Econometrics - Advanced Methods

(Research Metrics 2)

Maximum Likelihood

Prof. Dr. Derya Uysal
Department of Economics
University of Munich
Email: derya.uysal@econ.lmu.de

Winter 2024/25

Overview

What are we going to cover?

1. Maximum Likelihood Estimation
   - Parametric Model
   - Likelihood Function
   - Score, Hessian and Information
   - Asymptotic Properties of the MLE
2. Kullback-Leibler Divergence and Pseudo Maximum Likelihood
3. Hypothesis and Specification Tests

Literature:

- Wooldridge (2010), Ch. 13.1-13.7, 13.11
- Cameron & Trivedi (2005), Ch. 5.6, 5.7
- Hansen (2022), Ch. 10

Reading List

James P. Smith (2009): The impact of childhood health on adult labor market outcome. *Review of Economics and Statistics*, 91(3), 478–489

Rainer Winkelmann (2004): Health care reform and the number of doctor visits: An econometric analysis. *Journal of Applied Econometrics*, 19, 455–472.

Maximum Likelihood Estimation

Parametric Model

- First, we consider unconditional distributions, i.e. the probability function does not depend on conditioning variables.
- A parametric model for a random variable Y is a complete probability function depending on an unknown parameter vector $\boldsymbol{\theta}$.
- In the discrete case, we can write a parametric model as a probability mass function and in the continuous case we can write it as a density function: $f(y|\boldsymbol{\theta})$ (we will not differentiate the notation)
- The parameter $\boldsymbol{\theta}$ belongs to a set Θ called the parameter space.
- Example: One parametric model is $Y \sim N(\mu, \sigma^2)$, which has a density $f(y|\mu, \sigma^2) = \sigma^{-1}\phi((y-\mu)/\sigma)$ with $\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$ (pdf of a standard normal). The parameters are $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.
- A parametric model does not need to be one of textbook functional forms, it can be developed by a user of a specific application.

Parametric Model

- A parametric model specifies the distribution for all observations.
- We will (mostly) focus on random samples, i.e. the observations are i.i.d.

Definition 1.1 (Model)

A model for a random sample is the assumption that Y_i, $i = 1, \ldots, n$ are i.i.d. with known density function (or mass function) with unknown parameter $\boldsymbol{\theta} \in \Theta$.

- A model is correctly specified when there exists a parameter such that the model corresponds to the true data distribution.

Definition 1.2 (Correct Specification)

A model is correctly specified when there exists a unique parameter value $\boldsymbol{\theta}_0 \in \Theta$ such that $f(y|\boldsymbol{\theta}_0) = f(y)$, the true data distribution. The parameter value $\boldsymbol{\theta}_0$ is called the true parameter value. The parameter $\boldsymbol{\theta}_0$ is unique if there is no other $\boldsymbol{\theta}$ such that $f(y|\boldsymbol{\theta}_0) = f(y|\boldsymbol{\theta})$. A model is misspecified if there is no parameter value in $\boldsymbol{\theta} \in \Theta$ such that $f(y|\boldsymbol{\theta}) = f(y)$.

Example 1.1

- Assume the true density is $f(y) = 2\exp(-2y)$.

Example 1.1

- Assume the true density is $f(y) = 2\exp(-2y)$.
- The exponential model $f(y|\lambda) = \lambda^{-1}\exp(-y/\lambda)$ is a correctly specified model with $\lambda_0 = 1/2$.

Example 1.1

- Assume the true density is $f(y) = 2\exp(-2y)$.
- The exponential model $f(y|\lambda) = \lambda^{-1}\exp(-y/\lambda)$ is a correctly specified model with $\lambda_0 = 1/2$.
- The gamma model $f(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp(-y\beta)$ is also correctly specified with $\beta_0 = 2$ and $\alpha_0 = 1$.
- The log-normal model $f(y|\theta,\nu) = \frac{1}{\sqrt{2\pi\nu}} y^{-1}\exp\left(-\frac{(\log y-\theta)^2}{2\nu}\right)$, is misspecified, as there are no parameters such that the log-normal density is equal to $f(y) = 2\exp(-2y)$.

Example 1.2

Assume the true density is $f(y) = \phi(y)$ ((standard normal). Correctly specified models include the normal and the Student's t, but not, for example, the logistic.

Example 1.2

Assume the true density is $f(y) = \phi(y)$ ((standard normal). Correctly specified models include the normal and the Student's t, but not, for example, the logistic.

Because:

- Normal and Student's t models include $\phi(y)$ (exactly or as a limit).
- The logistic family cannot reproduce $\phi(y) \Rightarrow$ misspecified.

Example 1.3

- Consider the mixture of normals model
  $f(y, p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = p\phi_{\sigma_1}(y - \mu_1) + (1 - p)\phi_{\sigma_2}(y - \mu_2)$.

Example 1.3

- Consider the mixture of normals model $f(y, p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = p\phi_{\sigma_1}(y - \mu_1) + (1 - p)\phi_{\sigma_2}(y - \mu_2)$.
- This includes $\phi(y)$ as a special case, so it is a correct model, but "true" parameter is not unique, because $(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

$\rightarrow$ $= (p, 0, 1, 0, 1)$ for any p,

$\rightarrow$ $= (1, 0, 1, \mu_2, \sigma_2^2)$ for any μ_2 and σ_2^2,

$\rightarrow$ $= (0, \mu_1, \sigma_1^2, 0, 1)$ for any μ_1 and σ_1^2, we have $\phi(y)$.

