ECONOMETRICS

NEW SZÉCHENYI PLAN

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

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 6.

Multivariate regression III

Péter Elek, Anikó Bíró

Content

F-test (cont.),

Stability tests

Adjusted R^2, model selection

Dummy variables

F-test more generally

Joint test of r constraints in a regression with k explanatory variables

Nested hypotheses: the parameter set of the model is a subset of that of the original one

Example

$$U: y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$H_0: \beta_2 = 0, \beta_3 = 0$$

$$R: y = \alpha + \beta_1 x_1 + v$$

F-test, test statistic

Sum of squares decompositon

$$TSS = RESS + RRSS = RESS + (RRSS - URSS) + URSS$$

Degrees of freedom

$$n - 1 = (k - r) + (n - k + r - 1) = (k - r) + r + (n - k - 1)$$

$$F = \frac{(RRSS - URSS) / r}{URSS / (n - k - 1)} \sim F_{r,(n-k-1)}$$

$\sim \chi_r^2 / r$, in large samples, approximately the Wald-test)

$$RRSS = S_{yy}(1 - R_R^2) \quad URSS = S_{yy}(1 - R_U^2)$$

$$F = \frac{(R_U^2 - R_R^2) / r}{(1 - R_U^2) / (n - k - 1)} \sim F_{r,(n-k-1)}$$

Testing a linear function of the parameters

Example: Cobb-Douglas production function

$$\log X = \alpha + \beta_1 \log L + \beta_2 \log K + u$$

$$H_0: \beta_1 + \beta_2 = 1$$

t-test: $\theta = \beta_1 + \beta_2$ $\beta_2 = \theta - \beta_1$

$$\log X = \alpha + \beta_1(\log L - \log K) + \theta \log K + u$$

$$H_0: \theta = 1$$

t-test directly on $\beta_1 + \beta_2$, using that the variance of the sum is:

$$Var(\beta_1\hat{} + \beta_2\hat{}) = Var(\beta_1\hat{}) + Var(\beta_2\hat{}) + 2cov(\beta_1\hat{}, \beta_2\hat{})$$

F-test: $\beta_2 = 1 - \beta_1$

$$R: \log X - \log K = \alpha + \beta_1 (\log L - \log K) + u$$

Stability test: two independent data sets (sometimes referred to as Chow-test)

1. $y_i = \alpha_1 + \beta_{11} x_{1i} + \beta_{21} x_{2i} + \ldots + \beta_{k1} x_{ki} + u_i,\ i = 1 \ldots n_1$
2. $y_i = \alpha_2 + \beta_{12} x_{1i} + \beta_{22} x_{2i} + \ldots + \beta_{k2} x_{ki} + v_i,\ i = 1 \ldots n_2$

$$H_0: \alpha_1 = \alpha_2,\ \beta_{11} = \beta_{12},\ \ldots,\ \beta_{k1} = \beta_{k2}$$

$$F = \frac{(RRSS - RSS_1 - RSS_2) / (k+1)}{(RSS_1 + RSS_2) / (n_1 + n_2 - 2k - 2)} \sim F_{k+1,\, n_1 + n_2 - 2k - 2}$$

RRSS: from the merged data set,

RSS_1, RSS_2: from separate regressions

Stability test – Chow-test (predictive)

1. $y_i = \alpha_1 + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \ldots + \beta_{k1}x_{ki} + u_i,\ i = 1 \ldots n_1$
2. $y_i = \alpha_2 + \beta_{12}x_{1i} + \beta_{22}x_{2i} + \ldots + \beta_{k2}x_{ki} + v_i,\ i = 1 \ldots n$

$n - n_1 < k + 1$ is possible (in contrast to the previous one)

RSS_1: res. sum of squares based on the first n_1 observations

$RRSS$: res. sum of squares based on the model estimated from all ($n = n_1 + n_2$) observations

$$F = \frac{(RRSS - RSS_1)/(n - n_1)}{RSS_1/(n_1 - k - 1)} \sim F_{(n - n_1),(n_1 - k - 1)}$$

ELTE ECON

Adjusted R^2

$$\hat{\sigma}^2 = \frac{RSS}{\mathrm{df}}$$

Including new variables: RSS and the degree of freedom are both decreasing (the number of normal equations is increasing)

