Last Slides!!!

Unit 11 – Regression

Your Final Professor Colton

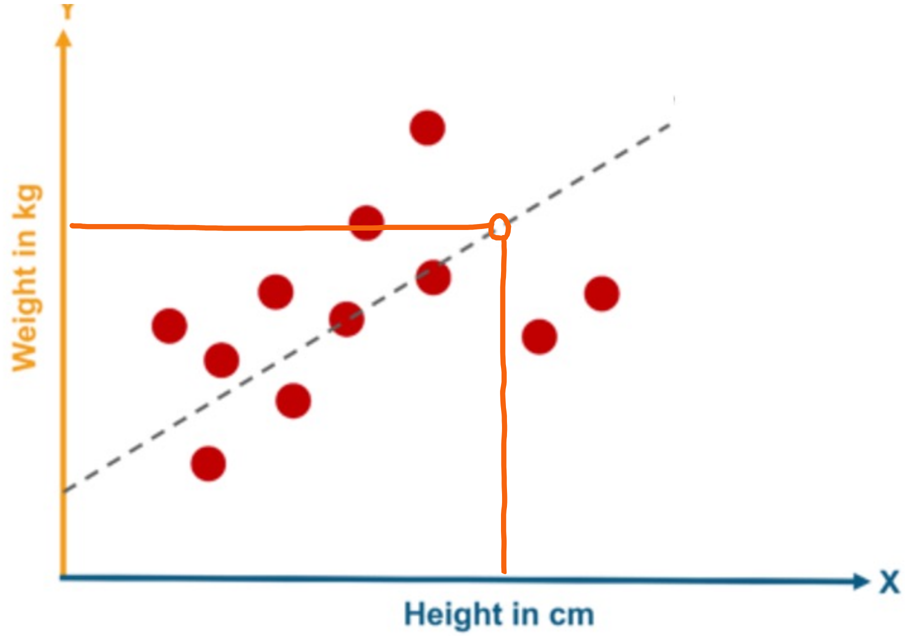
Unit 11 - Outline

<u>Unit 11 – Correlation and Regression</u>

- Regression Equation
- Predictions
- Coefficient of Determination

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**.

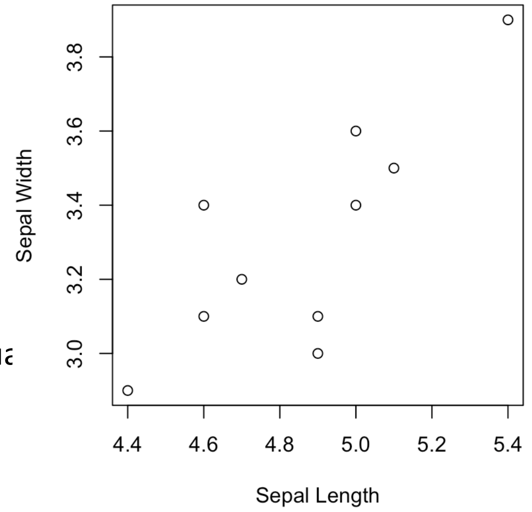


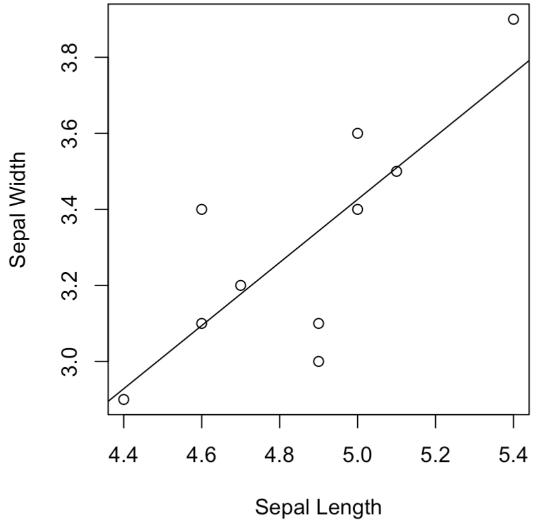
https://www.edureka.co/blog/linear-regression-in-python/

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 actua data points to the best fit line.





Equation

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

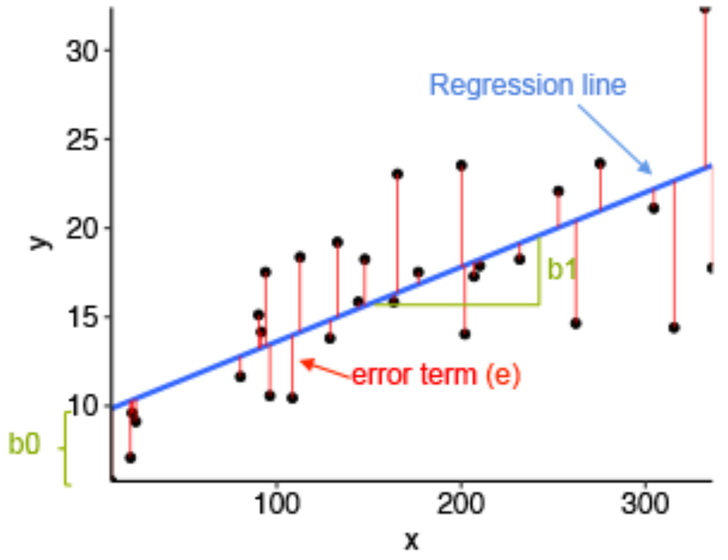
$$\hat{Y} = b_0 + b_1 X$$

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

Predicted Sepal Width = b_0 + b_1 (Sepal Length)

$(\widehat{Sepal\ Width})$

- b_0 = Y intercept
  - It is the location where the regression line crosses the Y-axis (value of Y when X = 0)
- b_1 = 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
- $\hat{\hat{Y}}$ = Predicted value of the response variable for the given X



