Multiple Hypothesis Testing: The F-test[*]

Matt Blackwell

December 3, 2008

1 A bit of review

When moving into the matrix version of linear regression, it is easy to lose sight of the big picture and get lost in the details of dot products and such. It is vital to take a step back and figure out where we are and what we are doing in order to keep ourselves grounded and understanding the material.

We start with a **population**, consisting of units (countries, registered voters, counties, &c). We obtain a sample from this population, which is our **data**. We want to learn something about the population from this sample. We call these parameters, or quantities of interest. In the beginning of class we were trying to find, say, the percent of registered voters who voted in Fulton County (our parameter of the population). We put our data into an **estimator** (the sample mean) and get out an **estimate** (.42). We can then use hypothesis tests and confidence intervals in deal with the uncertainty inherent in the sampling process.

Hypothesis testing has us ask this: if we suppose some null hypothesis is true, how likely is it that we would have obtained this result from random sampling? We reject the null hypothesis if we determine our estimate is unlikely (the probability is less than α, a small number) given the null. Confidence intervals collect all of the null hypotheses that we *cannot* reject at some level; that is, these are the values of the true parameters we think could have plausibly generated our observed data.

We said that we want to find out likely our data is under some hypothesis. But, you may ask, how do we know how likely our data is under some hypothesis? For example, we know that the sample mean, $\bar{X}$ tends to be Normally distributed around the true mean μ with standard error $\sigma/\sqrt{n}$. But don't actually know μ so we don't actually know this distribution. Promisingly, we do know the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

We know that this had a standard Normal distribution. Thus, we could calculate the Z for some proposed value of μ and see how likely that Z would be in the standard Normal. This is an example of a test statistic.

[*] For gov2k in Fall 2008. Parts are heavily borrowed (read: stolen) from past gov2k TFs, specifically Jens Hainmueller, Ryan Moore and Alison Post.

2 The F-test

We have seen our t-statistic follows a t distribution with a "degrees of freedom" parameter. This fact has been useful for hypothesis testing, both of sample means and of regression coefficients. We are able to test, say, the hypothesis that some variable has no effect on the dependent variable. All we do is calculate a t-statistic for this null hypothesis and our data and see if that test statistic is unlikely under the null distribution (the Student's t-distribution).

Unfortunately, when we have more complicated hypotheses, this test no longer works. Hypotheses involving multiple regression coefficients require a different test statistic and a different null distribution. We call the test statistics F_0 and its null distribution the F-distribution, after R.A. Fisher (we call the whole test an F-test, similar to the t-test). Again, there is no reason to be scared of this new test or distribution. We are still just calculating a test statistic to see if some hypothesis could have plausibly generated our data.

2.1 Usage of the F-test

We use the F-test to evaluate hypotheses that involved multiple parameters. Let's use a simple setup:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$$

2.1.1 Test of joint significance

Suppose we wanted to test the null hypothesis that all of the slopes are zero. That is, our null hypothesis would be

$$\begin{aligned} H_0 : & \beta_1 = 0 \text{and} \\ & \beta_2 = 0 \text{and} \\ & \beta_3 = 0. \end{aligned}$$

We often write this more compactly as $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Note that this implies the following alternative hypothesis:

$$\begin{aligned} H_1 : & \beta_1 \neq 0 \text{or} \\ & \beta_2 \neq 0 \text{or} \\ & \beta_3 \neq 0. \end{aligned}$$

This is a test of the null that none of the independent variables have predictive power. We could use another null such as $H_0 : \beta_1 = \beta_3 = 0$ to see if either X_1 or X_3 has predictive power, when controlling for X_2.

These are often substantively interesting hypotheses. For example, if we wanted to know how economic policy affects economic growth, we may include several policy instruments (balanced budgets, inflation, trade-openness, &c) and see if all of those policies are jointly significant. After all, our theories rarely tell us *which* variable is important, but rather a broad category of variables.