Example 1.3

- Consider the mixture of normals model
  $f(y, p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = p\phi_{\sigma_1}(y - \mu_1) + (1 - p)\phi_{\sigma_2}(y - \mu_2)$.
- This includes $\phi(y)$ as a special case, so it is a correct model, but "true" parameter is not unique, because $(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

$\rightarrow$ $= (p, 0, 1, 0, 1)$ for any p,

$\rightarrow$ $= (1, 0, 1, \mu_2, \sigma_2^2)$ for any μ_2 and σ_2^2,

$\rightarrow$ $= (0, \mu_1, \sigma_1^2, 0, 1)$ for any μ_1 and σ_1^2, we have $\phi(y)$.

- Thus: although the model is correct, it does not meet the definition of correctly specified.

Likelihood Function

- The likelihood is the joint density of the observations calculated using the model, i.e.

$$f(y_1, \ldots, y_n | \boldsymbol{\theta})$$

Likelihood Function

- The likelihood is the joint density of the observations calculated using the model, i.e.

$$f(y_1, \ldots, y_n|\boldsymbol{\theta})$$

- Independence of observations means that the joint density is the product of the individuals densities

$$f(y_1, \ldots, y_n|\boldsymbol{\theta}) = f(y_1|\boldsymbol{\theta})f(y_2|\boldsymbol{\theta}) \cdots f(y_n|\boldsymbol{\theta})$$

Likelihood Function

- The likelihood is the joint density of the observations calculated using the model, i.e.

$$f(y_1, \ldots, y_n|\boldsymbol{\theta})$$

- Independence of observations means that the joint density is the product of the individuals densities

$$f(y_1, \ldots, y_n|\boldsymbol{\theta}) = f(y_1|\boldsymbol{\theta})f(y_2|\boldsymbol{\theta}) \cdots f(y_n|\boldsymbol{\theta})$$

- Identical distributions means that all the densities are identical, so the joint density equals:

$$f(y_1, \ldots, y_n|\boldsymbol{\theta}) = f(y_1|\boldsymbol{\theta})f(y_2|\boldsymbol{\theta}) \cdots f(y_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

- The joint density evaluated at the observed data and viewed as a function of $\boldsymbol{\theta}$ is called the likelihood function.

Likelihood Function

Definition 1.3 (Sample Likelihood Function)

Let $L_n(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|y_1, \ldots, y_n) \equiv f(y_1, \ldots, y_n|\boldsymbol{\theta})$ be the joint density of a sample of n random variables characterized by a k-dimensional vector of parameters $\boldsymbol{\theta}$. Then $L(\boldsymbol{\theta})$ is called the sample likelihood function

Definition 1.4 (Sample Likelihood Function of i.i.d. Random Variables)

If $(y_1, \ldots, y_n)$ are iid random variables with density function $f(y_i|\boldsymbol{\theta})$, the sample likelihood function takes the following form:

$$L_n(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}),$$

with the sample log-likelihood function

$$\begin{aligned} \ln L_n(\boldsymbol{\theta}) &= \ln L(\boldsymbol{\theta}|y_1, \ldots, y_n) \\ &= \ln \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}). \end{aligned}$$

Example 1.4

Assume that an iid random sample of size n has been drawn from a Bernoulli distribution with parameter π.

- π is the probability of success and the counter-probability is $1 - \pi$.
- the probability mass function for each y_i can be written

$$f(y_i|\pi) = \pi^{y_i}(1-\pi)^{1-y_i} \quad y_i = 0, 1$$

- Then the likelihood function and the log-likelihood function have the form

$$L(\pi) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$$

$$\ln L(\pi) = \sum_{i=1}^{n} y_i \ln \pi + (1-y_i)\ln(1-\pi)$$

Example 1.4 (cont.'d)

- Figure 1 plots the likelihood function for two different samples of size $n = 5$.
- The first sample $(0, 0, 0, 1, 1)$ has the likelihood function ...
- The second sample $(0, 0, 1, 1, 1)$ has the likelihood function ...

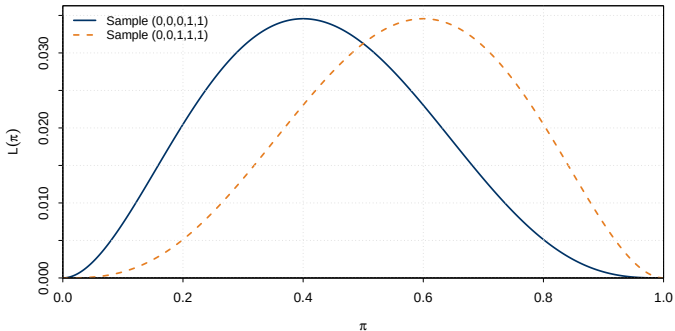


Figure: Likelihood Function for the Bernoulli Example

Maximum Likelihood Estimator

Definition 1.5 (Maximum Likelihood Estimator)

Let $L_n(\boldsymbol{\theta}|\boldsymbol{y})$ be the sample likelihood function. Then the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is the value of the parameter vector $\boldsymbol{\theta}$ that maximizes $L_n(\boldsymbol{\theta}|\boldsymbol{y})$, that is

$$\widehat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta}|\boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

Any positive monotone transformation of $L(\boldsymbol{\theta})$ is also maximized by $\widehat{\boldsymbol{\theta}}_n$. Hence

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta}|\boldsymbol{y}) &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n} \ln L_n(\boldsymbol{\theta}|\boldsymbol{y}) \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}). \end{aligned}$$

Maximum Likelihood Estimator

Definition 1.5 (Maximum Likelihood Estimator)

Let $L_n(\boldsymbol{\theta}|\boldsymbol{y})$ be the sample likelihood function. Then the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is the value of the parameter vector $\boldsymbol{\theta}$ that maximizes $L_n(\boldsymbol{\theta}|\boldsymbol{y})$, that is

$$\widehat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta}|\boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

Any positive monotone transformation of $L(\boldsymbol{\theta})$ is also maximized by $\widehat{\boldsymbol{\theta}}_n$. Hence

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta}|\boldsymbol{y}) &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n} \ln L_n(\boldsymbol{\theta}|\boldsymbol{y}) \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}). \end{aligned}$$

- Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely.
- In general, the MLE is a good estimator some of optimality properties discussed later.

