Linear Regression

Warm Up

1. Rank the following ten foods in order from the food you like most to the food you like least.

    1. Bagel
    2. Chocolate Chip Muffin
    3. Pasta
    4. Pizza
    5. French Fries
    6. Mozzarella Sticks
    7. Deli Sandwiches
    8. Hamburger/Cheeseburger
    9. Chicken Nuggets
    10. Chicken Sandwich

2. Record your preferences below:

    The number of the food I like the most is: _____

    The number of the food I like 2nd most is: _____

    The number of the food I like 3rd most is: _____

    The number of the food I like 4th most is: _____

    The number of the food I like 5th most is: _____

    The number of the food I like 6th most is: _____

    The number of the food I like 7th most is: _____

    The number of the food I like 8th most is: _____

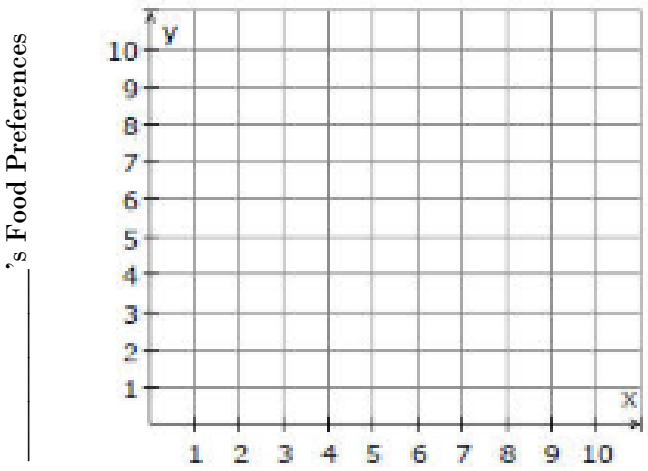
    The number of the food I like 9th most is: _____

    The number of the food I like the least is: _____

3. Find a partner. Write your responses for the 10 items above as ordered pairs. For example, if your 1st favorite food is pizza and your partner's favorite food is fries, then write the ordered pair (4, 5) which is (pizza, fries). Then go to the food you both like second most. You should have ten ordered pairs. Whoever uses their choice as the x-coordinate should remain the x-coordinate for all 10 ordered pairs.

List the ordered pairs below:

4. Plot your 10 points on the coordinates plane below.



_____'s Food Preferences

5. Analyze your data.

- The stronger the positive association, the more likely you and your partner would enjoy going out to eat together.
- The stronger the negative association, the less likely you and your partner would enjoy going out to eat together,
- If the association is week, then your agreement on dinner would be a hit or miss.

What conclusions can draw based upon your scatter plot?

Scatterplots

Say we wanted to know if there is a relationship between a person's shoe size and their height. We would need to look at both variables, shoe size and height. We could randomly select 50 people and make (x, y) points by using (shoe size, height), plot these 50 points. This would create a scatter plot.

Scatterplots may be the most common display for data. By just looking at them, you can see patterns, trends, relationships, and even the occasional outlier value sitting apart from the other.

Relationships between variables are often at the heart of what we'd like to learn from data:

- Are grades higher now than they used to be?
- Do people tend to reach puberty at a younger age then in previous generations?
- Does applying magnets to parts of the body relieve pain? If so, are stronger magnets more effective?
- Do students learn better with the use of computers?

Questions such as these relate two quantitative variables and ask whether there is an **association** between them. Scatterplots are the ideal way to picture such associations.

What should we look for in a scatterplot? We are going to look at direction, shape, strength and outliers. (SODS)

Direction:

Shape:

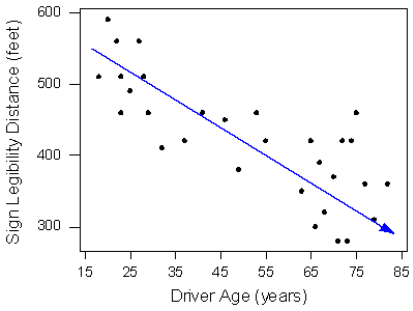
Strength:

Outliers:

Example 1: For each of the following scatterplots, describe the direction (positive or negative), form (linear or curved), and strength (strong, weak, or moderate) of each relationship in the context of the problem. Are there any outliers?

a.

b.



Variables:

What variable should go on the x-axis and which on the y-axis?

