

Thursday, September 24: Chapter 7: Describing Bivariate Data with Scatterplots

The first step when analyzing the relationship between two _____ variables is to graph the data using a _____.

In a scatterplot, the _____ variable should be on the x-axis and the _____ variable should be on the y-axis. The explanatory variable seeks to explain or predict changes in the response variable. Usually the explanatory variable comes first chronologically. For example, when comparing SAT score and college GPA, SAT would be the explanatory variable since colleges use it to make predictions about how successful students will be in college.

Each axis should be clearly _____ with the variable's name and unit. It should also have a well marked and uniform _____ on each axis, however, the scales do not need to be the same.

The axes often intersect at (0, 0) but this can change depending on the range of the data sets. Patterns are often more visible when there is less "empty space" and the data is more spread out.

The 4 key features of a scatterplot are: _____

1. DIRECTION:

- _____: Higher values of one variable are associated with higher values of the other variable.

- _____: Higher values of one variable are associated with lower values of the other variable.

- _____: Higher values of one variable do not give any information about the values of the other variable. In other words, the two variables are independent.

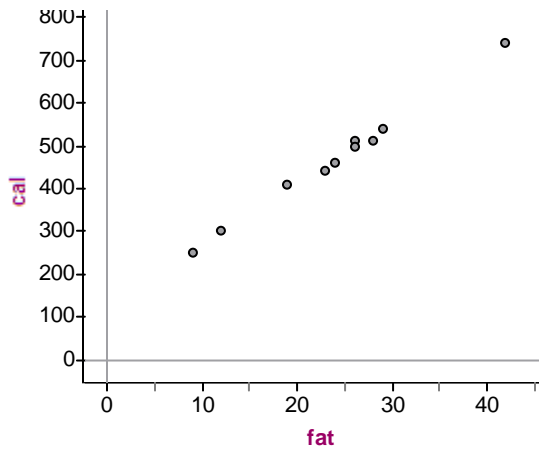
2. FORM:

3. STRENGTH (SCATTER)

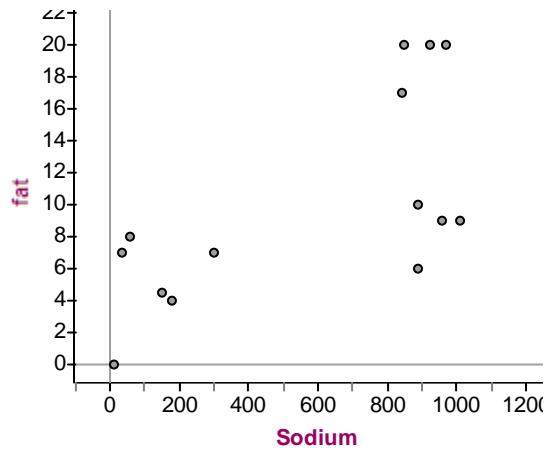
4. UNUSUAL VALUES:

- _____ that fall outside the pattern of the rest of the data and _____ of points that are isolated from the rest of the data. Always investigate these values!

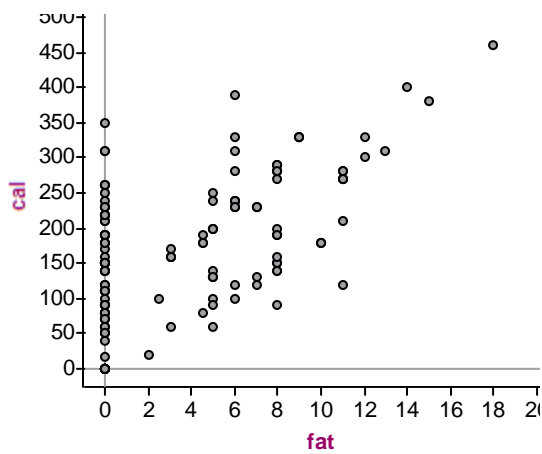
Describe the following scatterplots (from McDonald's Nutrition Facts):



Type = "beef"