Adjusted R^2: $1-\bar{R}^2 = \frac{n-1}{n-k-1}(1-R^2)$

$t < 1$: omitting a variable: $\bar{R}^2$ is increasing

$F < 1$: omitting more variables: $\bar{R}^2$ is increasing

Possible: different conclusions based on t and F (e.g. multicollinearity)

Model selection

Nested hypotheses: t- and F-test

Non-nested hyp., dependent variable is the same, e.g.:

$$R\&D = \alpha + \beta \log(revenue) + u$$

$$R\&D = \alpha + \beta_1\, revenue + \beta_2\, revenue^2 + u$$

based on adjusted R^2 or information criteria (e.g. AIC)

AIC (Akaike information criterion):

$$RSS \cdot exp(2(k + 1)/n)$$

Adjusted R^2, example

Wage survey (2003): does the experience or the age explain more in the wage equation?

Equation: EQ_EXP Workfile: TARIFA2003

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LOG(KER)
Method: Least Squares
Date: 11/23/08 Time: 08:59
Sample: 1 201971
Included observations: 201971

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.43052	0.005226	1995.866	0.0000
ISKVEG9	0.163335	0.000509	320.6889	0.0000
EXP01	0.020825	0.000368	56.66378	0.0000
EXP2	-0.000268	7.17E-06	-37.37194	0.0000

R-squared	0.341767	Mean dependent var	11.58047
Adjusted R-squared	0.341757	S.D. dependent var	0.592283
S.E. of regression	0.480532	Akaike info criterion	1.372175
Sum squared resid	46636.45	Schwarz criterion	1.372377
Log likelihood	-138565.8	F-statistic	34954.97
Durbin-Watson stat	0.922321	Prob(F-statistic)	0.000000

Equation: EQ_KOR Workfile: TARIFA2003

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LOG(KER)
Method: Least Squares
Date: 11/23/08 Time: 09:00
Sample: 1 201971
Included observations: 201971

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.986636	0.014081	709.2486	0.000
ISKVEG9	0.154711	0.000496	312.1709	0.000
KOR	0.031483	0.000672	46.84169	0.000
KOR^2	-0.000278	7.80E-06	-35.61278	0.000

R-squared	0.342334	Mean dependent var	11.5804
Adjusted R-squared	0.342324	S.D. dependent var	0.59228
S.E. of regression	0.480325	Akaike info criterion	1.37131
Sum squared resid	46596.24	Schwarz criterion	1.37151
Log likelihood	-138478.7	F-statistic	35043.2
Durbin-Watson stat	0.923453	Prob(F-statistic)	0.00000

Logarithmic forms

Log-log (loglinear) – elasticity

$$\ln(y) = \alpha + \beta \ln(x) + u$$

$$\%\Delta\hat{y} = \left(e^{\hat{\beta}} - 1\right) \cdot \%\Delta x \approx \hat{\beta} \cdot \%\Delta x$$

Partly logarithmic forms

$$y = \alpha + \beta \ln(x) + u \qquad \Delta\hat{y} \approx \frac{\hat{\beta}}{100} \%\Delta x$$

$$\ln(y) = \alpha + \beta x + u \qquad \%\Delta\hat{y} \approx 100\hat{\beta}\, \Delta x$$

Quadratic form

Increasing or decreasing partial effect

$$y = \alpha + \beta_1 x_1 + \beta_2 x_1^{\ 2} + u$$

$$\frac{\Delta \hat{y}}{\Delta x_1} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x_1$$

Example: wage survey (2003), quadratic function of experience, estimated equations:

log(Ker) = 9.83 + 0.135 ISKVEG9 + 0.0082 EXP

log(Ker) = 9.83 + 0.135 ISKVEG9 + 0.022 EXP – 0.00029 EXP^2

positive (but decreasing) partial effect for

0.022/(2*0.00029) = 39 years

Interactions

Partial effect depends on other explanatory variables as well:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_1 x_2 + u$$

$$\frac{\Delta \hat{y}}{\Delta x_1} \approx \hat{\beta}_1 + \hat{\beta}_2 x_2$$

Example: wage and education premium depend on the profitability of the firm (net sales revenue – material costs)

Log(wage) = 10.304 + 0.139 Educ9 + 0.092 Log(Profit)

Log(wage) = 10.597 + 0.079 Educ9 + 0.043 Log(Profit) + 0.010 (Educ9*Log(Profit))

Dummy variables on the right hand side

So far: mainly continuous variables (quantitative information) – e.g. wage, consumption, wealth, education (?)

