Econometrics 2 - Lecture 1

ML Estimation, Diagnostic Tests

Contents

Organizational Issues

Course schedule

Class	Date
1	Fr, Mar 9
2	Fr, Mar 16
3	Fr, Mar 23
4	Fr, Apr 6
5	Fr, Apr 20
6	Fr, Apr 27

Classes start at 10:00

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

Organizational Issues, cont'd

Literature

Course textbook

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

Suggestions for further reading

- W.H. Greene, *Econometric Analysis*. 7th Ed., Pearson International, 2012
- R.C. Hill, W.E. Griffiths, G.C. Lim, *Principles of Econometrics*, 4th Ed., Wiley, 2012

Aims and Content

Aims of the course

- Deepening the understanding of econometric concepts and principles
- Learning about advanced econometric tools and techniques
  - ML estimation and testing methods (MV, Cpt. 6)
  - Time series models (MV, Cpt. 8, 9)
  - Multi-equation models (MV, Cpt. 9)
  - Models for limited dependent variables (MV, Cpt. 7)
  - Panel data models (MV, Cpt. 10)
- Use of econometric tools for analyzing economic data: specification of adequate models, identification of appropriate econometric methods, interpretation of results
- Use of GRETL

Contents

- Organizational Issues
- Overview of Contents
- Linear Regression: A Review
- Estimation of Regression Parameters
- Estimation Concepts
- ML Estimator: Idea and Illustrations
- ML Estimator: Notation and Properties
- ML Estimator: Two Examples
- Asymptotic Tests
- Some Diagnostic Tests

Limited Dependent Variables: An Example

Explain whether a household owns a car: explanatory power have

- income
- household size
- etc.

Regression is not suitable!

Why?

Limited Dependent Variables: An Example

Explain whether a household owns a car: explanatory power have

- income
- household size
- etc.

Regression is not suitable!

- Owning a car has two manifestations: yes/no
- Indicator for owning a car is a binary variable

Models are needed that allow to describe a binary dependent variable or a, more generally, limited dependent variable

Cases of Limited Dependent Variable

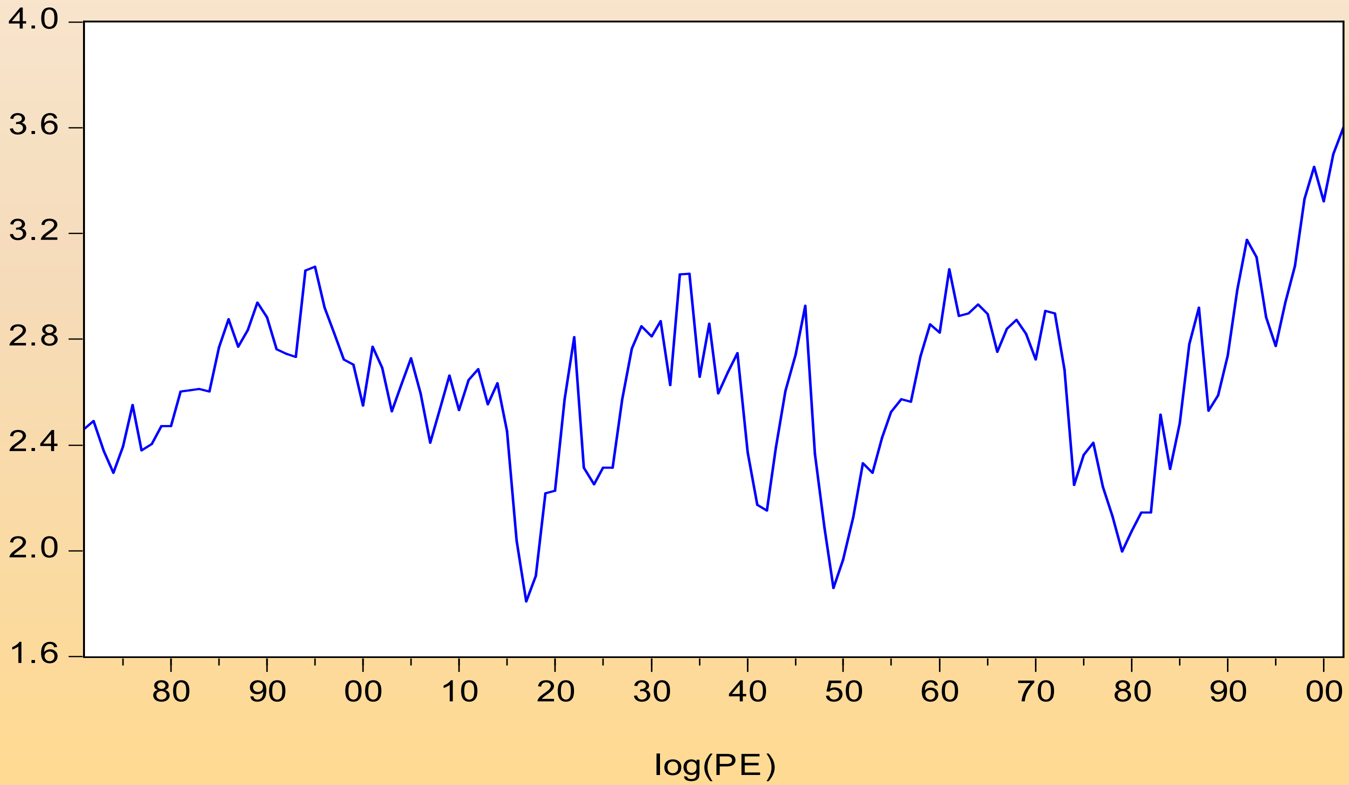
Typical situations: functions of explanatory variables are used to describe or explain

- Dichotomous dependent variable, e.g., ownership of a car (yes/no), employment status (employed/unemployed)
- Ordered response, e.g., qualitative assessment (good/average/bad), working status (full-time/part-time/not working)
- Multinomial response, e.g., trading destinations (Europe/Asia/Africa), transportation means (train/bus/car)
- Count data, e.g., number of orders a company receives in a week, number of patents granted to a company in a year
- Censored data, e.g., expenditures for durable goods, duration of study with drop outs

Time Series Example: Price/Earnings Ratio

Verbeek's data set PE: PE = ratio of S&P composite stock price index and S&P composite earnings of the S&P500, annual, 1871-2002

- Is the PE ratio mean reverting?
- log(PE)
  - Mean 2.63 (PE: 13.9)
  - Min 1.81 (6.1)
  - Max 3.60 (36.6)
  - Std 0.33



Time Series Models

Purpose of modelling

- Description of the data generating process
- Forecasting

Types of model specification

- Deterministic trend: a function $f(t)$ of the time t, describing the evolution of E{Y_t} over time

  $$Y_t = f(t) + \varepsilon_t,\ \varepsilon_t\text{: white noise}$$

  e.g., $Y_t = \alpha + \beta t + \varepsilon_t$
- Autoregression AR(1)

  $$Y_t = \delta + \theta Y_{t-1} + \varepsilon_t,\quad |\theta| < 1,\ \varepsilon_t\text{: white noise}$$

  generalization: ARMA(p,q)-process

  $$Y_t = \theta_1 Y_{t-1} + \dots + \theta_p Y_{t-p} + \varepsilon_t + \alpha_1\varepsilon_{t-1} + \dots + \alpha_q\varepsilon_{t-q}$$