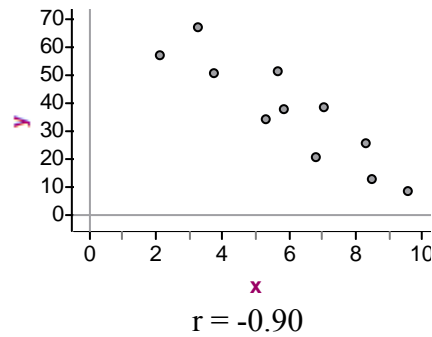
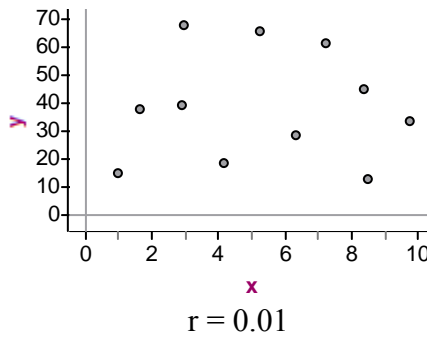
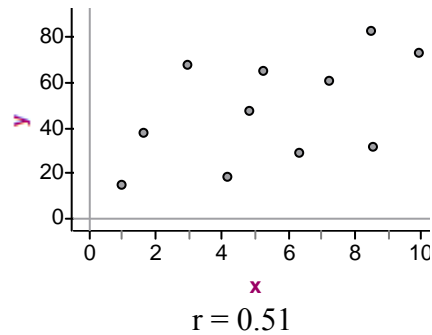
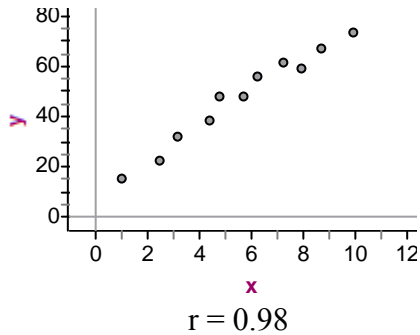
Type = "salad"



Type = "beverage"

A numerical way to help us quantify the amount of scatter in a scatterplot is to calculate the correlation coefficient, which measures the strength of the linear relationship between two quantitative variables.

The correlation coefficient, denoted r , will always be between $-1 \leq r \leq 1$. Here are some examples of scatterplots with their correlation coefficients:



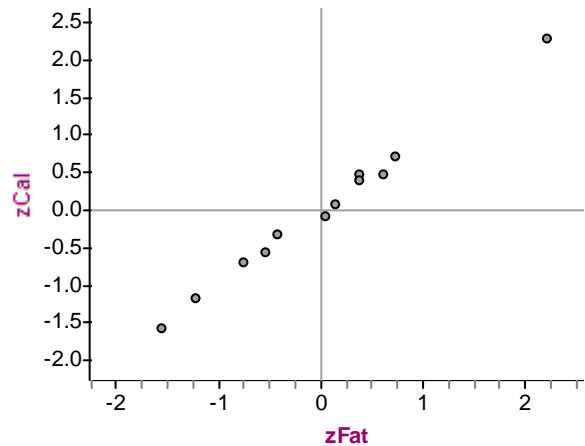
Calculating r :

Consider the scatterplot showing *calories* vs. *fat* for beef products at McDonalds. There seems to be a very strong association so the correlation would be very close to +1.

How does the relationship change if we standardized both variables?

$$\text{For fat, } z_x = \frac{x - \bar{x}}{s_x} = \frac{x - 22.67}{8.77}$$

$$\text{For calories, } z_y = \frac{y - \bar{y}}{s_y} = \frac{y - 450}{128}$$



Since having a majority of the points in quadrants I and III indicates a positive association, we will use the products $z_x z_y$ (which are always positive in QI and QIII and negative in QII and QIV) to help us calculate r .

To get an overall sense of the relationship, we add up these products: $\sum z_x z_y$. If the sum is positive, we have a positive association. If the sum is negative, we have a negative association. If there is no association, the sum should be close to 0. Also, since the size of this sum will get bigger the more data we have, we divide the sum by $n - 1$ to find the correlation coefficient:

$$r = \frac{\sum z_x z_y}{n - 1} = \frac{10.966}{12 - 1} = .997$$

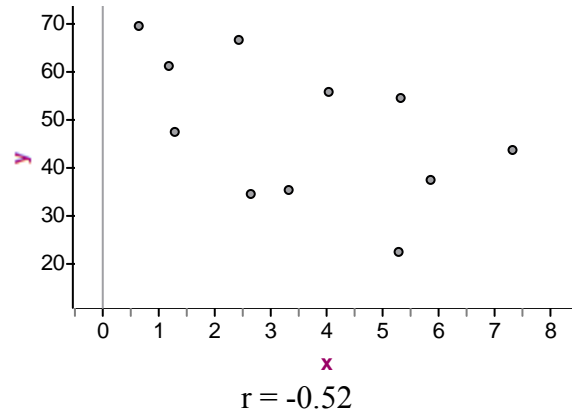
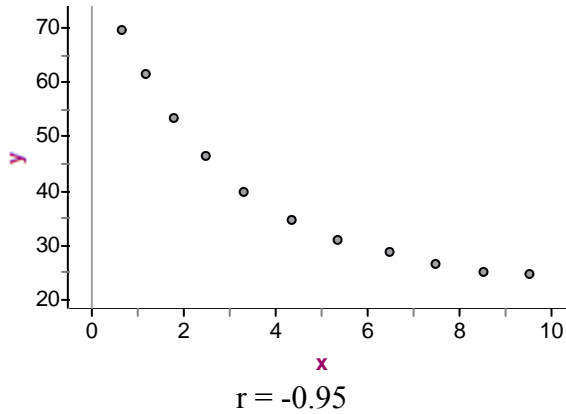
Note: When we use the word “correlation” in statistics, we are referring to the correlation coefficient. If you want to describe a relationship in a more casual way, use the word “association”.

