8.3 What's r^2 Anyway? Name ____________________

We're looking for a way to estimate the weight (*y*) of a mystery person. Sometimes we'll guess too high, other times too low. We seek a method that will minimize the errors we make *in the long run*. Absent any other information, what's the best weight to guess?

---

1. Paste the height-weight data into JMP (remember to Paste With Column Names).
2. The errors made using this guess-the-mean strategy would be $(y - \bar{y})$. Create a list of these errors in the next column over in JMP by adding a formula. Call the column *Errors*.
3. Now Analyze the Distribution of *Errors*.

   a. Where is the center of the distribution? _____________

   b. What is the spread? __________________

   c. What do the errors greater than 0 represent?

   d. What do the errors smaller than 0 represent?

4. We want a quantitative measure of the overall amount of error. Simply summing the errors won't work. Why not? What *must* that total be?

(Hint: Find the average of the Errors and then multiply it by how many samples there are).

5. You know the usual Statistics trick, of course: sum the *squares* of the errors. Squaring makes them all positive (better for adding) and places greater emphasis on the larger errors we hope to avoid. Calculate the squared errors in the next column using another formula in JMP, call it "*Errors Squared*" and then find the sum and record it below. (Hint: See the hint from the last step to find the sum).

6. Sum of the Errors Squared = ___________________*__________________ = ______________

   Mean of the Errors Squares   Total number of samples   Sum of Errors Squared

7. Now suppose that, rather than just having to guess blindly, we knew something useful about the mystery person. Let's say, um, height! Taking this additional information into account should enable us to make a better guess about weight.

   a. How strong is the relationship between Height and Weight? Find the correlation.

   (Analyze→Fit Y by X→choose the variables→Click ▼ and choose Density Ellipse→0.95→Click ▶ next to "*Bivariate Normal Ellipse P=0.95*")

   r =__________

   b. And humor me... r^2 = ______% (Click ▼ and choose Fit Line→Look for *RSquare*)

8. Write the equation that we'd use to predict weight from height.

9. So, how much better is using the regression line than the blind guess-the-average method?

   Let's look at the resulting errors, $(y - \bar{y})$ also known as ______________________________.

10. Save the Residuals (Click

 and choose "*Save Residuals*").
11. We hope this linear model method's errors are generally smaller than the first batch were because we are using *Height* to help make a prediction. How should the boxplot of these errors compare to the one we looked at before? Check that out by creating a parallel boxplot in Graph Builder. Happy?

12. The two boxplots show us that taking the mystery person's height into account generally reduces errors in our estimates of weight. Quantitatively, how much better is this method? To find out, find the sum of the squares of the residuals. Make a new column to calculate these as you did with the *Errors Squared*. Write the result below.

Sum of the $Residuals^2$= _____________________*____________________ = ________________

Mean of the $Residuals^2$ Total number of samples Sum of $Residuals^2$

13. Overall, error is now a lot smaller! Calculate the percent of the original error that still remains.

14. Calculate <u>what percent of the original error has been removed by using the regression line</u>.

15. Where have you seen this number before on the page? Allow yourself a second or two to freak out about how cool this actually is. Explain the connection below.

16. What percent of variation in a person's weight can be explained by the variation in height?

Here’s the data, in case you lost it.

HT(in)	WT(lb)
67	140
71	165
73	168
71	142
74	200
74	175
68	135
73	145
71	150
72	155
69	168
66	106
70	144
71	132
70	140
71	140
70	140
69	130
70	150
74	170
71	175
74	180
72	150
70	150
73	190