ECONOMETRICS

ECONOMETRICS

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

NEW SZÉCHENYI PLAN

HUNGARY'S RENEWAL

The project is supported by the European Union.

National Development Agency
www.ujszechenyiterv.gov.hu
06 40 638 638

HUNGARY'S RENEWAL

The projects have been supported by the European Union.

ELTE Faculty of Social Sciences, Department of Economics

ECONOMETRICS

Authors: Péter Elek, Anikó Bíró

Supervised by: Péter Elek

June 2010

ECONOMETRICS

Week 2.

Simple regression I.

Péter Elek, Anikó Bíró

Plan

Basics, examples

Assumptions of the regression model

Interpretation of the parameters

Estimation methods

- Optimal least squares (OLS)
- Method of moments
- (Maximum likelihood method)

Properties of the estimation, sampling distribution

Introduction

Simple regression

y = sales

x = expenditure on advertising

Multivariate regression

y = wage of employee

$x1$ = education

$x2$ = work experience

$x3$ = living area etc.

Aims

Analyse the effects of such decisions on y which change the x variables

Forecast y with the help of x

Decide if any x has significant effect on y

Simple (linear) regression: basics I.

$y_i = \alpha + \beta x_i + u_i$ (stochastic relationship)

y
- Forecasted variable
- Explained variable
- Dependent variable

x
- Forecasting variable
- Explanatory variable
- Independent variable
- Causal variable

u error term
- Random human reactions cannot be forecasted
- Effect of omitted variables
- Measurement error in y

Simple regression: basics II.

Regression parameters:

- Intercept
- Slope

Origin of regression: Francis Galton

- Relationship between the height of children (y) and of their parents (x)
- $y = m + \rho x$
- $\rho < 1$ found: “regression to the mean”

Assumptions

1. $E(u_i) = 0$
2. u_i, u_j independent for all $i \neq j$
3. x_i, u_j independent for all i, j

   Surely satisfied if x_i variables are not random
4. $Var(u_i) = \sigma^2$ for all i (homoscedasticity)
5. u_i normally distributed for all i: $N(0, \sigma^2)$

Interpretation

Consequence of (1) and (3):exogeneity, i.e.

$E(u_i|\, x_k) = 0$ for all i,k

So $E(y_i|x_i) = \alpha + \beta x_i$

Thus β can be interpreted as partial effect:

$$\beta = \frac{\partial E\left(y_i \mid x\right)}{\partial x_i}$$

Interpretation of α: $\alpha = E(y_i|x_i = 0)$

Estimation I

Optimal least squares (OLS)

$$\min_{\hat{\alpha},\hat{\beta}} Q = \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

Two normal equations:

$$\frac{\partial Q}{\partial \hat{\alpha}} = 0 \Rightarrow \sum_i 2(y_i - \hat{\alpha} - \hat{\beta} x_i)(-1) = 0$$

$$\frac{\partial Q}{\partial \hat{\beta}} = 0 \Rightarrow \sum_i 2(y_i - \hat{\alpha} - \hat{\beta} x_i)(-x_i) = 0$$

Estimation II

Method of moments (MM)

Method of moments: theoretical moments are equalized to observed moments

(e.g. expected value to sample mean, variance to sample variance)

Normal equations (same as before)

$E(u) = 0 \qquad \sum \hat{u}_i = 0$

$cov(u,x) = 0 \qquad \sum x_i \hat{u}_i = 0$

where $\hat{u}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$

System of equations:

$$\sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0$$

Estimation III

Maximum likelihood (ML) method

Reminder: based on the sample observations (y_i) we are searching such θ parameter for which the probability of observing the given sample is the highest

$$L(\theta) = \prod_{i=1}^{n} f_\theta(y_i) \to \max$$

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta(y_i) \to \max$$

$$\frac{\partial l}{\partial \theta} = 0$$

Results: same equations as under OLS (if error terms normally distributed)

$$L(\alpha,\beta,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[\frac{-(y_i-\alpha-\beta x_i)^2}{2\sigma^2}\right]$$

$$l(\alpha,\beta,\sigma)=\log L(\alpha,\beta,\sigma)=C-n\log\sigma-\frac{\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial\alpha}=\frac{\sum_{i=1}^{n}2\cdot(y_i-\alpha-\beta x_i)}{2\sigma^2}=0$$

$$\frac{\partial l}{\partial\beta}=\frac{\sum_{i=1}^{n}2\cdot(y_i-\alpha-\beta x_i)\cdot x_i}{2\sigma^2}=0$$

Estimators

$$\sum y_i = n\hat{\alpha} + \hat{\beta}\sum x_i \qquad \bar{y} = \hat{\alpha} + \hat{\beta}\cdot\bar{x}$$

$$\sum x_i y_i = \hat{\alpha}\sum x_i + \hat{\beta}\sum x_i^2 \qquad \sum x_i y_i = n\bar{x}(\bar{y} - \hat{\beta}\cdot\bar{x}) + \hat{\beta}\sum x_i^2$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\cdot\bar{x}$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}\cdot x_i$$

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}\cdot x_i$$

“Orthogonality” conditions

Normal equations in modified form:

$$0 = \sum \hat{u}_i$$

$$0 = \sum x_i \hat{u}_i$$

Therefore:

$$\sum \hat{u}_i y = 0$$

$$\sum \hat{u}_i \hat{y}_i = \sum \hat{u}_i \left(\hat{\alpha} + \hat{\beta} x_i \right) = 0$$

$$\sum \hat{u}_i \left(\hat{y}_i - y \right) = 0$$

Decomposition of the total sum of squares

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$TSS = \quad RSS \quad + \quad ESS$$