Maximum Likelihood Estimator

There are two inherent drawbacks associated with the general problem of finding the maximum of a function:

- Finding the global maximum and verifying that, indeed, a global maximum is found.
- Numerical sensitivity: how sensitive is the estimate to small changes in data? (rather a mathematical problem)

Maximum Likelihood Estimator

If the likelihood function is differentiable in the parameter, possible candidates for the MLE are the values that solve the first-order condition, i.e.

$$\frac{d}{d\boldsymbol{\theta}} L_n(\boldsymbol{\theta}) = 0$$

Note that:

- FOC being equal to 0 is only necessary for a maximum but not sufficient.
- The zeros of the first derivative locate only extreme points in the interior of the domain of a function. If the extrema occur on the boundary the first derivative might not be 0.

Maximum Likelihood Estimator

Another way is to abandon differentiation and proceed with a direct maximization.

- This is usually simpler algebraically, especially if the derivatives tend to get messy, but it is sometimes harder to implement because there are no set rules to follow.
- One general technique is find a global upper bound on the likelihood and then establish there is a unique point for which the upper bound is attained.

Example 1.5 (Maximum Likelihood Estimator)

- Density:
$$f(y|\lambda) = \lambda^{-1} \exp(-y/\lambda)$$
- The likelihood function is:
$$L_n(\lambda) = \prod_{i=1}^{n} \lambda^{-1} \exp\left(-\frac{y_i}{\lambda}\right) = \frac{1}{\lambda^n} \exp\left(-\frac{n\bar{Y}_n}{\lambda}\right)$$
- The first-order condition for the maximization is
$$0 = \frac{d}{d\lambda} L_n(\lambda) = -n\frac{1}{\lambda^{n+1}} \exp\left(-\frac{n\bar{Y}_n}{\lambda}\right) + \frac{1}{\lambda^n} \exp\left(-\frac{n\bar{Y}_n}{\lambda}\right) \frac{n\bar{Y}_n}{\lambda^2}$$
- Canceling the common terms and solving, we find the unique solution which is maximum likelihood estimator (MLE) for λ:
$$\hat{\lambda}_n = \bar{Y}_n$$

Example 1.5 (cont.'d)

- The log likelihood function is:

$$\ln L_n(\lambda) = \sum_{i=1}^{n} \left[-\ln \lambda - \frac{Y_i}{\lambda} \right] = -n \ln \lambda - \frac{n\bar{Y}_n}{\lambda}$$

- The first-order condition with respect tot he log likelihood

$$0 = \frac{d}{d\lambda} \ln L_n(\lambda) = -\frac{n}{\lambda} + \frac{n\bar{Y}_n}{\lambda^2}$$

- The unique solution is $\hat{\lambda}_n = \bar{Y}_n$.
- The second order condition is:

$$\frac{d^2}{d\lambda^2} \ln L_n(\hat{\lambda}_n) = \frac{n}{\hat{\lambda}^2} - 2\frac{n\bar{Y}_n}{\hat{\lambda}^3} = -\frac{n}{\bar{Y}_n^2} < 0$$

This condition verifies that $\hat{\lambda}_n$ is a maximizer rather than a minimizer.

Invariance Property

A special property of the MLE (not shared by all estimators) is that it is invariant to transformations.

Theorem: Invariance Property

If $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta} \in \mathbb{R}^m$, then for any transformation $\boldsymbol{\beta} = h(\boldsymbol{\theta}) \in \mathbb{R}^l$, the MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_n = h(\hat{\boldsymbol{\theta}}_n)$.

Invariance Property

Example 1.5 (cont.'d)

$$f(y|\lambda) = \lambda^{-1} \exp(-y/\lambda)$$

- We know $\hat{\lambda}_n = \bar{Y}_n$.
- Set $\beta = 1/\lambda$, so $h(\lambda) = 1/\lambda$.
- The log density of the reparametrized model is $\ln f(y|\beta) = \ln \beta - y\beta$.
- The log likelihood function is:

  $$\ln L_n(\beta) = n \ln \beta - \beta n \bar{Y}_n$$

  which has a maximizer $\hat{\beta}_n = 1/\bar{Y}_n = h(\bar{Y}_n)$, as the invariance property suggested.

Steps to find the MLE

1. Construct $f(y|\boldsymbol{\theta})$ as a function of y and $\boldsymbol{\theta}$.
2. Take the logarithm $\ln f(y|\boldsymbol{\theta})$
3. Evaluate at $y = y_i$ and sum over i : $\ln L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta})$
4. If possible, solve the first-order condition (FOC) o find the maximum.
5. Check the second-order condition to verify it is a maximum.
6. If solving the FOC is not possible, use other methods to maximize $\ln L_n(\boldsymbol{\theta})$

Score, Hessian and Information

Recall the log-likelihood function

$$\ln L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}).$$

Assume that $f(y_i|\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$.

Definition 1.6 (Score Function)

The first derivative of the log-likelihood function, $\frac{\partial \ln L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, is called the score function, or simply score, denoted by $\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y})$

Score, Hessian and Information

Recall the log-likelihood function

$$\ln L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}).$$

Assume that $f(y_i|\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$.

