LECTURE 6

Introduction to Econometrics

Hypothesis testing & Goodness of fit

October 25, 2016

On today's lecture

- We will explain how multiple hypotheses are tested in a regression model
- We will define the notion of the overall significance of a regression
- We will introduce a measure of the goodness of fit of a regression (R^2)
- Readings for this week:
  - Studenmund, Chapters 5.5 & 2.4
  - Wooldridge, Chapters 4 & 3

Testing multiple hypotheses

- Suppose we have a model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- Suppose we want to test multiple linear hypotheses in this model
- For example, we want to see if the following restrictions on coefficients hold jointly:

$$\beta_1 + \beta_2 = 1 \qquad \text{and} \qquad \beta_3 = 0$$

- We cannot use a t-test in this case (t-test can be used only for one hypothesis at a time)
- We will use an F-test

Restricted vs. unrestricted model

- We can reformulate the model by plugging the restrictions as if they were true (model under H_0)
- We call this model *restricted model* as opposed to the *unrestricted model*
- The unrestricted model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- We derive (on the lecture) the restricted model:

$$y_i^* = \beta_0 + \beta_1 x_i^* + \varepsilon_i \quad ,$$

where $y_i^* = y_i - x_{i2}$ and $x_i^* = x_{i1} - x_{i2}$

Idea of the F-test

- If the restrictions are true, then the restricted model fits the data in the same way as the unrestricted model
  - residuals are nearly the same
- If the restrictions are false, then the restricted model fits the data poorly
  - residuals from the restricted model are much larger than those from the unrestricted model
- The idea is thus to compare the residuals from the two models

Idea of the F-test

- How to compare residuals in the two models?
  - Calculate the sum of squared residuals in the two models
  - Test if the difference between the two sums is equal to zero (statistically)
  - H_0: the difference is zero (residuals in the two models are the same, restrictions hold)
  - H_A: the difference is positive (residuals in the restricted model are bigger, restrictions do not hold)

- Sum of squared residuals
  - $SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

F-TEST

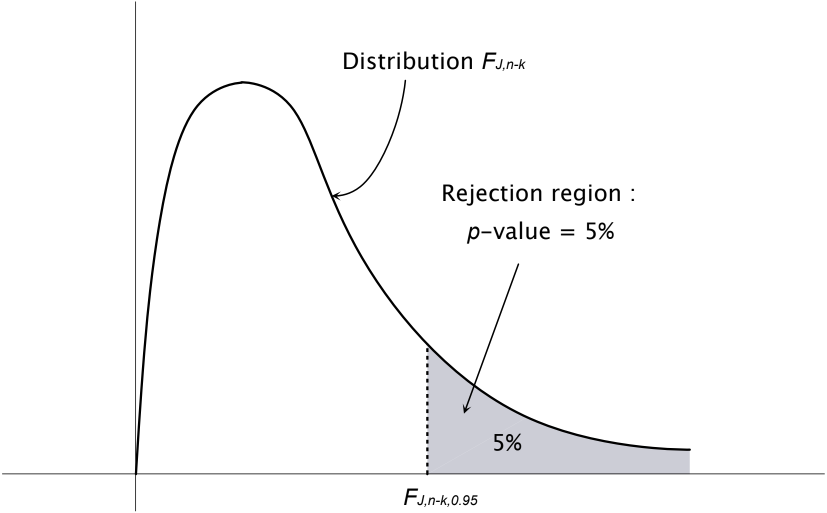
- The test statistic is defined as

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)} \sim F_{J,n-k} \ ,$$

where:

SSE_R	...	sum of squared residuals from the restricted model
SSE_U	...	sum of squared residuals from the unrestricted model
J	...	number of restrictions
n	...	number of observations
k	...	number of estimated coefficients (including intercept)

F-TEST



EXAMPLE

- We had the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- We wanted to test

$$H_0 : \begin{cases} \beta_1 + \beta_2 = 1 \\ \beta_3 = 0 \end{cases} \quad \text{vs.} \quad H_A : \begin{cases} \beta_1 + \beta_2 \neq 1 \\ \beta_3 \neq 0 \end{cases}$$

- Under H_0, we obtained the restricted model

$$y_i^* = \beta_0 + \beta_1 x_i^* + \varepsilon_i \quad ,$$

where $y_i^* = y_i - x_{i2}$ and $x_i^* = x_{i1} - x_{i2}$

Example

- We run the regression on the unrestricted model, we obtain SSE_U
- We run the regression on the restricted model, we obtain SSE_R
- We have $k = 4$ and $J = 2$
- We construct the F-statistic $F = \frac{(SSE_R - SSE_U)/2}{SSE_U/(n-4)}$
- We find the critical value of the F distribution with 2 and $n-4$ degrees of freedom at the 95% confidence level
- If $F > F_{2,n-4,0.95}$, we reject the null hypothesis
  - we reject that the restrictions hold jointly

