CHAPTER 8 Linear Models

Chapter Outline

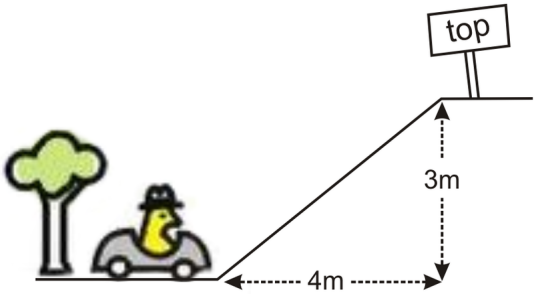
8.1 Review of Rate of Change

Learning Objectives

- Review rates of change.
- Interpret graphs and compare rates of change.

Introduction

We come across many examples of slope in everyday life. For example, a slope is in the pitch of a roof, the grade or incline of a road, and the slant of a ladder leaning on a wall. In math, we use the word **slope** to define steepness in a particular way.



$$\text{Slope} = \frac{\text{distance moved vertically}}{\text{distance moved horizontally}}$$

This is often reworded to be easier to remember:

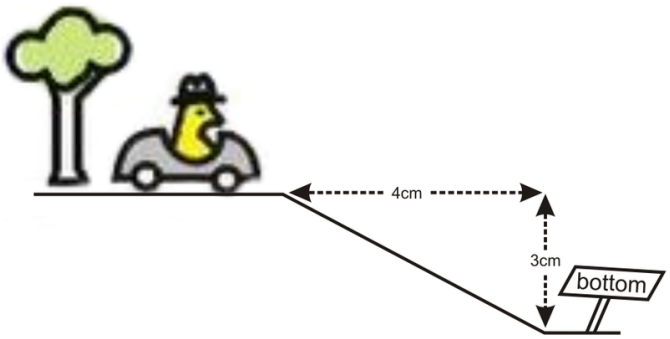
$$\text{Slope} = \frac{\text{rise}}{\text{run}}$$

Essentially, slope is the change in y if x increases by 1.

In the picture above, the slope would be the ratio of the **height** of the hill (the *rise*) to the horizontal **length** of the hill (the *run*).

$$\text{Slope} = \frac{rise}{run} = \frac{\triangle y}{\triangle x} = \frac{3}{4} = 0.75$$

If the car were driving to the *right* it would *climb* the hill. We say this is a positive slope. Anytime you see the graph of a line that goes up as you move to the right, the slope is **positive**.



If the car were to keep driving after it reached the top of the hill, it may come down again. If the car is driving to the *right* and *descending*, then we would say that the slope is **negative**. The picture above has a **negative slope** of -0.75.

So as we move from left to right, positive slopes increase while negative slopes decrease.

Find a Rate of Change

The slope of a function that describes real, measurable quantities is often called a **rate of change**. In that case, the slope refers to a change in one quantity (y) *per* unit change in another quantity (x).

Example A

Andrea has a part time job at the local grocery store. She saves for her vacation at a rate of $15 every week. Express this rate as money saved *per day* and money saved *per year*.

Converting rates of change is fairly straight forward so long as you remember the equations for rate (i.e. the equations for slope) and know the conversions. In this case 1 week $=$ 7 days and 52 weeks $=$ 1 year.

$$\text{rate} = \frac{\$15}{1 \text{ week}} \cdot \frac{1 \text{ week}}{7 \text{ days}} = \frac{\$15}{7 \text{ days}} = \frac{15}{7} \text{ dollars per day} \approx \$2.14 \text{ per day}$$

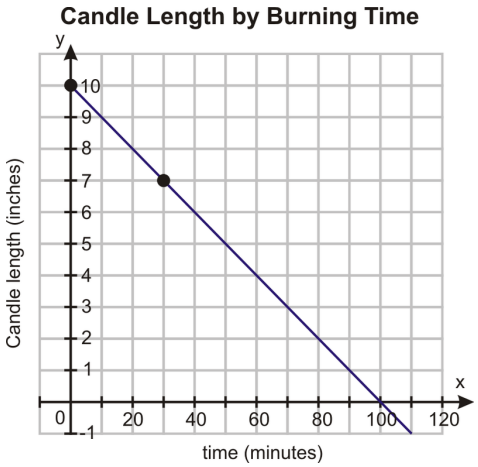
$$\text{rate} = \frac{\$15}{1 \text{ week}} \cdot \frac{52 \text{ week}}{1 \text{ year}} = \$15 \cdot \frac{52}{\text{year}} = \$780 \text{ per year}$$

Example B

A candle has a starting length of 10 inches. Thirty minutes after lighting it, the length is 7 inches. Determine the rate of change in length of the candle as it burns. Determine how long the candle takes to completely burn to nothing.

