

---

Econometrics - Lecture 1

# Introduction to Linear Regression

---

# Contents

- Organizational Issues
- Some History of Econometrics
- An Introduction to Linear Regression
  - OLS as an algebraic tool
  - The Linear Regression Model
  - Small Sample Properties of OLS estimator
- Introduction to GRETl

# Organizational Issues

## **Aims of the course**

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

# Organizational Issues, cont'd

## Literature

### Course textbook

- Marno Verbeek, *A Guide to Modern Econometrics*, 3<sup>rd</sup> Ed., Wiley, 2008

### Suggestions for further reading

- P. Kennedy, *A Guide to Econometrics*, 6<sup>th</sup> Ed., Blackwell, 2008
- W.H. Greene, *Econometric Analysis*. 6th Ed., Pearson International, 2008

# 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, histogram, frequency tables, 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 discussion, 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

---

- Organizational Issues
- **Some History of Econometrics**
- An Introduction to Linear Regression
  - OLS as an algebraic tool
  - The Linear Regression Model
  - Small Sample Properties of OLS estimator
- Introduction to GRETL

# 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
- Use of statistical methods for
  - measuring theoretical coefficients, e.g., demand elasticities,
  - representing business cycles
- Data: limited availability; time-series on agricultural commodities, foreign trade



# 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, 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: O.Morgenstern (head), A.Wald, G.Tintner
- Econometric Society, founded 1930 by R.Frisch et al.
  - Facilitates exchange of scholars from Europe and US
  - Covers econometrics and mathematical statistics

# First Steps

- R.Frisch, J.Tinbergen:
  - Macro-economic modeling 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 (1944)
  - Econometrics as a tool for testing economic theories
  - States assumptions needed for building and testing econometric models
  - 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
  - Methodology for macro-economic modeling 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 equations
- Haavelmo's "The Probability Approach in Econometrics"
  - Why is the probability approach indispensable?
  - Modeling procedure based on ML and hypothesis testing

# Haavelmo's Arguments for the Probabilistic Approach

- Economic variables in economic theory and econometric models
  - “Observational” vs. “theoretical” vs. “true” variables
  - Models have to take into account inaccurately measured data and passive observations
- Unrealistic assumption of permanence of economic laws
  - Ceteris paribus assumption
  - Economic time-series data
  - Simplifying economic theories
  - Selection of economic variables and relations out of the whole system of fundamental laws

# Cowles Commission Methodology

Assumptions based to macro-econometric modeling 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  $\alpha, \dots, \delta$  are independent of time  $t$
3. Parameters  $\alpha, \dots, \delta$  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 till ~1970
  - Multi-equation models for analyses and forecasting
  - Growing computing power
  - Development of econometric tools

- Scepticism

- 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
  - software

# The 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 (MV, Ch. 3): interpretation of the fitted model, selection of regressors, testing the functional form
4. Heteroskedascity and autocorrelation (Verbeek, Ch. 4): causes, 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 modeling

---

# The Next 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 as an algebraic tool
  - The Linear Regression Model
  - Small Sample Properties of OLS estimator
- Introduction to GRETl

