

Lectures 5 & 6: Hypothesis Testing

... in which you learn to apply the concept of “statistical significance” to OLS estimates, learn the concept of “t values” , how to use them in regression work and come across the idea of “confidence intervals” and “p values” in the process

Hypothesis Testing

Given now understand how to estimate the slope and intercept using OLS (and that OLS is the most efficient of all unbiased estimation procedures given the 4 Gauss-Markov assumptions hold). The next thing is to focus on the contribution of the individual right hand side variables. Just because a variable has a large coefficient does not necessarily mean its contribution is significant. This means we need to understand the ideas behind the t value and how to use t values in applied work

The screenshot shows the Stata/SE 10.0 interface with the following components:

- Command Window:**

```

1 use "C:\qm2\Lecture 2\bhatshow.dta"
2 reg hourpay age
3 reg hourpay age if 20<=age &
4 reg hourpay age if 20<=age &
5 reg hourpay age if 20<=age &
6 reg hourpay age if 21<=age &
7 reg hourpay age if 22<=age &
8 reg hourpay age if 22<=age &
9 reg hourpay age if 22<=age &
10 reg hourpay age
11 reg hourpay age if 22<=age &
12 use "C:\qm2\Lecture 2\bhatshow.dta"
13 reg hourpay age
14 reg hourpay age
15 reg hourpay age
16 reg hourpay age
17 use "C:\qm2\Lecture 2\bhatshow.dta"
18 reg hourpay age
19 reg hourpay age if 22<=age &
20 reg hourpay age if 22<=age &
21 use "C:\qm2\Lecture 2\bhatshow.dta"
22 reg hourpay age
23 reg hourpay age if 22<=age &

```
- Results Window:**

Regression 1: reg hourpay age

Source	SS	df	MS
Model	115.800815	1	115.800815
Residual	6010.5411	240	25.0439212
Total	6126.34191	241	25.4205059

Number of obs = 242
F(1, 240) = 4.62
Prob > F = 0.0325
R-squared = 0.0189
Adj R-squared = 0.0148
Root MSE = 5.0044

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0599578	.0278831	2.15	0.033	.005031 .1148847
_cons	6.481781	1.141873	5.68	0.000	4.232408 8.731155

Regression 2: reg hourpay age if 22<=age & age<=26

Source	SS	df	MS
Model	38.3499476	1	38.3499476
Residual	381.690212	24	15.9037588
Total	420.040159	25	16.8016064

Number of obs = 26
F(1, 24) = 2.41
Prob > F = 0.1335
R-squared = 0.0913
Adj R-squared = 0.0534
Root MSE = 3.988

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.034335	.6660834	1.55	0.134	-.3403938 2.409063
_cons	-17.31101	15.95395	-1.09	0.289	-50.23833 15.61632
- Variables Window:**

Name	Label	Type
edage	age when compld...	b
hourpay	gross hourly pay	c
age	age	t
female		b
grad		b

If wish to make inferences about how close an estimated value is to a hypothesised value or even to say whether the influence of a variable is not simply the result of statistical chance then need to make one additional assumption about the behaviour of the (true, unobserved) residuals in the model

We know already that $u_i \sim (0, \sigma^2_u)$

Now assume additionally that residuals follow a Normal distribution $u_i \sim N(0, \sigma^2_u)$

(Since residuals capture influence of many unobserved (random) variables, can use **Central Limit Theorem** which says that the sum of a set of random variables will have a normal distribution)

If u is normal, then it is easy to show that the OLS coefficients (which are a linear function of u) are also normally distributed with the means and variances that we derived earlier. So

$$\hat{\beta}_0 \sim N(\beta_0, \text{Var}(\hat{\beta}_0)) \quad \text{and} \quad \hat{\beta}_1 \sim N(\beta_1, \text{Var}(\hat{\beta}_1))$$

Can use this to test hypotheses about the values of individual coefficients

If a variable is normally distributed we know that the distribution of values is
a) symmetric and b) centred on its mean value

and that:

66% of values lie within mean ± 1 *standard deviation

95% of values lie within mean ± 1.96 *standard dev.

99% of values lie within mean ± 2.9 *standard dev.

Easier to work with the *standard normal distribution* which has a mean of 0 and variance of 1

- this can be obtained from any normal distribution by subtracting the mean and dividing by the standard deviation

