Almost Almost There!!

Unit 9 – Correlation and Regression

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

Unit 9 - Outline

Correlation and Regression

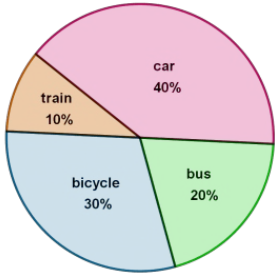
- Intro
- Scatterplots
- Correlation
- Regression Line
- Predictions
- Evaluating Predictions

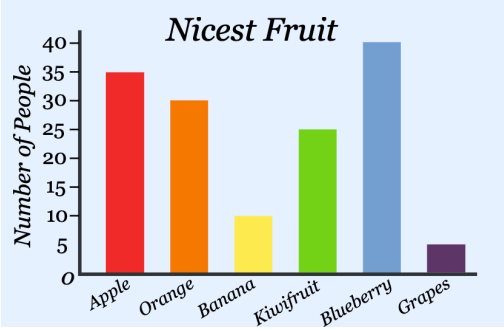
Review + New

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

Qualitative (Categorical) Data

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

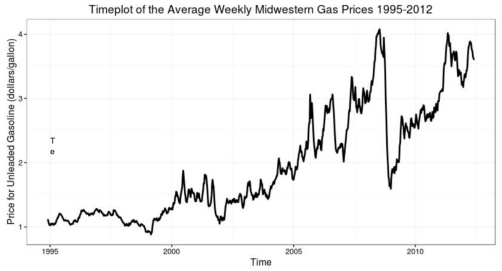




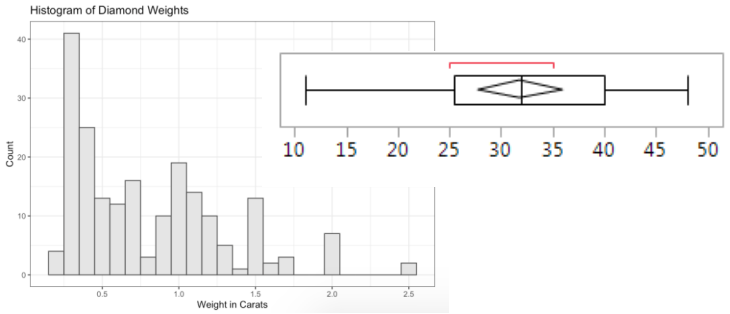
- Here we can describe the mode (the most common category)

Quantitative Data

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



- With line graphs, we described the trend and seasonal variation



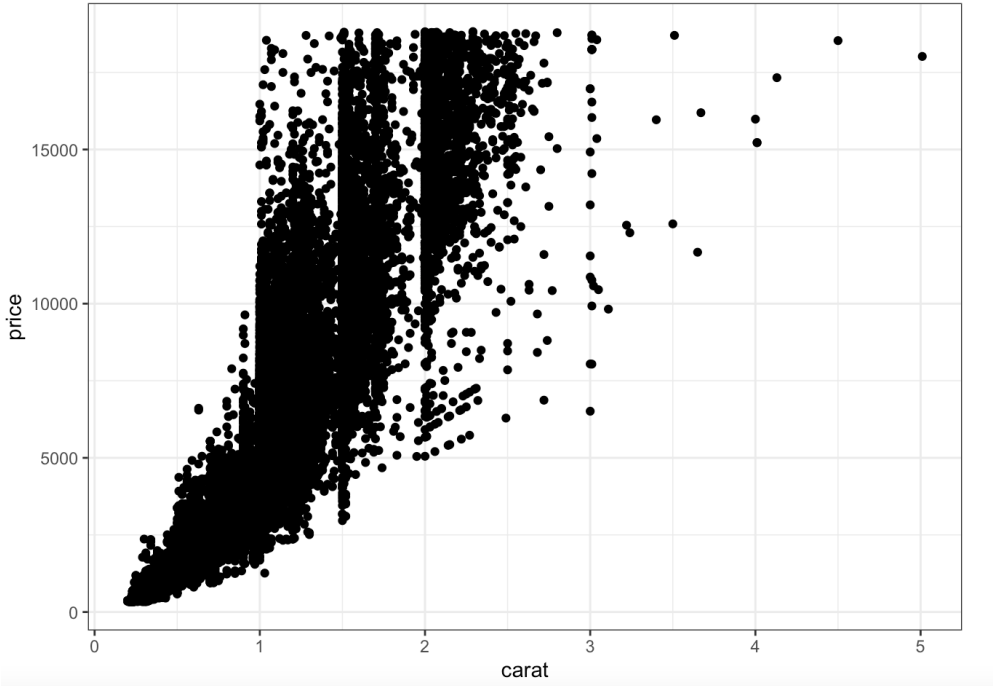
- With histograms and boxplots, we described the SOCS! Shape, Outliers, Center and Spread
- Also numerical summaries like mean, median, SD, IQR, etc.

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

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.

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

b) The weight and height of a person

c) The amount of yearly rainfall and crop yield

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

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.

<u>Overall Patter of a Scatterplot</u>

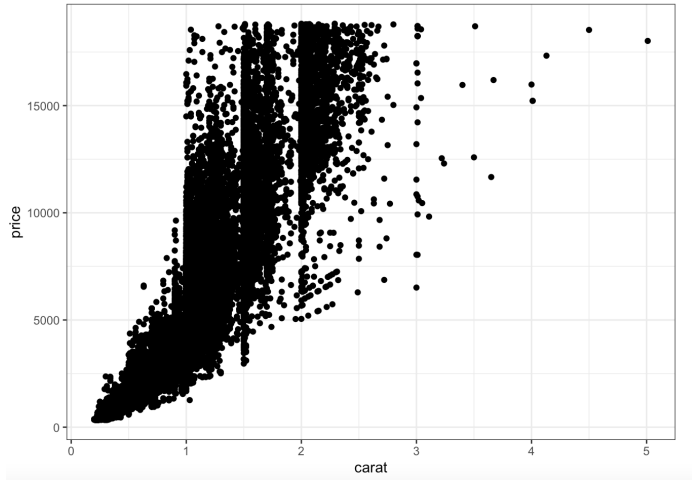
- We can describe the overall pattern of a scatterplot by looking at four characteristics:
    1) Form
    2) Direction
    3) Strength
    4) Outliers

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

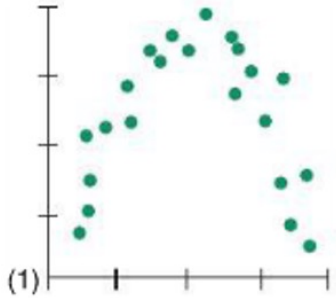
- Points are showing some evidence of curvature
- NOT a straight line → Any type of CLEAR curvature in the “best” line we could

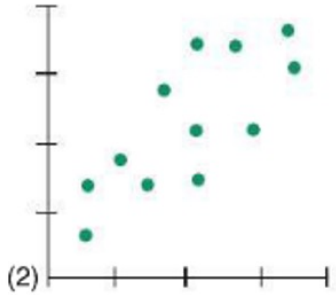
Random scatter

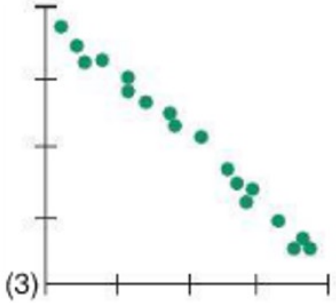
- There really is no pattern, points are just scattered about randomly kinda like a cloud of points

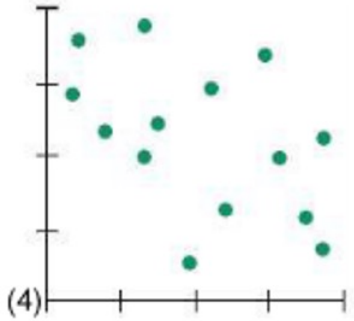
LCQ: Form

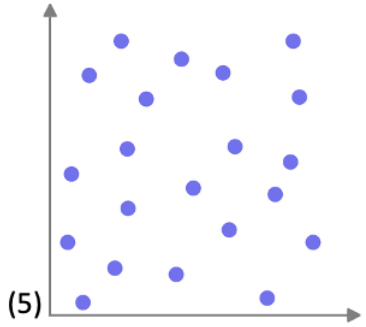
Problem: Determine the form for each of the following scatterplots.





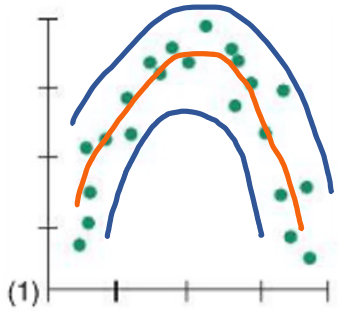






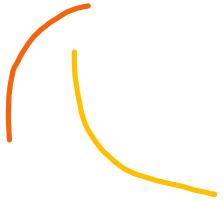
LCQ: Form

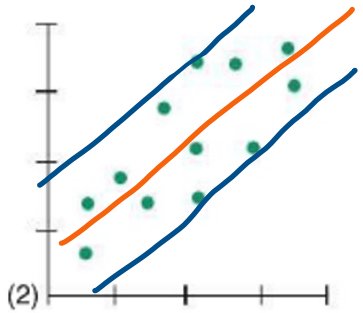
Problem: Determine the form for each of the following scatterplots.



