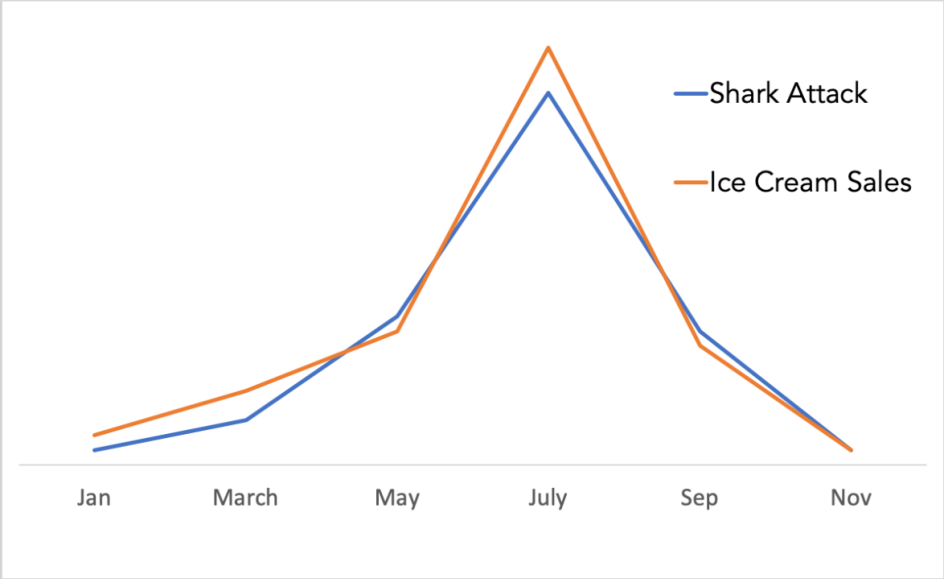
Unit 2: Exploring Two-Variable Data

Introduction

In the first unit, we focused on exploring one-variable (univariate) data, and we concluded by looking at comparing two independent sets of data. But what happens if we have two variables that are (potentially) related? In this unit we will be investigating that question. We need to look at the relationships between variables, and trying to draw conclusions based on those relationships.

A big part of this will be understanding the difference between correlation and causation – a lot of times we can find data that is highly correlated (which appears to have a strong statistical connection) but that has no causal connection. Recognizing **correlations** and interpreting them can sometimes allow us to ask better questions and design experiments that can ultimately lead us to conclusions about causation (but much more about that in the unit on collecting data and designing experiments.

An interesting example of correlation without causation is the fact that shark attacks have a strong, positive correlation with ice cream consumption. When you track data, you find that the more ice cream purchased by a population the higher the rate of shark attacks gets, and the less ice cream purchased by a population, the lower the rate of shark attacks.



Hopefully, it is obvious that eating ice cream does not cause shark attacks (or that more shark attacks don't cause populations to eat more ice cream). A more plausible explanation for the strong correlation is both ice cream sales and swimming in the ocean increase in the summer months, and therefore shark attacks happen more frequently as a result of more swimmers in the ocean. The truth is, of course, is that we don't actually **know** why these two things are so highly correlated, but just assuming that one causes the other is highly flawed logic.

Rather, we should use the data to make well-reasoned conjectures and think of *questions* around which we could design an *experiment* or series of experiments.

Summary:

- Don't assume **correlation** means **causation** – rather, we should use correlations that we find in data to ask good questions and ultimately design experiments to test for actual relationships in data.

2.1 Two Categorical Variables, Representations and Statistics

Objectives:

- Use graphical representations of two categorical variables to compare data and determine if variables are potentially associated.
- Evaluate data in two-way tables
- Generate joint relative frequencies from data in two-way tables.

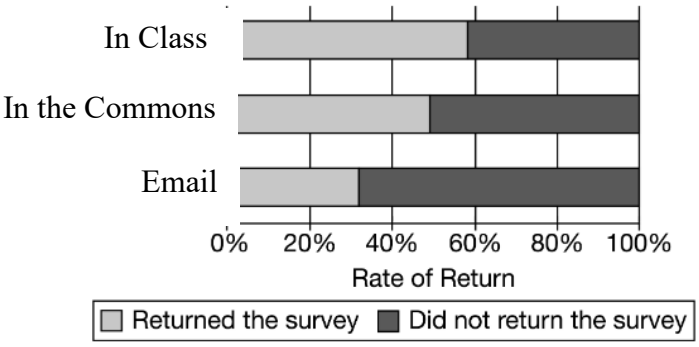
In the last section, we looked at comparing two sets of data based on the shapes of graphs or qualities we could interpret based on their statistics. We are going to do something that initially appears very similar, but we are going to focus on graphs and tables for a categorical variable broken down by categories of *another variable*.

For example, perhaps I survey a group of 100 people on whether they prefer to eat ice cream in a waffle cone or a regular cone. If I decide to break the data up and observe the results of that question for children and adults, I have broken down the initial variable (cone preference) into categories of a second variable (age).

Example 1: Mr. Murphy wanted to survey students at SI, but he wanted to know if different methods of administering the survey would lead to different return rates. He decided on the following three methods:

- **In Class**: Surveys were given during class and asked to return them when completed.
- **In the Commons**: Surveys were distributed in the Commons to students on their break, and they were asked to turn them in when they were completed.
- **Email**: Surveys were emailed to students, and they were asked to turn them in when they were completed.

Mr. Murphy collected the data and put together the following side-by-side, segmented bar graphs with the data he collected. Interpret the results.



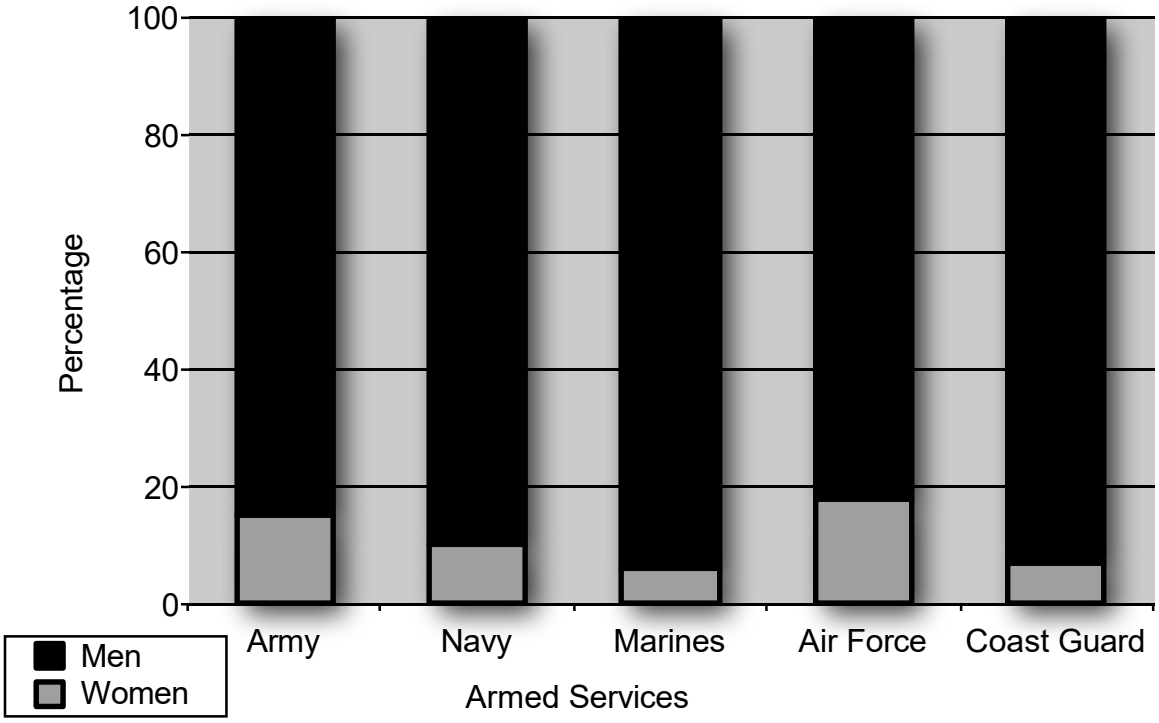
With these kinds of graphs, many people mistakenly try to overgeneralize and state "trends" (like "emailing surveys is a bad way to survey students"). You should not do that. Rather, focus on the data in the graph – do not try to project to a conclusion; with this kind of data you cannot make that kind of conclusion.

You could appropriately state: "In class has the highest rate of return for the students surveyed, and email has the lowest rate of return for the students surveyed."

The AP Test likes to make some multiple choice questions where the wrong answers have conclusions that are overstated compared to the conclusion you can actually make from the data, so be careful!

Example 2: The breakdown of percentages of men and women in the armed services in 2000 is given in the following **segmented bar graph**.

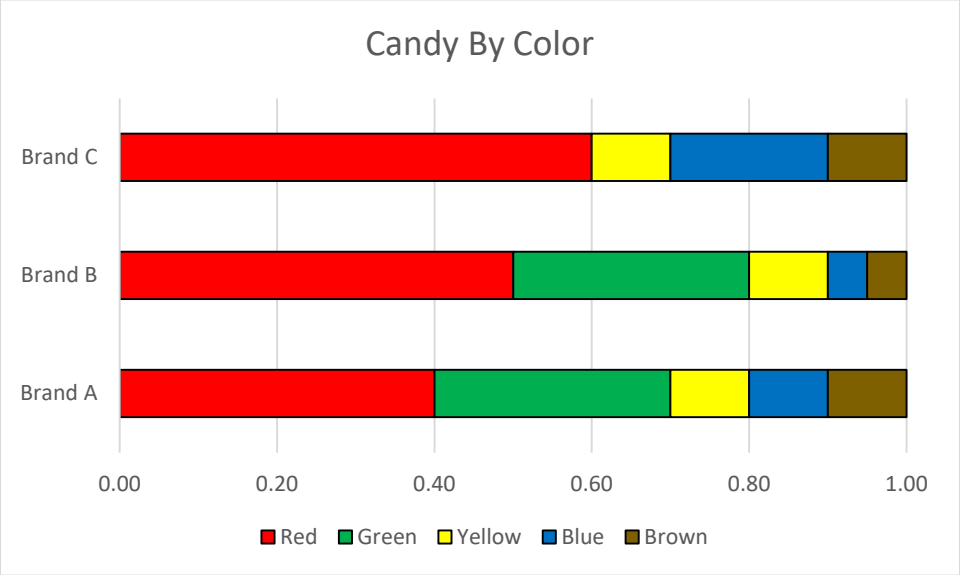
Which of the following can be stated from an observation of the chart?



a) The number of women in the Marines is less than the number of women in any other armed service.
b) The number of men in the Air Force is less than the number of men in any other armed service.
c) The percentage of women in the Marines is less than the percentage of women in any other armed service.
d) The proportion of men in each of the services is the same.
e) The percentage of women in the armed services is changing over time.

The correct answer was c). a) and b) are incorrect because they refer to the actual number of women, but this table only shows *relative frequencies* (note the *y*-axis is labeled "percentage"). d) is obviously wrong because all of the bars clearly show different proportions of men, and e) makes no sense because this is the data for one year – we know nothing about a change over time.

Example 3: There are three brands of colored candies that are sold in bags. Mr. Murphy buys one bag of each type and counts out the colors and puts the relative frequencies onto the segmented bar graph, below:

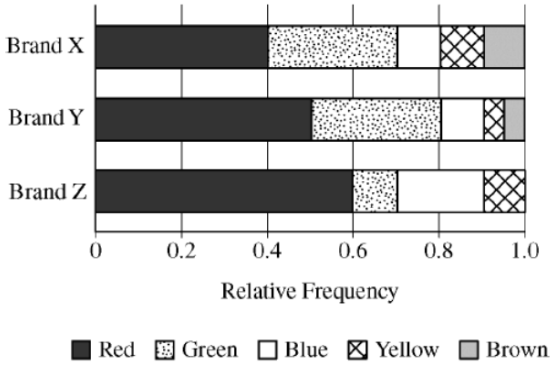


Which of the following statements must be true?

a) For Brand A, there were more green candy pieces than red candy pieces in the bag.
b) For Brand B, there were twice as many green candy pieces as blue pieces.
c) There were more green candy pieces in the Brand B bag than there were in the Brand A bag.
d) For Brand C, there were the same number of yellow candy pieces as brown candy pieces.
e) All three bags had the same number of yellow candy pieces.

The correct answer is d). a) and b) are incorrect by observing the proportions of the candies in the relevant bags. c) and e) are wrong because you cannot compare numbers of candies in the different bags – we only have relative frequencies; we do not know how many candies were in each bag.

I showed this particular problem in color to make the introduction to this a little easier, but on the AP Test you will see them in black and white with different shading or patterns to distinguish them. A similar question from the AP test might look like this:



Let's look at another way of displaying data – the two-way table. Suppose I ask students in grades 11 and 12 if they like English or Math better as a subject. I need to have the data broken down by grade and by subject preference (two separate ways, thus the name "two-way table". Here is a display of the data using this type of table:

	11^{th} Grade	12^{th} Grade	Total
English	27	32	59
Math	58	48	106
Total	85	80	165

A lot of information is contained in this table:

- There are 165 students total who responded to the question
  - 85 were juniors, 80 were seniors (grades 11 and 12)
- 59 preferred English, while 106 preferred math
- We could also list out the preferences by specific grade and subject (for example, 27 11^{th} graders preferred English)

We can also break these into proportions:

What portion of students were in 11^{th} grade and preferred Math?

$\frac{58}{165} = 0.352$ This is called a **joint relative frequency**.

- **Joint Relative Frequency:** A cell frequency divided by the total for the whole table.

Example 3: Mr. Maychrowitz is curious to find out if students in the gaming club prefer Elden Ring or Horizon: Forbidden West. Since Elden Ring is a video game that is rated for 17+ ages, he decides to only ask juniors and seniors, and he puts the results in a two-way table. Use that table to answer the questions below:

	11^{th} Grade	12^{th} Grade	Total
Elden Ring	8	9	17
Forbidden West	14	19	33
Total	22	28	50

a) What is the joint relative frequency of 12^{th} graders who preferred Forbidden West?

b) What is the proportion of students who preferred Elden ring, regardless of age?

c) What is the proportion of 11^{th} graders who responded to this survey?

d) What proportion of 12^{th} graders preferred Elden Ring?

Answers: a) 0.38 b) 0.34 c) 0.44 d) 0.321

(Obviously, Mr. M is horrified by these results and begins the hard work of indoctrinating his club into becoming avid fans of Elden Ring).

The answers to b), c), and d) are not *joint relative frequencies*. They are actually slightly different quantities:

- **Marginal Relative Frequency:** A row or column total that is divided by the table total. Parts c) and d) are both *marginal relative frequencies*.
- **Conditional Relative Frequency:** A relative frequency for a specific part of the table, like a cell frequency in a column (or row) divided by that column (or row). Part d) is a *conditional relative frequency*.

Example 4: Convert the two-way table in example 3 into a two-way relative frequency table

	11^{th} Grade	12^{th} Grade	Total
Elden Ring	8	9	17
Forbidden West	14	19	33
Total	22	28	50

Start by taking each cell, and divide by the total number of data points in the table (in this case, 50 people – highlighted in the table above)

	11^{th} Grade	12^{th} Grade	Total
Elden Ring	$\frac{8}{50}=0.16$	$\frac{9}{50}=0.18$	$\frac{17}{50}=0.34$
Forbidden West	$\frac{14}{50}=0.28$	$\frac{19}{50}=0.38$	$\frac{33}{50}=0.66$
Total	$\frac{22}{50}=0.44$	$\frac{28}{50}=0.56$	$\frac{50}{50}=1.00$

Notice that all of the values in the total column still do total to 1.00 (100% of the data).

	11^{th} Grade	12^{th} Grade	Total
Elden Ring	0.16	0.18	0.34
Forbidden West	0.28	0.38	0.66
Total	0.44	0.56	1.00

We can actually use two-way tables and two-way relative frequencies to determine if there may be a relationship between the two variables.

Example 5: Suppose we collect data on high school students who have a job and high school students who own a car. We want to see if there could be a relationship between the ownership of a car and having a job. We collect data from 75 students and find the following:

	Own a car	Do not own a car	Total
Have a job	21	9	30
Do not have a job	17	28	45
Total	38	37	75

If I think there may be a relationship between having a job and owning a car, I should compare what proportion of the students with a job own a car (the *conditional relative frequency*) with the proportion of students who own a car generally (the *marginal relative frequency*).

$$\frac{\text{own a car and have a job}}{\text{have a job}} = \frac{21}{30} = 0.70$$

$$\frac{\text{own a car}}{\text{total}} = \frac{38}{75} = 0.51$$

Since $0.70 > 0.51$ the data indicates a relationship between having a job and owning a car. This seems to indicate a positive association between the two variables. It does not guarantee a relationship, but it gives us an indication that we should further explore this question (by designing an experiment to test this supposition, if possible, or by collecting more data to explore the issue further).

