ØAMET4100 · Spring 2019
Lecture Note 3B

Instructor: Fenella Carpena

January 24, 2019

1 Hypothesis Testing in Multiple Regression Model

1.1 Overview

When testing hypotheses about the coefficients in a regression model with more than one right hand side variable, there are three different types of tests that you might encounter: test of a single coefficient, joint test of coefficients, and linear restrictions. For these types of tests, the 2 main tools we use are the t-statistic and F-statistic. The table below summarizes the types of tests and the test statistics.

Type	Example	Test Statistic
1. Single One restriction involving one parameter	$H_0 : \beta_1 = 4$ $H_1 : \beta_1 \neq 4$	t-stat
2. Joint Multiple restrictions	$H_0 : \beta_1 = 0, \beta_2 = 0$ $H_1 : \beta_1 \neq 0$ and/or $\beta_2 \neq 0$	F-stat **(cannot use t-stat)**
3. Linear Linear combination of coefficients	$H_0 : 4\beta_1 + 2\beta_2 = 5$ $H_1 : 4\beta_1 + 2\beta_2 \neq 5$ or $H_0 : \beta_1 - \beta_2 = 0$ $H_1 : \beta_1 - \beta_2 \neq 0$	(1) F-stat, (2) t-stat (by first transforming the regression

As you can see from the above table, we typically use the t-stat when there is only 1 equation ("restriction") in our null hypothesis, and we typically use the F-stat when there is more than 1.

1.2 Example: Using t-stat and F-stat in practice

Perhaps the easiest way to understand hypothesis testing is through an example, this lecture note will go through Exercises 1.1 to 1.6 in Worksheet 3B. The exercise is as follows.

Suppose we have the following model that explains baseball players' salaries.

$$salary_i = \beta_0 + \beta_1 years_i + \beta_2 gamesyr_i + \beta_3 bavg_i + \beta_4 hrunsyr_i + \beta_5 rbisyr_i + u_i \tag{1}$$

where for each player i, *salary* is the salary in 1993, *years* is years in the league, *gamesyr* is average games played per year, *bavg* is the career batting average, *hrunsyr* is the number of home runs per year, *rbisyr* is runs batted in per year. Further, suppose that we estimated the above equation using data we have on hand, and that we obtained the following regression results

$$\widehat{salary} = \underset{(0.29)}{11.10} + \underset{(0.0121)}{0.0689} \cdot years + \underset{(0.0026)}{0.0126} \cdot gamesyr + \underset{(0.0010)}{0.00098} \cdot bavg + \underset{(0.0161)}{0.0144} \cdot hrunsyr + \underset{(0.0072)}{0.0108} \cdot rbisyr$$

$$N = 353,\ SSR = 183.186,\ R^2 = 0.6278$$

How would we test the hypothesis that $H_0 : \beta_3 = 0, H_1 : \beta_3 \neq 0$ at the 5% level? We would calculate the t-statistic. t-stat $= (\hat{\beta}_3 - 0)/SE(\hat{\beta}_3) = 0.00098/0.0010 \approx 0.98 < 1.96$; we fail to reject the

null hypothesis.

How would we test the hypothesis that $H_0 : \beta_4 = 0, H_1 : \beta_4 \neq 0$ at the 5% level? We would calculate the t-statistic. t-stat $= (\hat{\beta}_4 - 0)/SE(\hat{\beta}_4) = 0.0144/0.0161 \approx 0.894 < 1.96$; we fail to reject the null hypothesis.

How would we test the hypothesis that $H_0 : \beta_5 = 0, H_1 : \beta_5 \neq 0$ at the 5% level? We would calculate the t-statistic. t-stat $= (\hat{\beta}_5 - 0)/SE(\hat{\beta}_5) = 0.0108/0.0072 = 1.5 < 1.96$; we fail to reject the null hypothesis.

How would we test the hypothesis that once years in the league and games per year have been controlled for, the variable *bavg*, *hrunsyr*, and *rbisyr* (which we can think of as measure of performance) have no effect on salary? (Assume we are carrying out this test at 5% significance level.)

