NEW SZÉCHENYI PLAN

ECONOMIC STATISTICS

Sponsored by a Grant TÁMOP-4.1.2-08/2/A/KMR-2009-0041

Course Material Developed by Department of Economics,

Faculty of Social Sciences, Eötvös Loránd University Budapest (ELTE)

Department of Economics, Eötvös Loránd University Budapest

Institute of Economics, Hungarian Academy of Sciences

Balassi Kiadó, Budapest

ELTE ECON

ELTE Társadalomtudományi Kar

KTI IE

BALASSI KIADÓ

HUNGARY'S RENEWAL

The project is supported by the European Union.

Author: Anikó Bíró

Supervised by Anikó Bíró

June 2010

Week 7

Omitted variables, multicollinearity, binary regressors – introduction

Simple vs. multivariate

Example: housing prices (CAD, source: hprice.xls)

- Multivariate:

$$\hat{P} = -4009.6 + 5.4\text{lot} + 2824.6\text{bedroom} + \\ +17105.2\text{bathroom} + 7634.9\text{stories}$$

- Univariate: $\hat{P} = 32794.0 + 27477.0\text{bathroom}$
- Larger estimated coefficient

Example, cont.

Explanation for different coefficients:

- Influence of several factors

- Correlation with the number of bathrooms
  - E.g. positive correlation between lot size – number of bathroom
- Univariate regression: cannot separate the effects

Omitted variables

- Bias due to omitted variables:

  Estimation is not correct if we omit such a variable which is correlated with the included explanatory variables
- Include those variables which have explanatory power!
- But: redundant variables – estimation precision decreases
  - General practice: omit the insignificant ones

Wage tariff example

- Simple:

	Coeff.	*Standard dev.*	*t stat.*	*P-value*
Intercept	-161796,32	9514,04	-17,01	0,00
Education	24855,33	707,51	35,13	0,00

- Multiple, corr(educ.,age)=-0.04

	Coeff.	Standard dev.	t stat.	P-value
Intercept	-328321,34	8040,13	-40,84	0,00
Education	27250,22	452,97	60,16	0,00

Age	3171,29	109,05	29,08	0,00

Multicollinearity

- Some of the explanatory variables are strongly correlated
- The effects of the regressors are difficult to separate

- Solution: omit some of the regressors – not always desirable!

- ”Symptoms”:
  - Low t-, high P-values
  - At the same time, R-squared is high
  - Coefficients are very sensitive to the inclusion of additional (collinear) variables
  - Estimated coefficients are very different from the expected values (clearly unreasonable coefficients)

Multicollinearity - example

Earnings regressions, corr(age, experience)=0,97

r-squared	0,468			
	Coeff.	*Standard dev.*	*t stat.*	*P-value*
Intercept	-1,7E+11	3,05E+10	-5,647	1,72E-08
Education	-2,9E+10	5,08E+09	-5,647	1,72E-08
Age	2,87E+10	5,08E+09	5,647	1,72E-08
Experience	-2,9E+10	5,08E+09	-5,647	1,72E-08

r-squared	0,465			
	Coeff.	*Standard dev.*	*t stat.*	*P-value*
Intercept	-328321	8040,126	-40,835	0
Education	27250,22	452,9723	60,159	0
Age	3171,293	109,0451	29,082	6,3E-172

Binary explanatory variables

- Qualitative, coding: 0 – 1
- Binary = dummy = dichotomous variable
- Examples:
  - Housing prices: is there garage, air conditioning, …
  - Wages: male – female
  - Medical expenditures: if insured or not
  - Etc.

