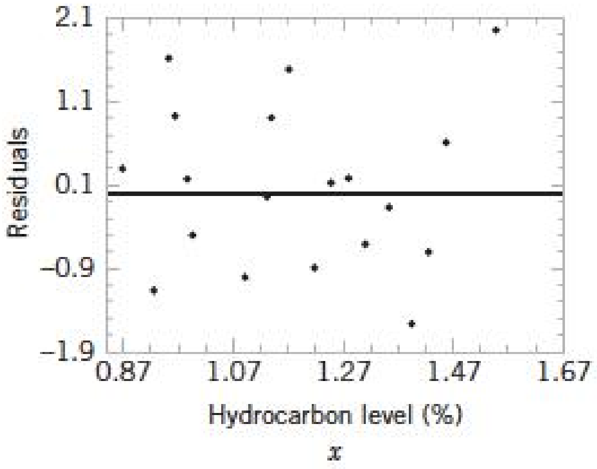
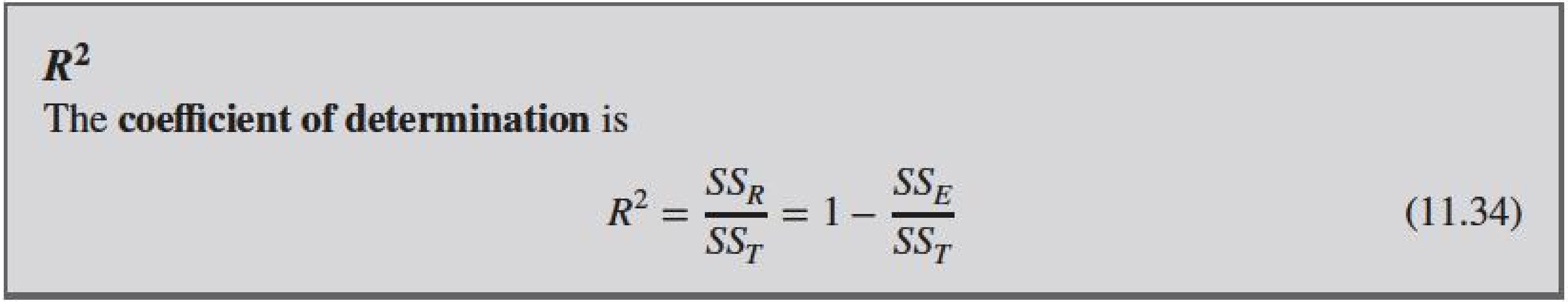


FIGURE 11.12

Plot of residuals versus hydrocarbon level x.

Coefficient of Determination (R^2)



R^2

The **coefficient of determination** is

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \quad (11.34)$$

Sec 11.7.2 Coefficient of Determination (R^2)

Correlation

- Recall that the correlation coefficient is defined as

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Where σ_{XY} is the covariance between *Y* and *X*
- The conditional distribution of *Y* for a given value of *X* = *x* is

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

where

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X} \qquad \beta_1 = \frac{\sigma_Y}{\sigma_X}\rho$$

Correlation

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

$$\hat{\rho} = \frac{\sum_{i=1}^{n} Y_i(X_i - \overline{X})}{\left[\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2\right]^{1/2}} = \frac{S_{XY}}{\left(S_{XX} SS_T\right)^{1/2}} \tag{11.43}$$

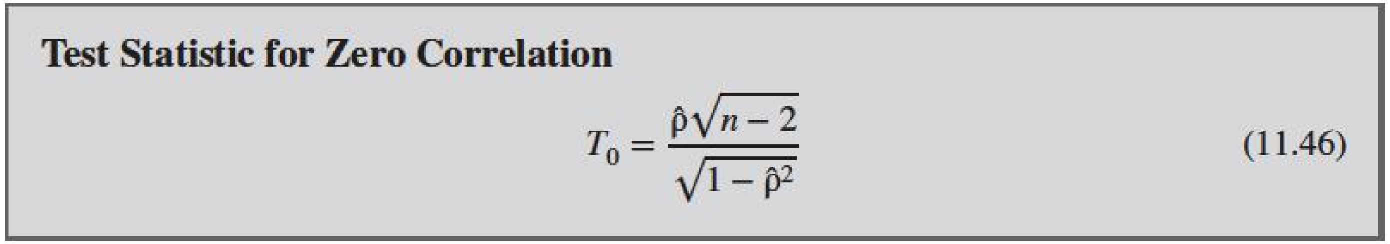
- Note tha $\hat{\beta}_1 = \left(\frac{SS_T}{S_{XX}}\right)^{1/2} \hat{\rho}$

