6 Building Empirical Models

REGRESSION MODELS

LEARNING OBJECTIVES

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

1. Use simple linear or multiple linear regression for building empirical models of engineering and scientific data.
2. Analyze residuals to determine if the regression model is an adequate fit to the data or to see if any underlying assumptions are violated.
3. Test statistical hypotheses and construct confidence intervals on regression model parameters.
4. Use the regression model either to estimate a mean or to make a prediction of a future observation.
5. Use confidence intervals or prediction intervals to describe the error in estimation from a regression model.
6. Comment on the strengths and weaknesses of your empirical model.

6-1 Introduction To Empirical Models

Table 6-1 Salt Concentration in Surface Streams and Roadway Area

Observation	Salt Concentration (y)	Roadway Area (x)
1	3.8	0.19
2	5.9	0.15
3	14.1	0.57
4	10.4	0.40
5	14.6	0.70
6	14.5	0.67
7	15.1	0.63
8	11.9	0.47
9	15.5	0.75
10	9.3	0.60
11	15.6	0.78
12	20.8	0.81
13	14.6	0.78
14	16.6	0.69
15	25.6	1.30
16	20.9	1.05
17	29.9	1.52
18	19.6	1.06
19	31.3	1.74
20	32.7	1.62

6-1 Introduction To Empirical Models

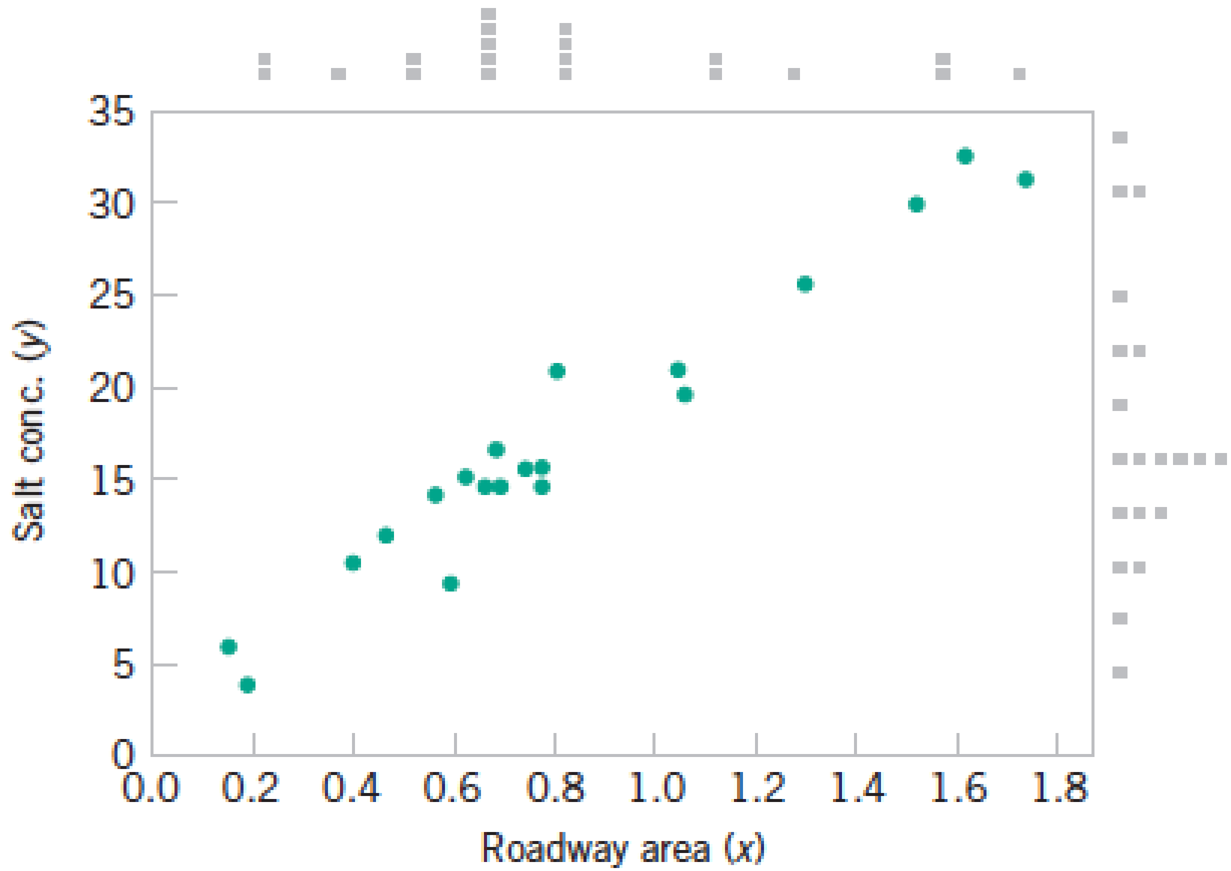
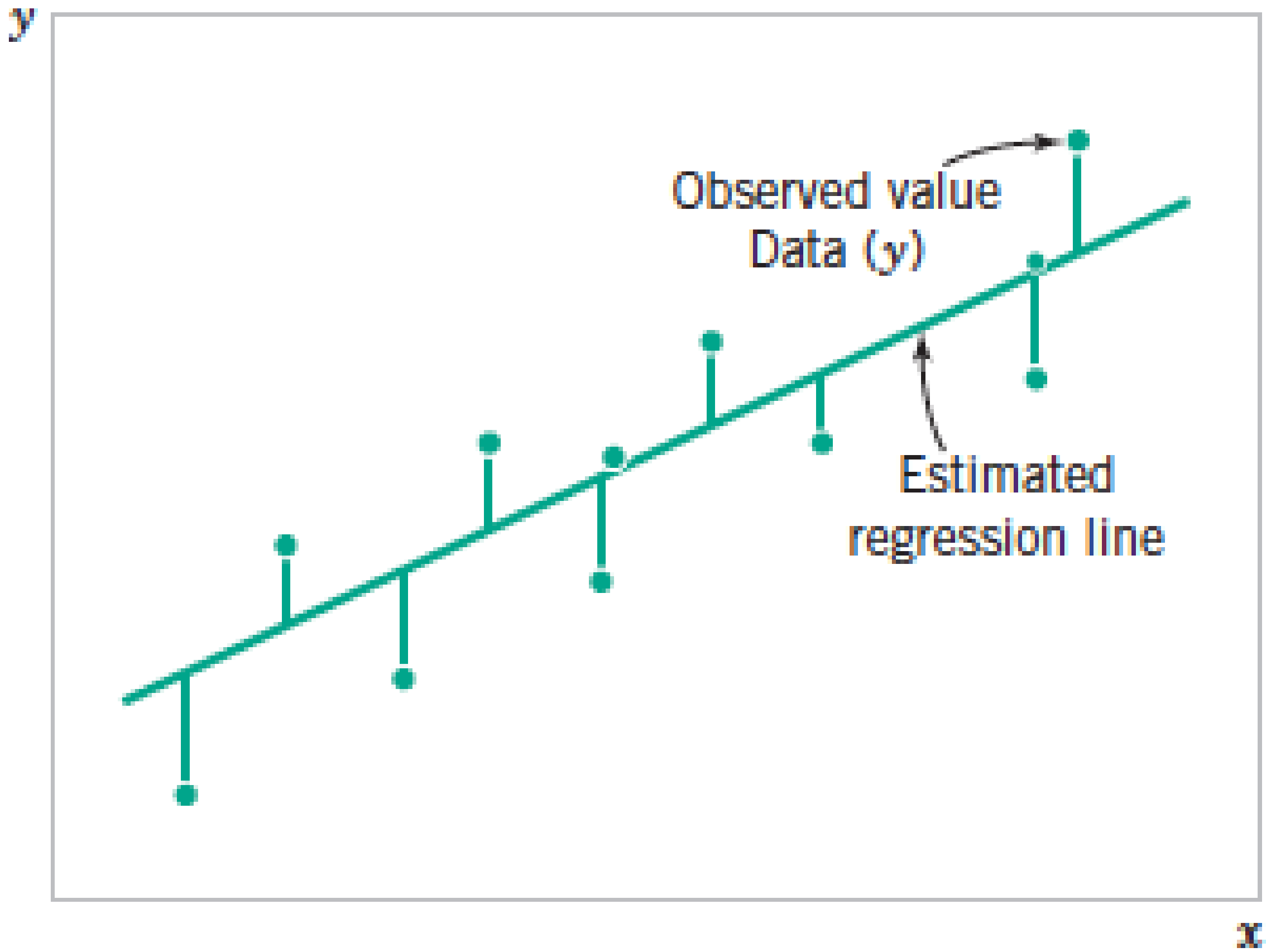


