Econometrics - Lecture 1

Econometrics – First Steps

Contents

- Organizational Issues
- Some History of Econometrics
- An Introduction to Linear Regression
  - OLS: An Algebraic Tool
  - The Linear Regression Model
  - Small Sample Properties of the OLS Estimator
- Introduction to GRETL

Organizational Issues

Course schedule

Class	Date
1	Fr, Sept 29
2	Fr, Oct 6
3	Fr, Oct 13
4	Fr, Oct 27
5	Fr, Nov 3
6	Fr, Nov 10

Time: 10:00-14:00

Organizational Issues, cont’d

Aims of the course

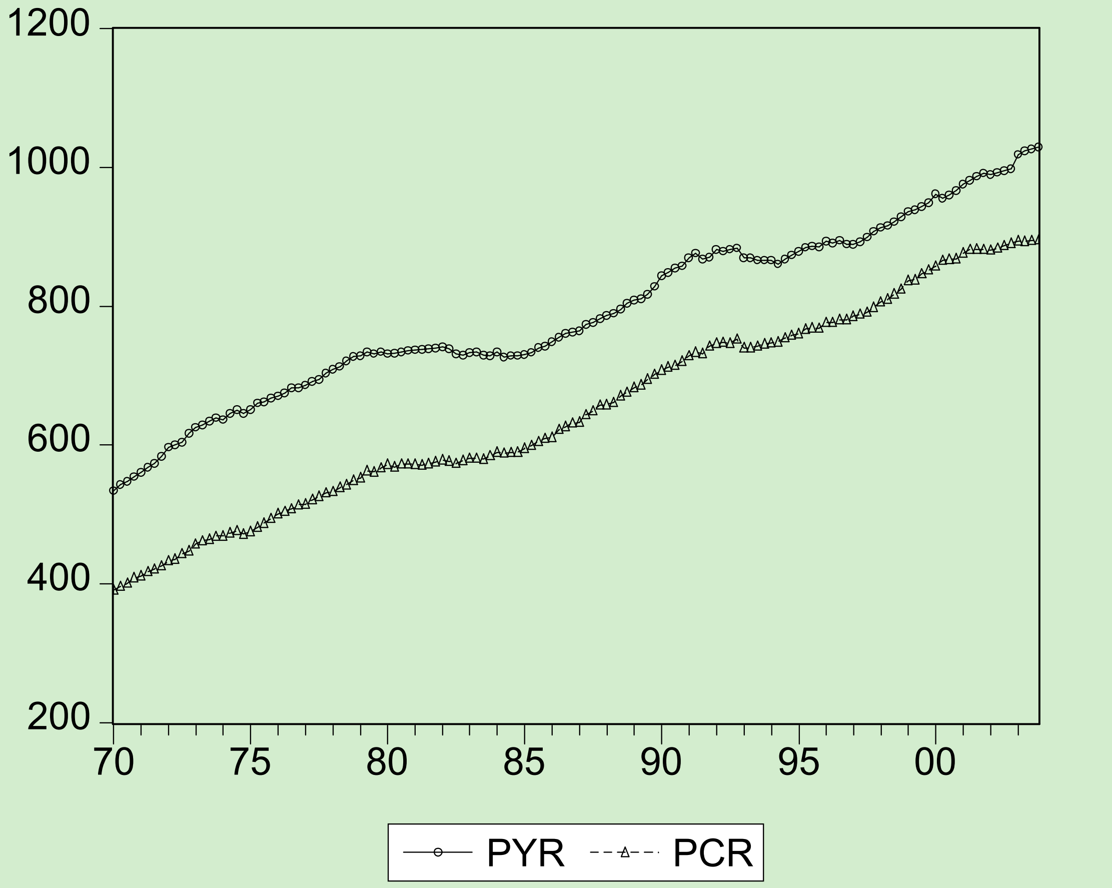
- Use of econometric tools for analyzing economic data: specification of adequate models, identification of appropriate econometric methods, estimation of model parameters, interpretation of results
- Introduction to commonly used econometric tools and techniques
- Understanding of econometric concepts and principles
- Use of GRETL

Example: Individual Wages

Sample (US National Longitudinal Survey, 1987)

- *N* = 3294 individuals (1569 females)
- Variable list
  - WAGE: wage (in 1980 $) per hour (p.h.)
  - MALE: gender (1 if male, 0 otherwise)
  - SCHOOL: years of schooling
  - EXPER: experience in years
  - AGE: age in years
- Questions of interest
  - Effect of gender on wage p.h.: Average wage p.h.: 6,31$ for males, 5,15$ for females
  - Effects of education, of experience, of interactions, etc. on wage p.h.

Example: Income and Consumption



PCR: Private Consumption, real, in bn. EUROs

PYR: Household's Disposable Income, real, in bn. EUROs

1970:1-2003:4
136 observations

Basis: 1995

Source: AWM-Database

Organizational Issues, cont'd

Literature

Course textbook

- Marno Verbeek, *A Guide to Modern Econometrics*, 4rd ed., Wiley, 2012

Suggestions for further reading

- Peter Kennedy, *A guide to econometrics*. 6th ed., Blackwell, 2008.
- William H. Greene, *Econometric Analysis*. 7th Ed., Prentice Hall, 2011

Organizational Issues, cont'd

Prerequisites

- Linear algebra: linear equations, matrices, vectors (basic operations and properties)
- Descriptive statistics: measures of central tendency, measures of dispersion, measures of association, frequency tables, histogram, scatter plot, quantile
- Theory of probability: probability and its properties, random variables and distribution functions in one and in several dimensions, moments, convergence of random variables, limit theorems, law of large numbers
- Mathematical statistics: point estimation, confidence interval, hypothesis testing, *p*-value, significance level

