CK-12 Probability and Statistics - Advanced (Second Edition)

Ellen Lawsky
Larry Ottman
Raja Almukkahal
Brenda Meery
Danielle DeLancey

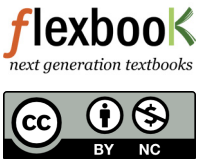
Chapter 9

Regression and Correlation

To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Printed: December 17, 2014



AUTHORS
Ellen Lawsky
Larry Ottman
Raja Almukkahal
Brenda Meery
Danielle DeLancey

CHAPTER 9 Regression and Correlation

Chapter Outline

9.1 Scatterplots and Linear Correlation

Learning Objectives

- Understand the concepts of bivariate data and correlation, and the use of scatterplots to display bivariate data.
- Understand when the terms 'positive', 'negative', 'strong', and 'perfect' apply to the correlation between two variables in a scatterplot graph.
- Calculate the linear correlation coefficient and coefficient of determination of bivariate data, using technology tools to assist in the calculations.
- Understand properties and common errors of correlation.

Introduction

So far we have learned how to describe distributions of a single variable and how to perform hypothesis tests concerning parameters of these distributions. But what if we notice that two variables seem to be related? We may notice that the values of two variables, such as verbal SAT score and GPA, behave in the same way and that students who have a high verbal SAT score also tend to have a high GPA (see table below). In this case, we would want to study the nature of the connection between the two variables.

TABLE 9.1: A table of verbal SAT values and GPAs for seven students.

Student	SAT Score	GPA
1	595	3.4
2	520	3.2
3	715	3.9
4	405	2.3
5	680	3.9
6	490	2.5
7	565	3.5

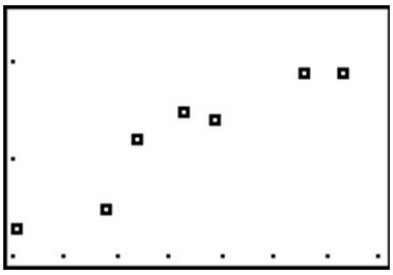
These types of studies are quite common, and we can use the concept of correlation to describe the relationship between the two variables.

Bivariate Data, Correlation Between Values, and the Use of Scatterplots

Correlation measures the relationship between bivariate data. *Bivariate data* are data sets in which each subject has two observations associated with it. In our example above, we notice that there are two observations (verbal SAT score and GPA) for each subject (in this case, a student). Can you think of other scenarios when we would use bivariate data?

If we carefully examine the data in the example above, we notice that those students with high SAT scores tend to have high GPAs, and those with low SAT scores tend to have low GPAs. In this case, there is a tendency for students to score similarly on both variables, and the performance between variables appears to be related.

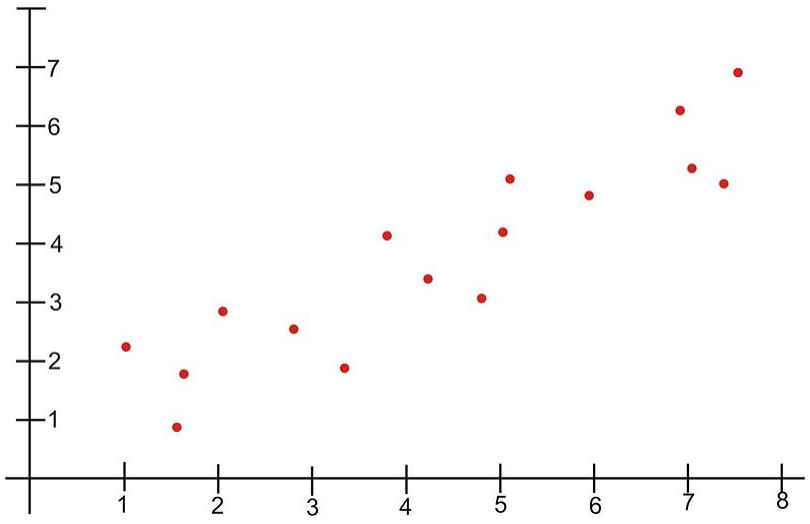
Scatterplots display these bivariate data sets and provide a visual representation of the relationship between variables. In a scatterplot, each point represents a paired measurement of two variables for a specific subject, and each subject is represented by one point on the scatterplot.



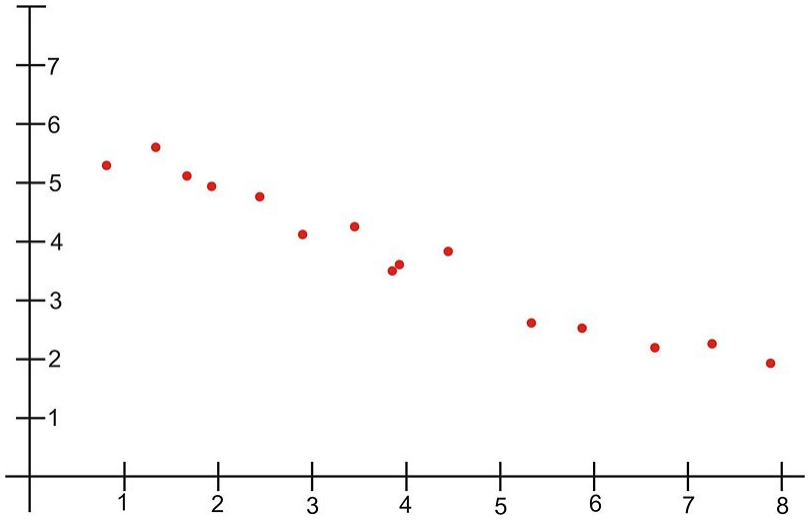
Correlation Patterns in Scatterplot Graphs

Examining a scatterplot graph allows us to obtain some idea about the relationship between two variables.

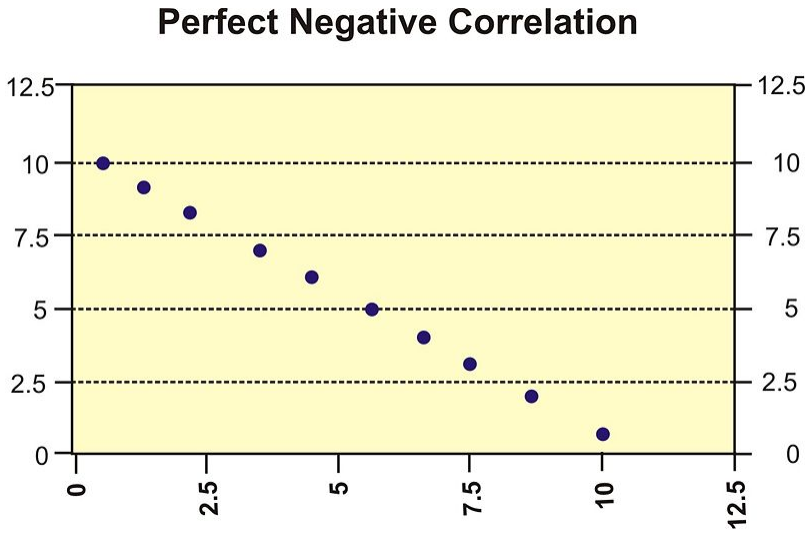
When the points on a scatterplot graph produce a lower-left-to-upper-right pattern (see below), we say that there is a *positive correlation* between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be high as well, and vice versa.



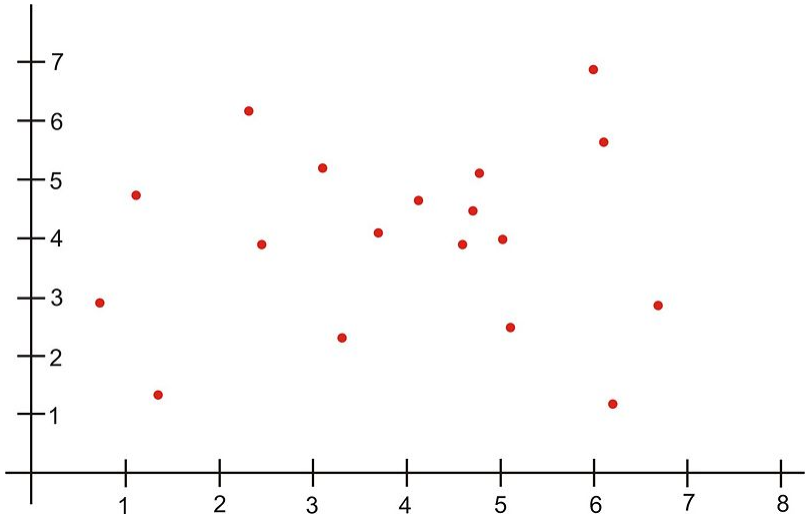
When the points on a scatterplot graph produce a upper-left-to-lower-right pattern (see below), we say that there is a *negative correlation* between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be low, and vice versa.



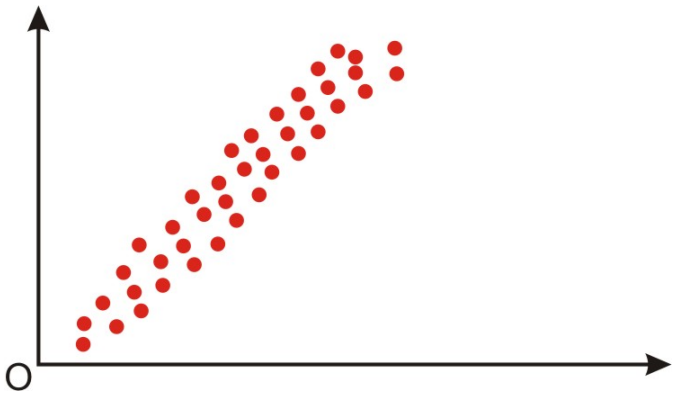
When all the points on a scatterplot lie on a straight line, you have what is called a *perfect correlation* between the two variables (see below).



