SUPPLEMENT FOR CHAPTER 8: WHY R^2 IS THE FRACTION OF VARIATION ACCOUNTED FOR BY REGRESSION

Note: This should be preceded by the Supplementary Exercises for Chapter 8.

This supplement will first discuss notation, then show the steps of the argument, with blanks for reasons to be filled in.

Notation is as in the Exercise for Chapter 8:

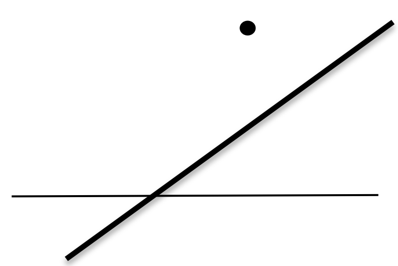
- We have values $x_1, x_2, \ldots, x_n$ for an explanatory variable x, and corresponding values $y_1, y_2, \ldots, y_n$ for the response variable y.
- The least squares line has equation $\hat{y} = b_0 + b_1x$.
- The predicted value for x_i is $\hat{y}_i = b_0 + b_1x_i$.
- The i^{th} residual is $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1x)$.
- $\bar{x}$ = the mean of the x_i's, and $\bar{y}$ = the mean of the y_i's

In addition, we define:

i) The *total sum of squares* as $SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
ii) The *residual sum of squares* as $SS_R = \sum_{i=1}^{n} e_i^2$
iii) The *model sum of squares* as $SS_M = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Draw a Picture: In the following diagram, the thick (slanting) line is the least squares line, the thin (horizontal) line is the line $y = \bar{y}$, and the dot is the point (x_i, y_i).

i. Label the least squares line, the line $y = \bar{y}$, and the point (x_i, y_i).
ii. Draw and label the point $(x_i, \hat{y}_i)$.
iii. Show and label each of the following in a way that shows the connection between these three quantities:
 $y_i - \hat{y}_i$ (also known as e_i), $\hat{y}_i - \bar{y}$, and $y_i - \bar{y}$



Comments:

- SS_{total} can be considered as a measure of the *total variability* of y.
  - *Think* about why this is reasonable. (Using the picture or adding more to it might help.)
- The variance of y is $s_y^2 = SS_{total}/(n-1)$, so SS_{total} can be considered as just a rescaling of s_y^2.

- SS_M can also be considered as a measure of the total variability of the predicted values $\hat{y}_i$.
  - *Think* about why this is reasonable.
  - Remembering from the exercises what the mean of the $\hat{y}_i$'s is may help.
  - Using the picture or adding to it may help.
- SS_{total} is sometimes called TSS or SST
- SS_M is sometimes called SSM or MSS. (It is also sometimes called the sum of squares for regression, which has abbreviations that can be confused with those for SS_R.)
- SS_R is sometimes called SSR or RSS. It is also sometimes called the error sum of squares, leading to abbreviations SS_E, SSE, and ESS.

Now the calculations, with blanks to fill in with reasons (mostly either definitions or results of exercises) for the non-algebra steps:

I. $SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ ________________________________

$= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$

$= \sum_{i=1}^{n}[(y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2]$

$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ ________________________________

$= \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

$= SS_R + SS_M$ ________________________________

This says that the total variation of y (SS_{total}) can be broken up into the sum of the residual sum of squares (SS_R) and the model sum of squares (SS_M). The residual sum of squares can be interpreted as the variation that is *left over* after accounting for a linear effect of x, so our equation says we can interpret SS_M as the amount of the total variation that is *accounted for* by regression on x.

II. $SS_M = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ ________________________________

$= b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2$ ________________________________

$= \left(r\frac{s_y}{s_x}\right)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$ ________________________________

$= \left(r\frac{s_y}{s_x}\right)^2 (n-1)s_x^2$ ________________________________

$= r^2(n-1)\, s_y^2$

$= r^2 \, SS_{total}$ ______________________________

Combining Parts I and II, we have:

The *proportion* of total variation of y that is accounted for by regression on x is

$\frac{SS_M}{SS_{total}}$ (Part ____)

$= \frac{r^2 SS_{total}}{SS_{total}}$ (Part ____)

$= r^2$