# 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

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

$\beta$ : coefficient of  $X$

$\alpha$ : intercept

multiple linear regression model

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

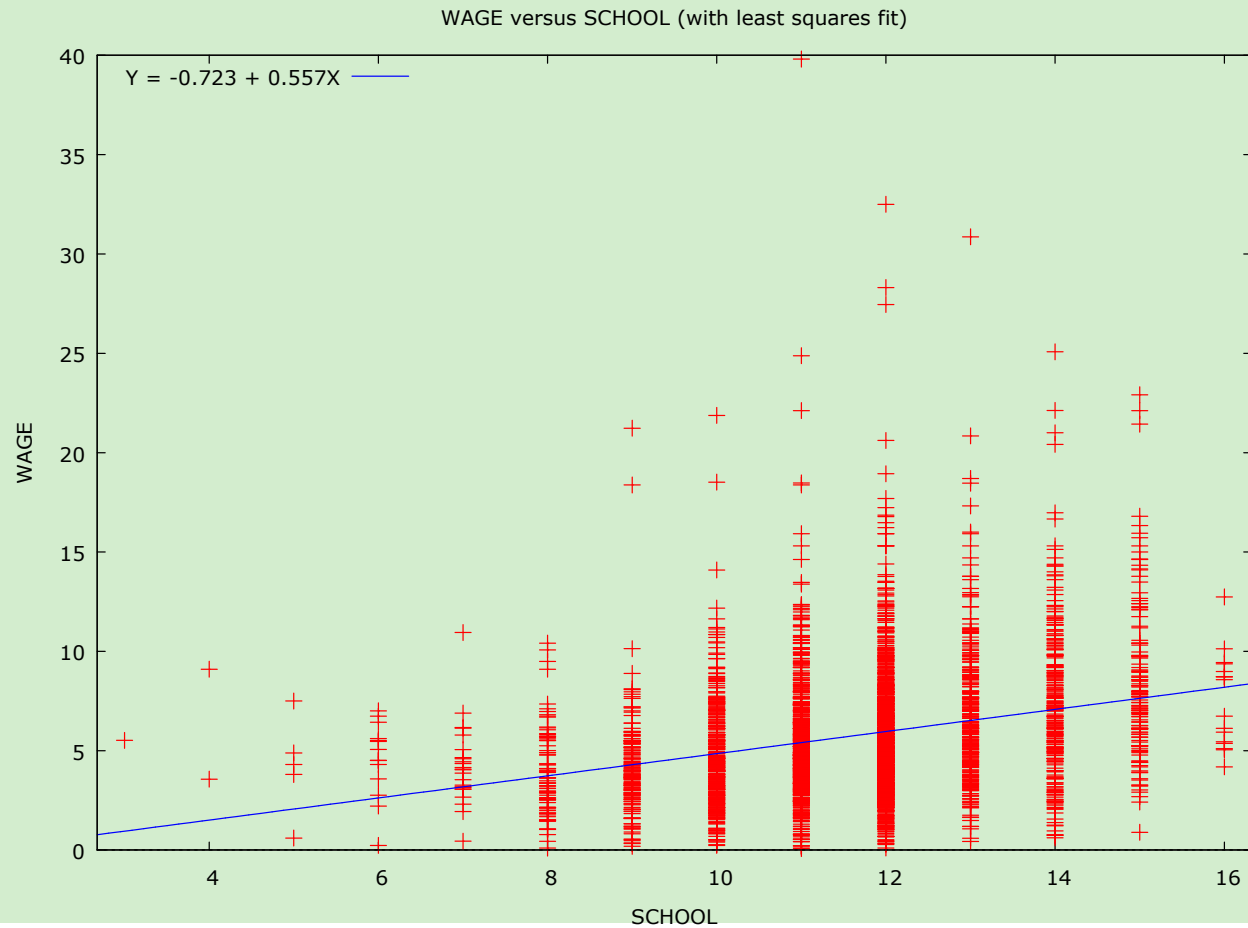
# Example: Individual Wages

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)
  - EXPER: experience in years
  - SCHOOL: years of schooling
  - 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



# Fitting a Model to Data

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

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

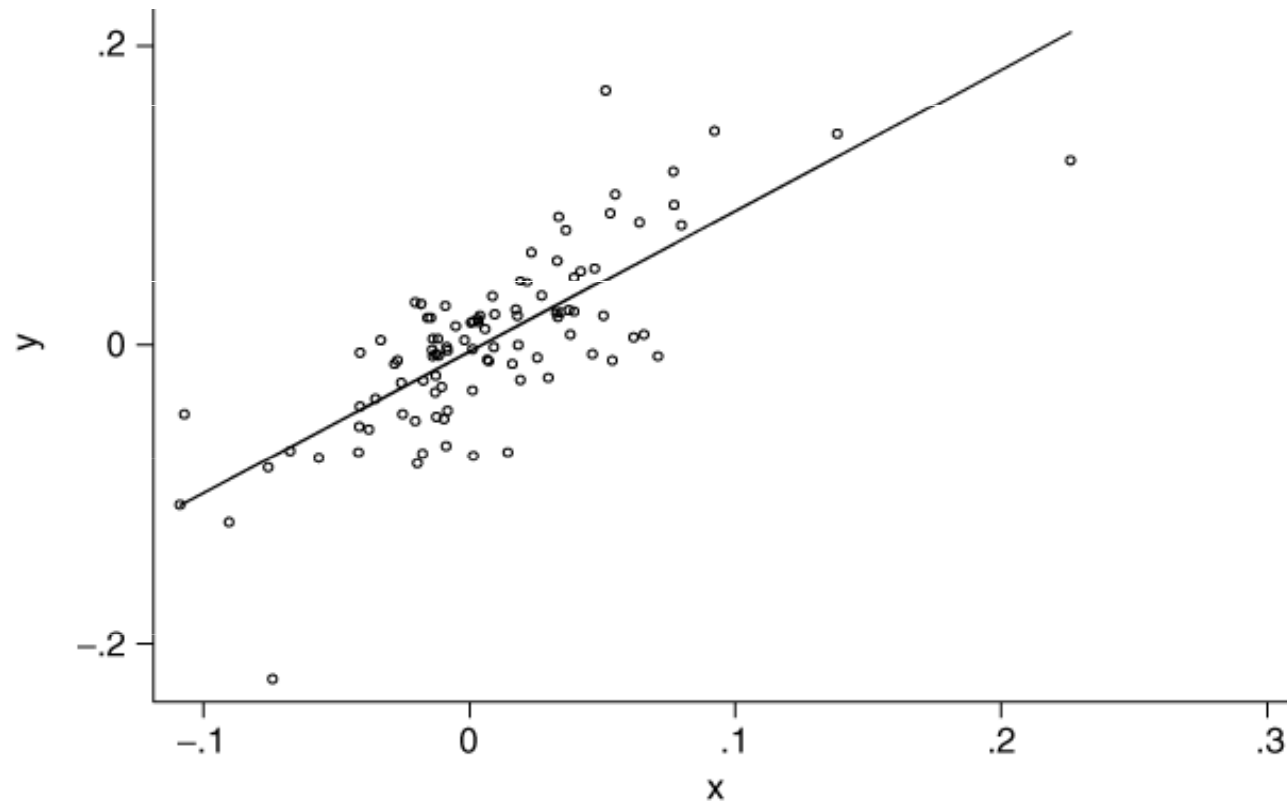
$$S(\beta_1, \beta_2) = \sum_i [y_i - (\beta_1 + \beta_2 x_i)]^2 = \sum_i \varepsilon_i^2$$