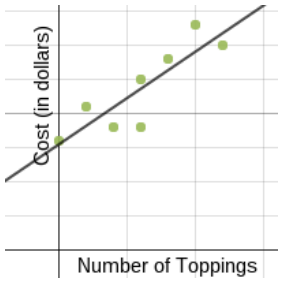
Response variable:

Explanatory variable:

Example 2: Identify the explanatory variable and the response variable for the following scatterplot that represents number of toppings and cost of a pizza. Then, write a sentence in the context of the problem that describes the pattern of change in the response variable as the explanatory variable increasese.

Explanatory Variable:

Response Variable:



Example 3: Use your calculator to create a scatter plot of the following data set. Then, use the scatter plot to describe direction, form and strength.

Hours Spent Studying	Exam Score
2	53
4.5	35
5	91
5	72
6	60
3	62
10	85
9.5	78
8	99

Correlation

Correlation coefficient (*r*).

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

Example: Let's look at the following data in order to calculate the correlation coefficient.

(4, -4) (3, -2) (0, 0) (-3, 2) (-4, 4)

Step 1: Make a scatter plot of your data on your calculator. Note its direction (positive or negative) and form (linear or curved).

Step 2: Calculate the mean and standard deviation of all the *x*'s and *y*'s.

$\bar{x} =$ _____ $s_x =$ _____ $\bar{y} =$ _____ $s_y =$ _____

Step 3: Calculate the z-score for each *x* and its corresponding *y*. Round each to two decimal places. Write your answers as points.

Step 4: Multiply the numbers in each of the points together. Add them up.

Step 5: Divide by n – 1 (the number of data values minus 1).

But, what does this mean?

Here is a useful list of facts about the correlation coefficients:

1. The sign of the correlation coefficient gives the direction of the linear association (think slope).

2. Correlation is always between -1 and 1. The closer r is to 1 or -1 the stronger the association, the closer to 0 the weaker the association.

3. A correlation of 1 or -1 rarely happens with real data because it would mean that all the data points fall exactly on a single straight line.

4. Correlation is commutative. The correlation of x with y is the same as the correlation of y with x. This means you will get the same correlation coefficient no matter what variable you assign to x or to y.

5. Correlation has no units, you are using standardized values to calculate it, which means there are no units.

6. Correlation measures the strength and direction of a *linear* association between two variables. This is why we always make a picture first. If your data is curved, r is not going to help you.

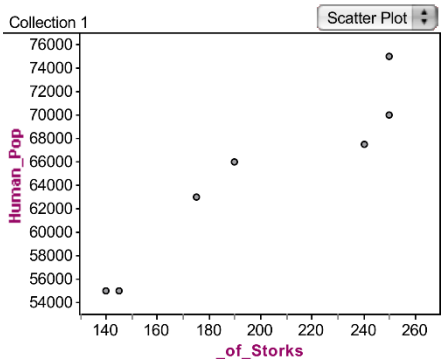
7. Correlation is sensitive to outliers. A single outlier can make a weak correlation stronger or a strong correlation weaker. As always, beware of outliers.

How strong is strong? You will usually see correlations described as weak, moderate, or strong, but be careful. There is no agreement among statisticians on what those terms mean. The same r-value might be considered strong in one context and weak in another. Using the words, weak, moderate or strong to describe a linear association can be useful additions to the numerical value that correlation provides. Be sure to include the correlation and show a scatterplot, so others can judge for themselves.

Correlation ≠ Causation

Whenever we have a strong correlation, it's tempting to try to explain it by imagining the explanatory variable has *caused* the response to change. Humans tend to see causes and effects in everything.

Example: A scatterplot of the human population (y) of Oldenburg, Germany, in the beginning of 1930 plotted against the number of storks nesting in the town (x) shows a tempting pattern.



The variables are obviously related to each other (the correlation is .97!), but that doesn't prove that storks bring more babies. It turns out that storks nest on house chimneys. More people mean more houses, more nesting sites, and so more storks. The causation is actually in the opposite direction, but you can't tell from the scatterplot or r-value. You need additional information – not just the data – to determine the real relationship.

Lurking Variable

A scatterplot on the damage (in dollars) caused to a house by fire would show a strong correlation with the number of firefighters at the scene. Surely damage doesn't cause firefighters to start appearing at fires. And firefighters do seem to cause damage, spraying water all around and chopping holes. Does that mean we should not call the fire department? Of course not. There is an underlying variable that leads to both more damage and more firefighters: the size of the blaze!!!

