Last Unit!!!

Unit 11 – Correlation and Regression

Your Ready-to-be-done-with-slides Professor Colton

Unit 11 - Outline

Unit 11 – Correlation and Regression

Intro

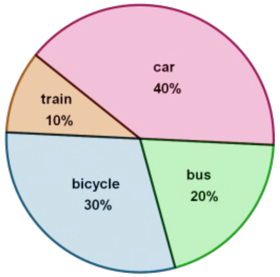
- Correlation Coefficient (r)
- Testing for Significant Linear Correlation

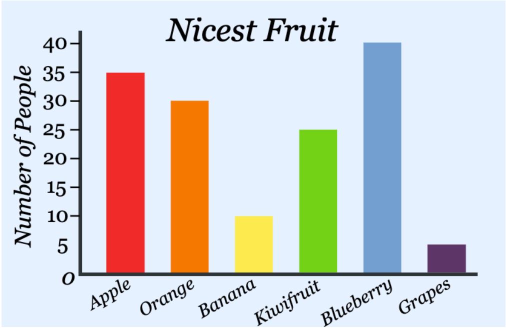
Review + New

- We have studied how to **display** and **describe** distributions depending on the type of data

<u>Qualitative (Categorical) Data</u>

- Non-Numerical data with different categories.
- Ex) States, letter grades, class standing, etc.

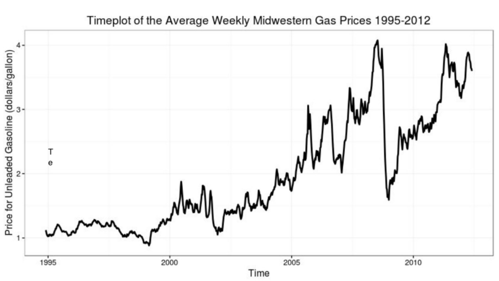


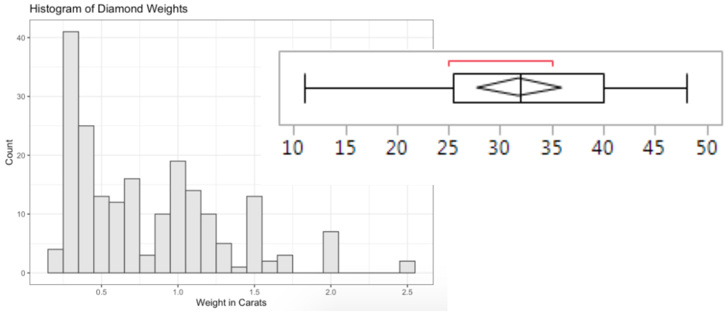


- Here we can describe the mode (the most common category)

<u>Quantitative Data</u>

- Numerical data, counts or measurements
- Arithmetic operations such as adding and averaging make sense
- Ex) Income, GPA, Height, Weight, etc.





- With line graphs (time series plots), we described the trend and seasonal variation

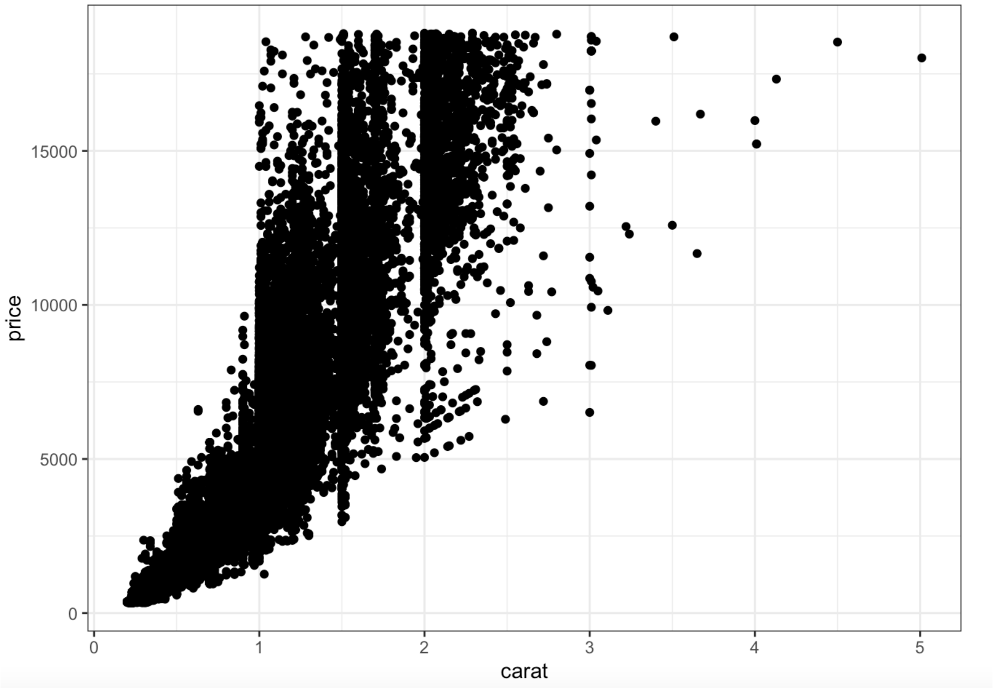
- With histograms and boxplots, we described the SOCS! Shape, Outliers, Center and Spread
- Also numerical summaries like mean, median, SD, IQR, etc.

- (Other than the line graph) these displays were all for ONE variable!
- Now we are going to display and describe the <u>relationship between TWO QUANITITATIVE variables</u>!

Scatterplots

Scatterplot

- Displays the relationship between **two quantitative** variables measured on the same individuals.
- Useful to determine if an association exists!
  - So is there a pattern where *some values of one variable* tend to occur with *some values of the other variable*
  - Example) Smaller carat diamonds tend to have lower prices, and as the carat increases prices tend to increase as well
- In some situations, such as experiments or studies, we want to see if one variable explains the variability of the other.
  - Can we use one variable to predict the other?
  - Example) If I have a 2.5 carat diamond, how much will it cost?



Setup

- Axes
  - The explanatory (independent) variable goes on the X (horizontal) axis
  - The response (dependent) variable goes on the Y (vertical) axis
    - Example) How large a diamond is impacts how much it costs! So the price would be the dependent (Y) and carat independent (X)
  - If there is no clear explanatory/response relationship, then it does not make a difference which variable goes on which axis.
- Points
  - Every individual in the data set has two measurements, one for each variable, and each individual appears as a dot on the plot.
  - Every point is an ordered pair (x,y)

LCQ – Explanatory vs Response

Problem: Determine which variable would go on the X-axis and the Y-axis of the scatterplot → *It is important to get this correct because it will impact the analyses we will learn later!!*

Strategy→ Think in terms of dependent and independent variables (response vs explanatory) ; or maybe chronologically, which comes first)

a) The amount of time spent studying for an exam and the exam score

X = Time spend studying → Chronologically, you study before the exam. So this would be the independent

Y = Exam score → Exam score definitely depends on the amount of time you spend studying

a) The weight and height of a person

X = Height → In general (overall), we can say that the height of someone helps determine their weight (the taller, the more weight) so it would come first

- *This pattern might not always be true, but in general we can say this*

Y = Weight

a) The amount of yearly rainfall and crop yield

X = Amount of yearly rainfall

Y = Crop yield → This definitely depends on the amount yearly rainfull, so it would be on the Y-axis

a) A student's math SAT score and the verbal SAT score

X = either or

Y = the other one

For this one there is no clear explanatory vs response. Can't really say that math score depends on verbal score or vice versa.

So we can choose which goes where, both would be correct

Interpreting Scatterplots

Just like there was a specific way we describe the SOCS of histogram for example, there are certain aspects that we look for when interpreting the relationship between two variables in a scatterplot.

Overall Patter of a Scatterplot

- We can describe the overall pattern of a scatterplot by looking at four characteristics:
    1) Form
    2) Direction
    3) Strength
    4) Outliers

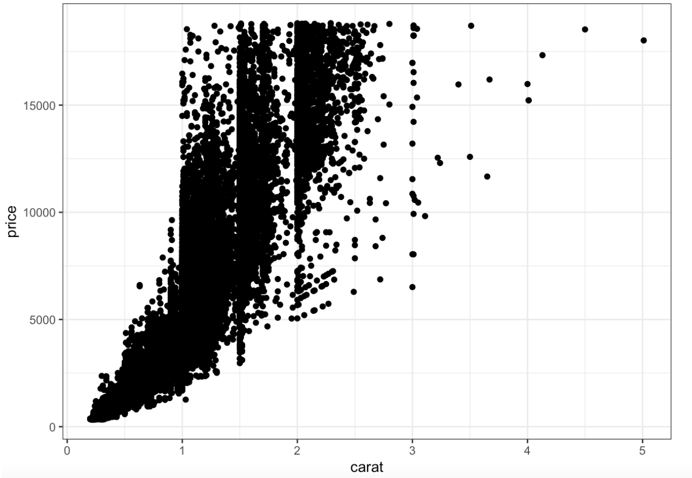
Form

Form

- This refers to the pattern of the dots
- Or we can think about it as the “best” type of line that we could draw to the data, if any
  - Imagine trying to draw a line that tries to go through the middle of all the data as best as possible, and maybe “bands” that surround the data
- It might not be super clear every time, so we are just looking in general!

- *Remember, we are trying to describe the relationship between our two variables, this is the GOAL!*

- There are three types of forms we will consider:



Linear

- Points are following a general linear trend → Straight line
- Note that it won’t be perfect, but a general linear trend is what we are looking for here
- A linear relationship occurs quite frequently in the real world

Curved

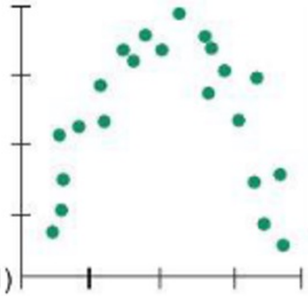
- Points are showing some evidence of curvature
- NOT a straight line → Any type of CLEAR curvature in the “best” line we could

Random scatter

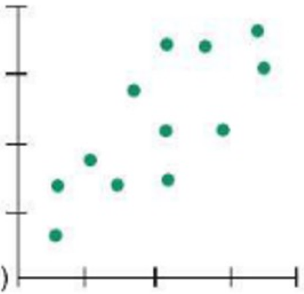
- There really is no pattern, points are just scattered about randomly kinda like a cloud of points

LCQ: Form

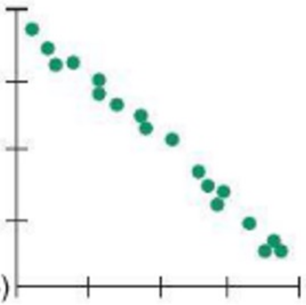
Problem: Determine the form for each of the following scatterplots.



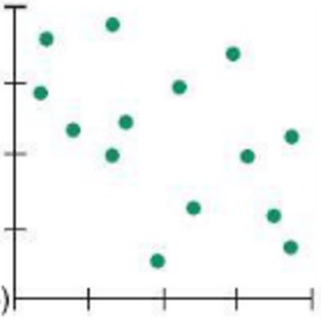
(1)



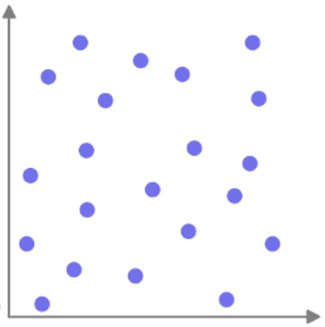
(2)



(3)



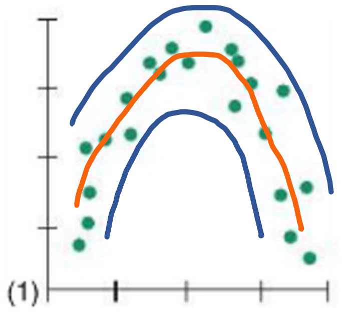
(4)



(5)

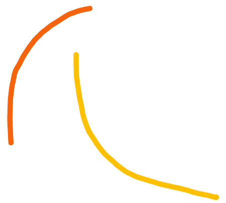
LCQ: Form

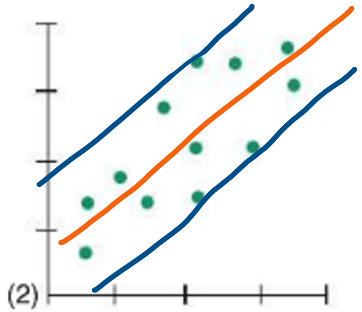
Problem: Determine the form for each of the following scatterplots.



