Chapter 5 – Linear Quantitative Relationships

Introduction: In chapter 1, we have looked at how to analyze a categorical data set. Then we looked at relationships between categorical data sets in chapter 2. In chapters 3 and 4, we looked at how to analyze quantitative data. It follows that now we are ready to look at relationships between quantitative data sets.

There are many different types of quantitative relationships that statisticians study. We will focus on the most common in this chapter, which is the study of linear relationships between quantitative variables.

Algebra requirement: Algebra classes study the subject of lines. However, they do not study lines the way statisticians and data scientists study lines. As with most things, statistics is the study of world around us with real data and real applications. For example, algebra classes may study slope between two points. A statistician studies the slope as an average rate of change between thousands or even millions of points based on real data. An algebra class may find the equation of a line between two points. A statistician studies linear prediction formulas created from thousands of points, uses those formulas to predict world climate changes, and studies the accuracy of those predictions with residual analysis.

The point is that while a basic understanding of slope and lines is helpful, it is not necessary to understanding this chapter. The study of linear quantitative relationships in statistics is extremely different from what someone studies in algebra. It almost feels like a completely new subject about how to see the world around us through fresh eyes.
Section 5A – Introduction to Quantitative Relationships, Explanatory and Response Variables, Scatterplots with Technology

Remember quantitative data is numerical measurement data, not categories. The numbers in the data set should measure something. They often have units and we should be able to take an average in context.

In this section, we will be focusing on two different quantitative variables with different units. It is much easier to compare the average salary in thousands of dollars from people in Arizona to the average salary in thousands of dollars from people in New Mexico. The two data sets have the same units and can be compared directly. For example, we can determine if the average salary of the people from Arizona is higher or lower than the average salary from the people from New Mexico.

When the units are different, you cannot just compare the centers or spreads. It becomes a much more complicated process. If you look at countries around the world and study the relationship between their unemployment rates and their national debts in millions of dollars, you cannot compare the national debt in millions of dollars to the unemployment rate percentage directly. They are completely different things.

So how do we analyze the relationships between two different quantitative variables? We will start by assigning one variable to be the explanatory variable and one variable to be the response variable.

Explanatory and Response Variables

In algebra classes, we are often given an X and a Y variable and asked to plot a couple points. In statistics, we know it is not so simple. In statistics, we often call the X variable the “explanatory variable” or “independent variable”. We call the Y variable the “response variable” or “dependent variable”. I prefer explanatory and response because the terms independent and dependent can be confusing to students when they study the subject of independence. Real quantitative relationship analysis requires some serious thought about which variable should be the explanatory variable (X) and which variable should be the response variable (Y).

Guidelines for choosing the explanatory (X) and the response (Y)

1. The response variable should respond.

Often business analysis involves studying the costs or profits of company over a period of several months or years. Should we assign the costs to be the explanatory variable (X) or the response variable (Y)? What about the time in months? Think of it this way. Does time respond to the costs of the company? Probably not. That does not sound right. Do the costs respond to time? That may be true. Whichever variable responds to the other should probably be your
response variable. In this case, I should assign the time (months) as my explanatory variable (X) and the costs (thousands of dollars) as my response variable (Y).

2. The response variable should be the focus of your study or the variable you may want to make predictions about.

Let us look at the example of the unemployment rates and national debts of various countries. Those variables may relate to each other. In other words either variable could be the response variable. In that case, pick the variable you are most interested in to be the response variable (Y). I was studying unemployment rates in various countries and wanted to see if the national debt was related to unemployment. I was also interested in trying to predict unemployment rates with my prediction equation. Since the focus of my study was unemployment and I wanted to eventually make predictions about unemployment, I let the unemployment rates be my response variable (Y). Therefore, my explanatory variable (X) will be the national debts.

Ordered Pairs
Once you have chosen which variable is X and which variable is Y, you will need to find ordered pair data. Ordered pair data pairs an X in the first data set with a Y value in the second data set. There needs to be some kind of relationship between them. For example, I do not want to pair 20 random unemployment rates with 50 national debts. First there needs to be the same number of X and Y values. Computer programs will give error messages if the frequency N for one quantitative variable is not the same as the frequency N for the other data set. If I want to study the relationship between national debt and unemployment rates, I do not want to pair any national debt with any unemployment rate. I want to collect the data together. The national debt and unemployment rate should come from the same country and hopefully the same year. I went from country to country and looked up their estimated national debt and their unemployment rate at the same time (August 2017). This is difficult to do. Getting data is a difficult job. Websites, articles and various sources often disagree with one another. Therefore, these values are just approximations and may not be perfectly accurate.

<table>
<thead>
<tr>
<th>Country</th>
<th>August 2017 National Debt (Billions of U.S. Dollars)</th>
<th>Unemployment Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>2472.9</td>
<td>9.6</td>
</tr>
<tr>
<td>Mexico</td>
<td>474.5</td>
<td>3.5</td>
</tr>
<tr>
<td>U.S.A</td>
<td>19873.5</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>9094.9</td>
<td>3.06</td>
</tr>
<tr>
<td>Canada</td>
<td>826.5</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>406.0</td>
<td></td>
</tr>
<tr>
<td>United Kingdom</td>
<td>2279.8</td>
<td></td>
</tr>
</tbody>
</table>
You can now write an ordered pair. They are often written in the form \((X, Y)\). So for Mexico’s data I would write the ordered pair as \((474.5 \text{ Billion } \$, 3.5\% \text{ unemployment})\) and Japan’s data as \((9094.9 \text{ Billion } \$, 3.06\% \text{ unemployment})\). Notice the first number in the pair describes the explanatory variable \((X)\) and is often called the “X coordinate”. The second number in the ordered pair describes the response variable \((Y)\) and is often called the “Y coordinate”.

**Rectangular Coordinate System**

Once you have your ordered pairs, you can graph them. To graph quantitative variables with different units, we will need both an X-axis and a Y-axis.

Notice to graph the point \((2, 3)\) we find 2 on the X-axis and 3 on the Y-axis and then the point would be where they meet. Notice that to make the ordered pair, a rectangle is created. That is why this system of graphing with X and Y-axes is often called the “rectangular coordinate system”.

The key is the units though. Always pay close attention to the units for your explanatory and response variables. For example, the x-axis could be describing temperature in degrees Fahrenheit and the y-axis could be describing profits in thousands of dollars. So \((2 , 3)\) is really describing the ordered pair \((2 \text{ degrees Fahrenheit}, 3 \text{ thousand dollars})\) and \((-3 , 1)\) is really describing the ordered pair \((-3 \text{ degrees Fahrenheit}, 1 \text{ thousand dollars})\).

This chapter is from *Introduction to Data Analysis*, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017
Scatterplots
There are many types of graphs statisticians look at when studying relationships between quantitative variables with different units. The most important graph though is the scatterplot. We said in the last couple of chapters that the first step when analyzing quantitative data is to find the shape. The scatterplot is a graph of the ordered pairs on the rectangular coordinate system. This graph shows the shape of the quantitative relationship.

We should again use technology to create a scatterplot. Once you have collected your ordered pair data and chosen which column will be the explanatory (X) and which will be the response (Y), you can create a scatterplot with any statistics software.

Create Scatterplot with Statcato:

Example 1
Let us look at the health data again. Statistics analysis always starts with a question, even if it is a question in your own mind. My first question was to see if there is a relationship between the weight of a man and his cholesterol.

Step 1: First notice that the health data does have ordered pair data containing the weight and cholesterol of the same 40 men. Having ordered pair data is vital to studying quantitative relationships. I cannot take a weight of one man and pair it with the cholesterol of a different man. The values need to come from the same person.

Here are the weights and cholesterol values for the first ten men in the data set so you can see the ordered pair structure.

<table>
<thead>
<tr>
<th>Men’s Weight (Lbs.)</th>
<th>Men’s Cholesterol (mg per deciliter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>169.1</td>
<td>522</td>
</tr>
<tr>
<td>144.2</td>
<td>127</td>
</tr>
<tr>
<td>179.3</td>
<td>740</td>
</tr>
<tr>
<td>175.8</td>
<td>49</td>
</tr>
<tr>
<td>152.6</td>
<td>230</td>
</tr>
<tr>
<td>166.8</td>
<td>316</td>
</tr>
<tr>
<td>135</td>
<td>590</td>
</tr>
<tr>
<td>201.5</td>
<td>466</td>
</tr>
<tr>
<td>175.2</td>
<td>121</td>
</tr>
<tr>
<td>139</td>
<td>578</td>
</tr>
</tbody>
</table>

Step 2: The next step is to choose which variable is to be the explanatory (X) and which variable should be the response variable (Y). These variables may respond to each other so I could
choose either variable to the response. I am interested in focusing on cholesterol. I also want to maybe predict cholesterol levels based on a man’s weight, so I will make cholesterol my response (Y). Therefore, I will make the weight the explanatory variable (X).

Step 3: Create a Scatterplot

Use technology to create the scatterplot. Statcato can make the scatterplot with the click of a couple buttons. Start by copy and pasting the two data sets into Statcato. Now click on the graph menu. Then click on scatterplot. You will need to choose which column is the X variable and which column is the Y-values. It is always important to label your graphs, especially the X and Y-axis, and to the give the graph a title. Statcato can find the best-fit line or curve that fits the scatterplot. These are sometimes called “regression lines” or “regression curves”. We have not discussed these yet, so let us uncheck the button that says, “Show regression curve”.

Creating a Scatterplot with Statcato: Graph => Scatterplot => Pick which column is X and which is Y => Add Series => Label your X-axis and Y-axis => Give graph a title => uncheck “show regression curve”

Here is the scatterplot with Statcato.
Shape of a scatterplot
When looking at the shape of a scatterplot, you have to forget about what you learned in Algebra classes. In algebra classes, we often start with a linear or curved function and find ordered pairs that lie on that line or curve. That is not how real data works in statistics. The dots will rarely go through a line or a curve. In some ways, statistics analysis is the opposite of algebra. Instead of focusing on a linear and curved function and then finding ordered pairs, in statistics, we start with a graph of all the real data ordered pairs and then find the line or curve that best fits all of the data. This can be a difficult process.

Key Question: Start by asking yourself a simple question. The points will not lie on a line or a curve, but can I imagine a line or a curve that the points could be relatively close to? That is the key. You have to take all of the points into account.

In the weight / cholesterol example, the points seem very scattered all over with no obvious pattern. Do not force a computer program like Statcato to draw a best-fit line or curve if there is no pattern in the data. The line or curve the computer draws will not be very accurate. This scatterplot tells that there is hardly any relationship between the weight of these men and the cholesterol of these men. Sometimes lighter men had a high cholesterol. Sometimes lighter men had a low cholesterol. Sometimes heavier men had a high cholesterol. Sometimes heavier men had a low cholesterol.

Example 2
Let us look at another example from the health data. This time I wanted to look at the weight of the men (in pounds) and the waist size of the men (in centimeters).

Step 1: Do I have ordered pair data? Yes. The health data contained the weights and waist sizes of the same 40 men.

Step 2: Pick which variable is the explanatory (X) and the response (Y). I was interested in predicting the weight of a man from his waist size. Remember the variable you want to predict and are most interested in should by your response (Y). So I picked the men’s waist size (in cm) to be the explanatory variable (X) and the men’s weight (in pounds) to be the response variable (Y).