Figure 6-1 Scatter diagram of the salt concentration in surface streams and roadway area data in Table 6-1.

Simple Regression Plot With Data

If we have n observations (x_1, y_1), (x_2, y_2), …, (x_n, y_n).



6-1 Introduction To Empirical Models

Simple Linear Regression Model

In the **simple linear regression model** the dependent variable, or **response,** is related to one independent, or **regressor variable,** as

(6-1)

where ϵ is the random error term. The parameters β_0 and β_1 are called **regression coefficients.**

6-1 Introduction To 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:

$$E(Y|x) = \mu_{Y|x} = \boxed{}$$

where the slope and intercept of the line are called **regression coefficients**.

The **simple linear regression model** is given by

$$Y = \boxed{}$$

where ε is the random error term.

6-1 Introduction To 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

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

The variance of Y given x is

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

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

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

- The **method of least squares** is used to estimate the parameters, β0 and β1 by minimizing the sum of the squares of the vertical deviations in Figure 6-6.

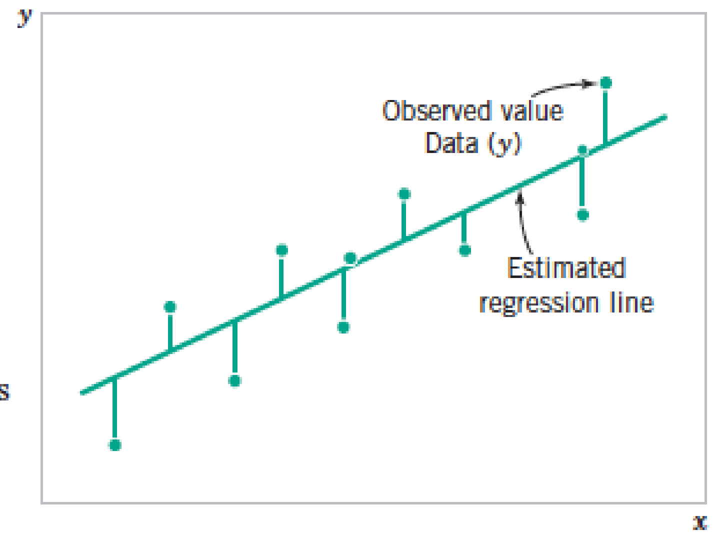


Figure 6-6 Deviations of the data from the estimated regression model.

6-1 Introduction To Empirical Models

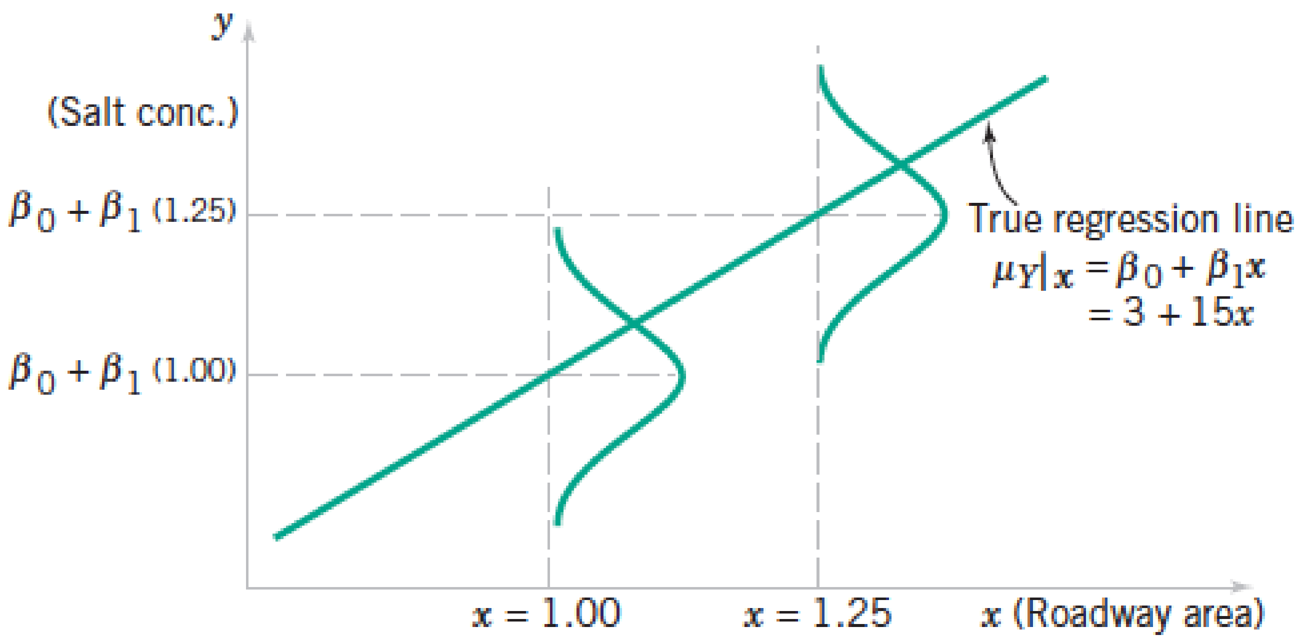


Figure 6-2 The distribution of Y for a given value of x for the salt concentration–roadway area data.

