Chapter 13
Linear Regression

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 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, developed the linear regression technique using basic algebra and the Pearson correlation coefficient. Norman R. Draper and Harry Smith, Jr. in 1966 published 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 invited speakers 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 which 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 in handling nominal and ordinal data, interaction effects, and non-linear effects.

The linear regression equation developed by Bottenberg and Ward was expressed as: $Y = a + bX + e$. The Y variable represented a continuous measure, which was referred to as the dependent variable. The X variable represented a continuous measure, which was called an independent variable, but later referred to as a predictor variable. The value *a* was termed the "intercept" and represented the value on the Y-axis where the least squares line crossed. The *b* value was a "weight," later

R. Schumacker and S. Tomek, *Understanding Statistics Using R*,
DOI 10.1007/978-1-4614-6227-9_13,

referred to as a regression weight or coefficient. The value of *e* was referred to as prediction error, which is calculated as the difference between the Y variable and the predicted Y value (Yhat) from the linear regression equation, given values for the intercept and regression weight. An example will illustrate the logic behind the linear regression equation.

Regression Equation

Given the following data pairs on the amount of recyclable aluminum in ounces (Y) and the number of aluminum cans (X), a linear regression equation can be created: Y = a + bX + e.

Recyclable Aluminum (Y)	Number of Aluminum Cans (X)
1	2
2	4
3	6
4	8
5	10
6	12
7	14

The regression intercept (*a*) indicates the point on the Y-axis where the least squares line crosses in the scatter plot. The "rise" and "run" or regression weight (*b*) determines the rate of change, which can be seen by the slope of the least squares line in the scatter plot. It is important to understand that a linear regression equation only refers to the range of values for the pairs of Y and X scores. Given the linear regression equation: Y = a + bX + e, the intercept and regression weight (slope) for the data can be calculated as:

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

$$a = \overline{Y} - b\overline{X}$$

The correlation coefficient for these data is r = +1.0, the mean of Y = 4 and the mean of X = 8. The standard deviation of Y values is 2, and the standard deviation of X values is 4. Placing these values in the intercept and regression weight formula results in: $a = 2 - (2/4)\ 4 = 0$ and $b = 1\ (2/4) = .5$, with a linear regression equation: Y = 0 + (1/2) X. Since the intercept is zero, the equation is simply Y = .5X. An inspection of the data reveals that 2 aluminum cans yields 1 ounce of recyclable aluminum, 4 aluminum cans yields 2 ounces of recyclable aluminum, and so forth because one-half the number of aluminum cans equals the number of ounces. A scatter plot of these data would indicate the "rise" and "run" of this relationship with the least squares line intersecting the Y-axis at a = 0. Notice that there is no error in the prediction since every Y value is perfectly predicted by knowledge of X, i.e., *e* = Y − Yhat = 0. Perfect relationships like this don't often occur with real data!

Regression Line and Errors of Prediction

A more realistic example will help to demonstrate the linear regression equation, least squares line, and error of prediction. The data for twenty student math achievement scores (Y) and days absent during the week from school (X) are summarized below.

Student	X	Y	X^2	Y^2	XY
1	2	90	4	8100	180
2	4	70	16	4900	280
3	3	80	9	6400	240
4	5	60	25	3600	300
5	1	95	1	9025	95
6	2	80	4	6400	160
7	5	50	25	2500	250
8	3	45	9	2025	135
9	2	75	4	5625	150
10	4	65	16	4225	260
11	5	45	25	2025	225
12	1	80	1	6400	80
13	4	80	16	6400	320
14	5	60	25	3600	300
15	1	85	1	7225	85
16	0	90	0	8100	0
17	5	50	25	2500	250
18	3	70	9	4900	210
19	4	40	16	1600	160
20	0	95	0	9025	0
Σ	59	1405	231	104575	3680

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

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

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

Recall from the previous chapter that the sum of products and sum of squares X and sum of squares Y were used in computing the correlation coefficient:

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

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

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

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

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

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

$$b = r_{XY}\frac{S_Y}{S_X} = -.804\left(\frac{17.58}{1.73}\right) = -8.16$$

These values will closely approximate (within rounding error) those output by a computer program. The prediction of Y given knowledge of X is then possible using the intercept and slope values in the following linear regression equation:

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

To determine the predicted Y values (Yhat) we would substitute each value of X into the linear regression equation. The resulting Y, Yhat, and errors of prediction are given below.

Y	Yhat	e(Y−Yhat)
90	78.00	12.00
70	61.68	8.32
80	69.84	10.16
60	53.52	6.48
95	86.16	8.84
80	78.00	2.00
50	53.52	−3.52
45	69.84	−24.84
75	78.00	−3.00
65	61.68	3.32
45	53.52	−8.52
80	86.16	−6.16
80	61.68	18.32
60	53.52	6.48
85	86.16	−1.16
90	94.32	−4.32
50	53.52	−3.52
70	69.84	.16
40	61.68	−21.68
95	94.32	.68

The error of prediction for the first student is computed as follows:

Step 1

$$\text{Yhat} = a + bX$$
$$\text{Yhat} = 94.32 + (-8.16 * 2) = 78$$

Step 2

$$e = Y - \text{Yhat}$$
$$e = 90 - 78 = 12$$

Check on linear equation:

$$Y = a + bX + e$$
$$90 = 94.32 + (-16.32) + 12.00$$

In this data example, the correlation coefficient is negative ($r = -.80$), which indicates that as days absent during the week increases (X), the math achievement scores decrease. This relationship would be depicted as a downward trend in the data points on a scatter plot. 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. We square the correlation coefficient value to obtain a variance accounted for interpretation, i.e., when $r = -.80$, $r^2 = .64$. Knowledge of the number of days absent accounts for 64% of the variance in the math achievement scores.

