Martin Luther University of Halle-Wittenberg
Department of Economics
Chair of Econometrics

Econometrics II
1. Maximum likelihood estimation

Christoph Wunder

Goals

- Introduce method of maximum likelihood (ML)
- ML algorithms
- Describe likelihood-based tests
- Model fit and model comparison
- Apply methods in R

Outline

Outline

1.1 Intuition

- Assume distribution of the data. (Note that the maximum likelihood method requires an assumption about the (conditional) distribution of the outcome variable.)
- Parameters of the distribution are unknown.
- Determine the likelihood of observing the data.
- Choose those values for the unknown parameters that give the observed data the highest likelihood (maximum likelihood estimates, MLE).

Example: tosses of a globe

McElreath (2016), Ch. 2.2

- How much of the surface of planet earth is covered in water?
- Strategy: toss a globe. After catching, record whether or not the surface under right index finger is water or land. Repeat.
- We define:

$$y_i = \begin{cases} 1, & \text{if the } i\text{th toss produces "water";} \\ 0, & \text{if the } i\text{th toss produces "land".} \end{cases} \tag{1.1}$$

- Suppose the sample contains $w = \sum_i y_i$ tosses with "water" and $n - w$ tosses with "land".

- Likelihood function: for one particular sample (in a given order), the probability of w water observations in n tosses, with probability π on each toss and n tosses total, is:

$$L(\pi) = \pi^w (1-\pi)^{n-w}. \tag{1.2}$$

- ML method: we choose the value of π such that the likelihood is maximal.
- Often more convenient to maximize the log-likelihood function:

$$\ln L(\pi) = w \ln \pi + (n-w) \ln(1-\pi). \tag{1.3}$$

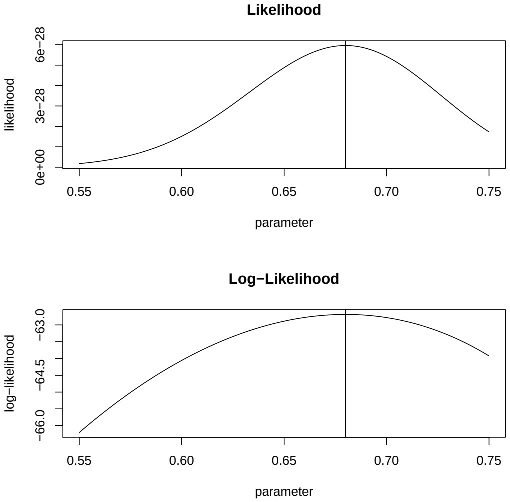
- Maximizing the log-likelihood function gives the first order condition:

$$\frac{d \ln L(\pi)}{d\pi} = \frac{w}{\pi} - \frac{n-w}{1-\pi} = 0. \tag{1.4}$$

- Solving for π gives the ML estimator

$$\hat{\pi} = w/n. \tag{1.5}$$

Numerical example: $n = 100$, $w = 68$.



Outline

1.2 Concepts and properties of ML estimation

- Assuming i.i.d. observations, the joint probability density function of the random sample $y_1, y_2, ..., y_n$ is the product of individual densities:

$$f(y_1, y_2, ..., y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta). \tag{1.6}$$

- The likelihood function, $L(\theta; \mathbf{y})$, is the probability density function of the observed data, viewed as a function of the unknown parameters

$$L(\theta; \mathbf{y}) = \prod_{i=1}^{n} L_i(\theta; y_i) = \prod_{i=1}^{n} f(y_i; \theta), \tag{1.7}$$

  where $L_i(\theta; y_i)$ denotes the individual likelihood contributions.
- Note that the joint density in 1.6 is a function of the data conditioned on the parameters while the likelihood function in 1.7 is a function of parameters conditioned on the data. The likelihood function is often denoted simply $L(\theta)$.

- It is numerically convenient to work with the log-likelihood function:

$$\ln L(\theta) = \sum_{i=1}^{n} \ln L_i(\theta) = \sum_{i=1}^{n} \ln f(y_i; \theta), \tag{1.8}$$

- The ML estimator, $\hat{\theta}$, is the solution to

$$\max_{\theta} \ln L(\theta). \tag{1.9}$$

  Since the logarithm is a monotonic transformation, any values $\hat{\theta}$ that maximize $\ln L(\theta)$ also maximize $L(\theta)$.
- The score vector is the vector of first derivatives of the log-likelihood function:

$$\mathbf{s}(\theta) \equiv \frac{\partial \ln L(\theta)}{\partial \theta}. \tag{1.10}$$

- The FOCs imply $\mathbf{s}(\hat{\theta}) = \mathbf{0}$.