In this case, we will graph the function to visualize what is happening.

We have two points to start with. We know that at the moment the candle is lit $(\text{time} = 0)$ the length of the candle is 10 inches. After thirty minutes $(\text{time} = 30)$ the length is 7 inches. Since the candle length is a function of time we will plot time on the horizontal axis, and candle length on the vertical axis. Here is a graph showing this information.



The rate of change of the candle is simply the slope. Since we have our two points $(x_1, y_1) = (0, 10)$ and $(x_2, y_2) = (30, 7)$ we can move straight to the formula.

$$\text{Rate of change} = \frac{\triangle y}{\triangle x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{(7\ inches) - (10\ inches)}{(30\ minutes) - (0\ minutes)} = \frac{-3\ inches}{30\ minutes} = -0.1\ inches\ per\ minute$$

The slope is negative. A negative rate of change means that the quantity is decreasing with time.

We can also convert our rate to inches per hour.

$$\text{rate} = \frac{-0.1\ inches}{1\ minute} \cdot \frac{60\ minutes}{1\ hour} = \frac{-6\ inches}{1\ hour} = -6\ inches\ per\ hour$$

To find the point when the candle burns to nothing, or reaches zero length, we can read off the graph (100 minutes). We can use the rate equation to verify this algebraically.

$$\text{rate} \times \text{time} = \text{Length burned}$$
$$0.1 \times 100 = 10$$

Since the candle length was originally 10 inches this confirms that 100 minutes is the correct amount of time.

Lesson Summary

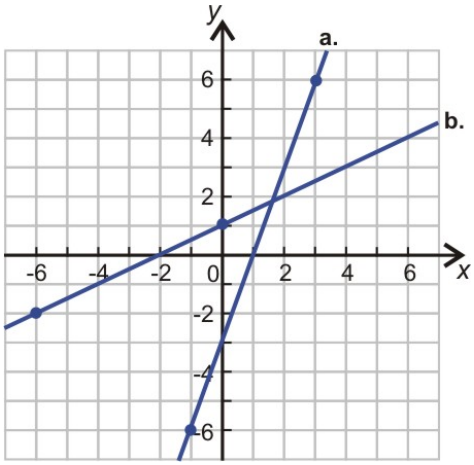
- Slope is a measure of change in the vertical direction for each step in the horizontal direction.
- Slope $= \frac{rise}{run}$ or $slope = \frac{\triangle y}{\triangle x}$
- The slope between two points (x_1, y_1) and $(x_2, y_2) = \frac{y_2 - y_1}{x_2 - x_1}$

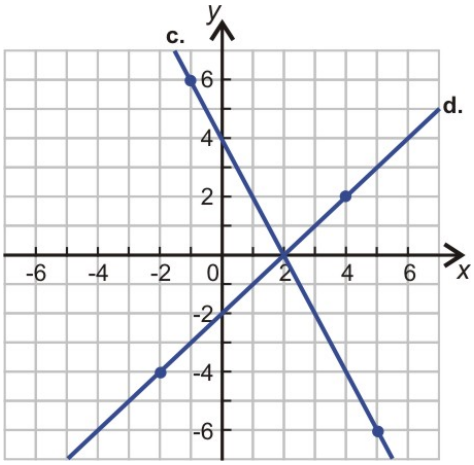
Review Questions

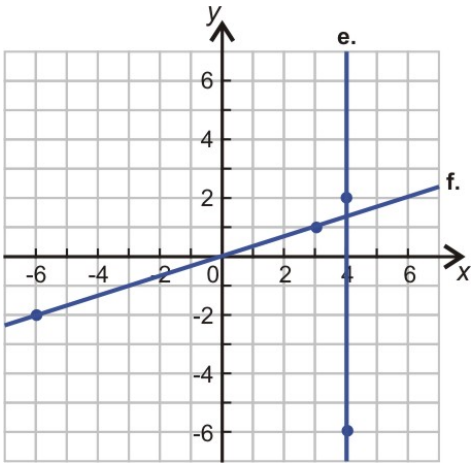
Use the slope formula to find the slope of the line that passes through each pair of points.

