Chapter 2. Simple Linear Regression

Outline

2.1. Covariance and Correlation Coefficient and Example

2.1 Covariance and Correlation Coefficient and Example

Introduction

*We start with the simple case of studying the relationship between a **response** variable Y and a **predictor** variable X_1 .*

Since we have only one predictor variable, we shall drop the subscript in X_1 and use X for simplicity.

*We discuss **covariance** and **correlation coefficient** as measures of the direction and strength of the **linear relationship** between the two variables.*

*A **simple linear regression** model is then formulated and the key **theoretical** results are given and **illustrated** by numerical examples.*

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

*Suppose we have **observations** on **n** subjects consisting of a dependent or **response** variable Y and an **explanatory** variable X. The observations are usually recorded as in the Table 2.1.*

Table 2.1 Notation for the Data Used in Simple Regression and Correlation

Observation Number	Response Variable Y	Predictor X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

Example

*To study the relationship between the heights of father and son, we surveyed **221** pairs of fathers and sons, measured their heights. Then*

$$n = 221; \quad \textbf{response} \text{ variable } \leftarrow \text{ son's height}; \quad \textbf{predictor} \leftarrow \text{ father's height}$$

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

*We wish to measure both the **direction** and the **strength** of the relationship between Y and X. Two related measures, known as the **covariance** and the **correlation** coefficient, are developed below.*

On the scatter plot of Y versus X, let us draw a vertical line at $\bar{x}$ and a horizontal line at $\bar{y}$, where

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2.1}$$

are the sample mean of Y and X, respectively. The two lines divide the graph into four quadrants.

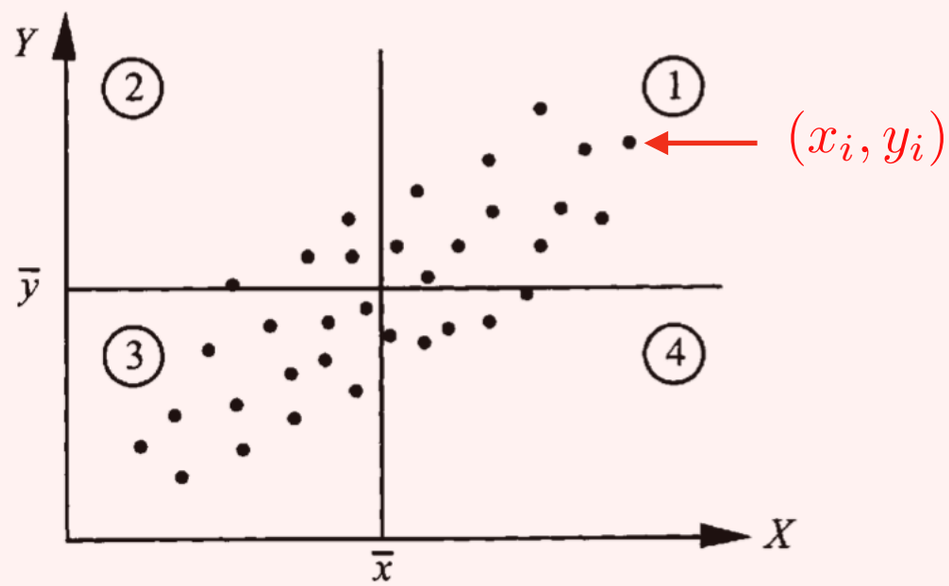


Figure 2.1 Graphical illustration of the correlation coefficient.

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

For each point i in the graph, compute the following quantities:

- $y_i - \bar{y}$, the deviation of each observation y_i from the mean of the response variable,
- $x_i - \bar{x}$, the deviation of each observation x_i from the mean of the predictor variable, and
- the product of the above two quantities, $(y_i - \bar{y})(x_i - \bar{x})$.

It is clear from the graph that the quantity $y_i - \bar{y}$ is positive for every point in the first and second quadrants and is negative for every point in the third and fourth quadrants. Similarly, the quantity $x_i - \bar{x}$ is positive for every point in the first and fourth quadrants and is negative for every point in the second and third quadrants.

Table 2.2 Algebraic Signs of the Quantities $(y_i - \bar{y})$ and $(x_i - \bar{x})$

Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
1	+	+	+
2	+	−	−
3	−	−	+
4	−	+	−

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

*If the linear relationship between Y and X is **positive** (as X increases Y also increases), then there are more points in the first and third quadrants **than** in the second and fourth quadrants. In this case, the sum of the last column in Table 2.2 is likely to be **positive** because there are more positive **than** negative quantities.*

*Conversely, if the relationship between Y and X is **negative** (as X increases Y decreases), then there are more points in the second and fourth quadrants than in the first and third quadrants. Hence the sum of the last column in Table 2.2 is likely to be **negative**. Therefore, the sign of the quantity*

$$\text{Cov}(Y,X) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{n-1}, \qquad (2.2)$$

*which is known as the covariance between Y and X, indicates the **direction** of the linear relationship between Y and X.*

If $\text{Cov(Y,X)} > 0$, then there is a positive relationship between Y and X;

but if $\text{Cov(Y,X)} < 0$, then the relationship is negative.

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

***Unfortunately**, Cov(Y, X) does not tell us much about the strength of such a relationship because it is affected by changes in the units of measurement. For example, we would get two different values for the Cov(Y,X) if we report Y and/or X in terms of thousands of dollars instead of dollars. To avoid this disadvantage of the covariance, we **standardize** the data before computing the covariance.*

*To **standardize** the Y data, we first subtract the mean from each observation then divide by the standard deviation, that is, we compute*

$$z_i = \frac{y_i - \bar{y}}{s_y}, \quad (2.3) \qquad \text{where} \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \quad (2.4)$$

is the sample *standard deviation* of *Y. It can be shown that the **standardized** variable Z has mean zero and standard deviation one. We standardize X in a similar way.*

*The covariance between the **standardized** X and Y data is known as the **correlation coefficient** between Y and X and is given by*

$$\text{Cor}(Y, X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s_y}\right)\left(\frac{x_i - \bar{x}}{s_x}\right). \quad (2.5)$$

Looks complicate? We have a simpler definition.

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

Equivalent formulas for the correlation coefficient are

$$\text{Cor}(Y,X) = \frac{\text{Cov}(Y,X)}{s_y s_x} \tag{2.6}$$

$$= \frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum(y_i-\bar{y})^2\sum(x_i-\bar{x})^2}}. \tag{2.7}$$

Thus, Cor(Y, X) can be interpreted either as the covariance between the standardized variables ***or*** *the* ***ratio*** *of the covariance to the standard deviations of the two variables.*

Correlation coefficient is symmetric, that is Cor(Y,X)=Cor(X,Y).

Unlike *Cov(Y, X), Cor(Y, X) is* ***scale invariant,*** *that is, it does not change if we change the units of measure.*

Further more, Cor(Y, X) satisfies $-1 \leq \text{Cor}(Y,X) \leq 1$ *. These properties make the Cor(Y, X) a useful* ***quantity*** *for measuring both the direction and the strength of the relationship between Y and X.*

The ***magnitude*** *of Cor(Y, X) measures the strength of the linear relationship between Y and X. The closer Cor(Y, X) is to 1 or -1, the* ***stronger*** *is the relationship between Y and X. The sign of Cor(Y, X) indicates the direction of the relationship between Y and X.*

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

Theorem $$-1 \leq \text{Cor}(Y, X) \leq 1 \qquad (2.8)$$

Proof: By Cauchy-Schwartz inequality, $$\left|\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right| \leq \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

*Note, however, that Cor(Y, X) =0 does not necessarily mean that Y and X are not related. It only implies that they are not **linearly** related because the correlation coefficient measures only linear relationships. In other words, the Cor(Y, X) can still be **zero** when Y and X are **nonlinearly** related.*
*For example, Y and X in Table 2.3 have the **perfect** nonlinear relationship $Y = 50 - X^2$ yet Cor(Y,X)=0.*

Table 2.3 Data Set with a Perfect Nonlinear Relationship Between Y and X, Yet $\text{Cor}(X, Y) = 0$

Y	X	Y	X	Y	X
1	−7	46	−2	41	3
14	−6	49	−1	34	4
25	−5	50	0	25	5
34	−4	49	1	14	6
41	−3	46	2	1	7

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

*Note, Furthermore, like many other summary statistics, the Cor(Y, X) can be **substantially** influenced by one or a few **outliers** in the data. To emphasize this point, Anscombe (1973) has constructed four data sets, known as Anscombe **quartet**, each with a **distinct** pattern, but each having the same set of summary statistics (e.g., the same value of the **correlation coefficient**).*
The data and graphs are reproduced in Table 2.4 and Figure 2.3.