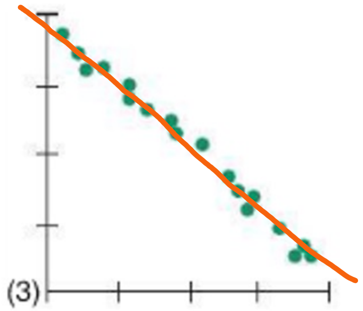
Curved → CLEAR curvature if we draw that line through the middle and if we trace the lower and upper edges

Curvature can also look like these where it kinda levels off, it doesn't have to be the full rainbow

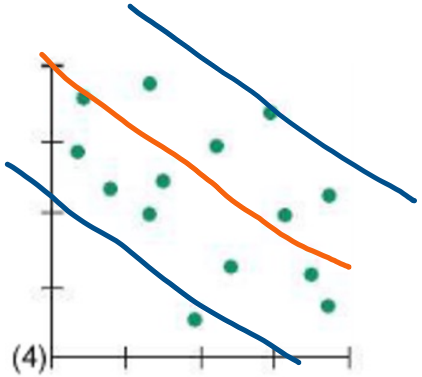




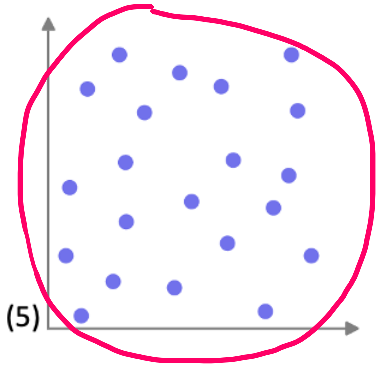
Linear → for the most part follows the straight diagonal line through the middle



Linear → dots more clearly follows a linear path



Linear-ish → not perfect (or as clear as the others), but these dots generally follow straight line path with the middle and outside lines



Random → there is no patter here, just scattered everywhere

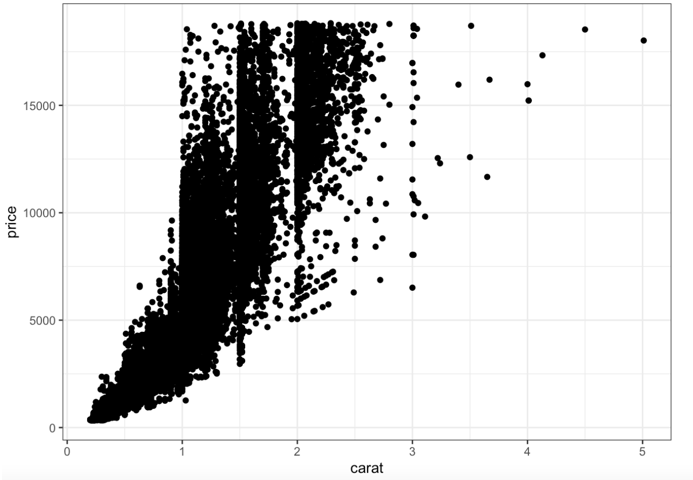
Direction

Direction

- This refers to the **direction** of the association between the two variables

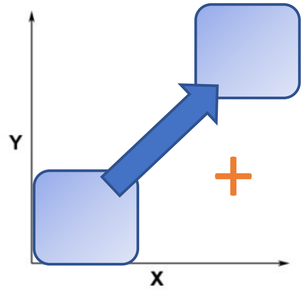
** This only applies to linear relationships as curved patterns can have both positive and negative directions in the same scatterplot, and random relationships are neither entirely positive or negative.

- We can think of this as the slope of our line!



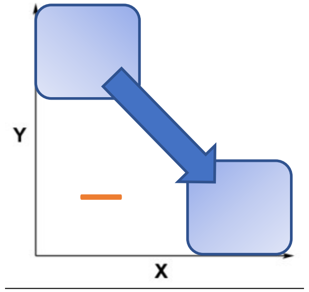
Positive Association

- High values of one variable correspond (tend to occur with) to high values in the other variable, and
- Low values of one variable correspond to low values in the other variable
- So as the values of X increase, the values of Y tend to increase as well → Positive slope!
  - Variables are moving in the same general direction



Negative Association

- High values of one variable correspond (tend to occur with) to low values of the other and vice versa
- So as the values of X increase, the values of Y tend to decrease → Negative slope!
  - Variables are moving in the opposite general direction

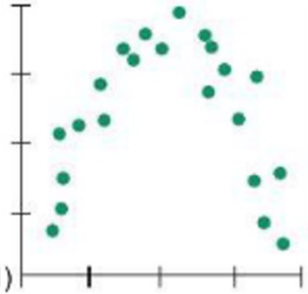


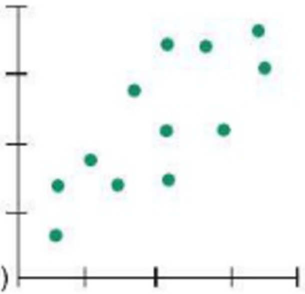
No Association

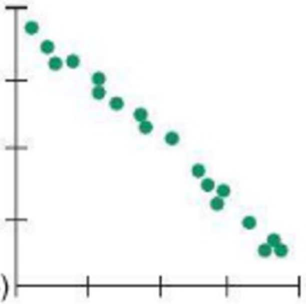
- There is no pattern or general trend
- Low values of one variable can have both high or low values of the other; high values of one variable can have both high or low values of the other
- Knowing if we have a high or low value of one variable gives us no indication whether the other variable’s value will be high or low

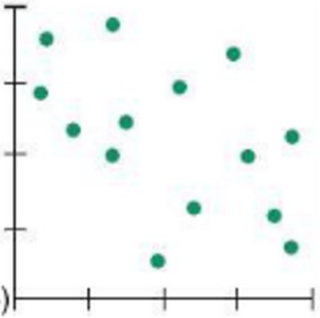
LCQ: Direction

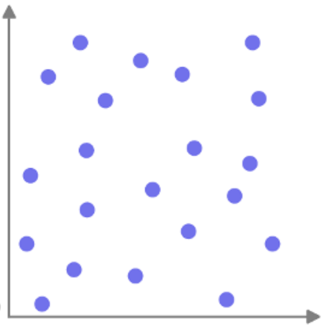
Problem: Determine the direction for each of the following scatterplots.


(1)


(2)

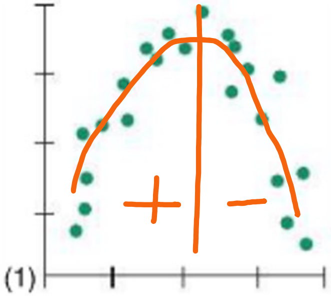

(3)


(4)


(5)

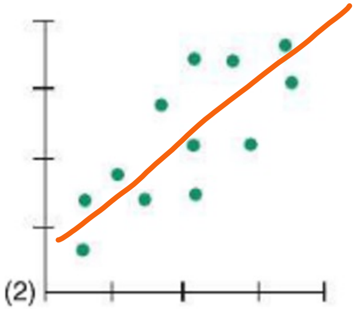
LCQ: Direction

Problem: Determine the direction for each of the following scatterplots.

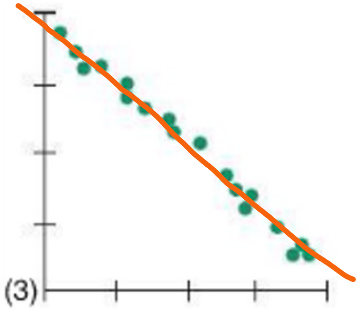


Starts positive, then goes negative; Not applicable → If we divide this plot in two, there are two different associations! This is because of the curved form.

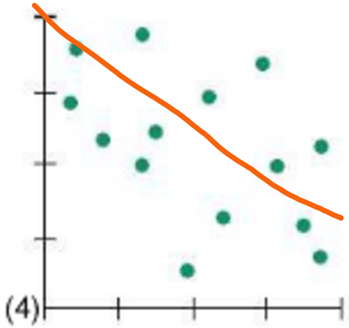
- *We are only going to apply direction to linear forms, so this wouldn't fit*



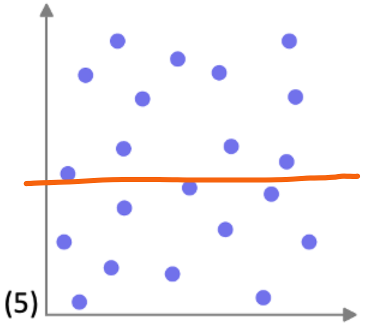
Positive → definitely a positive slope to the "best" line we could draw. And we see the points shifting in the same direction, both increasing together



Negative → clear negative slope, Ys decrease as Xs increase (opposite directions



Positive → Still a positive slope to the line through the middle of the points



No association → there is no clear shift in the positive or negative direction, points kinda stay flat

It turns out the "best" line that we can draw here is a horizontal line cutting them in half, which would have a slope of zero!

We can think of no association as a slope of zero

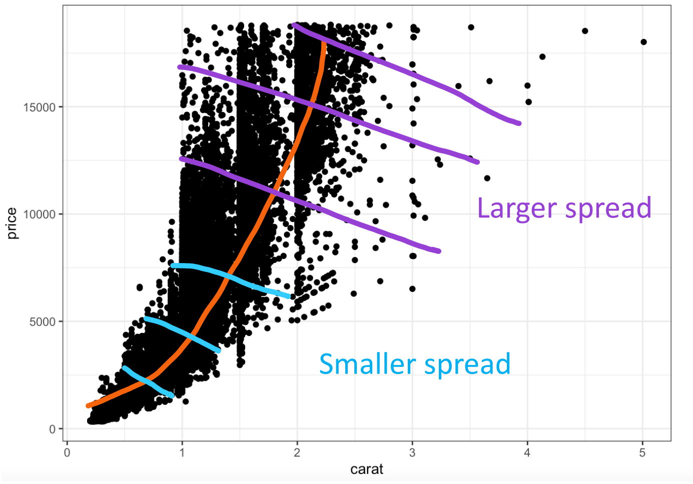
Strength

Strength

- This refers to how **strong** the association is
- In other words, how well the data fits the pattern?

- We can think of this as how CLOSE the data are to our “best” line (the form)!
  - So how small or large the spread is around our “best” line

- To visualize this, we can draw ovals centered on our “best” line that capture the majority of the points
  - Then look at how wide the oval is in the center perpendicular to the “best” line



** We are ONLY going to apply this to linear relationships

- We are going to classify strength in three categories:

High Strength (Strong)

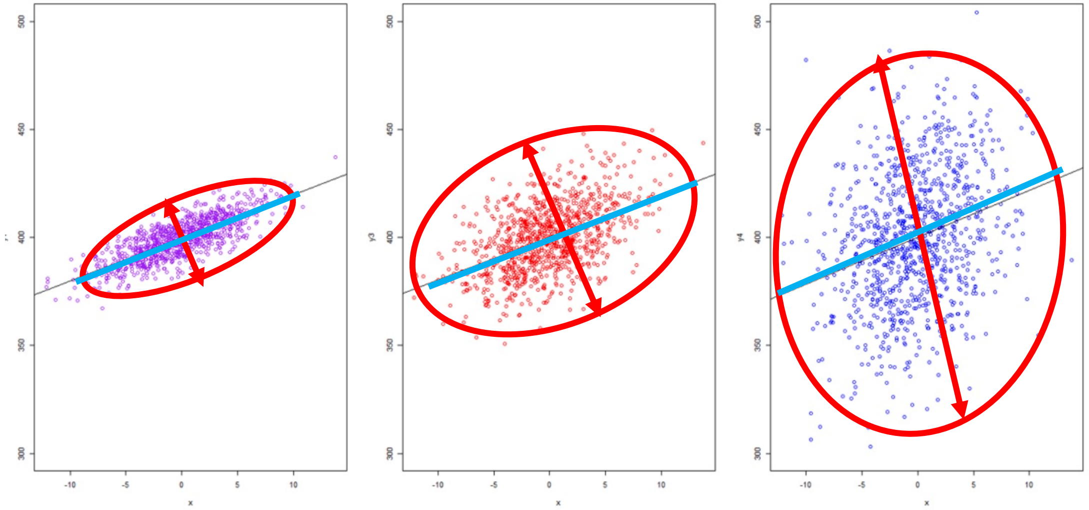
- The dots follow the linear pattern closely, with little scatter
- Relatively small width of our oval

Moderate Association

- The dots follow a general pattern, but not as tightly packed
- Relatively medium width of our oval, still looks like an oval

Low Strength (Weak)

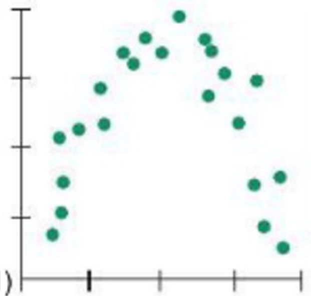
- The dots do not appear to be following a pattern
- Relatively large width of our oval, which starts looking more like a circle



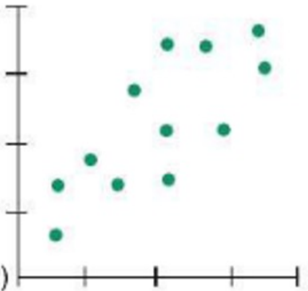
https://bookdown.org/yshang/book/correlation.html

LCQ: Strength

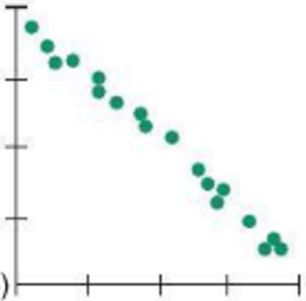
Problem: Determine the strength for each of the following scatterplots.



