Sum Squared Errors

Paul Glezen

August 2016

Abstract

A footnote in OpenIntro Statistics, 3rd Edition[1], Section 5.5.2, Analysis of Variance (ANOVA), relates an identity for the *sum of squared errors* (SSE).

$$\begin{aligned} \text{SSE} &= \text{SST} - \text{SSG} \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \end{aligned} \tag{1}$$

where s_i^2 is the sample variance for group i among k groups. This note fleshes out this identity in more detail.

1 Definitions

We define the following notation and definitions.

n — The number of elements in a sample.

k — The number of groups in the sample.

n_j — The number of elements in group j, $n = \sum_{j=1}^{k} n_j$.

x_i — The i^{th} element in the sample.

$\bar{x}$ — The sample average.

$\bar{x}_j$ — The average of elements in group j.

s_j^2 — The sample variance for group j.

SST — Sum of Squares Total $= \sum_{i=1}^{n} (x_i - \bar{x})^2$.

SSG — Sum of Squares between Groups $= \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$

SSE — Sum of Squared Errors = SST - SSG

[1] Available at `https://www.openintro.org/stat/textbook.php`

2 Group Sample Variance

The sample group variance is calculated like any other sample variance, except that the calculation is restricted to a group. We'll designate x_{ij} as element i in group j. Using this notation, the formula for the group average is $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$. So

$$n_j \bar{x}_j = \sum_{i=1}^{n_j} x_{ij} \tag{2}$$

This is often used in the calulations that follow. For the group variance of group j, we have

$$\begin{aligned}
s_j^2 &= \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 \\
&= \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij}^2 - 2x_{ij}\bar{x}_j + \bar{x}_j^2) \\
&= \frac{1}{n_j - 1}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{2\bar{x}_j}{n_j - 1}\sum_{i=1}^{n_j} x_{ij} + \frac{\bar{x}_j^2}{n_j - 1}\sum_{i=1}^{n_j} 1 \\
&= \frac{1}{n_j - 1}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{2\bar{x}_j n_j \bar{x}_j}{n_j - 1} + \frac{\bar{x}_j^2 n_j}{n_j - 1} \\
&= \frac{1}{n_j - 1}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{n_j}{n_j - 1}\bar{x}_j^2 \qquad (3)
\end{aligned}$$

We will use this identity in the following section.

3 Sum of Square Total

Let $G_j = \{x_{ij}, i = 1 \ldots n_j\}$ be the subset of $\{x_i\}$ for each group so that $G_i \bigcap G_j = \emptyset$ when $i \neq j$ and $\bigcup_{j=1}^{k} G_j$ is the whole sample. The sum of squares total can then be expressed as

$$\text{SST} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 \tag{4}$$

The point of regrouping the sums is to combine terms in the sum of squared errors expression.

4 Sum of Squared Errors

Now we can express the sum of squared errors as

$$
\begin{aligned}
\text{SSE} &= \text{SST} - \text{SSG} \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 - \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 - \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 \qquad (5) \\
&= \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 - n_j(\bar{x}_j - \bar{x})^2\right] \\
&= \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(x_{ij}^2 - 2x_{ij}\bar{x} + \bar{x}^2) - n_j(\bar{x}_j^2 - 2\bar{x}_j\bar{x} + \bar{x}^2)\right] \\
&= \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j} x_{ij}^2 - 2\bar{x}\sum_{i=1}^{n_j} x_{ij} + \bar{x}^2\sum_{i=1}^{n_j} 1 - n_j\bar{x}_j^2 + 2n_j\bar{x}_j\bar{x} - n_j\bar{x}^2\right] \\
&= \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j} x_{ij}^2 - 2\bar{x}(n_j\bar{x}_j) + \bar{x}^2 n_j - n_j\bar{x}_j^2 + 2n_j\bar{x}_j\bar{x} - n_j\bar{x}^2\right] \qquad (6) \\
&= \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j} x_{ij}^2 - n_j\bar{x}^2\right] \\
&= \sum_{j=1}^{k}(n_j - 1)\left[\frac{1}{n_j - 1}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{n_j}{n_j - 1}\bar{x}^2\right] \\
&= \sum_{j=1}^{k}(n_j - 1)s_j^2 \qquad (7)
\end{aligned}
$$

In step (5) we used the SST identity in (4). In step (6) we group average identity in (2). In the last step, we used the group variance expression from (3).