Organizational Issues, cont'd

Teaching and learning method

- Course in six blocks
- Class discussions, written homework (computer exercises, GRETL) submitted by groups of (3-5) students, presentations of homework by participants
- Final exam

Assessment of student work

- For grading, the written homework, presentation of homework in class, and a final written exam will be of relevance
- Weights: homework 40 %, final written exam 60 %
- Presentation of homework in class: students must be prepared to be called at random

Contents

Empirical Economics Prior to 1930ies

The situation in the early 1930ies

- Theoretical economics aims at “operationally meaningful theorems“; “operational” means purely logical mathematical deduction
- Economic theories or laws are seen as deterministic relations; no inference from data as part of economic analysis
- Ignorance of the stochastic nature of economic concepts
- Data: limited availability; time-series on agricultural commodities, foreign trade
- Use of statistical methods for
  - measuring theoretical coefficients, e.g., demand elasticities
  - representing business cycles

Early Institutions

- Applied demand analysis: US Bureau of Agricultural Economics
- Statistical analysis of business cycles: H.L.Moore (Columbia University): Fourier periodogram; W.M.Persons et al. (Harvard): business cycle forecasting; US National Bureau of Economic Research (NBER)
- Cowles Commission for Research in Economics
  - Founded 1932 by A.Cowles: determinants of stock market prices?
  - Formalization of econometrics, development of econometric methodology
  - R.Frisch, G.Tintner; European refugees
  - J.Marschak (head 1943-55) recruited people like T.C.Koopmans, T.M.Haavelmo, T.W.Anderson, L.R.Klein
  - Interests shifted to theoretical and mathematical economics after 1950

Early Actors

- R.Frisch (Oslo Institute of Economic Research): econometric project, 1930-35; T.Haavelmo, O.Reiersol
- J.Tinbergen (Dutch Central Bureau of Statistics, Netherlands Economic Institute; League of Nations, Genova): macro-econometric model of Dutch economy, ~1935; T.C.Koopmans, H.Theil
- Austrian Institute for Trade Cycle Research (Österreichisches Institut für Konjunkturforschung, 1927, F.v.Hayek, L.v.Mises): O.Morgenstern (head), A.Wald, G.Tintner
- Econometric Society, founded 1930 by R.Frisch et al.
  - Facilitates exchange of scholars from Europe and US
  - Dealing with econometrics and mathematical statistics

First Steps

- R.Frisch, J.Tinbergen:
  - Macro-economic modelling based on time-series, ~ 1935
  - Aiming at measuring parameters, e.g., demand elasticities
  - Aware of problems due to quality of data
  - Nobel Memorial Prize in Economic Sciences jointly in 1969 ("for having developed and applied dynamic models for the analysis of economic processes")
- T.Haavelmo
  - "The Probability Approach in Econometrics": PhD thesis (1946)
  - Econometrics as a tool for testing economic theories
  - Nobel Memorial Prize in Economic Sciences in 1989 ("for his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures")

First Steps, cont'd

- Cowles Commission
  - Formalization of econometrics, development of econometric methodology
  - Methodology for macro-economic modelling based on Haavelmo's approach
  - Cowles Commission monographs by G.Tintner, T.C.Koopmans, et al.

The Haavelmo Revolution

- Introduction of probabilistic concepts in economics
  - Obvious deficiencies of traditional approach: Residuals, measurement errors, omitted variables; stochastic time-series data
  - Advances in probability theory in early 1930ies
  - Fisher's likelihood function approach
- Haavelmo's ideas
  - Critical view of Tinbergen's macro-econometric models
  - Thorough adoption of probability theory in econometrics
  - Conversion of deterministic economic models into stochastic structural equation models
- Haavelmo's "The Probability Approach in Econometrics"
  - Why is the probability approach indispensable?
  - Modelling procedure based on ML estimation and hypothesis testing

Cowles Commission Methodology

Assumptions based to macro-econometric modelling and testing of economic theories

Time series model

$$Y_t = \alpha X_t + \beta W_t + u_{1t}$$

$$X_t = \gamma Y_t + \delta Z_t + u_{2t}$$

1. Specification of the model equation(s) includes the choice of variables; functional form is (approximately) linear
2. Time-invariant model equation(s): the model parameters α, …, δ are independent of time *t*
3. Parameters α, …, δ are structurally invariant, i.e., invariant wrt changes in the variables
4. Causal ordering (exogeneity, endogeneity) of variables is known
5. Statistical tests can falsify but not verify a model

Classical Econometrics and More

- "Golden age" of econometrics until ~1970
  - Multi-equation models for analyses and forecasting
  - Growing computing power
  - Development of econometric tools
- Skepticism
  - Poor forecasting performance
  - Dubious results due to
    - wrong specifications
    - imperfect estimation methods

Model	year	eq's
Tinbergen	1936	24
Klein	1950	6
Klein & Goldberger	1955	20
Brookings	1965	160
Brookings Mark II	1972	~200

- Time-series econometrics: non-stationarity of economic time-series
  - Consequences of non-stationarity: misleading *t*-, DW-statistics, R^2
  - Non-stationarity: needs new models (ARIMA, VAR, VEC); Box & Jenkins (1970: ARIMA-models), Granger & Newbold (1974, spurious regression), Dickey-Fuller (1979, unit-root tests)

