LECTURE 5

Introduction to Econometrics

Multiple Regression Analysis: Hypothesis Testing II

Hieu Nguyen

Fall semester, 2024

REVISION: the Classical Linear Model (CLM) Assumptions

1. Linearity: the regression model is linear in the parameters (coefficients)
2. Random sampling: the data is a random sample drawn from the population and each data point follows the population equation
3. No perfect collinearity: the values of explanatory variables are not all the same and no explanatory variable is a perfect linear function of any other explanatory variable(s)
4. Zero conditional mean: values of explanatory variables must contain no information about the mean of the unobserved factors - explanatory variables are uncorrelated with the error term
5. Homoskedasticity: the error term has a constant variance
6. Normality of the error term: the error term is normally distributed

TODAY'S LECTURE

e We continue our discussion on how hypotheses about coefficients can be tested in regression models

e We will explain what significance of coefficients mean

e We will learn how to read regression output

e Readings:

Wooldridge Chapter 4;
Studenmund Chapter 5.1-5.4

INFERENCE: The t Test

e „Statistically significant“ variables in a regression

- If a regression coefficient is different from zero in a two-sided test, the corresponding variable is said to be „statistically significant“
- If the number of degrees of freedom is large enough so that the normal approximation applies, the following rules of thumb apply:

$|t - ratio| > 1.645$ → „statistically significant at 10 % level“

$|t - ratio| > 1.96$ → „statistically significant at 5 % level“

$|t - ratio| > 2.576$ → „statistically significant at 1 % level“

INFERENCE: The t Test

- Guidelines for discussing economic and statistical significance
  - If a variable is statistically significant, discuss the magnitude of the coefficient to get an idea of its economic or practical importance
  - <u>The fact that a coefficient is statistically significant does not necessarily mean it is economically or practically significant!</u>
  - If a variable is statistically and economically important but has the „wrong“ sign, the regression model might be misspecified
  - If a variable is statistically insignificant at the usual levels (10%, 5%, 1%), one may think of dropping it from the regression
  - If the sample size is small, effects might be imprecisely estimated so that the case for dropping insignificant variables is less strong

INFERENCE: The t Test

- Testing more general hypotheses about a regression coefficient
- Null hypothesis

$$H_0 : \ \beta_j = a_j$$

Hypothesized value of the coefficient

- t-statistic

$$t = \frac{(estimate - hypothesized\ value)}{standard\ error} = \frac{(\hat{\beta}_j - a_j)}{se(\hat{\beta}_j)}$$

- The test works exactly as before, except that the hypothesized value is substracted from the estimate when forming the statistic

INFERENCE: The t Test

- **Example: Campus crime and enrollment**
  - An interesting hypothesis is whether crime increases by one percent if enrollment is increased by one percent

$$\widehat{\log}(crime) = \underset{(1.03)}{-\ 6.63} + \underset{(0.11)}{\boxed{1.27}} \log(enroll)$$

Estimate is different from one but is this difference statistically significant?

$$n = 97,\ R^2 = .585$$

$$H_0 : \beta_{\log(enroll)} = 1,\ H_1 : \beta_{\log(enroll)} \neq 1$$

$$t = (1.27 - 1)/.11 \approx 2.45 > 1.96 = c_{0.05}$$

The hypothesis is rejected at the 5% level

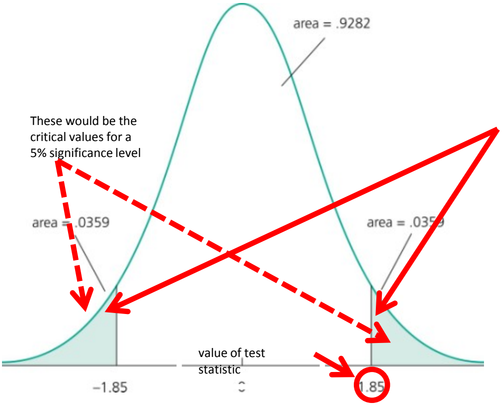
INFERENCE: The t Test

e **Computing p-values for t-tests**

- If the significance level is made smaller and smaller, there will be a point where the null hypothesis cannot be rejected anymore
- The reason is that, by lowering the significance level, one wants to avoid more and more to make the error of rejecting a correct H_0
- The smallest significance level at which the null hypothesis is still rejected, is called the p-value of the hypothesis test
- A small p-value is evidence against the null hypothesis because one would reject the null hypothesis even at small significance levels
- A large p-value is evidence in favor of the null hypothesis
- P-values are more informative than tests at fixed significance levels