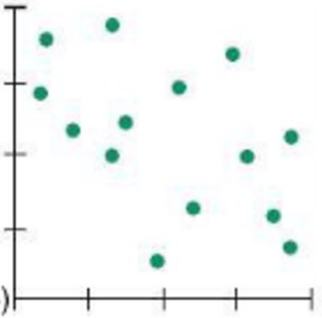
(1)



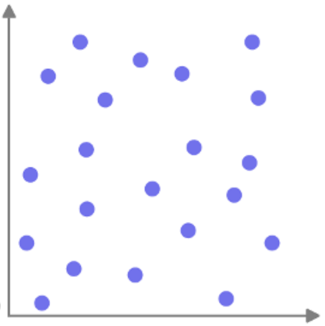
(2)



(3)



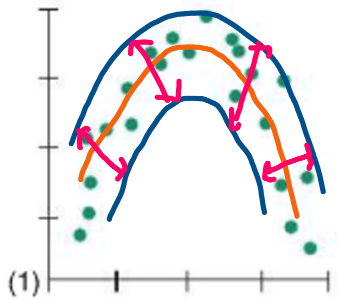
(4)



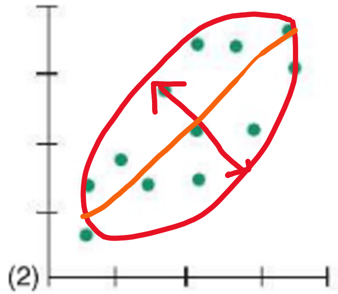
(5)

LCQ: Strength

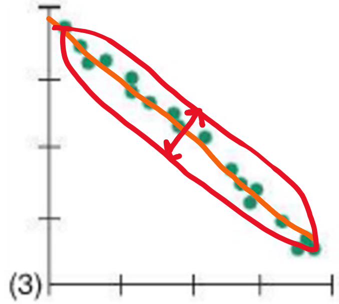
Problem: Determine the strength for each of the following scatterplots.



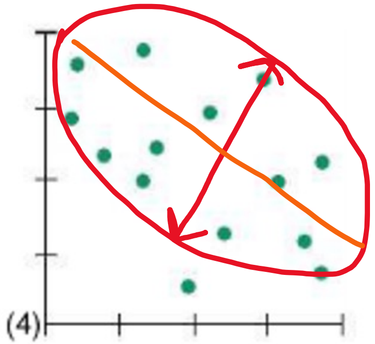
Starts, but not applicable → Points definitely follow this pattern very well, indicating a high strength. But because it has a curved form, we are not going to apply our definition of strength



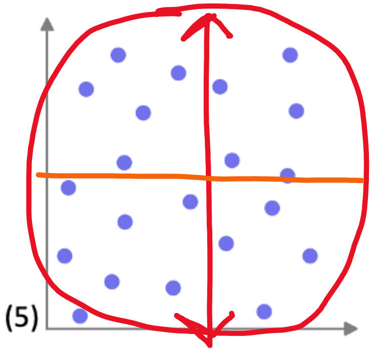
Moderate → Fairly smallish width of our oval, not too much spread around the "best" line



Strong → very small spread around the "best" line. Points follow this pattern very closley

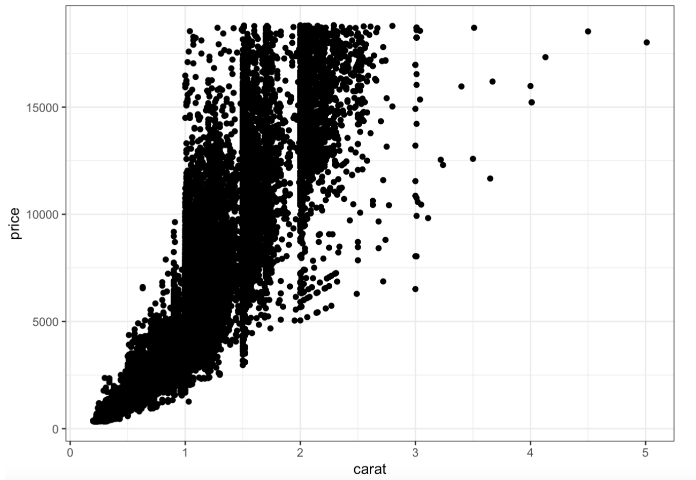


Moderate, maybe weak → medium-ish width of our oval. Hard to determine whether it would be moderate or weak. But definitely weaker than (2),



Super weak→ almost have a circle / rectangular-ish shape here because of the lack of association, very large spread too

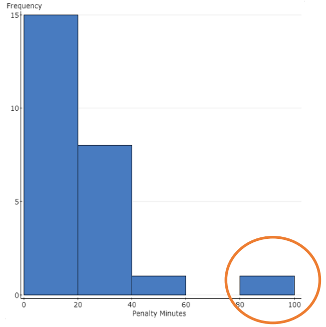
Outliers

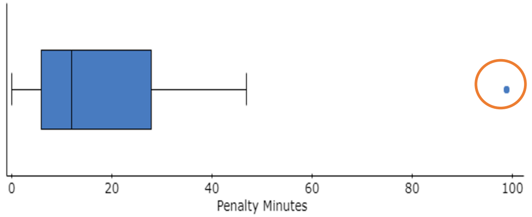


Outliers

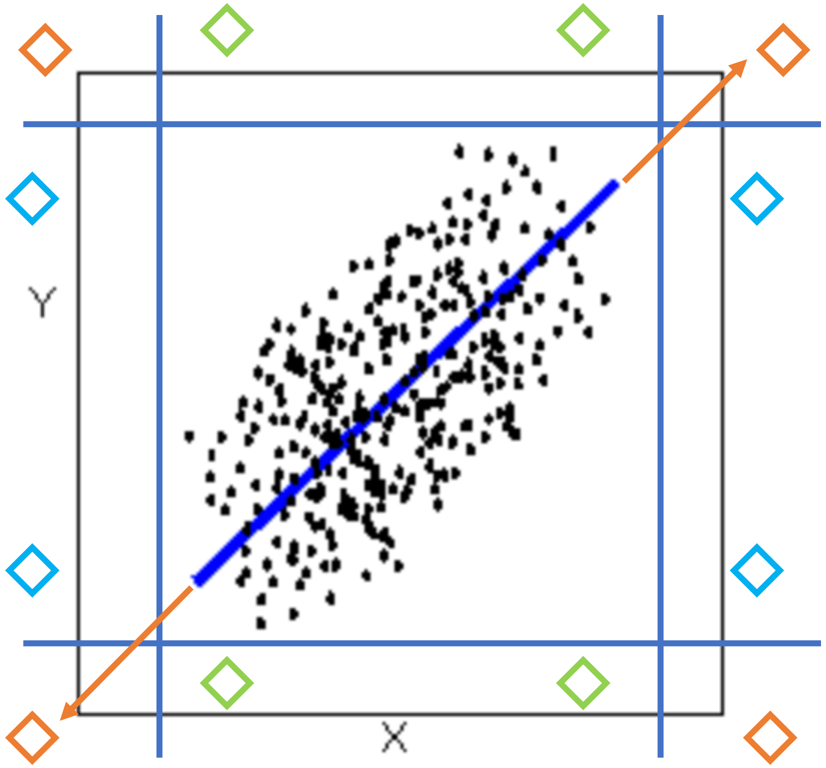
- Outliers are points that do not follow (deviate from) the overall / general pattern

- Recall outliers from histograms and boxplots





- Now we have two variables, so points can be an outlier in several different ways:
  - Only in the X direction → meaning the Y values are within the normal range, but the Xs are too different
  - Only in the Y direction → X values are within the normal range, but the Ys are too different
  - In both directions → Both X and Y are way different



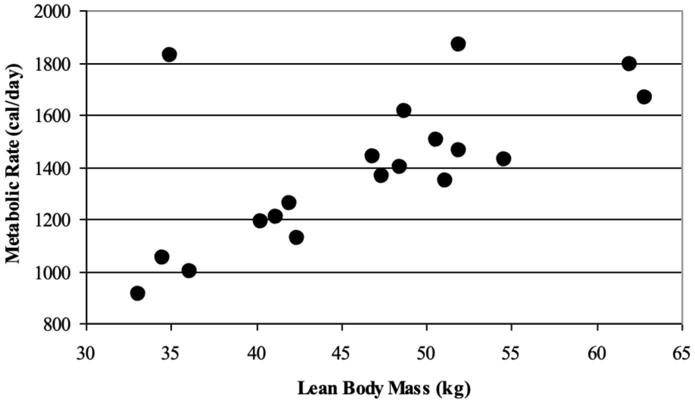
This one and the upper right corner are at least in the path if we extend the best line, the other orange outliers are VERY different than the all other points

https://bolt.mph.ufl.edu/6050-6052/unit-1/case-q-q/scatterplots/

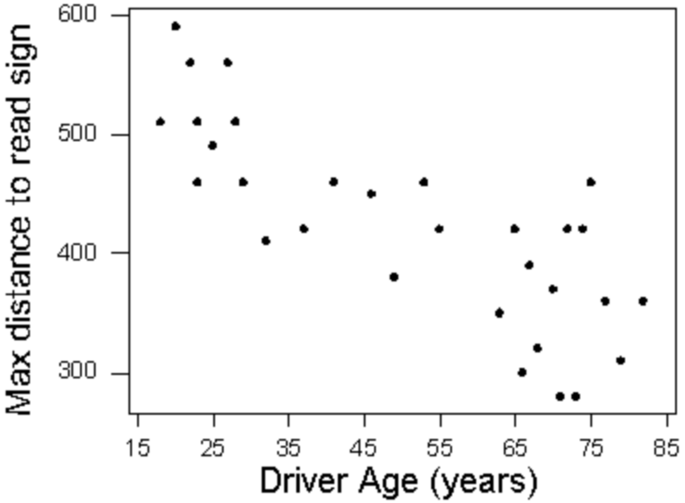
LCQ: Interpreting Scatterplots

Problem: Interpret the following scatterplots by discussing their Form, Direction, Strength and Outliers

a)



b)

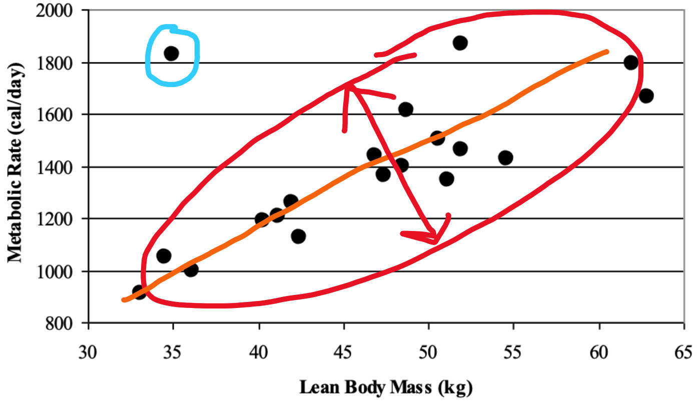


https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/scatterplots-3-of-5/

LCQ: Interpreting Scatterplots

Problem: Interpret the following scatterplots by discussing their Form, Direction, Strength and Outliers

a)



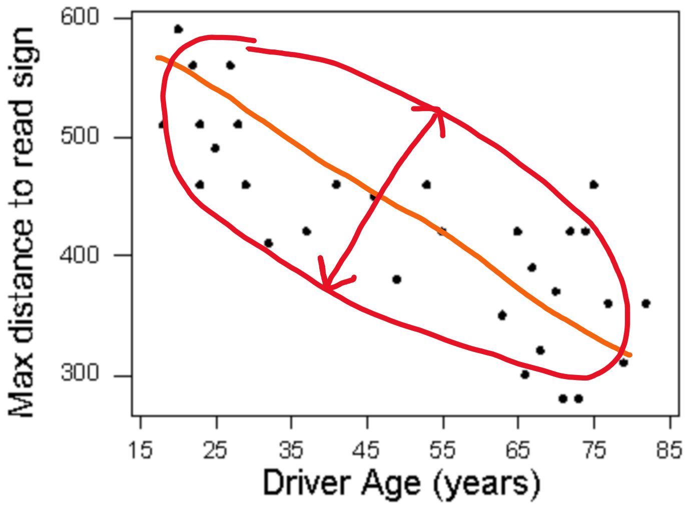
All 4 pieces:

Linear form, positive direction, moderate (strong?) strength, appears to be at least one outlier

If we wanted to write this interpretation in a nice sentence with context, it could go like this:

There is a positive, moderate strength, linear relationship between lean body mass and metabolic rate. There appears to be one outlier at ≈ (35, 1850)

b)



https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/scatterplots-3-of-5/

All 4 pieces:

Linear form, negative direction, moderate strength, and does not appear to be any outliers

If we wanted to write this interpretation in a nice sentence with context, it could go like this:

There is a negative, moderate strength, linear relationship between drivers age and max distance to read sign. There does not appear to be any outliers

Correlation

- We just learned how to summarize the key features of a scatterplot, but they were fairly subjective
- To get a more accurate representation, we need to quantify our description of the relationship between X and Y

Correlation

- The **correlation (*r*)** is an index that expresses the direction and strength of the relationship
  - It combines both of these aspects into a single number measure
  - Often referred to as the correlation coefficient (or Pearson's correlation)
- It's scale goes from -1 to 1 → $-1 \leq r \leq 1$
- Correlation (*r*) is a statistic that is used to estimate the parameter ρ (Greek letter "rho")
  - ρ represents the linear correlation coefficient between population X and Y values

Interpreting the Correlation

- The sign of the correlation coefficient indicates the direction of the association.
  - This will always be the same sign as the slope
- The absolute value of the correlation coefficient |r| indicates its strength.