Econometrics …

- … consists of the application of statistical data and techniques to mathematical formulations of economic theory. It serves to test the hypotheses of economic theory and to estimate the implied interrelationships. (Tinbergen, 1952)
- … is the interaction of economic theory, observed data and statistical methods. It is the interaction of these three that makes econometrics interesting, challenging, and, perhaps, difficult. (Verbeek, 2008)
- … is a methodological science with the elements
  - economic theory
  - mathematical language
  - statistical methods
  - computer science

  aiming to give empirical content to economic relations. (Pesaran, 1987)

Our Course

1. Introduction to linear regression (Verbeek, Ch. 2): the linear regression model, OLS method, properties of OLS estimators
2. Introduction to linear regression (Verbeek, Ch. 2): goodness of fit, hypotheses testing, multicollinearity
3. Interpreting and comparing regression models (Verbeek, Ch. 3): interpretation of the fitted model, selection of regressors, testing the functional form
4. Heteroskedascity and autocorrelation (Verbeek, Ch. 4): causes and consequences, testing, alternatives for inference
5. Endogeneity, instrumental variables and GMM (Verbeek, Ch. 5): the IV estimator, the generalized instrumental variables estimator, the generalized method of moments (GMM)
6. The practice of econometric modelling

Econometrics 2: An Advanced Course

- Univariate and multivariate time series models: ARMA-, ARCH-, GARCH-models, VAR-, VEC-models
- Models for panel data
- Models with limited dependent variables: binary choice, count data

Contents

- Organizational Issues
- Some History of Econometrics
- An Introduction to Linear Regression
  - OLS: An Algebraic Tool
  - The Linear Regression Model
  - Small Sample Properties of the OLS Estimator
- Introduction to GRETL

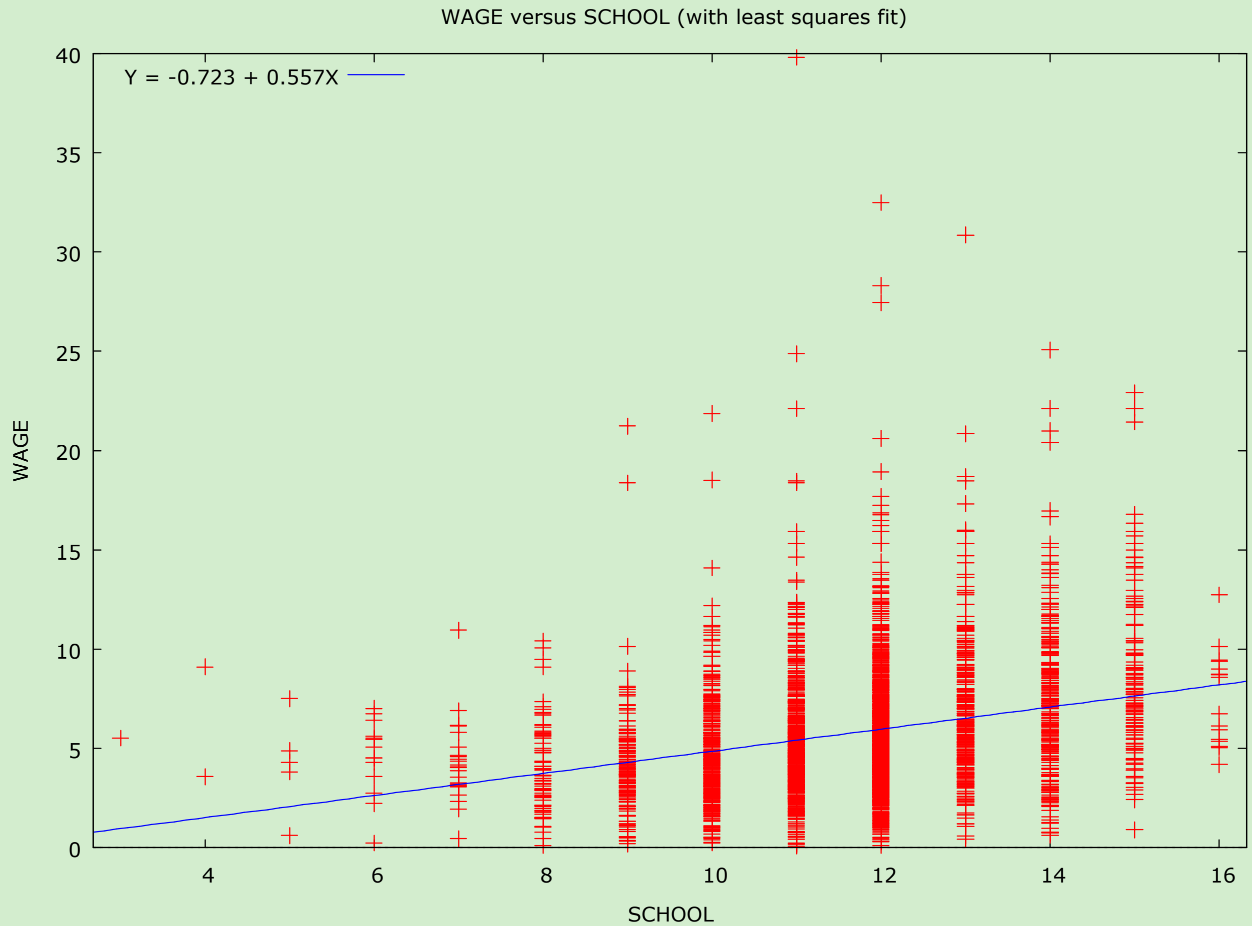
Example: Individual Wages

Sample (US National Longitudinal Survey, 1987)

- N = 3294 individuals (1569 females)
- Variable list
  - WAGE: wage (in 1980 $) per hour (p.h.)
  - MALE: gender (1 if male, 0 otherwise)
  - SCHOOL: years of schooling
  - EXPER: experience in years
  - AGE: age in years
- Possible questions
  - Effect of gender on wage p.h.: Average wage p.h.: 6,31$ for males, 5,15$ for females
  - Effects of education, of experience, of interactions, etc. on wage p.h.