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

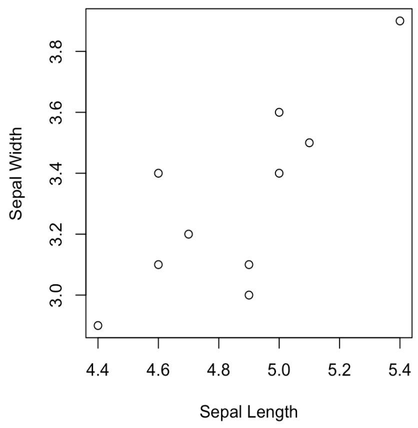
Parameters

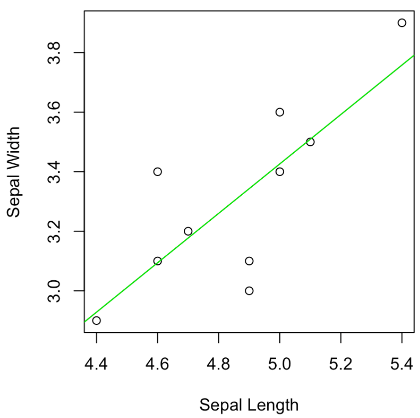
- b_0 and b_1 are statistics that are used as point estimates for the parameters β_0 and β_1 respectively.
- In the population, we have the regression line: $\hat{Y} = \beta_0 + \beta_1 X$
- Our equation above is an estimate of this based on our sample data!

Using Calc – Calculating Regression Line

GOAL: Calculate the Regression Line!

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





$\hat{Y} = b_0 + b_1 X$

b_0 = ?? and b_1 = ??

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 for the correlation test **
   e) RegEQ: *Leave blank for now*

   Calculate

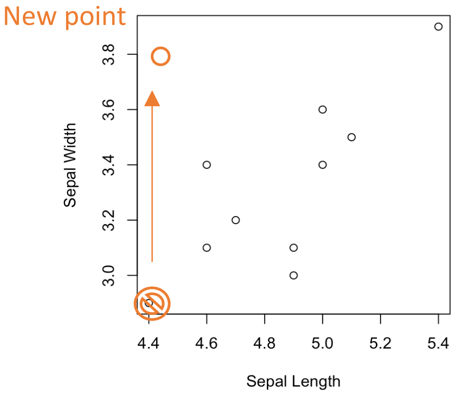
** The Alternative Hypothesis will NOT change the equation of the regression line, only the p-value of the correlation test

<u>Outliers Demonstration</u>

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

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

Now recalculate the equation!



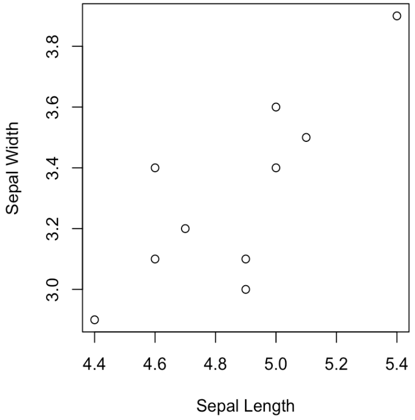
Using Calc – Calculating Regression Line

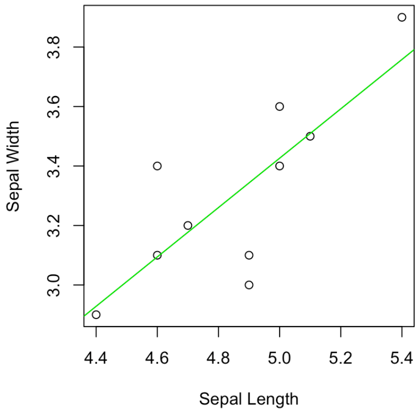
GOAL: Calculate the Regression Line!

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

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 for the correlation test **
   e) RegEQ: *Leave blank for now*
   Calculate

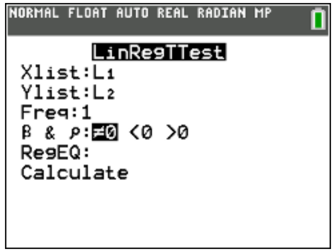
** The Alternative Hypothesis will NOT change the equation of the regression line, only the p-value of the correlation test

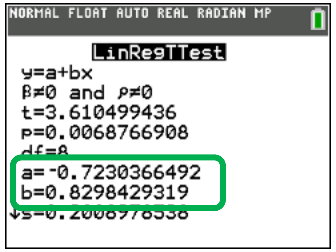




$\hat{Y} = b_0 + b_1X$

b_0 = ?? and b_1 = ??





Calculator Output

y = a + bx → calculator's notation for the regression equation

a = intercept b_0

b = slope b_1

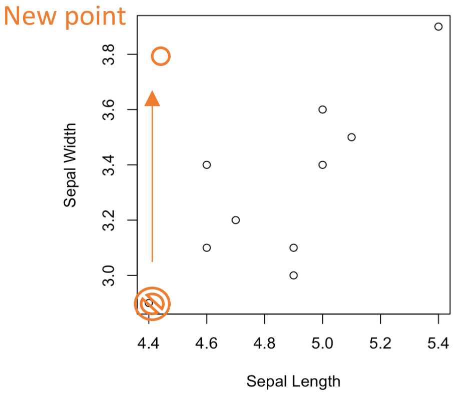
Regression Equation: $\hat{Y} = -0.723 + 0.83X$

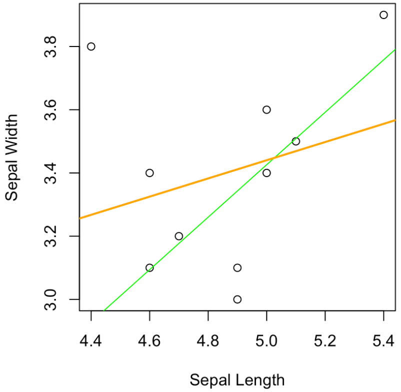
Outliers Demonstration

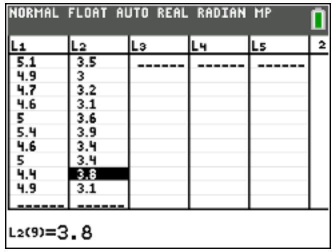
Let's change one data point to see the effects on the regression line:

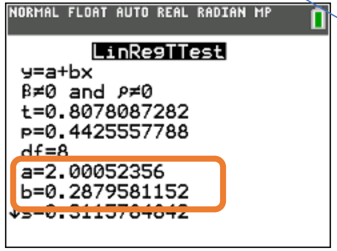
- 9^{th} observation: (4.4, 2.9) → (4.4, 3.8)

Now recalculate the equation!









