CHAPTER 16

LINEAR REGRESSION TESTS OF PREDICTION

In the late 1950s and early 1960s, the mathematics related to solving a set of simultaneous linear equations was introduced to the field of statistics in the United States. In 1961, Franklin A. Graybill published a definitive text on the subject, *An Introduction to Linear Statistical Models*, which piqued the curiosity of several scholars. A few years later, in 1963, Robert A. Bottenberg and Joe H. Ward Jr., who worked in the Aerospace Medical Division at Lackland Air Force Base in Houston, Texas, applied the linear regression technique using basic algebra and the Pearson correlation coefficient. Norman R. Draper and Harry Smith Jr. published in 1966 one of the first books on the topic, *Applied Regression Analysis*. In 1967, under a funded project by the U.S. Department of Health, Education, and Welfare, W. L. Bashaw and Warren G. Findley invited several scholars to the University of Georgia for a symposium on the general linear model approach to the analysis of experimental data in educational research. The five speakers invited were Franklin A. Graybill, Joe H. Ward Jr., Ben J. Winer, Rolf E. Bargmann, and R. Darrell Bock. Dr. Graybill presented the theory behind statistics, Dr. Ward presented the regression models, Dr. Winer discussed the relationship between the general linear regression model and the analysis of variance, Dr. Bargmann presented applied examples that involved interaction and random effects, and Dr. Bock critiqued the concerns of the others and discussed computer programs that would compute the general linear model and analysis of variance. Since the 1960s, numerous textbooks and articles in professional journals have painstakingly demonstrated that the linear regression technique, presented by Bottenberg and Ward, is the same as the analysis of variance. In recent years, multiple regression techniques have proven to be more versatile than analysis of variance, hence the two methods today are combined into the general linear model framework (McNeil, Newman, & Fraas, 2012).

● GAUSS-MARKOV THEOREM

Linear regression is based on the Gauss-Markov theorem, which states that if the errors of prediction are independently distributed, sum to zero, and have constant variance, then the least squares estimation of the **regression weight** (b) is the best linear unbiased estimator of the population *B*. The least squares criterion is sometimes referred to as BLUE, or best linear unbiased estimator. Basically, of all unbiased estimators that are **linear functions** of the observations, the least squares estimation of the regression weight yields the smallest sampling variance or errors of prediction.

We will demonstrate linear regression in this chapter and apply the least squares estimation method to determine the regression weight in the equation. This approach will also be applied in the next chapter on multiple regression, termed ordinary least squares regression. The individual observation on Y_i is composed of a constant value assigned to all individuals (*a*), plus a common regression weight (*b*) applied to each individual corresponding X_i value, with a residual or error term (*e*). Each individual observation, Y_i, is the sum of *a* plus bX_i plus *e*, which is expressed in equation format as $Y_i = a + bX_i + e$. The selection of a value for *b* (regression weight) is done to minimize the error (*e*). The error is the difference between the original individual observation Y_i and what the linear regression equation would predict for each individual, denoted by $\hat{Y}_i$. The error term is therefore computed as $e = Y_i - \hat{Y}_i$. The Gauss-Markov theorem provides the rule that justifies the selection of a regression weight based on minimizing the error of prediction, which gives the best prediction of *Y*. We refer to this as the *least squares criterion*, that is, selecting regression weights based on minimizing the sum of squared errors of prediction.

● LINEAR REGRESSION EQUATION

The linear regression equation used by Bottenberg and Ward was expressed as

$$Y = a + bX + e.$$

The *Y* variable represents a continuous measure that was referred to as the *dependent variable*. The *X* variable represents a continuous measure that was called an independent variable but was later referred to as a *predictor variable*. The value *a* was termed the *intercept* and represented the value on the *y*-axis where the regression line crossed. The *b* value was a weight, later referred to as a *regression weight* or coefficient. The value *e* was referred to as *prediction error*, which is calculated as the difference between the *Y* variable and the predicted *Y* value $(\hat{Y})$ from the linear regression equation. The predicted *Y* value is computed based on the values of the intercept and regression weight. An example will illustrate the logic behind the linear regression equation.

Time Spent Studying (X)	*Exam Scores (Y)*
1	70
2	75
3	80
4	85
5	85
6	90
7	98

Given the data pairs (to the left) on the amount of time, in hours, spent studying (*X*) and the corresponding exam scores (*Y*), a linear regression equation can be created:

The regression intercept (*a*) indicates the point on the *y*-axis where a regression line crosses in a scatterplot. The regression weight (*b*) determines the rate of change (sometimes called *rise* and *run*), which can be seen by the slope of the regression line in the scatterplot. Given the correlation and standard deviation values for *X* and *Y*, the **intercept** and **slope** (regression weight) for the data can be calculated:

$$b = r_{XY}\frac{S_Y}{S_X}$$

$$a = \bar{Y} - b\bar{X}.$$

The prediction of *Y* given knowledge of *X* is expressed in a linear regression prediction equation as

$$\hat{Y} = a + bX.$$

Notice that the error (*e*) is not expressed in the linear regression prediction equation. We predict Y ($\hat{Y}$) given the correlation of *X* with *Y*, so $e = Y - \hat{Y}$, which occurs when the Pearson correlation is not +1.00 or −1.00. When Pearson $r = +1.00$ or Pearson $r = -1.00$, then no prediction error exists, that is, $Y = \hat{Y}$, and prediction error = 0.

TIP

- ✓ The linear regression equation only relates to the range of values for the pairs of *Y* and *X* scores used to calculate the slope or regression weight (*b*) and the intercept (*a*).
- ✓ $R = r$ (Pearson correlation) when using a single predictor—called multiple correlation coefficient.
- ✓ $R = r\,(Y,\ \hat{Y})$—correlation between *Y* and predicted *Y* values.
- ✓ R^2 = Multiple correlation coefficient—coefficient of determination squared.

NOTE: Do not confuse the use of *R* in linear regression with the R software notation.

The calculation of the regression weight (*b*) requires the computation of the Pearson correlation coefficient and the standard deviation of scores on *Y* and *X*. The calculation of the intercept (*a*) requires the use of the regression weight (*b*) and the mean values for the *Y* and *X* scores. The intercept calculation should give us a clue—that if we only have the *Y* scores and no knowledge of *X*, then our best prediction of *Y* is the mean of *Y*. It is only when we have additional variables that correlate with *Y* that we can be predict variability in the *Y* scores.

The linear regression equation calculations for number of hours spent studying (*X*) predicting the exam score (*Y*), given the Pearson correlation, means, and standard deviations of *X* and *Y*, are

$$rXY = +0.983$$

$$\bar{Y} = 83.286, \quad SD(Y) = 9.34$$

$$\bar{X} = 4, \quad SD(X) = 2.16$$

$$b = r_{XY}\frac{S_Y}{S_X} = +0.983\left(\frac{9.34}{2.16}\right) = 4.25$$

$$a = \bar{Y} - b\bar{X} = 83 - 4.25(4) = 83.286 - 17 = 66.286.$$

The linear regression prediction equation is then created as

$$\hat{Y} = 66.286 + 4.25(X).$$

The linear regression prediction equation contains error (e) because the correlation between X and Y is not perfect. We use the equation to predict a Y score given the value of X. For example,

$$\hat{Y} = 66.286 + 4.25(X)$$
$$\hat{Y} = 66.286 + 4.25(1)$$
$$\hat{Y} = 70.536.$$

This indicates that for 1 hour of study time, the predicted exam score would be 70.536. The actual Y score was 70, so the linear regression equation overpredicted; that is, error = −0.536. The error is determined by $Y - \hat{Y}$. I have placed the Y, predicted Y, and error value in the table below for each value of X.