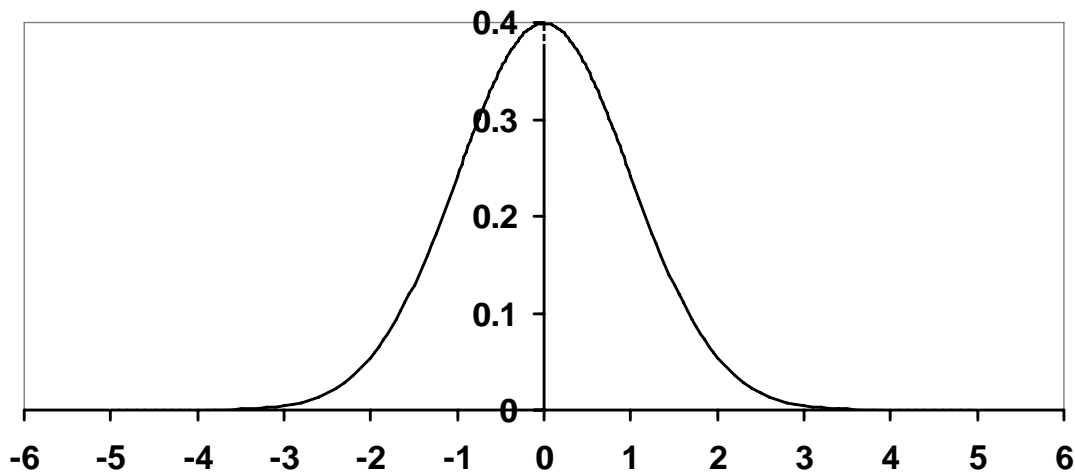
ie Given $\hat{\beta}_1 \sim N(\beta_1, \text{Var}(\hat{\beta}_1))$ then $z = \frac{\hat{\beta}_1 - \beta_1}{s.d.(\hat{\beta}_1)} \sim N(0,1)$

Since the mean of this variable is zero and because a normal distribution is symmetric and centred around its mean, and the standard deviation=1, we know that

66% of values lie within 0 ± 1

95% of values lie within 0 ± 1.96

99% of values lie within 0 ± 2.9



Here is a graph of a normal distribution with zero mean and unit variance

Can use this to construct *confidence intervals* of the form

$$\Pr[-\delta \leq z \leq \delta] = 1 - \alpha$$

which says that the probability of any value drawn from a standard normal distribution lying between the values $\pm \delta$ is $1 - \alpha$ %

α is the **size** or **significance level** of the test, $0 < \alpha < 1$

δ is the limit or **critical value** of the interval

For example, given a standard normal variable we can be 95% confident that all values will lie in the range ± 1.96

or

$$(1) \quad \Pr[-1.96 \leq z \leq 1.96] = 1 - .05 = 0.95$$

Why?

- since we know 95% of all values of a variable that has mean 0 and variance 1 will lie within

$$0 \pm 1.96 * \text{standard deviation}$$

If an estimate falls within this range it is said to lie in the **acceptance region**

(95 times out of a 100, estimates will lie in the range $0 \pm 1.96 \cdot \text{standard deviation}$)

Testing A Hypothesis Relating To A Regression Coefficient

Model: $Y = \beta_0 + \beta_1 X + u$

Null hypothesis: $H_0: \beta_1 = \beta_1^0$

Alternative hypothesis: $H_0: \beta_1 \neq \beta_1^0$

Example

$Cons = \beta_0 + \beta_1 Income + u$

Null hypothesis: $H_0: \beta_1 = 0$

Alternative hypothesis: $H_0: \beta_1 \neq 0$

In order to be able to say whether OLS estimate is close enough to hypothesized value so as to be acceptable, we take the distribution of estimates implied by the estimated OLS variance and look to see whether this range will contain the hypothesized value.

We now know that

$$z = \frac{\hat{\beta}_1 - \beta_1}{s.d.(\hat{\beta}_1)} \sim N(0,1)$$

sub. this into (1) gives

$$\Pr \left[-1.96 \leq \frac{\hat{\beta}_1 - \beta_1}{s.d.(\hat{\beta}_1)} \leq 1.96 \right] = 0.95$$

or equivalently multiplying the terms in square brackets in (2) by $s.d.(\hat{\beta}_1)$

$$-1.96 \cdot s.d.(\hat{\beta}_1) \leq \hat{\beta}_1 - \beta_1 \leq 1.96 \cdot s.d.(\hat{\beta}_1)$$

and taking $\hat{\beta}_1$ to the other sides of the equality gives

$$\Pr \left[\hat{\beta}_1 - 1.96 \cdot s.d.(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96 \cdot s.d.(\hat{\beta}_1) \right] = 0.95$$

which says that given an OLS estimate and its standard deviation we can be 95% confident that the true (unknown) value for β_1 will lie in this region

Unfortunately we never know the true standard deviation of β_1 , only ever have an estimate, the standard error

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{s^2}{N * Var(X)}}$$

So have to replace the standard deviation with this in the above. When we do this we no longer have a standard normal distribution, but instead the statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1 \sqrt{N * Var(X)}}{s} \sim t(N - k)$$

is said to follow a t distribution with N-k “degrees of freedom”

N = sample size

k = no. of right hand side **coefficients** in the model

(so includes the constant)

Now this distribution has its own set of critical values at given significance levels which vary, unlike the standard normal distribution, with the degrees of freedom in the model, (which in turn depends on the sample size and the number of variables in the model)

Nevertheless the general rules established above apply

The t distribution will be centred on zero (if the hypothesised value β_1^0 and estimated value $\hat{\beta}_1$ are the same then

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} \text{ will be zero })$$

then given a large sample (>150) we can be 95% confident that the true (unknown) value for β_1 will lie in the region

$$\Pr \left[\hat{\beta}_1 - 1.96 * s.e.(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96 * s.e.(\hat{\beta}_1) \right] = 0.95$$

To test if an estimated value comes close enough to a hypothesized value β_1^0 to be acceptable use

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} \sim t(N - k)$$

Intuitively the further the estimate from the hypothesised value, the larger the estimated t value

The standard error allows the “acceptable” distance between estimate and hypothesised value to vary with the precision of the estimate. The more precise (efficient) the estimate the smaller the distance needed to stay in the acceptance region.