1. (-5, 7) and (0, 0)
2. (-3, -5) and (3, 11)

3. (3, -5) and (-2, 9)
4. (-5, 7) and (-5, 11)
5. (9, 9) and (-9, -9)
6. (3, 5) and (-2, 7)
7. Use the points indicated on each line of the graphs to determine the slopes of the following lines.







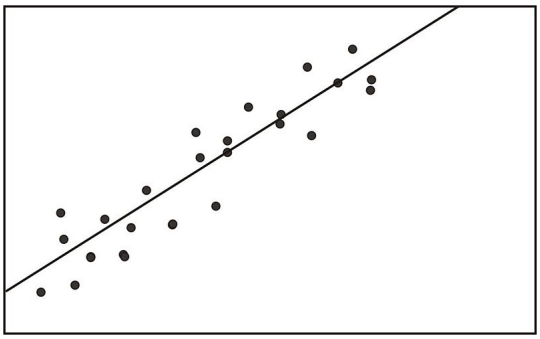
8.2 Linear Regression Models

Learning Objectives

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

Introduction

Earlier we learned that correlation could be used to assess the strength and direction of a linear relationship between two variables. We llustrated 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**, or the **line of best fit**.

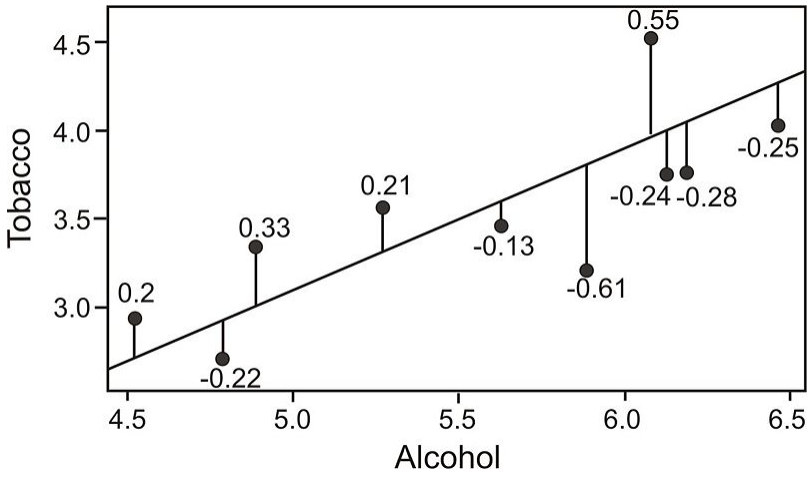


Generating and Graphing the Regression Line

Linear regression involves using data to calculate a line that best fits that data, and then using that line to predict scores on one variable from another. Prediction is simply the process of estimating scores of the **outcome** (or dependent) variable based on the scores of the **predictor** (or independent) variable.

To generate the regression line, we look for a line of best fit. There are many ways one could define this best fit. Statisticians define the best-fit line to be the one that minimizes the sum of the squared distances from the observed data to the 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**.

Our goal in the method of least squares is 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. The smaller the residuals, the better the fit.



As you can see, the regression line is a straight line that expresses the relationship between two variables. 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.

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:

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

where:

$\hat{Y}$ (pronounced Y-hat) is the score that we are trying to predict.

b is the slope of the line; it is also called the **regression coefficient**.

a is the y-intercept; it is also called the *regression constant*. It is the value of Y when the value of X is 0.

Linear Models

In the previous chapter, we discussed the average rate of change of a function on an interval. For many functions, the average rate of change is different on different intervals. A linear function, however, has the same average rate of change on every interval. When a linear model is used to describe data, it is assuming a **constant rate of change**.

