1. Maximum likelihood method (ML)

1.1 Likelihood function

Problem:

While the distribution and thus the probability or density function f(y; θ) of a random variable y (e.g. Bernoulli distribution, Poisson distribution, normal distribution) is known, some parameters of this distribution that are summarized in the vector $\theta = (\theta_1, \theta_2,\ldots, \theta_m)$‘ are unknown and have to be estimated

→ For this purpose a random sample $y_1,\ldots, y_n$ of n observations is drawn from the corresponding distribution where $f(y_i; \theta)$ is the probability or density function of y_i for observation i

Due to the independence of the y_i, the joint probability function (for discrete random variables) or density function (for continuous random variables) of the $y_1,\ldots, y_n$ is the product of the individual probability or density functions:

$$f\,(y_1,..., y_n; \theta) = f\left(y_1; \theta\right) \cdots f\left(y_n; \theta\right) = \prod_{i=1}^{n} f\left(y_i; \theta\right)$$

If this function is not considered as a function of the random sample $y_1,\ldots, y_n$ given the parameters in θ, but as a function of θ for a given random sample $y_1,\ldots, y_n$, it can be interpreted as a likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f\left(y_i; \theta\right)$$

It should be noted that the likelihood function as well as all following derived functions are random variables (or random vectors or random matrixes) before the drawing of a random sample.

The idea of the maximum likelihood method is to find the value $\hat{\theta}$ of θ that maximizes the likelihood function on the basis of the random sample $y_1,\ldots, y_n$. However, the maximization procedure is generally not based on the likelihood function, but on the log-likelihood function, i.e. the following natural logarithm:

$$\log L(\theta) = \log f\left(y_1; \theta\right) + \cdots + \log f\left(y_n; \theta\right) = \sum_{i=1}^{n} \log f\left(y_i; \theta\right)$$

The maximization of logL(θ) leads to the same estimator $\hat{\theta}$ as the maximization of L(θ). Advantages of the use of the log-likelihood function:

- The use of logL(θ) avoids extremely small values in the case of discrete random variables (due to the multiplication of probabilities and thus values that are smaller than one) and extremely high values in the case of continuous random variables (if values that are higher than one in the density functions are multiplied)
- Generally, the maximization process to receive $\hat{\theta}$ is much simpler for the log-likelihood function than for the likelihood function

Example: Bernoulli distribution

The probability function of a Bernoulli distributed random variable y with parameter p is:

$$f\left(y; p\right) = \begin{cases} 1 - p & \text{for } y = 0 \\ p & \text{for } y = 1 \\ 0 & \text{else} \end{cases}$$

or

$$f\left(y; p\right) = (1\text{-}p)^{1\text{-}y} p^{y} \quad \text{for } y = 0, 1$$

If a random sample $y_1,\ldots, y_n$ of n observations is drawn from a Bernoulli distribution with parameter p, this leads to the following likelihood and log-likelihood functions:

$$L(\theta) = L(p) = (1\text{-}p)^{1\text{-}y_1} p^{y_1} \cdots (1\text{-}p)^{1\text{-}y_n} p^{y_n} = \prod_{i=1}^{n} (1\text{-}p)^{1\text{-}y_i} p^{y_i}$$

$$\log L(\theta) = \log L(p) = \left[(1\text{-}y_1)\log(1\text{-}p) + y_1 \log p\right] + \cdots + \left[(1\text{-}y_n)\log(1\text{-}p) + y_n \log p\right]$$

$$= \sum_{i=1}^{n} \left[(1\text{-}y_i)\log(1\text{-}p) + y_i \log p\right]$$

Maximization approach:

$$\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m\right)' = \arg\max_{\theta}\left[\log L\left(\theta\right)\right] = \arg\max_{\theta}\left[\sum_{i=1}^{n} \log f\left(y_i;\, \theta\right)\right]$$

The ML estimator $\hat{\theta}$ is therefore the value of θ that maximizes the log-likelihood function.

