Analysis of Variance and Related Topics for Ordinary Least Squares Regression

The ANOVA Table for OLS Regression

The Analysis of Variance or ANOVA Table is a fundamental output from a fitted OLS regression model. The output from the ANOVA table is used for a number of purposes.

- Show the decomposition of the total variation.
- Compute the R-Squared and Adjusted R-Squared metrics.
- Perform the Overall F-test for a regression effect.
- Perform a F-test for nested models as commonly used in forward, backward, and stepwise variable selection.

Decomposing the Sample Variation

- The Total Sum of Squares is the total variation in the sample.
- The Regression Sum of Squares is the variation in the sample that has been explained by the regression model.
- The Error Sum of Squares is the variation in the sample that cannot be (or has not been) explained.

SST	$\sum_i^n (Y_i - \bar{Y})^2$	Total Sum of Squares
SSR	$\sum_i^n (\widehat{Y}_i - \bar{Y})^2$	Regression Sum of Squares
SSE	$\sum_i^n (Y_i - \widehat{Y}_i)^2$	Error Sum of Squares

The Coefficient of Determination - R-Squared

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{1}$$

- The Coefficient of Determination R^2 will take values $0 \leq R^2 \leq 1$ and represents the proportion of the variance explained by the regression model.

- Implicitly, R^2 is a function of the number of parameters in the model. For a nested subset of predictor variables $p_0 < p_1$, i.e. p_1 contains the original p_0 predictor variables and some new predictor variables, R^2 will have a monotonic relationship such that $R^2(p_0) \leq R^2(p_1)$.

Adjusted R-Squared

$$R^2_{ADJ} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} \tag{2}$$

- Note that standard regression notation uses k for the number of predictor variables included in the regression model and p for the total number of parameters in the model. When the model includes an intercept term, then $p = k + 1$. When the model does not include an intercept term, then $p = k$.

- The Adjusted R-Squared metric accounts for the *model complexity* of the regression model allowing for models of different sizes to be compared.

- The Adjusted R-Squared metric will not be monotonic in the number of model parameters.

The Overall F-Test for a Regression Effect

Consider the regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k. \tag{3}$$

The Overall F-Test for a regression effect is a joint hypothesis test that at least one of the predictor variables has a non-zero coefficient.

- The null and alternate hypotheses are given by

  $$H_0 : \beta_1 = \cdots = \beta_k = 0 \qquad \text{versus} \qquad H_1 : \beta_i \neq 0 \tag{4}$$

  for some $i \in \{1, \ldots, k\}$.

- The test statistic for the Overall F-test is given by

  $$F_0 = \frac{SSR/k}{SSE/(n-p)} \tag{5}$$

  which has a F-distribution with $(k, n-p)$ degrees-of-freedom for a regression model with k predictor variables and p total parameters. When the regression model includes an intercept, then $p = k + 1$. If the regression model does not include an intercept, then $p = k$.

The F-Test for Nested Models

For our discussion of nested models let's consider two concrete examples which we will refer to as the *full model* (FM)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{6}$$

and the *reduced model* (RM).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{7}$$

Notice that the predictor variables in the reduced model are a subset of the predictor variables in the full model, i.e. $RM \subset FM$.

- In this notation we say that the FM *nests* the RM, or the RM is *nested by* the FM.

- We only use the terms *full model* and *reduced model* in the context of nested models.

- We can use a F-test for nested models to decide whether or not to include an additional predictor variable in the final model.

Details for the F-Test for Nested Models

Given a full model and a reduced model we can perform a F-test for nested models for the exclusion of a single predictor variable or multiple predictor variables.

In the context of our concrete example, we could test either of these null hypotheses.

- Example 1: Test a Single Predictor Variable

$$H_0 : \beta_3 = 0 \qquad \text{versus} \qquad H_1 : \beta_3 \neq 0 \tag{8}$$

- Example 2: Test Multiple Predictor Variables

$$H_0 : \beta_2 = \beta_3 = 0 \qquad \text{versus} \qquad H_1 : \beta_i \neq 0 \tag{9}$$

  for some $i \in \{2, 3\}$.

Details for the F-Test for Nested Models - Continued

The test statistic for the F-test for nested models will always have this form in terms of the FM and the RM.

- Test Statistic for the Nested F-Test

$$F_0 = \frac{[SSE(RM) - SSE(FM)] \,/\, [dim(FM) - dim(RM)]}{SSE(FM) \,/\, [n - dim(FM)]} \tag{10}$$

- The test statistics is based on the reduction in the SSE obtained from adding additional predictor variables. Note that SSE(FM) is always less than SSE(RM).

- The *dimension* of a statistical model is the number of parameters.

Connection to Forward Variable Selection

The F-test for nested models is the standard statistical test implemented in most statistical software packages for performing forward and backward, and hence stepwise, variable selection.

Forward Variable Selection

- Given the model $Y = \beta_0 + \beta_1 X_1$ and a set of candidate predictor variables $Z_1, \ldots, Z_s$, how do we select the best Z_i to include in our model as X_2?

- In forward variable selection the FM will be $Y = \beta_0 + \beta_1 X_1 + \beta_2 Z_i$ and the RM will be $Y = \beta_0 + \beta_1 X_1$. The forward variable selection algorithm will select the Z_i with the largest F-statistic that is statistically significant at a predetermined level. The algorithm will continue to add predictor variables until there are no predictor variables that are statistically significant to the predetermined level.

Backward Variable Selection

- Given the model
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_s X_s \tag{11}$$
how do we eliminate predictor variables whose effects are not statistically significant?

- In backward variable selection the FM will be $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_s X_s$ and the RM will be $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{s-1} X_{s-1}$, for notational convenience. The backward variable selection algorithm will drop the X_i with the smallest F-statistic that is not statistically significant at a predetermined level. The algorithm will continue to drop predictor variables until there are no predictor variables that aren't statistically significant to the predetermined level.

- Note that both of the forward and backward variable selection procedures consider only one variable at each iteration.