Table 2.4 Anscombe Quartet: Four Data Sets Having Same Values of Summary Statistics

Y_1	X_1	Y_2	X_2	Y_3	X_3	Y_4	X_4
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

Source: Anscombe (1973).

2.1 Covariance and Correlation Coefficient and Example

Covariance and Correlation Coefficient

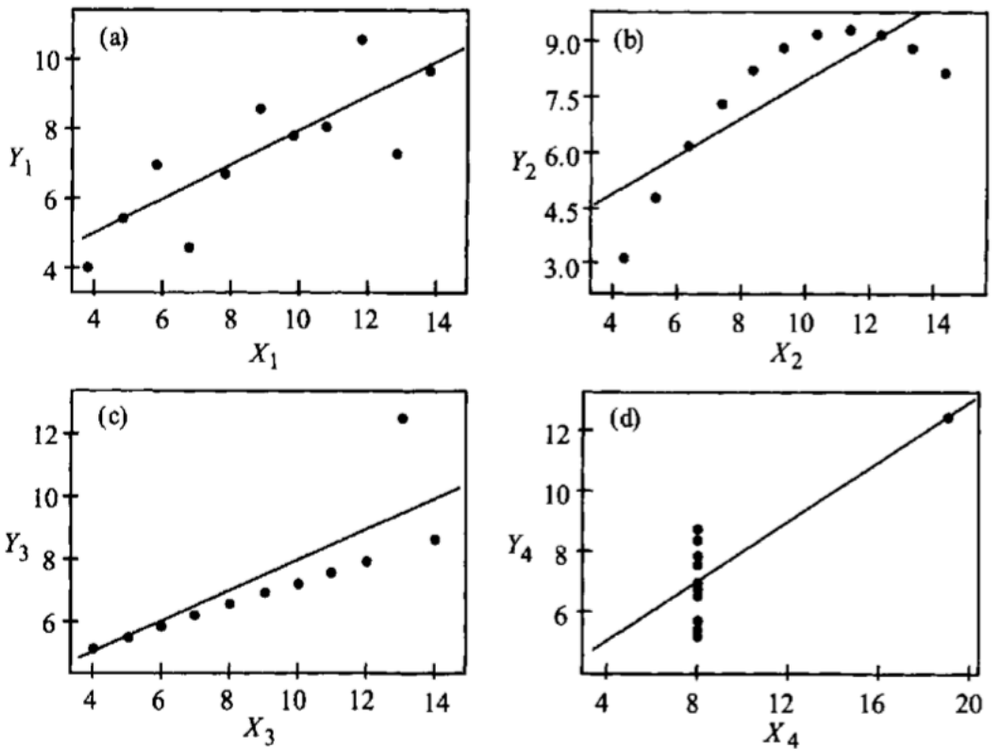


Figure 2.3 Scatter plots of the data in Table 2.4 with the fitted lines.

*An examination of Figure 2.3 shows that only the **first** set, whose plot is given in (a), can be described by a linear model. The plot in (b) shows the second data set is distinctly **nonlinear** and would be better fitted by a quadratic function. The plot in (c) shows that the third data set has one point that **distorts** the slope and the intercept of the fitted line. The plot in (d) shows that the fourth data set is unsuitable for linear fitting, the fitted line being determined essentially by one extreme observation. Therefore, it is **important** to examine the scatter plot of Y versus X before interpreting the numerical value of Cor(Y, X).*

2.1 Covariance and Correlation Coefficient and Example

Example: Computer Repair Data

*As an illustrative example, consider a case of a company that markets and repairs small computers. To study the relationship between the **length** of a service call and the **number** of electronic components in the computer that must be repaired or replaced, a sample of records on service calls was taken. The data consist of the length of service calls in minutes (the response variable) and the number of components repaired (the predictor variable). The data are presented in Table 2.5.*

Table 2.5 Length of Service Calls (in Minutes) and Number of Units Repaired

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

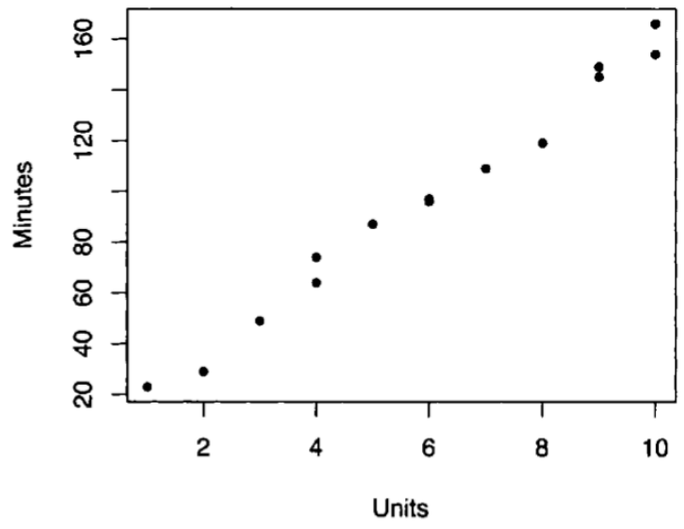


Figure 2.4 Computer Repair data: Scatter plot of Minutes versus Units.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1361}{14} = 97.21 \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{84}{14} = 6,$$

and

$$\text{Cov}(Y, X) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n-1} = \frac{1768}{13} = 136,$$

$$\text{Cor}(Y, X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}} = \frac{1768}{\sqrt{27768.36 \times 114}} = 0.996.$$

2.1 Covariance and Correlation Coefficient and Example

Example: Computer Repair Data

*Before drawing conclusions from this value of Cor(Y, X), we should **examine** the corresponding **scatter** plot of Y versus X. This plot is given in Figure 2.4. The high value of Cor(Y, X) = 0.996 is consistent with the **strong** linear relationship between Y and X exhibited in Figure 2.4. We therefore conclude that there is a strong positive relationship between repair time and units repaired.*

*Although Cor(Y, X) is a useful quantity for measuring the direction and the strength of linear relationships, it **cannot** be used for **prediction** purposes, that is, we cannot use Cor(Y,X) to predict the value of one variable given the value of the other. Furthermore, Cor(Y, X) measures only **pairwise** relationships.*

*Regression analysis, however, can be used to relate **one or more** response variable to one or more predictor variables. It can also be used in prediction. Regression analysis is an attractive extension to correlation analysis because it **postulates** a model that can be used not only to measure the direction and the strength of a relationship between the response and predictor variables, but also to **numerically** describe that relationship.*

2.2. Simple Linear Regression Model and Parameter Estimation

2.2 Simple Linear Regression Model and Parameter Estimation

Simple Linear Regression Model

The relationship between a response variable Y and a predictor variable X is postulated as a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (2.9)$$

where β_0 and β_1 are constants called the **model regression coefficients or parameters** , and ε is a random **error**

*It is assumed that in the range of observations studied, the linear equation (2.9) provides an acceptable approximation to the true relation between Y and X. In other words, Y is approximately a linear function of X, and ε measures the **discrepancy** in that approximation. In particular ε contains no **systematic** information for determining Y that is already **captured** in X.*

The coefficient β_1, called the **slope**, may be interpreted as the change in Y for unit change in X.

The coefficient β_0, called the **intercept**, is the predicted value of Y when X=0.

2.2 Simple Linear Regression Model and Parameter Estimation

Simple Linear Regression Model

According to the simple linear model, each observation in Table 2.1 can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n, \qquad (2.10)$$

where y_i represents the i-th value of the response variable Y, x_i the i-th value of the predictor variable X , and ε_i represents the error in the approximation of y_i.

*Regression analysis **differs** in an important way from correlation analysis. The correlation coefficient is symmetric in the sense that Cor(Y, X) is the same as Cor(X, Y). The variables X and Y are of **equal** importance. In regression analysis the response variable Y is of **primary** importance. The importance of the predictor X lies on its ability to **account** for the variability of the response variable Y and **not** in itself per se. Hence Y is of primary importance.*

Returning to the Computer Repair Data example, suppose that the company wants to forecast the number of service engineers that will be required over the next few years. A linear model,

$$\text{Minutes} = \beta_0 + \beta_1 \text{Units} + \varepsilon, \qquad (2.11)$$

is assumed to represent the relationship between the length of service calls and the number of electronic components in the computer that must be repaired or replaced.

2.2 Simple Linear Regression Model and Parameter Estimation

Parameter Estimation

Based on the available data, we wish to estimate the parameters β_0 and β_1. This is equivalent to finding the straight line that gives the best fit (representation) of the points in the scatter plot of the response versus the predictor variable (see Figure 2.4).