Definition 1.6 (Score Function)

The first derivative of the log-likelihood function, $\frac{\partial \ln L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, is called the score function, or simply score, denoted by $\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y})$

For k-dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, the score function, $\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y})$, is a $k \times 1$ column vector, i.e.:

$$\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_k} \end{pmatrix}$$

Score, Hessian and Information

Recall the log-likelihood function

$$\ln L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}).$$

Assume that $f(y_i|\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$.

Definition 1.6 (Score Function)

The first derivative of the log-likelihood function, $\frac{\partial \ln L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, is called the score function, or simply score, denoted by $\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y})$

For k-dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, the score function, $\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y})$, is a $k \times 1$ column vector, i.e.:

$$\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_k} \end{pmatrix}$$

The score tells us how sensitive the log-likelihood is to the parameter vector.

Score, Hessian and Information

Recall the log-likelihood function

$$\ln L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}).$$

Assume that $f(y_i|\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$.

Definition 1.6 (Score Function)

The first derivative of the log-likelihood function, $\frac{\partial \ln L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, is called the score function, or simply score, denoted by $\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y})$

For k-dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, the score function, $\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y})$, is a $k \times 1$ column vector, i.e.:

$$\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_k} \end{pmatrix}$$

The score tells us how sensitive the log-likelihood is to the parameter vector.
$S_n(\hat{\boldsymbol{\theta}}) = 0$, when $\hat{\boldsymbol{\theta}}$ is an interior solution.

Definition 1.7 (Hessian Matrix)

The second derivative of the log-likelihood function, $\frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$, is commonly referred to as the Hessian Matrix, or simply Hessian, denoted by $\mathcal{H}_n(\boldsymbol{\theta}|\boldsymbol{y})$

Hessian of a k-dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, is a $k \times k$ **symmetric** matrix, i.e.:

$$\mathcal{H}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1^2} & \frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_2 \partial \theta_k} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \ln L_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_k^2} \end{pmatrix}$$

If the log-likelihood function is concave in $\boldsymbol{\theta}$, $\mathcal{H}_n(\boldsymbol{\theta}|\boldsymbol{y})$ is said to be negative definite.
The Hessian indicates the degree of curvature in the log-likelihood function.

Note that due to the additivity of terms in the log-likelihood function, the first and second derivatives are additive functions as well:

$$\mathcal{S}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \mathcal{S}_i(\boldsymbol{\theta}|y_i) \quad \text{where } \mathcal{S}_i(\boldsymbol{\theta}|y_i) = \frac{\partial \ln f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\mathcal{H}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \mathcal{H}_i(\boldsymbol{\theta}|y_i) \quad \text{where } \mathcal{H}_i(\boldsymbol{\theta}|y_i) = \frac{\partial^2 \ln f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

Example 1.6

Assume that a random sample of size n has been drawn from a Bernoulli distribution, as before. Derive the score function, ML estimator and the Hessian.

Proposition: Properties of the Likelihood

Let $L(\boldsymbol{\theta})$ be a sample likelihood as defined in Definition 1.3. If the model is correctly specified, the support of Y does not depend on $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0$ lies in the interior of Θ,[1] then the following properties hold:

(i)

$$\mathrm{E}\left[\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = 0$$

(ii)

$$\mathrm{E}\left[\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] = -\mathrm{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \equiv \mathcal{I}(\boldsymbol{\theta}).$$

Note that the left hand side is the variance of the score (since the expectation of the score is 0). This relationship is known as the **information equality property** (ii) which states that the expectation of the outer product of the gradient (the variance of the score) is equal to the (Fisher) information matrix.

[1] regularity conditions

Example 1.7

Assume that a random sample of size n has been drawn from a Bernoulli distribution with true parameter π_0, as before.

- Show that the expected score is equal to zero at the true parameter value.
- Derive the variance of the score
- Show that information matrix equality holds

Proposition: Cramer-Rao Lower Bound

Let $L(\boldsymbol{\theta})$ be a likelihood fulfilling the certain regularity conditions. Then for the variance-covariance matrix of any unbiased estimator of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}} = T(y_1, y_2, \ldots, y_n)$, the following inequality holds:

$$\mathrm{V}\left[\widehat{\boldsymbol{\theta}}\right] \succeq \mathcal{I}(\boldsymbol{\theta})^{-1}, \quad \forall \boldsymbol{\theta} \in \Theta$$

where $\mathcal{I}(\boldsymbol{\theta}) \equiv -\mathrm{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$

Basic Concepts of Likelihood Theory

- So far, we have done the analysis in terms of the density of an observed random variable and a vector of parameters, $f(y_i|\boldsymbol{\theta})$.
- Usually, econometric models are about the parameter vector of a conditional model for y_i given $\boldsymbol{x}_i$.
- Assume y_i and $\boldsymbol{x}_i$ follow any joint density function $f(y, \boldsymbol{x}; \boldsymbol{\delta})$ with $\boldsymbol{\delta}$ being the parameter vector describing the joint process.
- Decompose the joint density in terms of the conditional density of y_i given $\boldsymbol{x}_i$ and the marginal density of $\boldsymbol{x}_i$:

$$f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\delta}) = f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\delta}) \cdot f(\boldsymbol{x}|\boldsymbol{\delta})$$

- The sample log likelihood is:

$$\ln L(\boldsymbol{\delta}|y_1,\ldots,y_n,x_1,\ldots,x_n) = \sum_{i=1}^{n} \ln f(y_i,\boldsymbol{x}_i|\boldsymbol{\delta})$$

$$= \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\delta}) + \sum_{i=1}^{n} \ln f(\boldsymbol{x}_i|\boldsymbol{\delta})$$

- Assume that the process generating $\boldsymbol{x}_i$ takes place outside the model of interest.
- That means that the parameters that appear in $f(\boldsymbol{x}_i|\boldsymbol{\delta})$ do not overlap with those that appear in $f(y_i|\boldsymbol{x}_i;\boldsymbol{\delta})$.