PE Ratio: Various Models

Diagnostics for various competing models: $\Delta y_t = \log PE_t - \log PE_{t-1}$

Best fit for

- BIC: MA(2) model $\Delta y_t = 0.008 + e_t - 0.250\, e_{t-2}$
- AIC: AR(2,4) model $\Delta y_t = 0.008 - 0.202\, \Delta y_{t-2} - 0.211\, \Delta y_{t-4} + e_t$
- Q_{12}: Box-Ljung statistic for the first 12 autocorrelations

Model	Lags	AIC	BIC	Q_{12}	*p*-value
MA(4)	1–4	-73.389	-56.138	5.03	0.957
AR(4)	1–4	-74.709	-57.458	3.74	0.988
MA	2, 4	-76.940	-65.440	5.48	0.940
AR	2, 4	**-78.057**	-66.556	4.05	0.982
MA	2	-76.072	**-67.447**	9.30	0.677
AR	2	-73.994	-65.368	12.12	0.436

Multi-equation Models

Economic processes: Simultaneous and interrelated development of a set of variables

Examples:

- Households consume a set of commodities (e.g., food, durables); the demanded quantities depend on the prices of commodities, the household income, the number of persons living in the household, etc.; a consumption model contains a set of dependent variables and a set of explanatory variables.
- The market of a product is characterized by (a) the demanded and supplied quantity and (b) the price of the product; a model for the market consists of equations representing the development and interdependencies of these variables.
- An economy consists of markets for commodities, labour, finances, etc.; a model for a sector or the full economy contains descriptions of the development of the relevant variables and their interactions.

Panel Data

Population of interest: individuals, households, companies, countries

Types of observations

- Cross-sectional data: Observations of all units of a population, or of a (representative) subset, at one specific point in time
- Time series data: Series of observations on units of the population over a period of time
- Panel data (longitudinal data): Repeated observations of (the same) population units collected over a number of periods; data set with both a cross-sectional and a time series aspect; multi-dimensional data

Cross-sectional and time series data are special cases of panel data

Panel Data Example: Individual Wages

Verbeek's data set "males"

- Sample of
  - 545 full-time working males
  - each person observed yearly after completion of school in 1980 till 1987
- Variables
  - *wage*: log of hourly wage (in USD)
  - *school*: years of schooling
  - *exper*: age – 6 – *school*
  - dummies for union membership, married, black, Hispanic, public sector
  - others

Panel Data Models

Panel data models allow

- controlling individual differences, comparing behaviour, analysing dynamic adjustment, measuring effects of policy changes
- more realistic models than cross-sectional and time-series models
- more detailed or sophisticated research questions

E.g.: What is the effect of being married on the hourly wage

Contents

The Linear Model

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

A simple linear regression model

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

β: coefficient of X
α: intercept

A 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), $i = 1,\ldots,N$

Model for observations: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1,\ldots,N$

Fitted values: $\hat{y}_i = b_1 + b_2 x_i$, $i = 1,\ldots,N$

Principle of (Ordinary) Least Squares gives the OLS estimators

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

Objective function: sum of the squared deviations

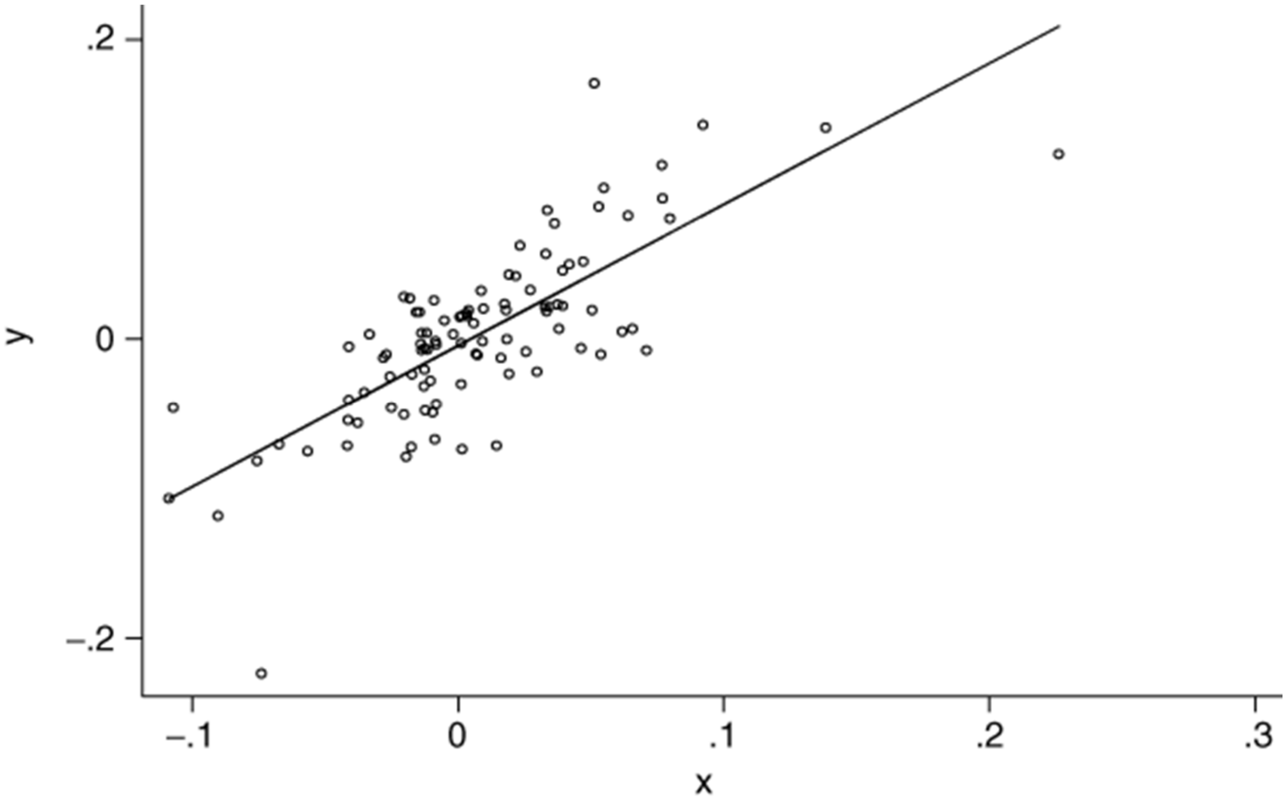
$$S(\beta_1, \beta_2) = \Sigma_i [y_i - (\beta_1 + \beta_2 x_i)]^2 = \Sigma_i \varepsilon_i^2$$

Deviations between observation and fitted values, residuals:

$$e_i = y_i - \hat{y}_i = y_i - (b_1 + b_2 x_i)$$

Observations and Fitted Regression Line

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



Contents

OLS Estimators

Equating the partial derivatives of S(β_1, β_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 $\bar{x}, \bar{y}$ and

and second moments

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

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

OLS Estimators: The 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

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

this results in the OLS estimators

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

In Matrix Notation

N observations

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

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

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

Gauss-Markov Assumptions

Observation y_i (i = 1, …, N) is a linear function

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

of observations x_{ik}, k =1, …, K, of the regressor variables and the error term ε_i

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

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)

Normality of Error Terms

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 $N(0,\sigma^2)$
- with mean 0
- and variance σ^2

Error terms are "normally and independently distributed" (NID, n.i.d.)