- The second derivative of the log-likelihood function is referred to as the (symmetric) Hessian matrix:

$$\mathbf{H}(\theta) = \frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'} = \begin{pmatrix} \frac{\partial^2 \ln L(\theta)}{(\partial\theta_1)^2} & \frac{\partial^2 \ln L(\theta)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 \ln L(\theta)}{\partial\theta_1\partial\theta_K} \\ \frac{\partial^2 \ln L(\theta)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 \ln L(\theta)}{(\partial\theta_2)^2} & \cdots & \frac{\partial^2 \ln L(\theta)}{\partial\theta_2\partial\theta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln L(\theta)}{\partial\theta_K\partial\theta_1} & \frac{\partial^2 \ln L(\theta)}{\partial\theta_K\partial\theta_2} & \cdots & \frac{\partial^2 \ln L(\theta)}{(\partial\theta_K)^2} \end{pmatrix} \quad (1.11)$$

- The score, $\mathbf{s}(\theta)$, and the Hessian, $\mathbf{H}(\theta)$, are random variables because they depend on the sample. The usually differ in repeated samples.
- The information matrix (Fisher information) is defined as the negative expectation of the Hessian matrix:

$$\mathbf{I}(\theta) = -E[\mathbf{H}(\theta)]. \quad (1.12)$$

- The inverse information matrix is the variance of the ML estimator:

$$Var(\hat{\theta}) = \mathbf{I}(\theta)^{-1}. \quad (1.13)$$

- Since $\mathbf{I}(\theta)^{-1}$ depends on the unknown true parameter vector θ, we use a consistent estimator of the variance matrix. Three possibilities:
  1. Using the expected Hessian:

$$\widehat{Var}(\hat{\theta})_1 = \{-E[\mathbf{H}(\hat{\theta})]\}^{-1} \tag{1.14}$$

  2. Using the observed Hessian (when the expected Hessian cannot be obtained, standard procedure in software):

$$\widehat{Var}(\hat{\theta})_2 = [-\mathbf{H}(\hat{\theta})]^{-1} \tag{1.15}$$

  3. Using the variance of the score (outer product of the gradient, Berndt-Hall-Hall-Hausman (BHHH) estimator, easiest to compute):

$$\widehat{Var}(\hat{\theta})_3 = \left[\sum_{i=1}^{n} \mathbf{s}_i(\hat{\theta})\mathbf{s}_i(\hat{\theta})'\right]^{-1}, \tag{1.16}$$

where $\mathbf{s}_i = \frac{\partial \ln L_i(\hat{\theta})}{\partial \hat{\theta}}$ are the individual score contributions.

The ML estimator is

1. consistent,
2. asymptotically efficient (reaches the Cramér Rao lower bound),
3. asymptotically normally distributed:

$$\hat{\theta} \overset{a}{\sim} N(\theta, -E[\mathbf{H}(\theta)]^{-1}) \tag{1.17}$$

4. invariant to one-to-one transformations of θ.
    - If $\hat{\theta}$ is the ML estimator of θ, then $h(\hat{\theta})$ is the ML estimator of $h(\theta)$.
    - Example: consider the linear model $y = \beta_0 + \beta_1 x + \epsilon$. The ML estimator of the ratio β_0/β_1 is equal to the ratio of the ML estimators $\hat{\beta}_0/\hat{\beta}_1$.

Outline

1.3 ML estimation for linear regression

- Assumptions:
  1. Data $(y_i, \mathbf{x}_i)$ on $i = 1, \ldots, n$ individuals.
  2. Model: $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$
  3. Shape of the distribution: $y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$
  4. Observations are independently distributed.
- Likelihood function:

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) \tag{1.18}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2} \cdot \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right) \tag{1.19}$$

- Log-likelihood function:

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) + \sum_{i=1}^{n}\left(-\frac{1}{2}\cdot\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right) \tag{1.20}$$

- Likelihood equations (score vector):

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}\left(-2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}\right) = \mathbf{0} \tag{1.21}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \tag{1.22}$$

$$\tag{1.23}$$

- The ML estimators for the linear model are

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.24}$$

$$\hat{\sigma}^2_{ML} = n^{-1}\mathbf{e}'\mathbf{e}, \tag{1.25}$$

where $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}$.