INFERENCE: The t Test

- How the p-value is computed (here: two-sided test)



These would be the critical values for a 5% significance level

value of test statistic

The p-value is the significance level at which one is indifferent between rejecting and not rejecting the null hypothesis.

In the two-sided case, the p-value is thus the probability that the t-distributed variable takes on a larger absolute value than the realized value of the test statistic, e.g.:

$$P(|t - ratio| > 1.85) = 2(.0.359) = .0718$$

From this, it is clear that <u>a null hypothesis is rejected if and only if the corresponding p-value is smaller than the significance level.</u>

For example, for a significance level of 5% the t-statistic would not lie in the rejection region.

INFERENCE: Confidence Intervals

- Confidence intervals
- Simple manipulation of the result in Theorem 4.2 implies that

$$P\Big(\underbrace{\hat{\beta}_j - c_{0.05}\cdot se(\hat{\beta}_j)}_{\text{Lower bound of the Confidence interval}} \leq \beta_j \leq \underbrace{\hat{\beta}_j + \overset{\text{Critical value of two-sided test}}{c_{0.05}}\cdot se(\hat{\beta}_j)}_{\text{Upper bound of the Confidence interval}}\Big) = \underset{\text{Confidence level}}{0.95}$$

- Interpretation of the confidence interval
  - The bounds of the interval are random
  - In repeated samples, the interval that is constructed in the above way will cover the population regression coefficient in 95% of the cases

INFERENCE: Confidence Intervals

e Confidence intervals for typical confidence levels

$$P\left(\widehat{\beta}_j - c_{0.01} \cdot se(\widehat{\beta}_j) \leq \beta_j \leq \widehat{\beta}_j + c_{0.01} \cdot se(\widehat{\beta}_j)\right) = 0.99$$

$$P\left(\widehat{\beta}_j - c_{0.05} \cdot se(\widehat{\beta}_j) \leq \beta_j \leq \widehat{\beta}_j + c_{0.05} \cdot se(\widehat{\beta}_j)\right) = 0.95$$

$$P\left(\widehat{\beta}_j - c_{0.10} \cdot se(\widehat{\beta}_j) \leq \beta_j \leq \widehat{\beta}_j + c_{0.10} \cdot se(\widehat{\beta}_j)\right) = 0.90$$

Use rules of thumb $c_{0.01} = 2.576, c_{0.05} = 1.96, c_{0.10} = 1.645$

e Relationship between confidence intervals and hypotheses tests

$a_j \notin interval \;\Rightarrow\;$ reject $H_0 : \beta_j = a_j$ in favor of $H_1 : \beta_j \neq 0$

INFERENCE: Confidence Intervals

- e Example: Model of firms‘ R&D expenditures

Spending on R&D

Annual sales

Profits as percentage of sales

$$\widehat{\log}(rd) = \underset{(.47)}{-\ 4.38} + \underset{(.060)}{1.084}\ \log(sales) + \underset{(0.0128\ )}{0.0217}\ profmarg$$

$$n = 32,\ R^2 = .918,\ df = 32-2-1 = 29 \Rightarrow c_{0.05} = 2.045$$

$$1.084 \pm 2.045(.060) = \boxed{(.961, 1.21)}$$

$$.0217 \pm 2.045\ (0.0128\ ) = \boxed{(-.0045, .0479)}$$

The effect of sales on R&D is relatively precisely estimated as the interval is narrow. Moreover, the effect is significantly different from zero because zero is outside the interval.

This effect is imprecisely estimated as the interval is very wide. It is not even statistically significant because zero lies in the interval.

