LECTURE 5

Introduction to Econometrics

Hypothesis testing

October 18, 2016

On today's lecture

- We are going to discuss how hypotheses about coefficients can be tested in regression models
- We will explain what significance of coefficients means
- We will learn how to read regression output
- Readings for this week:
  - Studenmund, Chapter 5.1 - 5.4
  - Wooldridge, Chapter 4

Questions we ask

- What conclusions can we draw from our regression?
- What can we learn about the real world from a sample?
- Is it likely that our results could have been obtained by chance?
- If our theory is correct, what are the odds that this particular outcome would have been observed?

Hypothesis testing

- We cannot prove that a given hypothesis is “correct” using hypothesis testing
- All that can be done is to state that a particular sample conforms to a particular hypothesis
- We can often reject a given hypothesis with a certain degree of confidence
- In such a case, we conclude that it is very unlikely the sample result would have been observed if the hypothesized theory were correct

Null and Alternative Hypotheses

- First step in hypothesis testing: state explicitly the hypothesis to be tested
- *Null hypothesis*: statement of the range of values of the regression coefficient that would be expected to occur if the researcher's theory were *not* correct
- *Alternative hypothesis*: specification of the range of values of the coefficient that would be expected to occur if the researcher's theory were correct
- In other words: we define the null hypothesis as the result we do not expect

Null and Alternative Hypotheses

- Notation:
  - H_0 ... null hypothesis
  - H_A ... alternative hypothesis
- Examples:
  - *One-sided test*

    $$\begin{aligned} H_0 : \quad \beta &\leq 0 \\ H_A : \quad \beta &> 0 \end{aligned}$$

  - *Two-sided test*

    $$\begin{aligned} H_0 : \quad \beta &= 0 \\ H_A : \quad \beta &\neq 0 \end{aligned}$$

Type I and type II errors

- It would be unrealistic to think that conclusions drawn from regression analysis will always be right

- There are two types of errors we can make
  - Type I : We reject a true null hypothesis
  - Type II : We do not reject a false null hypothesis

- Example:
  - $H_0 : \beta = 0$
  - $H_A : \beta \neq 0$

  - Type I error: it holds that $\beta = 0$, we conclude that $\beta \neq 0$
  - Type II error: it holds that $\beta \neq 0$, we conclude that $\beta = 0$

Type I and type II errors

- Example:
  - H_0 : The defendant is innocent
  - H_A : The defendant is guilty

  - Type I error = Sending an innocent person to jail
  - Type II error = Freeing a guilty person
- Obviously, lowering the probability of Type I error means increasing the probability of Type II error
- In hypothesis testing, we focus on Type I error and we ensure that its probability is not unreasonably large

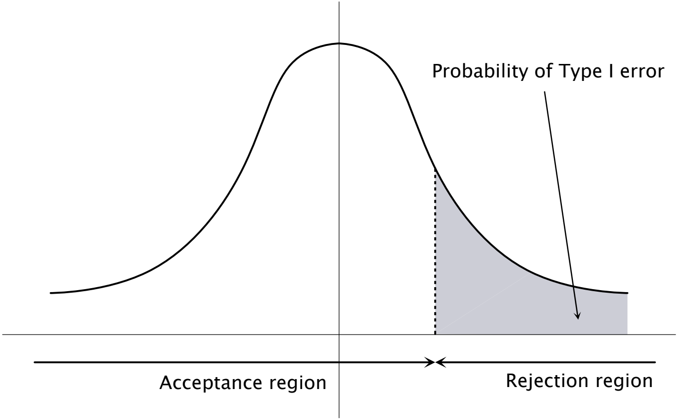
Decision rule

1. Calculate sample statistic
2. Compare sample statistic with the *critical value* (from the statistical tables)

- The critical value divides the range of possible values of the statistic into two regions: *acceptance region* and *rejection region*
  - If the sample statistic falls into the rejection region, we reject H_0
  - If the sample statistic falls into the acceptance region, we do not reject H_0
- The idea is that if the value of the coefficient does not support H_0, the sample statistic should fall into the rejection region

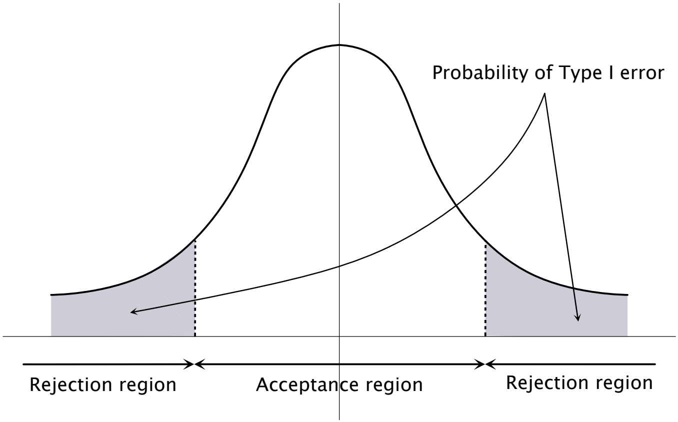
One-sided rejection region

- $H_0: \quad \beta \leq 0 \qquad \text{vs} \qquad H_A: \quad \beta > 0$
- Distribution of $\widehat{\beta}$:



Two-sided rejection region

- $H_0: \quad \beta = 0 \qquad \text{vs} \qquad H_A: \quad \beta \neq 0$
- Distribution of $\widehat{\beta}$:



THE t-TEST

- We use t-test to test hypothesis about individual regression slope coefficients
- Test of more than one coefficient at a time (joint hypotheses) are typically done with the F-test (see next lecture)
- The t-test is appropriate to use when the stochastic error term is normally distributed and when the variance of that distribution is unknown
  - These are the usual assumptions in regression analyses
- The t-test accounts for differences in the units of measurement of the variables

THE t-TEST

- Consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Suppose we want to test (b is some constant)

$$H_0: \ \beta_1 = b \quad \text{vs} \quad H_A: \ \beta_1 \neq b$$

- We know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, Var(\widehat{\beta}_1)\right) \quad \Rightarrow \quad \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{Var(\widehat{\beta}_1)}} \sim N(0,1)$$

The t-test

- Problem: $Var(\widehat{\beta}_1)$ depends on the variance of error term σ^2, which is unobservable and therefore unknown
- It has to be estimated as

$$\hat{\sigma}^2 := s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} \quad ,$$

  k is the number of regression coefficients (here $k = 3$)
  $\mathbf{e}$ is the vector of residuals

- We denote *standard error* of $\widehat{\beta}_1$ (sample counterpart of standard deviation $\sigma_{\widehat{\beta}_1}$) as $s.e.\left(\widehat{\beta}_1\right)$

The t-test

- We define the t-statistic

$$t \quad := \quad \frac{\widehat{\beta}_1 - \beta_1}{s.e.\left(\widehat{\beta}_1\right)} \quad \sim t_{n-k}$$

  where $\widehat{\beta}_1$ is the estimated coefficient and β_1 is the value of the coefficient that is stated in our hypothesis
- This statistic depends only on the estimate $\widehat{\beta}_1$, our hypothesis about β_1, and it has a known distribution

TWO-SIDED t-TEST

- Our hypothesis is

$$H_0: \quad \beta_1 = b \qquad \text{vs} \qquad H_A: \quad \beta_1 \neq b$$

- Hence, our t-statistic is

$$t = \frac{\widehat{\beta}_1 - b}{s.e.\left(\widehat{\beta}_1\right)}$$

  - where $\widehat{\beta}_1$ is the estimated regression coefficient of β_1
  - b is the constant from our null hypothesis
  - $s.e.\left(\widehat{\beta}_1\right)$ is the estimated standard error of $\widehat{\beta}_1$

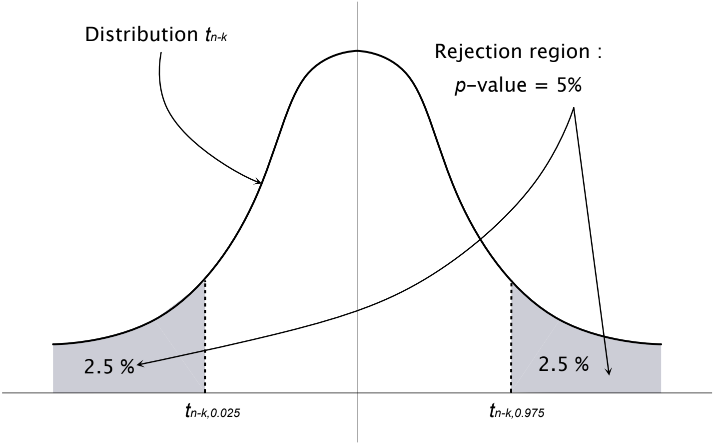
Two-sided t-test

How to determine the *critical value* for this test statistic?

- The critical value is the value that distinguishes the acceptance region from the rejection region

1. We set the probability of Type I error
   - Let's set the Type I. error to 5%
   - We say the p-value of the test is 5% or that we have a test at 95% confidence level
2. We find the critical values in the statistical tables: $t_{n-k,0.975}$ and $t_{n-k,0.025}$
   - The critical value depends on the chosen level of Type I error and $n-k$
   - Note that $t_{n-k,0.975} = -t_{n-k,0.025}$

TWO-SIDED t-TEST



- We reject H_0 if $|t| > t_{n-k,0.975}$

One-sided t-test

- Suppose our hypothesis is

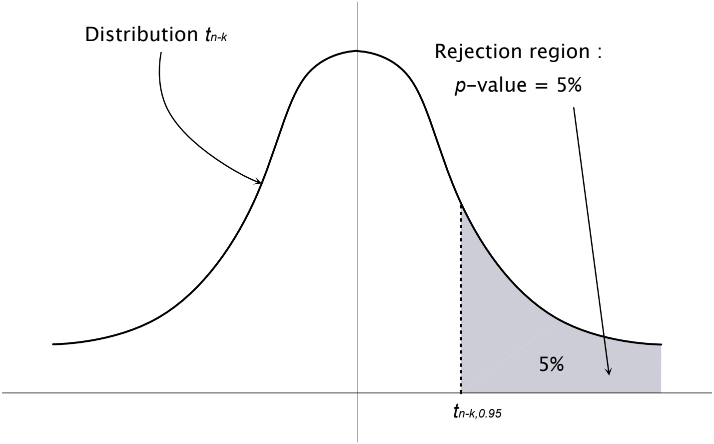
$$H_0: \quad \beta_1 \leq b \qquad \text{vs} \qquad H_A: \quad \beta_1 > b$$