Binary / dummy variables

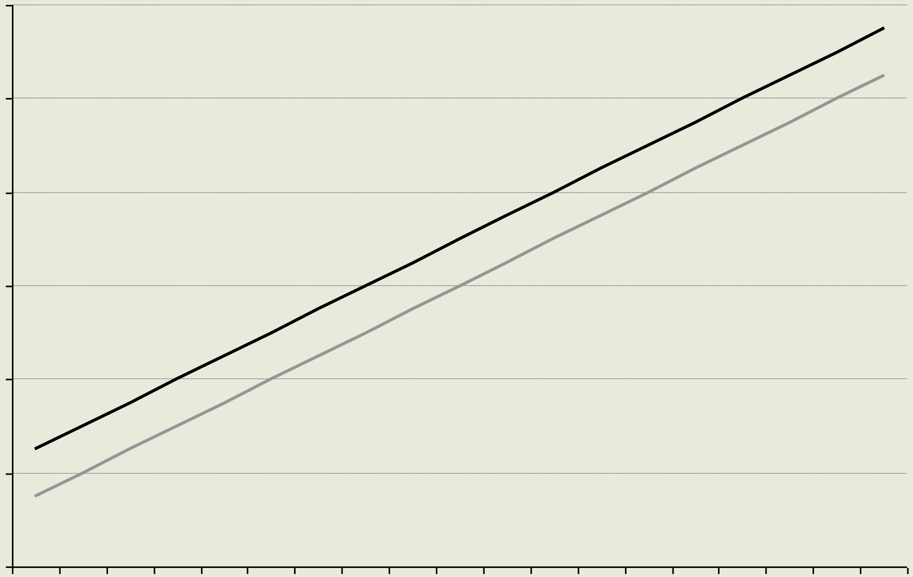
- Qualitative information
- Examples: gender, employed, country dummy…

Different intercepts – 2 groups

Example:
$$\log(wage_i) = \begin{cases} \alpha_1 + \beta Educ_i + u_i, \text{if from Budapest} \\ \alpha_2 + \beta Educ_i + u_i, \text{otherwise} \end{cases}$$

$$\Leftrightarrow \log(wage_i) = \alpha_1 + (\alpha_2 - \alpha_1) D_i + \beta Educ_i + u_i,$$

$$D_i = 0 \text{ if from Budapest}, D_i = 1 \text{ otherwise}$$



Different intercepts, example

Based on the 2003 wage survey

Dependent Variable: LOG(KER)
Method: Least Squares
Sample: 1 201971
Included observations: 199204
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.93263	0.003597	3039.442	0.0000
ISKVEG9	0.149569	0.000519	287.9107	0.0000
VIDEK	-0.164612	0.002906	-56.64754	0.0000

R-squared	0.331955	Mean dependent var	11.57456
Adjusted R-squared	0.331948	S.D. dependent var	0.590660

Log dependent variable

Estimated equation:

$$\log(Wage_i) = 10.93 - 0.16 Countryside_i + 0.15 Educ_i$$

Countryside: lower wage by approx. 16% (ceteris paribus)

Exact difference („log” is the natural logarithm in Eviews):

$$\log(Wage_1) - \log(Wage_0) = -0.16 = \log \frac{Wage_1}{Wage_0}$$

$$\Rightarrow \%\ \text{wage difference}: \left(e^{-0.16} - 1\right) \cdot 100 = -14.79$$

More than two groups

N groups (e.g. regions instead of Budapest / countryside)

$$y_i = \begin{cases} \alpha_1 + \beta x_i + u_i \text{ in Group 1} \\ \alpha_2 + \beta x_i + u_i \text{ in Group 2} \\ \vdots \\ \alpha_N + \beta x_i + u_i \text{ in Group N} \end{cases}$$

$$\Leftrightarrow y_i = \alpha_1 + (\alpha_2 - \alpha_1) D_2 + \ldots + (\alpha_N - \alpha_1) D_N + \beta x_i + u_i$$

$$D_2 = 1 \text{ in Group 2 (0 otherwise)}, \ldots,$$

$$D_N = 1 \text{ in Group N (0 otherwise)}$$

N-1 dummies in the regression (if there are N groups),

Group N: benchmark group!