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

• Using Equation 6-8, 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

$$\boxed{L = \sum_{i=1}^{n} \epsilon_i^2 =}$$

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

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

Determining the Coefficients – Differentiate and set the result to Zero

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

Simplifying these two equations yields

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{6-12}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Equations 6-12 are called the **least squares normal equations.** The solution to the normal equations results in the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

Computing Formulas for Simple Linear Regression

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6-13}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{6-14}$$

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

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (6\text{-}15)$$

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. Subsequently we will use the residuals to provide information about the **adequacy** of the fitted model.

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

Notation

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

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

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

EXAMPLE 6-1
Salt Concentration and Roadway Data

Fit a simple linear regression model to the data on salt concentration and roadway area in Table 6-1.

Solution. To build the regression model, the following quantities are computed:

$$n = 20 \quad \sum_{i=1}^{20} x_i = 16.480 \quad \sum_{i=1}^{20} y_i = 342.70 \quad \bar{x} = 0.824 \quad \bar{y} = 17.135$$

$$\sum_{i=1}^{20} y_i^2 = 7060.00 \quad \sum_{i=1}^{20} x_i^2 = 17.2502 \quad \sum_{i=1}^{20} x_i y_i = 346.793$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 17.2502 - \frac{(16.486)^2}{20} = 3.67068$$

and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 346.793 - \frac{(16.480)(342.70)}{20} = 64.4082$$

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

EXAMPLE 6-1

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{64.4082}{3.67068} = 17.5467$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 17.135 - (17.5467)0.824 = 2.6765$$

The fitted simple linear regression model is

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

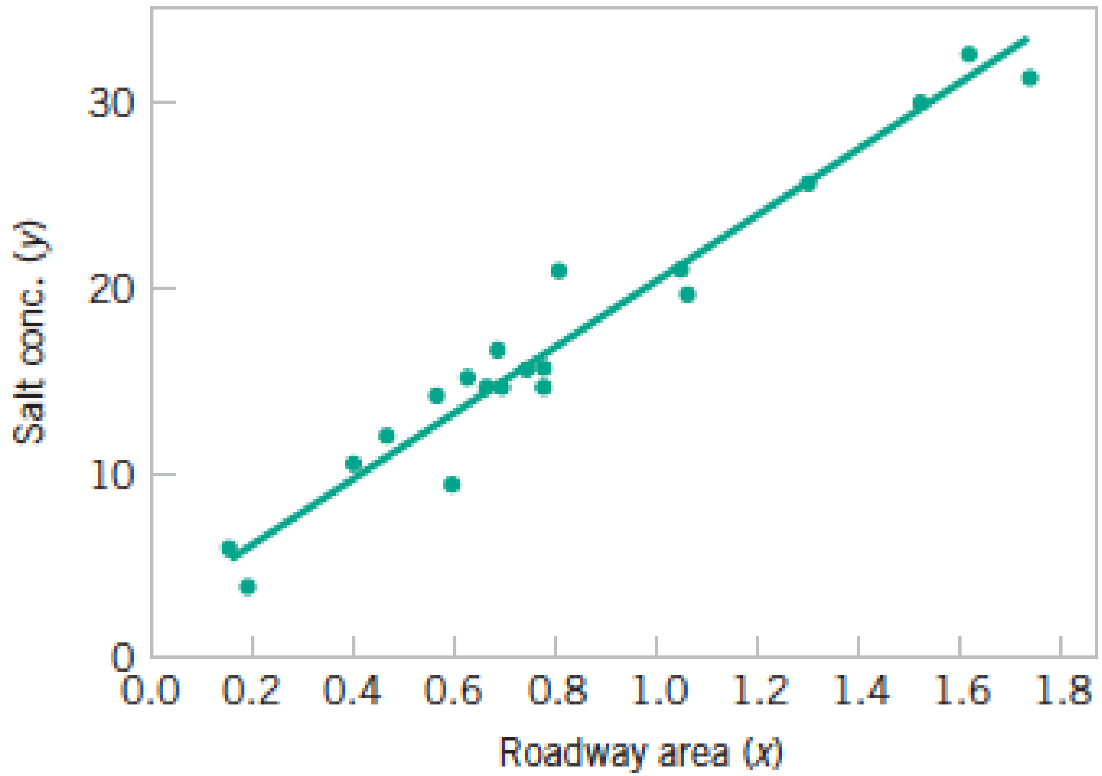


Figure 6-7 Scatter diagram of salt concentration y versus roadway area x and the fitted regression model.

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

Table 6-2 Minitab Regression Analysis Output for Salt Concentration and Roadway Data

Regression Analysis: Salt conc (y) versus Roadway area (x)

The regression equation is

Salt conc (y) = 2.68 + 17.5 Roadway area (x)

Predictor	Coef	SE Coef	T	P
Constant	2.6765 ← $\hat{\beta}_0$	0.8680	3.08	0.006
Roadway area	17.5467 ← $\hat{\beta}_1$	0.9346	18.77	0.000

S = 1.791 ← $\hat{\sigma}$ R-Sq = 95.1% R-Sq(adj) = 94.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1130.1 ← SS_R	1130.1	352.46	0.000
Residual Error	18	57.7 ← SS_E	3.2 ← $\hat{\sigma}^2$		
Total	19	1187.9 ← SS_T			

Model Parameters Evaluation

- Is the Regression Sufficient ? Significant ?
- Is this the right model to use ?
- How accurate are the coefficients
- How accurate is a predicted value of ‘y’ for a new ‘x’
- How far can I extrapolate the results

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

Error Sum of Squares

The **residual sum of squares** (sometimes called the **error sum of squares**) is defined as

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \qquad (6\text{-}16)$$

and for simple linear regression the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \qquad (6\text{-}17)$$

6-2 Simple Linear Regression

Regression Assumptions and Model Properties

Coefficient Estimators, Simple Linear Regression

1. Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased estimators** of the intercept and slope, respectively. That is, the distribution of $\hat{\beta}_1$ (and $\hat{\beta}_0$) is centered at the true value of β_1 (and β_0).
2. The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

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

3. The distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are **normal.**

6-2 Simple Linear Regression

Regression Assumptions and Model Properties

Standard Error of the Slope and Intercept, Simple Linear Regression

The standard errors of the slope and intercept in simple linear regression are

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{6-18}$$

and

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \tag{6-19}$$

respectively.

6-2 Simple Linear Regression

Regression and Analysis of Variance

ANOVA for Regression Analysis

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SS_R + SS_E \qquad (6\text{-}21)$$

Regression Sum of Squares

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

6-2 Simple Linear Regression

Regression and Analysis of Variance

Coefficient of Determination (R^2)

The coefficient of determination is defined as

$$R^2 = 1 - \frac{SS_E}{SS_T} \qquad (6\text{-}22)$$

It is interpreted as the proportion of variability in the observed response variable that is explained by the linear regression model. Sometimes the quantity reported is $100R^2$, and it is referred to as the percentage of variability explained by the model.

