Hypothesis Testing in the Multiple regression model

- Testing that individual coefficients take a specific value such as zero or some other value is done in exactly the same way as with the simple two variable regression model.

- Now suppose we wish to test that a number of coefficients or combinations of coefficients take some particular value.

- In this case we will use the so called “F-test”

- Suppose for example we estimate a model of the form

$$Y_i = a + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + b_5 X_{i5} + u_i$$

- We may wish to test hypotheses of the form {H0: b_1=0 and b2=0 against the alternative that one or more are wrong} or {H0: b_1=1 and b_2-b_3=0 against the alternative that one or more are wrong} or {H0: b_1+b_2=1 and a=0 against the alternative that one or more are wrong}
- This lecture is inference in this more general set up.
- We will not outline the underlying statistical theory for this. We will just describe the testing procedure.

Definitions

- **The Unrestricted Model:** This is the model without any of the restrictions imposed. It contains all the variables exactly as in the regression of the previous page

- **The Restricted Model:** This is the model on which the restrictions have been imposed. For example all regressors whose coefficients have been set to zero are excluded and any other restriction has been imposed.

Example 1

- Suppose we want to test that :**H0: b_1=0 and b2=0** against the alternative that one or more are wrong in:

$$Y_i = a + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + b_5 X_{i5} + u_i$$

- The above is the **unrestricted model**

- The **Restricted Model would be**

$$Y_i = a + b_3 X_{i3} + b_4 X_{i4} + b_5 X_{i5} + u_i$$

Example 2

- Suppose we want to test that : **H_0: b_1=1 and b_2-b_3=0** against the alternative that one or more are wrong :

$$Y_i = a + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + b_5 X_{i5} + u_i$$

- The above is the **unrestricted model**
- The **Restricted Model would be**

$$Y_i = a + X_{i1} + b_2 X_{i2} - b_2 X_{i3} + b_4 X_{i4} + b_5 X_{i5} + u_i$$

- Rearranging we get a model that uses new variables as functions of the old ones:

$$(Y_i - X_{i1}) = a + b_2 (X_{i2} - X_{i3}) + b_4 X_{i4} + b_5 X_{i5} + u_i$$

- Inference will be based on comparing the fit of the restricted and unrestricted regression.

- The unrestricted regression will **always fit at least as well as the restricted one.** The proof is simple: When estimating the model we minimise the residual sum of squares. In the unrestricted model we can always choose the combination of coefficients that the restricted model chooses. Hence the restricted model can never do better than the unrestricted one.

- So the question will be how much improvement in the fit do we get by relaxing the restrictions relative to the loss of precision that follows. The distribution of the test statistic will give us a measure of this so that we can construct a decision rule.

Further Definitions

- Define the **Unrestricted Residual Residual Sum of Squares (URSS)** as the residual sum of squares obtained from estimating the unrestricted model.
- Define the **Restricted Residual Residual Sum of Squares (RRSS)** as the residual sum of squares obtained from estimating the restricted model.
- Note that according to our argument above $RRSS \geq URSS$

- Define the **degrees of freedom** as N-k where $\boldsymbol{N}$ is the sample size and $\boldsymbol{k}$ is the number of parameters estimated in the unrestricted model (I.e under the alternative hypothesis)
- Define by $\boldsymbol{q}$ the number of restrictions imposed (in both our examples there were two restrictions imposed

The F-Statistic

- The Statistic for testing the hypothesis we discussed is

$$F = \frac{(RRSS - URSS)/q}{URSS/(N-K)}$$

- The test statistic is always positive. We would like this to be "small". The smaller the *F*-statistic the less the loss of fit due to the restrictions

- Defining "small" and using the statistic for inference we need to know its distribution.

The Distribution of the *F*-statistic

- As in our earlier discussion of inference we distinguish two cases:

Normally Distributed Errors

  - The errors in the regression equaion are distributed normally. In this case we can show that under the null hypothesis H_0 the *F*-statistic is distributed as an ***F*** **distribution with degrees of freedom (*q,N-k)*** .
  - The number of restrictions q are the **degrees of freedom of the numerator.**
  - *N-K* are the **degrees of freedom of the denominator.**

- Since the smaller the test statistic the better and since the test statistic is always positive we only have one critical value.
- For a test at the α level of significance we choose a critical value of $F_{1-\alpha,(q,N-k)}$

- If the test statistic is below the critical value we accept the null hypothesis.
- Otherwise we reject.