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

Using Calc – Plotting Regression Line

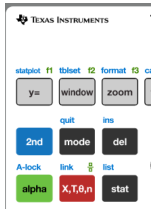
GOAL: Plotting the Regression Line!

1) Make Scatterplot
   a) Enter data: X (L_1) and Y (L_2)
   b) STAT PLOT
      - ON, Type = Scatter plot image, Xlist = L_1, Ylist = L_2
   c) ZOOM → 9:ZoomStat
      - This automatically zooms to whatever data range the stat plot requires

Two Options to Add Regression Line

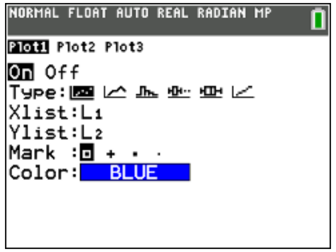
Option 1) Manually add regression line

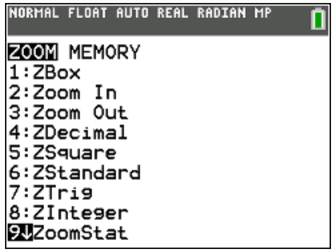
a) Get regression equation from LinRegTTest output
b) Type equation in Y= → Y_1
   - Use Red button to type in the Variable X
c) Graph



Using Calc – Plotting Regression Line

GOAL: Plotting the Regression Line!



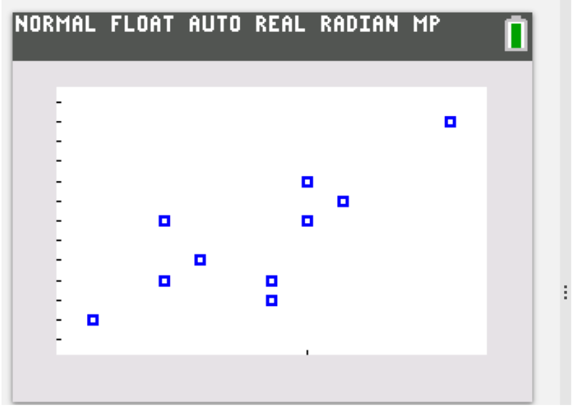


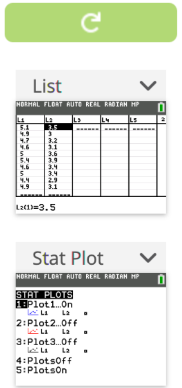
1) Make Scatterplot

a) Enter data: X (L_1) and Y (L_2)

b) STAT PLOT
   - ON, Type = Scatter plot image, Xlist = L_1, Ylist = L_2

c) ZOOM → 9:ZoomStat
   - This automatically zooms to whatever data range the stat plot requires





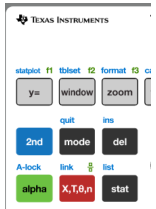
Two Options to Add Regression Line

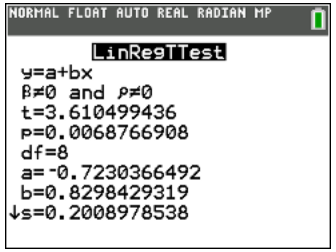
Option 1) Manually add regression line

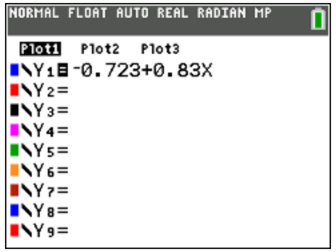
a) Get regression equation from LinRegTTest output

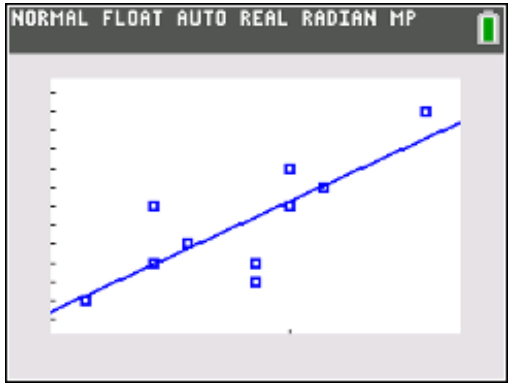
b) Type equation in Y= → Y_1
   - Use Red button to type in the Variable X

c) Graph









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

Using Calc – <u>Plotting</u> Regression Line

<u>GOAL</u>: Plotting the Regression Line!

<u>Option 2) Let calc add regression line</u>

a) LinRegTTest
   - RegEQ: Y_1
     - All other options are the same, we just need to tell our calculator to put the resulting regression equation in Y_1 for us!
   - To do this: Vars → Y-Vars → Function → Y_1
     - This should be a ONE TIME SETUP

   Calculate
   - If you look in Y= now, should see the exact regression equation from the output!

b) Graph

Using Calc – Plotting Regression Line

GOAL: Plotting the Regression Line!

