

Advanced Regression

The basic population regression function specification for the multiple regression model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i.$$

Where Y is the dependent variable and X1 to Xk are independent variables. This model has k-variables, and each coefficient is a partial regression coefficient, reflecting the relationship between a one unit increase in X and the dependent variable Y holding all other factors, or variables, constant. Remember that under the assumptions of the Gauss-Markov theorem, the partial regression estimates for β_0 to β_k are the “Best Linear Unbiased Estimates” among all unbiased linear estimators. (I leave it to you to list these assumptions.)

Interpretation of Coefficients

Given the assumptions of the Gauss-Markov theorem, the conditional expectation of Y is given as:

$$E(Y_i | X_{1i}, \dots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}.$$

In other words, the regression delivers the conditional mean of Y, or the expected value of Y, conditional on the given fixed values of X1 to Xk. You should always remember that a regression just estimates a conditional mean, the average value of Y given a set of observable factors such as X1 to Xk.

For a single partial regression coefficient, you interpret its coefficient simply as a partial derivative. For the coefficient on X1 above, β_1 represents the change in the mean value of Y per unit change in X1, holding all other variables constant (or we might say *ceteris paribus*).

This comes directly from:

$$\frac{\partial E(Y_i)}{\partial X_{1i}} = \beta_1.$$

Goodness of Fit

Remember for any regression model, the total sum of squares is just the sum of the explained sum of squares and the residual, or error, or unexplained, sum of squares. Mathematically this is expressed as:

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

From this, we can express the R-square measure as:

$$R^2 = \frac{\text{ExplainedSS}}{\text{TotalSS}} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Or alternatively, we can express this as:

$$R^2 = 1 - \left(\frac{\text{UnExplainedSS}}{\text{TSS}} \right) = 1 - \left(\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \right).$$

Note that the R-square ranges in value from 0 to 1, and is sometimes called the multiple correlation coefficient. Further remember that the R-square will rise even if only one statistically insignificant variable is added to the model. As an alternative, consider the adjusted R-square. This measure adjusts for degrees of freedom, and thus how many controls there are in the model. In practice, the R-square and adjusted R-square are close in value. I leave it to you to choose which you want to report. The adjusted R-square is given as:

$$\bar{R}^2 = 1 - \left(\frac{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n-k)}}{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n-1)}} \right).$$

Where n is sample size and k is the number of variables included in the model.

Hypothesis Testing

Test 1 – Testing the Significance of a particular partial regression coefficient

If

$$\varepsilon_i \sim N(0, \sigma^2),$$

Then we can use the t-test to examine the significance of an individual coefficient. In this case, the hypotheses of a two-tail test would be:

$$H_0: \beta_1 = 0;$$

$$H_1: \beta_1 \neq 0.$$

The test statistic for this hypothesis is given as:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}.$$

Which has n-k-1 degrees of freedom (where the “1” represents the intercept).

A 95% confidence interval for the estimated regression coefficient is given by:

$$\hat{\beta}_1 - t_{\alpha/2} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2} SE(\hat{\beta}_1).$$

Where alpha represents the level of statistical significance.

When interpreting a regression coefficient, you should discuss three things: size, sign, and significance. Perhaps of greatest interest to us as policy makers is the size of an estimated coefficient. It is possible, with a large sample size, to get a statistically significant result for any sized parameter, but the big question then becomes “is this coefficient estimate policy relevant.”

Test 2a – Overall Significance of a Regression

The hypotheses for this test are:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0.$$

$$H_1: \text{at least one coefficient} \neq 0.$$

The null hypothesis might also be expressed as “all coefficients are jointly zero”. This is essentially testing if the R-square is statistically different from zero. The test statistic, an F-Statistic, is given as:

$$F_{k,(n-k-1)} = \frac{\text{ExplainedSS}/k}{\text{UnExplainedSS}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}.$$

This statistic has k and n-k-1 degrees of freedom.

Test 2b – Comparing Constrained and Unconstrained Models (adding a small set of variables to a regression)

Suppose you want to know if a set of additional variables adds explanatory power to a model. In this case, you have two models:

$$\text{constrained: } Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i.$$

$$\text{unconstrained: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i.$$

In other words, does the addition of variables X2 and X3 add explanatory power to the constrained model. To assess this question, consider the following hypotheses:

$$H_0: \beta_2 = \beta_3 = 0$$

$$H_1: \text{at least one} \neq 0.$$

The test statistic is given as:

$$F_{(k_{uc}-k_c),(n-k_{uc}-1)} = \frac{(\text{ExplainedSS}_{uc} - \text{ESS}_c)/(k_{uc} - k_c)}{\text{UnExplainedSS}_{uc}/(n - k_{uc} - 1)} = \frac{(R_{uc}^2 - R_c^2)/(k_{uc} - k_c)}{(1 - R_{uc}^2)/(n - k_{uc} - 1)}.$$

Notice that the degrees of freedom for the numerator is simply the number of added variables.

Rejection of this null hypothesis would suggest that the addition of X2 and X3 to the constrained model adds explanatory power to the model.

Note: when adding only a single variable to a regression, the above F-statistic will be the square of the t-statistic for the test on the significance of the regression coefficient.

Test 3 – Testing the Equality of Two Coefficients

Consider the following regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i.$$

Suppose you wanted to do the following hypothesis test:

$$H_0: \beta_3 = \beta_4.$$

$$H_1: \beta_3 \neq \beta_4.$$

In other words, you want to test the equality of two coefficients. This is particularly helpful if you are testing a program treatment that has more than one effect (or more than one treatment). This is just a t-test, and can be calculated as:

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - 0}{SE(\hat{\beta}_3 - \hat{\beta}_4)}.$$

STATA calculates this nicely for you.