A hidden variable that stands behind a relationship is called a **lurking variable.** You can often debunk claims made about data by finding a lurking variable behind the scenes.

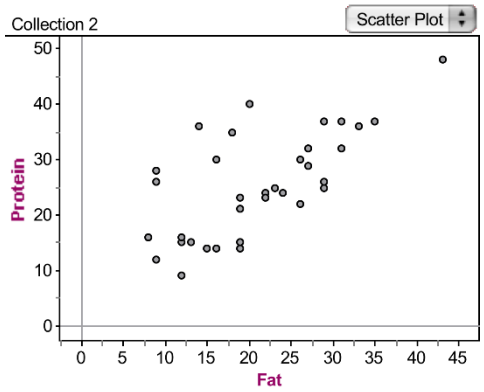
Example: Use your calculator to make a scatterplot of the data. Note its direction and form. Then, find the correlation coefficient. What does it tell you?

x	y
0	24.5
1	27.0
2	28.8
3	31.8
4	32.9
5	35.5

Linear Regression

The Big Mac has been one of McDonalds' signature sandwiches. One Big Mac provides 25 grams of protein – half the protein you would need in a day. It also supplies 550 calories and 29 grams of fat. That is 45% of the recommended daily intake of fat grams. So after eating a Big Mac the rest of your calories that day will need to be low fat!!!

Of course the Big Mac isn't the only item McDonalds sells. How are fat and protein related for the entire McDonald's menu? The scatterplot below of the Fat (in grams) versus the Protein (in grams) for foods sold at McDonalds.



It shows a positive, moderate (r = .61), linear relationship. If you want to consume only 15 grams of fat in your McDonalds lunch, how much protein would you consume? Now we need to model the relationship with a line and give its equation. The equation will let us predict the protein content for any McDonalds sandwich given the amount of fat it has. This equation is just the equation of a straight line that goes through the points, but no line can go through all the points. So what equation should we use?

Residuals

$$\text{residual} = \text{observed value} - \text{predicted value} = y - \hat{y}$$

<u>Example:</u> Suppose we come up the equation for the best fit line for the McDonald's data. This line allows us to predict the number protein grams given the number of fat grams in a McDonald's sandwich.

a. For a certain sandwich, the best fit line predicts 15 grams of protein but it actually has 17 grams. What's the residual?

b. For another sandwich the model predicts 40 grams of protein and we see that the residual is -8 grams. What's the actual protein content of this sandwich?

c. There's one McDonalds sandwich that has a residual of 0. Explain what that means.

The **line of best fit** is the line where the sum of the squared residuals is the smallest. We call this the **least squares regression line (LSR).** You might think finding this line is hard, surprisingly it is not. Below you will see the formula for finding the LSR.

$$\hat{y} = a + bx \qquad b = r\frac{s_y}{s_x} \qquad a = \bar{y} - b\bar{x}$$

Example: Nutritionists recommend that you have a good breakfast. They are especially concerned with "empty calories" in breakfast cereals. They recorded facts about 77 cereals, including their calories per serving and sugar content (in grams) per serving. The scatterplot that they plotted with Sugar as the explanatory and Calories as the response was positive, linear, and moderate.

The nutritionists calculated the following data. Remember, sugar is *x* and calories is *y*.

Sugar: $\bar{x} = 7$ grams $\quad s_x = 4.4$ grams

Calories: $\bar{y} = 107$ calories $\quad s_y = 19.5$ calories

Correlation: $r = 0.564$

a. Calculate the LSR line.

b. Interpret the slope and y-intercept in the context of the problem.

c. Does the y-intercept make sense?

Example: The linear model relating hurricanes' wind speeds (in knots) to their central pressure (in millibars) was $W = 955.27 - 0.897P$.

a. Hurricane Katrina had a central pressure measured at 920 millibars. What does our regression model predict for her maximum wind speed?

b. Katrina's actual wind speed was measured at 110 knots. How good was our prediction? (Hint: find the residual).

