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

The project is supported by the European Union.

Authors: Péter Elek, Anikó Bíró

Supervised by Péter Elek

June 2010

Week 3

Simple regression II.

Plan

Estimation of standard deviation
Hypothesis testing, confidence interval
Forecasting
Outliers, alternative functional forms

Reminder I

$y_i = \alpha + \beta x_i + u_i$

Assumptions:

1. $E(u_i) = 0$
2. $Var(u_i) = \sigma^2$ for all i
3. u_i, u_j independent for all $i \neq j$
4. x_i, u_j independent for all i, j
5. u_i normally distributed for all i: $N(0, \sigma^2)$

Reminder II

$y_i = \alpha + \beta x_i + u_i$

Estimation

Method of moments $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$

OLS

Maximum likelihood

Unbiased estimator – normality and homoscedasticity not needed!

Spurious regression

"Regression to the mean" for normally distributed variables with same standard deviation:

$$E(Y|X = x) - m_y = \rho(X - m_x),\ \rho<1$$

Coefficient of the regression: $\hat{\rho} = \frac{S_{xy}}{S_{xx}}$

Statistical consequence: coefficient less than 1!

Examples: height of parents and children, scores of first and second exams

Sampling distribution of coefficient estimates

$$\mathrm{Var}\left(\hat{\beta}\right) = \mathrm{Var}(S_{xy} / S_{xx}) = \mathrm{Var}\left(\sum \left(x_i - \bar{x}\right) y_i / S_{xx}\right) =$$
$$= \sigma^2 \left(\sum \left(x_i - \bar{x}\right)^2\right) / {S_{xx}}^2 = \sigma^2 / S_{xx}$$
$$\hat{\beta} \sim N\left(\beta, \mathrm{Var}\left(\hat{\beta}\right)\right)$$

Estimation of variance

$$u_i = y_i - (\alpha + \beta x_i) = \hat{u}_i + (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) x_i$$

$$0 = \sum \hat{u}_i \left[(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) x_i \right] \quad \text{from normal equations}$$

$$\sum u_i^2 = \sum \hat{u}_i^2 + \sum \left[(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) x_i \right]^2$$

$$Q = RSS + Q_2$$

$$\sigma^2 \chi_n^2 \sim \sigma^2 \chi_{n-2}^2 + \sigma^2 \chi_2^2$$

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \sim \sigma^2 \frac{\chi_{n-2}^2}{n-2}, \qquad \mathrm{E}(\hat{\sigma}^2) = \sigma^2$$

Chi-squared, t-distribution

$x_1, x_2, \ldots, x_n$ independent variables with standard norm. distribution

$$\mathrm{Z} = \sum_{i=1}^{n} x_i^{\,n} \sim \chi_n^{\,2}$$

$x \sim N(0,1)$ $\quad \mathrm{y} \sim \chi_{\mathrm{n}}^{\,2}$ independent $\Rightarrow$

$$\mathrm{Z} = \mathrm{x}/\sqrt{\mathrm{y}/\mathrm{n}} \sim t_n$$

Hypothesis testing, confidence interval

$$\hat{\sigma}^2/\sigma^2 \sim \frac{\chi^2_{n-2}}{n-2} \quad \text{independent of } \left(\hat{\alpha},\hat{\beta}\right) \text{ (no proof)}$$

$$\frac{\hat{\beta}-\beta}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1) \qquad \frac{\hat{\beta}-\beta}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$$

$$\frac{\hat{\alpha}-\alpha}{\sqrt{\sigma^2\left(1/n+\bar{x}^2/S_{xx}\right)}} \sim N(0,1) \qquad \frac{\hat{\alpha}-\alpha}{\sqrt{\hat{\sigma}^2\left(1/n+\bar{x}^2/S_{xx}\right)}} \sim t_{n-2}$$

Confidence interval, hypothesis testing

$$P\left(-t_{n-2}(1-\alpha/2) < \frac{\hat{\beta}-\beta}{SE\left(\hat{\beta}\right)} < t_{n-2}(1-\alpha/2)\right) = 1-\alpha$$

Analysis of variance

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

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

Previous slide: $RSS \sim \sigma^2\chi^2_{n-2}$

If $\beta = 0$, thus y_i independent N(α,σ^2) variables then

$TSS \sim \sigma^2\chi^2_{n-1}$ (Fisher-Bartlett theorem)

$ESS \sim \sigma^2\chi^2_1$

RSS and *ESS* independent

Analysis of variance (cont.)