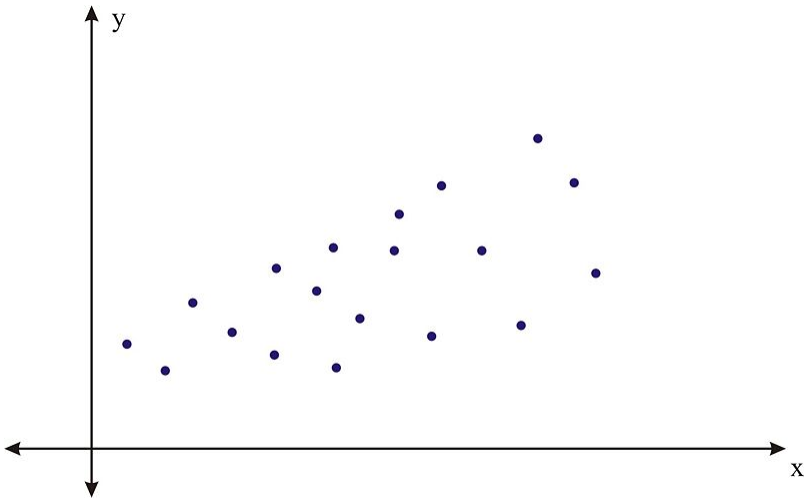
A scatterplot in which the points do not have a linear trend (either positive or negative) is called a *zero correlation* or a *near-zero correlation* (see below).



When examining scatterplots, we also want to look not only at the direction of the relationship (positive, negative, or zero), but also at the *magnitude* of the relationship. If we drew an imaginary oval around all of the points on the scatterplot, we would be able to see the extent, or the magnitude, of the relationship. If the points are close to one another and the width of the imaginary oval is small, this means that there is a strong correlation between the variables (see below).



However, if the points are far away from one another, and the imaginary oval is very wide, this means that there is a weak correlation between the variables (see below).



Correlation Coefficients

While examining scatterplots gives us some idea about the relationship between two variables, we use a statistic called the *correlation coefficient* to give us a more precise measurement of the relationship between the two variables. The correlation coefficient is an index that describes the relationship and can take on values between -1.0 and $+1.0$, with a positive correlation coefficient indicating a positive correlation and a negative correlation coefficient indicating a negative correlation.

The absolute value of the coefficient indicates the magnitude, or the strength, of the relationship. The closer the absolute value of the coefficient is to 1, the stronger the relationship. For example, a correlation coefficient of 0.20 indicates that there is a weak linear relationship between the variables, while a coefficient of -0.90 indicates that there is a strong linear relationship.

The value of a perfect positive correlation is 1.0, while the value of a perfect negative correlation is -1.0.

When there is no linear relationship between two variables, the correlation coefficient is 0. It is important to

remember that a correlation coefficient of 0 indicates that there is no *linear* relationship, but there may still be a strong relationship between the two variables. For example, there could be a quadratic relationship between them.

The Pearson product-moment correlation coefficient is a statistic that is used to measure the strength and direction of a linear correlation. It is symbolized by the letter r. To understand how this coefficient is calculated, let's suppose that there is a positive relationship between two variables, X and Y. If a subject has a score on X that is above the mean, we expect the subject to have a score on Y that is also above the mean. Pearson developed his correlation coefficient by computing the sum of cross products. He multiplied the two scores, X and Y, for each subject and then added these cross products across the individuals. Next, he divided this sum by the number of subjects minus one. This coefficient is, therefore, the mean of the cross products of scores.

Pearson used standard scores (z-scores, t-scores, etc.) when determining the coefficient.

Therefore, the formula for this coefficient is as follows:

$$r_{XY} = \frac{\sum z_X z_Y}{n-1}$$

In other words, the coefficient is expressed as the sum of the cross products of the standard z-scores divided by the number of degrees of freedom.

An equivalent formula that uses the raw scores rather than the standard scores is called the raw score formula and is written as follows:

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}}$$

Again, this formula is most often used when calculating correlation coefficients from original data. Note that n is used instead of $n-1$, because we are using actual data and not z-scores. Let's use our example from the introduction to demonstrate how to calculate the correlation coefficient using the raw score formula.

Example: What is the Pearson product-moment correlation coefficient for the two variables represented in the table below?

TABLE 9.2: The table of values for this example.

Student	SAT Score	GPA
1	595	3.4
2	520	3.2
3	715	3.9
4	405	2.3
5	680	3.9
6	490	2.5
7	565	3.5

In order to calculate the correlation coefficient, we need to calculate several pieces of information, including xy, x^2, and y^2. Therefore, the values of xy, x^2, and y^2 have been added to the table.

TABLE 9.3:

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2
1	595	3.4	2023	354025	11.56

TABLE 9.3: (continued)

Student	**SAT Score** (X)	**GPA** (Y)	xy	x^2	y^2
2	520	3.2	1664	270400	10.24
3	715	3.9	2789	511225	15.21
4	405	2.3	932	164025	5.29
5	680	3.9	2652	462400	15.21
6	490	2.5	1225	240100	6.25
7	565	3.5	1978	319225	12.25
Sum	3970	22.7	13262	2321400	76.01

Applying the formula to these data, we find the following:

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}} = \frac{(7)(13262) - (3970)(22.7)}{\sqrt{[(7)(2321400) - 3970^2][(7)(76.01) - 22.7^2]}}$$
$$= \frac{2715}{2864.22} \approx 0.95$$

The correlation coefficient not only provides a measure of the relationship between the variables, but it also gives us an idea about how much of the total variance of one variable can be associated with the variance of the other. For example, the correlation coefficient of 0.95 that we calculated above tells us that to a high degree, the variance in the scores on the verbal SAT is associated with the variance in the GPA, and vice versa. For example, we could say that factors that influence the verbal SAT, such as health, parent college level, etc., would also contribute to individual differences in the GPA. The higher the correlation we have between two variables, the larger the portion of the variance that can be explained by the independent variable.

The calculation of this variance is called the *coefficient of determination* and is calculated by squaring the correlation coefficient. Therefore, the coefficient of determination is written as r^2. The result of this calculation indicates the proportion of the variance in one variable that can be associated with the variance in the other variable.

On the Web

http://tinyurl.com/ylcyh88 Match the graph to its correlation.

http://tinyurl.com/y8vcm5y Guess the correlation.

http://onlinestatbook.com/stat_sim/reg_by_eye/index.html Regression by eye.

The Properties and Common Errors of Correlation

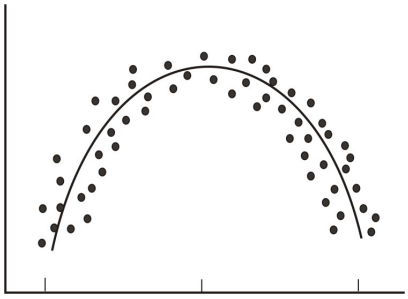
Correlation is a measure of the linear relationship between two variables-it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other things may be causing the two correlated variables to relate as they do. Therefore, it is important to remember that we are interpreting the variables and the variance not as causal, but instead as relational.

When examining correlation, there are three things that could affect our results: linearity, homogeneity of the group, and sample size.

Linearity

As mentioned, the correlation coefficient is the measure of the linear relationship between two variables. However, while many pairs of variables have a linear relationship, some do not. For example, let's consider performance anxiety. As a person's anxiety about performing increases, so does his or her performance up to a point. (We sometimes call this good stress.) However, at some point, the increase in anxiety may cause a person's performance to

go down. We call these non-linear relationships *curvilinear relationships*. We can identify curvilinear relationships by examining scatterplots (see below). One may ask why curvilinear relationships pose a problem when calculating the correlation coefficient. The answer is that if we use the traditional formula to calculate these relationships, it will not be an accurate index, and we will be underestimating the relationship between the variables. If we graphed performance against anxiety, we would see that anxiety has a strong affect on performance. However, if we calculated the correlation coefficient, we would arrive at a figure around zero. Therefore, the correlation coefficient is not always the best statistic to use to understand the relationship between variables.



Homogeneity of the Group

Another error we could encounter when calculating the correlation coefficient is homogeneity of the group. When a group is homogeneous, or possesses similar characteristics, the range of scores on either or both of the variables is restricted. For example, suppose we are interested in finding out the correlation between IQ and salary. If only members of the Mensa Club (a club for people with IQs over 140) are sampled, we will most likely find a very low correlation between IQ and salary, since most members will have a consistently high IQ, but their salaries will still vary. This does not mean that there is not a relationship-it simply means that the restriction of the sample limited the magnitude of the correlation coefficient.

Sample Size

Finally, we should consider sample size. One may assume that the number of observations used in the calculation of the correlation coefficient may influence the magnitude of the coefficient itself. However, this is not the case. Yet while the sample size does not affect the correlation coefficient, it may affect the accuracy of the relationship. The larger the sample, the more accurate of a predictor the correlation coefficient will be of the relationship between the two variables.

Lesson Summary

Bivariate data are data sets with two observations that are assigned to the same subject. Correlation measures the direction and magnitude of the linear relationship between bivariate data. When examining scatterplot graphs, we can determine if correlations are positive, negative, perfect, or zero. A correlation is strong when the points in the scatterplot lie generally along a straight line.

The correlation coefficient is a precise measurement of the relationship between the two variables. This index can take on values between and including -1.0 and $+1.0$.

