Practical Econometrics

for

Finance and Economics

(Econometrics 2)

Seppo Pynnönen and Bernd Pape
Department of Mathematics and Statistics,
University of Vaasa

1. Introduction

1.1 Econometrics

Econometrics is a discipline of statistics, specialized for using and developing mathematical and statistical tools for empirical estimation of economic relationships, testing economic theories, making economic predictions, and evaluating government and business policy.

Data: Nonexperimental (observational)

Major tool: Regression analysis (in wide sense)

1.2 Types of Economic Data

(a) Cross-sectional

Data collected at given point of time. E.g. a sample of households or firms, from each of which are a number of variables like turnover, operating margin, market value of shares, etc., are measured.

From econometric point of view it is important that the observations consist a *random sample* from the underlying *population*.

(b) Time Series Data

A time series consist of observations on a variable(s) over time. Typical examples are daily share prices, interest rates, CPI values.

An important additional feature over cross-sectional data is the *ordering* of the observations, which may convey important information.

An additional feature is *data frequency* which may require special attention.

(c) Pooled Cross-sections

Both time series and cross-section features.

For example a number of firms are randomly selected, say in 1990, and another sample is selected in 2000.

If in both samples the same features are measured, combining both years form a pooled cross-section data set.

Pooled cross-section data is analyzed much the same way as usual cross-section data.

However, it may be important to pay special attention to the fact that there are 10 years in between.

Usually the interest is whether there are some important changes between the time points. Statistical tools are usually the same as those used for analysis of differences between two independently sampled populations.

(d) Panel Data

Panel data (longitudinal data) consists of (time series) data for the same cross section units over time.

Allows to analyze much richer dependencies than pure cross section data.

Example 1.1: Job training data from Holzer et al. (1993) Are training subsidies effective? The Michigan experience, *Industrial and Labor Relations Review* 19, 625–636.

Excerpt from the data:

year	fcode	employ	sales	avgsal
1987	410032	100	4.70E+07	35000
1988	410032	131	4.30E+07	37000
1989	410032	123	4.90E+07	39000
1987	410440	12	1560000	10500
1988	410440	13	1970000	11000
1989	410440	14	2350000	11500
1987	410495	20	750000	17680
1988	410495	25	110000	18720
1989	410495	24	950000	19760
1987	410500	200	2.37E+07	13729
1988	410500	155	1.97E+07	14287
1989	410500	80	2.60E+07	15758
1987	410501	.	6000000	.
1988	410501	.	8000000	.
1989	410501	.	1.00E+07	.
etc				

1.3 The linear regression model

The linear regression model is the single most useful tool in econometrics.

Assumption: each observation i, $i = 1, \ldots, n$ is generated by the underlying process described by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i, \tag{1}$$

where y_i is the dependent or explained variable and $x_{i1}, x_{i2}, \ldots, x_{ik}$ are independent or explanatory variables, u is the error term, and $\beta_0, \beta_1, \ldots, \beta_k$ are regression coefficients (slope coefficients) (β_0 is called the intercept term or constant term).

A notational convenience:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \tag{2}$$

where $\boldsymbol{x}_i = (1, x_{1i}, \ldots, x_{ik})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ are $k+1$ column vectors.

Stacking the x-observation vectors to an $n \times (k+1)$ matrix

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_i' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{i1} & x_{i2} & \ldots & x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \tag{3}$$

we can write

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{4}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, and $\boldsymbol{u} = (u_1, \ldots, u_n)'$.

Example 1.2: In Example 1.1 the interest is whether grant for employee education decreases product failures. The estimated model is assumed to be

$$\log(\texttt{scrap}) = \beta_0 + \beta_1 \texttt{grant} + \beta_2 \texttt{grant}_{-1} + u, \tag{5}$$

where $\texttt{scrap}$ is scarp rate (per 100 items), $\texttt{grant} = 1$ if firm received grant in year t $\texttt{grant} = 0$ otherwise, and $\texttt{grant}_{-1} = 1$ if firm received grant in the previous year $\texttt{grant}_{-1} = 0$ otherwise.

The above model does not take into account that the data consist of three consecutive year measurements from the same firms (i.e., panel data).

