Econometrics
Chapter 5: Hypothesis Test and CI's in Simple OLS

In the class size example we estimated a β_1 - the effect of class size on test scores.

Say some tax payers are mad that the city is wasting money because they claim class size is not related to test scores ($\beta_1 = 0$), can we reject this assertion with our data?

Testing Hypothesis about Regression Coefficients

Testing is based on the sampling distribution of our estimator.

We know our estimator is (asymptotically) distributed normal (by the CLT).

We then 'standardized' our estimator by subtracting off the *hypothesised mean* (centering) and dividing by the standard error of the estimator - this gave us a random variable that was distributed 'standard normal'

This is our t-statistic, and this is used for testing and to compute our p-values.

$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}}$

And if doing two sided test then:

p-value $= 2\Phi(-|t^{act}|)$ (recall Φ is the notation for the CDF of the standard normal distribution).

So the only thing we are missing to do this here for our $\hat{\beta}_1$ is the standard error.
Remember the SE is an *estimate* of the standard deviation of the *sampling distribution* of $\hat{\beta}_1$.

$SE(\hat{beta_1}) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}}$

$\hat{\sigma}_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum(X_i - \overline{X})^2 \hat{u}_i^2}{[\frac{1}{n}\sum(X_i - \overline{X})^2]^2}$

Now this looks complicated, but the standard errors will be computed by the programs you use and given to you (R).
So in practice you do not need to know this equation, just know that you need to get the SE to get the t-stat to get the p-value.

Recall what is the p-value

The probability of observing a value of $\hat{\beta}$ at least as different from the null hypothesis ($\beta_{1,0}$) as your estimate ($\hat{\beta}^{act}$) *assuming the null is true*:

$$p - value = Pr_{H_o}[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \beta_{1,0}|] = Pr_{H_o}[|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}| > |\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}|] = Pr_{H_0}(|t| > |t^{act}|)$$

and again, since we know that t is distributed standard normal under the null hypothesis,

p-value $= Pr(|Z| > |t^{act}|) = 2\Phi(-|t^{act}|)$

And recall standard is to reject the null if p is less than 5%, which in large samples corresponds to a t-value of 1.96.

Software typically computes t's and p's based on the null being 0

t-stat and P-value for test scores

$\hat{\beta}_1 = 2.28$ with $SE(\hat{\beta}_1) = 0.52$.

Test a null of $\beta_1 = 0$

What is the t-stat? What is the p-value? Can you ***draw*** what the p-value is? Do you reject the null? What does this mean?

t-stat = -4.38

p-value of - so small not even reported

The area of both tails beyond +/- 4.38 on standard normal. So we can pretty surely reject the null!

When reporting results one typically reports the Est. and the SE in parenthesis below the est.

Many times one will 'star' (*) estimates according to significance levels - one for 10%, two for 5%, three for 1%.

Rarely do you see one sided tests, and we normally don't care about the intercept. But all what we did above is easy to do with the intercept and with one-sided tests if you wish.

Confidence Intervals (CIs) for Regression Coefficients

CI's give us more info than a simple test

What is our 95% confidence interval?:
Region of null hypotheses that cannot be rejected based on a 5% test

If I have a CI of (3,8), it is *not correct* to say you are 95% sure the true value is between 3 and 8

95% CI for $\beta_1 = [\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)]$

Recall that 1.96 is the critical value for a 5% test - the t-stat that would just give me a p-value of 5%.

CI for test scores

$\hat{\beta}_1 = 2.28$ with $SE(\hat{\beta}_1) = 0.52$.

What is the 95% CI? What about 90%?

CI(95%) = [-3.30, -1.26]

CI(90%) =

Regression When X is Binary

What if the regressor (X) only takes a 0 or 1 value: sex, urban or rural, foreign born etc.

We call these 'binary' or 'indicator' or 'dummy' variables.

All the mechanics described above hold and can do all the same.

Just interpretation is different (not really a 'slope' of a line) - it is equivalent to performing a difference of means test (Chapter 3).

Note we can write the this as two equations, one for each value of X:

$Y_i(D_i = 0) = \beta_0 + u_i$
$Y_i(D_i = 1) = \beta_0 + \beta_1 + u_i$

and so β_1 is the difference in the two groups. And a test against a null of 0 is a test for a difference in means (if β has any effect).

But otherwise all the above holds just the same.

Heteroskedasticity and Homoskedasticity

All we have assumed (in our assumptions) about the errors (u) are is $E(u|X) = 0$.

IF the variance of the u's does not depend on X ($E(u^2|X) = \sigma^2$), they are *homoskedastic*. This is a special case.

What if the spread of the distribution of u widens as X increases, then the u's are *heterskedastic*.

Example of male/female wages: $Wage_i = \beta_0 + \beta_1 Male + u_i$

so:
$Wage_i(women) = \beta_0 + u_i$
$Wage_i(man) = \beta_0 + \beta_1 + u_i$

Asking if the errors are homoskedastic is the same as asking if the variance of earnings is the same for men and women - in most countries they are not.

Regardless, the OLS est.'s are (under assumptions 1-3) unbiased, consistent, and asymptotically normally distributed.

However, if the errors are also homoskedastic, then the OLS estimators are also the most efficient among linear (in Y) unbiased estimators. (Gauss-Markov theorem). This is why sometimes you may hear that the OLS estimator is 'BLUE' - but this is a special case.

If the errors are also homoskedastic then the formulas for the SE's we saw earlier simplify, but they can ONLY be used in this case (homoskedastic only SEs).

If one uses the simplified homoskedastic only SE formula, and the errors are heteroskedastic, then the computed t value will not be distributed normal and testing/CI's will not be valid.

In general should default to the general case, and since the software will give you the answer anyway, who cares.

But note that for historical reasons most software reports as default the homoskedastic-only SE's and one must specifically ask for the heteroskedastic-robust SE's.