6-2 Simple Linear Regression

6-2.2 Testing Hypothesis in Simple Linear Regression

The Analysis of Variance Approach

Table 6-3 Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	1	MS_R	MS_R/MS_E
Error or residual	SS_E	$n - 2$	MS_E	
Total	SS_T	$n - 1$		

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

Table 6-2 Minitab Regression Analysis Output for Salt Concentration and Roadway Data

Regression Analysis: Salt conc (y) versus Roadway area (x)

The regression equation is

Salt conc (y) = 2.68 + 17.5 Roadway area (x)

Predictor	Coef	SE Coef	T	P
Constant	2.6765 ← $\hat{\beta}_0$	0.8680	3.08	0.006
Roadway area	17.5467 ← $\hat{\beta}_1$	0.9346	18.77	0.000

S = 1.791 ← $\hat{\sigma}$ R-Sq = 95.1% R-Sq(adj) = 94.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1130.1 ← SS_R	1130.1	352.46	0.000
Residual Error	18	57.7 ← SS_E	3.2 ← $\hat{\sigma}^2$		
Total	19	1187.9 ← SS_T			

6-2 Simple Linear Regression

6-2.2 Testing Hypothesis in Simple Linear Regression

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

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

6-2 Simple Linear Regression

6-2.2 Testing Hypothesis in Simple Linear Regression

Use of *t*-Tests

We would reject the null hypothesis if

6-2 Simple Linear Regression

6-2.2 Testing Hypothesis in Simple Linear Regression

Use of *t*-Tests

Suppose we wish to test

An appropriate test statistic would be

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

Reject Null Hypothesis if:

6-2 Simple Linear Regression

6-2.2 Testing Hypothesis in Simple Linear Regression

Use of *t*-Tests

An important special case of the hypotheses of Equation 6-23 is

$$H_0\colon \beta_1 = 0$$

$$H_1\colon \beta_1 \neq 0$$

These hypotheses relate to the ______________________________

Failure to reject H_0 is equivalent to concluding that there is no linear relationship between x and Y.

6-2 Simple Linear Regression

6-2.2 Testing Hypothesis in Simple Linear Regression

Use of *t*-Tests

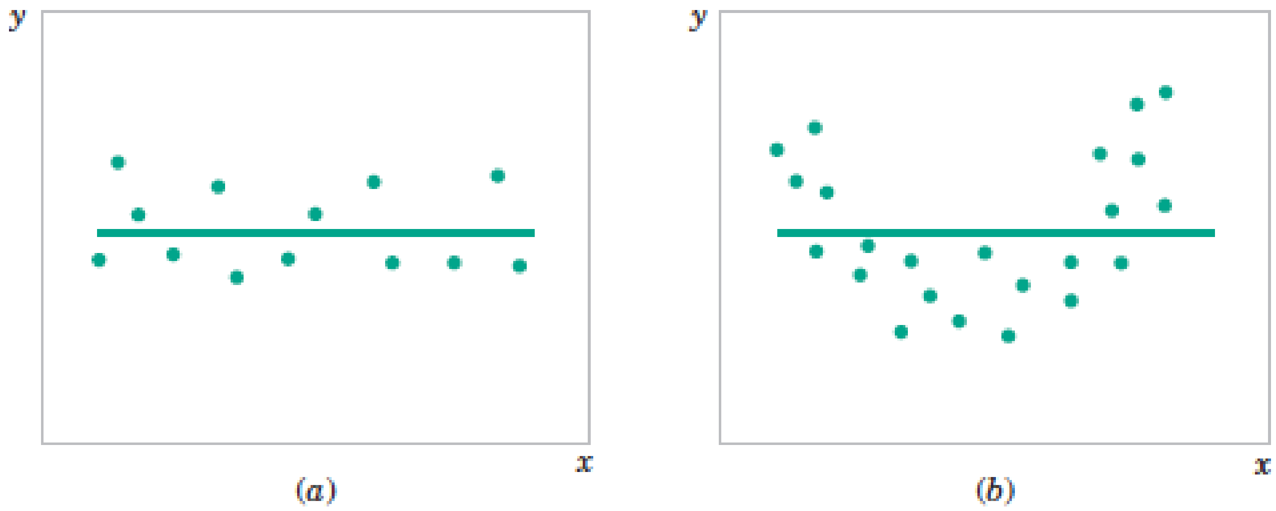


Figure 6-8 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

6-2 Simple Linear Regression

6-2.2 Testing Hypothesis in Simple Linear Regression

Use of *t*-Tests

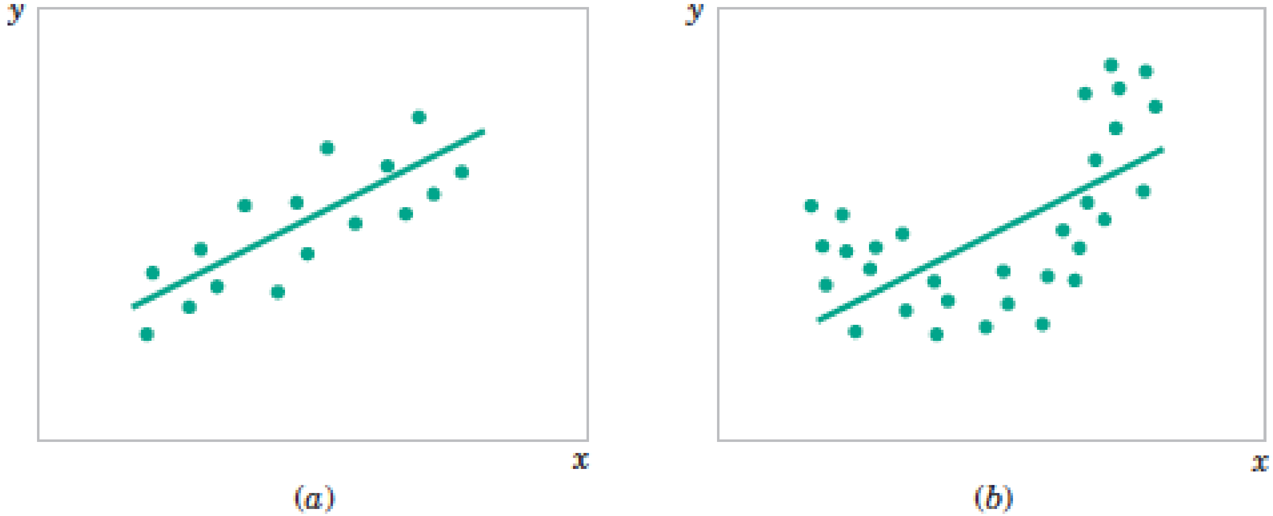


Figure 6-9 The hypothesis H_0: $\beta_1 = 0$ is rejected.