Option 2) Leta Calc add regression line

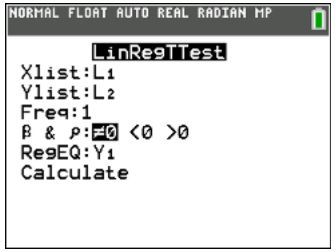
a) LinRegTTest

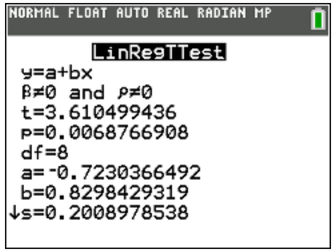
- RegEQ: Y_1
  - All other options are the same, we just need to tell our calculator to put the resulting regression equation in Y_1 for us!
- To do this: Vars → Y-Vars → Function → Y_1
  - This should be a ONE TIME SETUP

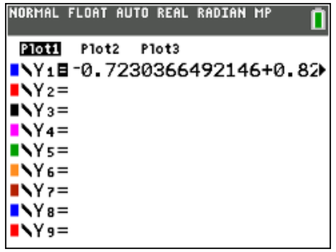
Calculate

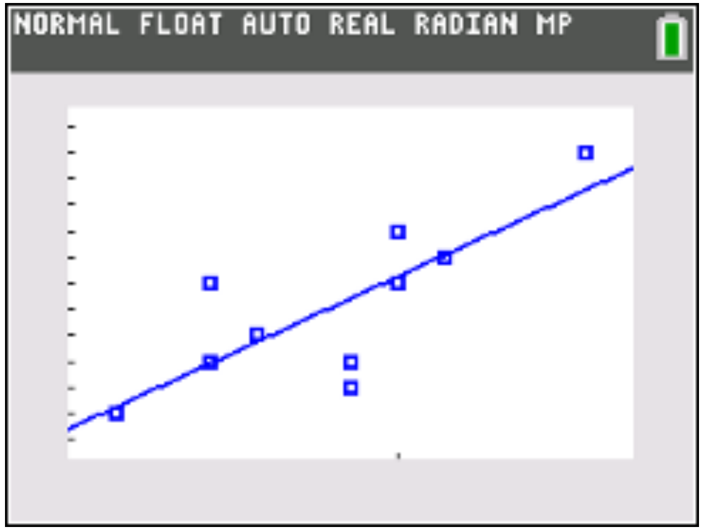
- If you look in Y= now, should see the exact regression equation from the output!

b) Graph





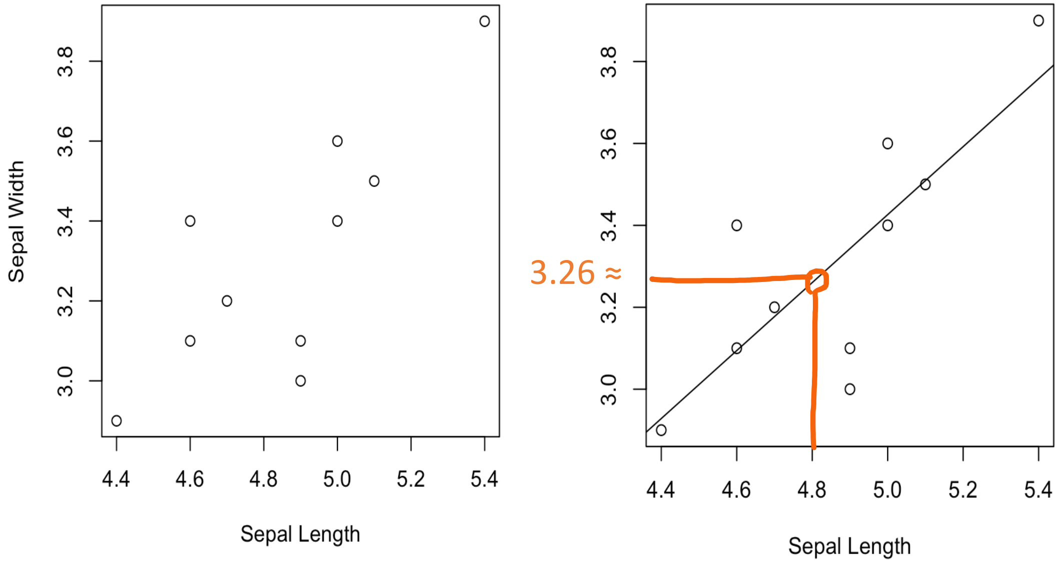




Predicting

<u>Predictions Using the Regression Equation</u>

- The primary use for a regression equation is to **predict** the value of the <u>dependent variable</u> for a value of the <u>independent variable</u>
  - We can think of our regression line, and specifically $\hat{Y}$, as <u>predicted or expected values of Y</u> for <u>all X values</u> *in the X range of our sample data*!
- This is another form of <u>inference</u>! 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, $\hat{Y} \approx$??)



Calculating Predictions

How to Calculate Predictions

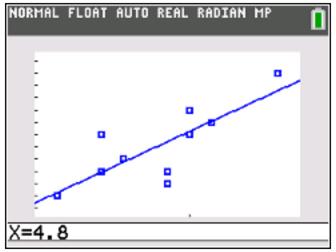
- This is simple, all we have to do is plug in the new X value to our equation and this will give us the predicted Y

Two Options

1) 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?
   - $(X, \hat{Y}) = (4.8, ??) \rightarrow \hat{Y} = ??$

2) Our calculator can do this for us!
   - IF we used the calculator to add the regression line to the graph, we can do the following:
   - GRAPH → CALC (2nd TRACE) → 1:Value → X = < type in X value of interest > → Enter
     - This will calculate the Y based on the equation that is entered in Y_1

** Note this will work even if we manually typed in the rounded regression equation to Y_1, but we might as well be more precise!*



$X = 4.8 \rightarrow \text{Predicted } Y = ??$

LCQ: Try for a new length: X = 5.3

Manual way:
???

Calc way:
???

Calculating Predictions

<u>How to Calculate Predictions</u>

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

<u>Two Options</u>

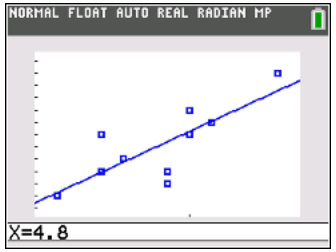
1) We can do this by hand quite easily!

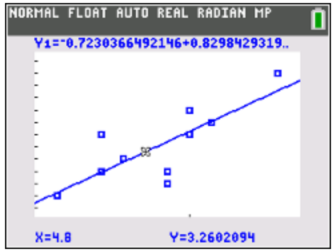
   - Ex) If I select a new flower that has a Sepal Length of 4.8, what will the Sepal Width be?
   - *(X, $\hat{Y}$) = (4.8, ??)* → $\hat{Y} = -0.723 + 0.83(4.8) = 3.261$ *Predicted Width*

2) Our calculator can do this for us!