- We may also write: $\hat{\rho}^2 = \hat{\beta}_1^2 \frac{S_{XX}}{SS_T} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T}$

Correlation

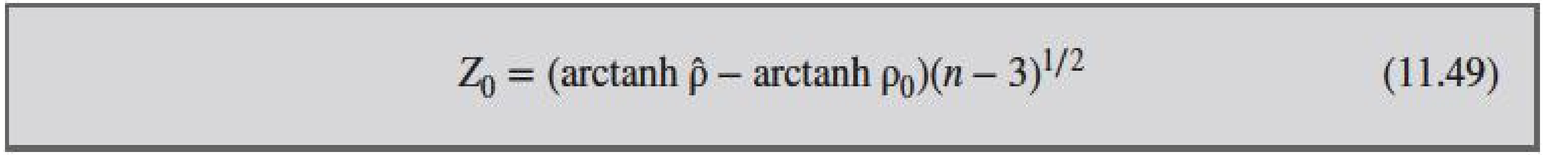
The test procedure for the hypotheses

$$H_0\text{: } \rho = \rho_0 \qquad H_1\text{: } \rho \neq \rho_0$$

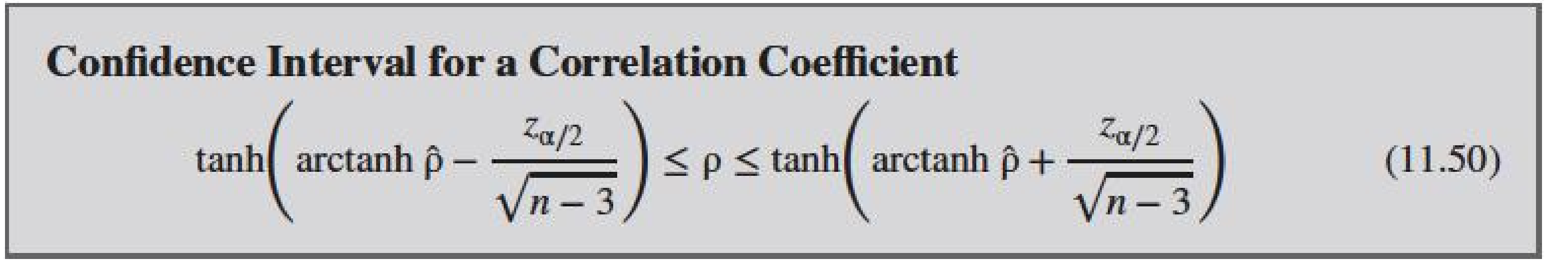


Test Statistic for Zero Correlation

$$T_0 = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \tag{11.46}$$



$$Z_0 = (\text{arctanh } \hat{\rho} - \text{arctanh } \rho_0)(n-3)^{1/2} \tag{11.49}$$



Confidence Interval for a Correlation Coefficient

$$\tanh\left(\text{arctanh } \hat{\rho} - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(\text{arctanh } \hat{\rho} + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \tag{11.50}$$

Example 11.8 | Wire Bond Pull Strength

In Chapter 1 (Sec 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 arc 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 displayed box plots of each individual variable on the scatter diagram. There is evidence of a linear relationship between the two variables.

Example 11.8b | Wire Bond Pull Strength

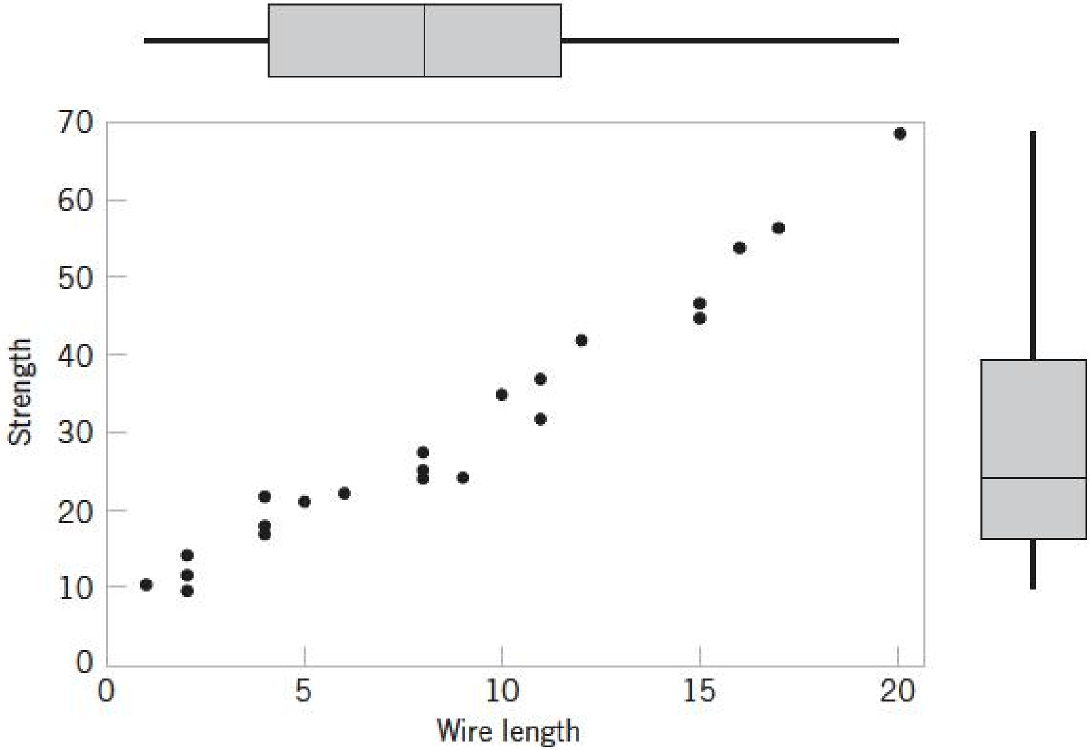


FIGURE 11.13

Scatter plot of wire bond strength versus wire length.

Example 11.8c | Wire Bond Pull Strength

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
```

Example 11.8d | Wire Bond Pull Strength

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

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

Note that $\hat{\rho}^2$ = $(0.9818)^2$ = 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.

Example 11.8e | Wire Bond Pull Strength

Now suppose that we wish to test the hypotheses

$$H_0\text{: } \rho = 0$$

$$H_1\text{: } \rho \neq 0$$

with $\alpha = 0.05$. We can compute the *t*-statistic of Equation 11.46 as

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

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

Example 11.8f | Wire Bond Pull Strength

Finally, we may construct an approximate 95% confidence interval on r from Equation 11.50. Since $\text{arctanh}\,\hat{\rho}$ = $\text{arctanh}\ 0.9818$ = 2.3452, Equation 11.50 becomes

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

which reduces to

$$0.9585 \leq \rho \leq 0.9921$$

Regression on Transformed Variables

- We occasionally find that the straight-line regression model $Y = \beta_0 + \beta_1 x + \varepsilon$ 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.**

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.

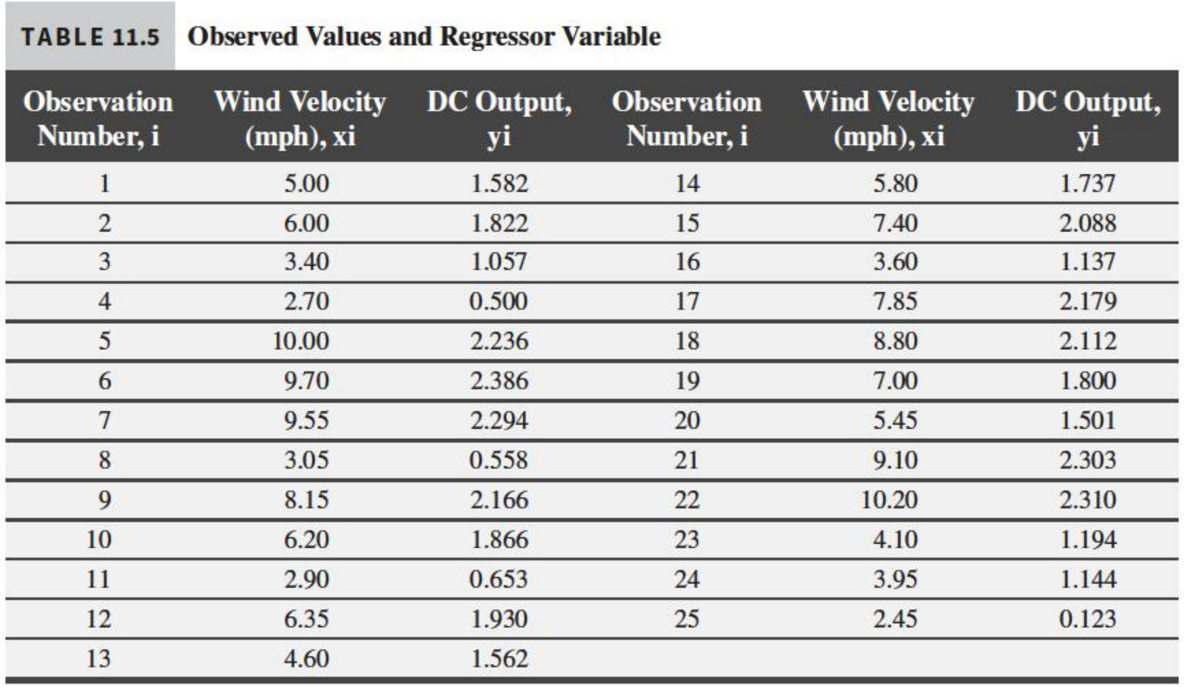


TABLE 11.5 **Observed Values and Regressor Variable**

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

Example 11.9b | Windmill Power

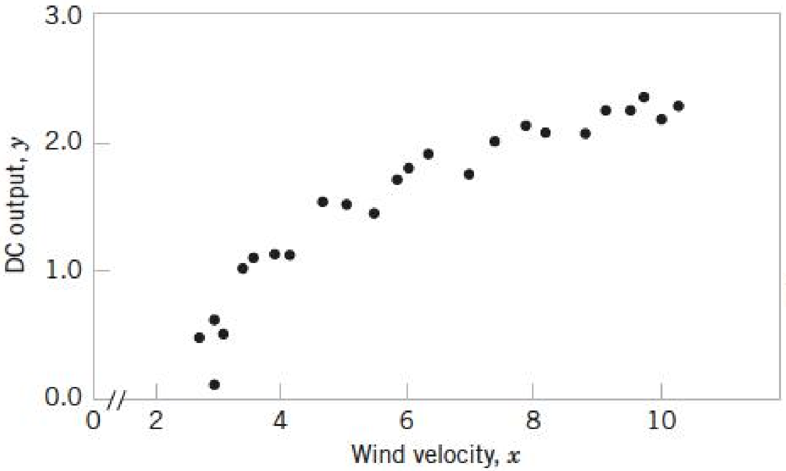


FIGURE 11.14

Plot of DC output y versus wind velocity x for the windmill data.

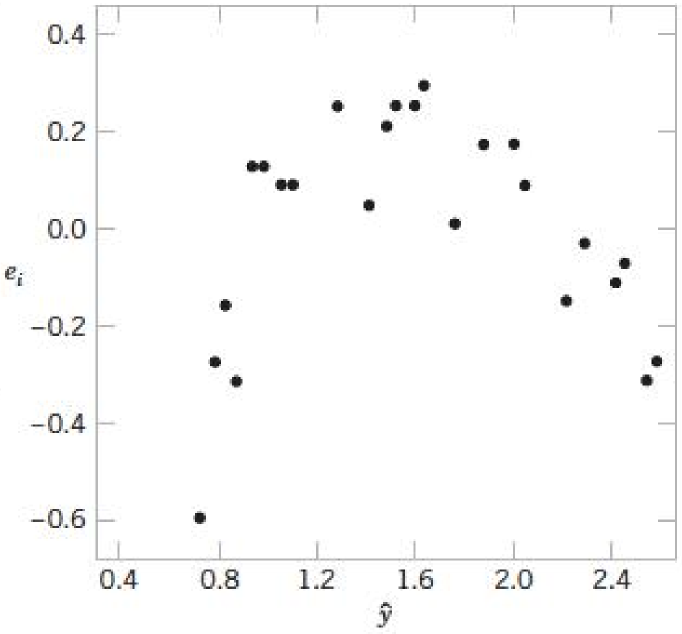


FIGURE 11.15

Plot of residuals e_i versus fitted values $\widehat{y}_i$ for the windmill data.

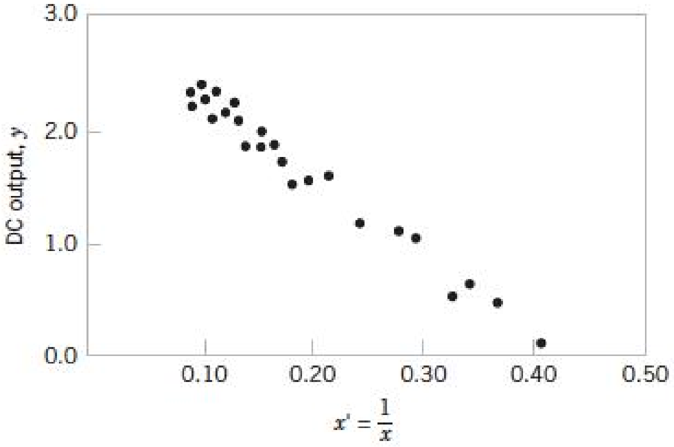


FIGURE 11.16

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

Example 11.9c | Windmill Power

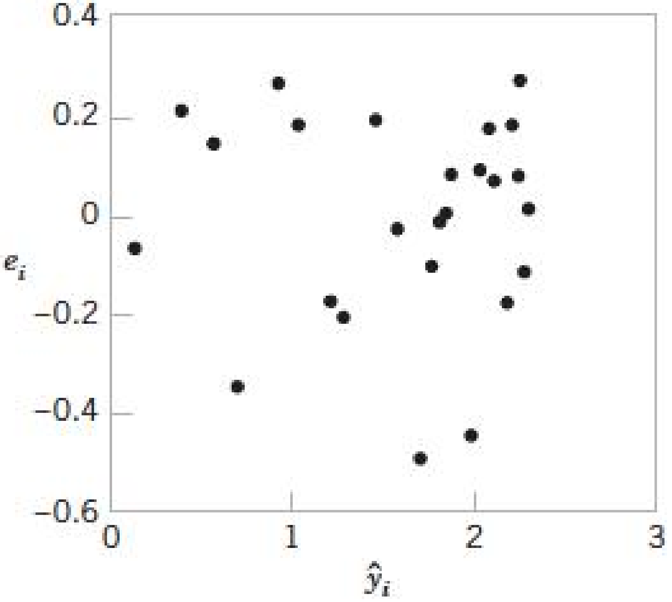


FIGURE 11.17

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

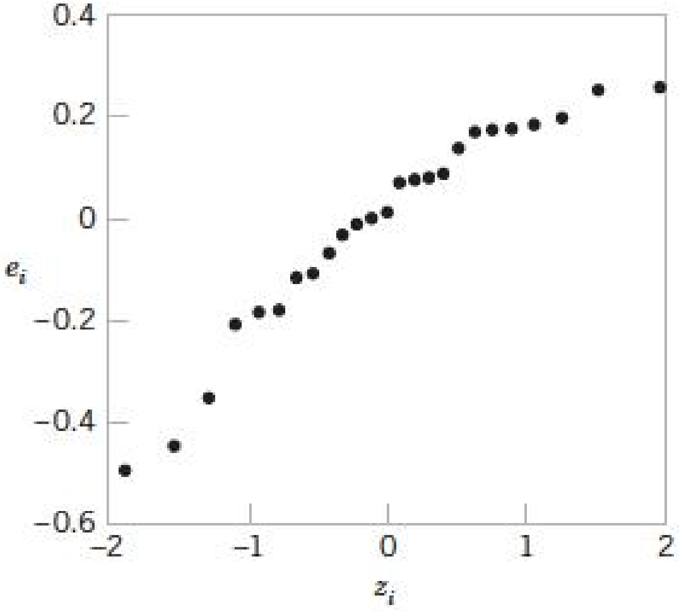


FIGURE 11.18

Normal probability plot of the residuals for the transformed model for the windmill data.

Logistic Regression

- Linear regression works well when the response variable is **quantitative**
- Logistic regression works in cases when the response variable takes on only two possible values, 0 and 1, an arbitrary assignment from observing a **qualitative response.**
- In logist

$$\frac{E(Y)}{1 - E(Y)} = \exp(\beta_0 + \beta_1 x) \qquad (11.54)$$

Important Terms and Concepts

- Analysis of variance table
- Coefficient of determination
- Confidence interval on the intercept
- Confidence interval on the mean response
- Confidence interval on the slope
- Correlation coefficient
- Empirical model
- Error sum of squares
- Intrinsically linear model
- Least squares
- Logistic regression
- Logit response function
- Mean squares
- Normal probability plot of residuals
- Odds ratio