In mathematical terms, this null hypothesis is expressed as

$$H_0 : \beta_3 = 0,\ \beta_4 = 0,\ \beta_5 = 0 \ vs.\ H_1 : \text{at least one of } \beta_3, \beta_4, \beta_5 \text{ is not equal to } 0 \tag{2}$$

A common pitfall is to think that we can test the above hypothesis by constructing a t-stat for $\beta_3, \beta_4, \beta_5$, and then using each t-stat, test whether we can reject the null that the coefficient is zero. This is **not appropriate**, because we want to look at all 3 coefficients simultaneously. Therefore, what we need is the **joint distribution** of $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$. If we looked at each of these coefficient one at a time by looking at the t-statistic, we will not be putting any restriction on the other parameters. Note that earlier, we looked at the t-stat for the separate (individual) hypothesis tests of $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$, and we saw that each t-stat that is less than 1.96 (so we failed to reject the null hypothesis in all cases). This might lead you to conclude that we should reject H_0 indicated in (2) above. But as we will see later, this conclusion turns out to be wrong.

To carry out the hypothesis test, we would need to look at $\beta_3, \beta_4, \beta_5$ jointly. For this, we will need to use the F-stat. **If the population regression errors are assumed to be homoskedastic,** we can use the following formula for the F-stat

$$F - stat = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}.$$

Let's break down the different parts of this formula.

- SSR_r is the sum of squared residuals (SSR) of the *restricted* regression (this is why I have the subscript r). The restricted regression is the regression that imposes the null hypothesis. In this example, it means exclude the 3 parameters that are assumed to be zero in the hypothesis test from the regression. In other words, the *restricted* regression is

$$salary_i = \beta_0 + \beta_1 years_i + \beta_2 gamesyr_i + u_i$$

- SSR_{ur} is the SSR of the *unrestricted* regression, i.e., the original regression model that we have which includes all parameters. That is, the *unrestricted* regression is the same as what we started with, equation (1) above.

- q is the number of restrictions in our hypothesis, indicated in (2). Here, $q = 3$.

- n is the sample size and k is the number of regressors in the unrestricted regression. So $n-k-1 = 353-5-1$.

- What is the intuition behind the F-stat? Looking at the formula, we see that the F-stat looks pretty close to $(SSR_r - SSR_{ur})/SSR_{ur}$ (i.e., if we remove q and $n-k-1$ in the formula). Note that $(SSR_r - SSR_{ur})/SSR_{ur}$ can be interpreted as the proportional change in the SSR when we move from the unrestricted to the restricted model. How does this relate to the hypothesis test? Recall that SSR is a measure of fit (with lower SSR corresponding to a better fit). If the restricted model (which assumes that H_0 is true) results in significantly higher residuals, we would get a high F-stat. This also means that the restricted model has a worse fit on the data relative to the unrestricted model. Thus, we have evidence that casts doubt of H_0 being true and we reject H_0.

- Other notes about the F-stat to help you understand it: (1) Since SSR_r is always greater than SSR_{ur} (why?), the F-stat is always positive. So if you are calculating the F-stat and you get a negative number, you're doing something incorrectly; (2) Also note that q is the degrees of freedom in the restricted model minus the degrees of freedom in the unrestricted model, i.e. $q = df_r - df_{ur} = (n - k - 1 + q) - (n - k - 1)$; (4) Finally, $n - k - 1$ is the degrees of freedom in the unrestricted model, i.e. df_{ur}.

Let's now calculate the F-stat. Suppose we are given that the SSR in the restricted model is 198.311. Then, the F-stat will be given by

$$F - stat = \frac{198.311 - 183.186}{183.186} \cdot \frac{353 - 5 - 1}{3} \approx 9.55.$$

The next step will be to compare the F-stat to the appropriate critical value. Generally, for large N, the F-stat has an approximate distribution of $F_{q,\infty}$, where in this case $q = 3$. The critical value for a test at 10% significance is approximately 2.08. Since $9.55 > 2.08$, we reject the null hypothesis.

