Chapter 6 – Linear Quantitative Relationships

Introduction: In previous sections, we have looked at how to analyze a categorical data set. Then we looked at relationships between categorical data sets. In chapters 4 and 5, 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 calculate 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. The slope calculation is much more complicated. 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. While a basic understanding of slope and lines is helpful, realize that the calculations will be much more complicated. The study of linear quantitative relationships in statistics is extremely different from algebra.

Technology: Many of the calculations in this chapter are extremely difficult to calculate by hand with a calculator. As with many statistics and graphs, we prefer to use a computer software like StatKey to calculate. Then we can focus on understanding the meaning behind these statistics and graphs and what they tell us about the world around us.

Terminology

Correlation: Statistical analysis that determines if there is a relationship between two different quantitative variables.

Regression: Statistical analysis that involves finding the line or model that best fits a quantitative relationship, using the model to make predictions, and analyzing error in those predictions.

Scatter Plot: A graph that shows the x-axis, y-axis and points at all of the ordered pairs in the data.

Slope (b_1): The average amount of increase or decrease in the y-variable for every one-unit increase in the x-variable.

Y-Intercept (b_0): The predicted y-value when the x-value is zero.

Regression Line ($\hat{y} = b_0 + b_1 x$): Also called the "Line of Least Squares" or the "Line of Best Fit". This line will be represented by a linear equation, but realize it is not a line between two points. It is the line that best fits many points.

Section 6A – 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 responses 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.

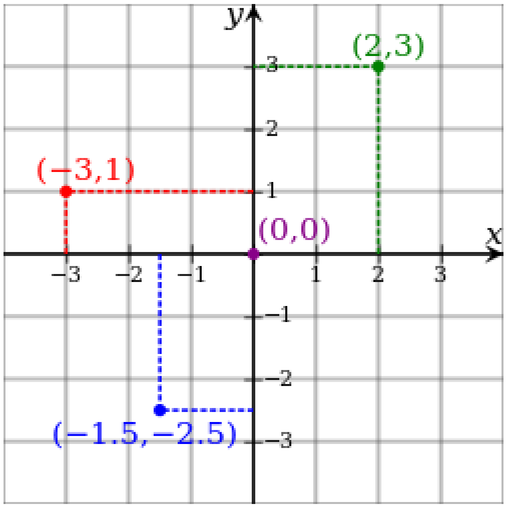
Ordered Pairs

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 difficult job. Websites, articles and various sources often disagree with one another. Therefore, these values are just approximations and may not be perfectly accurate.

Country	August 2017 National Debt (Billions of U.S. Dollars)	Unemployment Rate (%)
France	2472.9	9.6
Mexico	474.5	3.5
U.S.A	19873.5	
Japan	9094.9	3.06
Canada	826.5	
Australia	406.0	
United Kingdom	2279.8	

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 Billion $, 3.5% unemployment) and Japan's data as (9094.9 Billion $, 3.06% 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 degrees Fahrenheit , 3 thousand dollars) and (-3 , 1) is really describing the ordered pair (-3 degrees Fahrenheit , 1 thousand dollars).

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.

<u>How to create a Scatterplot with StatKey</u>:

- Open the data. Then open a new spreadsheet and paste the two quantitative data sets next to each other side by side. It is customary to have the explanatory column (X) on the left and the response column (Y) on the right. Then copy the two columns together.
- Now we will go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Quantitative Variables". Click on "Edit Data" at the top. Push Control A on your keyboard to highlight old data and then push "delete" on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the "Edit Data" field. If your data has a title, click the box that says "Data has header row". If your data does not have a title, do NOT check the box that says "Data has header row". Then press OK. The graph you see is the scatterplot.

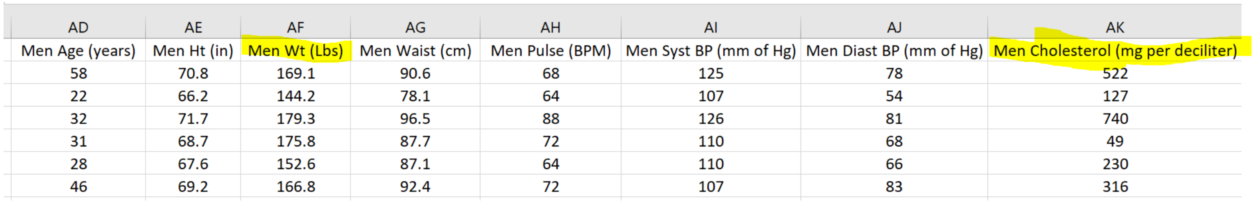
Note: *Your X variable should be on the horizontal axis and the Y variable should be on the vertical axis. If the X and Y variables are backward in the graph, simply click the "switch variables" button. It is also nice to check the "show regression line" box. The regression line is the line that best fits the points in the scatterplot. StatKey will also given us some statistics to help understand the relationship. These statistics we will explore in future sections.*

Note: *The titles of the quantitative columns of data can be problematic sometimes if they are too long. If StatKey gives an error message, it may be a problem with the title. If that is the case, you may need to either shorten the title or completely delete the title.*

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.

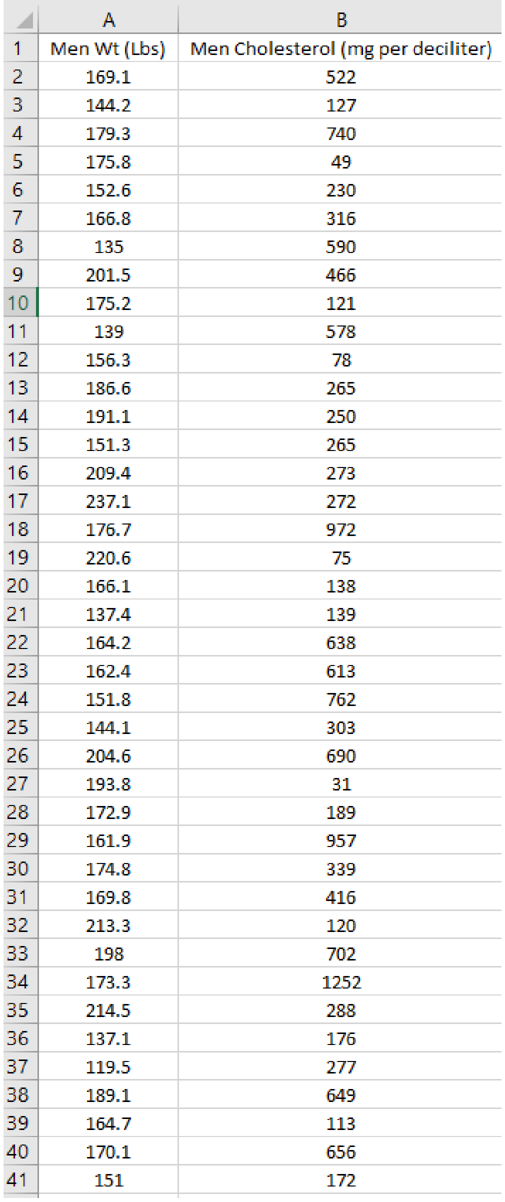
Step 1: First notice that the health data does have ordered pair data containing the weight and cholesterol of the same 40 men. Having ordered pair data is vital to studying quantitative relationships. I cannot take a weight of one man and pair it with the cholesterol of a different man. The weight and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



AD	AE	AF	AG	AH	AI	AJ	AK
Men Age (years)	Men Ht (in)	Men Wt (Lbs)	Men Waist (cm)	Men Pulse (BPM)	Men Syst BP (mm of Hg)	Men Diast BP (mm of Hg)	Men Cholesterol (mg per deciliter)
58	70.8	169.1	90.6	68	125	78	522
22	66.2	144.2	78.1	64	107	54	127
32	71.7	179.3	96.5	88	126	81	740
31	68.7	175.8	87.7	72	110	68	49
28	67.6	152.6	87.1	64	110	66	230
46	69.2	166.8	92.4	72	107	83	316

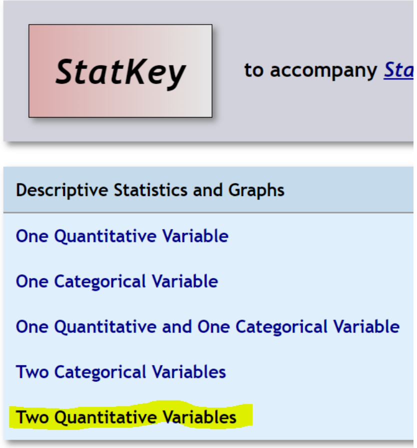
Step 2: The next step is to choose which variable is to be the explanatory (X) and which variable should be the response variable (Y). The variable you want to predict should be the response variable (Y). I want to maybe predict cholesterol levels based on a man's weight, so I will make cholesterol my response (Y). Therefore, I will make the weight the explanatory variable (X).

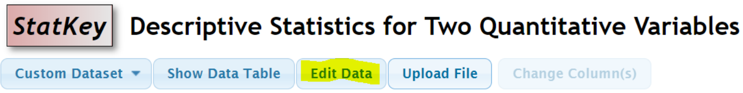
Step 3: Now we need to open a new excel spread sheet and copy and paste the men's weight and men's cholesterol data into two columns next to each other. It is common to put the X variable on the left and the Y variable on the right.

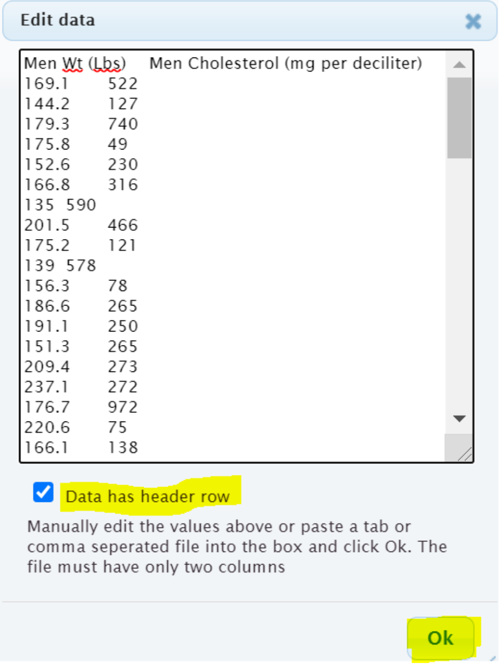


	A	B
1	Men Wt (Lbs)	Men Cholesterol (mg per deciliter)
2	169.1	522
3	144.2	127
4	179.3	740
5	175.8	49
6	152.6	230
7	166.8	316
8	135	590
9	201.5	466
10	175.2	121
11	139	578
12	156.3	78
13	186.6	265
14	191.1	250
15	151.3	265
16	209.4	273
17	237.1	272
18	176.7	972
19	220.6	75
20	166.1	138
21	137.4	139
22	164.2	638
23	162.4	613
24	151.8	762
25	144.1	303
26	204.6	690
27	193.8	31
28	172.9	189
29	161.9	957
30	174.8	339
31	169.8	416
32	213.3	120
33	198	702
34	173.3	1252
35	214.5	288
36	137.1	176
37	119.5	277
38	189.1	649
39	164.7	113
40	170.1	656
41	151	172

Step 4: Now we will go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Quantitative Variables". Click on "Edit Data" at the top. Push Control A on your keyboard and delete all old data in the edit data field. Then paste the two columns of quantitative data into the "Edit Data" field. Since these columns of data have titles, we will click the box that says "Data has header row". Then press OK.


StatKey
to accompany Sta
Descriptive Statistics and Graphs
One Quantitative Variable
One Categorical Variable
One Quantitative and One Categorical Variable
Two Categorical Variables
Two Quantitative Variables

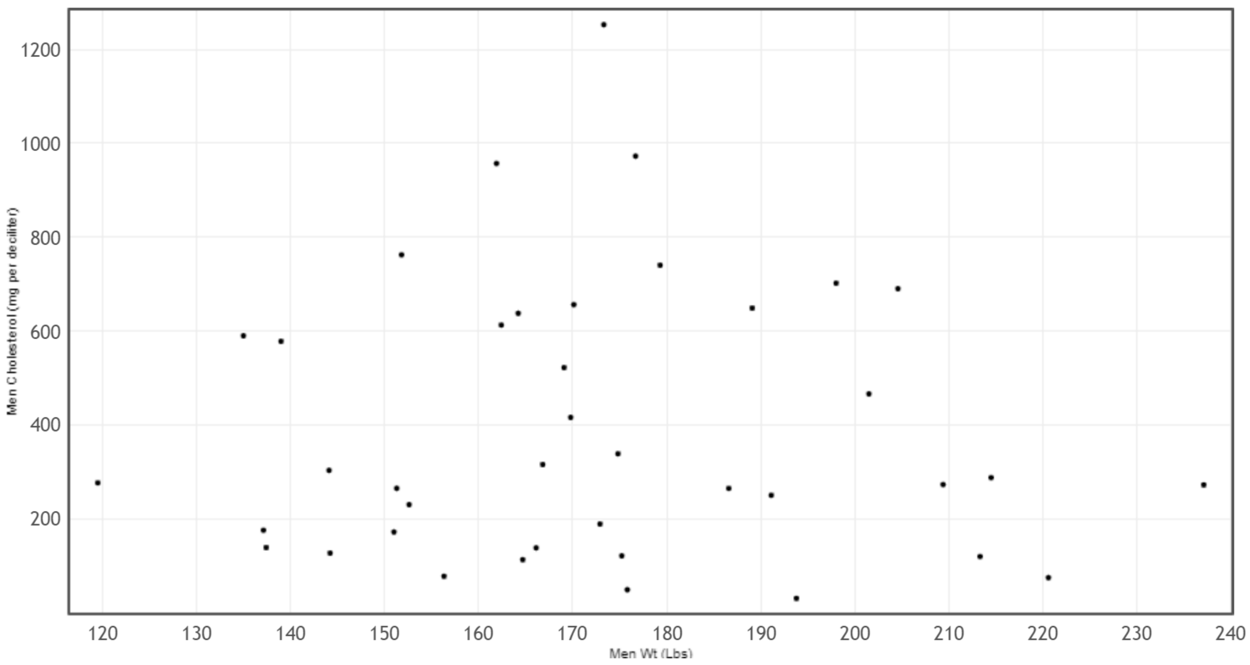




```
Men Wt (Lbs)    Men Cholesterol (mg per deciliter)
169.1       522
144.2       127
179.3       740
175.8       49
152.6       230
166.8       316
135  590
201.5       466
175.2       121
139  578
156.3       78
186.6       265
191.1       250
151.3       265
209.4       273
237.1       272
176.7       972
220.6       75
166.1       138
```

☑ Data has header row

Manually edit the values above or paste a tab or comma seperated file into the box and click Ok. The file must have only two columns



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.

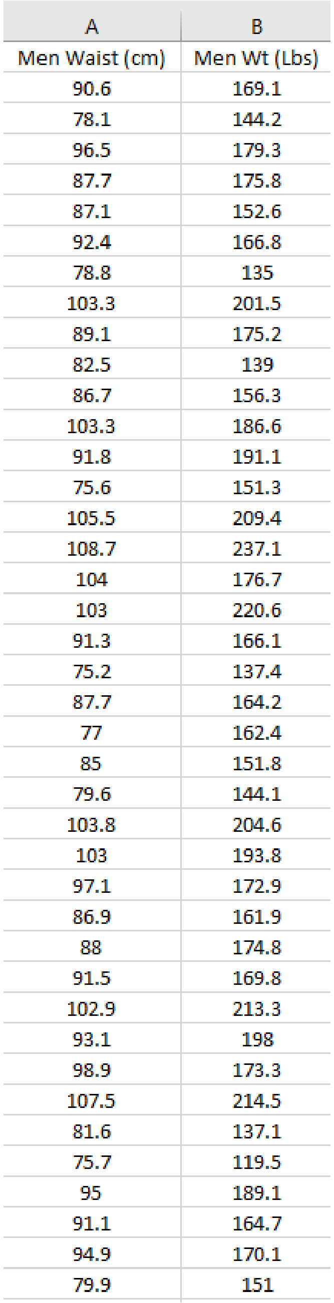
Key Question: Start by asking yourself a simple question. The points will not lie on a line or a curve, but can I imagine a line or a curve that the points could be relatively close to? That is the key. You have to take all of the points into account.

In the men's weight / men's cholesterol example, the points seem very scattered all over with no obvious pattern. Do not force a computer program like StatKey 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, since the points will not follow any particular pattern. 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. Sometimes we refer to this as "no correlation", "no relationship" or "no association".

Example 2

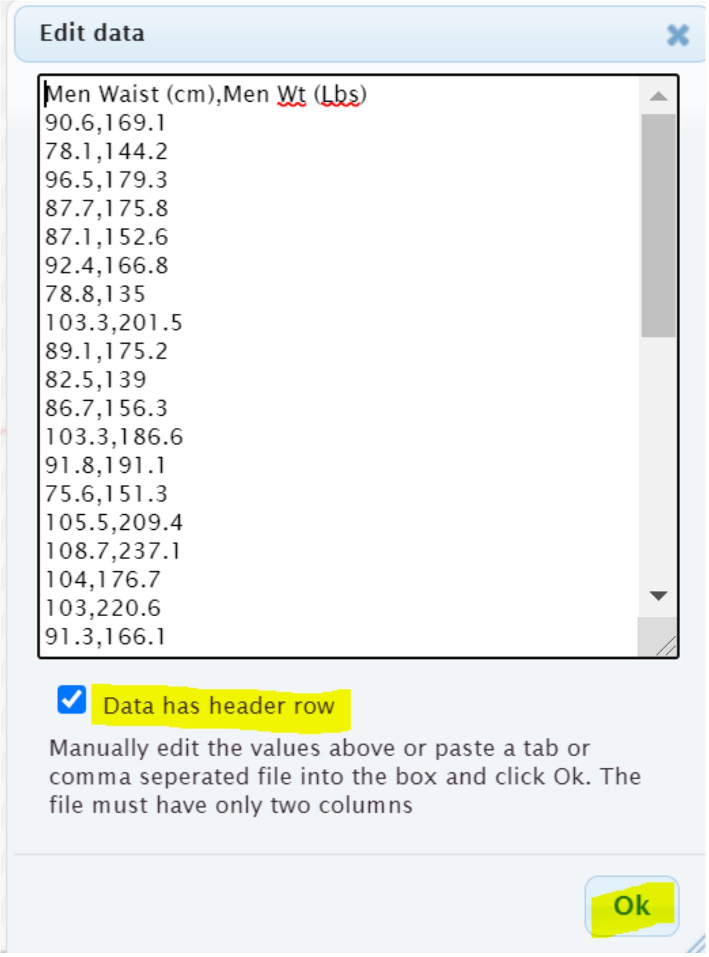
Let us look at another example from the health data. This time I wanted to look at the weight of the men (in pounds) and the waist size of the men (in centimeters).
Step 1: Do I have ordered pair data? Yes. The health data contained the weights and waist sizes of the same 40 men.