Interactions between binary variables

Example: male / female wage gap is not the same in Budapest and in the countryside

Four categories:

	Budapest	countryside
female		
male		

Benchmark group: females in Budapest

Two equivalent models

$$\log(wage_i) = \alpha_0 + \alpha_1 Countryside\_Fem_i + \alpha_2 Bp\_Male_i + \alpha_3 Countryside\_Male_i + \\ + \beta Educ_i + u_i$$

$$\log(wage_i) = \alpha_0 + \alpha_1^* \cdot Male_i + \alpha_2^* \cdot Countryside_i + \alpha_3^* \cdot Countryside_i \cdot Male_i + \\ + \beta Educ_i + u_i$$

Interactions, example

Wage survey estimates (benchmark: females in Bp.):

Dependent Variable: LOG(KER)
Method: Least Squares
Included observations: 199204
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.90119	0.004225	2579.942	0.0000
VIDEK*(1-FFI)	-0.172604	0.003569	-48.36810	0.0000
(1-VIDEK)*FFI	0.053984	0.005444	9.916527	0.0000
VIDEK*FFI	-0.102617	0.003795	-27.03654	0.0000
ISKVEG9	0.150625	0.000524	287.6727	0.0000

R-squared	0.335102	Mean dependent var	11.57456
Adjusted R-squared	0.335089	S.D. dependent var	0.590660

Dependent Variable: LOG(KER)
Method: Least Squares
Included observations: 199204
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.90119	0.004225	2579.942	0.0000
VIDEK	-0.172604	0.003569	-48.36810	0.0000
FFI	0.053984	0.005444	9.916527	0.0000
VIDEK*FFI	0.016003	0.005902	2.711195	0.0067
ISKVEG9	0.150625	0.000524	287.6727	0.0000

R-squared	0.335102	Mean dependent var	11.57456
Adjusted R-squared	0.335089	S.D. dependent var	0.590660

$$-0,1026 = -0.1726 + 0.0540 + 0.0165$$

Non-constant slope parameters

In case of two groups

$$y_i = \begin{cases} \alpha + \beta_{11} x_{1i} + \beta_2 x_{2i} + u_i \text{ for Group 1} \\ \alpha + \beta_{12} x_{1i} + \beta_2 x_{2i} + u_i \text{ for Group 2} \end{cases}$$

$$\Leftrightarrow y_i = \alpha + \beta_{11} x_{1i} + (\beta_{12} - \beta_{11}) D_i x_{1i} + \beta_2 x_{2i} + u_i$$

$$D_i = 0 \text{ for Group 1, } D_i = 1 \text{ for Group 2}$$

→ Interaction of a dummy variable and an explanatory variable

Example

Effect of education gender-dependent but the effect of age is not

$$\log(Wage_i) = \alpha_0 + \beta_1 Educ_i + \beta_2 Educ_i \cdot Male_i + \beta_3 Age_i + u_i$$

Dependent Variable: LOG(KER)
Method: Least Squares
Included observations: 201971
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.45207	0.004941	2115.328	0.0000
ISKVEG9	0.149156	0.000504	295.9727	0.0000
ISKVEG9*FFI	0.014087	0.000440	31.99000	0.0000
KOR	0.007764	9.57E-05	81.14545	0.0000

R-squared	0.342540	Mean dependent var	11.58047
Adjusted R-squared	0.342530	S.D. dependent var	0.592283

Examining the stability of the coefficients

Example: is the wage model the same for males and females?

Cross sectional analysis (time series: stability of the coefficient in time)

$$\log(Wage_i) = \alpha_1 + \alpha_2 Male_i + \beta_1 Educ_i + \beta_2 Educ_i \cdot Male_i + \beta_3 Age_i + \beta_4 Age_i \cdot Male_i + u_i$$

$$\mathrm{H}_0 : \alpha_2 = 0, \beta_2 = 0, \beta_3 = 0$$

F-test (also possible for a subset of the restrictions)

Problem if $N_2 < k \rightarrow$ Chow-test (predictive) can be used (see week 5)

Examining the stability of the coefficients, cont.