Individual Wages, cont'd

Wage per hour vs. Years of schooling



Linear Regression

Y: explained variable
X: explanatory or regressor variable
The linear regression model describes the data-generating process of *Y* under the condition *X*

simple linear regression model

$$Y = \beta_1 + \beta_2 X$$

β_2: coefficient of *X*
β_1 : intercept

multiple linear regression model

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

Fitting a Model to Data

Choice of values b_1, b_2 for model parameters β_1, β_2 of $Y = \beta_1 + \beta_2 X$, given the observations (y_i, x_i), $i = 1,\ldots,N$

Principle of (Ordinary) Least Squares or OLS:

$$b_i = \arg\min_{\beta 1,\ \beta 2} S(\beta_1, \beta_2),\ i = 1,2$$

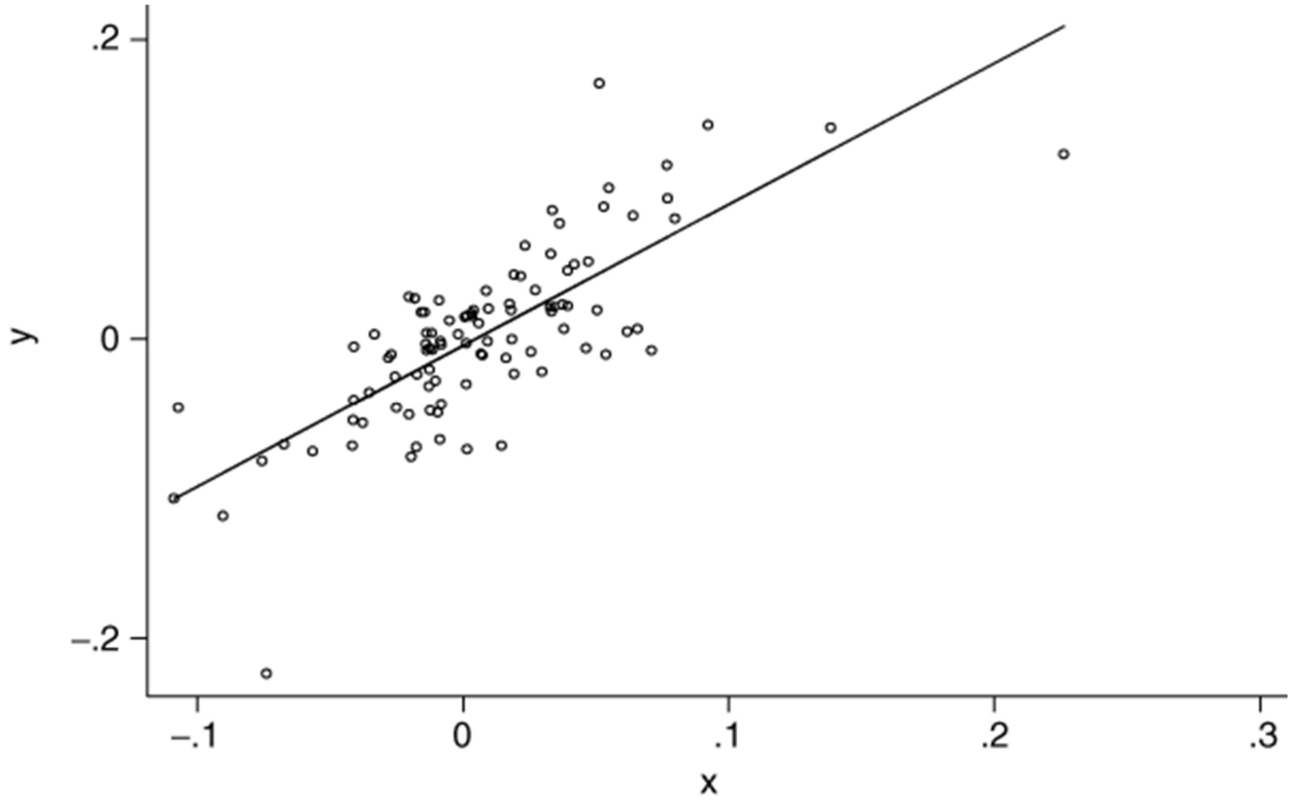
Objective function: sum of the squared deviations

$$S(\beta_1, \beta_2) = \Sigma_i\, [y_i - (\beta_1 + \beta_2 x_i)]^2 = \Sigma_i\, \varepsilon_i^2$$

Deviation between observation and fitted value: $\varepsilon_i = y_i - (\beta_1 + \beta_2 x_i)$

Observations and Fitted Regression Line

Simple linear regression: Fitted line and observation points (Verbeek, Figure 2.1)



OLS Estimators

Equating the partial derivatives of $S(\beta_1, \beta_2)$ to zero: **normal equations**

$$Nb_1 + b_2 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i$$

$$b_1 \sum_{i=1}^{N} x_i + b_2 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i$$

OLS estimators b_1 und b_2 result in

$$b_2 = \frac{s_{xy}}{s_x^2}$$

$$b_1 = \bar{y} - b_2 \bar{x}$$

with mean values $\bar{x}$ and $\bar{y}$

and second moments

$$s_{xy} = \frac{1}{N-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$s_x^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

Individual Wages, cont'd

Sample (US National Longitudinal Survey, 1987): wage per hour, gender, experience, years of schooling; N = 3294 individuals (1569 females)

Average wage p.h.: 6,31$ for males, 5,15$ for females

Model:

$$wage_i = \beta_1 + \beta_2 \, male_i + \varepsilon_i$$

$male_i$: male dummy, has value 1 if individual is male, otherwise value 0

OLS estimation gives

$$wage_i = 5{,}15 + 1{,}17 * male_i$$

Compare with averages!