Note: r is often called Pearson’s correlation coefficient

Properties of the Correlation Coefficient:

1. The value of r does not depend on the unit of measure since r is based on z-scores, which have no units. For example, the relationship between height and weight is equally strong if we use inches and pounds or centimeters and kilograms.
2. r has no units.
3. The value of r does not depend on which variable is x and which is y . The product $z_x z_y$ is the same as $z_y z_x$.
4. $-1 \leq r \leq 1$.
 - When $r > 0$, the relationship is positive.
 - When $r < 0$, the relationship is negative.
 - As $r \rightarrow \pm 1$, the relationship is stronger and has less scatter
 - As $r \rightarrow 0$, the relationship is weaker and has more scatter
5. $r = \pm 1$ only when the data are in a perfect line. This is the only case where the values of one variable can be completely determined by the values of the other variable.
6. The value of r is a measure of the strength of a _____ relationship. It measures how closely the data fall to a straight line. An r value near 0, however, does not imply that there is no relationship, only no linear relationship. For example, quadratic or sinusoidal data have an r close to 0, even though there may be a strong relationship present.

Also, even though r measures the strength of a linear relationship, it does NOT tell us if a linear model is appropriate. Only _____ can do that. The correlation coefficient just measures how much scatter there is from a line on a scale from -1 to 1 (whether or not it is appropriate to use a line to model the data).



Don't confuse correlation with cause-and-effect:

There is a strong positive association between monthly ice cream sales at Baskin Robbins and monthly drowning deaths. Should we close Baskin Robbins to save people from drowning?

Summary: You can never prove cause-and-effect from a scatterplot!!

Using the TI-83 to make scatterplots

- Enter data in L1 and L2
- Zoomstat
- Window
- Note: to sort bivariate data and keep the ordered pairs together, enter `SortA(L1, L2)`. This will sort the data by L1 and keep the pairs together.

Calculating r :

- *One time only*: Catalog: Diagnostic On: Enter: Enter
- Stat: Calc: 8: LinReg a + bx L1,L2

HW #18: SR (142-152), Required Reading (155-158), Problems page 160 (3, 5, 7, 11, 16, 23, 25, 26, 27, 29)

Monday, September 28: Chapter 10: Fitting a line to Bivariate data

When the form in a scatterplot is linear, we can use an equation in the form $\hat{y} = a + bx$ to model the relationship between the explanatory variable (x) and the response variable (y)

- a = y-intercept (constant)
- b = slope
- \hat{y} (“y hat”) signifies that the value of \hat{y} is an estimate or prediction
- Statisticians prefer the form $\hat{y} = a + bx$ instead of $\hat{y} = mx + b$, but they are equivalent.
- Some books use the notation $\hat{y} = b_0 + b_1x$

How can we find the best linear model? In other words, how can we know which line is “best?”

Since our goal is to make good predictions, we want to minimize the vertical deviations from the observations to the line. These vertical deviations are called prediction errors or _____.

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

The best fitting line is the line which minimizes the sum of the squared residuals, $\sum (y - \hat{y})^2$. This line is called the _____ (LSRL).

Applet: <http://www.rossmanchance.com/applets/Reg/index.html> (try this at home)

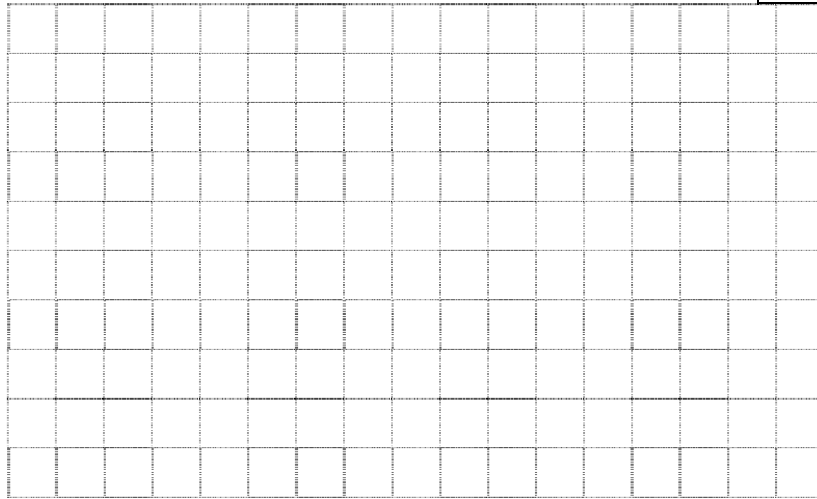
Using the TI-83 to calculate the LSRL:

