Matematické Metody v Ekonometrii 1.

The Structure of Econometrics Data, The Simple Linear Model, The Ordinary Least Square

Blanka Šedivá

KMA

zimní semestr 2016/2017

Mathematical Models in Econometrics KMA/MME

- Range: 2 + 1, 4 credits,
- Lecturers: RNDr. Blanka Šedivá, Ph.D.
- Email: `sediva@kma.zcu.cz`,
- Home page: http://home.zcu.cz/ sediva/
- Software: Important feature of this course is learning how to use mathematical software capable of econometrical modelling. Note that MATLAB shall be used for practical purposes within lectures throughout the semester. However, students should feel free to use any eligible mathematical software they prefer (e.g. Mathematica, R, Maple, etc.).
- Literature:
  - Wooldridge, J. M.: Introductory Econometrics: A Modern Approach. South-Western, Cengage Learning 2013.
  - Greene, W.H.: Econometric Analysis. Pearson Education, Inc. 2003.
  - Gujarati, D.N.: Basic econometrics, McGraw Hill 2003.

Econometrics

- Econometrics is the application of mathematics, statistical methods, and computer science, to economic data and is described as the branch of economics that aims to give empirical content to economic relations.
- . . . econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference. [1]
- Econometric theory uses statistical theory to evaluate and develop econometric methods.
- Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analyzing economic history, and forecasting. [2]

[1] P. A. Samuelson, T. C. Koopmans, and J. R. N. Stone, "Report of the Evaluative Committee for Econometrica," Econometrica, vol. 22, no. 2, April 1954, pp. 141–146.

[2] https://en.wikipedia.org/wiki/Econometrics

Mathematical economics x Econometric x Economic statistics

- Mathematical economics
  - Economic theory makes statements or hypotheses that are mostly qualitative in nature.
  - The main concern of mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory.
- Econometric
  - ...is mainly interested in the empirical verification of economic theory.
  - ...great deal of ingenuity and practical skill.
  - In econometrics the modeler is often faced with observational as opposed to experimental data.
- Economic statistics
  - ...concerned with collecting, processing, and presenting economic data in the form of charts and tables.

The Structure of Economic Data

- Cross-Sectional Data
  set consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time. Sometimes, the data on all units do not correspond to precisely the same time period.
- Time Series Data
  set consists of observations on a variable or several variables over time. Because past events can influence future events and lags in behaviour are prevalent in the social sciences, time is an important dimension in a time series data set.
- Pooled Cross Sections
  sets have both cross-sectional and time series features.
- Panel or Longitudinal Data
  set consists of a time series for each cross-sectional member in the data set.

Cross-Sectional Data

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

© Cengage Learning, 2013

Time Series Data

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

© Cengage Learning, 2013

Pooled Cross Sections

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

© Cengage Learning, 2013

Panel or Longitudinal Data

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

© Cengage Learning, 2013

Econometric Methodology

- Statement of theory or hypothesis.
- Specification of the mathematical model of the theory
- Specification of the statistical, or econometric, model
- Obtaining the data
- Estimation of the parameters of the econometric model
- Hypothesis testing
- Forecasting or prediction
- Using the model for control or policy purposes.

Statistical versus deterministic relationship

X and Y random variables with p.d.f f or p.c.f. F

- Independence of random variable $F(x, y) = F_1(x) \cdot F_2(y)$
- Transformation of random variable, for example $Y = X^2$
- Correlation between random variables $\Rightarrow$ correlation analysis, where the primary objective is to measure the strength or degree of linear association between two variables.
- Regression analysis deals with the dependence of one variable on other variables.
- In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated. The dependent variable is assumed to be statistical, random, or stochastic, that is, to have a probability distribution.
- A statistical relationship in itself cannot logically imply causation.