*We estimate the parameters using the popular **least squares method**, which gives the line that **minimizes** the sum of squares of the **vertical** distances from each point to the line. The vertical distances represent the errors in the response variable. These errors can be obtained by rewriting*

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = 1, 2, \cdots, n. \tag{2.12}$$

The sum of squares of these distances can then be written as

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{2.13}$$

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes $S(\beta_0, \beta_1)$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.14}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.15}$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the **least square estimates** because they are the solutions to the **least squares method**

2.2 Simple Linear Regression Model and Parameter Estimation

Parameter Estimation

*The **least squares regression line** is given by*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X. \tag{2.16}$$

*Note that a least squares line **always exists** because we can always find a line that gives the minimum sum of squares of the vertical distances. In fact, as we shall see later, in some cases a least squares line **may not be unique**. These cases are not common in practice.*

For each observation in our data we can compute

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \cdots, n. \tag{2.17}$$

*These are called the **fitted values**. Thus, the i-th fitted value, $\hat{y}_i$, is the point on the l**east squares regression line** corresponding to x_i. The **vertical distance** corresponding to the i-th observation is*

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n. \tag{2.18}$$

*These **vertical** distances are called the ordinary least squares **residuals**. One property of the residuals is that their **sum** is **zero**. This means that the sum of the distances above the line is **equal** to the sum of the distances below the line.*

2.2 Simple Linear Regression Model and Parameter Estimation

Example: Computer Repair Data

Using the Computer Repair data and the quantities in Table 2.6, we have

Table 2.6 Quantities Needed for Computation of Correlation Coefficient Between Length of Service Calls, Y, and Number of Units Repaired, X

i	y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})(x_i - \bar{x})$
1	23	1	−74.21	−5	5507.76	25	371.07
2	29	2	−68.21	−4	4653.19	16	272.86
3	49	3	−48.21	−3	2324.62	9	144.64
4	64	4	−33.21	−2	1103.19	4	66.43
5	74	4	−23.21	−2	538.90	4	46.43
6	87	5	−10.21	−1	104.33	1	10.21
7	96	6	−1.21	0	1.47	0	0.00
8	97	6	−0.21	0	0.05	0	0.00
9	109	7	11.79	1	138.90	1	11.79
10	119	8	21.79	2	474.62	4	43.57
11	149	9	51.79	3	2681.76	9	155.36
12	145	9	47.79	3	2283.47	9	143.36
13	154	10	56.79	4	3224.62	16	227.14
14	166	10	68.79	4	4731.47	16	275.14
Total	1361	84	0.00	0	27768.36	114	1768.00

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{1768}{114} = 15.509,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 97.21 - 15.509 \times 6 = 4.162.$$

2.2 Simple Linear Regression Model and Parameter Estimation

Example: Computer Repair Data

Then the equation of the least squares regression line is

$$\text{Minutes} = 4.162 + 15.509\text{Units}. \tag{2.19}$$

*This least squares line is shown together with the **scatter** plot of Minutes versus Units in Figure 2.5.*

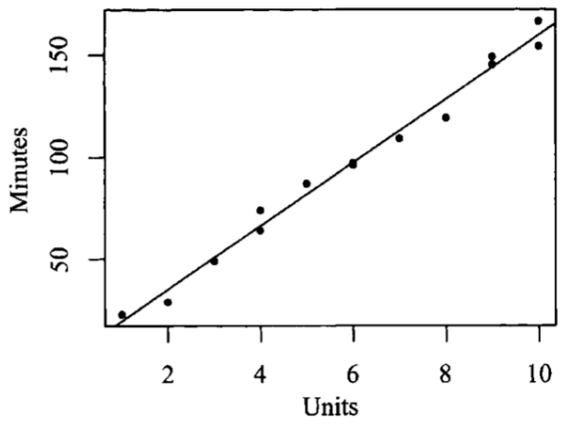


Figure 2.5 Plot of Minutes versus Units with the fitted least squares regression line.

These regression coefficients can be interpreted in physical terms. The constant term represents the setup or startup time for each repair and is approximately 4 minutes. The coefficient of Units represents the increase in the length of a service call for each additional component that has to be repaired. From the data given, we estimate that it takes about 15.5 minutes for each additional component that has to be repaired.

*For example, the length of a service call in which four components had to be repaired is obtained by substituting Units = 4 in the equation of the **regression line** and obtaining* $\hat{y} = 4.162 + 15.509 \times 4 = 66.20$

2.2 Simple Linear Regression Model and Parameter Estimation

Example: Computer Repair Data

Table 2.7 Fitted Values, $\hat{y}_i$, and Ordinary Least Squares Residuals, e_i, for Computer Repair Data

i	x_i	y_i	$\hat{y}_i$	e_i	i	x_i	y_i	$\hat{y}_i$	e_i
1	1	23	19.67	3.33	8	6	97	97.21	−0.21
2	2	29	35.18	−6.18	9	7	109	112.72	−3.72
3	3	49	50.69	−1.69	10	8	119	128.23	−9.23
4	4	64	66.20	−2.20	11	9	149	143.74	5.26
5	4	74	66.20	7.80	12	9	145	143.74	1.26
6	5	87	81.71	5.29	13	10	154	159.25	−5.25
7	6	96	97.21	−1.21	14	10	166	159.25	6.75

2.2 Simple Linear Regression Model and Parameter Estimation

An alternative formula for $\hat{\beta}_1$ can be expressed as

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \text{Cor}(Y, X)\frac{s_y}{s_x} \qquad (2.20)$$

form which it can be seen that $\hat{\beta}_1, \text{Cov}(Y, X),$ and $\text{Cor}(Y, X)$ ahve the same sign.

This makes intuitive sense because positive (negative) slope means positive (negative) correlation.

*So far in our analysis we have made only one assumption, namely, that Y and X are linearly related. This assumption is referred to as the **linearity** assumption.*

*This is merely an **assumption** or a hypothesis about the relationship between the response and predictor variables. An early step in the analysis should always be the **validation** of this assumption. We wish to determine if the data at hand support the assumption that Y and X are linearly related. An **informal** way to check this assumption is to examine the **scatter** plot of the response versus the predictor variable, preferably drawn with the least squares line superimposed on the graph (see Figure 2.5). If we observe a nonlinear pattern, we will have to take corrective action.*

2.2 Simple Linear Regression Model and Parameter Estimation

How the Least Square Solution is Obtained?

Theorem The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes $S(\beta_0, \beta_1)$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof Recall that $S(\beta_0, \beta_1) = \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2$

The minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ should vanish the derivative. So

$$\left.\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1}\right|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = 2\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i) x_i = 0$$

$$\left.\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = 2\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i) = 0$$

$\Rightarrow$ Solve for $\hat{\beta}_0$ and $\hat{\beta}_1$

2.3. Test of Hypothesis and Confidence Intervals

2.3 Test of Hypothesis and Confidence Intervals

Core Assumptions

*As stated earlier, the **usefulness** of X as a predictor of Y can be measured **informally** by examining the **correlation coefficient** and the corresponding **scatter** plot of Y versus X. A more formal way of measuring the usefulness of X as a predictor of Y is to conduct a **test of hypothesis** about the regression parameter β_1 . Note that the hypothesis $\beta_1 = 0$ means that there is **no linear relationship** between Y and X.*

A test of this hypothesis requires the following assumptions:

Core Assumption

For every fixed value of X, the error ε's are independent normal random variables with mean zero and variance σ^2

$$\varepsilon_1, \cdots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

2.3 Test of Hypothesis and Confidence Intervals

Distribution of Least Square Estimators

Since the error $\varepsilon's$ are random, the least squares **estimator** $\hat{\beta}_0$ and $\hat{\beta}_1$ are also random.

Then, what are the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$?

Theorem Under the core assumptions, $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed with

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right) \tag{2.21}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \tag{2.22}$$

Proof

2.3 Test of Hypothesis and Confidence Intervals

Distribution of Least Square Estimators

Remarks $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1.

*The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the **unknown** parameter σ^2. So we need to estimate σ^2 from the data, an **unbiased** estimate of σ^2 is given by*

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}, \qquad (2.23)$$

*where SSE is the **sum of squares of the residuals** (errors). The number n- 2 in the denominator is called the **degrees of freedom** (df). It is equal to the number of observations **minus** the number of estimated regression coefficients.*

Replacing σ^2 by $\hat{\sigma}^2$, we get **unbaised** estimates of the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$.

*An estimate of the **standard deviation** is called the **standard error** (s.e.) of the estimator.*

2.3 Test of Hypothesis and Confidence Intervals

Distribution of Least Square Estimators

Thus, the **standard errors** of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\text{s.e.}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}} \quad (2.24)$$

$$\text{s.e.}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}, \quad (2.25)$$

*respectively, where $\hat{\sigma}$ is the square root of $\hat{\sigma}^2$. The standard error $\hat{\beta}_1$ is a measure of how **precisely** the slope has been estimated. The **smaller** the standard error the more **precise** the estimator.*

*With the **sampling distributions** of $\hat{\beta}_0$ and $\hat{\beta}_1$, we are now in position to perform **statistical analysis** concerning the usefulness of X as a predictor of Y. Under the **normality assumption**, an appropriate test statistic for testing the **null** hypothesis $H_0 : \beta_1 = 0$ against the **alternative.** $H_1 : \beta_1 \neq 0$ is the **t-Test**,*

$$t_1 = \frac{\hat{\beta}_1}{\text{s.e.}(\hat{\beta}_1)}. \quad (2.26)$$

The statistic t_1 is distributed as a Student's t with $n - 2$ degrees of freedom.