Step 3: Make a scatterplot
I copy and pasted the men’s weight and waist size data in Statcato and then used the steps listed above to create the following scatterplot.
Step 4: Interpret the shape of the scatterplot.

Remember the points will not lie on a line or a curve. The key question is are they close to a line or a curve?

This scatterplot shows a distinct linear pattern. The points look like they are close to a line going up from left to right. This is often called a “positive linear relationship” or a “positive correlation”.

Example 3
To be run well, businesses often require a large amount of statistical analysis. Here is a scatterplot made from data describing the monthly costs of running a company. The data describes the month (X) and the costs (Y) in thousands of dollars. Notice month 0 is the initial cost of starting up the company.
Notice most of the points do seem to be close to a line. They are following a linear pattern. In fact, the linear pattern seems to be going down from left to right. We often call this a “negative linear relationship” or a “negative correlation”.

Unusual Value (Outlier): Notice there appears to be a point that does not follow the pattern. The company had an unusually high cost in the 15th month of operation. You may want to check with the company to determine if this was a mistake in the data. Maybe the cost was supposed to be 2.5 thousand dollars instead of 5.2 thousand dollars. However, this outlier was not a mistake. The company had some equipment break down and had to replace it. When studying quantitative relationships with scatterplots, it is important to look for these unusual values (outliers). Notice this point does not seem to follow the negative linear pattern.

Aside from the outlier, this graph is good news for the company. As the months increase, the costs of the company seem to be decreasing dramatically.

Example 4

A college started a solar energy program. Here is a scatterplot describing their data in 2009. The explanatory variable (X) was the # months in 2009 and the response variable was the solar energy generated from that month in kilowatt-hours (kWh).
Notice the points in the scatterplot do not seem to be close to a line, but there seems to be some relationship. It seems to follow a curve of some kind. You would have to use your imagination some, but can you draw a curve that might fit this data? Here is a possible curve.

Notice the curve seems to fit this data pretty well. The points are pretty close to the curve. You may remember from previous algebra classes that this “U” shaped curve is called a “parabola” or a “quadratic curve”.

Scatterplots and Shapes
We have seen in this section that to analyze two quantitative data sets with different units, we have to find ordered pair data and chose one variable to be the explanatory variable (X) and the
other variable to be the response variable (Y). We can then create a scatterplot to see if there is a relationship between the variables. We saw various possibilities for these quantitative relationships.

- No relationship at all: Points in scatterplot are spread out all over and do not seem to be close to any line or curve.
- Positive Linear Relationship: Points in the scatterplot seem to be close to a line that is going up from left to right (increasing). Look out for points that do not seem to fit the pattern (outliers).
- Negative Linear Relationship: Points in the scatterplot seem to be close to a line that is going down from left to right (decreasing). Look out for points that do not seem to fit the pattern (outliers).
- Curved Relationship: Points in the scatterplot seem to be close to a curve. There are many different types of curves possible when looking at data. Look out for points that do not seem to fit the pattern (outliers).

Problem Set Section 5A
Open the health data and create scatterplots with Statcato. Save the graphs on a word document or make a general sketch of the graph on a sheet of paper.

For each of the following problems:

- Tell which variable you chose to be the explanatory and which you chose to be the response variable and why.
- Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
- Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
- Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in Statcato and estimate the x and y coordinate for the outlier.

1. Explore the relationship between a woman’s height and weight.
2. Explore the relationship between a man’s height and weight.
3. Explore the relationship between a woman’s age and cholesterol.
4. Explore the relationship between a man’s age and cholesterol.

5. Explore the relationship between a woman’s weight and body mass index (BMI).

6. Explore the relationship between a man’s weight and body mass index (BMI).

7. Explore the relationship between a woman’s systolic blood pressure and her diastolic blood pressure.

8. Explore the relationship between a man’s systolic blood pressure and his diastolic blood pressure.

9. Explore the relationship between the length of a man’s leg and the circumference of his wrist.

10. Explore the relationship between a woman’s pulse and her cholesterol level.

---

**Section 5B – Strength and Direction of Linear Relationships and the Correlation Coefficient “r”**

We may be able to see if a scatterplot has a linear relationship, but it is hard to quantify how much of a linear relationship it has. Sometimes the scale can make it look like the points are not close to a line, when indeed they are. We need a way to measure the linear relationship.

Fortunately, there are ways statisticians measure the strength of a linear relationship. One of the statistics that measures quantitative relationships is the correlation coefficient “r”.

**Definition of the correlation coefficient (r):** The correlation coefficient “r” is a number between -1 and +1 that describes the strength and direction of the linear relationship. “r” values can tell us if the linear relationship is strong, moderate or weak, or does not exist. It can tell us if the linear relationship is positive (linear pattern going up from left to right) or negative (linear pattern going down from left to right).

**Interpreting the correlation coefficient “r”**

**Step 1:** Look at a scatterplot.

Always start by looking at a scatterplot. Have an idea of what the scatterplot looks like before you try to find and interpret the correlation coefficient.

**Step 2:** Calculate “r”

Once you have seen the scatterplot, use a statistics software to calculate the correlation coefficient “r”. Warning: The correlation coefficient “r” is extremely difficult and time
consuming to calculate. No data analyst or statistician calculate “r” with a formula and calculator, especially for big data sets. Always use a computer software program to do the difficult calculation and then focus on being able to interpret and explain the meaning of the correlation coefficient.

Statcato can calculate the “r” value very quickly. First copy and paste your ordered pair data into Statcato. Make sure you chose which variable will be X and which will be Y and then create a scatterplot to look at. Now go to the statistics menu and click on “correlation and regression”, then “linear (two variables)”. Now pick which column of data will be your explanatory variable (X) and which column of data will be your response variable (Y) and push the “add series” button. Now push “OK”.

To find the correlation coefficient “r” with Statcato: Statistics => Correlation and Regression => Linear (Two Variables) => Pick which column is X and which is Y => Add Series => OK

The printout from Statcato will give a lot of information about the relationship. Much of the information you will not understand until later on in your statistics class. Look for “r” under the “test statistic” table.

Step 3: Interpret what “r” is telling us about the quantitative relationship.

Let us see what the “r” value is telling us about the linear relationship. Correlation coefficients are difficult to read, but here are some general guidelines. These are not “set in stone” rules. The number of points in the data can make a difference in the interpretation of the correlation coefficient.

Notes about “r”

- r close to +1: This tells us that there is a strong positive correlation. Strong in the sense that the points are close to a line and positive means that the line is going up from left to right (increasing).
- r close to -1: This tells us that there is a strong negative correlation. Strong in the sense that the points are close to a line and negative means that the line is going down from left to right (decreasing).
- r close to 0: This tells us that there is no linear relationship between the variables.
- r values in between ± 0.6 to ± 1.0 are usually pretty strong. Again, the negative tells us the line is going down from left to right and the positive tells us the line is going up from left to right. The sign does not tell us the strength of the relationship.
- r values in between ± 0.4 to ± 0.5 are usually moderate in strength. This means there is a linear relationship, but it is not necessarily strong or weak. It is more in the middle.
- r values in between ± 0.2 to ± 0.3 are usually pretty weak. This means there is a linear relationship, but it is very weak.
• r values in between 0 to ± 0.1 usually tell us there is no linear relationship between the variables. Be careful of the signs when you get an “r” value close to zero. For example, an “r” value of -0.044 does not mean there is a negative linear relationship. Remember the “r” value usually needs to be around -0.2 to even be considered weak.

• The correlation coefficient will be the same if the X and Y are inversed. The calculation for “r” does not change if you make one variable the response or the other variable the response.

• I always find it is helpful to keep the following number line in mind when interpreting a correlation coefficient.

![Number line for correlation coefficients](image)

Example 1
In the previous section, we looked at men’s weight and cholesterol from the health data. I wanted to see if there is a relationship between the weight of a man and his cholesterol.

Since I was most interested in the cholesterol, I let the cholesterol be the response variable (Y) and the weight be the explanatory variable (X). We then used Statcato to create the following scatterplot. Here are the steps to creating a scatterplot in case you forgot.

Creating a Scatterplot with Statcato: Graph => Scatterplot => Pick which column is X and which is Y => Add Series => Label your X-axis and Y-axis => Give graph a title => uncheck “show regression curve”
In this scatterplot, the points seem very scattered all over with no obvious pattern. Now let us use Statcato to calculate the correlation coefficient \( r \).

To find the correlation coefficient \( r \) with Statcato: **Statistics \( \Rightarrow \) Correlation and Regression \( \Rightarrow \) Linear (Two Variables) \( \Rightarrow \) Pick which column is X and which is Y \( \Rightarrow \) Add Series \( \Rightarrow \) OK**

Statcato gave us a lot of information, most of which we have not learned yet. See if you can find the \( r \). You will find it in a little table. The correlation coefficient is the number next to the letter \( r \) and under the words “Test Statistic”.

<table>
<thead>
<tr>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r )</td>
</tr>
<tr>
<td>(-0.0259)</td>
</tr>
</tbody>
</table>

Interpretation: So what does this statistic of \( r = -0.0259 \) tell us. Looking at the number line, we see that the \( r \) value of \(-0.0259\) though negative is extremely close to zero. This does not tell us there is negative correlation. This statistics agrees with what we said earlier when we looked at the scatterplot. There seems to be no linear relationship (no correlation) between the weight and cholesterol of these men.
Example 2
In the last section, we also looked at the relationship between the weight of the men (in pounds) and the waist size of the men (in centimeters). I was interested in predicting the weight of a man from his waist size, so I picked the men’s waist size (in cm) to be the explanatory variable (X) and the men’s weight (in pounds) to be the response variable (Y).

We used Statcato to find the following scatterplot and correlation coefficient “r”.

![Scatterplot of men's waist size and weight](image)

The scatterplot shows a “positive linear relationship” or a “positive correlation”, but how strong is this relationship? To determine this we can look at the correlation coefficient “r”.

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8889</td>
</tr>
</tbody>
</table>

The correlation coefficient came out to be 0.8889. I like to put a positive sign in front of the correlation coefficient since 0.8889 really means +0.8889 and the sign of the r value is important to the interpretation.

Interpretation: “Strong, Positive Correlation”

Look again at the correlation coefficient number line. Notice that +0.8889 is a number very close to +1. That means that this correlation coefficient is telling us that there is a strong positive correlation between the waist size of a man (in cm) and his weight (in pounds). Therefore, it again confirms what our eyes were telling us when we looked at the scatterplot.
The points seem to be close to a line (strong) and that line is going up from left to right (positive).

Note: The question is often asked, what is the strength and direction of the linear relationship (correlation)? Remember the strength (strong, moderate, weak, none) is asking how close the points are to the line. The direction is asking if the line is going up or down from left to right (increasing or decreasing).

Example 3
In the last section, we also looked at an example with an outlier. The data describes the number of months in business (X) and the company costs (Y) in thousands of dollars. The scatterplot shown below shows that most of the data follows a negative linear pattern, but month 15 had a higher cost than expected and did not seem to follow the pattern.

When a scatterplot shows an unusual point (outlier), it is often asked, “How influential is that outlier?” In other words, is the outlier doing a lot of damage to the overall relationship?
Outliers can make a big difference to the strength of the relationship. Correlation coefficients can show a weak relationship with the outlier, but a very strong relationship without the outlier. When this happens, we call this an “influential outlier”.