6-2 Simple Linear Regression

The Analysis of Variance Approach

Testing for Significance of Regression in Simple Linear Regression

$$MS_R = \frac{SS_R}{1} \qquad MS_E = \frac{SS_E}{n-p} \tag{6-29}$$

Null hypothesis: H_0: $\beta_1 = 0$

Alternative hypothesis: H_1: $\beta_1 \neq 0$

Test statistic:
$$F_0 = \frac{MS_R}{MS_E} \tag{6-30}$$

Rejection criterion for a fixed-level test: $f_0 > f_{\alpha,1,n-2}$

P-value: Probability beyond f_0 in the $F_{1,n-2}$ distribution

6-2 Simple Linear Regression

6-2.2 Testing Hypothesis in Simple Linear Regression

The Analysis of Variance Approach

Table 6-3 Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	1	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-2$	MS_E	
Total	SS_T	$n-1$		

6-2 Simple Linear Regression

6-2.3 Confidence Intervals in Simple Linear Regression

Confidence Intervals on the Model Parameters in Simple Linear Regression

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

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\, se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\, se(\hat{\beta}_1) \qquad (6\text{-}31)$$

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

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\, se(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\, se(\hat{\beta}_1) \qquad (6\text{-}32)$$

where $se(\hat{\beta}_1)$ and $se(\hat{\beta}_0)$ are defined in equations 6-18 and 6-19, respectively.

6-2 Simple Linear Regression

6-2.3 Confidence Intervals in Simple Linear Regression

Confidence Interval on the Mean Response in Simple Linear Regression

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

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\, se(\hat{\mu}_{Y|x_0}) \le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\, se(\hat{\mu}_{Y|x_0}) \qquad (6\text{-}34)$$

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

6-2 Simple Linear Regression

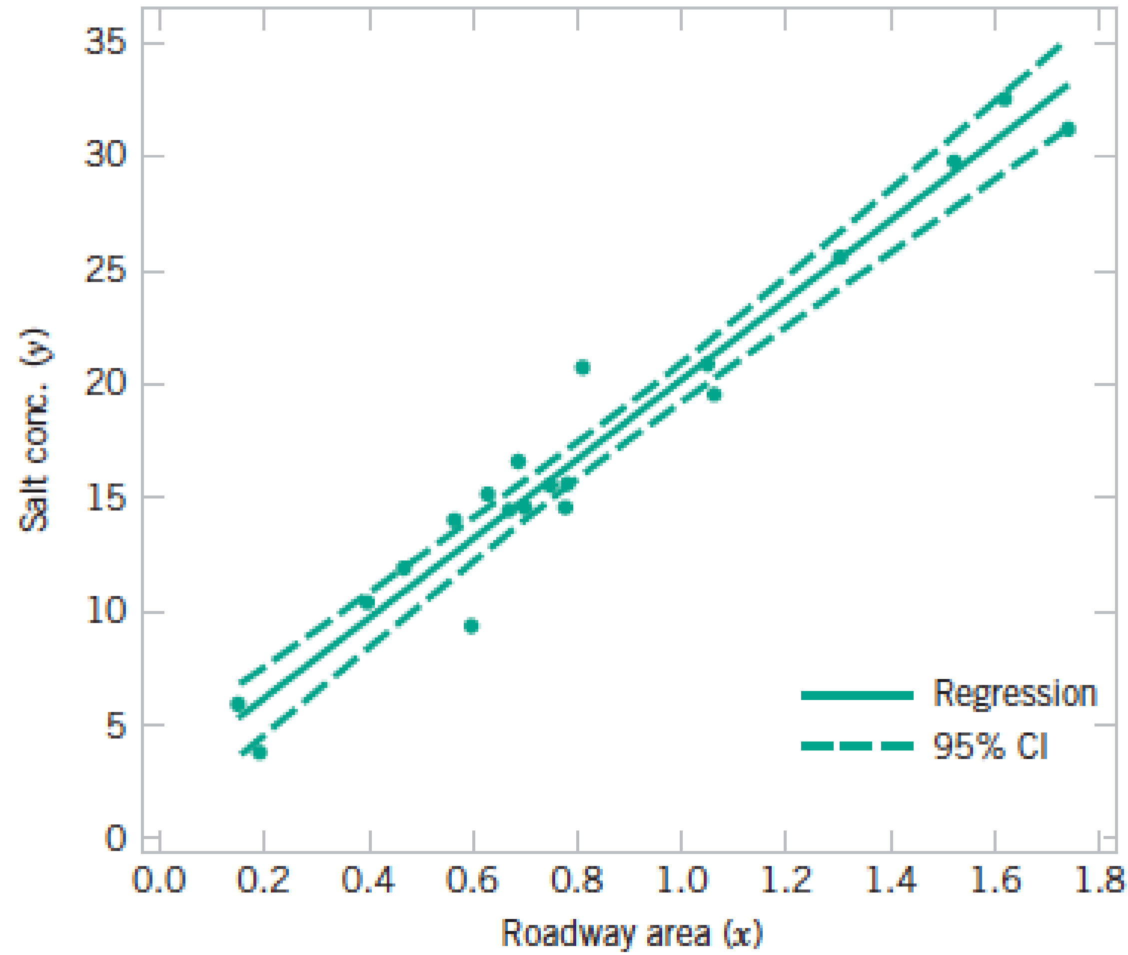


Figure 6-10 Scatter diagram of salt concentration and roadway area from Example 6-1 with fitted regression line and 95% confidence limits on $\mu_{Y|x_0}$.

6-2 Simple Linear Regression

6-2.4 Prediction of Future Observations

Prediction Interval on a Future Observation in Simple Linear Regression

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

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

$$\leq Y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \qquad (6\text{-}36)$$