Total $TSS = \sum (y_i - \bar{y})^2 = S_{yy}$

Explained $ESS = \sum (\hat{y}_i - \bar{y})^2 = \sum \left\{ \left(\hat{\alpha} + \hat{\beta} \cdot x_i \right) - \left(\hat{\alpha} + \hat{\beta} \cdot \bar{x} \right) \right\}^2 = \hat{\beta}^2 \cdot S_{xx} = \hat{\beta} \cdot S_{xy}$

Residual $RSS = \sum (y_i - \hat{y}_i)^2 = TSS - ESS = S_{yy} - \hat{\beta} \cdot S_{xy}$

In some textbooks the other way round ("regression" and "error")

Correlation, coefficient of determination

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \pm \sqrt{\frac{S_{xy}^2 / S_{xx}}{S_{yy}}} = \pm \sqrt{\frac{ESS}{TSS}}$$

r_{xy}: observed correlation between x_i and y_i

r_{xy}^2: coefficient of determination

Unbiasedness of the estimators

$$E(\hat{\beta}) = \frac{\sum (x_i - \bar{x}) E(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) \beta (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \beta$$

$$E(\hat{\alpha}) = E(\bar{y} - \hat{\beta} \cdot \bar{x}) = E(\bar{y}) - E(\hat{\beta})\bar{x} = \alpha$$

We assumed here that x_i variables are fixed, but results hold also if these are random variables

Not needed: normal distribution of error term, homoscedasticity

Optimality properties of the estimators

BLUE (best linear unbiased estimator): if homoscedasticity is assumed then our estimator has the smallest variance among the unbiased linear estimators (more details: multivariate case)

If the error term is normally distributed then it is the best estimator among ALL unbiased estimators

Example

2003 Wage Tariff, simple regression

$$\log(Wage_i) = \alpha + \beta_1 Edu_i + u_i$$

Dependent Variable: LOG(KER)
Method: Least Squares
Date: 02/17/09 Time: 16:15
Sample: 1 201971
Included observations: 201971

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.11927	0.005027	2013.122	0.0000
ISKEV	0.121868	0.000409	297.8910	0.0000

R-squared	0.305251	Mean dependent var	11.58047
Adjusted R-squared	0.305248	S.D. dependent var	0.592283
S.E. of regression	0.493679	Akaike info criterion	1.426146
Sum squared resid	49223.58	Schwarz criterion	1.426247
Log likelihood	-144018.0	Hannan-Quinn criter.	1.426175
F-statistic	88739.06	Durbin-Watson stat	0.888245
Prob(F-statistic)	0.000000		

Interpretation

1 more year of education increases log(wage) by 0.12

Thus wage is increased by 12%

Can be used for forecasting purposes

But: causality (exogeneity)?

Not sure, e.g.

- Work experience (observed)
- Abilities (difficult to observe)

Distribution of coefficient estimates (fix x_i variables)

In case of homoscedasticity

$$\mathrm{Var}\left(\hat{\beta}\right) = \mathrm{Var}(S_{xy} / S_{xx}) = \mathrm{Var}\left(\sum (x_i - \bar{x}) y_i / S_{xx}\right) =$$

$$= \sigma^2 \left(\sum (x_i - \bar{x})^2\right) / S_{xx}^2 = \sigma^2 / S_{xx}$$

$$\mathrm{Var}\left(\hat{\alpha}\right) = \sigma^2 \left(1/n + \bar{x}^2 / S_{xx}\right)$$

$$\mathrm{cov}\left(\hat{\alpha}, \hat{\beta}\right) = -\sigma^2 \bar{x} / S_{xx}$$

If error terms are normal

$$\hat{\beta} \sim N\left(\beta, \mathrm{Var}\left(\hat{\beta}\right)\right)$$

$$\hat{\alpha} \sim N\left(\alpha, \mathrm{Var}\left(\hat{\alpha}\right)\right)$$

but σ also has to be estimated…

Seminar

EViews, simple regression

EViews

Software for statistics – econometrics

User friendly, especially suitable for time series analysis

Help files (User's guide)

Stata

- Has more built-in procedures, easier to program
- Better for cross sectional and panel analysis

Gretl:

- Free to download, appropriate at BA level
- Deficiencies: panel and multivariate time series models

Statistical softwares: SPSS, R

Loading data I.

File/new/workfile - undated

Objects/new object/series - edit

Copy – paste

Name

Loading data II.

File/new/workfile – undated

Procs/Import/Read text-lotus-excel

- Source file should be closed!

Excel sheet name …

Names for series: pl. hours tax – reads both series

Manipulation of variables

- Open, descriptive statistics, graphs
  - View/Descriptive statistics
  - View/Graph
  - Multiple series can be selected (open as group)
- Generate variables (genr)
- Sample: smpl
  - smpl 1 20
  - smpl @all

  Or: quick/sample

Regression

Quick/estimate equation …

- Include constant! (c)
- Method: OLS is the default

Or:

equation name.ls …

Example 1 – public expenditures, GDP

- Eurostat data
- Why can be related? Direction of causality?
- Graph (scatterplot)
- Regression estimation
  - Interpretation of the coefficients
  - Significance – t-test, F-test (Wald)
  - Residuals: View/Resid.tests/histogram
- Problems?

Example 2 – hours worked, marginal tax rate

- OECD data
- Why can be related? Direction of causality? Expected sign of the slope coefficient?
- Graph (scatterplot)
- Estimation of the regression
  - Interpretation of the coefficients
  - Significance – t-test
  - Interpretation of R-squared
- Estimated value: forecast