Outline

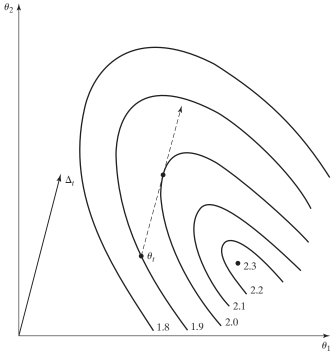
1.4 Algorithms

- Techniques:
  - Grid search
  - Analytical optimization (requires closed-form solution for FOC)
  - Numerical optimization
- General idea of an iterative algorithm:
  1. Begin with initial values for θ.
  2. Compute likelihood.
  3. Update values for θ.
  4. Compute likelihood: stop if optimum, go to step 3 if not.
- Convergence criterion: when are values optimal?
  - minimal change in log-likelihood
  - minimal change in $\hat{\theta}$
  - minimal change in $\mathbf{s}(\hat{\theta})$

Update values: if the value in iteration step t, θ_t, is not the optimal value for θ, then compute

$$\theta_{t+1} = \theta_t + \lambda_t \boldsymbol{\Delta}_t, \tag{1.26}$$

where λ_t is the step size and $\boldsymbol{\Delta}_t$ is the direction vector.



Source: Greene (2011), Figure E.2

Newton-Raphson algorithm

- Linear Taylor series approximation of the FOCs:

$$\frac{\partial \ln L(\theta_t)}{\partial \theta_t} \approx \mathbf{s}(\theta_{t-1}) + \mathbf{H}(\theta_{t-1})(\theta_t - \theta_{t-1}) \stackrel{!}{=} \mathbf{0}. \tag{1.27}$$

- Solving FOCs for θ_t,

$$\theta_t = \theta_{t-1} \underbrace{-[\mathbf{H}(\theta_{t-1})]^{-1}\mathbf{s}(\theta_{t-1})}_{\boldsymbol{\Delta}_{t-1}}, \tag{1.28}$$

and step size $\lambda_{t-1} = 1$.

- Statistical software packages offer alternative algorithms.
- Some algorithms may perform better than others for a given problem.
- Examples (see Gould et al. 2003, Ch. 1.3):
  - BHHH ("B-H-cubed") algorithm: Hessian is replaced by the outer product of the score
  - Davidon-Fletcher-Powell (DFP) algorithm
  - Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm
  - Steepest ascent method
  - Quadratic hill-climbing method

Outline

1.5.1 Likelihood ratio test

- ML estimation of the unrestricted model gives $\ln L_U$, ML estimation of the restricted model gives $\ln L_R$.
- Idea: if the restriction is valid, then the difference in log-likelihood functions, $\ln L_U - \ln L_R$, should be small.
- The likelihood ratio test statistic has, under H_0, a limiting χ^2 distribution with degrees of freedom equal to the number of restrictions.

$$LR = -2 \ln \left(\frac{L_R}{L_U} \right) = -2 \left(\ln L_R - \ln L_U \right) \tag{1.29}$$

1.5.2 Wald test

- Estimate unrestricted model using ML gives $\hat{\theta}$.
- We test the null hypothesis

$$\mathbf{c}(\theta) = \mathbf{q}. \tag{1.30}$$

- Idea: If the restriction is valid, then $\mathbf{c}(\hat{\theta}) - \mathbf{q}$ should be close to zero. Reject the hypothesis if this value is significantly different from zero.
- The Wald test statistic has, under H_0, a limiting χ^2 distribution with degrees of freedom equal to the number of restrictions.

$$W = \left[\mathbf{c}(\hat{\theta}) - \mathbf{q}\right]' \left(\text{Asy.Var}\left[\mathbf{c}(\hat{\theta}) - \mathbf{q}\right]\right)^{-1} \left[\mathbf{c}(\hat{\theta}) - \mathbf{q}\right] \tag{1.31}$$

- The variance of the restriction, $\text{Asy.Var}\left[\mathbf{c}(\hat{\theta}) - \mathbf{q}\right]$, is estimated as

$$\left(\frac{\partial \mathbf{c}(\hat{\theta})}{\partial \theta}\right)' Var(\hat{\theta}) \left(\frac{\partial \mathbf{c}(\hat{\theta})}{\partial \theta}\right). \tag{1.32}$$

