ECONOMIC STATISTICS

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

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

ECONOMIC STATISTICS

Author: Anikó Bíró

Supervised by Anikó Bíró

June 2010

ECONOMIC STATISTICS

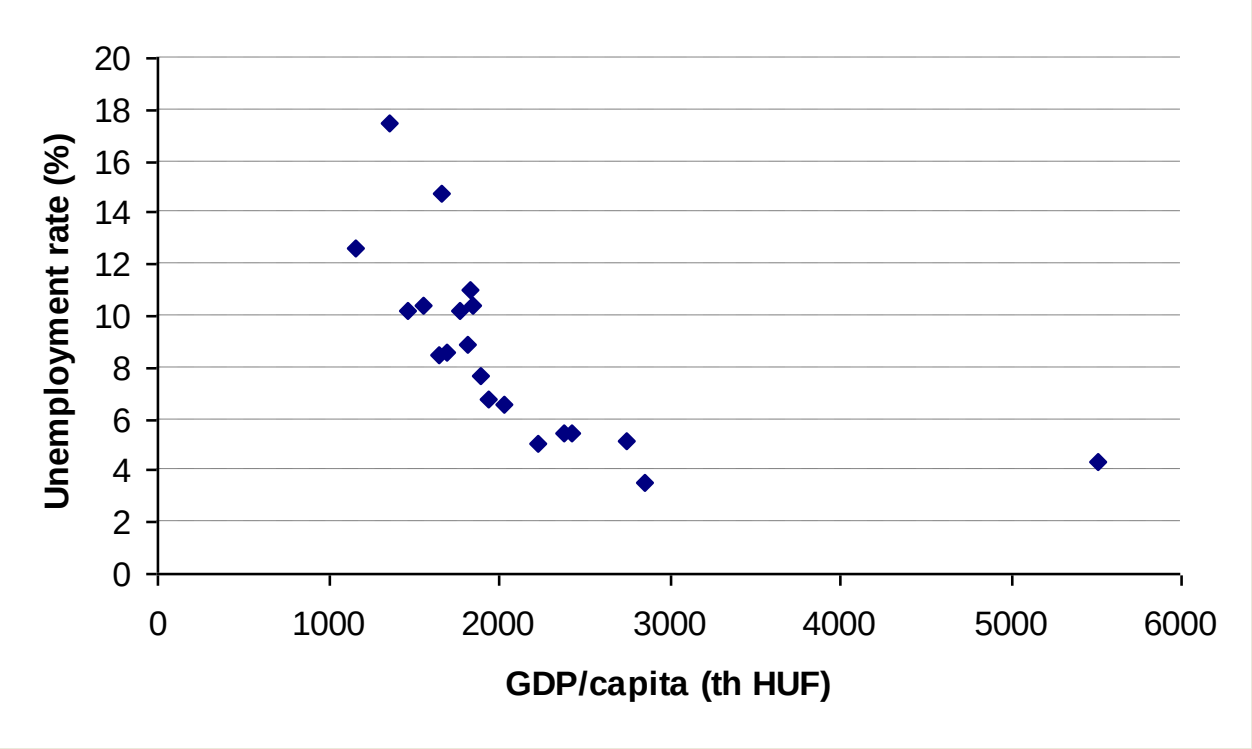
Week 3

Correlation, simple regression – introduction

Anikó Bíró

Example from week 2

Negative relationship between two variables – point diagram (KSH):



Correlation

- Relationship between two variables numerically
- Notation: correlation between X and Y is r_{XY}
- Square of correlation (r_{XY}^2): what percentage of Y's variation is explained by X = what percentage of X's variation is explained by Y

Supplement: formula of correlation

$$r = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$$

Properties of correlation

- Value between -1 and 1
- Positive value – positive relationship. r=0: no correlation between the variables
- Larger positive value – stronger positive relationship
- Correlation between X and Y = correlation between Y and X
- Correlation of a variable with itself = 1
- Correlation with a constant = 0

Example

Correlation between unemployment rate and GDP/capita = -0,62

- Negative relationship
- Higher GDP/capita – lower unemployment
- The standard deviation of county level GDP/capita explains 38% of the standard deviation of unemployment rate (0,62*0,62=0,384)

Causality?