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

6-2 Simple Linear Regression

6-2.4 Prediction of Future Observations

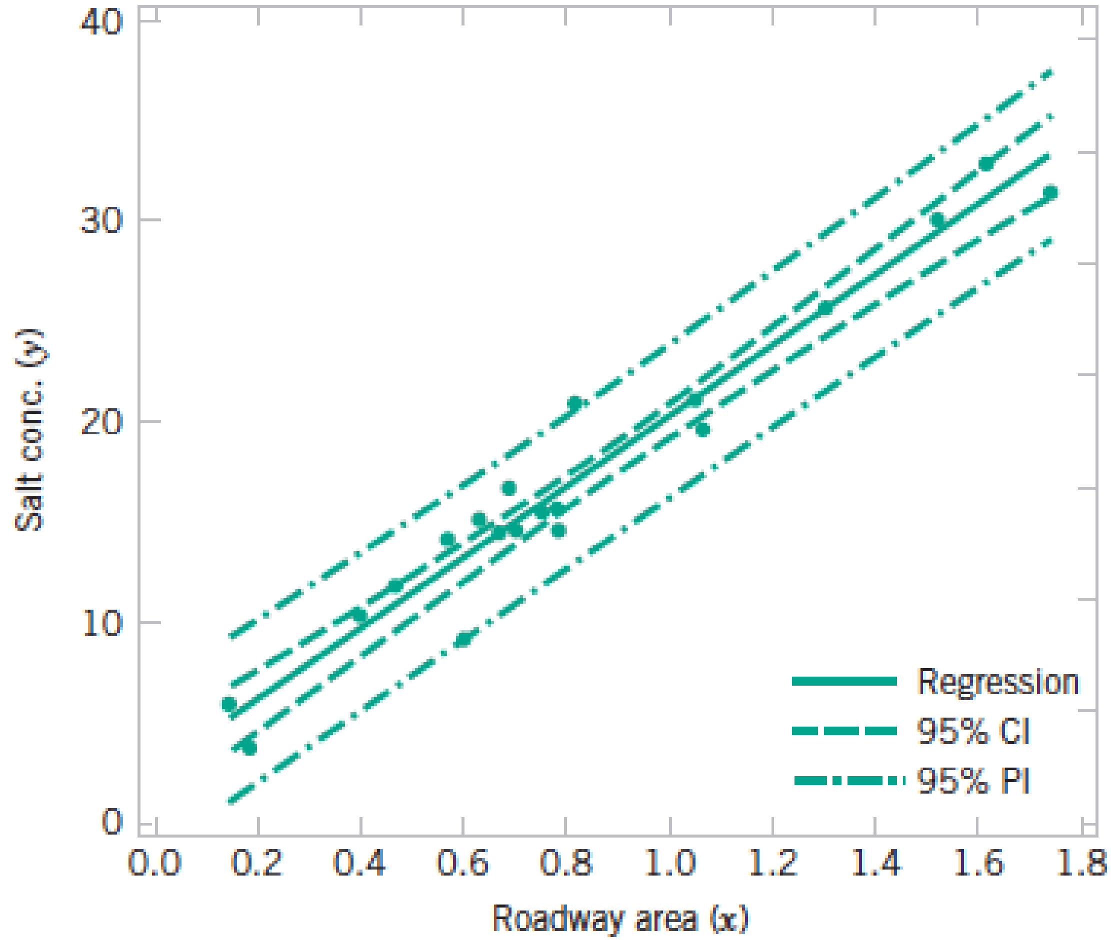


Figure 6-11 Scatter diagram of salt concentration–roadway area data from Example 6-1 with fitted regression line, 95% prediction limits (outer lines), and 95% confidence limits on $\mu_{Y|x_0}$.

6-2 Simple Linear Regression

6-2.5 Checking Model Adequacy

- 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

6-2 Simple Linear Regression

6-2.5 Checking Model Adequacy

- 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.

6-2 Simple Linear Regression

6-2.5 Checking Model Adequacy

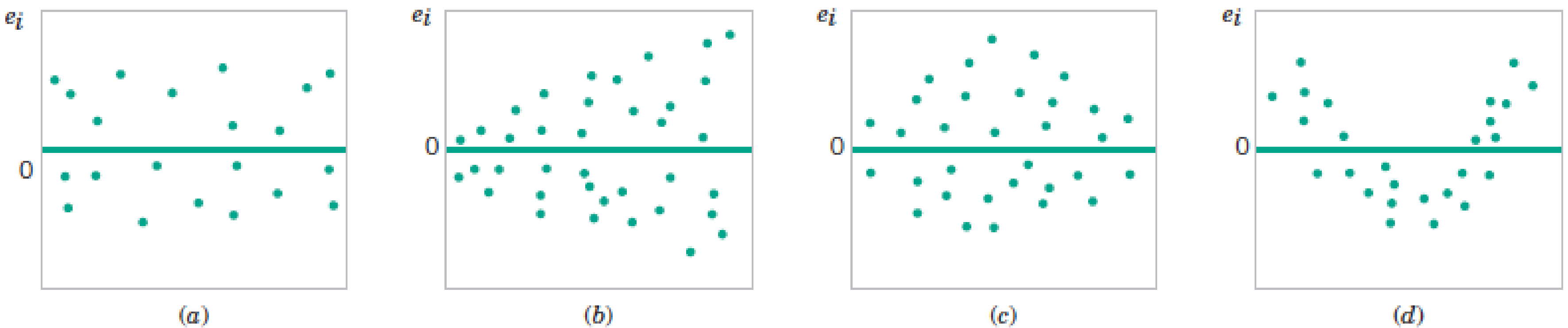


Figure 6-12 Patterns for residual plots: (a) satisfactory, (b) funnel, (c) double bow, (d) nonlinear. Horizontal axis may be time, $\hat{y}_i$, or x_i.

6-2 Simple Linear Regression

6-2.5 Checking Model Adequacy

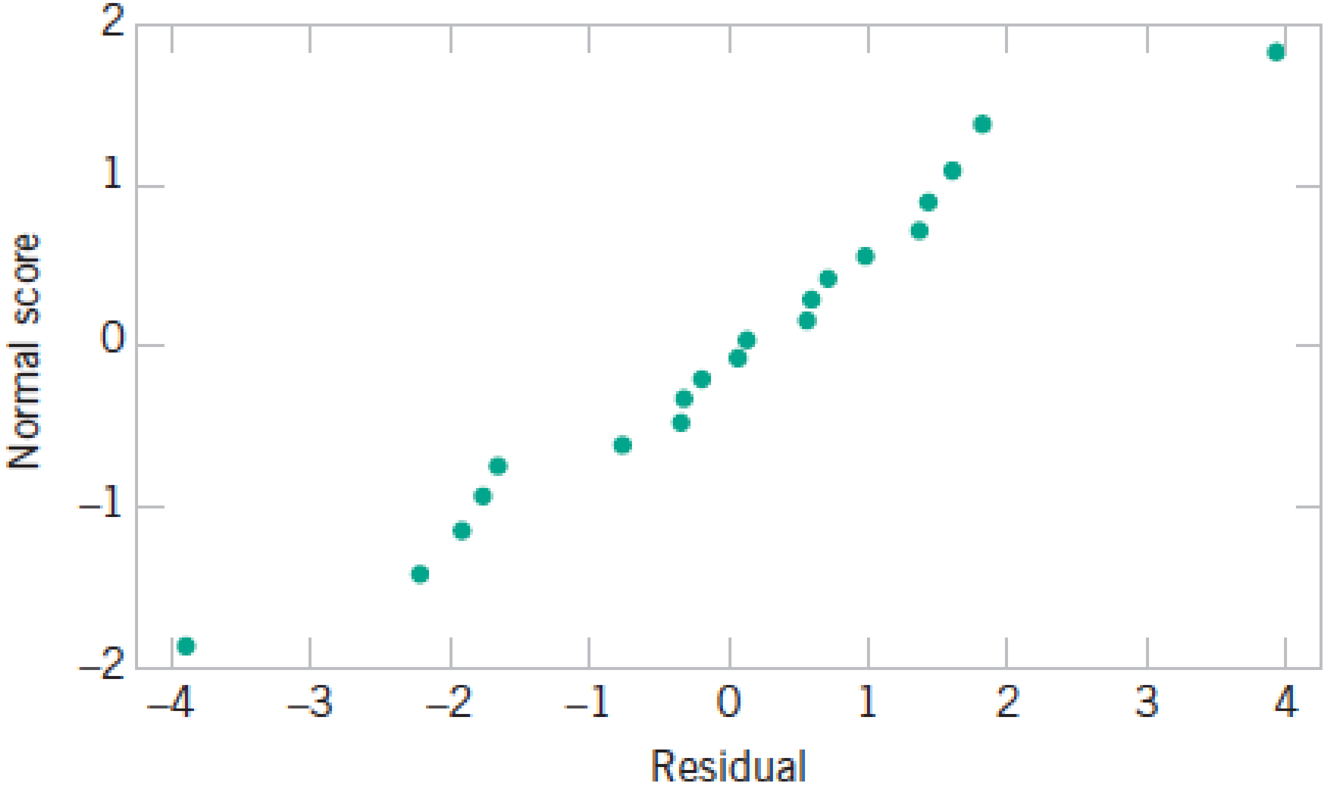


Figure 6-13 Normal probability plot of residuals from the salt concentration–roadway area regression model.