Regression function

- X Explanatory, Independent, Control variable, Regressor
- Y Dependent, Explained, Controlled variable, Regressand
- $\mathrm{E}\left(\mathrm{var}\left(Y|x\right)\right) \leq \mathrm{var}\, Y$,
- regression function $\mu(\boldsymbol{x}) = \mathrm{E}\left(Y|\boldsymbol{X} = \boldsymbol{x}\right)$ is conditional expectation function
- parametric regression function
  $\mathrm{E}\, Y_i = f(\boldsymbol{x}_i, \beta_1, \beta_2, \beta_3, \ldots, \beta_p)$, $i = 1, 2, \ldots, n$,
- parametric regression function with addition random part
  $Y_i = f(\boldsymbol{x}_i, \beta_1, \beta_2, \beta_3, \ldots, \beta_p) + \varepsilon_i$, $i = 1, 2, \ldots, n$,
- linear regression function
  $Y_i = \beta_1 + \beta_2\varphi_2(x_i) + \beta_3\varphi_3(x_i) + \cdots + \beta_p\varphi_p(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n$
- $y_1, y_2, \ldots, y_n$ samples of random variables Y_i
- the aim of regression analysis is to estimate unknown parameters $\beta_1, \beta_2, \ldots, \beta_p$

Simple linear regression model

- We assume a model $Y_i = \beta_1 + \beta_2\ x_i + \varepsilon_i, i = 1, 2, \ldots, n$
- Population regression line $f(x_i) = \beta_1 + \beta_2\ x_i$,
- The regression coefficients; β_1 and β_2 are also known as intercept and slope coefficients,
- ε_i stochastic disturbance or stochastic error term,
- $\mathrm{E}\,(Y|X = x) = \beta_1 + \beta_2\ x$ population regression function
- Samples $y_1, y_2, \ldots, y_n \Rightarrow$ the sample regression function
- Estimators of β_1, β_2 depend on samples $y_1, y_2, \ldots, y_n$
- The hat symbol denotes an estimated value $\hat{\beta}_1, \hat{\beta}_2 \Rightarrow$

$$\hat{\beta}_j = \hat{\beta}_j\,(y_1, y_2, \ldots, y_n)$$

- The sample regression function is based on the sample $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2\ x_i$

Fitting the model

- We will be using *least squares* regression. This measures the *goodness of fit* of a line by the sum of squared errors, *SSE*.
- Least squares regression chooses the line that minimizes

$$SSE(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 \cdot x_i)^2.$$

- In principle, we might measure goodness of fit differently:

$$SAD(\beta_1, \beta_2) = \sum_{i=1}^{n} |y_i - \beta_1 - \beta_2 \cdot x_i|.$$

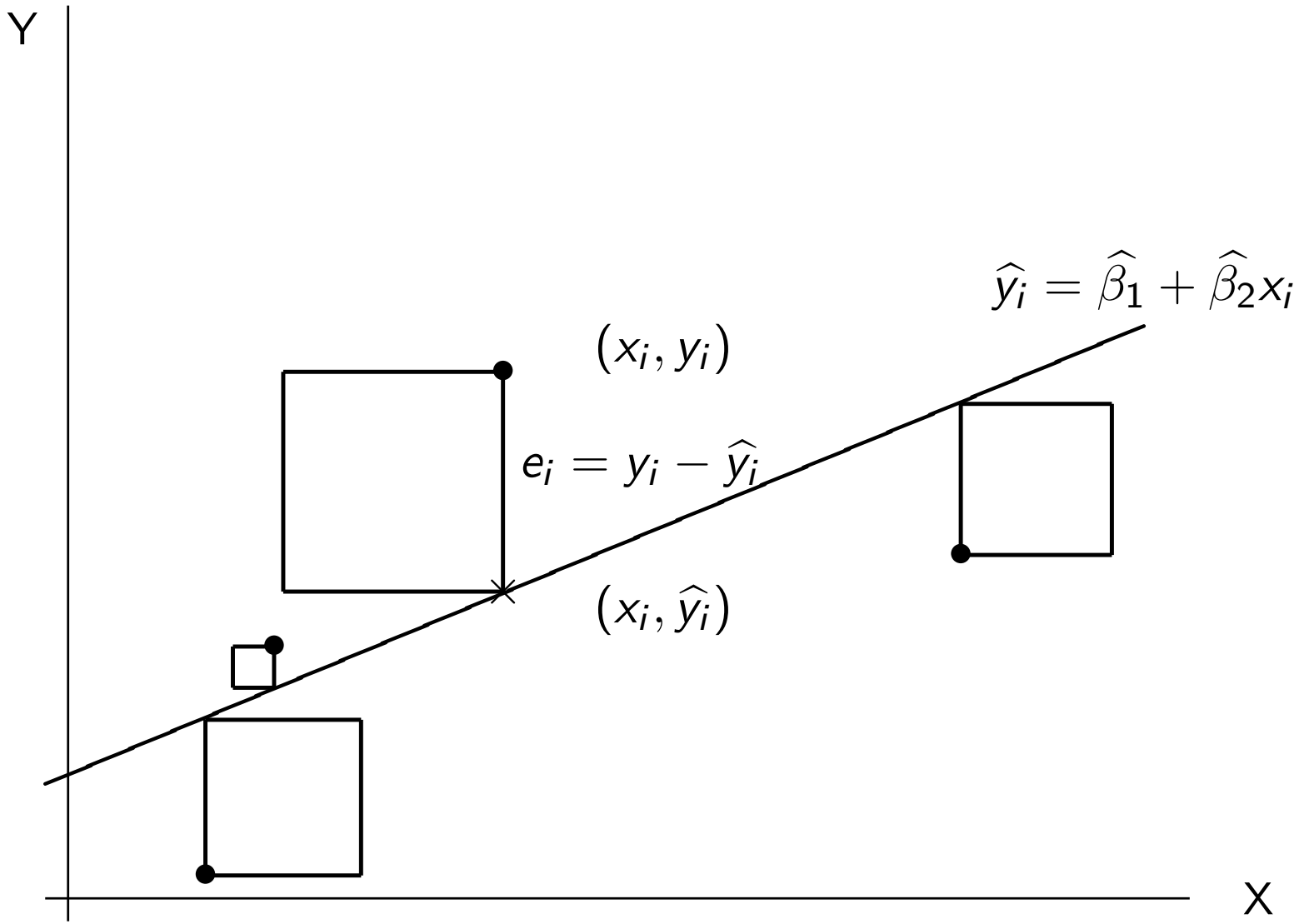
- For some *loss function* L we might try to minimize

$$L(\beta_1, \beta_2) = \sum_{i=1}^{n} L(y_i - \beta_1 - \beta_2 x_i)$$

Why least squares (LS)?

- With least squares, the minimizers have explicit formulae – not so important with today's computer power – especially when L is convex.
- Resulting formulae are *linear* in the outcome Y. This is important for inferential reasons. For only predictive power, this is also not so important.
- If assumptions are correct, then this is *maximum likelihood estimation*.
- Statistical theory tells us the *maximum likelihood estimators (MLEs)* are generally good estimators.

Geometric interpretation of LS



The Ordinal Least Squares estimators (OLS)

- $(\hat{\beta}_1, \hat{\beta}_2) = \operatorname{argmin}_{(\beta_1,\beta_2)\in\mathbf{R}^2} Q(\beta_1, \beta_2)$, where $Q(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - (\beta_1 + \beta_2 x_i))^2$ is criteria function
- short notation $b_1 = \hat{\beta}_1$, $b_2 = \hat{\beta}_2$

$$b_1 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} (x_i)^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \qquad b_2 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

Properties of the sample regression line

- $b_2 = \frac{s_{x\ y}}{s_x^2} = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i}{n} \frac{\sum_{i=1}^{n} y_i}{n}}{\frac{\sum_{i=1}^{n} (x_i)^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n} \right)^2}$, where

  $s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \overline{x}\ \overline{y} = \overline{xy} - \overline{x}\ \overline{y}$ and

  $s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i)^2 - (\overline{x})^2$
- $b_2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \frac{y_i - \overline{y}}{x_i - \overline{x}} = \sum_{i=1}^{n} w_i \tan(\alpha_i)$
- $\widehat{y} = b_1 + b_2 x = (\overline{y} - b_2 \overline{x}) + b_2 x = \overline{y} + b_2 (x - \overline{x})$
- $b_1 = \overline{y} - b_2 \overline{x}$