- Does one variable ”cause” the other?
- Correlation does not reveal the direction of causality
- There might be no causality at all
- Previous examples? (GDP – unemployment, GDP – number of enterprises)

Correlation between more variables

- M variables – M(M-1)/2 correlations
- Correlation matrix of 3 variables (X, Y, Z):

	X	Y	Z
X	1		
Y	r_{XY}	1	
Z	r_{XZ}	r_{ZY}	1

Strength of relationship graphically

- Point diagram between two variables
- See: textbook graphs
- ”How difficult is to draw a line fitting the points?”
- ”How scattered are the points?”

Correlation vs. regression

- Numerical analysis of the relationship between variables:
- Correlation:
  - Between 2 variables
  - Causality?
- Regression:
  - Complex relationships (more variables)
  - There might be an underlying economic model – causality
  - Examples: wage regression (education, …), inflation regression (interest rate, …)

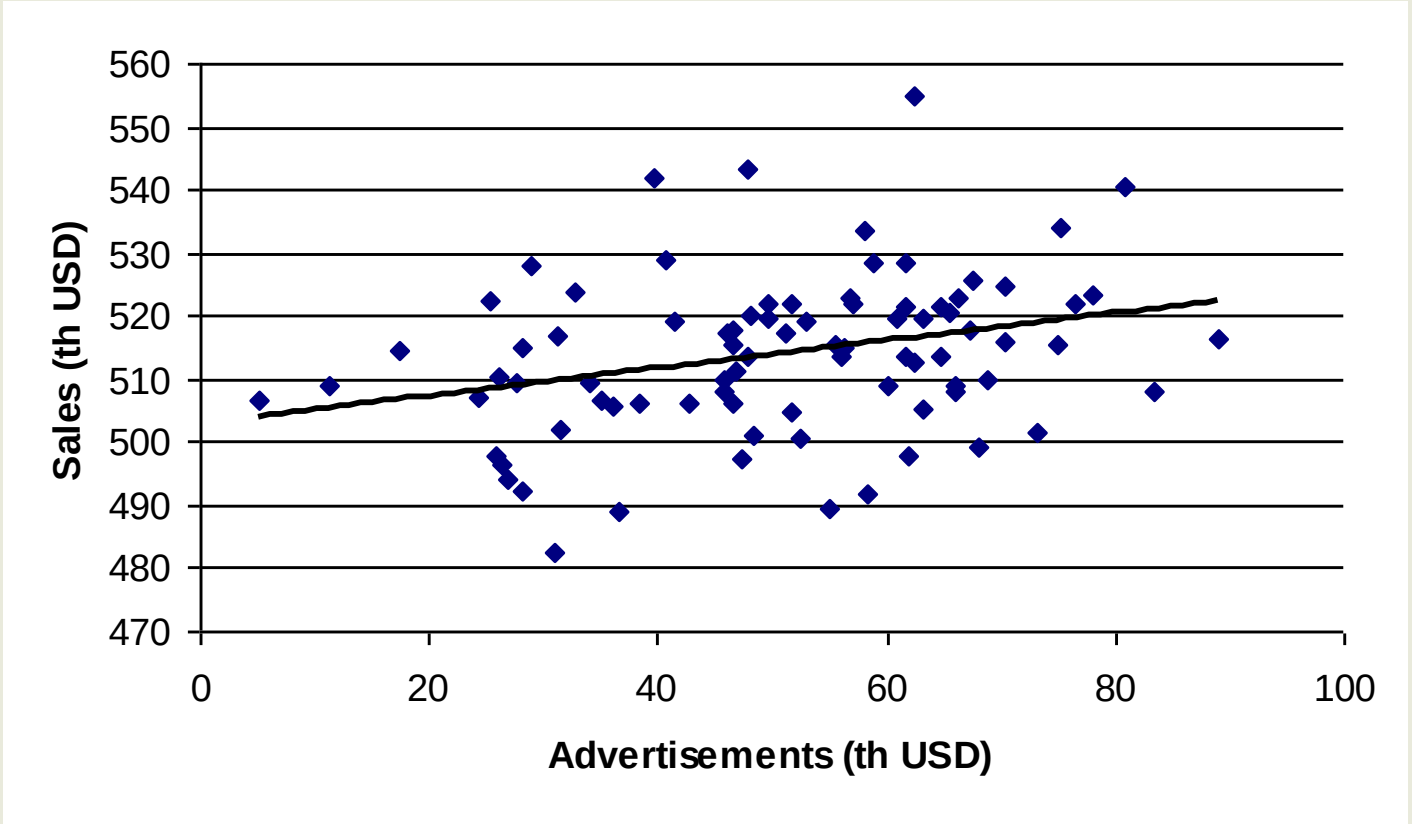
Simple (univariate) regression

- Y dependent variable, X explanatory variable (regressor)
- Assumption: linear relationship
- Regression line:

$$Y = \alpha + \beta X$$

- Reality: the data do not fit a line

Example: advertisement expenditures



Error term

- Linear regression: no perfect fit
  - Omitted, unobservable variables
  - Not linear relationship
- Regression model with error term:

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

- Error term (disturbance): distance between the data point and the regression line
- Causality (model)? Generalization of correlation?

Estimation

- Unknown value of coefficients
- Estimated coefficients: coefficients of the best fitting line
  - Notation:

$$\hat{\alpha}, \hat{\beta}$$

- Residual:

$$Y = \hat{\alpha} + \hat{\beta}X + u$$

$$u \neq e$$

OLS estimation

- Best fitting line – minimal sum of square of residuals

$$SSR = \sum_{i=1}^{N} u_i^2$$

- Ordinary least squares (OLS) estimation

Advertisement example, cont.

- Estimated coefficients:
- 502,92 – intercept parameter (constant);
- 0,22 – coefficient of advertisements (slope)
- Interpretation?
- Slope:
  - Average change in Y if X increases by one unit
  - Marginal effect

Summary

- Correlation:
  - Strength of relationship between two variables
  - Properties of correlation
  - Interpretation: square of correlation
- Linear regression (univariate):