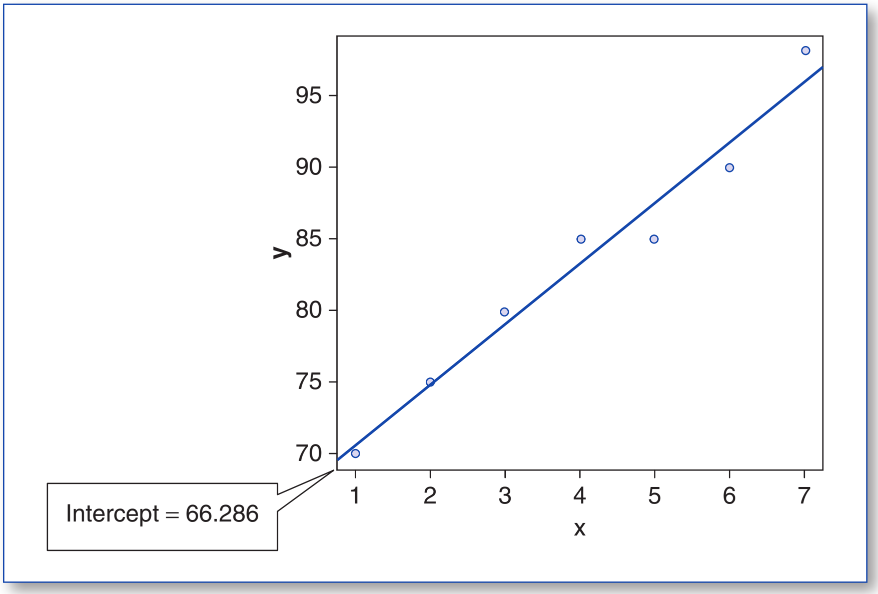
Y	$\hat{Y}$	*Error*
70	70.536	−0.536
75	74.786	0.214
80	79.036	0.964
85	83.286	1.714
85	87.536	−2.536
90	91.786	−1.786
98	96.036	1.964

Notice that the sum of the errors (*Error*) will always be 0. The reason is that some of the predictions are more than the actual Y value and some are less, but on average they will cancel out to zero.

A scatterplot of these data values would indicate the intercept and slope (rise and run) of this relationship, with the intercept intersecting the y-axis at a = 66.286. Perfect relationships between variables do not usually occur with real data, so the prediction of *Y* will generally involve some error of prediction. The predicted *Y* values form the regression line on the graph. The actual *Y* values fall near and far from this fitted regression line depending on the amount of error, or the difference between *Y* and the predicted *Y* value. The distance from the line for each *Y* therefore visually shows the amount of error in prediction.

We can see the linear relationship of *X* and *Y* by using a few R commands to create a scatterplot of the *X* and *Y* values with the linear regression line. After *X* and *Y* are placed in data vectors, the linear regression is computed, with the output saved in the file *model* using the *lm()* function. The *plot()* function graphs the *X* and *Y* values. The *abline()* function takes the *a* (intercept) and *b* (slope) values from the regression equation and uses them to draw a fitted regression line in the graph. The set of R commands are as follows:

```
> x = c(1,2,3,4,5,6,7)
> y = c(70,75,80,85,85,90,98)
> model = lm(y ~ x)
> plot(x,y)
> abline(model)
```



● LINEAR REGRESSION BY CALCULATOR

Another example will help demonstrate the linear regression prediction equation based on using summary values and a calculator. The data for 20 student math achievement scores (*Y*) and days absent from school (*X*) during the week are summarized below:

Student	X	Y	X^2	Y^2	XY
1	2	90	4	8,100	180
2	4	70	16	4,900	280
3	3	80	9	6,400	240
4	5	60	25	3,600	300
5	1	95	1	9,025	95
6	2	80	4	6,400	160
7	5	50	25	2,500	250
8	3	45	9	2,025	135
9	2	75	4	5,625	150
10	4	65	16	4,225	260
11	5	45	25	2,025	225
12	1	80	1	6,400	80
13	4	80	16	6,400	320
14	5	60	25	3,600	300
15	1	85	1	7,225	85
16	0	90	0	8,100	0
17	5	50	25	2,500	250
18	3	70	9	4,900	210
19	4	40	16	1,600	160
20	0	95	0	9,025	0
Σ	59	1,405	231	104,575	3,680

The summary statistics for these data can be calculated by hand as follows:

$$\overline{X} = \frac{\Sigma X}{N} = \frac{59}{20} = 2.95 \qquad S_X = \sqrt{\frac{SS_X}{N-1}} = \sqrt{\frac{56.95}{19}} = 1.73$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{1405}{20} = 70.25 \qquad S_Y = \sqrt{\frac{SS_Y}{N-1}} = \sqrt{\frac{5873.75}{19}} = 17.58$$

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{-464.75}{\sqrt{56.95(5873.75)}} = -.804.$$

TIP

✓ Recall from the previous chapter that the sum of products and sum of squares

$$SP = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N} = 3680 - \frac{(59)(1405)}{20} = -464.75$$

$$SS_Y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 104575 - \frac{(1405)^2}{20} = 5873.75$$

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 231 - \frac{(59)^2}{20} = 56.95$$

of X and sum of squares of Y were used in computing the correlation coefficient.

The intercept (a) and slope (b) in the linear regression prediction equation can now be computed:

$$b = r_{XY}\frac{S_Y}{S_X} = -0.804\left(\frac{17.58}{1.73}\right) = -8.16$$

$$a = \bar{Y} - b\bar{X} = 70.25 - [(-8.16)(2.95)] = 70.25 + 24.07 = 94.32.$$

The prediction of Y (math scores) given knowledge of X (days absent) is now possible, using the intercept and slope values, in the following linear regression prediction equation:

$$\hat{Y} = 94.32 + -8.16X.$$

To determine the predicted Y values, we would substitute each value of X into the linear regression equation. The resulting values for Y and $\hat{Y}$ and the errors of prediction are given below in a table format:

Student	X	Y	Ŷ	e
1	2	90	78	12
2	4	70	61.68	8.32
3	3	80	69.84	10.16
4	5	60	53.52	6.48
5	1	95	86.16	8.84
6	2	80	78.00	2.00

(Continued)

(Continued)

Student	X	Y	$\hat{Y}$	e
7	5	50	53.52	–3.52
8	3	45	69.84	–24.84
9	2	75	78.00	–3.00
10	4	65	61.68	3.32
11	5	45	53.52	–8.52
12	1	80	86.16	–6.16
13	4	80	94.32	18.32
14	5	60	53.52	6.48
15	1	85	86.16	–1.16
16	0	90	94.32	–4.32
17	5	50	53.52	–3.52
18	3	70	69.84	0.16
19	4	40	61.68	–21.68

NOTE: Errors of prediction can be positive or negative, but the sum (and mean) of the errors of prediction should be zero, $\sum e = 0$.

In this sample data set, the correlation coefficient is negative ($r = -.804$), which indicates that as the number of days absent during the week increases (*X*), the math achievement score (*Y*) decreases. We would conclude that absenteeism from school affects student test scores, which makes theoretical sense. This relationship would be depicted as a downward trend in the data points on a scatterplot. Also notice that the data points go together (covary) in a negative or inverse direction, as indicated by the negative sign for the sum of products in the numerator of the correlation coefficient formula.