The general formula of the linear model allows for easy calculation and interpretation of both the slope (the constant rate of change) and the y-intercept. In this model:

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

the slope, ***b***, tells us the change in the dependent (Y) variable for every unit change in the independent variable (X), and the y-intercept, **a,** tells us the value of Y when X=0.

Example A

Imagine that a town of 30,000 people grows by 2,000 people each year. Since the population, P, is growing at a constant rate of 2,000 people per year, P is a linear function of time, t. To generate the equation of this model, we calculate the slope and y-intercept.

Solution

What is the *slope*? Remember that the slope of a linear function is the average rate of change. We know that average rate of change of the population is 2,000 people per year. Therefore, b=2,000.

What is the *y-intercept*? The y-intercept is the size of the population at time=0. We should treat t=0 as the initial size of the town's population. So, in this problem, P=30,000.

This means our equation looks like this:

$$\hat{P} = 2,000t + 30,000$$

We can then easily calcuate the size of the population in 10 years. At a steady growth rate of 2,000 people per year, the town will grow from 30,000 to 50,000 over 10 years.

Generating a Linear Model from Raw Data

To calculate a regression line from our data, we need to find the values for b and a. To calculate the regression coefficient b (or slope), 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.

To calculate the regression constant a (or y-intercept), we use the following formula:

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

or

$$a = \bar{y} - b\bar{x}$$

Example B

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

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2

TABLE 8.1: (continued)

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.

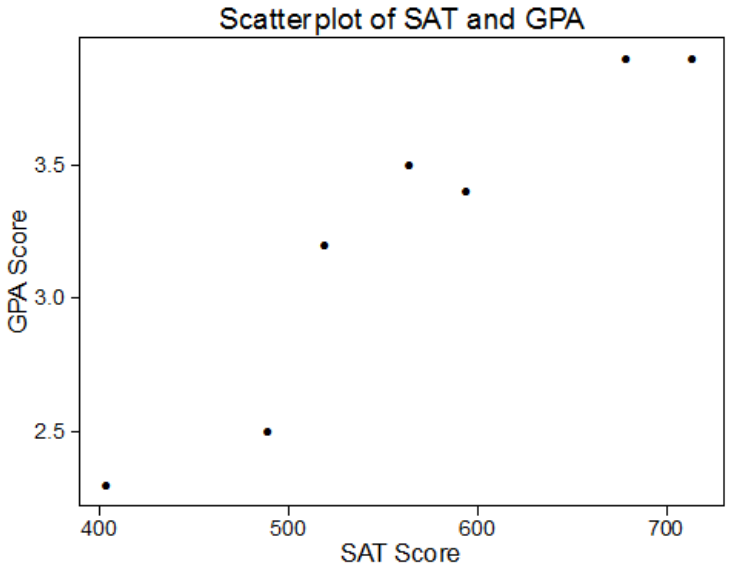


FIGURE 8.1

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, $\hat{Y} = 0.0056X + 0.094$.

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).

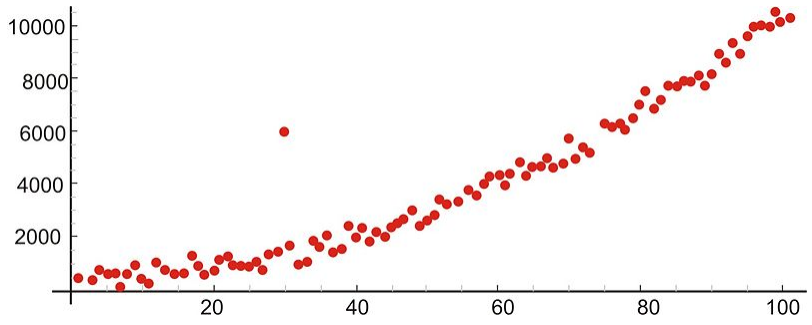
TABLE 8.2: GPA/SAT data, including predicted GPA values from the linear regression.