Ordinary Least Squares (OLS) Estimation yields (Stata):

```
regress lscrap grant grant_1

   Source |       SS       df       MS      Number of obs =     162
----------+------------------------------   F(  2,   159) =    0.30
    Model |  1.34805124     2   .67402562   Prob > F      =  0.7395
 Residual |  354.397022   159  2.22891209   R-squared     =  0.0038
----------+------------------------------   Adj R-squared = -0.0087
    Total |  355.745073   161  2.20959673   Root MSE      =   1.493

---------------------------------------------------
   lscrap |      Coef.   Std. Err.      t    P>|t|
--+------------------------------------------------
    grant |   .0543534    .310501     0.18   0.861
  grant_1 |  -.2652102     .36995    -0.72   0.474
    _cons |   .4150563    .139828     2.97   0.003
---------------------------------------------------
```

Neither of the coefficients are statistically significant and `grant` has even positive sign, although close to zero.

Dealing later with the panel estimation we will see that the situation can be improved.

The problem with the above estimation is that the OLS assumptions are usually not met in panel data. This will be discussed in the next chapter.

The OLS assumptions are:

(*i*) $\mathbb{E}[u_i|\mathbf{X}] = 0$ for all i
(*ii*) $\mathbb{V}\text{ar}[u_i|\mathbf{X}] = \sigma_u^2$ for all i
(*iii*) $\mathbb{C}\text{ov}[u_i, u_j|\mathbf{X}] = 0$ for all $i \neq j$,
(*iv*) $\mathbf{X}$ is a $n \times (k+1)$ matrix with rank $k+1$

Remark 1.1: Assumption (1) implies

(6) $$\mathbb{C}\text{ov}[u_i, \mathbf{X}] = 0,$$

which is crucial in OLS-estimation.

Under assumptions (*i*)–(*iv*) the OLS estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{7}$$

is the Best Linear Unbiased Estimator (BLUE) of the regression coefficients $\boldsymbol{\beta}$ of the linear model in equation (4).

This is known as the Gauss-Markov theorem.

The variance covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\sigma_u^2, \tag{8}$$

which depends upon the unknown variance σ_u^2 of the error terms u_i.

In order to obtain an estimator for $\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}]$, use the residuals

$$\hat{u} = y - \boldsymbol{X}\widehat{\beta} \tag{9}$$

in order to calculate

$$s_u^2 = \hat{u}'\hat{u}/(n-k-1), \tag{10}$$

which is an unbiased estimator of the error variance σ_u^2.

Then replace σ_u^2 in (8) with s_u^2 in order to obtain

$$\widehat{\mathbb{V}\text{ar}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}s_u^2 \tag{11}$$

as an unbiased estimator of $\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}]$.

1.4 Regression statistics

Sum of Squares (SS) identity:

$$\text{SST} = \text{SSR} + \text{SSE}, \tag{12}$$

where

$$\text{Total: } \text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{13}$$

$$\text{Model: } \text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \tag{14}$$

$$\text{Residual: } \text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{15}$$

with $\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample mean.

Goodness of fit:

R-square, R^2

(16) $$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

Adjusted R-square (Adj R-square), $\bar{R}^2$

(17) $$\bar{R}^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} = 1 - \frac{s_u^2}{s_y^2},$$

where

(18) $$s_u^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{\text{SSE}}{n-k-1}$$

is an estimator of the variance $\sigma_u^2 = \mathbb{V}\text{ar}[u_i]$ of the error term ($s_u = \sqrt{s_u^2}$, "Root MSE" in the Stata output), and

(19) $$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

is the sample variance of y.

1.5 Inference

Assumption

(v) $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I})$,

where $\mathbf{I}$ is an $n \times n$ identity matrix.

Individual coefficient restrictions:

Hypotheses are of the form

(20) $$H_0 : \beta_j = \beta_j^*,$$

where β_j^* is a given constant.

t-statistics:

(21) $$t = \frac{\hat{\beta}_j - \beta_j^*}{s.e(\hat{\beta}_j)},$$

where

(22) $$s.e(\hat{\beta}_{j-1}) = s_u \sqrt{(\mathbf{X}'\mathbf{X})^{jj}},$$

and $(\mathbf{X}'\mathbf{X})^{jj}$ is the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. (First diagonal element for $\hat{\beta}_0$, second diagonal element for $\hat{\beta}_1$, etc.)