The first derivative of the log-likelihood function is called score (function):

$$s(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \begin{pmatrix} \dfrac{\partial \log L(\theta)}{\partial \theta_1} \\ \dfrac{\partial \log L(\theta)}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial \log L(\theta)}{\partial \theta_m} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \sum_{i=1}^{n} \log f\left(y_i;\, \theta\right)}{\partial \theta_1} \\ \dfrac{\partial \sum_{i=1}^{n} \log f\left(y_i;\, \theta\right)}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial \sum_{i=1}^{n} \log f\left(y_i;\, \theta\right)}{\partial \theta_m} \end{pmatrix}$$

Due to the sum of the terms in the log-likelihood function, the score is also an additive function:

$$s(\theta) = \sum_{i=1}^{n} s_i(\theta) \;=\; \sum_{i=1}^{n} \frac{\partial \log f\left(y_i;\,\theta\right)}{\partial \theta}$$

Expectation of the score at the true, but unknown parameter vector θ:

$$E\left[s(\theta)\right] = 0$$

The second derivative of the log-likelihood function is called Hessian matrix:

$$H(\theta) = \frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'} = \begin{pmatrix} \frac{\partial^2 \log L(\theta)}{\left(\partial\theta_1\right)^2} & \frac{\partial^2 \log L(\theta)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial\theta_1\partial\theta_m} \\ \frac{\partial^2 \log L(\theta)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 \log L(\theta)}{\left(\partial\theta_2\right)^2} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial\theta_2\partial\theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\theta)}{\partial\theta_m\partial\theta_1} & \frac{\partial^2 \log L(\theta)}{\partial\theta_m\partial\theta_2} & \cdots & \frac{\partial^2 \log L(\theta)}{\left(\partial\theta_m\right)^2} \end{pmatrix}$$

Due to the sum of the terms in the log-likelihood function, the Hessian matrix is also an additive function:

$$\mathrm{H}(\theta) = \sum_{i=1}^{n} \mathrm{H}_i(\theta) \;=\; \sum_{i=1}^{n} \frac{\partial^2 \log f\left(y_i;\, \theta\right)}{\partial\theta\partial\theta'}$$

A necessary condition for the ML estimator $\hat{\theta}$ is that the score for this value of θ is zero:

$$\left.\frac{\partial \log L(\theta)}{\partial\theta}\right|_{\hat{\theta}} = s(\hat{\theta}) = \sum_{i=1}^{n} s_i(\hat{\theta}) \;=\; 0$$

As a consequence, the maximization process of the ML estimator can be characterized as follows:

$$\hat{\theta} = \arg \underset{\theta}{\text{solves}} \left[s(\theta) = \sum_{i=1}^{n} s_i(\theta) \;=\; \sum_{i=1}^{n} \frac{\partial \log f\left(y_i;\, \theta\right)}{\partial\theta} = 0 \right]$$

Additional necessary and sufficient condition for the ML estimator $\hat{\theta}$:

The Hessian matrix for this value of θ must be negative definite (if there is a solution at an inner point of the parameter space). The maximum can be local or global. In many simple cases the log-likelihood function is globally concave so that the solution of the first order condition leads to a unique and global maximum of the log-likelihood function.

Example: Bernoulli distribution

If a random sample $y_1,\ldots, y_n$ of n observations is drawn from a Bernoulli distribution with parameter p, this leads to the following score:

$$\frac{\partial \log L(p)}{\partial p} = s(p) = \sum_{i=1}^{n} s_i(p) = \sum_{i=1}^{n} \frac{\partial \left[(1\text{-}y_i)\log(1\text{-}p) + y_i \log p\right]}{\partial p} =$$

$$\sum_{i=1}^{n}\left[-\frac{1\text{-}y_i}{1\text{-}p} + \frac{y_i}{p}\right] = \sum_{i=1}^{n}\left[\frac{y_i(1\text{-}p) - p(1\text{-}y_i)}{p(1\text{-}p)}\right] = \sum_{i=1}^{n}\left[\frac{y_i\text{-}p}{p(1\text{-}p)}\right]$$

From the maximization of the log-likelihood function and thus equalizing the score with zero, it follows that the ML estimator of the probability p is equal to the sample mean, i.e. the proportion of ones in the sample:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

The following second derivative of the log-likelihood function and thus the Hessian matrix (i.e. scalar) is generally negative for all possible samples $y_1,\ldots, y_n$:

$$H(p) = \sum_{i=1}^{n}\left[-\frac{y_i}{p^2} - \frac{1\text{-}y_i}{(1\text{-}p)^2}\right]$$

All previous concepts for unconditional models with a random variable y can be simply transferred to conditional models and thus microeconometric models with a dependent variable y and k explanatory variables (and a constant) which are summarized in the vector x = (1, x_1,…, x_k)‘. x_i = (1, x_{i1},…, x_{ik})‘ is therefore the vector of explanatory variables (including a constant) for observation i.