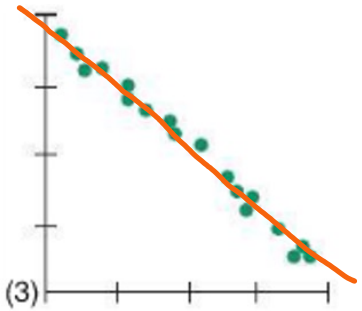
Curved → CLEAR curvature if we draw that line through the middle and if we trace the lower and upper edges

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

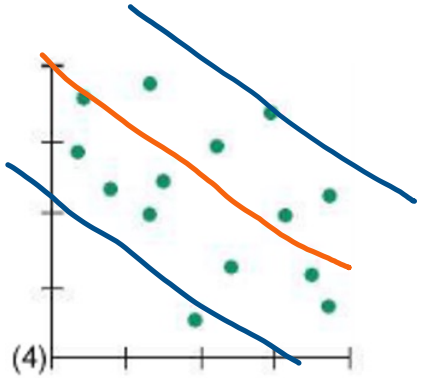




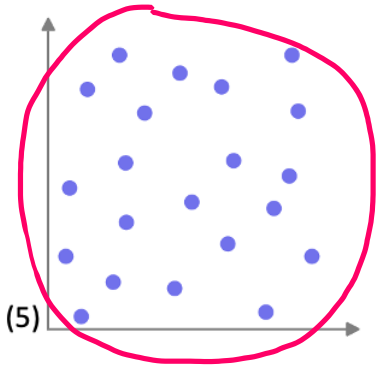
Linear → for the most part follows the straight diagonal line through the middle



Linear → dots more clearly follows a linear path



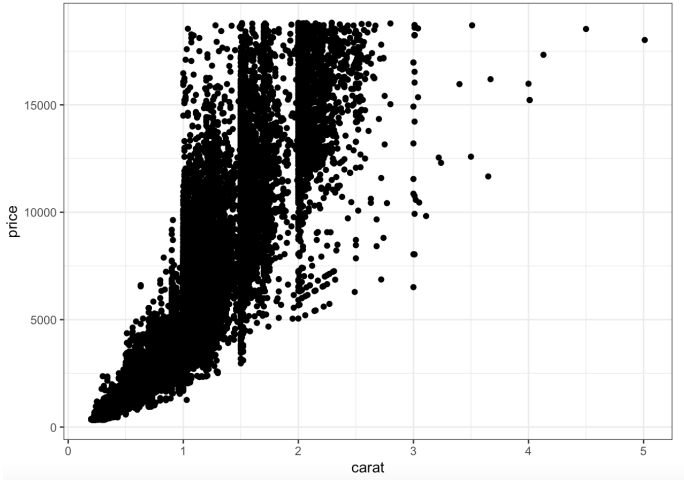
Linear-ish → not perfect (or as clear as the others), but these dots generally follow straight line path with the middle and outside lines



Random → there is no patter here, just scattered everywhere

Direction

Direction



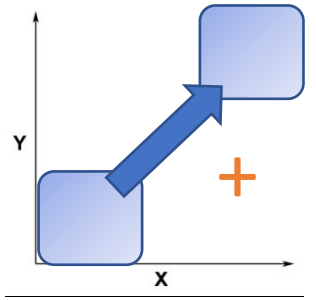
- This refers to the **direction** of the association between the two variables

** 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!

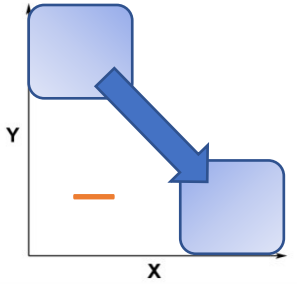
Positive Association

- 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

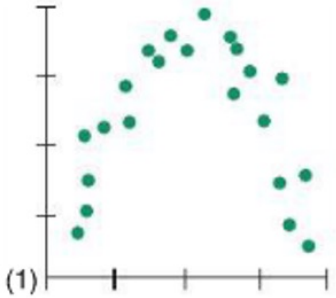


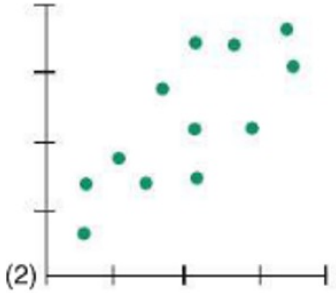
No Association

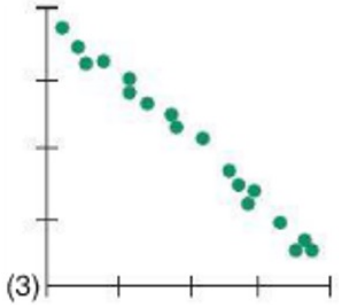
- There is no pattern or general trend
- Low values of one variable can have both high or low values of the other; high values of one variable can have both high or low values of the other
- Knowing if we have a high or low value of one variable gives us no indication whether the other variable's value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

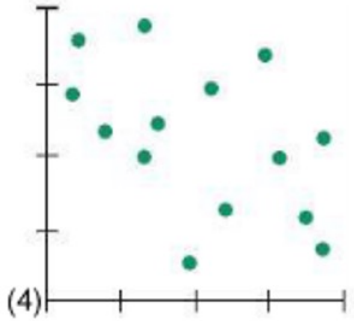
LCQ: Direction

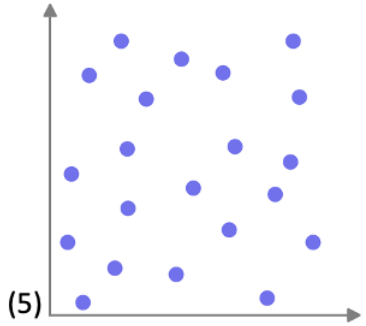
Problem: Determine the direction for each of the following scatterplots.





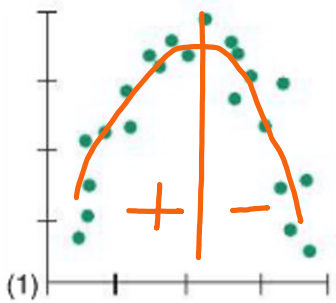






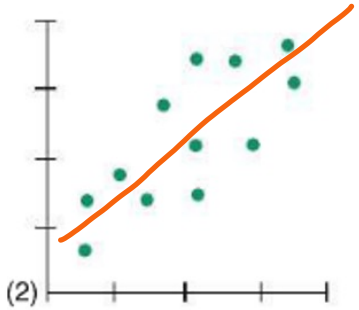
LCQ: Direction

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

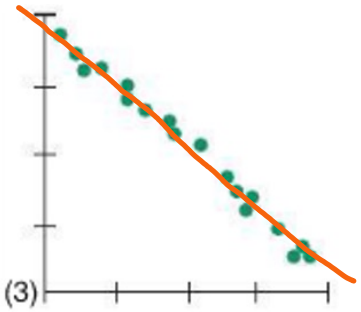


Starts positive, then goes negative; Not applicable → If we divide this plot in two, there are two different associations! This is because of the curved form.

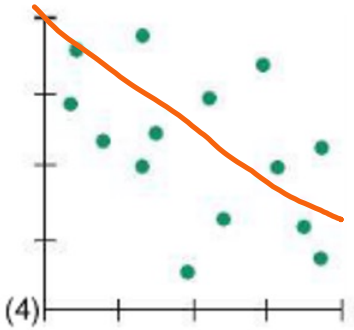
- *We are only going to apply direction to linear forms, so this wouldn't fit*



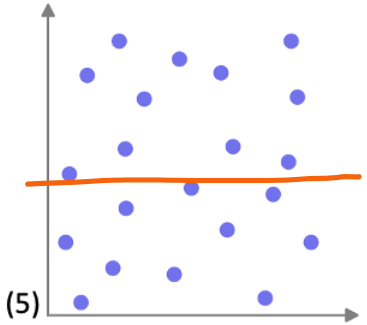
Positive → definitely a positive slope to the "best" line we could draw. And we see the points shifting in the same direction, both increasing together



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

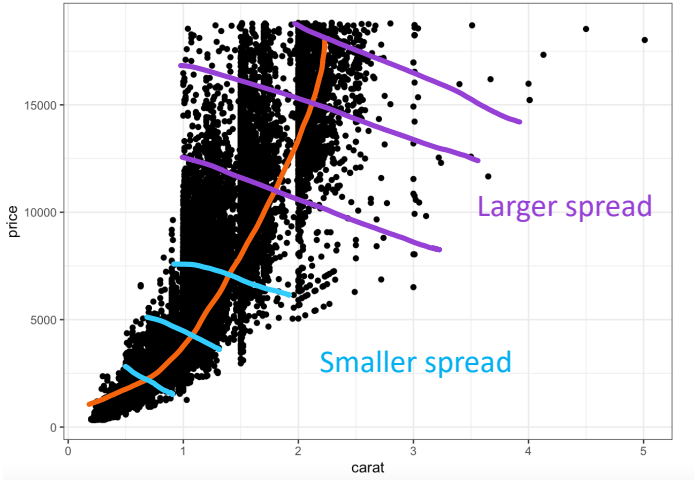
Strength

Strength

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

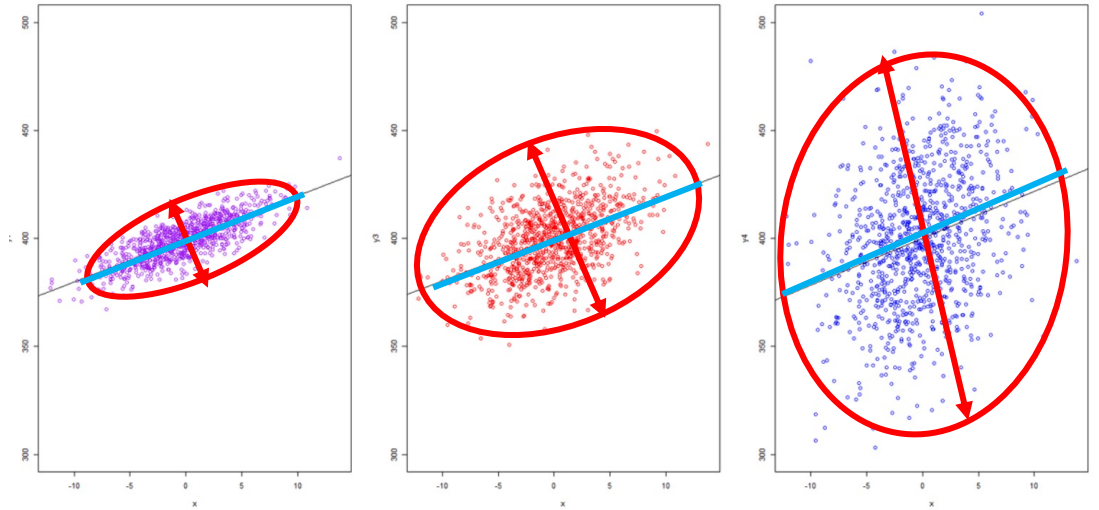
- We can think of this as how CLOSE the data are to our “best” line (the form)!
  - So how small or large the spread is around our “best” line