In addition, we may have a series of dummy variables that all measure some qualitative grouping. Suppose in the Fulton county data we had a dummy variable for each religion:

	Voted	Catholic	Protestant	Jewish	Other
1	1	0	1	0	0
2	1	0	0	0	1
3	0	1	0	0	0
4	1	0	1	0	0
5	0	0	0	1	0

We could run a regression with each dummy variable to see the rate at which each group votes (if this is confusing, take a look back at the lecture on dummy variables). The coefficients will always be in comparison to the omitted category, which may not be a useful test. It is usually more useful to test if there is any difference between any of the groups. We can do that with a null hypothesis that all of the religion coefficients are equal to zero.

We could also use these restrictions to test interaction or quadratic terms, as these will only have no effect at all when both coefficients are all of their constituent coefficients are equal to zero.

Note that we could replace 0 with any other number in our null hypothesis. Our theories often are not specific enough to test some other null, but it does arise. With logged dependent variables, authors sometimes test the null that the coefficients are 1 (since the effect on the unlogged variable would be 0).

2.1.2 Tests of linear restrictions

The joint significance tests of the previous section are important, but not the full extent of the F-test. We can test general linear restrictions. For instance, we may want to test if two coefficients are significantly different from one another. This null would be $H_0 : \beta_2 - \beta_1 = 0$ or, equivalently, $H_0 : \beta_2 = \beta_1$. Since we have shown that the scale of the independent variable affects the size of the coefficient, it is important to note that the independent variables for these coefficients should be on the same scale. For instance, you would not want to test the null that the effect of years of education on income equals the effect of gender as they are on completely different scales. You may want to test the difference between the effect of years of education and the effect of years of experience, though. Those are on the same scale and the test has substantive interest.

It is possible to have even more complicated linear restrictions, such as

$$\begin{aligned} H_0 : \beta_3 - 7 &= 3 \times \beta_2 \\ 3 \times \beta_2 &= \beta_1 - \beta_4. \end{aligned}$$

Again, we would usually write this as $H_0 : \beta_3 - 7 = 3 \times \beta_2 = \beta_1 - \beta_4$. These types of restrictions are obviously less common as our theories rarely give us such sharp predictions about our coefficients. These types of restrictions might be useful if we need to rescale some of the coefficients to make them comparable.

2.2 Calculating the F-statistic

We showed what kinds of hypotheses we can test with the F-test in the previous section, but now we need to actually calculate the test statistic. The motivation is that we want to know the distribution of the test statistic under the null hypotheses. Earlier we noted that $\frac{\hat{\beta}-\beta_{null}}{\hat{\sigma}/\sqrt{n}}$ follows a t-distribution under the null that true mean of $\hat{\beta}$ is β_{null}. This is the core of the t-test.

For the more complicated null hypotheses in the previous sections, we will calculate F_0, which will follow an F distribution under those nulls. We will deal with the simpler joint significance tests first, then move on to the more general linear restrictions.

2.2.1 F_0 for joint significance tests

If our null is of the form, $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, then we can write the test statistic in the following way:

$$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-(k+1))},$$

where SSR_r stands for the sum of the suared residuals of the restricted model and SSR_{ur} is the same for the unrestricted model. We also have that n is the number of observations, k is the number of independent variables in the unrestricted model and q is the number of restrictions (or the number of coefficients being jointly tested).

This terminology may seem a bit strange at first. We are "restricting" the general model by imposing supposing that the null is true and removing variables from the model. Thus, the difference $(SSR_r - SSR_{ur})$ is telling us how much bigger the residuals are in the model where the null hypothesis is true. If the residuals are a lot bigger in the restricted model, then F_0 will also be big. When the residuals are bigger, we know that this means the fit of the regression is worse. Thus, F_0 is big when the restriction makes the fit of the regression a lot worse which is exactly when we would question the null hypothesis. If these variables really had no predictive power, then removing them should not affect the residuals. We will discuss how big F_0 needs to be to reject the null hypothesis a bit later.