With the conditional probability or density function f(y|x; θ) of y and a random sample (y_i, x_i) (i = 1,…, n) it follows for the conditional joint probability or density function:

$$\mathrm{f}\,(\mathrm{y}_1,\ldots,\mathrm{y}_n|\,\mathrm{x}_1,\ldots,\mathrm{x}_n;\theta) = \mathrm{f}\left(\mathrm{y}_1|\,\mathrm{x}_1;\theta\right)\cdots\ \mathrm{f}\left(\mathrm{y}_n|\,\mathrm{x}_n;\theta\right) = \prod_{i=1}^{n}\mathrm{f}\left(\mathrm{y}_i|\,\mathrm{x}_i;\theta\right)$$

It follows for the log-likelihood function, the score, the Hessian matrix, and the maximization approach:

$$\mathrm{logL}(\theta) = \sum_{i=1}^{n}\mathrm{logf}\left(\mathrm{y}_i|\,\mathrm{x}_i;\theta\right)$$

$$\mathrm{s}(\theta) = \sum_{i=1}^{n}\frac{\partial\mathrm{logf}\left(\mathrm{y}_i|\,\mathrm{x}_i;\theta\right)}{\partial\theta}$$

$$\mathrm{H}(\theta) = \sum_{i=1}^{n}\frac{\partial^2\mathrm{logf}\left(\mathrm{y}_i|\,\mathrm{x}_i;\theta\right)}{\partial\theta\partial\theta'}$$

$$\hat{\theta} = \arg\max_{\theta}\left[\sum_{i=1}^{n}\mathrm{logf}\left(\mathrm{y}_i|\,\mathrm{x}_i;\theta\right)\right]$$

Example 1: General classical linear regression models (I)

With $x_i = (1, x_{i1},\ldots, x_{ik})'$ as the vector of k explanatory variables (including a constant) and the (k+1)-dimensional parameter vector $\beta = (\beta_0, \beta_1,\ldots, \beta_k)'$, the linear regression model has the following form (for i = 1,…, n):

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

Due to the normality assumption for the error term ε_i it follows:

$$y_i | x_i \sim N(\beta' x_i; \sigma^2)$$

Density function of y_i:

$$f\left(y_i | x_i; \beta, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \beta' x_i}{\sigma}\right)^2\right] = \left(2\pi\right)^{-\frac{1}{2}}\left(\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \beta' x_i}{\sigma}\right)^2\right]$$

Log-likelihood function on the basis of a random sample of n pairs of observations (y_i, x_i) (i = 1,…, n):

$$\log L(\theta) = \log L(\beta, \sigma^2) = \sum_{i=1}^{n}\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\left(\frac{y_i - \beta' x_i}{\sigma}\right)^2\right]$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta' x_i)^2$$

Example 1: General classical linear regression models (II)

First order conditions for maximizing the log-likelihood function:

$$\frac{\partial \log L(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i (y_i - \beta' x_i) = 0$$

$$\frac{\partial \log L(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} (y_i - \beta' x_i)^2 = 0$$

The partial derivative of logL(β, σ^2) with respect to β is a (k+1)-dimensional vector. It follows for the ML estimator of β:

$$\hat{\beta} = \underbrace{\left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}}_{(k+1)\times(k+1)} \underbrace{\left(\sum_{i=1}^{n} x_i y_i\right)}_{(k+1)\times 1}$$

The inverse of the (k+1)×(k+1) matrix exists if there is no exact linear relationship between the explanatory variables. The ML estimator $\hat{\beta}$ is therefore identical to the corresponding OLS estimator (in the classical linear regression model). The reason for this is that the first order conditions for optimizing the respective objective functions are identical.

Example 1: General classical linear regression models (III)

In contrast, the following ML estimator of σ² differs from the familiar OLS estimator of the variance of the error term:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}$$

While this estimator is biased, it is consistent and asymptotically efficient (see below).

