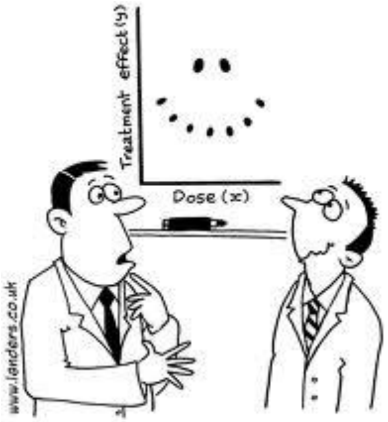
Section 1.5 – Linear Regression Using Technology

MDM4U
David Chen

Last class, you learned that by examining a scatter plot, you can see whether the relationship between two variables is strong or weak, positive or negative, linear or non-linear.



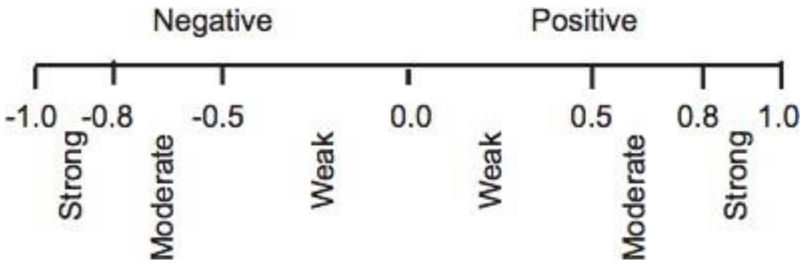
"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

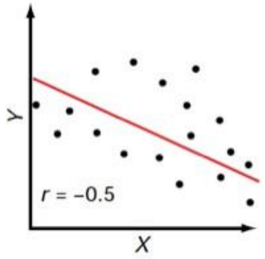
In this lesson, you will use technology that will allow you to quantify the linear correlation between two quantitative variables. We will be looking at four main statistics to describe the correlation:

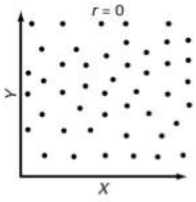
1. The correlation coefficient (r)
2. The coefficient of determination (r^2)
3. Regression line $\hat{y} = a + bx$
4. Residual values (observed y – predicted y)

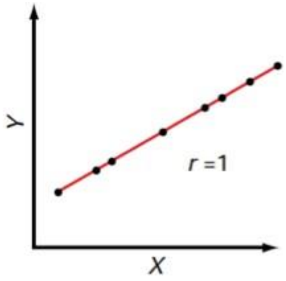
Part 1: The Correlation Coefficient r

The correlation coefficient, r, is a number between -1 and 1 that is an indicator of both the strength and direction of a linear relationship between two quantitative variables. A value of r = 0 indicates no correlation, while r = 1 or r = -1 indicates a perfect positive or negative correlation.









http://guessthecorrelation.com/

Part 2: The Coefficient of Determination r^2

The coefficient of determination r^2, is a number between 0 and 1 that is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

The coefficient of determination is a measure of how well the regression line (line of best fit) represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

Part 3: Regression Line (Line of Best Fit)

A regression line is a line that describes how a dependent (response) variable y changes as an independent (explanatory) variable x changes. We often use a regression line to predict the value of y for a given value of x.

The equation is of the form $\hat{y} = a + bx$

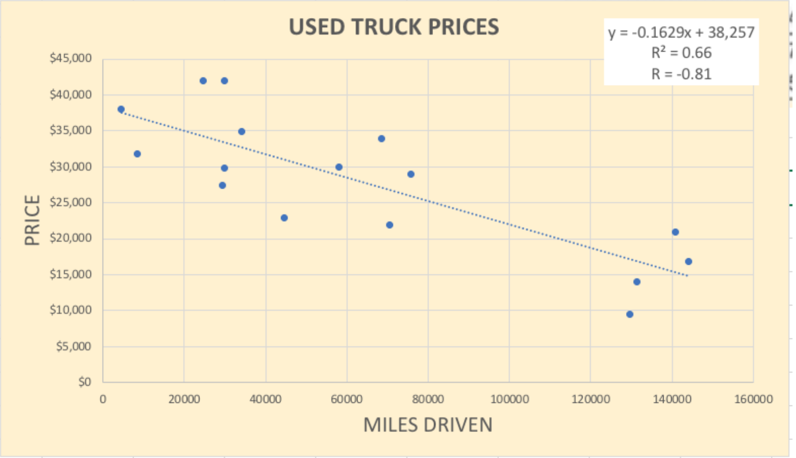
In this equation,

- $\hat{y}$ is the predicted value of the dependent variable y for a given value of x
- b is the slope, the amount by which y is predicted to change when x increases by one unit
- a is the y-intercept, the predicted value of y when $x = 0$