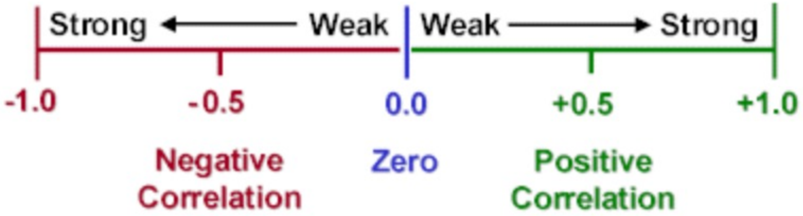
Properties of Correlation

- Both variables have to be quantitative
- *r* measures the strength of linear relationships
- *r* has no units of measurement
- *r* the units of measurement for one or both variables will not change the value of *r*
- Correlation is the same regardless of which variable you label as X and Y
- *r* is strongly affected by outliers
- Does NOT imply a cause-and-effect relationship

Formula

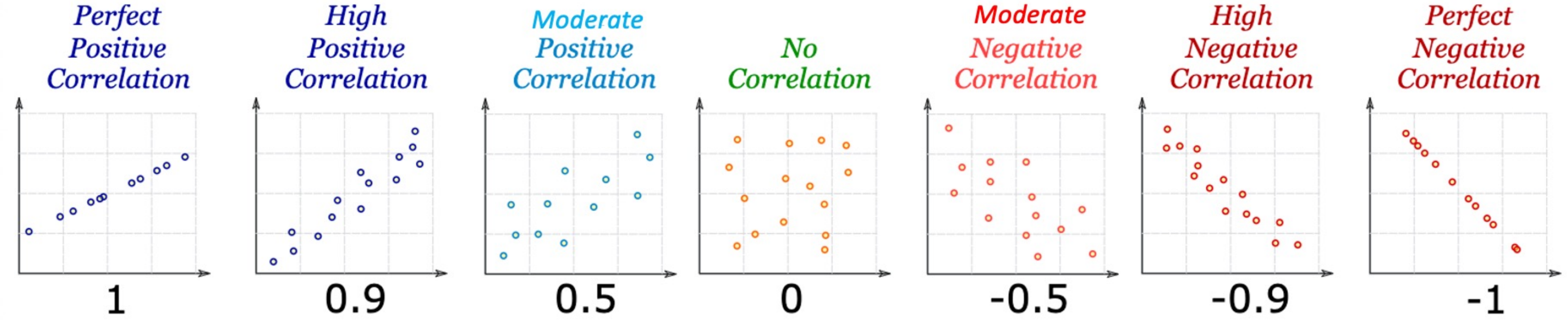
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- In words, we could describe this formula as an "average" of the product of the standardized values of the two variables.
- Our calculator will do this for us!



Correlation	Intepretation
$0.8 < r \leq 1$	Strong, positive relationship
$0.4 < r \leq 0.8$	Moderate, positive relationship
$0 < r \leq 0.4$	Weak, positive relationship
$-0.4 \leq r < 0$	Weak, negative relationship
$-0.8 \leq r < -0.4$	Moderate, negative relationship
$-1 \leq r < -0.8$	Strong, negative relationship

Correlation Visually



https://www.mathsisfun.com/data/scatter-xy-plots.html

With perfect correlations, there is an EXACT equation describing the data: such as Y = 10 – 2X

- If I know X, I automatically know where Y is going to be

As we decrease the strength of the correlation, there more and more uncertainty is introduced.

- So maybe the equation becomes Y = 10 – 2X ± 1 and then decrease again and it becomes Y = 10 – 2X ± 3
- Until ultimately as the correlation gets very close to zero, knowing X gives me no knowledge about Y

LCQ: Interpreting Correlation

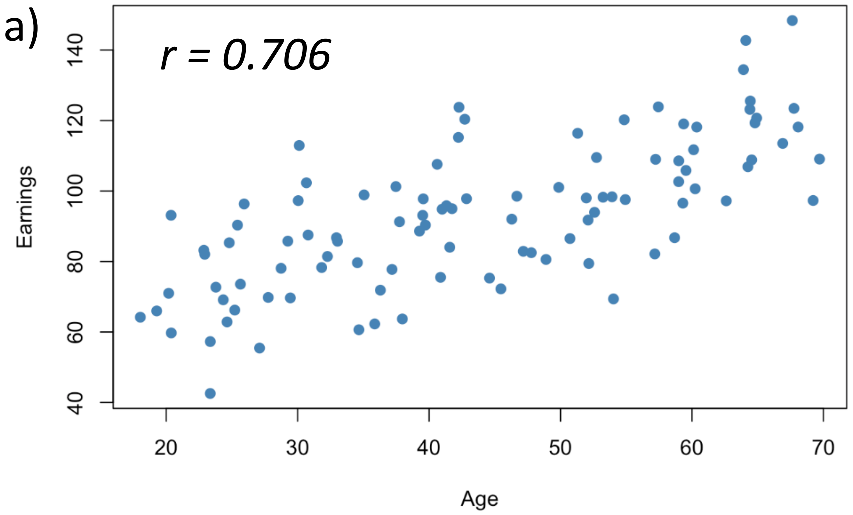
How to Interpret Correlation

- Have to mention the **magnitude** (which refers to the strength of linear relationship between X and Y) and **direction** (+/-)
- And **USE CONTEXT**!!!
- General structure → There is a < strong, moderate, weak >, < positive / negative > linear relationship between X and Y

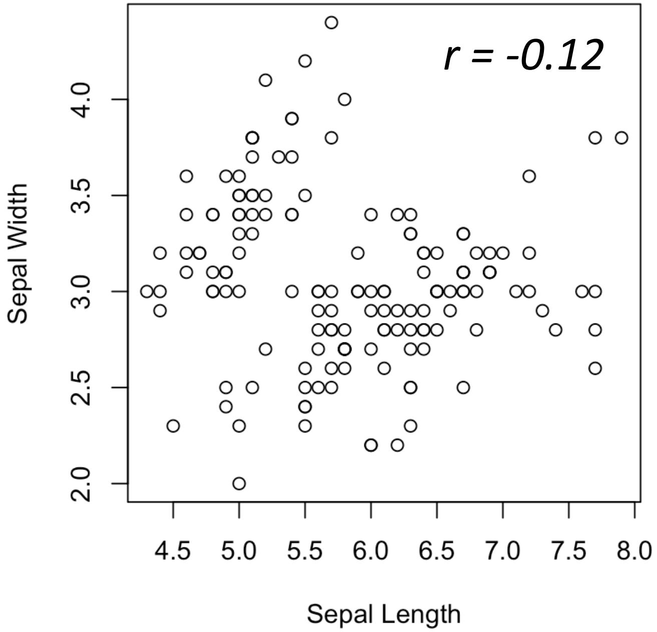
Problem: Interpret the correlations for the each of the following scatterplots

https://www.econometrics-with-r.org/3-7-scatterplots-sample-covariance-and-sample-correlation.html

a)



b)



LCQ: Interpreting Correlation

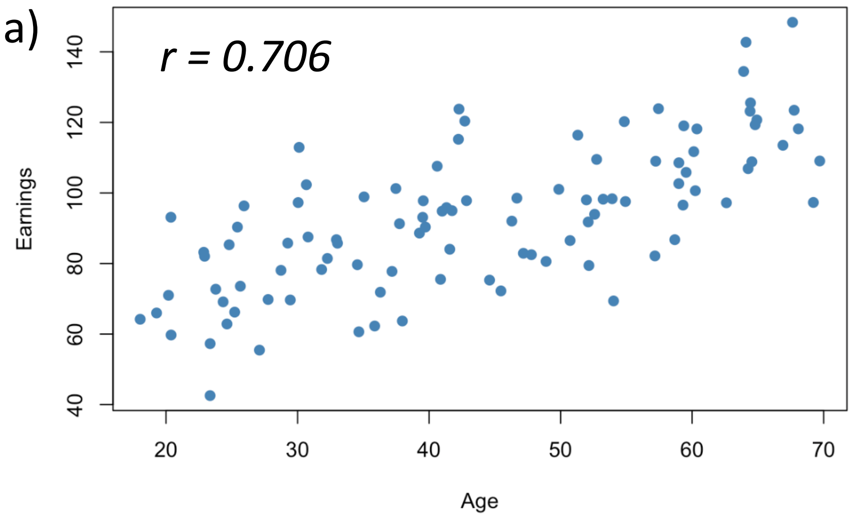
How to Interpret Correlation

- Have to mention the **magnitude** (which refers to the strength of linear relationship between X and Y) and **direction** (+/-)
- And **USE CONTEXT**!!!
- General structure → There is a < strong, moderate, weak >, < positive / negative > linear relationship between X and Y

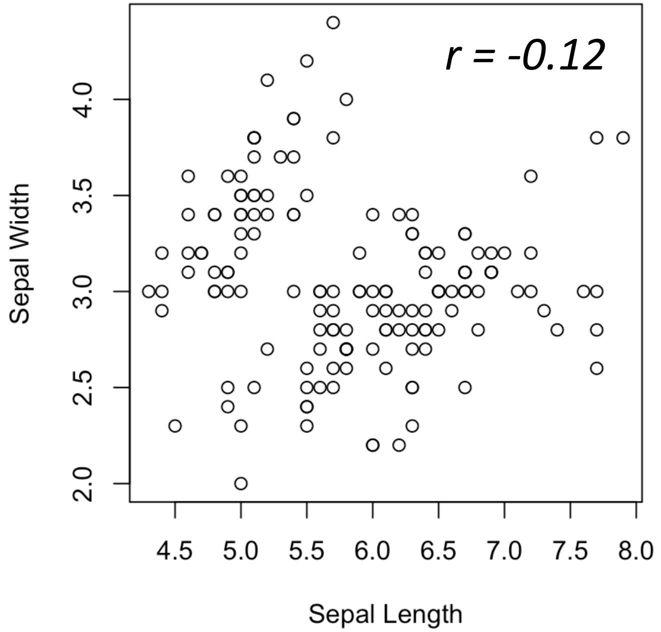
Problem: Interpret the correlations for the each of the following scatterplots

https://www.econometrics-with-r.org/3-7-scatterplots-sample-covariance-and-sample-correlation.html

a)



b)



Options

1) There is a moderate, positive linear correlation between x and y → MISSING CONTEXT!!! Use the variable names at least

2) There is a moderate, positive linear correlation (relationship) between age an earnings → PERFECT now! Could also say 'linear relationship' instead of 'correlation, both would be correct!

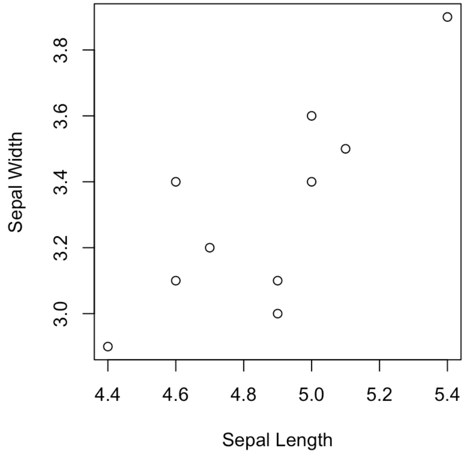
There is a weak, negative linear correlation between Sepal Length and Sepal Width → Remember the direction is based off the sign of the correlation and the strength is based on the magnitude (refer to the guidelines in the table!