Examples

- Examples of Critical values for 5% tests in a regression model with 6 regressors under the alternative
  - Sample size 18. One restriction to be tested: Degrees of freedom 1, 12: $F_{1-0.05,(1,12)} = 4.75$
  - Sample size 24. Two restrictions to be tested: degrees of freedom 2, 18: $F_{1-0.05,(2,18)} = 3.55$
  - Sample size 21. Three restrictions to be tested: degrees of freedom 3, 15: $F_{1-0.05,(3,15)} = 3.29$

Inference with non-normal errors

- When the regression errors are not normal (but satisfy all the other assumptions we have made) we can appeal to the central limit theorem to justify inference.
- In large samples we can show that the q times the *F* statistic is distributed as a random variable with a

$$qF \overset{\alpha}{\sim} \chi^2_q \text{ distribution}$$

Examples

- Examples of Critical values for 5% tests in a regression model with 6 regressors under the alternative. Inference based on large samples:

  – One restriction to be tested: Degrees of freedom 1. :

  $$\chi^2_{1-0.05,1} = 3.84$$

  – Two restrictions to be tested: degrees of freedom 2:

  $$\chi^2_{1-0.05,2} = 5.99$$

  – Three restrictions to be tested: degrees of freedom 3:

  $$\chi^2_{1-0.05,3} = 7.81$$

Example: The Demand for butter:

Hypothesis to be tested: Butter and margarine advertising do not change demand and income elasticity of butter is one: **Three restrictions**

Unrestricted Model

```
. regr lbp  lpbr lpsmr lryae ltba lrma

      Source |       SS       df       MS                      Number of obs =      51
-------------+------------------------------                   F(  5,    45) =   11.74
       Model |  .357443231     5  .071488646                   Prob > F      =  0.0000
    Residual |  .273965407    45   .00608812                   R-squared     =  0.5661
-------------+------------------------------                   Adj R-squared =  0.5179
       Total |  .631408639    50  .012628173                   Root MSE      =  .07803

------------------------------------------------------------------------------
  log butter purchases |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------+------------------------------------------------------
log price of butter    |   -.7297508   .1540721    -4.74   0.000    -1.040068   -.4194336
log price of margarine |    .7795654   .3205297     2.43   0.019     .1339856    1.425145
log real income        |    .9082464    .510288     1.78   0.082    -.1195263    1.936019
log butter advertising |   -.0167822   .0133142    -1.26   0.214    -.0435984    .0100339
log margarine advertising |-.0059832   .0166586    -0.36   0.721    -.0395353     .027569
Constant               |    6.523365   .8063481     8.09   0.000     4.899296    8.147433
------------------------------------------------------------------------------
```

Restricted Model

$$lbp = a + b_1 lpbr + b_2 lpsmr + 1 \times lryae + 0 \times ltba + 0 \times lrma + u$$

$$(lbp - lryae) = a + b_1 lpbr + b_2 lpsmr + u$$

. **gen lbpry=lbp-lryae**

. regr lbpry lpbr lpsmr

```
      Source |       SS       df       MS                  Number of obs =      51
-------------+------------------------------                F(  2,    48) =   43.74
       Model |  .523319203     2  .261659601                Prob > F      =  0.0000
    Residual |  .287162961    48  .005982562                R-squared     =  0.6457
-------------+------------------------------                Adj R-squared =  0.6309
       Total |  .810482164    50  .016209643                Root MSE      =  .07735
```

New dep var	lbpry	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
log price of butter	lpbr	-.7481124	.14332	-5.22	0.000	-1.036277	-.4599483
log price of margarine	lpsmr	.782316	.2466846	3.17	0.003	.2863234	1.278309
Constant	_cons	6.255797	.5969626	10.48	0.000	5.055523	7.456071

The Test

- The value of the test statistic is

$$F = \frac{(0.287 - 0.274)/3}{0.274/(51-6)} = 0.71$$

- The critical value for a 5% test wit (3,45) degrees of freedom is 2.81
- We accept the null hypothesis since 0.71<2.81.