- Thus, we partition $\boldsymbol{\delta}$ into $[\boldsymbol{\theta}, \boldsymbol{\gamma}]$ so that the log-likelihood function may be written

$$\ln L(\boldsymbol{\delta}|y_1, \ldots, y_n, x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(y_i, \boldsymbol{x}_i|\boldsymbol{\delta})$$
$$= \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \ln f(\boldsymbol{x}_i|\boldsymbol{\gamma})$$

- As long as $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ have no elements in common and no restrictions connect them (such as $\boldsymbol{\theta} + \boldsymbol{\gamma} = 1$), then the two parts of the log likelihood may be analyzed separately.
- In most cases, the marginal distribution of $\boldsymbol{x}_i$ will be of no (or secondary) interest.

Definition 1.8 (Conditional Likelihood Function)

The conditional likelihood function of $\boldsymbol{\theta}$ for a random variable y_i with density function $f(y \mid x, \boldsymbol{\theta})$ given the random variable $\boldsymbol{x}_i$ is

$$L_i(\boldsymbol{\theta} \mid y_i, \boldsymbol{x}_i) \equiv f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}).$$

We will denote the logarithm of the likelihood function, the conditional loglikelihood function, by

$$\ln L_i(\boldsymbol{\theta} | y_i, \boldsymbol{x}_i) = \ln f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}).$$

Consistency of the MLE

Theorem 13.1 in Wooldridge (2010, p. 475) establishes the consistency of MLE estimator. Without going into the technical details, the key assumptions are:

- The conditional density of y given $\boldsymbol{x}$ is known up to the parameter θ, i.e. the density function is correctly specified.
- θ is (globally) identified in Θ, i.e. for every $\theta_1 \in \Theta$, $\theta \neq \theta_1$ implies that

$$\Pr\left[f(y|\boldsymbol{x},\theta) \neq f(y|\boldsymbol{x},\theta_1)\right] > 0$$

- The log-likelihood function is continuous in θ.

Then, the MLE $\widehat{\boldsymbol{\theta}}_n$ is consistent. That is:

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$$

Asymptotic Properties of the MLE

Asymptotic Distribution of the MLE:

Under the regularity conditions, the MLE is asymptotically normal so that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathcal{I}_1(\boldsymbol{\theta}_0)^{-1})$$

Finally, the MLE is asymptotically efficient relative to all other consistent, uniformly asymptotically normal (CUAN) estimators.

VC-Matrix Estimation

- The variance-covariance matrix of the limiting distribution of the stabilizing transformation is $\mathcal{I}_1(\boldsymbol{\theta}_0)^{-1}$.
- For a finite n we obtain the following relationship:

$$\mathrm{V}\left[\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right] \approx \frac{1}{n}\mathcal{I}_1(\boldsymbol{\theta}_0)^{-1} = \mathcal{I}(\boldsymbol{\theta}_0)^{-1}$$

  where the equality results from $n\mathcal{I}_1 = \mathcal{I}$.
- Replacing $\mathcal{I}_1$ with a consistent estimator $\hat{\mathcal{I}}_1$ yields the estimator of the variance-covariance matrix:

$$\hat{\mathrm{V}}\left[\widehat{\boldsymbol{\theta}}_n\right] = \frac{1}{n}\hat{\mathcal{I}}_1(\widehat{\boldsymbol{\theta}}_n)^{-1} = \left(n\hat{\mathcal{I}}_1(\widehat{\boldsymbol{\theta}}_n)\right)^{-1} \tag{1}$$

VC-Matrix Estimation

Three alternatives for estimating $\mathcal{I}_1$ consistently:

1. **Expected Hessian Estimator**

$$\hat{\mathcal{I}}_1 = -\operatorname{E}\left[\frac{\partial^2 \ln f(y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$$

2. **Sample Hessian Estimator**

$$\hat{\mathcal{I}}_1 = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln f(y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$$

3. **Outer Product Estimator**

$$\hat{\mathcal{I}}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f(y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\cdot\frac{\partial \ln f(y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$$

VC-Matrix Estimation

- Estimator (1) comes closest to the definition of the information matrix, but it can only be computed when the Hessian is available as an explicit function of $\boldsymbol{\theta}$, which is not often.
- (2) This is the most common variance estimator. It is based on the formula for the expected Hessian. The second derivative matrix can be calculated analytically if the derivatives are known. Alternatively, it can be calculated using numerical derivatives.
- Alternative (3) is the OPG-form and uses the information equality property. This is the form which is computationally least burdensome, because it only requires the computation of first derivatives.

VC-Matrix Estimation

Using the three estimators for $\mathcal{I}_1$ yields the following estimators for the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}_n$:

$$\text{(i) } \hat{\mathrm{V}}\left[\widehat{\boldsymbol{\theta}}_n\right] = \mathcal{I}(\boldsymbol{\theta})^{-1}\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

$$\text{(ii) } \hat{\mathrm{V}}\left[\widehat{\boldsymbol{\theta}}_n\right] = -\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

$$\text{(iii) } \hat{\mathrm{V}}\left[\widehat{\boldsymbol{\theta}}_n\right] = \left[\sum_{i=1}^{n} \frac{\partial \ln f(Y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \cdot \frac{\partial \ln f(Y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

Additionally, one can also use the robust variance covariance matrix estimator:

$$\text{(iv) } \hat{\mathrm{V}}\left[\widehat{\boldsymbol{\theta}}_n\right] =$$

$$\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1} \left[\sum_{i=1}^{n} \frac{\partial \ln f(Y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \cdot \frac{\partial \ln f(Y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right] \left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

Example 1.8

Assume that a random sample of size n has been drawn from a Bernoulli distribution, as before. Show that estimators (i)-(ii) are identical.