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).



A	B
Men Waist (cm)	Men Wt (Lbs)
90.6	169.1
78.1	144.2
96.5	179.3
87.7	175.8
87.1	152.6
92.4	166.8
78.8	135
103.3	201.5
89.1	175.2
82.5	139
86.7	156.3
103.3	186.6
91.8	191.1
75.6	151.3
105.5	209.4
108.7	237.1
104	176.7
103	220.6
91.3	166.1
75.2	137.4
87.7	164.2
77	162.4
85	151.8
79.6	144.1
103.8	204.6
103	193.8
97.1	172.9
86.9	161.9
88	174.8
91.5	169.8
102.9	213.3
93.1	198
98.9	173.3
107.5	214.5
81.6	137.1
75.7	119.5
95	189.1
91.1	164.7
94.9	170.1
79.9	151

Step 3: Make a scatterplot with technology. We will use StatKey and follow the same steps as in example1.

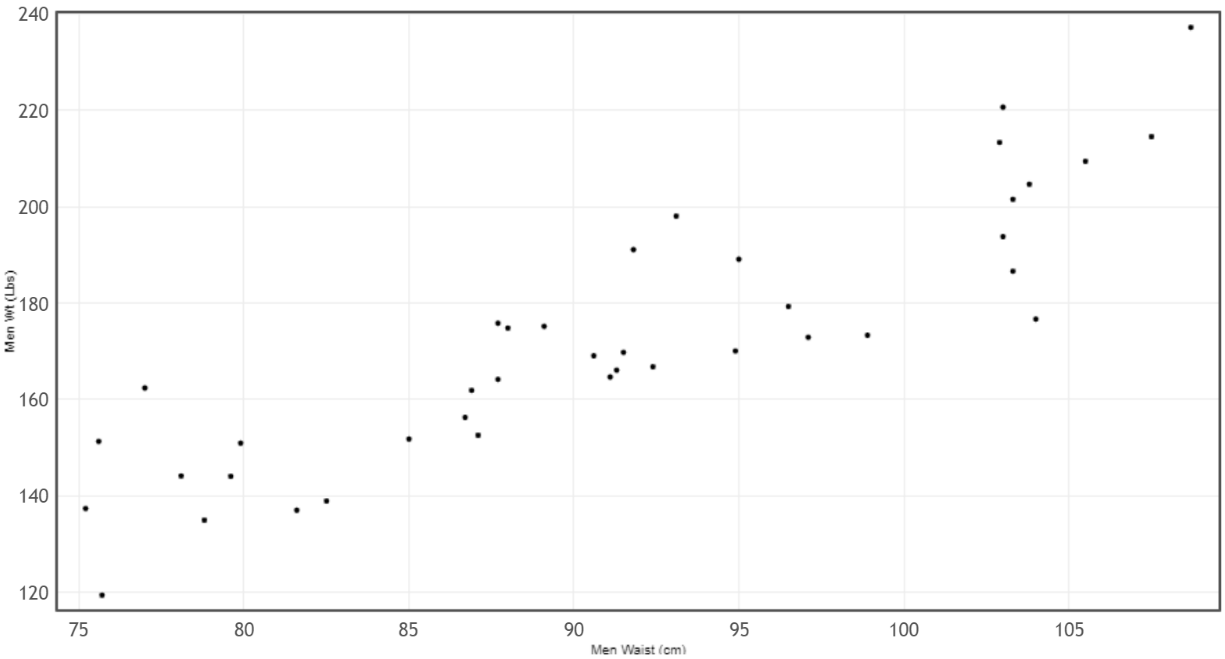


Edit data

```
Men Waist (cm),Men Wt (Lbs)
90.6,169.1
78.1,144.2
96.5,179.3
87.7,175.8
87.1,152.6
92.4,166.8
78.8,135
103.3,201.5
89.1,175.2
82.5,139
86.7,156.3
103.3,186.6
91.8,191.1
75.6,151.3
105.5,209.4
108.7,237.1
104,176.7
103,220.6
91.3,166.1
```

☑ Data has header row

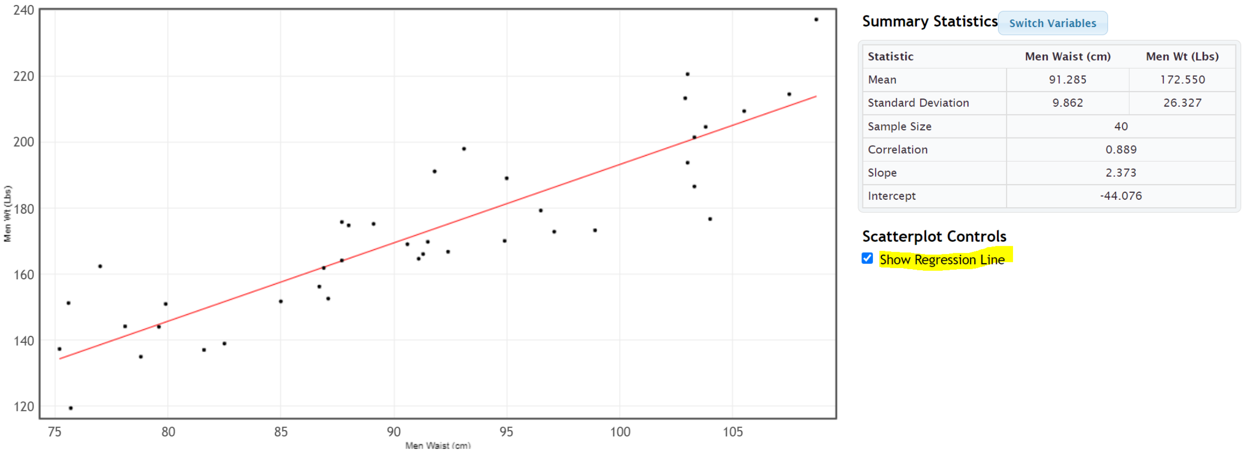
Manually edit the values above or paste a tab or comma seperated file into the box and click Ok. The file must have only two columns



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”. In fact we can have the computer find the line of best fit. This is called the “regression line”. By clicking on “Show Regression Line” we see the line of best fit drawn. Notice the points do not go through the line, but they are close to the line. Also notice the line goes up from left to right.



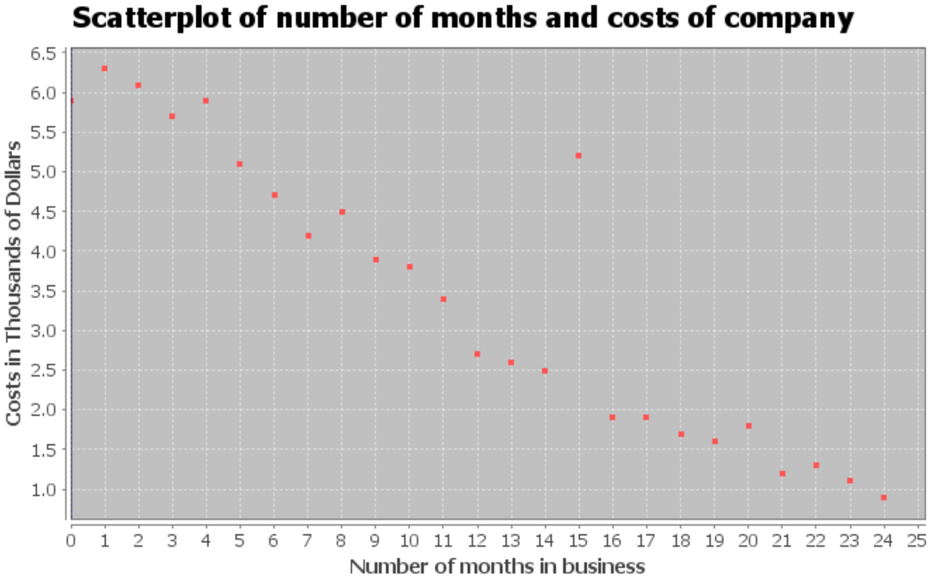
Summary Statistics

Statistic	Men Waist (cm)	Men Wt (Lbs)
Mean	91.285	172.550
Standard Deviation	9.862	26.327
Sample Size	40	
Correlation	0.889	
Slope	2.373	
Intercept	-44.076	

Scatterplot Controls

☑ Show Regression Line

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.



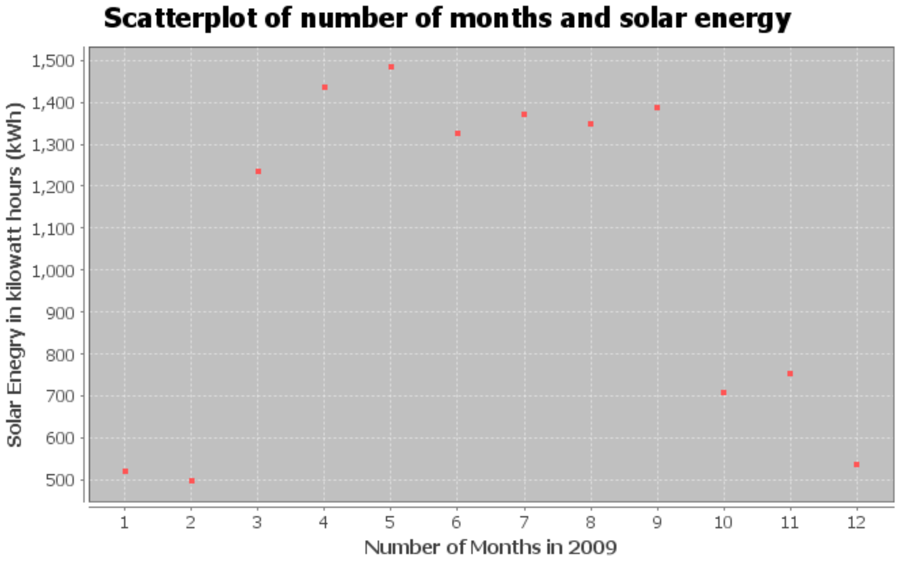
Notice most of the points do seem to be close to a line. They are following a linear pattern. In fact, the linear pattern seems to be going down from left to right. We often call this a “negative linear relationship” or a “negative correlation”.

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.

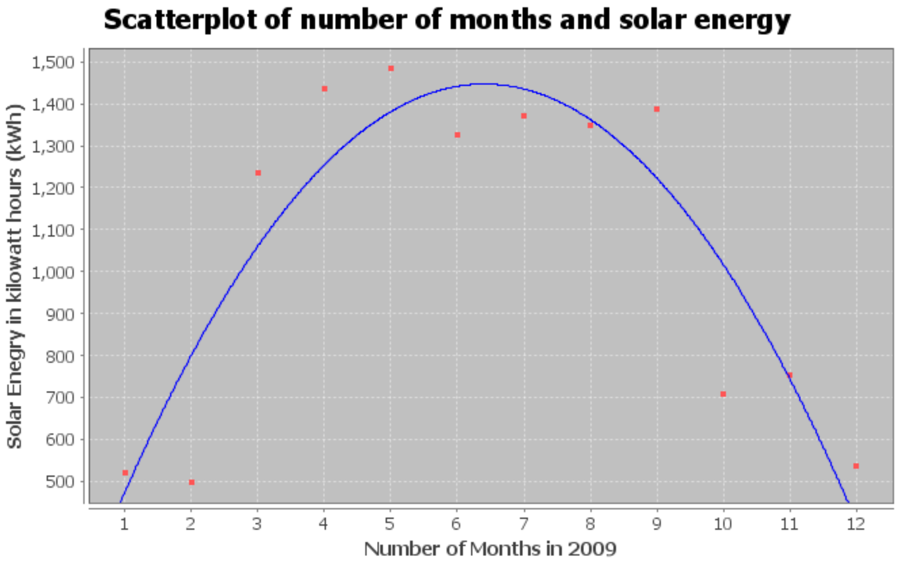
Aside from the outlier, this graph is good news for the company. As the months increase, the costs of the company seem to be decreasing dramatically.

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”. So this scatterplot indicates that there is no linear relationship, but there is a quadratic relationship between time in months and solar energy.

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 Correlation** (Positive Linear Relationship): Points in the scatterplot seem to be close to a line that is going up from left to right (increasing). This is sometimes called a “direct relationship” in mathematics. As the X variable increases, the Y variable also increases. Look out for points that do not seem to fit the linear pattern (outliers).
- **Negative Correlation** (Negative Linear Relationship): Points in the scatterplot seem to be close to a line that is going down from left to right (decreasing). This is sometimes called an “indirect relationship” or “inverse relationship” in mathematics. As the X variable increases, the Y variable decreases. Look out for points that do not seem to fit the linear pattern (outliers).
- **Curved Relationship** (Non-linear Relationship): Points in the scatterplot do not seem to follow any line, but do 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 curve pattern (outliers).

Note: The use of the word “correlation” denotes a linear relationship between two quantitative variables. If you have a relationship between categorical variables or between a categorical and quantitative variable, we usually refer to that as an “association” or a “relationship”.

Note: Correlation does not imply causation. It is wrong to state that one variable causes another just because they are related.

Problem Set Section 6A

Directions: Open the health data from Canvas or from www.matt-teachout.org. Use the indicated columns of data to create scatterplots with StatKey. Save the scatterplot on a word document or make a general sketch of the graph on a sheet of paper and answer the questions.

How to create a Scatterplot with StatKey:

- Open the data. Then open a new spreadsheet and paste the two quantitative data sets next to each other side by side. It is customary to have the explanatory column (X) on the left and the response column (Y) on the right. Then copy the two columns together.
- Now we will go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Quantitative Variables". Click on "Edit Data" at the top. Push Control A on your keyboard to highlight old data and then push "delete" on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the "Edit Data" field. If your data has a title, click the box that says "Data has header row". If your data does not have a title, do NOT check the box that says "Data has header row". Then press OK. The graph you see is the scatterplot. If your scatterplot has the X and Y variables backward, simply click the "Switch Variables" button in StatKey.

1. Explore the relationship between a woman's weight and height.

   a) Which variable did you chose to be the explanatory variable?
   b) Which variable did you chose to be the response variable?
   c) 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.
   d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
   e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

2. Explore the relationship between a man's weight and height.

   a) Which variable did you chose to be the explanatory variable?
   b) Which variable did you chose to be the response variable?
   c) 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.
   d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
   e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

3. Explore the relationship between a woman's cholesterol and age.

   a) Which variable did you chose to be the explanatory variable?
   b) Which variable did you chose to be the response variable?
   c) 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.
   d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
   e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

4. Explore the relationship between a man's cholesterol and age.

    a) Which variable did you chose to be the explanatory variable?
    b) Which variable did you chose to be the response variable?
    c) 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.
    d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
    e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

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

    a) Which variable did you chose to be the explanatory variable?
    b) Which variable did you chose to be the response variable?
    c) 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.
    d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
    e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

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

    a) Which variable did you chose to be the explanatory variable?
    b) Which variable did you chose to be the response variable?
    c) 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.
    d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
    e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

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

    a) Which variable did you chose to be the explanatory variable?
    b) Which variable did you chose to be the response variable?
    c) 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.
    d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
    e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

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

    a) Which variable did you chose to be the explanatory variable?
    b) Which variable did you chose to be the response variable?
    c) 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.
    d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
    e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

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

a) Which variable did you chose to be the explanatory variable?
b) Which variable did you chose to be the response variable?
c) 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.
d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

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

a) Which variable did you chose to be the explanatory variable?
b) Which variable did you chose to be the response variable?
c) 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.
d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

Section 6B – 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 linear 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.

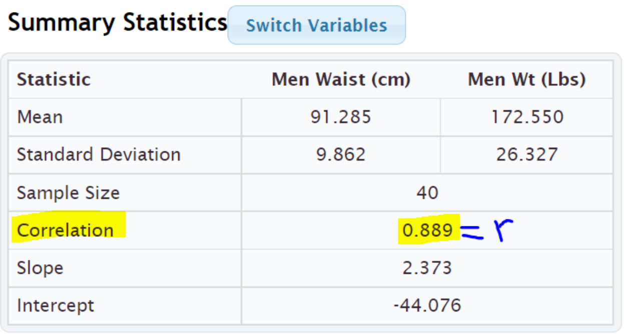
Always start by looking at a scatterplot. Have an idea of what the scatterplot looks like before you try to find and interpret the correlation coefficient.

Step 2: Calculate "r"

Once you have seen the scatterplot, use a statistics software like StatKey to calculate the correlation coefficient "r". Warning: The correlation coefficient "r" is extremely difficult and time consuming to calculate. No data analyst or statistician calculates "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.

Creating the Correlation Coefficient "r" with StatKey:

- To put the data into StatKey, you will want to open a fresh excel spreadsheet and paste the two quantitative data sets next to each other side by side. Then copy the two columns together.
- Now we will go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Quantitative Variables". Click on "Edit Data" at the top. Push Control A on your keyboard to highlight old data and then push "delete" on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the "Edit Data" field. If your data has a title, click the box that says "Data has header row". If your data does not have a title, do NOT check the box that says "Data has header row". Then press OK.
- You will see the correlation coefficient "r" under "Summary Statistics". Look next to "Correlation".