INFERENCE: Testing Hypotheses About a Linear Combination of Parameters

e Example: Return to education at 2 year vs. at 4 year colleges

Years of education at 2 year colleges

Years of education at 4 year colleges

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

Test $H_0 : \beta_1 - \beta_2 = 0$ against $H_1 : \beta_1 - \beta_2 < 0$.

A possible test statistic would be:

$$t = \frac{\boxed{\widehat{\beta}_1 - \widehat{\beta}_2}}{se(\widehat{\beta}_1 - \widehat{\beta}_2)}$$

The difference between the estimates is normalized by the estimated standard deviation of the difference. The null hypothesis would have to be rejected if the statistic is „too negative“ to believe that the true difference between the parameters is equal to zero.

INFERENCE: Testing Hypotheses About a Linear Combination of Parameters

e Impossible to compute with standard regression output because

$$se(\hat{\beta}_1-\hat{\beta}_2) = \sqrt{\widehat{Var}(\hat{\beta}_1-\hat{\beta}_2)} = \sqrt{\widehat{Var}(\hat{\beta}_1)+\widehat{Var}(\hat{\beta}_2)-2\widehat{Cov}(\hat{\beta}_1,\hat{\beta}_2)}$$

Usually not available in regression output

e Alternative method

Define $\theta_1 = \beta_1 - \beta_2$ and test $H_0 : \theta_1 = 0$ against $H_1 : \theta_1 < 0$.

$$\log(wage) = \beta_0 + (\theta_1 + \beta_2) jc + \beta_2 univ + \beta_3 exper + u$$

$$= \beta_0 + \theta_1 jc + \beta_2 \boxed{(jc + univ)} + \beta_3 exper + u$$

Insert into original regression

a new regressor (= total years of college)

INFERENCE: Testing Hypotheses About a Linear Combination of Parameters

- Estimation results

Total years of college

$$\widehat{\log}(wage) = \underset{(.021)}{1.472} - \underset{(.0069)}{.0102}\, jc + \underset{(.0023)}{.0769}\, totcoll + \underset{(.0002)}{.0049}\, exper$$

$$n = 6,763, \;\; R^2 = .222$$

Hypothesis is rejected at 10% level but not at 5% level

$$t = -.0102/.0069 = -1.48$$

$$p - value = P(t - ratio < -1.48) = .070$$

$$-.0102 \pm 1.96(.0069) = (-.0237, .0003)$$

- This method works <u>always</u> for single linear hypotheses

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

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

- Restricted model can be derived to have the following form:

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

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

$$SSR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

F-TEST

- The test statistic is defined as

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1} \quad ,$$

where:

SSR_r	…	sum of squared residuals from the restricted model
SSR_{ur}	…	sum of squared residuals from the unrestricted model
q	…	number of restrictions
n	…	number of observations
k	…	number of estimated coefficients

INFERENCE: The F Test

e Testing multiple linear restrictions: The F-test

e Testing exclusion restrictions

Salary of major league baseball player

Years in the league

Average number of games per year

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr$$

$$+\beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

Batting average

Home runs per year

Runs batted in per year

$\boxed{H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0}$ against $H_1 : H_0$ is not true

Test whether performance measures have no effect/can be exluded from regression.

INFERENCE: The F Test

e Estimation of the unrestricted model

$$\widehat{\log}(salary) = \underset{(0.29)}{11.19} + \underset{(.0121)}{.0689}\ years + \underset{(.0026)}{.0126}\ gamesyr$$

$$+ \underset{(.00110)}{.00098}\ \boxed{bavg} + \underset{(.0161)}{.0144}\ \boxed{hrunsyr} + \underset{(.0072)}{.0108}\ \boxed{rbisyr}$$

None of these variabels are statistically significant when tested individually

$$n = 353,\ SSR = \boxed{183.186},\ R^2 = .6278$$

<u>Idea:</u> How would the model fit be if these variables were dropped from the regression?