   ** Note this will work even if we manually typed in the rounded regression equation to Y_1, but we might as well be more precise!*

   - IF we used the calculator to add the regression line to the graph, we can do the following:
   - GRAPH → CALC (2nd TRACE) → 1:Value → X = < type in X value of interest > → Enter
     - This will calculate the Y based on the equation that is entered in Y_1





X = 4.8 → $Predicted\ Y = 3.2602$

** Difference between this answer and the previous is because of roundoff error*

<u>LCQ</u>: Try for a new length: X = 5.3

Manual way:

$\hat{Y} = -0.723 + 0.83(5.3) = 3.676$

Calc way:

X (Length) = 5.3 → Predicted Y (Width) = 3.6707

Interpolating vs Extrapolating

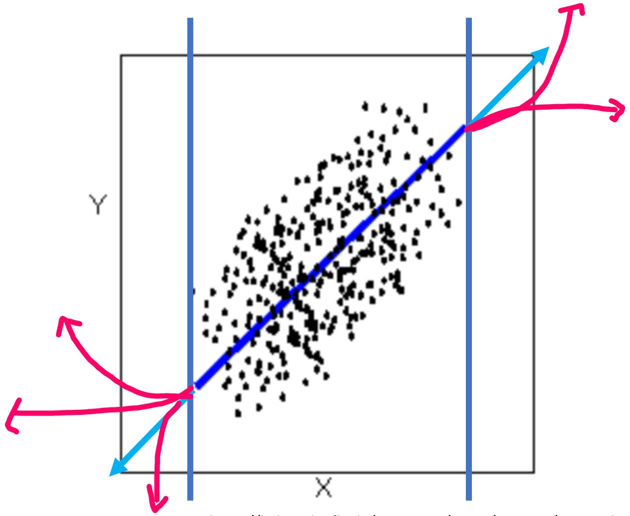
- When we predict, we are actually doing one of two things (one is good, one is bad):

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/

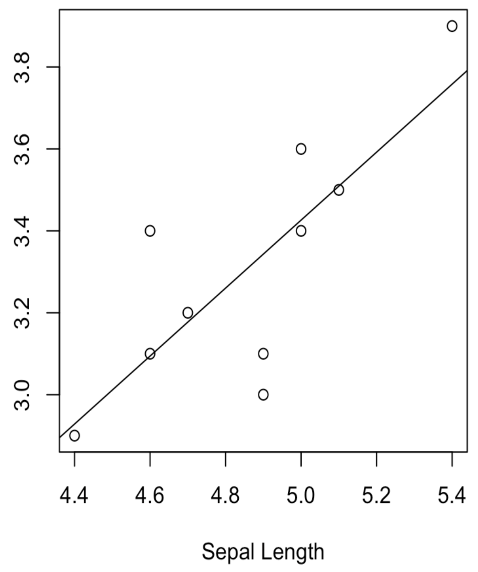
LCQ: Interpolating vs Extrapolating

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

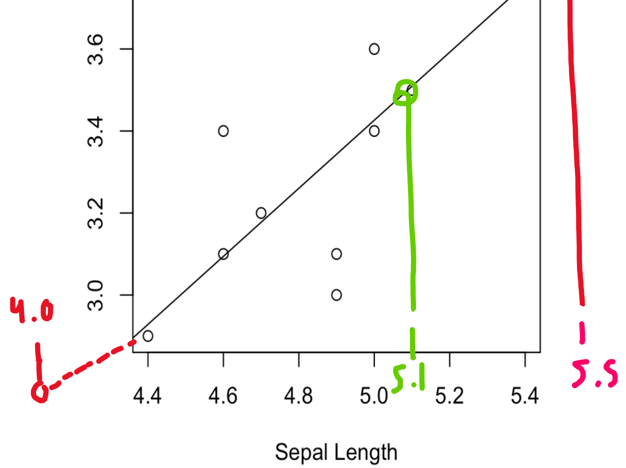
a) Predict the Sepal Width for a Sepal Length = 4.0

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

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



LCQ: Interpolating vs Extrapolating

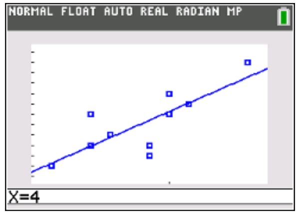


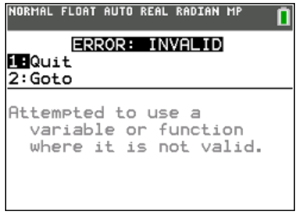
Problem: Determine if the following predictions are interpolating or extrapolating. Then calculate the prediction.

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

Using calc: *X = 4.0 → Predicted Y = ERROR*
→ Our calculator recognizes that our X value of interest is outside the range of the original data, so it gives us an error and does not give us a result





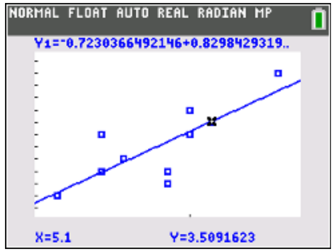
So have to calculate the prediction manually: $\hat{Y} = -0.723 + 0.83(4) = 2.597$
→ But we have to know that this result should be treated with caution because it is an extrapolation!

- *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 we won't have any issues calculating the prediction and no concerns in doing so

Using calc: *X = 5.1 → Predicted Y = 3.509*



c) Predict the Sepal Width for a Sepal Length = 5.5
Extrapolating → Even though this is very close to the max X value of 4.4, it is still outside the range

Manual calculation: $\hat{Y} = -0.723 + 0.83(5.5) = 3.842$
→ Shouldn't trust this prediction because we are extrapolating!