Individual Wages, cont'd

OLS estimated wage equation (Table 2.1, Verbeek)

Dependent variable: *wage*

Variable	Estimate	Standard error
constant	5.1469	0.0812
male	1.1661	0.1122
$s = 3.2174$	$R^2 = 0.0317$	$F = 107.93$

$wage_i = 5{,}15 + 1{,}17 * male_i$

estimated wage p.h for males: 6,313

for females: 5,150

OLS Estimators: General Case

Model for Y contains K-1 explanatory variables

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_K X_K = x'\beta$$

with $x = (1, X_2, \ldots, X_K)'$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_K)'$

Observations: $(y_i, x_i') = (y_i, (1, x_{i2}, \ldots, x_{iK}))$, $i = 1, \ldots, N$

OLS estimates $b = (b_1, b_2, \ldots, b_K)'$ are obtained by minimizing the objective function wrt the β_k's

$$S(\beta) = \sum_{i=1}^{N} (y_i - x_i'\beta)^2$$

this results in

$$-2\sum_{i=1}^{N} x_i (y_i - x_i'b) = 0$$

OLS Estimators: General Case, cont'd

or

$$\left(\sum_{i=1}^{N} x_i x_i'\right) b = \sum_{i=1}^{N} x_i y_i$$

the **normal equations**, a system of *K* linear equations for the components of *b*

Given that the symmetric *K*x*K*-matrix $\sum_{i=1}^{N} x_i x_i'$ has full rank *K* and is hence invertible, the OLS estimators are

$$b = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i y_i$$

Best Linear Approximation

Given the observations: $(y_i, x_i') = (y_i, (1, x_{i2}, \ldots, x_{iK}))$, $i = 1, \ldots, N$

For y_i, the linear combination or the fitted value

$$\hat{y}_i = x_i' b$$

is the best linear combination for Y from $X_2, \ldots, X_K$ and a constant (the intercept)

Some Matrix Notation

N observations

$(y_1,x_1), \dots , (y_N,x_N)$

Model: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1, \dots, N$, or

$$y = X\beta + \varepsilon$$

with

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

OLS Estimators in Matrix Notation

Minimizing

$$S(\beta) = (y - X\beta)' (y - X\beta) = y'y - 2y'X\beta + \beta' X'X\beta$$

with respect to β gives the normal equations

$$\frac{\partial S(\beta)}{\partial \beta} = -2(X'y - X'Xb) = 0$$

resulting from differentiating S(β) with respect to β and setting the first derivative to zero

The vector *b* of OLS solution or OLS estimators for β is

$$b = (X'X)^{-1}X'y$$

The best linear combinations or **predicted values** for *Y* given *X* or projections of *y* into the space of *X* are obtained as

$$\hat{y} = Xb = X(X'X)^{-1}X'y = P_x y$$

the *N*x*N*-matrix P_x is called the **projection matrix** or **hat matrix**

Residuals in Matrix Notation

The vector y can be written as $y = Xb + e = \hat{y} + e$ with residuals

$$e = y - Xb \text{ or } e_i = y_i - x_i'b,\ i = 1, \ldots, N$$

- From the normal equations follows

  $$-2(X'y - X'Xb) = -2\, X'e = 0$$

  i.e., each column of X is orthogonal to e
- With

  $$e = y - Xb = y - P_x y = (I - P_x)y = M_x y$$

  the **residual generating matrix** M_x is defined as

  $$M_x = I - X(X'X)^{-1}X' = I - P_x$$

  M_x projects y into the orthogonal complement of the space of X
- Properties of P_x and M_x: symmetry ($P'_x = P_x$, $M'_x = M_x$) idempotence ($P_x P_x = P_x$, $M_x M_x = M_x$), and orthogonality ($P_x M_x = 0$)

Properties of Residuals

Residuals: $e_i = y_i - x_i'b$, $i = 1, \ldots, N$

- Minimum value of objective function

  $$S(b) = e'e = \Sigma_i e_i^2$$

- From the orthogonality of $e = (e_1, \ldots, e_N)'$ to each $x_i = (x_{1i}, \ldots, x_{Ni})'$, $i = 1, \ldots, K$, i.e., $e'x_i = 0$, follows that

  $$\Sigma_i e_i = 0$$

  i.e., average residual is zero, if the model has an intercept

Contents

US Wages

US wages are gender-specific

The relation

$$wage_i = \beta_1 + \beta_2\, male_i + \varepsilon_i$$

with $male_i$: male dummy (equals 1 for males, otherwise 0)

- describes the wage of individual *i* as a function of its gender
- is assumed to be true for all US citizens

Given sample data ($wage_i$, $male_i$, $i = 1,\dots N$), OLS estimation of β_1 and β_2 may result in

$$wage_i = 5{,}15 + 1{,}17 * male_i$$

- This is not (only) a description of the sample!
- But reflects a general relationship