Finally, an important note: the formula we've used above for the F-stat is valid only under the assumption of homoskedasticity. Under heteroskedastic errors, we should use the heteroskedasticity-robust F-statistic (but deriving the formula for this F-stat is beyond the scope of this course).

We just rejected the joint hypothesis that $bavg$, $hrunsyr$, $rbisyr$ have no effect on salary. But if we had looked at each of these variables individually, we would have failed to reject each null hypothesis separately, because the individual t-stats are less than 1.96. What might explain the difference in these results?

In this example, the reason why the individual t-stats are low is because $hrunsyr$ and $rbsyr$ are highly correlated. Imperfect multicollinearity makes it difficult to estimate their coefficients precisely (why?) resulting in a low t-stat. Since the F-stat tests whether $bavg$, $hrunsyr$ and $rbsyr$ are jointly different from zero, the high correlation between $hrunsyr$ and $rbsyr$ does not have any role. Generally speaking, F-stats are useful for testing the significance of a group of variables when many of the variables in the group are highly correlated. For example, suppose we want to test whether firm performance affects CEO salary. Since there are many ways to measure firm performance, we might have multiple measures of firm performance that are highly correlated. The F-stat will allow us to test whether the measures of firm performance, taken as a group, has any effect on CEO salary.

1.3 Relationship between t-stat and F-stat

We generally use the t-stat when the hypothesis has one restriction, and we normally use the F-stat when the hypothesis has multiple restrictions. However, it is also possible to use the F-stat when there is only one restriction (so that $q = 1$), for example, to test the two-sided hypothesis that $\beta_1 = 0$. In this special case, the F-stat = (t-stat)^2, and both the F-stat and t-stat will lead to the same conclusion.

2 Control Variables and Conditional Mean Independence

Again, it's perhaps easier to understand these concepts using an example, so let's look at the following. Suppose that we want to understand the effect of cigarette smoking during pregnancy on birthweight. Consider the model

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 mothereduc_i + \beta_4 fathereduc_i + u_i$$

where $bwght$ is birth weigh of person i, $cigs$ is average number of cigarettes smoked per day during pregnancy, $mothereduc$ is mother's education, and $fathereduc$ is father's education.

Since our objective is to understand how cigarette smoking affects birthweight, $cigs$ is our **variable of interest** and β_1 is our **coefficient of interest**. $faminc$, $mothreduc$, $fathereduc$ are called **control variables** that we include in the study because of potential omitted variable bias.

Recall that the first least squares assumption in the multiple regression model is $E[u_i|X_{1i}, ..., X_{ki}] = 0$. This assumption gives us unbiasedness of $\beta_0, \beta_1, ..., \beta_k$. However, if we are only interested in one of the β's (for example, we only care about the coefficient on *cigs*, β_1, in the above model), then we can relax the first least square assumption, and instead require **conditional mean independence.**

In the birthweight and cigarette smoking set-up, if

$$E[u_i|cigs_i, faminc_i, mothereduc_i, fathereduc_i] = E[u_i|faminc_i, mothereduc_i, fathereduc_i],$$

then our OLS estimate of the coefficient on $cigs_i$ (that is, $\hat{\beta}_1$) will be unbiased. This conditional mean independence assumption is saying that if we control for $faminc$, $mothereduc$, and $fathereduc$, then the error term u and the variable *cigs* are uncorrelated. Another way of thinking about this is that holding $faminc$, $mothereduc$, and $fathereduc$ is as good as randomly assigned, so that $\hat{\beta}_1$ gives us a causal interpretation of the effect of cigarette smoking on birthweight. Furthermore, note that even though the conditional mean independence assumption tells us that $\hat{\beta}_1$ will be unbiased, it does not say anything about whether $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$ will also be unbiased.

Finally, please read page Section 7.5 "Model Specification for Multiple Regression" of the textbook. I think these concepts are very important for this class, especially to understand conditional mean independence and how to interpret R^2, $\bar{R}^2$.