- 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:



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

High Strength (Strong)

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

Moderate Association

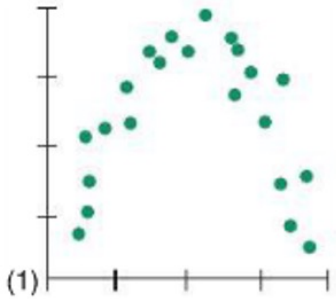
- The dots follow a general pattern, but not as tightly packed
- Relatively medium width of our oval, still looks like an oval

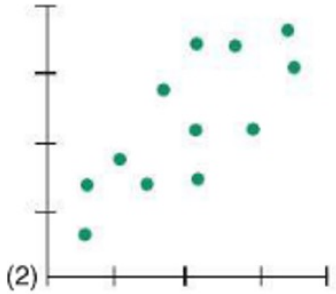
Low Strength (Weak)

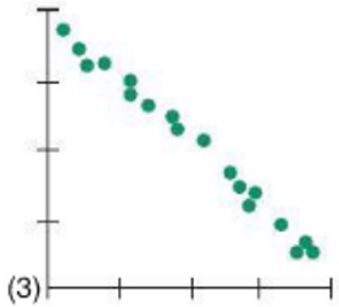
- The dots do not appear to be following a pattern
- Relatively large width of our oval, which starts looking more like a circle

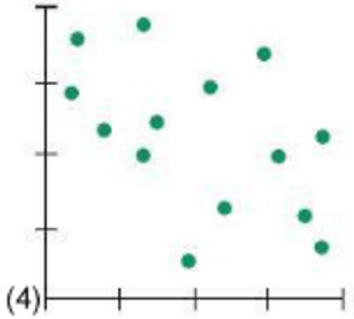
LCQ: Strength

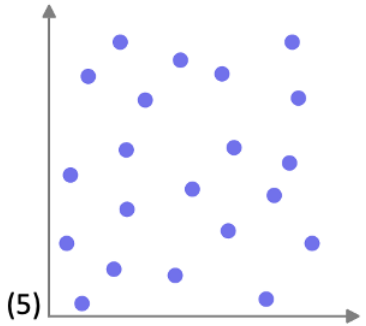
Problem: Determine the strength for each of the following scatterplots.





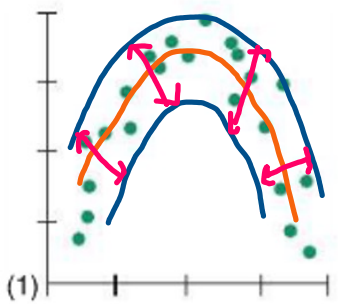




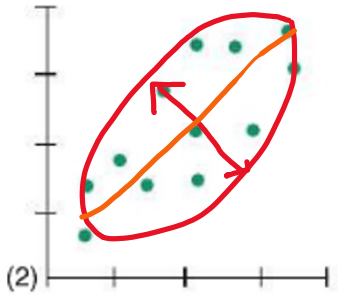


LCQ: Strength

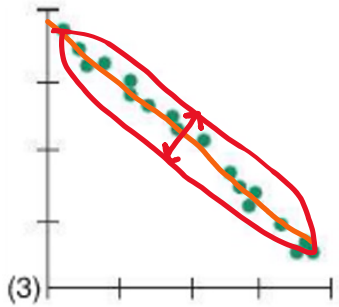
Problem: Determine the <u>strength</u> for each of the following scatterplots.



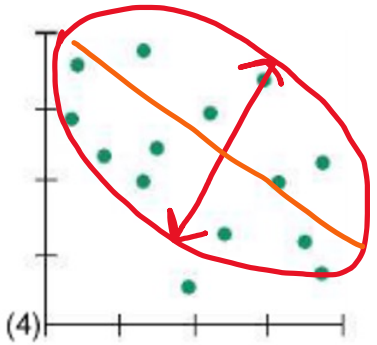
Strong, but not applicable → Points definitely follow this pattern very well, indicating a high strength. But because it has a curved form, we are not going to apply our definition of strength



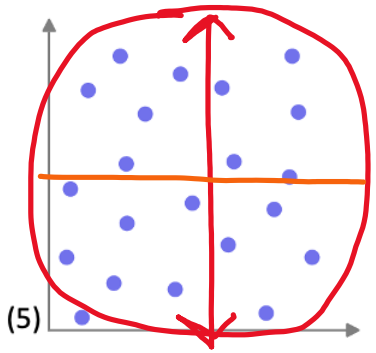
Moderate → Fairly smallish width of our oval, not too much spread around the "best" line



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),

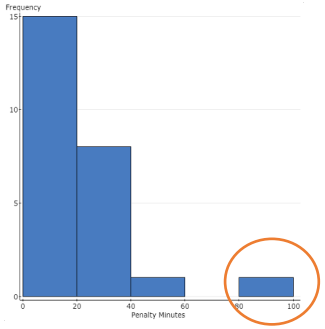


Super weak→ almost have a circle / rectangular-ish shape here because of the lack of association, very large spread too

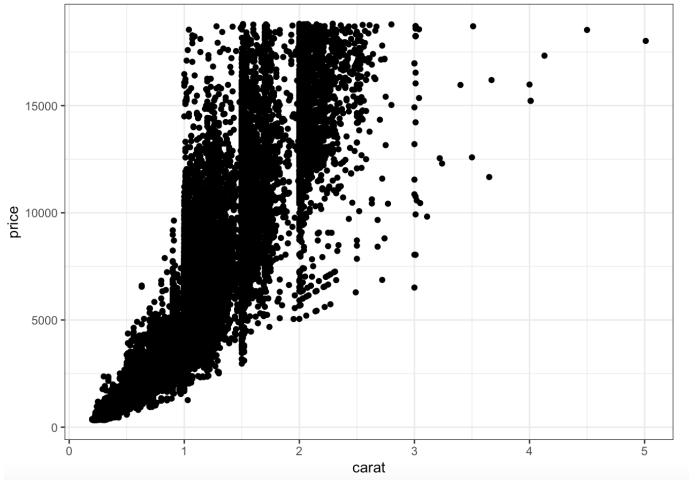
Outliers

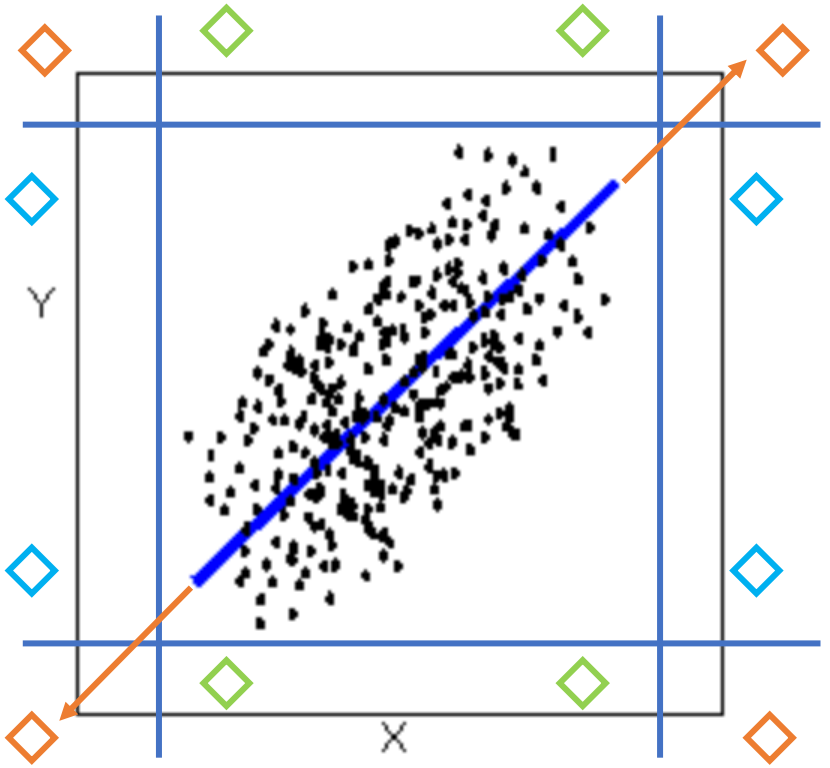
Outliers

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



- 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





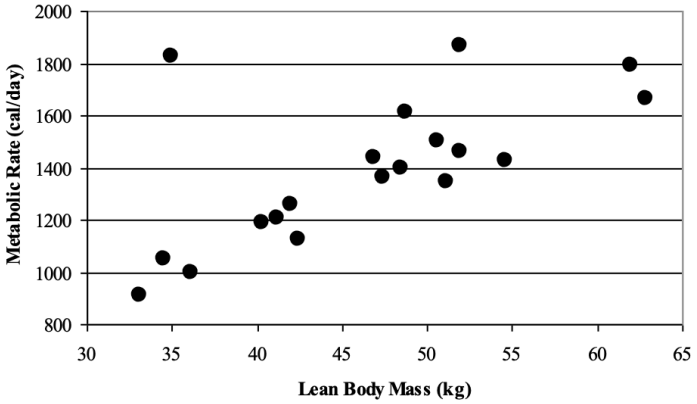
https://bolt.mph.ufl.edu/6050-6052/unit-1/case-q-q/scatterplots/