A Large sample example: Testing for seasonality in fuel expenditure

. regress wfuel logex spring summer autumn

```
      Source |       SS           df       MS                      Number of obs =    4785
-------------+------------------------------                       F(  4,  4780) =  100.33
       Model |  .549215033     4  .137303758                       Prob > F      =  0.0000
    Residual |  6.54124051  4780   .00136846                       R-squared     =  0.0775
-------------+------------------------------                       Adj R-squared =  0.0767
       Total |  7.09045554  4784  .001482119                       Root MSE      =  .03699
```

Share of Fuel in budget	wfuel |	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
Log real Expenditure	logex |	-.0116965	.0007125	-16.42	0.000	-.0130934	-.0102996
	spring |	.0064453	.0015151	4.25	0.000	.0034751	.0094155
	summer |	-.0020176	.0015453	-1.31	0.192	-.005047	.0010118
	autumn |	-.0099518	.001524	-6.53	0.000	-.0129396	-.0069641
	_cons |	.1167705	.003095	37.73	0.000	.110703	.1228381

The Restricted Model: Excludes the Seasonal Indicator

```
. regress wfuel logex

      Source |       SS       df       MS              Number of obs =    4785
-------------+------------------------------           F(  1,  4783) =  270.14
       Model |   .37905073     1   .37905073           Prob > F      =  0.0000
    Residual |  6.71140481  4783  .001403179           R-squared     =  0.0535
-------------+------------------------------           Adj R-squared =  0.0533
       Total |  7.09045554  4784  .001482119           Root MSE      =  .03746

------------------------------------------------------------------------------
Share of Fuel in budget       wfuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
log real expenditure           logex |  -.0118527   .0007211   -16.44   0.000    -.0132665   -.0104389
Constant                      _cons |   .1160574   .0030032    38.64   0.000     .1101697    .1219451
------------------------------------------------------------------------------
```

The Chi squared test statistic: (6.71- 6.54)/(6.54/4785)= 124.38

Critical Value for 5% test and three degrees of freedom 7.81

Hypothesis rejected since 124.38>7.81

Alternative form of the F-statistic using the R squared

- So long as the Total sum of squares is kept the same between models we can also write the F-statistic as

$$F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(N - k)}$$

- where U refers to the unrestricted model and R to the restricted model
- This will not work if we compute the R squared with different dependent variables in each case (e.g. because of transformations.

Heteroskedasticity

- Heteroskedasticity means that the variance of the errors is not constant across observations.
- In particular the variance of the errors may be a function of explanatory variables.
- Think of food expenditure for example. It may well be that the “diversity of taste” for food is greater for wealthier people than for poor people. So you may find a greater variance of expenditures at high income levels than at low income levels.

- Heteroskedasticity may arise in the context of a “random coefficients model.
- Suppose for example that a regressor impacts on individuals in a different way

$$Y_i = a + (b_1 + \varepsilon_i) X_{i1} + u_i$$

$$Y_i = a + b_1 X_{i1} + \varepsilon_i X_{i1} + u$$

- Assume for simplicity that å and u are independent.
- Assume that å and *X* are independent of each other.
- Then the error term has the following properties:

$$E(\varepsilon_i X_i + u_i \mid X) = E(\varepsilon_i X_i \mid X) + E(u_i \mid X) = E(\varepsilon_i \mid X) X_i = 0$$

$$Var(\varepsilon_i X_i + u_i \mid X) = Var(\varepsilon_i X_i \mid X) + Var(u_i \mid X) = X_i^2 \sigma_\varepsilon^2 + \sigma^2$$

- Where σ_ε^2 is the variance of å

Implications of Heteroskedasticity

- Assuming all other assumptions are in place, the assumption guaranteeing unbiasedness of OLS **is not violated**. Consequently **OLS is unbiased** in this model

- However the assumptions required to prove that OLS is efficient are violated. Hence **OLS is not BLUE** in this context

- We can devise an efficient estimator by reweighing the data appropriately to take into account of heteroskedasticity