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

HUNGARY'S RENEWAL

The project is supported by the European Union.

Author: Anikó Bíró
Supervised by Anikó Bíró
June 2010

Week 4

Simple regression – fit, nonlinearity, confidence interval

Simple regression – reminder

- Regression model:

$$Y_i = \alpha + \beta X_i + e_i$$
$$Y_i = \hat{\alpha} + \hat{\beta} X_i + u_i$$

- Estimation: OLS

Example

70 tropical countries, relationship between X: population density (capita/1000 ha), and Y: deforestation rate (%)

	Coefficients
Intercept	0,60
X variable	0,001

Interpretation?

Measure of fit

- OLS: finding the best fitting line
- How good is the fit?
  - Measure: R^2
  - Simple (univariate) regression:

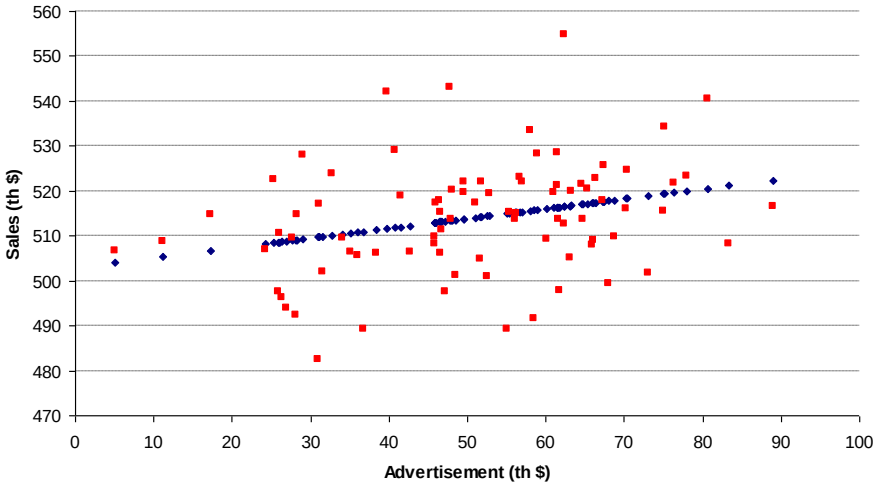
    square of correlation= R^2

Estimated value

- Regression line: $Y = \alpha + \beta X + e$
- Estimated/fitted/forecasted value: $\hat{Y} = \hat{\alpha} + \hat{\beta} X$

Comparison of the two – how good is the fit

Advertisement example:



Residual: $u = Y - \hat{Y}$

Residual vs. error term!

R2

To Total sum of squares : $TSS = \sum (Y_i - \bar{Y})^2$

Variance : TSS/(N -1)

Regression sum of squares : $RSS = \sum (\hat{Y}_i - \bar{Y})^2$

Sum of squared residuals : $SSR = \sum (Y_i - \hat{Y}_i)^2 = \sum u_i^2$

$$TSS = RSS + SSR$$

$$R^2 = 1 - \frac{SSR}{TSS} = \frac{RSS}{TSS}$$

Interpret R2

- What percentage of the variance of Y is explained by X

$$0 \leq R^2 \leq 1$$

- R^2=1 – perfect fit

Deforestation example

Regression statistics	
r-squared	0,434

ANALYSIS OF VARIANCIE

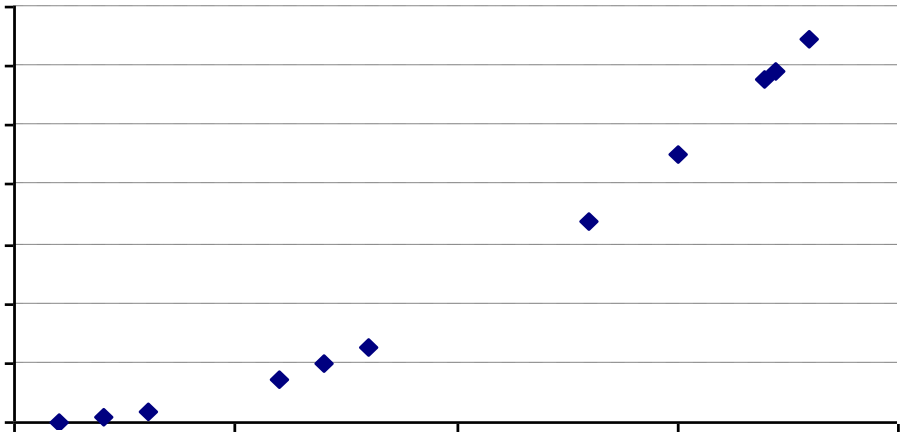
	df	*SS*
Regression	1	25,828
Residual	68	33,618
Total	69	59,446