2.3 Test of Hypothesis and Confidence Intervals

Test $H_0 : \beta_1 = 0$

*The test is carried out by comparing this observed value with the appropriate **critical value** obtained from the t-table, which is $t_{n-2,\alpha/2}$, where α is a specified significance level. Note that we divide α by 2 because we have a **two-sided** alternative hypothesis.*
*Accordingly, H_0 is to be **rejected** at the significance level α if*

$$|t_1| \geq t_{(n-2,\alpha/2)}, \qquad (2.27)$$

An equivalent criterion is to compare the p-value for the t-Test with α and reject H_0 if

$$2 \cdot P(T_{n-2} \geq |t_1|) \leq \alpha \qquad (2.28)$$

where T_{n-2} denotes the student t random variable with df=n-2.

$2 \cdot P(T_{n-2} \geq |t_1|)$ is called the **p-value**

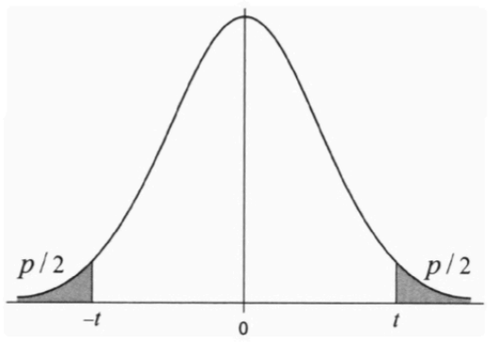


Figure 2.6 Graph of the probability density function of a t-distribution. The p-value for the t-Test is the shaded areas under the curve.

Figure 2.6 is a graph of the density function of a t-distribution. The p-value is the sum of the two shaded areas under the curve.

2.3 Test of Hypothesis and Confidence Intervals

Test $H_0 : \beta_1 =$ any specific value

*The previous t-Test can be **generalized** to test the more **general** hypothesis $H_0 : \beta_1 = \beta_1^0$, where β_1^0 is a constant chosen by the **investigator**, against the **two-sided alternative** $H_1 : \beta_1 \neq \beta_1^0$.*

The appropriate test statistic in this case is the t-Test:

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\text{s.e.}(\hat{\beta}_1)} . \qquad (2.29)$$

The statistic t_1 is distributed as a Student's t with $n-2$ degrees of freedom.

Thus, $H_0 : \beta_1 = \beta_1^0$ is rejected if $|t_1| \geq t_{(n-2,\alpha/2)}$ holds.

Example (computer repair data)

Suppose that the management expected the increase in service time for each additional unit to be repaired to be 12 minutes. Do the data support this conjecture? The answer may be obtained by testing $H_0 : \beta_1 = 12$ against $H_1 : \beta_1 \neq 12$. The appropriate statistic is

$$t_1 = \frac{\hat{\beta}_1 - 12}{\text{s.e.}(\hat{\beta}_1)} = \frac{15.509 - 12}{0.505} = 6.948,$$

with 12 degrees of freedom. The critical value for this test is $t_{(12,0.025)} = 2.18$. Since 6.948>2.18, the result is highly significant, leading to the rejection of the null hypothesis. The management's estimate of the increase in time for each additional component to be repaired is not supported by the data.
*Their estimate is too **low**.*

2.3 Test of Hypothesis and Confidence Intervals

Test $H_0 : \beta_0 =$ any specific value

The need for testing hypotheses regarding the regression ***parameter*** β_0 *may also arise in practice. More specifically, suppose we wish to test* $H_0 : \beta_0 = \beta_0^0$ *against* $H_1 : \beta_0 \neq \beta_0^0$ *, where* β_0^0 *is a constant chosen by the* ***investigator****. The appropriate test in this case is given by*

$$t_0 = \frac{\hat{\beta}_0 - \beta_0^0}{\text{s.e.}(\hat{\beta}_0)} . \qquad (2.30)$$

The ***least squares estimates*** *of the regression coefficients, their* ***standard errors****, the t-Tests for testing that the corresponding coefficient is zero, and the* ***p-values*** *are usually given as part of the regression output by statistical packages. These values are usually displayed in a table such as the one in Table 2.8. This table is known as the* ***coefficient table****.*

Table 2.8 Standard Regression Output.

Variable	Coefficient (Formula)	s.e. (Formula)	t-Test (Formula)	p-value
Constant	$\hat{\beta}_0$	s.e.$(\hat{\beta}_0)$	t_0	p_0
X	$\hat{\beta}_1$	s.e.$(\hat{\beta}_1)$	t_1	p_1

2.3 Test of Hypothesis and Confidence Intervals

Test $H_0 : \beta_0 =$ any specific value

A Test Using Correlation Coefficient

As mentioned above, a test of $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ can be thought of as a test for determining whether the response and the predictor variables are linearly related. We used the t-Test in (2.26) to test this hypothesis. An alternative test, which involves the correlation coefficient between Y and X, can be developed. Suppose that the population correlation coefficient between Y and X is denoted by ρ. If $\rho \neq 0$, then Y and X are linearly related. An appropriate test for testing $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ is given by

$$t_1 = \frac{\text{Cor}(Y, X)\sqrt{n-2}}{\sqrt{1 - [\text{Cor}(Y, X)]^2}}, \tag{2.32}$$

where $\text{Cor}(Y, X)$ is the sample correlation coefficient between Y and X, defined in (2.6), which is considered here to be an estimate of ρ. The t-Test in (2.32) is distributed as a Student's t with $n - 2$ degrees of freedom. Thus, $H_0 : \rho = 0$ is rejected if (2.27) holds [or, equivalently, if (2.28) holds]. Again if $H_0 : \rho = 0$ is rejected, it means that there is a statistically significant linear relationship between Y and X.

It is clear that if no linear relationship exists between Y and X, then $\beta_1 = 0$. Consequently, the statistical tests for $H_0 : \beta_1 = 0$ and $H_0 : \rho = 0$ should be identical. Although the statistics for testing these hypotheses given in (2.26) and (2.32) look different, it can be demonstrated that they are indeed algebraically equivalent. **How ??**

2.3 Test of Hypothesis and Confidence Intervals

Example (Computer Repair Data)

As an illustrative example, Table 2.9 shows a part of the regression output for the Computer Repair data in Table 2.5. Thus, for example, $\hat{\beta}_1 = 15.509$, the s.e.$(\hat{\beta}_1) = 0.505$, and hence $t_1 = 15.509/0.505 = 30.71$. The critical value for this test using $\alpha = 0.05$, for example, is $t_{(12,0.025)} = 2.18$. The $t_1 = 30.71$ is much larger than its critical value 2.18. Consequently, $H_0 : \beta_1 = 0$ is rejected, which means that the predictor variable Units is a statistically significant predictor of the response variable Minutes. This conclusion can also be reached by observing that the p-value ($p_1 < 0.0001$) is much less than $\alpha = 0.05$ indicating very high significance.

Table 2.9 Regression Output for Computer Repair Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	4.162	3.355	1.24	0.2385
Units	15.509	0.505	30.71	< 0.0001

2.3 Test of Hypothesis and Confidence Intervals

Confidence Interval

To construct confidence intervals for the regression parameters, we also need to assume that the ε's have a normal distribution, which will enable us to conclude that the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal, as discussed before. Consequently, the $(1-\alpha)\times 100\%$ confidence interval for β_0 is given by

$$\hat{\beta}_0 \pm t_{(n-2,\alpha/2)} \times \text{s.e.}(\hat{\beta}_0), \qquad (2.33)$$

where $t_{(n-2,\alpha/2)}$ is the $(1-\alpha/2)$ percentile of a t distribution with $n-2$ degrees of freedom. Similarly, limits of the $(1-\alpha)\times 100\%$ confidence interval for β_1 are given by

$$\hat{\beta}_1 \pm t_{(n-2,\alpha/2)} \times \text{s.e.}(\hat{\beta}_1). \qquad (2.34)$$

Interpretation of Confidence Interval

*If we were to take **repeated** samples of the same size at the **same** values of X and construct, for example, 95% confidence intervals for the **slope** parameter for **each** sample, then 95% of these intervals would be **expected** to contain the **true** value of the slope.*

From Table 2.9 we see that a 95% confidence interval for β_1 is

$$15.509 \pm 2.18 \times 0.505 = (14.408, 16.610). \qquad (2.35)$$

That is, the incremental time required for each broken unit is between 14 and 17 minutes. The calculation of confidence interval for β_0 in this example is left as an exercise

2.4. Prediction

2.4 Prediction

*The fitted regression equation can be used for **prediction**. We distinguish between **two** types of predictions:*

1. The prediction of the value of the response variable Y which corresponds to any chosen value, x_0, of the predictor variable.
2. The estimation of the mean response μ_0, when $X = x_0$.

