Likelihood based Statistical Inference

Dottorato in Economia e Finanza
Dipartimento di Scienze Economiche
Univ. di Verona

L. Pace, A. Salvan, N. Sartori

Udine, April 2008
Lecture 2
Likelihood: observed quantities, exact properties, inference

2.1 Likelihood: observed quantities
2.2 Likelihood: exact properties
2.3 Likelihood: inference (no nuisance parameters)
2.4 Reparameterizations
2.5 A two-parameter model
2.6 Example: linear model with normal errors
2.1 Likelihood: observed quantities

*PS96:/PS97 §1.4.2
*PS01: §§3.1–3.4


2.2 Likelihood: exact properties

*PS01: Ch. 5
*PS96/PS97: SS1.4.2–1.4.3, 9.2.1


2.3 Likelihood: inference (no nuisance parameters)

*PS01: §3.6
*PS96: §1.4.2–1.4.3, §3.5 (introd.)
*PS97: §1.4.2–1.4.3, §3.4 (introd.)


2.4 Reparameterizations

*PS96: §1.4.3


2.1 Likelihood: observed quantities

□ The **likelihood function** is

\[
L = L(\theta) = L(\theta; y) = c(y) \, p_Y(y; \theta),
\]

(1)

\(c(y) > 0\) arbitrary proportionality constant.

**Interpretation:** on the basis of data \(y\), the parameter value \(\theta' \in \Theta\) is more credible than \(\theta'' \in \Theta\) as an index of the data generating probability model if \(L(\theta') > L(\theta'')\).

\(\Rightarrow\) the empirical support received by \(\theta' \in \Theta\) from \(y\) is to be compared with that received by \(\theta'' \in \Theta\) on the basis of the **ratio** \(L(\theta')/L(\theta'')\).

\[L(\theta)\text{ is not a density on } \Theta!\]

Two likelihood functions are called **equivalent** if they agree up to a positive multiplicative constant (which can depend on \(y\) but not on \(\theta\)).
In many cases it is more convenient to consider the log-likelihood function

\[ l = l(\theta) = l(\theta; y) = \log L , \]  

we let \( l(\theta) = -\infty \) if \( L(\theta) = 0 \) (the log-likelihood is defined up to an additive constant, which only depends on \( y \)).

Independent observations:

\[ l(\theta) = \sum_{i=1}^{n} \log p_{Y_i}(y_i; \theta) . \]  

A value of \( \theta \) that maximizes \( L(\theta; y) \) over \( \Theta \), that is a value \( \hat{\theta} \) such that \( L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta) \), is called maximum likelihood estimate (m.l.e.) of \( \theta \).
Likelihood: observed quantities

- **Regular parametric model** if $\Theta$ is an open subset of $\mathbb{R}^p$ and the function $l(\theta)$ is continuously differentiable on $\Theta$ up to the third order. A major condition for a model to be regular is that all its probability distribution have the same support. Typically exponential families are regular models.

- **Log-likelihood derivatives.** In a regular parametric model,

\[
\begin{align*}
  l_r &= l_r(\theta; y) = \frac{\partial}{\partial \theta^r} l(\theta), \\
  l_{rs} &= l_{rs}(\theta; y) = \frac{\partial^2}{\partial \theta^r \partial \theta^s} l(\theta), \\
  l_{rst} &= l_{rst}(\theta; y) = \frac{\partial^3}{\partial \theta^r \partial \theta^s \partial \theta^t} l(\theta),
\end{align*}
\]
The score function is the vector $l_\star = l_\star(\theta; y) = (l_1, \ldots, l_p)$.

In the following, unless otherwise stated, it will be assumed that the m.l.e. is unique and that it is the unique solution to the likelihood equation

$$l_\star(\theta) = 0.$$  \hspace{1cm} (4)

Sufficient conditions exist for specific classes of parametric models (Ch.'s 5–7 of PS97).