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

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

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

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: <i>wage</i>		
Variable	Estimate	Standard error
constant	5.1469	0.0812
<i>male</i>	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

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

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

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

OLS estimates  $b = (b_1, b_2, \dots, b_K)'$  are obtained by minimizing the objective function wrt the  $\beta_k$ '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 so-called normal equations, a system of  $K$  linear equations for the components of  $b$

Given that the symmetric  $K \times K$ -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:  $(y_i, x_i') = (y_i, (1, x_{i2}, \dots, x_{iK}))$ ,  $i = 1, \dots, N$

For  $y_i$ , the linear combination or the fitted value

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

is the best linear combination for  $Y$  from  $X_2, \dots, X_K$  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  $\beta$  gives the normal equations

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

resulting from differentiating  $S(\beta)$  with respect to  $\beta$  and setting the first derivative to zero

The OLS solution or OLS estimators for  $\beta$  are

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

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  $N \times N$ -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

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

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

- Minimum value of objective function

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

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

$$\sum_i e_i = 0$$

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

# Contents

- Organizational Issues
- Some History of Econometrics
- An Introduction to Linear Regression
  - OLS as an algebraic tool
  - The Linear Regression Model
  - Small Sample Properties of OLS estimator
- Introduction to GRETL

# Economic Models

Describe economic relationships (not only 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  $y_i, x_{i2}, \dots, x_{iK}$ : observable, sample ( $i = 1, \dots, N$ ) from a well-defined population or universe
- Error term  $\varepsilon_i$  (disturbance term) contains all influences that are not included explicitly in the model; unobservable; assumption  $E\{\varepsilon_i | x_i\} = 0$  gives

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

the model describes the expected value of  $y$  given  $x$

- Unknown coefficients  $\beta_1, \dots, \beta_K$ : population parameters

# The Sampling Concept

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

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

A) Sampling process with fixed, non-stochastic  $x_i$ 's

- New sample: new error terms  $\varepsilon_i$  and new  $y_i$ 's
- Random sampling of  $\varepsilon_i$ ,  $i = 1, \dots, N$ : joint distribution of  $\varepsilon_i$ 's determines properties of  $b$  etc.

B) Sampling process with samples of  $(x_i, y_i)$  or  $(x_i, \varepsilon_i)$

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

# The Sampling Concept, 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

# The Ceteris Paribus Condition

- The linear regression model needs assumptions to allow interpretation
- Assumption for  $\varepsilon_i$ 's:  $E\{\varepsilon_i | x_i\} = 0$ ; exogeneity of variables  $X$
- This implies

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

i.e., the regression line describes the conditional expectation of  $y_i$  given  $x_i$

- Coefficient  $\beta_k$  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)
- Exogeneity can be restrictive



# Regression Coefficients

Linear regression model:

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

Coefficient  $\beta_k$  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:  $wage_i = \beta_1 + \beta_2 male_i + \beta_3 school_i + \beta_4 exper_i + \varepsilon_i$   
 $\beta_3$  measures the impact of one additional year at school upon a person's wage, keeping gender and years of experience fixed

# Estimation of $\beta$

Given a sample  $(x_i, y_i)$ ,  $i = 1, \dots, N$ , the OLS estimators for  $\beta$

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

can be used as an approximation for  $\beta$

- 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  $\varepsilon_i$ 's determine the quality, i.e., the statistical properties, of  $b$