- enter the data in L1 and L2
- stat: calc: 8: LinReg (a+bx) L1,L2 (Note: 4 and 8 are the same, just different forms)
- You should always use the TI83 to find the LSRL. Ignore any directions that say otherwise.

A random sample of 9 infants was taken and the age (in months) and height (in inches) of each infant was recorded:

- a. Sketch the scatterplot and describe what you see

age	height
1	20.5
1	19
2	21
4	22
4	23.5
6	22.5
7	23
7	24
9	26



- b. Calculate the value of the correlation coefficient. What does it tell you?
- c. Since the scatterplot shows a linear form, calculate the LSRL and graph it on the plot.
- d. Interpret the slope in the context of the problem

e. Interpret the y -intercept in the context of the problem

- f. If a child is 5 months old, how tall should he be, based on the model? In other words, how tall is an average 5 month old?
- g. Calculate the residual for a 5-month old child who is 23.9 inches tall. Interpret this value.
- h. Would you be willing to predict the height of a 10 year old child with this model? Why not?

Def: _____ is using your model to make predictions outside of the range of the data. It is very unreliable since the form of the data may not stay the same.

HW #19 SR (168-170), Problems page 189 (5, 13, 14, 21, 22, 31ab, 33)

Tuesday, September 29: Chapter 8: Assessing the Fit of a Line

After we find the Least Squares line, we should examine how well the model fits the data.

Important questions to consider are:

1. Is a linear model really appropriate, or would a curved model be better?
2. If we make predictions with the model, how accurate will our predictions be?
3. Are there any unusual aspects of the data set we need to consider before we make predictions with the model?

Question 1: Is the linear model appropriate?

Def: a _____ gives us a closer look at the pattern of the residuals. It plots the x-values on the x-axis and the residuals $(y - \hat{y})$ on the y-axis.

For the following data sets, sketch the original data with the least squares line. Then, sketch the residual plot and use it to decide if the linear model is appropriate.

x	y
1	1
2	5
3	7
4	8
5	12
6	12
7	17

x	y
1	.1
2	.8
3	2
4	3.3
5	5
6	7.3
7	9.9

Making residual plots on the TI-83:

L1 = x L2 = y L3 = \hat{y} (predicted values) L4 = $y - \hat{y}$ (residuals) Scatterplot L1, L4

For the second data set, the data is close to the line (not much scatter = high correlation) even though there is a obvious curve in the residual plot. The residual plot indicates that a line is not the best way to model this data. However, the lack of scatter means that the predictions using the linear model will still be fairly accurate within the range of our data, though not as accurate as with a curved model.

In conclusion, a residual plot will tell you if a linear model is the right type of model (has the right form) or if we should consider fitting a non-linear model. The correlation coefficient, however, tells us how much scatter there is from the LSRL, regardless of whether or not it is the best model.

NOTE: Sometimes a residual plot is made with the predicted values (\hat{y}) on the x-axis and the residuals on the y-axis. This is because computer software packages are built for multiple regression, which uses many different x-variables to predict y. Instead of using 1 of the x-variables and ignoring the others, they use the predicted values since they are a function of all the x's. However, the plot will still show the same characteristics and should be interpreted in the same way.

Question 2: How accurate will our predictions be?

Suppose that I randomly selected 10 students from CDO and recorded their weight (in pounds):
{103, 201, 125, 179, 150, 138, 181, 220, 113, 126}

If I were to randomly select one more student, what would be a good prediction for his or her weight?

Of course, this prediction is not likely to be correct.

Typically, how far are the observations from the mean? In other words, how far off should we expect to be?

Is there any way to improve our prediction? In other words, is there a way I can reduce the standard deviation?

Here are the heights (in inches) for the original 10 students:
{61, 68, 65, 69, 65, 61, 64, 72, 63, 62}

Sketch the scatterplot and calculate the LSRL using height as the explanatory variable.

Of course, the predictions using the regression line aren't perfect either.

