Chapter 3 Correlation and Simple Linear Regression

Chapter 3

Never make predictions, especially about the future.
– Casey Stengel

Cross Sectional Data

From the company files the forecast analyst collects a random sample of data from 10 stores from the 172. In the language of statistics, the forecast analyst has chosen a sample of $n = 10$ from a population of $N = 172$ stores. Since we are considering these two sets of data jointly, we term this set of cross-sectional data *bivariate data.*

Table 3.1

10 Randomly Chosen Stores
Sales Volume and the Corresponding Advertising Expenditures
in 1,000 Units

Store	Sales	Advertising
1	162.5	3.0
2	188.0	4.5
3	240.0	7.0
4	385.5	11.0
5	140.5	1.5
6	202.0	5.0
7	315.0	7.0
8	385.5	9.5
9	260.5	6.0
10	265.0	7.0

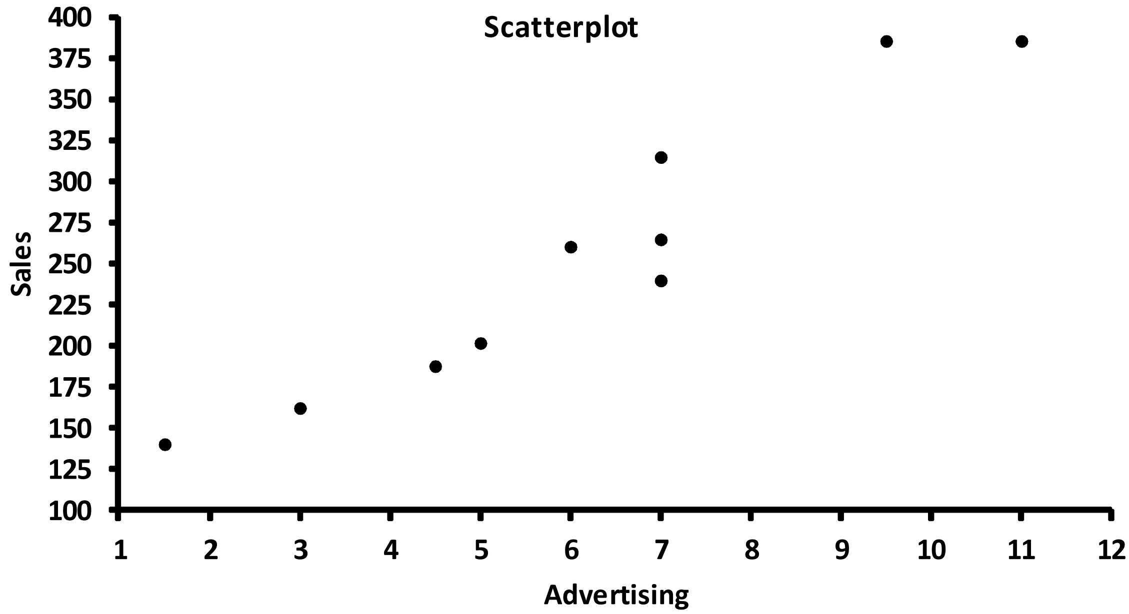
These data pairs are often referred to as “observations." We have a sample of 10 observations.

Scatter diagrams

The importance of graphing data cannot be over-emphasized. Often just a rough visual display of data is extraordinarily revealing. Plot the data! Graph the data!

We plot the Advertising values along the horizontal axis and the Sales numbers along the vertical axis. The horizontal axis is usually denoted the *X-axis,* and the numbers plotted along the X-axis are the *X-values* or *X variable.* Similarly, the vertical axis is usually denoted the *Y-axis,* and so the numbers plotted along the Y-axis are the *Y-values* or *Y variable.*

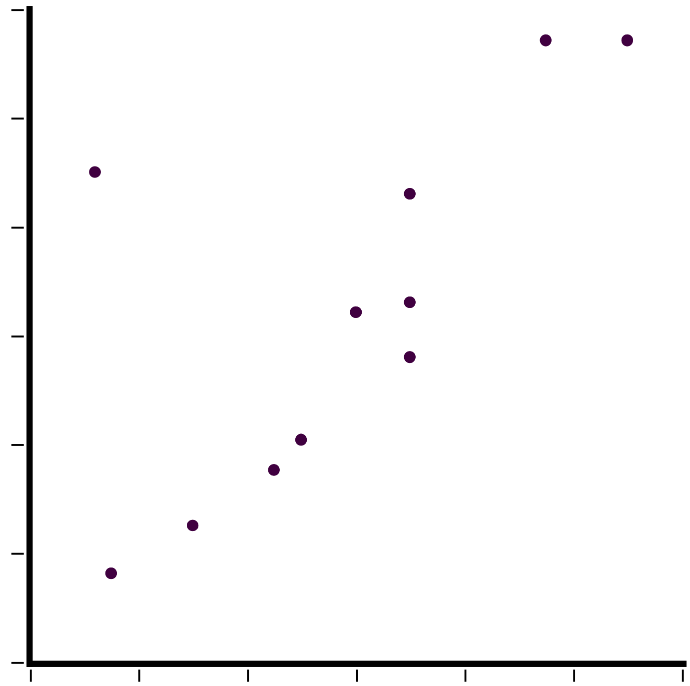
We denote with a subscript of *i* the i^{th} observation of each variable. Y_i and X_i are read "Y sub i" and "X sub i".

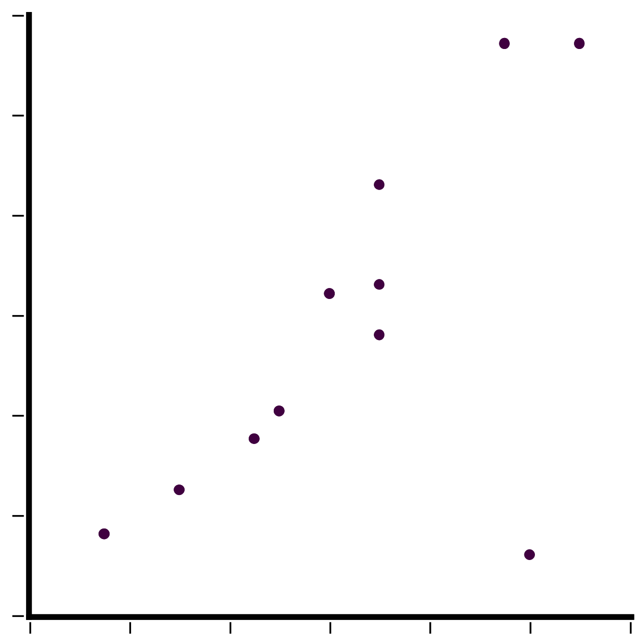


The scatter diagram quickly reveals that a relationship appears to exist between Sales Volume and Advertising Expenditures. We shall term this a *positive relationship* because when we observe an increase in Advertising, we observe, most often, a corresponding increase in Sales.

Outliers

The scatter diagram will also quickly reveal *outliers* or unusual points. By an outlier we mean a scatter diagram point that is not following the apparent pattern. Figures 3-2, 3-3 illustrate two examples of outlier points. At the very least, the forecast analyst should check for data entry or recording errors that may have caused the unusual numbers. If indeed, the points are not clerical errors, but are true outliers, there are methods for dealing with them. We shall discuss later how we deal with outliers, but for now, at this first data stage, our primary concern is to check for possible outliers.





Simple Statistics of the Data

We compute the sample means, variations, variances, and standard deviation of both the *X* and *Y* data sets.

Sample Mean of *X*

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \textbf{3.1}$$

Sample Variance of *X*

$$s_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n\text{-}1} \quad \textbf{3.2}$$

The terms $(X_i - \bar{X})^2$ are the squared deviations from the mean and their sum, $\sum(X_i - \bar{X})^2$, is the *Sum of the Squared Deviations of X*, the *SSX*, also called the *Variation of X*. Using the data from Table 3.1 we calculate the means, variations, and variances.

Table 3.2

	Mean	Variation	Variance	Standard Deviation
Advertising X	$\bar{X} = 6.15$	$SSX = 72.5250$	$s_X^2 = 8.06$	$s_X = \sqrt{8.06} = 2.84$
Sales Y	$\bar{Y} = 254.45$	$SSY = 66{,}997.2250$	$s_Y^2 = 7{,}441.91$	$s_Y = \sqrt{7{,}441.91} = 86.27$

The Covariance and Correlation between *X* and *Y*

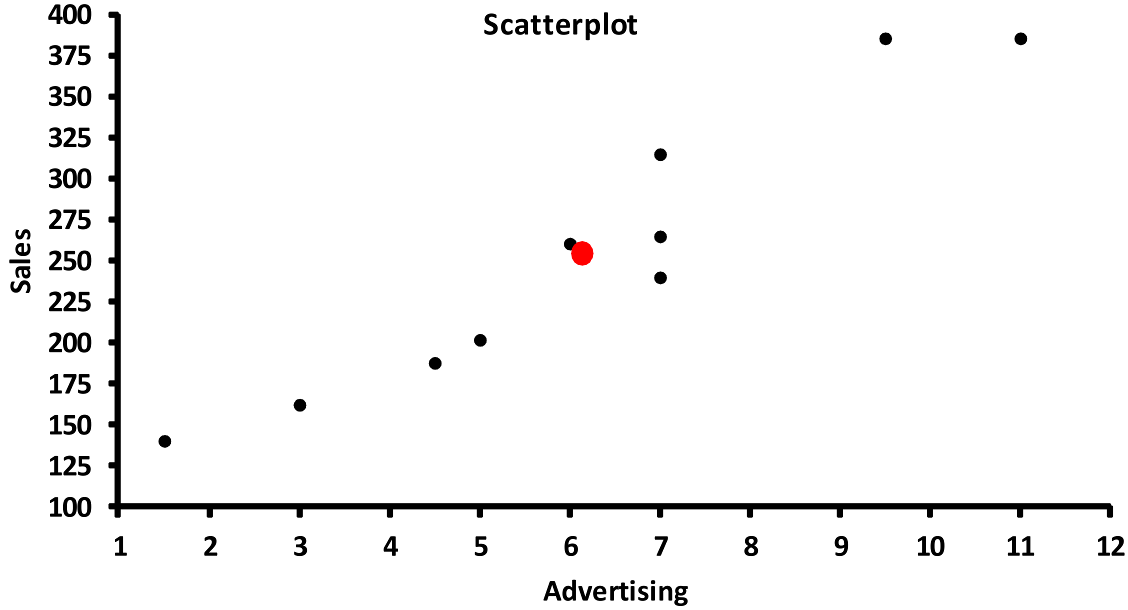
While the variance of *X* is a measure of the dispersion of the *X* data, and the variance of *Y* is a measure of the dispersion of the *Y* data, it is the *Covariance between X and Y* that is a measure of the way in which the two sets of the data move together.

Returning to the scatter diagram, we locate the pair of means

$$(\bar{X}, \bar{Y}) = (6.15, 254.45)$$

in the “center" of the data.

Now we consider the point $(9.5, 385.5)$, the Location 8 data values, which are up and to the right of the point of means.



Because the Location 8 data values are up and to the right of the point of means, their respective distances from the point of means (their deviations from the mean) are positive.

$$X_8 - \overline{X} = 9.50 - 6.15 \qquad = +3.35$$

and

$$Y_8 - \overline{Y} = 385.50 - 254.45 \qquad = +131.05$$

So the product of the deviations

$$(X_8 - \overline{X})(Y_8 - \overline{Y}) = (3.35)(131.05)$$

$$= 439.02$$

is again a positive number.

For the pair $(3.0, 162.5)$ of Location 1, which are down and to the left of the point of means their respective distances (deviations from the mean) are negative.

$$X_1 - \overline{X} = 3.00 - 6.15 \qquad = -3.15$$

and

$$Y_1 - \overline{Y} = 162.50 - 254.45 \qquad = -91.95$$

So the product of the negative deviations

$$(X_1 - \overline{X})(Y_1 - \overline{Y}) = (-3.15)(-91.95)$$

$$= 289.64$$

is again a positive number.

When data create a scatter diagram that tends upwards to the right then it will be generally the case that if an X-value is larger than $\overline{X}$, the corresponding Y-value will also be larger than $\overline{Y}$, so that the product of $(X_i - \overline{X})(Y_i - \overline{Y})$ is again positive.

Similarly, if the data tends upward and to the right, and an X-value is less than $\overline{X}$, then the corresponding Y-value will usually be less than $\overline{Y}$, so that their product

$$(X_i - \overline{X})(Y_i - \overline{Y})$$

is again positive.

These products $(X_i - \overline{X})(Y_i - \overline{Y})$ are called the *cross products.*

Consequently, the sum $\sum(X_i - \overline{X})(Y_i - \overline{Y})$ of positive cross products again will be positive.

Dividing the sum by $n - 1$, results in the *Covariance*, denoted *COV(X,Y)*, or s_{XY}.

Covariance between X and Y

$$\boldsymbol{COV(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}} \qquad \mathbf{3.3}$$

Using the results in Table 3.3 we compute the covariance between X and Y.

Table 3.3 Determining the Covariance between X and Y using Formula 3.4

X_i	Y_i	Deviation from $\overline{X}$ $X_i - \overline{X},\ \overline{X} = 6.15$	Deviation from $\overline{Y}$ $Y_i - \overline{Y},\ \overline{Y} = 254.45$	Product of the Deviations $(X_i - \overline{X})(Y_i - \overline{Y})$	
3.0	162.5	-3.15	-91.95	(-3.15)(-91.95)	= +289.6425
4.5	188.0	-1.65	-74.45	(-1.65)(-74.45)	= +109.6425
7.0	240.0	+0.85	-14.45	(+0.85)(-14.45)	= -12.2825
11.0	385.5	+4.85	+131.05	(+4.85)(+131.05)	= +635.5925
1.5	140.5	-4.65	-113.95	(-4.65)(-113.95)	= +529.8675
5.0	202.0	-1.15	-52.45	(-1.15)(-52.45)	= +60.3175
7.0	315.0	+0.85	+60.55	(+0.85)(+60.55)	= +51.4675
9.5	385.5	+3.35	+131.05	(+3.35)(+131.05)	= +439.0175
6.0	260.5	-0.15	+6.05	(-0.15)(+6.05)	= -0.9075
7.0	265.0	+0.85	+10.55	(+0.85)(+10.55)	= +8.9675
					2,111.3250

Sum of the Cross Products

$$SSXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = 2,111.3250$$

Covariance

$$COV(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{2{,}111.3250}{9} = 234.59$$

The Correlation between *X* and *Y*

Since the magnitude of the covariance is determined in part by the magnitude of the numbers being used, we normalize the covariance by dividing by the standard deviation of *X* and the standard deviation of *Y*. The covariance divided by the standard deviations is defined as the *correlation between X and Y.*

The correlation between *X* and *Y*, denoted $\hat{\rho}_{XY}$, is defined as

Correlation between *X* and *Y*

$$\hat{\rho}_{XY} = \frac{s_{XY}}{s_X s_Y} \qquad \mathbf{3.5}$$

An algebraically equivalent formula is

Correlation between *X* and *Y*

$$\hat{\rho}_{XY} = \frac{SSXY}{\sqrt{SSX}\,\sqrt{SSY}} \qquad \mathbf{3.5a}$$

In the present example, using formula 3.5a, the correlation is

$$\hat{\rho}_{XY} = \frac{SSXY}{\sqrt{SSX}\,\sqrt{SSY}} = \frac{2{,}111.3250}{\sqrt{72.5250}\,\sqrt{66{,}977.2250}} = .958$$

Correlation is a good measure of the linear relationship between two variables *X* and *Y*. It can be shown that ρ_{XY} always takes on values between -1 and +1.

$$-1 \leq \rho_{XY} \leq +1$$

Perfect linear correlation is either +1 or -1. If the scatter diagram forms a pattern as in Figure 3-6a, then $\rho_{XY} = +1$. Or, if the scatter diagram forms a pattern as in Figure 3-6b, then $\rho_{XY} = -1$.

If the scatter diagram forms patterns as in Figure 3-6c, or 3-6d, then $\rho_{XY} = 0$.

$\rho_{XY} = 0$ in Figure 3-6c is reasonable since there appears no linear pattern of the data.

$\rho_{XY} = 0$ in Figure 3-6d is also reasonable since, while there is a clear pattern of the data in the scatter diagram, it is not a *linear* pattern.

Figure 3-6

Distinguishing between Correlation and Causation

We must distinguish between *statistical correlation* and *causation.* The preceding sections on covariance and correlation dealt with the issue of quantifying the linear relationship between two variables. We are not quantifying the causal relationship (the cause-and-effect relationship) between X and Y. Ultimately, we would like to determine a mathematical relationship between the two economic variables of Advertising and Sales, but remember correlation is not a measure of their causal relationship.

For additional practice and review we shall always include one or more solved problems and to illustrate the ideas and concepts within the chapter. The reader may skip these practice problems without loss of information if he/she does not desire additional practice.

Solved Problem 1

Let us consider another set of bivariate data which has been collected for us. We have a set of 10 observations of Y and X and will practice the 6 Stages of Forecasting with these data throughout this Chapter.

Determine the simple graphical and numerical statistics of the following set of data of $n = 10$ observations.

Table 3.4

Observation	Y	X
1	17	2
2	8	7
3	12	5
4	5	10
5	6	9
6	9	9
7	16	1
8	17	2
9	13	7
10	17	3

Steps in the Solution

1 Construct a table of data.

2 Construct a scatter diagram of the data
Check for Outliers

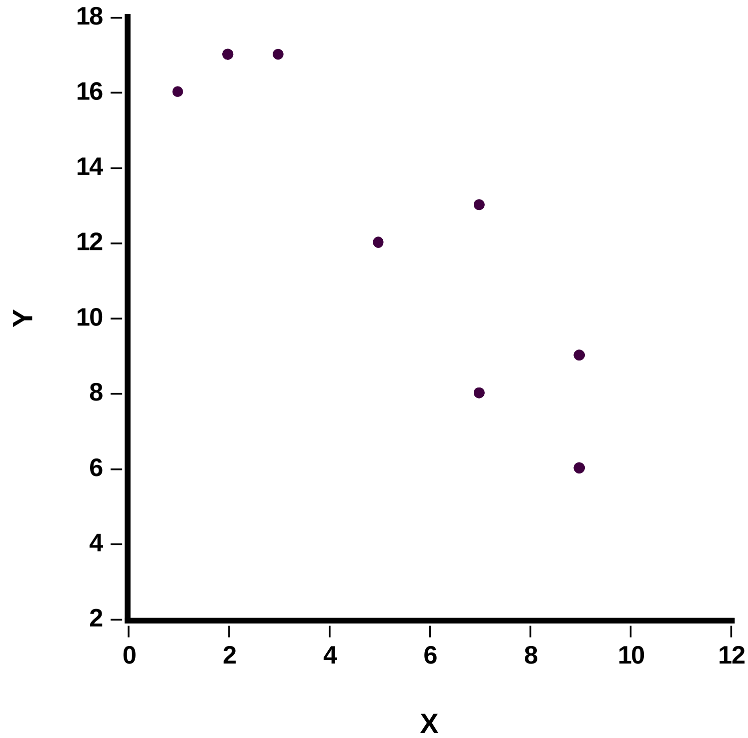
3 Calculate the Simple Statistics of the Data
Calculate the Covariance and Correlation between X and Y

1 Construct a table of the data

This has been done for us.

2 Construct the Scatter Diagram of the Data

We construct a scatter diagram for a quick view and visual check of the data.



We observe the scatter diagram describes a *negative relationship* between X and Y, as X increases, Y decreases.

Check for Outliers

There do not appear to be any outliers.

3 Calculate the Simple Statistics of the Data

Mean of X | Variance of X

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{55}{10} = 5.5 \qquad s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{10}(X_i - 5.5)^2}{9} = \frac{100.5}{9} = 11.2$$

Mean of Y | Variance of Y

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{120}{10} = 12.0 \qquad s_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{\sum_{i=1}^{10}(Y_i - 12.0)^2}{9} = \frac{202.0}{9} = 22.4$$

Create a table of summary information

Table 3.5

	Mean	Variation	Variance	Standard Deviation
X	$\bar{X} = 5.5$	$SSX = 100.5$	$s_X^2 = 11.2$	$s_X = \sqrt{11.2} = 3.35$
Y	$\bar{Y} = 12.0$	$SSY = 202.0$	$s_Y^2 = 22.4$	$s_Y = \sqrt{22.4} = 4.73$

Calculate the Covariance and Correlation between *X* and *Y*

Covariance between X and Y

$$COV(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{10}(X_i - 5.5)(Y_i - 12.0)}{9} = \frac{-133.0}{9} = -14.8$$

$$COV(X,Y) = \text{-14.8}$$

Note that $SSXY = -133.0$, the sum of the Cross Products

Correlation between X and Y

$$\hat{\rho}_{XY} = \frac{COV(X, Y)}{s_X s_Y} = \frac{-14.8}{(3.35)(4.73)} = -.933$$

$$\hat{\rho}_{XY} = -.933$$

With correlation at $-.933$ this suggests a fairly linear, negative relationship between X and Y.

Stage 2 Identification of the Basic Model

Dependent and Independent Variables

If we do believe that there exists a relationship between two economic variables, then one of the variables, like Sales Volume, is “dependent" on the other variable, Advertising.

Sales, Y_i, is the *dependent variable*, since we believe that Sales Volume at a particular Location depends on the amount of dollars spent on Advertising at that Location. Advertising, X_i, is the *independent variable* in that we may chose any amount of Advertising dollars for a Location's budget. This also referred to was “one-way causality" in that we believe that Advertising has a direct impact on Sales whereas Sales does not have an impact on Advertising.

Symbolically, $Y <$ X
Dependent Independent

Other terms used for dependent and independent variables are:

$Y <$	X
Explained $<$	Explanatory
Predicted $<$	Predictor
Regressed $<$	Regressor

The Linear Model

We now wish to determine a mathematical model linking the dependent and independent variables. Since a good approximation for this scatter diagram of data is a ***straight line*** through the data, a model using a straight line is the natural choice.

The equation of a straight line is often written

Equation of a Straight Line

$$\boldsymbol{Y = mX + b} \qquad \textbf{3.6}$$

where *m* denotes the *slope* or *steepness* of the line and *b* denotes the *y-intercept.*

Figure 3-8

In econometrics and forecasting, equation (3.6) is usually written in a slightly different form.

In the equation

$$Y = mX + b \qquad 3.6$$

the *mX* and *b* are transposed,

$$Y = b + mX \qquad 3.6a$$

and different letters are used.

Deterministic Equation

$$\boldsymbol{Y = \beta_0 + \beta_1 X} \qquad \textbf{3.7}$$

The Greek letters β_0 and β_1 are read

β_0 “beta-zero," “beta sub-zero," or “beta-nought."

β_1 “beta-one," or “beta sub-one."

It is termed a “Deterministic Equation" because for a chosen X value, we can determine exactly what the corresponding Y value is.

However, the data (X_i, Y_i) in the scatter diagram do not form a perfectly straight line, and the difference between the points on the straight line and the actual data points is the *random error term* or the *disturbance term,* which we denote with the Greek letter epsilon, ϵ_i. ϵ_i is pronounced “epsilon sub i" or just “epsilon i."

Stochastic Equation

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i} \qquad \textbf{3.8}$$

It is termed a “Stochastic Equation" because we cannot, for a chosen *X* value, determine exactly the corresponding *Y* value. There is an uncertain, random, or stochastic portion of the equation, and that is the ϵ_i.

Hence, we distinguish between the *actual value of Y at observation i,* Y_i, and the straight line value or the *expected value of* Y_i *at observation i*, given a particular value X_i at observation *i.*

The “expected value of Y_i, given X_i," we denote by $E(Y_i|\, X_i)$. It is beyond the scope of this textbook to develop a statistically rigorous definition of the “expected value of Y_i, given X_i" so we shall appeal to the reader's intuition for this concept. The expected value suggests that this is what we expect the value of *Y* to be, given a particular *X* in the regression equation. Or in other words, the value of *Y* is conditional on the regression equation and the chosen value of *X.*

The Expected Value of Y_i, given X_i

$$\boldsymbol{E(Y_i\,|X_i) = \beta_0 + \beta_1 X_i} \qquad \textbf{3.9}$$

By comparing equation 3.8 with equation 3.9 we see that the only distinction between the actual value Y_i, given X_i, and the expected value $E(Y_i|X_i)$ is the random error term, ϵ_i.

Figure 3-9

Equation 3.8 is called a *stochastic equation* in that the actual value Y_i is not necessarily limited to one value for one specific value of X_i, there is variability. It is also called a *probabilistic model* because there is a probability distribution associated with Y_i for each X_i.

As an illustration, in the set of data in Table 3.1 there are three instances when $X_i = 7$. When $i = 3$, $i = 7$, and $i = 10$. That is, there were three Locations when Advertising Expenditure averaged $7,000.

$X_3 = 7$
$X_7 = 7$
$X_{10} = 7$

Yet the corresponding Sales Volumes were not the same.

$X_3 = 7 \qquad Y_3 = 240$
$X_7 = 7 \qquad Y_7 = 315$
$X_{10} = 7 \qquad Y_{10} = 265$

Figure 3-10 below shows the three actual values of Y_i when X_i is 7 (the stochastic equation) and the Fitted Value, $\hat{Y}_i = 279.2$, (the value of the deterministic equation, which we shall derive shortly). The Fitted Value is an estimate of $E(Y_i|X_i = 7)$.

Figure 3-10

As mentioned before, the difference between the Actual Value and the Fitted Value is the error term, or fitted residual $\hat{e}_i$. Table 3.6 lists the Actuals and the Fitted Values, and the corresponding error terms

Table 3.6

Independent Value X_i	Dependent Actual Value Y_i	Dependent Fitted Value $\hat{Y}_i$	Error Term Actual − Fitted $\hat{e}_i = Y_i - \hat{Y}_i$
$X_3 = 7$	$Y_3 = 240$	$\hat{Y}_3 = 279.2$	$\hat{e}_3 = Y_3 - \hat{Y}_3 = -39.2$
$X_7 = 7$	$Y_7 = 315$	$\hat{Y}_7 = 279.2$	$\hat{e}_7 = Y_7 - \hat{Y}_7 = +35.8$
$X_{10} = 7$	$Y_{10} = 265$	$\hat{Y}_{10} = 279.2$	$\hat{e}_{10} = Y_{10} - \hat{Y}_{10} = -14.2$

The Method of Least Squares

The Theory

We have discussed only in general terms the straight line as being the "best fit" of the data. We now wish to make the meaning of "best fit" precise. As shown in Table 3.6, for each expected value there will be a corresponding error value.

If we wished to determine that line such that the simple sum of the errors is as small as possible, i.e. zero, there would not be a unique line. It can be shown that for any set of data points in a scatter diagram there are many lines (actually an infinite number) that will cause the simple sum of the errors to equal zero.

However, if we wish to ***minimize the Sum of the Squared Errors (the SSE)***, then it can be proven that there is ***only one such line*** that will minimize the sum of the squared errors. For historical reasons this line has the title as the "regression line."

Thus, the goal of estimation of the parameters of the linear model is to estimate the unknown parameters, β_0 and β_1, so that the regression line will minimize the sum of the squared errors.

We write the Sum of the Squared Errors as

$$SSE = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i \hat{\epsilon}_i^2 \qquad 3.10$$

Minimizing *SSE* is completely determined by the values of β_0 and β_1. The estimates of β_0 and β_1 will determine the minimum *SSE*, or, as it is often termed the "least squares", so that the estimates of β_0 and β_1 are usually called the "ordinary least squares" estimates. "Ordinary least squares" estimates is abbreviated as "OLS" estimates.

The OLS estimates of β_0 and β_1 are completely determined by the set of data, (X_i, Y_i). The OLS parameter estimates are given by the formulas below:

The Formulas

Parameter estimate of β_1

$$\hat{\beta}_1 = \frac{COV(X,Y)}{s_X^2} \qquad \mathbf{3.11}$$

Another algebraically equivalent formula for $\hat{\beta}_1$ which is sometimes easier to use is:

Parameter estimate of β_1

$$\hat{\beta}_1 = \frac{SSXY}{SSX} \qquad \mathbf{3.11a}$$

Parameter estimate of β_0

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \mathbf{3.12}$$

We have changed the notation slightly at this point. We are using the notations, $\hat{\beta}_1$ and $\hat{\beta}_0$. The carat over each Greek letter is read “hat" as in “beta-one hat," and “beta-nought hat." The distinction is that since we are dealing with sample data and not the whole population of data, we are constructing sample estimates of the true (but unknown) population parameters β_1 and β_0. $\hat{\beta}_1$ is the OLS estimate of β_1, and $\hat{\beta}_0$ is the OLS estimate of β_0.

In our chapter example, we shall use formulas (3.11a) and (3.12). These formulas require $SSXY = 2{,}111.325$ (page), $SSX = 72.5250$ (page), $\overline{Y} = 254.45$ and $\overline{X} = 6.15$ (page).

Thus, we have

$$\hat{\beta}_1 = \frac{SSXY}{SSX} = \frac{2{,}111.3250}{72.5250} = 29.112$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = 254.45 - (29.112)(6.15)$$

$$= 254.45 - 179.0388 = 75.411$$

The Y-intercept:

$$\hat{\beta}_0 = 75.41, \text{ the sample estimate of } \beta_0.$$

The slope:

$$\hat{\beta}_1 = 29.11, \text{ the sample estimate of } \beta_1.$$

The OLS estimate of the regression equation is thus:

OLS Estimate of the regression line

$$\boxed{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i} \qquad \textbf{3.13}$$

$$\hat{Y}_i = 75.41 + 29.11X_i$$

This equation is called the *fitted line (or equation)* or *predictor line (equation),* or *regression line (equation).*

Thus, for example, when $X_i = 7$, the fitted value of Y_i is

$$\hat{Y}_i = 75.41 + 29.11X_i = 75.41 + 29.11(7)$$

$$= 75.41 + 203.77 = 279.18$$

$\hat{Y}_i = 279.2$ is the termed *the fitted value of Y,* or *the mean value* or, *the expected value of Y* when $X_i = 7$.

We substitute each X_i in the regression equation and determine all the Fitted Y-values.

Table 3.7

i	X_i	Actual Y_i	Fitted $\hat{Y}_i$
1	3.0	162.5	$162.75 = 75.41 + 29.11(3.0)$
2	4.5	188.0	$206.41 = 75.41 + 29.11(4.5)$
3	7.0	240.0	$279.19 = 75.41 + 29.11(7.0)$
4	11.0	385.5	$395.64 = 75.41 + 29.11(11.0)$
5	1.5	140.5	$119.08 = 75.41 + 29.11(1.5)$
6	5.0	202.0	$220.97 = 75.41 + 29.11(5.0)$
7	7.0	315.0	$279.19 = 75.41 + 29.11(7.0)$
8	9.5	385.5	$351.97 = 75.41 + 29.11(9.5)$
9	6.0	260.5	$250.08 = 75.41 + 29.11(6.0)$
10	7.0	265.0	$279.19 = 75.41 + 29.11(7.0)$

$\boxed{\text{Figure 3-11}}$

Estimating the Basic Model

Use the formulas to determine the OLS Regression Equation

Use $\hat{\beta}_1 = \frac{COV(X,Y)}{s_X^2}$ or $\hat{\beta}_1 = \frac{SSXY}{SSX}$

and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Second Step in Stage 3

Create the OLS Regression Equation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Third Step in Stage 3

Create a Table of Fitted Values

Substituting the X-values into the formula a Table of Fitted Values should be created.

The Table should include the X-values, Actual Y-values, and Fitted Y-values.

Fourth Step in Stage 3

Plot the Fitted Y-values with the actual Y-values.

Solved Problem 2

Determine the Equation of the Regression Line of the Second Set of Data (Solved Problem 1).

Steps in the Solution

1 Use the OLS Regression Formulas
2 Create the OLS Regression line
3 Create a Table of Fitted Values
4 Plot the Fitted Values

1 Use the OLS Regression Formulas

Using the data and calculations from Table 3.5 we determine the parameters estimates of the regression line. In this example, using formulas (3.13a) and (3.14) we have

$SSXY =$ -133.0 and $SSX =$ 100.5 were determined on page 10.

$$\hat{\beta}_1 = \frac{SSXY}{SSX} = \frac{-133.0}{100.5} = -1.3234$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 12.0 - (-1.3234)(5.5) = 12.0 + 7.2786$$
$$= 19.2786$$

The Y-intercept: $\hat{\beta}_0 = 19.2786$, the sample estimate of β_0.

The slope: $\hat{\beta}_1 = -1.3234$, the sample estimate of β_1.

2 Create the OLS Regression line

The OLS estimate of the equation of the regression line is thus

$$\hat{Y}_i = 19.28 - 1.32X_i$$

3 Create a Table of Fitted Values

Table 3.8

Observation i	X_i	Actual Y_i	Fitted $\hat{Y}_i$	
1	2	17	16.63	$= 19.28 - 1.32(2)$
2	7	8	10.01	$= 19.28 - 1.32(7)$
3	5	12	12.66	$= 19.28 - 1.32(5)$
4	10	5	6.04	$= 19.28 - 1.32(10)$
5	9	6	7.37	$= 19.28 - 1.32(9)$
6	9	9	7.37	$= 19.28 - 1.32(9)$
7	1	16	17.96	$= 19.28 - 1.32(1)$
8	2	17	16.63	$= 19.28 - 1.32(2)$
9	7	13	10.01	$= 19.28 - 1.32(7)$
10	3	17	15.31	$= 19.28 - 1.32(3)$

4 Plot the Fitted Values

Figure 3-12

Before continuing we must clearly identify the underlying assumptions of the OLS linear regression model.

Assumptions for the Linear Regression Model

Assumption 1

Y_i can be modeled by the linear stochastic or probabilistic equation

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

in which each X_i is uncorrelated with the corresponding ϵ_i

Assumption 2

ϵ_i is a random variable which is

(a) normally distributed,
(b) with mean $= 0$,
(c) and variance, $\sigma^2 = \sigma_\epsilon^2$.

Assumption 3

For different observations, *i* and *i*′, ϵ_i and $\epsilon_{i'}$ are uncorrelated;
i.e. $COV(\epsilon_i, \epsilon_{i'}) = 0$

These three assumptions regarding the model and the disturbance term mean that we assume that the Y_i varies randomly around the regression line, and that the variance around the regression is the same regardless of the X_i point. The assumption about equal variance is termed *homoscedasticity*, a funny sounding word from the Greek language meaning “same scatter."

Figure 3-13 below illustrates the assumption of equal variance around the regression line.

Figure 3-13

At this point we have identified the regression model and have estimated the parameters of the model. The next stage is to perform diagnostics to check the statistical validity of the model.

Stage 4 Diagnostics and Residual Analysis

Diagnostics 1
Calculating MSE

The Theory

Calculating the ***Mean Squared Error*** of the model, the ***MSE***, or $\boldsymbol{s_\epsilon^2}$

Many of the diagnostics require the the estimation of σ_ϵ^2. We have assumed that the linear model can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

σ_ϵ^2 is the variance of the ϵ_i's; a measure of the variability of the Y_i's around the regression line.

As β_0 and β_1 are unknown parameters, estimated by OLS by calculating $\hat{\beta}_0$ and $\hat{\beta}_1$, similarly, since ϵ_i is an unknown value for each *i* they must be estimated by the *residual* or *error term*, denoted $\hat{\epsilon}_i$. $\hat{\epsilon}_i$ is the difference between the Actual, Y_i, and the Fitted or Predicted, $\hat{Y}_i$.

The Formulas

Residual

$$\boldsymbol{Error = Actual - Fitted} \qquad \textbf{3.15}$$

Residual

$$\boldsymbol{\hat{\epsilon}_i = Y_i - \hat{Y}_i} \qquad \textbf{3.16}$$

Thus, the sum of the squared errors
SSE

$$\boldsymbol{SSE = \sum \hat{\epsilon}_i^2} \qquad \textbf{3.17}$$

From the *SSE* we calculate the *MSE* or s_ϵ^2.

$\boldsymbol{s_\epsilon^2}$ or MSE

$$s_\epsilon^2 = \frac{SSE}{n-2} \qquad \textbf{3.18}$$

We divide by two less than sample size since we estimated two parameters, β_0 and β_1 and lost two degrees of freedom.

The Practice in Calculator Format

Listed below in Table 3.9 are the actual and fitted values for each X_i, given the equation $Y_i = 75.41 + 29.11X_i$, the error or residual, $\hat{\epsilon}_i$, and the error squared, $\hat{\epsilon}_i^2$.

Table 3.9

Observation i	X_i	Actual Y_i	Fitted $\hat{Y}_i$	Error $Y_i - \hat{Y}_i = \hat{\epsilon}_i$	Squared Error $\hat{\epsilon}_i^2$
1	3.0	162.5	$162.75 = 75.41 + 29.11(3.0)$	-0.25	.06
2	4.5	188.0	$206.41 = 75.41 + 29.11(4.5)$	-18.41	339.08
3	7.0	240.0	$279.19 = 75.41 + 29.11(7.0)$	-39.19	1,536.17
4	11.0	385.5	$395.64 = 75.41 + 29.11(11.0)$	-10.14	102.86
5	1.5	140.5	$119.08 = 75.41 + 29.11(1.5)$	+21.42	458.90
6	5.0	202.0	$220.97 = 75.41 + 29.11(5.0)$	-18.97	359.86
7	7.0	315.0	$279.19 = 75.41 + 29.11(7.0)$	+35.81	1,282.07
8	9.5	385.5	$351.97 = 75.41 + 29.11(9.5)$	+33.53	1,123.99
9	6.0	260.5	$250.08 = 75.41 + 29.11(6.0)$	+10.42	108.53
10	7.0	265.0	$279.19 = 75.41 + 29.11(7.0)$	-14.19	201.47
			Sums	0	5,513

Sum of Residuals (Errors)

$$\sum \hat{\epsilon}_i = 0$$

Sum of Squared Errors

$$SSE = \sum \hat{\epsilon}_i^2 = 5{,}513.00$$

Calculating *MSE* or s_ϵ^2

$$s_\epsilon^2 = \frac{SSE}{n-2} = \frac{5{,}513.00}{8} = 689.12$$

We remind the reader that the $SSE = 5{,}513.00$ is the smallest such sum of squared errors since the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ were constructed so as to minimize the *SSE*. There are many other pairs of parameters that would cause the sum of the errors to equal 0, but their *sum of squared errors* would be larger than the *SSE* of 5,513.00.

$$s_\epsilon = \sqrt{689.12} = 26.25$$

s_ϵ is called the *standard error of the model* or the *standard error of the estimate.*

Solved Problem 3

Determine the Residuals and MSE of the Second Regression Line

Steps in the Solution

1 Use the OLS Regression Equation to determine the Fitted Values.
2 From the Fitted Values create a Table of Fits, Residuals and Squared Residuals
3 Determine the Sum of the Residuals and Squared Residuals
4 MSE is then the Sum of the Squared Residuals divided by n $-$ 2.

1 Use the OLS Regression Equation

Using the regression equation, $\hat{Y}_i = 19.28 - 1.32X_i$ we calculate the Fitted values for each X_i (See Table 3.10).

2 Determine the Residuals and the Squared Residuals

We compare the Fitted Values with the Actual Values to determine the error or residual, $\hat{\epsilon}_i$, and the error squared, $\hat{\epsilon}_i^2$ (See Table 3.10).

Table 3.10

Observation i	X_i	Actual Y_i	Fitted $\hat{Y}_i$		Error $Y_i - \hat{Y}_i = \hat{\epsilon}_i$	Squared Error $\hat{\epsilon}_i^2$
1	2	17	16.63	$= 19.28 - 1.32(2)$	0.37	0.14
2	7	8	10.01	$= 19.28 - 1.32(7)$	-2.01	4.06
3	5	12	12.66	$= 19.28 - 1.32(5)$	-0.66	0.44
4	10	5	6.04	$= 19.28 - 1.32(10)$	-1.04	1.09
5	9	6	7.37	$= 19.28 - 1.32(9)$	1.37	1.87
6	9	9	7.37	$= 19.28 - 1.32(9)$	1.63	2.66
7	1	16	17.96	$= 19.28 - 1.32(1)$	-1.96	3.82
8	2	17	16.63	$= 19.28 - 1.32(2)$	0.37	0.14
9	7	13	10.01	$= 19.28 - 1.32(7)$	2.99	8.91
10	3	17	15.31	$= 19.28 - 1.32(3)$	1.69	2.86
				Sums	0.00	25.99

3 Determine the Sum of Residuals and the Sum of the Squared Residuals

Sum of Errors $\sum \hat{\epsilon}_i = 0$

Sum of Squared Errors $\sum \hat{\epsilon}_i^2 = 25.99$

4 Calculate MSE

Calculating *MSE* or s_ϵ^2 $\quad s_\epsilon^2 = \frac{SSE}{n-2} = \frac{25.99}{8} = 3.25$

Diagnostics 2
Diagnostic check of β_1

The Theory

Hypothesis Testing of β_1

β_1 is the slope of the regression equation, and for the regression equation to be useful requires that the slope be non-zero. A slope of zero, $\beta_1 = 0$, means that the regression line is a horizontal line, and would not be a useful line for forecasting purposes.

Consequently, we must test the hypothesis that β_1 is statistically significantly different from zero (And trying repeating "statistically significantly" three times without stopping.) In other words, the null hypothesis that we establish is:

1 $$H_0: \quad \beta_1 = 0$$

And the alternate hypothesis is that β_1 is different from zero.

2 $$H_a: \quad \beta_1 \neq 0$$

The hypothesis test statistic is the t-ratio

3 $$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

We compare that calculated t-ratio against the 5% two-tail t-critical value at $n - 2 = 8$ degrees of freedom.

4 $$\text{t-critical value} = 2.306$$

If our calculated t-ratio is greater in absolute value than the t-critical value, then β_1 is statistically significantly different from zero. To be different from zero requires it to be at least 2.306 standard deviations from 0. The sample data imply that the population parameter β_1 is significantly different from 0.

5

Hence, if we reject the null hypothesis that β_1 is equal to zero, we say the data implies that β_1 is not zero. With β_1 not equal to zero means that we have a regression line that is not horizontal.

The Formulas

The variance of β_1 is determined by the formula:

Variance of β_1

$$s^2_{\hat{\beta}_1} = \frac{s^2_\epsilon}{(n-1)s^2_X} \quad \textbf{3.19}$$

$$s^2_{\hat{\beta}_1} = \frac{s^2_\epsilon}{(n-1)s^2_x} = \frac{689.12}{(9)(8.06)} = 9.501$$

The standard error of β_1

$$s_{\hat{\beta}_1} = \sqrt{s^2_{\hat{\beta}_1}} \quad \textbf{3.20}$$

$$s_{\hat{\beta}_1} = \sqrt{9.501} = 3.08$$

The Practice in Calculator Format

The hypothesis test statistic is the t-ratio

3 $$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{29.11}{3.08} = 9.45$$

We compare that calculated t-ratio against the 5% two-tail t-critical value at $n - 2 = 8$ degrees of freedom.

4 t-critical value $= 2.306$

Our calculated t-ratio is greater in absolute value than the t-critical value.[1] This means that $\hat{\beta}_1$ is statistically significantly different from zero. To be different from zero requires it to be at least 2.306 standard deviations from 0; our calculated value is 9.45 standard deviations from 0. It is significantly different from 0.

5 Hence, we reject the null hypothesis that β_1 is equal to zero. With β_1 not equal to zero means that we have regression line that is not horizontal.

Figure 3-14

1

Solved Problem 4

Perform a Diagnostic Check of β_1 in the Second Regression Equation

$$\hat{Y}_i = 19.28 - 1.323X_i$$

Steps in the Solution

1 Determine the standard error of $\hat{\beta}_1$.
2 Calculate the t-value of $\hat{\beta}_1$.
3 Compare the t-value of $\hat{\beta}_1$ with the critical value from the t-table.
4 Decide on β_1.

1 Determine the standard error of $\hat{\beta}_1$

Using $s_\epsilon^2 = 3.25$ and $s_x^2 = 11.2$, we determine the standard error of $\hat{\beta}_1$. We use the information from Table 3.2 page 4.

variance of $\hat{\beta}_1$ $$s^2_{\hat{\beta}_1} = \frac{s_\epsilon^2}{(n-1)s_x^2} = \frac{3.25}{(9)(11.2)} = .0322$$

standard error of $\hat{\beta}_1$ $$s_{\hat{\beta}_1} = \sqrt{.0322} = .1796$$

2 Calculate the t-ratio of $\hat{\beta}_1$

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{-1.323}{0.1796} = -7.369$$

3 Determine the t-critical value

At $\alpha = 5\%$ level of significance and $10 - 2 = 8$ degrees of freedom, t-critical value is 2.306

4 Decide on β_1

The *t*-ratio is much larger in absolute value than 2 so it allows us to easily reject the implicit null hypothesis that β_1 is zero.

Diagnostics 3
Goodness of Fit **The Coefficient of Determination**

The Theory

Having determined $\hat{\beta}_1$ is a non-zero parameter estimate, we wish to examine how “good a fit" the regression line is to the actual data. The residuals of the fitted line will provide a good measure of how well the regression line fits the data. Large residuals imply a poor linear fit, while small residuals imply a good fit. However, “large" and “small" residuals are relative to the variability of *Y*. Consequently we will construct a measure of goodness of fit through the variability of *Y* and the variability of the residuals.

Consider Figure 3-15 below.

Figure 3-15

Point A, $(7, 315)$ in Figure 3-16 represents the actual value of Y_i when $X_i = 7$ for a particular observation, (X_i, Y_i).

Point B, $(7, 279.2)$ is the fitted value of Y_i, $\hat{Y}_i$, when $X_i = 7$ for this observation, $(X_i, \hat{Y}_i)$.

Point C, (7, 254.5) is the mean value of Y, $\bar{Y}$, for this observation (or for any observation), $(X_i, \bar{Y}_i)$.

Now, for the regression equation to have explanatory power, the fitted value $\hat{Y}_i$, in general, should be closer to the actual value, Y_i, than the mean value, $\bar{Y}$, is to the actual value. In other words, regression should forecast better than the simple average of the Y's

For this observation the difference between Actual and Fitted, the Residual, is

$$Y_i - \hat{Y}_i = 315 - 279.2 = 35.80$$

While the difference between the Actual and the Mean is

$$Y_i - \bar{Y}_i = 315 - 254.5 = 60.50$$

Notice also that geometrically, the distance from A to C is equal to the sum of the distances from A to B and from B to C.

Figure 3-16

In this figure,

Distance from A to C = (Distance from A to B) + (Distance from B to C)

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

It is an algebraic fact (which we won't prove in this book) that the above equation holds for the sum of the squares of each term.

That is, because,

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \tag{3.21}$$

then,

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \qquad 3.22$$

This equation is often titled the

Sum of Squares Partitioning

$$\boxed{\boldsymbol{\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \qquad 3.23}}$$

The terms of the equation are denoted by

SSY, the Total Variation

$$\boxed{\boldsymbol{\sum(Y_i - \bar{Y})^2 = SSY \qquad 3.24}}$$

SSE, the Variation due to Error

$$\boxed{\boldsymbol{\sum(Y_i - \hat{Y}_i)^2 = SSE \qquad 3.25}}$$

We define $\sum(\hat{Y}_i - \bar{Y}_i)^2$ as the *Sum of Squares Due to Regression*, or the *Variation due to Regression.* $\sum(\hat{Y}_i - \bar{Y})^2$ is a measure of variation of the regression fits around $\bar{Y}$. We denote it as *SSR*

SSR, the Variation due to Regression

$$\boxed{\boldsymbol{\sum(\hat{Y}_i - \bar{Y})^2 = SSR \qquad 3.26}}$$

Thus,

$$\boxed{\boldsymbol{SSY = SSR + SSE \qquad 3.27}}$$

$\sum(\hat{Y}_i - \bar{Y})^2$, or *SSR*, is the *Variation around $\bar{Y}$ due to the Regression.* It is called the “Explained Variation," because it is the variation explained by the regression equation.

$\sum(Y_i - \hat{Y}_i)^2$, or *SSE*, is the *Variation around Y_i due to the Residuals or Errors. SSE* is the “Unexplained Variation."

Thus, equation 3.31 may be written as

$$\boxed{\boldsymbol{Total\ Variation = Explained\ Variation + Unexplained\ Variation \qquad 3.28}}$$

Our interest lies with the ratio of the Explained Variation to the Total Variation, which we define as R^2.

The Formulas

The Coefficent of Determination, $\mathbf{R^2}$

$$R^2 = \frac{\textit{Explained Variation}}{\textit{Total Variation}} \qquad \mathbf{3.29}$$

$\mathbf{R^2}$

$$R^2 = \frac{SSR}{SSY} = 1 - \frac{SSE}{SSY} \qquad \mathbf{3.30}$$

The Practice in Calculator Format

Determining R^2

From the formula

$$SSY = SSR + SSE$$

we have

$$SSR = SSY - SSE$$

We know from before that

$$SSY = 66{,}997.2250 \qquad \text{(page } \quad)$$

$$SSE = 5{,}513.00 \qquad \text{(page } \quad)$$

so

$$SSR = 66{,}997.2250 - 5{,}513.00 = 61{,}484.2250$$

$$R^2 = \frac{SSR}{SSY} = \frac{61{,}484.2250}{66{,}997.2250} = .918$$

Hence, $R^2 = 91.8\%$

R^2 is called the *Coefficient of Determination.* R^2 is a measure of the *goodness of fit* of the regression equation. R^2 ranges from 0 to 1 so that a perfect fit causes $R^2 = 1$ and no fit causes $R^2 = 0$.

R^2 is a measure of the variation in Y due to the variation in X. We interpret $R^2 =$ 92% to mean that 92% of the variation in Y is due to the variation in X.

For example, if we increase X one unit from $X_i = 7$ to $X_i = 8$, then $\hat{Y}_i$ changes from $\hat{Y}_i = 279.2$ to $\hat{Y}_i = 308.3$. Nearly 92 percent of the change in $\hat{Y}_i$ of 29.1 units can be explained by the one unit change in X_i.

Solved Problem 5

Determine R^2 of the Second Regression Line

Steps in the Solution

1 Use SSY and SSE to determine SSR
2 Use the formula to determine R^2

1 Use SSY and SSE to determine SSR

We use the formula,

$$SSY = SSR + SSE$$

$$202.0 = SSR + 25.99$$

$$SSR = 202.00 - 25.99 = 176.01$$

2 Use the formula to determine R^2

$$R^2 = \frac{SSR}{SSY} = \frac{176.01}{202.00} = .871$$

$$R^2 = 87.1\%$$

Diagnostics 4

***ANOVA* and the *F* test**

The Theory

Analysis of Variance (ANOVA) is a way of determining if two means are statistically significantly different. In the case of regression analysis and forecasting we are testing if the “mean" $E(Y_i|X_i)$, for each *i*, is statistically significantly different from $\bar{Y}$. Again, we are testing if our model has more explanatory power than just using $\bar{Y}$ as a forecast.

Simple *ANOVA* uses the partitioning of the total variation, *SSY*, into the variation due to treatments plus the variation due to error.

$$SSY = SST + SSE$$

In the setting of a regression model, the variation due to treatment is the variation due to regression, *SSR*. Hence, as in equation (3.31), (page)

$$SSY = SSR + SSE \qquad 3.31$$

The Formulas

The Explained Variance, Mean Square due to Regression, MSR

$$\textbf{\textit{Explained Variance}} = \boldsymbol{MSR} = \frac{\boldsymbol{SSR}}{\boldsymbol{k}} \qquad \textbf{3.31}$$

where k is the number of independent variables in the regression model. In this case $k = 1$, for simple linear regression.

The Unexplained Variance, Mean Square due to Error, MSE

$$\textbf{\textit{Unexplained Variance}} = \boldsymbol{MSE} = \frac{\boldsymbol{SSE}}{\boldsymbol{n - (k + 1)}} \qquad \textbf{3.32}$$

The F-statistic

$$\boldsymbol{F} = \frac{\boldsymbol{MSR}}{\boldsymbol{MSE}} \qquad \textbf{3.33}$$

The Practice in Calculator Format

From equation (3.31), we have

$$SSY = SSR + SSE$$

or

$$SSR = SSY - SSE$$

$$SSR = 66977 - 5513$$

$$SSR = 61464$$

Thus, $MSR = \frac{SSR}{k} = \frac{61,464}{1} = 61,464$

and $MSE = \frac{SSE}{n-(k+1)} = \frac{5513}{8} = 689.125$

$$F = \frac{MSR}{MSE} = \frac{61,464}{689.125} = 89.19$$

The complete *ANOVA* table thus is

Table 3.11

Variation due to		Degrees of Freedom	Variances	
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n - (k + 1)$	$MSE = \frac{SSE}{n-(k+1)}$	

Total	SSY	$n - 1$

Table 3.12

Variation due to		Degrees of Freedom	Variances
Regression	SSR	$k = 1$	$MSR = \frac{SSR}{k} = \frac{61{,}464}{1} = 61{,}464$
Error	SSE	$n - (k + 1) = 8$	$MSE = \frac{SSE}{n-(k+1)} = \frac{5{,}513}{8} = 689.125$
Total	SSY	$n - 1 = 9$	

$$F = \frac{MSR}{MSE} = \frac{61{,}464}{689.125} = 89.19$$

Testing at $\alpha = .05$, the F critical value at 1 and 8 degrees of freedom is

$$F_{1,8} = 5.32$$

The computed F statistic

$$F = 89.21$$

We reject the implicit null hypothesis that there is no difference among the means, $E(Y_i|X_i)$ and $\bar{Y}$.

Solved Problem 6

Determine F and the *ANOVA* table of the Second Regression Line
(See page 34 for data and sums)

Steps in the Solution

1 Use the values of SSR and SSE, determine MSR and MSE
2 Create the ANOVA table
3 Determine the F value
4 Compare it to the F critical value

1 Use the values of SSR and SSE in the Table

2 Create the ANOVA table

3 Determine the F value

See Table 3.13 below

Table 3.13

Variation due to	Degrees of Freedom	Variances

Regression	$SSR = 176.01$	$k = 1$	$MSR = \frac{SSR}{k} = \frac{176.01}{1} = 176.01$
Error	$SSE = 25.99$	$n - (k+1) = 8$	$MSE = \frac{SSE}{n-(k+1)} = \frac{25.99}{8} = 3.249$
Total	$SSY = 202.00$	$n - 1 = 9$	

$$F = \frac{MSR}{MSE} = \frac{176.01}{3.249} = 54.175$$

4 Compare to the F critical value

Testing at $\alpha = .05$, the F critical value at 1 and 8 degrees of freedom is

$$F_{1,8} = 5.32$$

The computed F statistic

$$F = 54.175$$

We reject the implicit null hypothesis that there is no difference among the means, $E(Y_i|X_i)$ and $\overline{Y}$.

Summarizing the Diagnostics of a Regression Model

A useful summary method of listing the diagnostic statistics with the regression model is

$$\hat{Y}_i = \underset{}{75.41} + \underset{(9.45)}{29.11}X_i \qquad R^2 = .918 \qquad F_{1,8} = 89.21$$

or,

$$\widehat{SALES}_i = 75.41 + \underset{(9.45)}{29.11}(ADVER_i) \qquad R^2 = .918 \qquad \mathrm{F}_{1,8} = 89.21$$

The t-ratios are listed in parentheses below the estimated parameters. In general, if the t-ratio is greater than 2, we then conclude that the parameters are non-zero and may be retained in the model.

The $R^2 = .918$ indicates that the regression equation explains almost 92% of the variation in the dependent variable. And the $F = 89.21$ allows us to reject the null hypothesis that there is no relationship between Sales and Advertising.

Solved Problem 7

Summarize the Diagnostics of the Second Regression Model

Steps in the Solution

1 Reproduce the Regression Equation
2 Add the t-ratios, the R^2, and the F values

$$\hat{Y}_i = \underset{(16.91)}{19.28} - \underset{(-7.37)}{1.32}X_i \qquad R^2 = .87 \qquad F_{1,8} = 54.175$$

Diagnostics

The Fourth Stage in the Forecasting Process is to perform statistical tests to check the statistical validity of the forecasting model.

The Practice in Calculator Format

First Step in Stage 4

- Determine the *MSE*, s_ϵ^2 of the Regression Equation

This is a measure of the dispersion of the data around the regression line.

Second Step in Stage 4

- Determine the standard deviation of $\hat{\beta}_1$.

Third Step in Stage 4

- Determine the t-ratio of $\hat{\beta}_1$ and look up the t-critical value in the t-table.

Use the t-ratio to statistically test if $\hat{\beta}_1$ is significant.

Fourth Step in Stage 4

- Determine the R^2 of the model.

Look for a reasonably high R^2.

Fifth Step in Stage 4

- *Create the ANOVA* table and determine the *F* statistic of the model

Check the *F* statistic against the critical value for the *F*.

Sixth Step in Stage 4

- Summarize the diagnostics

Make sure all looks good before going on to forecasting.

Stage 5 **Forecasting and Confidence Intervals of the Model**

The Theory

The Difference between a Forecasted Value and a Fitted Value

We have established a model linking Sales Volume and Advertising.

$$\hat{Y}_i = 75.41 + 29.11X_i$$

We forecast Sales when $X = \$10{,}000$. That is when, $X_i = 10$, then Sales Volume, Y_i, is forecasted to be

$$\hat{Y}_i = 75.41 + 29.11(10) = 366.51$$

$$\widehat{SALES}_i = \$366{,}510$$

Using $X = 10$ to determine $\hat{Y} = 366.51$ is a **Forecasted Value** because there was no instance in our data when $X = 10$. $X = 10$ is a new value for X. This is also called an ***interpolation*** because the X value we chose was between actual X values from the data. In our data on X, we have values for Y when X is 9.5 and when X is 11.0. $\hat{Y} = 366.51$ is an interpolation when $X = 10$.

Using $X = 7$ to determine $\hat{Y} = 279.1$ is a **Fitted Value** because we had three instances in which $X = 7$ and we had a corresponding actual Y value for comparison to the Fitted Y Value.

Forecasting the Expected Value (Mean Value) of Y_i given X_i

Recall that the regression equation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

is the best estimate of the expected value equation

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

Figure 3-17

So that when $X_i = 10$, and correspondingly $\hat{Y}_i = 366.51$, then 366.51 is the best estimate of $E(Y_i|X_i = 10)$;

$$E(Y_i|X_i = 10) = 366.51$$

This means that the *expected value* or *mean value* of Y_i is 366.51 when $X_i = 10$. In other words, the *mean value* of Y_i is 366.51 over all occasions when X_i is 10.

Confidence Intervals of Forecast

Now $\hat{\beta}_0$ and $\hat{\beta}_1$ are based on a sample of bivariate data. Thus, if we were to re-sample the bivariate data we would obtain different estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ (a different regression line), and correspondingly a different value for $\hat{Y}_i$. This is often termed the "sampling error of estimate."

Thus, we wish to construct a confidence interval around the forecasted expected value of Y_i which takes into account the sampling error of the estimates.

To construct a confidence interval around $\hat{Y}_i$ requires a variance of forecast.

The Formulas

Variance of mean forecast

$$s^2_{\hat{Y}_i} = s^2_{\epsilon}\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}\right] \qquad \textbf{3.34}$$

By convention we name the square root of the variance of forecast, the *standard error of forecast.*

Standard Error of Forecast

$$s_{\hat{Y}_i} = \sqrt{s^2_{\epsilon}\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}\right]} \qquad \textbf{3.35}$$

Equation (3.34) reveals that the variance of forecast is dependent on the variability of the mean of the data around the regression line as represented by s^2_{ϵ}. In addition, the variance of the forecast is dependent on the distance that X_i is from the mean value, $\bar{X}$. The further the chosen X_i is from $\bar{X}$, the greater the forecast variance.

Figure 3-18 below illustrates this variance. The forecast confidence interval will form a curved envelope with minimum interval at the mean $(\bar{X}, \bar{Y})$.

Figure 3-18

Mean Forecast Confidence Interval

$$\hat{Y}_i \pm t_{\alpha/2,n-(k+1)} s_{\hat{Y}_i} \qquad \mathbf{3.36}$$

The Practice in Calculator Format

For example, in the case where $X_i = 10$

mean:

$$\hat{Y}_i = 75.41 + 29.11(10) = 366.51$$

variance of mean forecast:

$$s^2_{\hat{Y}_i} = s^2_\epsilon \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} \right] \qquad 3.39$$

$$= 689.12 \left[\frac{1}{10} + \frac{(10 - 6.15)^2}{72.5250} \right] = 209.78$$

standard error of forecast:

$$s_{\hat{Y}_i} = \sqrt{209.78} = 14.48$$

A 95% confidence interval (level of significance α = 5%) with 10 observations uses

$$t_{\alpha/2,n-(k+1)} = t_{.05/2,8} = t_{.025,8} = 2.306$$

With a forecast of $\hat{Y}_i = 366.51$ and a standard error of forecast $s_{\hat{Y}_i} = 14.48$, a 95% confidence interval around $\hat{Y}_i = 366.51$ is

$$366.51 \pm (2.306)(14.48)$$

$$366.51 \pm 33.40$$

Thus, a 95 percent confidence interval around the *mean value forecast,* given $X_i =$ 10 is

$$333.11 \leq Y_i \leq 399.91$$

$$\$333{,}110 \leq Y_i \leq \$399{,}910, \text{ given } X_i = \$10{,}000$$

Figure 3-19

Forecasting an *Individual* Value Y_i given X_i

The Theory

If we wish a forecast of a particular value of Y_i given X_i, then we are using the regression equation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad 3.15$$

as the best estimate of

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad 3.7$$

For $X_i = 10$, the forecast will be the same as before,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 75.41 + 29.11(10) = 366.51$$

However, the forecast variance will be greater since we must now also take into account the variance of the ϵ's. Hence, the usual equation for the variance of forecast for a individual value, $\hat{Y}_i$ is

The Formulas

Variance of Individual Forecast

$$\boxed{s^2_{\hat{Y}_i} = s^2_\epsilon + s^2_\epsilon \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} \right] \qquad \mathbf{3.37}}$$

Standard Error of Individual Forecast

$$\boxed{s_{\hat{Y}_i} = \sqrt{s^2_\epsilon + s^2_\epsilon \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} \right]} \qquad \mathbf{3.38}}$$

The Practice in Calculator Format

In our example, then, when $X_i = 10$, $\hat{Y}_i = 366.51$ and

Variance of Individual Forecast*:*

$$s^2_{\hat{Y}_i} = s^2_\epsilon + s^2_\epsilon \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} \right] = 689.2 + 689.2 \left[\frac{1}{10} + \frac{(10 - 6.15)^2}{72.5250} \right] = 898.99$$

Standard Error of Forecast for the Individual:

$$s_{\hat{Y}_i} = \sqrt{898.99} = 29.98$$

With a forecast of $\hat{Y}_i = 366.51$ and a standard error of forecast $s_{\hat{Y}_i} = 29.98$, we form a forecast confidence interval around $Y_i = 366.51$ as

$$366.51 \pm (2.306)(29.98)$$

$$366.51 \pm 69.14$$

$$297.37 \le Y_i \le 435.65$$

When $X_i = \$10{,}000$ on a single, individual occasion, we forecast, with 95 percent confidence, that the Sales Volume will be somewhere between $297,370 and $435,650.

This confidence interval is considerably wider than the previous,

$$\$333{,}110 \le Y_i \le \$399{,}910, \quad \text{given } X_i = \$10{,}000$$

The confidence interval around an individual estimate is always larger than the confidence interval around a mean estimate. Figure 3-20 illustrates this point.

Figure 3-20

Solved Problem 8

Produce Forecasts and Confidence Intervals using the Second Regression Line

Steps in the Solution

1 Substitute in values for the independent variable
2 Determine the confidence intervals of forecast

1 Substitute in values for the independent variable

The regression equation of the second example is

$$\hat{Y}_i = 19.28 - 1.32X_i$$

For example, if $X_i = 4$, then Y_i, is forecasted to be

$$\hat{Y}_i = 19.28 - 1.32(4) = 14.0$$

We construct a 95% confidence interval around the forecast of $\hat{Y}_i = 14.0$

Variance of the mean forecast

$$s^2_{\hat{Y}_i} = s^2_\epsilon \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} \right]$$

where $n = 10$, $\bar{X} = 5.5$, $s^2_\epsilon = 3.25$, and $SSX = 100.5$.

$$s^2_{\hat{Y}_i} = s^2_\epsilon \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} \right] = 3.25 \left[\frac{1}{10} + \frac{(4 - 5.5)^2}{100.5} \right] = 0.3978$$

$$s_{\hat{Y}_i} = \sqrt{.3978} = 0.6307$$

Mean Forecast Confidence Interval

$$\hat{Y}_i \pm t_{\frac{\alpha}{2}, n-(k+1)} s_{\hat{Y}_i}$$

A 95% confidence interval around the mean value forecast, using

$$t_{\frac{\alpha}{2}, n-(k+1)} = t_{\frac{.05}{2}, 8} = t_{.025,8} = 2.306$$

$$14.0 \pm (2.306)(0.6307)$$

$$14.0 \pm 1.4544$$

Thus, a 95% confidence interval around the mean value forecast, given $X_i = 4$ is

$$12.5 \leq Y_i \leq 15.5$$

Figure 3-21

The estimate of the variance of the *individual* forecast has an additional variance term

$$s^2_{\hat{Y}_i} = s^2_\epsilon + s^2_\epsilon \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} \right]$$

Hence,

$$s^2_{\hat{Y}_i} = s^2_\epsilon + s^2_\epsilon\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}\right] = 3.25 + 3.25\left[\frac{1}{10} + \frac{(4-5.5)^2}{100.5}\right] = 3.25 + 0.3978 = 3.6478$$

$$s_{\hat{Y}_i} = \sqrt{3.6478} = 1.9099$$

$$14.0 \pm (2.306)(1.9099)$$

$$14.0 \pm 4.4043$$

Thus, a 95% confidence interval around the individual forecast, given $X_i = 4$ is

$$9.6 \leq Y_i \leq 18.4$$

Figure 3-22

Forecasting and Confidence Intervals

The Fifth Stage in the Forecasting Process is the fun part of producing forecasts.

The Practice in Calculator Format

First Step in Stage 5

- Determine the forecasted Y-values by using the appropriate X-values.

Second Step in Stage 5

- Determine the variance of forecast

Remember to distinguish between a forecast of the mean from the forecast of individual since the variance of forecast differs.

Third Step in Stage 5

- Determine standard error of forecast, look up the appropriate t-value, and calculate the 95% confidence interval.

We are essentially determining the high and low value of forecast.

A Final Note on the Importance of Graphing Data

Given the power and availability of computer software to compute regression estimates, it is often easy to forget to just look at the data. The following example is a famous morality tale about relying too heavily on numerical output and not graphically inspecting the data. We consider the following four sets of bivariate data.

Table 3.14

I		II		III		IV	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

We put all four sets through a software package to determine a number of numerical measures. It turns out that for all four data sets. . .

The sample size $n = 11$.

The means are the same.

$$\bar{X} = 9.0 \qquad \bar{Y} = 7.5$$

The regression equation for all four is

$$\hat{Y}_i = 3 + .5X_i$$

And for all four sets of data

$$s_{\hat{\beta}_1} = 0.118$$

$$t = 4.24$$

$SSX = 110.0 \quad SSR = 27.50 \quad SSE = 13.75 \quad \rho_{XY} = .82 \quad R^2 = .67$

With all these equal numerical measures we would expect the data sets to look much alike. Now, consider the scatter diagrams of data sets I, II, III, and IV in Figure 3-23.

Figure 3-23

Stage 6 **Evaluation of the Performance of the Forecasting Model**

We believe that the Evaluation of the Performance of the Forecasting Model is a very important, and yet, too often neglected Stage in the Forecasting Process. Consequently, we shall try to include within each chapter some discussion of Forecast Evaluation, and we are including a full chapter, Chapter 14, specifically on Forecast Evaluation. The reader may thus refer to Chapter 14 at any point for additional and expanded information on Forecast Evaluation.

We begin here with some of the basic concepts of Forecast Evaluation. There are several methods of evaluating forecast errors. After producing a set of forecasts and then comparing the forecast with the actual to determine the errors, we have a set of forecast errors. For example, let us suppose on three different occasions we forecasted Sales $=$ 366.51 (using Advertising $=$ 10), and that on three other occasions we forecasted Sales $=$ 308.29 (using Advertising $=$ 8). We then compare the six forecasts with the six actual values for those six particular occasions.

The difference between e_i and $\hat{e}_i$

It is important to understand the difference between e_i and $\hat{e}_i$. e_i is the Error of Forecast, the difference between the Actual Value and the Forecasted Value. Whereas $\hat{e}_i$ is the Residual, the difference between the Actual Value and the Fitted Value.

Forecast Error

$$\textit{Forecast Error} = \textit{Actual} - \textit{Forecast} \qquad \textbf{3.39}$$

Forecast Error

$$e_i = Y_i - \hat{Y}_i \qquad \textbf{3.40}$$

Table 3.15

	Actual	Forecast	Forecast Error
i	Y_i	$\hat{Y}_i$	e_i
1	384.84	366.51	$+$ 18.33
2	410.49	366.51	$+$ 43.98
3	274.38	308.29	$-$ 33.91
4	292.88	308.29	$-$ 15.41
5	311.53	366.51	$-$ 54.98
6	366.87	308.29	$+$ 58.58

Forecast error measures

Bias

Bias is just the mean of the forecast errors.

Bias

$$BIAS = \frac{\sum e}{n} \qquad \textbf{3.41}$$

BIAS describes the average size and direction of the forecast error. *BIAS* can be either positive, negative, or zero. Notice that a *positive* error means an *underforecast,* while a *negative* error means an *overforecast.*

$$BIAS = \frac{\sum e}{6} = \frac{+16.62}{6} = +2.77$$

MAD

MAD is the abbreviation for ***Mean Absolute Deviation (Error)***. It is similar to *BIAS*, however, the absolute value of the forecast error is used. The absolute value of the forecast error is simply the magnitude of the forecast error with no consideration to the direction of the forecast error.

MAD

$$\boldsymbol{MAD = \frac{\sum |e|}{n}} \qquad \textbf{3.42}$$

Using the above example.

Table 3.16

	Actual	Forecast	Forecast Error	Absolute Forecast Error
i	Y_i	$\hat{Y}_i$	e_i	$\lvert e_i \rvert$
1	384.84	366.51	+ 18.33	18.33
2	410.49	366.51	+ 43.98	43.98
3	274.38	308.29	− 33.91	33.91
4	292.88	308.29	− 15.41	15.41
5	311.53	366.51	− 54.98	54.98
6	366.87	308.29	+ 58.58	58.58

$$MAD = \frac{\sum |e|}{6} = \frac{225.19}{6} = 37.53$$

MAD describes the average magnitude of error.

MSE

MSE is the abbreviation for ***Mean Squared Error***. Instead of absolute value the forecast errors are squared.

MSE

$$\boldsymbol{MSE = \frac{\sum e^2}{n}} \qquad \textbf{3.43}$$

Using the above example

Table 3.17

Actual	Forecast	Error	Squared Error

i	Y_i	$\hat{Y}_i$	e_i	e_i^2
1	384.84	366.51	$+$ 18.33	335.99
2	410.49	366.51	$+$ 43.98	1,934.24
3	274.38	308.29	$-$ 33.91	1,149.89
4	292.88	308.29	$-$ 15.41	237.47
5	311.53	366.51	$-$ 54.98	3,022.80
6	366.87	308.29	$+$ 58.58	3,431.62
			Sum	10,112.01

$$MSE = \frac{\sum e^2}{6} = \frac{10{,}112.01}{6} = 1{,}685.34$$

MAPE

MAPE is the abbreviation for ***Mean Absolute Percent Error***. *MAPE* is useful for comparing forecast accuracy of different data sets. It two data sets are of different magnitudes of numbers, *MAD* and *MSE* would not be comparable. By using *MAPE* we can compare the forecast accuracy of different data sets because *MAPE* converts them all to a percentage basis.

Percent error $$PE = \frac{\text{Forecast Error}}{\text{Actual}} \times 100$$

MAPE

$$\boxed{\boldsymbol{MAPE = \frac{\sum |PE|}{n}} \qquad \textbf{3.44}}$$

Using the above example numbers.

Table 3.18

	Actual	Forecast	Error	Percent Error	Absolute Percent Error
i	Y_i	$\hat{Y}_i$	e_i	PE	$\lvert PE \rvert$
1	384.84	366.51	$+$ 18.33	+ 5%	5
2	410.49	366.51	$+$ 43.98	+11%	11
3	274.38	308.29	$-$ 33.91	-12%	12
4	292.88	308.29	$-$ 15.41	- 5%	5
5	311.53	366.51	$-$ 54.98	- 18%	18
6	366.87	308.29	$+$ 58.58	+16%	16
					67

$$\text{MAPE} = \frac{\sum |PE|}{6} = \frac{67}{6} = 11.17\%$$

MAPE, in this case, reveals that average percent error of forecast is about 11.2%. In other words, the six forecasts were within about 11.2% of the target, on the average.

PROBLEMS AND QUESTIONS

Stage 1 Collection and Evaluation of Data

3.1 Sales and Advertising Revisited.

The table below lists Sales Volumes and Advertising Expenditures for a corporation from ten randomly selected Locations.

Location i	Sales x 1,000 units	Advertising x \$10,000
1	101	1.2
2	92	0.8
3	110	1.0
4	120	1.3
5	90	0.7
6	93	1.0
7	82	0.8
8	75	0.6
9	91	0.9
10	105	1.1

Construct a scatter diagram of the data.

3.2 Sales and Price

The table below lists 10 randomly chosen weeks of orange juice sales and the corresponding price.

Week	Sales x 1,000 units	Price per unit
1	10	\$1.30
2	6	2.00
3	5	1.70
4	12	1.50
5	10	1.60
6	15	1.20
7	5	1.60
8	12	1.40
9	17	1.00
10	20	1.10

Construct a scatter diagram of the data.

3.3 Evaluation of Data: Simple Statistics

Refer to Problem 3.1. Determine the simple statistics of the data set.

a. Determine the Mean, Median, and Mode of X and of Y
b. Determine the Variance and Standard Deviation of X and of Y
c. Are there any unusual values (outliers) of X or of Y? Are there any unusual pairs (X, Y)?
d. Determine the Covariance and the Correlation between X and Y
e. Construct the Covariance Matrix and the Correlation Matrix between X and Y.

3.4 Evaluation of Data: Simple Statistics

Refer to Problem 3.2. Determine the simple statistics of the data set.

Repeat parts a-e as in Problem 3.3.

Stage 2 Identification and the Basic Model

3.5 Refer to Problem 3.3.
Consider the scatter diagram, if it appears <u>linear</u>, then sketch a straight line,“by sight," through the points.

3.6 Refer to Problem 3.2.
Consider the scatter diagram, if it appears <u>linear</u>, then sketch a straight line, “by sight," through the points.

3.7 Linear models
a. What is the difference between a deterministic mathematical model and a probabilistic mathematical model?
b. Explain the purpose of the random error term ϵ in a probabilistic mathematical model.

3.8 Graphing linear models
For each of the linear equations below, determine the slope and intercept, and graph each equation.
a. $Y = 3 + 2X$
b. $Y = 12.5 - 4X$
c. $Y = -.2X + 2$
d. $2X + 5Y = 30$

Stage 3 Estimation of Parameters

3.9 Refer to Problem 3.1.
a. Determine the Ordinary Least Squares (OLS) regression line of the data.
b. As a check on the calculations in part a., plot the ten points and graph the least squares regression line. Does the line appear to be a good fit to the data?
c. Refer to Problem 3.1. again. Compare the straight line drawn “by sight" with the graph of the least squares regression line in part b. above. Is your “by sight" line close to the the regression line?

3.10 Refer to Problem 3.2.
Repeat parts a-c as in Problem 3.9.

Stage 4 Diagnostics and Residual Analysis

3.11 Residuals and the MSE
Refer to Problem 3.9.
a. Using the OLS regression equation substitute the X values in to determine the corresponding fitted $\hat{Y}$ values. Find the difference between the actual Y values and the fitted $\hat{Y}$ values, the residuals.
b. Using the residuals determined in part a. find the MSE of the model.
c. Find the standard error of the model.

3.12 Refer to Problem 3.10 model. Repeat Problem 3.11 questions with the Problem 3.10 model.

3.13 Diagnostics of the Parameters.

Refer to Problem 3.9.

a. Determine the Variance and Standard Deviation of $\hat{\beta}_1$.

b. Determine a 95% confidence interval for $\hat{\beta}_1$.

c. Conduct the standard statistical test of $\hat{\beta}_1$.

3.14 Refer to Problem 3.10. Repeat Problem 3.13 questions with Problem 3.10.

3.15 Determining R^2.
Refer to Problem 3.9. Determine R^2 of the model.
Compare this with the correlation of the data derived in Problem 3.3.

3.16 Refer to Problem 3.10. Determine R^2 of the model.
Compare this with the correlation of the data derived in Problem 3.4.

3.17 The *ANOVA* table.
Construct an *ANOVA* table of the model from Problem 3.9.
Conduct the appropriate F test.

3.18 Construct an *ANOVA* table of the model from Problem 3.10.
Conduct the appropriate F test.

3.19 Summarizing the statistics of a model.
Present the summary statistics for the model of Problem 3.9 in the manner of page 44.

3.20 Present the summary statistics for the model of Problem 3.10 in the manner of page 44.

Stage 5 Forecasting and Confidence Intervals

3.21 Forecasting with SLR.

a. Forecast sales with the Problem 3.9 model if Advertising is set for $8,000.

b. Determine the 95% confidence interval for the forecast of the mean

c. Determine the 95% confidence interval for the forecast of the individual.

d. Repeat parts a-c with Advertising set at $11,000.

3.22 Forecasting with SLR.

a. Forecast sales with the Problem 3.10 model if Price is set at $1.35.

b. Determine the 95% confidence interval for the forecast of the mean

c. Determine the 95% confidence interval for the forecast of the individual.

d. Repeat parts a-c with Price set at $2.00.