Confidence intervals:

A $100(1-\alpha)\%$ confidence interval for a single parameter is of the form

$$\widehat{\beta}_j \pm t_{\alpha/2} s.e(\widehat{\beta}_j), \tag{23}$$

where $t_{\alpha/2}$ is the $1-\alpha/2$ percentile of the t-distribution with $df = n-k-1$ degrees of freedom, which may be obtained from excel with the command TINV(α, df).

F-test:

The overall hypothesis that none of the explanatory variables influence the y-variable, i.e.,

(24) $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

is tested by an F-test of the form

(25) $$F = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)},$$

which is F-distributed with degrees of freedom $f_1 = k$ and $f_2 = n - k - 1$ if the null hypothesis is true.

General (linear) restrictions:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \boldsymbol{q}, \tag{26}$$

where $\mathbf{R}$ is a fixed $m \times (k+1)$ matrix and $\boldsymbol{q}$ is a fixed m-vector.

m indicates the number of independent linear restrictions imposed to the coefficients.

The alternative hypothesis is

$$H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{q}. \tag{27}$$

The null hypothesis in (26) can be tested with an F-statistic of the form

$$F = \frac{(SSE_R - SSE_U)/m}{SSE_U/(n-k-1)}, \tag{28}$$

which under the null hypothesis has the F-distribution with degrees of freedom $f_1 = m$ and $f_2 = n - k - 1$. SSE_R and SSE_U denote the residual sum of squares obtained in the restricted and unrestricted models, respectively.

Examle 1.3: Consider model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u. \tag{29}$$

In terms of the general linear hypothesis (26) testing for single coefficients, e.g.,

$$H_0 : \beta_1 = 0 \tag{30}$$

is obtained by selecting

$$\mathbf{R} = (0\ 1\ 0\ 0\ 0) \quad \text{and} \quad \boldsymbol{q} = 0. \tag{31}$$

The null hypothesis in (24), i.e.,

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \tag{32}$$

is obtained by selecting

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{q} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{33}$$

$$H_0 : \beta_1 + \beta_2 = 1,\ \beta_3 = \beta_4 \tag{34}$$

corresponds to

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{q} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{35}$$

Example 1.4. Consider the following consumption function (C =consumption, Y = disposable income):

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 C_{t-1} + u_t. \tag{36}$$

Then $\beta_1 = dC_t/dY_t$ is called the short-run MPC (marginal propensity to consume).

The long-run MPC $\beta_{\texttt{lrmpc}} = dE(C)/dE(Y)$ is

$$\beta_{\texttt{lrmpc}} = \frac{\beta_1}{1-\beta_2}. \tag{37}$$

Test the hypothesis whether the long run MPC = 1, i.e.,

$$H_0 : \frac{\beta_1}{1-\beta_2} = 1. \tag{38}$$

This is equivalent to $\beta_1 + \beta_2 = 1$.

Thus, the non-linear hypothesis (38) reduces in this case to the linear hypothesis

$$H_0 : \beta_1 + \beta_2 = 1, \tag{39}$$

and we can use the general linear hypothesis of the form (26) with

$$\mathbf{R} = (0\ \ 1\ \ 1) \quad \text{and} \quad \boldsymbol{q} = 1. \tag{40}$$

Remark 1.2: Hypotheses of the form (39) can be easily tested with the standard t-test by re-parameterizing the model.

Defining $Z_t = C_{t-1} - Y_t$, equation (36) is (statistically) equivalent to

$$C_t = \beta_0 + \gamma Y_t + \beta_2 Z_t + u_t, \tag{41}$$

where $\gamma = \beta_1 + \beta_2$.

Thus, in terms of (41) testing hypothesis (38) reduces to testing

$$H_0 : \gamma = 1, \tag{42}$$

which can be worked out with the usual t-statistic.

$$t = \frac{\hat{\gamma} - 1}{s.e(\hat{\gamma})}. \tag{43}$$

Example 1.5: Generalized Cobb-Douglas production function in transportation industry* Y_i = value added (output), L = labor, K = capital, and N = the number of establishments in the transportation industry.