Example 1: The equation of the regression line for the scatterplot shown to the right is $\widehat{price} = 38257 - 0.1629(miles\ driven)$. Identify the slope and y-intercept of the regression line. Interpret each value in context.

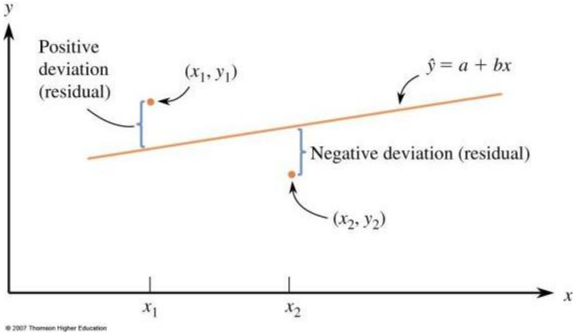
The slope $b = -0.1629$ tells us that the price is predicted to go down by 0.1629 dollars for each additional mile driven.

The y-intercept $a = 38257$ is the predicted price for a truck that has been driven 0 miles.



Part 4: Residual Values

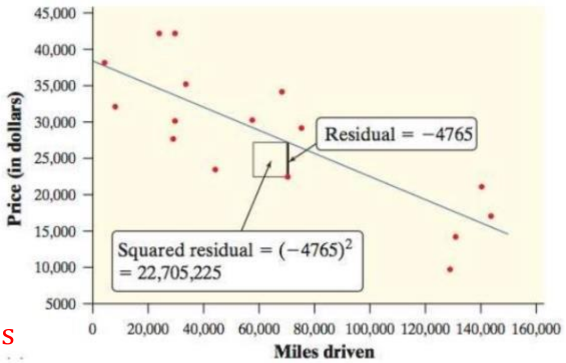
A residual is the difference between an observed value of y and the value predicted by the regression line ($\hat{y}$). The residual value tells us how far off the linear regression's prediction is at a given point.



Residual = observed y – predicted y

$= y -\hat{y}$

Example 2: Using the regression equation from example 1, find and interpret the residual for a truck that had 70583 miles driven and a price of $21994.



Solution:

The regression line predicts a price of

$$\widehat{price} = 38257 - 0.1629(70583)$$
$$= \$26759$$

But for this truck, its actual price was $21994. The truck's residual is

Residual = observed y – predicted y

$= y -\hat{y}$

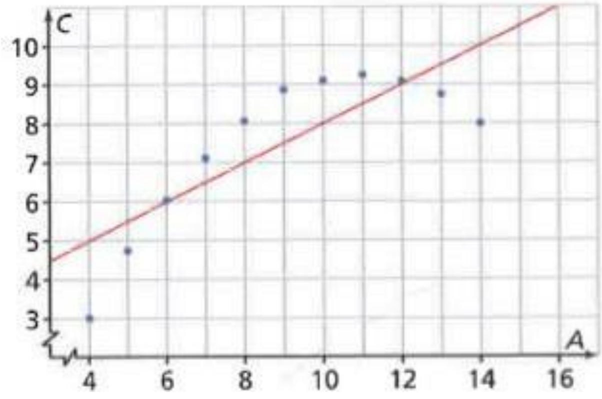
$= 21994 - 26759$

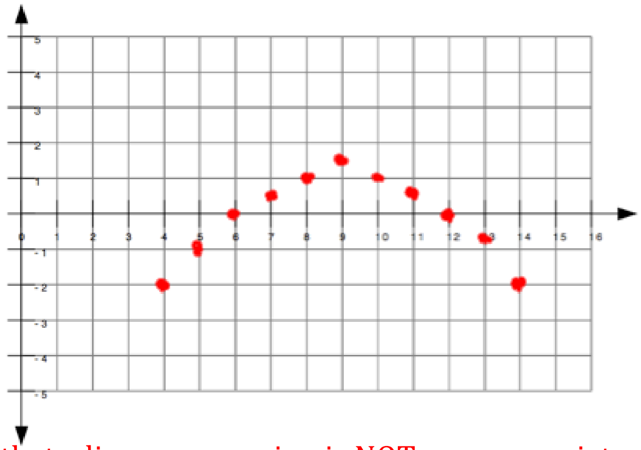
$= -4765$

This tells us that the actual price of this truck is $4765 lower than expected based on its mileage. Graphically speaking, the point is 4765 units below the line of best fit.

Note: If the regression model is a good fit, the residuals should be fairly small, and there should be no noticeable pattern. Large residuals or a noticeable pattern are indicators that another model may be more appropriate.

Example 3: Sketch the residual plot for the following graph and comment about what it tells you.





The distinguishable pattern in the residual plot shows that a linear regression is NOT an appropriate regression model in this situation.

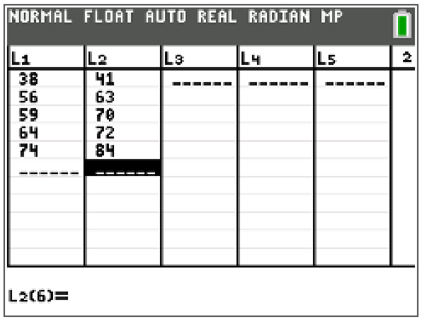
Part 5: Linear Regression Using a Ti-83/84 Calculator

Example 4: Archaeopteryx is an extinct beast having feathers like a bird but teeth and a tail like a reptile. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they are different species rather than individuals from the same species. If the specimens belong to the same species and differ in size because some are younger than others, there should be a positive linear relationship between the lengths of a pair of bones from all individuals. An outlier from this relationship would suggest a different species. Here are data on the lengths in centimeters of the femur and the humerus for the five specimens that preserve both bones.

Femur (x)	38	56	59	64	74
Humerus (y)	41	63	70	72	84

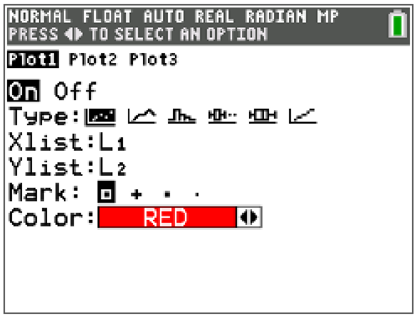
a) Make a scatterplot of the data

- Turn on diagnostics: 2nd → 0 → diagnosticON → ENTER
- Input data in to L1 and L2: STAT → ENTER
- Turn on statplot: 2nd → y= → ENTER → ON (make sure scatter plot is chosen)
- View graph: GRAPH → ZOOM → ZOOMSTAT

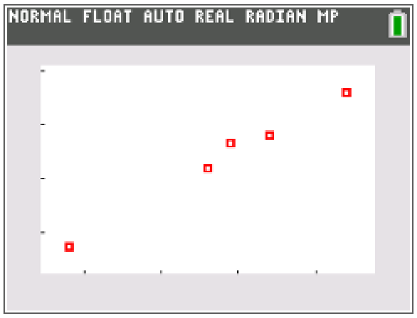


```
NORMAL FLOAT AUTO REAL RADIAN MP
L1      L2      L3      L4      L5      2
38      41      ------  ------  ------
56      63
59      70
64      72
74      84
------  ------

L2(6)=
```

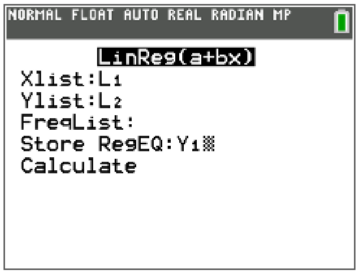


```
NORMAL FLOAT AUTO REAL RADIAN MP
PRESS ◀▶ TO SELECT AN OPTION
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:L2
Mark: □ + · ·
Color: RED ◀▶
```

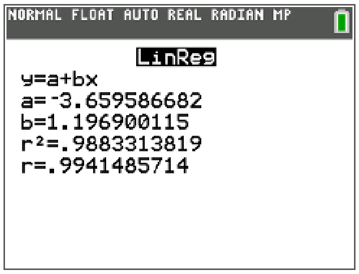


b) Find the equation of the regression line and interpret the slope and y-intercept in context.