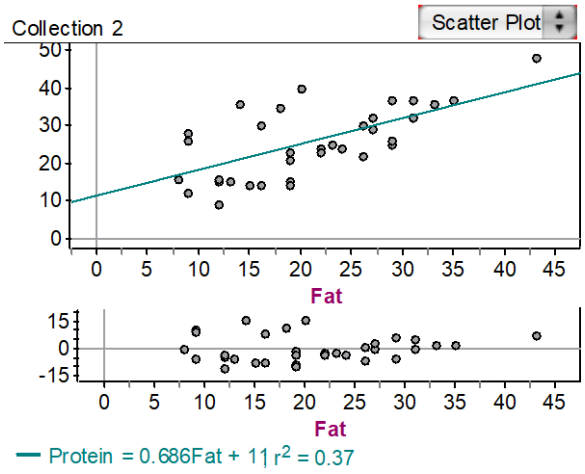
Example: Find the LSR line for the set of data to the right.

x	y
0	24.5
1	27.0
2	28.8
3	31.8
4	32.9
5	35.5

Residual Plots and Extrapolation

Residual Plot

A **residual plot** is a scatterplot with the points (x, residual) plotted. Below you will see the McDonald's data and the residual plot below it.



A residual plot should be the most boring scatterplot you've seen. It shouldn't have any interesting features, like direction or shape. It should have about the same amount of scatter throughout. It should show no bends and it should show no outliers. Our scatterplot above is a good boring scatterplot!!

Example: The table shows the change in tuition costs at Arizona State University during the 1990s.

a. Use your calculator to make a scatter plot of the data.

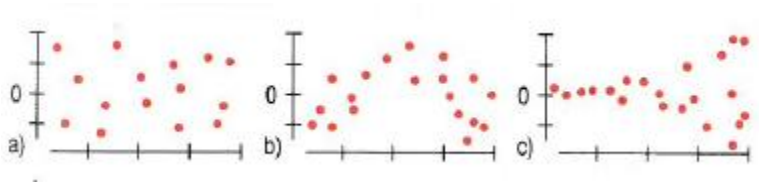
b. Use your calculator to find the LSR line.

Year	Tuition
0	6546
1	6996
2	6996
3	7350
4	7500
5	7978
6	8377

c. Use your calculator to find the correlation coefficient.

d. Make a residual plot. What does it tell you about the appropriateness of the linear model?

Example: Tell what each of the residual plots below indicates about the appropriateness of the linear model that was fit to the data.



Extrapolation

Association vs. Causation

Outliers and Influential Points

Outlier:

Influential Point:

Example 1: An emergency service wishes to see whether a relationship exists between the high outside temperature on a given day and the number of emergency calls it receives. They examine data from 10 randomly selected days last year. The data is as follows:

Temperature	74	82	88	67	93	99	101	78	85	90
# of Calls	4	8	10	8	11	14	13	6	8	10

a. Use your calculator to make a scatter plot of your data. Sketch a picture below including labels and scale.

b. Find the LSR line. State the equation below and draw the line on your scatterplot.

c. Find and interpret the value of *r*.

d. Circle the point on your scatterplot which represents the day in which the temperature was 67 degrees. Describe below how this point differs from the rest of the points in the plot.

e. Remove the point you circled from your calculator (delete both x and y) and calculate the regression line without the 67 degree day. Write the equation below, then draw this new line on your scatterplot in part (a). Label which line is for part (b) and which line is for part (e).

f. Find and interpret the new value of *r*.

Example 2: The number of passes completed and the total number of passing yards was recorded for NFL quarterback Brett Farve for each of the 16 regular season games in the fall of 2006. The data is shown below:

Completions	15	31	25	22	22	19	17	28	24	5	22	24	22	20	26	21
Yards	170	340	340	205	220	206	180	287	347	73	266	214	293	174	285	285

a. Use your calculator to make a scatter plot of your data. Sketch a picture below including labels and scale.

b. Find the LSR line. State the equation below and draw the line on your scatterplot.

c. Find and interpret the value of *r*.

d. Circle the point on your scatterplot which represents the game in which Brett Favre only had 5 completions. Describe below how this point differs from the rest of the points in the plot.

e. Remove the point you circled from your calculator (delete both x and y) and calculate the regression line without the 5 completion game. Write the equation below, then draw this new line on your scatterplot in part (a). Label which line is for part (b) and which line is for part (e).

f. Find and interpret the new value of *r*.

Example 3: In regression, an influential observation is one which has a great influence on the regression line. In each of the previous exercises, you removed a point that was different from the rest of the data set. Which point (the 67 degree day or the 5 day completion game) would you consider to be a more influential observation? Explain.