Let us use Statcato to calculate the correlation coefficient “r” and shed some light on this relationship. It seems like the relationship should be pretty strong, but how is the outlier effecting the overall strength of the relationship?

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.9409</td>
</tr>
</tbody>
</table>

Interpretation: Look at the correlation coefficient number line again. An r value of -0.9409 is very close to -1. That means that despite the outlier, the correlation is still very strong. This tells us that the outlier is not very influential. The overall interpretation is still strong and negative. Therefore, there is a strong negative correlation between the number of months in business and the costs of the company in thousands of dollars.

Example 4
Curved relationships can be tricky. Remember if you do have a curved relationship, you really want to ask the computer to draw a curve that best fits the data, not a line. Some students get into trouble because they look at the r value for a line and try to apply it to a curve. When you ask a program like Statcato to calculate “r” under the linear menu, you are asking the computer how close your points are to a line, not a curve!

In the last section, we looked at a scatterplot that showed a curved pattern. The explanatory variable (X) was the # months in 2009 and the response variable was the solar energy generated from that month in kilowatt-hours (kWh).
Notice the points in the scatterplot do not seem to be close to a line, but there seems to be a curved relationship.

If we went to Statcato and had the computer calculate the correlation coefficient “r” under the linear menu, we would get the following.

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.0568</td>
</tr>
</tbody>
</table>

This is why it is so important to look at scatterplot before calculating statistics. We need to know what shape we are dealing with. This r value is very close to 0, meaning that there is no linear relationship. If a student looked at just the r value without looking at a scatterplot, they may incorrectly think that there is no relationship between time (months) and solar energy. There is actually a strong relationship, but it is just not linear.

Interpretation: “Strong Curved Relationship”

It is important to know the shape before calculating statistics. Calculating statistics for a linear relationship when it is really curved can be very misleading. We will learn in chapter six, that we can calculate statistics that measure the strength of curved relationships.
Problem Set Section 5B
Open the health data, create scatterplots, and find the correlation coefficient “r” with Statcato. Save the “r” value and graphs on a word document or make a general sketch of the graph on a sheet of paper.

For each of the following problems:

- Tell which variable you chose to be the explanatory and which you chose to be the response variable and why.
- Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
- Calculate the correlation coefficient “r” with Statcato.
- Interpret the strength and direction of the linear relationship.
- Are there any outliers that do not seem to fit the pattern? Use the r value to determine if the outlier is influential or not.

1. Explore the relationship between a woman’s height and weight.
2. Explore the relationship between a man’s height and weight.
3. Explore the relationship between a woman’s age and cholesterol.
4. Explore the relationship between a man’s age and cholesterol.
5. Explore the relationship between a woman’s weight and body mass index (BMI).
6. Explore the relationship between a man’s weight and body mass index (BMI).
7. Explore the relationship between a woman’s systolic blood pressure and her diastolic blood pressure.
8. Explore the relationship between a man’s systolic blood pressure and his diastolic blood pressure.
9. Explore the relationship between the length of a man’s leg and the circumference of his wrist.
10. Explore the relationship between a woman’s pulse and her cholesterol level.
Section 5C – Confounding Variables, r-squared, Correlation is not Causation, and Multivariable Studies

Correlation is NOT Causation
There is a famous saying in statistics, “correlation is not causation”. We saw this when we looked at categorical relationships in chapter 2. Just because there is a relationship between two variables, does not mean that one variable causes the other.

Why? Why doesn’t a relationship imply causation?

The real reason is confounding variables. Confounding variables (also called “lurking variables”) are other variables that might be related to the response variable other than the explanatory variable you are looking at. It helps to look at an example.

In the previous section, we found that there is a strong positive linear relationship (correlation) between the waist size and weight of forty men in the health data. Does that mean that the waist size of a man causes them to have a certain weight? No, it does not. The weight of a man is influenced by many factors other than just his waist size. Can you think of any?

Height of the man
Genetics (How tall are his parents?)
Amount of Exercise
Quality of his diet
Body Mass Index
Amount of Muscle
Amount of Fat

These are called confounding variables. Many variables influence a man’s weight other than just his waist size. That is why it is wrong to say things like, “a small waist size causes a man to not weigh very much”. (Athletes often have a lot of muscle mass and may weigh a lot, but have a small waist size.)

Multivariable Relationships and r-squared
Suppose we are studying the weight of a man and what to know which variables have the strongest relationship with weight. An important statistic often used in studies like this is the square of the correlation coefficient (r-squared).

Definition of r-squared: The square of the correlation coefficient. Tells us the percentage of variability in the response variable (Y) that can be explained by the relationship with the explanatory variable (X).
Notes about r-squared

- Calculating r-squared is usually given in correlation and regression printouts from statistics programs like Statcato. If you have the correlation coefficient “r”, you can also push the square button on a calculator or multiply the “r” value by itself.
- r-squared is positive. Remember when you square a number (even a negative number) the result will be positive. r-squared is never negative.
- r-squared is a percentage. Make sure to take the r-squared value in the computer and multiply it by 100% to convert it into a percentage.
- Do not convert the correlation coefficient r into a percentage. r is not a percentage. It is a decimal number between -1 and +1.
- The higher the r-squared percentage is, the stronger the relationship. The lower the r-squared percentage is, the weaker the relationship.
- r-squared can be calculated for lines and curves. r-squared is a great statistic for quantitative relationship studies because it can be calculated for linear relationships and for curved relationships. (As long as you tell the computer what relationship to calculate.)

Calculating r-squared with Statcato: Statistics => Correlation and Regression => Linear (Two Variables) => Pick which column is X and which is Y => Add Series => OK

Note: The r-square value can be found at the bottom of the Statcato printout where it says “Coefficient of determination r-square =”. You can also square the r value yourself, but the computer knows you need it and does it automatically.

Example 1
Let us use Statcato to calculate r-squared. In the last section, we found that there was strong negative correlations between the number of months a company had been in business and their costs. Here is the scatterplot again.
Using Statcato, I calculated the r-squared value for this data. This is what the printout looked like on Statcato.

Coefficient of determination \( r^2 = 0.8853 \)

As with all statistics, we should ask ourselves what this means. What does it tell us? Start by converting \( r \)-squared into a percentage by multiplying by 100%.

\[ r^2 = 0.8853 \approx 88.5\% \]

In this problem and in this chapter, our \( r \)-squared values were describing linear relationships. We will see in the next chapter that the sentence if it is a curved relationship we just need to change the sentence slightly to say, “Curved relationship with \( X \)”.

Interpretation: So 88.5% of the variability in \( Y \) (costs in thousands of dollars) for this company can be explained by the linear relationship with \( X \) (time in months). This percentage is very high (close to 100%) indicated this is an extremely strong linear relationship. Time is a good explanatory variable for predicting costs. However this does not indicate that time causes costs to decrease (correlation is not causation). There are many confounding variables involved.
Example 2: Multiple Variable Quantitative Relationship Studies
What variables have the strongest relationship with the weight of a man? (What variables are most important to study when looking at men’s weight?) This kind of study is sometimes called “multiple regression”.

The health data has several variables that we might look at, but which ones have the strongest relationships with men’s weight? This is actually not as difficult as it seems. We will need to choose a different explanatory variable (X) each time and then use a statistics software program like Statcato to calculate r-squared for each variable with the men’s weight.

Response Variable: Weight of Men (in pounds)
What variables that might be related to the weight of the men?
For this example, we will focus on the variables in the health data and on linear relationships only. Linear relationships are the most common type of quantitative relationship statisticians study.

Men’s Age (years)
Men’s Height (inches)
Men’s Waist Size (cm)
Men’s Pulse (beats per minute)
Men’s Systolic Blood Pressure (mm of Hg)
Men’s Diastolic Blood Pressure (mm of Hg)
Men’s Cholesterol (mg per deciliter)
Men’s Body Mass Index (BMI) (kg per m^2)
Men’s Leg Length (inches)
Men’s Elbow Circumference (Inches)
Men’s Wrist Circumference (Inches)
Men’s Arm Length (Inches)

Letting each of these variables be the explanatory variable, we can calculate the r-squared value for each. Remember we need to keep the men’s weight as the response variable though, since that is the variable we are studying.

Plugging in each of the variables into Statcato for X and letting the men’s weight always be the Y variable, we get the following r-squared values. Remember to multiply the decimals by 100% to convert into a percentage.
Men’s Age (years) and Weight (pounds):  $r$-squared = 0.0815 $\approx$ 8.2%

Men’s Height (inches) and Weight (pounds):  $r$-squared = 0.2727 $\approx$ 27.3%

Men’s Waist Size (cm) and Weight (pounds):  $r$-squared = 0.7902 $\approx$ 79.0%

Men’s Pulse (beats per minute) and Weight (pounds):  $r$-squared = 0.0031 $\approx$ 0.31%

Men’s Systolic Blood Pressure (mm of Hg) and Weight (pounds):  $r$-squared = 0.1240 $\approx$ 12.4%

Men’s Diastolic Blood Pressure (mm of Hg) and Weight (pounds):  $r$-squared = 0.1503 $\approx$ 15.0%

Men’s Cholesterol (mg per deciliter) and Weight (pounds):  $r$-squared = 0.0007 $\approx$ 0.07%

Men’s Body Mass Index (BMI) (kg per m$^2$) and Weight (pounds):  $r$-squared = 0.6395 $\approx$ 64.0%

Men’s Leg Length (inches) and Weight (pounds):  $r$-squared = 0.1380 $\approx$ 13.8%

Men’s Elbow Circumference (Inches) and Weight (pounds):  $r$-squared = 0.4034 $\approx$ 40.3%

Men’s Wrist Circumference (Inches) and Weight (pounds):  $r$-squared = 0.2696 $\approx$ 27.0%

Men’s Arm Length (Inches) and Weight (pounds):  $r$-squared = 0.6750 $\approx$ 67.5%

Multiple Regression Interpretation:

So which variables had the strongest relationship with the weight of the men?

Height (27.3%), Waist Size (79.0%), Body Mass Index (64.0%), Elbow Circumference (40.3%), Wrist Circumference (27.0%) and Arm Length (67.5%) all showed a linear relationship to the weight of the men. Waist size, body mass index and arm length showed very strong linear relationships, but all of these variables showed some correlation and we should think about all of them if we want to study men’s weights.

What about the other variables?

Pulse (0.31%) and cholesterol (0.07%) showed no relationship at all. (Notice their $r$-squared values are very close to zero.)
Surprisingly, the following variables showed a weak linear relationship.

Age (8.2%), Systolic Blood Pressure (12.4%), Diastolic Blood Pressure (15.0%), Leg Length (13.8%),

**Explaining r-squared**

Let us practice writing a sentence to explain the meaning of some of the r-squared values in the previous example. As with most statistics, focus more on understanding and being able to explain statistics, than on calculation. Remember r-squared is the percentage of explained variability in Y (weight) that can be explained by the relationship with (X).

Sentence to explain r-square for men’s age and weight: The r-squared value for the men’s age and weight was 0.0815 or 8.2%. Explain this statistic to someone.

8.2% of the variability in the men’s weight (in pounds) can be explained by the linear relationship with their ages (in years). This tells us that there is a weak linear relationship between the age of these men and their weights.