Using Calc - Correlation

GOAL: Calculate the Correlation Coefficient!

1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinRegTTest
   a) Xlist = L_1
   b) Ylist = L_2
   c) Freq = 1
   d) β & ρ: *Ignore for now*
   e) RegEQ: *Leave blank for now*
   Calculate

X	Y
Sepal Length	**Sepal Width**
5.1	3.5
4.9	3
4.7	3.2
4.6	3.1
5	3.6
5.4	3.9
4.6	3.4
5	3.4
4.4	2.9
4.9	3.1

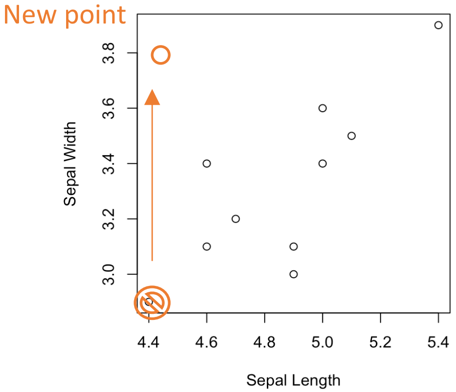


Outliers Demonstration

Let's change one data point to see the effects on the correlation:

- 9th observation: (4.4, 2.9) → (4.4, 3.8)

Now recalculate the correlation!



Using Calc - Correlation

GOAL: Calculate the Correlation Coefficient!

1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinRegTTest
   a) Xlist = L_1
   b) Ylist = L_2
   c) Freq = 1
   d) β & ρ: Ignore for now
   e) RegEQ: *Leave blank for now*

   Calculate

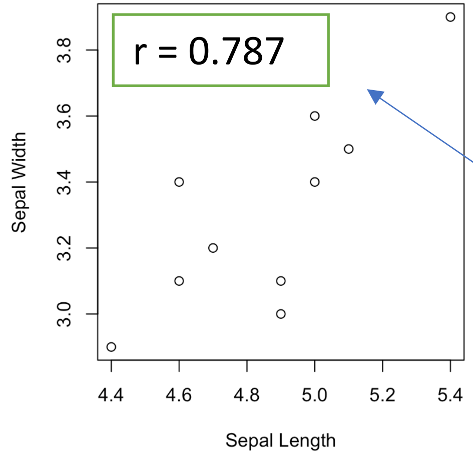
X	Y
Sepal Length	**Sepal Width**
5.1	3.5
4.9	3
4.7	3.2
4.6	3.1
5	3.6
5.4	3.9
4.6	3.4
5	3.4
4.4	2.9
4.9	3.1

Outliers Demonstration

Let's change one data point to see the effects on the correlation:

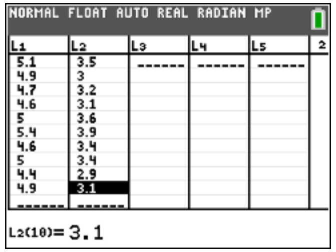
- 9th observation: (4.4, 2.9) → (4.4, 3.8)

Now recalculate the correlation!

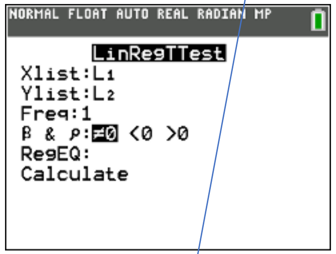


Impact on Correlation

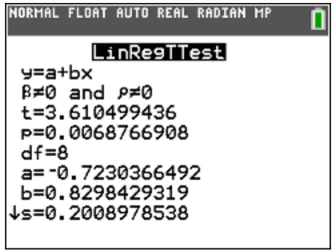
- Changing one point decreased the correlation by 0.512 (which is slightly more than a quarter of the total scale (total range is from -1 to 1, length 2)
- This illustrates how sensitive the correlation can be, especially with small sample sizes



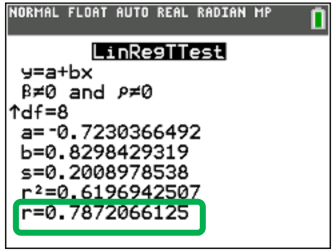
```
NORMAL FLOAT AUTO REAL RADIAN MP
L1     L2     L3      L4      L5     2
5.1    3.5    ------  ------  ------
4.9    3
4.7    3.2
4.6    3.1
5      3.6
5.4    3.9
4.6    3.4
5      3.4
4.4    2.9
4.9    3.1
------ ------
L2(10)=3.1
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
        LinRegTTest
Xlist:L1
Ylist:L2
Freq:1
β & ρ:≠0 <0 >0
RegEQ:
Calculate
```



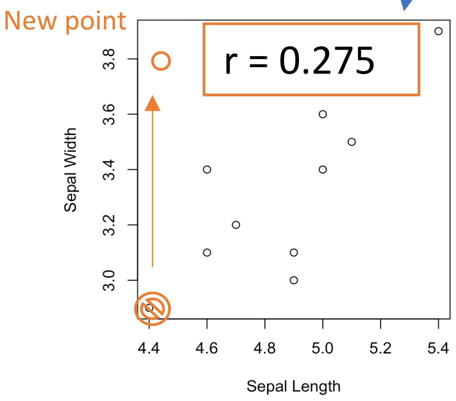
```
NORMAL FLOAT AUTO REAL RADIAN MP
        LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
 t=3.610499436
 p=0.0068766908
 df=8
 a=⁻0.7230366492
 b=0.8298429319
↓s=0.2008978538
```

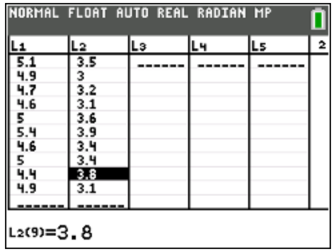


```
NORMAL FLOAT AUTO REAL RADIAN MP
        LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑df=8
 a=⁻0.7230366492
 b=0.8298429319
 s=0.2008978538
 r²=0.6196942507
 r=0.7872066125
```

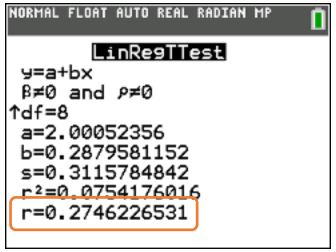
r = Correlation

This is the only output we need for now!





```
NORMAL FLOAT AUTO REAL RADIAN MP
L1     L2     L3      L4      L5     2
5.1    3.5    ------  ------  ------
4.9    3
4.7    3.2
4.6    3.1
5      3.6
5.4    3.9
4.6    3.4
5      3.4
4.4    3.8
4.9    3.1
------ ------
L2(9)=3.8
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
        LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑df=8
 a=2.00052356
 b=0.2879581152
 s=0.3115784842
 r²=0.0754176016
 r=0.2746226531
```

Test for a Linear Relationship via Correlation

Test for a Linear Relationship

- In order to test for a significant linear relationship between two quantitative variables, we can do a test on the correlation
    - Recall Correlation measures the strength of a LINEAR relationship
    - So if the correlation is large enough in magnitude, we can conclude that a linear relationship is indeed present

- If we have a linear relationship between our independent and dependent variable, we can estimate our “best” line and use that to make predictions!
    - This is ultimate goal!

Process

- Before we run our test, we have to look at the scatterplot!
    - ** This is kinda going to be our assumption for this test! (along with having a random sample two quantitative variables! Although we won’t have to mention these)*

    - We are looking for the **form**!

- If the data look at all linear, then we can conduct a hypothesis test using the same procedures we have learned before

Hypotheses

Logic

- This is pretty much the SAME thing that we have been doing with Hypothesis Tests for Proportions and Means
- Except now we are studying a different parameter ρ!

- Recall ρ represents the linear correlation coefficient between population X and Y values

Hypotheses

- Null Hypothesis → We always start by **assuming ρ = 0**
  - There **not a significant linear correlation**, which means there is **NOT a linear relationship** between our X and Y variables

- Alternative Hypothesis → Then we are trying to show the opposite, that **ρ does not equal zero**
  - There is a **significant linear correlation**, which means that that there actually is a **linear relationship**!
  - We can be more specific if we want and try to show that there is specifically a positive or specifically a negative linear correlation as well!

Null Hypothesis

$H_0: \rho = 0$ → NO linear relationship

Alternative Hypothesis

$H_A: \rho \neq 0$ → Linear relationship

$H_A: \rho > 0$ → Positive Linear

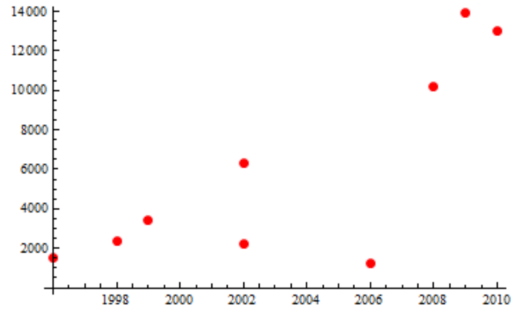
*** We will keep it simple, so do NOT need to define our parameter*

→ But know that there is context involved when we make our interpretation!

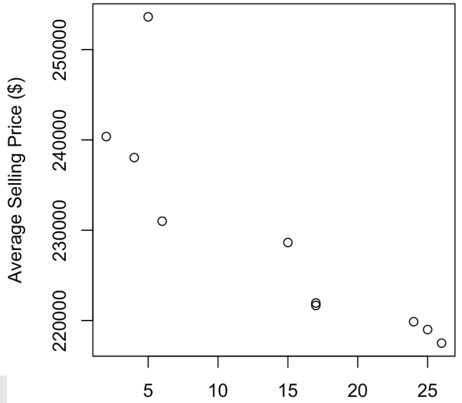
LCQ – Hypotheses

Problem: 1) Determine if it is appropriate to run a test on the Correlation and 2) State the Null and Alternative Hypotheses for each of the following scenarios.

a) You are interested in purchasing a new Honda Civic, so you collect a random sample of 9 Honda Civics being sold on Craiglist. Determine if there is a significant linear correlation between the model year of a Honda Civic being sold on Craigslist and its selling price. Use a 5% significance level.



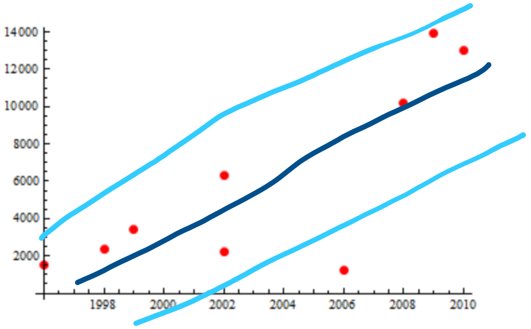
b) A prospective homebuyer is trying to decide which neighborhood to move into. The real estate agent says that subdivisions with more houses for sale typically have homes for lower prices and provides a random sample of 10 neighborhoods with and all available listings. Determine if there is a significant negative linear correlation between number of homes on the market in a subdivision and average sale price at a 1% level of significance.



LCQ – Hypotheses

Problem: 1) Determine if it is appropriate to run a test on the Correlation and 2) State the Null and Alternative Hypotheses for each of the following scenarios.

a) You are interested in purchasing a new Honda Civic, so you collect a random sample of 9 Honda Civics being sold on Craiglist. Determine if there is a significant linear correlation between the model year of a Honda Civic being sold on Craigslist and its selling price. Use a 5% significance level.



1) Determine if appropriate

Scatterplot looks roughy linear, can run test on ρ → Not a perfect linear relationship, but definitely enough to warrant a test! We are just mainly looking for the NOT curvature and any slight signs of linear

2) Hypotheses

H_0: $\rho = 0$ → ALWAYS start with NO linear correlation, so = 0. This of course means starting with the assumption of NO linear relationship, and then trying to show otherwise

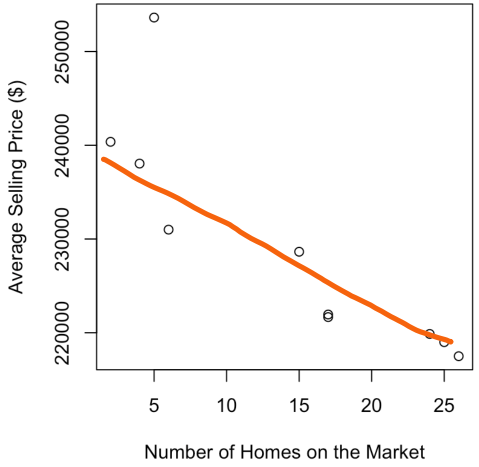
H_A: $\rho \neq 0$ → In the underlined wording, there was no indication whether we were trying to show a positive or negative correlation. We just wanted there to be some correlation.