MLE of Linear Model

Consider the standard linear regression model

$$y_i = \boldsymbol{x}_i'\beta + \varepsilon_i \quad i = 1, \ldots, n$$

where $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma^2\right)$

- Write down the likelihood and loglikelihood function of the sample.
- Derive the maximum likelihood estimator of the model parameters.
- The density of normally distributed random variable Z with mean μ and variance σ^2 is given by:

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}$$

MLE of Linear Model

- Setting these to zero yields two expressions that look familiar from OLS estimation,

$$\hat{\boldsymbol{\beta}}_{ML} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i y_i$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} \right)^2$$

- The only difference to the OLS expressions is that the estimator of $\hat{\sigma}^2_{ML}$ does not involve a degrees of freedom adjustment.

- This implies that ML estimation yields a smaller variance of ε than OLS, but the difference vanishes as the sample size increases.

Kullback-Leibler Divergence and Pseudo Maximum Likelihood

Kullback-Leibler Divergence

Definition 1.9 (Kullback-Leibler Divergence)

The Kullback-Leibler divergence between densities $f(x)$ and $g(x)$ is

$$KLIC(f,g) = \int f(y) \ln\left(\frac{f(y)}{g(y)}\right) dy$$

- The Kullback-Leibler divergence is also known as the "Kullback-Leibler Information Criterion," and hence the acronym KLIC.
- The KLIC distance is not symmetric, thus, $KLIC(f,g) \neq KLIC(g,f)$.

Kullback-Leibler Divergence

Theorem: Properties of KLIC

1. $KLIC(f, f) = 0$
2. $KLIC(f, g) \geq 0$
3. $f = \arg\min_g KLIC(f, g)$

Kullback-Leibler Divergence

Theorem: Properties of KLIC

1. $KLIC(f, f) = 0$
2. $KLIC(f, g) \geq 0$
3. $f = \arg\min_g KLIC(f, g)$

Let $f_\theta = f(y|\theta)$ be a parametric family with $\theta \in \Theta$. From property 3, we deduce that θ_0 minimizes the KLIC divergence between f and f_θ.

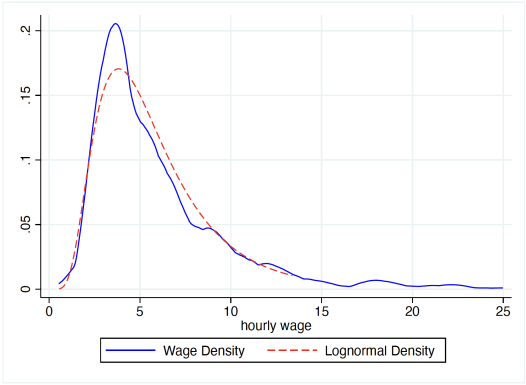
Theorem: Minimum of KLIC divergence

If $f(y) = f(y|\theta_0)$ for some $\theta \in \Theta_0$, then

$$\theta_0 = \arg\min_{\theta \in \Theta} KLIC(f, f_\theta)$$

This simply points out that since the KLIC divergence is minimized by setting the densities equal, the KLIC divergence is minimized by setting the parameter equal to the true value.

Approximating Models



- wage density: $f(y)$, log-normal density: $f(y|\theta)$
- log-normal parametric model appears to be close to wage density but there are differences.
- parametric model is likely misspecified, but it could be a good approximation-approximating model

Approximating Models

- Given the concept of an approximating model, how should we choose its parameter?
- One solution is to minimize a measure of the divergence between densities

Definition 1.10 (Pseudo-true parameter)

The pseudo-true parameter θ^* for a model f_θ that best fits the true density f based on Kullback-Leibler divergence is

$$\theta^* = \arg\min_{\theta \in \Theta} KLIC(f, f_\theta).$$

- This definition corresponds to the true parameter value if the model is correctly specified.
- The name "pseudo-true parameter" refers to the fact that when f_θ is a misspecified parametric model, there is no true parameter, but there is a parameter value that produces the best fitting density.

Notice that we can write the KLIC divergence as

$$\begin{aligned} KLIC(f, f_\theta) &= \int f(y) \ln f(y) dy - \int f(y) \ln f(y|\theta) dy \\ &= \int f(y) \ln f(y) dy - \mathrm{E}\left[\ln f(y|\theta)\right] \end{aligned}$$

Since only second part depends on θ, θ^* can be found by maximizing $\mathrm{E}\left[\ln f(y|\theta)\right]$.

Thus, the pseudo-true parameter, just like the true parameter, is an analog of the MLE that maximizes the sample counterpart of $\mathrm{E}\left[\ln f(y|\theta)\right]$.

Connection to the MLE

Maximum likelihood has twin roles:

1. It is an estimator of the true parameter θ_0 when the model $f(y|\theta)$ is correctly specified.
2. It is an estimator of the pseudo-true parameter θ^* when the model is not correctly specified.

If the model is correct, the MLE will produce an estimator of the true distribution, but otherwise it will produce an approximation that produces the best fit as measured by the Kullback-Leibler divergence.

Pseudo-Maximum Likelihood

- The pseudo-MLE is the estimator which maximizes a loglikelihood function using a misspecified density function.
- The pseudo-MLE, $\hat{\boldsymbol{\theta}}_{PMLE}$ converges in probability to the pseudo-true value $\boldsymbol{\theta}^*$, i.e.

$$\hat{\boldsymbol{\theta}}_{PMLE} \xrightarrow{p} \boldsymbol{\theta}^*$$

- If the true data-generating process (dgp) differs from the assumed density, then usually $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$ and the pseudo-MLE is inconsistent.