  - Underlying economic model
  - Error term
  - Residual
  - Estimation: OLS

Correlation, simple regression – introduction

Seminar 3

Correlation

- Relationship between two variables numerically
- Square of correlation (r_{XY}^2): what percentage of Y's variation is explained by X = what percentage of X's variation is explained by Y

- Excel: CORREL() function

Properties of correlation

- Value between -1 and 1
- Positive value – positive relationship. r=0: no correlation between the variables
- Correlation between X and Y = correlation between Y and X
- Correlation of a variable with itself = 1

Examples

Correlation and squared correlation?

- KSH county level data: unemployment and GDP/capita?
- KSH county level data: GDP/capita and number of registered enterprises?
- MNB: HUF/EUR and HUF/USD?

Correlation between more variables - example

- European sample, women aged 50+ (SHARE subsample)
  - Education (0–4 scale)
  - If ever smoked daily
  - Malignant tumor (cancer)
- Qualitative data
- What kind of correlation expected?

Example, cont.

- Immediate causality: smoking – cancer
- Proximate causality: higher education level – cancer

	Educ.	Smoke	Cancer
Educ.	1		
Smoke	0,18	1	
Cancer	0,01	0,04	1

Simple (univariate) regression

- Y dependent variable, X explanatory variable
- Assumption: linear relationship
- Regression line:

$$Y = \alpha + \beta X$$

- Error term vs. residual

Example: advertisement expenditures

Koop: Advert.xls file

- Correlation?
- Point diagram
- Regression line with Excel: Diagram/Trend line

OLS estimation

- Best fitting line – minimal sum of square of residuals

$$SSR = \sum_{i=1}^{N} u_i^{\,2}$$

- Ordinary least squares (OLS) estimation
- Excel: Data analysis/Regression – estimate and interpret coefficients of the advertisement examples
- Sensitivity of the coefficients to scaling

(unit of measurement)?

Further examples

- KSH county level data: regression of unemployment rate on the number of registered enterprises

  Y: unemployment

  X: enterprises

- Forest.xls: effect of population growth on deforestation?

Homework 2 (individual)

Choose 3 variables from a database among which correlation is expected

- What kind of relationship is expected? Explain.
- Descriptive statistics (graphical + numerical)
- Analysis of correlations