- Advantage: only estimation of the unrestricted model is required.
- Disadvantage: the Wald statistic is not invariant to the formulation of the restrictions.
- Example:
    - Model: $E[y|\mathbf{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
    - Hypothesis:

$$H_0 : \beta_1 = \beta_2 \cdot \beta_3 \tag{1.33}$$

can be tested as

$$H_0 : \beta_1 - \beta_2 \cdot \beta_3 = 0 \tag{1.34}$$

or

$$H_0 : \frac{\beta_1}{\beta_3} - \beta_2 = 0 \tag{1.35}$$

1.5.3 Lagrange multiplier (LM) test (score test)

- LM test is based on the restricted model.
- Idea: if the restriction is valid, then the slope of the log-likelihood of the unrestricted model should be near zero at the restricted estimator (i.e. the derivatives of the unrestricted log-likelihood evaluated at the restricted parameter vector will be approximately zero).
- The LM statistic has, under H_0, a limiting χ^2 distribution with degrees of freedom equal to the number of restrictions.

$$LM = \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right)' \left[\mathbf{I}(\hat{\theta}_R) \right]^{-1} \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right) \tag{1.36}$$

Outline

1.6 Model fit

- There is no obvious measure of explained variation (such as R^2 for the linear model).
- Pseudo R^2 (likelihood ratio index, McFadden R^2):

$$\text{Pseudo } R^2 = 1 - \frac{\ln L}{\ln L_0}, \tag{1.37}$$

  where $\ln L$ is the log-likelihood for the full model, $\ln L_0$ is the log-likelihood for the model with only a constant term.
  - If $\ln L \gg \ln L_0$, then Pseudo R^2 approaches 1.
  - If $\ln L = \ln L_0$, then Pseudo $R^2 = 0$.

Outline

1.7 Comparing models

- When the models are nested, likelihood-based tests can be used.
- When the models are nonnested, information criteria are used.
- Akaike information criterion (AIC):

$$AIC = -2 \ln L + 2K \tag{1.38}$$

- Bayes information criterion (BIC):

$$BIC = -2 \ln L + K \ln n, \tag{1.39}$$

  where K is the number of parameters in the model.
- The model with the lowest AIC or BIC is usually preferred.

Outline

1.8 Example: tosses of a globe

Example in R

- Simulate some fake data:

```
set.seed(647382)
y <- rbinom( n = 100, size = 1, prob = 0.7)
head(y, 20)
## [1] 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 0 0 1 1 1
```

- Define log-likelihood function

```
log.likelihood <- function(pi, y) {
    sum ( y*log(pi) + (1-y)*log(1-pi) )
}
```

or

```
log.likelihood <- function(pi, y) {
    sum ( dbinom( x = y, size = 1, prob = pi, log = TRUE ) )
}
```

- Maximize log-likelihood

```
result <- optim(
  par = 0.5,             # initial value
  fn = log.likelihood, # function to be maximized
  y = y,                 # supply data
  method = "L-BFGS-B", lower = 1e-10, upper = (1 - 1e-10),
  control = list(fnscale = -1), # maximization problem
  hessian = TRUE)
```

- ML estimate

```
round( result$par, 4 )
## [1] 0.68
```

- Compute variance and s.e. using the inverse of the negative actual Hessian:

```
round( c( (-result$hessian)^(-1),
          sqrt( (-result$hessian)^(-1) ) ), 4 )
## [1] 0.0022 0.0466
```

- Value of log-likelihood function

```
result$value
## [1] -62.68695
```

- The Newton-Raphson algorithm is available in the R package `maxLik`:

```
library(maxLik)
mle.nr <- maxLik( logLik = log.likelihood, start = c(pi=0.5), y=y )
summary(mle.nr)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 2 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: -62.68695
## 1  free parameters
## Estimates:
##    Estimate Std. error t value Pr(> t)
## pi  0.68000    0.04665   14.58  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

References

Reading list: Verbeek (2012), Chapter 6, Winkelmann and Boes (2006), Chapter 3, Greene (2011), Chapter 14.

Gould, W., Pitblado, J., and Sribney, W. (2003). *Maximum Likelihood Estimation with Stata*. Stata Press, College Station.

Greene, W. H. (2011). *Econometric Analysis*. Prentice Hall International, Upper Saddle River, New Jersey.

McElreath, R. (2016). *Statistical Rethinking. A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC Press, Boca Raton.

Verbeek, M. (2012). *A Guide to Modern Econometrics*. John Wiley & Sons, Chichester.

Winkelmann, R. and Boes, S. (2006). *Analysis of Microdata*. Springer, Berlin, Heidelberg.