To calculate the correlation coefficient, we most often use the raw score formula, which allows us to calculate the coefficient by hand.

This formula is as follows: $r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}}$.

When calculating the correlation coefficient, there are several things that could affect our computation, including curvilinear relationships, homogeneity of the group, and the size of the group.

Multimedia Links

For an explanation of the correlation coefficient **(13.0)**, see kbower50, The Correlation Coefficient (3:59).

MEDIA

Click image to the left or use the URL below.

URL: http://www.ck12.org/flx/render/embeddedobject/1104

Review Questions

1. Give 2 scenarios or research questions where you would use bivariate data sets.
2. In the space below, draw and label four scatterplot graphs. One should show:
   a. a positive correlation
   b. a negative correlation
   c. a perfect correlation
   d. a zero correlation
3. In the space below, draw and label two scatterplot graphs. One should show:
   a. a weak correlation
   b. a strong correlation.
4. What does the correlation coefficient measure?
5. The following observations were taken for five students measuring grade and reading level.

TABLE 9.4: A table of grade and reading level for five students.

Student Number	Grade	Reading Level
1	2	6
2	6	14
3	5	12
4	4	10
5	1	4

(a) Draw a scatterplot for these data. What type of relationship does this correlation have?

(b) Use the raw score formula to compute the Pearson correlation coefficient.

6. A teacher gives two quizzes to his class of 10 students. The following are the scores of the 10 students.

TABLE 9.5: Quiz results for ten students.

Student	Quiz 1	Quiz 2
1	15	20
2	12	15

TABLE 9.5: (continued)

Student	Quiz 1	Quiz 2
3	10	12
4	14	18
5	10	10
6	8	13
7	6	12
8	15	10
9	16	18
10	13	15

(a) Compute the Pearson correlation coefficient, r, between the scores on the two quizzes.

(b) Find the percentage of the variance, r^2, in the scores of Quiz 2 associated with the variance in the scores of Quiz 1.

(c) Interpret both r and r^2 in words.

7. What are the three factors that we should be aware of that affect the magnitude and accuracy of the Pearson correlation coefficient?

9.2 Least-Squares Regression

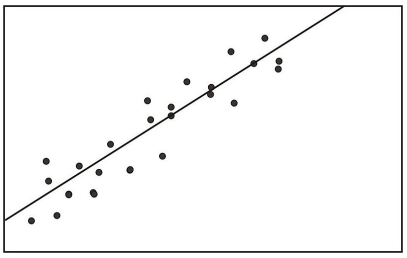
Learning Objectives

- Calculate and graph a regression line.
- Predict values using bivariate data plotted on a scatterplot.
- Understand outliers and influential points.
- Perform transformations to achieve linearity.
- Calculate residuals and understand the least-squares property and its relation to the regression equation.
- Plot residuals and test for linearity.

Introduction

In the last section, we learned about the concept of correlation, which we defined as the measure of the linear relationship between two variables. As a reminder, when we have a strong positive correlation, we can expect that if the score on one variable is high, the score on the other variable will also most likely be high. With correlation, we are able to roughly predict the score of one variable when we have the other. Prediction is simply the process of estimating scores of one variable based on the scores of another variable.

In the previous section, we illustrated the concept of correlation through scatterplot graphs. We saw that when variables were correlated, the points on a scatterplot graph tended to follow a straight line. If we could draw this straight line, it would, in theory, represent the change in one variable associated with the change in the other. This line is called the *least squares line*, or the *linear regression line* (see figure below).



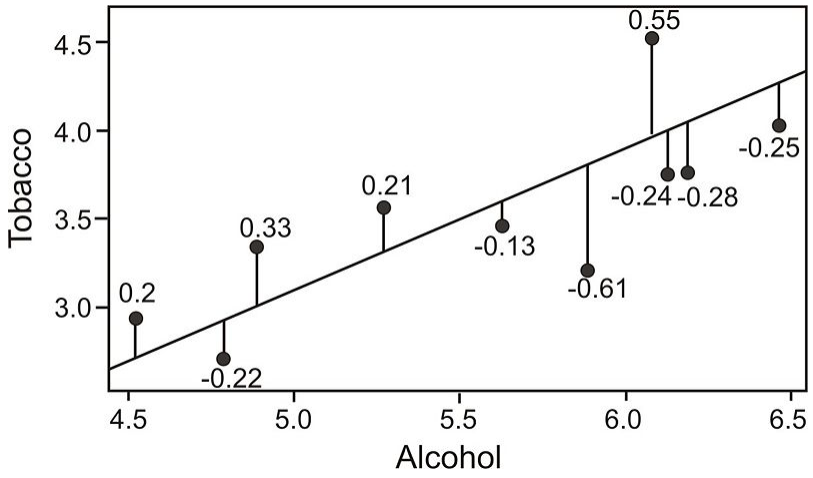
Calculating and Graphing the Regression Line

Linear regression involves using data to calculate a line that best fits that data and then using that line to predict scores. In linear regression, we use one variable (the *predictor variable*) to predict the outcome of another (the *outcome variable*, or *criterion variable*). To calculate this line, we analyze the patterns between the two variables.

We are looking for a line of best fit, and there are many ways one could define this best fit. Statisticians define this line to be the one which minimizes the sum of the squared distances from the observed data to the line.

To determine this line, we want to find the change in X that will be reflected by the average change in Y. After we calculate this average change, we can apply it to any value of X to get an approximation of Y. Since the regression line is used to predict the value of Y for any given value of X, all predicted values will be located on the regression

line, itself. Therefore, we try to fit the regression line to the data by having the smallest sum of squared distances possible from each of the data points to the line. In the example below, you can see the calculated distances, or residual values, from each of the observations to the regression line. This method of fitting the data line so that there is minimal difference between the observations and the line is called the *method of least squares*, which we will discuss further in the following sections.



As you can see, the regression line is a straight line that expresses the relationship between two variables. When predicting one score by using another, we use an equation such as the following, which is equivalent to the slope-intercept form of the equation for a straight line:

$$Y = bX + a$$

where:

Y is the score that we are trying to predict.

b is the slope of the line.

a is the y-intercept, or the value of Y when the value of X is 0.

To calculate the line itself, we need to find the values for b (the *regression coefficient*) and a (the *regression constant*). The regression coefficient explains the nature of the relationship between the two variables. Essentially, the regression coefficient tells us that a certain change in the predictor variable is associated with a certain change in the outcome, or criterion, variable. For example, if we had a regression coefficient of 10.76, we would say that a change of 1 unit in X is associated with a change of 10.76 units of Y. To calculate this regression coefficient, we can use the following formulas:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

or

$$b = (r)\frac{s_Y}{s_X}$$

where:

r is the correlation between the variables X and Y.

s_Y is the standard deviation of the Y scores.

s_X is the standard deviation of the X scores.

In addition to calculating the regression coefficient, we also need to calculate the regression constant. The regression constant is also the y-intercept and is the place where the line crosses the y-axis. For example, if we had an equation with a regression constant of 4.58, we would conclude that the regression line crosses the y-axis at 4.58. We use the following formula to calculate the regression constant:

$$a = \frac{\sum y - b\sum x}{n} = \bar{y} - b\bar{x}$$

Example: Find the least squares line (also known as the linear regression line or the *line of best fit*) for the example measuring the verbal SAT scores and GPAs of students that was used in the previous section.

TABLE 9.6: SAT and GPA data including intermediate computations for computing a linear regression.

Student	**SAT Score** (X)	**GPA** (Y)	xy	x^2	y^2
1	595	3.4	2023	354025	11.56
2	520	3.2	1664	270400	10.24
3	715	3.9	2789	511225	15.21
4	405	2.3	932	164025	5.29
5	680	3.9	2652	462400	15.21
6	490	2.5	1225	240100	6.25
7	565	3.5	1978	319225	12.25
Sum	3970	22.7	13262	2321400	76.01

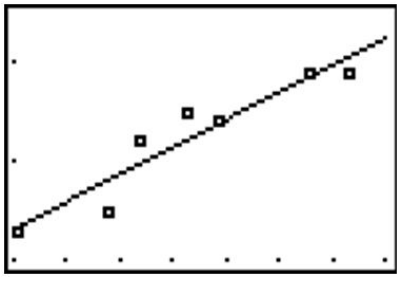
Using these data points, we first calculate the regression coefficient and the regression constant as follows:

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2} = \frac{(7)(13,262) - (3,970)(22.7)}{(7)(2,321,400) - 3,970^2} = \frac{2715}{488900} \approx 0.0056$$

$$a = \frac{\sum y - b\sum x}{n} \approx 0.094$$

Note: If you performed the calculations yourself and did not get exactly the same answers, it is probably due to rounding in the table for xy.

Now that we have the equation of this line, it is easy to plot on a scatterplot. To plot this line, we simply substitute two values of X and calculate the corresponding Y values to get two pairs of coordinates. Let's say that we wanted to plot this example on a scatterplot. We would choose two hypothetical values for X (say, 400 and 500) and then solve for Y in order to identify the coordinates (400, 2.334) and (500, 2.89). From these pairs of coordinates, we can draw the regression line on the scatterplot.



Predicting Values Using Scatterplot Data

One of the uses of a regression line is to predict values. After calculating this line, we are able to predict values by simply substituting a value of a predictor variable, X, into the regression equation and solving the equation for the outcome variable, Y. In our example above, we can predict the students' GPA's from their SAT scores by plugging in the desired values into our regression equation, $Y = 0.0056X + 0.094$.

For example, say that we wanted to predict the GPA for two students, one who had an SAT score of 500 and the other who had an SAT score of 600. To predict the GPA scores for these two students, we would simply plug the two values of the predictor variable into the equation and solve for Y (see below).