Estimation, coefficients

- OLS method unchanged, different interpretation of coefficients
- Simple regression:

$$Y = \alpha + \beta D + e$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta} D$$

$$\hat{Y} = \hat{\alpha}, \text{if D} = 0$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}, \text{if D} = 1$$

- Mean of two subgroups

Examples

1. Housing prices

$$\hat{P} = 59\,885 + 25\,996 Cond$$

- Mean price with air conditioning: 85 881 CAD

2. Earnings (Wage tariff 2003 subsample)

$$\hat{W} = 159\,289 + 66\,854 male$$

- Average earnings, males: 226 142 Ft

  Average earnings, females: 159 289 Ft

More binary variables

$$Y_i = \alpha + \beta_1 D_{i1} + \ldots + \beta_k D_{ik} + e_i$$

- Number of groups: 2^k
- Group means: sum of respective coefficients
- Interpretation of coefficients: partial effect

Binary and continuous explanatory variables

- Only binary: different means
- Binary and not binary: different intercept
- Simplest model:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + e_i$$
$$\text{Intercept}: \alpha \text{ or } \alpha + \beta_1$$

Binary regressors – example

Hprice.xls – housing price regression:

	Coeff.	*Standard dev.*	*t stat.*	*P-value*
Intercept	30555,75	2289,991	13,34317	2,59E-35
Air cond.	19268,8	1909,658	10,09018	4,72E-22
Recreation room	7395,032	2462,386	3,003198	0,002795
Basement	6187,162	1945,687	3,179937	0,001557
Lot size	5,433193	0,410367	13,23985	7,35E-35

Wage tariff (gross monthly earnings) example

	Coeff.	*Standard dev.*	*t stat.*	*P-value*
Intercept	159288,68	1823,60	87,35	0,00
Male	66853,52	3249,19	20,58	0,00

	Coeff.	*Standard dev.*	*t stat.*	*P-value*
Intercept	-296984,11	7674,03	-38,70	0,00
Male	24708,10	2547,18	9,70	0,00
Education	29187,63	482,57	60,48	0,00
Experience	3033,58	108,97	27,84	0,00

Summary

- Omitted variables
- Redundant variables
- Multicollinearity
- Binary regressors

Omitted variables, multicollinearity, binary regressors – introduction

Seminar 7

Omitted variables

- Bias due to omitted variables:
  - Estimation is not correct if we omit such a variable which is correlated with the included explanatory variables
- Include those variables which have explanatory power!
- But: redundant variables – estimation precision decreases
- General practice: omit the insignificant ones

Omitted variables – example

Electricity firms (electric.xls), regression of total production cost, logarithmic form

- Coefficients of labor and capital unit cost are insignificant
  - Explanation? Small importance, small variance, …
- How do the coefficients of output and fuel cost change if the other regressors are omitted?

Multicollinearity

- Some of the explanatory variables are strongly correlated
- The effects of the regressors are difficult to separate
- ”Symptoms”:
  - Low t-, high P-values
  - At the same time, R-squared is high
- Solution: omit some of the regressors – not always desirable!

Multicollinearity, example

Textbook example 6.3 (forest.xls)

Binary regressors

$$Y_i = \alpha + \beta_1 D_{i1} + \ldots + \beta_k D_{ik} + e_i$$

- Number of groups: 2^k
- Group means: sum of respective coefficients
- Interpretation of coefficients: partial effect

Binary and continuous explanatory variables

- Only binary: different means
- Binary and not binary: different intercept
- Simplest model:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + e_i$$
$$\text{Intercept}: \alpha \text{ or } \alpha + \beta_1$$

Example 1

Housing prices (hprice.xls)

- Explanatory variables: lot size, air conditioning, recreation room, basement
- Coefficient of lot size (sq. foot)? – Same for all subgroups!
- Coefficients of the binary variables?

Example 2

Earnings regression based on Wage tariff data

- Regressors: male, years of schooling (education), experience

$$\hat{W} = -296\,984 + 24\,708\text{male} + 29\,188\text{educ} + 3\,034\exp$$
$$\hat{W} = 159\,289 + 66\,854\text{male}$$

- Explanation for different estimated coefficient?

Homework 4 (groups)

Estimation of a macroeconomic model (similar to the example in seminar 6) with current data.

Use a cross sectional sample of a group of countries. Analyze the GDP growth averaged over a selected period.

- Specify a multivariate regression model with brief reasoning.
- Estimate the model, interpret the coefficients, analyze their significance.
- Omit a significant variable. Analyze the effect of omission.