14.3: Partitioning Sums of Squares

Learning Objectives

- Compute the sum of squares Y
- Convert raw scores to deviation scores
- Compute predicted scores from a regression equation
- Partition sum of squares Y into sum of squares predicted and sum of squares error
- Define r^2 in terms of sum of squares explained and sum of squares Y

One useful aspect of regression is that it can divide the variation in Y into two parts: the variation of the predicted scores and the variation of the errors of prediction. The variation of Y is called the sum of squares Y and is defined as the sum of the squared deviations of Y from the mean of Y. In the population, the formula is

$$SSY = \sum (Y - \mu_Y)^2 \tag{14.3.1}$$

where SSY is the sum of squares Y, Y is an individual value of Y, and μ_y is the mean of Y. A simple example is given in Table 14.3.1. The mean of Y is 2.06 and SSY is the sum of the values in the third column and is equal to 4.597.

Table 14.3.1: *Example of SSY*

Y	Y-M_y	(Y-M_y)^2
1.00	-1.06	1.1236
2.00	-0.06	0.0036
1.30	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

When computed in a sample, you should use the sample mean, M, in place of the population mean:

$$SSY = \sum (Y - M_Y)^2 \tag{14.3.2}$$

It is sometimes convenient to use formulas that use deviation scores rather than raw scores. Deviation scores are simply deviations from the mean. By convention, small letters rather than capitals are used for deviation scores. Therefore the score y indicates the difference between Y and the mean of Y. Table 14.3.2 shows the use of this notation. The numbers are the same as in Table 14.3.1.

Table 14.3.2: *Example of SSY using Deviation Scores*

Y	**y**	**y^2**
1.00	-1.06	1.1236
2.00	-0.06	0.0036
1.30	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

The data in Table 14.3.3 are reproduced from the introductory section. The column X has the values of the predictor variable and the column Y has the criterion variable. The third column, y, contains the differences between the column Y and the mean of Y.

Table 14.3.4: *Example data (The last row contains column sums)*

X	**Y**	**y**	**y^2**	**Y'**	**y'**	**y'^{2}**	**Y-Y'**	**(Y-Y')^2**

X	Y	y	y^2	Y'	y'	y'^2	Y-Y'	$(Y-Y')^2$
1.00	1.00	-1.06	1.1236	1.210	-0.850	0.7225	-0.210	0.044
2.00	2.00	-0.06	0.0036	1.635	-0.425	0.1806	0.365	0.133
3.00	1.30	-0.76	0.5776	2.060	0.000	0.0000	-0.760	0.578
4.00	3.75	1.69	2.8561	2.485	0.425	0.1806	1.265	1.600
5.00	2.25	0.19	0.0361	2.910	0.850	0.7225	-0.660	0.436
Sums								
15.00	10.30	0.00	4.597	10.300	0.000	1.806	0.000	2.791

The fourth column, y^2, is simply the square of the y column. The column Y' contains the predicted values of Y. In the introductory section, it was shown that the equation for the regression line for these data is

$$Y' = 0.425X + 0.785. \tag{14.3.3}$$

The values of Y' were computed according to this equation. The column y' contains deviations of Y' from the mean of Y' and y'^2 is the square of this column. The next-to-last column, $Y - Y'$, contains the actual scores (Y) minus the predicted scores (Y'). The last column contains the squares of these errors of prediction.

We are now in a position to see how the SSY is partitioned. Recall that SSY is the sum of the squared deviations from the mean. It is therefore the sum of the y^2 column and is equal to 4.597. SSY can be partitioned into two parts: the sum of squares predicted (SSY') and the sum of squares error (SSE). The sum of squares predicted is the sum of the squared deviations of the predicted scores from the mean predicted score. In other words, it is the sum of the y'^2 column and is equal to 1.806. The sum of squares error is the sum of the squared errors of prediction. It is therefore the sum of the $(Y - Y')^2$ column and is equal to 2.791. This can be summed up as:

$$SSY = SSY' + SSE \tag{14.3.4}$$

$$4.597 = 1.806 + 2.791 \tag{14.3.5}$$

There are several other notable features about Table $14.3.3$. First, notice that the sum of y and the sum of y' are both zero. This will always be the case because these variables were created by subtracting their respective means from each value. Also, notice that the mean of $Y - Y'$ is 0. This indicates that although some Y values are higher than their respective predicted Y values and some are lower, the average difference is zero.

The SSY is the total variation, the SSY' is the variation explained, and the SSE is the variation unexplained. Therefore, the proportion of variation explained can be computed as:

$$\text{Proportion explained} = \frac{SSY'}{SSY} \tag{14.3.6}$$

Similarly, the proportion not explained is:

$$\text{Proportion not explained} = \frac{SSE}{SSY} \tag{14.3.7}$$

There is an important relationship between the proportion of variation explained and Pearson's correlation: r^2 is the proportion of variation explained. Therefore, if $r = 1$, then, naturally, the proportion of variation explained is 1; if $r = 0$, then the proportion explained is 0. One last example: for $r = 0.4$, the proportion of variation explained is 0.16.

Since the variance is computed by dividing the variation by N (for a population) or $N - 1$ (for a sample), the relationships spelled out above in terms of variation also hold for variance. For example,

$$\sigma^2_{total} = \sigma^2_{Y'} + \sigma^2_e \tag{14.3.8}$$

where the first term is the variance total, the second term is the variance of Y', and the last term is the variance of the errors of prediction ($Y - Y'$). Similarly, r^2 is the proportion of variance explained as well as the proportion of variation explained.

Summary Table

It is often convenient to summarize the partitioning of the data in a table. The degrees of freedom column (df) shows the degrees of freedom for each source of variation. The degrees of freedom for the sum of squares explained is equal to the number of predictor variables. This will always be 1 in simple regression. The error degrees of freedom is equal to the total number of observations minus 2. In this example, it is $5 - 2 = 3$. The total degrees of freedom is the total number of observations minus 1.

Table 14.3.4: *Summary Table for Example Data*

Source	Sum of Squares	df	Mean Square
Explained	1.806	1	1.806
Error	2.791	3	0.930
Total	4.597	4	

Contributor

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.