The observed information matrix is

$$j = j(\theta) = \begin{pmatrix} -l_{11} & \cdots & -l_{1p} \\ \vdots & \ddots & \vdots \\ -l_{p1} & \cdots & -l_{pp} \end{pmatrix}.$$ 

or, in compact notation, $j = [-l_{rs}]$; here and below the matrix with elements $a_{rs}$ is indicated by $[a_{rs}]$. 
Likelihood: observed quantities

- The elements of the inverse of a matrix \([a_{rs}]\) will be denoted by upper indices, that is, \([a_{rs}]^{-1} = [a^{rs}]\); in particular \(j^{-1} = [j^{rs}]\) and \(i^{-1} = [i^{rs}]\), when these inverses exist.

- Quantities like \([l_{rst}]\) are called arrays.

- Sometimes the notations \(\hat{l} = l(\hat{\theta})\), \(\hat{j} = j(\hat{\theta})\), \(\hat{i} = i(\hat{\theta})\), etc., are used to denote likelihood quantities evaluated at \(\hat{\theta}\). By slightly adjusting the terminology, if \(\theta = (\tau, \zeta)\) and \(\hat{\theta} = (\hat{\tau}, \hat{\zeta})\), we say that \(\hat{\tau}\) is the maximum likelihood estimate of \(\tau\).

- In regular models, \(\hat{\theta}\) and partial derivatives of \(l(\theta)\) with respect to components of \(\theta\) give us information about the behaviour of the log-likelihood function.

With \(p = 1\),

\[
\begin{align*}
l(\theta) &= l(\hat{\theta}) + (\theta - \hat{\theta})l_*(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2l_{**}(\hat{\theta}) + \ldots \\
&= l(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2j(\hat{\theta}) + \ldots .
\end{align*}
\]
Hence, the relative log-likelihood $l(\theta) - l(\hat{\theta})$ may be expanded as

$$l(\theta) - l(\hat{\theta}) = -\frac{1}{2} (\theta - \hat{\theta})^2 j(\hat{\theta}) + \ldots.$$ 

Thus, the larger $j(\hat{\theta})$ is, the more rapidly values of $\theta$ different from $\hat{\theta}$ lose empirical support. In other words, the larger $j(\hat{\theta})$ is, the more the likelihood is concentrated around $\hat{\theta}$. For this reason, $j(\hat{\theta})$ is a measure of information about $\theta$ contained in the data. When $p > 1$, $j(\hat{\theta})$, is positive-definite. Expansion:

$$l(\theta) - l(\hat{\theta}) = - (\theta - \hat{\theta})^\top j(\hat{\theta})(\theta - \hat{\theta})/2 + \ldots,$$

$\Rightarrow$ $l(\theta) - l(\hat{\theta})$ can be approximated around $\hat{\theta}$ by a negative-definite quadratic form.

To sum up, in a regular model the log-likelihood function has approximately a parabolic behaviour, at least in the region with largest likelihood values.
The practise of examining the log-likelihood function is to be recommended and it is facilitated by the graphic options of many popular statistics packages. For an approach to inference which is based entirely on the likelihood function, see Edwards (1972) and Royall (1997, *Statistical Evidence*, Chapman and Hall, London).
2.2 Likelihood: exact properties

**Behaviour under one-to-one transformations**

Let $\mathcal{F} = \{p_Y(y; \theta), \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a parametric stat. model for data $y$.

- **Invariance under one-to-one data transformations.**
  Let $t = g(y)$ be a one-to-one function of $y$ defined on $\mathcal{Y}$ with range $\mathcal{T}$, a subset of the same Euclidean space, and let us denote by $\mathcal{F}_T = \{p_T(t; \theta), \theta \in \Theta \subseteq \mathbb{R}^p\}$ the statistical model for $T = g(Y)$.
  Information about $\theta$ contained in $y$ is equivalent to information contained in $t$: only the measurement of data changes. Likelihood is invariant under one-to-one transformations of the data. Indeed the likelihood for $\theta$ based on the statistical model $\mathcal{F}_T = \{p_T(t; \theta), \theta \in \Theta \subseteq \mathbb{R}^p\}$ and on data $t = g(y)$ is equivalent to that based on $\mathcal{F}$ and on data $y$:

$$L^T(\theta; t(y)) = c(y)L^Y(\theta; y).$$
**Likelihood: exact properties**

- **Invariance under reparameterizations.**
  Let \( \psi = \psi(\theta) \) be a reparameterization of \( \mathcal{F} \) (hereafter a one-to-one smooth function from \( \Theta \subseteq \mathbb{R}^p \) to \( \Psi \subseteq \mathbb{R}^p \), infinitely differentiable together with its inverse).
  In the new parameterization:
  
  \[
  \mathcal{F} = \{ p_Y(y; \psi), \psi \in \Psi \subseteq \mathbb{R}^p \},
  \]
  
  with \( \Psi = \{ \psi \in \mathbb{R}^p : \psi = \psi(\theta), \theta \in \Theta \} \).
  The statistical model as well as the inference problem are left unchanged. Hence, inference conclusions should not depend on the parameterization.
The likelihood function is indeed \textit{invariant} under reparameterizations of the model: the likelihood function $L^\Psi(\psi)$ in the new parameterization and the original likelihood $L^\Theta(\theta)$ are related as

$$L^\Psi(\psi) = L^\Theta(\theta(\psi)),$$

so that

$$l^\Psi(\psi) = l^\Theta(\theta(\psi)),$$

$$\Rightarrow \hat{\psi} = \psi(\hat{\theta}).$$

This means that the m.l.e. is \textit{equivariant under reparameterizations}.

In general, likelihood inference about $\psi$ is simply the translation in the new parameter space of conclusions about $\theta$. 
Repeated sampling principle (r.s.p.) ⇒ study of the probability distribution of the random variable $l(\theta) = l(\theta; Y)$, or of related quantities, for $\theta$ fixed and as $y$ varies in the sample space $\mathcal{Y}$ according to a density $p_Y(y; \tilde{\theta})$ in $\mathcal{F}$, where $\tilde{\theta} \in \Theta$ is a parameter value not necessarily equal to $\theta$.

Null distribution if $\tilde{\theta} = \theta$.
Non-null distribution if $\tilde{\theta} \neq \theta$.

Analogously, we refer to null moments (evaluated with respect to a null distribution), and to non-null moments (evaluated with respect to a non-null distribution).
Symbols such as $E_\theta(\cdot)$, $\text{Var}_\theta(\cdot)$ are used to indicate expectation, variance (or covariance matrix), etc., calculated with reference to $p_Y(y; \theta)$.


\[ E_\theta(l(\theta; Y)) > E_\theta(l(\theta'; Y)), \quad \theta' \neq \theta. \] (7)
To prove (7), note first that because of parameter identifiability

\[ Pr_\theta \left( \frac{p_Y(Y; \theta')} {p_Y(Y; \theta)} = 1 \right) < 1, \]

that is the r.v. \( p_Y(Y; \theta')/p_Y(Y; \theta) \) is non-degenerate for any pair of values \( \theta \) and \( \theta' \) in \( \Theta \) with \( \theta \neq \theta' \). Moreover, \( \log(\cdot) \) is a strictly concave function, so that, by Jensen’s inequality,

\[ E_\theta \left( \log \frac{p_Y(Y; \theta')} {p_Y(Y; \theta)} \right) < \log \left\{ E_\theta \left( \frac{p_Y(Y; \theta')} {p_Y(Y; \theta)} \right) \right\} = 0, \tag{8} \]

taking into account that

\[ E_\theta \left( \frac{p_Y(Y; \theta')} {p_Y(Y; \theta)} \right) = \int_Y \frac{p_Y(y; \theta')} {p_Y(y; \theta)} p_Y(y; \theta) \, dy = 1. \]

(Notation for the continuous case has been used without loss of generality). Inequality (7) follows from (8).
□ **Expectation of the score function.**

For a regular model, assuming that the conditions which permit differentiation and integration to be interchanged are satisfied,

\[
E_\theta(l_*(\theta)) = E_\theta(l_*(\theta; Y)) = 0 , \quad \theta \in \Theta .
\]

(9)

Thus, in agreement with the property that log-likelihood functions associated to data from \( p_Y(y; \theta^0) \) have, on average, maximum at \( \theta^0 \), the components of the score function evaluated at \( \theta^0 \), are, on average, equal to zero.