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

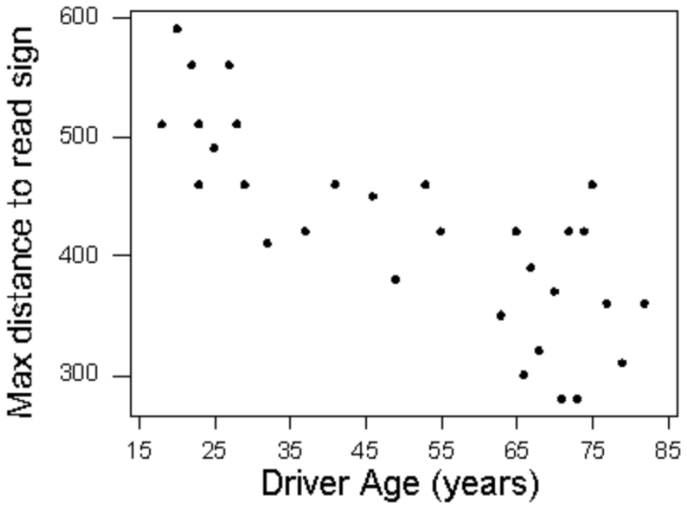
LCQ: Interpreting Scatterplots

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

a)



b)

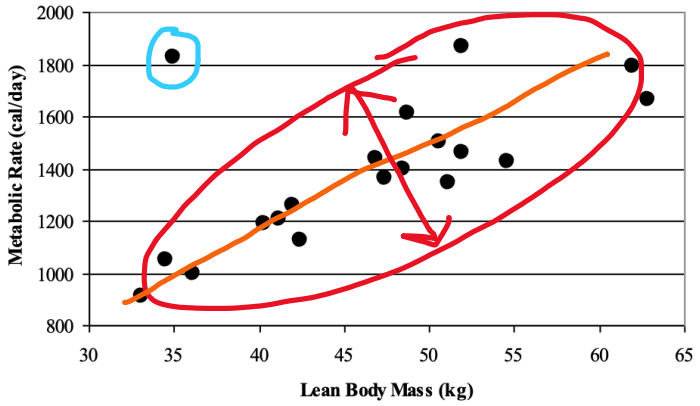


https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/scatterplots-3-of-5/

LCQ: Interpreting Scatterplots

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

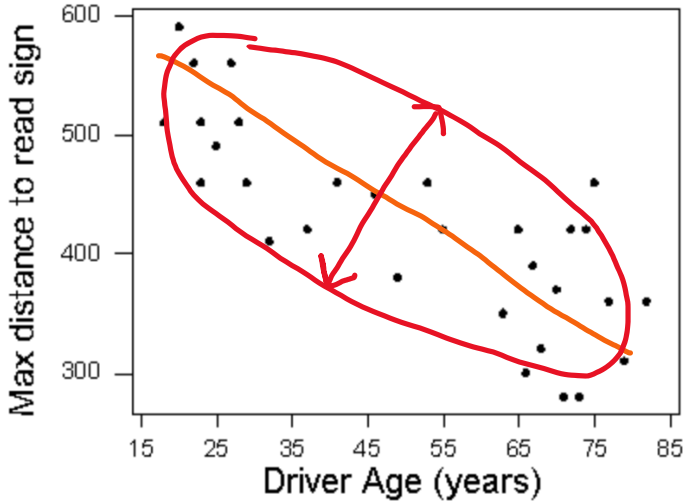
a)



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)



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

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

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 → $-1 \leq r \leq 1$

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.

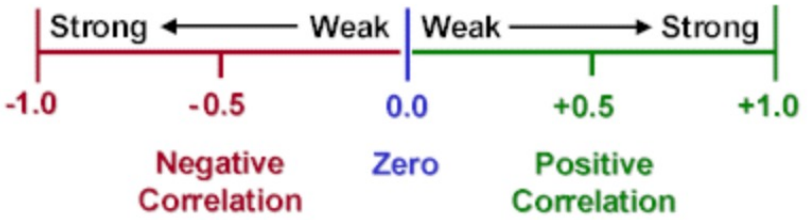
Properties of Correlation

- Both variables have to be quantitative
- *r* measures the strength of linear relationships
- *r* has no units of measurement
- Changing 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

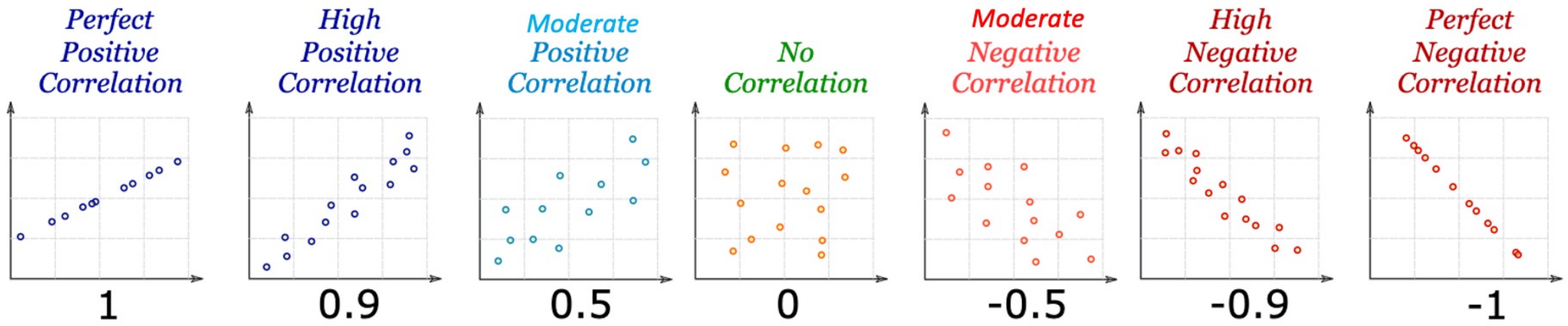
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- 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

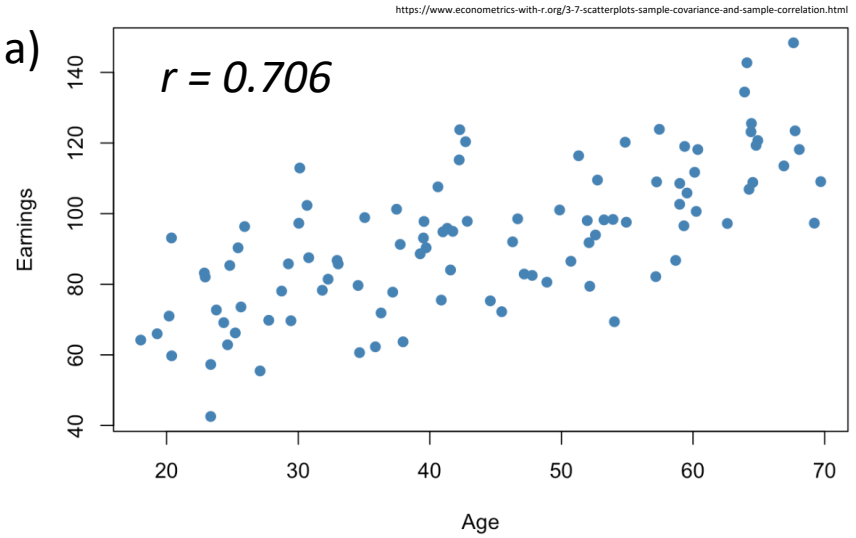
LCQ: Interpreting Correlation

How to Interpret Correlation

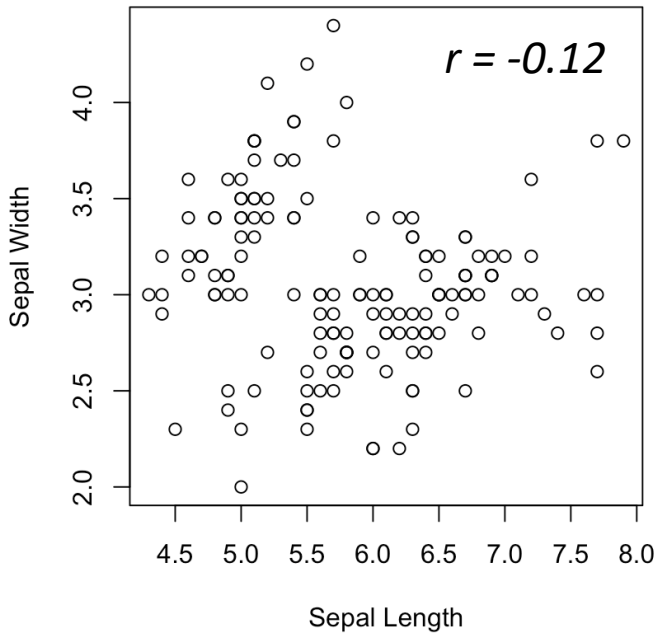
- Have to mention the **magnitude** (which refers to the strength of linear relationship between X and Y) and **direction** (+/-)
- And **USE CONTEXT**!!!
- General structure → There is a < strong, moderate, weak >, < positive / negative > linear relationship between < X and Y >