6-2 Simple Linear Regression

6-2.5 Checking Model Adequacy

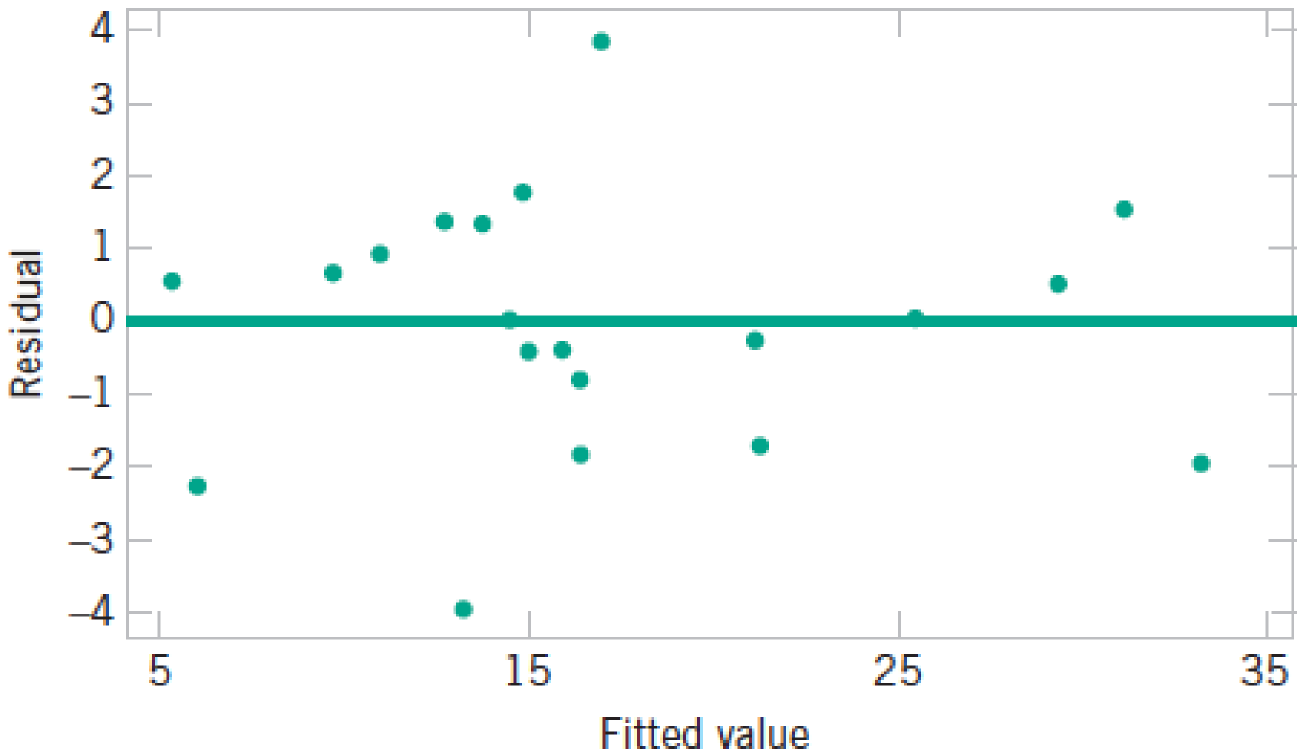


Figure 6-14 Plot of residuals versus fitted values $\hat{y}$ for the salt concentration–roadway area regression model.

6-2 Simple Linear Regression

6-2.5 Checking Model Adequacy

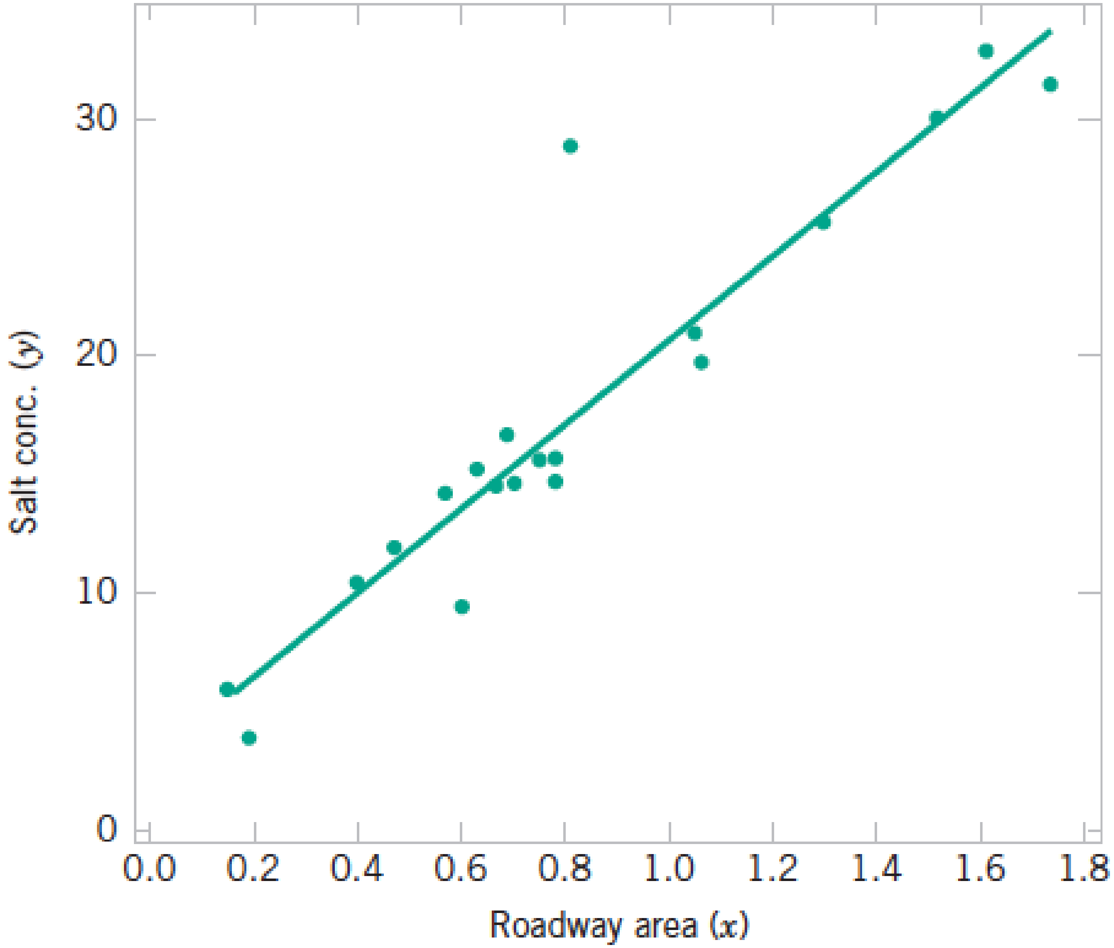


Figure 6-15 Effect of an outlier.