$\beta = 0$ hypothesis

Source of var.	Sum of squares	D. of freed.	Mean squares	F
Regr.	$ESS = r^2 S_{yy}$ $= \hat{\beta} S_{xy} = \hat{\beta}^2 S_{xx}$	1	$MS_1 = ESS/1$ $\sim \chi_1^2/1$	$F = MS_1/MS_2 = (n-2)r^2/(1-r^2)$ $\sim F_{1,n-2}$ $= \hat{\beta}^2 / \hat{\sigma}^2 / S_{xx}$ $\sim t_{n-2}^2$
Residual	$RSS = (1-r^2)S_{yy}$ $= (n-2)\hat{\sigma}^2$	$n-2$	$MS_2 = RSS/(n-2)$ $\sim \chi_{n-2}^2/(n-2)$	
Total	S_{yy}	$n-1$		

Forecasting

$$y_0 = \alpha + \beta x_0$$

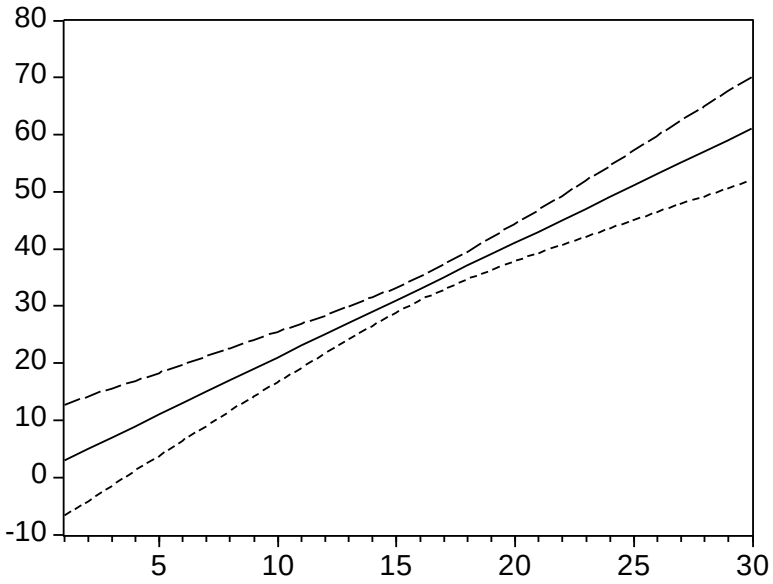
$$\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$$

$$E(\hat{y}_0 - y_0) = E(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) x_0 = 0 \text{ (unbiased)}$$

$$\begin{aligned} \mathrm{Var}(\hat{y}_0 - y_0) &= \mathrm{Var}(\hat{\alpha} - \alpha) + x_0^2 \mathrm{Var}(\hat{\beta} - \beta) \\ &\quad + 2x_0 \,\mathrm{cov}(\hat{\alpha} - \alpha, \hat{\beta} - \beta) + \mathrm{Var}\, u_0 \\ &= \sigma^2 (1 + 1/n + (x_0 - \bar{x})^2 / S_{xx}) \end{aligned}$$

If $\bar{x} = x_0$ then it is minimal.

