Understanding r^2

Consider the following data set:

x	3	1	2	5	4
y	6	2	4	15	8

A Simple Model

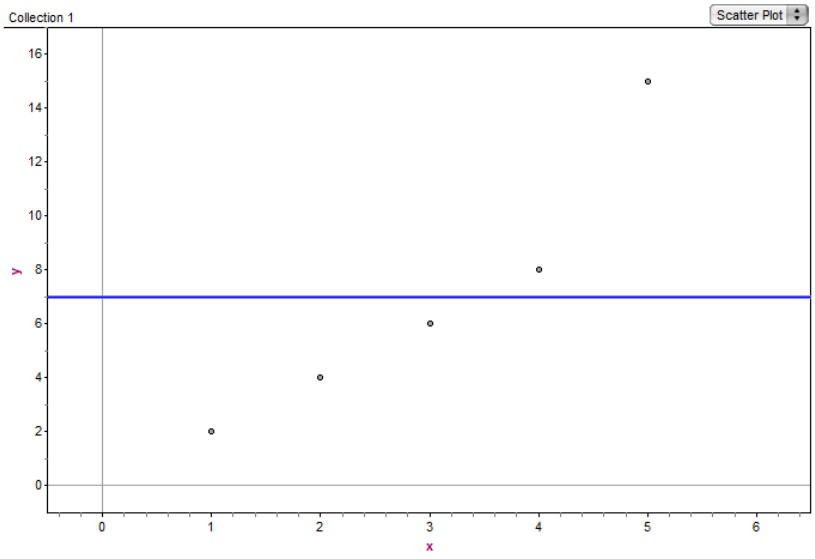
Suppose that an additional point is going to be added to the data set and we want to predict the y-coordinate of the new point. Furthermore, suppose we do not have an explanatory variable, or we have an explanatory variable but don't know how to perform linear regression. Since the mean of the y-coordinates ($\bar{y}$) is 7, the best model for making predictions would be:

$$\hat{y} = 7$$

1. Of course, predictions made using this model won't always be correct. For each of the points in the table above, regardless of what x is, the model predicts that y will be 7. Find the error in our prediction for y (observed y – predicted y) for each of the points in the table below. Then find the square of each error.

		$(3,6)$	$(1,2)$	$(2,4)$	$(5,15)$	$(4,8)$
error from the model $\hat{y} = 7$	$y - \bar{y}$					
square of the error from the model $\hat{y} = 7$	$(y - \bar{y})^2$					

2. For each data point in the graph below, draw a segment representing the error if we use the model $\hat{y} = 7$ for predicting y.



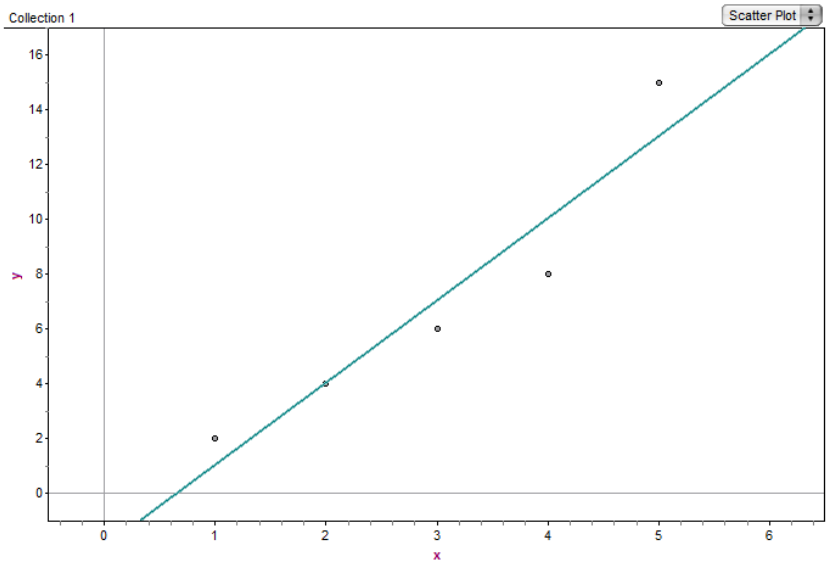
3. Draw shaded squares on the plot to represent the squares of the errors. Some squares may overlap.

4. Find $\Sigma(y - \bar{y})^2$. This quantity is called the **"total sum of squares about the mean $\bar{y}$," or SST**.

A Better Model

The goal of our modeling efforts is to find a better model for predicting y than $\hat{y} = 7$. Let's look at the least-squares regression line for this data set. The plot below shows the same data as before with the LSRL.

5. Of course, there is still error in this new model. How can you tell by looking at the graph?



6. For each of the points in the table below, find the error in our predictions if we use the equation:

$$\hat{y} = 3x - 2$$

.

Then find the square of each error.

		$(3, 6)$	$(1, 2)$	$(2, 4)$	$(5, 15)$	$(4,8)$
error from the model $\hat{y} = 3x - 2$	$y - \hat{y}$					
square of the error from the model $\hat{y} = 3x - 2$	$(y - \hat{y})^2$					

7. For each data point in the graph above, draw a segment representing the error if we use the model $\hat{y} = 3x - 2$ for predicting y.

8. Draw shaded squares on the plot to represent the squares of the errors with this new model.

9. Find $\Sigma(y - \hat{y})^2$. This quantity is called the **"sum of squares for error," or SSE**.

Comparing the Models

The natural next step is to find a number which gives us a sense of how the accuracy of predictions with the equation $\hat{y} = 3x - 2$ compares with the accuracy of predictions with the equation $\hat{y} = 7$. Of course, we would like this number to work out to be r^2. For this problem r^2 equals 0.9 or 90%.

10. We will start by measuring how much less error there is using $\hat{y} = 3x - 2$ to make predictions instead of $\hat{y} = 7$. Do this by using the sum of the squared errors for each equation.

$$SST - SSE =$$

11. Now express the reduction in error from #10 as a proportion of the sum of the squared errors with the original model $\hat{y} = 7$. Your answer should equal r^2.

$$\frac{(SST - SSE)}{SST} =$$