Property (9) could also be phrased as:
the likelihood equation \( l_*(\theta) = 0 \) is an unbiased estimating equation.
In general, an estimating equation is an equation in $\theta$ of the form $q(\theta; y) = 0$. Its solution gives an estimate $\hat{\theta}^q$ of $\theta$. The estimating equation $q(\theta; y) = 0$ is called unbiased $E_\theta(q(Y; \theta)) = 0$ for all $\theta \in \Theta$. Unbiasedness of an estimating equation is a key property to show consistency of the corresponding estimator (see e.g. § 6.2 in PS01; Serfling, 1980, § 7.2.1; van der Vaart, 1998, § 5.2).
We call expected information or Fisher information the quantity

$$i(\theta) = E_\theta (j(\theta)) = \left[-E_\theta \left( \frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right) \right],$$

(null) expectation of the observed information. It is a $p \times p$ matrix, $i(\theta) = [i_{rs}(\theta)]$. Under random sampling of size $n$, $i(\theta)$ is proportional to $n$. Indeed, $i(\theta) = ni_1(\theta)$, where $i_1(\theta)$ is the expected information for one observation.
Covariance matrix of the score function.

Under regularity conditions (see e.g. Azzalini, 1996, section 3.2.4), the information identity

\[ E_{\theta} \left( l_\ast(\theta) \left( l_\ast(\theta) \right)^\top \right) = i(\theta) , \quad \theta \in \Theta \]  

(10)

holds. In other words, the expected information matrix is the null second moment of the score and, as such, is a non-negative definite matrix. In other words, the log-likelihood has Hessian matrix which is, on average, non-positive definite at the true parameter value.
2.3 Likelihood: inference (no nuisance parameters)

From a likelihood function $L(\theta)$ it is natural:

\textit{i)} to use as a point estimate for $\theta$ the value $\hat{\theta} = \hat{\theta}(y)$ having maximum likelihood;

\textit{ii)} to consider confidence regions of the form

$$\hat{\Theta}(y) = \{\theta \in \Theta : L(\theta) \geq cL((\hat{\theta}))\},$$

with $0 < c < 1$, made up of parameter values having large likelihood;
iii) to test an hypothesis about $\theta$, $H_0 : \theta \in \Theta_0 \subset \Theta$, by comparing the maximum of $L(\theta)$ for $\theta \in \Theta_0$

with $L(\hat{\theta})$, the maximum of $L(\theta)$ for $\theta \in \Theta$.

Let us denote by $L(\hat{\theta}_0)$ the maximum of $L(\theta)$ for $\theta \in \Theta_0$. This leads to the test statistic

$$\frac{L(\hat{\theta})}{L(\hat{\theta}_0)}$$

(11)

with large values significant. This statistic can be equivalently expressed using the log-likelihood, as $l(\hat{\theta}) - l(\hat{\theta}_0)$. 
Likelihood ratio tests and confidence regions

□ Likelihood ratio test

Suppose we want to test the simple hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. Twice the logarithm of (11) is

$$W(\theta_0) = 2 \left\{ l(\hat{\theta}) - l(\theta_0) \right\},$$

also called log likelihood ratio statistic.
Large values of $W(\theta_0)$ are critical against $H_0$. 
Sometimes $W(\theta_0)$ is a monotonically increasing function of a statistic $t(y)$ having known distribution under $H_0$. It is thus easy to compute the observed significance level ($p$-value) associated to the observed $y^{obs}$ as

$$\alpha^{obs} = \Pr_{\theta_0}(W(\theta_0) \geq W(\theta_0)^{obs}) = \Pr_{\theta_0}(t(Y) \geq t(y^{obs})) .$$

Usually, however, one has to resort to an approximation for the null distribution of $W(\theta_0)$.

Under regularity conditions, the asymptotic null distribution of $W(\theta_0)$ is chi-squared on $p$ degrees of freedom, where $p$ is the number of scalar components of $\theta$.