Summary Statistics

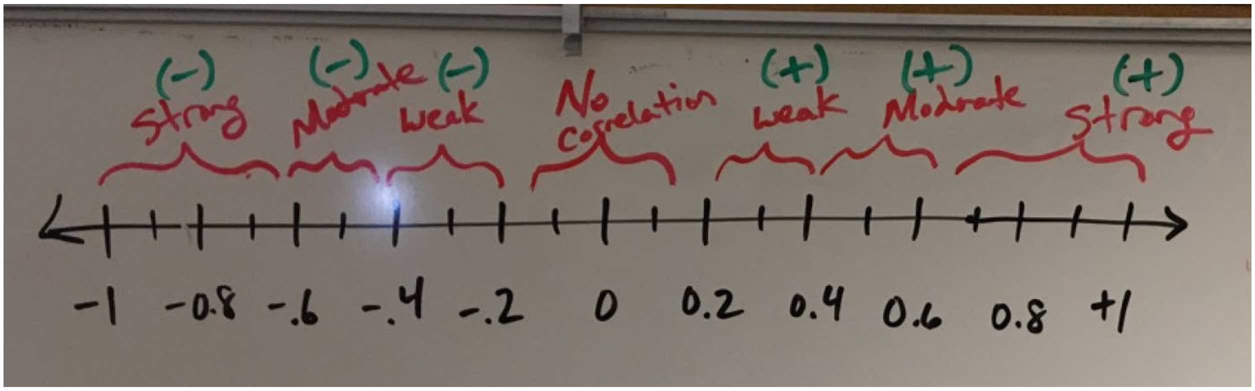
Statistic	Men Waist (cm)	Men Wt (Lbs)
Mean	91.285	172.550
Standard Deviation	9.862	26.327
Sample Size	40	
Correlation	0.889 = r	
Slope	2.373	
Intercept	-44.076	

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 the regression line and positive means that the regression 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 the regression line and negative means that the regression 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 switched. The calculation for “r” does not depend on which variable is X or Y.
- I always find it is helpful to keep the following number line in mind when interpreting a correlation coefficient.

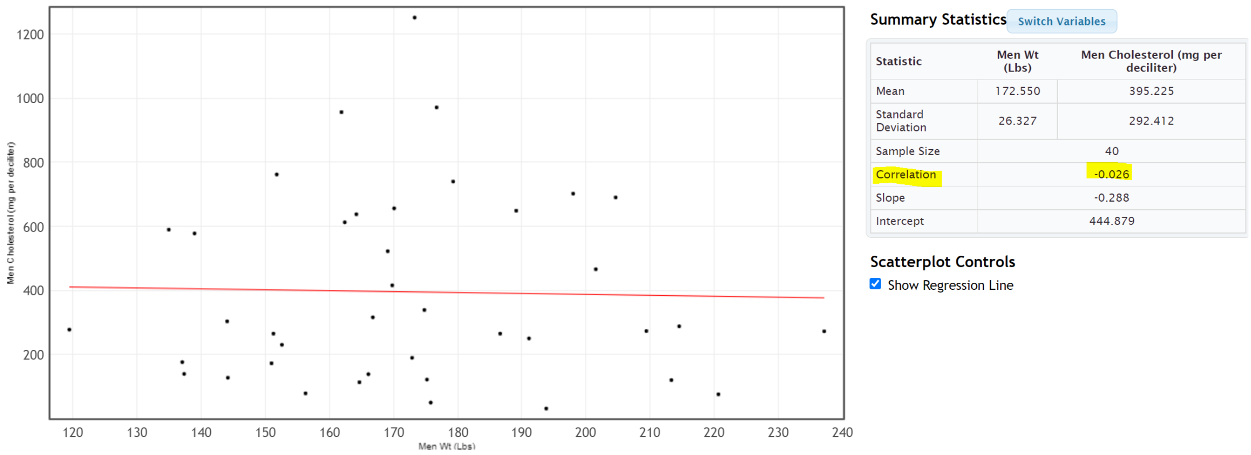


Note: The question is often asked, what is the strength and direction of the linear relationship (correlation)? Remember the strength (strong, moderate, weak, none) is asking how close the points are to the line. The direction is asking if the line is going up or down from left to right (increasing or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

Example 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 StatKey to create the following scatterplot and summary statistics.



Summary Statistics

Statistic	Men Wt (Lbs)	Men Cholesterol (mg per deciliter)
Mean	172.550	395.225
Standard Deviation	26.327	292.412
Sample Size	40	
Correlation	-0.026	
Slope	-0.288	
Intercept	444.879	

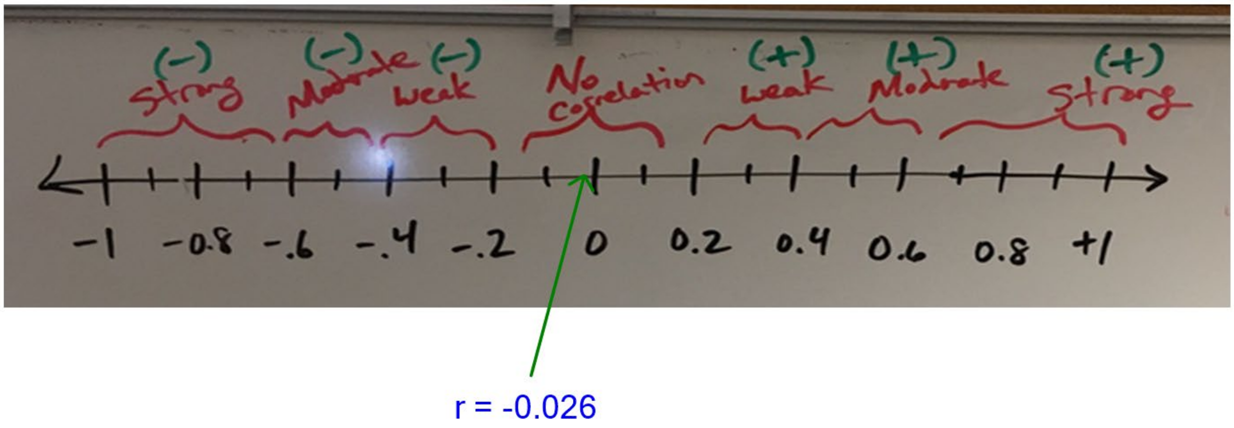
Scatterplot Controls

☑ Show Regression Line

In this scatterplot, the points seem very scattered all over with no obvious pattern. The points do not seem to follow the regression line. Visually we think that there may be no linear relationship. Does the correlation coefficient confirm this suspicion?

You will find the correlation coefficient “r” under “Summary Statistics”. Look for the number next to “Correlation”. We see that the correlation coefficient r = −0.026

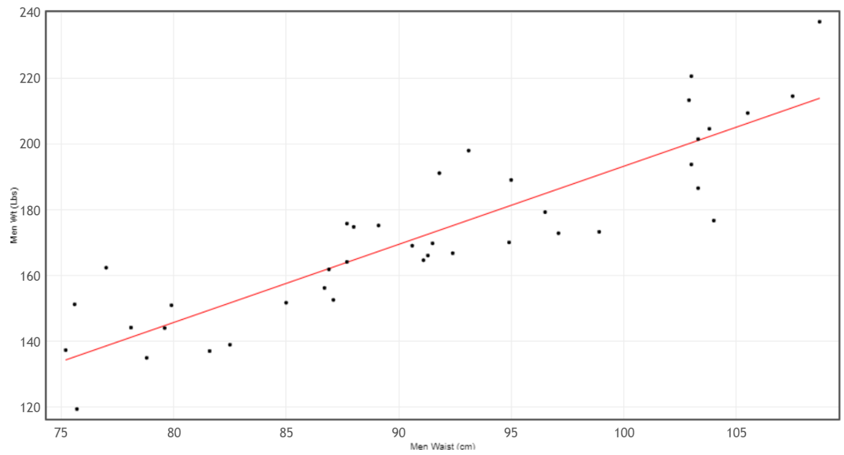
Interpretation: So what does this statistic of r = −0.026 tell us? Looking at the number line, we see that the r value of −0.026 though negative is extremely close to zero on the number line. This does <u>not</u> tell us there is negative correlation. This statistics agrees with what we said earlier when we looked at the scatterplot. There seems to be no linear relationship (no correlation) between the weight and cholesterol of these men.

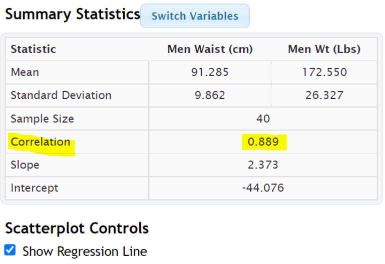


Example 2

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 StatKey to find the following scatterplot and correlation coefficient “r”.




Summary Statistics Switch Variables

Statistic	Men Waist (cm)	Men Wt (Lbs)
Mean	91.285	172.550
Standard Deviation	9.862	26.327
Sample Size	40	
Correlation	0.889	
Slope	2.373	
Intercept	-44.076	

Scatterplot Controls

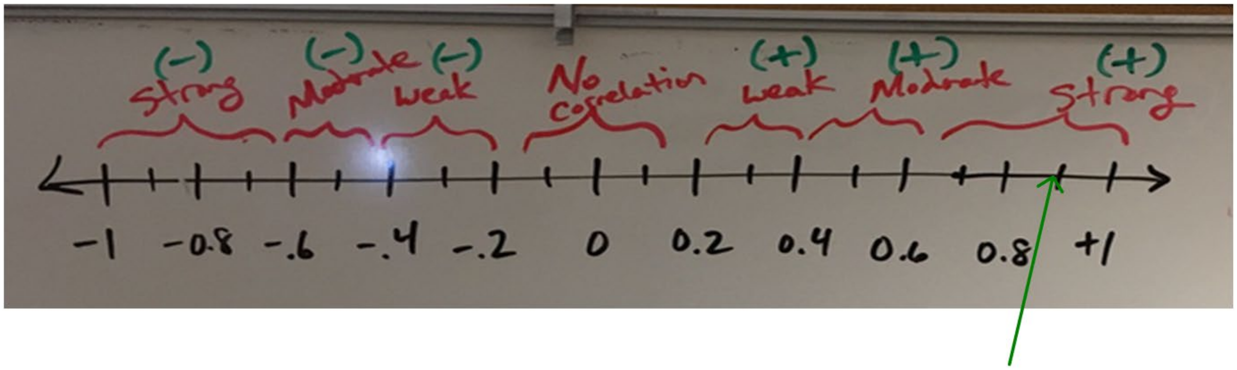
☑ Show Regression Line

The points in the scatterplot seem to be close to the regression line and the line is going up from left to right. 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".

The correlation coefficient came out to be 0.889. I like to put a positive sign in front of the correlation coefficient since 0.889 really means +0.889 and the sign of the correlation coefficient is important to the interpretation.

Interpretation: "Strong, Positive Correlation"

Look again at the correlation coefficient number line. Notice that +0.889 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).

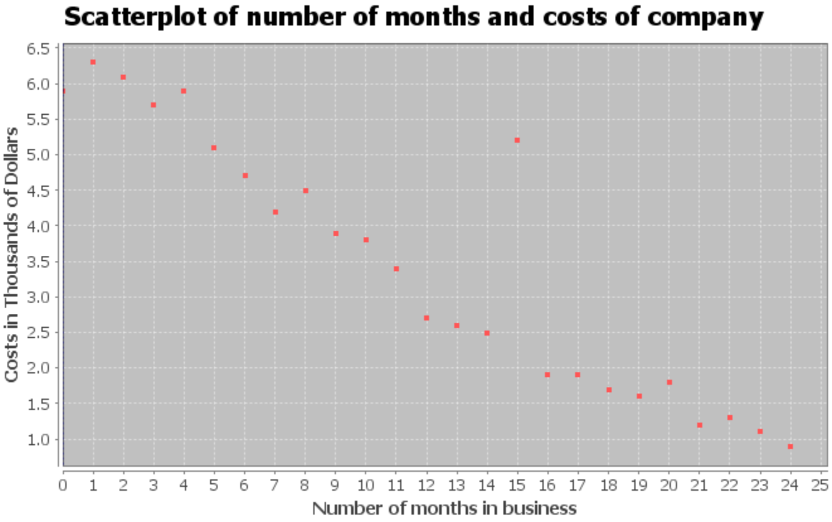


r = +0.889

Important Note: Correlation does not imply causation. Just because there is a strong correlation between the waist size and weight of the these 40 men, it does NOT imply that waist size CAUSES a man to have a certain weight. There are other factors involved. In order to prove cause and effect, the data must be collected with experimental design and the confounding variables must be controlled. Neither is the case in this example.

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. The computer also gave us the correlation coefficient “r”. Let’s see if we can interpret it.

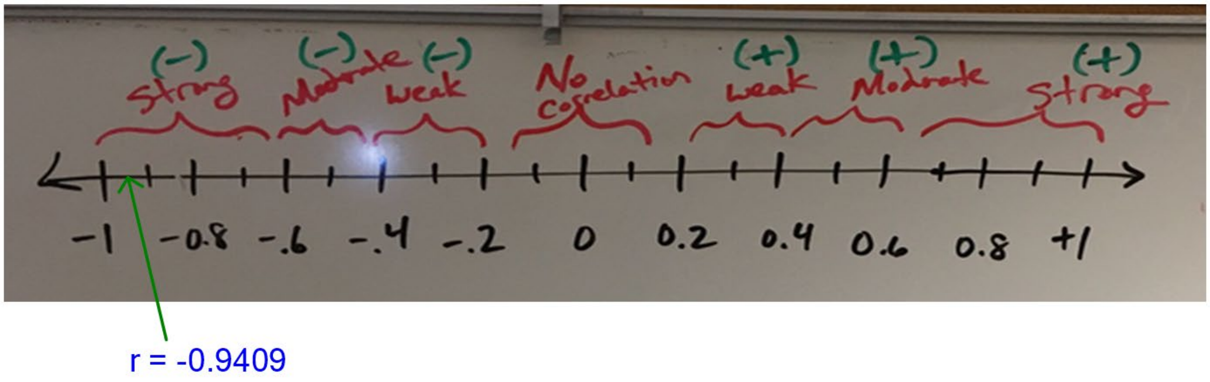


	Correlation Coefficient
r	−0.9409

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 the correlation coefficient “r” to shed some light on this relationship. Is the outlier influential? It seems like the relationship should be pretty strong, but how is the outlier effecting the overall strength of the relationship?

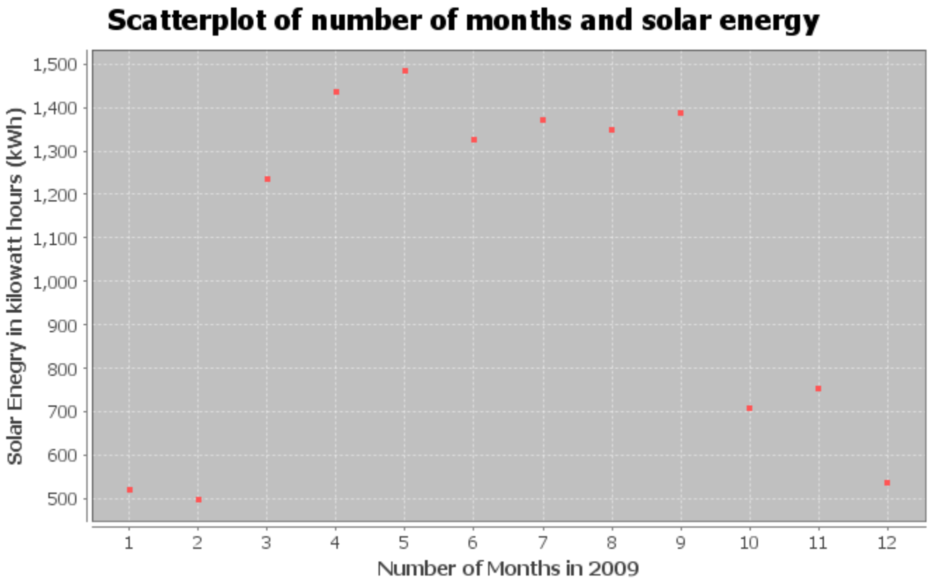
Interpretation: Look at the correlation coefficient number line again. The correlation coefficient of r = −0.9409 is very close to −1. That means that despite the outlier, the correlation is still very strong. This tells us that the outlier is not very influential. The overall interpretation is still 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.



Important Note: Correlation does not imply causation. Just because there is a strong correlation between time in months and costs for this business, it does NOT imply that time CAUSES the company to have certain costs. There are other factors involved. In order to prove cause and effect, the data must be collected with experimental design and the confounding variables must be controlled. Neither is the case in this example.

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 to calculate “r”, 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.

	Test Statistic
r	−0.0568

This is why it is so important to look at scatterplot before interpreting the correlation coefficient. We need to know what shape we are dealing with. This correlation coefficient 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 at all between time (months) and solar energy. There is actually a strong relationship, but it is just not linear.

Interpretation: “Strong Curved Relationship”

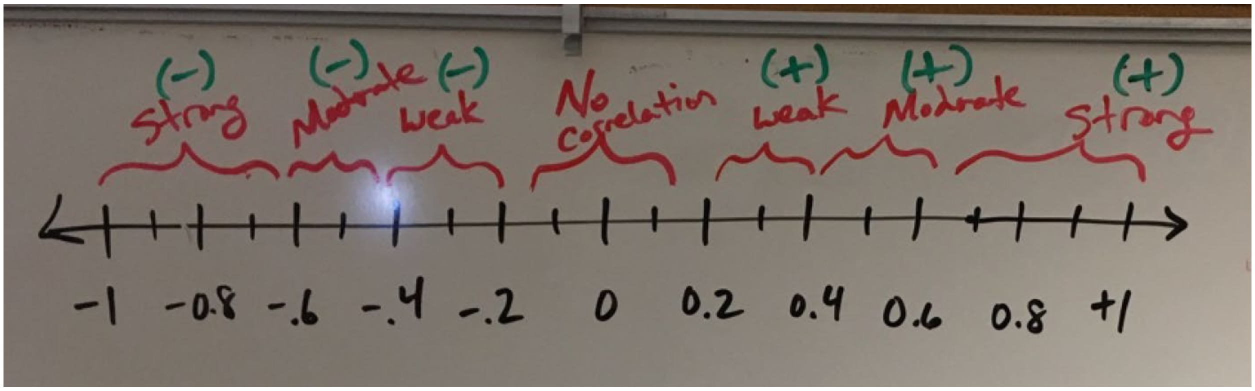
It is important to know the shape before calculating statistics. Calculating statistics for a linear relationship when it is really curved can be very misleading. We will learn in our next chapter how to analyze curved relationships.

Problem Set Section 6B

Directions: Open the "Health Data" from Canvas or from www.matt-teachout.org. Use the indicated columns of data to create scatterplots with StatKey and calculate the correlation coefficient "r".

How to create a Scatterplot and calculate the correlation coefficient "r" with StatKey:

- Open the data. Then open a new spreadsheet and paste the two quantitative data sets next to each other side by side. It is customary to have the explanatory column (X) on the left and the response column (Y) on the right. Then copy the two columns together.
- Now we will go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Quantitative Variables". Click on "Edit Data" at the top. Push Control A on your keyboard to highlight old data and then push "delete" on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the "Edit Data" field. If your data has a title, click the box that says "Data has header row". If your data does not have a title, do NOT check the box that says "Data has header row". Then press OK. The graph you see is the scatterplot. If your scatterplot has the X and Y variables backward, simply click the "Switch Variables" button in StatKey. The correlation coefficient "r" is listed "Summary Statistics" where it says "Correlation".



1. Explore the relationship between a woman's weight and height.

   f) What is the correlation coefficient "r".

   g) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

2. Explore the relationship between a man's weight and height.

   a) What is the correlation coefficient "r".

   b) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

3. Explore the relationship between a woman's cholesterol and age.

   f) What is the correlation coefficient "r".

   g) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

4. Explore the relationship between a man’s cholesterol and age.

   a) What is the correlation coefficient “r”.
   b) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

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

   a) What is the correlation coefficient “r”.
   b) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

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

   a) What is the correlation coefficient “r”.
   b) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

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

   a) What is the correlation coefficient “r”.
   b) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

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

   a) What is the correlation coefficient “r”.
   b) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

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

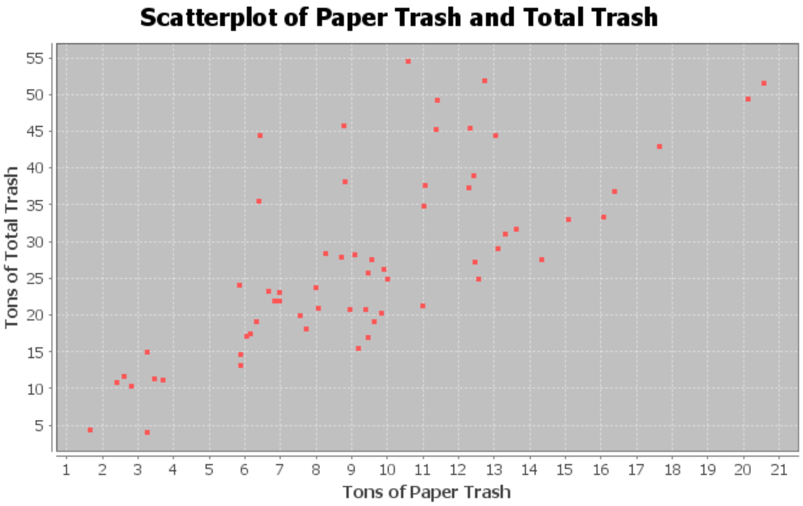
   a) What is the correlation coefficient “r”.
   b) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

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

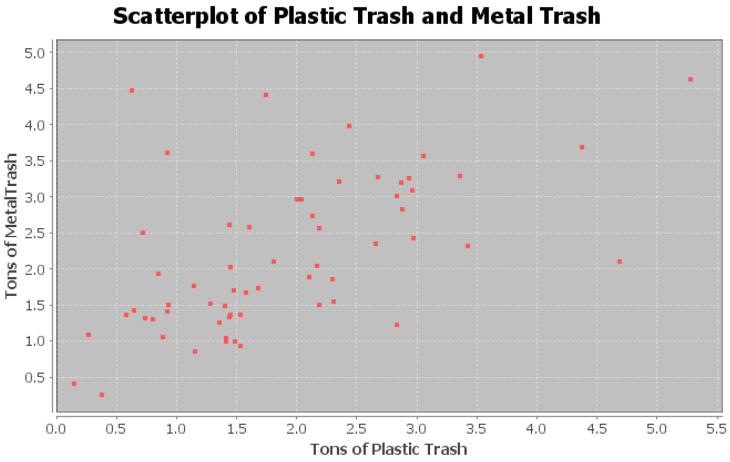
    a) What is the correlation coefficient “r”.
    b) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

(#11-17) Directions: Use the given scatterplot and correlation coefficient r to determine the strength and direction of the correlation.

11. 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) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

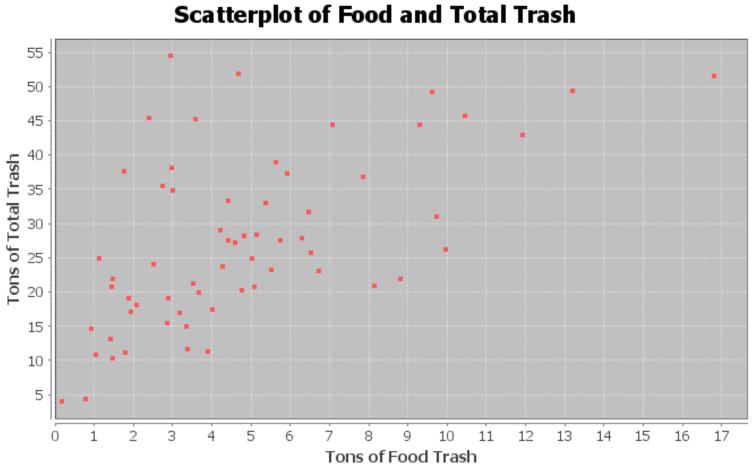


12. The x variable is describing the number of tons of plastic trash and the y variable is the number of tons of metal trash. (r = 0.5862) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

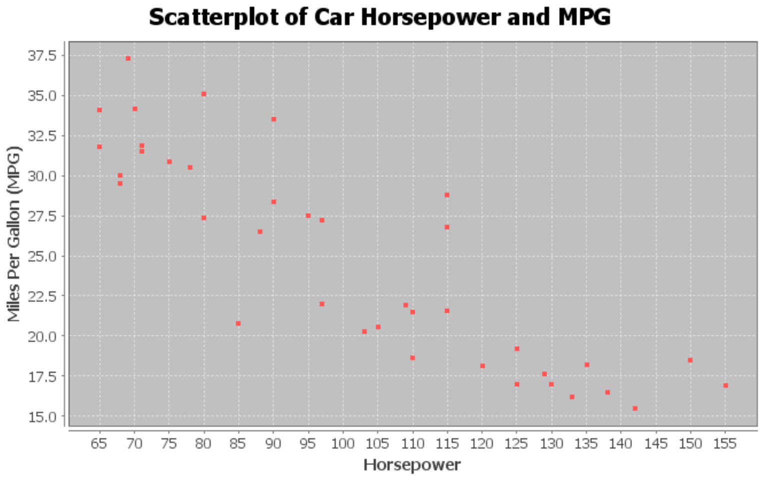


13. The x variable is describing the number of tons of food trash and the y variable is the number of tons of total trash. (r = 0.5833) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate

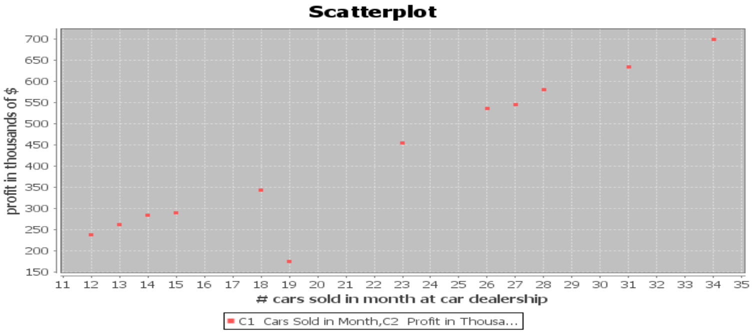
negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.


Scatterplot of Food and Total Trash

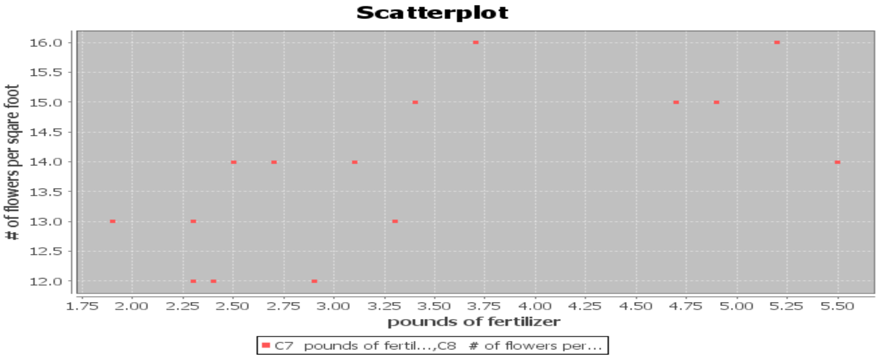
14. The x variable is describing the horsepower of an automobile and the y variable is describing the miles per gallon. (r = -0.8713) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.


Scatterplot of Car Horsepower and MPG

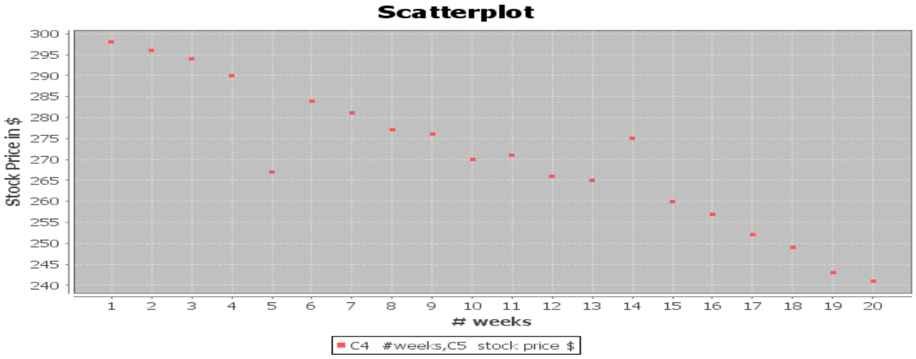
15. The x variable is the number of cars sold and the y variable is the total profit in thousands of dollars. (r = 0.9404) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.



16. 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) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.



17. The x variable is the week and the y variable is the stock price in dollars. (r = -0.9429) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.



Section 6C – 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. 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.)

The Coefficient of Determination (r^2)

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 “coefficient of determination” (r^2). The coefficient of the determination (r^2) is calculated by squaring the r – value (squaring the correlation coefficient).

Definition of the Coefficient of Determination (r^2)**:** 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

- R-squared is often given in correlation and regression printouts. StatKey however does not provide r-squared in its printout. It is easy to calculate though if you have the correlation coefficient “r”. Just push the square button on a calculator or multiply the “r” value by itself.
- R-squared is always 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 proportion that can be converted into 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.)

Looking at r and r-squared

Note: The following table is not exact. R and r-squared interpretations can differ depending on the data and how many points you have. These are just general guidelines.

	Correlation Coefficient (r)	Coefficient of Determination (r-squared)
No Correlation	≈ 0 → ± 0.19	≈ 0 → 3%
Weak Correlation	≈ ± 0.2 → ± 0.39	≈ 4% → 15%
Moderate Correlation	≈ ± 0.4 → ± 0.59	≈ 16% → 35%
Strong Correlation	≈ ± 0.6 → ± 1.0	≈ 36% → 100%

Calculating the Coefficient of Determination (r^2) with StatKey:

- To put the data into StatKey, you will want to open a fresh excel spreadsheet and paste the two quantitative data sets next to each other side by side. Then copy the two columns together.
- Now we will go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Quantitative Variables". Click on "Edit Data" at the top. Push Control A on your keyboard to highlight old data and then push "delete" on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the "Edit Data" field. If your data has a title, click the box that says "Data has header row". If your data does not have a title, do NOT check the box that says "Data has header row". Then press OK.
- You will see the correlation coefficient "r" under "Summary Statistics". Look next to "Correlation".
- Square the r-value by either multiplying "r" by itself or using the square button on a calculator.

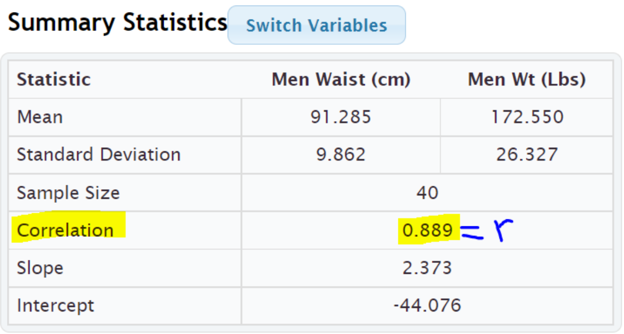
Example 1

In our last section, we used the health data to calculate the correlation coefficient "r" with StatKey. This helped us see that there was a strong positive correlation between the waist size and weight of these men. The computer calculated that r = 0.889. From this, we can calculate r-squared.

Coefficient of Determination (r-squared) = 0.889 x 0.889 = 0.790321 ≈ 79.0%.

Sentence explaining r-squared in context: 79.0% of the variability in the men's weights can be explained by the linear relationship with the men's waist size.

Interpretation: This percentage is very high indicated this is an extremely strong linear relationship. Waist size is a good explanatory variable for predicting weight. However this does not indicate that waist size causes weight to increase (correlation is not causation). There are many confounding variables involved.

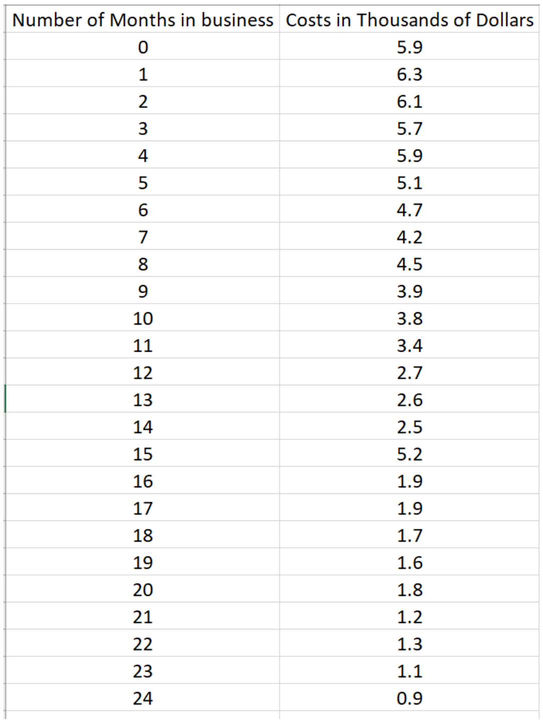


Summary Statistics

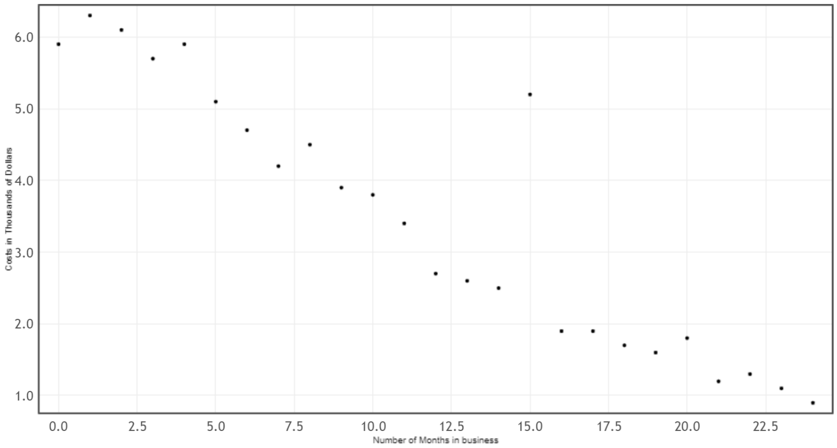
Statistic	Men Waist (cm)	Men Wt (Lbs)
Mean	91.285	172.550
Standard Deviation	9.862	26.327
Sample Size	40	
Correlation	0.889 = r	
Slope	2.373	
Intercept	-44.076	

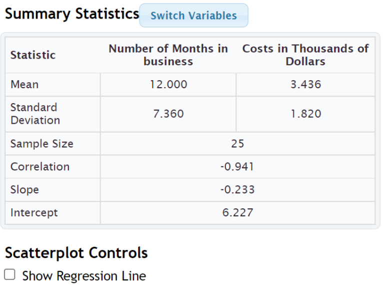
Example 2

Let’s analyze the following time and cost data for a company. Use StatKey to calculate the correlation coefficient (r) and then use “r” to calculate the coefficient of determination (r-squared).



Number of Months in business	Costs in Thousands of Dollars
0	5.9
1	6.3
2	6.1
3	5.7
4	5.9
5	5.1
6	4.7
7	4.2
8	4.5
9	3.9
10	3.8
11	3.4
12	2.7
13	2.6
14	2.5
15	5.2
16	1.9
17	1.9
18	1.7
19	1.6
20	1.8
21	1.2
22	1.3
23	1.1
24	0.9





Summary Statistics

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000	3.436
Standard Deviation	7.360	1.820
Sample Size	25	
Correlation	-0.941	
Slope	-0.233	
Intercept	6.227	

Scatterplot Controls

□ Show Regression Line

Coefficient of Determination (r-squared) = (−0.941) x (−0.941) = +0.885481 ≈ 88.5%.

Sentence explaining r-squared in context: 88.5% of the variability in costs for the company can be explained by the linear relationship with time in months.

Interpretation: This percentage is very high 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 3: 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 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.

Men's Age (years) and Weight (pounds): r-squared = 0.0815 ≈ 8.2%

Men's Height (inches) and Weight (pounds): r-squared = 0.2727 ≈ 27.3%

Men's Waist Size (cm) and Weight (pounds): r-squared = 0.7902 ≈ 79.0%

Men's Pulse (beats per minute) and Weight (pounds): r-squared = 0.0031 ≈ 0.31%

Men's Systolic Blood Pressure (mm of Hg) and Weight (pounds): r-squared = 0.1240 ≈ 12.4%

Men's Diastolic Blood Pressure (mm of Hg) and Weight (pounds): r-squared = 0.1503 ≈ 15.0%

Men's Cholesterol (mg per deciliter) and Weight (pounds): r-squared = 0.0007 ≈ 0.07%

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

Men's Leg Length (inches) and Weight (pounds): r-squared = 0.1380 ≈ 13.8%

Men's Elbow Circumference (Inches) and Weight (pounds): r-squared = 0.4034 ≈ 40.3%

Men's Wrist Circumference (Inches) and Weight (pounds): r-squared = 0.2696 ≈ 27.0%

Men's Arm Length (Inches) and Weight (pounds): r-squared = 0.6750 ≈ 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 with the men's weight.

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

Problem Set Section 6C

(#1-10) Answer the following questions to explain each of the following r-squared values. In each of these examples the weight was the response variable (Y).

1. Men's Waist Size (cm) and Weight (pounds): r-squared = 0.7902

   a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
   b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
   c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

2. Men's Pulse (beats per minute) and Weight (pounds): r-squared = 0.0031

   a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
   b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
   c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

3. Men's Systolic Blood Pressure (mm of Hg) and Weight (pounds): r-squared = 0.1240

   a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
   b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
   c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

4. Men's Diastolic Blood Pressure (mm of Hg) and Weight (pounds): r-squared = 0.1503

   a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
   b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
   c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

5. Men's Cholesterol (mg per deciliter) and Weight (pounds): r-squared = 0.0007

   a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
   b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
   c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

6. Men's Body Mass Index (BMI) (kg per m^2) and Weight (pounds): r-squared = 0.6395

   a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
   b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
   c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

7. Men’s Leg Length (inches) and Weight (pounds): r-squared = 0.1380

a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

8. Men’s Elbow Circumference (Inches) and Weight (pounds): r-squared = 0.4034

a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

9. Men’s Wrist Circumference (Inches) and Weight (pounds): r-squared = 0.2696

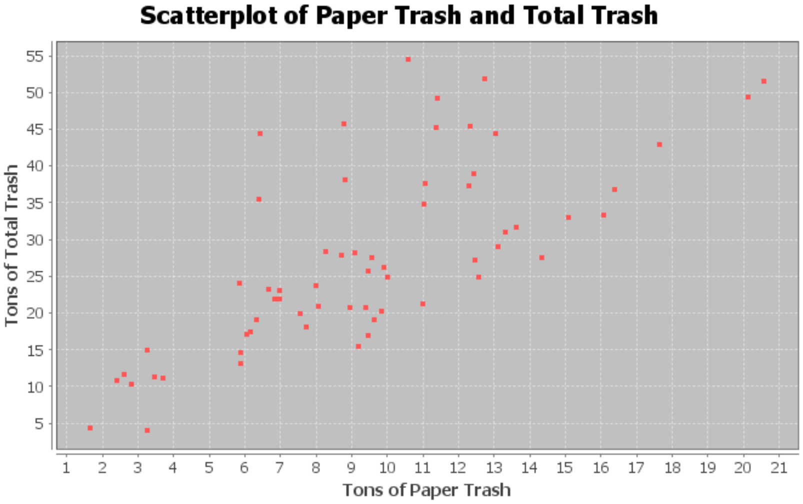
a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

10. Men’s Arm Length (Inches) and Weight (pounds): r-squared = 0.6750

a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

(#11-17) Directions: Use the given scatterplots and correlation coefficients “r” to answer the following questions for each problem.

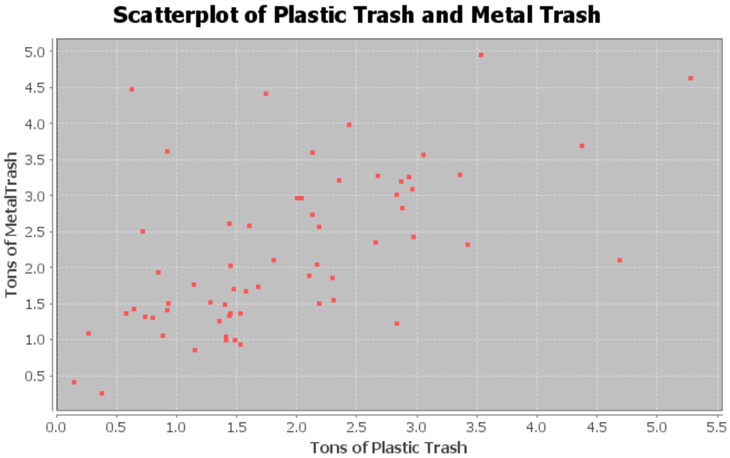
11. The x variable is describing the number of tons of paper trash and the y variable is the number of tons of total trash. (r = 0.7287)



a) Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying r x r.

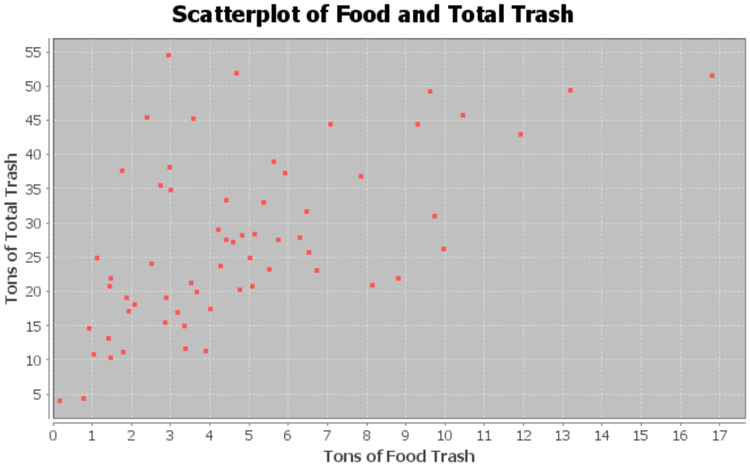
b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
d) List other possible confounding variables that may also account for the variability in y.
e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

12. 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)



a) Find the value of r-squared by squaring the correlation coefficient "r" with your calculator or by multiplying r x r.
b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
d) List other possible confounding variables that may also account for the variability in y.
e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

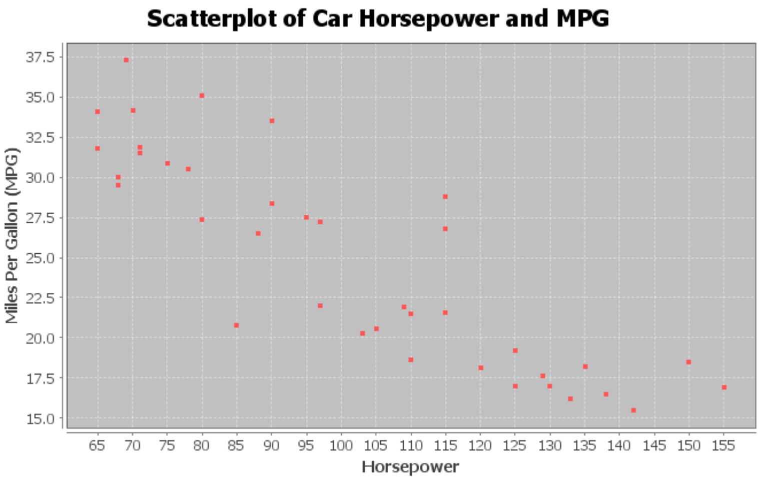
13. The x variable is describing the number of tons of food trash and the y variable is the number of tons of total trash. (r = 0.5833)



a) Find the value of r-squared by squaring the correlation coefficient "r" with your calculator or by multiplying r x r.
b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.

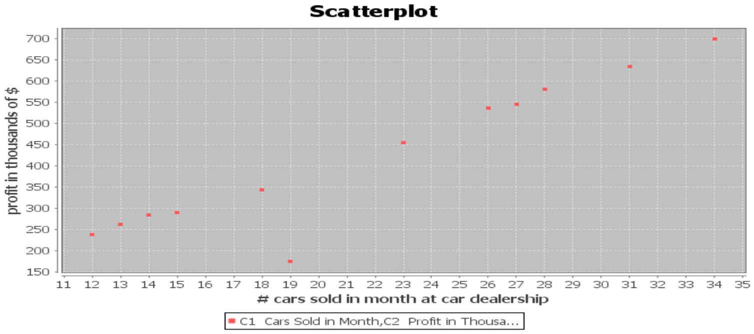
c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
d) List other possible confounding variables that may also account for the variability in y.
e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

14. The x variable is describing the horsepower of an automobile and the y variable is describing the miles per gallon. (r = -0.8713)



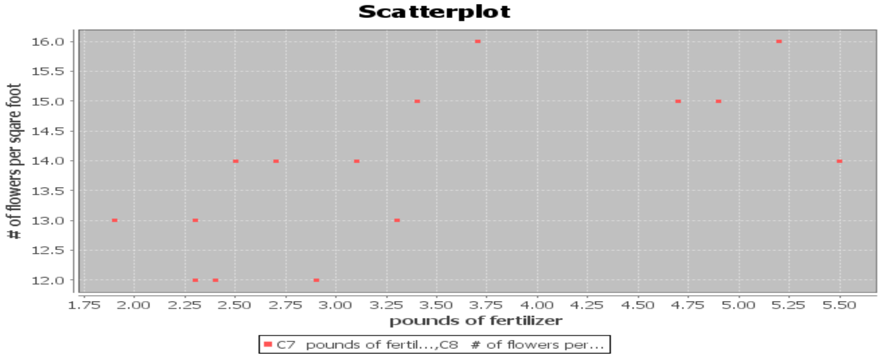
a) Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying r x r.
b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
d) List other possible confounding variables that may also account for the variability in y.
e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

15. The x variable is the number of cars sold and the y variable is the total profit in thousands of dollars. (r = 0.9404)



a) Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying r x r.
b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
d) List other possible confounding variables that may also account for the variability in y.
e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

16. 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)



a) Find the value of r-squared by squaring the correlation coefficient "r" with your calculator or by multiplying r x r.

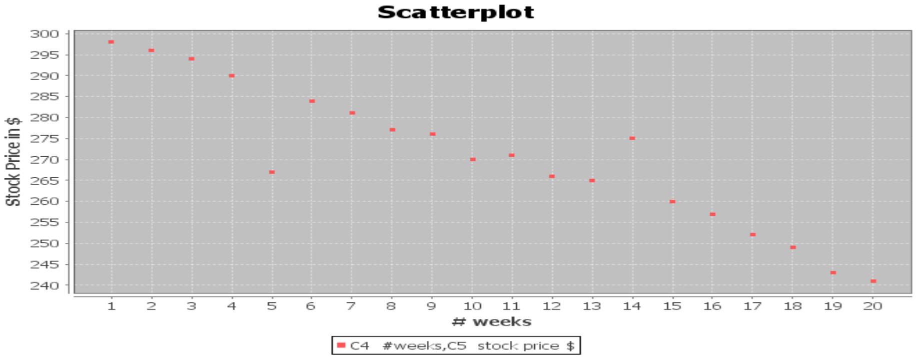
b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.

c) Write the r-squared definition sentence in context to explain the r-squared in this problem.

d) List other possible confounding variables that may also account for the variability in y.

e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

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



a) Find the value of r-squared by squaring the correlation coefficient "r" with your calculator or by multiplying r x r.

b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.

c) Write the r-squared definition sentence in context to explain the r-squared in this problem.

d) List other possible confounding variables that may also account for the variability in y.

e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

Section 6D – Best Fit Regression Line with Technology, 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.

In this section, we will discuss how to find a line that minimizes the distance between itself and thousands or millions of points in a scatterplot. As you can imagine, it is much more complicated than finding a line between two points.

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 slope and y-intercept of the regression line with StatKey:

- To put the data into StatKey, you will want to open a fresh excel spreadsheet and paste the two quantitative data sets next to each other side by side. Then copy the two columns together.
- Now we will go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Quantitative Variables". Click on "Edit Data" at the top. Push Control A on your keyboard to highlight old data and then push "delete" on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the "Edit Data" field. If your data has a title, click the box that says "Data has header row". If your data does not have a title, do NOT check the box that says "Data has header row". Then press OK.
- You will see the slope and y-intercept of the regression line under "Summary Statistics". Look next to "Slope" and "Intercept".
- To see the regression line on the scatterplot, check the box that says "Show Regression Line".

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 StatKey. We will show the process of how the line is calculated though.

To calculate the slope (b_1) and the y-intercept (b_0) of the regression line, StatKey will 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:

$\hat{y}$ = (Y-intercept) + (slope) x

$\hat{y} = b_0 + b_1$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 (b_1)

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.

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

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.

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

Calculate the Y-intercept (b_0)

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 (mean of X, 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.

$\bar{x}$: mean average of the explanatory data set (X)

$\bar{y}$: mean average of the response data set (Y)

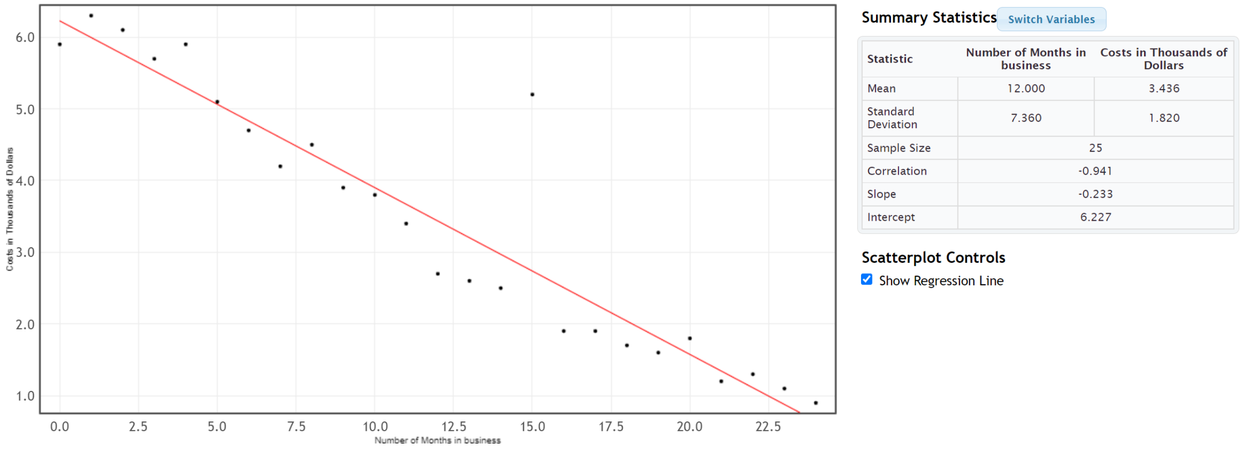
Best Fit Y-intercept = (mean of Y values) – (slope) (mean of X values)

$$\text{Best Fit Y-intercept} = \overline{\text{y}} - \textit{slope}\ (\overline{\text{x}})$$

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 using the regression line. If there is no correlation, then the regression line is not accurate. Here is the scatterplot and statistics from StatKey.



Summary Statistics

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000	3.436
Standard Deviation	7.360	1.820
Sample Size	25	
Correlation	-0.941	
Slope	-0.233	
Intercept	6.227	

Scatterplot Controls

☑ Show Regression Line

The scatter plot and correlation coefficient (r = −0.941) indicated that this data has a strong negative correlation. Notice the slope and y-intercept have already been calculated. Since there is a strong correlation, the slope and y-intercept are pretty accurate. Let's discuss how the computer calculated the slope and y-intercept.

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 slope and y-intercept for the regression line.

$\overline{x}$: mean average of the explanatory data set (X)

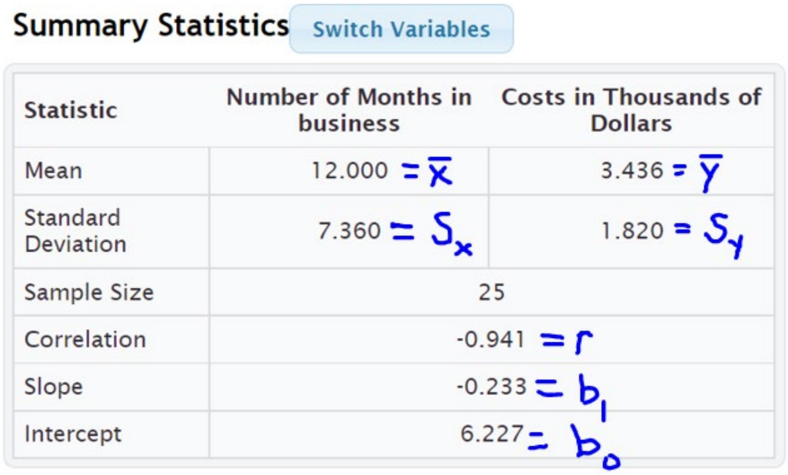
$\overline{y}$: mean average of the response data set (Y)

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

These are listed in the StatKey printout.



Summary Statistics

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000 = $\bar{x}$	3.436 = $\bar{y}$
Standard Deviation	7.360 = S_x	1.820 = S_y
Sample Size	25	
Correlation	-0.941 = r	
Slope	-0.233 = b_1	
Intercept	6.227 = b_0	

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).

Best Fit Slope (b_1) = $\frac{r \times S_y}{S_x}$ = $\frac{-0.941 \times 1.820}{7.360}$ ≈ −0.232692934 ≈ −0.233 *(Same as StatKey)*

Now that we have the slope (b_1), we can use the mean averages to calculate the best fit Y-intercept (b_0). It is better to use the

Best Fit Yintercept (b_0) = $\bar{y} - (slope \times \bar{x})$ = 3.436 – (−0.232692934 × 12.000) = 3.436 – (−2.792315217) = 3.436 + (+2.792315217) ≈ 6.228 *(Close to what StatKey gave. Computer programs will always be more accurate than hand calculations since they keep more decimal places and round less.)*

Now that we know the best-fit slope (b_1) and Y-intercept (b_0), we can write the equation of the regression line.

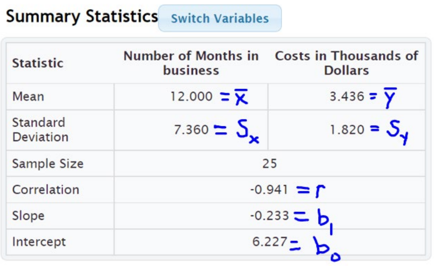
Equation of the Regression Line:

$\hat{y}$ = (Y-intercept) + (slope) x

$\hat{y}$ = 6.227 + (−0.233) x

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 for the time/cost data. This information may be very important for the company.



Summary Statistics Switch Variables

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000 = $\bar{x}$	3.436 = $\bar{y}$
Standard Deviation	7.360 = S_x	1.820 = S_y
Sample Size	25	
Correlation	-0.941 = r	
Slope	-0.233 = b_1	
Intercept	6.227 = b_0	

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.

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 the X variable.

The slope was −0.233. 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.

Slope = −0.233 = $\frac{-0.233\ thousand\ dollars}{+1\ month}$

First, recognize what the slope is <u>not</u> saying. The slope is <u>not</u> 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.

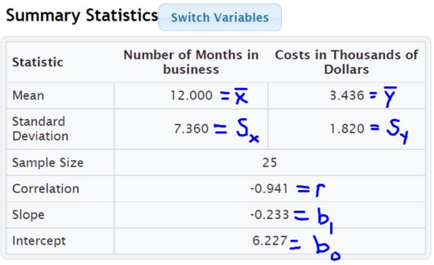
Since this slope was negative, it is a decrease. The slope is telling us that monthly costs are decreasing about 0.233 thousand dollars per month on average. Looking at the units, we can also explain it this way: Monthly costs are decreasing about $233 per month on average. *(Notice we did not say that the costs were decreasing −0.233 or decreasing −$233. The negative is described by the word "decrease".)*

Sentence Explaining the Slope in Context: Monthly costs for this company are decreasing about $233 per month on average.

Interpretation of Y-intercept

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 = 6.227.

Definition of Y-intercept of the Regression Line: The predicted Y-value when X is zero.



Summary Statistics Switch Variables

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000 = $\bar{x}$	3.436 = $\bar{y}$
Standard Deviation	7.360 = S_x	1.820 = S_y
Sample Size	25	
Correlation	-0.941 = r	
Slope	-0.233 = b_1	
Intercept	6.227 = b_0	

A Y-intercept is the predicted Y value when X is zero. Do not forget to include the units. Therefore, the Y-intercept of 6.227 really represents the ordered pair (0 months, 6.227 thousand dollars).

Sentence Explaining the Y-intercept in Context: At the start of the company (month 0), we predict there was an average initial cost of approximately 6.227 thousand dollars (or $6,227).

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. When zero is not in the scope of the X values, 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.*

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 6D

(#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.

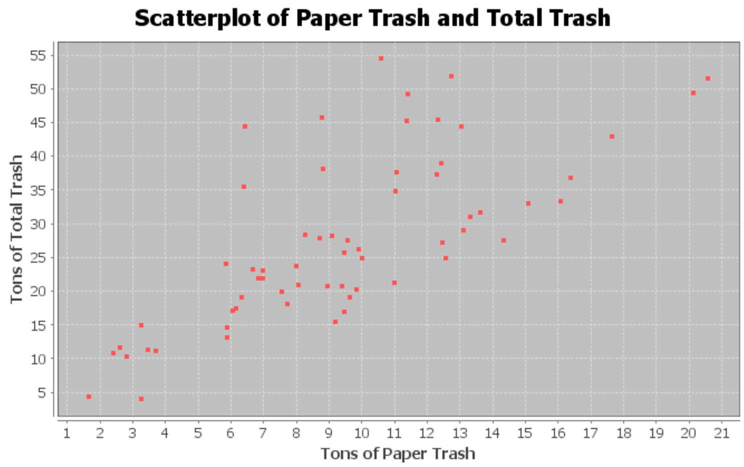
- Slope of the Regression Line = (multiply correlation coefficient r times the standard deviation of Y values) ÷ standard deviation of X values

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

  $\hat{y}$ = *(Y-intercept)* + *(slope)* x

1.



r = 0.7287

	Mean	StDev
(x)Paper Trash	9.428	4.168
(y)Total Trash	27.44	12.46

Slope = ___________________

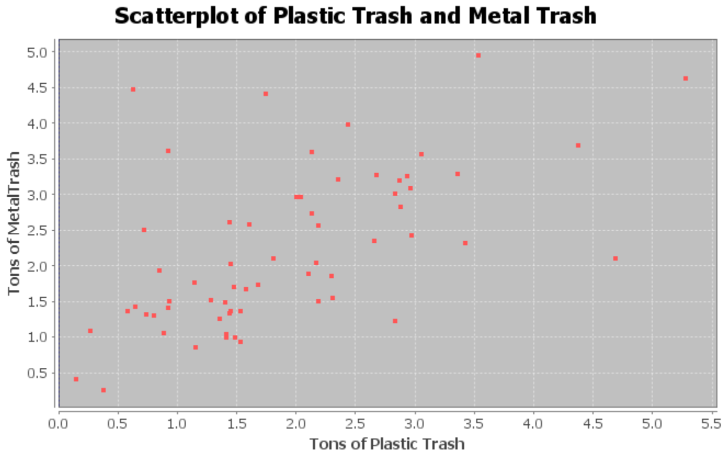
Sentence Explaining the Slope:

Y – Intercept = _________________

Sentence Explaining the Y-intercept:

Equation of the Regression Line: Y = __________ + __________ X

2.



r = 0.5862

	Mean	StDev
(x)Plastic Trash	1.911	1.065
(y) Metal Trash	2.218	1.091

Slope = ___________________

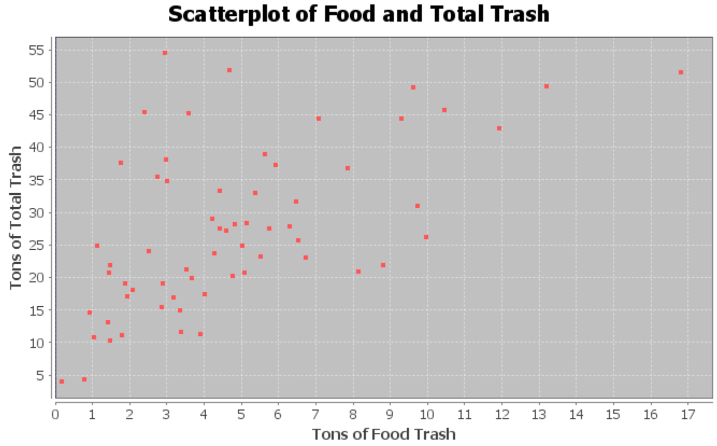
Sentence Explaining the Slope:

Y – Intercept = _________________

Sentence Explaining the Y-intercept:

Equation of the Regression Line: Y = __________ + __________ X

3.



r = 0.5833

	Mean	StDev
(x)Food Trash	4.816	3.297
(y)Total Trash	27.44	12.46

Slope = ___________________

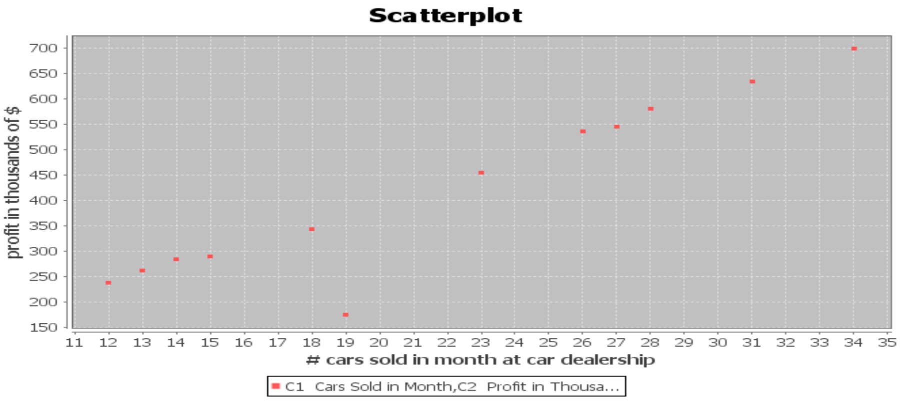
Sentence Explaining the Slope:

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

	Mean	Standard Deviation
(x) Cars Sold in Month	21.667	7.512
(y) Profit in Thousands of Dollars	420.25	175.615

Slope = ___________________

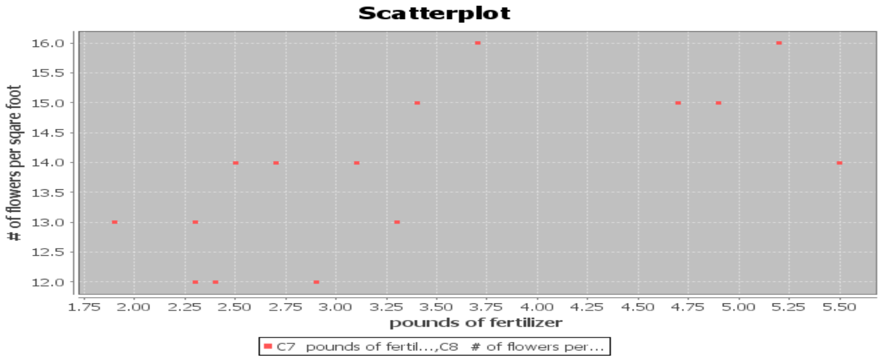
Sentence Explaining the Slope:

Y – Intercept = _________________

Sentence Explaining the Y-intercept:

Equation of the Regression Line: Y = __________ + __________ X

5. This data describes the relationship between the number of pounds of fertilizer used and the number of flowers per square foot.



r = 0.6727

Variable	Mean	Standard Deviation
(x) pounds of fertilizer used	3.387	1.165
(y) # of flowers per sq. ft.	13.867	1.356

Slope = ___________________

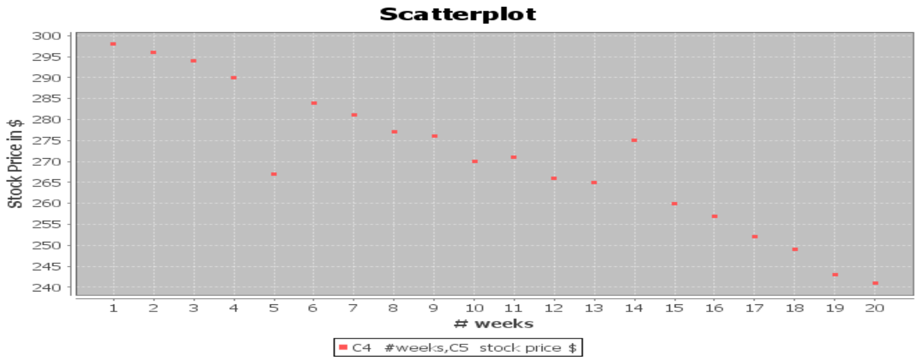
Sentence Explaining the Slope:

Y – 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

Variable	Mean	Standard Deviation
(x) #weeks	10.5	5.916
(y) stock price $	270.6	17.031

Slope = ___________________

Sentence Explaining the Slope:

Y – Intercept = _________________

Sentence Explaining the Y-intercept:

Equation of the Regression Line: Y = __________ + __________ X

(#7-12) Directions: For the following problems, use the indicated data and StatKey to calculate the correlation coefficient “r” and the slope and y-intercept of the regression line. Then answer the questions.

Calculating the slope and y-intercept of the regression line with StatKey:

- To put the data into StatKey, you will want to open a fresh excel spreadsheet and paste the two quantitative data sets next to each other side by side. Then copy the two columns together.
- Now we will go to www.lock5stat.com and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Quantitative Variables”. Click on “Edit Data” at the top. Push Control A on your keyboard to highlight old data and then push “delete” on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the “Edit Data” field. If your data has a title, click the box that says “Data has header row”. If your data does not have a title, do NOT check the box that says “Data has header row”. Then press OK.
- You will see the slope and y-intercept of the regression line under “Summary Statistics”. Look next to “Slope” and “Intercept”.
- The correlation coefficient “r” will be listed under “Summary Statistics”. Look under “Correlation”.
- To see the regression line on the scatterplot, check the box that says “Show Regression Line”.

7. Open the “Cigarette Data” from Canvas or from www.matt-teachout.org. 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. You do NOT need to save or copy the scatterplot. Give the r-value and 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) What is the equation of the regression line?

8. Open the "Cigarette Data" from Canvas or from www.matt-teachout.org. 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. You do NOT need to save or copy the scatterplot. Give the r-value and 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) What is the equation of the regression line?

9. Open the "Health Data" from Canvas or from www.matt-teachout.org. 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 body mass index (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. You do NOT need to save or copy the scatterplot. Give the r-value and 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) What is the equation of the regression line?

10. Open the "Health Data" from Canvas or from www.matt-teachout.org. 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. You do NOT need to save or copy the scatterplot. Give the r-value and 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) What is the equation of the regression line?

11. Open the "Bear Data" from Canvas or from www.matt-teachout.org. 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. You do NOT need to save or copy the scatterplot. Give the r-value and 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) What is the equation of the regression line?

12. Open the "Bear Data" from Canvas or from www.matt-teachout.org. 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. You do NOT need to save or copy the scatterplot. Give the r-value and 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) What is the equation of the regression line?

Section 6E – Residuals, Standard Deviation of the Residual Errors (S_e), Residual Plots and Histogram of the Residuals

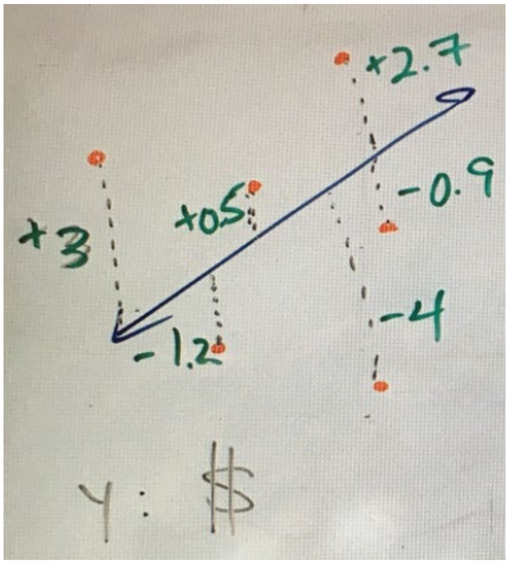
Note about section 6E: StatKey does not calculate residual plots, histogram of the residuals, or the standard deviation of the residual errors. We will focus on interpreting residuals in this section. Students will not be asked to calculate them with computer software.

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 be 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 be above the actual point meaning that the prediction for that particular x value is too high.



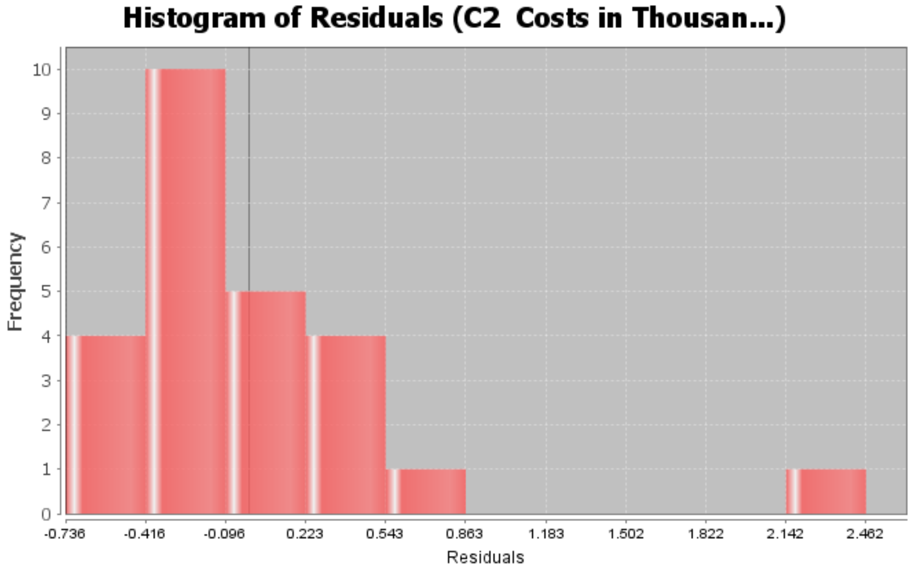
The picture above shows the idea of residuals. In this example, the response variable was in dollars so all of the residual errors are in dollars. A point that has a residual of +3 means the point was 3 dollars above the regression line. A point that has a residual of -4 means that the point was 4 dollars 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".

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.

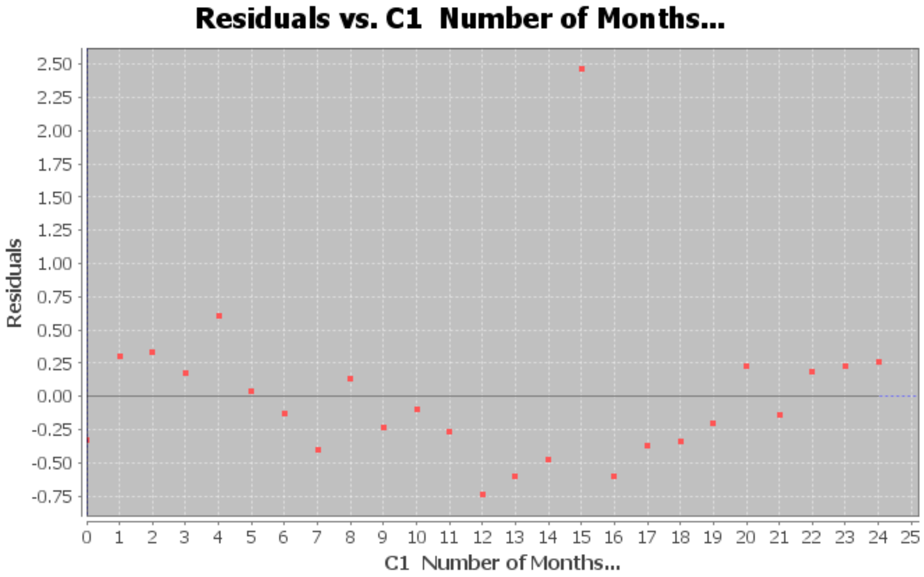
Interpretation of the histogram of the residuals: In the histogram shown above describing the residuals for the cost data, we see that the graph is not very bell shaped. In fact, it is skewed right. We also see that the zero line is a little off from the highest bar. This tells us the graph is not centered 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.

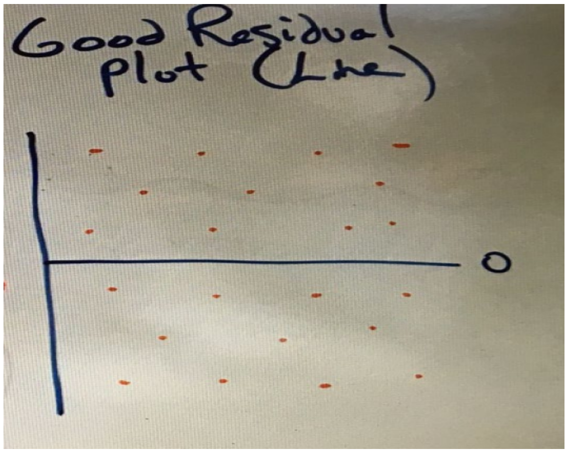


Let us see if we can understand what we are looking at. The horizontal line at zero represents the regression line itself. Notice the vertical scale on the left is no longer the same as the scatterplot. It has positive numbers above zero and negative numbers below zero. The units for the vertical access 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.

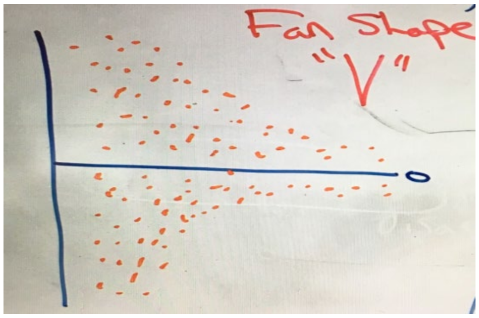
Interpreting a Residual Plot

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.

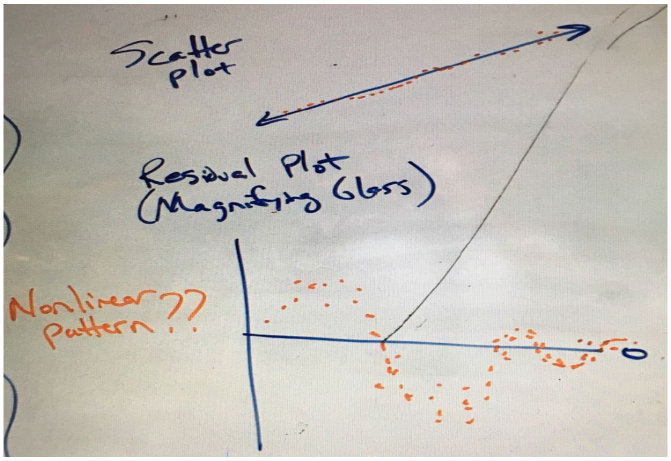
Good Residual Plot (Evenly Spread Out)



Bad Residual Plot ("V" Shape, Not 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.

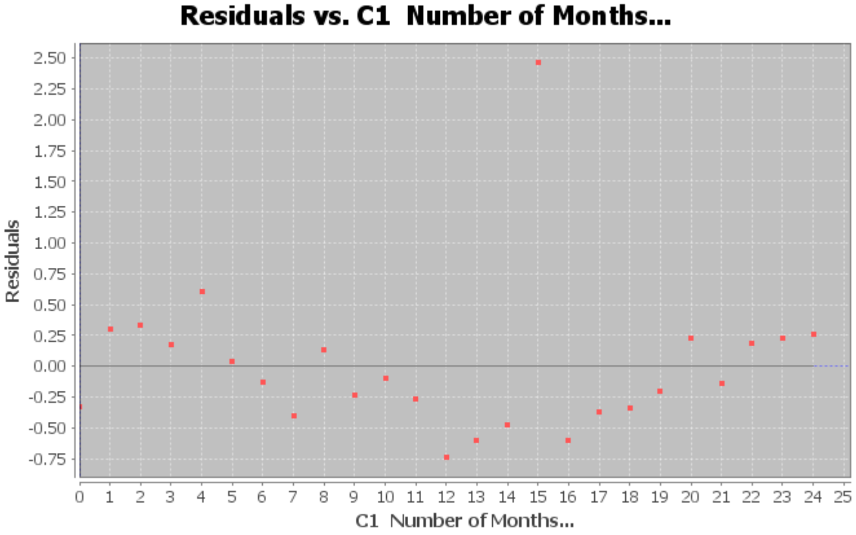


Note about terminology: *The term "nonlinear" can refer to a curved pattern in the data, but can be misleading. Statisticians include many curved patterns under the heading of "linear regression" since they focus on the study of transformations from curves to lines.*

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.

Interpretation of the residual plot: 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.

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 (normal) 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”.

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

Standard Deviation of the Residual Errors (S_e) = 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

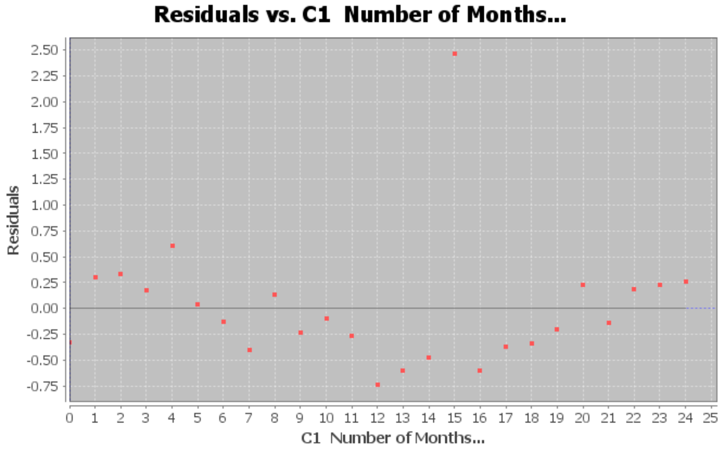
- *Standard deviation is a measure of spread for bell shaped (normal) data sets. If the histogram of the residuals is not bell shaped, then the standard deviation is not as accurate.*
- *Do not confuse the standard deviation of the X values, the standard deviation of the Y values and the standard deviation 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". This can be confusing, because the term "standard error" is used to describe the standard deviation of a sampling distribution.*

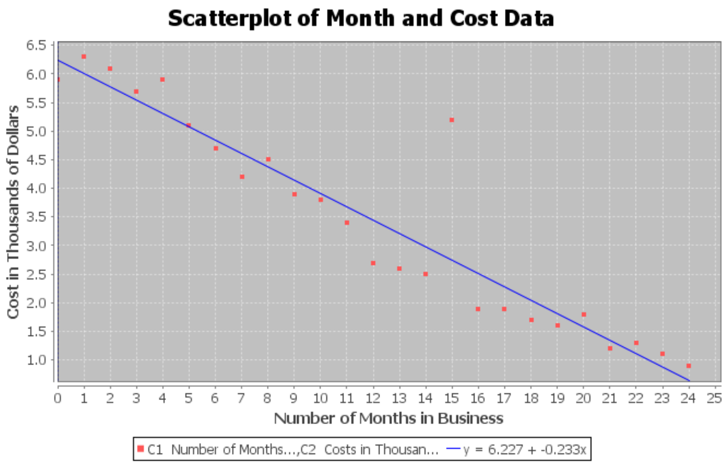
Example 1 continued

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.

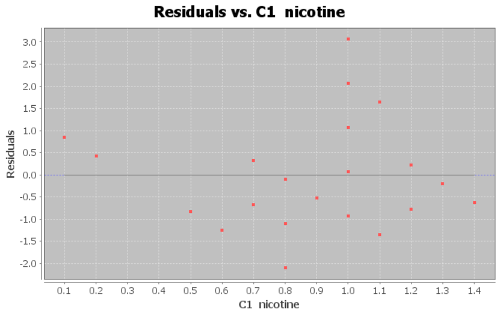
Remember that the accuracy of the regression line and the standard deviation of the residuals is tied to the residual plot being bell shaped (normal). In this example, the histogram of the residuals was skewed right. This tells us that the regression line and standard deviation will not be quite as accurate.

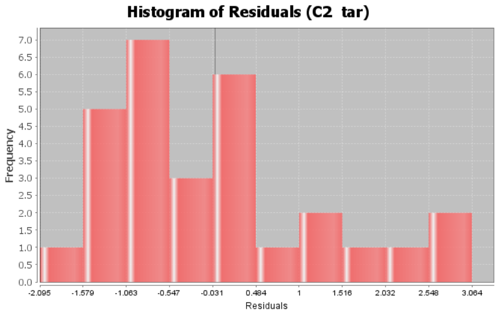
Problem Set Section 6E

Note about the problems in section 6E: StatKey does not calculate residual plots, histogram of the residuals, or the standard deviation of the residual errors. We will focus on interpreting residuals in this section. Students will not be asked to calculate them with computer software.

1. Mg of nicotine is the explanatory variable and mg of tar is the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors (S_e) to answer the following questions.

S_e = 1.2984 mg of tar





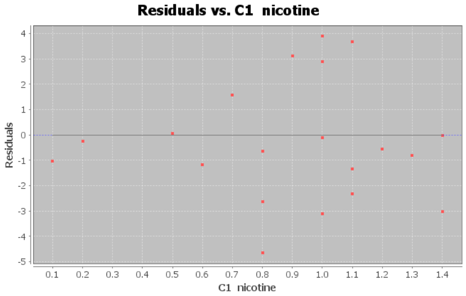
a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

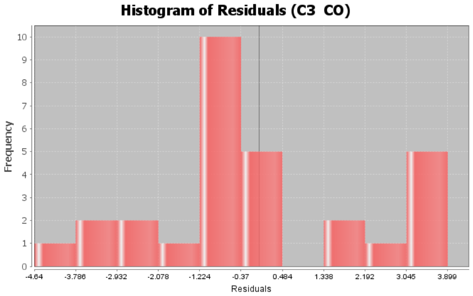
b) Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).

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

2. Mg of nicotine is the explanatory variable and mg of carbon monoxide (CO) is the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors (S_e) to answer the following questions.

S_e = 2.2961 PPM





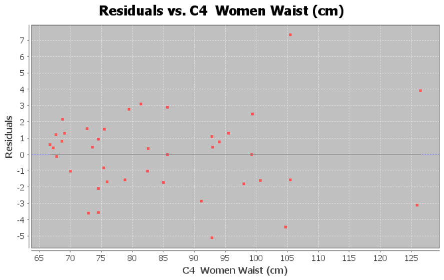
a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

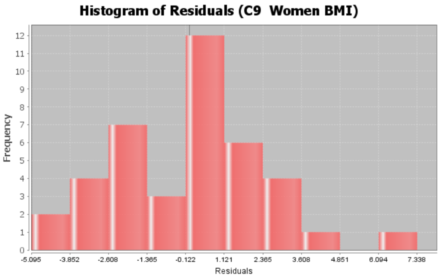
b) Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).

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

3. Women's waist size is the explanatory variable and women's body mass index (BMI) is the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors (S_e) to answer the following questions.

S_e = 2.4761 kg/m^2


Residuals vs. C4 Women Waist (cm)


Histogram of Residuals (C9 Women BMI)

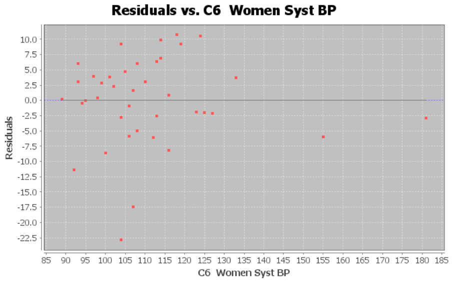
a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

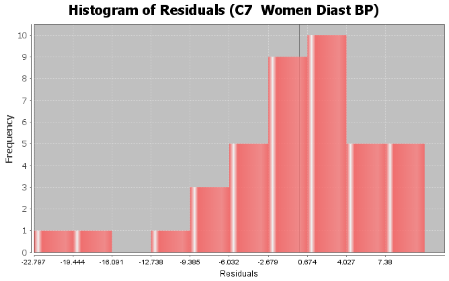
b) Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).

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

4. Women's systolic blood pressure is the explanatory variable and women's diastolic blood pressure is the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors (S_e) to answer the following questions.

S_e = 7.2912 mm of Hg


Residuals vs. C6 Women Syst BP


Histogram of Residuals (C7 Women Diast BP)

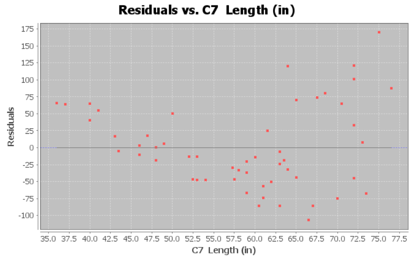
a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

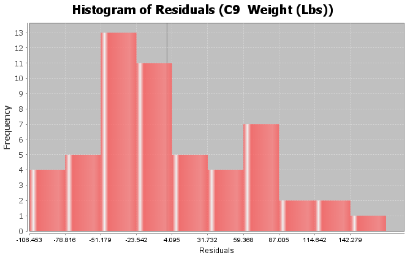
b) Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).

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

5. Bear length in inches is the explanatory variable and bear weight in pounds the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors (S_e) to answer the following questions.

S_e = 61.8272 pounds


Residuals vs. C7 Length (in)


Histogram of Residuals (C9 Weight (Lbs))

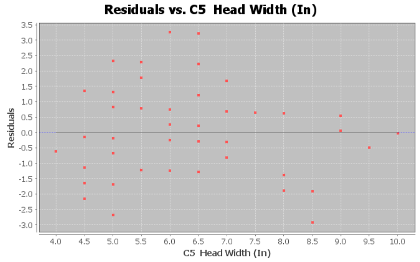
a) Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?

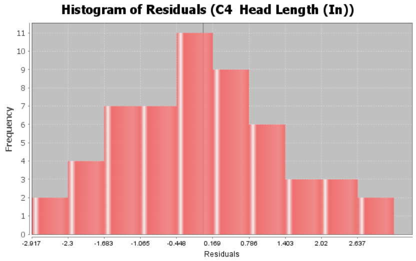
b) Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).

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

6. Bear head width in inches is the explanatory variable and bear head length in inches is the response variable. Use the given residual plot verses the x-variable, histogram of the residuals, and the standard deviation of the residual errors (S_e) to answer the following questions.

S_e = 1.4231 inches


Residuals vs. C5 Head Width (In)


Histogram of Residuals (C4 Head Length (In))

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 6F – Predictions, Scope of the X-values, Extrapolation, and Using the Standard Deviation of the Residual Errors

Note about section 6F: StatKey can calculate the scatterplot, the correlation coefficient, and the slope and y-intercept of the regression line. We expect students to be able to use StatKey to calculate these. However, StatKey does <u>not</u> calculate the standard deviation of the residual errors. We will focus on interpreting the standard deviation of the residual errors and students will <u>not</u> be asked to calculate it.

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.

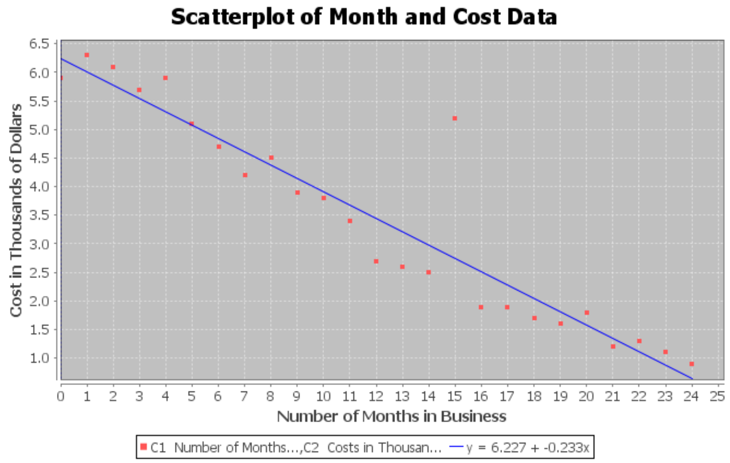
Notes about Making Predictions with the Regression Line

1. <u>There should be some correlation between the variables.</u>

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 with the scatterplot and the correlation coefficient "r" <u>before</u> using the regression line to make a prediction.

2. <u>The scope of the X values and "Extrapolation"</u>

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.



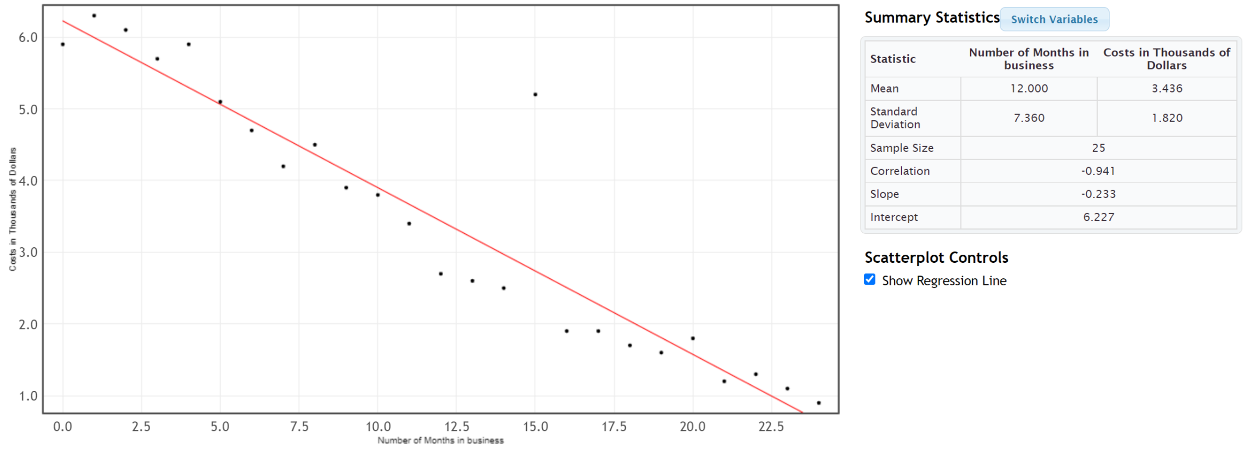
Notice that the X-axis of the scatterplot represents the number of months the company has been in business. The X values go from 0 to 24 months. This is called the "scope of the X values" or the "scope of the data". Recall that this regression line fits the data really well and had a strong linear relationship 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.

For example, let us look at scatterplot and statistics for the time (months) and costs (thousands of dollars) for a company.



Summary Statistics

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000	3.436
Standard Deviation	7.360	1.820
Sample Size	25	
Correlation	-0.941	
Slope	-0.233	
Intercept	6.227	

Scatterplot Controls

☑ Show Regression Line

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.

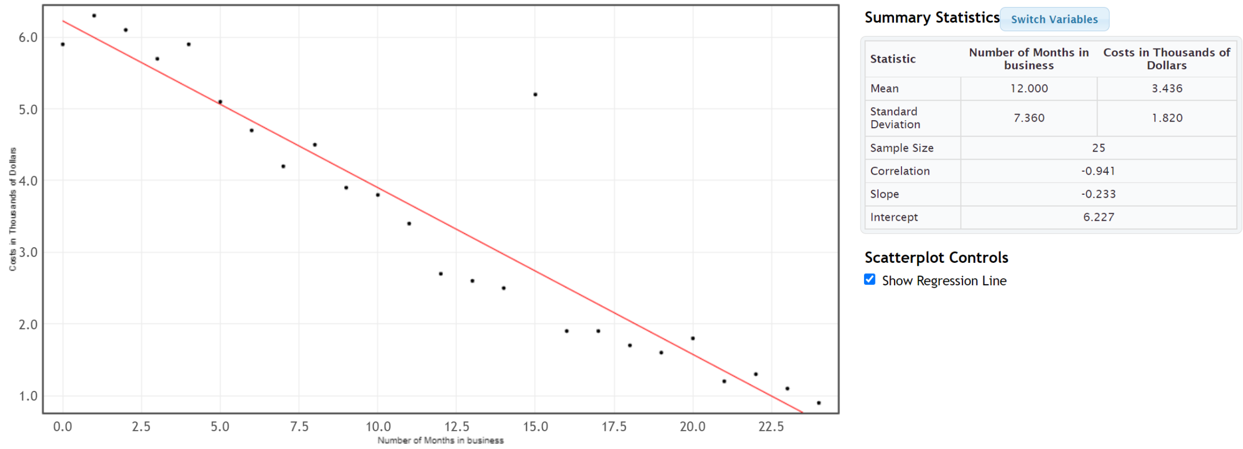
3. <u>Making the Prediction</u>

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, StatKey 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.

Example 1

Let us look at scatterplot and statistics for the the time (months) and costs (thousands of dollars) for a company. 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.



Summary Statistics

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000	3.436
Standard Deviation	7.360	1.820
Sample Size	25	
Correlation	-0.941	
Slope	-0.233	
Intercept	6.227	

Scatterplot Controls

☑ Show Regression Line

First notice that the regression line does fit the data in scatterplot and the correlation coefficient r is close to −1. This means there is a strong (negative) correlation and the regression line formula is likely to be more accurate. First we need to plug in the slope and y-intercept into the regression line formula.

$\hat{y}$ = Y-intercept + (Slope) x

$\hat{y} = 6.227 + (-0.233)\ x$

This formula can now be used to make predictions about y. The symbol $\hat{y}$ means "predicted y value". When we plug in a number for x and find $\hat{y}$, we are making a prediction based on data.

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.

$\hat{y} = 6.227 + (-0.233)\ x$

$\hat{y} = 6.227 + (-0.233)\ (15)$

$\hat{y} = 6.227 + (-3.495)$

$\hat{y} = +2.732$ thousand dollars

Therefore, if the company had not had to replace those parts, we predict their average costs would have been about 2.732 thousand dollars ($2,732) in their 15th month in business.

How much error could there be in that prediction? 0.6295 thousand dollars ($629.50)

Remember, the standard deviation of the residual errors tells us how much prediction error we have.

Standard Deviation of the Residual Errors = 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).

So we predict that in month 15, the companies costs would have been about $2732. This prediction could be off by about $629.50 on average.

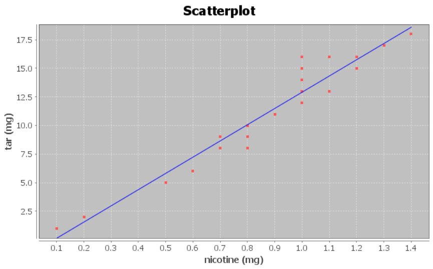
Example 2
Can we use the month and cost regression line from Example 1 to predict the costs in month 50?

No. This is an extrapolation. Our prediction may be very off. Also, our standard deviation of the residual errors would not be accurate. We are out of the scope of the data.

Problem Set Section 6F

Directions: You will not need to use StatKey for these problems. Graphs and statistics were already calculated for you. Use the given scatterplot, correlation coefficient (r), standard deviation of residual errors, and the equation of the regression line to answer the following questions.

1. 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.



Correlation Coefficient r = 0.9614

Regression Line Equation: Y = −1.2713 + 14.2076 X

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?

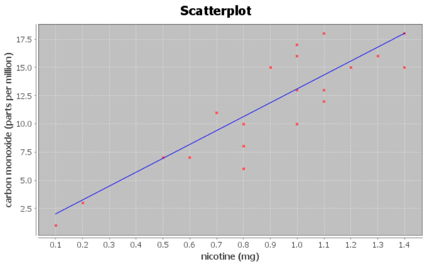
c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

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. Explore the relationship between mg of nicotine and carbon monoxide (CO) in part per million (ppm) 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 + 12.3057 X

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?

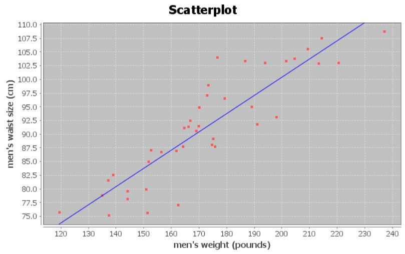
c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

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. 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 + 0.3330 X

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?

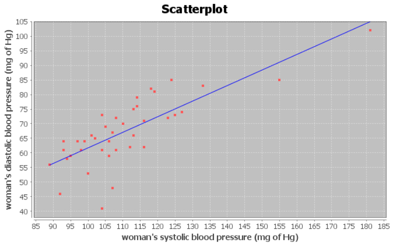
c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

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 confounding variables that might influence waist size other than weight?

4. Explore the relationship between a woman's systolic blood pressure (mm of Hg) and her diastolic blood pressure (mm of Hg). 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 + 0.5335 X

s_e = 7.2912 mm of Hg

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

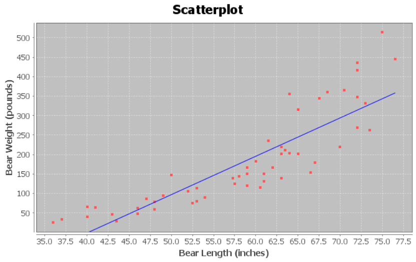
b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?

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. 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 + 9.8390 X

s_e = 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?

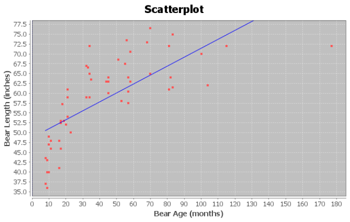
c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?

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. 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 + 0.2281 X

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?

Chapter 6 Review Sheet

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.

Problems for Chapter 6 Review Sheet

1. Define each of the following.

   a) Explanatory variable
   b) Response variable
   c) Correlation Coefficient "r"
   d) r-squared
   e) slope
   f) y-intercept
   g) residual
   h) standard deviation of the residual errors.

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.

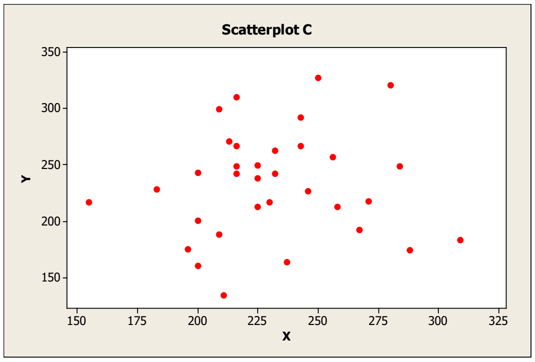
a) Which variable should be the explanatory variable (x)?
(Number of migraines or hours of sleep)

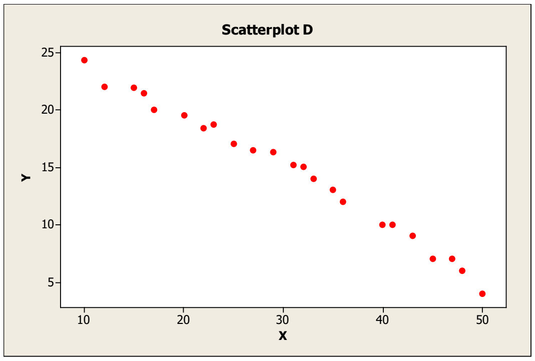
b) Which variable should be the response variable (y)?
(Number of migraines or hours of sleep)

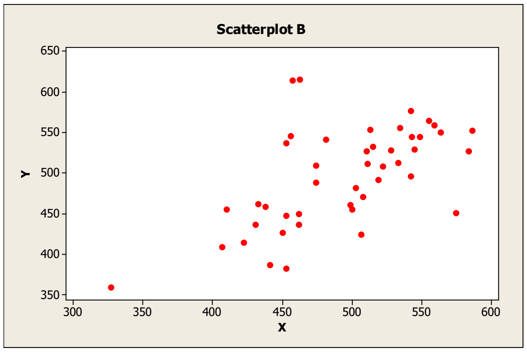
c) Sanvi is hoping to prove that lack of sleep causes migraines. If the data showed a strong correlation between migraines and sleep, would this prove that lack of sleep causes people to get migraines? Why or why not?

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, r = -0.993 , 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.

$$\left(r = 0.819\right)$$

	Mean	Standard Deviation
(x) Explanatory Variable	$\overline{x} = 19.18$	$s_x = 3.83$
(y) Response Variable	$\overline{y} = 82.55$	$s_y = 11.64$

a)

$$\text{slope}: \ m = \frac{r \cdot s_y}{s_x}$$

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

Slope = ______________________

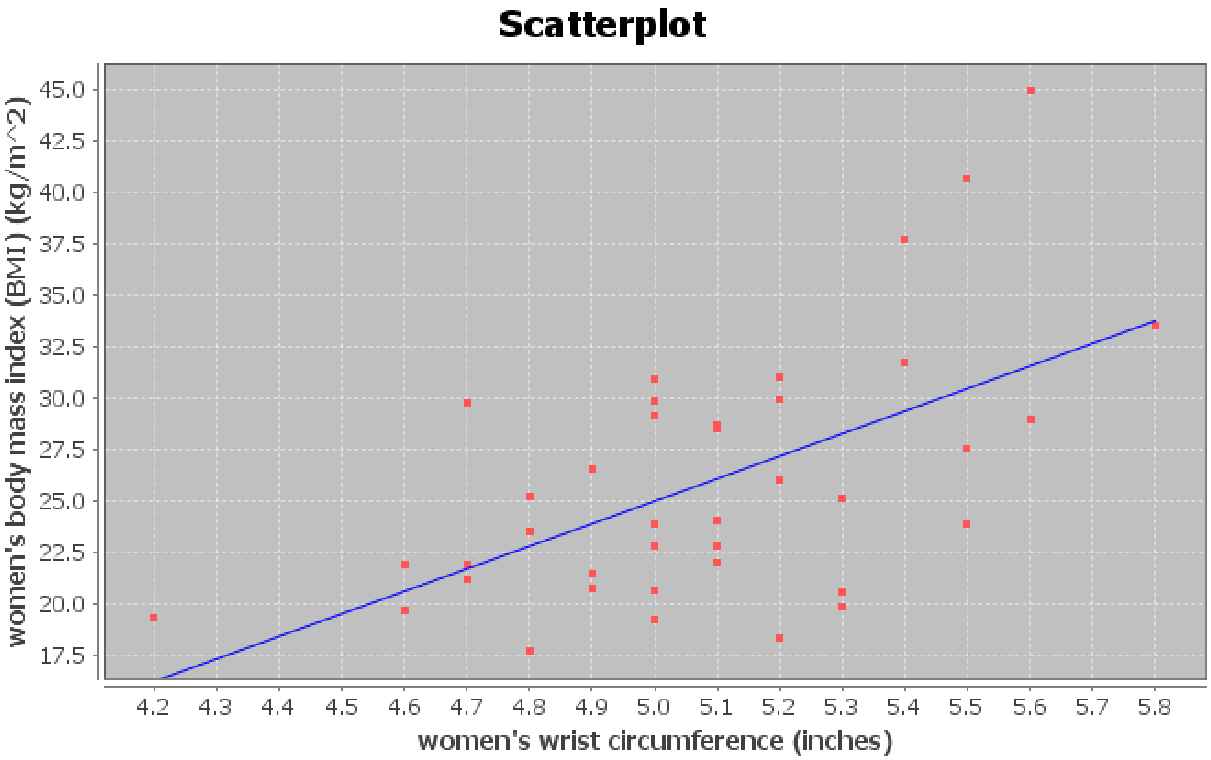
b)

$$\text{y-intercept}: \ b = \overline{y} - m(\overline{x})$$

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

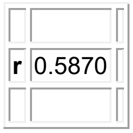
Y – Intercept = ______________________

(#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.

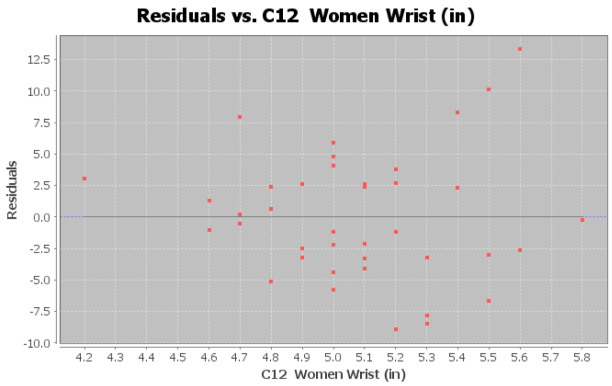
11. List three confounding variables that might influence the body mass index of a woman other than size of her wrist.

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.

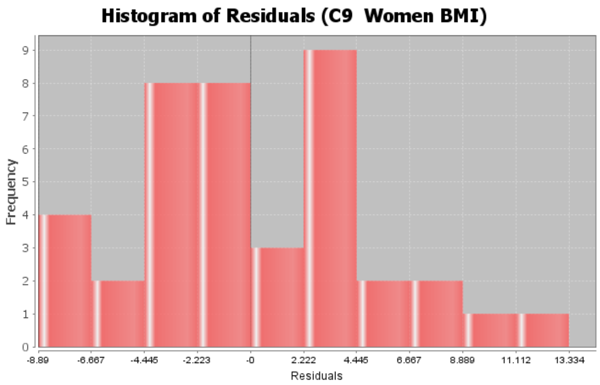
13. What was the standard deviation of the residual errors? What units does the standard deviation of the residuals have in this problem?

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



Residuals vs. C12 Women Wrist (in)



Histogram of Residuals (C9 Women BMI)

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 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?

Chapter 6 Project
Linear Quantitative Relationships

Online Class Directions: *This will be an individual project. Each student is required to analyze a pair of quantitative data sets from the following topic list. You can find your two columns of data in the "Math 075 Project Data Correlation Regression". You can find this data in Canvas or at www.matt-teachout.org. Put your two columns of data into StatKey to calculate, create a poster summarizing their findings, and email a picture of your poster to your instructor.*

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) and the units for x and y.**
- **Why this topic was important or interesting to your group.**
- **Go to www.lock5stat.com. Click on "Two quantitative variables" under the "descriptive statistics and graphs" menu. Click on "edit data" button in StatKey. Copy and paste the two columns together into StatKey. If your two columns have titles, click the box that says "data has a header row". Push OK. Click the box that says "Show Regression Line".**
- **Look at the scatterplot carefully. Make sure the x variable you picked is on the horizontal axis and the response variable (y) is on the vertical axis. If they are not push the "Switch Variables" button.**
- **Draw a rough sketch of the scatterplot on your poster. Label your axes and draw the regression line also.**
- **What is the correlation coefficient (r) for your data? You will find this where it says "Correlation" under "Summary Statistics".**
- **Use the r-value to classify the linear relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no correlation.**
- **Square your r-value tor multiply r x r in order to calculate the coefficient of determination (r-squared). Put r-squared on your poster.**
- **Write a sentence describing its meaning of r-square in context.**
- **What is the slope of the regression line? You will find this where it says "Slope" under "Summary Statistics" in StatKey. Put the slope on your poster.**
- **Write a sentence describing the meaning of the slope in context.**
- **What is the Y-intercept of the regression line? You will find this where it says "Intercept" under "Summary Statistics" in StatKey. Put the Y-intercept on your poster.**
- **Write a sentence describing the meaning of the y-intercept in context.**
- **Put in the slope and Y-intercept into the regression line equation $\hat{y} = (\boldsymbol{Y - intercept}) + (\boldsymbol{Slope})\boldsymbol{X}$. Put the regression equation on your poster.**
- **Find the scope of the x-values? (Estimate two numbers on the X-axis that the points on the scatterplot are in between.)**
- **Pick any x-value in the scope. Plug in that x value into the regression line equation and predict the y value.**
- **Decorate your poster to spark interest.**
- **Now take a picture of your poster project and submit the picture to your instructor in Canvas.**
- **After submitting the picture of the poster to your instructor, go to the discussion menu in Canvas and complete the "Chapter 6 Project Discussion". You will be discussing your findings with other students in the class.**

Face to face Class Directions: *The class will be separated into groups. Each group is required to pick a "team name" for their group and analyze a pair of quantitative data sets from the following topic list. You can find your two columns of data in the "Math 075 Project Data Correlation Regression". You can find this data in Canvas or at www.matt-teachout.org. Put your two columns of data into StatKey to calculate, create a poster summarizing their findings, and present the poster to 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) and the units for x and y.**
- **Why this topic was important or interesting to your group.**
- **Go to www.lock5stat.com. Click on "Two quantitative variables" under the "descriptive statistics and graphs" menu. Click on "edit data" button in StatKey. Copy and paste the two columns together into StatKey. If your two columns have titles, click the box that says "data has a header row". Push OK. Click the box that says "Show Regression Line".**
- **Look at the scatterplot carefully. Make sure the x variable you picked is on the horizontal axis and the response variable (y) is on the vertical axis. If they are not push the "Switch Variables" button.**
- **Draw a rough sketch of the scatterplot on your poster. Label your axes and draw the regression line also.**
- **What is the correlation coefficient (r) for your data? You will find this where it says "Correlation" under "Summary Statistics".**
- **Use the r-value to classify the linear relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no correlation.**
- **Square your r-value tor multiply r x r in order to calculate the coefficient of determination (r-squared). Put r-squared on your poster.**
- **Write a sentence describing its meaning of r-square in context.**
- **What is the slope of the regression line? You will find this where it says "Slope" under "Summary Statistics" in StatKey. Put the slope on your poster.**
- **Write a sentence describing the meaning of the slope in context.**
- **What is the Y-intercept of the regression line? You will find this where it says "Intercept" under "Summary Statistics" in StatKey. Put the Y-intercept on your poster.**
- **Write a sentence describing the meaning of the y-intercept in context.**
- **Put in the slope and Y-intercept into the regression line equation $\hat{y} = (Y - intercept) + (Slope)X$. Put the regression equation on your poster.**
- **Find the scope of the x-values? (Estimate two numbers on the X-axis that the points on the scatterplot are in between.)**
- **Pick any x-value in the scope. Plug in that x value into the regression line equation and predict the y value.**
- **Decorate your poster to spark interest.**

Presentation
Make sure each person on the team understands the poster and can present your findings. Bring your poster to a designated presentation area in the classroom and hang or tape your poster to a wall. One person at a time will present the poster. We will then rotate so that each member of the team gets to present. Everyone else will listen to presentations and give feedback.