Properties of OLS Estimators

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

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

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

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

Properties

- 1., 2., and 3. follow from Gauss-Markov assumptions
- 4. needs in addition the normality assumption (A5)

Distribution of *t*-statistic

t-statistic

$$t_k = \frac{b_k}{se(b_k)}$$

with the standard error $se(b_k)$ of b_k follows

1. the *t*-distribution with *N*-*K* d.f. if the Gauss-Markov assumptions (A1) - (A4) and the normality assumption (A5) hold
2. approximately the *t*-distribution with *N*-*K* d.f. if the Gauss-Markov assumptions (A1) - (A4) hold but not the normality assumption (A5)
3. asymptotically ($N \to \infty$) the standard normal distribution N(0,1)
4. Approximately, for large *N*, the standard normal distribution N(0,1)

The approximation error decreases with increasing sample size *N*

OLS Estimators: Consistency

The OLS estimators b are consistent,

$$\text{plim}_{N \to \infty} b = \beta,$$

if one of the two sets of conditions are fulfilled:

- (A2) from the Gauss-Markov assumptions and the assumption (A6), or
- the assumption (A7), which is weaker than (A2), and the assumption (A6)

Assumptions (A6) and (A7):

A6	$1/N\ \Sigma^N_{i=1}\ x_i\ x_i'$ converges with growing N to a finite, nonsingular matrix Σ_{xx}
A7	The error terms have zero mean and are uncorrelated with each of the regressors: $E\{x_i\ \varepsilon_i\} = 0$

Contents

Estimation Concepts

OLS estimator: Minimization of objective function $S(\beta) = \Sigma_i \varepsilon_i^2$ gives

- K first-order conditions $\Sigma_i (y_i - x_i'b) x_i = \Sigma_i e_i x_i = 0$, the normal equations
- OLS estimators are solutions of the normal equations
- Moment conditions

$$E\{(y_i - x_i'\beta) x_i\} = E\{\varepsilon_i x_i\} = 0$$

- Normal equations are sample moment conditions (times N)

IV estimator: Model allows derivation of the moment conditions

$$E\{(y_i - x_i'\beta) z_i\} = E\{\varepsilon_i z_i\} = 0$$

which are functions of

- observable variables y_i, x_i, instrument variables z_i, and unknown parameters β
- Moment conditions are used for deriving IV estimators
- OLS estimators are special case of IV estimators

Estimation Concepts, cont'd

GMM estimator: generalization of the moment conditions

$$E\{f(w_i, z_i, \beta)\} = 0$$

- with observable variables w_i, instrument variables z_i, and unknown parameters β; f: multidimensional function with as many components as moment conditions
- Allows for non-linear models
- Under weak regularity conditions, the GMM estimators are
  - consistent
  - asymptotically normal

Maximum likelihood estimation

- Basis is the distribution of y_i conditional on regressors x_i
- Depends on unknown parameters β
- The estimates of the parameters β are chosen so that the distribution corresponds as good as possible to the observations y_i and x_i

Contents

Example: Urn Experiment

The experiment:

- The urn contains red and white balls
- Proportion of red balls: p (unknown)
- N random draws
- Random draw i: $y_i = 1$ if ball in draw i is red, $y_i = 0$ otherwise; $P\{y_i=1\} = p$
- Sample: N_1 red balls, $N-N_1$ white balls
- Probability for this result:

$$P\{N_1 \text{ red balls}, N-N_1 \text{ white balls}\} \approx p^{N1} (1 - p)^{N-N1}$$

Likelihood function $L(p)$: The probability of the sample result, interpreted as a function of the unknown parameter p

$$L(p) = p^{N1} (1 - p)^{N-N1} , 0 < p < 1$$

Urn Experiment: Likelihood Function and LM Estimator

Likelihood function: (proportional to) the probability of the sample result, interpreted as a function of the unknown parameter *p*

$$L(p) = p^{N1} (1 - p)^{N-N1} \text{ , } 0 < p < 1$$

Maximum likelihood estimator: that value $\hat{p}$ of *p* which maximizes *L*(*p*)

$$\hat{p} = \arg\max_p L(p)$$

Calculation of $\hat{p}$: maximization algorithm

- As the log-function is monotonous, coordinates *p* of the extremes of *L*(*p*) and log *L*(*p*) coincide
- Use of log-likelihood function is often more convenient

$$\log L(p) = N_1 \log p + (N - N_1) \log (1 - p)$$

Urn Experiment: Likelihood Function, cont'd

Verbeek, Fig.6.1

p	log *L*(*p*)
0.1	-107.21
0.2	-83.31
0.3	-72.95
0.4	-68.92
0.5	-69.31
0.6	-73.79
0.7	-83.12
0.8	-99.95
0.9	-133.58