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

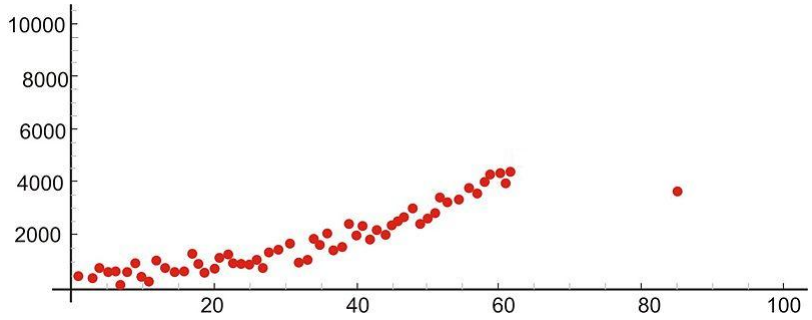
An **outlier** is an extreme observation that does not fit the general correlation or regression pattern (see figure below). In the regression setting, outliers will be far away from the regression line in the y-direction. Since it is an unusual observation, the inclusion of an outlier may affect the 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.

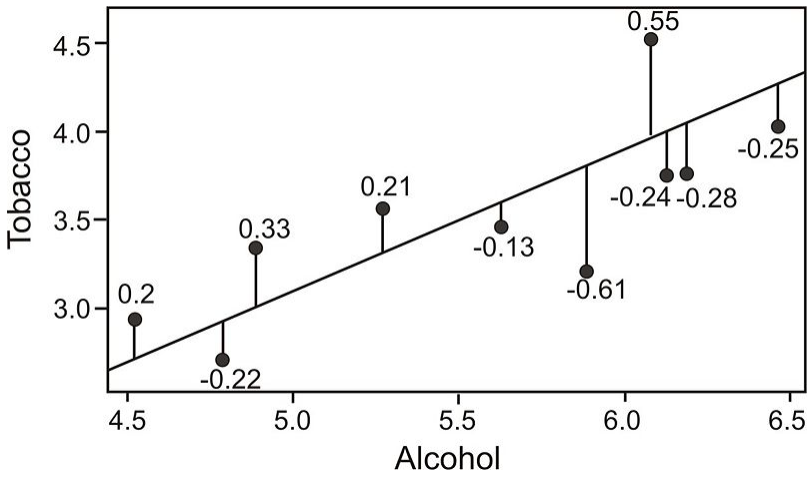
An **influential point** in regression is one whose removal would greatly impact the equation of the regression line. Usually, an influential point will be separated in the x direction from the other observations. It is possible for an outlier to be an influential point. However, there are some influential 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.

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, since, again, 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 C

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

TABLE 8.3: 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.

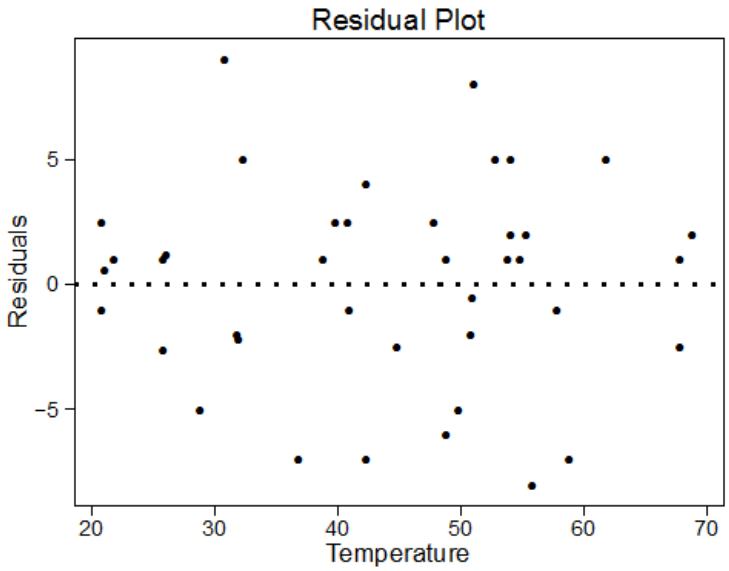


FIGURE 8.2

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 $\hat{y}$.

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.

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 8.4: 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
11	4	2
12	1	8
13	1	10
14	1	12
15	2	8

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