*For the **first** case, the **predicted** value* $\hat{y}_0$ *is* $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$ (2.36)

The standard error of this prediction is $$\text{s.e.}(\hat{y}_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\,.$$ (2.37)

The confidence interval for the predicted value with confidence coefficient $(1-\alpha)$ *are given by*

Often called prediction interval $$\hat{y}_0 \pm t_{(n-2,\alpha/2)}\ \text{s.e.}(\hat{y}_0).$$ (2.38)

*For the **second** case, the **mean response** value* μ_0 *is estimated by* $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$ (2.39)

The standard error of this estimate is $$\text{s.e.}(\hat{\mu}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\,,$$ (2.40)

from which it follows that the confidence interval for μ_0 *with confidence coefficient* $(1-\alpha)$ *are given by*

Often called confidence interval $$\hat{\mu}_0 \pm t_{(n-2,\alpha/2)}\ \text{s.e.}(\hat{\mu}_0).$$ (2.41)

2.4 Prediction

Note that the point estimate of μ_0 is identical to the predicted response $\hat{y}_0$.

*The standard error of $\hat{\mu}_0$ is, however, **smaller** than the standard error of $\hat{y}_0$ and can be seen by comparing with two **s.e. equations**.*

*Intuitively, this makes sense. There is greater **uncertainty** (variability) in predicting one observation (the next observation) than in estimating the **mean** response when $X = x_0$.*

*The averaging that is implied in the **mean** response reduces the variability and uncertainty associated with the estimate.*

Example (computer repair data)

Suppose that we wish to predict the length of a service call in which four components had to be repaired. If $\hat{y}_4$ denotes the predicted value, then

$$\hat{y}_4 = 4.162 + 15.509 \times 4 = 66.20,$$

with a standard error as

$$\text{s.e.}(\hat{y}_4) = 5.392\sqrt{1 + \frac{1}{14} + \frac{(4-6)^2}{114}} = 5.67.$$

On the other hand, if the service department wishes to estimate the expected (mean) service time for a call that needed four components repaired. Denoting by μ_4 , the expected service time for a call that needed four components to be repaired, we have

$$\hat{\mu}_4 = 4.162 + 15.509 \times 4 = 66.20,$$

with a standard error as

$$\text{s.e.}(\hat{\mu}_4) = 5.392\sqrt{\frac{1}{14} + \frac{(4-6)^2}{114}} = 1.76.$$

2.4 Prediction

As can be seen , the standard error of prediction increases the farther the value of the predictor variable is from the center of the actual observations. Care should be taken when predicting the value of Minutes corresponding to a value for Units that does not lie close to the observed data. There are two dangers in such predictions. First, there is substantial uncertainty due to the large standard error. More important, the linear relationship that has been estimated may not hold outside the range of observations. Therefore, care should be taken in employing fitted regression lines for prediction far outside the range of observations. In our example we would not use the fitted equation to predict the service time for a service call which requires that 25 components be replaced or repaired. This value lies too far outside the existing range of observations.

2.5. Measuring the Quality of Fit

2.5 Measuring the Quality of Fit

__After__ fitting a linear model relating Y to X, we are interested not only in knowing whether a linear relationship __exits__, but also in measuring the __quality__ of the fit of the model to the data. The quality of the fit can be assessed by one of the following __highly related__ (hence, somewhat redundant) ways:

1. When using the tests in $t_1 = \hat{\beta}_1/\text{s.e.}(\hat{\beta}_1)$, if H_0 is rejected, the magnitude of the values of the test (or the corresponding p-values) gives us information about the *strength* (not just the existence) of the linear relationship between Y and X. Basically, the larger the t (in absolute value) or the smaller the corresponding p-value, the stronger the linear relationship between Y and X. These tests are objective but they require all the assumptions stated earlier, specially the assumption of normality of the ε's.

2. The strength of the linear relationship between Y and X can also be assessed directly from the examination of the scatter plot of Y versus X together with the corresponding value of the correlation coefficient $\text{Cor}(Y, X)$
The closer the set of points to a straight line [the closer $\text{Cor}(Y, X)$ to 1 or -1], the stronger the linear relationship between Y and X. This approach is informal and subjective but it requires only the linearity assumption.

2.5 Measuring the Quality of Fit

3. Examine the scatter plot of Y versus $\hat{Y}$. The closer the set of points to a straight line, the stronger the linear relationship between Y and X. One can measure the strength of the linear relationship in this graph by computing the correlation coefficient between Y and $\hat{Y}$, which is given by

$$\text{Cor}(Y, \hat{Y}) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}, \qquad (2.42)$$

where $\bar{y}$ is the mean of the response variable Y and $\bar{\hat{y}}$ is the mean of the fitted values. In fact, the scatter plot of Y versus X and the scatter plot of Y versus $\hat{Y}$ are redundant because the patterns of points in the two graphs are identical. The two corresponding values of the correlation coefficient are related by the following equation:

Why ??

$$\text{Cor}(Y, \hat{Y}) = |\text{Cor}(Y, X)|. \qquad (2.43)$$

absolute value

Note that $\text{Cor}(Y, \hat{Y})$ cannot be negative (why?), but $\text{Cor}(Y, X)$ can be positive or negative $[-1 \leq \text{Cor}(Y, X) \leq 1]$. Therefore, in simple linear regression, the scatter plot of Y versus $\hat{Y}$ is redundant. However, in multiple regression, the scatter plot of Y versus $\hat{Y}$ is not redundant.

2.5 Measuring the Quality of Fit

4. Although scatter plots of Y versus $\hat{Y}$ and $\text{Cor}(Y, \hat{Y})$ are redundant in simple linear regression, they give us an indication of the quality of the fit in both simple and multiple regression. Furthermore, in both simple and multiple regressions, $\text{Cor}(Y, \hat{Y})$ is related to another useful measure of the quality of fit of the linear model to the observed data. This measure is developed as follows. After we compute the least squares estimates of the parameters of a linear model, let us compute the following quantities:

$$\begin{aligned} \text{SST} &= \sum (y_i - \bar{y})^2, \\ \text{SSR} &= \sum (\hat{y}_i - \bar{y})^2, \\ \text{SSE} &= \sum (y_i - \hat{y}_i)^2, \end{aligned} \tag{2.44}$$

where SST stands for the total sum of squared deviations in Y from its mean $\bar{y}$, SSR denotes the sum of squares due to regression, and SSE represents the sum of squared residuals (errors). The quantities $(\hat{y}_i - \bar{y})$, $(\hat{y}_i - \bar{y})$, and $(y_i - \hat{y}_i)$ are depicted in Figure 2.7 for a typical point (x_i, y_i). The line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the fitted regression line based on all data points (not shown on the graph) and the horizontal line is drawn at $Y = \bar{y}$. Note that for every point (x_i, y_i), there are two points, $(x_i, \hat{y}_i)$, which lies on the fitted line, and $(x_i, \bar{y})$ which lies on the line $Y = \bar{y}$.