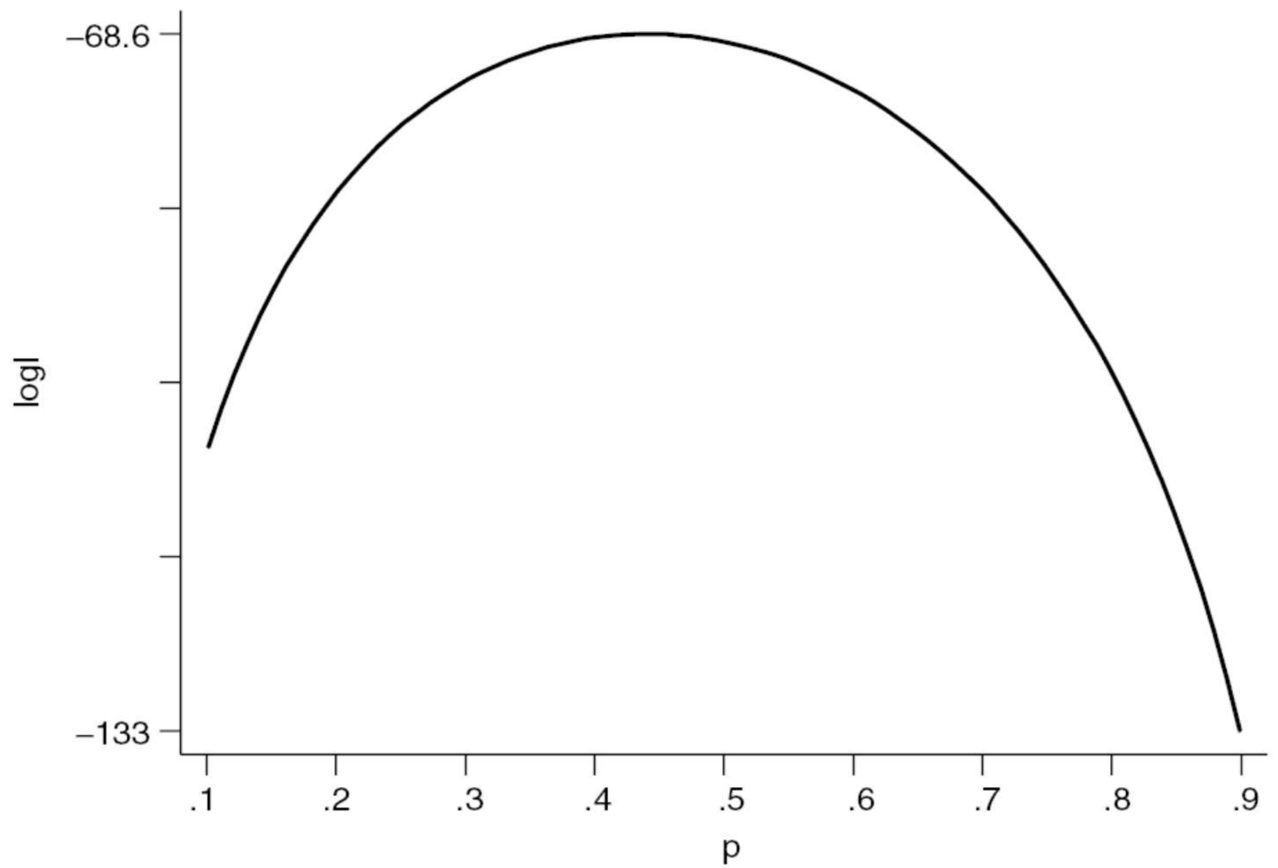


Figure 6.1 Sample loglikelihood function for $N = 100$ and $N_1 = 44$

Urn Experiment: ML Estimator

Maximizing log *L(p)* with respect to *p* gives the first-order condition

$$\frac{d \log L(p)}{dp} = \frac{N_1}{p} - \frac{N - N_1}{1 - p} = 0$$

Solving this equation for *p* gives the maximum likelihood estimator (ML estimator)

$$\hat{p} = \frac{N_1}{N}$$

For $N = 100$, $N_1 = 44$, the ML estimator for the proportion of red balls is $\hat{p} = 0.44$

Maximum Likelihood Estimator: The Idea

- Specify the distribution of the data (of *y* or *y* given *x*)
- Determine the likelihood of observing the available sample as a function of the unknown parameters
- Choose as ML estimates those values for the unknown parameters that give the highest likelihood
- Properties: In general, the ML estimators are
  - consistent
  - asymptotically normal
  - efficient

  provided the likelihood function is correctly specified, i.e., distributional assumptions are correct

Example: Normal Linear Regression

Model

$y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$

with assumptions (A1) – (A5)

From the normal distribution of ε_i follows: contribution of observation *i* to the likelihood function:

$$f(y_i | X_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i - \beta_1 - \beta_2 X_i)^2}{\sigma^2} \right\}$$

$L(\beta,\sigma^2) = \prod_i f(y_i | x_i;\beta,\sigma^2)$ due to independent observations; the log-likelihood function is given by

$$\log L(\beta, \sigma^2) = \log \prod_i f(y_i | X_i; \beta, \sigma^2)$$

$$= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_1 - \beta_2 X_i)^2$$

Normal Linear Regression, cont'd

Maximizing log $L(\beta,\sigma^2)$ with respect to β and σ^2 gives the ML estimators

$$\hat{\beta}_2 = Cov\{y,x)\}/V\{x\}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

which coincide with the OLS estimators, and

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i e_i^2$$

which is biased and underestimates σ^2!

Remarks:

- The results are obtained assuming normally and independently distributed (NID) error terms
- ML estimators are consistent but not necessarily unbiased; see the properties of ML estimators below

Contents

ML Estimator: Notation

Let the density (or probability mass function) of y_i, given x_i, be given by $f(y_i|x_i,\theta)$ with K-dimensional vector θ of unknown parameters

Given independent observations, the likelihood function for the sample of size N is

$$L(\theta \mid y, X) = \prod_i L_i(\theta \mid y_i, x_i) = \prod_i f(y_i \mid x_i; \theta)$$

The ML estimators are the solutions of

$$\max_\theta \log L(\theta) = \max_\theta \Sigma_i \log L_i(\theta)$$

or the solutions of the K first-order conditions

$$s(\hat{\theta}) = \frac{\partial \log L(\theta)}{\partial \theta}\Big|_{\hat{\theta}} = \sum_i \frac{\partial \log L_i(\theta)}{\partial \theta}\Big|_{\hat{\theta}} = \sum_i s(\theta)\Big|_{\hat{\theta}} = 0$$

$s(\theta) = \Sigma_i s_i(\theta)$, the K-vector of gradients, also denoted *score vector*

Solution of $s(\theta) = 0$

- analytically (see examples above) or
- by use of numerical optimization algorithms

Matrix Derivatives

The scalar-valued function

$$\log L(\theta \mid y, X) = \prod_i \log L_i(\theta \mid y_i, x_i) = \log L(\theta_1, \ldots, \theta_K \mid y, X)$$

or – shortly written as log $L(\theta)$ – has the K arguments $\theta_1, \ldots, \theta_K$

- K-vector of partial derivatives or gradient vector or score vector or gradient

$$\frac{\partial \log L(\theta)}{\partial \theta} = \left(\frac{\partial \log L(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \log L(\theta)}{\partial \theta_K} \right)' = s(\theta)$$

- KxK matrix of second derivatives or Hessian matrix

$$\frac{\partial^2 \log L(\theta)}{\partial \theta \, \partial \theta'} = \begin{pmatrix} \frac{\partial^2 \log L(\theta)}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial \theta_1 \partial \theta_K} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\theta)}{\partial \theta_K \partial \theta_1} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial \theta_K \partial \theta_K} \end{pmatrix}$$

ML Estimator: Properties

The ML estimator is

1. Consistent
2. asymptotically efficient
3. asymptotically normally distributed:

$$\sqrt{N}(\hat{\theta}-\theta) \rightarrow \mathrm{N}(0,V)$$

V: asymptotic covariance matrix of $\sqrt{N}\hat{\theta}$

The Information Matrix

Information matrix $I(\theta)$

- $I(\theta)$ is the limit (for $N \rightarrow \infty$) of

$$\bar{I}_N(\theta) = -\frac{1}{N} E\left\{\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right\} = -\frac{1}{N}\sum_i E\left\{\frac{\partial^2 \log L_i(\theta)}{\partial\theta\partial\theta'}\right\} = \frac{1}{N}\sum_i I_i(\theta)$$

- For the asymptotic covariance matrix V can be shown: $V = I(\theta)^{-1}$
- $I(\theta)^{-1}$ is the lower bound of the asymptotic covariance matrix for any consistent, asymptotically normal estimator for θ: Cramèr-Rao lower bound

Calculation of $I_i(\theta)$ can also be based on the outer product of the score vector

$$J_i(\theta) = E\{s_i(\theta)s_i(\theta)'\} = -E\left\{\frac{\partial^2 \log L_i(\theta)}{\partial\theta\partial\theta'}\right\} = I_i(\theta)$$

for a miss-specified likelihood function, $J_i(\theta)$ can deviate from $I_i(\theta)$