Pseudo-Maximum Likelihood

- Huber (1967) and White (1982) showed that the asymptotic distribution of the pseudo-MLE is similar to that for the MLE, except that it is centered around $\boldsymbol{\theta}^*$ and the information matrix equality no longer holds. Then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{PMLE} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathbf{N}(0, \boldsymbol{A}^{*-1}\boldsymbol{B}^*\boldsymbol{A}^{*-1})$$

with

$$A^* \equiv -\operatorname{E}\left[\mathcal{H}_i(\boldsymbol{\theta}^*)\right]$$
$$B^* \equiv \operatorname{E}\left[\mathcal{S}_i(\boldsymbol{\theta}^*)\mathcal{S}_i(\boldsymbol{\theta}^*)'\right]$$

Linear Exponential Family

- Pseudo ML estimation becomes particularly attractive if the pseudo-true distribution (approximating model) used for ML estimation is taken from the linear exponential family of distributions.
- An LEF density can be expressed as

$$f(y|\mu) = \exp\left(a(\mu) + b(y) + c(\mu)y\right)$$

  with $\mathrm{E}[y] = \mu$.
- Most important members of the LEF are the standard normal, the Poisson distribution and the exponential distribution.
- The pseudo-MLE based on an LEF density is consistent provided only that the conditional mean of y given x is correctly specified.
- Note that the actual dgp for y need not be LEF. It is the specified density, potentially incorrectly specified, that is LEF.

Linear Exponential Family

Distribution	$f(y) = \exp\{a(\cdot) + b(y) + c(\cdot)y\}$	$\mathrm{E}[y]$	$\mathrm{V}[y] = [c'(\mu)]^{-1}$
Normal	$\exp\{\frac{-\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{\mu}{\sigma^2}y\}$	μ	σ^2
Bernoulli	$\exp\{\ln(1-p) + \ln(p/(1-p))y\}$	$\mu = p$	$\mu(1-\mu)$
Exponential	$\exp\{\ln\lambda - \lambda y\}$	$\mu = 1/\lambda$	μ^2
Poisson	$\exp\{-\lambda - \ln y! + y\ln\lambda\}$	$\mu = \lambda$	μ

Table: Linear Exponential Family Densities: Leading Examples

Hypothesis and Specification Tests

Formalism for hypotheses

- Linear hypothesis:

$$\begin{aligned} H_0 : \boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r} &= \mathbf{0} \\ H_1 : \boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r} &\neq \mathbf{0} \end{aligned}$$

- The number of (linear) restrictions is h. This is also the number of equations in the hypothesis. The number of unknown parameters in the model is k. In general, $h \leq k$.

- Example: $k = 3$, $H_0 : \beta_1 = 1 \wedge \beta_2 - \beta_3 = 2$. Thus $h = 2$ and

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix} \text{ and } \boldsymbol{r} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Formalism for hypotheses

- Nonlinear hypothesis:

$$\begin{aligned} H_0 : \boldsymbol{h}(\boldsymbol{\theta}) &= \mathbf{0} \\ H_1 : \boldsymbol{h}(\boldsymbol{\theta}) &\neq \mathbf{0} \end{aligned}$$

- $\boldsymbol{h}(\cdot)$ is a vector-valued function. The number of restrictions is h. This is also the number of rows in the hypothesis. The number of unknown parameters in the model is q. In general, $h \neq q$
- Examples: $H_0 : \theta_j = 0$; $H_0 : \theta_1 = \theta_2$; $H_0 : \frac{\theta_1}{\theta_2} = 1$

Unrestricted and Restricted Models

- When we test restrictions on parameters, we can estimate two models: the unrestricted model and the restricted model.
- Let's consider an example. Suppose we estimate a model in which the explanatory variables enter via the single index $\boldsymbol{x}_i'\boldsymbol{\beta}$ with maximum likelihood.
- To keep things simple, assume that the model has only two explanatory variables, x_{1i} and x_{2i}, and two parameters, β_1 and β_2.
- We would like to test $H_0 : \beta_1 = c\beta_2$ where c is a real number. Thus, we have one linear restriction.

Unrestricted and Restricted Models

- In the unrestricted model, we use the linear index $x_{1i}\beta_1 + x_{2i}\beta_2$.
- We obtain unrestricted estimates of β_1 and β_2 and the unrestricted value of the likelihood function at its maximum, $\hat{L}_U = L(\widehat{\boldsymbol{\beta}}_U)$

Unrestricted and Restricted Models

- In the restricted model, we impose the restriction we want to test. One way to do this is to rewrite the linear index as

$$x_{1i}\beta_1 + x_{2i}\beta_2 = x_{1i}\beta_1 + x_{2i}\frac{1}{c}\beta_1 = \left(x_{1i} + \frac{x_{2i}}{c}\right)\beta_1$$

- Effectively, we use only one explanatory variable, namely, $\left(x_{1i} + \frac{x_{2i}}{c}\right)$, and estimate its coefficient $\hat{\beta}_1$.
- The value of the likelihood at its maximum in this restricted model we denote by $\hat{L}_R = L(\widehat{\boldsymbol{\beta}}_R)$.
- Of course, we can also recover $\hat{\beta}_2 = \hat{\beta}_1/c$ from the restricted model.

Unrestricted and Restricted Models

In general case; we use the following notation:

- $\widehat{\boldsymbol{\theta}}_U$ and $\widehat{\boldsymbol{\theta}}_R$ are the unrestricted and restricted estimates, respectively.
- $\hat{L}_U$ and $\hat{L}_R$ are the values of the likelihood functions at their respective maximums in the unrestricted and restricted models.
- h is the number of restrictions imposed by $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$ (in the above example, $h = 1$.)

