Multiple regression analysis: Inference, Part II

Prof. Sebastián Vallejo Vera
University of Western Ontario

Confidence intervals

Under the classical linear model assumptions (CLM), we can easily construct a **confidence interval (CI)** for the population parameter β_j.

Using the fact that $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j)$ has a t distribution with $n-k-1$ degrees of freedom, simple manipulation leads to a CI for the unknown β_j: a 95% confidence interval, given by

$$\hat{\beta}_j \pm c \cdot se(\hat{\beta}_j)$$

where the constant c is the 97.5^{th} percentile in t_{n-k-1} distribution (because it comes from a two-tailed test).

Confidence Intervals

More precisely, lower and upper bounds of the confidence interval are given by

$$\underline{\beta}_j \equiv \hat{\beta}_j - c \cdot se(\hat{\beta}_j)$$

$$\bar{\beta}_j \equiv \hat{\beta}_j + c \cdot se(\hat{\beta}_j)$$

Confidence intervals

A CI means that if random samples were obtained over and over again, with $\underline{\beta}_j$ and $\bar{\beta}_j$ computed each time, then the (unknown) population value β_j would lie in the interval $(\underline{\beta}_j, \bar{\beta}_j)$ for 95% of the samples.

Unfortunately, for the single sample that we use to construct the CI, we do not know whether β_j is actually contained in the interval. We hope we have obtained a sample that is one of the 95% of all samples where the interval estimate contains β_j, but we have no guarantee.

Finally, note that once a CI is constructed, it is easy to carry out two-tailed hypotheses tests. If the null hypothesis is $H_0 : \beta_j = a_j$, then H_0 is rejected against $H1 : \beta_j \neq a_j$ at (say) the 5% significance level if, and only if, a_j is *not* in the 95% confidence interval.

Testing hypotheses about a single linear combination of the parameters

Although in most applications we perform hypothesis testing or construct confidence intervals to test hypotheses about a single β_j, at times, we may be interested in testing hypotheses involving more than one of the population parameters. Suppose the following regression equation:

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

where:

- *jc* is the number of years attending a two-year college
- *univ* is the number of years at a four-year college
- *exper* is the number of months in the workforce

Testing hypotheses about a single linear combination of the parameters

The hypothesis of interest is whether one year at a junior college is worth one year at a university:

$$H_0 : \beta_1 = \beta_2$$

Under H_0, another year at a junior college and another year at a university lead to the same *ceteris paribus* percentage increase in wage. For the most part, the alternative of interest is one-sided: a year at a junior college is worth less than a year at a university. This is stated as:

$$H_1 : \beta_1 < \beta_2$$

Testing hypotheses about a single linear combination of the parameters

The difficulty arises from the fact that the hypotheses concern *two* parameters, β_1 and β_2. The two hypotheses can be rewritten as follows:

$$H_0 : \beta_1 - \beta_2 = 0$$
$$H_1 : \beta_1 - \beta_2 < 0$$

and we can proceed as usual using the following t statistic:

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2)}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$

The calculation of $(\hat{\beta}_1 - \hat{\beta}_2)$ is straightforward but that of $se(\hat{\beta}_1 - \hat{\beta}_2)$ is much more complicated. Thankfully, statistical softwares can do these calculations for us.

Testing multiple linear restrictions: The F test

Frequently, we wish to test multiple hypotheses about the underlying parameters β_1, β_2, ..., β_k.

We already know how to test whether a particular variable has no partial effect on the dependent variable: use the t statistic. Now, we want to test whether a *group* of variables has no effect on the dependent variable. More precisely, the null hypothesis is that a set of variables has no effect on y, once another set of variables has been controlled.

This is referred to as *testing exclusion restrictions*.

Testing multiple linear restrictions: The F test

Suppose the following *unrestricted* model with k independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + u \qquad (1)$$

The number of parameters in the unrestricted model is $k + 1$.

Suppose now that we have q exclusion restrictions to test: that is, the null hypothesis states that q of the variables in equation (1) have zero coefficients. For notational simplicity, assume that it is the last q variables in the list of independent variables: x_{k-q+1}, ..., x_k. The null hypothesis is thus stated as:

$$H_0 : \beta_{k-q+1} = 0, ..., \beta_k = 0$$

which puts q exclusion restrictions on the model (1).

Testing multiple linear restrictions: The F test

The alternative is simply that H_0 is false. In other words:

$$H_1 : H_0 \textit{ is not true}$$

This means that at least one of the parameters in H_0 is different from zero (when tested *jointly*).

When we impose the restrictions under H_0, we are left with the *restricted* model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-q} x_{k-q} + u$$

Testing multiple linear restrictions: The F test

The appropriate test to evaluate such exclusion restrictions is the **F statistic** (or *F ratio*) and is defined by:

$$F \equiv \frac{(RSS_r - RSS_{ur})/q}{RSS_{ur}/(n-k-1)}$$

where RSS_r is the sum of squared residuals from the restricted model and RSS_{ur} is the sum of squared residuals from the unrestricted model.

Remember that, because the OLS estimates are chosen to minimize the sum of squared residuals, the *RSS always increases* when variables are dropped from the model.

The question is whether this increase in the *RSS* from the restricted model (RSS_r) is large enough relative to the *RSS* in the unrestricted model (RSS_{ur}) to warrant rejecting the null hypothesis. This is exactly what the *F* test does.