Thus, under such conditions, $W(\theta)$ is an asymptotically pivotal quantity.
Confidence regions

Confidence regions with nominal level (level = null coverage probability) $1 - \alpha$ are

$$\hat{\Theta}(y) = \{ \theta \in \Theta : W(\theta) < \chi^2_{p;1-\alpha} \}.$$ 

The region $\hat{\Theta}(y)$ can be written as

$$\hat{\Theta}(y) = \{ \theta \in \Theta : l(\theta) > l(\hat{\theta}) - \frac{1}{2} \chi^2_{p;1-\alpha} \}.$$ 

With $p = 1$ or 2, graphical representation of likelihood confidence regions is rather simple starting from the log-likelihood plot. Alternatively, the region $\hat{\Theta}(y)$ might also be read from the plot of $W(\theta)$. 

Likelihood: observed quantities, exact properties, inference
One-sided version of $W(\theta)$

When $\theta$ is a scalar, one might be interested in testing $H_0 : \theta = \theta_0$ against the one-sided alternative $H_{1^{dx}} : \theta > \theta_0$ or $H_{1^{sx}} : \theta < \theta_0$.

The natural test statistic based on the likelihood is the signed root of $W(\theta_0)$,

$$r(\theta_0) = \text{sgn}(\hat{\theta} - \theta_0) \sqrt{W(\theta_0)},$$

where $\text{sgn}(\cdot)$ is the sign function: $\text{sgn}(x) = +1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$, $\text{sgn}(x) = 0$ if $x = 0$. Moreover the non-negative square root of $W(\theta_0)$ is to be considered.
One-sided version of $W(\theta)$

When $\theta$ is a scalar, one might be interested in testing $H_0 : \theta = \theta_0$ against the one-sided alternative $H_{1dx} : \theta > \theta_0$ or $H_{1sx} : \theta < \theta_0$.

The natural test statistic based on the likelihood is the signed root of $W(\theta_0)$,

$$r(\theta_0) = \text{sgn}(\hat{\theta} - \theta_0) \sqrt{W(\theta_0)},$$

where $\text{sgn}(\cdot)$ is the sign function: $\text{sgn}(x) = +1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$, $\text{sgn}(x) = 0$ if $x = 0$. Moreover the non-negative square root of $W(\theta_0)$ is to be considered.

The test based on $r(\theta_0)$ rejects $H_0$ for large (small) values when the alternative is $H_{1dx}$ ($H_{1sx}$).

For some models $r(\theta_0)$ is a monotonically increasing function of a statistic $t(y)$ having known distribution under $H_0$.

Usually, however, under regularity conditions, the approximation $r(\theta_0) \sim N(0, 1)$ is used for the null distribution.
\( r(\theta_0) \) with two-sided and symmetric around zero critical region is equivalent to the test based on \( W(\theta_0) \).

\[
\hat{\Theta}(y) = \{ \theta \in \Theta : -z_{1-\alpha/2} < r(\theta) < z_{1-\alpha/2} \}.
\]

For graphical illustrations of construction of confidence regions, see e.g. Example 3.9 in PS01.
Asymptotically equivalent versions of $W(\theta)$ and $r(\theta)$

The statistic

$$W(\theta_0) = 2 \left\{ l(\hat{\theta}) - l(\theta_0) \right\},$$

is also called Wilks’ test statistic, from S. Wilks, who obtained its chi-squared asymptotic null distribution in 1938. $W(\theta_0)$ is a likelihood-based measure of the distance between $\hat{\theta}$ and $\theta_0$. Due to the parabolic approximation for the relative log-likelihood, the asymptotic null distribution of $W(\theta_0)$ coincides with the asymptotic null distribution of

$$W_e(\theta_0) = (\hat{\theta} - \theta_0)^\top i(\theta_0)(\hat{\theta} - \theta_0).$$

The test statistic $W_e(\theta_0)$ is called Wald test statistic. It measures the distance between $\hat{\theta}$ and $\theta_0$ directly on $\Theta$ using the metric defined by the expected information.
A third asymptotically equivalent statistic is the score test, also called Rao’s test. It is given by

\[ W_u(\theta_0) = l_*(\theta_0)^\top i(\theta_0)^{-1} l_*(\theta_0). \]

