Econometrics

Chapter 7: Hypothesis Tests and CI's in Multiple Regression

Hypothesis Tests and CI's for a Single Coefficient

Everything is the same as we did when we were in the single regressor case - just the formula for the SE is different, but software will give it to us.

Other than that the concepts, meanings, and steps to do hypothesis testing and CI's for a single coefficient are just as we did before.

Test of Joint Hypothesis

Now that we have more than one coefficient, perhaps we want to test something about two or more coefficients at once - joint hypothesis.

To do this we will use a new statistic - the F-statistic.

For example maybe we want to test whether *both* β_1 *and* β_2 are zero.

$H_0 : \beta_1 = 0; \beta_2 = 0$
$H_1 : \beta_1 \neq 0 - and/or - \beta_2 \neq 0$

In general the null is a list of q joint restrictions on the coefficients and the alternative is that at least one restriction is false.

Why not do one at a time?

Lets think about what the test is doing and what our rejection criterion means.

It means we set the test so that we reject the null when it is true only 5% of the time (and so fail to reject a true null 95%).

Now recall the two t's (if we were to test each separately just using our previous t tests) are jointly normal by CLT and also both marginally normally distributed.

So we reject the null that both are zero effects if either of the t-stats is greater than 1.96.

Lets look at simplest example when the t-stats are uncorrelated.

What is the chance that we *fail to reject* both of the nulls when joint null is true (they both really are zero effects)?

Well since they are independent, its 95% x 95% = 90.25%.

So we reject at least one individually, and so also the joint, almost 10% of the time when the joint is true.

And if they are correlated this becomes even more complicated - and obviously wrong. So lets do something different.

The F-Statistic

With q restrictions the F is distributed $F_{q,\infty}$ in large samples. Find in Table 4, so big F rejects the null that all q restrictions are true

We will not worry about the general equation - software will give it to you.

"Overall F" is the test that all slopes are zero, this is reported normally by software (beware homo vs. heterskedastic)

The homoskedstic only F-Statistic

Now, this may not be valid, but since the general heteroskedastic F-stat is too complicated, maybe this simple one can give some intuition.

First, note you have two competing regressions - an unrestricted (all βs) and a restricted one (with some of the βs set to zero).

Basically what we are saying is that if the 'fit' is improved enough (in terms of sum of squared residuals) then we will reject the null (that all the new regressors are zero):

(Homoskedastic only) $F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{unrestricted} - 1)}$

also:

(Homoskedastic only) $F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$

Testing Single Restriction Involving Multiple Coefficients

Maybe want to test something like: $H_0 : \beta_1 = \beta_2$ vs. not equal

One way is some software have built in F tests to do this.

We will see just how to do this in R, it is easy and very flexible

Model Specification for Multiple Regression

What to include in the model?

Need knowledge of the problem - 'institutional knowledge' - and focus on omitted variable bias.

Do not just rely on R^2 statistics.

Also beware of 'model searching' and 'model uncertainty'.

Technically, if you run a regression, and then see one variable is not significant and so drop it. Then run a new regression without it, the SEs on your coefficients will not be correct. Because it does not take into account the 'specification testing' you just did.

There are ways to take this into account - but its complicated.

Most would say pick a model based on economic theory and just stick with it.

Omitted Variable Bias

Recall this is usually a big concern. If an omitted variable is a determinant of Y, and is correlated with one of the regressors.

Control Variables

Control variables are not the thing of interest.

Generally we have a question: effect of class size on test scores, but if we ignore other things will have OVB. So we add the other control variables in order to estimate our effect of interest.

Maybe we are worried about omitted variables like private tutoring or other outside learning opportunities likely correlated with class size (because of the income relation) that affect test scores even after including % English speakers. Think about the timing - is this an issue?

But we don't have measures for these. But maybe we can get somewhere if we have another variable that can 'control' for these learning opportunities and make our class size variable 'as if' randomly assigned (with regards to error term).

For example: percent of students who get reduced or free lunch. This is correlated with income effect of the area and so is likely correlated with these outside learning opportunities. And if we think this is a sufficient 'control' we can get somewhere with a slight modification of our Assumption #1.

Replace $E[u|X] = 0$ with $E[u|X_1, X_2] = E[u|X_2]$.

So our error term is mean independent of our variable of interest (class size) once we *control* for these other things.

In this case we can interpret this coefficient as a causal relationship. But that does not mean we can interpret the coefficient on 'free lunch' as causal. The whole reason of including this is it is related to other things not measured in u that effect Y.

Note that this 'control' variable is essentially endogengous - thats why we cannot put a causal meaning to its coefficient. But it helps us put a causal interpretation to our variable of interest

Importantly, again we need to worry about timing. Here '% free lunch' again is something that is determined before class size.

Interpreting R stats

High R's do not mean that your regressors are 'causing' your Y!

They do not tell you if you have omitted variable bias!

They do not tell you you have the best set of regressor in your model!

All it tells you is that your regressors are good at predicting (within sample) your Y's

Specifications/Units?

As noted above, start with a model based on theory. Then maybe as a 'robustness' check, try some plausible alternative specifications. If your estimates change a lot then you maybe have a problem

Units? Just choose so easy to read. If money and $\beta = 0.00000045$ maybe convert to millions for example.

Presentation of Results:

Results are normally presented as follows, with different regressions in different columns, the independent variables in the rows, and SEs in parenthesis under estimated coefficients, and with stars indicating p-values. You should have a descriptive title and any needed notes under the table.

Table 1: Regression results: Dependent Variable is Children's Economic Status

	Percentile Ranking		Log Earnings	
	Base	Full	Base	Full
Parental economic status	0.2920***	0.1756***	0.3540***	0.2286***
	(0.0241)	(0.0247)	(0.0308)	(0.0320)
Cohort * Parental	-0.0276	0.0020	-0.0615	-0.0325
economic status	(0.0364)	(0.0375)	(0.0439)	(0.0459)
AFQT score		0.0659***		0.1744***
		(0.0081)		(0.0243)
Cohort * AFQT score		-0.0322***		-0.0397
		(0.0120)		(0.0359)

Note: Dependent variable is child's economic status (percentile ranking or log earnings). Robust standard errors for OLS estimates are in parenthesis. All regressions control for cohort, race, sex, parental age, parental age squared, age, age squared, and interactions between child age variables and cohort. Significance levels are denoted: *** for 1%, ** for 5% and * for 10%.