Figure 2.7 →

2.5 Measuring the Quality of Fit

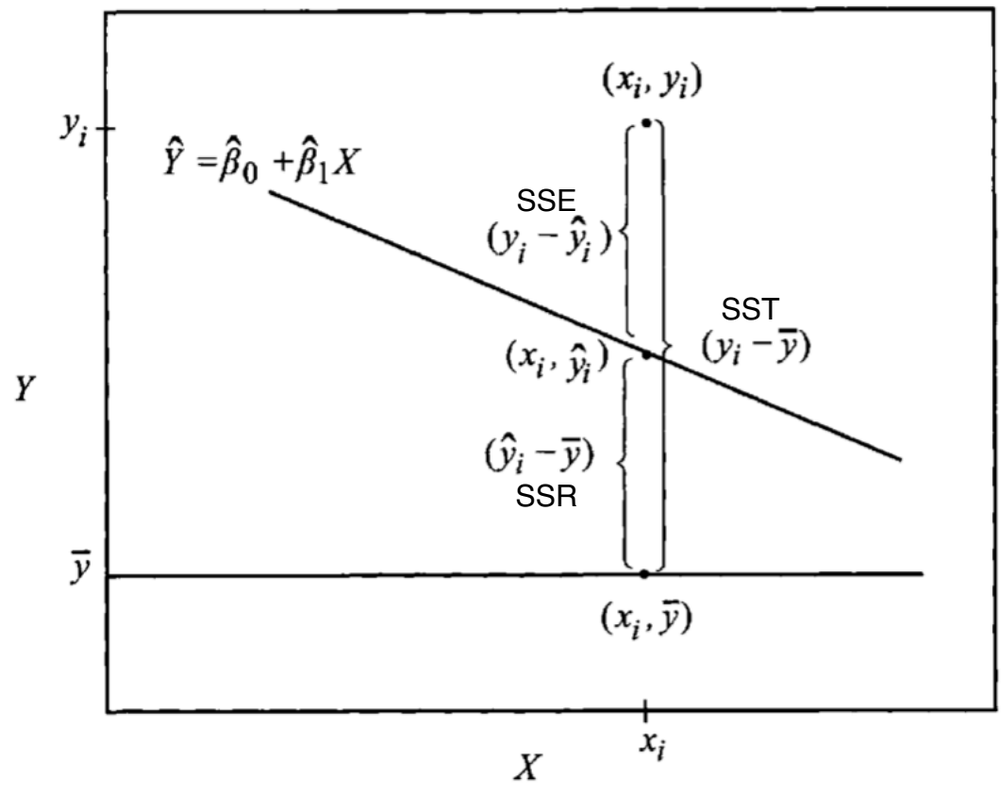


Figure 2.7 Graphical illustration of various quantities computed after fitting a regression line to data.

A fundamental equality, in both simple and multiple regressions, is given by

$$\text{SST} = \text{SSR} + \text{SSE}. \qquad (2.45)$$

Why ??

This equation arises from the description of an observation

y_i	=	$\hat{y}_i$	+	$(y_i - \hat{y}_i)$
Observed	=	Fit	+	Deviation from fit.

Subtracting from both sides, we obtain

$y_i - \bar{y}$	=	$(\hat{y}_i - \bar{y})$	+	$(y_i - \hat{y}_i)$
Deviation from mean	=	Deviation due to fit	+	Residual.

2.5 Measuring the Quality of Fit

Class Discussion

$\hat{\mathbf{Y}} = (\hat{y}_1, \cdots, \hat{y}_n)'$ is orthogonal to $\hat{\mathbf{Y}} - \mathbf{Y} = (\hat{y}_1 - y_1, \cdots, \hat{y}_n - y_n)'$

$$\text{So } \sum_{i=1}^{n} \hat{y}_i(\hat{y}_i - y_i) = 0$$

2.5 Measuring the Quality of Fit

Class Discussion

Show the equivalence between $\frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$ and $\frac{\text{Cor}(Y,X)\sqrt{n-2}}{\sqrt{1-[\text{Cor}(Y,X)]^2}}$, eq. (2.26) and (2.32)

2.5 Measuring the Quality of Fit

*Accordingly, the total sum of squared **deviations** in Y can be **decomposed** into the sum of two quantities, the first, **SSR**, measures the quality of X as a predictor of Y, and the second, **SSE**, measures the error in this prediction. Therefore, the **ratio** R^2 =SSR/SST can be interpreted as the **proportion** of the total variation in Y that is **accounted** for by the predictor variable X.*

We can rewrite R^2 as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}. \qquad (2.46)$$

Additionally, it can be shown that $[\text{Cor}(Y, X)]^2 = [\text{Cor}(Y, \hat{Y})]^2 = R^2.$ (2.47) **Why ??**

*In **simple linear regression**, R^2 is equal to the square of the **correlation coefficient** between the response variable Y and the predictor X or to the square of the **correlation coefficient** between the response variable Y and the **fitted** values $\hat{Y}$.*

*The **goodness-of-fit** index, R^2 , may be interpreted as the **proportion** of the total variability in the **response** variable Y that is **accounted** for by the predictor variable X. Note that $0 \leq R^2 \leq 1$ because $\text{SSE} \leq \text{SST}$. If R^2 is near 1, then X accounts for a large part of the variation in Y. For this reason, R^2 is known as the **coefficient of determination** because it gives us an idea of how the predictor variable X accounts for (**determines**) the response variable.*

2.5 Measuring the Quality of Fit

Example: Computer Repair Data

Using the Computer Repair data, the fitted values, and the residuals in Table 2.7, the reader can verify that $\text{Cor}(Y, X) = \text{Cor}(Y, \hat{Y}) = 0.994$, from which it follows that $R^2 = (0.994)^2 = 0.987$. The same value of R^2 can be computed as follows.. Verify that SST = 27768.348 and SSE = 348.848. So that

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{348.848}{27768.348} = 0.987.$$

The value $R^2 = 0.987$ indicates that nearly 99% of the total variability in the response variable (Minutes) is accounted for by the predictor variable (Units). The high value of R^2 indicates a strong linear relationship between servicing time and the number of units repaired during a service call.

2.5 Measuring the Quality of Fit

Core Assumption is Crucial

We reemphasize that the regression assumptions should be checked before drawing statistical conclusions from the analysis (e.g., conducting tests of hypothesis or constructing confidence or prediction intervals) because the validity of these statistical procedures hinges on the validity of the assumptions. Chapter 4 presents a collection of graphical displays that can be used for checking the validity of the assumptions. We have used these graphs for the computer repair data and found no evidence that the underlying assumptions of regression analysis are not in order. In summary, the 14 data points in the Computer Repair data have given us an informative view of the repair time problem. Within the range of observed data, we are confident of the validity of our inferences and predictions.

2.6. Regression Line Through the Origin

2.6 Regression Line Through the Origin

Special Case of Simple Linear Model

We have considered fitting the model

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad (2.48)$$

which is a regression line with an intercept. Sometimes, it may be necessary to fit the model

$$Y = \beta_1 X + \varepsilon, \qquad (2.49)$$

a line passing through the origin. This model is also called the *no-intercept* model. The line may be forced to go through the origin because of subject matter theory or other physical and material considerations. For example, distance traveled as a function of time should have no constant.

For **no-intercept model**, the **least squares estimate of** β_1 is $\hat{\beta}_1 = \dfrac{\sum y_i x_i}{\sum x_i^2}.$ (2.50)

The i-th fitted value is $\hat{y}_i = \hat{\beta}_1 x_i, \quad i = 1, 2, \cdots, n,$ (2.51)

and the corresponding residual is $e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n.$ (2.52)

*Note that the degrees of freedom for SSE is **n-1, not** n-2 ,as is the case for a model with an intercept.*

The standard error of the $\hat{\beta}_1$ is $\text{s.e.}(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{\sum x_i^2}},$ (2.53) where $\hat{\sigma} = \sqrt{\dfrac{\sum e_i^2}{n-1}} = \sqrt{\dfrac{\text{SSE}}{n-1}}.$ (2.54)

2.6 Regression Line Through the Origin

Special Case of Simple Linear Model

*Note that the residuals do **not** necessarily **add up to zero** as is the case for a model **with an intercept** Also, the fundamental identity **SST=SSR+SSE** is **no** longer true in general. For this reason, some quality measures for models with an intercept such as* R^2 *are **no longer appropriate** for models **with no intercept**. The **appropriate** identity for the case of models with **no intercept** is obtained by replacing* $\bar{y}$ *by zero. Hence, the **fundamental** identity becomes*

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2, \qquad (2.55)$$

from which R^2 is redefined as $$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}. \qquad (2.56)$$

This is the appropriate form of R^2 for models with no intercept.

*Note however, that the **interpretations** for the two formulas of* R^2 *are **different**. In the case of models **with an intercept**,* R^2 *can be interpreted as the proportion of the variation in Y that is accounted for by the predictor variable X after **adjusting** Y by its mean. For models **without an intercept**, no adjustment of Y is made.*

The formula used for testing hypothesis for and constructing confidence interval for β_1 still holds but with the new formula for $\hat{\beta}_1$ and $s.e.(\hat{\beta}_1)$

2.7. An Example Using R

2.7 An Example Using R

Let's consider a simple example of how the speed of a car affects its stopping distance, that is, how far it travels before it comes to a stop. To examine this relationship, we will use the `cars` *dataset which, I s a default* `R` *dataset. Thus, we don't need to load a package first; it is immediately available.*

To get a first look at the data you can use the `View()` function inside RStudio.

```
View(cars)
```

We could also take a look at the variable names, the dimension of the data frame, and some sample observations with `str()`.

```
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

2.7 An Example Using R

There are a number of additional functions to access some of this information directly.

```
dim(cars)
```

```
## [1] 50  2
```

```
nrow(cars)
```

```
## [1] 50
```

```
ncol(cars)
```

```
## [1] 2
```

Other than the two variable names and the number of observations, this data is still just a bunch of numbers, so we should probably obtain some context.

```
?cars
```

2.7 An Example Using R

Reading the documentation we learn that this is data gathered during the 1920s about the speed of cars and the resulting distance it takes for the car to come to a stop. The interesting task here is to determine how far a car travels before stopping, when traveling at a certain speed. So, we will first plot the stopping distance against the speed.