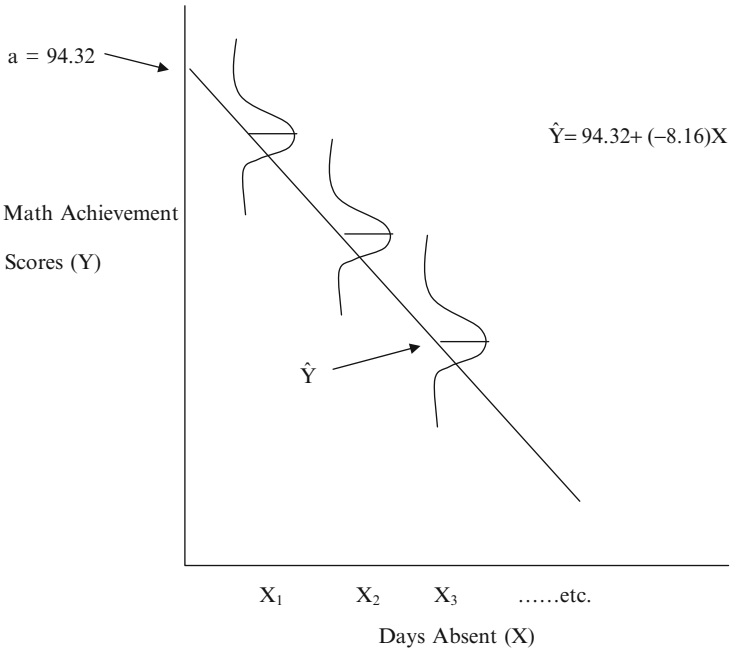
The errors of prediction also serve to identify the accuracy of the regression equation. Notice that some of the errors are positive and some of the errors are negative, consequently the sum (and mean) of the errors should be zero. We expect the Y scores to be normally distributed around Yhat for each value of X so that the variability of these errors indicate the standard deviation of the Y scores around Yhat for each value of X. The standard deviation of the Y scores around Yhat is called a standard error of estimate. It is computed as:

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

Another approach using the standard deviation of Y, the correlation coefficient, and sample size is computed as:

$$S_{Y.X} = S_Y \sqrt{1 - r^2} \sqrt{n - 1 / n - 2} = 17.58 \sqrt{1 - (-.804)^2} \sqrt{20 - 1 / 20 - 2} = 10.75$$

A graph of the Y score distribution around each individual X score will help to better understand the interpretation of the standard error of estimate and the concept of homoscedasticity (equal variance of Y scores around Yhat for each X score along the line of least squares). For each value of X, there is a distribution of Y scores around each Yhat value.



The $S_{Y.X} = 10.75$ is the standard deviation of the Y scores around the predicted Yhat score for each X score. This standard deviation is assumed to be the same for each distribution of Y scores along the least squares line, i.e., homoscedasticity of variance along the least squares line. The predicted Yhat is the mean of the distribution of Y scores for each value of X. The standard error of estimate is therefore calculated as the square root of the sum of the squared differences between Y and Yhat, i.e., $(Y - Yhat)^2$, divided by $N - 2$. Since it is assumed that different values of Y vary in a normal distribution around Yhat, the assumption of equal variance in Y across the least squares line is important because you want an accurate mean squared error!

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 "rise" to "run" interpretation of the z-values. The standard scores (z-scores), as they are sometimes called, can be converted back to their respective raw score. The z-score formula subtracts the mean from each score and divides by the standard deviation. The formula you may recall is:

$$z = \frac{X - \overline{X}}{S}$$

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 zero and standard deviations for X and Y are one:

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

$$b = r_{XY}\frac{S_Y}{S_X} = -.804\left(\frac{1}{1}\right) = -.804$$

Because the mean and standard deviation of z-scores are zero (0) and one (1), respectively, the least squares line would pass through the origin ($Y = 0$ and $X = 0$) of the scatter plot with the Y and X axes labeled in z-score units. Notice that the correlation coefficient captures the slope of the least squares line. The regression equation in z-score form is written as: $Z_Y = \beta Z_X$, with $\beta = -.804$, the Pearson correlation coefficient.

The use of linear regression in applied research is very popular. For example, admission into graduate school is based on the prediction of grade point average using the Graduate Record Exam (GRE) score. Colleges and Universities predict budgets and enrollment from 1 year to the next based on previous attendance data. The Pearson correlation coefficient played an important role in making these predictions possible. A statistically significant correlation between Y and X will generally indicate a good prediction is possible because the difference between the observed Y values and the predicted Yhat 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 regression line, which minimizes 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 regression formula using z-scores, we see the unique role that the Pearson correlation coefficient plays.