It measures the distance between \( l_*(\theta_0) \) and \( E_{\theta_0}(l_*(\theta_0)) = 0 \), using the metric defined by the inverse of the expected information. When \( p = 1 \), it is possible to define one-sided versions of \( W_e(\theta_0) \) and of \( W_u(\theta_0) \), analogous to \( r(\theta_0) \). For instance,

\[ r_e(\theta_0) = \sqrt{i(\theta_0)}(\hat{\theta} - \theta_0). \]

The asymptotic null distribution of both \( W_e(\theta_0) \) and \( W_u(\theta_0) \) is \( \chi_p^2 \). The asymptotic null distribution of the one-sided versions (\( p = 1 \)) is \( N(0, 1) \). The same asymptotic results hold if the expected information is replaced by \( i(\hat{\theta}) \) or \( j(\hat{\theta}) \).
Comparisons

- Wald test has a rather simple expression and interpretation, it is widely used in applications. However it does depend on the parameterization of the statistical model (see the following section). On the other hand, $W_u$ and $W$ do not depend on the parameterization.

- If compared with the other two versions, $W$ does respect more closely the shape of the log-likelihood, especially when this is far from being parabolic.

- Confidence regions based on $W$ are necessarily included in the parameter space. This is not always true for confidence regions based on the other two versions.

- $W_u$ is simple to compute and is related to (locally) optimal tests (cf. § 3.5.3 of PS97).
2.4 Reparameterizations

Likelihood and log-likelihood do not depend on the parameterization of \( \mathcal{F} \). Let \( \psi = \psi(\theta) \) be a one-to-one smooth function from \( \Theta \subseteq \mathbb{R}^p \) to \( \Psi \subseteq \mathbb{R}^p \), infinitely differentiable together with its inverse. Then \( \psi \) defines an alternative parameterization of the model. Since \( \theta \) and \( \psi(\theta) \) identify the same element of \( \mathcal{F} \), we have (recall (12) and (13)):

\[
L^\Psi(\psi) = L^\Theta(\theta(\psi)), \quad l^\Psi(\psi) = l^\Theta(\theta(\psi)).
\]

\[
L^\Psi(\psi) = L^\Theta(\theta(\psi)) \quad (12)
\]

\[
l^\Psi(\psi) = l^\Theta(\theta(\psi)) \quad (13)
\]

\( \Rightarrow \) likelihood is an intrinsic function, i.e. it does not depend on the coordinate system expressed by the parameterization of \( \mathcal{F} \).
Reparameterizations

Important consequence for the m.l.e.: **equivariance under reparameterization**. If $\psi$ is an alternative parameterization of $\mathcal{F}$, we will have $\hat{\psi} = \psi(\hat{p}(\cdot)) = \psi(\hat{\theta})$ and $\hat{\theta} = \theta(\hat{\psi})$.

On the other hand, log-likelihood derivatives $l_r$, $l_{rs}$, ..., and their expectations depend on the parameterization. Let us denote by $\psi^a$, $\psi^b$, ... ($a, b = 1, \ldots, p$), the generic components of $\psi$, whereas $\theta^r$, $\theta^s$, ... denote the generic components of $\theta$.

According to the differentiation rule for composite functions (chain rule),

$$\bar{l}_a = \sum_{r=1}^{p} l_r \theta^r_a , \quad (14)$$

where

$$\bar{l}_a = \frac{\partial}{\partial \psi^a} l^\Psi(\psi) , \quad (15)$$

and $l_r = l_r(\theta(\psi))$. 

Likelihood: observed quantities, exact properties, inference
Reparameterizations

Let

\[ \bar{l}_{ab} = \frac{\partial^2}{\partial \psi^a \partial \psi^b} l^\Psi(\psi), \quad \bar{l}_{abc} = \frac{\partial^3}{\partial \psi^a \partial \psi^b \partial \psi^c} l^\Psi(\psi), \] \tag{16} 

e tc., by re-applying the chain rule, we have

\[ \bar{l}_{ab} = \sum_{r,s=1}^p l_{rs} \theta^r_a \theta^s_b + \sum_{r=1}^p l_r \theta^r_{ab}, \] \tag{17} 

where

\[ l_{rs} = l_{rs}(\theta(\psi)) \]

\[ \theta^r_{ab} = \left( \frac{\partial^2}{\partial \psi^a \partial \psi^b} \right) \theta^r(\psi). \]
Reparameterizations