Example: Normal Linear Regression

Model

$y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$

with assumptions (A1) – (A5) fulfilled

The score vector with respect to β = (β₁,β₂)' is – using $x_i = (1, X_i)'$ –

$$s_i(\beta) = \frac{\partial}{\partial \beta} \log L_i(\beta, \sigma^2) = \frac{1}{\sigma^2} \varepsilon_i x_i$$

The information matrix is obtained both via Hessian and outer product

$$I_{i,11}(\beta, \sigma^2) = -E\left\{ \frac{\partial^2 \log L_i(\theta)}{\partial \beta \partial \beta'} \right\} = E\{s_i s_i'\}$$

$$= \frac{1}{\sigma^4} E\{\varepsilon_i^2 x_i x_i'\} = \frac{1}{\sigma^2} x_i x_i' = \frac{1}{\sigma^2} \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}$$

Covariance Matrix *V*: Calculation

Two ways to calculate *V*:

- Estimator based on the information matrix $I(\theta)$

$$\hat{V}_H = \left(-\frac{1}{N} \sum_i \frac{\partial^2 \log L_i(\theta)}{\partial\theta\,\partial\theta'} \Big|_{\hat{\theta}} \right)^{-1} = \bar{I}_N(\hat{\theta})^{-1}$$

  index "H": the estimate of *V* is based on the Hessian matrix

- Estimator based on the score vector

$$\hat{V}_G = \left(\frac{1}{N} \sum_i s_i(\hat{\theta}) s_i(\hat{\theta})' \right)^{-1} = \left(\frac{1}{N} \sum_i J_i(\hat{\theta}) \right)^{-1}$$

  with score vector $s(\theta)$; index "G": the estimate of *V* is based on gradients
  - also called: OPG (outer product of gradient) estimator
  - also called: BHHH (Berndt, Hall, Hall, Hausman) estimator
  - E{$s_i(\theta)$ $s_i(\theta)$'} coincides with $I_i(\theta)$ if $f(y_i | x_i,\theta)$ is correctly specified

Contents

Again the Urn Experiment

Likelihood contribution of the *i*-th observation

$$\log L_i(p) = y_i \log p + (1 - y_i) \log (1 - p)$$

This gives scores

$$\frac{\partial \log L_i(p)}{\partial p} = s_i(p) = \frac{y_i}{p} - \frac{1-y_i}{1-p}$$

and

$$\frac{\partial^2 \log L_i(p)}{\partial p^2} = -\frac{y_i}{p^2} - \frac{1-y_i}{(1-p)^2}$$

With $E\{y_i\} = p$, the expected value turns out to be

$$I_i(p) = E\left\{-\frac{\partial^2 \log L_i(p)}{\partial p^2}\right\} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

The asymptotic variance of the ML estimator $V = I^{-1} = p(1-p)$

Urn Experiment and Binomial Distribution

The asymptotic distribution is

$$\sqrt{N}(\hat{p} - p) \rightarrow N\left(0, p(1-p)\right)$$

- Small sample distribution:

  $N\hat{p} \sim B(N, p)$

- Use of the approximate normal distribution for portions $\hat{p}$
  - rule of thumb for using the approximate distribution

    $N\,p\,(1-p) > 9$

Test of H_0: $p = p_0$ can be based on test statistic

$$(\hat{p} - p_0) / se(\hat{p})$$

Example: Normal Linear Regression

Model

$y_i = x_i'\beta + \varepsilon_i$

with assumptions (A1) – (A5)

Log-likelihood function

$$\log L(\beta, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y_i - x_i'\beta)^2$$

Scores of the i-th observation

$$s_i(\beta, \sigma^2) = \begin{pmatrix} \dfrac{\partial \log L_i(\beta, \sigma^2)}{\partial \beta} \\ \dfrac{\partial \log L_i(\beta, \sigma^2)}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \dfrac{y_i - x_i'\beta}{\sigma^2} x_i \\ -\dfrac{1}{2\sigma^2} + \dfrac{1}{2\sigma^4}(y_i - x_i'\beta)^2 \end{pmatrix}$$

Normal Linear Regression: ML-Estimators

The first-order conditions – setting both components of $\Sigma_i s_i(\beta,\sigma^2)$ to zero – give as ML estimators: the OLS estimator for β, the average squared residuals for σ²

$$\hat{\beta} = \left(\sum\nolimits_i x_i x_i'\right)^{-1} \sum\nolimits_i x_i y_i, \ \hat{\sigma}^2 = \frac{1}{N} \sum\nolimits_i (y_i - x_i'\hat{\beta})^2$$

Asymptotic covariance matrix: Contribution of the *i*-th observation ($E\{\varepsilon_i\} = E\{\varepsilon_i^3\} = 0$, $E\{\varepsilon_i^2\} = \sigma^2$, $E\{\varepsilon_i^4\} = 3\sigma^4$)

$$I_i(\beta,\sigma^2) = E\{s_i(\beta,\sigma^2) s_i(\beta,\sigma^2)'\} = \text{diag}\left(\frac{1}{\sigma^2} x_i x_i', \frac{1}{2\sigma^4}\right)$$

gives

$$V = I(\beta,\sigma^2)^{-1} = \text{diag}\,(\sigma^2 \Sigma_{xx}^{-1}, 2\sigma^4)$$

with $\Sigma_{xx} = \lim\,(\Sigma_i x_i x_i')/N$

Normal Linear Regression: ML- and OLS-Estimators

The ML estimate for β and σ² follow asymptotically

$$\sqrt{N}(\hat{\beta}-\beta) \rightarrow \mathrm{N}(0, \sigma^2 \Sigma_{xx}{}^{-1})$$

$$\sqrt{N}(\hat{\sigma}^2-\sigma^2) \rightarrow \mathrm{N}(0, 2\sigma^4)$$

For finite samples: Covariance matrix of ML estimators for β

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2 \left(\sum_i x_i x_i'\right)^{-1}$$