- This is similar-ish wording to trying to show the average is different than 50 (we don't care if μ is above or below, just not equal to 50)
- Except now with correlation, our "benchmark" is zero, which is kind of an absence of correlation. So showing that there IS a significant correlation means not equal to zero

LCQ – Hypotheses

Problem: 1) Determine if it is appropriate to run a test on the Correlation and 2) State the Null and Alternative Hypotheses for each of the following scenarios.

b) A prospective homebuyer is trying to decide which neighborhood to move into. *The real estate agent says that subdivisions with more houses for sale typically have homes for lower prices* and provides a random sample of 10 neighborhoods with and all available listings. Determine if there is a significant negative linear correlation between number of homes on the market in a subdivision and average sale price at a 1% level of significance.



1) Determine if appropriate

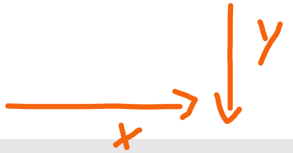
Scatterplot looks linear, can run test on ρ → the outlier in the upper left maybe throws off the pattern, but definitely worth running the test for linearity

2) Hypotheses

H_0: $\rho = 0$ → ALWAYS start with NO linear correlation, so rho = 0

H_A: $\rho < 0$ → In the underlined wording here, there IS an indication that we want a NEGATIVE correlation! So not only do we want a significant correlation, we want rho to be less than zero

- *We could also get this from the previous beliefs italicized in the wording. The real estate agent believes that neighborhoods with MORE houses available sell for LESS on average*
- *So our X is increasing and our Y is decreasing! These are moving in opposite directions → NEGATIVE association*



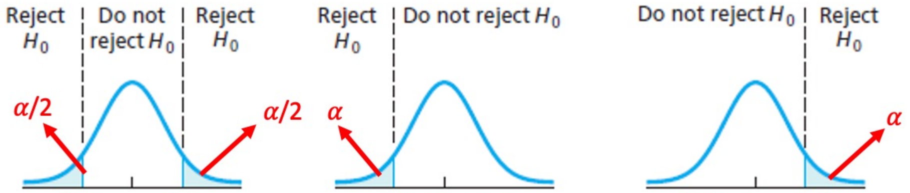
Rejection Region

Rejection Region for Test on Correlation

Degrees of freedom from days past
Before with means we had: CV = t with df = n - 1 (sample size – 1)*

- This will actually be based on the ***t*-distribution** again!
  - But now we have new degrees of freedom! ***df = n – 2***
    - This is because we have two variables now!
  - So all of our Critical Values will be **t*s** with the **new *df***!
- Using calc:

Left-Tailed: $t^* = invT(area = \alpha, df = n - 2)$
Right-Tailed: $t^* = invT(area = 1 - \alpha, df = n - 2)$
Two-Tailed: $t^* = invT(area = \frac{\alpha}{2}, df = n - 2)$



Type of Test:	(a) Two tailed	(b) Left tailed	(c) Right tailed
Critical Value(s):	-t* +t*	-t*	+t*
Rejection Region:	Both sides	Left side	Right side

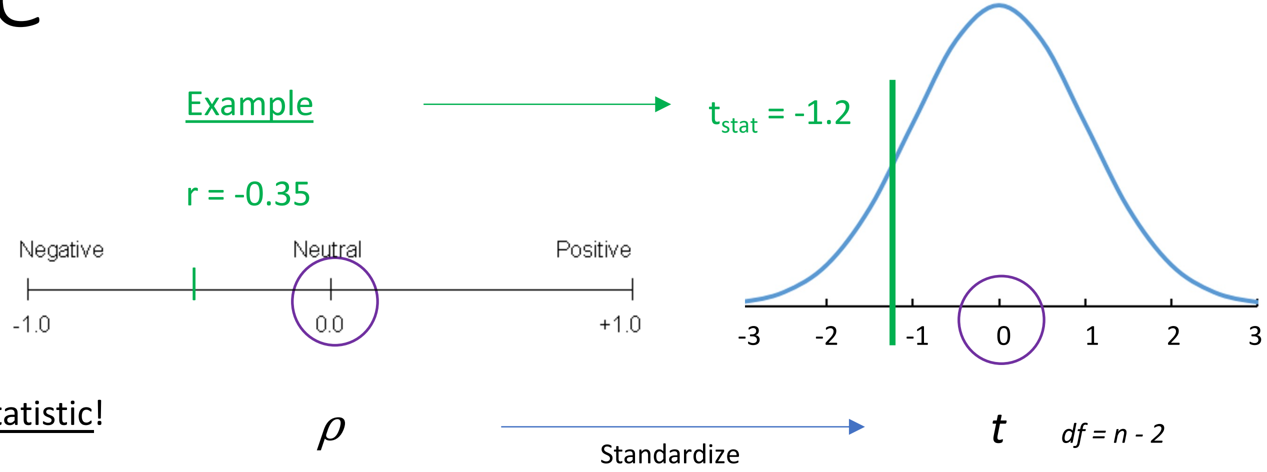
Test Statistic and P-value

Test Statistic

- The **Test Statistic** t_{stat} has the following formula:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad df = n - 2$$

- The **sign** of the correlation will ALWAYS **match** the sign of the Test Statistic!



Example

r = -0.35

t_{stat} = -1.2

ρ Standardize t $df = n - 2$

Logic

- Every bivariate (two variable) sample has a correlation r → this is our statistic, just like $\bar{x}$ or $\hat{p}$
  - We are implicitly comparing this to our parameter with the null value: $r - \rho$, where we start with $\rho = 0$ *(just like $\bar{x}$ - μ or $\hat{p}$ - p)*
- So we standardize this, which is mainly taking into account the sample size!
  - With small samples, the correlation is VERY UNSTABLE (sensitive to small changes in the data), just like we saw in the outliers demo
- So by standardizing, we can tell if the correlation is **strong enough** given the amount of data to conclude if there is really an underlying linear relationship between X and Y

P-Value

- The p-value is the probability of getting our Test Statistic or more extreme, depending on the alternative hypothesis
- Stronger sample correlations (further from zero) will result in smaller p-values!

Decisions

- We can make our decisions to Reject or Fail to Reject using both the traditional method and the p-value method!

Using Calc – Test on Correlation Coefficient

GOAL: Conduct a Test on the Correlation Coefficient!

1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinRegTTest
   a) Xlist = L_1
   b) Ylist = L_2
   c) Freq = 1
   d) β & ρ = Alternative Hypothesis ***
   e) RegEQ: *Leave blank for now*

   Calculate

****** Alternative Hypothesis in Calc***

ρ is the correlation between the X and Y values in the population
β is the true slope of our best line if we had population data

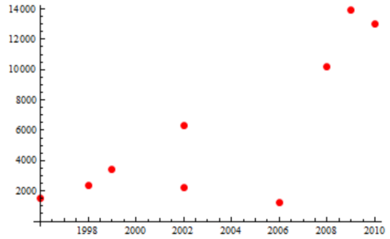
Recall the correlation will always have the same sign as our slope!

** Data table organization

Whenever data is presented in a table like it is here, generally speaking:
- For Vertical tables, the X variable will be on the left and Y on the right
- For Horizontal tables, the X will be the top and Y the bottom

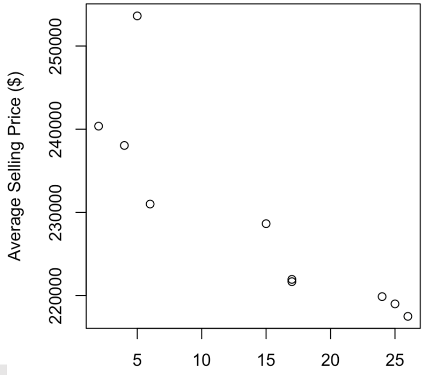
Setup: a) Determine if there is a significant linear correlation between the model year of a Honda Civic being sold on Craigslist and its selling price. Use a 5% significance level.

X	Y
Year	**Price $**
2002	2,200
2009	13,900
2002	6,300
2010	13,000
1996	1,500
1998	2,400
2008	10,200
1999	3,400
2006	12,500



Setup: b) Determine if there is a significant negative linear correlation between number of homes on the market in a subdivision and average sale price at a 1% level of significance.

X	**Number of Homes on the Market**	26	6	17	24	4	17	25	2	5	15
Y	**Average Selling Price $**	217,500	231,000	221,946	219,873	238,045	221,670	218,999	240,367	253,622	228,642



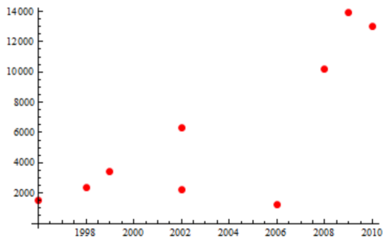
Using Calc – Test on Correlation Coefficient

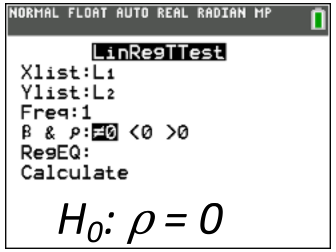
<u>GOAL</u>: Conduct a Test on the Correlation Coefficient!

1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinRegTTest
   a) Xlist = L_1
   b) Ylist = L_2
   c) Freq = 1
   d) β & ρ = Alternative Hypothesis ***
   e) RegEQ: *Leave blank for now*
   Calculate

Setup: a) Determine if there is a significant linear correlation between the model year of a Honda Civic being sold on Craigslist and its selling price. Use a 5% significance level.

X	Y
Year	**Price $**
2002	2,200
2009	13,900
2002	6,300
2010	13,000
1996	1,500
1998	2,400
2008	10,200
1999	3,400
2006	12,500

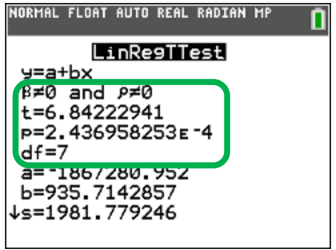




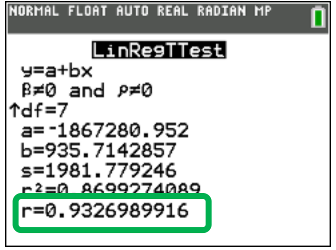
```
NORMAL FLOAT AUTO REAL RADIAN MP
        LinRegTTest
 Xlist:L1
 Ylist:L2
 Freq:1
 β & ρ:≠0 <0 >0
 RegEQ:
 Calculate
```

H_0: $\rho = 0$

H_A: $\rho \neq 0$



```
NORMAL FLOAT AUTO REAL RADIAN MP
        LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
 t=6.84222941
 p=2.436958253E-4
 df=7
 a=-1867280.952
 b=935.7142857
↓s=1981.779246
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
        LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑df=7
 a=-1867280.952
 b=935.7142857
 s=1981.779246
 r²=0.8699274089
 r=0.9326989916
```

<u>Calc Output we want</u>

β & ρ = Alternative hypothesis
t = t_{stat} (Test Statistic)
p = p-value
df = degrees of freedom (n – 2)

r = sample correlation

*** *<u>Alternative Hypothesis in Calc</u>*

ρ is the correlation between the X and Y values in the population
β is the true slope of our best line if we had population data

Recall the <u>correlation will always have the same sign as our slope!</u>

** <u>Data table organization</u>

Whenever data is presented in a table like it is here, generally speaking:

- For Vertical tables, the X variable will be on the left and Y on the right
- For Horizontal tables, the X will be the top and Y the bottom

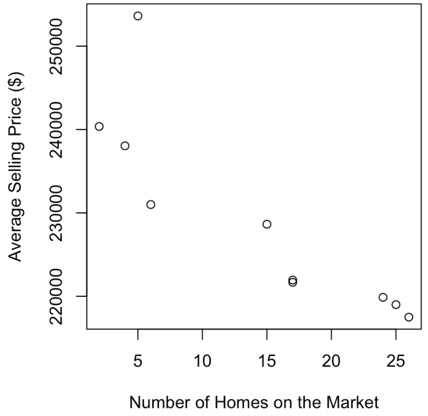
Using Calc – Test on Correlation Coefficient

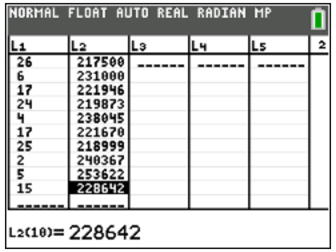
GOAL: Conduct a Test on the Correlation Coefficient!