$$\log(Y/N) = \beta_0 + \beta_1 \log(K/N) + \beta_2 \log(L/N) + u. \tag{44}$$

Estimation results:

```
Dependent Variable: LOG(VALUEADD/NFIRM)
Method: Least Squares
Sample: 1 25
Included observations: 25
==============================================================
Variable             Coefficient   Std. Error  t-Statistic   Prob.
--------------------------------------------------------------
C                    2.293263    0.107183    21.39582    0.0000
LOG(CAPITAL/NFIRM)   0.278982    0.080686    3.457639    0.0022
LOG(LABOR/NFIRM)     0.927312    0.098322    9.431359    0.0000
==============================================================
R-squared            0.959742  Mean dependent var      0.771734
Adjusted R-squared   0.956082  S.D. dependent var      0.899306
S.E. of regression   0.188463  Akaike info criter.    -0.387663
Sum squared resid    0.781403  Schwarz criterion      -0.241398
Log likelihood       7.845786  Hannan-Quinn criter.   -0.347095
F-statistic          262.2396  Durbin-Watson stat      1.937830
Prob(F-statistic)    0.000000
```

*Zellner, A and N. Revankar (1970). Generalized production functions, *Review of Economic Studies* 37, 241–250. *Data Source*: http://people.stern.nyu.edu/wgreene/Text/econometricanalysis.htm Table F7.2

According to the results the capital elasticity is 0.279 and the labor elasticity is 0.927, thus labor intensive.

Remark 1.3: Estimation of the regression parameters under the restrictions of the form $\mathbf{R}\boldsymbol{\beta} = \boldsymbol{q}$ are obtained by using restricted Least Squares, provided by modern statistical packages.

Let us test for the constant return to scale, i.e.,

$$H_0 : \beta_1 + \beta_2 = 1. \tag{45}$$

The general restricted hypothesis method (26) yields

```
=============================================
Test statistics            df         p-value
---------------------------------------------
F-statistic 14.82203    (1, 22)      0.0009
=============================================
```

which rejects the null hypothesis.

In order to demonstrate the re-parametrization approach, define the regression model

$$\log(Y/N) = \beta_0 + \gamma \log(K/N) + \beta_2 \log(L/K) + u \tag{46}$$

Estimation of the specification yields

```
Dependent Variable: LOG(VALUEADD/NFIRM)
Method: Least Squares
Sample: 1 25
Included observations: 25
==============================================================
Variable             Coefficient  Std. Error t-Statistic    Prob.
--------------------------------------------------------------
C                       2.293263    0.107183    21.39582   0.0000
LOG(CAPITAL/NFIRM)      1.206294    0.053584    22.51232   0.0000
LOG(LABOR/CAPITAL)      0.927312    0.098322    9.431359   0.0000
==============================================================
R-squared               0.959742  Mean dependent var       0.771734
Adjusted R-squared      0.956082  S.D. dependent var       0.899306
S.E. of regression      0.188463  Akaike info criter.     -0.387663
Sum squared resid       0.781403  Schwarz criterion       -0.241398
Log likelihood          7.845786  Hannan-Quinn criter.    -0.347095
F-statistic             262.2396  Durbin-Watson stat       1.937830
Prob(F-statistic)       0.000000
```

All the goodness-of-fit of these models are exactly the same, indicating the equivalence of the models in a statistical sense.

The null hypothesis of the constant returns to scale in terms of this model is

$$H_0 : \gamma = 1. \tag{47}$$

The t-value is

$$t = \frac{\widehat{\gamma} - 1}{s.e(\widehat{\gamma})} = \frac{1.206294 - 1}{0.053584} \approx 3.85 \tag{48}$$

with $p - \text{value} = 0.0009$, exactly the same as above, again rejecting the null hypothesis.

1.6 Nonlinear hypotheses

Economic theory implies sometimes nonlinear hypotheses.

In fact, the long-run MPC example is an example of non-linear hypothesis, which we could transform to a linear hypothesis.

This is not always possible.

For example a hypothesis of the form

$$H_0 : \beta_1\beta_2 = 1 \tag{49}$$

is nonlinear.

Non-linear hypotheses can be tested using Wald-test, Lagrange multiplier test, or Likelihood Ratio (LR) test.

All of these are under the null hypothesis asymptotically χ^2-distributed with degrees of freedom equal to the number of imposed restrictions on the parameters. These tests will be considered more closely later, after a brief discussion of maximum likelihood estimation.