Coefficient of Determination

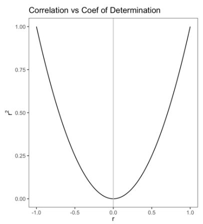
- In addition to correlation, we can also <u>assess the strength</u> of the relationship using another measure called the **Coefficient of Determination**

Coefficient of Determination

- The **Coefficient of Determination** (r^2) is the square of the correlation
  - It measures the usefulness of the regression line in making predictions
  - Specifically, it determines the percent of the variation in the Y variable that can be explained by the linear relationship with the X variable

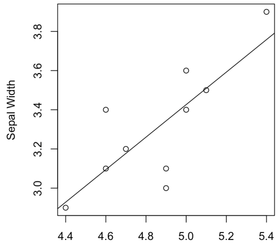
Properties of r^2

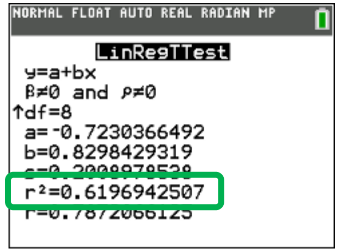
- Range from 0% to 100%
- The closer r^2 is to 100%, the stronger the relationship between X and Y
- As r gets closer to -1 or 1, r^2 increases



Calculating r^2

- Our calculator gives us this when we find the regression line!





Interpreting r^2

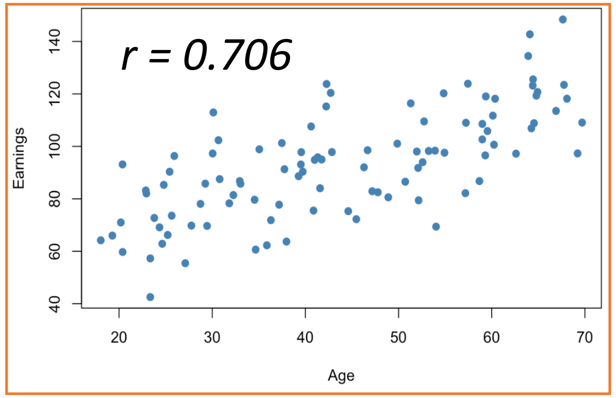
- We have a general structure for how to interpret this measure:

USING CONTEXT!

- r^2 % of the variation Y can be explained by the linear relationship with X

Example

- 62% of the variation in Sepal Width can be explained by the linear relationship with Sepal Length



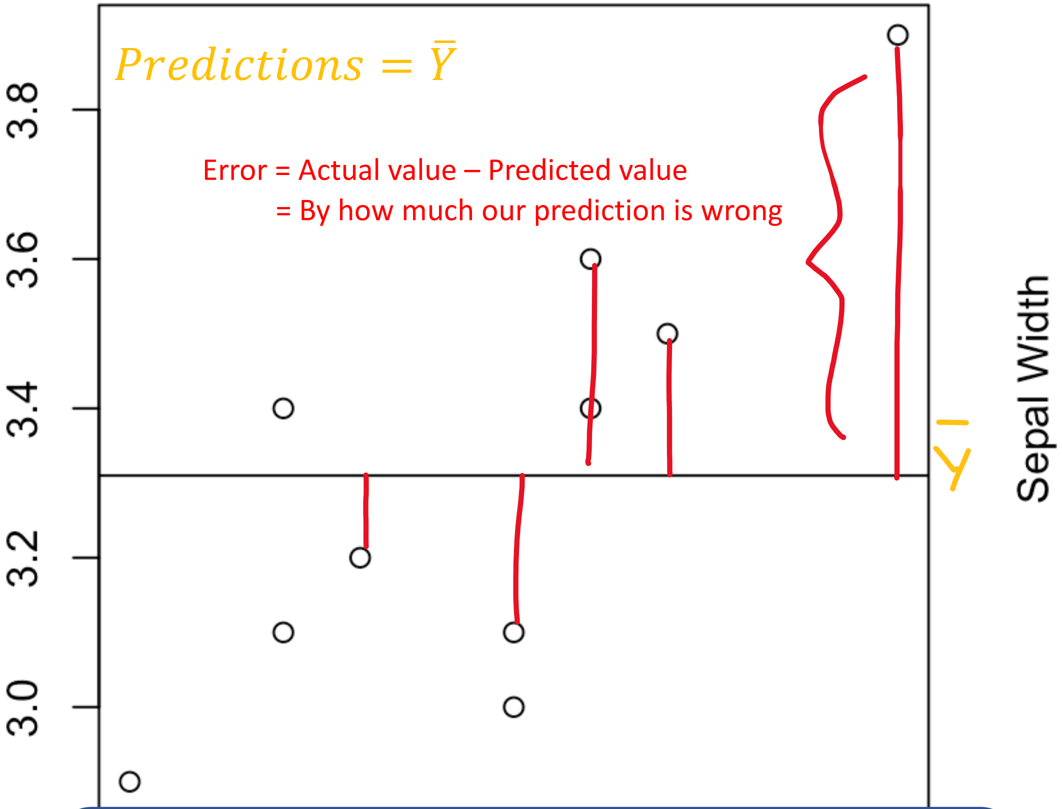
LCQ: Interpret r^2

First calculate $r^2 = (0.706)^2 = 0.498$

49.8% of the variation in Earnings can be expla
by the linear relationship to Age →
This is just using the general structure and fillir
the value and context for this scenario!

