Partitioning of Sums of Squares in Simple Linear Regression

George H Olson, Ph. D.
Leadership and Educational Studies
Appalachian State University

The *parametric* model for the regression of *Y* on *X* is given by

$$Y_i = \alpha + \beta X_i + \varepsilon_i. \tag{1}$$

The model for the regression of *Y* on *X* in a *sample* is

$$Y_i = a + bX_i + e_i. \tag{2}$$

Calculation of the constants in the model:

the *slope* (*b*) is given by

$$b = \frac{\sum x_i y_i}{\sum x_i^2}, \tag{3}$$

where $x_i = \left(X_i - \overline{X}\right)$,

and the intercept by

$$a = \overline{Y} - b\overline{X}. \tag{4}$$

Using the coefficients, *a* and b, we can construct an equation for estimating (or *predicting*) an individual's score on *Y*:

$$\hat{Y}_i = a + bX_i \tag{5}$$

A closer look at the regression equation, (5), and using (3) and (4) leads us to

$$\begin{aligned}\hat{Y}_i &= a + bX_i \\ &= (\bar{Y} - b\bar{X}) + bX_i \\ &= \bar{Y} + b(X_i - \bar{X}) \\ &= \bar{Y} + bX_i.\end{aligned} \tag{6}$$

Partitioning the *Sum of Squares*, $\sum(\hat{Y}_i - \bar{Y})^2$.

First, consider the following identity

$$Y_i = \bar{Y} + (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i). \tag{7}$$

If we subtract $\bar{Y}$ from each side of the equation, we obtain

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \tag{8}$$

After squaring and summing, we have

$$\begin{aligned}\sum(Y_i - \bar{Y})^2 &= \sum[(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2 \\ &= \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 \\ &\quad +2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)\end{aligned} \tag{9}$$

or, after simplifying[1],

$$\begin{aligned}\sum y_i^2 &= \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y})^2 \\ &= SS_{reg} + SS_{res}\end{aligned} \tag{10}$$

Where SS_{reg} = regression sum of squares, and SS_{res} = residual sum of squares.

Dividing (10) through by the total sum of squares, SS_{tot} (= $\sum y^2$) gives

$$\frac{\sum y_i^2}{\sum y_i^2} = \frac{SS_{reg}}{\sum y_i^2} + \frac{SS_{res}}{\sum y_i^2} \tag{11}$$

or

[1] See the appendix to see how the equation simplifies.

$$1 = \frac{SS_{reg}}{\sum y_i^2} + \frac{SS_{res}}{\sum y_i^2} \qquad (12)$$

A computational example

It is often useful to devise simple computational examples, such as the following:

Y	*X*
3	1
1	0
0	1
4	-1
5	2

The means of the two variables, *Y* and *X* are

$$\bar{X} = \frac{\sum X_i}{n} = \frac{3}{5} = .6$$

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{13}{5} = 2.6$$

Having computed the means, we now compute the deviations, squares of deviations, and cross-products of deviations

Deviations, squares, and cross-products

Y	*y*	y^2	*X*	*x*	x^2	*xy*
3	.4	.16	1	.4	.16	.16
1	-1.6	2.56	0	-.6	.36	.96
0	-2.6	6.76	1	-1.6	2.56	4.16
4	1.4	1.96	-1	.4	.16	.56
5	2.4	5.76	2	1.4	1.96	3.36

The *sums* of squares and cross-products are computed as

$$\sum x_i^2 = 5.2$$
$$\sum y_i^2 = 17.2$$
$$\sum x_i y_i = 9.2$$

and the regression coefficients as

$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{9.2}{5.2} = 1.769,$$

and

$a = \overline{Y} - b\overline{X}$ = 2.6 – 1.769 * .6 = 2.6 – 1.061 = 1.539.

The *regression equation* can now be written as

$$Y_i' = 1.539 + 1.769 X_i$$

From an earlier equation (10) we obtain

$$\begin{aligned} SS_{reg} &= \sum(\hat{Y}_i - \overline{Y})^2 \\ &= \sum[(\overline{Y} + bx_i) - \overline{Y}]^2 \\ &= \sum(\overline{Y} + bx_i - \overline{Y})^2 \\ &= \sum(bx_i)^2 \\ &= b^2 \sum x_i^2 \\ &= (\frac{\sum x_i y_i}{\sum x^2}) \sum x_i^2 \\ &= \frac{(\sum x_i y_i)^2}{\sum x^2} \\ &= \frac{84.64}{5.2} \\ &= 16.28. \end{aligned} \tag{13}$$

Note that we could also have computed

$$\begin{aligned} SS_{reg} &= b \sum x_i y_i \\ &= 1.769 * 9.2 \\ &= 16.28. \end{aligned} \tag{14}$$

An alternative calculation is given by

$$\begin{aligned} SS_{reg} &= b^2 \sum x_i^2 \\ &= (1.769)^2 * 5.2 \\ &= 3.1294 * 5.2 \\ &= 16.28 \end{aligned} \tag{15}$$

Hence,

$$SS_{res} = \sum y_i^2 - SS_{reg} \tag{16}$$
$$= 17.2 - 16.28$$
$$= .92$$

The equation for the Pearson correlation is

$$r_{xy}^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2} \tag{17}$$

Therefore, SS_{res} can also be computed as

$$\begin{aligned} r_{xy}^2 \sum y_i^2 &= \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2} \sum y_i^2 \\ &= \frac{(\sum x_i y_i)^2}{\sum x_i^2} \\ &= SS_{reg} \end{aligned} \tag{18}$$

Appendix

Showing the Simplification of the Partitioning of SS_Y

Beginning with,

$$\begin{aligned} \sum (Y_i - \overline{Y})^2 &= \sum [(\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)]^2 \\ &= \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \\ &\quad +2\sum (\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i), \end{aligned}$$

we need to show that

$$2\sum (\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i) = 0.$$

Recalling (4, 5 & 6) we can write,

$$
\begin{aligned}
2\sum(\hat{Y}_i-\bar{Y})(Y_i-\hat{Y}) &= 2\sum((\bar{Y}-b\bar{X}+bX_i)-\bar{Y})(Y_i-\hat{Y}_i) \\
&= 2\sum((\bar{Y}+b(X_i-\bar{X})-\bar{Y})(Y_i-\hat{Y}_i)) \\
&= 2\sum(b(X_i-\bar{X})(Y_i-\hat{Y}_i)) \\
&= 2\sum b(X_i-\bar{X})(Y_i-(\bar{Y}-b\bar{X})+bX_i) \\
&= 2b\sum(X_i-\bar{X})(Y_i-(\bar{Y}+b(X_i-\bar{X}))) \\
&= 2b\sum(X_i-\bar{X})((Y_i-\bar{Y})-b(X_i-\bar{X})) \\
&= 2b\sum((X_i-\bar{X})(Y_i-\bar{Y})-b(X_i-\bar{X})^2) \\
&= 2b\sum(x_iy_i-bx_i^2) \\
&= 2b\left[\sum x_iy_i-\left(\frac{\sum x_iy_i}{\sum_i x^2}\right)\sum x_i^2\right] \\
&= 2b(\sum x_iy_i-\sum x_iy_i) \\
&= 0.
\end{aligned}
$$