This gives rise to a general rule for hypothesis testing

“If **absolute** value of $\hat{t} >$ critical value, reject the null hypothesis at the $\alpha\%$ significance level.”

So if

$$\left| \hat{t} \right| = \left| \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} \right| > t_{N-k}^\alpha \quad \text{reject the null}$$

$$\left| \hat{t} \right| = \left| \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} \right| \leq t_{N-k}^\alpha \quad \text{accept the null}$$

A large estimated t value means that $\hat{\beta}_1$ is sufficiently different from the hypothesised value, even allowing for random statistical variation (through the standard error), to be acceptable with $\alpha\%$ degree of confidence

(ie $\hat{\beta}_1$ lies outside the $\alpha\%$ acceptance region generated by the hypothesised value β_1^0)

The most common levels of significance of a test are

$\alpha = 0.05$ (5% significance)
 $\alpha = 0.01$ (1% significance)

- most econometric computer packages routinely report the t value for a null hypothesis that that particular coefficient is zero (the variable has no effect)

Usual convention is to base a t-test using the critical values at the 5% level of significance.

- This in part depends on sample sizes and also the prevalence of Type I and Type II error

By **reducing** the size of the test (α) we **increase** the acceptance region for a given t estimate (and reduce the range of estimate that fall in the rejection region)

The danger of this is that increase the chance that the null hypothesis could be false, but that we accept it.

This called Type II error

By **increasing** the size of the test (α) we **reduce** the acceptance region for a given t estimate (and increase the range of estimate that fall in the rejection region)

The danger of this is that increase the chance that reject the null hypothesis even though it is true.

This called Type I error

Clearly to reduce Type I error we increase the acceptance region by choosing a smaller size of the test, but the cost of this is that it increase the risk of Type II error

So what to do?

- depends on which is the most serious
- the more you believe that avoiding Type 1 error is more important than avoiding Type II error, then the larger the acceptance region (confidence interval) and the smaller the size of the test (α).

Since rarely know the costs of these errors, have to settle on critical values which balance the two errors

- which is why the 5% level of significance is the one usually used.

It is true however that there is some **trade-off** between the sample size and the significance level of the test.

(It is more common to use the term "significance level" of the test than "size".

Unfortunately it is sometimes confusing, since to increase the significance level of the test usually means reduce the size of the test ie go from 5% to 1% level)

Since we know $s.e.(\hat{\beta}_1) = \sqrt{\frac{s^2}{N * Var(X)}}$

Then $\uparrow N \rightarrow \downarrow s.e.(\hat{\beta}_1)$

and hence because $\hat{t} = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)}$ this will \uparrow estimated t value

So estimates from larger samples will tend to have smaller standard errors and larger t values than estimates from smaller samples, other things equal.

Sometimes therefore might want to impose different significance levels depending on the sample size.

There are no rules on this but might want to think about the following guidelines:

N in the 10's	use $\alpha = 10\%$
N in the 100's	use $\alpha = 5\%$
N in the 1000's	use $\alpha = 1\%$

A useful corollary to the discussion on hypothesis testing is to be aware of the **p value** of a test

The p value is the lowest significance level at which the null hypothesis can be rejected (the exact probability of committing Type I error)

In practice this amounts to finding the significance level α which, given sample size N and no. right hand side coefficients k, equates the critical and estimated t values

$$\hat{t} = t_{N-k}^{\alpha/2}$$

Intuitively if can $\uparrow\alpha$ a lot (and in so doing reduce the acceptance region) and still accept the null hypothesis, this suggests the null hypothesis is likely to be true.

So a **high** p value (high α) is evidence **in favour** of the null and a **low** p value is evidence **in against** the null

More formally if in the regression output

p < chosen level of α (say 0.05 = 5%)
reject the null hypothesis

Statistical v. Practical Significance

Just because a variable is statistically significant in a regression does not mean that it has a large economic impact on the dependent variable

May have a large t values (and \therefore be statistically significant from zero) but if the estimated coefficient

If $\hat{\beta} = \frac{\partial y}{\partial X}$ is small then

the impact of a change in X on y $\hat{\beta} \partial X = \partial y$
would also be very small

Moral: Significance and size of effect are both important. When reporting the effect of coefficients

1. Check variable is statistically significant from zero
2. If it is then (and only then) discuss the size of the effect as implied by the regression coefficient