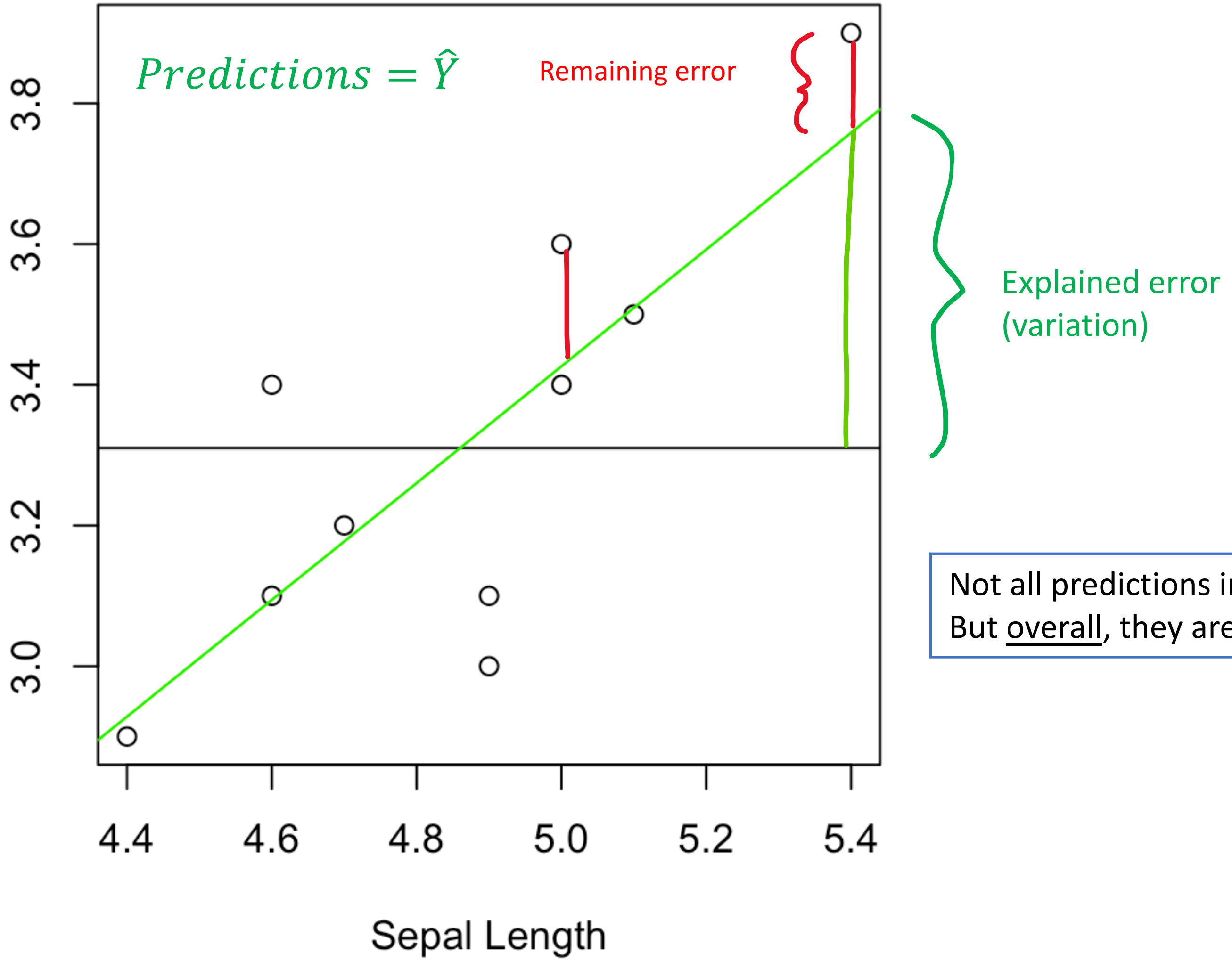
Idea Behind Coefficient of Determination: Explained and Unexplained Variation

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}$$



We start by ONLY looking at the Y variable

- Our best prediction would the mean $\bar{Y}$!
- We see our predictions aren't very good…



Not all predictions improve
But <u>overall</u>, they are better!

Now we <u>update our predictions</u> using <u>X knowledge</u>, which gives us the regression line $\hat{Y}$

- Our predictions have improved! We are wrong by less

Problem Session!!!

Example 2

temp	depth
-6	21.1
-3	26
-2	18
1	19.2
6	16.9
10	18.1
11	16.8
19	11.8

The article "Snow Cover and Temperature Relationships in North America and Eurasia" (Journal of Climate and Applied Meteorology [1983]: 460-469) explored the relationship between October-November continental snow cover and December-February temperature.

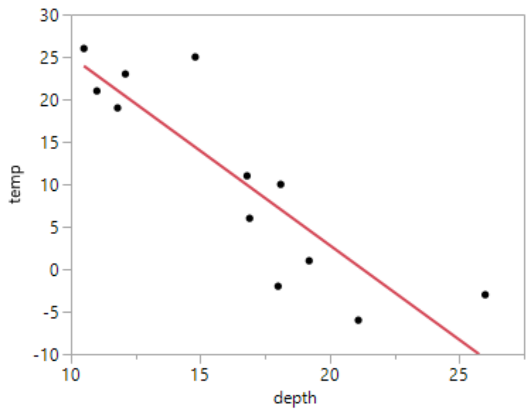
a) Does there seem to be a positive association, a negative association, or no association from the scatter plot?

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

c) Find and interpret the correlation coefficient, r, and r^2.

d) Find and interpret the equation for the line of best fit.

e) What temperature will the model predict if we have 12 inches of snow? If we have 36 inches



Some Solutions

c) $r^2 = 0.7674$ -> The coefficient of determination of 0.7674 indicates that approximately 76.74% of the variability in temperature can be predicted by snow depth.

r = -0.8760

- The correlation coefficient of -0.8760 indicates that there is a strong, negative linear relationship between snow cover and temperature; as snow cover increases, temperature decreases.

d) Predicted Temp = 47.296 – 2.224(snow depth)

e)

- For 12 inches of snow, the model predicts a temperature of

Temp = 47.296 – 2.224(12) = **20.608 degrees**

- For 36 inches of snow, the model predicts a temperature of

Temp = 47.296 – 2.224(36) = **-32.768 degrees**

Problem #13

For the bookstore sales data in Exercise 1, the manager wants to predict *Sales* from *Number of Sales People Working*.

a) Find the slope estimate, b_1.

b) Find the intercept, b_0.

c) Write down the equation that predicts *Sales* from *Number of Sales People Working*.

d) If 18 people are working, what *Sales* do you predict?

Problem #13 Output

a) Find the slope estimate.

b) Find the intercept.