Sentence to explain r-square for men’s height and weight: The r-squared value for the men’s height and weight was 0.2727 or 27.3%. Explain this statistic to someone.

27.3% of the variability in the men’s weight (in pounds) can be explained by the linear relationship with their heights in inches. This tells us that there is moderate linear relationship between the height of these men in inches and their weights in pounds.

---

**Problem Set Section 5C**

1. Write a couple sentences to explain each of the following r-squared values from the multiple variable example in this section. In each of these examples the weight was the response variable (Y).

   a) Men’s Waist Size (cm) and Weight (pounds): r-squared = 0.7902 \approx 79.0%

   b) Men’s Pulse (beats per minute) and Weight (pounds): r-squared = 0.0031 \approx 0.31%

   c) Men’s Systolic Blood Pressure (mm of Hg) and Weight (pounds): r-squared = 0.1240 \approx 12.4%

   d) Men’s Diastolic Blood Pressure (mm of Hg) and Weight (pounds): r-squared = 0.1503
e) Men’s Cholesterol (mg per deciliter) and Weight (pounds): r-squared = 0.0007 \approx 0.07%

f) Men’s Body Mass Index (BMI) (kg per m^2) and Weight (pounds): r-squared = 0.6395 \approx 64.0%

g) Men’s Leg Length (inches) and Weight (pounds): r-squared = 0.1380 \approx 13.8%

h) Men’s Elbow Circumference (Inches) and Weight (pounds): r-squared = 0.4034 \approx 40.3%

i) Men’s Wrist Circumference (Inches) and Weight (pounds): r-squared = 0.2696 \approx 27.0%

j) Men’s Arm Length (Inches) and Weight (pounds): r-squared = 0.6750 \approx 67.5%

Use the given graphs and r values to complete the following for each problem.

- Find the value of r-squared by squaring the r value with your calculator.
- Write a sentence interpreting the r-squared percentage in the context of data. Be sure to include the appropriate units.
- List other possible confounding variables that may also account for the variability in y.
- Since there are confounding variables, is it ok to say that the X variable causes the Y?

2. The x variable is describing the number of tons of paper trash and the y variable is the number of tons of total trash. (r = 0.7287)
3. The x variable is describing the number of tons of plastic trash and the y variable is the number of tons of metal trash. \( r = 0.5862 \)

4. The x variable is describing the number of tons of food trash and the y variable is the number of tons of total trash. \( r = 0.5833 \)
5. The x variable is describing the horsepower of an automobile and the y variable is describing the miles per gallon. \( r = -0.8713 \)

6. The x variable is the number of cars sold and the y variable is the total profit in thousands of dollars. \( r = 0.9404 \)
7. The x variable is the number of pounds of fertilizer used and the y variable is the number of flowers per square foot. \( r = 0.6727 \)

8. The x variable is the week and the y variable is the stock price in dollars. \( r = -0.9429 \)

9. We want to study Body Mass Index (BMI) for men in (kilograms per meter squared) and what variables have the strongest relationship with BMI for men.
Open the health data in Statcato. For each of the following explanatory variables, calculate the r-squared value and convert it into a percentage. Label which variables have the strongest relationship to men’s BMI and which variables do not seem to have much of a relationship.

Men’s Age (years) and men’s BMI (kg/m^2): r-squared (as %) = _________________

Men’s Height (inches) and men’s BMI (kg/m^2): r-squared (as %) = _________________

Men’s Weight (pounds) and men’s BMI (kg/m^2): r-squared (as %) = _________________

Men’s waist size (cm) and men’s BMI (kg/m^2): r-squared (as %) = _________________

Men’s cholesterol (mg/deciliter) and men’s BMI (kg/m^2): r-squared (as %) = _________________

Follow up questions:

Of the five explanatory variables, which had the strongest relationship to BMI?

Were there any of the r-squared values that surprised you? Explain why.

10. We want to study the weight of bears and what variables have the strongest relationship with the bear weights.

Open the bear data in Statcato. For each of the following explanatory variables, calculate the r-squared value and convert it into a percentage. Label which variables have the strongest relationship to bear weights and which variables do not seem to have much of a relationship.

Bear’s age (months) and weights (pounds): r-squared (as %) = _________________

Bear’s head length (inches) and weights (pounds): r-squared (as %) = _________________

Bear’s head width (inches) and weights (pounds): r-squared (as %) = _________________
Bear neck circumference (inches) and weights (pounds):
r-squared (as %) = _________________

Bear chest size (inches) and weights (pounds): r-squared (as %) = _________________

Bear length (inches) and weights (pounds): r-squared (as %) = _________________

Follow up questions:

Of the six explanatory variables, which had the strongest relationship a bear’s weight?

In the wild, it may be difficult to get an accurate weight of a bear. A park ranger wants to take measurements that may be used to estimate the weight of the bear. Which measurements should the ranger focus on?

Were there any of the r-squared values that surprised you? Explain why.

Section 5D – Best Fit Regression Line with Technology with
Slope and Y-intercept Interpretation

We said that one of the main differences lines in algebra classes and lines in statistics is the number of points. Algebra talks about the equation of a line between two points, while in statistics we talk about the line that best fit thousands or even millions of points.

How is this done? How do you find a line that minimizes the distance between itself and thousands of points in a scatterplot?
This line of best fit is often called the “regression line”.

Definition of the “regression line”: This line best fits all the points in the scatterplot by minimizing the vertical distance between itself and all of the points in the scatterplot. It is also called the “line of best fit” or the “line of least squares”. It also is sometimes called a “prediction line” because if the two quantitative data sets have correlation, then the regression line can become a formula for making predictions. Note: If there is no linear relationship (no correlation), then we should not use the regression line to make predictions.

Calculating the Regression Line
As with most statistics, the regression line from the points in your scatterplot is very complex and time-consuming calculation. It is best to calculate the line with a statistics software program like Statcato. We will show the process of how the line is calculated though.

To calculate a regression line you will need to calculate five different statistics, each of which is a very time consuming calculation. However, if you have these statistics already calculated, you can use them to get the regression line with a couple formulas.

\( \bar{x} \): mean average of the explanatory data set (X)  
\( \bar{y} \): mean average of the response data set (Y)  
\( S_x \): standard deviation of the explanatory data set (X)  
\( S_y \): standard deviation of the response data set (Y)  
\( r \): correlation coefficient between X and Y

Recall that the equation of a line is made up of two values, the slope of the line and the y-intercept. If you can find the slope and the y-intercept, you can write the formula for the equation of the line.

Equation of the Regression Line:
\( Y = (Y\text{-intercept}) + (\text{slope})X \)

Notice that the order of the slope and Y-intercept are backwards from algebra classes you may
have seen. Algebra usually writes the equation with the slope first. In statistics, we prefer to write the Y-intercept first. The reason why is that Y-intercepts are usually initial values and the slope is how much the variables change after this initial Y-intercept value. Therefore, it makes sense that the initial value comes first in the formula. Just remember, whether you are looking at a line from algebra or statistics, the number in front of the “X” is the slope.

**Calculate the Slope**

We start by calculating the slope of the regression line. How do you find the slope that best fits thousands of data points? Start with the definition of slope. Slope is defined as the rate of change between the Y variable and X variable. In algebra, they often define slope as “rise over run” or “change in Y / change in X”. It is easy to measure this change when you have two points, but how do you measure this change when you have thousands of points? Think of change in Y as the variability in Y and change in X as the variability in X. So we need a measure of the variability (spread) of X and Y. In regression line calculations, we use the standard deviation as our measure of spread.

\[
\text{Slope} \approx \frac{\text{Variability in } Y}{\text{Variability in } X} \approx \frac{\text{Standard Deviation of all the } Y \text{ values (} S_y \text{)}}{\text{Standard Deviation of all the } X \text{ values (} S_x \text{)}}
\]

Now there are two problems with leaving the formula like this. The first is that standard deviation is a distance calculation and is always positive. If all we do is divide the standard deviations, it will be impossible to get a negative slope (which happen all the time). The second problem is we need to take into account the strength of the correlation. It turns out that both of these problems can be solved by multiplying this ratio by the correlation coefficient. Remember the correlation coefficient measure the strength and direction (negative or positive) of the linear relationship.

\[
\text{Best Fit Slope} = r \left( \frac{S_y}{S_x} \right) = \frac{r \left( S_y \right)}{S_x}
\]

**Calculate the Y-intercept**

If you recall from your algebra classes, you can calculate the Y-intercept if you know the slope and point on the line. This is true for statistics as well, but what point should we use? There may be thousands of points in a scatterplot and the regression line does not have to go through any of them. The regression line gets close as possible to all of them.

Point on the regression line: It turns out the point we want to use for the regression line calculation is not any of the points in the scatterplot. Remember we want the line to go through the center of the spread of points. The mean average is a measure of spread, so we
like to use the ordered pair \((\text{mean of } X, \text{mean of } Y)\) to calculate the Y-intercept. Hence, to calculate the Y-intercept for the best-fit line, we will use the mean of all the X values, the mean of the Y values, and the best-fit slope.

\[
\bar{x} : \text{mean average of the explanatory data set (X)} \\
\bar{y} : \text{mean average of the response data set (Y)}
\]

Best Fit Y-intercept = \((\text{mean of } Y \text{ values}) – (\text{slope}) (\text{mean of } X \text{ values})\)

Best Fit Y-intercept = \(\bar{y} – \text{slope} (\bar{x})\)

When calculating the Y-intercept, we need to multiply the slope times the mean average of the explanatory data set (X). Then subtract the answer from the mean average of the response data set (Y).

**Example 1**

Let us look again at an example from the last section. We looked at some data that gave the number of months a company had been in business and their costs in thousands of dollars. We found that there was an outlier, but that overall there was a strong negative correlation. This is important. Always check to see if there is correlation, before taking time to calculate the regression line. Here is the scatterplot for the data.

![Scatterplot of number of months and costs of company](image)

Note: Computer programs like Statcato give you the regression line in the printout though, so a better way to think of it is do not use the regression line equation if there was no correlation. The equation does not fit the data.
Since this data did show a strong correlation, let us look at how the best-fit regression line would be calculated for this data.

Explanatory Variable (X): Number of months the company is in business
Response Variable (Y): Company costs in thousands of dollars

We will need the statistics listed above if we are going to calculate the regression line.

\[\bar{X}: \text{mean average of the explanatory data set (X)}\]
\[\bar{Y}: \text{mean average of the response data set (Y)}\]
\[S_x: \text{standard deviation of the explanatory data set (X)}\]
\[S_y: \text{standard deviation of the response data set (Y)}\]
\[r: \text{correlation coefficient between X and Y}\]

I used Statcato to calculate the statistics.

**Descriptive Statistics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Months in business</td>
<td>12.0</td>
<td>7.360</td>
</tr>
<tr>
<td>Costs in Thousands of Dollars</td>
<td>3.436</td>
<td>1.820</td>
</tr>
</tbody>
</table>

\[r = 0.9409\]

To calculate the slope of the best fit regression line, we will take the correlation coefficient “r”, multiply by the standard deviation of the Y values (costs) and then divide by the standard deviation of the X values (months).
Now that we have the slope, we can use the mean averages to calculate the best fit Y-intercept.

\[
\text{Best Fit Y-intercept} = \bar{y} - \text{slope} \times \bar{x} \\
\approx 3.436 - (-0.232668206 \times 12.0) \\
\approx 3.436 - (-2.792018478) \approx 3.436 + (2.792018478) \\
\approx +6.228018 \approx +6.228
\]

Now that we know the best-fit slope and Y-intercept, we can write the equation of the regression line.

