

ECO383: Supplementary Note on Inference

Once we know the distribution of b^{OLS} , we are in a position to carry out hypothesis testing.

Note that under modified Assumption 4, we have that

$$u_i \sim \text{i.i.d. } N(0, \sigma^2), \text{ for all } i.$$

This assumption (clearly) implies that

$$E(u_i) = 0, \text{ for all } i \text{ (Assumption 1)}$$

$$\text{Var}(u_i) = \sigma^2, \text{ for all } i \text{ (Assumption 2)}$$

$$E(u_i u_j) = 0, \text{ for } i \neq j \text{ (Assumption 3).}$$

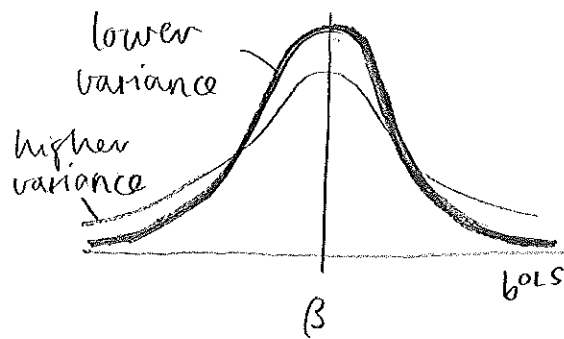
Given our underlying model, it also implies that

$$b^{OLS} \sim N\left(\beta, \frac{\sigma^2}{\sum_i (c_i - \bar{c})^2}\right).$$

This statement implies that b^{OLS} has a normal p.d.f., as follows:

$$f(b^{OLS}) = \frac{1}{\sqrt{2\pi \left(\frac{\sigma^2}{\sum_i (c_i - \bar{c})^2}\right)}} e^{-\frac{(b^{OLS} - \beta)^2}{2 \frac{\sigma^2}{\sum_i (c_i - \bar{c})^2}}} \quad (1)$$

Once we know that density, we can plot the distribution of our random variable, b^{OLS} . It will follow the usual normal bell-shape, centred around β and with a 'spread' that is increasing in the variance, $\sigma^2 / \sum_i (c_i - \bar{c})^2$. Plotting



two illustrative distributions, both centred around β , we see the lower variance distribution (say, with a smaller σ^2 value) is more concentrated around the mean.

Note one very important point: we do not actually know what β is ^{in terms of its value}. That is why we go to great lengths to try and estimate it, based on samples drawn from the overall population.

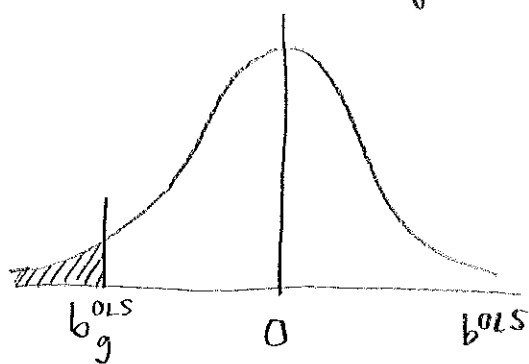
Given the structure we have put in place, though, we can do a neat thing. We can say, suppose β took on a particular value. (This would re-centre the normal

distribution associated with β - the variance would remain the same.) Often, to be concrete, we will be interested to see whether β could really be zero, given the sample slopes that we have obtained. Let us write that possibility (the 'null' hypothesis)

$$H_0: \beta = 0$$

This is equivalent to a 'no effect of class size' scenario that, among other things, would be very relevant for policy.

Under H_0 , we can plot the implied distribution of our slope estimator:



Having done that, we can then compute the probability that our sample estimate is

equal to, say, b_g^{OLS} or even more negative under the null hypothesis.

Why might that be interesting?

Given our distributional assumption in (1),

we are in a position to say what the probability is, under H_0 , that we would draw a random variable (a particular value of our slope estimator) that was less than or equal to (say) the particular value b_9^{OLS} in the diagram above. The associated probability is given by the shaded area.

[Aside: that probability is known as the "P-value" associated with the estimate b_9^{OLS} .]

Given that implied probability (and we could obtain a similar reading for any sample quantity), we can then say how likely it would be to obtain such an estimate if the true value were really zero. If this were very unlikely, then intuitively speaking, this would provide grounds for thinking the null hypothesis probably was not true.

Practical implementation

In the above discussion, we have worked entirely with the specific distribution given in (1). If our sample changed, we would likely have a different ' $\sum_i (c_i - \bar{c})^2$ ' value, and would need to re-tabulate the implied normal distribution. Further, working out probabilities associated with the general normal distribution is not immediately straightforward, as one needs to approximate the relevant 'shaded area' if you like. (Formally speaking, there is no closed-form solution to integrals involving the normal p.d.f.) One could write (or use) a program that generated normal 'tables' for any arbitrary settings of the distribution parameters. But there is a much more common (and more convenient) approach that is completely equivalent. This involves standardizing the random variable - in this instance, b^{OLS} - by deducting off the mean

and dividing through by the square root of the variance. Doing so gives us a test statistic, known as a z-score, which we will write as

$$Z = \frac{b^{OLS} - \beta}{\sqrt{\sigma^2 / \sum_i (c_i - \bar{c})^2}}$$

Now, we have already shown that if $b^{OLS} \sim N\left(\beta, \frac{\sigma^2}{\sum_i (c_i - \bar{c})^2}\right)$, then Z defined above will follow a normal $(0, 1)$ distribution, so

$$Z \sim N(0, 1).$$

[Actually, we showed that $E(Z) = 0$ and $\text{Var}(Z) = 1$. Then we just need the additional fact that a linear function of a normal variable — and note that Z is a linear function of b^{OLS} — is itself normally distributed.]