The following Hessian matrix is generally negative definite so that these parameter values actually refer to the maximum of the log-likelihood function:

$$H(\beta, \sigma^2) = \begin{bmatrix} -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i' & -\frac{1}{\sigma^4}\sum_{i=1}^{n} x_i (y_i - \beta' x_i) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n} (y_i - \beta' x_i) x_i' & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n} (y_i - \beta' x_i)^2 \end{bmatrix}$$

Example 2: Determinants of (the logarithm of) wages (I)

By using a classical linear regression model, the effect of the years of education (educ), the years of labor market experience (exper), and the years with the current employer (tenure) on the logarithm of hourly wage (logwage) is examined on the basis of n = 526 workers:

$$\text{logwage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + \varepsilon$$

However, instead of an OLS estimation, we now consider an ML estimation of the parameters. Such estimations can be conducted with the Stata command ("ml model") including the lf method, which assumes the independence of the observations and uses numerical first and second derivatives of the log-likelihood function. For different ML estimations with Stata, separate programs have to be written which specify the log-likelihood functions.

Components of the program ("linearregression", which is an arbitrary term):

- Three arguments: Logarithm of individual density function ("loglikelihood"), $\beta' x_i$ ("betaxi"), σ^2 ("sigmasquared")
- Logarithm of density function of y_i (i.e. individual log-likelihood function):

$$\log f\left(y_i | x_i; \beta, \sigma^2\right) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - \beta' x_i)^2$$

Example 2: Determinants of (the logarithm of) wages (II)

The Stata command “ml model lf” includes the definition of the dependent and explanatory variables and the Stata command “ml maximize” finally presents the estimation results. This ML estimation leads to the following results:

```
program linearregression
args loglikelihood betaxi sigmasquared
quietly replace `loglikelihood' = (-1/2)*ln(2*_pi) - (1/2)*ln(`sigmasquared')-
(1/(2*`sigmasquared'))*($ML_y1-`betaxi')^2
end
ml model lf linearregression (Betas: logwage = educ exper tenure) (Variance:)
ml maximize
```

```
                                                  Number of obs   =        526
                                                  Wald chi2(3)    =     243.02
Log likelihood = -313.54779                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Betas        |
        educ |    .092029    .007302    12.60   0.000     .0777173    .1063406
       exper |   .0041211   .0017167     2.40   0.016     .0007564    .0074858
      tenure |   .0220672   .0030819     7.16   0.000     .0160269    .0281076
       _cons |   .2843595   .1037935     2.74   0.006      .080928    .4877909
-------------+----------------------------------------------------------------
Variance     |
       _cons |   .1928813   .0118936    16.22   0.000     .1695704    .2161923
------------------------------------------------------------------------------
```

Example 2: Determinants of (the logarithm of) wages (III)

In contrast, the OLS estimation with Stata leads to the following results (see also chapter 0):

```
reg logwage educ exper tenure
```

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   80.39
       Model |  46.8741806     3  15.6247269           Prob > F      =  0.0000
    Residual |  101.455582   522  .194359353           R-squared     =  0.3160
-------------+------------------------------           Adj R-squared =  0.3121
       Total |  148.329763   525  .282532881           Root MSE      =  .44086