1.7 Maximum Likelihood Estimation

Likelihood Function

Generally, suppose that the probability distribution of a random variable Y depends on a set of parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$, then the probability density for the random variable is denoted as $f_Y(y; \boldsymbol{\theta})$. If for example $Y \sim N(\mu, \sigma^2)$, then $\boldsymbol{\theta} = (\mu, \sigma^2)$ and

$$f_Y(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{50}$$

In probability calculus we consider $\boldsymbol{\theta}$ as given and use the density f_Y in order to calculate the probability that Y attains a value near y as

$$P(y - \Delta y \le Y \le y + \Delta y) = \int_{y-\Delta y}^{y+\Delta y} f_Y(u; \boldsymbol{\theta}) du.$$

In maximum likelihood estimation we consider the data point y as given and ask which parameter set $\boldsymbol{\theta}$ most likely produced it. In that context $f_Y(y; \boldsymbol{\theta})$ is called the *likelihood* of observation y on the random variable Y.

In statistical analysis we may regard a sample of observations $y_1, \ldots, y_n$ as realisations (observed values) of independent random variables $Y_1, \ldots, Y_n$. If all random variables are identically distributed, that is, they all share the same density $f(y_i; \boldsymbol{\theta})$, then the *likelihood function* of $(y_1, \ldots, y_n)$ is the product of the likelihoods of each point, that is,

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}). \tag{51}$$

Taking (natural) logarithms on both sides, we get the *log likelihood function*

$$\ell(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}). \tag{52}$$

Denoting the log-likelihoods of individual observations as $\ell_i(\boldsymbol{\theta}) = \log f(y_i; \boldsymbol{\theta})$, we can write (52) as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}). \tag{53}$$

Example 1.6: Under the normality assumption of the error term u_i in the regression

(54) $$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

(55) $$u_i \sim N(0, \sigma_u^2).$$

It follows that given $\mathbf{x}_i$

(56) $$y_i|\mathbf{x}_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_u^2).$$

Thus, with $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2)'$, the (conditional) density function is

(57) $$f(y_i|\mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_u^2}} e^{-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma_u^2}},$$

(58) $$\ell_i(\boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma_u^2 - \frac{1}{2}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma_u^2},$$

and

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma_u^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma_u^2}.$$

(59)

In matrix form (59) becomes

(60)

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma_u^2 - \frac{1}{2\sigma_u^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Maximum Likelihood Estimate

We say that the parameter vector $\boldsymbol{\theta}$ is *identified* or *estimable* if for any other parameter vector $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, for some data data $\mathbf{y}$, $L(\boldsymbol{\theta}^*; \mathbf{y}) \neq L(\boldsymbol{\theta}; \mathbf{y})$.

Given data $\mathbf{y}$ the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ of the parameter for which

$$L(\widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}), \tag{61}$$

i.e., the parameter value that maximizes the likelihood function.

The MLE of a parameter vector $\boldsymbol{\theta}$ solves

$$L(\boldsymbol{\theta}; \mathbf{y})/\partial\theta_i = 0 \quad (i = 1, \ldots, q), \tag{62}$$

provided the matrix of second derivatives is negative definite.

In practice it is usually more convenient to maximize the log-likelihood, such that the MLE of $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ which satisfies

$$l(\widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}} \ell(\theta). \tag{63}$$

Example 1.7.
Consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i, \tag{64}$$

with $u_i \sim N(0, \sigma_u^2)$.

Given a sample of observations (y_1, x_1), $(y_2, x_2), \ldots, (y_n, x_n)$, the log likelihood is

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma_u^2 - \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2/\sigma_u^2, \tag{65}$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_u^2)$.

The maximum of (65) can be found by setting the partial derivatives to zero, that is,

$$\begin{aligned}
\frac{\partial\ell}{\partial\beta_0} &= \textstyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)/\sigma_u^2 = 0,\\
\frac{\partial\ell}{\partial\beta_1} &= \textstyle\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)/\sigma_u^2 = 0,\\
\frac{\partial\ell}{\partial\sigma_u^2} &= -\frac{n}{2\sigma_u^2} + \frac{1}{2(\sigma_u^2)^2}\textstyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = 0.
\end{aligned} \tag{66}$$

Solving these gives

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{67}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{68}$$

and

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2, \tag{69}$$

where

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{70}$$

is the regression residual and $\bar{y}$ and $\bar{x}$ are the sample means of y_i and x_i.

In this particular case the ML estimators of the regression parameters, β_0 and β_1 coincide with the OLS estimators.