Testing multiple linear restrictions: The F test

To use the F statistic, we must know its sampling distribution under the null in order to choose critical values and rejection rules. It can be shown that, under H_0 (and assuming the CLM assumptions hold), F is distributed as an F random variable with $(q,\ n-k-1)$ degrees of freedom. We write this as:

$$F \sim F_{q,n-k-1}$$

Testing multiple linear restrictions: The F test

We reject H_0 in favor of H_1 when F is sufficiently "large." How large depends on our chosen significance level. At the 5% level test, we have that c is the 95^th percentile in the $F_{q,n-k-1}$ distribution.

Once c has been obtained, we reject H_0 in favor of H_1 at the chosen significance level (here 5%) if

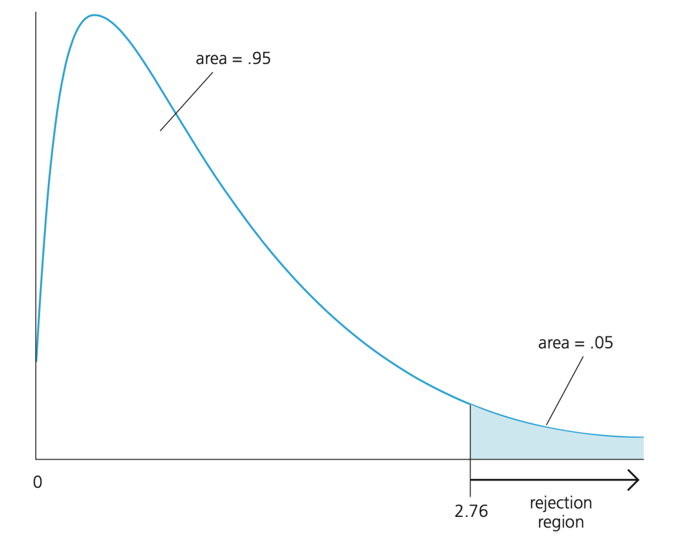
$$F > c$$

If H_0 is rejected, then we say that x_{k-q+1}, ..., x_k are **jointly statistically significant** (or just *jointly significant*) at the appropriate significance level. This test alone does not allow us to say which of the variables has a partial effect on y; they may all affect y or maybe only one affects y.

If the null is not rejected, then the variables are **jointly insignificant**, which may justify dropping them from the model.

Illustration of the F test

FIGURE 4.7 The 5% critical value and rejection region in an $F_{3,60}$ distribution.



The R-Squared form of the F Statistic

For testing exclusion restrictions, one can also use the R^2s from the restricted and unrestricted models to compute the F statistic:

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)}$$

This is called the **R-squared form of the F statistic**.

Note the order of the R^2s in the numerator: the unrestricted R^2 comes first (contrast this with the formula using the RSSs). Because $R^2_{ur} > R^2_r$, this shows again that F will always be positive.

Note also that when estimating the restricted model to compute an F test, we must use the *same* observations to estimate the unrestricted model; otherwise, the test is not valid. This is important because the restricted model can sometimes have a higher number of observations because it relies on a smaller number of independent variables, some of which presumably have missing values.

Computing p-values for F test

In the F testing context, the p-value is defined as

$$p - value = P(\mathcal{F} > F)$$

where we let $\mathcal{F}$ denote an F random variable with $(q, n-k-1)$ degrees of freedom, and F is the actual value of the test statistic.

The p-value still has the same interpretation as it did for t statistics: it is the probability of observing a value of F at least as large as we did, given that the null hypothesis is true. A small p-value is evidence against H_0.

The F statistic for overall significance of a regression

A special set of exclusion restrictions is routinely tested by most regression packages. These restrictions have the same interpretation, regardless of the model. In the model with k independent variables, we can write the null hypothesis as

$$H_0 : x_1, x_2, ..., x_k \textit{ do not help explain } y$$

This null hypothesis is, in a way, very pessimistic. It states that *none* of the explanatory variables has an effect on y. Stated in terms of the parameters, the null is that all slope parameters are zero:

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

and the alternative is that at least one of the β_j is different from zero.

The F statistic for overall significance of a regression

In this case, the unrestricted model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + u$$

and the restricted model is simply

$$y = \beta_0 + u \tag{2}$$

The R^2 from estimating equation (2) is zero; none of the variation in y is being explained because there are no explanatory variables. Therefore, the F statistic for testing H_0 can be written as:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

where R^2 is from the unrestricted model of regressing y on x_1, x_2, ..., x_k.

The F statistic for overall significance of a regression

Testing:

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

is sometimes called determining the **overall significance of the regression**.

If we fail to reject H_0, then there is no evidence that any of the independent variables help to explain y. This usually means that we must look for other variables to explain y.

In-class exercise, part 1

Using the 'fertil2' dataset from 'wooldridge' on women living in the Republic of Botswana in 1988, estimate the regression equation of the number of children on education (*educ*), age of the mother (*age*) and its square, electricity (*electric*), husband's education (*heduc*), and whether the women has a radio (*radio*) and/or a TV (*tv*) at home. Construct a 90% and 95% confidence interval for *electric*. Interpret.

In-class exercise, part 2

Evaluate the following hypotheses at the 5% and 1% levels of significance. Interpret.

1. $H_0 : \beta_{educ} = \beta_{heduc}$
   $H_1 : \beta_{educ} < \beta_{heduc}$

2. $H_0 : \beta_{radio} = 0, \beta_{tv} = 0$
   $H_1 : H_0 \ is \ not \ true$

3. $H_0 : \beta_{educ} = \beta_{age} = \beta_{electric} = \beta_{heduc} = \beta_{radio} = \beta_{tv} = 0$
   $H_1 : H_0 \ is \ not \ true$