Lecture 4
Hypothesis Testing

- We may wish to test *prior* hypotheses about the coefficients we estimate.
- We can use the estimates to test whether the data rejects our hypothesis.
- An example might be that we wish to test whether an elasticity is equal to one.
- We may wish to test the hypothesis that X has no impact on the dependent variable Y.
- We may wish to construct a confidence interval for our coefficients.

- A hypothesis takes the form of a statement of the true value for a coefficient or for an expression involving the coefficient.
- The hypothesis to be tested is called the null hypothesis.
- The hypothesis which it is tested again is called the alternative hypothesis.
- Rejecting the null hypothesis <u>does not imply accepting the alternative</u>
- We will now consider testing the simple hypothesis that the slope coefficient is equal to some fixed value.

Setting up the hypothesis

- Consider the simple regression model:

$$Y_i = a + bX_i + u_i$$

- We wish to test the hypothesis that $b=d$ where d is some known value (for example zero) against the hypothesis that *b is not equal to zero*. We write this as follows
- We write

$$H_0 : b = d$$

$$H_a : b \neq d$$

- To test the hypothesis we need to know the way that our estimator is distributed.
- We start with the simple case where we assume that the error term in the regression model is a *normal* random variable with mean zero and variance σ^2. This is written as $u \sim N(0, \sigma^2)$
- Now recall that the OLS estimator can be written as

$$\hat{b} = b + \sum_{i=1}^{N} w_i u_i$$

- Thus the OLS estimator is equal to a constant (*b*) plus a weighted sum of normal random variables
- <u>Weighted sums of normal random variables are also normal</u>

The distribution of the OLS slope coefficient

- It follows from the above that the OLS coefficient is a Normal random variable.
- What is the mean and what is the variance of this random variable?
- Since OLS is unbiased the mean is *b*
- We have derived the variance and shown it to be

$$Var(\hat{b}) = \frac{1}{N}\frac{\sigma^2}{Var(X)}$$

- Since the OLS estimator is Normally distributed this means that

$$z = \frac{\hat{b} - b}{\sqrt{Var(\hat{b})}} \sim N(0,1)$$

- The difficulty with using this result is that we do not know the variance of the OLS estimator because we do not know σ^2
- This needs to be estimated
- An unbiased estimator of the variance of the residuals is the residual sum of squares divided by the number of observations minus the number of estimated parameters. This quantity (*N-2)* in our case is called the degrees of freedom. Thus

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N-2}$$

- Return now to hypothesis testing. Under the null hypothesis *b*=*d*. Hence it must be the case that

$$z = \frac{\hat{b} - d}{\sqrt{Var(\hat{b})}} \sim N(0,1)$$

- We now replace the variance by its estimated value to obtain a <u>*test statistic*</u>:

$$z^* = \frac{\hat{b} - d}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{N} (X_i - \overline{X})^2}}}$$

- This test statistic is no longer Normally distributed, but follows the t-distribution with <u>*N-2 degrees of freedom.*</u>

Testing the Hypothesis

- Thus we have that *under the null hypothesis*

$$z^* = \frac{\hat{b} - d}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{N}(X_i - \overline{X})^2}}} \sim t_{N-2}$$

- The next step is to choose the size of the test (significance level). This is the probability that we reject a correct hypothesis.
- The conventional size is 5%. We say that the size $\alpha = 0.05$
- We now find the *critical values* $t_{\alpha/2,N}$ and $t_{1-\alpha/2,N}$

- We accept the null hypothesis if the test statistic is between the critical values corresponding to our chosen size.
- Otherwise we reject.
- The logic of hypothesis testing is that if the null hypothesis is true then the estimate will lie within the critical values $100 \times (1 - a)\%$ of the time.
- The ability of a test to reject a hypothesis is called the power of the test.

Confidence Interval

- We have argued that

$$z^* = \frac{\hat{b} - d}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{N}(X_i - \overline{X})^2}}} \sim t_{N-2}$$

- This implies that we can construct an interval such that the chance that the true *b* lies within that interval is some fixed value chosen by us. Call this value

$$1 - \alpha$$

- For a 95% confidence interval say this would be 0.95.

- From statistical tables we can find critical values such that any random variable which follows a t-distribution falls between these two values with a probability of $1-\alpha$. Denote these critical values by $t_{\alpha/2,N}$ and $t_{1-\alpha/2,N}$

- For a t random variable with 10 degrees of freedom and a 95% confidence these values are (2.228,-2.228).
- Thus

$$pr(t_{\alpha/2} < z^{*} < t_{1-\alpha/2}) = 1-\alpha$$

- With some manipulation we then get that

$$pr(\hat{b} - se(\hat{b}) \times t_{\alpha/2} < b < \hat{b} + se(\hat{b}) \times t_{\alpha/2}) = 1-\alpha$$

- The term in the brackets is the confidence interval.