# Contents

- Organizational Issues
- Some History of Econometrics
- An Introduction to Linear Regression
  - OLS as an algebraic tool
  - The Linear Regression Model
  - **Small Sample Properties of OLS estimator**
- Introduction to GRETL

# 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

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

# 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 of all possible samples
- Observations are generated by some random process

Distribution of the OLS estimator

- Actual distribution not known
- Theoretical distribution determined by assumptions on
  - model specification
  - the error term  $\varepsilon_i$  and regressor variables  $x_i$

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

# Gauss-Markov Assumptions

Observation  $y_i$  is a linear function

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

of observations  $x_{ik}$ ,  $k = 1, \dots, K$ , of the regressor variables and the error term  $\varepsilon_i$

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

A1	$E\{\varepsilon_i\} = 0$ for all $i$
A2	all $\varepsilon_i$ are independent of all $x_i$ (exogeneous $x_i$ )
A3	$V\{\varepsilon_i\} = \sigma^2$ for all $i$ (homoskedasticity)
A4	$\text{Cov}\{\varepsilon_i, \varepsilon_j\} = 0$ 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  $E\{y_i | x_i\}$  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  $E\{\varepsilon | X\} = E\{\varepsilon\} = 0$  and  $V\{\varepsilon | X\} = V\{\varepsilon\} = \sigma^2 I_N$

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

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

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 nice properties; see below
- Gauss-Markov assumptions are very strong and often not satisfied
- Relaxations of the Gauss-Markov assumptions and consequences of such relaxations are important topics



# Properties of OLS Estimators

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

Needs assumptions (A1) and (A2)

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

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

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

3. Gauss-Markov theorem: The OLS estimator  $b$  is a BLUE (best linear unbiased estimator) for  $\beta$

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

# The Gauss-Markov Theorem

OLS estimator  $b$  is BLUE (best linear unbiased estimator) for  $\beta$

- Linear estimator:  $b^* = Ay$  with any full-rank  $K \times N$  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 of the OLS estimators:

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

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

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

Under assumptions (A1)-(A4),  $s^2$  is unbiased for  $\sigma^2$

Attention: the estimator  $(N - 1)^{-1} \sum_i e_i^2$  is biased

- Estimated variance (covariance matrix) of  $b$ :

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

# Standard Errors of OLS Estimators, cont'd

- Variance of the OLS estimators:

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

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

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

- 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  $\beta$  is

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

$b$  is the more accurate, the larger  $N$  and  $s_x^2$  and the smaller  $\sigma^2$

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  $\beta_2$  is

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

$b_2$  is most accurate if  $X_2$  and  $X_3$  are uncorrelated

# Normality of Error Terms

For statistical inference purposes, a distributional assumption for the  $\varepsilon_i$ 's is needed

A5	$\varepsilon_i$ normally distributed for all $i$
----	--

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

$\varepsilon_i \sim \text{NID}(0, \sigma^2)$  for all  $i$

i.e., all  $\varepsilon_i$  are

- independent drawings
- from the *normal* distribution
- with mean 0
- and variance  $\sigma^2$

Error terms are “normally and independently distributed”

# 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  $\beta$

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

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

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

Needs assumptions (A1) - (A5)

# Example: Individual Wages

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

What do the assumptions mean?

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

(A2):  $x_i$  uncorrelated with  $\varepsilon_i$  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  $\varepsilon_i$ : is this realistic? (Would allow, e.g., for negative wages)



# Individual Wages, cont'd

OLS estimated wage equation (Table 2.1, Verbeek)

Dependent variable: <i>wage</i>		
Variable	Estimate	Standard error
constant	5.1469	0.0812
<i>male</i>	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  $\beta_1$ :  $4,988 \leq \beta_1 \leq 5,306$

# Your Homework

1. Verbeek's data set "WAGES" contains for a sample of 3294 individuals the wage and other variables. Using GRETL, draw box plots (a) for all wages p.h. in the sample, (b) for wages p.h. of males and of females.
2. For Verbeek's data set "WAGES", calculate, using GRETL, the mean wage p.h. (a) of the whole sample, (b) of males and females, and (c) of persons with schooling between (i) 0 and 6 years, (ii) 7 and 12 years, and (iii) 13 and 16 years.
3. For the simple linear regression ( $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ ): write the OLS estimator  $b = (X'X)^{-1}X'y$  in summation form; how is the matrix  $X$  (of order  $N \times 2$ ) related to the  $X_i$ ,  $i = 1, \dots, N$ ?
4. Show that the  $N$ -vector  $e$  of residuals fulfills the relation  $X'e = 0$ .
5. Show that  $P_x M_x = 0$  for the hat matrix  $P_x$  and the residual generating matrix  $M_x$ .