Estimation results and test statistic

Dependent Variable: LOG(KER)
Method: Least Squares
Included observations: 201971
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.39458	0.006437	1614.696	0.0000
FFI	0.104694	0.009702	10.79114	0.0000
ISKVEG9	0.150973	0.000580	260.0792	0.0000
ISKVEG9*FFI	0.011260	0.001075	10.47171	0.0000
KOR	0.008858	0.000124	71.71457	0.0000
KOR*FFI	-0.002065	0.000191	-10.81957	0.0000

R-squared	0.342951	Mean dependent var	11.58047
Adjusted R-squared	0.342935	S.D. dependent var	0.592283

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	438.0596	(3, 201965)	0.0000
Chi-square	1314.179	3	0.0000

Dummy dependent variable

Examples:

- Labour market: employed or not
- Consumption: real estate owner or not
- Finance: bankruptcy of the borrower or not

Binary dependent variable ↔ linear model?

Linear probability model I.

y binary variable

$$E(y \mid x) = P(y = 1 \mid x) = \alpha + \beta x$$

$$\text{Estimated model}: y_i = \alpha + \beta x_i + u_i$$

$$\hat{y}: y = 1 \text{ estimated prob.}$$

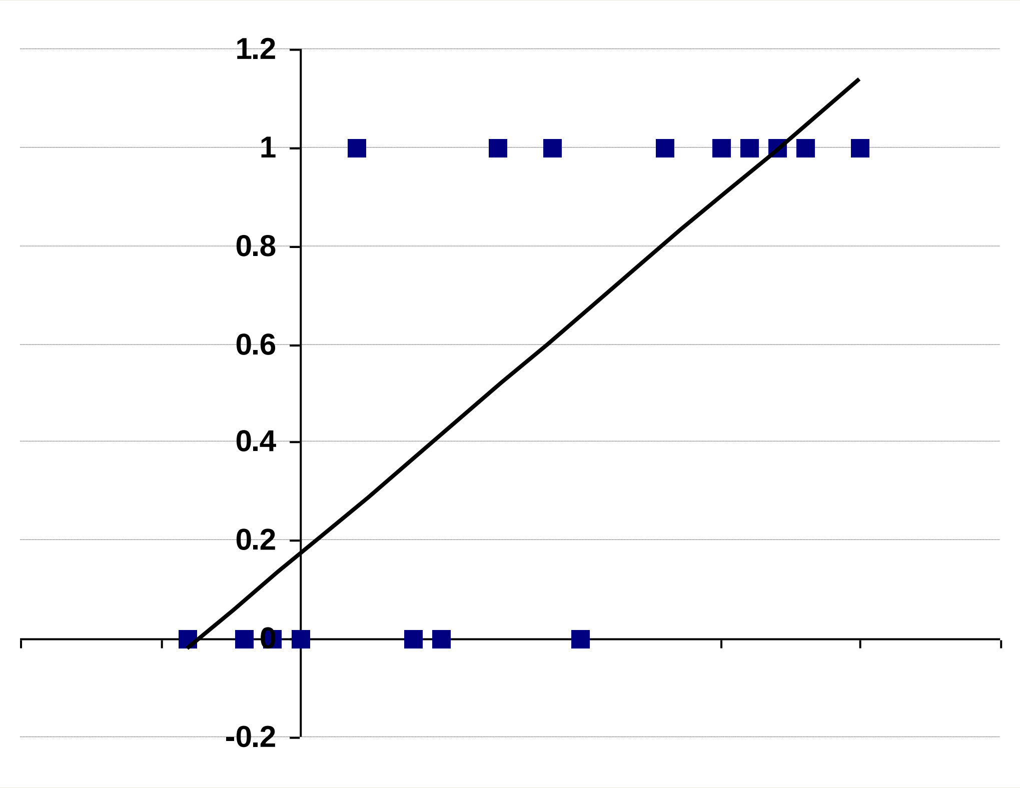
Nonlinear model: $P(y = 1 \mid x) = F(\beta x)$

If F is the Gaussian distribution function then the probit,

if $F(z) = e^z/(1 + e^z)$ then the logit model is obtained.

Linear probability model II.

Problem 1: estimated probability may lie outside the [0,1] interval



Linear probability model III.