Relationships established by data in two-way tables:

- If a *conditional relative frequency* is significantly different than a *marginal relative frequency* this **indicates a relationship between the data**.
    - If the conditional relative frequency *is greater*, that indicates **a positive relationship.**
    - If the conditional relative frequency *is less*, that indicates **a negative relationship.**
- If a *conditional relative frequency* is <u>not</u> different than a *marginal relative frequency* this **indicates <u>no</u> relationship between the data**.

Of course, what does "significantly different" mean. For right now, it will just be something that is obviously very different from the other value.

Example 6: Suppose I collect data on political party affiliation, and I believe there be a difference between party affiliation and gender. I collect data from 1000 people and the data is listed in the two-way table below:

	Men	Women	Total
Party A	150	250	400
Party B	225	375	600
Total	375	625	1000

Since I am looking for any relationship, I decide to see if women are a greater proportion of Party A than of the general survey group:

$$\text{Proportion of women in Party A} = \frac{250}{400} = 0.625$$

$$\text{Proportion of women surveyed} = \frac{625}{1000} = 0.625$$

Since these values are the same, it indicates no relationship.

Summary:

- **Two-Way Table:** A list of two variable data on a table.
  - **Joint Relative Frequency:** A cell frequency divided by the total for the whole table.
  - **Marginal Relative Frequency:** A row or column total that is divided by the table total.
  - **Conditional Relative Frequency:** A relative frequency for a specific part of the table, like a cell frequency in a column (or row) divided by that column (or row).
- If a *conditional relative frequency* is significantly different than a *marginal relative frequency* this **indicates a relationship between the data**.
  - If the conditional relative frequency *is greater*, that indicates **a positive relationship.**
  - If the conditional relative frequency *is less*, that indicates **a negative relationship.**
- If a *conditional relative frequency* is <u>not</u> different than a *marginal relative frequency* this **indicates <u>no</u> relationship between the data**.

Checkpoint 2.1

Multiple Choice:

Problems 1 to 3 deal with the following table and information:

A study has been done to determine whether or not a certain drug leads to an improvement in symptoms for patients for a particular medical condition. The results are shown in the following table:

	Improvement	No Improvement	Total
Drug	270	530	800
No Drug	120	480	600
Total	390	1010	1400

1. Based on this table, what is the probability that a patient shows improvement if it is known that the patient was given the drug?

   (a) 0.3250
   (b) 0.3375
   (c) 0.2250
   (d) 0.4355
   (e) None of the above

2. Based on this table, what is the probability that a patient shows improvement if it is known that they were not given the drug?

   (a) 0.200
   (b) 0.250
   (c) 0.279
   (d) 0.501
   (e) None of the above

3. Based on this table, what is the probability that a patient shows improvement if they participated in this study?

   (a) 0.200
   (b) 0.250
   (c) 0.279
   (d) 0.501
   (e) None of the above

Free Response:

1. A study has been done to determine whether or not a certain drug leads to an improvement in symptoms for patients for a particular medical condition. The results are shown in the following table:

	Improvement	No Improvement	Total
Drug	400	650	1050
No Drug	375	375	750
Total	775	1025	1800

Find each of the following:

(a) Find the conditional relative frequency of people who took the drug and showed improvement in the study.

(b) Find the marginal relative frequency of people who showed improvement in the study.

(c) Does the answer from (a) and (b) indicate that the drug has a positive impact, a negative impact, or no impact at all? Explain your reasoning.

2. A proficiency examination for a certain skill was given to 100 employees of a firm. Forty of the employees were men. Sixty of the employees passed the examination (by scoring above a present level for satisfactory performance). The breakdown of test results among men and women are shown in the accompanying diagram.

	Male (M)	Female (F)	Total
Pass (P)	24	36	60
Fail (F)	16	24	40
Total	40	60	100

Suppose that an employee is selected at random from among the 100 who took the examination.

a. Find the probability that the employee passed, given that he was a man.

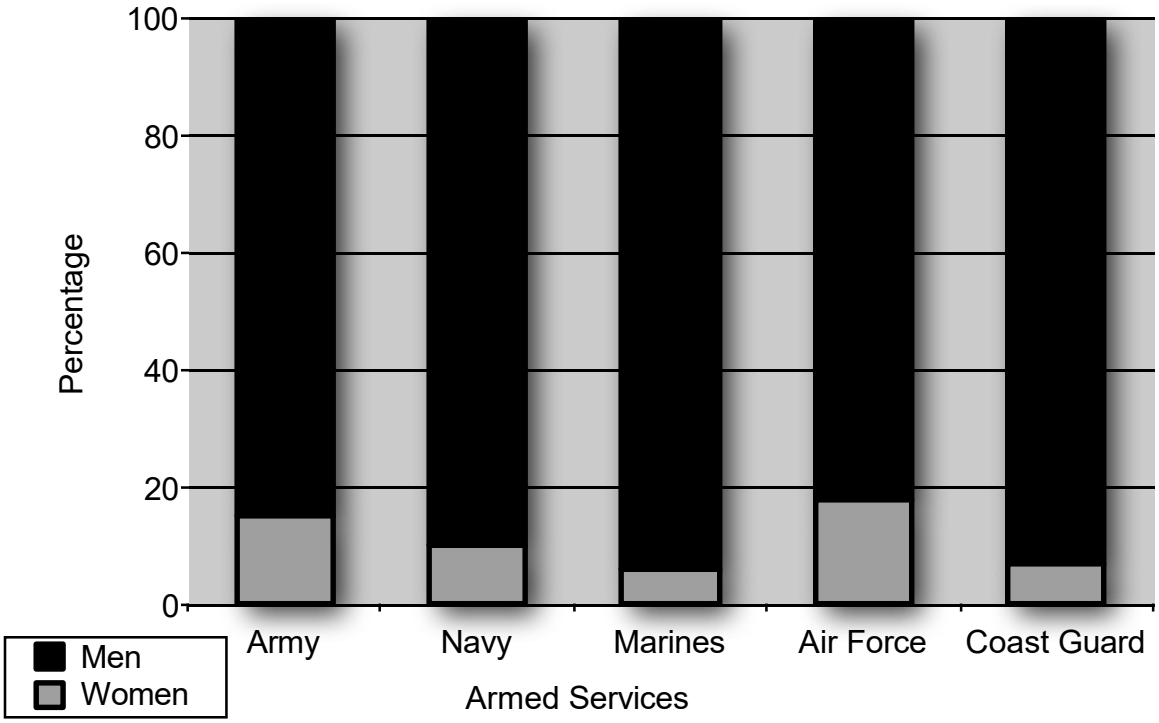
b. Find the probability that the employee was a man, given that a passing grade was received.

c. Are the events passing the exam and male positively related, negatively related, or not related?

d. Are the events passing the exam and female positively related, negatively related, or not related?

2.1 Homework

1) the breakdown of percentages of men and women in the armed services in 2000 is given in the following **segmented bar graph**.



(a) Which branch of the armed services has the highest proportion of women members?

(b) Which branch of the armed services has the highest number of women members?

(c) Which branch had the highest proportion of male members? Which one had the highest number of male members?

(d) Given that there were approximately 482,000 members of the Army in this year, and 356,000 members of the Air Force in the same year, which branch had more women?

(e) Given that the Coast Guard had 34,800 active members that year, and the Marines had 173,300 active members, were there more female Marines or male Coast Guards?

(f) Given that the Coast Guard had 34,800 active members that year, and the Marines had 173,300 active members, were there more female Marines or male Coast Guards?

(g) Given that the Coast Guard had 34,800 active members that year, and the Army had 482,000 active members, were there more female Marines or male Coast Guards?

2) The following two-way table shows the breakdown of SI students in AP Calculus (AB or BC) and AP Statistics in 2022 – 2023 school year*. Use that information to calculate each of the following:

	Male	Female	Total
AP Calculus	95	130	225
AP Statistics	72	69	141
Total	167	199	366

(a) Find the conditional relative frequency of students who are male and in AP Calculus.
(b) Find the conditional relative frequency of students who are female and in AP Calculus.
(c) Find the conditional relative frequency of students who are male and in AP Statistics.
(d) Find the conditional relative frequency of students who are female and in AP Statistics.
(e) Find the marginal relative frequency of female students in AP mathematics.
(f) Find the marginal relative frequency of male students in AP mathematics.
(g) Find the joint relative frequency of female AP Calculus students.
(h) Find the joint relative frequency of male AP Calculus students.
(i) Given the answer to (a) and (h), does there appear to be a relationship between being male and being in AP Calculus for this school year? If there is, what kind of relationship is it? Explain.
(j) Given the answer to (b) and (g), does there appear to be a relationship between being female and being in AP Calculus for this school year? If there is, what kind of relationship is it? Explain.

*Note that this data is not strictly accurate because some students are "double-counted" – that is, they are in both AP Calculus and AP Statistics. However, for illustrative purposes of the idea of a two-way table, it is a useful example without overly complicating.

2.2 Scatterplots: Representing the Relationship Between Two Variables

Objectives:

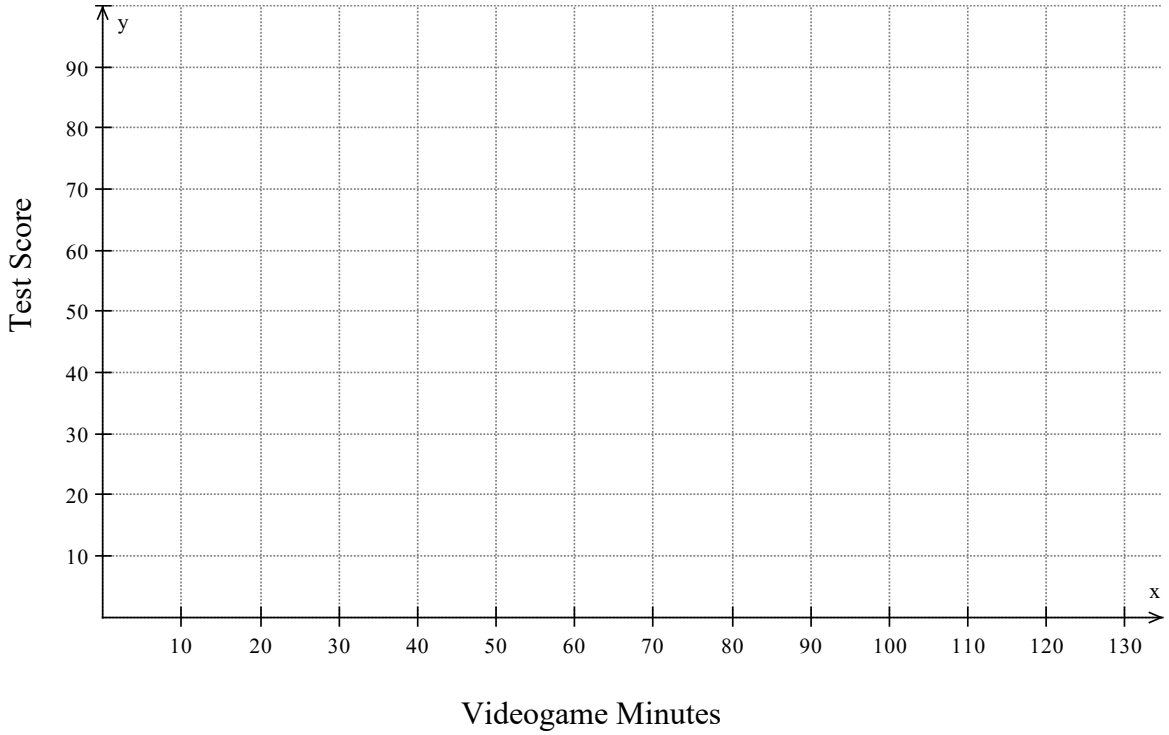
- Represent bivariate data with a scatterplot
- Describe the relationship in a scatterplot
- Calculate and interpret the correlation coefficient, *r*, in context.

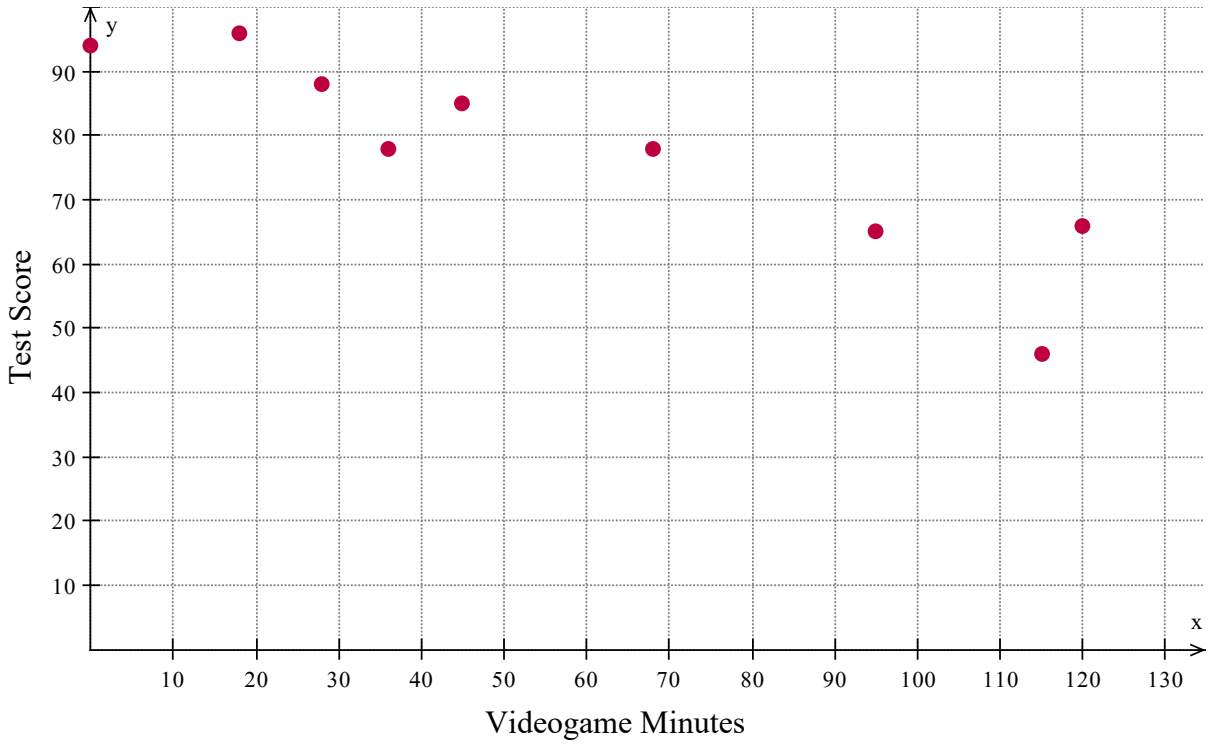
A common way of representing bivariate (two-variable) data is called a *scatterplot*.

- **Scatterplot**: A display of data that shows the relationship between two numerical variables.
    - Each member of the dataset is plotted as a point with *x*-*y* coordinates corresponding to the values of the two variables.

Example 1: Given the dataset below which tracked the minutes playing a videogame the night before a test and performance on the test, make a scatterplot of the data:

Minutes playing videogame	45	18	0	36	28	68	95	120	115
Score on test	85	96	94	78	88	78	65	66	46





- **Always label your variables!** Here we did it by labeling the axes, but anytime you draw any kind of graph, it is essential that you assign the variables (for example, x = videogame minutes, y = test score)

Looking at this set of data on the graph, it looks *roughly linear*. That is, you could draw a line through the "center" of the data and it looks like a generally negatively sloping line. These data points look fairly close together, so we would say the relationship was *strong*.

Therefore, we could conclude that there is a **strong, negative, linear correlation.**

- **Correlation:** The strength and direction of the *linear* relationship between two quantitative variables. Correlation is usually represented by the variable ***r***.
- The correlation between x and y is given by this formula:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

That formula looks a little complicated, and in truth, we will use a calculator to calculate r. Notice that the formula is product of the *deviation* of each variable divided by the *standard deviation* of each variable (note that this is also the product of the z–scores). All those products are then averaged, and that is r (again, note that it is not a true average, as we are dividing by $n - 1$ for the sample r value).

If we calculated r for the dataset in example 1, we would get $r = -0.91$. We will discuss more of what this means later in the section.