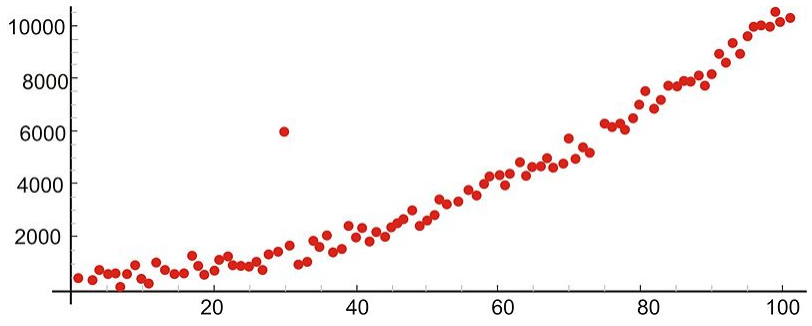
TABLE 9.7: GPA/SAT data, including predicted GPA values from the linear regression.

Student	**SAT Score** (X)	**GPA** (Y)	**Predicted GPA** ($\hat{Y}$)
1	595	3.4	3.4
2	520	3.2	3.0
3	715	3.9	4.1
4	405	2.3	2.3
5	680	3.9	3.9
6	490	2.5	2.8
7	565	3.5	3.2
Hypothetical	600		3.4
Hypothetical	500		2.9

As you can see, we are able to predict the value for Y for any value of X within a specified range.

Outliers and Influential Points

An *outlier* is an extreme observation that does not fit the general correlation or regression pattern (see figure below). In the regression setting, outliers will be far away from the regression line in the y-direction. Since it is an unusual observation, the inclusion of an outlier may affect the slope and the y-intercept of the regression line. When examining a scatterplot graph and calculating the regression equation, it is worth considering whether extreme observations should be included or not. In the following scatterplot, the outlier has approximate coordinates of (30, 6,000).

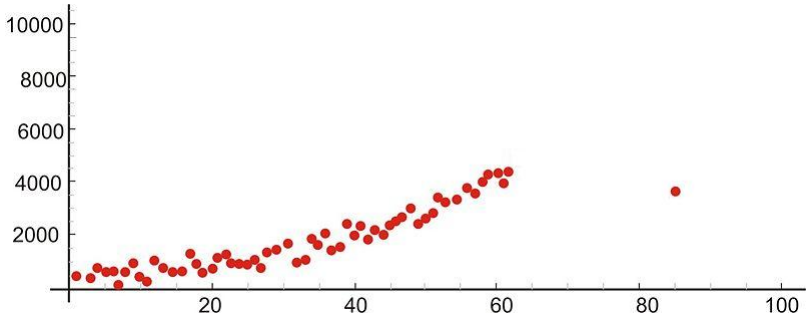


Let's use our example above to illustrate the effect of a single outlier. Say that we have a student who has a high GPA but who suffered from test anxiety the morning of the SAT verbal test and scored a 410. Using our original regression equation, we would expect the student to have a GPA of 2.2. But, in reality, the student has a GPA equal to 3.9. The inclusion of this value would change the slope of the regression equation from 0.0055 to 0.0032, which is quite a large difference.

There is no set rule when trying to decide whether or not to include an outlier in regression analysis. This decision depends on the sample size, how extreme the outlier is, and the normality of the distribution. For univariate data, we

can use the IQR rule to determine whether or not a point is an outlier. We should consider values that are 1.5 times the inter-quartile range below the first quartile or above the third quartile as outliers. Extreme outliers are values that are 3.0 times the inter-quartile range below the first quartile or above the third quartile.

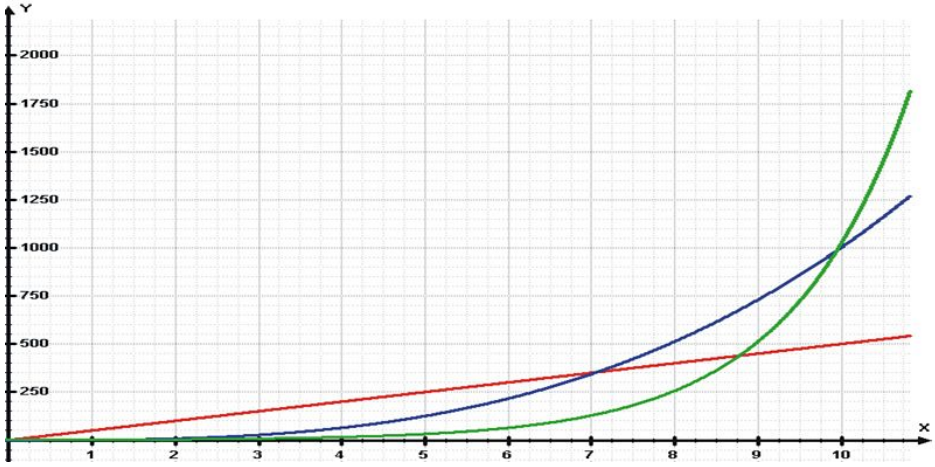
An *influential point* in regression is one whose removal would greatly impact the equation of the regression line. Usually, an influential point will be separated in the x direction from the other observations. It is possible for an outlier to be an influential point. However, there are some influential points that would not be considered outliers. These will not be far from the regression line in the y-direction (a value called a residual, discussed later) so you must look carefully for them. In the following scatterplot, the influential point has approximate coordinates of (85, 35,000).



It is important to determine whether influential points are 1) correct and 2) belong in the population. If they are not correct or do not belong, then they can be removed. If, however, an influential point is determined to indeed belong in the population and be correct, then one should consider whether other data points need to be found with similar x-values to support the data and regression line.

Transformations to Achieve Linearity

Sometimes we find that there is a relationship between X and Y, but it is not best summarized by a straight line. When looking at the scatterplot graphs of correlation patterns, these relationships would be shown to be curvilinear. While many relationships are linear, there are quite a number that are not, including learning curves (learning more quickly at the beginning, followed by a leveling out) and exponential growth (doubling in size, for example, with each unit of growth). Below is an example of a growth curve describing the growth of a complex society:



Since this is not a linear relationship, we cannot immediately fit a regression line to this data. However, we can perform a *transformation* to achieve a linear relationship. We commonly use transformations in everyday life. For

example, the Richter scale, which measures earthquake intensity, and the idea of describing pay raises in terms of percentages are both examples of making transformations of non-linear data.

Consider the following exponential relationship, and take the log of both sides as shown:

$$\begin{aligned} y &= ab^x \\ \log y &= \log(ab^x) \\ \log y &= \log a + \log b^x \\ \log y &= \log a + x \log b \end{aligned}$$

In this example, a and b are real numbers (constants), so this is now a linear relationship between the variables x and $\log y$.

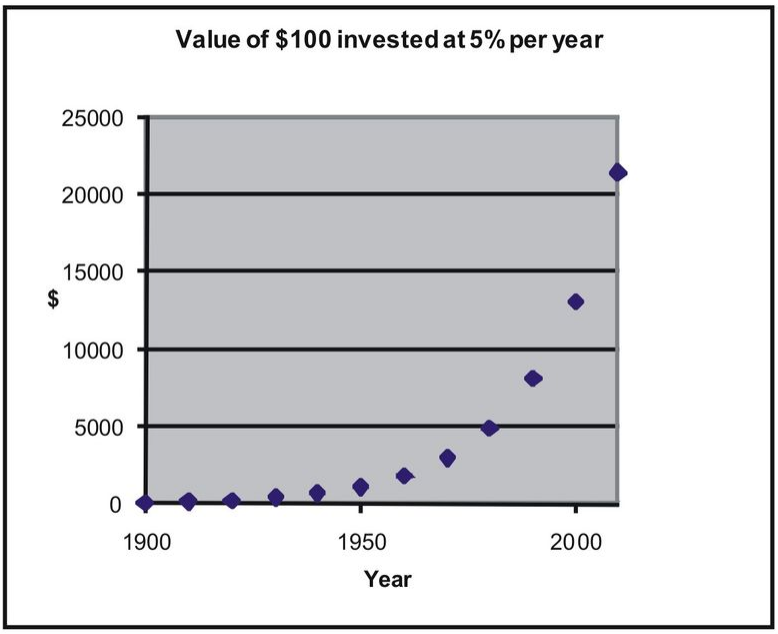
Thus, you can find a least squares line for these variables.

Let's take a look at an example to help clarify this concept. Say that we were interested in making a case for investing and examining how much return on investment one would get on $100 over time. Let's assume that we invested $100 in the year 1900 and that this money accrued 5% interest every year. The table below details how much we would have each decade:

TABLE 9.8: Table of account growth assuming $100 invested in 1900 at 5% annual growth.

Year	Investment with 5% Each Year
1900	100
1910	163
1920	265
1930	432
1940	704
1950	1147
1960	1868
1970	3043
1980	4956
1990	8073
2000	13150
2010	21420

If we graphed these data points, we would see that we have an exponential growth curve.