In OLS the error variance σ_u^2 estimator is

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{n}{n-2}\hat{\sigma}_u^2. \tag{71}$$

Properties of Maximum Likelihood Estimators

Let $\boldsymbol{\theta}_0$ be the population value of the parameter (vector) $\boldsymbol{\theta}$ and let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}_0$.

Then

(a) *Consistency*: $\operatorname{plim} \widehat{\theta} = \boldsymbol{\theta}_0$, i.e., $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$

(b) *Asymptotic normality*: $\widehat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}_0, \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\right)$ asymptotically, where

$$\mathbf{I}(\boldsymbol{\theta}_0) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]_{\theta=\theta_0}. \tag{72}$$

That is, $\widehat{\boldsymbol{\theta}}$ is *asymptotically normally distributed.* $\mathbf{I}(\boldsymbol{\theta}_0)$ is called the *Fisher information matrix* and

$$\mathbf{H} = \frac{\partial^2 \ell(\theta)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \tag{73}$$

is called the *Hessian* of the log-likelihood.

(c) Asymptotic efficiency: $\widehat{\boldsymbol{\theta}}$ is asymptotically efficient. That is, in the limit as the sample size grows, MLE is unbiased and its (limiting) variance is smallest among estimators that are asymptotically unbiased.

(d) *Invariance*: The MLE of $\gamma_0 = g(\boldsymbol{\theta}_0)$ is $g(\widehat{\boldsymbol{\theta}})$, where g is a (continuously differentiable) function.

Example 1.8: In Example 1.7 the MLE of the error variance σ_u^2 is given by $\widehat{\sigma}_u^2$ defined in equation (69). Using property (d), the MLE of the standard deviation $\sigma_u = \sqrt{\sigma_u^2}$ is $\widehat{\sigma}_u = \sqrt{\widehat{\sigma}_u^2}$.

Remark 1.5: The inverse of the Fisher information matrix defined in (72), $\mathbf{I}(\boldsymbol{\theta}_0)^{-1}$, plays a similiar role in MLE as does $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ in OLS. I.e. it may be used to find the standard errors of the ML estimators.

Example 1.9: Consider MLE from n observations on a normally distributed random variable with unknown parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)$. The log likelihood function is in analogy to (65)

(74)

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2/\sigma^2.$$

The first partial derivatives are

(75)

$$\frac{\partial\ell}{\partial\mu} = \frac{\sum(y_i-\mu)}{\sigma^2}, \qquad \frac{\partial\ell}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Setting these equal to zero yields the ML estimators

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2. \tag{76}$$

The second derivatives are

$$\frac{\partial^2\ell}{\partial\mu^2} = -\frac{n}{\sigma^2}, \qquad \frac{\partial^2\ell}{(\partial\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{\sum(y_i-\mu)^2}{(\sigma^2)^3}, \tag{77}$$

$$\text{and} \quad \frac{\partial^2\ell}{\partial\mu\partial\sigma^2} = \frac{\partial^2\ell}{\partial\sigma^2\partial\mu} = -\frac{\sum(y_i-\mu)}{(\sigma^2)^2},$$

such that the Hessian matrix becomes

$$\mathbf{H} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{\sum(y_i-\mu)}{(\sigma^2)^2} \\ -\frac{\sum(y_i-\mu)}{(\sigma^2)^2} & \frac{n}{2(\sigma^2)^2} - \frac{\sum(y_i-\mu)^2}{(\sigma^2)^3} \end{pmatrix}. \tag{78}$$

Taking expectations and multiplying with -1 yields the information matrix

$$\mathbf{I}(\boldsymbol{\theta}_0) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{pmatrix} \tag{79}$$

with inverse

$$\mathbf{I}(\boldsymbol{\theta}_0)^{-1} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}. \tag{80}$$

The standard errors of the ML estimators are found by taking the square roots of the diagonal elements of $\mathbf{I}(\hat{\boldsymbol{\theta}}_0)^{-1}$, that is

$\dfrac{\hat{\sigma}}{\sqrt{n}}$ is the standard error of $\hat{\mu}$, and

$\dfrac{\hat{\sigma}^2}{\sqrt{2/n}}$ is the standard error of $\hat{\sigma}^2$.

These may be used to construct confidence intervals in the usual way.

1.8 Likelihood Ratio, Wald, and Lagrange Multiplier tests

For testing general restrictions of the form

$$H_0 : c(\boldsymbol{\theta}) = 0, \tag{81}$$

where $c(\cdot)$ is some (vector valued) function, there are three general purpose test methods to be discussed on the following slides.