6-2 Simple Linear Regression

6-2.5 Checking Model Adequacy

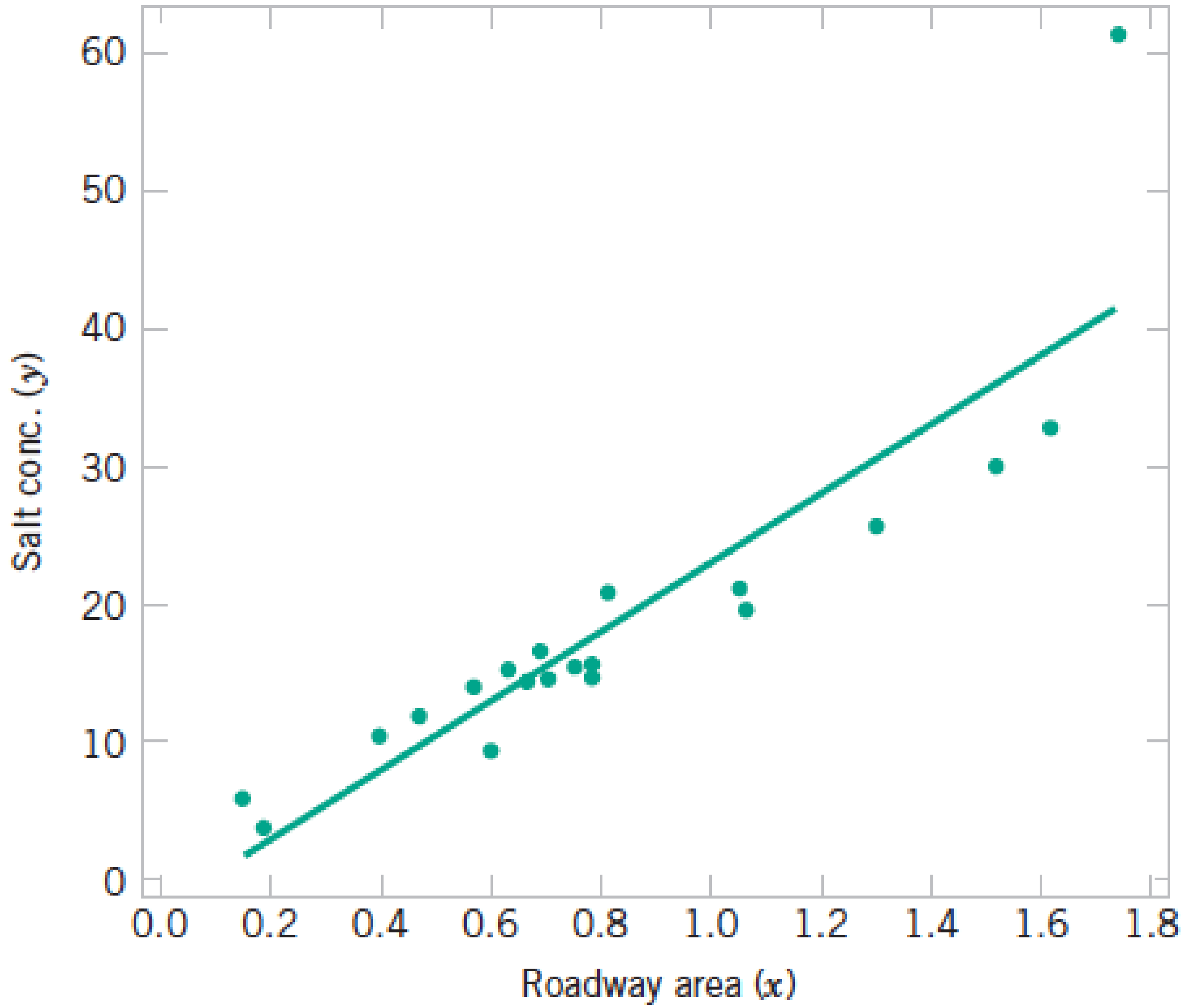


Figure 6-16 Effect of an influential observation.

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

Table 6-2 Minitab Regression Analysis Output for Salt Concentration and Roadway Data

Regression Analysis: Salt conc (y) versus Roadway area (x)

The regression equation is

Salt conc (y) = 2.68 + 17.5 Roadway area (x)

Predictor	Coef	SE Coef	T	P
Constant	2.6765 ← $\hat{\beta}_0$	0.8680	3.08	0.006
Roadway area	17.5467 ← $\hat{\beta}_1$	0.9346	18.77	0.000

S = 1.791 ← $\hat{\sigma}$ R-Sq = 95.1% R-Sq(adj) = 94.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1130.1 ← SS_R	1130.1	352.46	0.000
Residual Error	18	57.7 ← SS_E	3.2 ← $\hat{\sigma}^2$		
Total	19	1187.9 ← SS_T			

6-2 Simple Linear Regression

6-2.1 Least Squares Estimation

Obs	Roadway area	Salt conc	Fit	SE Fit	Residual
1	0.19	3.800	6.010	0.715	−2.210
2	0.15	5.900	5.309	0.746	0.591
3	0.57	14.100	12.678	0.465	1.422
4	0.40	10.400	9.695	0.563	0.705
5	0.70	14.600	14.959	0.417	−0.359
6	0.67	14.500	14.433	0.425	0.067
7	0.63	15.100	13.731	0.440	1.369
8	0.47	11.900	10.923	0.519	0.977
9	0.75	15.500	15.837	0.406	−0.337
10	0.60	9.300	13.205	0.452	−3.905
11	0.78	15.600	16.363	0.403	−0.763
12	0.81	20.800	16.889	0.401	3.911
13	0.78	14.600	16.363	0.403	−1.763
14	0.69	16.600	14.784	0.420	1.816
15	1.30	25.600	25.487	0.599	0.113
16	1.05	20.900	21.101	0.453	−0.201
17	1.52	29.900	29.347	0.764	0.553
18	1.06	19.600	21.276	0.457	−1.676
19	1.74	31.300	33.208	0.945	−1.908
20	1.62	32.700	31.102	0.845	1.598

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	24.610	0.565	(23.424, 25.796)	(20.665, 28.555)

Values of Predictors for New Observations

New Obs	Roadway area
1	1.25