ECON-122
Introduction to Econometrics

Agnieszka Postepska

July 26, 2011

What is inference?

- Inference is one of the main idea of the exercise we are doing in this class
- we want to be able to test hypothesis about the parameters of the model describing the underlying population
- to do that, econometricians developed set of tests, which we can apply to our estimated models
- recall that the *estimates* differ across random samples from the underlying population - they are realizations of a random variable which is defined by the formulas for *estimators*
- now, we are going to take a closer look at the *sampling distribution* of the OLS estimator.

Inference in the context of parameter estimation

- we want to be able to make statements about the *true* values of the $\beta's$
- simplest statement of interest is of the following form $\beta_j = \mu$, where μ is some constant
- the statement above is called the **null hypothesis,** H_0 - this statement we will test it in our sample
- so, in our example the null is: $H_0 : \beta_j = \mu$
- the statement against which we are testing is called **alternative hypothesis,** H_1
- the simplest version of the alternative hypothesis is $H_1 : \beta_j \neq \mu$
- note that the alternative is much more general than the null and this is the case in general as it incorporates all possible values the parameter can take on if null is not true
- also note that both, the null and the alternative, are the statements about the true parameter

Testing hypothesis

- in order to test hypothesis we need to construct a *test statistic* - expression that we can compute in our sample
- then we need to know how this statistic behaves - so find its sample distribution
- to proceed we need to find the sample distribution of the $\hat{\beta}'s$ (the OLS estimators for $\beta's$)
- what are the two things that we already know that characterize the distribution of the $\hat{\beta}'s$?

Sampling distribution of the OLS estimators

- as we condition everything on the $x's$, the values of the independent variable, the only random thing that is left is the error term
- it is straightforward then that the distribution of the OLS estimators depends on the distribution of the errors
- Thus, we need to state another assumption:
  A6: Normality: the error u is independent of the $x's$ and is normally distributed with zero mean and variance σ^2:
  $u \sim \text{Normal}\ (0, \sigma^2)$
- note that A6 implies A4 and A5 and much more as we are assuming the whole shape of the distribution now

Classical Linear Regression Model

- **CLRM is defined by A1-A6**
- in application: whether we can or cannot assume normality of errors is an empirical question- there is no theory that tells us how errors are distributed (use common sense, reasoning etc...)
- moreover, sometimes the distribution might be truncated because of the nature of the data (minimum wages, minimum height etc) - that's when we can use some transformation (such as take logs) to get a distribution closer to normal

What does the normality of the error term give us?

Normality of the error term translates into normal sampling distribution of the OLS estimator:

Normal Sampling Distributions:

Under the assumptions A1-A6, conditional on the sample values of the independent variables,

$$\hat{\beta}_j \sim \text{ Normal } [\beta_j, Var(\hat{\beta}_j)],$$

where $Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_i (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)}$. Therefore,

$$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}j) \sim \text{ Normal}(0, 1).$$

t-test of OLS parameters

Now we are almost ready for some hypothesis testing. We will start with testing individual parameters of the OLS regression. We are not yet completely ready because we do not know the variance of the sampling distribution of OLS estimator. We need the following result first:

t distribution for standardized estimators:

Under the assumptions A1-A6,

$$(\hat{\beta}_j - \beta_j)/se(\hat{\beta}j) \sim t_{n-k-1},$$

where $k+1$ is the number of unknown parameters in the population model (k slopes parameters and the intercept) and $(n-k-1)$ is the number of degrees of freedom.

Now we are ready for hypothesis testing!

Testing hypothesis about individual coefficient

In most applications, primary interest is in testing the following null hypothesis:

$$H_0 : \beta_j = 0$$

The null is that explanatory variable x_j has no effect on the explained variable, once all x_{-j} are accounted for.

Cook book procedure for a simple hypothesis testing

- Specify the null and the alternative
  - Two tailed test: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$
  - One tailed test: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \geq 0$ or $H_1 : \beta_j \leq 0$
- Define test statistic: $t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t_{n-k-1}$ (you want a test statistic of which distribution is known under the null)
- Pick the **decision rule**: if distribution is true than you would expect to find most realization of a r.v. close to zero - so:
  - if the particular realization you get is close to 0, you cannot reject the null,
  - if it is far away from 0, you reject the null

Exclusion test

- by testing whether a parameter is significantly different from zero, we are basically testing whether it can be excluded from the model this test is reported automatically by most of the statistical softwares
- if $\hat{\beta}_j = 0$ once all other $x's$ have been accounted for, it means that x_j has no effect on y
- it is important not to include irrelevant variables as it decreases the precision of other estimators (increase the variance of estimators of other $\hat{\beta}_j's$)
- in practice: pay attention to relative size of the coefficient to its standard error - if coefficient is large, even though it is insignificant it still makes sense to include it in the regression; if, however, parameter estimate is small, even if it is significant, it makes little practical difference to the model

Two types of errors one can make when testing a hypothesis

- TYPE 1 ERROR: reject the null when it is true
- TYPE 2 ERROR: fail to reject the null when the null is false

TYPE 1 ERROR is considered more severe, so pick the confidence level to minimize the probability of this error (usually 1 percent, 5 percent or 10 percent).

One-tailed test

$H_1 : \beta_j > 0$

- not interested in possibility of true values less than hypothesized value for some reason (maybe based on economic theory or common sense) - we are ruling out possibility of β taking on values below 0, the hypothesized value
- we reject the null when

$$t > c_\alpha$$

where c_α is the **critical value** defined by choosing the confidence level (α) (see graph!!!!)

- since the t-distribution is symmetric, to test the reverse hypothesis $H_1 : \beta_j < 0$, use the reverse rejection rule $\rightarrow$ reject the null if $t < -c_\alpha$

Two-tailed test

$H_1 : \beta_j \neq 0$

- now we are interested in possibility that true values are less or greater than hypothesized value - we are not ruling any values of the parameters out
- we reject the null when

$$|t| > c_{\frac{\alpha}{2}}$$

where c_α is the **critical value** defined by choosing the confidence level (α) (see graph!!!!)

Testing other hypothesis about β_j

- usually we are interested in testing $H_0 : \beta_j = 0$, however sometimes it is useful to test the null of the following form:

$$H_0 : \beta_j = a_j,$$

where a_j is the hypothesized value of β_j

- the appropriate test statistic is then:

$$t = \frac{(\hat{\beta}_j - a_j)}{se(\hat{\beta}_j)}$$

- to sum up, the general form of t-statistic is:

$$t = \frac{(\text{estimate} - \text{hypothesized value})}{\text{standard error}}$$

Some relevant issues

- **Comparison of one- and two-tailed test:** one tailed test has a larger rejection zone for the same significance level because it doesn't have an alternative rejection zone at the other tail of the distribution
- critical values are found in tables of distributions
- rather say "fail to reject the null" than "accept the alternative"

p-values

When we are rejecting the null, we do not differentiate between whether the test statistic is just outside of the rejection region, or way beyond the range. An alternative measure to look at is p-value:
p-value - is the smallest significance level at which the null would be rejected given the observed value of test statistic

- another interpretation: p -value is the probability of observing a t statistic as extreme as we did if the null hypothesis is true
- p-value is a probability, so its value must be between 0 and 1
- ONE-TAILED t-TEST: $p_1 = P(T > t)$ or $p_1 = P(T < t)$
- TWO-TAILED t-TEST: $p_2 = P(T > t \text{ or } T < -t) = P(|T| > |t|)$
- note that p-value from a two-tailed test will always be twice the p-value for a one-tailed test: $p_2 = 2p_1$
- p-value is most useful when rejection decision depends on choice of significance level