INFERENCE: The F Test

e Estimation of the restricted model

$$\begin{array}{lllll}\widehat{\log}(salary) = & 11.22 & + .0713\ years & + & .0202\ gamesyr \\ & (0.11) & (.0125) & & (.0013)\end{array}$$

$$n = 353,\ SSR = 198.311,\ R^2 = .5971$$

The sum of squared residuals necessarily increases, but is the increase statistically significant?

e Test statistic

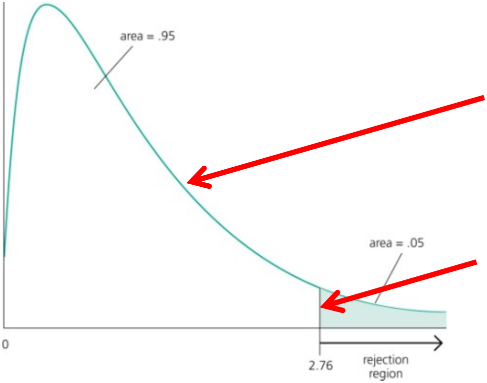
Number of restrictions

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

The relative increase of the sum of squared residuals when going from H_1 to H_0 follows a F-distribution (if the null hypothesis H_0 is correct)

INFERENCE: The F Test

e Rejection rule



A F-distributed variable only takes on positive values. This corresponds to the fact that the sum of squared residuals can only increase if one moves from H_1 to H_0.

Choose the critical value so that the null hypothesis is rejected in, for example, 5% of the cases, although it is true.

INFERENCE: The F Test

e Test decision in example

Number of restrictions to be tested

$$F = \frac{(198.311 - 183.186)/\enclose{circle}{3}}{183.186/\boxed{(353 - 5 - 1)}} \approx 9.55$$

Degrees of freedom in the unrestricted model

$$F \sim F_{3,347} \;\Rightarrow\; c_{0.01} = 3.78$$

$$P(F - statistic > 9.55) = 0.000$$

The null hypothesis is overwhelmingly rejected (even at very small significance levels).

e Discussion

- The three variables are „jointly significant“
- They were not significant when tested individually
- The likely reason is multicollinearity between them

INFERENCE: The F Test

e Test of overall significance of a regression

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u$$

$$\boxed{H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0}$$

The null hypothesis states that the explanatory variables are not useful at all in explaining the dependent variable

$$y = \beta_0 + u$$

Restricted model
(regression on constant)

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

e The test of overall significance is reported in most regression packages; the null hypothesis is usually overwhelmingly rejected

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:
  - $SST = \sum_{i=1}^{n} \left(y_i - \overline{y}_n\right)^2 \ldots$ *Total Sum of Squares*
  - $SSE = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2 \ \ldots$ *Regression (Explained) Sum of Squares*
- Define the measure of the goodness of fit:

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

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:

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

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

VARIANCE DECOMPOSITION AND R^2

- Variance decomposition: $SST = SSE + SSR$

- Intuition: total variation can be divided between the explained variation and the unexplained variation

  the true value y is a sum of estimated (explained) $\hat{y}$ and the

  residual e_i (unexplained part)

  $y_i = \widehat{y}_i + e_i$

- We can rewrite R^2:

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

ADJUSTED R^2

- The sum of squared residuals (SSR) decreases when additional explanatory variables are introduced in the model, whereas total sum of squares (SST) remains the same

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

$$R^2_{adj} = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}} \quad \cdot \leq R^2$$

(k is the number of coefficients)

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

FOUR IMPORTANT SPECIFICATION CRITERIA

Does a variable belong to the equation?

1. *Theory:* Is the variable's place in the equation unambiguous and theoretically sound? Does intuition tells you it should be included?

2. *t-test:* Is the variable's estimated coefficient significant in the expected direction?

3. R^2*:* Does the overall fit of the equation improve (enough) when the variable is added to the equation?

4. *Bias:* Do other variables' coefficients change significantly when the variable is added to the equation?

FOUR IMPORTANT SPECIFICATION CRITERIA

- If all conditions hold, the variable belongs in the equation

- If none of them holds, the variable is irrelevant and can be safely excluded

- If the criteria give contradictory answers, most importance should be attributed to theoretical justification

  Therefore, if theory (intuition) says that variable belongs to the equation, we include it (even though its coefficients might be insignificant!).