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 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 least squares line from sample data are estimates of the population intercept and slope. 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 Yhat.

REGRESSION R Program

A true population linear regression equation is specified based on the amount of overtime worked (X) and bonus points received (Y). The true population linear

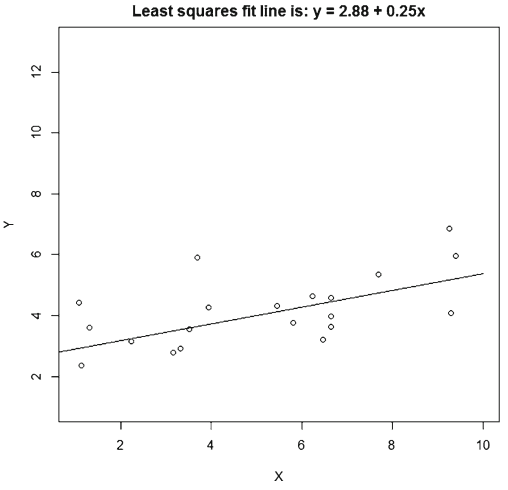
regression equation can be expressed as: Y=3+.25X+e. This equation indicates that if there are no overtime hours worked (X=0), the number of bonus points is 3.0, plus some random error (*e*) that is due to chance, which can be either positive or negative. The bonus points are increased by .25 for each hour of overtime worked. It is assumed that for each X, the Y values are normally distributed around predicted Y, their mean on the line Y=3+.25X, and that these normal distributions of Y values around their individual predicted Y have the same variance. The data will be selected at random from a normal population. The scatter plot of X and Y data points is produced by the **plot** function and the least squares regression line is drawn using the **lines** function each time you run the program. The linear regression equation is listed at the top of the scatter plot. The program uses the least squares **lsfit** function in R to calculate the intercept and slope of the regression equation. The program prints the intercept and slope of the true regression equation in the population and the regression equation based on the sample data. The pairs of X and Y values are printed for reference. The correlation coefficient is printed to reference the slope for a regression equation using standard scores for Y and X, which is Beta or the standardized regression coefficient.

REGRESSION Program Output

```
Scatterplot Data Points

(3.51,3.56) (3.16,2.79) (9.29,4.08) (1.13,2.36) (3.69,5.9)
(7.69,5.37) (1.08,4.43) (1.32,3.61) (3.32,2.93) (6.64,4.6)
(6.65,3.64) (5.8,3.77)  (6.24,4.65) (3.95,4.28) (6.64,3.99)
(2.23,3.15) (9.4,5.96)  (5.46,4.33) (6.47,3.21) (9.26,6.86)

True regression line is: y = 3 + 0.25x
Least squares fit line is: y = 2.88 + 0.25x
r = 0.59 (slope using standard scores for X and Y)
```



REGRESSION Exercises

1. Run REGRESSION and enter the 20 pairs of observations in the following spaces.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

   a. What is the equation for the true population least squares line? _____
   b. What is the equation of the sample least squares line? _____
   c. What is the slope of the equation if using z-scores? (Correlation coefficient)_____
   d. Print the scatter plot produced by the REGRESSION program.

2. Run REGRESSION program five more times, and record the sample regression equations.
   TRUE REGRESSION EQUATION: $Y = 3 + .25X$

   SAMPLE EQUATION
   Run 1 _____
   Run 2 _____
   Run 3 _____
   Run 4 _____
   Run 5 _____

   a. What is the slope of the true regression equation? _____
   b. What is the Y-intercept of the true regression equation? _____
   c. Calculate the error in the estimates of the sample slopes and intercepts for the five runs.
      Enter the values in the table below. (Error = Sample Estimate – True Population Value)

	ERROR IN SLOPE	ERROR IN INTERCEPT
RUN 1	_____	_____
RUN 2	_____	_____
RUN 3	_____	_____
RUN 4	_____	_____
RUN 5	_____	_____

d. Are the errors in the slopes in the same direction (positive versus negative)?
YES _____ NO _____
e. Are the errors in the Y-intercepts in the same direction (positive versus negative)?

YES _____ NO _____

3. Run the REGRESSION program and determine for a given value of X (overtime hours worked), what is the bonus received (Y)? Use the following values:

```
bTrue <- .50
aTrue <- 10
sampleSize <- 100
```

a. If X = 4, find the bonus (predicted Y) using the true population regression equation.

_____

b. If X = 4, find the bonus (predicted Y) using the sample equation.

_____

True or False Questions

Linear Regression

T F a. If the equation of a least squares line is Y = 4 − .5X, then the slope of the line is 4.
T F b. Prediction of Y given knowledge of X is the purpose of linear regression analysis.
T F c. The slope of the linear regression equation in z-score form is the Pearson correlation coefficient.
T F d. A regression equation from sample data will usually differ from the true population regression equation.
T F e. Y values are normally distributed around their Yhat means on the least squares line and the Y distributions have the same variance.