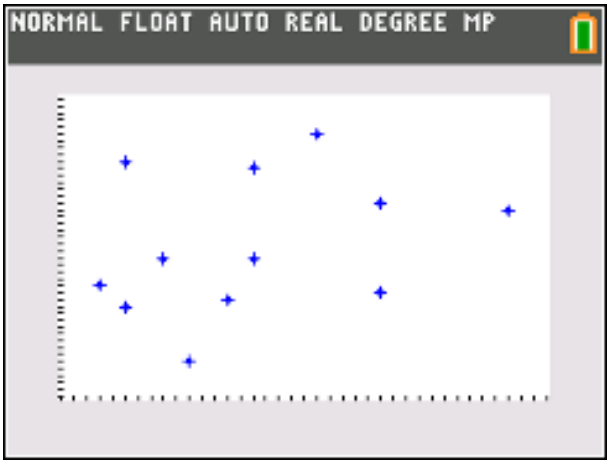
Example 2: Are more expensive bike helmets safer than less expensive ones? The accompanying data on x = price and y = quality rating for 12 different brands of bike helmets is given below. Quality rating was a number from 0 (worst possible rating) to 100, and was determined based on factors that included how well the helmet absorbed the force of an impact, the strength of the helmet, ventilation, and ease of use. Use your calculator to generate a scatterplot and find r.

Price	Quality Rating
35	65
20	61
30	60
40	55
50	54
23	47
30	47
18	43
40	42
28	41
20	40
25	32

First, enter this data as L_1 and L_2 in your calculator.

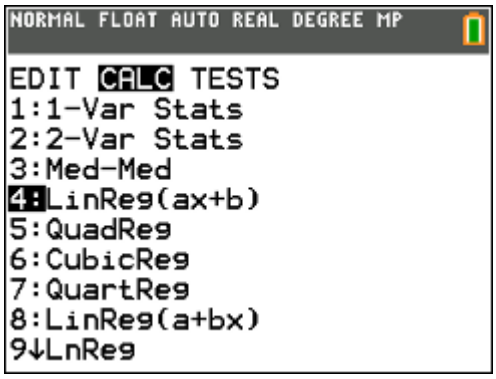
Important: Keep the data in the order it is in the tables – they are ordered pairs connected to each other. If you move values around, you will change the results of your analysis.

First go to **stat plot** and select the first option under type (it looks like a series of dots – just what we would want for a scatterplot) and the use ZoomStat.

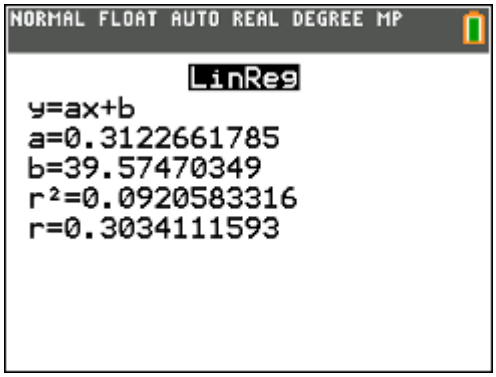


These dots look kind of all over the place, but there may be a *weak*, *positive*, *linear* correlation as they generally tend to look like they go up from left to right.

Now go to the **STAT** menu again, and go to the CALC menu:



We want option 4 (or option 8) – the calculator will perform a linear regression for us (more on this in the next section). Make sure you enter L_1 and L_2 as the lists for your regression.



Note that $r = 0.303$. This is a weak, positive correlation, just as we surmised looking at the scatterplot.

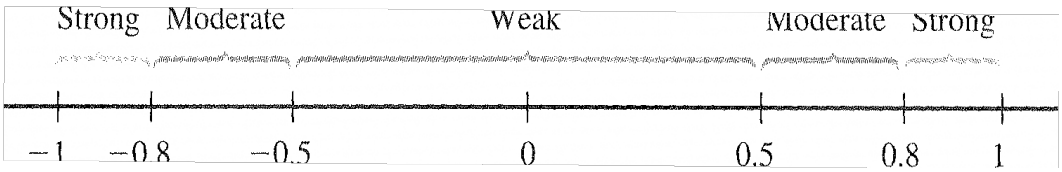
You may have noticed that I have used the words **weak** and **strong** with regards to correlation. This simply refers to how linear the relationship is.

- **Weak correlation**: the linear relationship between the data is slight.
- **Moderate correlation**: the linear relationship between the data is moderate.
- **Strong correlation**: the linear relationship between the data is strong.

Kind of obvious when you think about it, right? But what establishes strength? We cannot just go by an "eye-test" every time, because that is very subjective. Instead, we look at our calculated *r* values.

Properties of *r*:

- The value of *r* is between –1 and +1.
  - *r* = –1 is *perfectly linear* (all the points lie on a line) and has a *negative slope*.
  - *r* = +1 is *perfectly linear* (all the points lie on a line) and has a *positive slope*.
- The value of *r* does not depend on the unit of measurement for either variable.
- The value of *r* does not depend on which of the variables is considered *x* or *y*.



- The value of *r* is a measure of the extent to which *x* and *y* are linearly related.
- When interpreting *r* there are three qualifiers you must always mention:
  - Positive or negative
    - Positive means that as *x* increases, *y* increases.
    - Negative means that as *x* increases, *y* decreases.
  - Strength
  - Linear

Example 3: The following data on the average finishing time by age group for female participants in the New York City marathon is given below. Find and interpret r. Create a scatterplot of the data (either by hand or on the calculator) and find and interpret *r*.

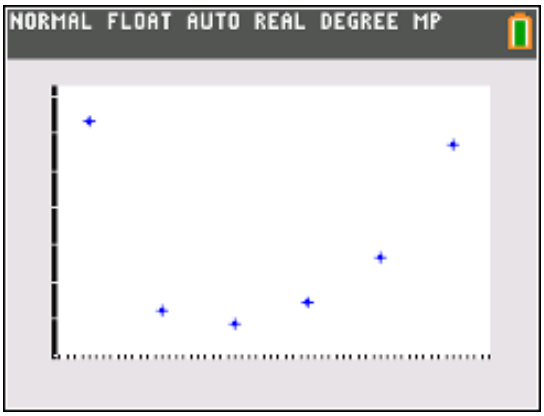
Age Group	Representative Age	Average Finish Time
10 - 19	15	302.38
20 - 29	25	193.63
30 - 39	35	185.46
40 - 49	45	198.49
50 - 59	55	224.3
60 - 69	65	288.71

Answer: Notice that this is categorical data for the "Age Group" so we convert it to numerical data by assigning a "Representative Age". So we have two variables:

x = Representative age $\qquad$ y = Average finish time

Running the data through the calculator we find an $r = 0.038$. This means that the data is *positive* and *weakly linear.*

In fact, this correlation is incredibly weak – it is close to 0. There may be a correlation that is not linear, or there may be no relationship at all. Let's take a look at our scatterplot:



This does not look linear at all, and the relationship (which seems to be there) might be quadratic or some other function. We will look at linear and non-linear correlations in subsequent sections.

An interesting thing to note, however, is that just because we have an r value that is close to 1 or –1 does not mean that a linear function is actually the best model for the data.

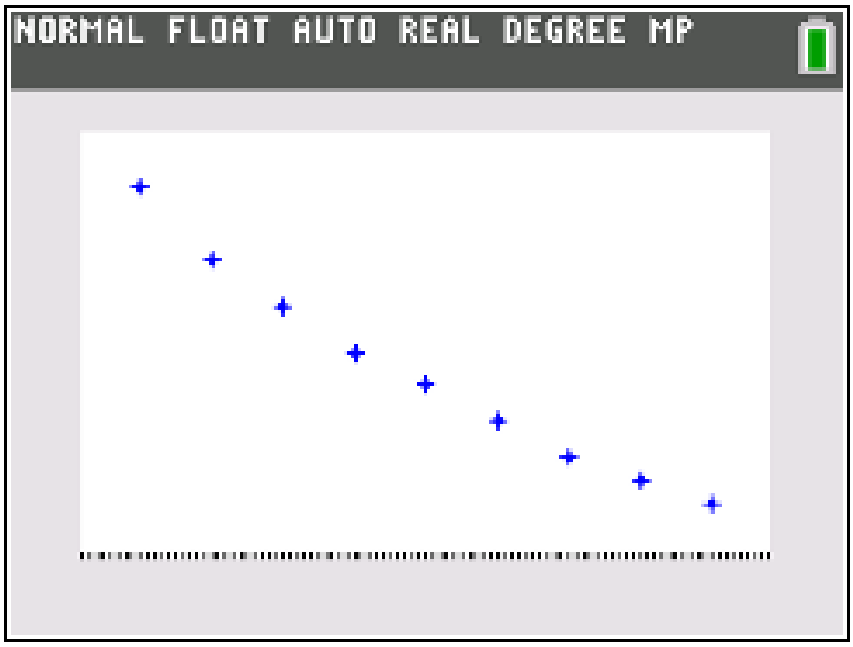
Example 4: An AP Chemistry student is collecting data on the decomposition of crystal violet solution using a photospectrometer. She collects the data and organizes it on the chart below:

Time (seconds)	10	20	30	40	50	60	70	80	90
Absorbance	0.90	0.78	0.70	0.62	0.57	0.51	0.45	0.41	0.37

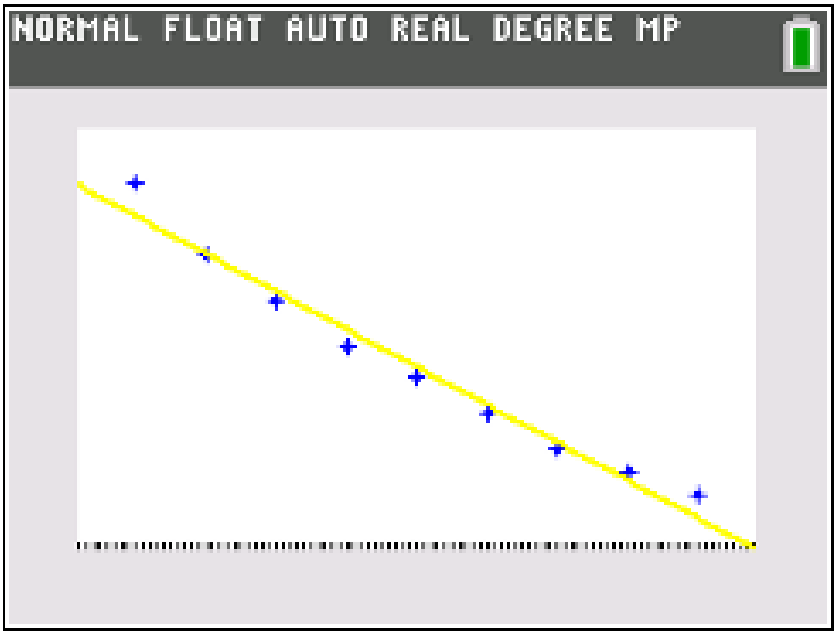
Calculate the r value and interpret its meaning.

When we use the calculator, we get $r = –0.987$. This is really close to –1, so if we never looked at a graph of the scatterplot, we would assume that we have a *strong*, *negative*, *linear correlation.*

However, if we take a look at the scatterplot, we see that the data is pretty clearly curving and not linear. We cannot simply rely on an *r* value to let us know that the relationship is linear – scatterplots must often pass an "eye-test" (there are other ways to check this, which we will examine in section 2.5)

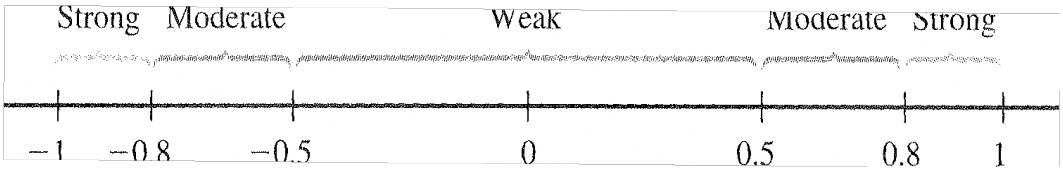


The curve becomes more apparent if I superimpose a line over the function. We can see that the data points start above the line, dip below it, and come back above it – very characteristic of data that is "curved".



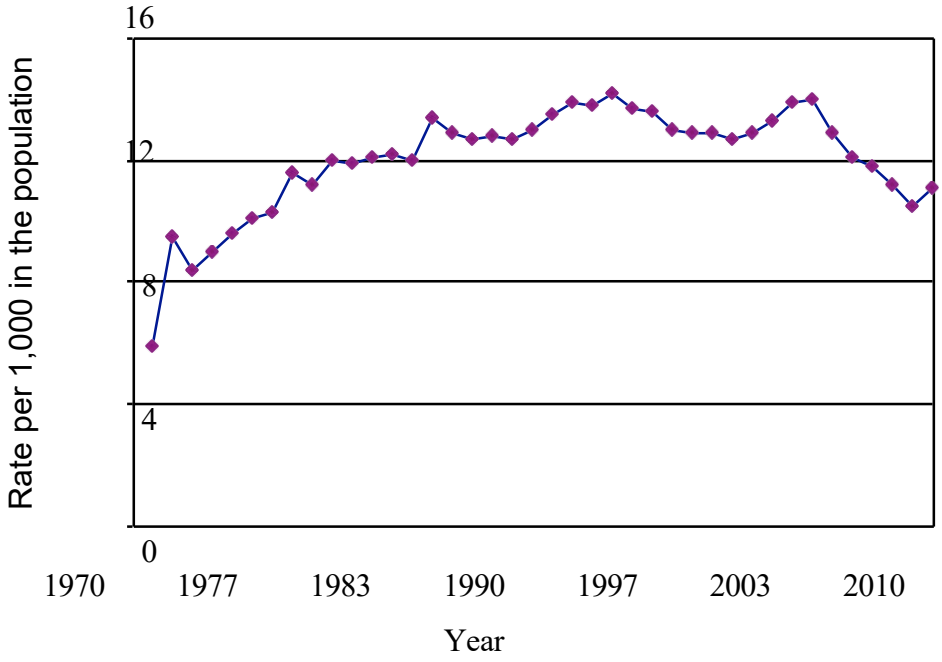
Summary:

- **Scatterplot**: A display of data that shows the relationship between two numerical variables.
    - Each member of the dataset is plotted as a point with *x*-*y* coordinates corresponding to the values of the two variables.
- **Sample Correlation (*r*):** Measures how strongly *x* and *y* in a ***sample*** of pairs are linearly related to each other.
- **Correlation does not imply causation:** Just because two variables are highly correlated does not mean that one causes the other.
- **Interpreting *r*:** You must mention all of the following:
    - **Strength of correlation:** weak, moderate, or strong.
    - **Direction:** whether the correlation is positive or negative.
    - **Linear:** you must state that it is a *linear relationship*.



- Just because data has an *r* value close to ± 1 does not guarantee a linear relationship.

Checkpoint 2.2



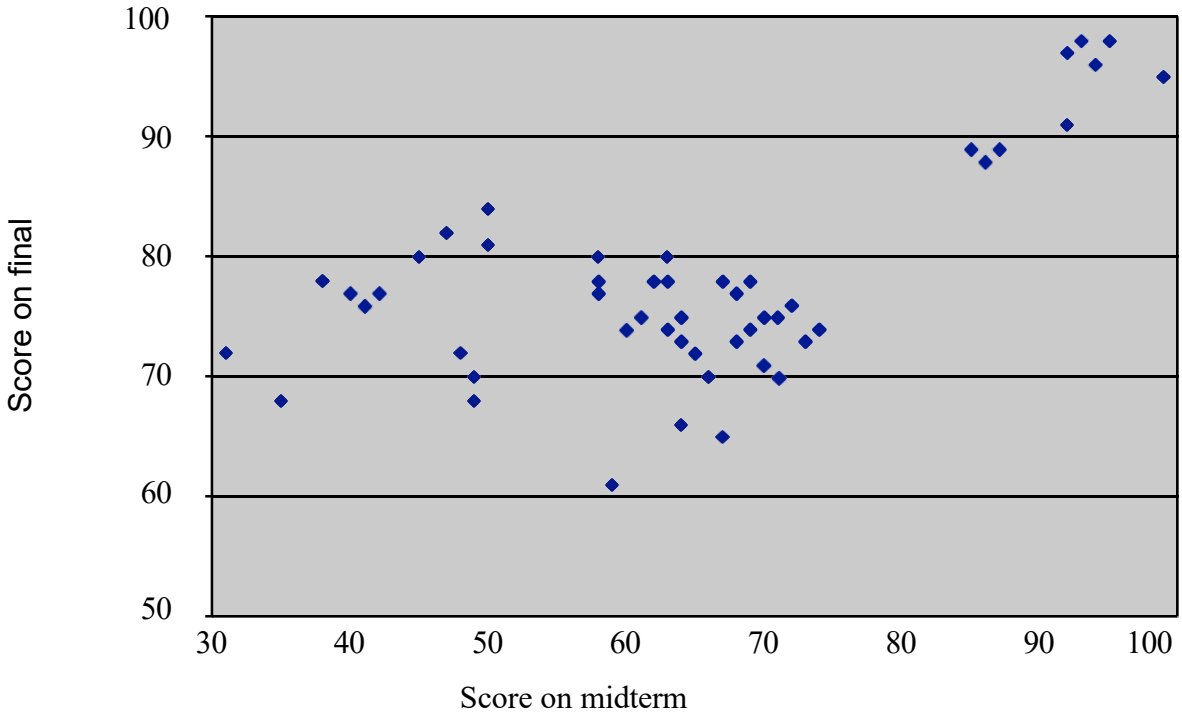
1. The following time-series graph displays the divorce rate in the United States (per 1000 in the population from 1971 to 2010).