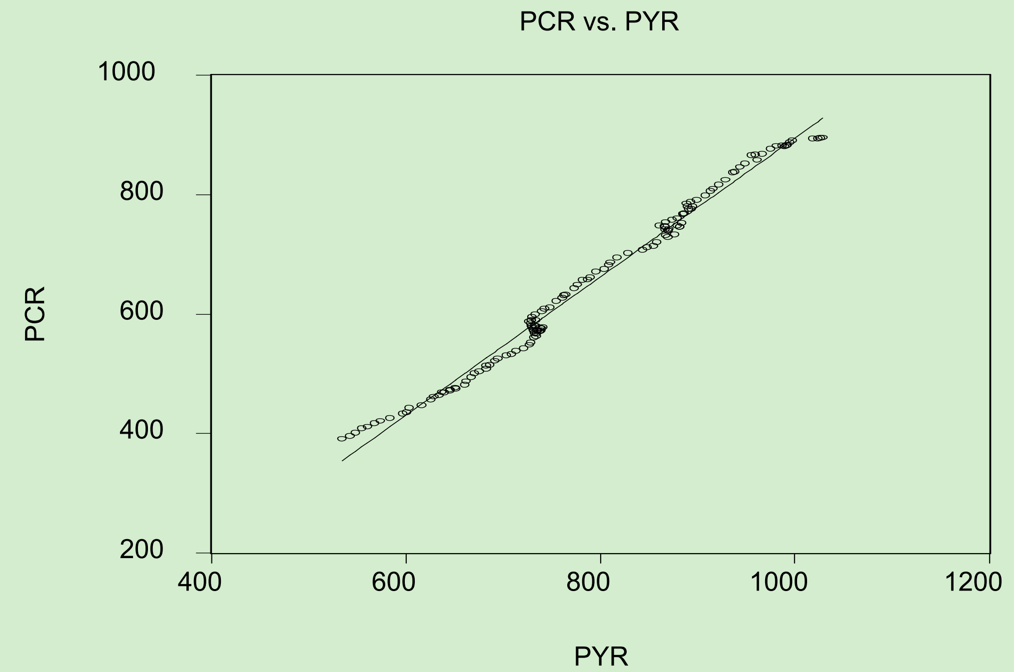
Income and Consumption

PCR: private consumption, real
PYR: household's disposable income, real,
AWM-Database, 1970:1-2003:4, Basis: 1995

Consumption function

$$PCR_t = \beta_1 + \beta_2 \, PYR_t + \varepsilon_t$$

- describes consumption in the Euro-zone



Economic Models

Describe economic relationships (not just a set of observations), have an economic interpretation

Linear regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i = x_i'\beta + \varepsilon_i$$

- Variables $Y, X_2, \dots, X_K$: observable
- Error term ε_i (disturbance term) contains all influences that are not included explicitly in the model; not observable; assumption $E\{\varepsilon_i \mid x_i\} = 0$ gives

  $$E\{y_i \mid x_i\} = x_i'\beta$$

  the model describes the expected value of y given x
- Sample ($y_i, x_{i2}, \dots, x_{iK}, i = 1, \dots, N$) from a well-defined population
- Unknown coefficients $\beta_1, \dots, \beta_K$: population parameters

Sampling in the Economic Context

The regression model $y_i = x_i'\beta + \varepsilon_i$, $i = 1, \ldots, N$; or $y = X\beta + \varepsilon$

describes one realization out of all possible samples of size N from the population

A) Sampling process with fixed, i.e., non-stochastic x_i's

- New sample: new error terms ε_i, $i = 1, \ldots, N$, and, hence, new y_i's
- Joint distribution of ε_i's determines properties of b etc.
- A laboratory setting, does not apply to the economic context

B) Sampling process with samples of (x_i, y_i) or (x_i, ε_i)

- New sample: new error terms ε_i and new x_i, $i = 1, \ldots, N$
- Random sampling of (x_i, ε_i), $i = 1, \ldots, N$: joint distribution of (x_i, ε_i)'s determines properties of b etc.

Sampling in the Economic Context, cont'd

- The sampling with fixed, non-stochastic x_i's is not realistic for economic data
- Sampling process with samples of (x_i, y_i) is appropriate for modeling cross-sectional data
  - Example: household surveys, e.g., US National Longitudinal Survey, EU-SILC
- Sampling process with samples of (x_i, y_i) from time-series data: sample is seen as one out of all possible realizations of the underlying data-generating process
  - Example: time series PYR and PCR of the AWM-Database

Assumptions of the Linear Regression Model

The linear regression model $y_i = x_i'\beta + \varepsilon_i$ makes use of assumptions

- Assumption for ε_i's: $E\{\varepsilon_i \mid x_i\} = 0$; exogeneity of variables X
  - X contains no information on the error term ε
  - $E\{\varepsilon_i \mid x_i\} = 0$ implies that ε_i and x_i are uncorrelated
- This implies

  $$E\{y_i \mid x_i\} = x_i'\beta$$

  i.e., the regression line describes the conditional expectation of y_i given x_i
- Coefficient β_k measures the change of the expected value of Y if X_k changes by one unit and all other X_j values, $j \neq k$, remain the same (ceteris paribus condition)

Regression Coefficients

Linear regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i = x_i'\beta + \varepsilon_i$$

Coefficient β_k measures the change of the expected value of Y if X_k changes by one unit and all other X_j values, $j \neq k$, remain the same (ceteris paribus condition); marginal effect of changing X_k on Y

$$\frac{\partial E\{y_i | x_i\}}{\partial x_{ik}} = \beta_k$$

Example

- Wage equation: $wage_i = \beta_1 + \beta_2\, male_i + \beta_3\, school_i + \beta_4\, exper_i + \varepsilon_i$

  β_3 measures the impact of one additional year at school upon a person's wage, keeping gender and years of experience fixed