Standard Deviation of the Residuals:

To get a sense of how close the points are to the line, we can calculate the standard deviation of the residuals, which gives an estimate of the average distance each observation is from the line (in other words, the average residual).

$$s = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{SS_{Resid}}{n-2}}$$

Note: “SS” = “Sum of Squares” so SS_{Resid} is the sum of squared residuals

Note: s is also called “root mean square error” (RMSE) or s_e (standard deviation of the errors).

Calculate the standard deviation of the residuals for this data:

Calculate the value of s for the two data sets on page 57:

HW #20: SR (178-179), problems page 189 (3, 4, 35 [skip part b but calculate and interpret s], 43 [also calculate and interpret s], 44 [also calculate and interpret s], 49)

Thursday, October 1: Chapters 8-9: R², Unusual Values, and Computer Output

Another way to evaluate how well a LSRL models a set of data is to consider the value of r^2 .

Def: The coefficient of determination, r^2 , is a measure of the proportion of variability in the y variable that can be accounted for by the linear relationship between x and y.

For example, at a pizzeria the price of a large pizza is a variable. That is, there is variability in the cost of a large pizza depending on how many items are included on the pizza. Suppose that large pizzas sell for \$8 plus \$1.50 per topping. If we were to plot the points (0, 8.00), (1, 9.50), (2, 11.00) they would fall *exactly* on a line. In this case, the number of toppings accounts for 100% (all) of the variability in price. Thus, $r^2 = 1$, or 100%.

Calculate the coefficient of determination for the height and weight data:

To measure the total variability in the y variable (weight), we measure the variability of y from its mean:

$$SSTotal = \sum (y - \bar{y})^2 =$$

- Note: We do not consider the x variable at all when we calculate SSTotal.
- Note: This is the same quantity that we use when we calculate s, the sample standard deviation for one variable (weight alone).

We can also measure the variability in y (weight) that still remains after we factor in x (height):

$$SSResid = \sum (y - \hat{y})^2 =$$

- Note: this is the same quantity that we used when we calculated s, the standard deviation of the residuals

The ratio of these two quantities, $\frac{SSResid}{SSTotal} =$ _____, describes the proportion of the variation in weight that is still unaccounted for after using height to predict weight. Using additional variables to predict weight, (e.g. waist size or gender) would make this ratio even smaller.

Thus, the proportion of the variation in weight that IS accounted for by weight is:

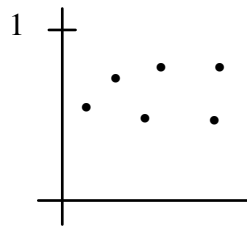
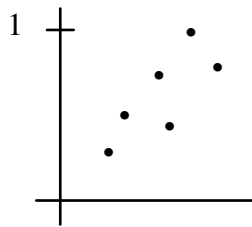
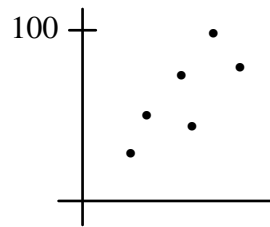
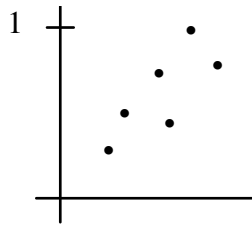
$$r^2 = 1 - \frac{SSResid}{SSTotal} =$$

So, the linear relationship between height and weight accounts for _____ % of the variability in weight.

Using the TI-83 to calculate r^2 :

What is the relationship between r^2 and s ?

- Both measure how well a line models the data
- r^2 has no unit and is usually expressed as a percent between 0% and 100%
- s is expressed in the same units as the response (y) variable



Question 3: Are there any unusual aspects of the data set we need to consider before we make predictions with the model?

Summary: Any point that stands apart from the others is called an _____. Since the LSRL must pass through the point (\bar{x}, \bar{y}) , points that are separated in the x-direction can be particularly _____. We say they have high _____.

When a point with high leverage lines up with the rest of the data, the LSRL won't change very much, but the correlation will be stronger.

When a point with high leverage does not line up with the rest of the data, it can have a large effect on both the line and the correlation. Note: Points with high leverage often do not have large residuals, since they pull the line close to them.

Points that are near \bar{x} will usually not be very influential.

Applet: <http://statweb.calpoly.edu/chance/applets/LRApplet.html>

Allows you to move points around to see changes in r, LSRL. Dynamic!

Understanding Computer Output:

On the AP Exam, questions about regression frequently include output from computer software, such as Minitab. Here is the output for the height and weight data we have been working with.