   The scatterplot shows

   (a) a downward trend in the divorce rate.
   (b) an upward trend in the divorce rate.
   (c) a uniform divorce rate.
   (d) a normal divorce rate.
   (e) none of the above.

2. Refer to the graph in Question 1. Which year had the highest divorce rate? The lowest divorce rate?
   (a) 2004, 1973
   (b) 1971, 2004
   (c) 1984, 1971
   (d) 1980, 2001
   (e) 1994, 1971

3. Extra study sessions were offered to students after the midterm to help improve their understanding of statistics. Student scores on the midterm and the final exam were recorded. The following scatterplot shows the final test scores against the midterm scores.



Which of the following statement correctly interprets the scatterplot?

(a) All students have shown a significant improvement in the final exam scores as a result of the extra study sessions.
(b) The extra study sessions were no help. Each student's final exam score was about the same as his or her score on the midterm.
(c) The extra study sessions further confused students. All student scores decreased from the midterm to final exam.
(d) Students that scored below 55 on the midterm showed considerable improvement on the final exam; those who scored between 55 and 80 on the midterm showed minimal improvement on the final exam; and those who scores above 80 on the midterm showed almost no improvement on the final exam.
(e) Students that scored below 55 on the midterm showed minimal improvement on the final exam; those who scored between 55 and 80 on the midterm showed moderate improvement on the final exam; and those who scores above 80 on the midterm showed considerable improvement on the final exam.

4. If there is a very strong correlation between two variables, then the correlation coefficient should be

   (a) close to +1
   (b) close to -1
   (c) close to -1 or +1
   (d) close to zero
   (e) There is no way to determine the correlation coefficient.

5. You are given the following set of observations for variables *x* and *y*.

x	-3	-1	1	3
y	8	4	5	-1

The correlation coefficient is:

(a) -1.0 (b) -0.8971 (c) +1 (d) 0.8971 (e) .2349

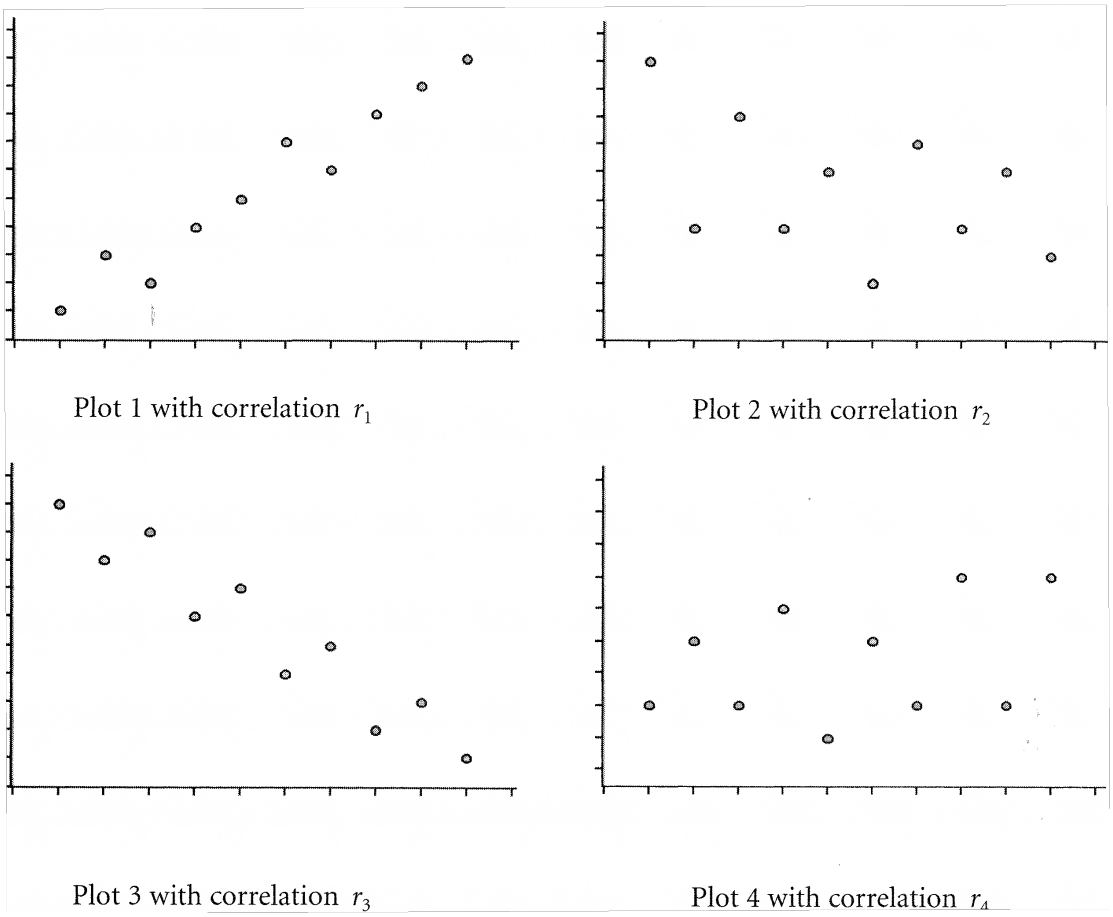
6. Pearson's correlation coefficient (*r*) is considered a symmetric measure because:

   (a) its values range from 0 to 1.
   (b) it indicates the causal relationship between two variables.
   (c) the sign of r is the same as the sign of the slope.
   (d) it will be the same regardless of which variable is the x and which is the y.
   (e) None of the above.

7. Suppose the correlation between two variables is $r = 0.23$. What will be the new correlation if 0.14 is added to all values of the *x* variable, every value for the *y* variable is doubled, and the two variables are interchanged?

   (a) 0.23 (b) 0.37 (c) 0.74 (d) -0.23 (e) -0.74

8. Order the correlation coefficients from least to greatest for the given scatterplots.



Plot 1 with correlation r_1

Plot 2 with correlation r_2

Plot 3 with correlation r_3

Plot 4 with correlation r_4

(a) $r_4 < r_3 < r_2 < r_1$

(b) $r_4 < r_2 < r_3 < r_1$

(c) $r_3 < r_2 < r_4 < r_1$

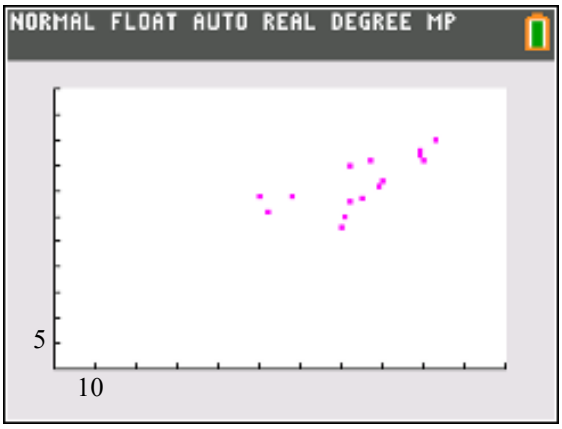
(d) $r_2 < r_3 < r_4 < r_1$

(e) $r_1 < r_2 < r_3 < r_4$

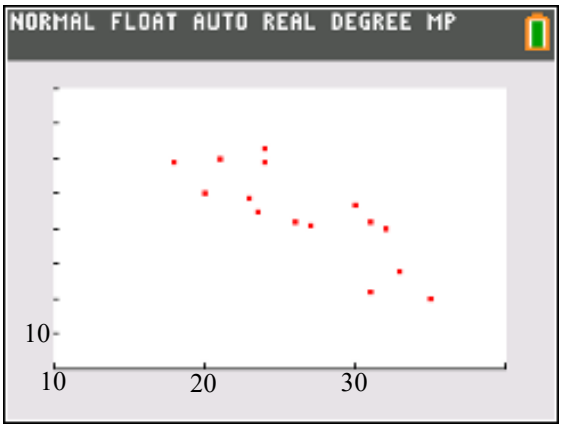
2.2 Homework

1. For each of the scatterplots show below, answer the following questions:
    a. Does there appear to be a relationship between x and y? If there is a relationship, does it appear to be linear?
    b. If there appears to be a linear relationship, would you describe it as positive or negative?

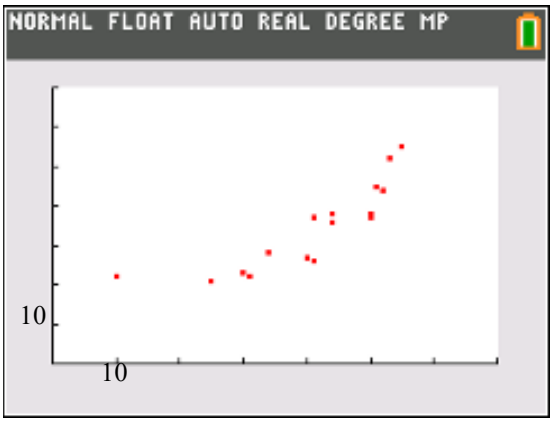
Scatterplot 1:



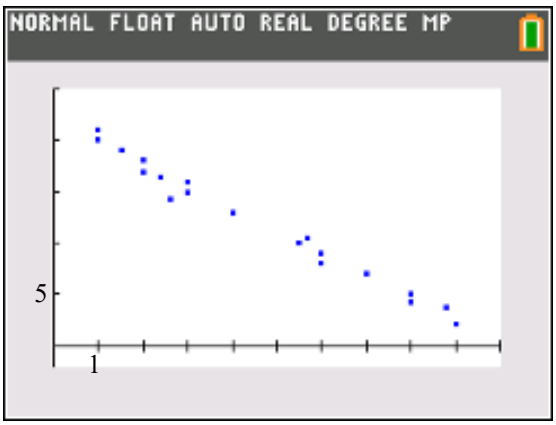
Scatterplot 2:



Scatterplot 3:



Scatterplot 4:



2. For each of the following pairs of variables, state whether you would expect a positive correlation, a negative correlation, or a correlation close to 0. Explain your choice.
    a. The maximum daily temperature in a region and energy costs for cooling a home in that region.
    b. Height of an individual and the IQ score of that individual.
    c. Height of an individual and shoe size of that individual.
    d. Score of an individual on the Math Section of the SAT and score of the same individual on the Math Section of the ACT.

3. The data in the table below are x = cost (in cents per serving) and y = fiber content (in grams per serving) for 18 high fiber cereals.

Cost per Serving	Fiber per Serving	Cost per Serving	Fiber per Serving
33	7	53	13
46	10	53	10
49	10	67	8
62	7	43	12
41	8	48	7
19	7	28	14
77	12	54	7
71	12	27	8
30	8	58	8

   a. Calculate the value of the correlation coefficient, r, and interpret its value for this data set.
   b. For these cereals, the serving size varies from ½ cup to 1¼ cups. Converting the price and fiber content to "per cup" instead of "per serving" gives the data displayed in the table below. Calculate and interpret the correlation coefficient, r, for this data set, and interpret its value. Is it greater than or less than the correlation coefficient for the data calculated in part a? Give a possible reason for the difference in the correlation coefficient (if there is a difference).

Cost per cup	Fiber per cup	Cost per cup	Fiber per cup
44.0	9.3	53.0	13.0
46.0	10.0	53.0	10.0
49.0	10.0	67.0	8.0
62.0	7.0	43.0	12.0
32.8	6.4	48.0	7.0
19.0	7.0	56.0	28.0
77.0	12.0	54.0	7.0
56.8	9.6	54.0	16.0
30.0	8.0	77.3	10.7

   c. Create a scatterplot for both sets of data and comment on any similarities or differences in the plots.

4. The authors of the scientific paper, **"Flat-footedness Is Not a Disadvantage for Athletic Performance in Children Aged 11 to 15 Years" (*Pediatrics* [2009]: e386–e392)**, studied the relationship between y = arch height of the foot and the scores on several different motor ability tests for 218 children. After recording and analyzing the data they collected, they reported the following correlation coefficients:

Motor Ability Test	Correlation between Test Score and Arch Height
Height of Counter Movement Jump	–0.02
Hopping: Average Height	–0.10
Hopping: Average Power	–0.09
Balance, Closed Eyes, One Leg	0.04
Toe Flexion	0.05

   a. Interpret the value of the correlation coefficient between Average Hopping Power and Arch Height. What does the fact that the r value is negative indicate about the relationship between arch height and hopping power?
   b. The title of the paper suggests that having a small arch height (i.e. being flat-footed) is not a disadvantage for motor skills and athletic abilities for this age group. Do the correlation coefficients given above seem to support this title or not? Explain briefly.

5. Mr. Maychrowitz was interested in seeing the relationship between x = the time he spent in a play session of Elden Ring (in minutes) and y = the amount of runes (experience) gained by his character in the game. He collected for each play session over a 15 day period, and the results are listed in the table below:

Time Played (minutes)	Runes Acquired	Time Played (minutes)	Runes Acquired	Time Played (minutes)	Runes Acquired
114	16,382	57	10,806	88	18,433
63	11,340	42	8,566	95	33,752
245	89,544	133	108,331	68	68,448
18	32,233	46	111,680	84	165,106

   a. Create a scatterplot for the data and calculate the correlation coefficient for the data set. Interpret the value of the correlation coefficient, r, that you calculated.
   b. Many people would assume that there would be a strong linear correlation between time playing a game and experience earned in the game, but this does not seem to be the case for this data. What might be the cause for this apparent discrepancy (keep in mind that experience in games like this is generally earned for defeating "monsters" in the game and that this is an "open-world" exploration style game as well).

2.3 Linear Regression Models

Objectives:

- Write the equation for a regression line.
- Interpret the slope and *y*-intercept of a regression line in context.
- Detect when extrapolation occurs.

When we looked for *r* in the last section, you may have noticed that we used a function denoted as "LinReg" on the calculator. This stands for *Linear Regression*.

- **Linear Regression:** Using a linear function to model the relationship between two numerical variables.

In algebra, *x* was the *independent variable* and *y* was the *dependent variable.* These were so named because, when lines are in $y = mx + b$ form, when we select *x* values, the *y* values depend on the selected *x* values. In statistics, we use slightly different vocabulary:

- **Explanatory variable:** The *x* variable, also called the independent or predictor variable
- **Response variable:** The *y* variable, also called the dependent variable.

Statistics also uses a different format for lines*:

- $y = ax + b$ is the common way of expressing the equation of a line in statistics
  - a = the slope of the line
  - b = the *y*-intercept of the line.

The full name for the line that we generate is called the **Least Squares Regression Line**. Essentially, the process of generating the line creates a line that is essentially an average distance from each of the points. This is not absolutely technically correct, but it is a good way of conceptualizing the process.

- Slope of the **Least Squares Regression Line** (sometimes called the LSRL):

$$a = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

Again, we will normally have the calculator calculate this value for us via LinReg.

*Many statisticians use $y = a + bx$ for the equation of a line, where a = the *y*-intercept, and b = the slope. The calculator has both options (under the STAT Calc menu, options 4 and 8). There is no particular advantage for either one, it is simply personal preference.

- The equation of the least squares regression line is as follows:

$$\hat{y} = ax + b$$

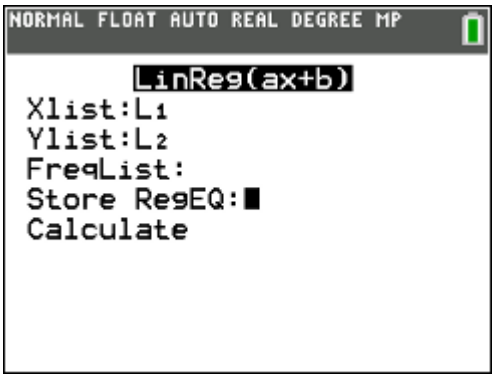
When the ^ is above the variable ($\hat{y}$ is read as "*y*-hat"), it indicates that it is a **prediction** for *y* when we put a particular *x* into our equation.