similar to OLS results

Contents

Diagnostic Tests

Diagnostic (or specification) tests based on ML estimators

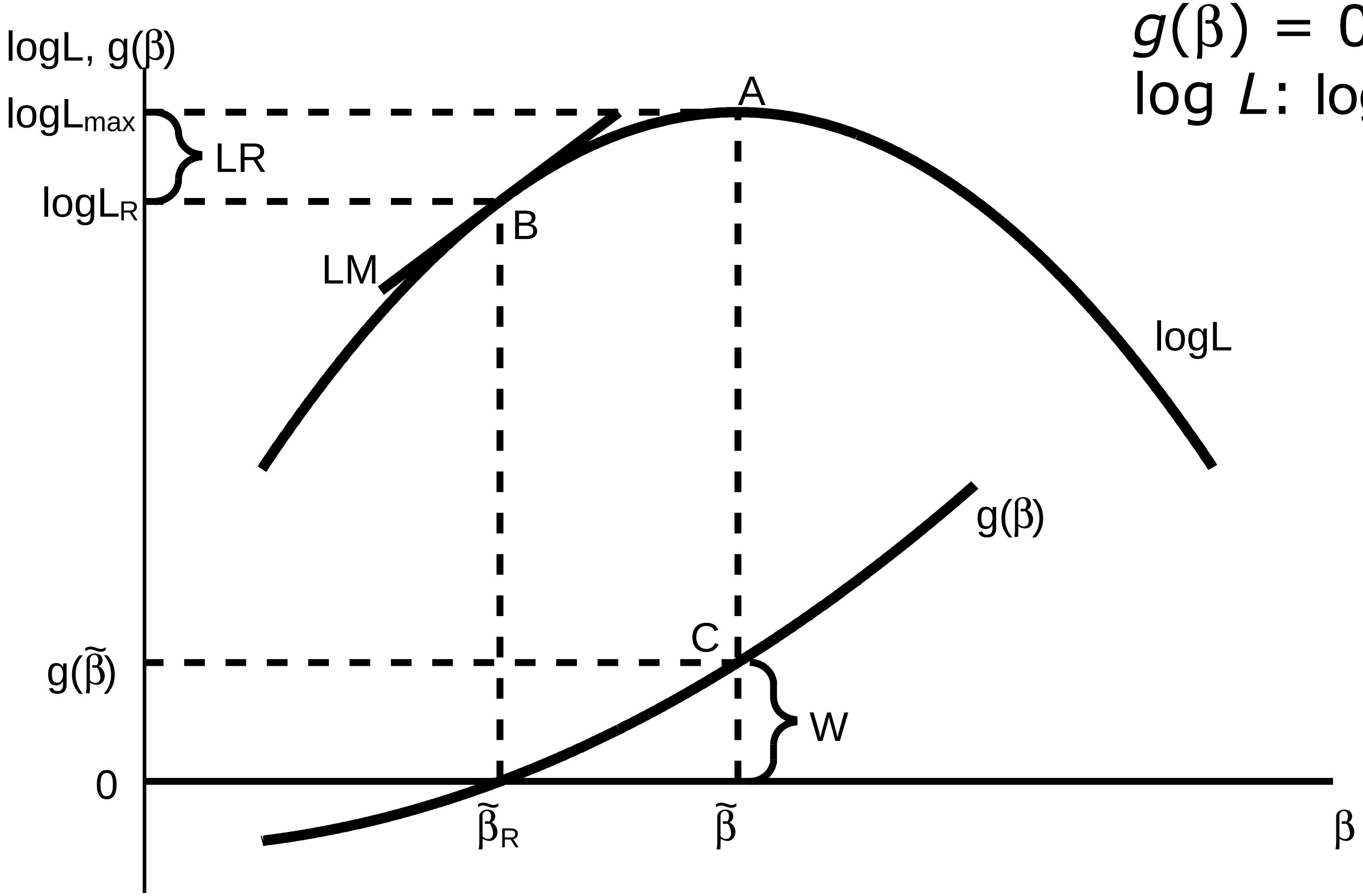
Test situation:

- K-dimensional parameter vector $\theta = (\theta_1, \ldots, \theta_K)'$
- $J \geq 1$ linear restrictions ($K \geq J$)
- H_0: $R\theta = q$ with JxK matrix R, full rank; J-vector q

Test principles based on the likelihood function:

1. Wald test: Checks whether the restrictions are fulfilled for the unrestricted ML estimator for θ; test statistic ξ_W
2. Likelihood ratio test: Checks whether the difference between the log-likelihood values with and without the restriction is close to zero; test statistic ξ_{LR}
3. Lagrange multiplier test (or score test): Checks whether the first-order conditions (of the unrestricted model) are violated by the restricted ML estimators; test statistic ξ_{LM}

Likelihood and Test Statistics



$g(\beta) = 0$: restriction
$\log L$: log-likelihood

The Asymptotic Tests

Under H_0, the test statistics of all three tests

- follow asymptotically, for finite sample size approximately, the Chi-square distribution with *J* d.f.
- The tests are asymptotically (large *N*) equivalent
- Finite sample size: the values of the test statistics obey the relation

$$\xi_W \geq \xi_{LR} \geq \xi_{LM}$$

Choice of the test: criterion is computational effort

1. Wald test: Requires estimation only of the unrestricted model; e.g., testing for omitted regressors: estimate the full model, test whether the coefficients of potentially omitted regressors are different from zero
2. Lagrange multiplier test: Requires estimation only of the restricted model; preferable if restrictions complicate estimation
3. Likelihood ratio test: Requires estimation of both the restricted and the unrestricted model

Wald Test

Checks whether the restrictions are fulfilled for the unrestricted ML estimator for θ

Asymptotic distribution of the unrestricted ML estimator:

$$\sqrt{N}(\hat{\theta}-\theta) \rightarrow N(0,V)$$

Hence, under H_0: $R\,\theta = q$,

$$\sqrt{N}(R\hat{\theta}-R\theta) = \sqrt{N}(R\hat{\theta}-q) \rightarrow N(0,RVR')$$

The test statistic

$$\xi_W = N(R\hat{\theta}-q)'\left[R\hat{V}R'\right]^{-1}(R\hat{\theta}-q)$$

- under H_0, ξ_W is expected to be close to zero
- *p*-value to be read from the Chi-square distribution with *J* d.f.

Wald Test, cont'd

Typical application: tests of linear restrictions for regression coefficients

- Test of H_0: $\beta_i = 0$

  $$\xi_W = b_i^2/[se(b_i)^2]$$

  - ξ_W follows the Chi-square distribution with 1 d.f.
  - ξ_W is the square of the *t*-test statistic
- Test of the null-hypothesis that a subset of *J* of the coefficients β are zeros

  $$\xi_W = (e_R'e_R - e'e)/[e'e/(N-K)]$$

  - e: residuals from unrestricted model
  - e_R: residuals from restricted model
  - ξ_W follows the Chi-square distribution with *J* d.f.
  - ξ_W is related to the *F*-test statistic by $\xi_W = FJ$

Likelihood Ratio Test

Checks whether the difference between the ML estimates obtained with and without the restriction is close to zero

for nested models

- Unrestricted ML estimator: $\hat{\theta}$
- Restricted ML estimator: $\tilde{\theta}$; obtained by minimizing the log-likelihood subject to $R\theta = q$

Under H_0: $R\theta = q$, the test statistic

$$\xi_{LR} = 2\left(\log L(\hat{\theta}) - \log L(\tilde{\theta})\right)$$

- is expected to be close to zero
- *p*-value to be read from the Chi-square distribution with *J* d.f.