Equation of the Regression Line:

\[
Y = (Y\text{-intercept}) + (\text{slope}) \times X \\
Y = 6.228 + (-0.233) \times X
\]

Statcato calculates this for us. Here is the printout from Statcato.

Regression
Regression equation \( Y = b_0 + b_1 x \)
\( b_0 = 6.2274 \)
\( b_1 = -0.2326 \)

Notice Statcato uses the following letters. These are common variables used in statistics for the slope and the Y-intercept.

\( b_0 \) :  Y-intercept
\( b_1 \) :  Slope

To get the formula, just replace the variables with the numbers they equal.

In addition, notice the accuracy was better than our calculation. Computer programs use a lot more decimal places when they calculate than we can with a calculator.
Calculating the Regression Line
Always use a statistics software program to calculate the regression line. Here are the directions for using Statcato to calculate the regression line. Notice it is the same directions as when we found “r” and “r-squared”.

To find the slope, Y-intercept and regression line equation with Statcato:
Statistics => Correlation and Regression => Linear (Two Variables) => Pick which column is X and which is Y => Add Series => OK

Interpreting the Slope and Y-intercept
Remember, calculating a regression line is not enough. We need to know what the slope and Y-intercept tell us about the relationship between the real data variables. We need to understand these statistics and be able to explain them to others.

Example 1 (Interpretation)
Let us see if we can explain what the slope and the Y-intercept to the employees of this company as it may be important information to them.

Let us look at the slope, Y-intercept and regression line equation that Statcato gave us. (Our calculation was not as accurate.)

Regression
Regression equation \( Y = b_0 + b_1x \)
\( b_0 = 6.2274 \)
\( b_1 = -0.2326 \)

Interpreting the slope
To interpret the slope, you have to remember that the slope measures the change in Y divided by the change in X. In other words, you cannot interpret the slope without thinking of it as a fraction and including the units.

The slope was -0.2326 (\( b_1 \) in the printout). A good way to think of this decimal as a fraction is to put it over +1. Then put the units of Y in the numerator and the units of X in the denominator.

\[-0.2326 = \frac{-0.2326 \text{ thousand dollars}}{+1} = \frac{-0.2326 \text{ thousand dollars}}{+1 \text{ month}}\]
First, recognize what the slope is not saying. The slope is not saying that the company had a cost of -0.2326 thousand dollars in the first month. (In fact, the company had a cost of over six thousand dollars in its first month.)

So what is the slope telling us? You also have to remember that slope is a rate of change (increase or a decrease). If it is negative, it is an average decrease and if it is positive, it is an average increase.

Since this slope was negative, it is a decrease. The slope is telling us that monthly costs are decreasing about 0.2326 thousand dollars per month on average. Looking at the units, we can also explain it this way: Monthly costs are decreasing about $233 per month on average. (Notice we did not say that the costs were decreasing -0.2326 or decreasing -$233. The negative is described by the word “decrease”.)

**Definition of Slope of the Regression Line:** A rate of change that measures the average increase or decrease in the Y variable per 1 unit of X.

**Interpretation of Y-intercept**
Statcato calculated the Y-intercept for the month/cost data. Let us see if we can interpret and explain the Y-intercept.

**Statcato Printout**
Regression equation \( Y = b_0 + b_1x \)
\( b_0 = 6.2274 \)
\( b_1 = -0.2326 \)

Remember the slope is the number in front of the X. The Y-intercept is the initial number in the regression line formula that is by itself. So the Y-intercept for the cost data is \( b_0 = 6.2274 \).

A Y-intercept is the predicted Y value when X is zero. Do not forget to include the units. Therefore, the Y-intercept of 6.2274 really represents the ordered pair (0 months, 6.2274 thousand dollars).

Interpretation of Y-intercept for Cost Data: At the start of the company (month 0), we predict an average initial cost of approximately 6.2274 thousand dollars (or $6,227).

**Definition of Y-intercept of the Regression Line:** The predicted Y-value when X is zero.
Note about Y-intercept Interpretations: The regression line is meant to apply to the X and Y values in the two data sets. Zero is often not represented in the X values of many quantitative data sets. We zero is not in the scope of the X values. When this happens, the Y-intercept will not make a whole lot of sense. It is still an important number in the formula if our predictions are to be accurate, but the formula may not be designed to plug in zero for X.

Making a Scatterplot with the Regression Line
It is often nice to see the scatterplot with the regression line drawn to see how well the data fitting the line. Statistics software programs can create the scatterplot with the regression line drawn. There are two ways of getting this graph with Statcato.

Making a scatterplot with regression line in the graph menu of Statcato: Graph => Scatterplot => Pick which column is X and which is Y => Add Series => Label your X-axis and Y-axis => Give graph a title => check the box “show regression curve” and then “Linear: y = ax+b”

In a quantitative relationship study, we often want the equation of the regression line, slope, Y-intercept, r, r-squared, etc. Statcato allows you to make the scatterplot with the regression line drawn from this menu as well. Just go to the statistics menu and click on correlation and regression. Then click on linear and pick your columns for X and Y. Push the add series button. Now click on the box that says, “Show a scatterplot for all pairs of data values”. Under the options, click on the box that says

Making a scatterplot with regression line from the linear correlation and regression menu in Statcato: Statistics => Correlation and Regression => Linear (Two Variables) => Pick which column is X and which is Y => Add Series => Check the box that says “show a scatterplot” => Label your X and Y axis and give the graph a title => OK
Here is the scatterplot with the regression line drawn for the month and cost data in this section. Notice it not only draws the regression line, but Statcato gives the equation of the line at the bottom of the graph. The slope and Y-intercept are rounded values though. Note that the dash in front of the formula does not mean negative Y. It is simply a divider telling where the equation begins. (The equation is $Y = 6.227 + -0.223X$)

**Scatterplot of months and cost data**
Important Note about Shape: We have seen in this section that the mean and standard deviations of the two quantitative variables are used to calculate the regression line. Remember that mean averages and standard deviations are only accurate if the data is bell shaped. Therefore, our regression line is not very accurate when the data is not bell shaped. We will see in the next section that to verify the shape requirement we will look at a special histogram called the “histogram of the residuals” to check this bell shaped requirement.

Problem Set Section 5D
(#1-6) Use the regression line formulas below and a calculator to calculate the slope, Y-intercept, and equation of the regression line for the following ordered pair data. The correlation coefficient (r), mean averages and standard deviations for both X and Y variables have been provided.

- Slope of the Regression Line = (correlation coefficient) (standard deviation of Y values) / standard deviation of X values

- Y-intercept of the Regression Line (multiply before you subtract.)
  = (mean of Y values) – (slope) (mean of X values)

- Equation of the Regression Line (plug in the slope and Y-intercept but leave the X and Y in the formula)
  
  \[ Y = (\text{Y-intercept}) + (\text{slope}) X \]
1.

Scatterplot of Paper Trash and Total Trash

\[ r = 0.7287 \]

\[
\begin{array}{lll}
(x) & \text{Paper Trash} & 9.428 & \text{Mean} \\
& \text{StDev} & 4.168 \\
(y) & \text{Total Trash} & 27.44 & \text{Mean} \\
& & 12.46 & \text{StDev}
\end{array}
\]

Slope = _________________

Sentence Explaining the Slope:

Y – Intercept = _________________

Sentence Explaining the Y-intercept:

Equation of the Regression Line: \( Y = \_\_\_\_\_\_\_ + \_\_\_\_\_\_ \times X \)
2.

$r = 0.5862$

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) Plastic Trash</td>
<td>1.911</td>
<td>1.065</td>
</tr>
<tr>
<td>(y) Metal Trash</td>
<td>2.218</td>
<td>1.091</td>
</tr>
</tbody>
</table>

Slope = __________________

Sentence Explaining the Slope:

$Y - \text{ Intercept} = ____________$

Sentence Explaining the Y-intercept:

Equation of the Regression Line: $Y = _________ + _________ X$
3.

$r = 0.5833$

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) Food Trash</td>
<td>4.816</td>
<td>3.297</td>
</tr>
<tr>
<td>(y) Total Trash</td>
<td>27.44</td>
<td>12.46</td>
</tr>
</tbody>
</table>

Slope = ________________

Sentence Explaining the Slope:

$Y – Intercept = ________________$

Sentence Explaining the Y-intercept:

Equation of the Regression Line: $Y = __________ + __________ X$
4. This data describes the relationship between the number of cars sold and total profit in thousands of dollars.

\[ r = 0.9404 \]

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) Cars Sold in Month</td>
<td>21.667</td>
<td>7.512</td>
</tr>
<tr>
<td>(y) Profit in Thousands of Dollars</td>
<td>420.25</td>
<td>175.615</td>
</tr>
</tbody>
</table>

Slope = ________________

Sentence Explaining the Slope:

\[ Y - \text{Intercept} = \textbf{ ________________} \]

Sentence Explaining the Y-intercept:

Equation of the Regression Line: \[ Y = \textbf{ __________} + \textbf{ __________} \times X \]
5. This data describes the relationship between the number of pounds of fertilizer used and the number of flowers per square foot.

![Scatterplot](image)

\[ r = 0.6727 \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) pounds of fertilizer used</td>
<td>3.387</td>
<td>1.165</td>
</tr>
<tr>
<td>(y) # of flowers per sq. ft.</td>
<td>13.867</td>
<td>1.356</td>
</tr>
</tbody>
</table>

Slope = ________________

Sentence Explaining the Slope:

\[ Y \text{ – Intercept} = ________________ \]

Sentence Explaining the Y-intercept:

Equation of the Regression Line: \[ Y = __________ + __________ X \]
6. The following data describes the Price of a stock over the first 20 weeks of this year.

\[ r = -0.9429 \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) #weeks</td>
<td>10.5</td>
<td>5.916</td>
</tr>
<tr>
<td>(y) stock price $</td>
<td>270.6</td>
<td>17.031</td>
</tr>
</tbody>
</table>

Slope = __________________

Sentence Explaining the Slope:

\[ Y - \text{Intercept} = \text{______________} \]

Sentence Explaining the Y-intercept:

Equation of the Regression Line: \( Y = \text{___________} + \text{___________} X \)

For the following problems, get the linear correlation and regression printout from Statcato. Include the scatterplot with the regression line drawn and use it to answer the following questions.
7. Open the cigarettes data set. Explore the relationship between mg of nicotine and mg of tar in cigarettes.

a) Let nicotine be the explanatory variable and tar be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. Save the scatterplot with the regression line drawn on a word document or draw a rough sketch of it on a piece of paper. Describe the strength and direction of the linear relationship. *(This tells us how well the line fits the data.)*

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.

c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) Find the equation of the regression line.

8. Open the cigarettes data set. Explore the relationship between mg of nicotine and the carbon monoxide (CO) (parts per million PPM) in cigarettes.

a) Let nicotine be the explanatory variable and CO be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. Save the scatterplot with the regression line drawn on a word document or draw a rough sketch of it on a piece of paper. Describe the strength and direction of the linear relationship. *(This tells us how well the line fits the data.)*

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.

c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) Find the equation of the regression line.