It is essential that you always include the "hat" on your response variable when you write up a least squares regression line. You will lose points on the AP Test if you neglect this.

Example 1: You are given the following set of observations for variables *x* and *y*. Use this and your calculator to create a linear regression line.

x	-3	-1	1	3
y	8	4	5	-1

Use the exact same process from the last section: STAT → Calc → 4:LinReg(ax+b)

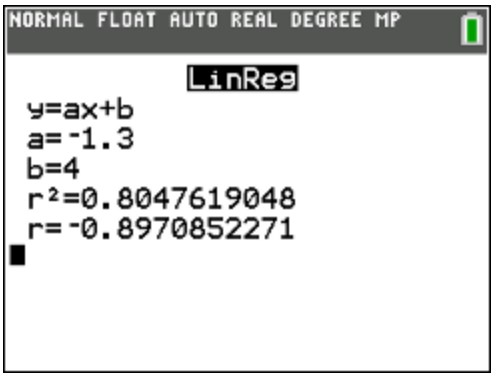


The only difference is that in the "Store RegEQ:" spot, we are going to have our calculator put the equation in Y_1 in our graphing screen.

To do this hit the **VARS** button (just below the directional keys)

1) Move to the Y-VARS menu
2) Select "1:Function"
3) Select Y_1

Now the calculator will take the regression line and plot it on the graph when you run the calculation.

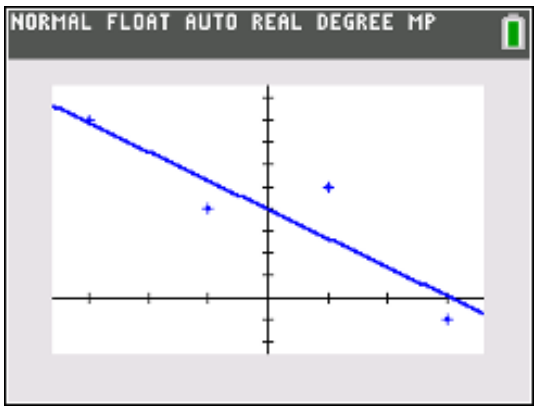


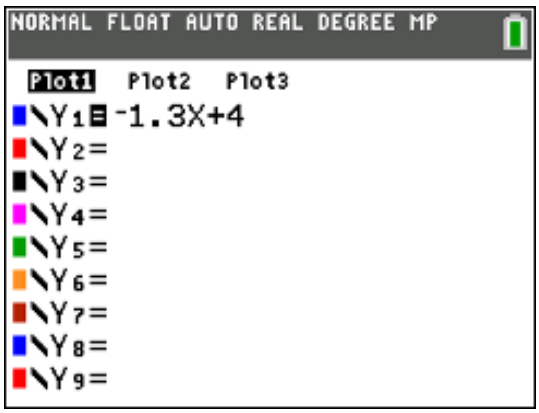
This indicates our least squares regression line $\hat{y} = -1.3x + 4$.

$r = -0.897$

This indicates a strong, negative linear correlation.

If we ZoomStat now, we will see both the scatterplot and the line. If we look at our "y=" screen, you will see the equation has been placed there for us.





Example 2: Studies have shown that people who suffer sudden cardiac arrest (SCA) have a better chance of survival if a defibrillator shock is administered very soon after cardiac arrest. How is survival rate related to the time between when cardiac arrest occurs and when the defibrillator shock is delivered?

Data are given below where y = survival rate (percent) and x = mean call-to-shock time (minutes) for a cardiac rehabilitation center (where cardiac arrests occurred while victims were hospitalized and so the call-to-shock time tended to be short) and for four communities of different sizes:

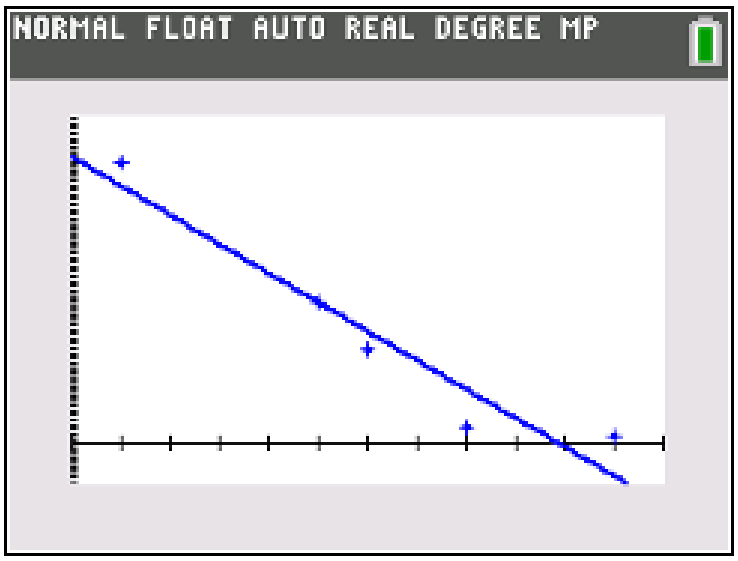
Mean call-to-shock time, x	2	6	7	9	12
Survival rate, y	90	45	30	5	2

Find the least square regression line, r, and the scatterplot. Interpret the slope and the y-intercept in the context of this problem.

Answer:

x = mean call-to-shock time y = survival rate

$\hat{y} = -9.2956x + 101.3285$ $r = -0.96$ (this is a strong negative linear correlation)



Slope: For every 1 minute increase in mean call-to-shock time, there is a decrease of 9.2956 percent decrease is survival rate.

***y*-intercept:** When x is 0 minutes, the predicted y-value is 101.3285 percent.

Notice that we did not simply report a number for the slope. Slope is $\frac{\Delta y}{\Delta x}$ from algebra, so it is the change in y for every 1 unit change in x. We have to format our answers very particularly for the AP test:

- **Slope: ALWAYS** use the words **"predicted average"** for free response questions, as in: *for every one unit increase in x the "predicted average" increase/decrease in y is _____(units).*
- ***y*-intercept: ALWAYS** use the word **"predicted"** for free response, as in: *when x = 0 (units) the predicted y value is _____(units)*

It is important to realize that this is a **prediction line**. Our data points do not necessarily lie on this line, and any value we put in for x will only get us a predicted value for y. Predicting values outside of the data set can cause problems and requires analysis to decide whether or not it is an appropriate use of the prediction line:

- **Extrapolation:** Using a prediction line to project values *outside* the initial dataset.

Example 3: Use the information and your work from example 2 to find the following:

a) Predict the SCA survival rate for a community with a mean call-to-shock time of 5 minutes.

$$\hat{y} = -9.2956(5) + 101.3285 = 54.850365$$

When the call-to-shock time is 5 minutes, the predicted average survival rate is 54.85 percent.

b) Should this line be used to predict the SCA survival rate for a community with a mean call-to-shock time of 20 minutes? Why/why not?

If we were to put in 20 minutes for x, we would get a negative value for y. Since we cannot have a less than 0 percent survival rate, we cannot extrapolate this far from this dataset.

c) Does the y-intercept represent a possible real world value for this situation?

No, it does not – you cannot have a call-to-shock time of 0 (care would have to be given at the moment the SCA occurred – which is impossible); in addition, it is impossible to have a survival rate above 100 percent, so this does not represent a real world value.

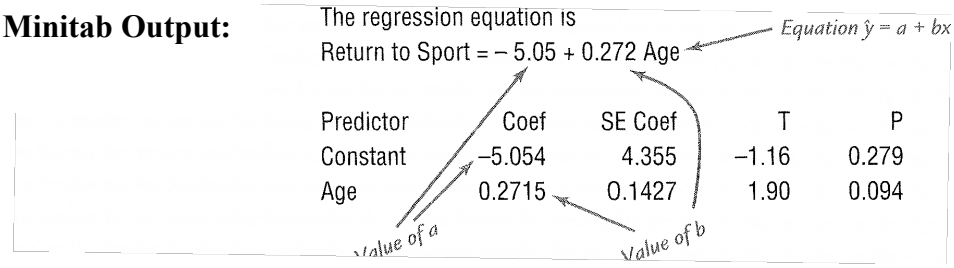
Example 4: How quickly can athletes return to their sport following injuries requiring surgery? We are given the following data on x = age and y = days after arthroscopic shoulder surgery before being able to return to their sport, for 10 weight lifters:

x	33	31	32	28	33	26	34	32	28	27
y	6	4	4	1	3	3	4	2	3	2

Find the least square regression line, r, and the scatterplot. Interpret the slope and the y-intercept in the context of this problem.

Take a look at the following summary. This is called a MINITAB output. MINITAB is a very popular and widely used statistical program.

Partial Minitab Output:



Note that in this particular example, we are using $\hat{y} = a + bx$. In addition, instead of using variables, the MINITAB output used the actual words that the variables would have represented. In fact, it is not uncommon to even see the "hat" notation over a word:

$$\widehat{\text{Return to Sport}} = -5.05 + 0.272\ \text{Age}$$

Sometimes, they will stretch the hat out to cover the whole word, other times it will just be a small hat centered over a word.

Example 5: A random sample of moving times (in minutes) and weights (in pounds) were recorded for 20 moving jobs requiring three-man crews, and the results of the regression analysis are shown below. The equation for the LSRL is

Predictor	Coef	StDev	T	P
Constant	21.84	25.54	0.86	0.404
Weight	0.036538	0.002977	12.27	0.000

S = 30.32 R – Sq = 89.3% R – Sq(adj) = 88.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	138434	138434	150.60	0.000
Residual Error	18	16546	919		
Total	19	154980			

(a) $\hat{\text{Weight}} = 21.84 + 0.037\,(\text{Time})$

(b) $\hat{\text{Time}} = 21.84 + 0.037\,(\text{Weight})$

(c) $\hat{\text{Weight}} = 25.54 + 0.003(\text{Time})$

(d) $\hat{\text{Time}} = 25.54 + 0.003(\text{Weight})$

(e) $\hat{\text{Time}} = 0.037 + 21.84\,(\text{Weight})$

There are a few more key facts to know about least squares regression lines:

- **Slope (*a*):** Slope can be found by the following equation as well:

$$a = r\left(\frac{s_y}{s_x}\right)$$

  - r is the correlation between x and y.
  - s_y is the standard deviation of the response variable (y)
  - s_x is the standard deviation of the explanatory variable (x)

- The least squares regression line always passes through the point $(\bar{x}, \bar{y})$
  - That means that the mean x and the mean y for the datasets lies on the regression line.

- r^2 is the **coefficient of determination.**
  - It is simply the *correlation coefficient* (r) squared.
  - This tells you the proportion of the variation of the response variable **explained by** the explanatory variable.

Example 5: Using the MINITAB output and line generated in example 4, find the value of r and r^2. Explain the meaning of each of these in context.

Predictor	Coef	StDev	T	P
Constant	21.84	25.54	0.86	0.404
Weight	0.036538	0.002977	12.27	0.000

S = 30.32 R – Sq = 89.3% R – Sq(adj) = 88.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	138434	138434	150.60	0.000
Residual Error	18	16546	919		
Total	19	154980			

$$\widehat{\text{time}} = 21.84 + 0.036538(\text{weight})$$

You may have notice on the output "R – Sq" (circled above). This is the r^2 value.

$$r^2 = 0.893$$
$$r = \sqrt{0.893} = 0.945$$

This is a **strong, positive, linear correlation** (because of r)

89.3% of the change in **time** ($\hat{y}$) is predicted by weight (x).

Example 6: The heart disease death rates per 100,000 people in the United States for certain years, as reported by the National Center for Health Statistics, were

Year:	1950	1960	1970	1975	1980
Death rate:	307.6	286.2	253.6	217.8	202.0

Which of the following is the correct interpretation of the coefficient of determination?

a) The heart disease rate per 100,000 people has been dropping on average of 3.627 per year.
b) The baseline heart disease rate is 7386.87
c) The regression line explains 96.28% of the variation in heart disease rates over the years.
d) The regression explains 98.12% of the variation in heart disease rates over the years.
e) Heart disease will be cured in the year 2036.

Answer: The first thing you have to do is enter your list and do a linear regression.

That gives us an $r^2 = 0.9628$, so the answer is c), 96.28% of the variation is explained by the explanatory variable. a) is incorrect because it is an interpretation of slope, b) is an interpretation of the *y*-intercept, d) incorrectly uses *r* instead of r^2, and e) is an incorrect extrapolation.

Summary:

- **Explanatory variable:** The *x* variable, also called the independent or predictor variable
- **Response variable:** The *y* variable, also called the dependent variable.
- **Extrapolation:** Using a prediction line to project values <u>*outside*</u> the initial dataset.
  - You will often be asked the appropriateness of extrapolation.
- Get used to your calculator and how it works to do linear regressions.
- **Slope and Intercept:** Make sure to follow the AP format for interpreting these:
  - **Slope: ALWAYS** use the words **"predicted average"** for free response questions, as in: *for every one unit increase in x the "predicted average" increase/decrease in y is _____(units).*
  - ***y*-intercept: ALWAYS** use the word **"predicted"** for free response, as in: *when x = 0 (units) the predicted y value is _____(units)*
- **Slope** of the linear regression is also $a = r\left(\frac{s_y}{s_x}\right)$
- $(\overline{x}, \overline{y})$ is on the least squares regression line
- r^2 tells you the proportion of the variation of the response variable **explained by** the explanatory variable.

Checkpoint 2.3

Multiple Choice

1. If you're attempting to predict a value of the response variable using a value of x that is outside the range of observed x values in your data set, you're conducting a process of:

   (a) predicting the slope of the regression line.
   (b) interpolation.
   (c) computing residuals.
   (d) extrapolation.
   (e) slope interpretation.

Free Response

1. (1999 Q1) Lydia and Bob were searching the internet to find information on air travel in the United States. They found data on the number of commercial aircraft in the United States during the years 1990 - 1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The computer output for their regression are given below.

Predictor	Coef	Stdev	t-ratio	p
Constant	2939.93	20.55	143.09	0.000
Years	233.517	4.316	54.11	0.000
s = 33.43				

(a) What is the value of the slope of the least squares regression line? Interpret the slope in the context of this situation.
(b) What is the value of the intercept of the least squares regression line? Interpret the intercept in the context of this situation.
(c) What is the predicted number of commercial aircraft flying in 1992?

2.3 Homework

1. Randomly selected alumni from many universities were asked if they agreed that their college education was worth the expense, and a portion of the data is displayed below **(Gallup-Purdue Index 2015 Report)**. One variable is the percentage of students who *strongly agreed* with the statement above, and the other variable is the ranking of the college as determined by *U.S. News and World Report*.

Ranking	Percentage of Alumni Who Strongly Agree
28	53
29	58
30	62
37	55
45	54
47	62
52	55
54	62
57	70
60	58
65	66
66	55
72	65
75	58
82	67
88	59
98	75

   a. Find the value of r and the least squares regression line that would allow you to predict the percentage of alumni who would strongly agree that their education was worth the cost, using ranking as the independent variable. What proportion of the variation is explained by the independent variable?
   b. Predict the percentage of alumni who would agree that their education was worth the cost for a university with a ranking of 50.
   c. Explain why it may not be a good idea to use this linear function to predict the percentage of alumni who would strongly agree that their education was worth the expense for a university with a ranking of 10.

2. There are various instruments designed to measure a person's ability to exhale – this measurement is incredibly useful as a diagnostic tool for identifying certain lung disorders. Two of these are the Wright Meter and the Mini-Wright Meter. The Wright Meter is generally considered to provide a better measure of air flow, while the Mini-Wright Meter is more portable and significantly easier to use. 17 subjects participated in a study where their exhales were measured once on each device. The data are recorded in the table below:

Mini-Wright Meter	Wright Meter	Mini-Wright Meter	Wright Meter
512	494	445	433
430	395	432	417
520	516	620	656
428	434	260	267
500	476	477	478
600	557	259	178
364	413	350	423
380	442	451	427
658	650		

a. If these two types of meters produce different results, but the results are highly correlated, it would be possible to use a reading from a Mini-Wright Meter to predict the reading from a Wright-Meter. Determine the correlation coefficient, r, for the data in the table above, and write the equation for the least squares regression line for this relationship (use the Mini-Wright Meter reading for x and the Wright Meter reading for y). Does the value of r seem to indicate that it would be reasonable to use this linear function to predict values for the Wright Meter or not? Explain briefly.
b. What proportion of the variation is explained by the independent variable?
c. What would you predict for the Wright Meter if the Mini-Wright Meter reads 525? If it reads 400?

