Understanding r-squared

Goal: To understand how r^2, the coefficient of determination, describes the **strength** of a linear model. As Rossman points out, r^2 "measures how closely the points fall to the least squares line and thus also provides an indication of how confident one can be of predictions made with the line"[1] Or to paraphrase Moore "When you report a regression, give r^2 as a measure of how successful the model is explaining the response [for a given explanatory value].[2]

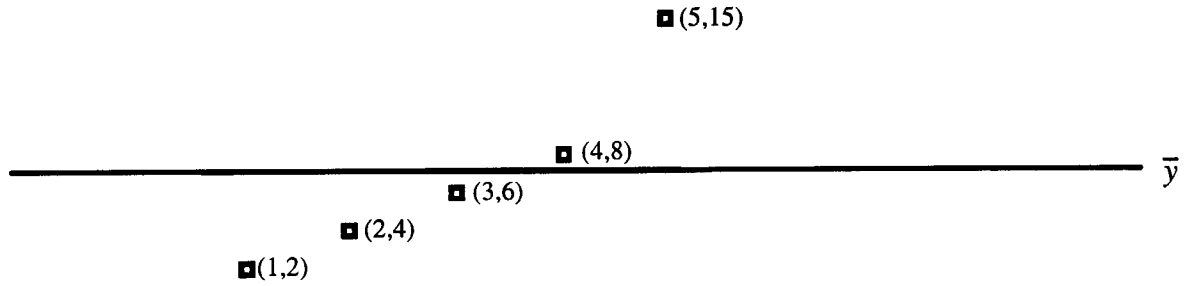
Goal of Modeling

1. To understand r^2 we need to ask the following question.

   If we did not know any modeling techniques such as regression what would the best model be?

   The answer is that, lacking any sophisticated techniques, our best model is the horizontal line containing the mean of the response values, i.e. $y = \bar{y}$.

2. Let's look at an example. Assume the data from a bivariate experiment are the points shown below.[3] On your calculator, draw a scatterplot of the data and a horizontal line representing $\bar{y}$. Your plot should look similar to the picture below.



   a. Is there error in this model?

   b. Draw a segment from each data point showing the error with respect to the model $y = \bar{y}$.

   c. Fill in the following chart:

		(1,2)	(2,4)	(3,6)	(4,8)	(5,15)
error with respect to the model y = 7	$y_i - \bar{y}$					
Square of the errors from the model y = 7	$(y_i - \bar{y})^2$					

   d. Draw shaded "squares" on your plot to represent the squared values just computed (note that since the scales of the axes are not the same, the "squares" will look like rectangles). Some may overlap.

   e. $\sum (y_i - \bar{y})^2 =$ _____.

---

1 *Workshop Statistics.* Allen Rossman. page 139

2 *Basic Practice of Statistics.* David S. Moore. p 127

3 Data set from Gretchen Davis at Santa Monica HS via Eric Mulfinger, Westridge School, Pasadena, CA

Understanding r-squared

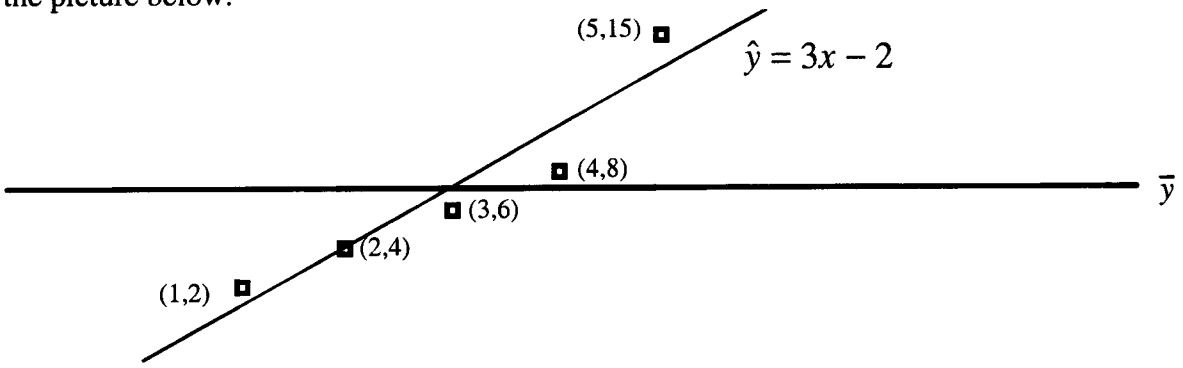
3. So what is the goal of our modeling efforts?

The goal of our modeling effort is to find a better model than the mean of the response variable.

We certainly would want the *sum of the squares of the errors from the new model* to be less than that of *the model using the mean of the response variable.*

A Better Model

1. So let's look for a better model. How about a Least Squares Regression Line (LSR line)? Using the same data as before, add the graph of the LSR line to your plot. Your plot should look like the picture below.



2. Is there still error in this new model? Comparing the two models, which appears to have lower error?

3. Draw a segment from each data point showing the error to the model $\hat{y} = 3x - 2$.

4. What does the "hat" symbol mean on $\hat{y}$?

5. Fill in the following chart:

		(1,2)	(2,4)	(3,6)	(4,8)	(5,15)
error with respect to the model $\hat{y} = 3x - 2$	$y_i - \hat{y}$					
Square of the errors with respect to the model $\hat{y} = 3x - 2$	$(y_i - \hat{y})^2$					

6. Draw shaded "squares" on your plot to represent the values just computed. Does the sum of the areas of these squares seem smaller than those from the model using the response mean?

7. $\sum (y_i - \hat{y})^2 =$ _____.

Understanding r-squared

9. You have determined that $\sum(y_i - \bar{y})^2 = 100$ and $\sum(y_i - \hat{y})^2 = 10$. This information suggests what conclusion?

Comparing Models

1. The natural next step is to find a number which gives us a sense how our new model compares with the model of the mean of the response variable. Of course we would like this number to be r^2.

$$\frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} = \_\_\_\_\_.$$

The answer should be 0.1 . Which of the following correctly describes the proportion you just computed?

a. The proportion of how much error there is in the new model with respect to the error in the mean model.

b. The proportion of how well the new model fits the data.

2. The correct answer to the previous question was *The proportion of how much error there is the new model with respect to the error in the mean model.* Remembering that we want r^2 to show us how *well* our model measures how closely the observed values fall to the least squares line. How would we compute r^2 from the proportion of error in the model? See footnote [4] for a hint.

3. So $r^2 = 1 - \frac{1}{10} = 0.9$. Check this against the r^2 your calculator computed.

So $r^2 = 1 - \dfrac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$

In other words, you find r^2 by finding:

$$1 - \frac{\text{the sum of the squares of the error of the observed data with respect to the model}}{\text{the sum of the squares of the error of the observed data with respect to the mean of the response variable}}$$

4. Experiment with the following Geometer' Sketchpad files located on the file server in the folder APSTATS- GSP: rsqr1pt.gps & rsqr3pts.gsp. Can you explain what each square represents to your teacher?

[4] What is the correct value of r-squared? See your calculator.