```r
plot(dist ~ speed, data = cars,
     xlab = "Speed (in Miles Per Hour)",
     ylab = "Stopping Distance (in Feet)",
     main = "Stopping Distance vs Speed",
     pch  = 20,
     cex  = 2,
     col  = "grey")
```

2.7 An Example Using R

To keep some notation consistent with above mathematics, we will store the response variable as ***y*** *and the predictor variable as* ***x***

```
x = cars$speed
y = cars$dist
```

Least Squares Estimate by Direct Computation

```
Sxy = sum((x - mean(x)) * (y - mean(y)))
Sxx = sum((x - mean(x)) ^ 2)
Syy = sum((y - mean(y)) ^ 2)
c(Sxy, Sxx, Syy)
```

```
## [1]  5387.40  1370.00 32538.98
```

Then finally calculate $\hat{\beta}_0$ and $\hat{\beta}_1$.

```
beta_1_hat = Sxy / Sxx
beta_0_hat = mean(y) - beta_1_hat * mean(x)
c(beta_0_hat, beta_1_hat)
```

```
## [1] -17.579095   3.932409
```

2.7 An Example Using R

Prediction by Direct Computation

We can now write the **fitted** or estimated line,

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

In this case,

$$\hat{y} = -17.58 + 3.93x.$$

We can now use this line to make predictions. First, let's see the possible x values in the `cars` dataset. Since some x values may appear more than once, we use the `unique()` to return each unique value only once.

```
unique(cars$speed)
```

```
##  [1]  4  7  8  9 10 11 12 13 14 15 16 17 18 19 20 22 23 24 25
```

Let's make a prediction for the stopping distance of a car traveling at 8 miles per hour.

$$\hat{y} = -17.58 + 3.93 \times 8$$

```
beta_0_hat + beta_1_hat * 8
```

```
## [1] 13.88018
```

This tells us that the estimated mean stopping distance of a car traveling at 8 miles per hour is 13.88.

2.7 An Example Using R
Use lm function

So far we have done regression by deriving the least squares estimates, then writing simple `R` commands to perform the necessary calculations. Since this is such a common task, this is functionality that is built directly into `R` via the `lm()` command.

*The `lm()` command is used to fit **linear models** which actually account for a broader class of models than simple linear regression, but we will use SLR as our first demonstration of `lm()`. The `lm()` function will be one of our most commonly used tools, so you may want to take a look at the documentation by using `?lm`. You'll notice there is a lot of information there, but we will start with just the very basics. This is documentation you will want to return to often.*

We'll continue using the `cars` data, and essentially use the `lm()` function to check the work we had previously done

```
stop_dist_model = lm(dist ~ speed, data = cars)
```

This line of code fits our very first linear model. The syntax should look somewhat familiar. We use the `dist ~ speed` syntax to tell `R` we would like to model the response variable `dist` as a linear function of the predictor variable `speed`. In general, you should think of the syntax as `response ~ predictor`. The `data = cars` argument then tells `R` that that `dist` and `speed` variables are from the dataset `cars`. We then store this result in a variable `stop_dist_model`.

2.7 An Example Using R
Use lm function

The variable `stop_dist_model` now contains a wealth of information, and we will now see how to extract and use that information. The first thing we will do is simply output whatever is stored immediately in the variable `stop_dist_model`.

```
stop_dist_model
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

We see that it first tells us the formula we input into `R`, that is `lm(formula = dist ~ speed, data = cars)`. We also see the coefficients of the model. We can check that these are what we had calculated previously. (Minus some rounding that `R` is doing when displaying the results. They are stored with full precision.)

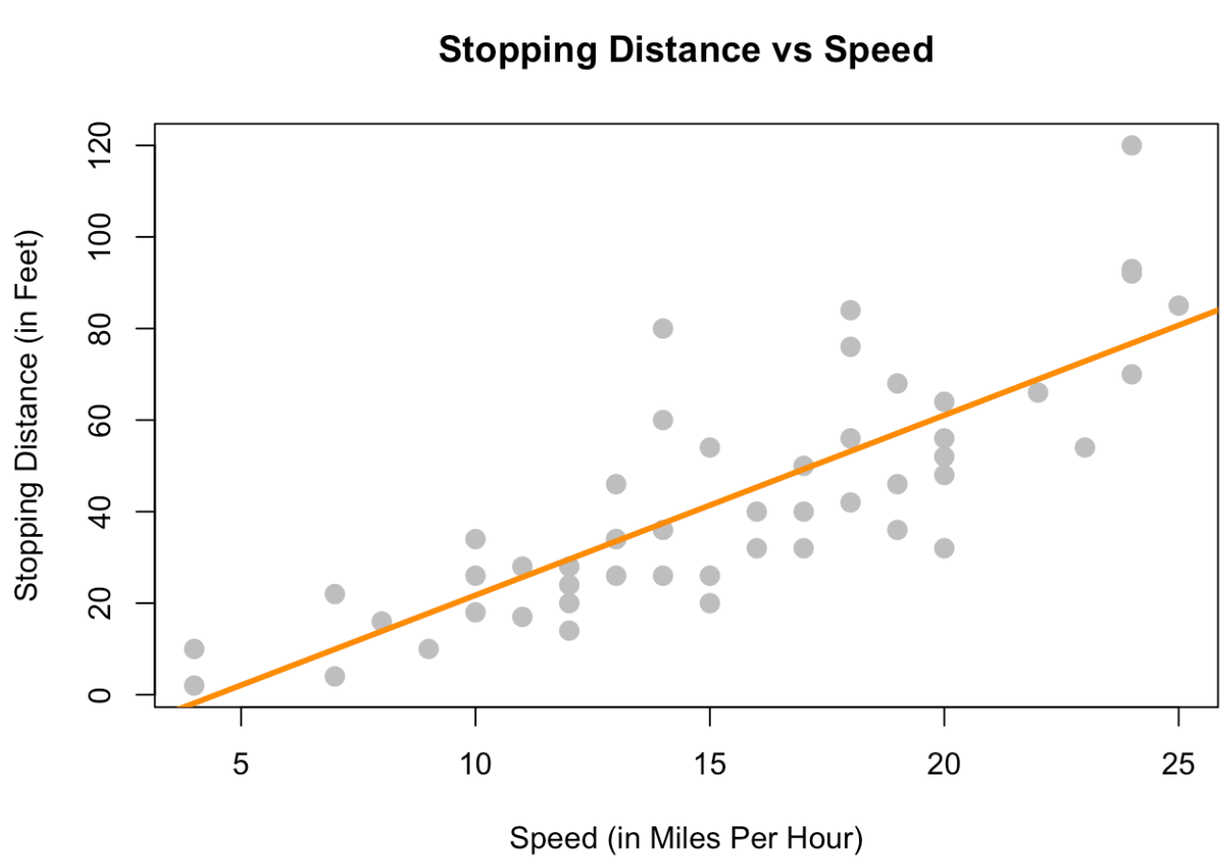
```
c(beta_0_hat, beta_1_hat)
```

```
## [1] -17.579095   3.932409
```

2.7 An Example Using R
Use lm function

Next, it would be nice to add the fitted line to the scatterplot. To do so we will use the `abline()` function.

```
plot(dist ~ speed, data = cars,
     xlab = "Speed (in Miles Per Hour)",
     ylab = "Stopping Distance (in Feet)",
     main = "Stopping Distance vs Speed",
     pch  = 20,
     cex  = 2,
     col  = "grey")
abline(stop_dist_model, lwd = 3, col = "darkorange")
```



Use lm function

2.7 An Example Using R

The `abline()` function is used to add lines of the form $a + bx$ to a plot. (Hence `ab line`.) When we give it `stop_dist_model` as an argument, it automatically extracts the regression coefficient estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$) and uses them as the slope and intercept of the line. Here we also use `lwd` to modify the width of the line, as well as `col` to modify the color of the line.

The "thing" that is returned by the `lm()` function is actually an object of class `lm` which is a list. The exact details of this are unimportant unless you are seriously interested in the inner-workings of `R`, but know that we can determine the names of the elements of the list using the `names()` command.

```
names(stop_dist_model)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

We can then use this information to, for example, access the residuals using the `$` operator.