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

a)



b)



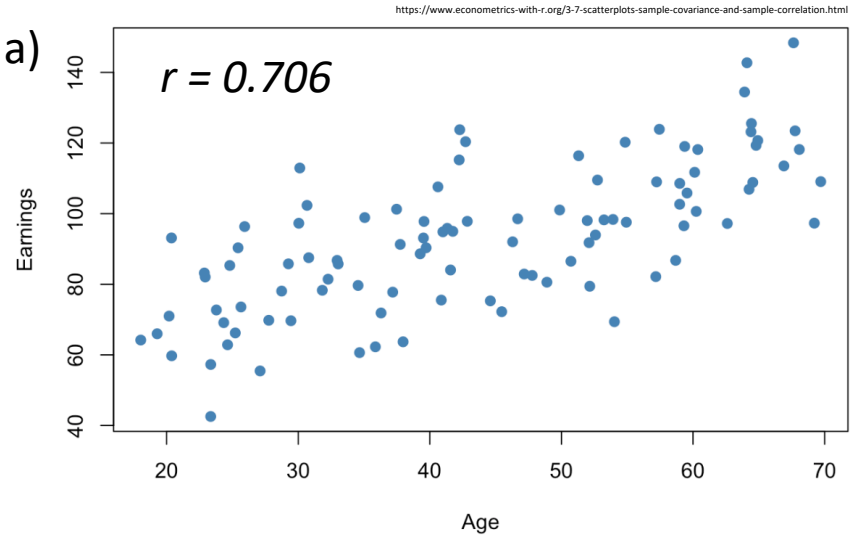
LCQ: Interpreting Correlation

How to Interpret Correlation

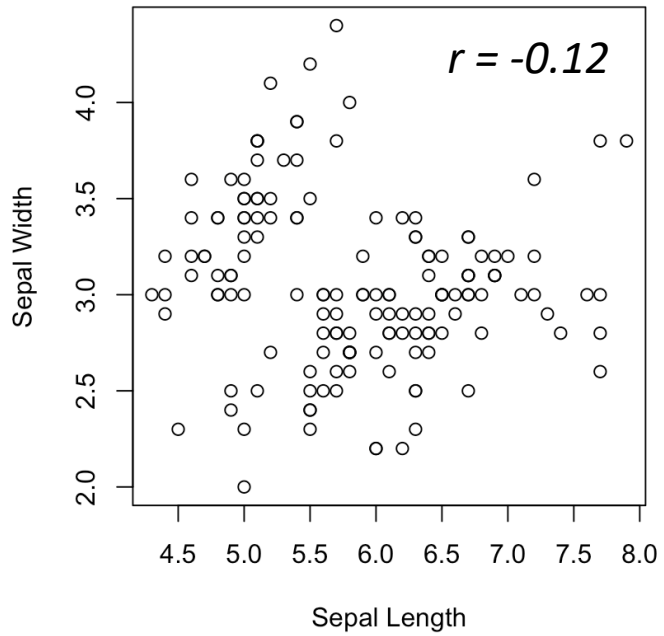
- Have to mention the **magnitude** (which refers to the strength of linear relationship between X and Y) and **direction** (+/-)
- And **USE CONTEXT**!!!
- General structure → There is a < strong, moderate, weak >, < positive / negative > linear relationship between < X and Y >

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

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

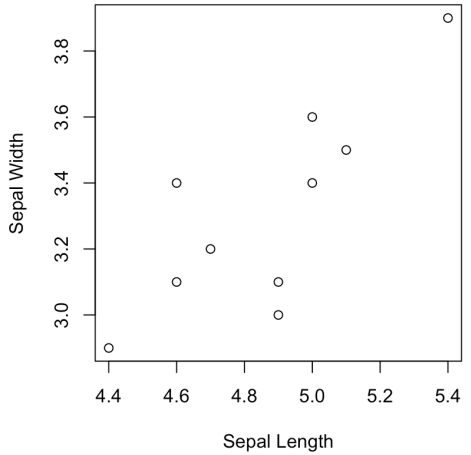
*** *One time setup: 2nd → Catalog → DiagnosticOn → Enter*

GOAL: Calculate the Correlation Coefficient!

1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinReg(a+bx) *STAT → CALC → LinReg(a+bx)*
   a) Xlist = L_1
   b) Ylist = L_2
   c) FreqList: *Leave blank*
   d) Store RegEQ: *Leave blank*

   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

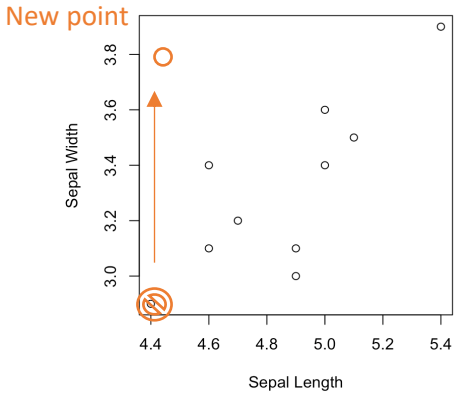


Outliers Demonstration

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

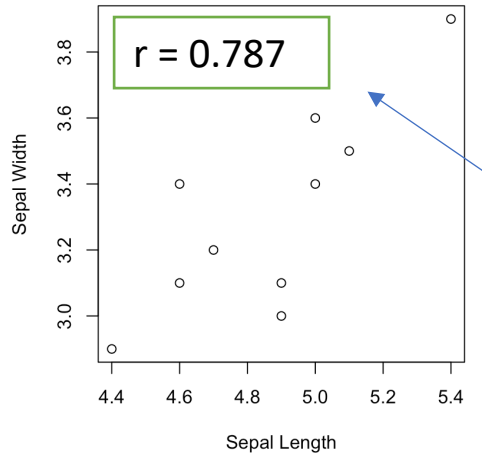
*** One time setup: 2nd → Catalog → DiagnosticOn → Enter

GOAL: Calculate the Correlation Coefficient!

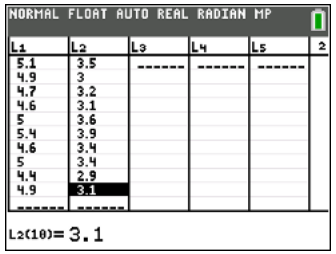
1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinReg(a+bx) *STAT → CALC → LinReg(a+bx)*
   a) Xlist = L_1
   b) Ylist = L_2
   c) FreqList: *Leave blank*
   d) Store RegEQ: *Leave blank*

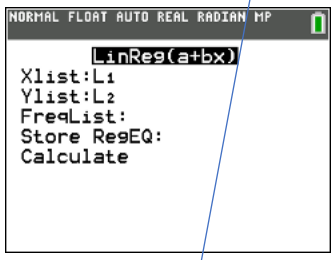
   Calculate

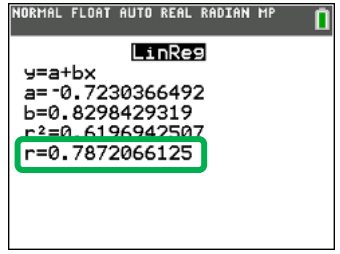
X	Y
Sepal Length	Sepal Width
5.1	3.5
4.9	3
4.7	3.2
4.6	3.1
5	3.6
5.4	3.9
4.6	3.4
5	3.4
4.4	2.9
4.9	3.1



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







r = Correlation

This is the only output we need for now!

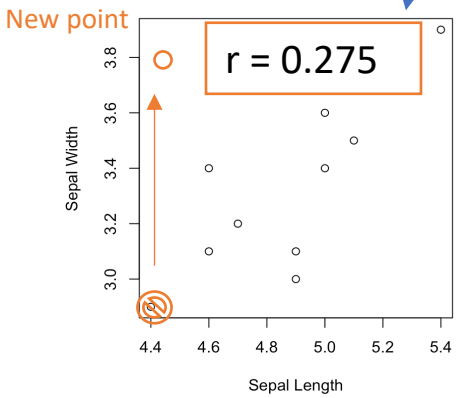
Show work: LinReg(a+bx)(Xlist = L1, Ylist = L2)

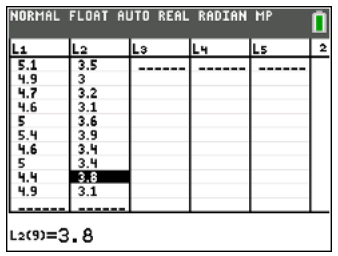
Outliers Demonstration

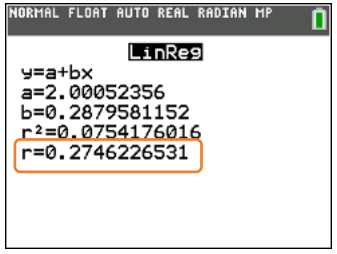
Let's change one data point to see the effects on the correlation:
- 9th observation: (4.4, 2.9) → (4.4, 3.8)

Now recalculate the correlation!

