Analysis of Variance Approach to Regression Analysis

Partitioning of Total Sum of Squares

The analysis of variance approach is based on the partitioning of sum of squares and degrees of freedom associated with the response variable Y.

The measure of total variation, denoted by **SSTO**, is the sum of squared deviations.

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{3.19}$$

Alternative formula:

$$SSTO = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \tag{3.20}$$

Take Note!!!

- If all Y_i observations are the same, $SSTO = 0$.
- The greater the variation among the Y_i observations, the larger $SSTO$.

The measure of variation in the Y_i observations that is present when the predictor variable X is taken into account is the sum of of squared deviations which is known as **SSE**. (It is the same in **3.15**)

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 = \sum_{i=1}^{n} e_i^2$$

Take Note!!!

- If all Y_i observations fall on the fitted regression line, $SSE = 0$.
- The greater the variation of the Y_i observation around the fitted regression line, the larger the SSE.

Each deviations is simply the difference between the fitted value on the regression line and the mean of the fitted values $\bar{Y}$. Sum of Squared Regression **(SSR)** may considered a measure of the part of the variability of the Y_i which is associated with the regression line.

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{3.21}$$

Exercise:

Show that $SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$. *(Alternative formula)*

Take Note!!!

- The larger SSR in the relation to $SSTO$, the greater the effect of the regression relation in accounting for the total variation in the Y_i observation.

Formal Development of Partitioning

The total deviation $Y_i - \bar{Y}$, used in the measure of the total variations Y_i without taking the predictor variable into account, can be decomposed into two components.

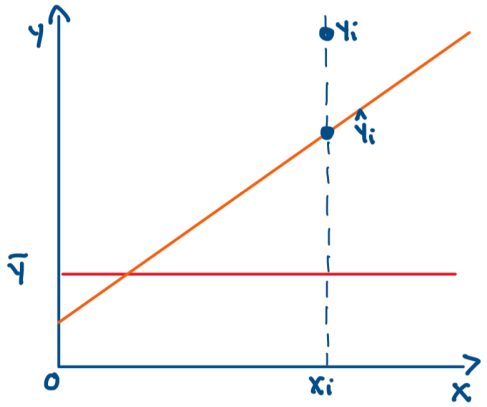
$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

1. The deviation of the fitted value $\hat{Y}_i$ around the mean $\bar{Y}$.
2. The deviation of Y_i around the regression line.

It is remarkable property that the sums of these squared deviations have the same relationship:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad \textbf{(3.22)}$$

SSTO = SSR + SSE



Exercise:

Prove this basic result in the analysis of variance: $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Breakdown of Degrees of Freedom

1. $n-1$ df for $SSTO$. One degrees of freedom is lost because the deviations $Y_i - \bar{Y}$ are subject to one constraint: they must sum to zero. Equivalently, one degree of freedom is lost because the sample mean $\bar{Y}$ is used to estimate the population mean.
2. $n-2$ for SSE. Two degrees of freedom are lost because the two parameters β_0 and β_1 are estimated in obtaining the fitted values $\hat{Y}_i$.
3. 1 df for SSR. Although there are n deviations $\hat{Y}_i - \bar{Y}$, all fitted values $\hat{Y}_i$ are calculated from the same estimated regression line.

Note that the degrees of freedom (df) are additive.

$$SSTO = SSR + SSE$$

$$n - 1 = 1 + n - 2$$

$$n - 1 = n - 1$$

Mean Square

A sum of squares divided by its associated degrees of freedom is called a **Mean Square.**

The regression mean square is denoted by, MSR:

$$MSR = \frac{SSR}{1} = SSR$$

The error mean square is denoted by, MSE. It is a variance error of the estimator mentioned in the previous topic **(3.17)**.

$$MSE = \frac{SSE}{n-2} = \hat{\sigma}^2$$

Two mean squares MSR and MSE do not add to $\frac{SSTO}{n-1}$. Thus, mean squares are not additive.