Example

- Consider the regression of log quantity of butter on the log price again
- regr lbp lpbr

Number of obs = 51

lbp	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
log price	-.8421586	.1195669	-7.04	0.000	-1.082437	-.6018798
_cons	4.52206	.1600375	28.26	0.000	4.200453	4.843668

- The statistic for the hypothesis that the elasticity is equal to one is

$$z = \frac{-0.84-(-1)}{0.12} = \frac{0.16}{0.12} = 1.33$$

- Critical values for the t distribution with 51-2 = 49 degrees of freedom (51 observations, 2 coefficients estimated) and significance level of 0.05 is approximately (2,-2)(from stat tables)
- Since -1.33 lies within this range we accept the null hypothesis
- The 95% **confidence interval** is

$$-0.84 \pm 2 \times 0.12 = (-1.08, -0.6)$$

- Thus the true elasticity lies within this range with 95% probability.
- Everything we have done is of course applicable to the constant as well. The variance formula is different however.

- Do we need the assumption of normality of the error term to carry out inference (hypothesis testing)?
- Under normality our test is *exact.* This means that the test statistic has exactly a *t distribution.*
- We can carry out tests based on *asymptotic approximations* when we have large enough samples.
- To do this we will use Central limit theorem results that state that in large samples weighted averages are distributed as normal variables.

A Central limit theorem

- Suppose we have a set of independent random numbers $v_i, \quad i = 1,\ldots,N$ all with constant variance s^2 and mean μ . Then

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (v_i - \mu) \quad \overset{\alpha}{\sim} \quad N(0, s^2)$$

- Where the symbol $\overset{\alpha}{\sim}$ reads “distributed asymptotically”, i.e. as the sample size N tends to infinity.

- This extends to weighted sums. Let μ=0. So we also have that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(w_i v_i) \overset{\alpha}{\sim} N\left(0, s^2 p\lim_{N\to\infty}\left(\frac{1}{N}\sum w_i^2\right)\right)$$

where $$p\lim_{N\to\infty}\frac{1}{N}\sum w_i^2$$

is the *probability limit* of the sum of squares of the weights. It is a limit for sums of random variables. This limit can be estimated in practice by the sum itself:

$$\frac{1}{N}\sum w_i^2$$

We require the limit to be finite: $p\lim_{N\to\infty}\frac{1}{N}\sum w_i^2 < \infty$

Applying the CLT to the slope coefficient for OLS

- Recall that the OLS estimator can be written as

$$\hat{b} - b = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(u_i - \overline{u})}{\sum_{i=1}^{N} (X_i - \overline{X})^2} = \sum_{i=1}^{N} w_i u_i$$

- This is a weighted sum of random variables as in the previous case.

The Central limit theorem applied to the OLS estimator

- We can apply the central limit theorem to the OLS estimator.
- Thus according to the central limit theorem we have that

$$\sqrt{N}(\hat{b}-b) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(X_i-\bar{X})u_i}{\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2} \overset{\alpha}{\sim} N\left(0,\sigma^2 p\lim_{N\to\infty}\left(\frac{1}{\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2}\right)\right)$$

- Comparing with the previous slide the weights are

$$w_i = \frac{(X_i-\bar{X})}{\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2}$$

- The implication is that the statistic we had before has *a normal distribution in large samples* irrespective of how the error term is distributed if it has a constant variance Assumption 2 - homoskedasticity.

$$z^* = \frac{\hat{b} - d}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{N} (X_i - \overline{X})^2}}} \overset{\alpha}{\sim} N(0,1)$$

- Note how the *N*s cancel from the top and bottom. In fact the test statistic is *identical* to the one we used under normality. The only difference is that now we will use the critical values of the Normal distribution. For a size of 5% these are +1.96 and – 1.96.

- The expression on the denominator is nothing but the standard error of the estimator.

- The test statistic for the special case when we are testing that the coefficient is in fact zero (no impact on *Y*) is often called the **t-statistic**.
- For a large sample test we can accept the hypothesis that a coefficient is zero with a 5% level of significance if the **t-statistic** is between (-1.96,1.96)

Example

lmap	Coef.	Std. Err.	**t**	P>|t|	**[95% Conf. Interval]**	
lpsmr	-.6856449	.2841636	**-2.41**	0.020	**-1.256693**	**-.1145967**
_cons	4.183766	.534038	7.83	0.000	3.110577	5.256956

- Regression of log margarine purchases on the log price.
- Test that the price effect is zero. Assume large enough sample and use the critical values from the Normal distribution.
- T-statistic = -0.69/0.28=-2.41
- 95% Normal critical values are –1.96,1.96
- The hypothesis is rejected
- The 95% confidence interval is (–1.26,-0.115) Quite wide which implies that the coefficient is not very precisely estimated.

Summary

- When the error term is normally distributed we can carry out exact tests by comparing the test statistic to critical values from the t-distribution
- If the assumption of normality is not believed to hold we can still carry out inference when our sample is large enough.
- In this case we simply use the normal distribution.