Confidence interval of forecasts



Forecasting expected value

$$
\begin{aligned}
E(y_0) &= \alpha + \beta x_0 \\
\hat{E}(y_0) &= \hat{\alpha} + \hat{\beta} x_0 \quad (= \hat{y}_0) \\
\mathrm{Var}\left(\hat{E}(y_0) - E(y_0)\right) &= \mathrm{Var}(\hat{\alpha} - \alpha) + x_0^2 \,\mathrm{Var}(\hat{\beta} - \beta) \\
&\quad + 2x_0 \,\mathrm{cov}(\hat{\alpha} - \alpha, \hat{\beta} - \beta) \\
&= \sigma^2 \left\{ 1/n + (x_0 - \bar{x})^2 / S_{xx} \right\} \\
&< \mathrm{Var}(\hat{y}_0 - y_0)
\end{aligned}
$$

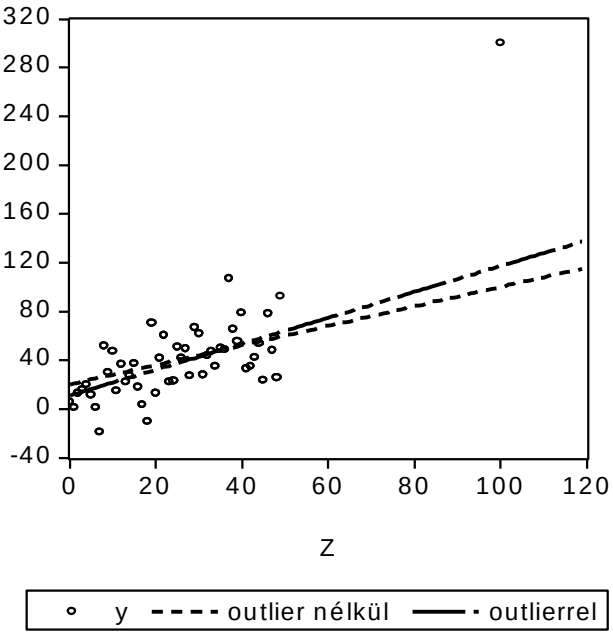
Outliers

Outlier: lies far from the other observations

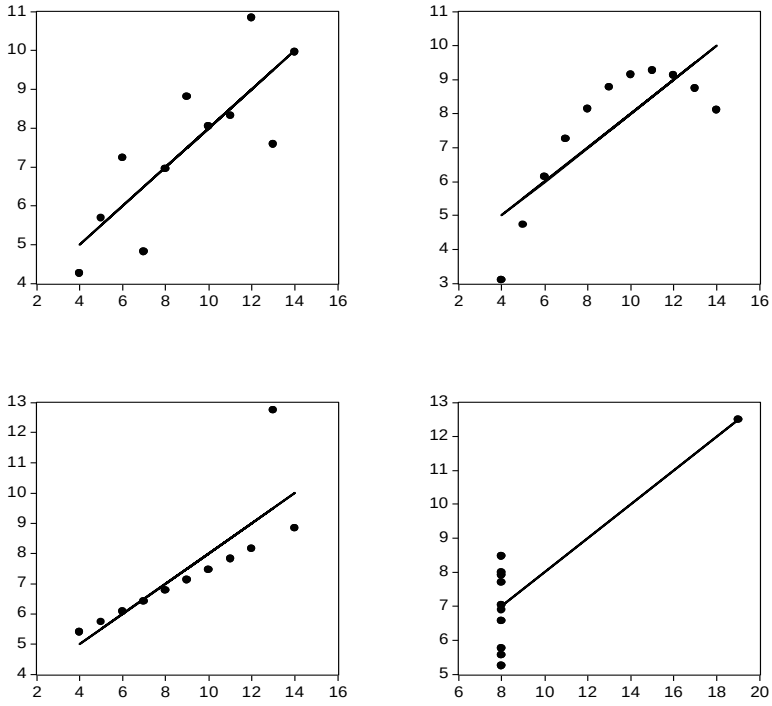
Can change the regression line

Reasons and handling:

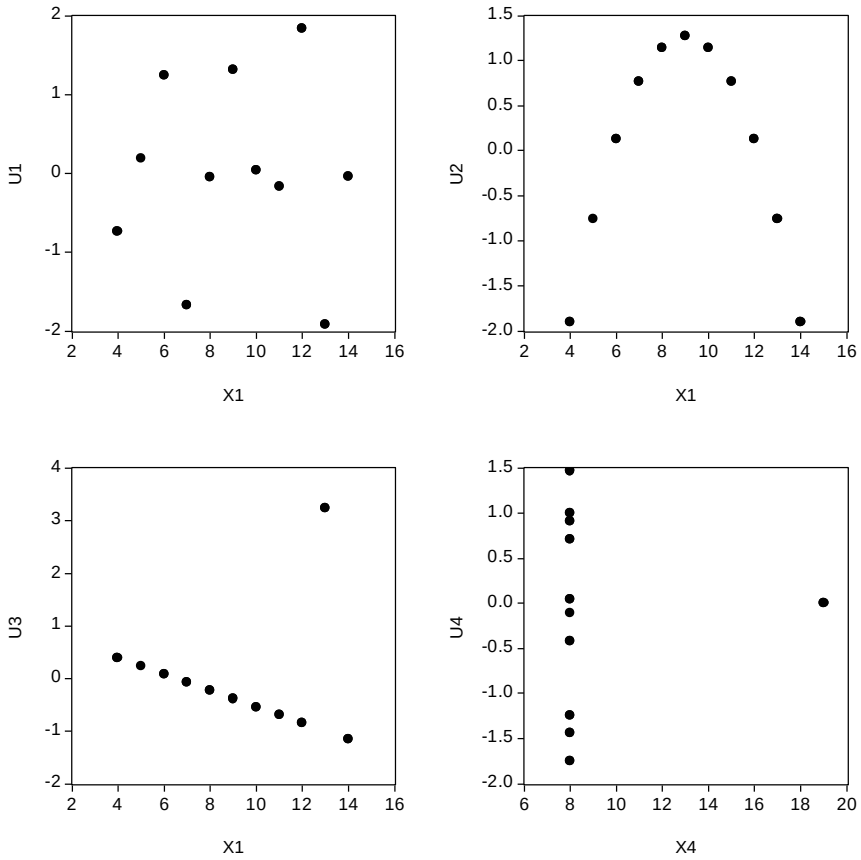
- Data error (omit the data)
- Special case (individual analysis)
- Same mechanisms, but outlier data (analyze with the other observations)



Outliers (cont.): same regression lines, but different relationships



Outliers (cont.): analysis of the residuals (will be important in multivariate case)



Alternative functional forms

$y = Ae^{\beta x}$ $\log(y) = \log(A) + \beta x$

Form of error term matters:

$y = Ae^{\beta x}e^{u}$ $\log(y) = \log(A) + \beta x + u$

$E(e^{u}) \neq e^{E(u)} = 1$, thus $E(y) \neq Ae^{\beta x}$

Other examples

$y = Ax^{\beta}$ $\log(y) = \log(A) + \beta log(x)$

$y = A + B/x$ (here only the explanatory variable has to be transformed)

Example: relationship between earnings and education

$\log(earn_i) = \alpha + \beta_1 educ_i + u_i$, 2003 Wage Tariff
(F-test is the square of t-test in univariate case)

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		

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	88739.06	(1, 201969)	0.0000
Chi-square	88739.06	1	0.0000

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(2)	0.121868	0.000409

Restrictions are linear in coefficients.

Example (cont.): Forecasting

How much earnings we expect with 15 years of education?
Uncertainty is relatively large.