Regression Analysis: weight versus height

The regression equation is
weight = - 448 + 9.25 height

Predictor	Coef	SE Coef	T	P
Constant	-447.6	130.0	-3.44	0.009
height	9.250	1.997	4.63	0.002

S = 21.8772 R-Sq = 72.8% R-Sq(adj) = 69.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	10267	10267	21.45	0.002
Residual Error	8	3829	479		
Total	9	14096			

The least squares regression line is given at the top, but on many AP questions, they will delete this part. Make sure you can find the equation from the second table!

Other notes:

- To find the correlation coefficient (r), take the square root of R-Sq (+ or -)
- You can ignore RSquare Adj (this is a multiple regression topic)
- S is the standard deviation of the residuals
- The Analysis of Variance table has all of the Sums of Squares needed to calculate r^2 and s.

HW #21 SR (179-180, 183-185, 188), problems page 189 (7, 9, 23, 41, 47-interpret s also), page 215 (11, 13, 15)

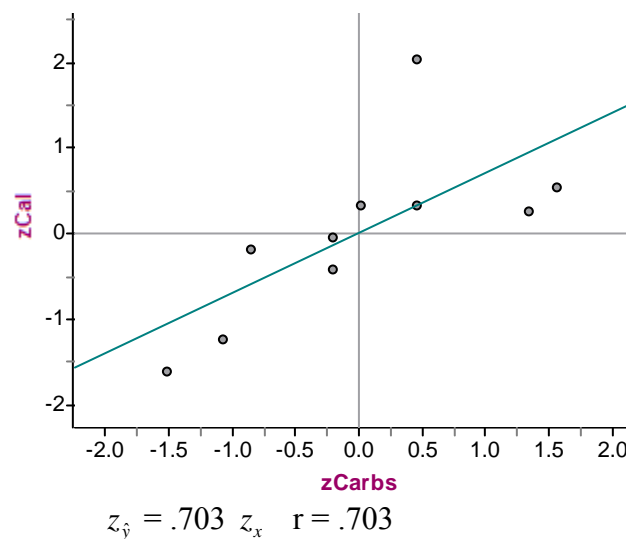
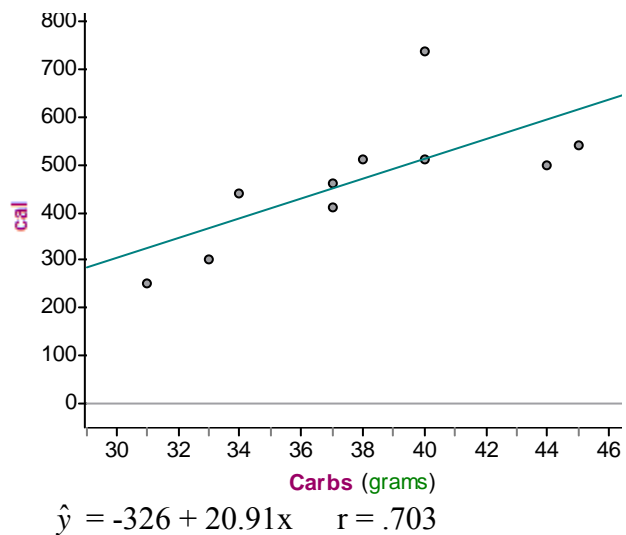
Monday, October 5: Chapter 8: Regression

Earlier we learned why the Least Squares Regression Line includes the words “Least Squares.” Today, we will learn why it includes the word “Regression.”

Thinking about the McDonald’s beef example, if we had a burger with an average amount of carbs, it would be reasonable to predict that it would also have an average number of calories.

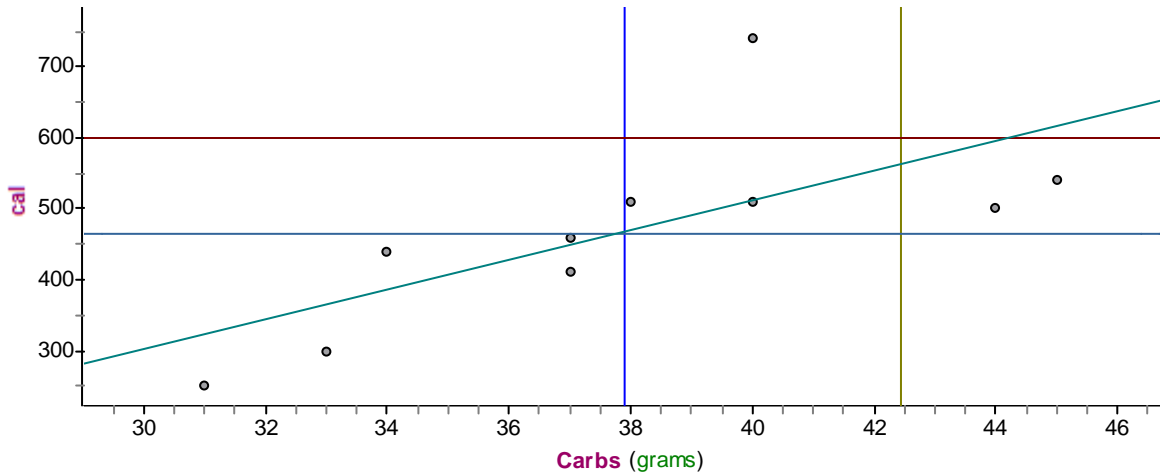
This illustrates an important property of the LSRL: It will always go through the point _____.