3. A sample of 548 randomly selected students from Massachusetts were followed over a 19 month period from 1995 to 1997 in a study of the relationship between TV viewing and eating habits **(*Pediatrics* [2003]: 1321-1326)**. For each additional hour of television viewed each day, the number of fruit and vegetable servings was found to decrease on average 0.14 servings.
a. For this study, what is the independent variable? What is the dependent variable?
b. Would the least squares regression line for predicting the number of servings of fruits and vegetables from the hours of TV viewed have a positive or a negative slope? Explain.
c. If $r = 0.64$ for this study, what proportion of the variation is explained by the independent variable?
d. If the average number of servings of fruits and vegetables was 4.45 for 1 hours of TV viewed, write the least squares regression line for this relationship.

4. Many studies have shown that people who suffer from a cardiac arrest event have a significantly better chance of survival if a defibrillator shock is administered soon after the cardiac arrest event. The question of the relationship between survival rate and time until a defibrillator shock is administered was explored in the paper **"Improving Survival from Sudden Cardiac Arrest: The Role of Home Defibrillators" (by J.K. Stross, University of Michigan, February 2002).** The data below give x = the mean time before administering the defibrillator shock (in minutes) and y = the survival rate (expressed as a percent). The data was collected at a cardiac rehabilitation center where cardiac arrests occurred while the patient was hospitalized, so response times tended to be short.

Mean time to Defibrillator Use, *x*:	2	6	7	9	12
Survival Rate, *y*:	90	45	30	5	2

a. Construct a scatterplot for these data. Describe the relationship between these two variables.
b. Find the equation of the least-squares regression line, and find the values of r and r^2.
c. What portion of the survival rate can be attributed to the mean time to defibrillator use?
d. Use your regression line to predict the survival rate for a group with a mean time to defibrillator use of 10 minutes.
e. Use your regression line to predict the survival rate for a group with a mean time to defibrillator use of 0.10 minutes and 11 minutes. What are the problems with extrapolating to these values? What might that indicate about the use of a linear regression?

2.4 Residuals and Residual Plots

Objectives:

- Calculate residuals and make a residual plot.
- Determine if regression model is a good fit for data.
- Interpret, in context, the correlation coefficient and the coefficient of determination.
- Interpret s_e , the standard error.
- Analyze the impact of influential points on a regression

As we noted in an earlier section, sometimes the *r* value can be close to ± 1 and our data still might not be linear. One of the ways that we could tell that the scatterplot was not linear was by looking at the data and seeing if it looked like it was curving rather than linear.

Of course, our eyes are not always reliable tools for this, so we need another tool. This tool is a **residual**.

- **Residual:** the difference between an observed *y* value and a predicted *y* value ($\hat{y}$)

$$\text{residual} = y_i - \hat{y}_i$$

  - Observed *y* value just means the actual *y* value from the dataset.
  - Predicted *y* value is the $\hat{y}$ value you get when you plug in the *x* corresponding with the observed *y* value (i.e. the *y* value that your regression line predicts for a given *x* value).
- Essentially, *residuals* are just how far off of the prediction line your data points are.

Example 1: You are given the following set of observations for variables *x* and *y*. Use this and your calculator to create a linear regression line. Then find each residual.

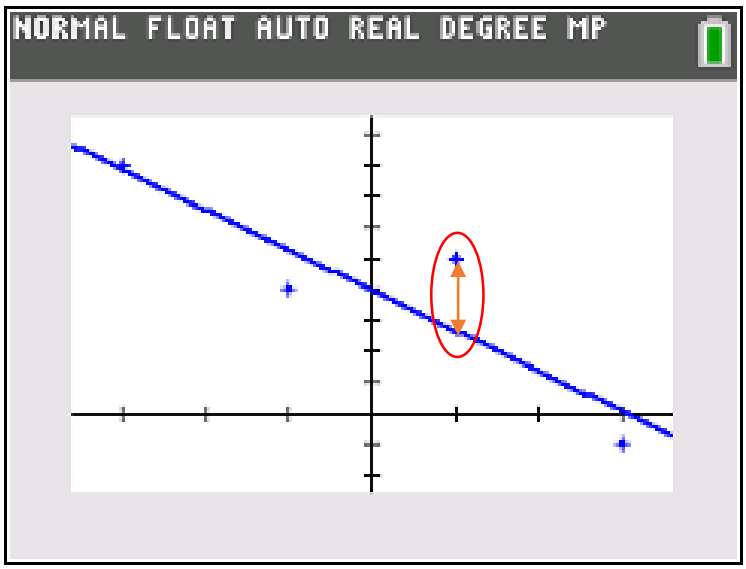
x	-3	-1	1	3
y	8	4	5	-1

$\hat{y} = -1.3x + 4$

Simply plug in each *x* and get a $\hat{y}$. The difference is the value of the residual.

x	-3	-1	1	3
y	8	4	5	-1
$\hat{y}$	7.9	5.3	2.7	0.1
Resid.	0.1	-1.3	2.3	-1.1

The positive residuals lie above the regression line, and the negative ones lie below it.



The residual for $x = 1$ is indicated by the double arrow in the picture (circled above).

Example 2: Consider the data on x = height (in inches) and y = weight (in pounds) for American females, age 30 – 39.

x	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
y	113	115	118	121	124	128	131	134	137	141	145	150	153	159	164

Find the least squares regression line and use it to calculate the following:

The residual for a height of 63 inches, and the residual for a height of 70 inches.

Regression line: $\hat{y} = 3.596x - 98.235$

To find the residuals, simply plug in the x values to find your $\hat{y}$ value and subtract it from the value corresponding to that x on the table. (Using the TABLE function on the calculator to find predicted y values from the equation you have stored as y_1).

Residual_{63} = 128 pounds – 128.34 pounds = –0.34 pounds

Residual_{70} = 153 pounds – 153.52 pounds = –0.52 pounds

Since both of these residuals are negative, we could conclude that they are below the line.

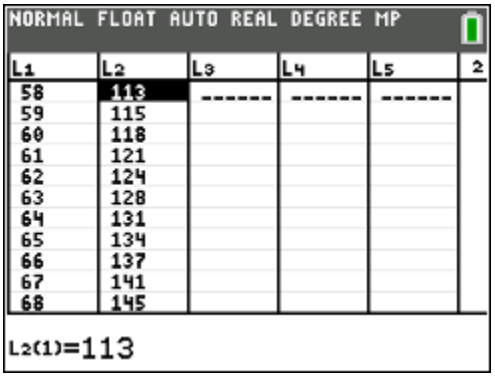
One of the ways to tell if a linear model is a good fit is to create a **residual plot**.

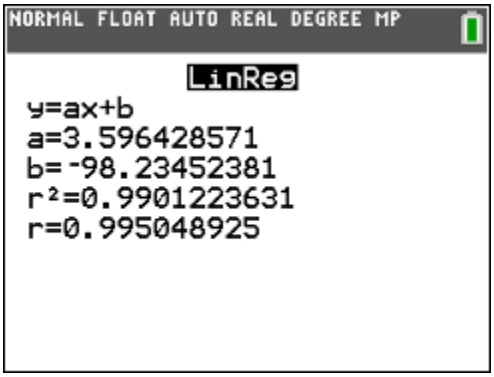
- **Residual Plot**: a scatterplot of the explanatory variable, x, and the residual values.
  - $(x, \text{residual})$
  - If the *residual plot* is scattered (has no pattern), then the model that created the residual is a good fitting model.
  - If the *residual plot* has a pattern, then there is a better regression model than the one you used to create the residual.

Of course, to create all those residuals by hand would be unpleasant, so we use our calculator to do it quickly. There are a couple of ways to do this. The easiest one is to just use a STAT PLOT and our lists. In order for this to work, however, I have to have already used two lists to create a regression line.

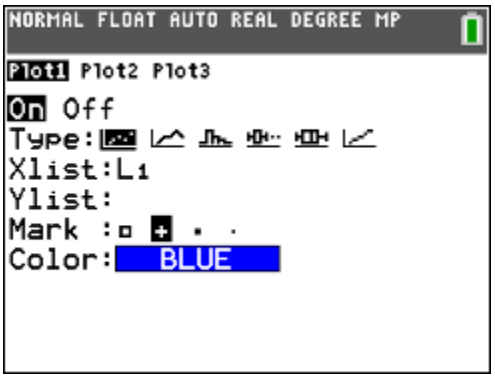
Example 3: Create a residual plot using the data from Example 2.

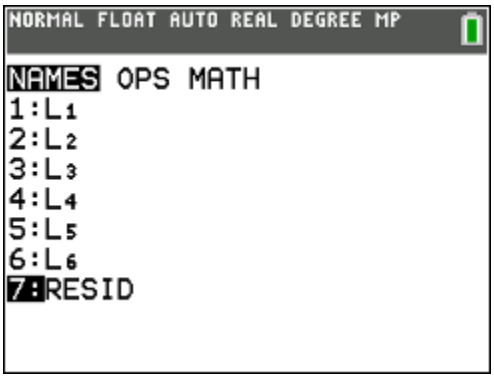
Since we already have the lists entered, and have run the regression, we should already have seen these:



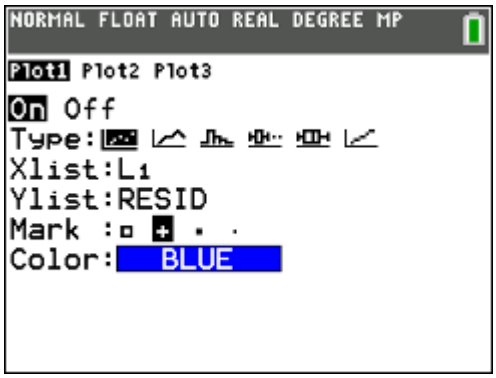


Now go to your STAT PLOT menu and create the residual. Select L_1 for your Xlist, and when you go to the list menu, you should see Residual as an option at the bottom of the list:

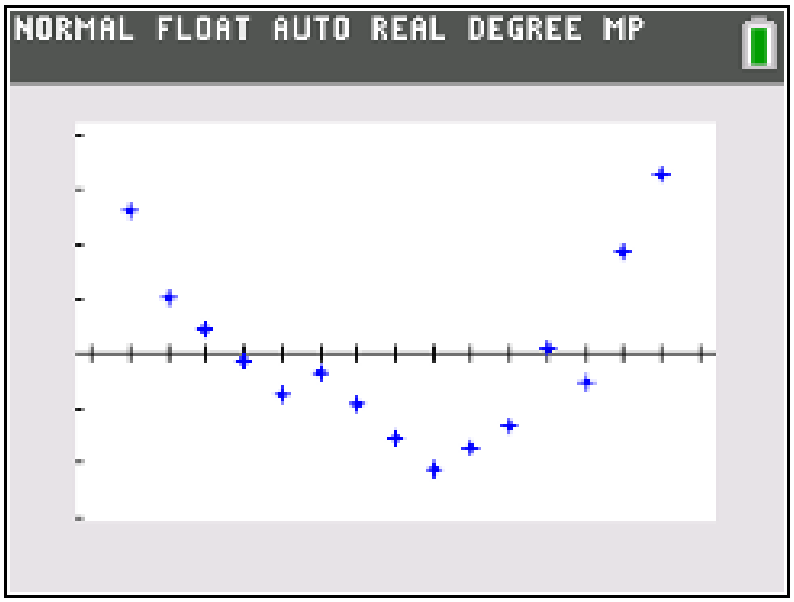




Now you should see that represented in your STAT PLOT menu, then just do your standard ZoomStat, and you have your residual plot:



And our residual plot…



There definitely seems to be a pattern here (it's not perfect, but it looks somewhat like a parabolic pattern). This would indicate **a linear model is not a good fit**, despite r being very close to 1 ($r = 0.995$, if you recall).

Example 4: One measure of success of knee surgery is post-surgical range of motion for the knee joint. Post-surgical range of motion was recorded for 12 patients who had surgery following a knee dislocation. The age of each patient was also recorded.

The data are given in the chart, and here is a partial MINITAB output:

Patient	Age (x)	Range of Motion (y)
1	35	154
2	24	142
3	40	137
4	31	133
5	28	122
6	25	126
7	26	135
8	16	135
9	14	108
10	20	120
11	21	127
12	30	122

Regression Analysis

The regression equation is

Range of motion = 108 + 0.871(Age)

Predictor	Coef	StDev	T	P
Constant	107.58	11.12	9.67	0.000
Age	0.8710	0.4146	2.10	0.062

s = 10.42 R-Sq = 30.6% R-Sq(adj) = 23.7%

Find the equation for the least squares regression line.

Use your calculator to construct a scatterplot and a residual plot.

Is this regression line a reasonable fit for the data? Explain why or why not.

Patient	Age (x)	Range of Motion (y)
1	35	154
2	24	142
3	40	137
4	31	133
5	28	122
6	25	126
7	26	135
8	16	135
9	14	108
10	20	120
11	21	127
12	30	122

Regression Analysis

The regression equation is

Range of motion = 108 + 0.871(Age)

Predictor	Coef	StDev	T	P
Constant	107.58	11.12	9.67	0.000
Age	0.8710	0.4146	2.10	0.062

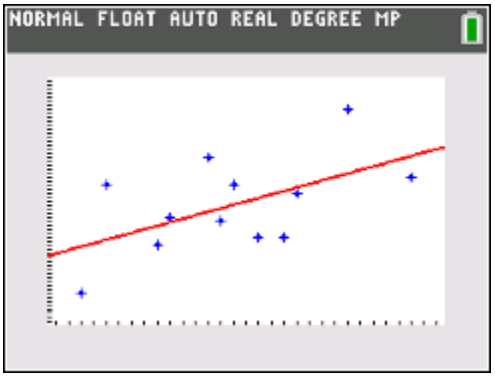
s = 10.42 R-Sq = 30.6% R-Sq(adj) = 23.7%

Find the equation for the least squares regression line.

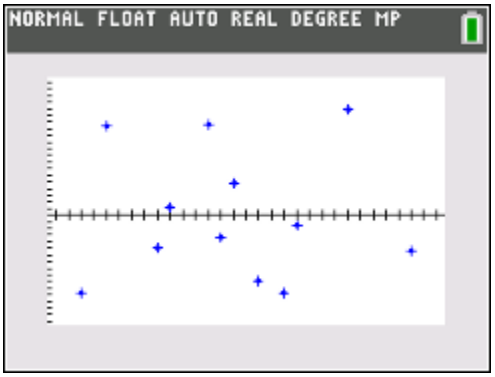
The regression line is given in the MINITAB: $\widehat{\text{range of motion}} = 108 + 0.871\ (\text{age})$

Use your calculator to construct a residual plot.

I put the data in L_1 and L_2, so my residual plot uses the L_1 and Resid lists:



Scatterplot



Residual Plot

(I included the regression line on the scatterplot to help with interpreting the reasonableness of the fit)

Is this regression line a reasonable fit for the data? Explain why or why not.

The scatterplot looks like a moderate positive linear correlation, and $r = 0.553$ confirms this. There is no pattern in the residual, so the least squares regression seems like a reasonable fit for the data. Only 30.6% of the response variable is predicted by the explanatory variable, but because there is no pattern in the residual, it is likely the best that we can do.

How to determine if a linear model is a good fit:

You should consider the following 4 factors in your decision:

1. Does the scatterplot look linear?
2. r should be close to 1 or –1 for a good fit.
3. The residual plot should be scattered for a good fit.
4. s_e (standard deviation of the residuals) should be small for a good fit.

The **standard deviation of the residual** (also referred to as the standard deviation about the least squares regression line) is denoted by s_e. We want s_e to be small compared to the values in the data table (though it is often difficult to interpret this). Even if r^2 is close to 1 or –1, we want the s_e to be small. In the MINITAB output, s_e is just denoted as s.

Regression Analysis

The regression equation is

Range of motion = 108 + 0.871(Age)

Predictor	Coef	StDev	T	P
Constant	107.58	11.12	9.67	0.000
Age	0.8710	0.4146	2.10	0.062

s = 10.42 R-Sq = 30.6% R-Sq(adj) = 23.7%

Let's look at another MINITAB output that we saw in the last section:

Predictor	Coef	StDev	T	P
Constant	21.84	25.54	0.86	0.404
Weight	0.036538	0.002977	12.27	0.000

S = 30.32 R – Sq = 89.3% R – Sq(adj) = 88.7%

Analysis of Variance

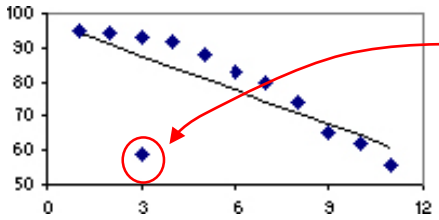
Source	DF	SS	MS	F	P
Regression	1	138434	138434	150.60	0.000
Residual Error	18	16546	919		
Total	19	154980			

The regression line for this data was $\widehat{\text{time}} = 21.84 + 0.036538(\text{weight})$, and the $r^2 = 0.893$, with an $r = 0.945$. This would make us think that it has a strong, positive, linear correlation.