New point



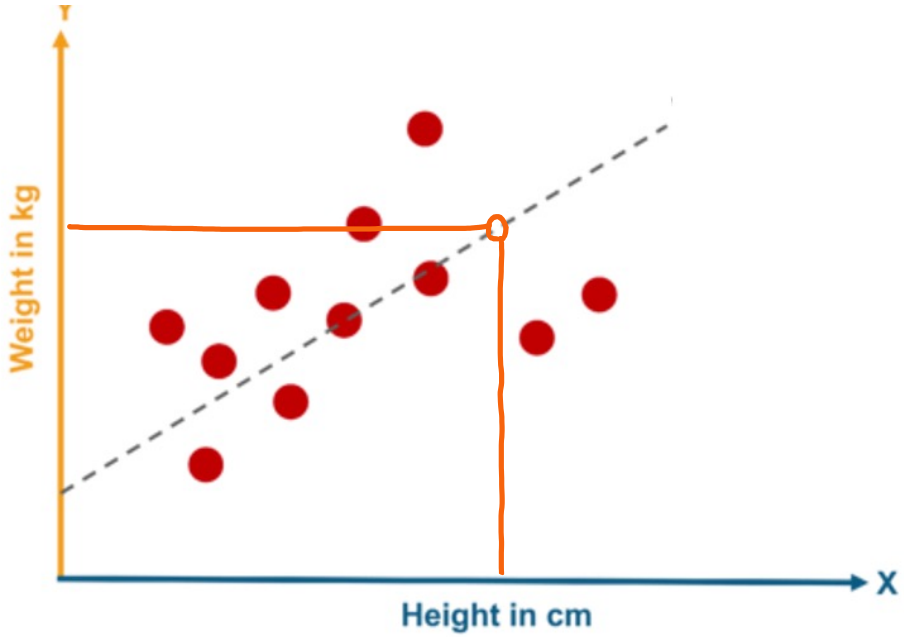




REGRESSION

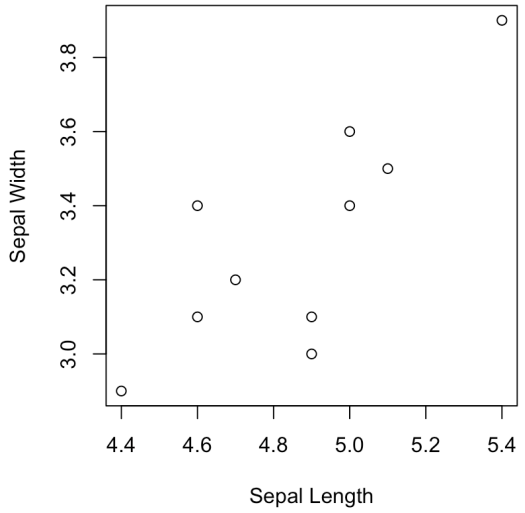
Motivation

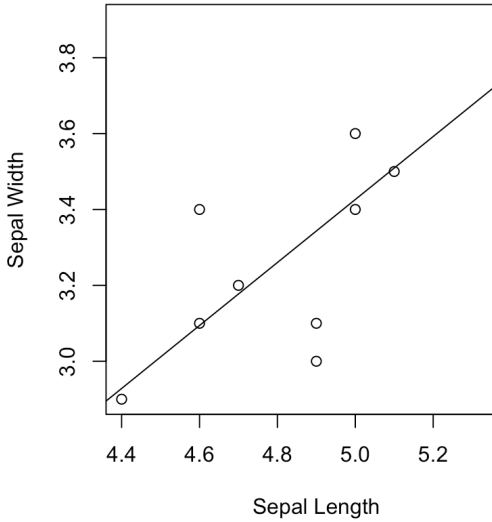
- In the case where our data does show evidence of a significant linear correlation, we would like to **model that relationship**!
- Modeling the relationship will allow us to **predict** Y values for new X values.
- The process is called **linear regression**.



https://www.edureka.co/blog/linear-regression-in-python/

Regression Line





Regression Line

- The **regression line** is a linear equation that *fits our data best*
  - Also called the "line of best fit"
- There is ONLY one "best" line for every dataset!
  - Technically, this is the line that minimizes the sum of the vertical distances from the actual data points to the best fit line.

Equation

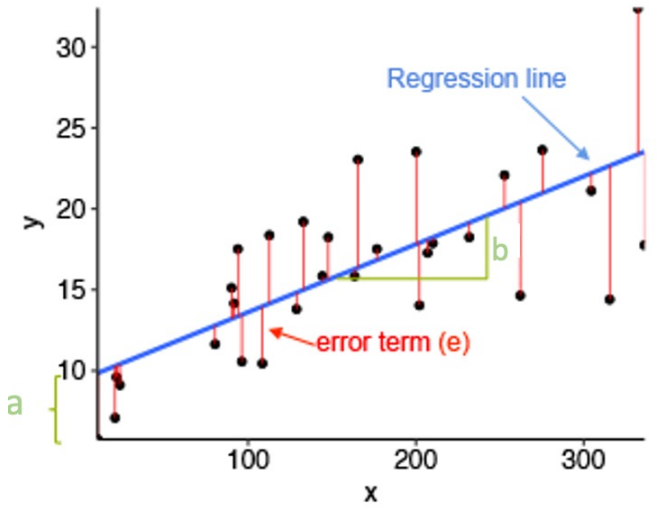
- Here is the form of our linear equation (written in slope-intercept form):

$$\hat{Y} = a + bX$$

It's important to get the X and Y variables correct or else our equation's variables will be backwards!

Predicted Sepal Width = a + b (Sepal Length)

$(\widehat{Sepal\ Width})$



https://www.analyticsvidhya.com/blog/2021/06/25-questions-to-test-your-skills-on-linear-regression-algorithm/

- a = Y intercept
  - It is the location where the regression line crosses the Y-axis (value of Y when X = 0)
- b = Slope
  - It measures the direction and steepness of the line
- X = Value of the explanatory variable
  - Doesn't have to be an X value that was included in the sample data
- $\hat{Y}$ = Predicted value of the response variable for the given X

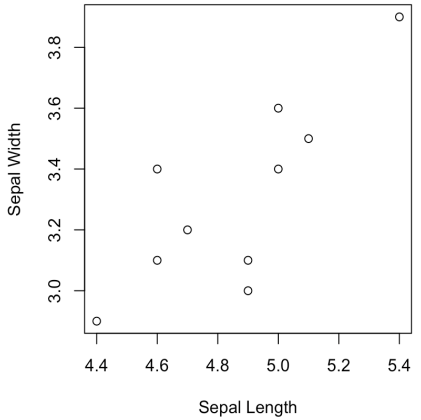
Using Calc – Regression Line

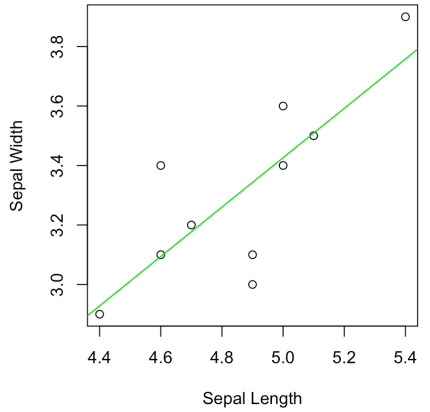
GOAL: Calculate the Regression Line!

1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinReg(a+bx) *STAT → CALC → LinReg(a+bx)*
   a) Xlist = L_1
   b) Ylist = L_2
   c) FreqList: *Leave blank*
   d) Store RegEQ: *Leave blank*

   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





$\hat{Y} = a + bX$
a = ?? and b = ??

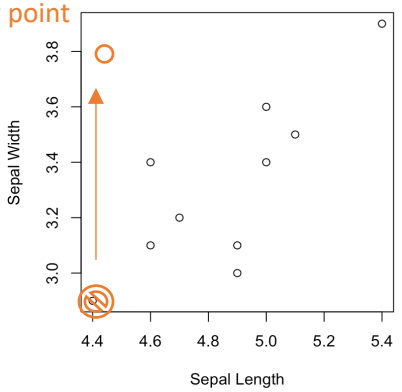
Outliers Demonstration

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

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

Now recalculate the equation!

New point



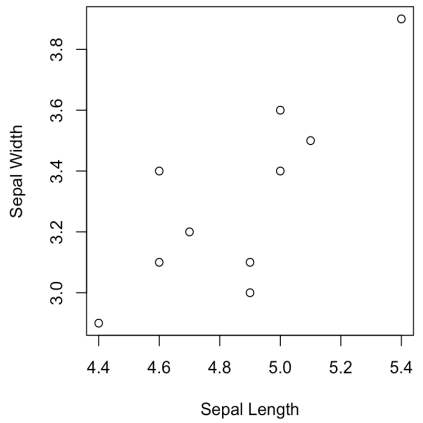
Using Calc – Regression Line

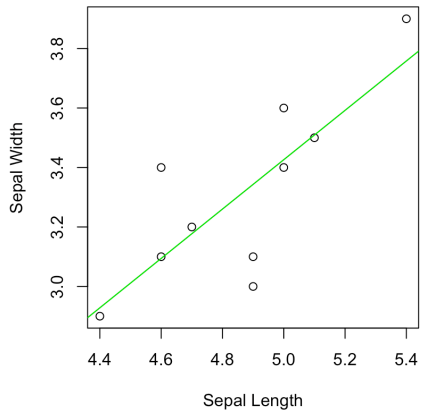
GOAL: Calculate the Regression Line!