Estimation of β

Given a sample (x_i, y_i), i = 1, …, N, the OLS estimators for β

$$b = (X'X)^{-1}X'y$$

can be used as an approximation for β

- The vector b is a vector of numbers, the estimates
- The vector b is the realization of a vector of random variables
- The sampling concept and assumptions on ε_i's determine the quality, i.e., the statistical properties, of b

Contents

Fitting Economic Models to Data

Observations allow

- to estimate parameters
- to assess how well the data-generating process is represented by the model, i.e., how well the model coincides with reality
- to improve the model if necessary

Fitting a linear regression model to data provides

- parameter estimates $b = (b_1, \ldots, b_K)'$ for coefficients $\beta = (\beta_1, \ldots, \beta_K)'$
- standard errors $se(b_k)$ of the estimates b_k, $k=1,\ldots,K$
- t-statistics, F-statistic, R^2, Durbin Watson test-statistic, etc.

Individual Wages, cont'd

Wage equation with three regressors (Table 2.2, Verbeek)

Table 2.2 OLS results wage equation

Dependent variable: *wage*

Variable	Estimate	Standard error	t-ratio
constant	−3.3800	0.4650	−7.2692
male	1.3444	0.1077	12.4853
school	0.6388	0.0328	19.4780
exper	0.1248	0.0238	5.2530

$s = 3.0462 \quad R^2 = 0.1326 \quad \bar{R}^2 = 0.1318 \quad F = 167.63$

OLS Estimator and OLS Estimates *b*

OLS estimates *b* are a realization of the OLS estimator

The OLS estimator is a random variable

- Observations are a random sample from the population
- Observations are generated by some random sampling process

Distribution of the OLS estimator

- Actual distribution not known
- Distribution determined by assumptions on
  - model specification
  - the error term ε_i and regressor variables x_i

Quality criteria (bias, accuracy, efficiency) of OLS estimates are determined by the properties of this distribution

Gauss-Markov Assumptions

Observation y_i is a linear function

$$y_i = x_i'\beta + \varepsilon_i$$

of observations x_{ik} of the regressor variables X_k, $k = 1, \ldots, K$, and the error term ε_i

for $i = 1, \ldots, N$; $x_i' = (x_{i1}, \ldots, x_{iK})$; $X = (x_{ik})$

A1	$E\{\varepsilon_i\} = 0$ for all i
A2	all ε_i are independent of all x_i (exogenous x_i)
A3	$V\{\varepsilon_i\} = \sigma^2$ for all i (homoskedasticity)
A4	$Cov\{\varepsilon_i, \varepsilon_j\} = 0$ for all i and j with $i \neq j$ (no autocorrelation)

In matrix notation: $E\{\varepsilon\} = 0$, $V\{\varepsilon\} = \sigma^2 I_N$

Systematic Part of the Model

The systematic part $E\{y_i \mid x_i\}$ of the model $y_i = x_i'\beta + \varepsilon_i$, given observations x_i, is derived under the Gauss-Markov assumptions as follows:

(A2) implies $E\{\varepsilon \mid X\} = E\{\varepsilon\} = 0$ and $V\{\varepsilon \mid X\} = V\{\varepsilon\} = \sigma^2 I_N$

- Observations x_i, $i = 1, \ldots, N$, do not affect the properties of ε
- The systematic part

  $$E\{y_i \mid x_i\} = x_i'\beta$$

  can be interpreted as the conditional expectation of y_i, given observations x_i

Is the OLS Estimator a Good Estimator?

- Under the Gauss-Markov assumptions, the OLS estimator has favourable properties; see below
- Gauss-Markov assumptions are very strong but not always satisfied
- Relaxations of the Gauss-Markov assumptions and consequences of such relaxations are important topics in econometrics

Properties of OLS Estimators

1. The OLS estimator b is unbiased: $E\{b \mid X\} = E\{b\} = \beta$

   Needs assumptions (A1) and (A2)

2. The variance of the OLS estimator b is given by

   $$V\{b \mid X\} = V\{b\} = \sigma^2(\Sigma_i x_i x_i')^{-1} = \sigma^2(X'X)^{-1}$$

   Needs assumptions (A1), (A2), (A3) and (A4)

3. Gauss-Markov Theorem: The OLS estimator b is a BLUE[1)] (best linear unbiased estimator) for β

   Needs assumptions (A1), (A2), (A3), and (A4) and requires linearity in parameters

---

1) OLS estimator is most accurate among linear unbiased estimators; see next slide

The Gauss-Markov Theorem

OLS estimator b is BLUE (best linear unbiased estimator) for β

- Linear estimator: $b^* = Ay$ with any full-rank KxN matrix A
- b^* is an unbiased estimator: $E\{b^*\} = E\{Ay\} = \beta$
- b is BLUE: $V\{b^*\} - V\{b\}$ is positive semi-definite, i.e., the variance of any linear combination $d'b^*$ is not smaller than that of $d'b$

  $$V\{d'b^*\} \geq V\{d'b\}$$

  e.g., $V\{b_k^*\} \geq V\{b_k\}$ for any k
- The OLS estimator is most accurate among the linear unbiased estimators