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .092029   .0073299    12.56   0.000     .0776292    .1064288
       exper |   .0041211   .0017233     2.39   0.017     .0007357    .0075065
      tenure |   .0220672   .0030936     7.13   0.000     .0159897    .0281448
       _cons |   .2843595   .1041904     2.73   0.007     .0796755    .4890435
------------------------------------------------------------------------------
```

The ML and OLS estimates of the parameters in β are therefore completely identical (the estimation of the variance σ^2 is no component of the OLS estimation). Only the estimated standard deviations of the estimated parameters are slightly different due to imprecisions in the maximization process of the log-likelihood function.

1.2 Properties of the ML estimator

Finite sample properties of the ML estimator $\hat{\theta}$ of θ:

- $\hat{\theta}$ is often a biased estimator of θ (e.g. the expectation of $\hat{\sigma}^2$ is not equal to σ^2 in the classical linear regression model, whereas $\hat{\beta}$ is unbiased in this case which is, however, an exception)
- $\hat{\theta}$ is generally not normally distributed (the normal distribution of $\hat{\beta}$ in the classical linear regression model is again an exception)
- The generally unknown small sample properties of ML estimators can be examined by Monte Carlo simulations for specific (microeconometric) models and specific parameter values

Asymptotic properties of the ML estimator $\hat{\theta}$ of θ (under several regularity conditions and if the underlying model is correctly specified):

- Consistency, i.e. $P(|\hat{\theta} - \theta| > \xi)$ converges (for $\xi > 0$) to zero for $n \rightarrow \infty$ or $\text{plim}(\hat{\theta}) = \theta$. This means that the asymptotic distribution of $\hat{\theta}$ is centered at θ and its variance goes to zero. An alternative notation for the consistency is:

$$\hat{\theta} \xrightarrow{p} \theta$$

- Asymptotic normality
- Asymptotic efficiency

The asymptotic normality does not directly refer to the ML estimator $\hat{\theta}$, but to $\sqrt{n}(\hat{\theta}\text{-}\theta)$:

$$\sqrt{n}(\hat{\theta} - \theta) \overset{a}{\sim} N\left(0; I(\theta)^{-1}\right) \quad \text{or} \quad \sqrt{n}(\hat{\theta} - \theta) \overset{d}{\to} N\left(0; I(\theta)^{-1}\right)$$

This means that $\sqrt{n}(\hat{\theta}\text{-}\theta)$ converges in distribution to the normal distribution, i.e. is asymptotically normally distributed with an expectation vector zero and the variance covariance matrix $I(\theta)^{-1}$. The matrix $I(\theta)$ is called information matrix and has the following form:

$$I(\theta) = -E\left(\frac{\partial^2 \log f_i(y_i;\theta)}{\partial\theta\partial\theta'}\right) = -E\left[H_i(\theta)\right]$$

The inverse of the information matrix is the Cramer Rao lower bound which implies that the difference between this matrix and the corresponding variance covariance matrix for any other consistent estimator of θ, for which $\sqrt{n}(\hat{\theta}\text{-}\theta)$ is asymptotically normally distributed, is negative definite. Since $\sqrt{n}(\hat{\theta}\text{-}\theta)$ reaches this lower bound, the ML estimator is asymptotically efficient.

Information matrix equality at the true, but unknown parameter vector θ:

$$I(\theta) = -E\left[H_i(\theta)\right] = E\left[s_i(\theta)s_i(\theta)'\right] = Var\left[s_i(\theta)\right]$$

This equality means that the variance covariance matrix of the score for observation i is identical to the negative expectation of the Hessian matrix for i and thus the information matrix.

From the asymptotic normality of $\sqrt{n}(\hat{\theta}-\theta)$ it follows that the ML estimator $\hat{\theta}$ is approximately normally distributed for large but finite samples of n observations:

$$\hat{\theta} \overset{\text{appr}}{\sim} N\left(\theta; \left[nI(\theta)\right]^{-1}\right)$$

The variance covariance matrix of $\hat{\theta}$ thus has the following form:

$$\mathrm{Var}(\hat{\theta}) = \begin{pmatrix} \mathrm{Var}(\hat{\theta}_1) & \mathrm{Cov}(\hat{\theta}_1,\hat{\theta}_2) & \vdots & \mathrm{Cov}(\hat{\theta}_1,\hat{\theta}_m) \\ \mathrm{Cov}(\hat{\theta}_2,\hat{\theta}_1) & \mathrm{Var}(\hat{\theta}_2) & \vdots & \mathrm{Cov}(\hat{\theta}_2,\hat{\theta}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{\theta}_m,\hat{\theta}_1) & \mathrm{Cov}(\hat{\theta}_m,\hat{\theta}_2) & \cdots & \mathrm{Var}(\hat{\theta}_m) \end{pmatrix} = \left[nI\left(\theta\right)\right]^{-1} = -E\left[nH_i\left(\theta\right)\right]^{-1}$$

However, the information matrix (and this symmetric and positive definite variance covariance matrix) is unknown in practice since it depends on the unknown θ and thus has to be (consistently) estimated, e.g. for statistical tests and the construction of confidence intervals. $\mathrm{Var}(\hat{\theta})$ can be estimated by including the ML estimator $\hat{\theta}$ instead of the true parameter vector:

$$\mathrm{V\hat{a}r}(\hat{\theta}) = -E\left[nH_i(\hat{\theta})\right]^{-1}$$

However, it is often not possible to obtain an exact expression for the expectation.

In practice the following two estimators for the variance covariance matrix (including consistent estimators of the information matrix) are commonly used:

$$\hat{\text{Var}}(\hat{\theta})_1 = \left[n\hat{I}(\hat{\theta})_1\right]^{-1} = -\left[n\frac{1}{n}\sum_{i=1}^{n} H_i(\hat{\theta})\right]^{-1} = -\left[\sum_{i=1}^{n} H_i(\hat{\theta})\right]^{-1}$$

$$\hat{\text{Var}}(\hat{\theta})_2 = \left[n\hat{I}(\hat{\theta})_2\right]^{-1} = \left[n\frac{1}{n}\sum_{i=1}^{n} s_i(\hat{\theta})s_i(\hat{\theta})'\right]^{-1} = \left[\sum_{i=1}^{n} s_i(\hat{\theta})s_i(\hat{\theta})'\right]^{-1}$$

The first estimator exclusively incorporates the Hessian matrix at the ML estimator for the observations, whereas the second approach exclusively includes the score at the ML estimator and is called the outer product of the gradient.

A third approach follows from the quasi maximum likelihood theory and includes a robust estimator of the information matrix that can even be consistent with respect to the true information matrix if the underlying model is misspecified. It includes both the Hessian matrix and the score at the ML estimator:

$$\hat{\text{Var}}(\hat{\theta})_3 = \left[n\hat{I}(\hat{\theta})_3\right]^{-1} = \left[\sum_{i=1}^{n} H_i(\hat{\theta})\right]^{-1}\left[\sum_{i=1}^{n} s_i(\hat{\theta})s_i(\hat{\theta})'\right]\left[\sum_{i=1}^{n} H_i(\hat{\theta})\right]^{-1}$$

$$= \hat{\text{Var}}(\hat{\theta})_1 \hat{\text{Var}}(\hat{\theta})_2^{-1} \hat{\text{Var}}(\hat{\theta})_1$$

All three estimators of the information matrix are asymptotically equivalent, but can be very different in small samples.

Invariance principle:

The ML estimator of any function h(θ) of θ is the function $h(\hat{\theta})$ of the ML estimator $\hat{\theta}$. Since it is an ML estimator, it is consistent, but usually not unbiased.

Delta method:

$$\hat{\mathrm{Var}}\left[h(\hat{\theta})\right] = \left[\frac{\partial h(\theta)}{\partial \theta}\right]_{\hat{\theta}}' \hat{\mathrm{Var}}(\hat{\theta}) \left[\frac{\partial h(\theta)}{\partial \theta}\right]_{\hat{\theta}}$$

The estimator of the variance covariance matrix of the ML estimator $h(\hat{\theta})$ of the function h(θ) is a quadratic form of the derivatives of h(θ) and an estimator of the variance covariance matrix of $\hat{\theta}$.

Problem of maximization of the log-likelihood function:

A very efficient way to find the ML estimator $\hat{\theta}$ in the maximization process is the analytical optimization by equaling the score to zero and solving for the maximizing parameters. However, this approach requires a closed form solution of the first order condition for maximizing the log-likelihood function. Indeed, in many cases it is not available due to the non-linearity of the log-likelihood function so that iterative numerical maximization algorithms have to be applied such as the Newton Raphson algorithm or the BHHH algorithm.

1.3 Statistical testing

Problem:

The ML estimation of (microeconometric) models leads to a corresponding point estimate, but does not account for the sampling variability which is the basis for the construction of confidence intervals and for statistical tests

Testable hypotheses refer to restrictions on the parameter space. The following general null and alternative hypotheses are based on q such restrictions:

$$H_0\text{: } c(\theta) = 0 \quad \Leftrightarrow \quad \begin{cases} c_1(\theta) = 0 \\ \vdots \\ c_q(\theta) = 0 \end{cases}$$

$$H_1\text{: } c(\theta) \neq 0$$

In the simplest case, the dimension of the function c(θ) is q = 1 and c(θ) refers to specific values of one parameter θ_l of the vector θ which leads to the following null hypothesis (with an arbitrary constant a):

$$H_0\text{: } \theta_l = a$$

On the basis of the ML estimation of (microeconometric) models, these null hypotheses can be statistically tested with the Wald test, the likelihood ratio test, or the score test which are asymptotically equivalent.

Restricted ML estimation:

An ML estimator $\hat{\theta}_r$ is called restricted ML estimator if the underlying ML estimation is based on specific restrictions for the unknown parameters. In the simplest case the restriction refers to specific values for unknown parameters. For example, if θ = (α, β)‘, a possible restriction is α = 1 so that the restricted ML estimator is $\hat{\theta}_r = (1, \hat{\beta}_r)$‘, whereas the unrestricted ML estimator is $\hat{\theta}_u = (\hat{\alpha}_u, \hat{\beta}_u)$‘.

Wald test procedure:

This test is based on the unrestricted ML estimator $\hat{\theta}$. Under the null hypothesis H_0: c(θ) = 0 it follows that $c(\hat{\theta})$ converges stochastically to zero since it is a consistent estimator of c(θ). Therefore, the alternative hypothesis implies that $c(\hat{\theta})$ strongly differs from the null vector.