1. Enter data
   a) X data in L_1
   b) Y data in L_2
2. LinReg(a+bx) *STAT → CALC → LinReg(a+bx)*
   a) Xlist = L_1
   b) Ylist = L_2
   c) FreqList: *Leave blank*
   d) Store RegEQ: *Leave blank*

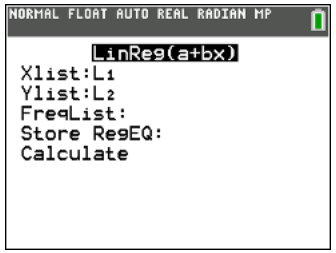
   Calculate

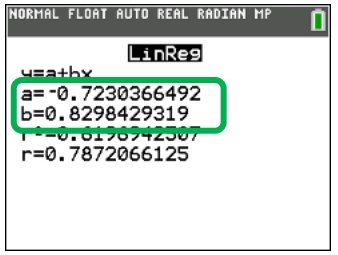
X	Y
Sepal Length	Sepal Width
5.1	3.5
4.9	3
4.7	3.2
4.6	3.1
5	3.6
5.4	3.9
4.6	3.4
5	3.4
4.4	2.9
4.9	3.1





$\hat{Y} = a + bX$
a = ?? and b = ??





Calculator Output

a = intercept b_0
b = slope b_1

Regression Equation: $\hat{Y} = -0.723 + 0.83X$

Show work: LinReg(a+bx)(Xlist = L1, Ylist = L2)

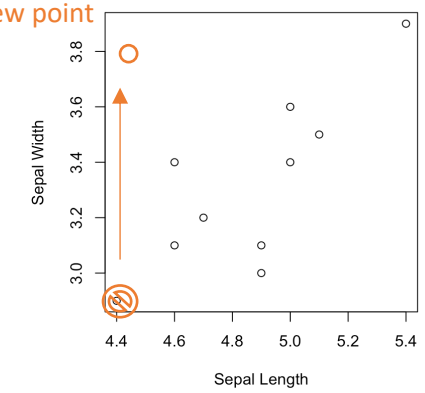
BIG d
equa
- *N*
  be
- *Bu*
  ou
  co

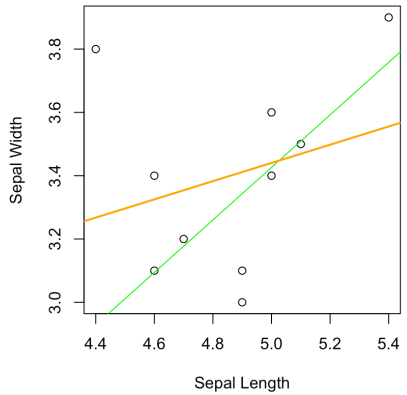
Outliers Demonstration

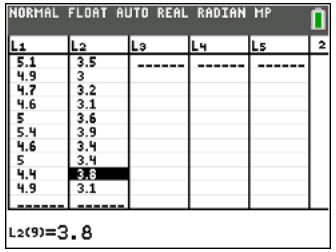
Let's change one data point to see the effects on the regression line:
- 9th observation: (4.4, 2.9) → (4.4, 3.8)

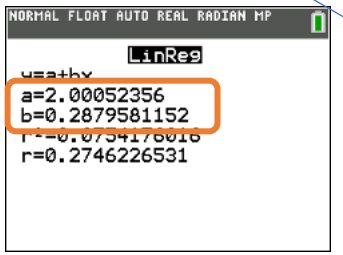
Now recalculate the equation!

New point







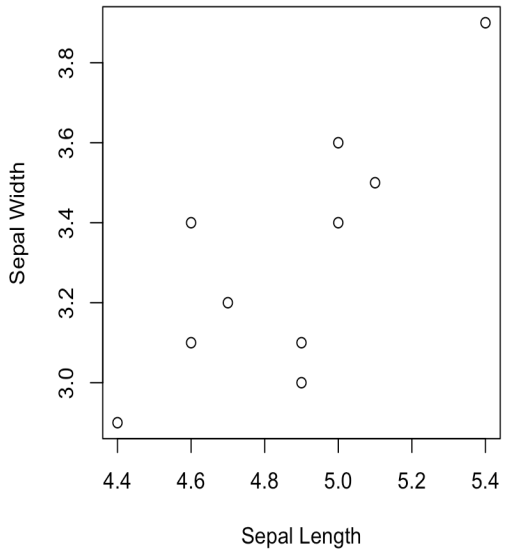


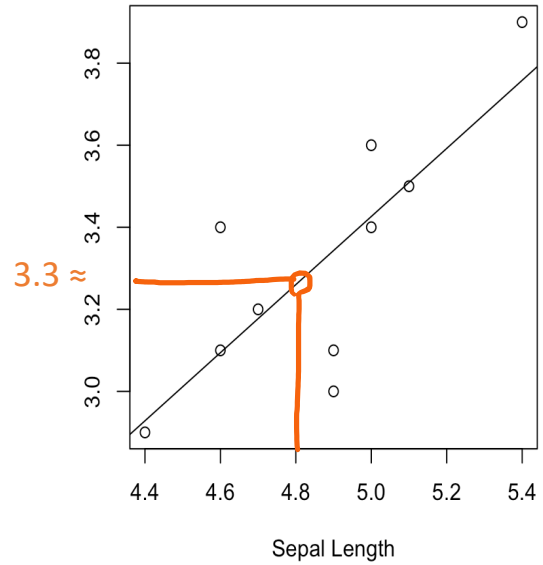
New Regression Equation: $\hat{Y} = 2.001 + 0.288X$

Predicting

Predictions Using the Regression Equation

- The primary use for a regression equation is to **predict** the value of the dependent variable for a value of the independent variable
    - We can think of our regression line, and specifically $\hat{Y}$, as predicted or expected values of Y for all X values *in the X range of our sample data*!
- This is another form of inference! We are using our sample data to make educated guesses about new data!
    - We can use our equation to answer a question like → If I select a new flower that has a Sepal Length of 4.8, what will the Sepal Width be?
    - Visually we could estimate this! (X = 4.8, $\hat{Y} \approx$?? Maybe 3.3)





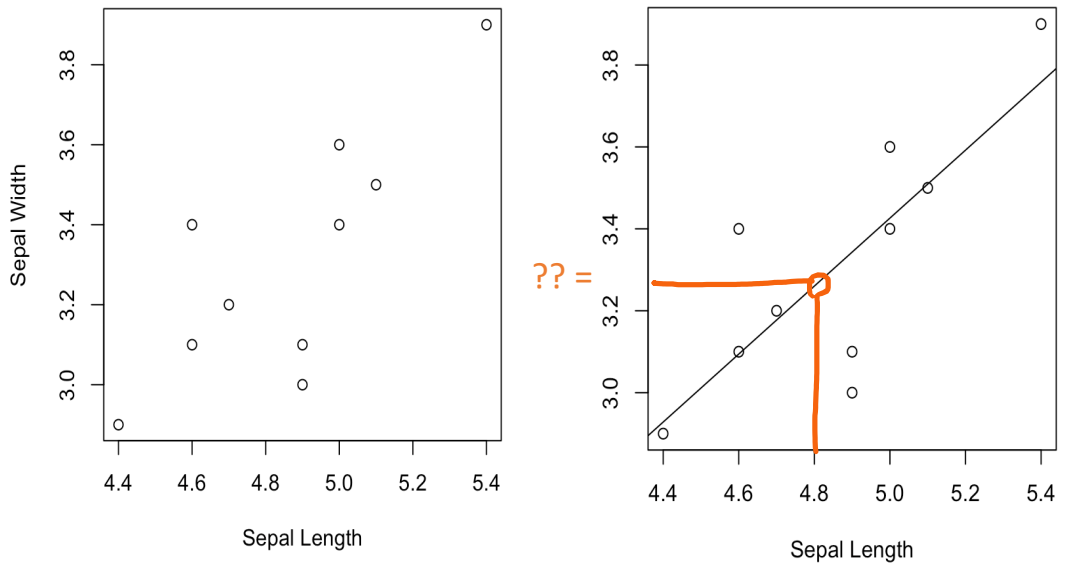
Calculating Predictions

How to Calculate Predictions

- This is simple, all we have to do is plug in the new X value to our equation and this will give us the predicted Y
- We can do this by hand quite easily!
  - Ex) If I select a new flower that has a Sepal Length of 4.8, what will the Sepal Width be?
  - (X, $\hat{Y}$) = (4.7, ??) → $\hat{Y} = ??$
  - Try for a new width: X = 5.3

Overestimate or Underestimate

- By comparing our prediction to the observed value, we can see if our model over or underestimated
- At X = 5, which point was overestimated by our model and which point was underestimated?