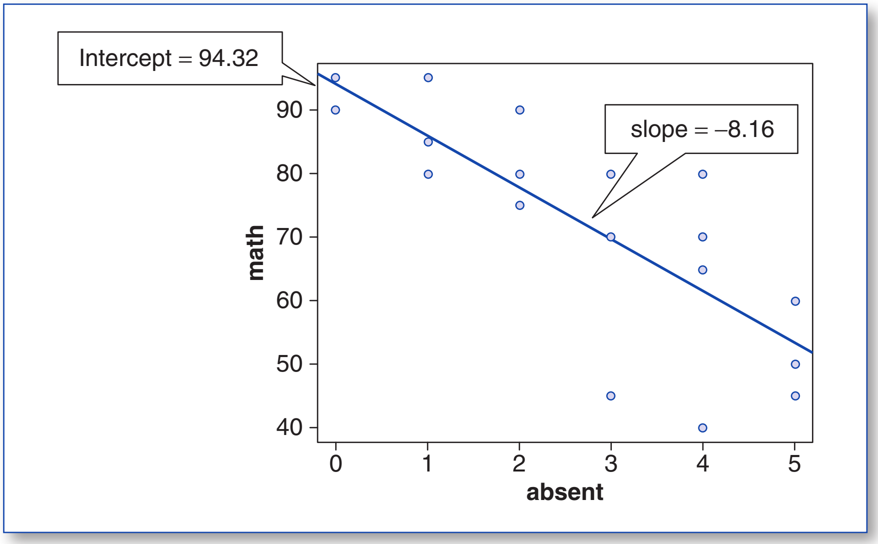
The set of R commands to analyze and verify our hand calculations would be as follows:

```
> absent = c(2,4,3,5,1,2,5,3,2,4,5,1,4,5,1,0,5,3,4,0)
> math = c(90,70,80,60,95,80,50,45,75,65,45,80,80,60,85,90,50,70,40,95)
> results = lm(math ~ absent)
> summary(results)

Call:
lm(formula = math ~ absent)

Residuals:
  Min       1Q     Median   3Q      Max
-24.842  -3.721   0.417    6.939  18.319
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    94.324    4.842     19.479 1.52e-13 ***
absent         -8.161    1.425     -5.727 1.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 18 degrees of freedom
Multiple R-squared: 0.6457, Adjusted R-squared: 0.626
F-statistic:  32.8 on 1 and 18 DF,  p-value: 1.979e-05
```

The intercept (94.32) and regression weight (–8.16) match our hand calculations. Next, use the following R commands to plot the X and Y values and include the regression prediction line:

```
> plot(absent, math)
> abline(results)
```



We square the correlation coefficient value to obtain a variance accounted for interpretation; that is, $r = -.804$, so the multiple R-squared (R^2) value = r^2 = .6457. Knowledge of the number of days absent accounts for 64.6% of the variance in the math achievement scores. The variance of the math achievement scores is $S_Y^2 = (17.58)^2 = 309.0564 \sim 309.06$, so the amount of variance explained is 0.646 * 309.0564 = 199.65. The amount of variance not explained is $(1 - r^2) \times S_Y^2 = 0.354 \times 309.0564 = 109.41$. The total S_Y^2 = %Explained + %Unexplained = 199.65 + 109.41 = 309.6. The errors of prediction help identify the accuracy of the regression equation. If the errors of prediction are small, the amount of variance unexplained is less; thus, more variance in *Y* is explained. Our results show 64.6% explained and 35.4% unexplained variance, most likely due to another predictor variable.

We expect the *Y* scores to be normally distributed around each predicted *Y* value for a given value of *X*, so that the variability of the errors of prediction indicates the standard deviation of the *Y* scores around the predicted *Y* value for each value of *X*. The standard deviation of the *Y* scores around each predicted *Y* value is called a *standard error of estimate*, or **standard error of prediction**. It is computed as

$$S_{Y.X} = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{2081.08}{18}} = 10.75.$$

Another approach, using the standard deviation of *Y*, the correlation coefficient, and the sample size, computes the standard error of estimate as follows:

$$S_{Y.X} = S_Y\sqrt{1-r^2}\sqrt{(n-1)/(n-2)}$$
$$S_{Y.X} = 17.58\sqrt{1-(-0.804)^2}\sqrt{(20-1)/(20-2)}$$
$$S_{Y.X} = 10.75.$$

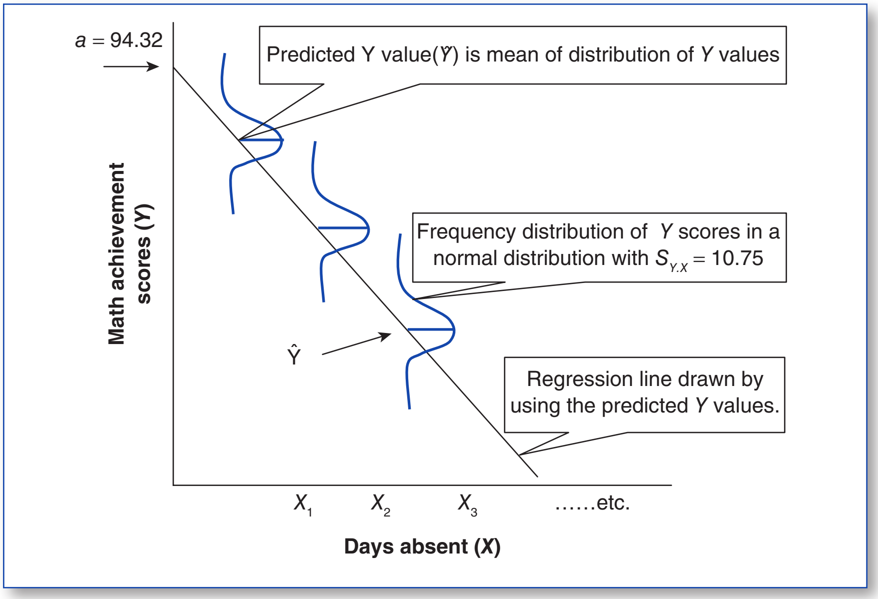
Note: The S_{YX} value is reported in the output as "Residual standard error: 10.75 on 18 degrees of freedom."

● GRAPH OF THE FITTED REGRESSION LINE

A graph of each frequency distribution of *Y* scores around each predicted *Y* value for each individual *X* score aids in the interpretation of the standard error of estimate. For each value of *X*, there is a distribution of *Y* scores around the predicted *Y* value. This standard deviation is assumed to be the same for each distribution of *Y* scores along the regression line, which is referred to as the **homoscedasticity of variance** along the regression line (equal variance of *Y* scores around a predicted *Y* value for each *X* score along the regression line). For the sample of data, $S_{Y.X}$ = 10.75, which is the standard deviation of the *Y* scores around the predicted *Y* value for each *X* score—assumed to be the same for each frequency distribution along the regression line—and is formed by

$$\hat{Y} = 94.32 + (-8.16)X.$$

The predicted *Y* value is the mean of the distribution of *Y* scores for each value of *X*. The predicted *Y* values are also used to draw the regression prediction line through the scatterplot of data points for *X* and *Y*. Since it is assumed that different values of *Y* vary in a normal distribution around the predicted *Y* values, the assumption of equal variance in *Y* across the regression line is important when using the regression prediction equation to predict *Y* for a given *X* value.


a = 94.32
Predicted Y value($\hat{Y}$) is mean of distribution of Y values
Math achievement scores (Y)
Frequency distribution of Y scores in a normal distribution with $S_{Y.X}$ = 10.75
$\hat{Y}$
Regression line drawn by using the predicted Y values.
X_1
X_2
X_3
……etc.
Days absent (X)