Expected Mean Squares

In order to make inferences based on the analysis of variance approach, we need to know the expected value of each of the mean squares.

The expected value of a mean square is the mean of its sampling distribution and tell us what is being estimated by the mean square. We stated earlier that MSE is an unbiased estimator of the error variance **(3.17)**.

$$E(MSE) = \sigma^2 \quad \text{or} \quad E(\hat{\sigma}^2) = \sigma^2 \tag{3.23}$$

The expected value of MSR is:

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{3.24}$$

Exercise:

Show that $E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$.

Important implications of the expected mean squares:

1. The mean of the sampling distribution of MSE is σ^2 whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$.
2. The mean of the sampling distribution of MSR is also σ^2 when $\beta_1 = 0$. Hence, when $\beta_1 = 0$, the sampling distribution of MSR and MSE are located identically and MSR and MSE will tend to be of the same order of magnitude.
3. When $\beta_1 \neq 0$, the mean of the sampling distribution of MSR is greater than σ^2 since the term $\beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$ must be positive. Thus, when $\beta_1 \neq 0$, the mean of the sampling distribution of MSR is located to the right of the MSE and hence, *MSR will tend to be larger than MSE.*

4. Comparison of MSE and MSR is useful for testing whether or not $\beta_1 = 0$.
5. If MSR and MSE are of the same order of magnitude, this would suggest that $\beta_1 = 0$.
6. If MSR is substantially greater than MSE, this would suggest that $\beta_1 \neq 0$.

ANOVA Test for Simple Linear Regression

Table 1. ANOVA table for Simple Linear Regression

Source of Variation	Sum of Square (SS)	Degrees of Freedom (df)	Mean Square (MS)	F - value	Critical Value
Regression	SSR	$df_1 = 1$	MSR	F	$F_{1-\alpha,df_1,df_2}$
Error/Residual	SSE	$df_2 = n - 2$	MSE		
Total	SSTO	$df_3 = n - 1$			

The general analysis of variance approach provides us with a battery of highly useful tests for regression models (and other linear statistical models). For the simple regression case considered here, the analysis of variance provides us with a test for:

$$H_0 : \beta_1 = 0 \qquad \text{and} \qquad H_1 : \beta_1 \neq 0 \tag{3.25}$$

Test statistic: The test statistic for the analysis of variance approach is denoted by F. As just mentioned, it compares MSR and MSE in the following fashion:

$$F = \frac{MSR}{MSE} \tag{3.26}$$

Take Note!!!

The large values of F support H_1 and values of F near 1 support H_0. In other words, the appropriate test is an upper-tail one.

General Linear Test Approach

The analysis of variance test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is an example of the general test for a linear regression model.

Full Model

For the simple linear regression case, the full model is the normal error regression model stated in **(3.1)**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \textbf{Full or Unrestricted Model}$$

We shall denote this sum of squares by $SSE(F)$ to indicate that it is the error sum of squares of the full model.

$$SSE(F) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2 = SSE$$

Thus, for the full model, the error sum of squares (SSE), which measures the variability of the Y_i observations around the fitted regression line.

Reduced Model

The model when H_0 holds, is called the reduced or restricted model. When $\beta_1 = 0$. We shall denote this sum of squares by $SSE(R)$.

$$SSE(R) = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = SSTO$$

Exercise:

Show that $SSE(R) = SSTO$.

If we compare $SSE(F)$ and $SSE(R)$. It can be seen that:

- $SSE(F) \leq SSE(R)$. The more parameters are in the model, the better one can fit the data and the smaller are deviations around the fitted regression function.
- When $SSE(F)$ is close to $SSE(R)$, the added parameters in the full model really do not help to reduce the variation in the Y_i about the fitted regression function. Thus, a small difference $SSE(R) - SSE(F)$ suggest that H_0 holds.
- A large difference H_1 holds because the additional parameters in the model help to reduce substantially the variation of the observations Y_i around the fitted regression function.