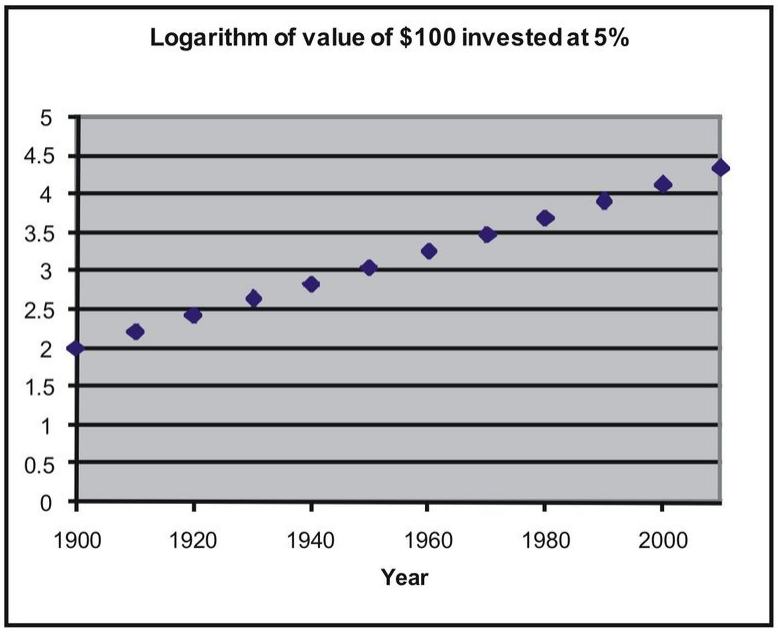


Say that we wanted to fit a linear regression line to these data. First, we would transform these data using logarithmic transformations as follows:

TABLE 9.9: Account growth data and values after a logarithmic transformation.

Year	Investment with 5% Each Year	Log of amount
1900	100	2
1910	163	2.211893
1920	265	2.423786
1930	432	2.635679
1940	704	2.847572
1950	1147	3.059465
1960	1868	3.271358
1970	3043	3.483251
1980	4956	3.695144
1990	8073	3.907037
2000	13150	4.118930
2010	21420	4.330823

If we plotted these transformed data points, we would see that we have a linear relationship as shown below:



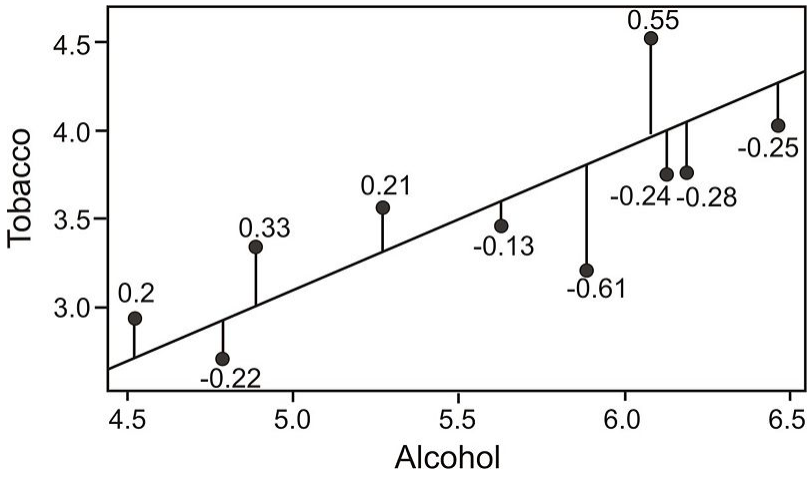
We can now perform a linear regression on (year, log of amount). If you enter the data into the TI-83/84 calculator, press **[STAT]**, go to the **CALC** menu, and use the 'LinReg(ax+b)' command, you find the following relationship:

$$Y = 0.021X - 38.2$$

with X representing year and Y representing log of amount.

Calculating Residuals and Understanding their Relation to the Regression Equation

Recall that the linear regression line is the line that best fits the given data. Ideally, we would like to minimize the distances of all data points to the regression line. These distances are called the error, e, and are also known as the *residual values*. As mentioned, we fit the regression line to the data points in a scatterplot using the least-squares method. A good line will have small residuals. Notice in the figure below that the residuals are the vertical distances between the observations and the predicted values on the regression line:



To find the residual values, we subtract the predicted values from the actual values, so $e = y - \hat{y}$. Theoretically, the sum of all residual values is zero, since we are finding the line of best fit, with the predicted values as close as

possible to the actual value. It does not make sense to use the sum of the residuals as an indicator of the fit, as the negative and positive residuals always cancel each other out to give a sum of zero. Therefore, we try to minimize the sum of the squared residuals, or $\sum(y-\hat{y})^2$.

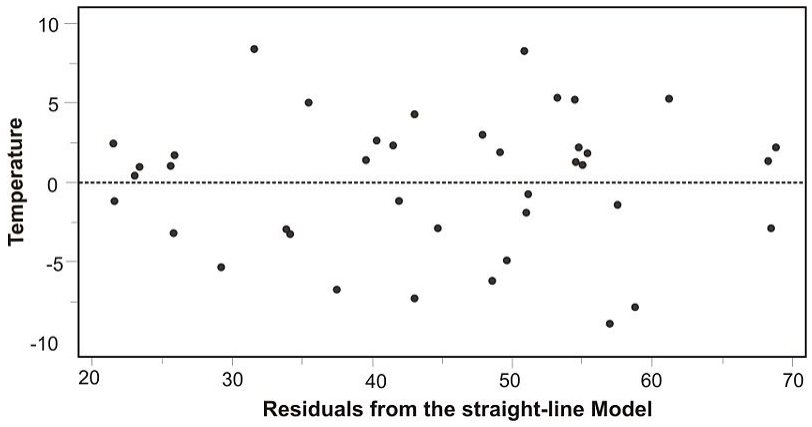
Example: Calculate the residuals for the predicted and the actual GPA's from our sample above.

TABLE 9.10: SAT/GPA data, including residuals.

Student	SAT Score (X)	GPA (Y)	Predicted GPA $(\hat{Y})$	Residual Value	Residual Value Squared
1	595	3.4	3.4	0	0
2	520	3.2	3.0	0.2	0.04
3	715	3.9	4.1	−0.2	0.04
4	405	2.3	2.3	0	0
5	680	3.9	3.9	0	0
6	490	2.5	2.8	−0.3	0.09
7	565	3.5	3.2	0.3	0.09
$\sum(y-\hat{y})^2$					0.26

Plotting Residuals and Testing for Linearity

To test for linearity and to determine if we should drop extreme observations (or outliers) from our analysis, it is helpful to plot the residuals. When plotting, we simply plot the x-value for each observation on the x-axis and then plot the residual score on the y-axis. When examining this scatterplot, the data points should appear to have no correlation, with approximately half of the points above 0 and the other half below 0. In addition, the points should be evenly distributed along the x-axis. Below is an example of what a residual scatterplot should look like if there are no outliers and a linear relationship.



If the scatterplot of the residuals does not look similar to the one shown, we should look at the situation a bit more closely. For example, if more observations are below 0, we may have a positive outlying residual score that is skewing the distribution, and if more of the observations are above 0, we may have a negative outlying residual score. If the points are clustered close to the y-axis, we could have an x-value that is an outlier. If this occurs, we may want to consider dropping the observation to see if this would impact the plot of the residuals. If we do decide to drop the observation, we will need to recalculate the original regression line. After this recalculation, we will have a regression line that better fits a majority of the data.

Lesson Summary

Prediction is simply the process of estimating scores of one variable based on the scores of another variable. We use the least-squares regression line, or linear regression line, to predict the value of a variable.

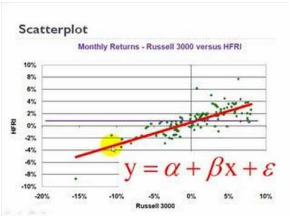
Using this regression line, we are able to use the slope, y-intercept, and the calculated regression coefficient to predict the scores of a variable. The predictions are represented by the variable $\hat{y}$.

When there is an exponential relationship between the variables, we can transform the data by taking the log of the dependent variable to achieve linearity between x and $\log y$. We can then fit a least squares regression line to the transformed data.

The differences between the actual and the predicted values are called residual values. We can construct scatterplots of these residual values to examine outliers and test for linearity.

Multimedia Links

For an introduction to what a least squares regression line represents **(12.0)**, see bionicturtledotcom, Introduction to Linear Regression (5:15).



MEDIA

Click image to the left or use the URL below.

URL: http://www.ck12.org/flx/render/embeddedobject/1105

Review Questions

1. A school nurse is interested in predicting scores on a memory test from the number of times that a student exercises per week. Below are her observations:

TABLE 9.11: A table of memory test scores compared to the number of times a student exercises per week.

Student	Exercise Per Week	Memory Test Score
1	0	15
2	2	3
3	2	12
4	1	11
5	3	5
6	1	8
7	2	15
8	0	13
9	3	2
10	3	4

TABLE 9.11: (continued)

Student	Exercise Per Week	Memory Test Score
11	4	2
12	1	8
13	1	10
14	1	12
15	2	8

(a) Plot this data on a scatterplot, with the x-axis representing the number of times exercising per week and the y-axis representing memory test score.

(b) Does this appear to be a linear relationship? Why or why not?

(c) What regression equation would you use to construct a linear regression model?

(d) What is the regression coefficient in this linear regression model and what does this mean in words?

(e) Calculate the regression equation for these data.

(f) Draw the regression line on the scatterplot.

(g) What is the predicted memory test score of a student who exercises 3 times per week?

(h) Do you think that a data transformation is necessary in order to build an accurate linear regression model? Why or why not?

(i) Calculate the residuals for each of the observations and plot these residuals on a scatterplot.

(j) Examine this scatterplot of the residuals. Is a transformation of the data necessary? Why or why not?