- STAT → CALC → LinReg (a+bx) → xlist: L1 → ylist: L2 → store RegEQ: Y1 → CALCULATE



```
NORMAL FLOAT AUTO REAL RADIAN MP
LinReg(a+bx)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:Y1
Calculate
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
LinReg
y=a+bx
a=-3.659586682
b=1.196900115
r²=.9883313819
r=.9941485714
```

equation: predicted humerus length = -3.66 + 1.20 (femur length)

y-intercept: when femur length is zero, the model predicts a humerus length of –3.66 cm

slope: for every one cm increase in femur length, the model predicts an average increase in humerus length of 1.2 cm

c) Find and interpret correlation coefficient, *r*.

an r of .994 indicates a strong, positive linear correlation between femur and humerus length

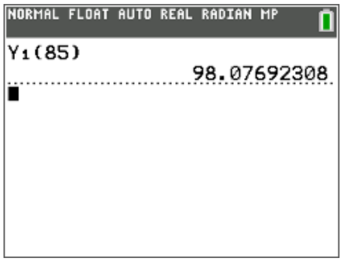
d) Find the coefficient of determination, r^2. Interpret it in the context of this data.

approximately 98.8% of the variation in humerus length can be explained by the approximate linear correlation with femur length

e) Use your equation to predict the humerus length for a femur that is 85 cm.

- VARS → Y-VARS → Y1 → (85) → ENTER

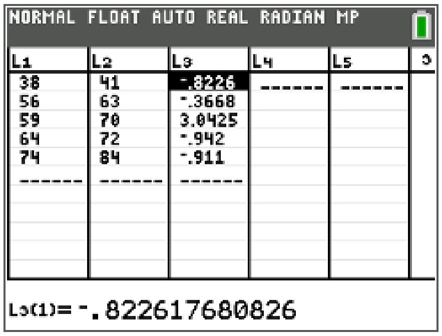
the model predicts a humerus length of 98.08 cm for a femur length of 85 cm.



```
NORMAL FLOAT AUTO REAL RADIAN MP
Y1(85)
                   98.07692308
```

f) Calculate the residual values and analyze the residual plot

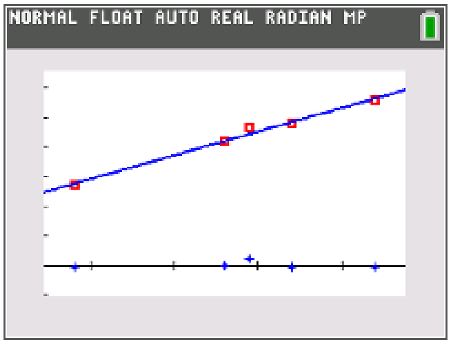
- Put residual values in to L3: STAT → ENTER → scroll to highlight L3 → ENTER → 2nd VARS → RESID → ENTER
- View residual plot: turn on plot 2 → Ylist: resid → GRAPH → ZOOMSTAT



NORMAL FLOAT AUTO REAL RADIAN MP

L1	L2	L3	L4	L5
38	41	⁻.8226	------	------
56	63	⁻.3668		
59	70	3.0425		
64	72	⁻.942		
74	84	⁻.911		
------	------	------		

L3(1)=⁻.822617680826



NORMAL FLOAT AUTO REAL RADIAN MP

Notice that the residual values are small and there is no noticeable pattern on the residual plot. This indicates that the regression line is a good model for the data.

Phrases to Use in Your Answers

Underlined words/phrases or blanks indicate context is needed.

regression: interpretation, in context, of

1. ***r*** – positive or negative, weak or strong linear correlation between explanatory variable and response variable

2. ***r^2*** – about x percent of the variation in the response variable can be explained by the approximate linear relationship with the explanatory variable.

3. **slope** – for every 1 unit increase in the explanatory variable, our model predicts an average increase of *y* units in the response variable.

4. ***y*-intercept** – at an explanatory variable value of 0 units, our model predicts