● LINEAR REGRESSION FUNCTION

The linear regression function in the script file (chap16a.r) involves taking a random sample of data and calculating a regression equation. The sample estimates for the intercept and slope are inferred to be the corresponding population parameters. A researcher typically does not know the true population parameters, which is why a random sample of data is used to provide estimates of the population parameters. The function permits a hypothetical comparison between a known population regression equation and the sample regression equation. You will need to enter the true values for the slope (*bTrue*) and intercept (*aTrue*) along with the sample size prior to running the function.

```
> bTrue = .25
> aTrue = 3
> sampleSize = 20
> chap16a(bTrue,aTrue,sampleSize)
```

The function will list the data points, the mean of X and Y, and the Pearson correlation with degrees of freedom and then compare the true regression equation (population) with the sample regression equation. The sample intercept and slope values are estimates of the population intercept and slope values. In practice, a researcher would not know the true population parameter values but would instead interpret the sample statistics as estimates of the population intercept and slope values.

```
Scatterplot Data Points

  X    Y
1.98 4.24
8.91 5.70
```

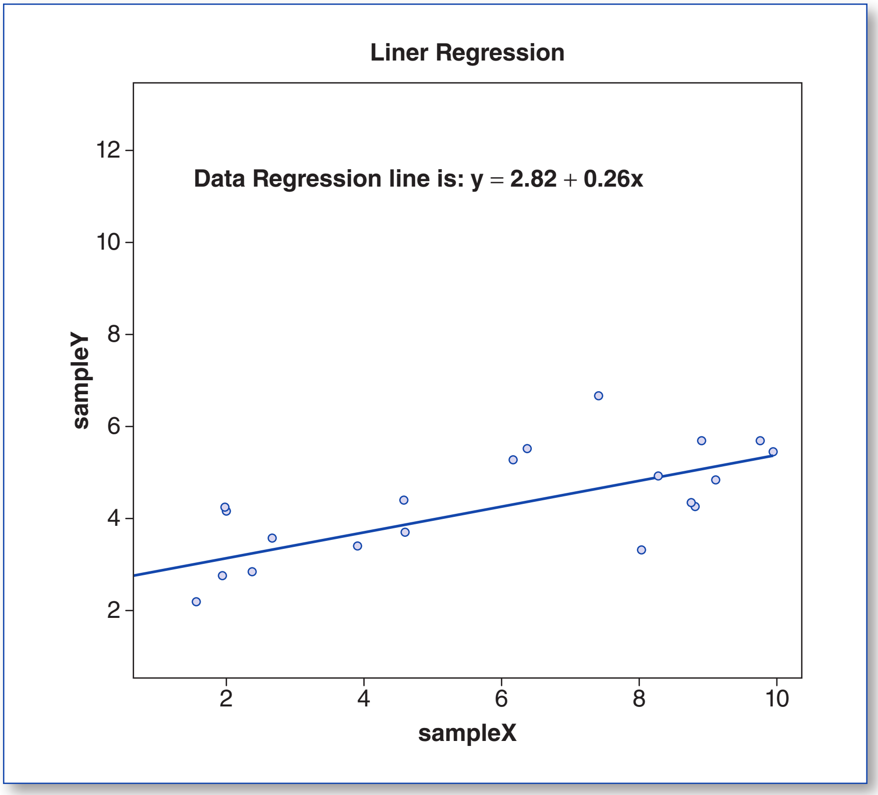
```
1.95 2.72
9.12 4.84
4.60 3.68
3.91 3.42
1.57 2.18
9.75 5.68
8.77 4.33
2.00 4.16
9.94 5.45
8.81 4.25
8.04 3.31
6.37 5.49
6.17 5.26
8.27 4.90
2.37 2.83
7.41 6.66
4.59 4.39
2.67 3.57

 Descriptive Statistics

   Mean   SD
X 5.860 3.037
Y 4.353 1.160

Pearson r =             0.684
Degrees of Freedom =  19.000

True Regression line is: y = 3 + 0.25x
Data Regression line is: y = 2.82 + 0.26x
```

Note: I used the *set.seed()* function so that you can replicate these results. Remove this function to obtain different results each time you run the function.

The *linear regression function* lists the sample data linear regression equation, while the output provides the true population values for the equation. This permits a comparison of how a random sample of data can provide a good estimate of population parameters.

● LINEAR REGRESSION WITH STANDARD SCORES

In some instances, the *Y* and *X* scores are converted to *z* scores or standard scores to place them both on the same measurement scale. This permits an equivalent interpretation of the change or



slope of z values for Y and X starting at 0. The z-score formula subtracts the mean from each score and divides by the standard deviation. The formula, you may recall, is

$$z_{X_i} = \frac{X_i - \overline{X}}{S_X}$$

and

$$z_{Y_i} = \frac{Y_i - \overline{Y}}{S_Y}.$$

The standard scores (z scores) can be converted back to their respective raw scores by computing

$$X_i = \overline{X} - S_X(z_{Xi})$$

and

$$Y_i = \overline{Y} - S_Y(z_{Yi}).$$

As a result of placing *Y* and *X* scores on the *z*-score scale, the intercept (*a*) and slope (*b*) in the linear regression equation are simplified because the mean values for *X* and *Y* are 0 and the standard deviations for *X* and *Y* are 1. This is indicated in their calculations:

$$a = \bar{Y} - b\bar{X} = (0\,) - b(0) = 0$$

and

$$b = r_{XY}\,\frac{S_Y}{S_X} = -0.804\left(\frac{1}{1}\right) = -0.804.$$

Because the mean and standard deviation of the *X* and *Y z* scores are 0 and 1, respectively, the regression line will pass through the origin of the scatterplot, where $Y = 0$ and $X = 0$, with the *Y* and *X* axes labeled in *z*-score units rather than raw score units.

The correlation coefficient captures the slope of the regression prediction line. The regression prediction equation in *z*-score form is

$$Z_{Y_i} = \beta Z_{X_i}$$

$$Z_{Y_i} = -0.804(Z_{X_i}),$$

where *b* (the raw score regression weight) is replaced with β (the standard score regression weight).

β will always equal the Pearson correlation coefficient in a single-predictor equation.

The use of linear regression in applied research is very popular. For example, admission to graduate school is based on the prediction of grade point average using the Graduate Record Exam score. Colleges and universities predict budgets and enrollment from one year to the next based on previous attendance data. The Pearson correlation coefficient plays an important role in making these predictions possible.

A statistically significant correlation between *Y* and *X* will generally indicate that a good prediction is possible, because the difference between the observed *Y* values and the predicted *Y* $(\hat{Y})$ values are kept to a minimum. The least squares line is fitted to the data to indicate the prediction trend. The least squares line is a unique fitted regression line based on minimizing the sum of the squared differences between the observed *Y*s and the predicted *Y*s, thus keeping prediction error to a minimum by the selection of values for the intercept (*a*) and slope (*b*). In the single-predictor regression formula that uses *z* scores, we see the unique role that the Pearson correlation coefficient plays as the slope value. A researcher should report the unstandardized linear regression values for the intercept (*a*) and slope (*b*) that we previously calculated, as well as the standardized regression weight (β).