Likelihood Ratio Test, cont'd

Test of linear restrictions for regression coefficients

- Test of the null-hypothesis that *J* linear restrictions of the coefficients β are valid

  $$\xi_{LR} = N \log(e_R'e_R/e'e)$$

  - *e*: residuals from unrestricted model
  - e_R: residuals from restricted model
  - ξ_{LR} follows the Chi-square distribution with *J* d.f.
- Requires that the restricted model is nested within the unrestricted model

Lagrange Multiplier Test

Checks whether the derivative of the likelihood for the restricted ML estimator is close to zero

Based on the Lagrange constrained maximization method

Lagrangian, given $\theta = (\theta_1', \theta_2')'$ with restriction $\theta_2 = q$, *J*-vectors θ_2, q, λ

$$H(\theta, \lambda) = \Sigma_i \log L_i(\theta) - \lambda'(\theta_2 - q)$$

First-order conditions give the restricted ML estimators $\tilde{\theta} = (\tilde{\theta}_1', q')'$ and $\tilde{\lambda}$

$$\sum_i \frac{\partial \log L_i(\theta)}{\partial \theta_1} |_{\tilde{\theta}} = \sum_i s_{i1}(\tilde{\theta}) = 0$$

$$\tilde{\lambda} = \sum_i \frac{\partial \log L_i(\theta)}{\partial \theta_2} |_{\tilde{\theta}} = \sum_i s_{i2}(\tilde{\theta})$$

λ measures the extent of violation of the restrictions, basis for ξ_{LM}

s_i are the scores; LM test is also called *score test*

Lagrange Multiplier Test, cont'd

For $\tilde{\lambda}$ can be shown that $N^{-1}\tilde{\lambda}$ follows asymptotically the normal distribution N(0,V_λ) with

$$V_\lambda = I_{22}(\theta) - I_{21}(\theta) I_{11}{}^{-1}(\theta) I_{22}(\theta) = [I^{22}(\theta)]^{-1}$$

i.e., the inverted lower block diagonal (dimension *J* x *J*) of the inverted information matrix

$$I(\theta)^{-1} = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix}^{-1} = \begin{pmatrix} I^{11}(\theta) & I^{12}(\theta) \\ I^{21}(\theta) & I^{22}(\theta) \end{pmatrix}$$

The Lagrange multiplier test statistic

$$\xi_{LM} = N^{-1} \tilde{\lambda}' \hat{I}^{22}(\tilde{\theta}) \tilde{\lambda}$$

has under H_0 an asymptotic Chi-square distribution with *J* d.f.

$\hat{I}^{22}(\tilde{\theta})$ is the lower block diagonal of the estimated inverted information matrix, evaluated at the restricted estimators for θ

The LM Test Statistic

Outer product gradient (OPG) of ξ_{LM}

- Information matrix estimated on basis of scores (cf. slide 48)

$$\hat{I}(\tilde{\theta}) = N^{-1}\sum_i s_i(\tilde{\theta})s_i(\tilde{\theta})' = N^{-1}diag\left\{0, \sum_i s_{i2}(\tilde{\theta})s_{i2}(\tilde{\theta})'\right\}$$

- With

$$\tilde{\lambda} = \sum_i s_{i2}(\tilde{\theta})$$

- the LM test statistics can be written as

$$\xi_{LM} = \sum_i s_{i2}(\tilde{\theta})'\left(\sum_i s_{i2}(\tilde{\theta})s_{i2}(\tilde{\theta})'\right)^{-1}\sum_i s_{i2}(\tilde{\theta})$$

With the *N*x*K* matrix of first derivatives *S* = [$s_1(\tilde{\theta})$, …, $s_N(\tilde{\theta})$]‘

$$\hat{I}(\tilde{\theta}) = N^{-1}\sum_i s_i(\tilde{\theta})s_i(\tilde{\theta})' = N^{-1}S'S$$

- and – with the *N*-vector *i* = (1, …, 1)’

$$\xi_{LM} = \sum_i s_{i2}(\tilde{\theta})'\left(\sum_i s_{i2}(\tilde{\theta})s_{i2}(\tilde{\theta})'\right)^{-1}\sum_i s_{i2}(\tilde{\theta})$$

$$= \sum_i s_i(\tilde{\theta})'\left(\sum_i s_i(\tilde{\theta})s_i(\tilde{\theta})'\right)^{-1}\sum_i s_i(\tilde{\theta}) = i'S(S'S)^{-1}S'i$$

Calculation of the LM Test Statistic

Auxiliary regression of a N-vector $i = (1, \ldots, 1)'$ on the scores $s_i(\tilde{\theta})$, i.e., on the columns of S; no intercept

- Predicted values from auxiliary regression: $S(S'S)^{-1}S'i$
- Sum of squared predictions: $i'S(S'S)^{-1}S'S(S'S)^{-1}S'i = i'S(S'S)^{-1}S'i$
- Total sum of squares: $i'i = N$
- LM test statistic

$$\xi_{LM} = i'S(S'S)^{-1}S'i = i'S(S'S)^{-1}S'i\ (i'i)^{-1}N = N\ \text{unc}R^2$$

with the uncentered R^2 of the auxiliary regression with residuals e

Remember: For the regression $y = X\beta + \varepsilon$

- OLS estimates for β: $b = (X'X)^{-1}X'y$
- the predictions for y: $\hat{y} = X(X'X)^{-1}X'y$
- uncentered R^2: $\text{unc}R^2 = \hat{y}'\hat{y}/y'y$

Also: $\sum_i s_i(\theta) = S'i$ and $\sum_i s_i(\theta)\, s_i(\theta)' = S'S$

The Urn Experiment: Three Tests of H_0: $p=p_0$

The urn experiment: test of H_0: $p = p_0$

The likelihood contribution of the *i*-th observation is

$$\log L_i(p) = y_i \log p + (1 - y_i) \log (1 - p)$$

This gives

$$s_i(p) = y_i/p - (1-y_i)/(1-p) \text{ and } I_i(p) = [p(1-p)]^{-1}$$

Wald test (with the unrestricted estimators $\hat{\theta}$ and $\hat{p}$)

$$\xi_W = N(R\hat{\theta} - q)\,[RV^{-1}R]^{-1}\,(R\hat{\theta} - q) = N(\hat{p} - p_0)\,[\hat{p}(1-\hat{p})]^{-1}\,(\hat{p} - p_0)$$

with $J = 1$, $R = I$; this gives

$$\xi_W = N\frac{(\hat{p} - p_0)^2}{\hat{p}(1-\hat{p})} = N\frac{(N_1 - Np_0)^2}{N(N - N_1)}$$

Example: In a sample of $N = 100$ balls, $N_1 = 40$ are red, i.e., $\hat{p} = 0.40$

- Test of H_0: $p_0 = 0.5$ results in

  $\xi_W = 4.167$, corresponding to a *p*-value of 0.041

The Urn Experiment: LR Test of H_0: $p=p_0$

Likelihood ratio test:

$$\xi_{LR} = 2\left(\log L(\hat{p}) - \log L(\tilde{p})\right)$$

with

$$\log L(\hat{p}) = N_1 \log(N_1 / N) + (N - N_1)\log(1 - N_1 / N)$$

$$\log L(\tilde{p}) = N_1 \log(p_0) + (N - N_1)\log(1 - p_0)$$

unrestricted estimator $\hat{p}$ and restricted estimator $\tilde{p}$

Example: In the sample of N = 100 balls, N_1 = 40 are red

- $\hat{p}$ =0.40, $\tilde{p}$ = p_0 = 0.5
- Test of H_0: p_0 = 0.5 results in

  ξ_W = 4.027, corresponding to a p-value of 0.045

The Urn Experiment: LM Test of H_0: $p=p_0$

Lagrange multiplier test:

with $$\tilde{\lambda} = \sum_i s_i(p)|_{p_0} = \frac{N_1}{p_0} - \frac{N - N_1}{1 - p_0} = \frac{N_1 - Np_0}{p_0(1 - p_0)}$$

and the inverted information matrix $[I(p)]^{-1} = p(1-p)$, calculated for the restricted case, the LM test statistic is

$$\xi_{LM} = N^{-1}\tilde{\lambda}[p_0(1-p_0)]\tilde{\lambda} = N(\hat{p} - p_0)[p_0(1-p_0)]^{-1}(\hat{p} - p_0)$$

$$= N\frac{(\hat{p} - p_0)^2}{p_0(1-p_0)}$$

Comparison of the test results

	Wald	LR	LM
Test statistic	4.167	4.027	4.000
p-value	0.041	0.045	0.046

Example:

- In the sample of N = 100 balls, 40 are red
- LM test of H_0: p_0 = 0.5 gives ξ_{LM} = 4.000 with *p*-value of 0.044

Contents

- Organizational Issues
- Overview of Contents
- Linear Regression: A Review
- Estimation of Regression Parameters
- Estimation Concepts
- ML Estimator: Idea and Illustrations
- ML Estimator: Notation and Properties
- ML Estimator: Two Examples
- Asymptotic Tests
- Some Diagnostic Tests

Normal Linear Regression: Scores

Log-likelihood function

$$\log L(\beta,\sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y_i - x_i'\beta)^2$$

Scores:

$$s_i(\beta,\sigma^2) = \begin{pmatrix} \dfrac{\partial \log L_i(\beta,\sigma^2)}{\partial\beta} \\ \dfrac{\partial \log L_i(\beta,\sigma^2)}{\partial\sigma^2} \end{pmatrix} = \begin{pmatrix} \dfrac{y_i - x_i'\beta}{\sigma^2} x_i \\ -\dfrac{1}{2\sigma^2} + \dfrac{1}{2\sigma^4}(y_i - x_i'\beta)^2 \end{pmatrix}$$

Covariance matrix

$$V = I(\beta,\sigma^2)^{-1} = \text{diag}(\sigma^2\Sigma_{xx}^{-1}, 2\sigma^4)$$

Testing for Omitted Regressors

Model: $y_i = x_i'\beta + z_i'\gamma + \varepsilon_i$, $\varepsilon_i \sim NID(0,\sigma^2)$; sample size N

Test whether the J regressors z_i are erroneously omitted:

- Fit the restricted model
- Apply the LM test to check H_0: $\gamma = 0$

First-order conditions give the scores

$$\frac{1}{\tilde{\sigma}^2}\sum_i \tilde{\varepsilon}_i x_i = 0, \quad \frac{1}{\tilde{\sigma}^2}\sum_i \tilde{\varepsilon}_i z_i, \quad -\frac{N}{2\tilde{\sigma}^2} + \frac{1}{2}\sum_i \frac{\tilde{\varepsilon}_i^2}{\tilde{\sigma}^4} = 0$$

with restricted ML estimators for β and σ^2; ML-residuals $\tilde{\varepsilon}_i = y_i - x_i'\hat{\beta}$

- Auxiliary regression of N-vector $i = (1, \ldots, 1)'$ on the scores gives the uncentered R^2
- The LM test statistic is $\xi_{LM} = N \,\text{unc}R^2$
- An asymptotically equivalent LM test statistic is NR_e^2 with R_e^2 from the regression of the ML residuals on x_i and z_i

Testing for Heteroskedasticity

Model: $y_i = x_i'\beta + \varepsilon_i$, $\varepsilon_i \sim NID$, $V\{\varepsilon_i\} = \sigma^2 h(z_i'\alpha)$, $h(.) > 0$ but unknown, $h(0) = 1$, $\partial/\partial\alpha\{h(.)\} \neq 0$, J-vector z_i

Test for homoskedasticity: Apply the LM test to check H_0: $\alpha = 0$

First-order conditions with respect to σ² and α give the scores

$$\tilde{\varepsilon}_i^2 - \tilde{\sigma}^2, \quad (\tilde{\varepsilon}_i^2 - \tilde{\sigma}^2) z_i'$$

with restricted ML estimators for β and σ²; ML-residuals $\tilde{\varepsilon}_i$

- Auxiliary regression of N-vector $i = (1, \ldots, 1)'$ on the scores gives the uncentered R^2
- LM test statistic $\xi_{LM} = N \text{ unc}R^2$; a version of Breusch-Pagan test
- An asymptotically equivalent version of the Breusch-Pagan test is based on NR_e^2 with R_e^2 from the regression of the squared ML residuals on z_i and an intercept
- Attention! No effect of the functional form of $h(.)$

Testing for Autocorrelation

Model: $y_t = x_t'\beta + \varepsilon_t$, $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, $v_t \sim NID(0,\sigma^2)$

LM test of H_0: $\rho = 0$

First-order conditions give the scores with respect to β and ρ

$$\tilde{\varepsilon}_t x_t', \quad \tilde{\varepsilon}_t \tilde{\varepsilon}_{t-1}$$

with restricted ML estimators for β and σ^2

- The LM test statistic is $\xi_{LM} = (T-1)\ \text{unc}R^2$ with the uncentered R^2 from the auxiliary regression of the N-vector $i = (1,\ldots,1)'$ on the scores
- If x_t contains no lagged dependent variables: products with x_t can be dropped from the regressors; $\xi_{LM} = (T-1)\ R^2$ with R^2 from $i = (1, \ldots, 1)'$ on the scores $\tilde{\varepsilon}_t \tilde{\varepsilon}_{t-1}$

An asymptotically equivalent test is the Breusch-Godfrey test based on NR_e^2 with R_e^2 from the regression of the ML residuals on x_t and the lagged residuals

Your Homework

1. Open the Greene sample file "greene7_8, Gasoline price and consumption", offered within the Gretl system. The dataset contains time series of annual observations from 1960 through 1995.The variables to be used in the following are: G = total U.S. gasoline consumption, computed as total expenditure of gas divided by the price index; Pg = price index for gasoline; Y = per capita (p.c.) disposable income; Pnc = price index for new cars; Puc = price index for used cars; Pop = U.S. total population in millions. Perform the following analyses and interpret the results:
   a. Produce and discuss a time series plot of the gasoline consumption (G), the disposable income (Y), and the U.S. total population (Pop).
   b. Produce and interpret the scatter plot of the p.c. gasoline consumption (Gpc) over the p.c. disposable income (Y).
   c. Fit the linear regression of log(Gpc) on the regressors log(Y) and Pg and give an interpretation of the outcome.

Your Homework, cont'd

d. Test for autocorrelation of the error terms using the LM test statistic $\xi_{LM} = (T\text{-}1)\, R^2$ with the uncentered R^2 from the auxiliary regression of the vector of ones $i = (1, \ldots, 1)'$ on the scores $(e_t{*}e_{t-1})$.

e. Test for autocorrelation using the Breusch-Godfrey test, the test statistic being TR_e^2 with R_e^2 from the regression of the residuals on the regressors and the lagged residuals e_{t-1}.

f. Use the Chow test to test for a structural break between 1979 and 1980.

2. Assume that the errors ε_t of the linear regression $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$ are NID(0, σ^2) distributed. (a) Determine the log-likelihood function of the sample for $t = 1, \ldots, T$; (b) derive (i) the first-order conditions and (ii) the ML estimators for β_1, β_2, and σ^2; (c) derive the asymptotic covariance matrix of the ML estimators for β_1 and β_2 on the basis (i) of the information matrix and (ii) of the score vector.