LR, Wald, and LM tests

- For testing single parameter hypotheses after ML estimation, t-statistics can be constructed.
- The asymptotic standard errors can be computed from the diagonal elements of the asymptotic variance matrix of the parameter estimates.
- For testing parameter restrictions of the form $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$, three equivalent tests exist for maximum likelihood estimation.
- These three tests, sometimes referred to as the trinity of tests.

Likelihood Ratio (LR) Test

- The likelihood ratio (LR) test statistic is

$$LR = \frac{\hat{L}_R}{\hat{L}_U} \text{ with } -2\ln LR \sim \chi^2_{(h)}.$$

- There is a slightly different, but obviously equivalent, version of this test statistic:

$$LR = 2\left(\ln L(\hat{\boldsymbol{\theta}}_U) - \ln L(\hat{\boldsymbol{\theta}}_R)\right) \sim \chi^2_{(h)}.$$

LR, Wald, and LM tests

Likelihood Ratio (LR) Test

- The intuition of the likelihood ratio test is simple.
- If the restriction is true, then both the restricted and the unrestricted model should explain the data equally well, and thus the two likelihoods should be similar.
- If the restriction is violated, the likelihood of the restricted model will be significantly smaller and the value of the LR statistic will reflect that.

Wald Test

- The Wald test statistic is given by

$$W = \left(\boldsymbol{h}(\widehat{\boldsymbol{\theta}})\right)' \left[\mathrm{V}\left[\boldsymbol{h}(\widehat{\boldsymbol{\theta}})\right]\right]^{-1} \left(\boldsymbol{h}(\widehat{\boldsymbol{\theta}})\right) \sim \chi^2_{(h)},$$

  where we can estimate the weighting matrix by

$$\hat{\mathrm{V}}\left[\boldsymbol{h}(\widehat{\boldsymbol{\theta}})\right] = \left[\frac{\partial \boldsymbol{h}(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}'}\right] \hat{\mathrm{V}}\left[\widehat{\boldsymbol{\theta}}\right] \left[\frac{\partial \boldsymbol{h}(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}'}\right]'$$

LR, Wald, and LM tests

- The Wald test for parameter restrictions follows the well-known principle of all Wald-type test statistics.
- Under the null hypothesis, all the elements of the vector $\boldsymbol{h}(\boldsymbol{\theta})$ are zero.
- To test this, we estimate the elements of this vector using $\widehat{\boldsymbol{\theta}}$, square each element (after normalizing such that each element is standard normal), and then sum these squared, normalized elements to obtain a single test statistic that has a χ^2 distribution.

Lagrange multiplier (LM) or score test

- The Lagrange multiplier (LM) or score test has the test statistic

$$LM = \left(\frac{\partial \ln L(\widehat{\boldsymbol{\theta}}_R)}{\partial \boldsymbol{\theta}} \right)' \left[\mathcal{I}(\widehat{\boldsymbol{\theta}}_R) \right]^{-1} \left(\frac{\partial \ln L(\widehat{\boldsymbol{\theta}}_R)}{\partial \boldsymbol{\theta}} \right) \sim \chi^2_{(h)}.$$

LR, Wald, and LM tests

- We know that the likelihood function is maximized by the population parameter values and that the first derivatives (scores) are zero at the population parameter values.
- If the restriction imposed under the null is correct, the scores evaluated at the restricted parameters should be zero. If it is not correct, the scores will be different from zero. In other words, the restricted model yields estimates that do not maximize the likelihood of the sample (they only maximize the restricted likelihood function).
- The LM test thus checks whether the scores evaluated at the restricted estimates are jointly zero, using a quadratic form with appropriate weighting matrix (similar to a Wald test statistic).

Practical Considerations

- Which of these three equivalent tests should we use in practice? There are three main considerations.
- The first two are only mentioned for the sake of completeness, the third one is most relevant.
  1. Small-sample properties: Even though the three statistics are asymptotically equivalent, they have different small-sample properties.
  2. Invariance: Only the LR statistic is invariant to re-parameterizing the conditional density (see C&T, section 7.2.9). Problems arise particularly in small samples.

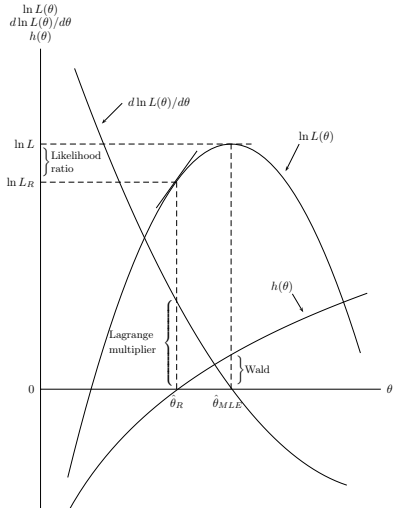
LR, Wald, and LM tests

3. Computational issues:
   - In those cases in which both the unrestricted and the restricted model can be estimated easily, the LR statistic is straightforward to compute.
   - The Wald statistic requires only estimation of the unrestricted model (and then imposes the restrictions on these estimates).
   - Conversely, the LM statistic requires only estimation of the restricted model.

The LM statistic is thus convenient whenever the restricted model is easier to compute, for instance in cases in which the restrictions make a non-linear model linear.

Hypothesis and Specification Tests

We consider maximum likelihood estimation of a parameter θ and a test of the hypothesis $H_0 : h(\theta) = 0$



- Likelihood ratio test: If the restriction $h(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function.
- Wald test: If the restriction is valid, then $h(\hat{\theta}_{MLE})$ should be close to zero because the MLE is consistent.
- Lagrange multiplier test: If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood.