11

Simple Linear Regression and Correlation

FPT UNIVERSITY
DREAM OF INNOVATION

CHAPTER OUTLINE

LEARNING OBJECTIVES

After careful study of this chapter, you should be able to do the following:

1. Use simple linear regression for building empirical models to engineering and scientific data
2. Understand how the method of least squares is used to estimate the parameters in a linear regression model
3. Analyze residuals to determine if the regression model is an adequate fit to the data or to see if any underlying assumptions are violated
4. Test statistical hypotheses and construct confidence intervals on regression model parameters
5. Use the regression model to make a prediction of a future observation and construct an appropriate prediction interval on the future observation
6. Apply the correlation model
7. Use simple transformations to achieve a linear regression model

11-1 Empirical Models

- Regression analysis is the process of building mathematical models or mathematical functions that can describe, predict or control of a variable from one or more other variables.

- 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.

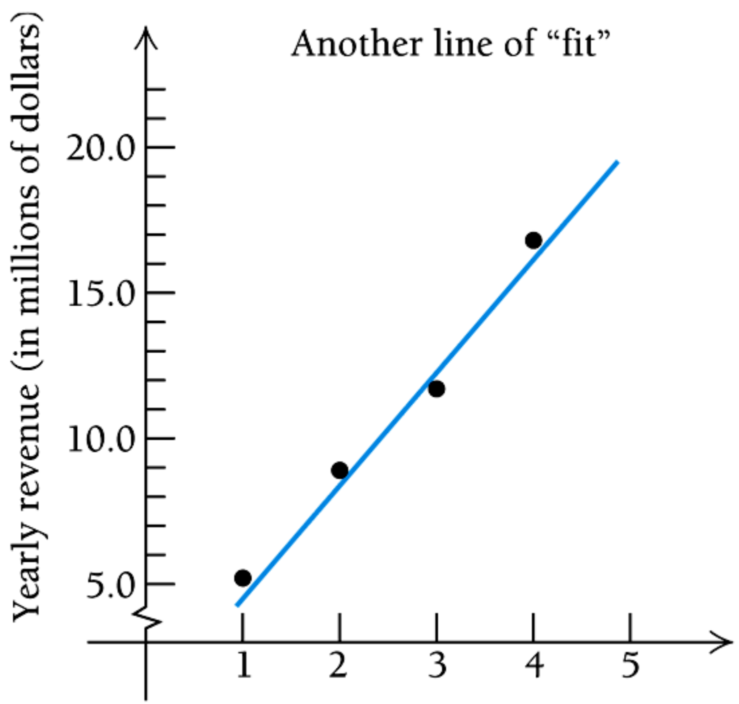
11-1 Empirical Models

For example , suppose that a car rental company that offers hybrid vehicles charts its revenue as shown below. How best could we predict the company's revenue for the year 2016?

Year, *x*	1996	2001	2006	2011	2016
Yearly Revenue, *y* (in millions of dollars)	5.2	8.9	11.7	16.8	?

11-1 Empirical Models

Suppose that we plot these points and try to draw a line through them that fits. Note that there are several ways in which this might be done. (See the graphs below.) Each would give a different estimate of the company’s total revenue for 2016.



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.

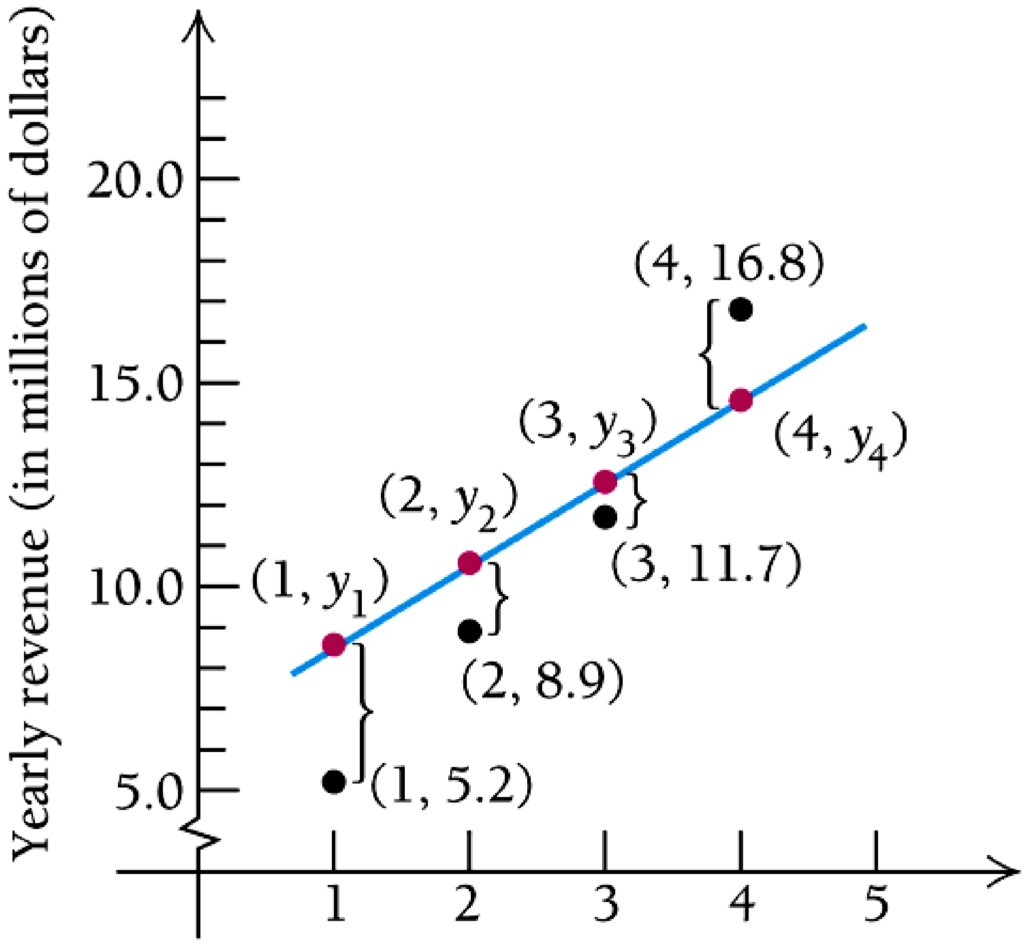
- Also, the variability of Y at a particular value of x is determined by the error variance, σ^2.
- This implies there is a distribution of Y-values at each x and that the variance of this distribution is the same at each x.

11-1 Empirical Models

To determine the equation of the line that "best" fits the data, we note that for each data point there will be a deviation, or error, between the *y*-value at that point and the *y*-value of the point on the line that is directly above or below the point.

Those deviations, in the example, $y_1 – 5.2$, $y_2 – 8.9$, $y_3 – 11.7$, and $y_4 – 16.8$, will be positive or negative, depending on the location of the line.

11-1 Empirical Models



11-2 Simple Linear Regression

We wish to fit these data points with a line,

$$y = \beta_1 x + \beta_0 \ ,$$

that uses values of β_1 and β_0 that, somehow, minimize the deviations in order to have a good fit.

One way of minimizing the deviations is based on the *least-squares assumption*.

11-2 Simple Linear Regression

Note that squaring each *y*-deviation gives us a sum of nonnegative terms. Were we to simply add the deviations, positive and negative deviations would cancel each other out.

Using the least-squares assumption with the yearly revenue data, we want to minimize.

$$(y_1 - 5.2)^2 + (y_2 - 8.9)^2 + (y_3 - 11.7)^2 + (y_4 - 16.8)^2$$

11-2 Simple Linear Regression

Also, since the points (1, y_1), (2, y_2), (3, y_3), and (4, y_4) must be solutions of $y = \beta_1 x + \beta_0$, it follows that

$$\begin{aligned}
y_1 &= \beta_1(1) + \beta_0 &&= \beta_1 + \beta_0 \\
y_2 &= \beta_1(2) + \beta_0 &&= 2\beta_1 + \beta_0 \\
y_3 &= \beta_1(3) + \beta_0 &&= 3\beta_1 + \beta_0 \\
y_4 &= \beta_1(4) + \beta_0 &&= 4\beta_1 + \beta_0
\end{aligned}$$

Substituting these values for each y in the previous equation, we now have a function of two variables.

11-2 Simple Linear Regression

$$L(\beta_1, \beta_0) = (\beta_1 + \beta_0 - 5.2)^2 + (2\beta_1 + \beta_0 - 8.9)^2 + (3\beta_1 + \beta_0 - 11.7)^2 + (4\beta_1 + \beta_0 - 16.8)^2$$

Thus, to find the regression line for the given set of data, we must find the values of β_0 and β_1 that minimize the function *L* given by the sum above.

We first find $\partial L / \partial \beta_0$ and $\partial L / \partial \beta_1$.

11-2 Simple Linear Regression

$$\frac{\partial L}{\partial \beta_0} = 2(\beta_1 + \beta_0 - 5.2) + 2(2\beta_1 + \beta_0 - 8.9)$$

$$+ 2(3\beta_1 + \beta_0 - 11.7) + 2(4\beta_1 + \beta_0 - 16.8)$$

$$= 20\beta_1 + 8\beta_0 - 85.2$$

and

$$\frac{\partial L}{\partial \beta_1} = 2(\beta_1 + \beta_0 - 5.2) + 2(2\beta_1 + \beta_0 - 8.9)2$$

$$+ 2(3\beta_1 + \beta_0 - 11.7)3 + 2(4\beta_1 + \beta_0 - 16.8)4$$

$$= 60\beta_1 + 20\beta_0 - 250.6$$

11-2 Simple Linear Regression

We set the derivatives equal to 0 and solve the resulting system:

$$\begin{aligned} 20\beta_1 + 8\beta_0 - 85.2 &= 0 \\ 60\beta_1 + 20\beta_0 - 250.6 &= 0 \end{aligned}$$

It can be shown that the solution to this system is

$$\beta_0 = 1.25, \qquad \beta_1 = 3.76.$$

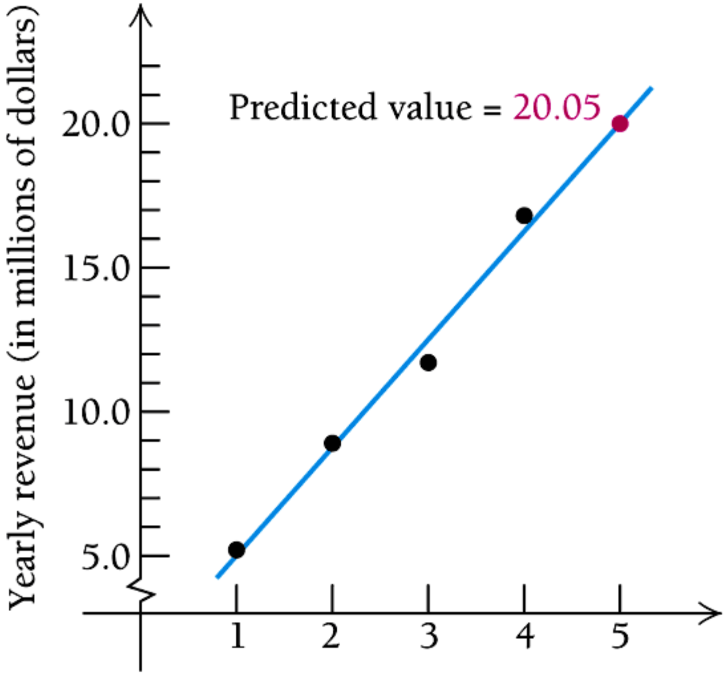
We leave it to the student to complete the *D*-test to verify that (1.25, 3.76) does, in fact, yield a minimum of *L*.

11-2 Simple Linear Regression

There is no need to compute $L(1.25, 3.76)$. The values of β_1 and β_0 are all we need to determine $y = \beta_1 x + \beta_0$. The regression line is

$$y = 3.76x + 1.25.$$

The graph of this "best-fit" regression line together with the data points is shown below. Compare it to the graphs before.



11-2 Simple Linear Regression

Now, we can use the regression equation to predict the car rental company's yearly revenue in 2016.

y = 3.76(5) + 1.25 = 20.05 or about $20.05 million.

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

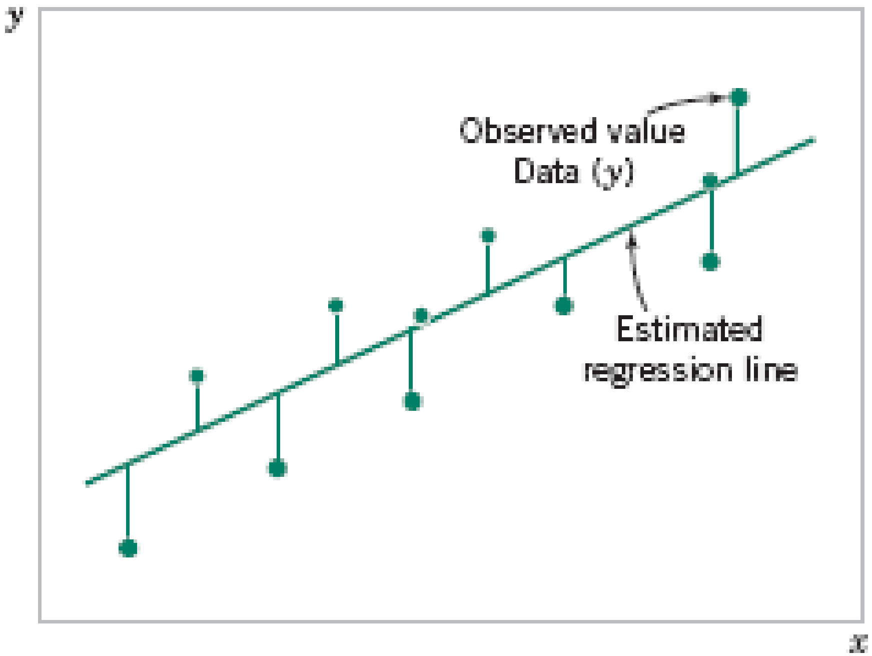


Figure 11-3 Deviations of the data from the estimated regression model.

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.

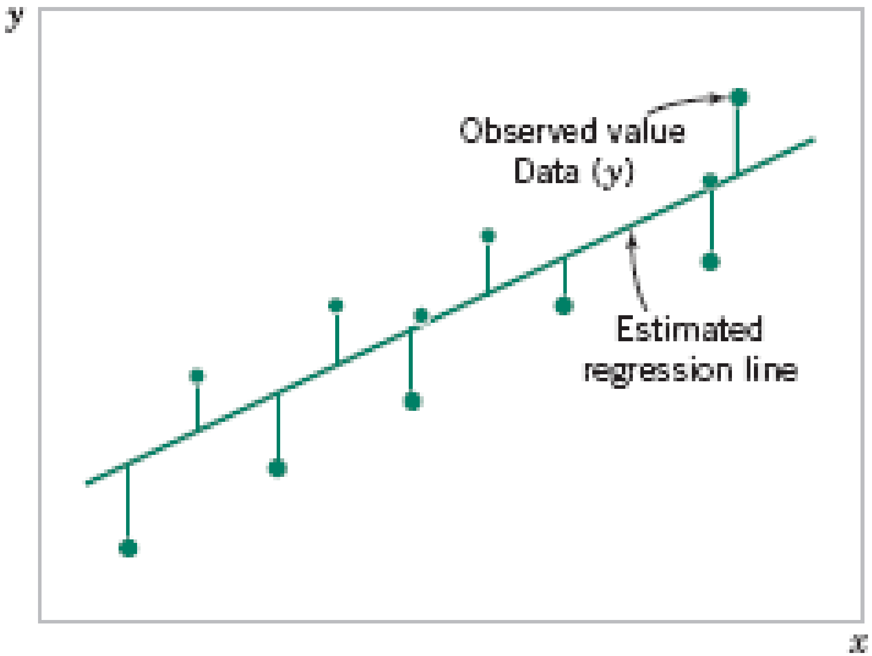


Figure 11-3 Deviations of the data from the estimated regression model.

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

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

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

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (11\text{-}9)$$

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.

Example

To study the relationship between ticket prices and number of passengers on each flight, research, 11 commercial flights, we have the following data table:

Number of passengers	Cost (1000$)
61	4.28
63	4.08
69	4.17
70	4.48
74	4.30
76	4.82
81	4.70
86	5.11
91	5.13
95	5.64
97	5.56

Find regression line of the number of passengers in term of ticket prices.

$y= - 24.53+21.67.x$

11-2 Simple Linear Regression

Notation

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

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

11-2 Simple Linear Regression

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

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

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

$$SS_T = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

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

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

11-3 Properties of the Least Squares Estimators

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

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

respectively, where $\hat{\sigma}^2$ is computed from 11 -13.

11-4 Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

Suppose we wish to test

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

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

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

An appropriate test statistic would be

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

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

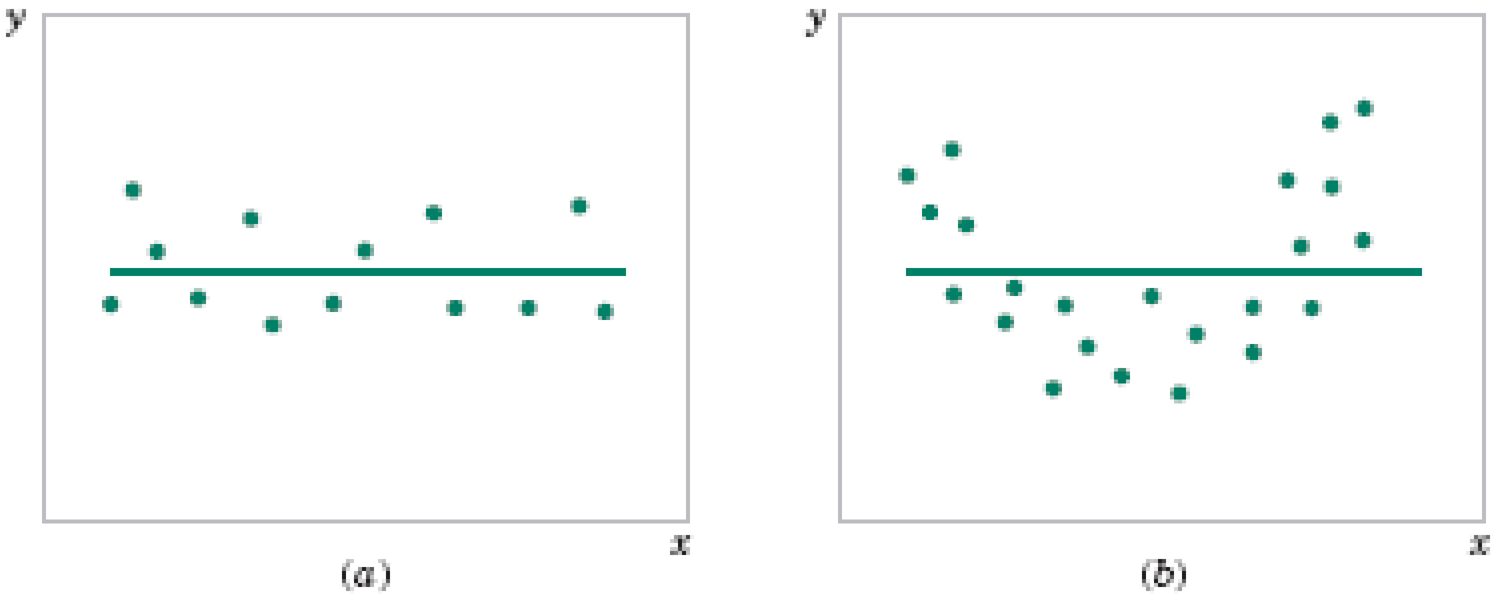


Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

11-4 Hypothesis Tests in Simple Linear Regression

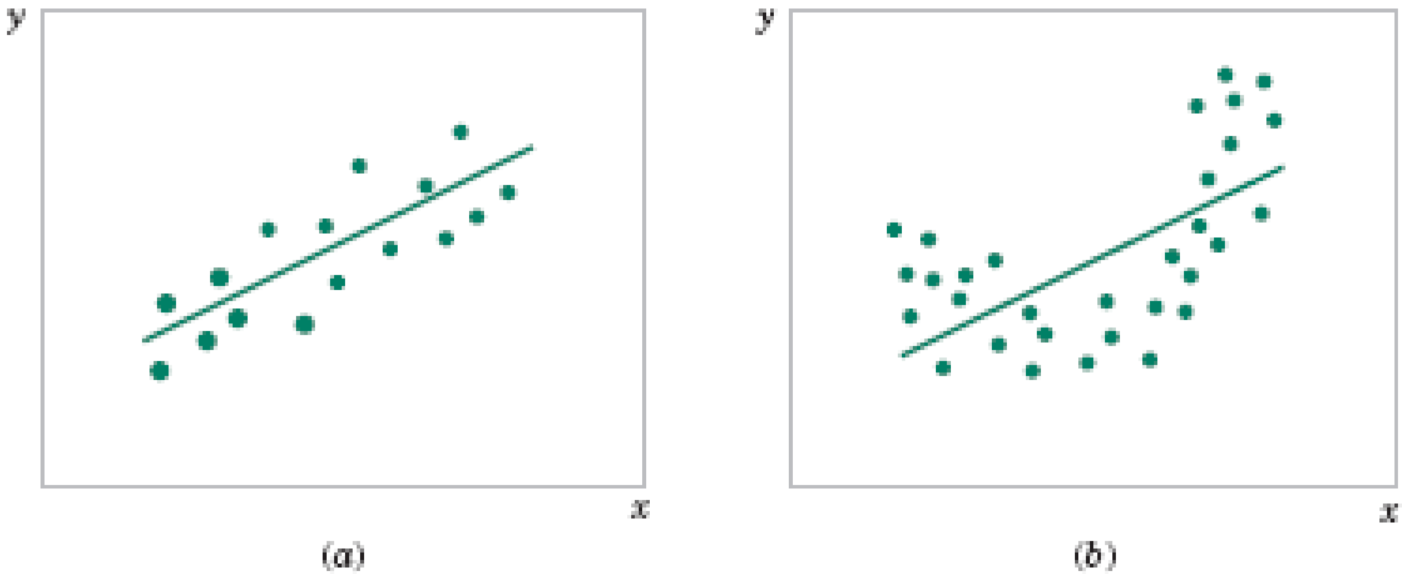


Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

11-4 Hypothesis Tests in Simple Linear Regression

- 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

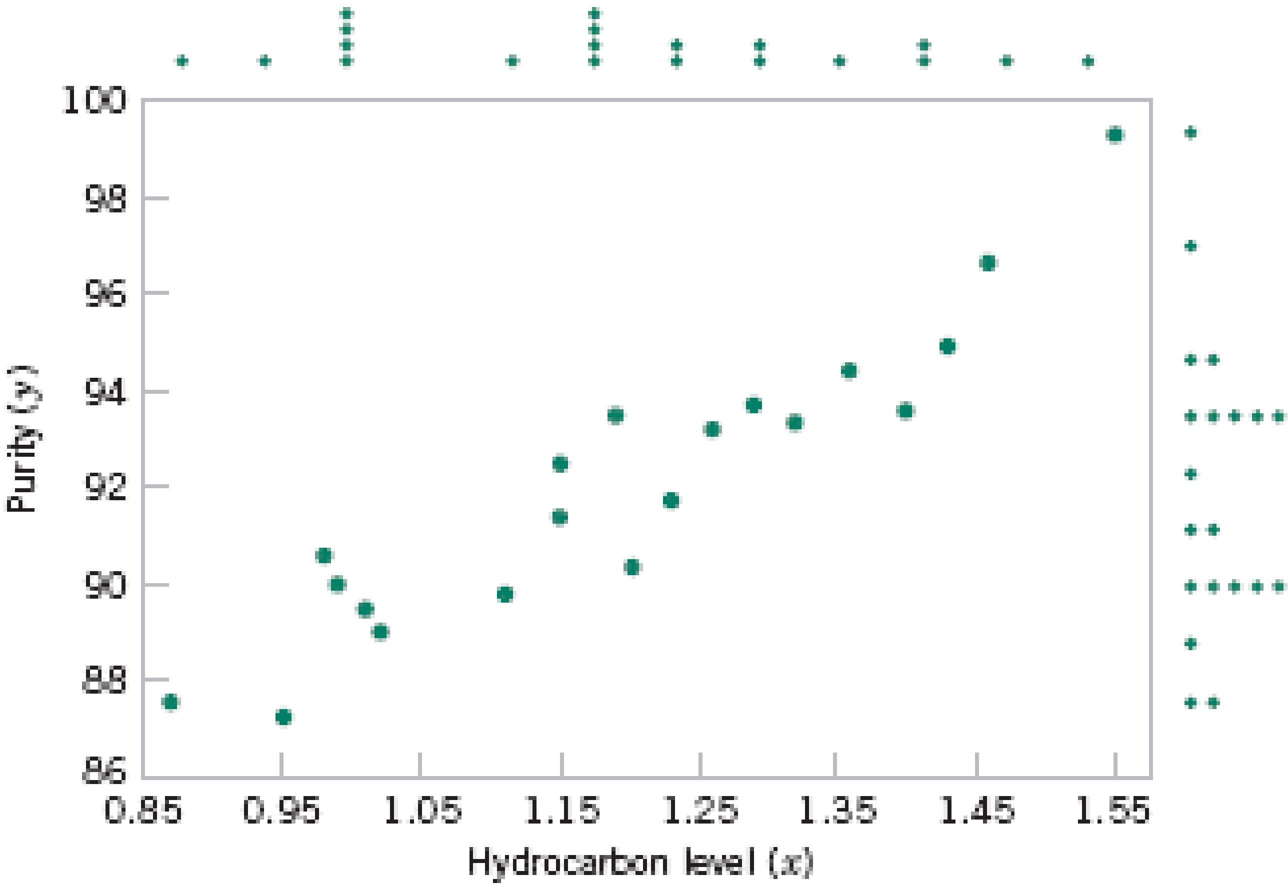


Figure. Scatter Diagram of oxygen purity versus hydrocarbon level from Table 11-1.

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

11-4 Hypothesis Tests in Simple Linear Regression

Example 11-2

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

$$\hat{\beta}_1 = 14.97 \quad n = 20, \quad S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$$

so the t-statistic in Equation 10-20 becomes

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

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. The P-value for this test is $P \simeq 1.23 \times 10^{-9}$. This was obtained manually with a calculator.

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

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad (11\text{-}35)$$

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

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

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

where

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X} \qquad (11\text{-}37)$$

$$\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho \qquad (11\text{-}38)$$

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

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

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

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

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

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

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

11-8 Correlation

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

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

where $\tanh u = (e^u - e^{-u})/(e^u + e^{-u})$.

11-8 Correlation

Example 11-8

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.

FPT UNIVERSITY
DREAM OF INNOVATION

Table 1-2 Wire Bond Pull Strength Data

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50
2	24.45	8	110
3	31.75	11	120
4	35.00	10	550
5	25.02	8	295
6	16.86	4	200
7	14.38	2	375
8	9.60	2	52
9	24.35	9	100
10	27.50	8	300
11	17.08	4	412
12	37.00	11	400
13	41.95	12	500
14	11.66	2	360
15	21.65	4	205
16	17.89	4	400
17	69.00	20	600
18	10.30	1	585
19	34.93	10	540
20	46.59	15	250
21	44.88	15	290
22	54.12	16	510
23	56.63	17	590
24	22.13	6	100
25	21.15	5	400

11-8 Correlation

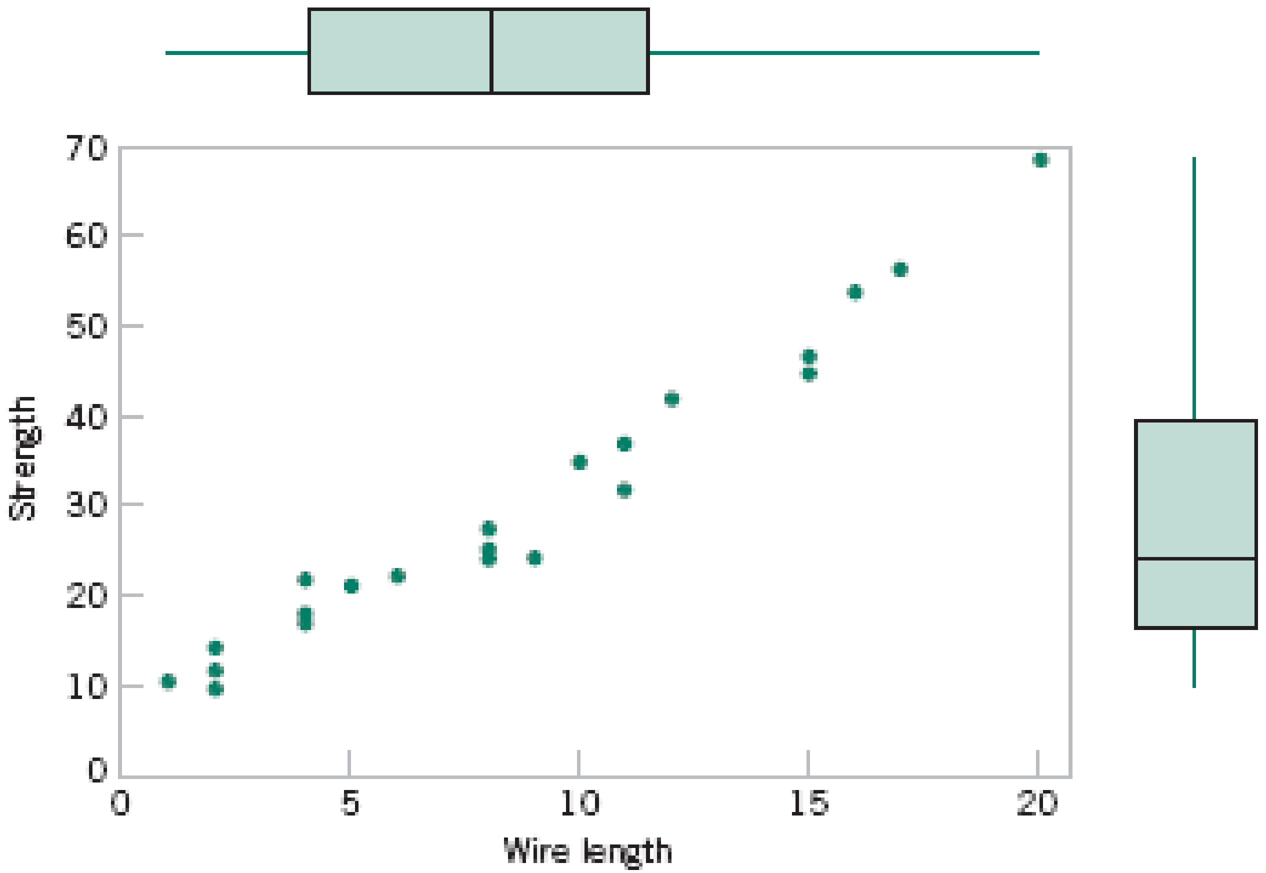


Figure 11-13 Scatter plot of wire bond strength versus wire length, Example 11-8.

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 hypothesis

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

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 10-57. 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

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

IMPORTANT TERMS AND CONCEPTS

- 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
- Logistic regression
- Model adequacy checking
- Regression analysis
- Odds ratio
- Prediction interval on a future observation
- Residual plots
- Residuals
- Scatter diagram
- Significance of regression
- Simple linear regression model standard errors
- Statistical tests on model parameters
- Transformations