Standard Errors of OLS Estimators

- Variance (covariance matrix) of the OLS estimators:

  $$V\{b\} = \sigma^2(X' X)^{-1} = \sigma^2(\Sigma_i x_i x_i')^{-1}$$

- Standard error of OLS estimate b_k: The square root of the k^{th} diagonal element of $V\{b\}$
- $V\{b\}$ is proportional to the variance σ^2 of the error terms
- Estimator for σ^2: sampling variance s^2 of the residuals e_i

  $$s^2 = (N - K)^{-1} \Sigma_i e_i^2$$

  Under assumptions (A1)-(A4), s^2 is unbiased for σ^2

  Attention: the estimator $(N - 1)^{-1} \Sigma_i e_i^2$ is biased
- Estimated variance (covariance matrix) of b:

  $$\tilde{V}\{b\} = s^2(X' X)^{-1} = s^2(\Sigma_i x_i x_i')^{-1}$$

Estimated Standard Errors of OLS Estimators

- Variance (covariance matrix) of the OLS estimators:

  $$V\{b\} = \sigma^2(X'X)^{-1} = \sigma^2(\Sigma_i x_i x_i')^{-1}$$

- Standard error of OLS estimate b_k: The square root of the k^{th} diagonal element of $V\{b\}$

  $$\sigma\sqrt{c_{kk}}$$

  with c_{kk} the k-th diagonal element of $(X'X)^{-1}$

- Estimated variance (covariance matrix) of b:

  $$\tilde{V}\{b\} = s^2(X'X)^{-1} = s^2(\Sigma_i x_i x_i')^{-1}$$

- Estimated standard error of b_k:

  $$se(b_k) = s\sqrt{c_{kk}}$$

Two Examples

1. Simple regression $y_i = \alpha + \beta x_i + \varepsilon_t$

   The variance for the OLS estimator of β is

$$V\{b\} = \frac{\sigma^2}{Ns_x^2}$$

   b is the more accurate, the larger N and s_x^2 and the smaller σ^2

2. Regression with two regressors:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_t$$

   The variance for the OLS estimator of β_2 is

$$V\{b_2\} = \frac{1}{1-r_{23}^2} \frac{\sigma^2}{Ns_{x2}^2}$$

   r_{23}^2: correlation coefficient between X_2 and X_3

   b_2 is most accurate if X_2 and X_3 are uncorrelated

Normality of Error Terms

For the purpose of statistical inference, a distributional assumption for the ε_i's is needed

A5	ε_i normally distributed for all *i*

Together with assumptions (A1), (A3), and (A4), (A5) implies

$\varepsilon_i \sim NID(0,\sigma^2)$ for all *i*

i.e., all ε_i are

- independent drawings
- from the *normal* distribution
- with mean 0
- and variance σ^2

Error terms are "normally and independently distributed" (NID)

Properties of OLS Estimators

1. The OLS estimator b is unbiased: $E\{b\} = \beta$
2. The variance of the OLS estimator is given by

$$V\{b\} = \sigma^2(X'X)^{-1}$$

3. The OLS estimator b is a BLUE (best linear unbiased estimator) for β

4. The OLS estimator b is normally distributed with mean β and covariance matrix $V\{b\} = \sigma^2(X'X)^{-1}$

$$b \sim N(\beta, \sigma^2(X'X)^{-1}), \; b_k \sim N(\beta_k, \sigma^2 c_{kk})$$

with c_{kk}: the k-th diagonal element of $(X'X)^{-1}$

Needs assumptions (A1) - (A5)

Individual Wages: Relevance of Assumptions

$wage_i = \beta_1 + \beta_2 * male_i + \varepsilon_i$

What do the assumptions mean?

(A1): $\beta_1 + \beta_2 * male_i$ contains the entire systematic part of the model; no other regressors besides gender are relevant?

(A2): x_i uncorrelated with ε_i for all i: knowledge of a person's gender provides no information about further variables which affect the person's wage; is this realistic?

(A3) $V\{\varepsilon_i\} = \sigma^2$ for all i: variance of error terms (and of wages) is the same for males and females; is this realistic?

(A4) $Cov\{\varepsilon_{i,}, \varepsilon_j\} = 0$, $i \neq j$: implied by random sampling

(A5) Normality of ε_i: is this realistic? (Would allow, e.g., for negative wages)

Individual Wages, cont'd

OLS estimated wage equation (Table 2.1, Verbeek)

Dependent variable: *wage*		
Variable	Estimate	Standard error
constant	5.1469	0.0812
male	1.1661	0.1122
$s = 3.2174$	$R^2 = 0.0317$	$F = 107.93$

b_1 = 5,1479, se(b_1) = 0,0812: mean wage p.h. for females: 5,15$, with std.error of 0,08$

b_2 = 1,166, se(b_2) = 0,112

95% confidence interval for β_1: $4{,}988 \le \beta_1 \le 5{,}306$

Your Homework

1. Verbeek's data set "wages1" contains for a sample of 3294 individuals the wage p.h. (*wage*) and other variables. Calculate, using GRETL, for the variable *school* (years of schooling) the mean (a) of the whole sample, (b) of males and females, and (c) the standard deviation of the years of schooling for males and for females.
2. For Verbeek's data set "wages1", using GRETL, (a) cross-tabulate the variable *school* (years of schooling) over *male* for individuals with *school* at least 8 years; compare (b) the mean values of the males and females; draw for the whole population (c) scatter plots of *wage* over *school* and *exper*; and (d) a factorized box plot of *wage* over *school*. Discuss the results.

Your Homework, cont'd

3. For the simple regression $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1,\ldots,N$, show that the variance of the OLS estimate for β is $\sigma^2/(Ns_x^2)$, where σ^2 is the error term variance, s_x^2 the variance of the x_i's.
4. For the sample (y_i, x_i), $i = 1,\ldots,N$, and the linear regression ($y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$): (a) write out the matrices $X'X$ and $X'y$; (b) write out the determinant $det[(X'X)^{-1}]$, the matrix $(X'X)^{-1}$, and the OLS estimator $b = (X'X)^{-1}X'y$.