<u>Remark 1.6</u>: We could specify the above hypothesis alternatively as

$$H_0 : r(\boldsymbol{\theta}) = \mathbf{q}, \tag{82}$$

where $r(\cdot)$ is some function and $\mathbf{q}$ is some constant.

Defining $c(\boldsymbol{\theta}) = r(\boldsymbol{\theta}) - \mathbf{q}$ reduces then to hypothesis (81) stated above.

1. Likelihood ratio test (LR-test)

$$LR = -2\log\left(\frac{L_R}{L_U}\right) = -2(\ell_R - \ell_U), \tag{83}$$

where

$$L_R = \max_{\theta, c(\theta)=0} L(\theta)$$

is the maximum of the likelihood under the restriction of hypothesis (81),

$$L_U = \max_{\theta} L(\theta)$$

is the unrestricted maximum of the likelihood function ($\ell_U = \log L_U$ and $\ell_R = \log L_R$).

Remark 1.7: Use of the LR test requires computing both the restricted MLE of $\boldsymbol{\theta}$ (to compute ℓ_R) and the unrestricted MLE (to compute ℓ_U).

Example 1.10.
The LR test for the linear regression model is

$$(84) \qquad LR = n(\log SSE_R - \log SSE_U),$$

where SSE_R and SSE_U are the residual sum of squares for the restricted and unrestricted model respectively, where this time arbitrary (not only linear) restrictions are allowed.

To see that (84) holds, insert the residual sum of squares $SSE = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$ and the ML estimate $\hat{\sigma}_u^2 = SSE/n$ into (60):
(85)

$$\begin{aligned}\ell(\boldsymbol{\theta}_0) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{SSE}{n}\right) - \frac{1}{2}\frac{n}{SSE}\cdot SSE \\ &= -\frac{n}{2}\left(1 + \log(2\pi) + \log\left(\frac{SSE}{n}\right)\right).\end{aligned}$$

Hence

$$\begin{aligned}LR &= -2(\ell_R - \ell_U) \\ &= n[\log(SSE_R/n) - \log(SSE_U/n)] \\ &= n(\log SSE_R - \log SSE_U).\end{aligned}$$

2. Wald test

$$W = c(\widehat{\boldsymbol{\theta}})'\mathbf{V}^{-1}c(\widehat{\boldsymbol{\theta}}), \tag{86}$$

where $\mathbf{V}$ is the asymptotic variance covariance matrix of $c(\widehat{\boldsymbol{\theta}})$.

Remark 1.8: Use of the Wald test requires only to find the unrestricted MLE.

The Wald test for linear regression with normally distributed errors is

$$W = \frac{SSE_R - SSE_U}{SSE_U/(n-k-1)}, \tag{87}$$

where k is the number of regressors (without the constant).

3. Lagrange multiplier test (LM)

(88) $$LM = \left(\frac{\partial \ell(\hat{\boldsymbol{\theta}}_R)}{\partial \boldsymbol{\theta}}\right)' \left[\mathbf{I}(\hat{\boldsymbol{\theta}}_R)\right]^{-1} \left(\frac{\partial \ell(\hat{\boldsymbol{\theta}}_R)}{\partial \boldsymbol{\theta}}\right),$$

where $\hat{\boldsymbol{\theta}}_R$ is the restricted MLE satisfying the restriction $c(\hat{\boldsymbol{\theta}}_R) = 0$ of the general hypothesis (81).

Remark 1.9: Use of the LM test requires only the restricted MLE.

The Lagrange multiplier test for linear regression with normally distributed errors is

(89) $$LM = \frac{SSE_R - SSE_U}{SSE_R/(n - k + q - 1)},$$

where k is the number of regressors (without constant) and q is the number of restrictions.

Under the null hypothesis (81) each of these test statistics is asymptotically χ^2-distributed with degrees of freedom equal to the number of restrictions q.

Thus, they are asymptotically equivalent. In small samples numerical values may differ, however. Usually the LR test is preferred, because it can be shown under fairly general conditions to be the most powerful test.

Bear in mind that while the tests can be developed for arbitrary distributions of the error term, their exact form depends upon that distribution. I.e. the test statistics (84), (87) and (89) apply only for regressions with normally distributed error terms.

Also bear in mind that the tests apply only in large samples.