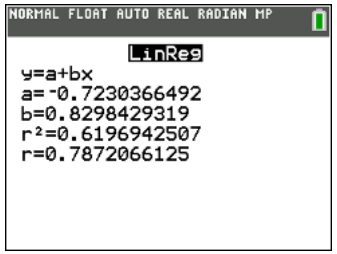


Calculating Predictions

How to Calculate Predictions

- This is simple, all we have to do is plug in the new X value to our equation and this will give us the predicted Y

- We can do this by hand quite easily!
  - Ex) If I select a new flower that has a Sepal Length of 4.8, what will the Sepal Width be?
  - *(X, $\hat{Y}$) = (4.8, ??)* → $\hat{Y} = -0.723 + 0.83(4.8) = 3.261$ *Predicted Width*
  - Try for a new length: X = 5.3 → $\hat{Y} = -0.723 + 0.83(5.3) = 3.676$

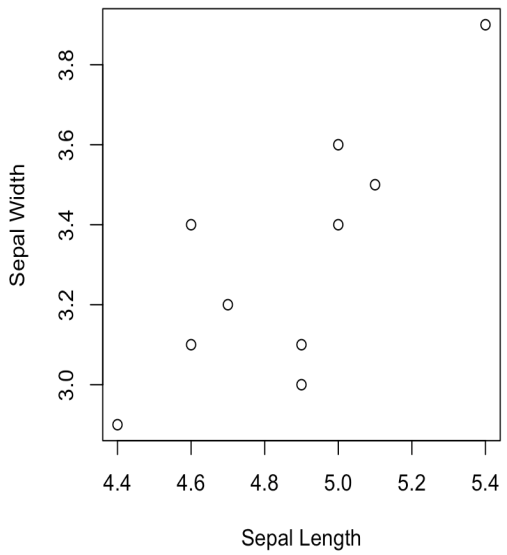


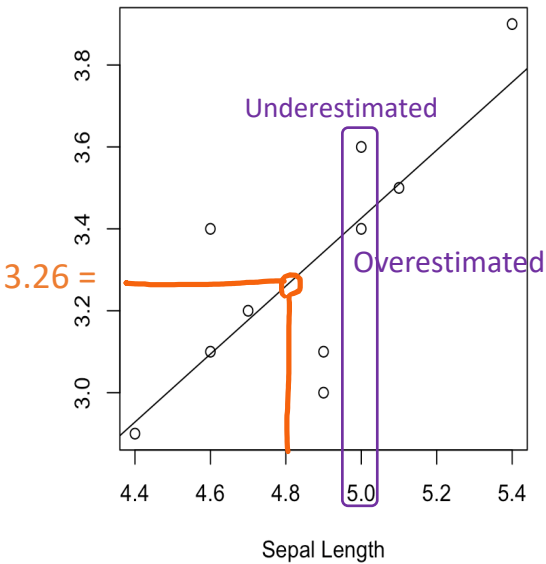
Regression Equation: $\hat{Y} = -0.723 + 0.83X$

In words (or typing) $\hat{Y}$ = Y hat or Predicted Y

Overestimate or Underestimate

- By comparing our prediction to the observed value, we can see if our model over or underestimated
- At X = 5, for one data point we overestimated (our regression line is above the actual value) and for the other we underestimated (the regression line is below the actual value)
  - TIP: Over / Under is from the perspective of the model





Evaluation Predictions

Evaluating Predictions

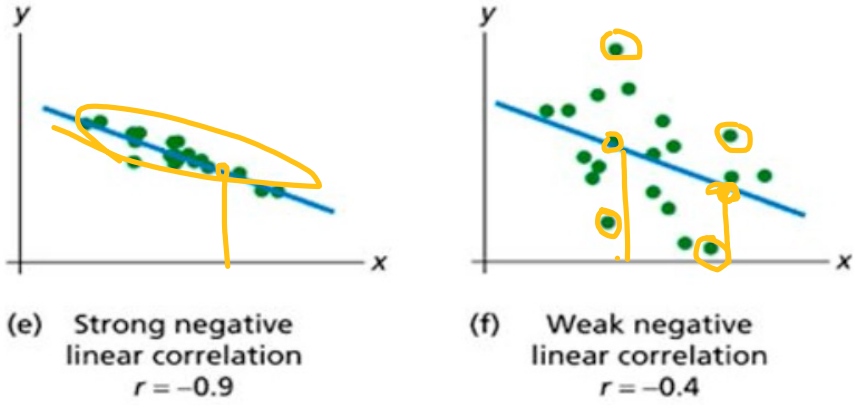
- In addition to fitting the line, we also want to think about how good our predictions will be based on this line
- One way we can **evaluate** the quality of our predictions (so how much we can trust them) is by the Correlation

High Correlation

- Indicates that there is a strong linear relationship between the two variables
- And therefore the line will fit well and the predictions will be accurate.

Low Correlation

- Indicates that there is a weak linear relationship between the two variables
- And therefore the line will not fit well and the predictions will not be accurate



(e) Strong negative linear correlation $r = -0.9$

(f) Weak negative linear correlation $r = -0.4$

https://slideplayer.com/slide/6204684/

Interpolating vs Extrapolating

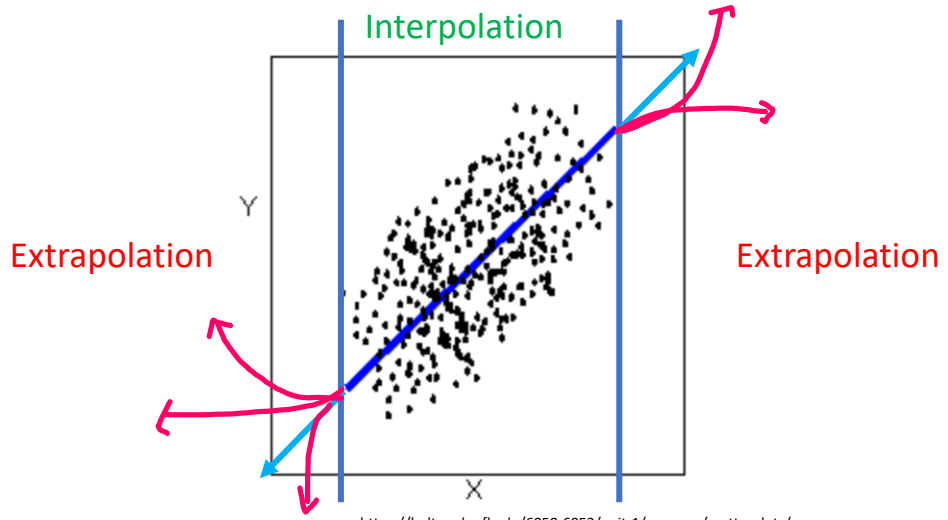
- When we predict, we are actually doing one of two things (one is good, one is bad):

Interpolation

- Interpolation results when the X value of interest falls between given values of X in our original data set
- Generally interpolation is considered a safe prediction method because we have already shown that our data behaves in a linear way within the range that we used to come up with the regression equation

Extrapolation

- Extrapolation results when the X value of interest falls outside the range of values for X in our original data set
- Extrapolation is considered riskier than interpolation because we have no way of knowing what the behavior of the data will be outside of the range we studied.
- It is a BIG assumption to think the regression line will continue in the EXACT same pattern (It could level off, or curve, or anything)



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

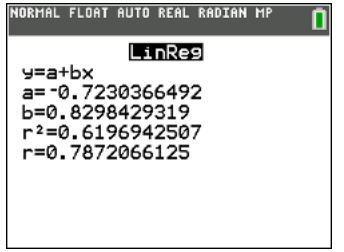
LCQ: Interpolating vs Extrapolating

Problem: Determine if the following predictions are interpolating or extrapolating. Then calculate the prediction.

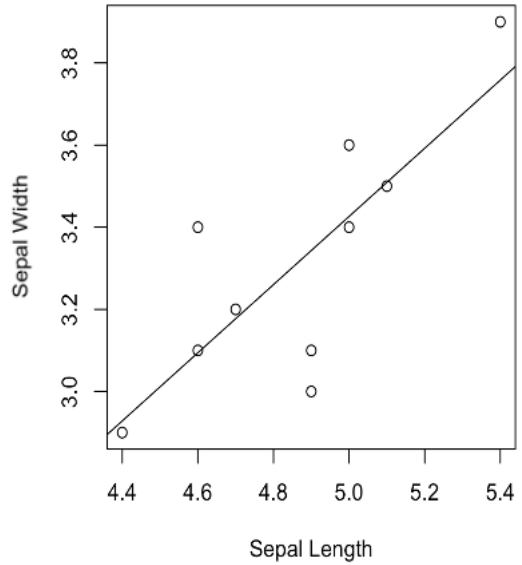
a) Predict the Sepal Width for a Sepal Length = 4.0

b) Predict the Sepal Width for a Sepal Length = 5.1

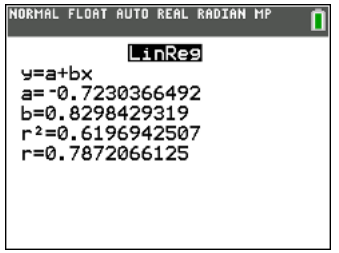
c) Predict the Sepal Width for a Sepal Length = 5.5



Regression Equation: $\hat{Y} = -0.723 + 0.83X$



LCQ: Interpolating vs Extrapolating



Problem: Determine if the following predictions are interpolating or extrapolating. Then calculate the prediction.

a) Predict the Sepal Width for a Sepal Length = 4.0

Extrapolating → X data ranges from 4.4 to 5.4 based on the scatter plot. Thus 4.0 is below the range

Regression Equation: $\hat{Y} = -0.723 + 0.83X$

$\hat{Y} = -0.723 + 0.83(4) = 2.597$

→ But we have to know that this result should be treated with caution because it is an extrapolation!

- *It is important to recognize that our equation will ALWAYS give us a result, even if I enter -10 or 1000!*
- *But contextually, some values are not going to make any sense… Can we have a negative length?? NO! So we have to be careful when using our equation to make predictions*

b) Predict the Sepal Width for a Sepal Length = 5.1

Interpolating → This is well within the X range of the original data that our regression equation was built on! So no concerns with this prediction

$\hat{Y} = -0.723 + 0.83(5.1) = 3.51$

c) Predict the Sepal Width for a Sepal Length = 5.5

Extrapolating → Even though this is very close to the max X value of 5.4, it is still outside the range

$\hat{Y} = -0.723 + 0.83(5.5) = 3.842$

→ Shouldn't trust this prediction because we are extrapolating!