Nonlinearity

Nonlinear relationship between X and Y

Common examples:

- Quadratic:



- Logarithmic

Logarithmic form

- Can ensure linear relationship
- Easy to interpret – elasticity:

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

$$\beta = \frac{d \ln Y}{d \ln X}$$

If X increases by one %, Y increases by beta % on average

- Unit of measurement does not matter
- Approximation of % change: $100 \cdot d \ln Y$

Logarithmic form, cont.

How to interpret the slope coefficients?

$$Y_i = \alpha + \beta \ln X_i + e_i$$

$$\ln Y_i = \alpha + \beta X_i + e_i$$

Uncertainty

- Real values of the coefficients are unknown
- Estimated based on a sample
  - Estimated value is not exactly equal to the true value
- Point estimation: does not reveal the uncertainty

Factors influencing the precision of the OLS estimation

See textbook graphs

- More observations – more precise estimation
- Smaller error terms – more precise estimation
- Larger variance of X – more precise estimation
  - Example: effect of education level on income

Confidence interval

$$\left(\hat{\beta} - t_b s_b, \hat{\beta} + t_b s_b\right)$$

$$s_b = \sqrt{\frac{SSR}{(N-2)\sum (X_i - \overline{X})^2}}$$

s_b : standard deviation of $\hat{\beta}$

t_b : Student's t - distribution

Larger confidence level $\rightarrow$ larger t_b

More observations $\rightarrow$ smaller t_b

Interpretation

- Most common:
  95% confidence interval
  ”There is 95% chance that the true value of the coefficient lies in the given interval”
- Large N, 95%: t=1,96
- Table of t-distribution

- Excel: confidence level can be chosen

Deforestation example

	Coefficients	*Standard dev.*	*Bottom 95%*	*Top 95%*
Intercept	0,6000	0,1123	0,3758	0,8241
X variable	0,0008	0,0001	0,0006	0,0011

Summary

- Interpretation of estimated coefficients
- R-squared
- Nonlinearity, logarithmic form
- Uncertainty, confidence interval

Simple regression – fit, nonlinearity, confidence interval

Seminar 4

R2

$$\text{Total sum of squares}: TSS = \sum (Y_i - \bar{Y})^2$$

$$\text{Variance}: \text{TSS/(N -1)}$$

$$\text{Regression sum of squares}: RSS = \sum (\hat{Y}_i - \bar{Y})^2$$

$$\text{Sum of squared residuals}: SSR = \sum (Y_i - \hat{Y}_i)^2 = \sum u_i^2$$

$$TSS = RSS + SSR$$

$$R^2 = 1 - \frac{SSR}{TSS} = \frac{RSS}{TSS}$$

Interpret R2

- What percentage of the variance of Y is explained by X

$$0 \leq R^2 \leq 1$$

- R^2=1 – perfect fit
- Examples: advertisement regression, KSH unemployment rate regression

Examples for nonlinearity

Textbook: 4.5, 4.6

Uncertainty

- Real values of the coefficients are unknown
- Estimated based on a sample
  - Estimated value is not exactly equal to the true value
- Point estimation: does not reveal the uncertainty
- Confidence interval:

$$\left(\hat{\beta} - t_b s_b, \hat{\beta} + t_b s_b\right)$$

$$s_b = \sqrt{\frac{SSR}{(N-2)\sum (X_i - \overline{X})^2}}$$

Examples

- Advertisement – sales example: confidence interval of the estimated slope parameter (various confidence levels)

- Real estate prices – lot size (hprice.xls)

Homework 3 (groups)

- Analyze the relationship between two variables from a cross sectional sample (KSH, Eurostat, OECD, Penn World tables)
  - Descriptive statistics of both variables
  - Correlation
  - Regression
    - Functional form?
    - Fit?
    - Interpret the estimation results (confidence interval, as well)