Problem 2: heteroscedasticity

$$y_i = \beta x_i + u_i, \; E(u_i) = 0$$

$$u_i = \begin{cases} -\beta x_i & (1-\beta x_i) \text{ prob.} \\ 1-\beta x_i & \beta x_i \text{ prob.} \end{cases}$$

$$\Rightarrow \mathrm{Var}(u_i) = (1-\beta x_i)(-\beta x_i)^2 + \beta x_i (1-\beta x_i)^2 = \beta x_i (1-\beta x_i) = \hat{y}_i (1-\hat{y}_i)$$

Solution:

Using robust SE

Weighted LS: $w_i = \hat{y}_i (1-\hat{y}_i)$

Linear probability model, example

Dependent variable: whether the person has a private health insurance or not (SHARE database)

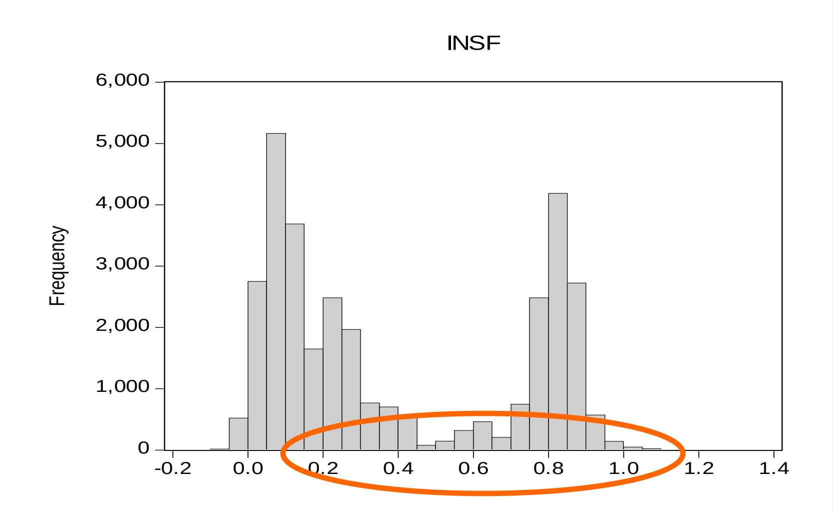
Explanatory variables: wealth, income, age, education, country dummies

Dependent Variable: INS
Method: Least Squares
Included observations: 32405
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.438462	0.015836	27.68759	0.0000
HNETW/10^6	0.008275	0.002079	3.980982	0.0001
HGTINC/10^6	0.445824	0.041066	10.85636	0.0000
AGE	-0.003281	0.000192	-17.05036	0.0000
EDU	0.000697	0.000280	2.487641	0.0129
COUNTRY=GERM	-0.026880	0.011713	-2.294858	0.0217
COUNTRY=SWE	-0.156186	0.010539	-14.81915	0.0000
COUNTRY=NETH	0.568132	0.011364	49.99388	0.0000
COUNTRY=SPAIN	-0.135671	0.010888	-12.46093	0.0000
COUNTRY=ITALY	-0.190969	0.010249	-18.63311	0.0000
COUNTRY=FRA	0.596495	0.011187	53.32010	0.0000
COUNTRY=DENM	0.115167	0.014577	7.900561	0.0000
COUNTRY=GRE	-0.186449	0.010237	-18.21386	0.0000
COUNTRY=SWITZ	0.345743	0.017138	20.17374	0.0000
COUNTRY=BELG	0.546137	0.011148	48.98888	0.0000

R-squared	0.464462	Mean dependent var	0.395957
Adjusted R-squared	0.464230	S.D. dependent var	0.489063

Histogram of predicted prob.



Seminar

Multivariate regression III.

Exercise: estimation of a wage equation based on a small sample from the wage survey

Variables

- Educ (years of education)
- Exp (experience)
- Wage
- Typ (type of settlement – qualitative variable)
- Bp (Budapest dummy)
- Male (male dummy)

Estimation of the wage equation I.

Model 1: modelling log(wage) in the private sector with the educ, exp, exp^2, bp, male variables and with the interaction of educ, exp with male

- Does the equation for males differ significantly from the equation for females?
  - Joint test of Male, Educ*Male and Exp*Male
- Experience-profile for Budapest males with 12 years of experience
  - Where is the maximum?
  - Graphical presentation of the experience-profile with confidence interval

Estimation of the wage equation II.

Model 2: previous model + dummies for “chief town of the county” and for “other town”

- Testing the equality of the two new coefficients with three methods
  - Directly
  - By a t-test after transformation
  - By comparing the R^2 of the restricted and unrestricted model
- Testing heteroscedasticity with the White- and Breusch–Pagan-tests
- Calculating robust standard errors and comparing them with the non-robust ones

Exercise: estimation of labour market participation with a linear probability model

Dependent variable: economically active or not

Explanatory variables: education, experience, age, kid below / over 6 years

OLS estimation with usual and with robust standard errors, WLS estimation

Forecasting probabilities