9.3 Inferences about Regression

Learning Objectives

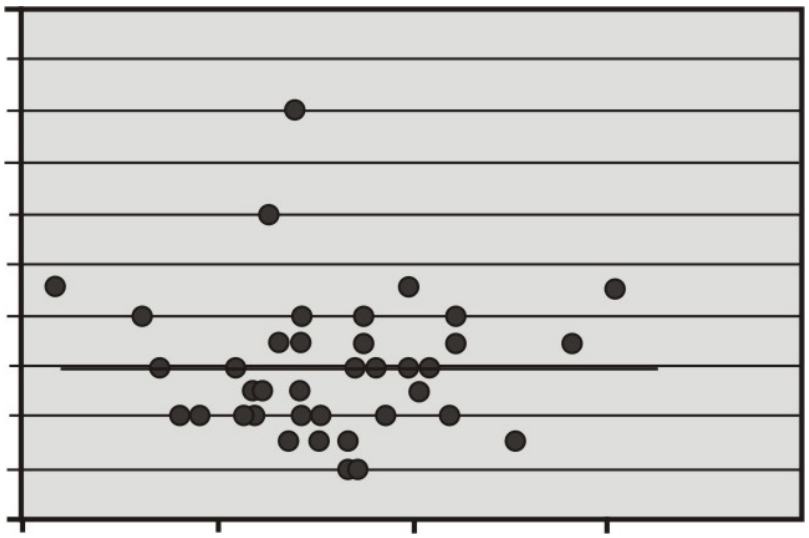
- Make inferences about regression models, including hypothesis testing for linear relationships.
- Make inferences about regression and predicted values, including the construction of confidence intervals.
- Check regression assumptions.

Introduction

In the previous section, we learned about the least-squares model, or the linear regression model. The linear regression model uses the concept of correlation to help us predict the score of a variable based on our knowledge of the score of another variable. In this section, we will investigate several inferences and assumptions that we can make about the linear regression model.

Hypothesis Testing for Linear Relationships

Let's think for a minute about the relationship between correlation and the linear regression model. As we learned, if there is no correlation between the two variables X and Y, then it would be nearly impossible to fit a meaningful regression line to the points on a scatterplot graph. If there was no correlation, and our correlation value, or r-value, was 0, we would always come up with the same predicted value, which would be the mean of all the predicted values, or the mean of $\hat{Y}$. The figure below shows an example of what a regression line fit to variables with no correlation $(r = 0)$ would look like. As you can see, for any value of X, we always get the same predicted value of Y.



Using this knowledge, we can determine that if there is no relationship between X and Y, constructing a regression line doesn't help us very much, because, again, the predicted score would always be the same. Therefore, when we

estimate a linear regression model, we want to ensure that the regression coefficient, β, for the population does not equal zero. Furthermore, it is beneficial to test how strong (or far away) from zero the regression coefficient must be to strengthen our prediction of the Y scores.

In hypothesis testing of linear regression models, the null hypothesis to be tested is that the regression coefficient, β, equals zero. Our alternative hypothesis is that our regression coefficient does not equal zero.

$$H_0: \beta = 0$$
$$H_a: \beta \neq 0$$

The test statistic for this hypothesis test is calculated as follows:

$$t = \frac{b - \beta}{s_b}$$
$$\text{where} \quad s_b = \frac{s}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s}{\sqrt{SS_X}},$$
$$s = \sqrt{\frac{SSE}{n-2}}, \text{ and}$$
$$SSE = \text{sum of residual error squared}$$

Example: Let's say that a football coach is using the results from a short physical fitness test to predict the results of a longer, more comprehensive one. He developed the regression equation $Y = 0.635X + 1.22$, and the standard error of estimate is 0.56. The summary statistics are as follows:

Summary statistics for two foot ball fitness tests.

$n = 24$	$\sum xy = 591.50$
$\sum x = 118$	$\sum y = 104.3$
$\bar{x} = 4.92$	$\bar{y} = 4.35$
$\sum x^2 = 704$	$\sum y^2 = 510.01$
$SS_X = 123.83$	$SS_Y = 56.74$

Using $\alpha = 0.05$, test the null hypothesis that, in the population, the regression coefficient is zero, or $H_0: \beta = 0$.

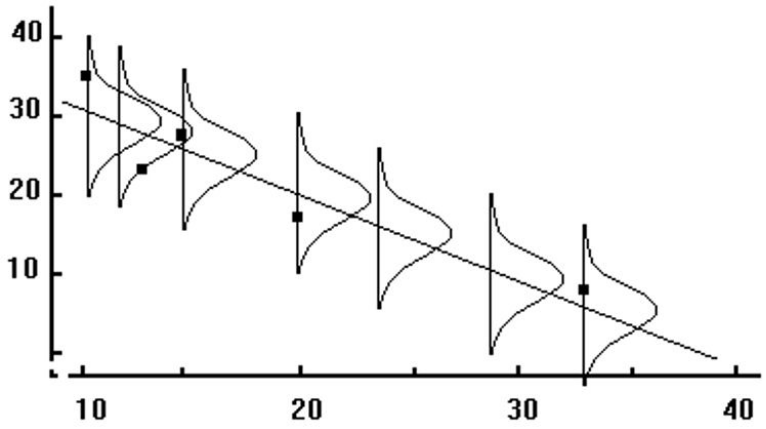
We use the t-distribution to calculate the test statistic and find that the critical values in the t-distribution at 22 degrees of freedom are 2.074 standard scores above and below the mean. Also, the test statistic can be calculated as follows:

$$s_b = \frac{0.56}{\sqrt{123.83}} = 0.05$$
$$t = \frac{0.635 - 0}{0.05} = 12.70$$

Since the observed value of the test statistic exceeds the critical value, the null hypothesis would be rejected, and we can conclude that if the null hypothesis were true, we would observe a regression coefficient of 0.635 by chance less than 5% of the time.

Making Inferences about Predicted Scores

As we have mentioned, a regression line makes predictions about variables based on the relationship of the existing data. However, it is important to remember that the regression line simply infers, or estimates, what the value will be. These predictions are never accurate 100% of the time, unless there is a perfect correlation. What this means is that for every predicted value, we have a normal distribution (also known as the *conditional distribution*, since it is conditional on the X value) that describes the likelihood of obtaining other scores that are associated with the value of the predictor variable, X.



If we assume that these distributions are normal, we are able to make inferences about each of the predicted scores. We can ask questions like, "If the predictor variable, X, equals 4, what percentage of the distribution of Y scores will be lower than 3?"

The reason why we would ask questions like this depends on the scenario. Suppose, for example, that we want to know the percentage of students with a 5 on their short physical fitness test that have a predicted score higher than 5 on their long physical fitness test. If the coach is using this predicted score as a cutoff for playing in a varsity match, and this percentage is too low, he may want to consider changing the standards of the test.

To find the percentage of students with scores above or below a certain point, we use the concept of standard scores and the standard normal distribution.

Since we have a certain predicted value for every value of X, the Y values take on the shape of a normal distribution. This distribution has a mean (the regression line) and a standard error, which we found to be equal to 0.56. In short, the conditional distribution is used to determine the percentage of Y values above or below a certain value that are associated with a specific value of X.

Example: Using our example above, if a student scored a 5 on the short test, what is the probability that he or she would have a score of 5 or greater on the long physical fitness test?

From the regression equation $Y = 0.635X + 1.22$, we find that the predicted score when the value of X is 5 is 4.40. Consider the conditional distribution of Y scores when the value of X is 5. Under our assumption, this distribution is normally distributed around the predicted value 4.40 and has a standard error of 0.56.

Therefore, to find the percentage of Y scores of 5 or greater, we use the general formula for a z-score to calculate the following:

$$z = \frac{Y - \hat{Y}}{s} = \frac{5 - 4.40}{0.56} = 1.07$$

Using the z-distribution table, we find that the area to the right of a z-score of 1.07 is 0.1423. Therefore, we can conclude that the proportion of predicted scores of 5 or greater given a score of 5 on the short test is 0.1423, or 14.23%.

Prediction Intervals

Similar to hypothesis testing for samples and populations, we can also build a confidence interval around our regression results. This helps us ask questions like "If the predictor variable, X, is equal to a certain value, what are the likely values for Y?" A confidence interval gives us a range of scores that has a certain percent probability of including the score that we are after.

We know that the standard error of the predicted score is smaller when the predicted value is close to the actual value, and it increases as X deviates from the mean. This means that the weaker of a predictor that the regression line is, the larger the standard error of the predicted score will be. The formulas for the standard error of a predicted score and a confidence interval are as follows:

$$s_{\hat{Y}} = s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}}$$
$$CI = \hat{Y} \pm ts_{\hat{Y}}$$

where:

$\hat{Y}$ is the predicted score.

t is the critical value for $n-2$ degrees of freedom.

$s_{\hat{Y}}$ is the standard error of the predicted score.

Example: Develop a 95% confidence interval for the predicted score of a student who scores a 4 on the short physical fitness exam.

We calculate the standard error of the predicted score using the formula as follows:

$$s_{\hat{Y}} = s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}} = 0.56\sqrt{1+\frac{1}{24}+\frac{(4-4.92)^2}{123.83}} = 0.57$$

Using the general formula for a confidence interval, we can calculate the answer as shown:

$$\begin{aligned} CI &= \hat{Y} \pm ts_{\hat{Y}} \\ CI_{0.95} &= 3.76 \pm (2.074)(0.57) \\ CI_{0.95} &= 3.76 \pm 1.18 \\ CI_{0.95} &= (2.58, 4.94) \end{aligned}$$

Therefore, we can say that we are 95% confident that given a student's short physical fitness test score, X, of 4, the interval from 2.58 to 4.94 will contain the student's score for the longer physical fitness test.

Regression Assumptions

We make several assumptions under a linear regression model, including:

At each value of X, there is a distribution of Y. These distributions have a mean centered at the predicted value and a standard error that is calculated using the sum of squares.