- Our t-statistic still is

$$t = \frac{\widehat{\beta}_1 - b}{s.e.\left(\widehat{\beta}_1\right)}$$

- We set the probability of Type I error to 5%
- We compare our statistic to the critical value $t_{n-k,0.95}$

ONE-SIDED t-TEST



- We reject H_0 if $t > t_{n-k,0.95}$

Significance of the coefficient

- The most common test performed in regression is

$$H_0 : \quad \beta = 0 \qquad \text{vs} \qquad H_A : \quad \beta \neq 0$$

  with the t-statistic

$$t = \frac{\widehat{\beta}}{s.e.\left(\widehat{\beta}\right)} \sim t_{n-k}$$

- If we reject $H_0 : \quad \beta = 0$, we say the coefficient β is *significant*

- This t-statistic is displayed in most regression outputs

THE p-VALUE

- Classical approach to hypothesis testing: first choose the significance level, then test the hypothesis at the given level of significance (e.g. 5%)
  - However, there is no "correct" significance level.
- Or we can ask a more informative question:
  - **What is the smallest significance level at which the null hypothesis would still be rejected?**
  - This level of significance is known as the p-value.
  - Remember that the significance level describes the probability of type I. error. The smaller the p-value, the smaller the probability of rejecting the true null hypothesis (the bigger the confidence the null hypothesis is indeed correctly rejected).
  - The p-value for $H_0 : \beta = 0$ is displayed in most regression outputs

Example

- Let us study the impact of years of education on wages:

$$wage = \beta_0 + \beta_1 education + \beta_2 experience + \varepsilon$$

- Output from Gretl:

```
Model 3: OLS, using observations 1-526
Dependent variable: wage

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const      -3.39054      0.766566     -4.423    1.18e-05 ***
  educ        0.644272     0.0538061    11.97     2.28e-29 ***
  exper       0.0700954    0.0109776     6.385    3.78e-10 ***

Mean dependent var   5.896103   S.D. dependent var   3.693086
Sum squared resid    5548.160   S.E. of regression   3.257044
R-squared            0.225162   Adjusted R-squared   0.222199
F(2, 523)            75.98998   P-value(F)           1.07e-29
Log-likelihood      -1365.969   Akaike criterion     2737.937
Schwarz criterion    2750.733   Hannan-Quinn         2742.948
```

Confidence interval

- A 95% confidence interval of β is an interval centered around $\widehat{\beta}$ such that β falls into this interval with probability 95%

$$P\left(\widehat{\beta} - c < \beta < \widehat{\beta} + c\right) =$$

$$= P\left(-\frac{c}{s.e.\left(\widehat{\beta}\right)} < \frac{\widehat{\beta} - \beta}{s.e.\left(\widehat{\beta}\right)} < \frac{c}{s.e.\left(\widehat{\beta}\right)}\right) = 0.95$$

- Since $\frac{\widehat{\beta}-\beta}{s.e.\left(\widehat{\beta}\right)} \sim t_{n-k}$, we derive the confidence interval:

$$\widehat{\beta} \pm t_{n-k,0.975} \cdot s.e.\left(\widehat{\beta}\right)$$

Confidence interval

- Output from Gretl (wage regression):

```
Model 3: OLS, using observations 1-526
Dependent variable: wage

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const      −3.39054      0.766566     −4.423    1.18e-05 ***
  educ        0.644272     0.0538061    11.97     2.28e-29 ***
  exper       0.0700954    0.0109776     6.385    3.78e-10 ***

Mean dependent var   5.896103   S.D. dependent var   3.693086
Sum squared resid    5548.160   S.E. of regression   3.257044
R-squared            0.225162   Adjusted R-squared   0.222199
F(2, 523)            75.98998   P-value(F)           1.07e-29
Log-likelihood      −1365.969   Akaike criterion     2737.937
Schwarz criterion    2750.733   Hannan-Quinn         2742.948
```

- Confidence interval for coefficient on education: $\widehat{\beta} \pm t_{n-k,0.975} \cdot s.e.\left(\widehat{\beta}\right) = 0.644 \pm 1.960 \cdot 0.054$
- $\widehat{\beta} \in [0.538; 0.750]$ with 95% probability

SUMMARY

- We discussed the principle of hypothesis testing
- We derived the t-statistic
- We defined the concept of the p-value
- We explained what significance of a coefficient means
- We observed a regression output on an example