● R FUNCTIONS FOR LINEAR REGRESSION

The *stats* package contains two different functions that can be used to estimate the intercept and slope in the linear regression equation. The two different R functions are *lm()* and *lsfit()*. The *lm()* function is preferred over the *lsfit()* function. The *lm()* function uses a data frame, whereas the *lsfit()* function uses a matrix or data vector. Also, the *lm()* function outputs an intercept term, which has meaning when interpreting results in linear regression. The linear regression command will therefore be of the form

```
> sampleReg = lm(sampleY ~ sampleX, data = out1)
```

The *lm()* function can also specify an equation with *no intercept* of the form

```
> sampleReg = lm(sampleY ~ 0 + sampleX, data = out1)
```

or

```
> sampleReg = lm(sampleY ~ sampleX - 1, data = out1)
```

The *lm()* function with *X* and *Y* data will be used in a single-predictor regression equation with and without an intercept term in the next two sections.

Linear Regression With Intercept

The *summary()* function returns the results of the linear regression equation. The intercept value of 94.32 is reported in the output.

```
> library(stats)
> sampleX = c(2,4,3,5,1,2,5,3,2,4,5,1,4,5,1,0,5,3,4,0)
> sampleY = c(90,70,80,60,95,80,50,45,75,65,45,80,80,60,85,90,50,70,40,95)
> sampleReg = lm(sampleY ~ sampleX)
> summary(sampleReg)

Call:
lm(formula = sampleY ~ sampleX)

Residuals:
   Min      1Q     Median    3Q      Max
-24.842  -3.721   0.417    6.939   18.319

Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept)   94.324      4.842    19.479 1.52e-13 ***
sampleX       -8.161      1.425    -5.727 1.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 18 degrees of freedom
Multiple R-squared: 0.6457, Adjusted R-squared: 0.626
F-statistic:  32.8 on 1 and 18 DF,  p-value: 1.979e-05
```

Linear Regression Without Intercept

The *lm()* function now contains the prediction equation with 0 as the intercept term. The *summary()* function now outputs results that do not contain the intercept value. Notice that the regression weight does not equal the correlation coefficient (–.8035), which can be obtained from the *cor(sampleX, sampleY)* command. You would need to use *z* scores for the *X* and *Y* values to achieve this standardized linear regression solution.

```
> library(stats)
> sampleX = c(2,4,3,5,1,2,5,3,2,4,5,1,4,5,1,0,5,3,4,0)
> sampleY = c(90,70,80,60,95,80,50,45,75,65,45,80,80,60,85,90,50,70,40,95)
> sampleReg = lm(sampleY ~ 0 + sampleX)
> summary(sampleReg)
```

These results would not be correct because standardized values for *X* and *Y* should be used.

```
Call:
lm(formula = sampleY ~ 0 + sampleX)

Residuals:
   Min    1Q     Median   3Q     Max
-34.65 -19.65   19.24   59.62   95.00

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
sampleX   15.931     3.236    4.924   9.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.18 on 19 degrees of freedom
Multiple R-squared: 0.5606, Adjusted R-squared: 0.5375
F-statistic: 24.24 on 1 and 19 DF,  p-value: 9.433e-05
```

The linear regression analysis with an intercept matches the previous example. The linear regression analysis without an intercept term does not yield the correct results. We would expect a standardized beta weight to equal the Pearson correlation coefficient in a single-predictor case. Another example will further demonstrate the use of the *lm()* linear regression function.

Linear Regression Example

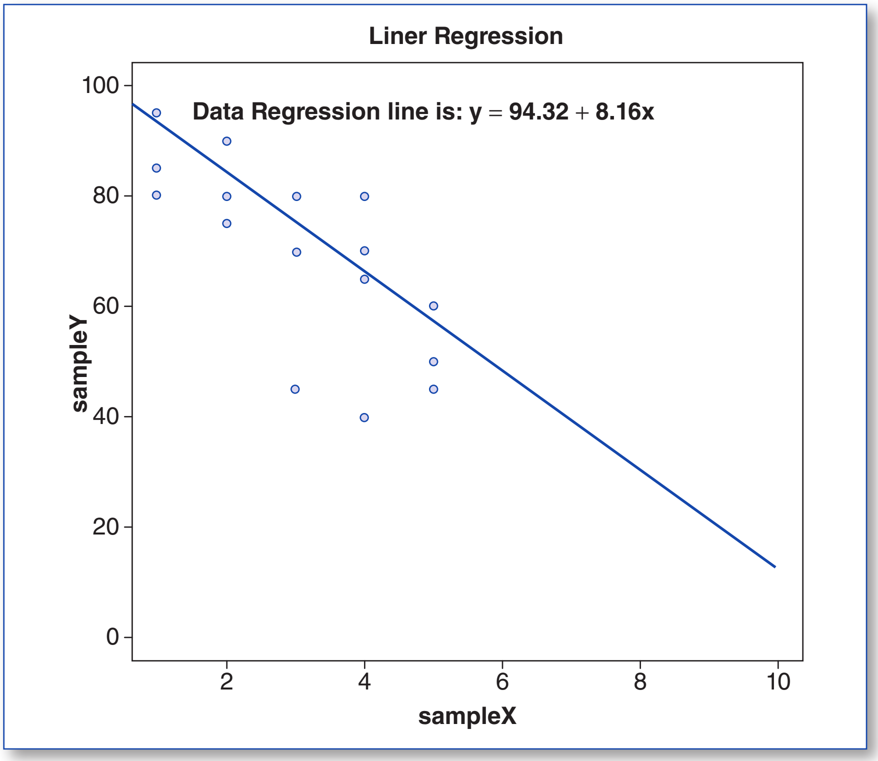
The *linear regression example function* in the script file (chap16b.r) uses the 20 student math achievement scores (*Y*) and the number of days absent from school during the week (*X*). The function will list the *X* and *Y* values, calculate the descriptive statistics (mean and standard deviation), calculate the Pearson *r* and degrees of freedom, and finally list the linear regression equation with the intercept and slope values. The results should be exactly the same as those calculated before by hand. You will need to enter the data vectors *sampleX* and *sampleY* prior to running the function.

```
> sampleX = c(2,4,3,5,1,2,5,3,2,4,5,1,4,5,1,0,5,3,4,0)
> sampleY = c(90,70,80,60,95,80,50,45,75,65,45,80,80,60,85,90,50,70,40,95)
> chap16b(sampleX,sampleY)
```

The results list the pairs of *X* and *Y* data values followed by the descriptive statistics and Pearson *r* value. The linear regression equation is then printed with the intercept and slope value. Finally, a scatterplot of the *X* and *Y* data values with the regression line is shown.

```
PROGRAM OUTPUT

Scatterplot Data Points

 X Y
 2 90
 4 70
 3 80
 5 60
 1 95
 2 80
 5 50
 3 45
 2 75
 4 65
 5 45
 1 80
 4 80
 5 60
 1 85
 0 90
 5 50
 3 70
 4 40
 0 95

Descriptive Statistics

   Mean     SD
X  2.95  1.731
Y 70.25 17.583


Pearson r =            -0.804
Degrees of Freedom =  19.000

Data Regression line is: y = 94.32 + -8.16x
```

The scatterplot shows the regression line crossing at 94.32 (intercept), then descending in a negative trend at –8.16 (slope).