9. Open the Health data set. Explore the relationship between a woman’s waist size in cm and her body mass index (BMI) in kg per m^2.

a) Let waist size be the explanatory variable and BMI be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. Save the scatterplot with the regression line drawn on a word document or draw a rough sketch of it on a piece of paper. Describe the strength and direction of the linear relationship. *(This tells us how well the line fits the data.)*
b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable. 
(Notice the Y-intercept does not make much sense since it is impossible for the explanatory variable to be zero.)

c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) Find the equation of the regression line.

10. Open the Health data set. Explore the relationship between a woman’s systolic blood pressure and her diastolic blood pressure.

a) Let systolic blood pressure be the explanatory variable and diastolic blood pressure be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. Save the scatterplot with the regression line drawn on a word document or draw a rough sketch of it on a piece of paper. Describe the strength and direction of the linear relationship. (This tells us how well the line fits the data.)

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable. 
(Notice the Y-intercept does not make much sense since it is impossible for the explanatory variable to be zero.)

c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) Find the equation of the regression line.

11. Open the Bears data set. Explore the relationship between the length of a bear in inches and the weight of the bear in pounds.

a) Let length be the explanatory variable and weight be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. Save the scatterplot with the regression line drawn on a word document or draw a rough sketch of it on a piece of paper. Describe the strength and direction of the linear relationship. (This tells us how well the line fits the data.)

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.
c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) Find the equation of the regression line.

12. Open the Bears data set. Explore the relationship between the head length of a bear in inches and the head width of the bear in inches.

a) Let the head width be the explanatory variable and head length be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. Save the scatterplot with the regression line drawn on a word document or draw a rough sketch of it on a piece of paper. Describe the strength and direction of the linear relationship. *(This tells us how well the line fits the data.)*

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable. *(Notice the Y-intercept does not make much sense since it is impossible for the explanatory variable to be zero.)*

c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) Find the equation of the regression line.

---

**Section 5E – Residuals, Standard Deviation of the Residual Errors (Se), Residual Plots and Histogram of the Residuals with Technology**

Statisticians often look deeper when they study quantitative relationships. One topic that is often explored is the study of “residuals”.

**Definition of Residual:** A “residual” or “residual error” is a measure of the vertical distance that each point in the scatterplot is above or below the line. It measures the difference between predicted Y values from the regression line and the actual Y values in the response data. If the
residual is positive, then the point is above the line. If the residual is negative, then the point is below the line.

**Notes about Residuals**

- If the residual is positive, then the point is above the line. If the residual is negative, then the point is below the line.
- The residual measures vertical distance to the line.
- The units of the residual are always the same as the response variable (Y).
- Since the regression line itself is sometimes used to make predictions, the residuals measure the amount of prediction error for each X value in the explanatory data. That is why residuals are often called “residual errors”.
- If the residual is positive, then the point is above the regression line. The line is where the predicted values are. Therefore, a positive residual tells us the line is beneath the actual point meaning that the prediction for that particular X value is too low.
- If the residual is negative, then the point is below the regression line. The line is where the predicted values are. Therefore, a negative residual tells us the line is above the actual point meaning that the prediction for that particular X value is too high.

The picture above shows the idea of residuals. In this example, the response variable was in dollars so all of the residual errors are in dollars. A point that has a residual of +3 means the point was 3 dollars above the regression line. A point that has a residual of -4 means that the point was 4 dollars below the regression line. What does this tell us about the predicted values on the line itself? The +3 residual tells us the point is 3 dollars above the line and that the line
is 3 dollars below the point. That means the predicted value is 3 dollars too low. The -4 residual tells us the point is 4 dollars below the line and that the line is 4 dollars above the point. That means the predicted value is 4 dollars too high.

Histogram of the Residuals
We saw at the end of the last section that since the regression line is based on the mean and standard deviations of the quantitative data sets, we need to check if the data is bell shaped. What we really need to check is to see if the “residuals” are bell shaped. To that end, a graph that is often looked at is the “histogram of the residuals”.

Statistics software programs like Statcato usually include the histogram of the residuals in the correlation and regression menu. To make a histogram of the residuals in Statcato, use the following steps.

Making a histogram of the residual errors in Statcato: Statistics => Correlation and Regression => Linear (Two Variables) => Pick which column is X and which is Y => Add Series => Check the box that says “show residual plots” => Check the box that says “Histogram of Residuals” => OK
Example 1
In this chapter, we have been looking at the months in business and cost data in thousands of dollars. Using the steps above, I made the following histogram of the residuals for the month and cost data.

Interpreting a Histogram of the Residual Errors
There are two things we like to check when we look at the histogram of the residual errors. As we have said before, we want to check to see if the data is close to bell shaped. It does not have to be a perfect bell shape, but it should not be radically skewed. The other factor is that that we want the center to be close to zero. The dark vertical line in the histogram is a marker for zero. An easy way to check this requirement is to see if the zero line is close to the highest bar. When the residuals are not bell shaped or if the graph is not centered at zero, then our regression line will not be as accurate as we think.

Two Requirements to Check when looking at a histogram of the residual errors:

1. The histogram should be close to bell shaped and not radically skewed.
2. The histogram should be centered close to zero. The zero line should coincide with the highest bar.

Example 1 continued
In the histogram shown above describing the residuals for the cost data, we see that the graph is not very bell shaped. In fact, it is skewed right. We also see that the zero line is a little off from the highest bar. This tells us the graph is not centered at zero as well as we would like.
This would be a red flag for a statistician or data analyst that the formula for the regression line is not as accurate as we think.

As with most things in statistics, there is a lot of grey area. In this last example, the zero line was not dramatically off from the highest hill.

**Residual Plot verses the X Variable**

Another graph that statisticians often look at is the “residual plot”. A residual plot is a graph of the residuals showing how far each point is from the regression line. Residuals are positive if the point is above the regression line and negative if the point is below the regression line. Therefore, in the residual plot, you will see negative and positive numbers. The horizontal zero line represents the regression line itself. Though there are many more advanced types of residual plots, we will focus on the “residual plot verses the X variable”. This graph shows the residuals with the original explanatory variable as the X-axis.

Here is a residual plot verses the X variable for the month and cost data again.

Let us see if we can understand what we are looking at. The horizontal line at zero represents the regression line itself. Notice the vertical scale on the left is no longer the same as the scatterplot. It has positive numbers above zero and negative numbers below zero. The units for the vertical access are still the same as Y variable (costs in thousands of dollars), but now it is showing the residual (how far each point is above or below the regression line). The X-axis is the same as the scatterplot, showing the number of months the company has been in business.
Statistics software programs like Statcato can make several types of residual plots in the correlation and regression menu. To make a residual plot verses the X variable in Statcato, use the following steps.

Making a histogram of the residual errors in Statcato: Statistics => Correlation and Regression => Linear (Two Variables) => Pick which column is X and which is Y => Add Series => Check the box that says “show residual plots” => Check the box that says “Residuals vs. X Variable” => OK

Interpreting a Residual Plot
There are many things statisticians look at when they study residual plots. We will focus on two in particular. The first is to look and see if the points are evenly spread out from zero line. We do not want to see a “V” or “fan” shape in the residual plot.

Good Residual Plot (Evenly Spread Out)
You can also look for curved (nonlinear) patterns in the residual plot. I like to think of residual plots as a magnifying glass. Points in a scatterplot can often look so small that it is difficult to see patterns in the data other than just the line. You can see the distances really well in a residual plot, which I find makes it easier to see curved relationships.
Note about terminology: The term “nonlinear” can refer to a curved pattern in the data, but can be misleading. Statisticians include many curved patterns under the heading of “linear regression” since they focus on the study of transformations from curves to lines.

Interpreting a Residual Plot
- The points should be evenly spread out from the zero line. You should not see a “V” shape.
- Look for curved patterns in the data that may not have been apparent in the scatterplot. If we see a curved pattern, we may want to use some type of regression curve, instead of the regression line.

Example 1 Continued
Here is the residual plot again for the month and cost data. Let us see if we can see any key features from this graph.
Notice we can see the outlier at 15 months, which is close to 2.5 thousand higher than what the regression line might predict. You may wish to set that outlier apart. It was an unusual event where the company had to purchase replacement equipment.

*Note: You have to be careful judging outliers from residual plots. The magnifying glass of a residual plot tends to make a lot of the ordered pairs look like outliers. I prefer to judge possible outliers with the scatterplot and correlation coefficient r.*

Note the curved “U” shaped pattern in the residual plot. This is an indication that a quadratic curve (parabola) may be a better fit for the data than the line was. (See chapter six.)

How about the even spread requirement? If we take the outlier out of the data. The rest of the dots have a pretty even spread from the zero line. We do not see any “V” shape.

**Standard Deviation of the Residual Errors (Se)**

Recall that the standard deviation of a single quantitative data set is a statistic that tells us how far typical values in the data set are from the mean in bell shaped data. Standard deviation is used in many different contexts in statistics. In regression theory, we can calculate the how far typical points are from the line or curve. We call this the “standard deviation of the residuals” or the “standard deviation of the residual errors”.

Standard deviation is always a difficult calculation, so we will use technology to calculate it. The standard deviation of the residual errors is given in the linear correlation and regression printout from Statcato.
To Calculate the Standard Deviation of the Residual Errors in Statcato:
*Statistics => Correlation and Regression => Linear (Two Variables) =>
Pick which column is X and which is Y => Add Series => OK

You will find the standard deviation at the very bottom of the correlation and regression printout under the r-squared value. Look for something that says “Standard error of estimate”.

Here is the standard deviation of the residuals for the month and cost data at the bottom of the Statcato printout.

Standard error of estimate = 0.6295

Interpreting the Standard Deviation of the Residual Errors
The standard deviation of the residual errors tells us two important things. Like the standard deviation for a single data set, the standard deviation for the residuals tells us how far typical points are from the regression line on average. Think of it as a measure of the average distance from the line. If there is correlation, we can use the regression line as a formula to make predictions. Therefore, the standard deviation gives us a measure of the average prediction error. If we use the regression line to make a prediction, then the standard deviation of the residuals tells us how far off the prediction might be on average.

1. The average distance that the points are from the regression line.
2. The average amount of prediction error.

Notes about Standard Deviation of the Residuals
- **Standard deviation is a measure of spread for bell shaped data sets. If the histogram of the residuals is not bell shaped, then the standard deviation is not as accurate.**
- **Do not confuse the standard deviation of the X values, the standard deviation of the Y values and the standard deviation of the residual errors. These are three different standard deviations that measure different things. The standard deviation of the Y values is how far typical values are on average from the mean of the Y values (response variable). The standard deviation of the X values is how far typical values are on average from the mean of the X values (explanatory variable). The standard deviation of the residual errors measures how far typical points in the scatterplot are on average from the regression line.**
- **Terminology: Some shorten the name for the “Standard Deviation of the Residual Errors” to just “Standard Error”. Even Statcato does this. This can be confusing, because the term “standard error” is used to describe the standard deviation of a sampling distribution.**
Example 1 continued

Here again is the standard deviation of the residuals for the month and cost data.

Standard error of estimate = 0.6295

What does this tell us about the data? It is good to look at the scatterplot and residual plot to give us a visual.