Overall significance of the regression

- Usually, we are interested in knowing if the model has some explanatory power, i.e. if the independent variables indeed "explain" the dependent variable

- We test this using the F-test of the joint significance of all $(k-1)$ slope coefficients:

$$H_0 : \begin{cases} \beta_1 = 0 \\ \beta_2 = 0 \\ \vdots \\ \beta_{k-1} = 0 \end{cases} \quad \text{vs.} \quad H_A : \begin{cases} \beta_j \neq 0 \\ \\ \text{for at least one } j = 1, \ldots, k-1 \end{cases}$$

Overall significance of the regression

- Unrestricted model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{k-1} x_{ik-1} + \varepsilon_i$$

- Restricted model:

$$y_i = \beta_0 + \varepsilon_i$$

- F-statistic:

$$F = \frac{(SSE_R - SSE_U)/(k-1)}{SSE_U/(n-k)} \sim F_{k-1,n-k}$$

  - Number of restrictions = $k - 1$
  - This F-statistic and the corresponding p-value are part of the regression output

EXAMPLE

```
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
```

GOODNESS OF FIT MEASURE

- We know that education and experience have a significant influence on wages
- But how important are they in determining wages?
- How much of difference in wages between people is explained by differences in education and in experience?
- How well variation in the independent variable(s) explains variation in the dependent variable?
- This are the questions answered by the goodness of fit measure - R^2

Total and explained variation

- **Total variation** in the dependent variable:

$$\sum_{i=1}^{n} (y_i - \overline{y}_n)^2$$

- Predicted value of the dependent variable = part that is explained by independent variables:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

(case of regression line - for simplicity of notation)

- **Explained variation** in the dependent variable:

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2$$

Goodness of fit - R^2

- Denote:
  - $SST = \sum_{i=1}^{n} \left(y_i - \overline{y}_n\right)^2$ ... *Total Sum of Squares*
  - $SSR = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2$ ... *Regression Sum of Squares*

- Define the measure of the goodness of fit:

$$R^2 = \frac{SSR}{SST} = \frac{\text{Explained variation in } y}{\text{Total variation in } y}$$

GOODNESS OF FIT - R^2

- In all models: $0 \leq R^2 \leq 1$
- R^2 tells us what percentage of the total variation in the dependent variable is explained by the variation in the independent variable(s)
  - $R^2 = 0.3$ means that the independent variables can explain 30% of the variation in the dependent variable
- Higher R^2 means better fit of the regression model (not necessarily a better model!)

Decomposing the variance

- For models with intercept, R^2 can be rewritten using the decomposition of variance.
- Variance decomposition:

$$\sum_{i=1}^{n} \left(y_i - \overline{y}_n\right)^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2 + \sum_{i=1}^{n} e_i^2$$

- $SST = \sum_{i=1}^{n} \left(y_i - \overline{y}_n\right)^2$ ... *Total Sum of Squares*
- $SSR = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2$ ... *Regression Sum of Squares*
- $SSE = \sum_{i=1}^{n} e_i^2$ ... *Sum of Squared Residuals*

Variance decomposition and R^2

- Variance decomposition: $SST = SSR + SSE$
- Intuition: total variation can be divided between the explained variation and the unexplained variation
  - the true value y is a sum of estimated (explained) $\widehat{y}$ and the residual e_i (unexplained part)
  - $y_i = \widehat{y}_i + e_i$
- We can rewrite R^2:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

ADJUSTED R^2

- The sum of squared residuals (SSE) decreases when additional explanatory variables are introduced in the model, whereas total sum of squares (SST) remains the same
  - $R^2 = 1 - \frac{SSE}{SST}$ increases if we add explanatory variables
  - Models with more variables automatically have better fit.
- To deal with this problem, we define the *adjusted* R^2:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-k}}{\frac{SST}{n-1}} \quad (\leq R^2)$$

(k is the number of coefficients including intercept)

- This measure introduces a "punishment" for including more explanatory variables

EXAMPLE

```
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
```

F-TEST - REVISITED

- Let us recall the F-statistic:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)} \sim F_{J,n-k}$$

- We can use the formula $R^2 = 1 - \frac{SSE}{SST}$ to rewrite the F-statistic in R^2 form:

$$F = \frac{(R_U^2 - R_R^2)/J}{(1 - R_U^2)/(n-k)} \sim F_{J,n-k}$$

- We can use this R^2 form of F-statistic under the condition that $SST_U = SST_R$ (the dependent variables in restricted and unrestricted models are the same)

Summary

- We showed how restrictions are incorporated in regression models
- We explained the idea of the F-test
- We defined the notion of the overall significance of a regression
- We introduced the measure or the goodness of fit - R^2
- We learned how total variation in the dependent variable can be decomposed