1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinRegTTest
   a) Xlist = L_1
   b) Ylist = L_2
   c) Freq = 1
   d) β & ρ = Alternative Hypothesis ***
   e) RegEQ: *Leave blank for now*
   Calculate

Setup: b) Determine if there is a significant negative linear correlation between number of homes on the market in a subdivision and average sale price at a 1% level of significance.

X	**Number of Homes on the Market**	26	6	17	24	4	17	25	2	5	15
Y	**Average Selling Price $**	217,500	231,000	221,946	219,873	238,045	221,670	218,999	240,367	253,622	228,642

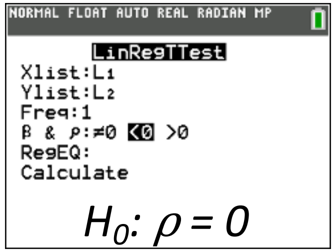



```
NORMAL FLOAT AUTO REAL RADIAN MP
L1     L2      L3      L4      L5      2
26     217500  ------  ------  ------
6      231000
17     221946
24     219873
4      238045
17     221670
25     218999
2      240367
5      253622
15     228642
------ ------
L2(10)= 228642
```

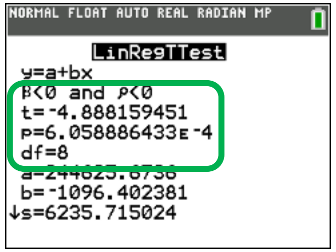
****** Alternative Hypothesis in Calc***

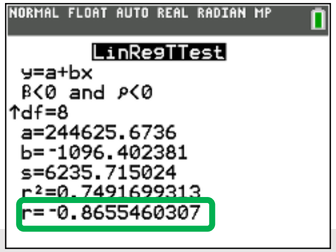
ρ is the correlation between the X and Y values in the population
β is the true slope of our best line if we had population data

Recall the correlation will always have the same sign as our slope!


```
NORMAL FLOAT AUTO REAL RADIAN MP
       LinRegTTest
 Xlist:L1
 Ylist:L2
 Freq:1
 β & ρ:≠0 <0 >0
 RegEQ:
 Calculate
```

H_0: $\rho = 0$
H_A: $\rho < 0$


```
NORMAL FLOAT AUTO REAL RADIAN MP
       LinRegTTest
 y=a+bx
 β<0 and ρ<0
 t=-4.888159451
 p=6.058886433E-4
 df=8
 a=244625.6736
 b=-1096.402381
↓s=6235.715024
```


```
NORMAL FLOAT AUTO REAL RADIAN MP
       LinRegTTest
 y=a+bx
 β<0 and ρ<0
↑df=8
 a=244625.6736
 b=-1096.402381
 s=6235.715024
 r²=0.7491699313
 r=-0.8655460307
```

** Data table organization

Whenever data is presented in a table like it is here, generally speaking:

- For Vertical tables, the X variable will be on the left and Y on the right
- For Horizontal tables, the X will be the top and Y the bottom

LCQ – Conclusions and Interpretations

Problem: Write the conclusions and interpretations for the previous scenarios using our results.

a) Use the Traditional Method
Setup: a) Determine if there is a significant linear correlation between the model year of a Honda Civic being sold on Craigslist and its selling price. Use a 5% significance level.

b) Use the P-Value Method
Setup: b) Determine if there is a significant negative linear correlation between number of homes on the market in a subdivision and average sale price at a 1% level of significance.

LCQ – Conclusions and Interpretations

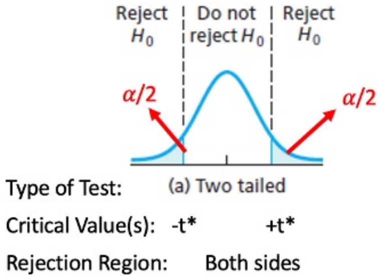
Problem: Write the conclusions and interpretations for the previous scenarios using our results.

a) Use the Traditional Method
Setup: a) Determine if there is a significant linear correlation between the model year of a Honda Civic being sold on Craigslist and its selling price. Use a 5% significance level. (Recall there was a sample size of n = 9)

Hypotheses
$H_0: \rho = 0$
$H_A: \rho \neq 0$

Test Statistic
Traditional Method → Need to find Critical Value first !!

CV = t* = invT(area = 0.05/2, df = 9 – 2) = -2.365
→ TWO tailed test, so we have to split alpha in two and we know that we will be comparing the absolute values of or CV and TS!



Type of Test: (a) Two tailed
Critical Value(s): -t* +t*
Rejection Region: Both sides

Showing work for Test Statistic we did earlier
Entered X data in L_1 and Y data in L_2
TS = t_{stat} = LinRegTTest(Xlist = L_1, Ylist = L_2, Freq = 1, β & ρ ≠ 0) = 6.842
df = 7

$|t_{stat}| = 6.842 > 2.365 = |t^*|$ → Reject H_0!

Conclusion and Interpretation

Since the absolute value of our TS t_{stat} = 6.82 is greater than the absolute value our CV t = 2.365, we reject the null hypothesis.*

We have sufficient evidence to conclude that there is a significant (positive) linear correlation between the model year of a Honda Civic and its selling price → PERFECT

- *Remember to put the values for the TS and CV in our conclusion, be specific!*
- *Our alternative is that there is a significant linear correlation!! And it is between our two variables IN CONTEXT!!!!! Don't just say X and Y*
  - *We could also say significant linear relationship, both would be correct!*
- *Additionally, we can write the extra info that this is a positive linear relationship, which we know because the Test Statistic was positive (and the sample correlation = 0.933)*

LCQ – Conclusions and Interpretations

Problem: Write the conclusions and interpretations for the previous scenarios using our results.

b) Use the P-Value Method

Setup: b) Determine if there is a significant negative linear correlation between number of homes on the market in a subdivision and average sale price at a 1% level of significance.

Hypotheses

H_0: $\rho = 0$

H_A: $\rho < 0$

P-Value Method

Showing work for p-value we did earlier

Entered X data in L_1 and Y data in L_2

P-value = LinRegTTest(Xlist = L_1, Ylist = L_2, Freq = 1, β & $\rho < 0$) ≈ 0

TS = $t_{stat} = -4.888$

df = 7

p-value $\approx 0 < 0.01 = \alpha$ → Reject H_0!

Really Small Probabilities

Remember what the notation our calculator uses for really small numbers!

P-value = 6.059 E-4

= 6.059×10^{-4}

= 6.059/10000

= 0.0006059

≈ 0

Conclusion and Interpretation

Since the p-value ≈ 0 is less than the significance level = 0.01, we reject the null hypothesis.

We have sufficient evidence to conclude that there is a significant negative linear correlation between the number of homes for sale in a neighborhood and the average selling price → PERFECT

- *Remember to put the values for the p-value and significance level, be specific!*
- *Again, USE CONTEXT!!!!! Don't just say X and Y*
- *And negative linear relationship because of the negative TS and r = -0.866*

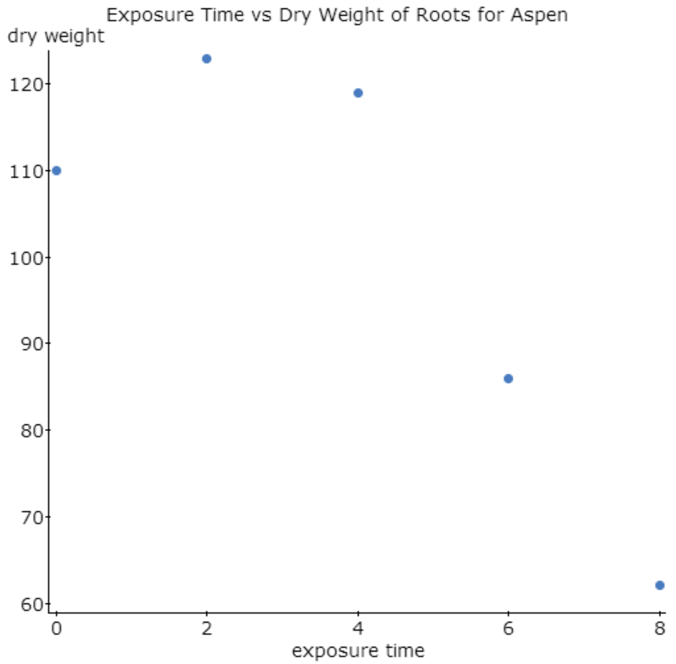
Problem Session!!!

Example 1

- The article “Effects of Gamma Radiation on Juvenile and Mature Cuttings of Quaking Aspen” (Forest Science [1967]: 240-245) reported data on x = exposure time to radiation and y = dry weight of roots.

a) Does there seem to be a positive association, a negative association, or no association?

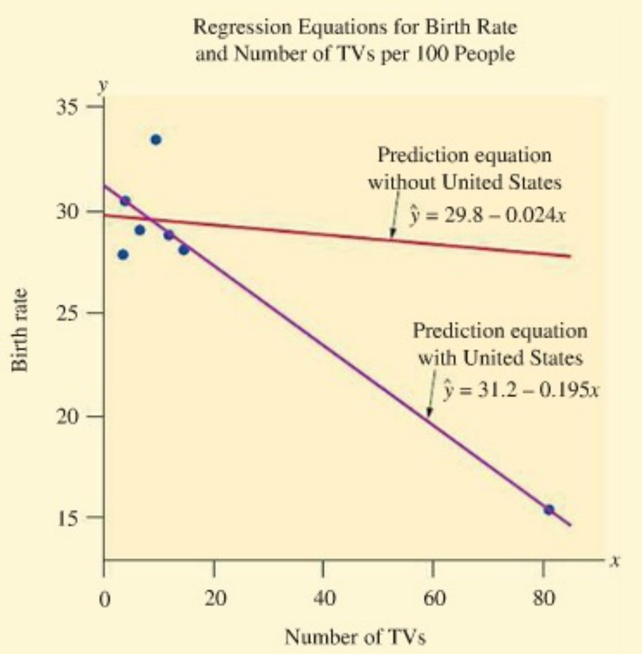
b) Can the trend in the data points be approximated reasonably well by a straight line?



TV Watching and Birth Rate

The figure shows recent data on x = the number of TVs per 100 people and y = the birth rate (number of births per 1000 people) for 6 African and Asian nations. The regression line $\hat{y} = 29.8 - 0.024x$ applies to the data for these six countries. For illustration, another point is added at (81, 15.2) which is the observation for the United States. The regression line for all 7 points is $\hat{y} = 31.2 - 0.195x$.

- Does the U.S. observation appear to be an outlier on x? an outlier on y? or both?



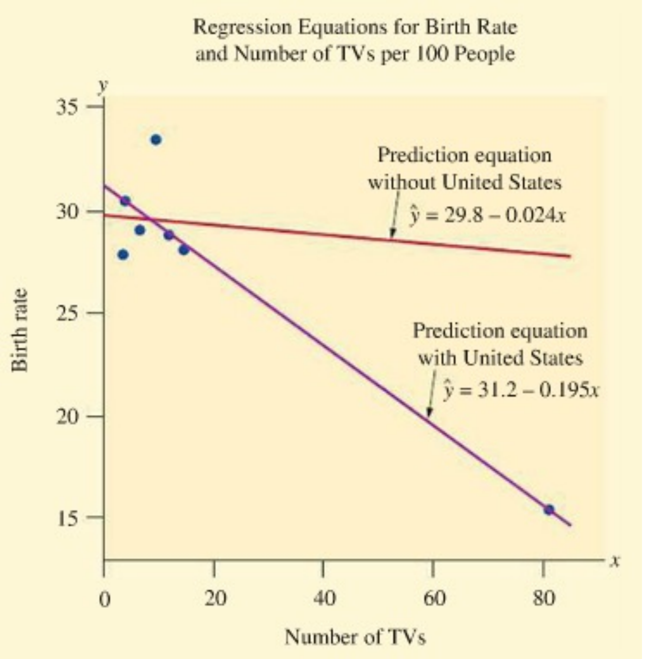
Agresti, A., & Franklin, C. (2013). *Statistics: The Art and Science of Learning from Data*, 3rd edition. Boston, MA: Pearson.

TV Watching and Birth Rate

The figure shows recent data on x = the number of TVs per 100 people and y = the birth rate (number of births per 1000 people) for 6 African and Asian nations. The regression line $\hat{y} = 29.8 - 0.024x$ applies to the data for these six countries. For illustration, another point is added at (81, 15.2) which is the observation for the United States. The regression line for all 7 points is $\hat{y} = 31.2 - 0.195x$.

- Does the U.S. observation appear to be an outlier on x? an outlier on y? or both?