The standard deviation of the residuals always has the same units as the Y variable. If you look at the vertical axis on the scatterplot, you can see that the Y variable is the cost in thousands of dollars. So the standard deviation = 0.6295 thousand dollars (or $629.50).
The standard deviation tells us two important things in this example. The first is that the points in the scatterplot are on average about 0.6295 thousand dollars from the line. The second is that the average prediction error for this regression line will be about 0.6295 thousand dollars (or $629.50). If we use the regression line to make a prediction in the scope of the X values, our prediction could be off by about $629.50 on average.

Remember that the accuracy of the regression line and the standard deviation of the residuals is tied to the residuals being bell shaped. In this example, the histogram of the residuals was skewed right. This tells us that the regression line and standard deviation will not be quite as accurate.

Problem Set Section 5E
For the following problems, get the linear correlation and regression printout from Statcato. Create a residual plot verse the X variable and the histogram of the residual errors and use them to answer the following questions.

1. Open the cigarettes data set. Explore the relationship between mg of nicotine and mg of tar in cigarettes. Let nicotine be the explanatory variable and tar be the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors to answer the following questions.

   \[ S_e = 1.2984 \text{ mg of tar} \]

   a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

   b) Was the points in the residual plot evenly spaced out or was there a sideways “V” shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern (“U” shaped or “S” shaped for example).

   c) Interpret the standard deviation in context by describing the two meanings of the standard deviation.
2. Open the cigarettes data set. Explore the relationship between mg of nicotine and mg of carbon monoxide (CO) in cigarettes. Let nicotine be the explanatory variable and CO be the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors to answer the following questions.

\[ S_e = 2.2961 \text{ PPM} \]

a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

b) Was the points in the residual plot evenly spaced out or was there a sideways “V” shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern (“U” shaped or “S” shaped for example).

c) Interpret the standard deviation in context by describing the two meanings of the standard deviation.

3. Open the Health data set. Explore the relationship between a woman’s waist size in cm and her body mass index (BMI) in kg per m\(^2\). Let waist size be the explanatory variable and body mass index (BMI) be the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors to answer the following questions.

\[ S_e = 2.4761 \text{ kg/m}^2 \]
a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

b) Was the points in the residual plot evenly spaced out or was there a sideways “V” shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern (“U” shaped or “S” shaped for example).

c) Interpret the standard deviation in context by describing the two meanings of the standard deviation.

4. Open the Health data set. Explore the relationship between a woman’s systolic blood pressure and her diastolic blood pressure. Let systolic blood pressure be the explanatory variable and diastolic blood pressure be the response variable. Use the given residual plot versus the x-variable, histogram of the residuals, and the standard deviation of the residual errors to answer the following questions.

\[ S_e = 7.2912 \text{ mm of Hg} \]

a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

b) Was the points in the residual plot evenly spaced out or was there a sideways “V” shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern (“U” shaped or “S” shaped for example).

c) Interpret the standard deviation in context by describing the two meanings of the standard deviation.

5. Open the Bears data set. Explore the relationship between the length of a bear in inches and the weight of the bear in pounds. Let length be the explanatory variable and weight be the response variable. Use the given residual plot versus the x-variable, histogram of the residuals, and the standard deviation of the residual errors to answer the following questions.

\[ S_e = 61.8272 \text{ pounds} \]
a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

b) Was the points in the residual plot evenly spaced out or was there a sideways “V” shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).

c) Interpret the standard deviation in context by describing the two meanings of the standard deviation.

6. Open the Bears data set. Explore the relationship between the head length of a bear in inches and the head width of the bear in inches. Let the head width be the explanatory variable and head length be the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors to answer the following questions.

\[ S_e = 1.4231 \text{ inches} \]

a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

b) Was the points in the residual plot evenly spaced out or was there a sideways “V” shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).

c) Interpret the standard deviation in context by describing the two meanings of the standard deviation.
Section 5F – Predictions, Scope of the X-values, Extrapolation, and Using the Standard Deviation of the Residuals

In this chapter, we have seen that we can use scatterplots, r, and r-squared to analyze and measure linear relationships (correlation) between two different quantitative variables. We also found that if there is a linear relationship, we could find the line of best fit (regression line). If there is a correlation between the variables, then we can use the regression line to predict Y values. In this section, we will look at how to make predictions and guidelines for interpreting those predictions.

Notes about Making Predictions with the Regression Line

1. There should be some correlation between the variables.

If there is no linear relationship between the variables, then the regression line does not fit the data, meaning our predictions will not be accurate. Always check correlation before using the regression line to make a prediction.

2. The scope of the X values and “Extrapolation”

In general, we like to use X values in the scope of the data. What do we mean by this? Look at the following scatterplot of the month and cost data.

Notice that the X-axis of the scatterplot represents the number of months the company has been in business. The X values go from 0 to 24 months. This is called the “scope of the X values” or the “scope of the data”. Recall that this regression line fits the data really well and
had a strong linear relationship with a high r-squared value. This tells us the regression line should be pretty accurate for predicting costs. The thing to remember is that this regression line and the standard deviation of the residuals (prediction error) was based on the X values of 0 to 24 months. Plugging in X values between 0 and 24 will give pretty accurate predictions.

If you plug in values outside of the scope, you are using the formula for something it was not designed for. Many people love to use regression to make predictions about the future. Remember, there is no guarantee that the data will follow this pattern into the future. Also the standard deviation no longer applies to predictions outside the scope, so we cannot measure how much prediction error we may have.

**Definition of “Extrapolation”:** Plugging in an X value outside the scope of the data into a regression line or curve in order to make a prediction. The predictions outside the scope may have a significant increase in error and we will be unable to measure the error.

Extrapolating outside the scope of the X values can lead to large errors in your prediction. Also, remember the standard deviation prediction error only applies if your X value is in the scope of the data.

Look again at the month and cost data.

The average decrease in costs per month (slope) is 0.233 thousand dollars. At this rate, the line will begin to predict negative costs for the company, something that is very unlikely to happen. In fact, if we extrapolate and plug in 28 months for X into the formula, we would get a prediction of about -0.3 thousand dollars (negative $300). The costs of the company will likely not drop to negative $300.

This problem with extrapolation is a good reason why we need to recalculate with new data every few years.
I will not say never extrapolate. Many data analysts extrapolate. People are always interested in what the data tells us about the future. I would say if you do extrapolate, do not extrapolate too much (excessive extrapolation). In the last example, we may like to extrapolate a little and predict the costs in month 25, but I would not use this equation to predict the costs in month 48.

Keep the scope of the X values in mind whenever you are making predictions with a regression line. Remember if you do extrapolate, proceed with caution. You may be telling someone a predicted Y value that is very wrong.

Note about the Y-intercept: The Y-intercept of the regression line is the predicted Y value when X is zero. Therefore, the Y intercept is the prediction you would get if you plug in zero for X in the formula. In the last section, we said that the Y-intercept often does not make sense when we try to explain it. The reason for this is that sometime zero is an extrapolation. Zero may not be in the scope of the X values. In other words, if zero is not in the scope, then the formula was never designed for you to plug in zero for X. Therefore, it is not surprising that the Y-intercept value does not make much sense in context. Extrapolations can give very unusual answers sometimes.

3. Making the Prediction

To make the prediction, plug in an X value into the regression line equation for X and use order of operations and a calculator to calculate the corresponding Y value. This Y value is the prediction. Many statistics software programs have prediction functions where it will calculate the prediction automatically. Unfortunately, Statcato does not have the prediction function at this time.

Make sure to follow the order of operations when you make your prediction. Multiply the X value by the slope first. Then add the Y-intercept. Be careful of negative number mistakes.

Let us look at the month and cost regression line again.
First, you may want to use the formula in the correlation and regression printout instead of the formula at the bottom of the scatterplot. The printout gives better accuracy and your prediction will have less rounding error.

Regression Printout from Statcato
Regression equation $Y = b_0 + b_1x$

$b_0 = 6.2274$

$b_1 = -0.2326$

Example 1
Recall that in month 15, the company had an unusually high cost due to having to buy some replacement parts. Use the regression line to predict the average costs of the company in month 15 if they had not had to replace those parts.

Remember, the X value is the month and Y value is the costs in thousands of dollars. Always keep your units in mind. To make the prediction, plug in 15 for X into the regression line formula. Remember to follow the order of operations and multiply first before adding. Be careful of making a calculation error with the negative numbers.

$Y = 6.2274 + -0.2326 X$

$Y = 6.2274 + -0.2326 (15)$

$Y = 6.2274 + -3.489$

$Y = 2.7384$ thousand dollars

Therefore, if the company had not had to replace those parts, we predict their average costs would have been about 2.7384 thousand dollars ($2,738) in their 15th month in business.
How much error could there be in that prediction? 0.6295 thousand dollars (≈$630)

Remember, the standard deviation of the residuals was already calculated in the correlation and regression printout.

Standard error of estimate = 0.6295

This is the average prediction error for any prediction in the scope of the X values. Month 15 was in the scope of the X values. The units of the standard deviation of the residual errors is the same as the predicted Y value (thousands of dollars).

Example 2
Let us use the month and cost regression line to make another prediction. Predict the costs in month 25.

First, recognize that this is an extrapolation. We may prefer not to extrapolate since our answer may be very off. If you do make the prediction, here is how we would calculate it.

\[ Y = 6.2274 + -0.2326 \times \]
\[ Y = 6.2274 + -0.2326 \times 25 \]
\[ Y = 6.2274 + -5.815 \]
\[ Y = 0.4124 \text{ thousand dollars} \]

We predict that the average costs would have been about 0.4124 thousand dollars (≈$412) in their 25th month in business.

How much average error could there be in that prediction? We do not know. It could be a lot. This was an extrapolation (out of the scope of the data). The standard deviation of the residuals only applies to predictions in the scope of the data. My guess is that the costs will not continue to drop at this rate and will level off into a more consistent cost at some point.

Problem Set Section 5F
Use the scatterplot, correlation coefficient (r), standard deviation of residual errors, and the equation of the regression line to answer the following questions. Use these to answer the following questions.
Besides Nicotine, two dangerous chemicals found in cigarettes are Tar and Carbon Monoxide (CO).

1. Let nicotine be the explanatory variable and tar be the response variable.

![Scatterplot]

Correlation Coefficient \( r = 0.9614 \)

Regression Line Equation: \( Y = -1.2713 + 14.2076X \)

\( s_e = 1.2984 \) mg of tar

a) Is there correlation between the variables? Explain why.

b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?

c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

d) Provided the regression line is suitable for making predictions, predict the amount of tar we can expect to have in a cigarette that has 0.8 mg of nicotine. How far off on average could our prediction be?

e) One company is working on a cigarette with 4.75 mg of nicotine in it. Would it be ok to predict the amount of tar for this new cigarette? Why or why not?

f) Why do you think it is important that people know how much tar is in cigarettes?

2. Open the cigarettes data set. Explore the relationship between mg of nicotine and mg of carbon monoxide (CO) in cigarettes. Let nicotine be the explanatory variable and CO be the response variable.
Correlation Coefficient $r = 0.8633$

Regression Line Equation: $Y = 0.7950 + 12.3057X$

$s_e = 2.2961$ parts per million (ppm)

a) Is there correlation between the variables? Explain why.

b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?

c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

d) Provided the regression line is suitable for making predictions, predict the amount of Carbon Monoxide (CO) we can expect to have in a cigarette that has 1.2 mg of nicotine. How far off on average could our prediction be?

e) One company is working on a cigarette with 4.75 mg of nicotine in it. Would it be ok to predict the amount of carbon monoxide released from this new cigarette? Why or why not?

f) Why do you think it is important that people know how much carbon monoxide is released when a cigarette is smoked?