But the value of $s_e = 30.32$ – this seems large, but as we do not know our dataset, we cannot actually conclude that. What if our weights were in tens of thousands of pounds, and our times were therefore in the hundreds to thousands of minutes? Then this standard error might be small. We wouldn't be able to tell without actually looking at the residuals, however.

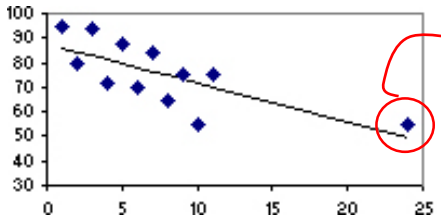
Points that are far away from the rest of the data can have a large and distorting effect on the least squares regression line. The two types of points that concern us are **outliers** and **high leverage points**.

- **Outlier:** a data point separated from the rest of the dataset; that is, a data point with a large residual.
  - Outliers tend to have a distorting effect on r and r^2.



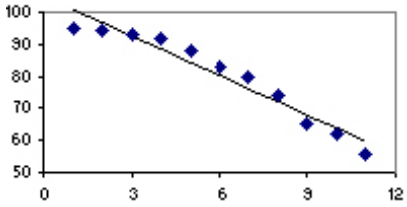
Outlier – notice how large the residual is compared to the rest of the data?

- **High-Leverage Point:** a data point has an x-value substantially larger or smaller than the other observations have.
  - They cause distortions in a and b (slope and y-intercept of the regression line)



High-Leverage Point – notice how it tends to pull the slope of the line up?

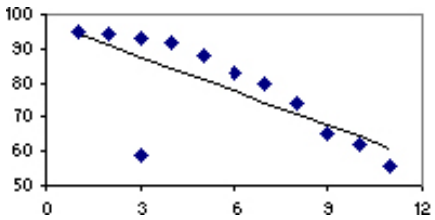
Take a look at this scatterplot:



Regression equation: $\hat{y} = 104.78 - 4.10x$

Coefficient of determination: $r^2 = 0.94$

Now here is what will happen if we add in an outlier:

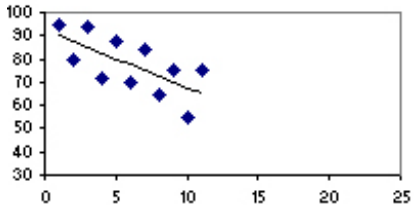


Regression equation: $\hat{y} = 97.51 - 3.32x$

Coefficient of determination: $r^2 = 0.55$

Notice how much the r^2 value dropped? From 0.94 to 0.55 – that is the impact of an outlier.

Now take a look at this scatterplot:

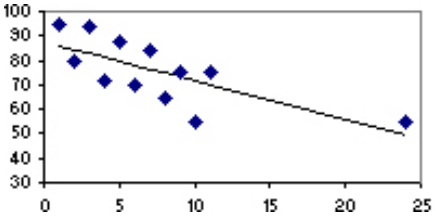


Regression equation: $\hat{y} = 92.54 - 2.5x$

Slope = –2.5

Coefficient of determination: $r^2 = 0.46$

And if we add a high leverage point:



Regression equation: $\hat{y} = 87.59 - 1.6x$

Slope = –1.6

Coefficient of determination: $r^2 = 0.52$

Notice the change in the slope? From –2.5 to –1.6.

Any point in a least squares regression that, if it was removed, would change the relationship between the explanatory and response variables significantly is called an **Influential Point**. An influential point will cause a much different slope, *y*-intercept, and/or correlation coefficient.

Outliers and *high-leverage points* are often influential points.

Summary:

- **Residual:** the difference between an observed *y* value and a predicted *y* value ($\hat{y}$)

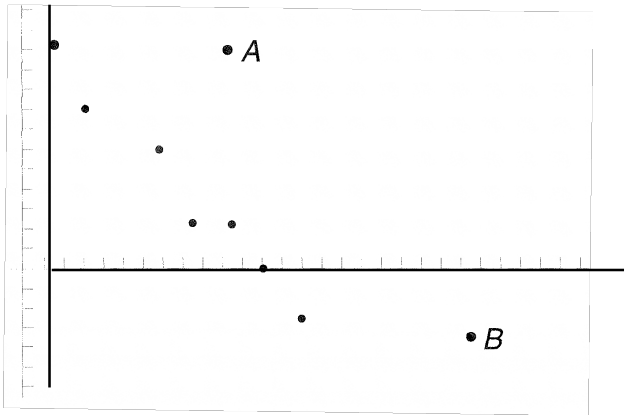
  $$\text{residual} = y_i - \hat{y}_i$$

- **Residual Plot**: a scatterplot of the explanatory variable, *x*, and the residual values.
- **A Linear Model is a good fit** if it meets the following criteria:
  1. Does the scatterplot look linear?
  2. *r* should be close to 1 or –1 for a good fit.
  3. The residual plot should be scattered for a good fit.
  4. s_e (standard deviation of the residuals) should be small for a good fit.
- **Outlier:** a data point separated from the rest of the dataset
- **High-Leverage Point:** a data point has an *x*-value substantially larger or smaller than the other observations have.
- **Influential Point:** Any point in a least squares regression that, if it was removed, would change the relationship between the explanatory and response variables significantly. *Outliers* and *High-Leverage Points* are often influential points.

Checkpoint 2.4

Multiple Choice

1. In the graph below how would the regression line computed with point B (not including point A), differ from the regression line using the original data points (excluding points A and B)?



(a) The y intercept of the line with point B would be greater.
(b) The r^2 value of the line with point B would be larger.
(c) The slope of the line with point B would be less steep than the slope of the line without point B.
(d) The y intercept of the line with point B would be zero.
(e) None of the above.

2. A study found correlation $r = 0.61$ between the sex of a worker and his or her income. You conclude that

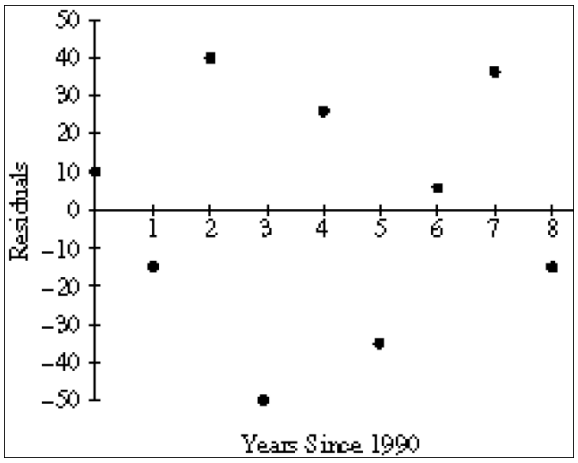
(a) women earn more than men on the average.
(b) women earn less than men on the average.
(c) an arithmetic mistake was made; this is not a possible value of r.
(d) this is nonsense because r makes no sense here.
(e) the correlation should have been $r = -0.61$.

3. Which of the following statements is/are true?
   I. Correlation and regression require explanatory and response variables.
   II. Scatterplots require that both variables be quantitative.
   III. Every least squares regression line passes through $(\overline{x}, \overline{y})$

(a) I, II (b) I, III (c) II, III (d) I, II, and III (e) None of these

Free Response

1. (1999 Q1) Lydia and Bob were searching the internet to find information on air travel in the United States. They found data on the number of commercial aircraft in the United States during the years 1990 - 1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.



Predictor	Coef	Stdev	t-ratio	p
Constant	2939.93	20.55	143.09	0.000
Years	233.517	4.316	54.11	0.000
s = 33.43				

(a) Is a line an appropriate model to use for these data? What information tells you this?

(b) What is the value of the slope of the least squares regression line? Interpret the slope in the context of this situation.

(c) What is the value of the intercept of the least squares regression line? Interpret the intercept in the context of this situation.

(d) What is the predicted number of commercial aircraft flying in 1992?

(e) What was the actual number of commercial aircraft flying in 1992?

2.4 Homework

1. An article on the cost of housing in California that appeared in the ***San Luis Obispo Tribune*** **(March 30, 2001)** included the following statement: "In Northern California, people from the San Francisco Bay Area pushed into the Central Valley, benefiting on average $4000 for every mile traveled east of the Bay Area."
    a. If this statement is correct, what is the slope of the least-squares regression line?
    b. If the median Bay Area home price was used as the value for this equation when a distance of 0 miles was used, and the median home price at that time was $450,000, write the least-squares regression line for this situation (y = home price and x = distance from the Bay Area).
    c. Calculate the value predicted by this regression line for a home that is 30 miles from the Bay Area. If the actual data point used to create the regression had a cost of $315,500, calculate the residual value for that data point.
    d. A researcher looked at the plot of the residuals and found that there was no pattern to the residual – what does this tend to indicate about the linear trend?

2. The data in the table below is similar to a table from the paper **"Six-Minute Walk Test in Children and Adolescents" (*The Journal of Pediatrics* [2007]: 395-399)**. Two hundred eighty boys completed a test that measured how far they could walk in 6 minutes (on a flat, hard surface). The table below shows the median walk distance for each age group (represented by a "representative age" number on the table).

Representative Age (years)	4.0	7.0	10.0	13.5	17.0
Median Walk Distance (meters)	544.3	584.0	667.3	701.1	727.6

    a. Using x = Representative Age and y = Median Walk Distance, construct a scatterplot by hand and on the calculator. Does the pattern appear linear?
    b. Use your calculator to find the least-squares line that describes the relationship between median walk distance and representative age. What is the r value for the relationship? What percent of the variation in walk distance is attributable to representative age?
    c. Calculate the five residuals and make a residual plot by hand and using your calculator. Are there any unusual features of this residual plot?
    d. What is the y-intercept for your regression line? Does this value have any meaning in the real world? Explain.
    e. Do you think that you could use your line to extrapolate a median walk distance for someone who is 45? Would you be able to calculate a residual for that value? Explain.

3. A certain type of algae is known to have the potential to damage river ecosystems. The data below on density of an algal colony of this type is for a set of nine rivers. The explanatory variable is rock surface area (x) and the response variable is algae colony density (y).

x	50	55	50	79	44	37	70	45	49
y	152	48	22	35	38	171	13	185	25

   a. Create a scatterplot and find the least squares regression line for this data.
   b. What is the value of r^2 for this data set? What percentage of the change in algae colony density is explained by rock surface area?
   c. What colony density would your regression line predict for a rock surface area of 50?
   d. What are the residual values for the two data points, (50, 152) and (50, 22)?
   e. How would you describe the linear relationship between the rock surface area and algae colony density?
   f. Based on your scatterplot, do any of the points appear to be influential? Does it appear to be an outlier or a high-leverage point?
   g. Calculate a least squares regression line without the influential point(s) identified in part f. How do the values of the slope and y-intercept compare to the values in the regression line from part a?

4. The chemical acrylamide is a product of frying foods, and there had been some suspicions that it was a potential carcinogen (extremely high doses in animals have been shown to cause cancer, but these levels are much higher than human exposure, and there is no conclusive evidence that there is a link in humans – as a result, the FDA still has guidelines for safe consumption of acrylamide). In 2012, a research team created a statistical model to estimate the acrylamide concentrations in French fries. The data below are approximated values from their study. This study investigated x = frying time (in seconds) and y = acrylamide concentration (in μg acrylamide per kg of French fries).

Frying Time	150	240	240	270	300	300
Acrylamide Concentration	155	120	190	185	140	270

a. Create a scatterplot and least squares regression line.
b. What are the values of r and r^2? How would you describe the relationship based on this?
c. What does your model predict for a frying time of 240 seconds? Calculate the residuals for both points with a frying time of 240 seconds, (240, 120) and (240, 190).
d. Create a residual plot of this data. Does the plot indicate that a linear model might be a good fit or not?

2.5 Analyzing Departures from Linearity

Objectives:

- Perform a power regression.
- Perform transformation regressions.

Often, the data we collect does not follow a strictly linear pattern – in fact, false assumptions of linearity are one of the most common mistaken statistical assumptions. It is an easy mistake to make – if something is good, for example, wouldn't twice as much be better?

Take for example, soaking in a hot tub. If the temperature of the water is 95°F, it's pretty nice. At 103°F you are having a very comfortable soak.

If the relationship were truly linear, increasing the temperature would increase your enjoyment of the hot tub. Obviously, that's not true (even 10 to 20 more degrees could potentially cause serious injuries – thirty seconds of exposure to 130° water will cause 3rd degree burns). Not every data set is linear.

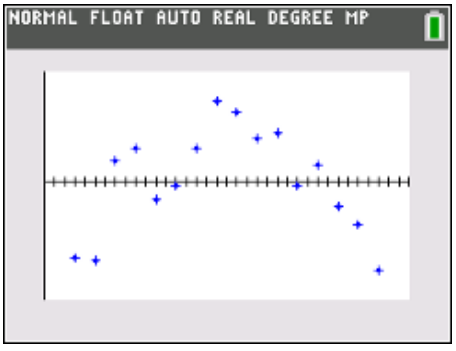
We will look at two common methods of fitting non-linear relationships: power regressions and transformations:

Polynomial (power) Regression:

Example 1: The focus of many agricultural experiments is to study how the yield of a crop varies with the time at which it is harvested. Accompanying data is given where the variables are x = time between flowering and harvesting (days) and y = yield of paddy, a field where rice is farmed (in kilograms per hectare):

Plot the data on your calculator; the residual plot is shown below:

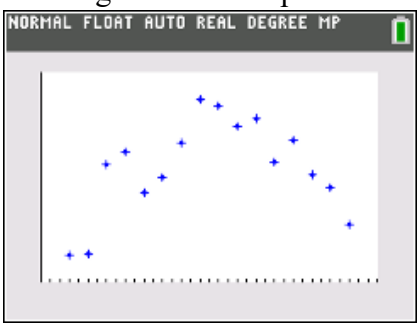
x	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46
y	2508	2518	3304	3423	3057	3190	3500	3883	3823	3646	3708	3333	3517	3214	3103	2776



Does a linear relationship seem to be a reasonable fit?

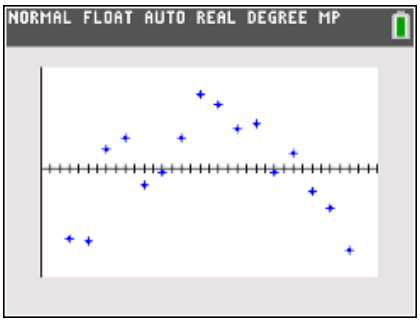
What kind of fit seems more reasonable?

1. Looking at the scatterplot that we create on our calculator, it does not look linear:



2. The r value is not close to 1 ($r = 0.274$ from our regression).

3. We have a pattern in the residual; in fact, it is very similar to the initial pattern we see in the scatterplot.

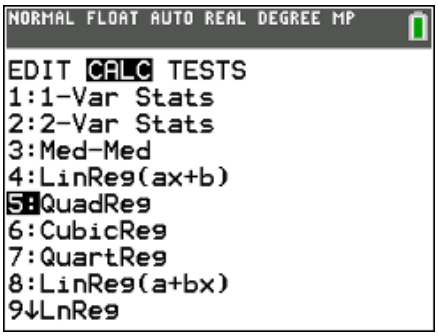


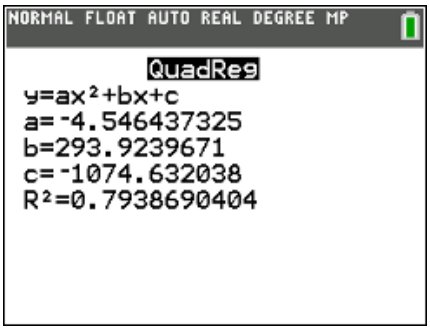
4. $S_e = 9.52$, not terribly large compared to the residual list, but all the other information makes it seem like a linear regression is not a good fit.

Therefore, a linear relationship does not appear to be a good fit.

This looks much more like a parabola, so we should see if a quadratic function fits the data better.

In fact, if we just go to the STAT button on the calculator again, and move to the CALC menu, we can see "5:QuadReg" – which is a quadratic regression.

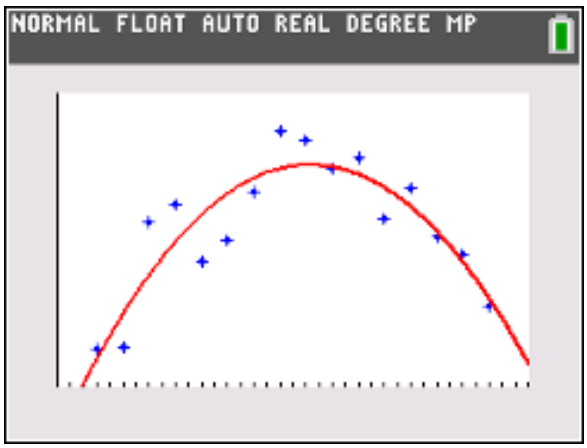




This means that we have a quadratic regression of $\hat{y} = -4.546x^2 + 293.924x - 1074.632$

The calculator gave an $r^2 = 0.794$, which would indicate 79.4% of the variation in y can be explained by x using this model.