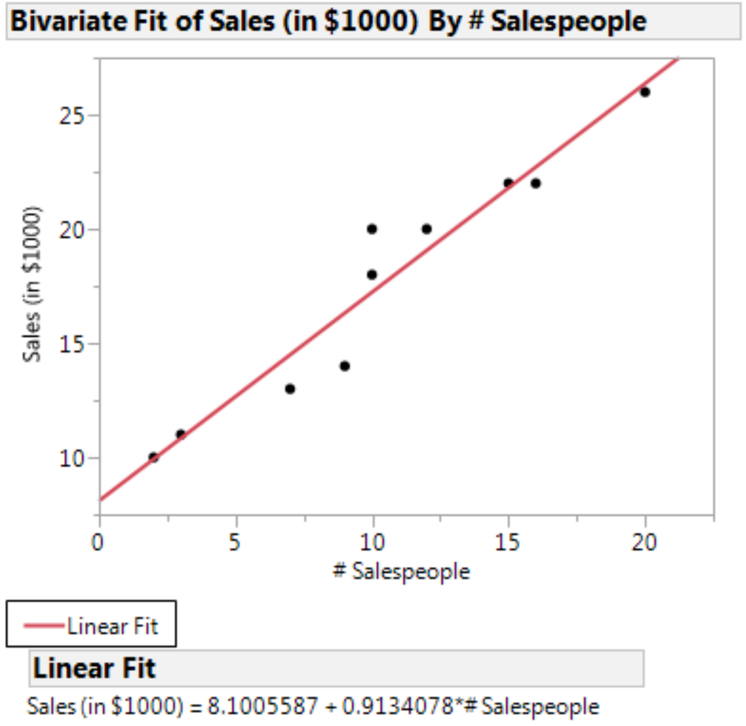
Linear Fit

Sales (in $1000) = 8.1005587 + 0.9134078*# Salespeople

Problem #13 Solution

a) Find the slope estimate, **b_1 = 0.913**

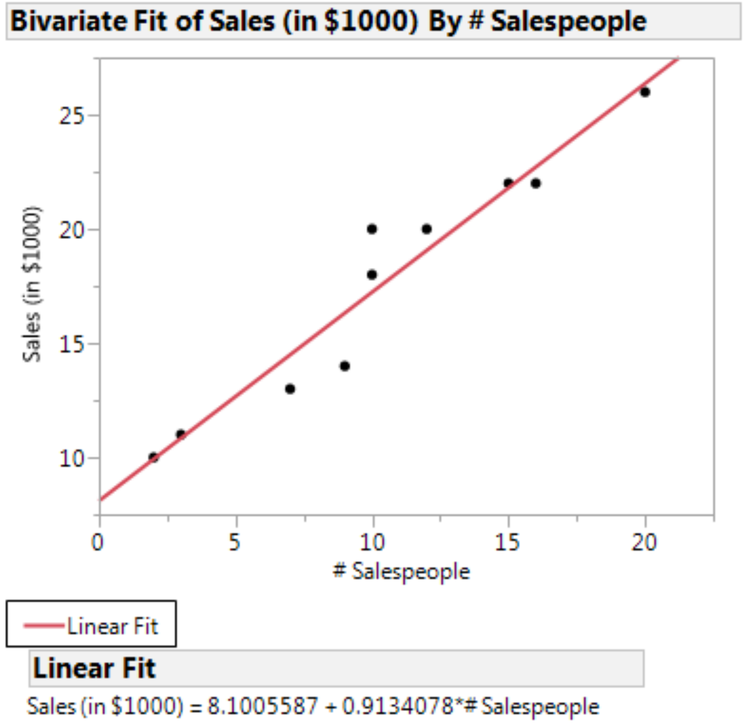
b) Find the intercept, **b_0= 8.101**



Problem #13, cont.

c) Write down the equation that predicts *Sales* from *Number of Sales People Working.*

d) If 18 people are working, what *Sales* do you predict?



Problem #13 Solution

c) Write down the equation that predicts *Sales* from *Number of Sales People Working*. $\widehat{\boldsymbol{Sales}} = \mathbf{8.101} + \mathbf{0.913(\#\boldsymbol{Salespeople})}$

d) If 18 people are working, what *Sales* do you predict? $\widehat{\boldsymbol{Sales}} = \mathbf{8.101 + 0.913(18) = \$24.535}$

Linear Fit

Sales (in $1000) = 8.1005587 + 0.9134078*# Salespeople

Problem #19

For the regression model for the bookstore of Exercise 1, what is the value of R^2 and what does it mean?

Linear Fit

Sales (in $1000) = 8.1005587 + 0.9134078*# Salespeople

Summary of Fit

RSquare	0.931932

Problem #19 Solution

For the regression model for the bookstore of Exercise 1, what is the value of R^2 and what does it mean?

R^2 = 0.9319, which indicates that approximately 93.19% of the variability in Sales can be predicted by the linear relationship between the number of salespeople and sales.

Linear Fit

Sales (in $1000) = 8.1005587 + 0.9134078*# Salespeople

Summary of Fit	
RSquare	0.931932

Problem #31

A linear model fit to predict weekly *Sales* of frozen pizza (in pounds) from the average *Price* ($ per unit) charged by a sample of stores in the city of Dallas in 39 recent weeks is:

$$\widehat{Sales} = 141{,}865.53 - 24{,}369.49 Price$$

a) What is the explanatory variable?

b) What is the response variable?

c) What do you predict the sales to be if the average price charged was $3.50 for a pizza?

Problem #31 Solution

$$\widehat{Sales} = 141{,}865.53 - 24{,}369.49Price$$

a) What is the explanatory variable? **Average price ($ per unit)**

b) What is the response variable? **Weekly sales of frozen pizza (in pounds)**

c) What do you predict the sales to be if the average price charged was $3.50 for a pizza?

$$\widehat{Sales} = 141{,}865.53 - 24{,}369.49(3.50) = \mathbf{56,572.315}\ \textbf{pounds}$$