● INTERPRETATION OF LINEAR REGRESSION OUTPUT

The *summary()* function can be used after obtaining results from the *lm()* function to output more diagnostic information. This is done by including the *lm()* function results in the *summary()* function. For example, add these two command lines to the above function in the Output section:

```
> data4 = summary(sampleReg)
> print(data4)
```

The resulting output would be

```
Call:
lm(formula = sampleY ~ sampleX, data = out1)

Residuals:
    Min      1Q  Median    3Q      Max
-24.842  -3.721  0.417  6.939  18.319
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     94.324      4.842   19.479 1.52e-13 ***
sampleX         -8.161      1.425   -5.727 1.98e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.75 on 18 degrees of freedom
Multiple R-squared: 0.6457, Adjusted R-squared: 0.626
F-statistic:  32.8 on 1 and 18 DF,  p-value: 1.979e-05
```

The coefficients are in raw score form, which shows the intercept, a = 94.324, and regression weight, b = –8.161. In addition, the standard error (*SE*) is displayed, which is used to calculate a t test of statistical significance. For the intercept, the t test would be computed as

$$t = \frac{\text{Coefficient}}{\text{Error}} = \frac{b_{\text{int}}}{SE_{b_{\text{int}}}} = \frac{94.324}{4.842} = 19.48.$$

For the regression weight, the t test would be computed as

$$t = \frac{\text{Coefficient}}{\text{Error}} = \frac{b_x}{SE_{b_x}} = \frac{-8.161}{1.425} = -5.727.$$

The intercept and slope are both statistically significant, as indicated by their respective p values, which are given in scientific notation. This was corrected by the command *options(scipen = 999)* in the function. The regression weight, for example, is statistically significant beyond a .05 level of probability at p = .0000198—not 1.98e–05 (scientific notation).

The *multiple R-squared* value is .6457. This indicates the amount of variance explained in Y given knowledge of X. In a single-predictor linear regression equation, this would be the same value as squaring the Pearson r coefficient between X and Y—that is,

$$R^2 = r^2$$
$$R^2 = (-.804)^2$$
$$R^2 = .646.$$

We can test the multiple R-squared value from the regression equation for statistical significance by using the F test. The F test was designed to test the ratio of two variances; in this case, it is the ratio of explained variance to unexplained variance in the regression equation. The F test is computed as follows:

$$F = \frac{R^2 / (k)}{1 - R^2 / (N - k - 1)}$$
$$F = \frac{.6457 / (1)}{.3543 / (20 - 1 - 1)}$$
$$F = 32.8.$$

where N = sample size and k = number of predictor variables. The number of predictor variables (k) is the degrees of freedom in the numerator of the F formula; that is, $df_1 = 1$. The expression ($N - k - 1$) is the degrees of freedom in the denominator of the F formula, which is $df_2 = (20 - 1 - 1) = 18$. The numerator and denominator degrees of freedom are used in Table 5 (.05 level of significance) or Table 6 (.01 level of significance) in Appendix A to obtain the respective F value we would expect by chance in the sampling distribution of F values. For the .05 level of probability, $F_{.05;df=1,18} = 4.41$, and for the .01 level of probability, $F_{.01;df=1,18} = 8.29$. The F value from our regression equation ($F = 32.8$) is greater than either F value at the .05 or .01 level of probability, which is reported as $p = .00001978$. The *multiple R-squared* value is therefore statistically significant, which means that the amount of variance in Y predicted by X is statistically significant.

We also know from the *multiple R-squared* value the amount of variance in Y that is *not* explained, that is, $1 - R^2$, or $1 - .646 = .354$, or 35% unexplained variation in Y scores. Recall that the amount of variance explained plus the amount of variance not explained equals the total amount of variance in Y ($S_Y^2 = 309.161$). So 100% of variance = 65% explained + 35% not explained (rounded up). In practical terms, we partition the actual variance in Y scores, as follows:

$$S_r^2 = (17.583)^2 = 309.161$$

$$S_{r\,\text{Explained}}^2 = .65(309.161) = 200.955$$

$$S_{r\,\text{Not explained}}^2 = .35(309.161) = 108.206.$$

A researcher would typically reduce the amount of variance *not* explained (108.206) by including additional predictor variables in the regression equation. We would want any additional predictor variables to be significantly correlated with Y but not have high correlation with the other predictor variables.

The *adjusted R-squared* value is less than the *multiple R-squared* value because it makes an adjustment for both the number of predictors and small sample sizes. The multiple R-squared value can become inflated or spuriously large when the number of predictors is similar to the sample size. The adjusted R-squared value makes an adjustment for the number of predictors in the regression equation, especially with small sample sizes. The formula is as follows:

$$\text{Adj}R^2 = 1 - [(1 - R^2)\left(\frac{N-1}{N-k-1}\right)]$$

$$\text{Adj}R^2 = 1 - [(.3543)(1.055)]$$

$$\text{Adj}R^2 = 1 - [.3737]$$

$$\text{Adj}R^2 = .626.$$

TIP

✓ *Multiple R* is the correlation between the Y scores and the predicted Y scores ($\hat{Y}$) from the regression equation. The square of multiple R is the *multiple R-squared* value (R^2).

$$\text{Multiple } R = -.804$$

$$\text{Multiple } R^2 = .646.$$

✓ The *adjusted R-squared* value makes an adjustment for the number of predictors in the equation, especially with small sample sizes.

$$\text{Adjusted } R^2 = .626.$$

Note: You would have to run the *cor.test()* function to test the statistical significance of the correlation coefficient—and thus a test of the regression weight being statistically significant in linear regression with a single predictor. These R commands would be as follows:

```
    > sampleX = c(2,4,3,5,1,2,5,3,2,4,5,1,4,5,1,0,5,3,4,0)
    > sampleY = c(90,70,80,60,95,80,50,45,75,65,45,80,80,60,85,90,50,70,40,95)
    > cor.test(sampleX, sampleY,alternative = "two.sided", method = "pearson",
conf.level = 0.95)

    Pearson's product-moment correlation

    data:  sampleX and sampleY
    t = -5.7275, df = 18, p-value = 0.00001979

    alternative hypothesis: true correlation is not equal to 0

    95 percent confidence interval:
     -0.9192124 -0.5602524

    sample estimates:
          cor
    -0.8035535
```

● HYPOTHESIS TESTING IN LINEAR REGRESSION

We use the same five-step hypothesis-testing approach to outline our test of a null hypothesis in linear regression. An example will illustrate the five-step hypothesis-testing approach.

Step 1. State the research question.

Can I statistically significantly positively predict IQ scores given knowledge of Reading scores?

Step 2. State the null and alternative statistical hypotheses.

$$H_0\text{: } \beta_{\text{IQ,READ}} = 0$$

$$H_A\text{: } \beta_{\text{IQ,READ}} \geq .389$$

The *r* value is from the table of critical values at the one-tailed test, alpha = .05 level

IQ	*Read*
118.00	66.00
99.00	50.00
118.00	73.00
121.00	69.00
123.00	72.00
98.00	54.00
131.00	74.00
121.00	70.00
108.00	65.00
111.00	62.00
118.00	65.00
112.00	63.00
113.00	67.00
111.00	59.00
106.00	60.00
102.00	59.00
113.00	70.00
101.00	57.00

Step 3. State the region of rejection, alpha level, direction of the hypothesis, and sample size.

Sample size = 18, $\alpha = .05$, one-tailed test, $df = 17$.

Region of rejection: $\beta_{\text{IQ,READ}} \geq .389$.

Step 4. Collect data, and compute the sample coefficient, $\beta_{\text{IQ,Read}}$, and summary statistics.