This would give an $r = 0.891$, indicating a strong quadratic relationship. The graph seems to fit the data much better as well:

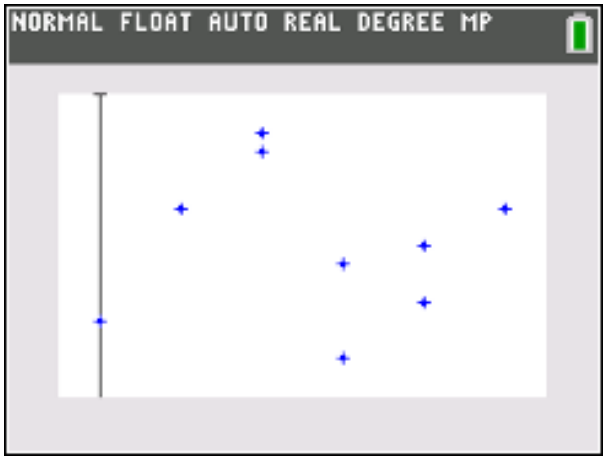


Example 2: Researchers have examined a number of climatic variables in an attempt to understand the mechanisms that govern rainfall runoff. A study examined the relationship between x = cloud cover index and y = sunshine index. Suppose that the cloud cover index can have values between 0 and 1. Consider the accompanying data.

Plot the data. Does a linear relationship seem to be a reasonable fit? Why or why not? What kind of fit seems more reasonable?

X	Y
0.2	10.98
0.5	10.94
0.3	10.91
0.1	10.94
0.2	10.97
0.4	10.89
0.0	10.88
0.4	10.92
0.3	10.86

If we look at the scatterplot of the data, we see it definitely does not look linear.



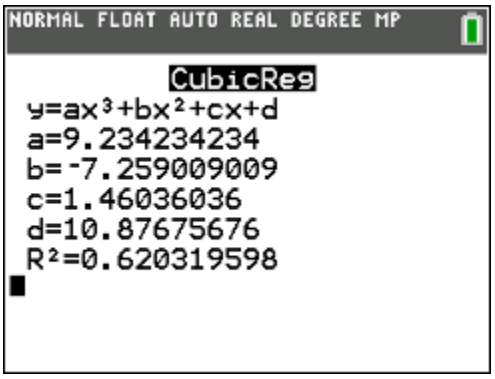
The plot has a distinct "up-and-down" look, so I would think that a cubic regression might work. You could do that on your calculator, and I have also included a MINITAB output for this:

Polynomial Regression

The regression equation is

y = 10.8768 + 1.46036x − 7.25901x**2 + 9.23423x**3

S = 0.0315265 R-Sq = 62.0% R-Sq (adj) = 39.3%



Use the cubic regression equation to predict the sunshine index for a day when the cloud cover index is 0.45.

Transformation Regression:

Sometimes we don't actually do a regression based on another function, instead we will *transform* the data. There are a lot of different transformations we can do – we could square or square root, log or natural log, we could even do sinusoidal transformations.

On the AP test **IF** they even ask this question, they will tell you what transformation to do and then ask if that transformation seems to be a good choice.

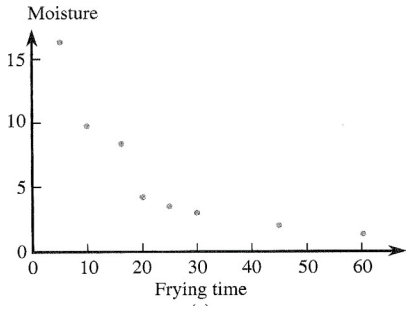
When you perform a transformation, it is a good fit if we meet the following criteria:

- The data looks more linear.
- A residual of the new plot has **no** pattern.

Example 3: No tortilla chip lover likes soggy chips, so it is important to find characteristics of the production process that produce chips with an appealing texture. The following data on x = frying time (in seconds) and y = moisture content (%) are provided.

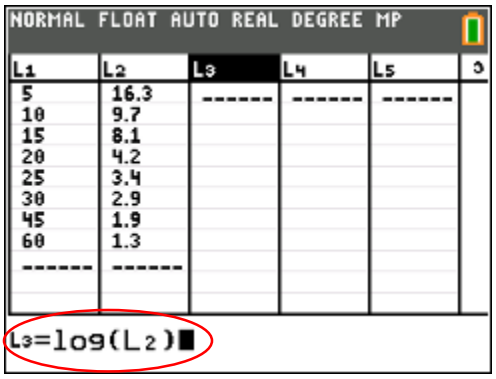
Frying time, x	5	10	15	20	25	30	45	60
Moisture content, y	16.3	9.7	8.1	4.2	3.4	2.9	1.9	1.3

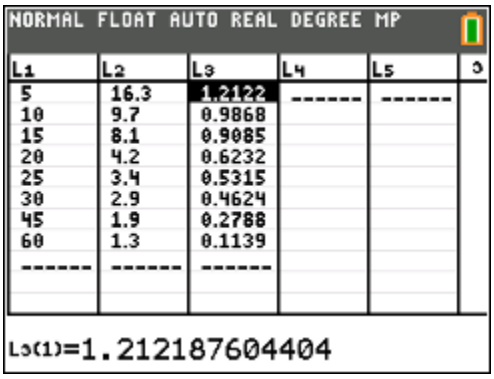
The scatterplot of the data is shown below:



Since there is a pretty clear pattern to the data, a linear model does not seem reasonable for this. It looks like an exponential function, so we might try a logarithmic transformation.

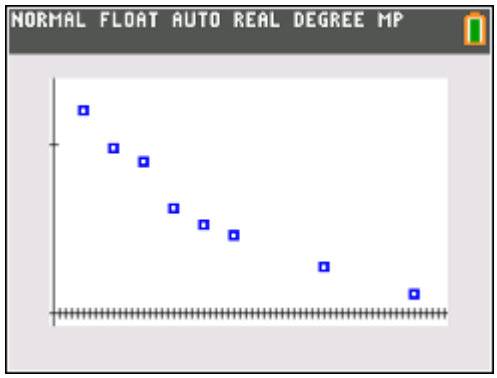
What we should do to achieve that is go back to the lists we entered in our calculator. Go to the top of L_3 and type in $\log(L_2)$ as shown below:





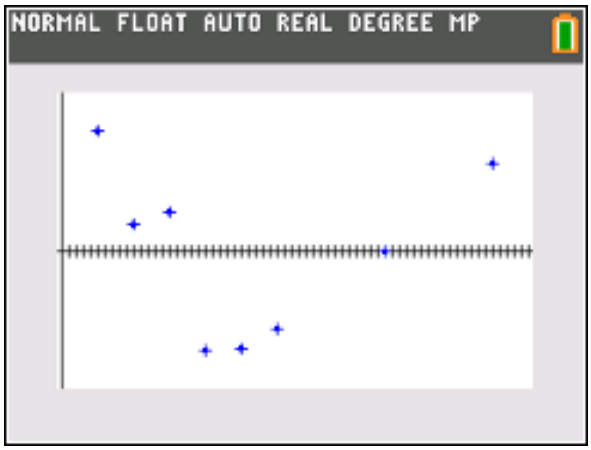
Click enter after doing this, and you will see L_3 populate with the log of the data from L_2.

Doing a scatterplot of L_1 and L_3, and we see the following:



That still looks pretty curved, so let's try a residual plot to see if this data has a pattern to it.

Running a linear regression on our calculator, we find $r = -0.95$, but we should look at our residual plot to be sure.



This still looks like it has a pattern, so this was probably not the best transformation.

Below is the MINITAB and graph of the logged data:

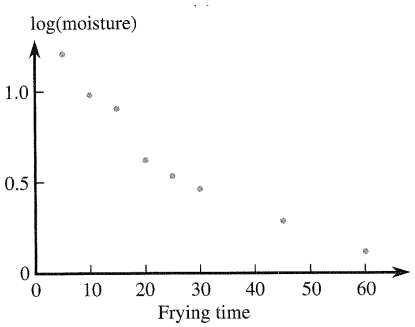
The regression equation is
log(moisture) = 1.14 − 0.0192 frying time

Predictor	Coef	StDev	T	P
Constant	1.14287	0.08016	14.26	0.000
frying t	−0.019170	0.002551	−7.52	0.000

S = 0.1246 R-Sq = 90.4% R-Sq(adj) = 88.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.87736	0.87736	56.48	0.000
Residual Error	6	0.09320	0.01553		
Total	7	0.97057			



If we wrote out this as a transformed line, we would get the following:

$$\log(\hat{y}) = -0.0192x + 1.14$$

We could transform this equation by taking both sides to the power of 10:

$$\hat{y} = 10^{-0.0192x+1.14}$$

Let's use this to predict the moisture content at $x = 35$ seconds:

$$\hat{y} = 10^{-0.0192(35)+1.14} = 2.694\%$$

But, as we said, this wasn't a great model because of the pattern in the residual. (It turns out that if we did log(log(y)), we would have had a transformation that was much better).

Example 4: A response variable appears to be exponentially related to the explanatory variable. The natural logarithm of each y-value is taken and the least-squares regression line is found to be $\ln(\hat{y}) = 1.64 - 0.88x$. Rounded to two decimal places, what is the predicted value of $x = 3.1$?

a) –1.09
b) –0.34
c) 0.34
d) 0.082
e) 1.09

Summary:

- If data has a pattern, we try to do a **regression based on the function it looks like**.
  - We can do quadratic, cubic, and quartic regressions – these are *power regressions*
- **Transformations:** we can transform the explanatory variable, the response variable, or both with any number of different functions to try to get our data to appear linear.
  - This is an effort to find the correct relationship between the data.
  - The AP Test will tell us what transformations to do, then we have to interpret if the transformed model fits well.
- A **transformed regression** is a good fit if it meets the following criteria:
  - The data looks more linear.
  - A residual of the new plot has **no** pattern.

Checkpoint 2.5

Multiple Choice

1. Using least-squares regression, I determine that the logarithm (base 10) of the population of a country is approximately described by this equation: **Log (population) = –13.5 + 0.01(year)**

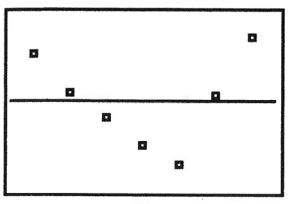
   Based on this equation, the population of the country in the year 2000 should be about
   (a) 6.5
   (b) 665
   (c) 2,000,000
   (d) 3,162,277
   (e) None of the above

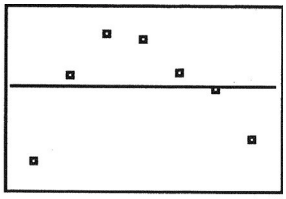
2. Suppose that the scatterplot of X and log Y shows a strong positive correlation close to 1. Which of the following is true?

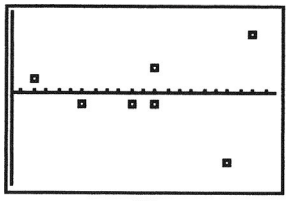
   I. The variables X and Y also have a correlation close to 1.
   II. A scatterplot of the variables X and Y shows a strong nonlinear pattern.
   III. The residual plot of the variables X and Y shows a random pattern.

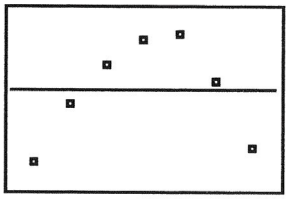
   (a) I only (b) II only (c) III only (d) I and II (e) I, II, and III

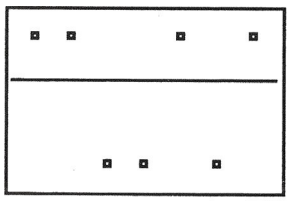
3. A researcher made a scatterplot from some previously collected data. The data was clearly nonlinear in shape. The researcher then tried a variety of transformations on the data in an attempt to linearize the results. The residual plot for each is shown below.


#1


#2


#3


#4


#5

Which of the transformations was best at linearizing the data?
(a) #1 (b) #2 (c) #3 (d) #4 (e) #5

4. A residual:

   (a) is the amount of variation explained by the least-squares regression line of y on x.
   (b) is how much an observed y value differs from a predicted y value.
   (c) predicts how well x explains y.
   (d) is the total variation of the data points.
   (e) should be smaller than the mean of y.

5. Which of the following would indicate the strongest relationship between two variables?
   (a) $r = 0.35$
   (b) $r = -.28$
   (c) $r = .21$
   (d) $r^2 = .01$
   (e) $r^2 = .23$

6. The coefficient of determination, r^2, between two variables is computed to be 81%. Which of the following statements must be true?

   (a) Large values of the explanatory variable correspond with large values of the response variable.
   (b) Large values of the explanatory variable correspond with small values of the response variable.
   (c) A cause and effect relationship exists between the explanatory and response variables.
   (d) There is a strong, positive, linear relationship between the explanatory and response variables.
   (e) Approximately 81% of the variability in the response variable is explained by regression on the explanatory variable.

7. If the model for the relationship between the score on the AP Statistics Exam (y) and the number of hours spent preparing for the test (x) was $\log \hat{y} = 0.1 + 1.9 \log x$, determine the residual if a student studied 9 hours and earned an 85.

   (a) 6.53
   (b) 3.14
   (c) 15.23
   (d) 0
   (e) -4.86

2.5 Homework

1. The paper, **"Aspects of Food Finding by Wintering Bald Eagles" (*The Auk* [1983]: 477-484)**, looked at the relationship between x = relative food availability and y = the time eagles spent flying in search of food (by the percentage of eagles soaring). The accompanying data is taken from that paper:

x	0	0	0.2	0.5	0.5	1.0	1.2	1.9	2.6	3.3	4.7	6.5
y	28.2	69.0	27.0	38.5	48.4	31.1	26.9	8.2	4.6	7.4	7.0	6.8

   a. Draw a scatterplot for this data. Would you describe the pattern you see in the plot as linear or as curved?
   b. A possible transformation that could lead to a linear pattern would be using $\sqrt{x}$ and $\sqrt{y}$. Construct a scatterplot for this transformation. Based on the scatterplot, would you say that this is a more reasonable choice for this data than the line in part a?
   c. What could be another potential transformation for x and/or y? Plot the data with your new transformation and assess whether it would be better than the transformation in part b.

2. The paper, **"Population Pressure and Agricultural Intensity" (*Annals of the American Geographers* [1977]: 384-386)**, reported a positive trend between x = population density and y = agricultural intensity for 18 different subtropical locations. The data are listed on the table below:

x	1.0	26.0	1.1	101.0	14.9	143.7	3.0	5.7	7.6
y	9	7	6	50	5	100	7	14	14

x	25.0	143.0	27.5	103.0	180.0	49.6	140.6	140.0	233.0
y	10	50	14	50	150	10	67	100	100

   a. Construct a scatterplot that uses x^2 and y. Does this plot look like the transformation resulted in a linear pattern?
   b. Draw a scatterplot that uses x and $\log(y)$. How does this plot compare to the one in part a? Does the plot look more or less linear?
   c. Now create a scatterplot using x^2 and $\log(y)$. Is this effective in creating a linear pattern in the plot?

3. The table below gives the number of heart transplants that occurred in the United States each year from 2006 to 2015 (numbered years 1 to 10, respectively, in the table below). The data is taken from the U.S. Department of Health and Human Services website.

Year	1	2	3	4	5	6	7	8	9	10
Number of Transplants	2193	2209	2163	2211	2332	2322	2378	2531	2655	2804

a. Construct a scatterplot for the data above and describe how the number of heart transplants has changed over the years 2006 to 2015.
b. Create a least-squares regression line that describes the relationship between x = year and y = number of transplants. Does the linear function seem like a good fit for this scatterplot?
c. Construct a residual plot for this data. Are there any features of the residual plot that indicate that there may be a transformation that could better describe the relationship?
d. Find a transformation of x and/or y that straightens this plot. Construct a scatterplot for your transformed variables.
e. Create a least-squares regression line for your transformation and use it to predict the number of heart surgeries in 2016.
f. The prediction that you made in part e involves predictions outside of the range of the x values for the data. What assumption must you be willing to make for this to be reasonable? Do you think that this assumption would still be valid for a prediction value for the year 2036 instead of 2016? Explain.

Unit 2 Practice Test