$$y_0 = \log(earn) = 10.12 + 15 \cdot 0.122 = 11.95$$

$$earn = 154{,}800 \text{ Ft (in 2003, but it is not unbiased)}$$

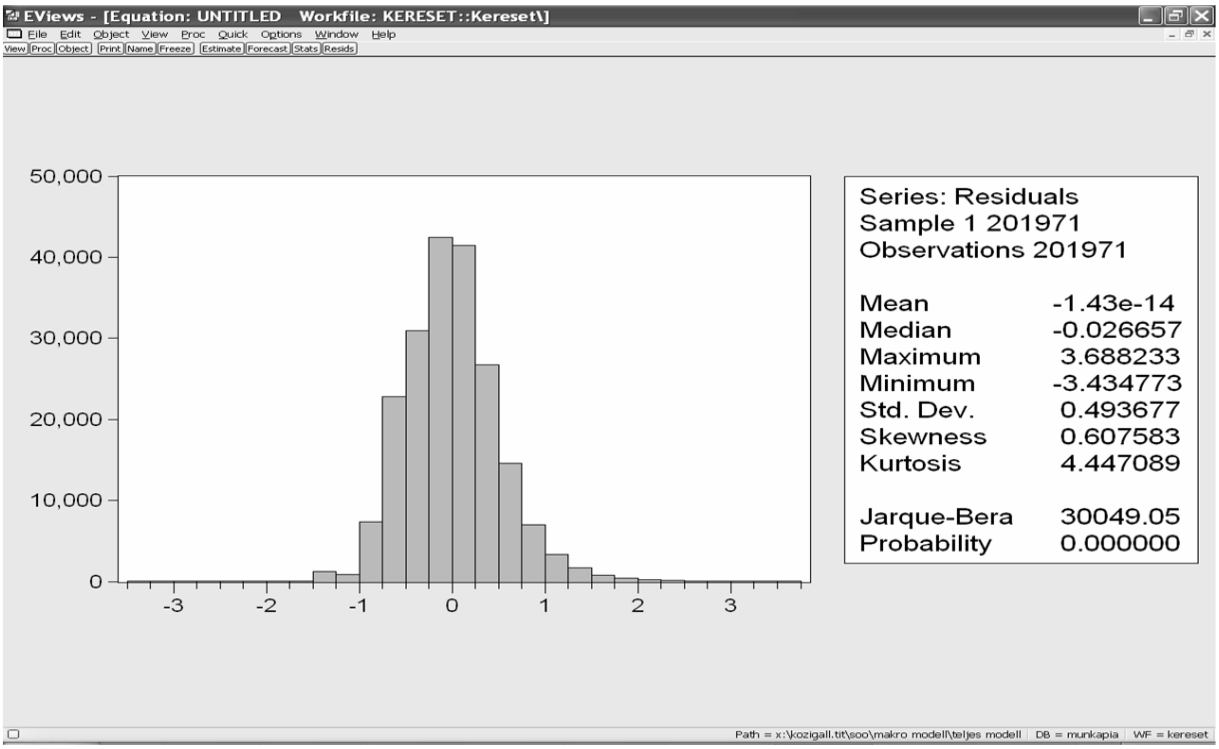
$$\sqrt{\mathrm{Var}(y_0)} = \sqrt{\mathrm{Var}(\hat{\alpha}) + 15^2 \cdot \mathrm{Var}(\hat{\beta}) + 2 \cdot 15 \cdot \mathrm{cov}(\hat{\alpha}, \hat{\beta}) + \mathrm{Var}(u)} = 0.4937$$

Assuming normally distributed error terms,

$$0.95 = P\left[\hat{E}(y_0) \in (11.95 \pm 1.96 \cdot 0.4937)\right]$$

$$0.95 = P\left[\text{ker} \in (58{,}800\text{Ft},\ 407{,}400\text{Ft})\right]$$

Normality of error terms: slight deviation from normal distribution



Simple regression, summary

Assumptions

Estimation and its properties (unbiased), interpretation of estimated coefficients

Hypothesis testing

Problem of outliers

Seminar
Simple regression II

Estimating marginal propensity of consumption

FOGYJOV file
CONS = $\alpha + \beta \cdot INC + u$
Sample size: 900
Interpretation of coefficients, calculation of marginal and average propensity of consumption
Interpretation of *t*-statistic, *p*-value, R^2, *RSS*
Testing $\beta = 1$ hypothesis
95% and 99% confidence intervals for β
Analysis of significance for a subsample of 30 observations
Forecasting for 1.5 m Ft annual income