```
stop_dist_model$residuals
```

```
##          1          2          3          4          5          6          7
##   3.849460  11.849460  -5.947766  12.052234   2.119825  -7.812584  -3.744993
##          8          9         10         11         12         13         14
##   4.255007  12.255007  -8.677401   2.322599 -15.609810  -9.609810  -5.609810
##         15         16         17         18         19         20         21
##  -1.609810  -7.542219   0.457781   0.457781  12.457781 -11.474628  -1.474628
##         22         23         24         25         26         27         28
##  22.525372  42.525372 -21.407036 -15.407036  12.592964 -13.339445  -5.339445
##         29         30         31         32         33         34         35
## -17.271854  -9.271854   0.728146 -11.204263   2.795737  22.795737  30.795737
##         36         37         38         39         40         41         42
## -21.136672 -11.136672  10.863328 -29.069080 -13.069080  -9.069080  -5.069080
```

Use lm function

2.7 An Example Using R

Another way to access stored information in `stop_dist_model` are the `coef()`, `resid()`, and `fitted()` functions. These return the coefficients, residuals, and fitted values, respectively.

```
coef(stop_dist_model)
```

```
## (Intercept)       speed
##  -17.579095    3.932409
```

```
fitted(stop_dist_model)
```

```
##         1         2         3         4         5         6         7         8
## -1.849460 -1.849460  9.947766  9.947766 13.880175 17.812584 21.744993 21.744993
##         9        10        11        12        13        14        15        16
## 21.744993 25.677401 25.677401 29.609810 29.609810 29.609810 29.609810 33.542219
##        17        18        19        20        21        22        23        24
## 33.542219 33.542219 33.542219 37.474628 37.474628 37.474628 37.474628 41.407036
##        25        26        27        28        29        30        31        32
## 41.407036 41.407036 45.339445 45.339445 49.271854 49.271854 49.271854 53.204263
##        33        34        35        36        37        38        39        40
## 53.204263 53.204263 53.204263 57.136672 57.136672 57.136672 61.069080 61.069080
##        41        42        43        44        45        46        47        48
## 61.069080 61.069080 61.069080 68.933898 72.866307 76.798715 76.798715 76.798715
##        49        50
## 76.798715 80.731124
```

Use lm function

2.7 An Example Using R

An `R` function that is useful in many situations is `summary()`. We see that when it is called on our model, it returns a good deal of information. By the end of the course, you will know what every value here is used for. For now, you should immediately notice the coefficient estimates, and you may recognize the R^2 value we saw earlier.

```
summary(stop_dist_model)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Use lm function 2.7 An Example Using R

The `summary()` command also returns a list, and we can again use `names()` to learn what about the elements of this list.

```
names(summary(stop_dist_model))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

So, for example, if we wanted to directly access the value of R^2, instead of copy and pasting it out of the printed statement from `summary()`, we could do so.

```
summary(stop_dist_model)$r.squared
```

```
## [1] 0.6510794
```

Another value we may want to access is s_e, which `R` calls `sigma`.

```
summary(stop_dist_model)$sigma
```

```
## [1] 15.37959
```

Use lm function

2.7 An Example Using R

Another useful function, which we will use almost as often as `lm()` is the `predict()` function.

```
predict(stop_dist_model, newdata = data.frame(speed = 8))
```

```
##        1
## 13.88018
```

The above code reads "predict the stopping distance of a car traveling 8 miles per hour using the `stop_dist_model`." Importantly, the second argument to `predict()` is a data frame that we make in place. We do this so that we can specify that `8` is a value of `speed`, so that predict knows how to use it with the model stored in `stop_dist_model`. We see that this result is what we had calculated "by hand" previously.

We could also predict multiple values at once.

```
predict(stop_dist_model, newdata = data.frame(speed = c(8, 21, 50)))
```

```
##         1         2         3
##  13.88018  65.00149 179.04134
```

$$\hat{y} = -17.58 + 3.93 \times 8 = 13.88$$
$$\hat{y} = -17.58 + 3.93 \times 21 = 65$$
$$\hat{y} = -17.58 + 3.93 \times 50 = 179.04$$

Use lm function

2.7 An Example Using R

We will now discuss the results displayed called `Coefficients`. First recall that we can extract this information directly.

```
names(summary(stop_dist_model))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

```
summary(stop_dist_model)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -17.579095  6.7584402 -2.601058 1.231882e-02
## speed         3.932409  0.4155128  9.463990 1.489836e-12
```

The `names()` function tells us what information is available, and then we use the `$` operator and `coefficients` to extract the information we are interested in. Two values here should be immediately familiar.

$$\hat{\beta}_0 = -17.5790949$$

and

$$\hat{\beta}_1 = 3.9324088$$

which are our estimates for the model parameters β_0 and β_1.

Use lm function

2.7 An Example Using R

Let's now focus on the second row of output, which is relevant to β_1.

```
summary(stop_dist_model)$coefficients[2,]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 3.932409e+00 4.155128e-01 9.463990e+00 1.489836e-12
```

Again, the first value, `Estimate` is

$$\hat{\beta}_1 = 3.9324088.$$

The second value, `Std. Error` , is the standard error of $\hat{\beta}_1$,

$$\text{SE}[\hat{\beta}_1] = \frac{s_e}{\sqrt{S_{xx}}} = 0.4155128.$$

The third value, `t value` , is the value of the test statistic for testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$,

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}[\hat{\beta}_1]} = \frac{\hat{\beta}_1 - 0}{s_e/\sqrt{S_{xx}}} = 9.46399.$$

Lastly, `Pr(>|t|)` , gives us the p-value of that test.

$$\text{p-value} = 1.4898365 \times 10^{-12}$$

Note here, we are specifically testing whether or not $\beta_1 = 0$.

The first row of output reports the same values, but for β_0.

```
summary(stop_dist_model)$coefficients[1,]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## -17.57909489   6.75844017  -2.60105800   0.01231882
```

2.7 An Example Using R

Confidence interval for coefficient parameter

Using `R` we can very easily obtain the confidence intervals for β_0 and β_1.

```
confint(stop_dist_model, level = 0.99)
```

```
##                 0.5 %    99.5 %
## (Intercept) -35.706610 0.5484205
## speed         2.817919 5.0468988
```

This automatically calculates 99% confidence intervals for both β_0 and β_1, the first row for β_0, the second row for β_1.

For the `cars` example when interpreting these intervals, we say, we are 99% confident that for an increase in speed of 1 mile per hour, the average increase in stopping distance is between 2.8179187 and 5.0468988 feet, which is the interval for β_1.

2.7 An Example Using R

Confidence interval for mean response

To find confidence intervals for the mean response using `R`, we use the `predict()` function. We give the function our fitted model as well as new data, stored as a data frame. (This is important, so that `R` knows the name of the predictor variable.) Here, we are finding the confidence interval for the mean stopping distance when a car is travelling 5 miles per hour and when a car is travelling 21 miles per hour.

```
new_speeds = data.frame(speed = c(5, 21))
predict(stop_dist_model, newdata = new_speeds,
        interval = c("confidence"), level = 0.99)
```

```
##         fit       lwr      upr
## 1  2.082949 -10.89309 15.05898
## 2 65.001489  56.45836 73.54462
```

2.7 An Example Using R

Prediction interval for New Observation

To calculate this for a set of points in `R` notice there is only a minor change in syntax from finding a confidence interval for the mean response.

```
predict(stop_dist_model, newdata = new_speeds,
        interval = c("prediction"), level = 0.99)
```

```
##         fit       lwr       upr
## 1  2.082949 -41.16099  45.32689
## 2 65.001489  22.87494 107.12803
```

Also notice that these two intervals are wider than the corresponding confidence intervals for the mean response.

2.7 An Example Using R

Confidence and Prediction Bands

Often we will like to plot both confidence intervals for the mean response and prediction intervals for all possible values of x. We calls these confidence and prediction bands.

```
speed_grid = seq(min(cars$speed), max(cars$speed), by = 0.01)
dist_ci_band = predict(stop_dist_model,
                       newdata = data.frame(speed = speed_grid),
                       interval = "confidence", level = 0.99)
dist_pi_band = predict(stop_dist_model,
                       newdata = data.frame(speed = speed_grid),
                       interval = "prediction", level = 0.99)
```

```
plot(dist ~ speed, data = cars,
     xlab = "Speed (in Miles Per Hour)",
     ylab = "Stopping Distance (in Feet)",
     main = "Stopping Distance vs Speed",
     pch  = 20,
     cex  = 2,
     col  = "grey",
     ylim = c(min(dist_pi_band), max(dist_pi_band)))
abline(stop_dist_model, lwd = 5, col = "darkorange")

lines(speed_grid, dist_ci_band[,"lwr"], col = "dodgerblue", lwd = 3, lty = 2)
lines(speed_grid, dist_ci_band[,"upr"], col = "dodgerblue", lwd = 3, lty = 2)
lines(speed_grid, dist_pi_band[,"lwr"], col = "dodgerblue", lwd = 3, lty = 3)
lines(speed_grid, dist_pi_band[,"upr"], col = "dodgerblue", lwd = 3, lty = 3)
points(mean(cars$speed), mean(cars$dist), pch = "+", cex = 3)
```

2.7 An Example Using R

Confidence and Prediction Bands