Using a regression model to predict scores only works if the regression line is a good fit to the data. If this relationship is non-linear, we could either transform the data (i.e., a logarithmic transformation) or try one of the other regression equations that are available with Excel or a graphing calculator.

The standard deviations and the variances of each of these distributions for each of the predicted values are equal. This is called *homoscedasticity*.

Finally, for each given value of X, the values of Y are independent of each other.

Lesson Summary

When we estimate a linear regression model, we want to ensure that the regression coefficient for the population, β, does not equal zero. To do this, we perform a hypothesis test, where we set the regression coefficient equal to zero and test for significance.

For each predicted value, we have a normal distribution (also known as the conditional distribution, since it is conditional on the value of X) that describes the likelihood of obtaining other scores that are associated with the value of the predictor variable, X. We can use these distributions and the concept of standardized scores to make predictions about probability.

We can also build confidence intervals around the predicted values to give us a better idea about the ranges likely to contain a certain score.

We make several assumptions when dealing with a linear regression model including:

At each value of X, there is a distribution of Y.

A regression line is a good fit to the data. There is homoscedasticity, and the observations are independent.

Review Questions

1. A college counselor is putting on a presentation about the financial benefits of further education and takes a random sample of 120 parents. Each parent was asked a number of questions, including the number of years of education that he or she has (including college) and his or her yearly income (recorded in the thousands of dollars). The summary data for this survey are as follows:

$$n = 120 \quad r = 0.67 \quad \sum x = 1,782 \quad \sum y = 1,854 \quad s_X = 3.6 \quad s_Y = 4.2 \quad s_{XY} = 3.12 \quad SS_X = 1542$$

(a) What is the predictor variable? What is your reasoning behind this decision?

(b) Do you think that these two variables (income and level of formal education) are correlated? Is so, please describe the nature of their relationship.

(c) What would be the regression equation for predicting income, Y, from the level of education, X?

(d) Using this regression equation, predict the income for a person with 2 years of college (13.5 years of formal education).

(e) Test the null hypothesis that in the population, the regression coefficient for this scenario is zero.

- First develop the null and alternative hypotheses.
- Set the critical value to $\alpha = 0.05$.
- Compute the test statistic.
- Make a decision regarding the null hypothesis.

(f) For those parents with 15 years of formal education, what is the percentage who will have an annual income greater than $18,500?

(g) For those parents with 12 years of formal education, what is the percentage who will have an annual income greater than $18,500?

(h) Develop a 95% confidence interval for the predicted annual income when a parent indicates that he or she has a college degree (i.e., 16 years of formal education).

(i) If you were the college counselor, what would you say in the presentation to the parents and students about the relationship between further education and salary? Would you encourage students to further their education based on these analyses? Why or why not?

9.4 Multiple Regression

Learning Objectives

- Understand a multiple regression equation and the coefficients of determination for correlation of three or more variables.
- Calculate a multiple regression equation using technological tools.
- Calculate the standard error of a coefficient, test a coefficient for significance to evaluate a hypothesis, and calculate the confidence interval for a coefficient using technological tools.

Introduction

In the previous sections, we learned a bit about examining the relationship between two variables by calculating the correlation coefficient and the linear regression line. But, as we all know, often times we work with more than two variables. For example, what happens if we want to examine the impact that class size and number of faculty members have on a university's ranking. Since we are taking multiple variables into account, the linear regression model just won't work. In multiple linear regression, scores for one variable are predicted (in this example, a university's ranking) using multiple predictor variables (class size and number of faculty members).

Another common use of multiple regression models is in the estimation of the selling price of a home. There are a number of variables that go into determining how much a particular house will cost, including the square footage, the number of bedrooms, the number of bathrooms, the age of the house, the neighborhood, and so on. Analysts use multiple regression to estimate the selling price in relation to all of these different types of variables.

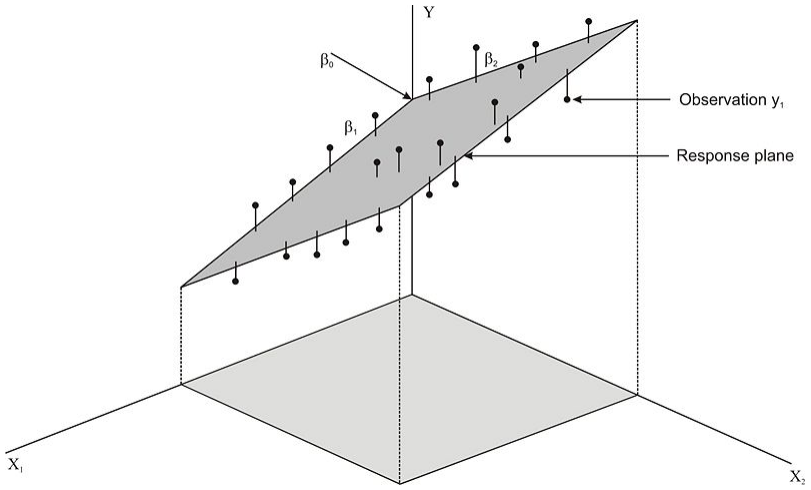
In this section, we will examine the components of a multiple regression equation, calculate an equation using technological tools, and use this equation to test for significance in order to evaluate a hypothesis.

Understanding a Multiple Regression Equation

If we were to try to draw a *multiple regression* model, it would be a bit more difficult than drawing a model for linear regression. Let's say that we have two predictor variables, X_1 and X_2, that are predicting the desired variable, Y. The regression equation would be as follows:

$$\hat{Y} = b_1 X_1 + b_2 X_2 + a$$

When there are two predictor variables, the scores must be plotted in three dimensions (see figure below). When there are more than two predictor variables, we would continue to plot these in multiple dimensions. Regardless of how many predictor variables there are, we still use the least squares method to try to minimize the distance between the actual and predicted values.



When predicting values using multiple regression, we first use the standard score form of the regression equation, which is shown below:

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

where:

$\hat{Y}$ is the predicted variable, or criterion variable.

β_i is the i^{th} regression coefficient.

X_i is the i^{th} predictor variable.

To solve for the regression and constant coefficients, we need to determine multiple correlation coefficients, r, and coefficients of determination, also known as proportions of shared variance, r^2. In the linear regression model, we measured r^2 by adding the squares of the distances from the actual points to the points predicted by the regression line. So what does r^2 look like in the multiple regression model? Let's take a look at the figure above. Essentially, like in the linear regression model, the theory behind the computation of a multiple regression equation is to minimize the sum of the squared deviations from the observations to the regression plane.

In most situations, we use a computer to calculate the multiple regression equation and determine the coefficients in this equation. We can also do multiple regression on a TI-83/84 calculator. (This program can be downloaded.)

Technology Note: Multiple Regression Analysis on the TI-83/84 Calculator

http://www.wku.edu/~david.neal/manual/ti83.html

Download a program for multiple regression analysis on the TI-83/84 calculator by first clicking on the link above.

It is helpful to explain the calculations that go into a multiple regression equation so we can get a better understanding of how this formula works.

After we find the correlation values, r, between the variables, we can use the following formulas to determine the regression coefficients for the predictor variables, X_1 and X_2:

$$\beta_1 = \frac{r_{Y1} - (r_{Y2})(r_{12})}{1 - r_{12}^2}$$
$$\beta_2 = \frac{r_{Y2} - (r_{Y1})(r_{12})}{1 - r_{12}^2}$$

where:

β_1 is the correlation coefficient for X_1.

β_2 is the correlation coefficient for X_2.

r_{Y1} is the correlation between the criterion variable, Y, and the first predictor variable, X_1.

r_{Y2} is the correlation between the criterion variable, Y, and the second predictor variable, X_2.

r_{12} is the correlation between the two predictor variables, X_1 and X_2.

After solving for the beta coefficients, we can then compute the b coefficients by using the following formulas:

$$b_1 = \beta_1 \left(\frac{s_Y}{s_1} \right)$$
$$b_2 = \beta_2 \left(\frac{s_Y}{s_2} \right)$$

where:

s_Y is the standard deviation of the criterion variable, Y.

s_1 is the standard deviation of the particular predictor variable (1 for the first predictor variable, 2 for the second, and so on).

After solving for the regression coefficients, we can finally solve for the regression constant by using the formula shown below, where k is the number of predictor variables:

$$a = \bar{y} - \sum_{i=1}^{k} b_i \bar{x}_i$$

Again, since these formulas and calculations are extremely tedious to complete by hand, we usually use a computer or a TI-83/84 calculator to solve for the coefficients in a multiple regression equation.

Calculating a Multiple Regression Equation using Technological Tools

As mentioned, there are a variety of technological tools available to calculate the coefficients in a multiple regression equation. When using a computer, there are several programs that help us calculate the multiple regression equation, including Microsoft Excel, the Statistical Analysis Software (SAS), and the Statistical Package for the Social Sciences (SPSS). Each of these programs allows the user to calculate the multiple regression equation and provides summary statistics for each of the models.

For the purposes of this lesson, we will synthesize summary tables produced by Microsoft Excel to solve problems with multiple regression equations. While the summary tables produced by the different technological tools differ slightly in format, they all provide us with the information needed to build a multiple regression equation, conduct hypothesis tests, and construct confidence intervals. Let's take a look at an example of a summary statistics table so we get a better idea of how we can use technological tools to build multiple regression equations.