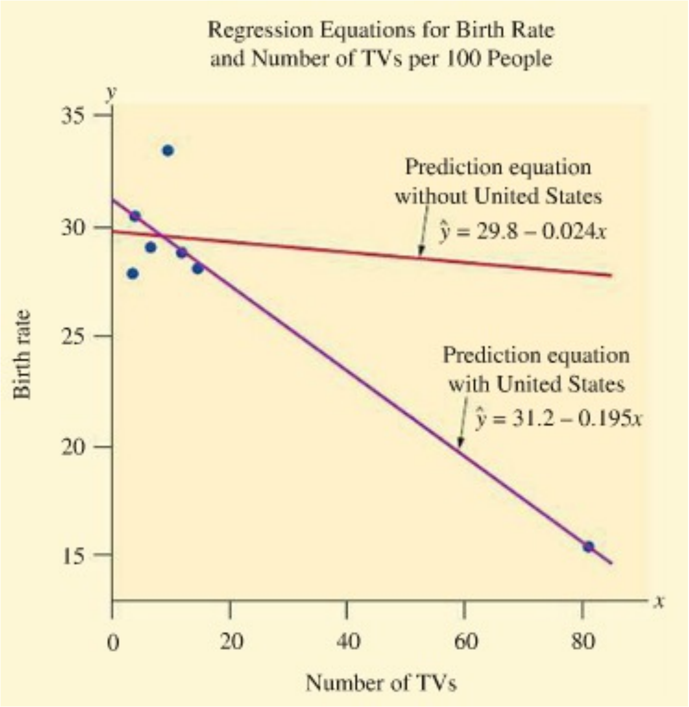
BOTH, very far in the X and Y from the rest of the data



Agresti, A., & Franklin, C. (2013). *Statistics: The Art and Science of Learning from Data*, 3rd edition. Boston, MA: Pearson.

TV Watching and Birth Rate, p.4

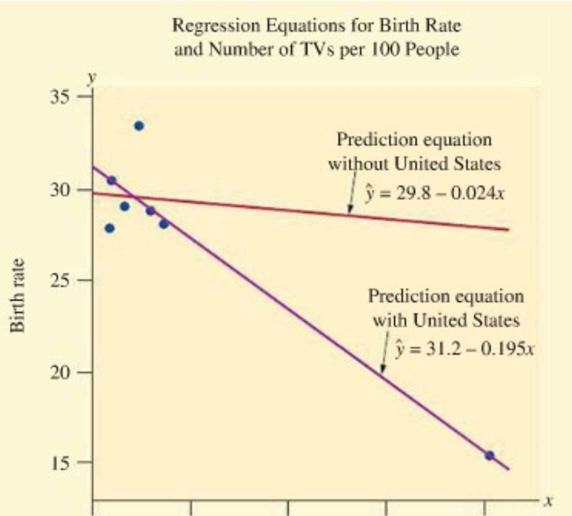
The correlation r = -0.051 without the U.S. but with the U.S., r = -0.935. Why you would conclude that the association between birth rate and number of TVs is very weak without the U.S. point and very strong with the U.S. point?



TV Watching and Birth Rate, p.5

The correlation r = -0.051 without the U.S. but with the U.S., r = -0.935. Why you would conclude that the association between birth rate and number of TVs is very weak without the U.S. point and very strong with the U.S. point?

- The association is very weak w/o the U.S. because the 6 countries all have very few TVs. The association is very strong w/the U.S. because the U.S. is so much higher in #of TVs and so much lower on birth rate that it makes the 2 variables seem related.



Problem #3

The human resources department at a large multinational corporation wants to be able to predict average salary for a given number of years' experience. Data on salary (in $1000s) and years of experience were collected for a sample of employees.

a) Which variable is the explanatory or predictor variable?

b) Which variable is the response variable?

c) Which variable would you plot on the y axis?

Problem #3 Solution

The human resources department at a large multinational corporation wants to be able to predict average salary for a given number of years' experience. Data on salary (in $1000s) and years of experience were collected for a sample of employees.

a) Which variable is the explanatory or predictor variable? **Years of experience**

b) Which variable is the response variable? **Salary**

c) Which variable would you plot on the y axis? **Salary**

Problem #1

Consider the data from a small bookstore.

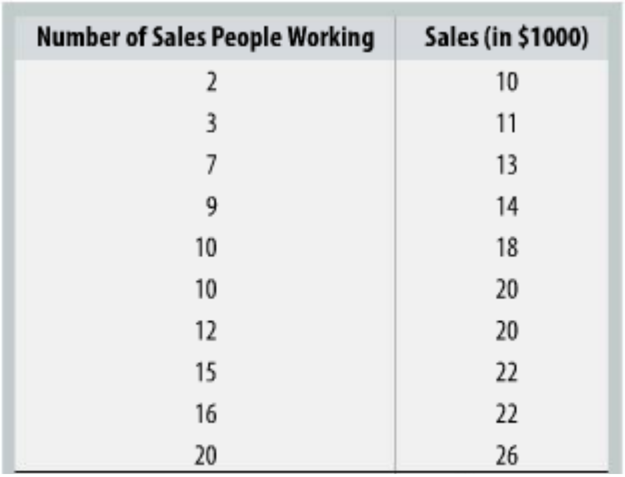
a) Prepare a scatterplot of Sales against Number of Sales People Working.

b) What can you say about the direction of the association?

c) What can you say about the form of the relationship?

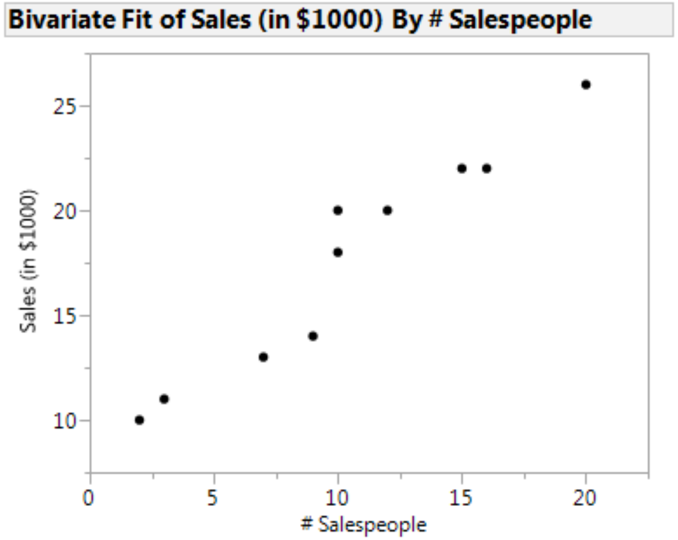
d) What can you say about the strength of the relationship?

e) Does the scatterplot show any outliers?



Number of Sales People Working	Sales (in $1000)
2	10
3	11
7	13
9	14
10	18
10	20
12	20
15	22
16	22
20	26

Problem #1 Solution



b) There is a **positive association** between number of sales people working and sales.

c) The relationship appears to be **linear**.

d) The relationship appears to be **strong**.

e) The scatterplot does **not** appear to show any **outliers**.

Problem #7

A larger firm is considering acquiring the bookstore of Exercise 1. An analyst for the firm, noting the relationship see in Exercise 1, suggests that when they acquire the store they should hire more people because that will drive higher sales.

- Is his conclusion justified?
- What alternative explanations can you offer?

Problem #7 Solution

Correlation does not imply causation. The analysts argument is that sales staff cause sales. However, the data may reflect the store hiring more people as sales increase, so any causation would run the other way.

Problem #25

Which of the scatterplots show:

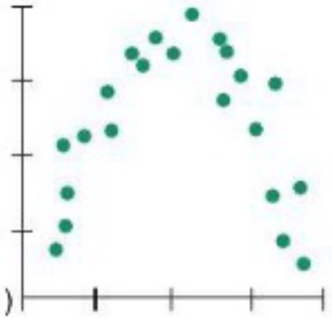
a) Little or no association?

b) A negative association?

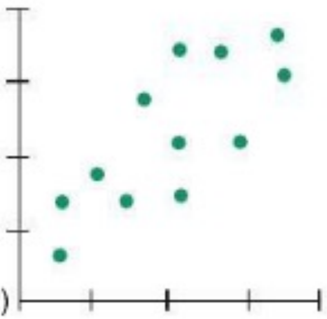
c) A linear association?

d) A moderately strong association?

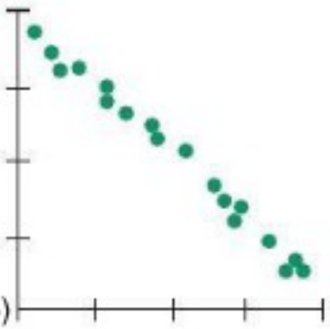
e) A very strong association?



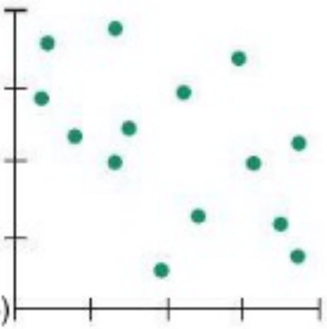
(1)



(2)



(3)



(4)

Problem #25 Solution

Which of the scatterplots show:

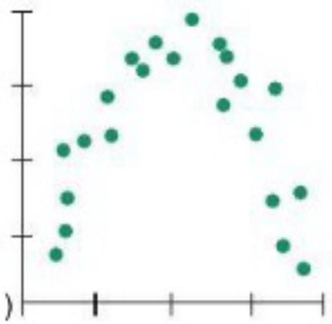
a) Little or no association? **none**

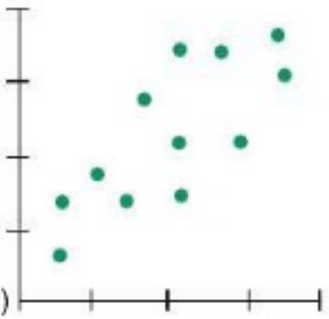
b) A negative association? **3 and 4**

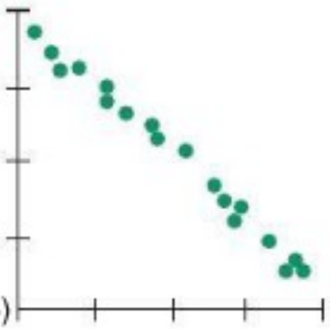
c) A linear association? **2, 3, and 4**

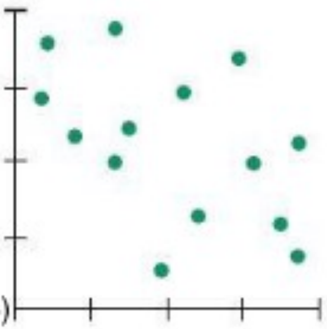
d) A moderately strong association? **2 and 4**

e) A very strong association? **1 and 3**


(1)


(2)


(3)


(4)

Problem #29

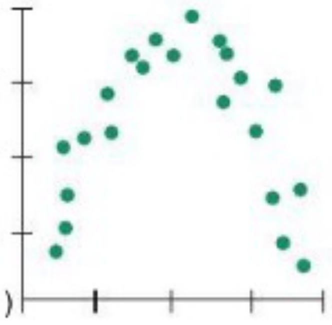
Match the scatterplots to the correlation coefficients. The calculated correlation coefficients are:

a) -0.923

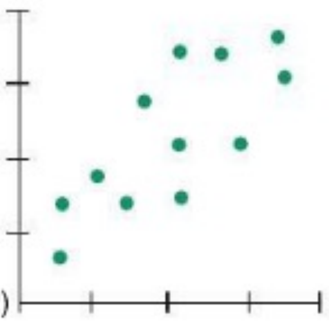
b) -0.487

c) 0.006

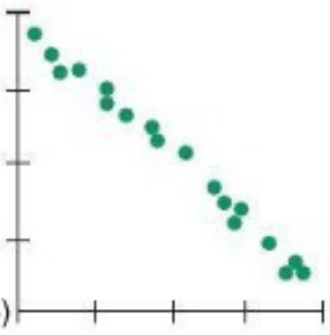
d) 0.777



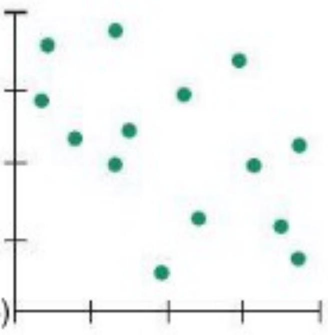
(1)



(2)



(3)



(4)

Problem #29 Solution

Match the scatterplots to the correlation coefficients.
The calculated correlation coefficients are:

a) -0.923 matches Graph 3

b) -0.487 matches Graph 4

c) 0.006 matches Graph 1

d) 0.777 matches Graph 2