Wald test statistic:

$$\mathrm{WT} = n c(\hat{\theta})' \left[\frac{\partial c(\hat{\theta})}{\partial \theta'} \hat{I}(\hat{\theta})^{-1} \frac{\partial c(\hat{\theta})'}{\partial \theta} \right] c(\hat{\theta})$$

$\hat{I}(\hat{\theta})$ is a consistent estimator of the information matrix and can e.g. be estimated on the basis of the three versions as discussed above.

If H_0: $c(\theta) = 0$ is true, it follows that the Wald test statistic is asymptotically χ^2 distributed with q degrees of freedom, i.e.:

$$WT \xrightarrow{d} \chi^2_q$$

Thus, the null hypothesis is (for a large sample size n) rejected in favor of the alternative hypothesis at the significance level α if:

$$WT > \chi^2_{q;1-\alpha}$$

In the case of the specific simple null hypothesis H_0: $\theta_l = a$ it follows that the corresponding Wald test statistic is asymptotically χ^2 distributed with one degree of freedom. Therefore, it follows the simplest and most important version of a Wald test statistic, namely the z statistic (or t statistic), which is asymptotically standard normally distributed:

$$WT = z = \frac{\hat{\theta}_l - a}{\sqrt{\hat{Var}(\hat{\theta}_l)}} \xrightarrow{d} N(0; 1)$$

The estimated variance of $\hat{\theta}_l$ is the l-th diagonal element of the estimated variance covariance matrix of $\hat{\theta}$ (e.g. on the basis of the three versions as discussed above). The null hypothesis is thus (for a large sample size n) rejected at the significance level α if:

$$|z| > z_{1-\alpha/2}$$

Likelihood ratio test procedure:

This test is based on both the unrestricted ML estimator $\hat{\theta}_u$ and the restricted ML estimator $\hat{\theta}_r$. In the following, the value of the log-likelihood function at the restricted ML estimator is denoted by $\log L(\hat{\theta}_r)$ and the value of the log-likelihood function at the unrestricted ML estimator is denoted by $\log L(\hat{\theta}_u)$, where $\log L(\hat{\theta}_r) \leq \log L(\hat{\theta}_u)$. The null hypothesis H_0: $c(\theta) = 0$ implies that these values are very similar, whereas the alternative hypothesis implies that the values of the restricted and unrestricted log-likelihood functions are strongly different.

Likelihood ratio test statistic:

$$\mathrm{LRT} = 2\left[\log L(\hat{\theta}_u) - \log L(\hat{\theta}_r)\right]$$

If H_0: $c(\theta) = 0$ is true, it follows that the likelihood ratio test statistic is asymptotically χ^2 distributed with q degrees of freedom, i.e.:

$$\mathrm{LRT} \overset{d}{\to} \chi^2_q$$

Thus, the null hypothesis is (for a large sample size n) rejected in favor of the alternative hypothesis at the significance level α if:

$$\mathrm{LRT} > \chi^2_{q;1-\alpha}$$

The main advantage of this test is that it is easy to perform. The practical disadvantage is that two models have to be estimated separately.

Score test procedure:

This test is only based on the restricted ML estimator $\hat{\theta}_r$. It is further based on the property that the expectation of the score at the true, but unknown parameter vector θ is zero and that the score (i.e. the first derivative of the log-likelihood function) at the unrestricted ML estimator $\hat{\theta}_u$ is zero (necessary condition for the ML estimator). The score test therefore considers the difference between the score at the restricted ML estimator $\hat{\theta}_r$ and the null vector. The null hypothesis H_0: c(θ) = 0 implies that these values are very similar, whereas the alternative hypothesis implies that the difference is large.

Score test statistic:

$$\mathrm{ST} = \frac{1}{n}\frac{\partial \log L(\hat{\theta}_r)}{\partial \theta'}\hat{I}(\hat{\theta}_r)^{-1}\frac{\partial \log L(\hat{\theta}_r)}{\partial \theta}$$

Under H_0: c(θ) = 0, $\hat{I}(\hat{\theta}_r)$ is a consistent estimator of the information matrix and can e.g. be estimated on the basis of the three versions as discussed above. If H_0 is true, it follows that the score test statistic is asymptotically χ^2 distributed with q degrees of freedom, i.e.:

$$\mathrm{ST} \xrightarrow{d} \chi^2_q$$

Thus, the null hypothesis is (for a large sample size n) rejected in favor of the alternative hypothesis at the significance level α if:

$$\mathrm{ST} > \chi^2_{q;1-\alpha}$$

Example: Determinants of (the logarithm of) wages (I)

Based on the discussed ML estimation of the classical linear regression model, which considers the effect of the years of education (educ), the years of labor market experience (exper), and the years with the current employer (tenure) on the logarithm of hourly wage (logwage), the null hypothesis that neither educ nor exper has any effect on logwage, i.e. that the two parameters of educ and exper are both zero, is tested. The command for the Wald test in Stata is:

```
test educ=exper=0

 ( 1)  [Betas]educ - [Betas]exper = 0
 ( 2)  [Betas]educ = 0

           chi2(  2) =  161.52
         Prob > chi2 =    0.0000
```

The application of the likelihood ratio test requires both the unrestricted and restricted ML estimation. After the unrestricted ML estimation the following Stata command for saving the estimation results is necessary (the choice of the name “unrestricted” is arbitrary):

```
estimates store unrestricted
```

After the restricted ML estimation a corresponding Stata command for saving the respective estimation results is necessary (the choice of the name “restricted” is again arbitrary).

Example: Determinants of (the logarithm of) wages (II)

```
ml model lf linearregression (Betas: logwage = tenure) (Variance:)
ml maximize

                                                  Number of obs   =        526
                                                  Wald chi2(1)    =      62.35
Log likelihood = -383.97797                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Betas        |
      tenure |   .0239514   .0030333     7.90   0.000     .0180063    .0298965
       _cons |   1.501007   .0268149    55.98   0.000     1.448451    1.553563
-------------+----------------------------------------------------------------
Variance     |
       _cons |   .2521113   .0155458    16.22   0.000      .221642    .2825806
------------------------------------------------------------------------------

estimates store restricted
```

Command for the likelihood ratio test in Stata:

```
lrtest unrestricted restricted

Likelihood-ratio test                                 LR chi2(2)  =    140.86
(Assumption: restricted nested in unrestricted)       Prob > chi2 =    0.0000
```

Both the Wald test and the likelihood ratio test therefore lead to the rejection of the null hypothesis at very low significance levels.

Model selection:

The underlying hypotheses in the three tests imply two (microeconometric) models that are nested, i.e. one model is a restricted version of the other model. If two models are non-nested, a selection between them can be based on the higher value of the log-likelihood function at the ML estimators. However, these values depend on the number of parameters in the models so that this number should be considered in a model selection. Two common approaches in this respect are the Akaike information and Schwarz information criteria.

Goodness-of-fit:

In linear regression models, the coefficient of determination R^2 is an appropriate goodness-of-fit measure. However, in many microeconometric models this measure is not available. Based on the value $\log L(\hat{\theta}_u)$ of the log-likelihood function at the ML estimator of the full microeconometric model with all explanatory variables (including a constant) and the value $\log L(\hat{\theta}_r)$ at the restricted ML estimator of the corresponding model that only includes the constant, the following pseudo R^2 with $\log L(\hat{\theta}_r) \leq \log L(\hat{\theta}_u)$ can be calculated:

$$R^2_{pseudo} = 1 - \frac{\log L(\hat{\theta}_u)}{\log L(\hat{\theta}_r)}$$

In microeconometric models with discrete dependent variables R^2_{pseudo} varies between zero and one and is often interpreted as goodness-of-fit measure. However, the analysis of tests is generally more important than this measure.