Example: Suppose we want to predict the amount of water consumed by football players during summer practices. The football coach notices that the water consumption tends to be influenced by the time that the players are on the field and by the temperature. He measures the average water consumption, temperature, and practice time for seven practices and records the following data:

TABLE 9.12:

Temperature (degrees F)	Practice Time (hrs)	H_2O Consumption (in ounces)
75	1.85	16
83	1.25	20
85	1.5	25
85	1.75	27
92	1.15	32
97	1.75	48
99	1.6	48

Figure: Water consumption by football players compared to practice time and temperature.

Technology Note: Using Excel for Multiple Regression

- Copy and paste the table into an empty Excel worksheet.
- Click the Data choice on the toolbar, then select 'Data Analysis,' and then choose 'Regression' from the list that appears (Note, if Data Analysis does not appear as a choice on your Data page need to follow the add-in instructions below).
- Place the cursor in the 'Input Y range' field and select the third column.
- Place the cursor in the 'Input X range' field and select the first and second columns.
- Place the cursor in the 'Output Range' field and click somewhere in a blank cell below and to the left of the table.
- Click 'Labels' so that the names of the predictor variables will be displayed in the table.
- Click 'OK', and the results shown below will be displayed.

Note: In Excel 2007, to add **Data Analysis** to your **Data** page, perform the following functions. Click the **Microsoft Office Button** in the upper left, then click on **Excel Options**. Click on **Add-ins**, then highlight the **Analysis ToolPak**, click **Go**, make sure the **Analysis ToolPak box** is checked off, and then click **OK**. The **Data Analysis** choice should now appear on your Excel **Data** page. Follow the remaining instructions above.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.996822
R Square	0.993654
Adjusted R Square	0.990481
Standard Error	1.244877
Observations	7

TABLE 9.13:

	Df	SS	MS	F	**Significance F**
Regression	2	970.6583	485.3291	313.1723	4.03E-05
Residual	4	6.198878	1.549719		
Total	6	976.8571			

TABLE 9.14:

	Coefficients	Standard Error	*t* Stat	*P*- value	Lower 95%	Upper 95%
Intercept	−121.655	6.540348	−18.6007	4.92e-05	−139.814	−103.496
Temperature	1.512364	0.060771	24.88626	1.55E-05	1.343636	1.681092
Practice Time	12.53168	1.93302	6.482954	0.002918	7.164746	17.89862

In this example, we have a number of summary statistics that give us information about the regression equation. As you can see from the results above, we have the regression coefficient and standard error for each variable, as well as the value of r^2. We can take all of the regression coefficients and put them together to make our equation.

Using the results above, our regression equation would be $\hat{Y} = -121.66 + 1.51(\text{Temperature}) + 12.53(\text{Practice Time})$.

Each of the regression coefficients tells us something about the relationship between the predictor variable and the predicted outcome. The temperature coefficient of 1.51 tells us that for every 1.0-degree increase in temperature, we predict there to be an increase of 1.5 ounces of water consumed, if we hold the practice time constant. Similarly, we find that with every one-hour increase in practice time, we predict players will consume an additional 12.53 ounces of water, if we hold the temperature constant. That equates to about 2.1 extra ounces of water for every 10 minutes increase in practice time.

With a value of 0.99 for r^2, we can conclude that approximately 99% of the variance in the outcome variable, Y, can be explained by the variance in the combined predictor variables. With a value of 0.99 for r^2, we can conclude that almost all of the variance in water consumption is attributed to the variance in temperature and practice time.

Testing for Significance to Evaluate a Hypothesis, the Standard Error of a Coefficient, and Constructing Confidence Intervals

When we perform multiple regression analysis, we are essentially trying to determine if our predictor variables explain the variation in the outcome variable, Y. When we put together our final equation, we are looking at whether or not the variables explain most of the variation, r^2, and if this value of r^2 is statistically significant. We can use technological tools to conduct a hypothesis test, testing the significance of this value of r^2, and construct confidence intervals around these results.

Hypothesis Testing

When we conduct a hypothesis test, we test the null hypothesis that the multiple r-value in the population equals zero, or $H_0 : r_{\text{pop}} = 0$. Under this scenario, the predicted values, or fitted values, would all be very close to the mean, and the deviations, $\hat{Y} - \bar{Y}$, and the sum of the squares would be close to 0. Therefore, we want to calculate a test statistic (in this case, the *F-statistic*) that measures the correlation between the predictor variables. If this test statistic is beyond the critical values and the null hypothesis is rejected, we can conclude that there is a nonzero relationship between the criterion variable, Y, and the predictor variables. When we reject the null hypothesis, we can say something like, "The probability that r^2 having the value obtained would have occurred by chance if the null hypothesis were true is less than 0.05 (or whatever the significance level happens to be)." As mentioned, we can use computer programs to determine the F-statistic and its significance.

Let's take a look at the example above and interpret the F-statistic. We see that we have a very high value of r^2 of 0.99, which means that almost all of the variance in the outcome variable (water consumption) can be explained by the predictor variables (practice time and temperature). Our ANOVA (ANalysis Of VAriance) table tells us that we have a calculated F-statistic of 313.17, which has an associated probability value of 4.03e-05. This means that the probability that 99 percent of the variance would have occurred by chance if the null hypothesis were true (i.e., none of the variance was explained) is 0.0000403. In other words, it is highly unlikely that this large level of variance was by chance. F-distributions will be discussed in greater detail in a later chapter.

Standard Error of a Coefficient and Testing for Significance

In addition to performing a test to assess the probability of the regression line occurring by chance, we can also test the significance of individual coefficients. This is helpful in determining whether or not the variable significantly contributes to the regression. For example, if we find that a variable does not significantly contribute to the regression, we may choose not to include it in the final regression equation. Again, we can use computer programs to determine the standard error, the test statistic, and its level of significance.

Example: Looking at our example above, we see that Excel has calculated the standard error and the test statistic (in this case, the t-statistic) for each of the predictor variables. We see that temperature has a t-statistic of 24.88 and a corresponding P-value of 1.55e-05. We also see that practice time has a t-statistic of 6.48 and a corresponding P-value of 0.002918. For this situation, we will set α equal to 0.05. Since the P-values for both variables are less than $\alpha = 0.05$, we can determine that both of these variables significantly contribute to the variance of the outcome variable and should be included in the regression equation.

Calculating the Confidence Interval for a Coefficient

We can also use technological tools to build a confidence interval around our regression coefficients. Remember, earlier in the chapter we calculated confidence intervals around certain values in linear regression models. However, this concept is a bit different when we work with multiple regression models.

For a predictor variable in multiple regression, the confidence interval is based on a t-test and is the range around the observed sample regression coefficient within which we can be 95% (or any other predetermined level) confident that the real regression coefficient for the population lies. In this example, we can say that we are 95% confident that the population regression coefficient for temperature is between 1.34 (the Lower 95% entry) and 1.68 (the Upper 95% entry). In addition, we are 95% confident that the population regression coefficient for practice time is between 7.16 and 17.90.

Lesson Summary

In multiple linear regression, scores for the criterion variable are predicted using multiple predictor variables. The regression equation we use for two predictor variables, X_1 and X_2, is as follows:

$$\hat{Y} = b_1 X_1 + b_2 X_2 + a$$

When calculating the different parts of the multiple regression equation, we can use a number of computer programs, such as Microsoft Excel, SPSS, and SAS.

These programs calculate the multiple regression coefficients, the combined value of r^2, and the confidence intervals for the regression coefficients.

On the Web

www.wku.edu/~david.neal/web1.html

Manuals by a professor at Western Kentucky University for use in statistics, plus TI-83/84 programs for multiple regression that are available for download.

http://education.ti.com/educationportal/activityexchange/activity_list.do

Texas Instrument Website that includes supplemental activities and practice problems using the TI-83 calculator.

Review Questions

1. A lead English teacher is trying to determine the relationship between three tests given throughout the semester and the final exam. She decides to conduct a mini-study on this relationship and collects the test data (scores for Test 1, Test 2, Test 3, and the final exam) for 50 students in freshman English. She enters these data into Microsoft Excel and arrives at the following summary statistics:

Multiple R	0.6859
R Square	0.4707
Adjusted R Square	0.4369
Standard Error	7.5718
Observations	50

TABLE 9.15: ANOVA

	Df	*SS*	*MS*	*F*	**Significance *F***
Regression	3	2342.7228	780.9076	13.621	0.0000
Residual	46	2637.2772	57.3321		
Total	49	4980.0000			

TABLE 9.16:

	Coefficients	**Standard Error**	***t* Stat**	***P* -value**
Intercept	10.7592	7.6268		
Test 1	0.0506	0.1720	0.2941	0.7700
Test 2	0.5560	0.1431	3.885	0.0003
Test 3	0.2128	0.1782	1.194	0.2387

(a) How many predictor variables are there in this scenario? What are the names of these predictor variables?

(b) What does the regression coefficient for Test 2 tell us?

(c) What is the regression model for this analysis?

(d) What is the value of r^2, and what does it indicate?

(e) Determine whether the multiple r-value is statistically significant.

(f) Which of the predictor variables are statistically significant? What is the reasoning behind this decision?

(g) Given this information, would you include all three predictor variables in the multiple regression model? Why or why not?

Keywords

Bivariate data

Coefficient of determination

Conditional distribution

Correlation

Correlation coefficient

Criterion variable

Curvilinear relationship

e

F-statistic

Homoscedasticity

Least squares line

Line of best fit

Linear regression

Linear regression line

Magnitude

Method of least squares

Multiple regression

Near-zero correlation

Negative correlation

Outcome variable

Outlier

Pearson product-moment correlation coefficient

Perfect correlation

Positive correlation

Predictor variable

r

r^2

Regression coefficient

Regression constant

Residual values

Scatterplots

Transformation

Zero correlation