2.2.2 F_0 for general linear restrictions

The general linear restrictions we wrote about can all be written in the following matrix form:

$$H_0 : \mathbf{L}\beta = \mathbf{c}$$

where we can form the matrices $\mathbf{L}$ and $\mathbf{c}$ to fit our hypothesis. Adam covered many examples of these in lecture, so I won't repeat them here. You also get practice of this in the homework. With this null hypothesis, we can write the test statistic as

$$F_0 = \frac{(\mathbf{L}\hat{\beta} - \mathbf{c})'[\hat{\sigma}^2\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{L}\hat{\beta} - \mathbf{c})}{q}$$

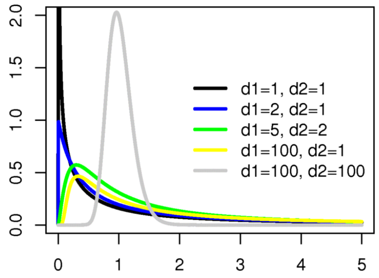
where q is the number of restrictions (the rows of $\mathbf{L}$ and $\mathbf{c}$). It seems like this obtuse piece of junk would be very hard to get intuition about and that is correct, but we can try. Note that $(\mathbf{L}\hat{\beta} - \mathbf{c})$ measure how different our observed coefficients differ from the hypothesis. If the null hypothesis were true, then this discrepancy would be 0 and our F_0 statistic would also be 0. Any deviation from 0 would be due to random chance or sampling. So, $(\mathbf{L}\hat{\beta} - \mathbf{c})'$ and $(\mathbf{L}\hat{\beta} - \mathbf{c})$ are squaring the deviations from the hypothesized value. The middle part $[\hat{\sigma}^2\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}$ is normalizing those deviations to be on a constant unit scale. You can see this by noting that it is simply the variance of the deviations:

$$\text{var}[\mathbf{L}\hat{\beta} - \mathbf{c}] = \mathbf{L}\text{var}[\hat{\beta}]\mathbf{L}' = \hat{\sigma}^2\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$$

Thus, F_0 will be big when the deviations from the hypothesis are big compared to what we expect the deviations to look like. This makes sense as this is exactly the times when we think the hypothesis is not plausible.

2.3 The null distribution: $\mathcal{F}$

Now that we have calculated the test statistic, F_0, we need to see if it is bigger than we would expect by random chance. To do this, we must know its distribution under the null hypothesis. For reasons that you need to take more classes to learn, we know that F_0 will be distributed $\mathcal{F}$ with degrees of freedom q and $n - (k + 1)$. The $\mathcal{F}$-distribution has two parameters, both called the degrees of freedom (usually they are called df1 and df2, or numerator df and denominator df). We can see how it changes over different values of each parameter:



For our purposes, the first d.f. will be q, the number of restrictions and the second will be $n - (k + 1)$, which is the number of observations minus the number of columns in the model matrix. Unless you would like to get deep into probability theory, there's no real need to know why this is true, think of it being similar to how we compute the t-distribution. Thus, if we 2 restrictions, 100 observations and 5 independent variables, we would know that F_0 is distributed $\mathcal{F}_{2,100-(5+1)}$ or $\mathcal{F}_{2,94}$ under the null hypothesis. If we found that $F_0 = 4.52$ we could see how unlikely that would be under the null by using `pf()` in R:

```
pf(x=4.51, df1=2, df2=94)
```

3 Confidence Ellipses

One way to think of confidence intervals are the set of parameter values, x, for which we cannot reject the null hypothesis $H_0 : \beta_j = x$. In this sense, we are "inverting" the test to find the confidence intervals. We would reject any null hypothesis outside of the interval. Again, confidence intervals represent all the plausible null hypotheses. We would like to get a similar interval for these more complicated null hypotheses. One way to do this is to find the set of parameter values such that we cannot reject them as the null. Let's write this out explicitly. First, let $f^{\alpha}_{q,n-(k+1)}$ be the critical value for the $\mathcal{F}_{q,n-(k+1)}$ at the α level (remember this is like 1.96 for the Normal at $\alpha = .05$). We can write:

$$\begin{aligned}\Pr[F_0 \leq f^{\alpha}_{q,n-(k+1)}] &= 1-\alpha \\ &= \Pr[\frac{(\mathbf{L}\hat{\beta}-\mathbf{c})'[\hat{\sigma}^2\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{L}\hat{\beta}-\mathbf{c})}{q} \leq f^{\alpha}_{q,n-(k+1)}] \\ &= \Pr[(\mathbf{L}\hat{\beta}-\mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{L}\hat{\beta}-\mathbf{c}) \leq \hat{\sigma}^2 q f^{\alpha}_{q,n-(k+1)}]\end{aligned}$$

which means that we can define a $1-\alpha$ confidence region by the following inequality:

$$(\mathbf{L}\hat{\beta}-\mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{L}\hat{\beta}-\mathbf{c}) \leq \hat{\sigma}^2 q f^{\alpha}_{q,n-(k+1)}.$$

This inequality defines an ellipse which contains all of the sets of parameters for which we cannot reject the null hypothesis. For more information on this, you should see Adam's handout on confidence regions.

4 Multiple Testing

The problem of multiple testing basically boil down to this: you are much more likely to falsely reject the null hypothesis (that is, get false significance) when you do multiple tests at once. As usual, it is easiest to see using a coin flipping example (due to Wikipedia).

Say that we have a coin and we want to determine if it is fair. We flip it 10 times and it comes out heads 9 times. Under the null hypothesis of a fair coin, this has probability $11 \times (.5)^{10} \approx 0.01$ of happening. This is relatively unlikely and we would reject this null hypothesis at $\alpha = 0.05$.

Now, suppose instead that I have 100 fair coins (a fact that I do not know) and I want to flip them all 10 times to see if they are fair. It is true that if I pick a particular (that is, pre-selected) coin, it has probability 0.01 of having 9 out of ten 9 heads. It is clear, though, that there is a good (greater than .5) chance that *some* coin in the 100 will get 9 heads. In fact, the probability of this happening it $1-(1-0.01)^{100} \approx 0.66$. Thus, we have a *very* good chance of finding one "unfair" coin if we are searching and not pre-selecting.

This same logic applies to hypothesis tests of significance in regressions and confidence intervals. If we choose a significance level for each comparison, α, then the significance level for looking at each test separately is $1-(1-\alpha)^q$ where q is the number of comparisons (this is when we compare independent tests, which is often not true of regression tests). For instance, let's say that our confidence level is 95%, so $\alpha = .05$. Then if we were testing 2 variables, then our actual confidence would be $.95^2 = .903$, so we would find a false rejection about 10% of the time. For 3 comparisons, it would be $.95^3 = .857$, so we would find a false rejection about 15% of the time. These are both clearly lower than the 5% we originally started with.

There are a couple of ways to get around this problem, none of which are incredibly awesome. Note that the F test cannot handle this problem. An F-test will allow us to test the null that all of the coefficients are equal zero; that is, are these variables jointly significance? What it will not tell us is *which* individual variables are significant.

4.1 Bonferroni's Correction

If we say that α_g is the Type I error rate (probability of a false significance) for the multiple tests and α is the same rate for each individual test, then Bonferroni's inequality tells us that

$$\alpha_g \leq \alpha \times q$$

where q is the number of tests. So if we wanted to assure ourselves that the multiple tests had at least some significance level, we could use more stringent tests for the each individual test. Suppose we want make sure $\alpha_g \leq 0.05$, then we could simply test each of the q variables with a $\frac{0.05}{q}$ level. Of course, this means that if we had 10 variables in our regression we would have look for significance levels of $0.05/10 = 0.005$, which would make a lot of results go away.

Note the $\leq$ in the above work. The Bonferroni correction will guarantee that the multiple tests will have the correct, but there may be a lower individual test level that will get the same level for the group of tests. It is often true that this correction will be overly conservative.

The Scheffé method described in lecture and Adam's notes is another way to get the correct intervals.