3. Open the Health data set. Explore the relationship between a man’s weight in pounds and his waist size in inches. Let weight be the explanatory variable and waist size be the response variable.
Correlation Coefficient $r = 0.8889$

Regression Line Equation: $Y = 33.8291 + 0.3330X$

$s_e = 4.5763$ cm

a) Is there correlation between the variables? Explain why.

b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?

c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

d) Provided the regression line is suitable for making predictions, predict the waist size of a man that weighs 200 pounds. How far off on average could our prediction be?

e) Should we use the regression line equation to predict the waist size of a man that weighs 400 pounds? Why or why not?

f) Can you think of any lurking variables that might influence waist size other than weight?

4. Open the Health data set. Explore the relationship between a woman’s systolic blood pressure and her diastolic blood pressure. Let systolic blood pressure be the explanatory variable and diastolic blood pressure be the response variable.

Correlation Coefficient $r = 0.7854$

Regression Line Equation: $Y = 8.3079 + 0.5335X$

$s_e = 7.2912$ mm of Hg

a) Is there correlation between the variables? Explain why.

b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?
c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

d) Provided the regression line is suitable for making predictions, predict the diastolic blood pressure of a person with a systolic blood pressure of 135. How far off on average could our prediction be?

e) One women with hypertension has a systolic blood pressure of 240. Would the regression line equation give an accurate prediction of her diastolic blood pressure? Why or why not?

5. Open the Bears data set. Explore the relationship between the length of a bear in inches and the weight of the bear in pounds. Let length be the explanatory variable and weight be the response variable.

Correlation Coefficient r = 0.8644

Regression Line Equation: \( Y = -393.8391 + 9.8390X \)

\( s_\epsilon = 61.8272 \) pounds

a) Is there correlation between the variables? Explain why.

b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?

c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

d) Provided the regression line is suitable for making predictions, predict the weight of a bear that is 72 inches long. How far off on average could our prediction be?

e) A young bear is only 18 inches long. Should we use the regression line equation to estimate the weight of this young bear? Why or why not?

f) Why do you think it would be useful to a researcher to be able to estimate the weight of a bear in the wild by measuring its length?
6. Open the Bears data set. Explore the relationship between the age of a bear in months and the length of the bear in inches. Let age be the explanatory variable and length be the response variable.

Correlation Coefficient $r = 0.7188$

Regression Line Equation: $Y = 48.6903 + 0.2281X$

$S_e = 7.5109$ inches

a) Is there correlation between the variables? Explain why.

b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?

c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

d) Provided the regression line is suitable for making predictions, predict the length of a bear that is 120 months old (10 years old). How far off on average could our prediction be?

e) Will this formula give accurate predictions of the length of newborn bears? Why or why not?

f) Why do you think it would be useful to a researcher to be able to estimate the length of a bear from its age?

Chapter 5 Review

In this chapter, we looked at finding a linear relationship (correlation) between two different quantitative variables with different units.

- When analyzing two different quantitative data sets, start by choosing one variable to be the response variable ($Y$) and the other to be the explanatory variable ($X$). In general,
the response variable (Y) should respond to the explanatory variable (X). If both variables respond, chose the variable you are more interested in and want to make predictions about to be the response variable (Y).

- The scatterplot and correlation coefficient “r” can tell us the strength of the linear relationship (strong, moderate, weak, or none) and the direction of the linear relationship (positive or negative). If r is close to +1, it is strong positive correlation. If r is close to -1, it is strong negative correlation. If r is close to zero, there is no correlation.
- R-squared is the percentage of variability in the y variable that can be explained by the relationship with the x variable. The higher the percentage, the stronger the relationship.
- Correlation is not causation. There are many other confounding variables that might influence the response variable other than the explanatory variable being studied.
- The regression line is the line that best fits the data and minimizes the vertical distances from all the points in the scatterplot to the line.
- Slope is the increase or decrease in the Y variables for every 1-unit increase in the X variable.
- Y-intercept is the predicted Y value when X is zero.
- The standard deviation of the residual errors (Se) gives the average vertical distance that the points are from the line. It also tells us the average prediction error.
- Residuals are the vertical distance that each point is above or below the line. Points above the line have a positive residual. Points below the line have a negative residual.
- A histogram of the residuals should be bell shaped (normal) and centered close to zero.
- A residual plot verses the x variable should be evenly spread out from the zero line and not fan shaped (not “V” shaped).
- To make a prediction with the regression line, first determine if the line fits the data. You should not make predictions when there is no correlation. If there is correlation, plug in the X value you want to predict in for X in the formula and solve for Y. The X value you plug in should be in the scope of the X values. Plugging in X values out of the scope is called “extrapolation” and can result in huge prediction errors.

**Problem Set Chapter 5 Review**

1. Define each of the following.
   a) Explanatory variable
   b) Response variable
   c) Correlation Coefficient “r”
2. When doing a correlation study with two quantitative variables, explain how we can tell which variable should be the explanatory variable (x) and the response variable (y).

3. Sanvi is a medical student studying sleep patterns and migraine headaches. She is hoping to use the hours someone sleeps in order to predict the number of migraines. She went to a clinic that specializes in treating people who suffer with headaches and recorded some data. When a patient with migraines came into the clinic, Sanvi recorded the average number of migraines they have each month and how many hours per night the person sleeps on average.

   a) Which variable should be the explanatory variable (x)?
      (Number of migraines or hours of sleep)

   b) Which variable should be the response variable (y)?
      (Number of migraines or hours of sleep)

   c) Sanvi is hoping to prove that lack of sleep causes migraines. If the data showed a strong correlation between migraines and sleep, would this prove that lack of sleep causes people to get migraines? Why or why not?

4. Match the correlation coefficients (r) with their scatterplots. (Each r value corresponds to only one graph.) For each graph describing the strength and direction of the linear relationship (correlation).
   \[ r = 0.592, \quad r = -0.993, \quad r = 0.023 \]
5. Use the following formulas to compute the slope and y-intercept for the regression line. Show your work and Round your answers to the hundredths place.

\( r = 0.819 \)

<table>
<thead>
<tr>
<th>(x) Explanatory Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) Explanatory Variable</td>
<td>( \bar{x} = 19.18 )</td>
<td>( s_x = 3.83 )</td>
</tr>
<tr>
<td>(y) Response Variable</td>
<td>( \bar{y} = 82.55 )</td>
<td>( s_y = 11.64 )</td>
</tr>
</tbody>
</table>
a) 

\[
slope : \quad m = \frac{r \cdot s_y}{s_x}
\]

To calculate the slope: $r$ times standard deviation of $y$ values, then divide by standard deviation of $x$ values.

Slope = ________________

b) 

\[
y\text{-intercept} : \quad b = \bar{y} - m(\bar{x})
\]

To calculate the $y$ – intercept (slope times the mean of $x$ values, then subtract the answer from the mean of the $y$-values).

$Y$ – Intercept = ________________

(#6-19) Let us look at the relationship between the wrist circumference of a woman (in inches) and her body mass index (BMI) in kg per meters squared. We used the health data and Statcato to create the following graphs and statistics. The explanatory variable (X) is the wrist circumference and the response variable (Y) is the body mass index.
**x = Women Wrist (in)**
**y = Women BMI (kg/m^2)**
**Sample size n = 40**

**Correlation**

\[
 r = 0.5870
\]

**Regression**

Regression equation \( Y = b_0 + b_1x \)

\[ b_0 = -29.7018 \]
\[ b_1 = 10.9407 \]

\[
 r^2 = 0.3446
\]

**Standard Deviation of the Residual Errors** = 5.0568

6. Use the scatterplot and the correlation coefficient “\(r\)” to describe the strength and direction of the linear relationship.
7. Look at the scatterplot. Estimate the scope of the x-values and put your answer below. Just give approximate values.

_________________ ≤ Wrist Circumference (Inches) ≤ ___________________

8. What was the slope of the regression line? Write a sentence to explain the slope in context.

9. What was r-squared? Convert the r-squared value into a percentage.

10. Write a sentence to explain r-squared in context.

11. List three confounding variables that might influence the body mass index of a woman other than size of her wrist.

12. Does this study prove that the size of these women’s wrist causes them to have a certain body mass index? (Yes or No) Explain why.

13. What was the standard deviation of the residual errors? What units does the standard deviation of the residuals have in this problem?

14. Explain the two meanings of the standard deviation of the residuals in the context of the wrist and BMI data.

Residual Plot and Histogram of the Residuals
15. Does the residual plot above on the left show a “V” shape or is it evenly spread out?

16. Does the residual plot above on the left show a curved pattern in the data?

17. Is the Histogram of the Residuals bell shaped? (Yes or No)

18. Is the histogram of the residuals centered close to zero? (Yes or No)

19. Use your calculator and the regression equation below to predict the body mass index for a woman with a wrist circumference of 4.5 inches. (Plug in 4.5 for X.) Show work.

\[ Y = 48.802 - 8.367 \times X \]

20. In the previous problem, how far off could our BMI prediction be on average (prediction error)? (No calculation needed)

21. Will this formula give an accurate prediction for the body mass index for a female child with a wrist circumference of 3.1 inches? Why or why not?
Project Chapter 5 - Linear Quantitative Relationships

Group Project Instructions- Your team will analyze the linear relationship between two quantitative data sets. The ordered pair data can be found at matt-teachout.org under “Int Alg for Stats” and “Data Sets”. Look for the data set that says, “Correlation data project#5”. Make sure your poster has the following items. You will be presenting your poster to the other students in the class.

Topics: IQ/Brain Volume, MLB Runs Scored/Attendance, MLB Runs Allowed/Wins, Price Item/Customer Satisfaction, Meat/illness, CEO Golf Score/Stock Price, Swim time/Pulse, Boats/Manatee Deaths, Cost of Living/Aviation Pay, Poverty/BMI, Alcohol/Tobacco in England

- Pick one of the pairs of quantitative variables and pick which should be X and which should be Y. The poster should give the explanatory variable (x) and response variables (y), what the units are for x and y.
- Why this topic was important or interesting to your group.
- Poster should have a scatterplot with regression line.
- Poster should have the correlation coefficient (r) and a sentence describing its meaning.
- Poster should have r-squared and a sentence describing its meaning.
- Poster should have the regression line equation
- Write a sentence describing the meaning of the slope
- Write a sentence describing the meaning of the y-intercept.
- Poster should have the Standard Deviation of the Residual Errors and two sentences describing both of the meanings of Standard Deviation.
- What is the scope of the x-values?
- Pick any x-value in the scope. Plug in that x value into the regression line equation and predict the y value. How much error could be in that prediction?
- Poster should have a residual plot verses the x-value. Is the residual plot “V” shaped or evenly spread out?
- Poster should have the histogram of the residuals. Is the histogram bell shaped? Is the histogram centered at zero?
- Decorate your poster to spark interest.
- Give a good presentation of the poster to classmates explaining everything.