Once we have the distribution of the z-score — it is a standard normal — then

we can readily implement the following testing procedure:

- i/ Compute the slope estimate for a given sample. Label this b_g^{OLS} , and note the associated sample value $\sum_i (c_i - \bar{c})^2$.
- ii/ Construct what is called our test statistic under the null hypothesis

$$Z_g = \frac{b_g^{OLS} - 0}{\sqrt{\sigma^2 / \sum_i (c_i - \bar{c})^2}} = \frac{b_g^{OLS}}{\sqrt{\sigma^2 / \sum_i (c_i - \bar{c})^2}}$$

This is a number.

- iii/ Next, given that the test statistic follows a $N(0, 1)$ distribution, we can compute the probability that we would see a value less than or equal to Z_g under the null.

Typically, a slightly different approach is taken at this third step. It involves constructing a 'rejection region' for the null.

Let us consider two types of (familiar) rejection regions: first for what is known as a two-sided test, the second for a one-sided test.

For a two-sided test, we compare

$$H_0: \beta = 0 \text{ (say)}$$

versus

$$H_1: \beta \neq 0.$$

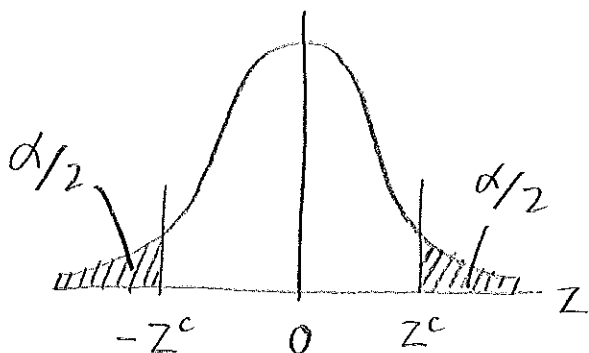
We set a 'level' of the test at a percentage α . Then, in the symmetric case, figure out a pair of critical values, indexed by z^c , such that

$$\Pr(z < -z^c) = \Pr(z > z^c) = \alpha/2.$$

[Alternatively, we could write (more compactly)]

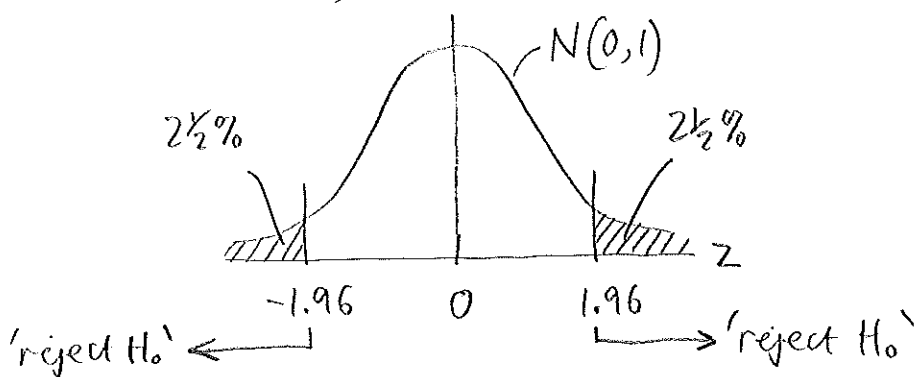
$$\Pr(|z| > z^c) = \alpha.]$$

We can illustrate these as:



The shaded areas are each equal to $\alpha/2$, the respective probabilities that z would fall into each region.

Typically, α will be set at a small number, like 5%. That being the case, the critical values will be $-Z_c = -1.96$ and $Z_c = 1.96$. If we set those critical values, then under the standard normal distribution, the probabilities that the standardized random variable Z falls to the left and right of the critical values (beneath the shaded areas below) equal $2\frac{1}{2}$ percent respectively.



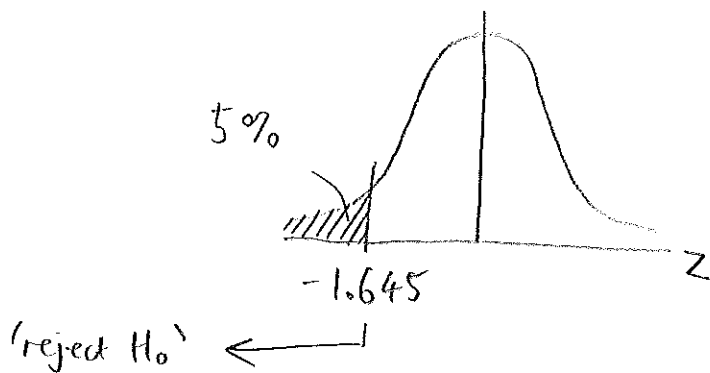
How would we implement our hypothesis test? We would reject H_0 in favour of the two-sided alternative H_1 if our test statistic Z_g fell into the rejection region (which has two portions), illustrated above.

For a one-sided test, we compare $H_0: \beta = 0$ versus $H_1: \beta < 0$ (say).

If we set $\alpha = 5\%$, then the one-sided critical value in this instance would involve $-z^c = -1.645$. This means that

$$\Pr(z < -z^c) = \Pr(z < -1.645) = 5\%.$$

To illustrate:



In words, if we found $z_0 < -1.645$, then we would reject H_0 in favour of the one-sided alternative, H_1 .

Looking ahead:

There are two issues still to address:

- i) If we set a particular 'level' α of our test, then there is a non-zero probability that we make an incorrect decision. We need to be aware of that.
- ii) We have implicitly assumed we know σ^2 . Typically, we will have to estimate it. That changes the distribution of our test statistic. (more to follow...)