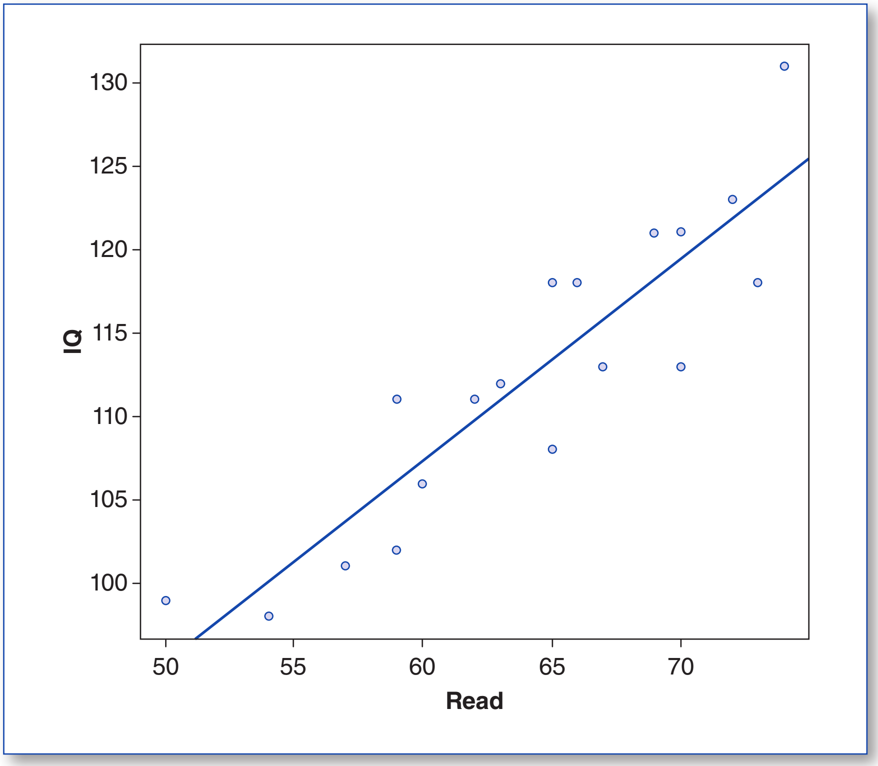
The set of R commands are in the Hypothesis Testing Linear Regression Example script file (chap16c.r), which are as follows:

```
   > IQ = c(118,99,118,121,123,98,131,121,1
08,111,118,112,113,111,106,102,113,101)
   > Read = c(66,50,73,69,72,54,74,70,65,62
,65,63,67,59,60,59,70,57)
   > model = lm(IQ ~ Read)
   > mean(IQ);sd(IQ)
   > mean(Read);sd(Read)
   > summary(model)
   > plot(Read, IQ)
   > abline(model)
```

The summary statistics, linear regression results, and scatterplot are output as follows:

```
> mean(IQ);sd(IQ)
[1] 112.4444
[1] 9.043829

> mean(Read);sd(Read)
[1] 64.16667
[1] 6.741007

> cor(IQ, Read)
[1] 0.8999127

> summary(model)

Call:
lm(formula = IQ ~ Read)

Residuals:
   Min     1Q   Median   3Q      Max
-6.487 -2.847 1.031  3.187  6.683
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.9738     9.4338    3.707  0.00191 **
Read           1.2073     0.1463    8.255  3.68e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.065 on 16 degrees of freedom
Multiple R-squared: 0.8098, Adjusted R-squared: 0.798
F-statistic: 68.14 on 1 and 16 DF,  p-value: 3.684e-07
```



Step 5. State the conclusion, and interpret the results.

The IQ mean = 112.44, *SD* = 9.04; the Read mean = 64.16, *SD* = 6.74. The Pearson $r = \beta_{\text{IQ,Read}}$, which is reported as .8999 (rounded to .90). Since our sample β = .90 is greater than the tabled β = .389, we reject the null and accept the alternative hypothesis at the .05 level of significance. The multiple *R*-squared value indicated that .8098 or 81% of the variance in IQ scores is explained or predicted by knowledge of Reading scores. $F = 68.14$, $df = 1, 16$ was statistically significant at $p = 3.684\text{e–}07 = .0000003684$. $t = (1.2073/0.1463) = 8.255$ is greater than the tabled $t = 1.74$ at the .05 level, one-tailed level of significance, so the b_{Read} parameter (unstandardized estimate) was statistically significant. A scatterplot with the fitted regression line for the equation IQ = 34.9738 + 1.2073 (Reading) is provided by the *plot()* and *abline()* functions. It visually shows that as Reading scores increase, IQ scores increase.

Our summary conclusions are based on comparing the sample regression weight with what is expected by chance given our degrees of freedom and alpha level of significance. We reject the null hypothesis and accept the alternative hypothesis based on this comparison at a specific level of probability. We are also able to test the statistical significance of the unstandardized regression estimate (b) using a t test. The multiple R-squared value is further tested for statistical significance using the F test. In addition, we are able to interpret the multiple R-squared value as the amount of variance explained in IQ scores given knowledge of Reading scores. Therefore, 81% of $(SD = 9.04)^2$ is explained by knowledge of Reading scores, which is .81(81.7216) = 66.194496, or 66.19. The amount of unexplained variance is $(1 - R^2)(81.7216) = 0.19(81.7216) = 15.527104$, or 15.51. The total variance in IQ scores, $(9.04)^2$, is equal to 66.19 (explained variance) + 15.51 (unexplained variance). It may be possible to further explain the variance in IQ scores, and thus reduce the amount of unexplained variance, by hypothesizing another predictor variable. Basically, having more than one predictor variable could yield a better prediction equation, and thus explain more of the variance in the Y scores.

We can also report the power for these results, given that $R^2 = .8098$ is a measure of effect size. The *pwr.f2.test()* function in the *pwr* package provides power calculations for the general linear model. In our example, the R command would be

```
> library(pwr)
> ?pwr.f2.test
pwr.f2.test(u = NULL, v = NULL, f2 = NULL, sig.level = 0.05, power = NULL)
```

Arguments

u	degrees of freedom for numerator
v	degrees of freedom for denominator
f2	effect size
sig.level	Significance level (Type I error probability)
Power	Power of test (1 minus Type II error probability)

```
> pwr.f2.test (u = 1, v = 16, .8098, .05, power = NULL)

Multiple regression power calculation

u = 1
v = 16
f2 = 0.8098
sig.level = 0.05
power = 0.9474134
```

The results indicate sufficient power for the effect size; that is, power = .947 for an effect size of $R^2 = .8098$.

● LINEAR REGRESSION AND ANALYSIS OF VARIANCE

When the intercept term is removed from the regression equation, it provides a comparison with analysis of variance. The regression formula has the form $y \sim x - 1$ or $y \sim 0 + x$, where y is the dependent variable and x, the independent variable. The tilde (~) sign is used in the regression formula to indicate that y is regressed on x. The 0 or −1 specifies that there is no intercept. We can show the differences and similarity between linear regression and analysis of variance by running the *aov()* and *lm()* functions, with associated output from the *summary()* function.

```
> sampleX = c(2,4,3,5,1,2,5,3,2,4,5,1,4,5,1,0,5,3,4,0)
> sampleY = c(90,70,80,60,95,80,50,45,75,65,45,80,80,60,85,90,50,70,40,95)
> sampleaov = aov(sampleY ~ 0 + sampleX)
> summary(sampleaov)
> samplereg = lm(sampleY ~ sampleX)
> summary(samplereg)

Analysis of Variance Output

sampleX     1   58625    58625   24.241 9.433e-05 ***
Residuals 19   45950     2418
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Linear Regression Output

Call:
lm(formula = sampleY ~ 0 + sampleX)

Residuals:
   Min    1Q   Median     3Q      Max
-34.65 -19.65   19.24   59.62   95.00

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
sampleX    15.931     3.236     4.924   9.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.18 on 19 degrees of freedom
Multiple R-squared: 0.5606, Adjusted R-squared: 0.5375
F-statistic: 24.24 on 1 and 19 DF,  p-value: 9.433e-05
```

The multiple *R*-squared value of .5606 can be computed from the analysis of variance sum of squares value in the summary table. Recall that $SS_T = SS_{\text{Regression}} + SS_{\text{Residual}}$ (104575 = 58625 + 45950). The multiple *R*-squared value due to regression is therefore computed as $SS_{\text{Regression}}/SS_T$ = 58625/104575 = .5606.