If we think about the scatterplot of (z_x, z_y) , this means that the line will go through $(0, 0)$. In other words, the y-intercept will be 0. Furthermore, it can be proven that the slope of the LSRL on the standardized plane is equal to the correlation coefficient (r). Thus, on the standardized plane, the LSRL is: $\hat{z}_y = 0 + r(z_x)$. So, for each standard deviation above the mean of x , the predicted value of y will be r standard deviations above the mean of y .



Note: Regression slopes don’t tell us the strength of an association since they are dependent on units. For example, if we measured carbs in mg, this would stretch the data out 1000 times on the x-axis and the slope will be 1000 times flatter.

Now, here is the original data again, with vertical lines showing \bar{x} and $\bar{x} + s_x$, horizontal lines showing \bar{y} and $\bar{y} + s_y$, and the least squares line.



- If the value of $x = \bar{x}$, the predicted value of y is \bar{y} .
- If the value of x is one standard deviation above the mean ($\bar{x} + s_x$), the predicted y value is NOT one standard deviation above the mean y -value ($\bar{y} + s_y$). Instead, the predicted y value is .703 standard deviations above the y -value ($\bar{y} + .703s_y$).
- If the value of x was 2 standard deviations below the mean, where would the predicted value of y be located?

Thus, since $-1 \leq r \leq 1$, the predicted value of y will almost always be closer to its mean than the x -value is to its mean (in terms of standard deviations). This concept is called “regression to the mean.” Furthermore, the weaker the association the more the predicted value of y will regress to its mean. This makes sense as we are less likely to make bold predictions when there is little association between the variables. We are more likely to play it safe and guess something close to the mean.

Dave Bock email idea

This concept was first publicized by Francis Galton, who noticed that tall fathers had tall sons, but not quite as tall on average, and that short fathers had short sons, but not as short on average. Imagine what would happen if this wasn’t true!

So far, we have been working with standardized values. What if we wanted to use the original units?

$$\hat{z}_y = rz_x$$

$$\frac{\hat{y} - \bar{y}}{s_y} = r \left(\frac{x - \bar{x}}{s_x} \right)$$

$$\hat{y} - \bar{y} = r \left(\frac{s_y}{s_x} \right) (x - \bar{x})$$

$$\hat{y} - \bar{y} = r \left(\frac{s_y}{s_x} \right) x - r \left(\frac{s_y}{s_x} \right) \bar{x}$$

$$\hat{y} = r \left(\frac{s_y}{s_x} \right) x - r \left(\frac{s_y}{s_x} \right) \bar{x} + \bar{y}$$

Since the slope of the regression line is the coefficient of x, $b = r \left(\frac{s_y}{s_x} \right)$.

Once we know the slope, we can use the fact that (\bar{x}, \bar{y}) is on the line to find the y-intercept: $\bar{y} = a + b\bar{x}$

For example, find the LSRL if $\bar{x} = 65, s_x = 4, \bar{y} = 150, s_y = 30, r = .62$.

Note: Unlike the correlation coefficient x and y are not interchangeable when calculating the LSRL. Therefore, we should never use a y-value to try to predict an x-value.

Note: if $r = 0$, then the slope = 0 and the LSRL is horizontal: $\hat{y} = \bar{y}$. If there is no linear association, knowing x won't help us predict y!

HW #22: SR (170-178), RR (198-210), problems page 189 (11, 15, 17, 19, 25, 27)

Tuesday, October 6: Review chapters 7-9

TED VIDEO

Chapters 7-9: The BIG Picture

The purpose of this unit is to investigate the relationship between 2 numerical variables. This relationship can be summarized by addressing:

Direction: positive or negative (determined by the slope, b , or correlation coefficient, r)

Form: linear or non-linear (we use a mathematical equation to model the form of the data. We use a residual plot to check if the model we chose is appropriate). If the form is linear, we can describe specifically how the y -values change with x by interpreting the slope. Data = Form + Scatter

Scatter (or strength): Are the observed values close to the values predicted by the model? The correlation coefficient (r), the coefficient of determination (r^2), and the standard deviation (s) measure this in slightly different ways. If you want to know how far off your predictions will be, use s , since it is measured in the units of y . On the other hand, r and r^2 are standardized values without units. If you tell a statistician that $r = 0.9$, he will know approximately what the scatterplot will look like. However, telling a statistician that $s = 0.9$ is meaningless if there are no units included.

Unusual values: Are there unusual values that influence the measures described above? Always graph the data first or risk being misled by unusual values.

Thursday, October 8: Test Chapters 7-9

Consider working ahead on the HW for after break. It is review for the midterm.

Monday, October 19: Review for Midterm

Discuss Projects: Proposal due Monday

HW #23: problems page 321 (23, 24-skip part a, 25, 26, 29, 31, 32, 37)

Tuesday, October 20: Review for Midterm

Something fun...

HW #24: problems page 131 (3, 6abc, 11, 14, 20, 26, 28, 30)

Thursday, October 22: Midterm

Proposals due on Monday!

Ask me for the Rubric for the Project.

AP Statistics First Semester Project: Response Bias

The Project: You and your partner (or you by yourself) will design and conduct an experiment to investigate the effects of response bias in surveys. You may choose the topic for your surveys, but you must design your experiment so that it can answer at least one of the following questions:

- Can the wording of a question create response bias?
- Do the characteristics of the interviewer create response bias?
- Does anonymity change the responses to sensitive questions?
- Does manipulating the answer choices change the response?

Proposal (20 points):

- The proposal is due: Monday, October 26. Late work will be penalized 20% per day, even if you are absent.
- The proposal will be worth 20% of the grade, so don't treat it casually.
- If the proposal isn't approved the first time, you will need to resubmit it for a reduced grade. You must attach the original proposal to any resubmissions.

In your proposal, you should:

- Describe your topic and state which type of bias you are investigating
- Describe how you will obtain your subjects (minimum sample size is 50). This must be practical!! Your population does not need to be from CDO nor should you interrupt any classes.
- Describe what your questions will be and how they will be asked, including how you will incorporate the principles of a good experiment and avoid potentially confounding variables. You should also indicate what your hypotheses are. Convince me that you have a good design!

Poster (80 points):

- The poster is due: _____. Late work will be penalized 20% per day, even if you are absent.
- The key to a good statistical poster is communication and organization. Make sure all components of the poster are focused on answering the question of interest.
- The poster should be standard sized and not on foam board. Make sure the poster is light enough to be hung on the wall.

The poster should include:

- Title (in the form of a question).
- Introduction. In the introduction you should discuss what question you are trying to answer, why you chose this topic, and what your hypotheses are.
- Data Collection. In this section you will describe how you obtained your data. Be specific.
- Graphs and Summary Statistics. Make sure the graphs are well labeled, easy to compare, and help answer the question of interest.
- Discussion and Conclusions. In this section, you will state your conclusions. You should also discuss any errors you made, what you could do to improve the study next time, and any other comments based on your own critical reflection on the project.
- Live action pictures of your data collection in progress.

Presentation: Each pair (or individual) will be required to give a 5 minute oral presentation to the class. Both members need to participate equally and should be prepared to answer questions.

Examples of Successful Projects:

“Cartoons”, by Sean Wu and Brian Hartzheim

1. “Do you watch cartoons?” (90% yes)
2. “Do you *still* watch cartoons?” (60% yes)

“Milk vs. Orange Juice”, by Angela Chen and Sharon Lai

1. “Which do you prefer, milk or orange juice, as a breakfast drink?” (milk: 14%)
2. “Milk contains high levels of vitamin D and calcium. Do you prefer milk or orange juice as a breakfast drink?” (milk: 64%)

“Cheating”, by Wilson Kurniawidjaja, Oliver Lee, and Charlene Wang

1. “Do you cheat in class?” (anonymous: 47% would)
2. “Do you cheat in class?” (not anonymous: 15% would)

“Make-Up”, by Caryn Suryamega and Trisha Tsuno

(all questions asked to males)

1. “Do you find females who wear makeup attractive?” (wearing makeup: 75% yes)
2. “Do you find females who wear makeup attractive?” (without wearing makeup: 30% yes)

“Time Online”, by Yale Lee and Helen Theung

1. “On average, how many hours do you spend online each week: 0-5, 6-10, 11-16, 17-25, 26- 35, or more?”
2. “On average, how many hours do you spend online each week: 0-5, 6-10, 11-16, or more?”
-For this question, the students anticipated that subjects would be embarrassed to put “more”.
In the first question, 50% answered over 17 hours, but in the second question, 0% did.