The multiple *R*-squared value is the amount of variance explained due to regressing *Y* on *X*. Therefore, $R^2 = 58625/104575 = .5606$. We also know that $1 - R^2 = 45950/104575 = .4394$. Thus, $R^2 + (1 - R^2) = 1$ (100% variance in *Y*), which is also expressed as $SS_Y = .5606 + .4394$. These two parts are interpreted as the amount of variance explained (R^2) and the amount of variance unexplained ($1 - R^2$), which can be represented in a Venn diagram. Notice that the *F* test, *df*, and *p* values are identical for both types of data analysis. Today, many statistical packages are dropping the separate analysis of variance and multiple regression routines in favor of the general linear model; for example, the IBM SPSS version 21 outputs both results.

SUMMARY

A brief history of multiple regression helps our understanding of this popular statistical method. It expands the early use of analysis of variance, especially analysis of covariance, into what is now called the general linear model. Over the years, researchers have come to learn that multiple regression yields similar results as analysis of variance but also provides many more capabilities in the analysis of data from research designs.

This chapter provided the basic understanding of linear regression. The following concepts were presented:

- The *a* in the regression equation is the intercept of the least squares line.
- The *b* coefficient in the regression equation is the slope of the least squares line.
- The intercept in the regression equation is called the *y*-intercept, the point at which the least squares line crosses the *y*-axis.
- In the linear regression equation, *X* is the independent variable and *Y* is the dependent variable.
- The linear regression equation using *z* scores for *X* and *Y* has a slope equal to the Pearson correlation coefficient.
- The intercept and slope of the linear regression prediction line from sample data are estimates of the population intercept and slope, respectively.
- The purpose of linear regression is to predict *Y* from knowledge of *X* using a least squares criterion to select an intercept and slope that will minimize the difference between *Y* and predicted *Y*, that is, error of prediction.

This chapter presented the basic linear regression equation for predicting *Y* from knowledge of *X* using the *lm()* function, with output provided by the *summary()* function. It is always recommended that a scatterplot of data be viewed to examine the relationship between *Y* and *X* using the *plot()* function. The basic algebra indicates that $Y = a + bX + e$, where *a* is an intercept term or the point where the fitted regression line crosses the *y*-axis; *b* is a regression weight determined by minimizing *e*, the error term; and *e* is the difference between *Y* and the predicted *Y* value. The process of estimating a regression weight (*b*) that minimizes the error is called the *least squares criterion*. The predicted *Y* values are computed using the *fitted()* function. A standard error of prediction is used to form confidence intervals around the predicted *Y* values along the fitted regression line, which indicates the distribution of *Y* values for each *X* value. The *e* values are computed by subtracting the predicted *Y* values from the *Y* values. The more *X* and *Y* are related or correlated, the better the prediction, and thus the less the error.

We learned that the regression weight should be reported in unstandardized and standardized formats. The *scale()* function creates variables in a standardized format, that is, mean = 0 and standard deviation = 1. A *t* test for the significance of the regression coefficient is computed using the unstandardized regression weight and associated standard error. The *R*-squared value is tested for statistical significance using the *F* test. An adjusted *R*-squared value was presented, which adjusts R^2 for the number of predictors and small sample sizes.

Hypothesis testing is conducted as with the other statistics. A null and an alternative hypothesis are stated, with the regression weight, *R*-squared value, or *F*-test value for a given degree of freedom and sample size. The other remaining hypothesis steps are followed to make a final decision regarding rejection or retention of the null hypothesis. The researcher then interprets the statistical significance of the regression weight, the *R*-squared value, and the *F* test to determine whether *X* predicts *Y*. The *aov()* and *lm()* functions provide the statistical results for analysis of variance and linear regression, respectively, which permits a comparison and helpful interpretation of the results.

TIP

✓ A list of statistical functions can be obtained by issuing the following command:

```
> library(help = “stats”)
```

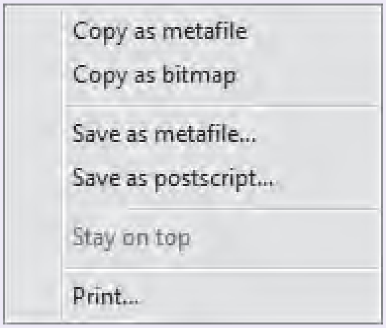
✓ The following R commands provide the function arguments and online documentation for the linear regression function:

```
> args(lm)

> help(lm)
```

✓ You can omit missing data that affect Pearson *r* and regression by using the following R command:

```
> data.new = na.omit(data.old)
```

✓ Right click your mouse on the scatterplot to save the image to the clipboard or print. The selection menu (to the right) will appear in the scatterplot:



✓ You can copy and paste scatterplots into Word documents by using the Alt+PrtScr keyboard keys; then in the Word document, use the Ctrl+V keyboard keys to paste the image. You can resize the image in the Word document by selecting the Format tab, which appears when the image is selected. The Crop tool option will permit easy resizing of the image.

EXERCISES

1. Briefly explain the regression terms in the equation $Y = a + bX + e$.
2. Given the following summary statistics for Y and X, calculate by hand the intercept and slope, then write out the regression equation. Show the formula and your work with four decimal places.

Variable	*Mean*	SD	*Correlation*
Y	6.6	2.702	.948
X	14.8	3.962	

3. Given Y and X below, use the regression equation to compute the predicted Y values and the prediction error values. Show your work with four decimal places. Enter the predicted Y values and their prediction error values in the table.

$$\hat{Y} = -2.9682 + .6465(X).$$

a. Is Y = Predicted Y + Prediction error?

b. Does the sum of the prediction error values equal 0?

Y	X	*Predicted Y*	*Prediction Error*
10	20		
8	15		
7	17		
5	12		
3	10		

4. Use the *lm()* and *abline()* functions to compute values, then plot the fitted regression line for the Y and X values in #3 above. Show the R commands and plot (use the Alt+PrtScr keyboard keys to copy and paste the plot).
5. Given the following data vectors for Y and X, use the *lm()* and *summary()* functions to compute the regression equation and descriptive output. Show the R commands and output.

```
y = c(10,8,7,5,3)
x = c(20,15,17,12,10).
```

R commands:

Output:

Interpretation:

a. Is the regression weight for X statistically significant at the p = .05 level of probability?

b. Is $F = t^2$?

c. What does the multiple R^2 value imply?

TRUE OR FALSE QUESTIONS

T F a. The Y mean is our best prediction given no knowledge of the X predictor variable.

T F b. The sum of the prediction errors will always equal 0.

T F c. The standard deviation of the Y scores around each predicted Y score is called the standard error of estimate or standard error of prediction.

T F d. The intercept value = 0 when using z scores for Y and X.

T F e. Pearson correlations in a matrix are in a nonstandardized (non–z score) format.

WEB RESOURCES

Chapter R script files are available at http://www.sagepub.com/schumacker

Hypothesis Testing Linear Regression Example R script file: chap16c.r

Linear Regression Example Function R script file: chap16b.r

Linear Regression Function R script file: chap16a.r