Transformation rules for observed and expected information:

\[
\bar{J}_{ab} = \sum_{r,s=1}^{p} j_{rs} \theta_a^r \theta_b^s - \sum_{r=1}^{p} l_r \theta_{ab}^r ,
\]

(18)

\[
\bar{i}_{ab} = \sum_{r,s=1}^{p} i_{rs} \theta_a^r \theta_b^s .
\]

(19)

Perhaps (14) and (19) look more familiar in matrix notation:

\[
l^\Psi(\psi) = [\theta_a^r]^T \ l^\Theta(\theta(\psi))
\]

and

\[
i^\Psi(\psi) = [\theta_a^r]^T \ i^\Theta(\theta(\psi)) \ [\theta_a^r] ;
\]

(20)

\([\theta_a^r] : p \times p \text{ matrix with } (r, a) \text{ element } \theta_a^r .\)
Let

\[ y = (0.446, 0.604, 2.137, 0.737, 0.996, 1.152, 1.124, 0.137, 0.982, 1.196, \\
0.841, 0.636, 0.459, 0.947, 0.037, 1.307, 0.858, 0.164, 0.444, 0.623) \]

be a random sample of size \( n = 20 \) from a Weibull \( W(\gamma, \lambda) \) distribution, with
p.d.f. for one observation

\[ p_{Y_i}(y_i; \gamma, \lambda) = \lambda \gamma y_i^{\gamma-1} e^{-\lambda y_i^\gamma}, \quad \gamma, \lambda > 0, y_i > 0. \]

Likelihood equation:

\[
\begin{align*}
\frac{n}{\gamma} + \sum \log y_i - \lambda \sum y_i^\gamma \log y_i &= 0 \\
\frac{n}{\lambda} - \sum y_i^\gamma &= 0
\end{align*}
\]
A two-parameter model

The solution (obtained numerically) is \((\hat{\gamma}, \hat{\lambda}) = (1.642, 1.24)\) and

\[
\hat{j} = \begin{pmatrix} 11.76 & 1.63 \\ 1.63 & 13.02 \end{pmatrix}.
\]

\[
W(\gamma, \lambda) = 2 \left\{ n \log \left( \frac{\hat{\lambda}}{\lambda} \right) + n \log \left( \frac{\hat{\gamma}}{\gamma} \right) + (\hat{\gamma} - \gamma) \sum \log y_i - \hat{\lambda} \sum y_i^{\hat{\gamma}} + \lambda \sum y_i^{\gamma} \right\}
\]

Parabolic approximation (Wald test):

\[
W(\gamma, \lambda) \doteq W_e(\gamma, \lambda) = (\theta - \hat{\theta})^\top \hat{j}(\theta - \hat{\theta}), \quad \theta = (\gamma, \lambda)^\top.
\]

For \(H_0: \gamma = 1, \lambda = 1.2\) against \(H_1: \bar{H}_0\), we get

\[
W(1, 1.2) = 6.068 \quad \text{with} \quad \alpha^{\text{oss}} = 0.048,
\]

\[
W_e(1, 1.2) = 4.94 \quad \text{with} \quad \alpha^{\text{oss}} = 0.084,
\]

\[
(\chi_2^2;0.95 = 5.99).
\]
A two-parameter model

Confidence regions for \((\gamma, \lambda)\) based on \(W(\gamma, \lambda)\)

\((1 - \alpha) \in \{0.25, 0.50, 0.75, 0.90, 0.95, 0.99\}\)
A two-parameter model

Confidence regions for $(\gamma, \lambda)$ based on $W(\gamma, \lambda)$

$(1 - \alpha) \in \{0.25, 0.50, 0.75, 0.90, 0.95, 0.99\}$
2.6 Example: linear model with normal errors

Let \( y = (y_1, \ldots, y_n) \) be a realization of \( Y = (Y_1, \ldots, Y_n) \) with \( Y_i \) independent, \( i = 1, \ldots, n \).

Let us assume that

\[
Y_i \sim N(\mu_i, \sigma^2)
\]

where the unknown \( \mu_i \) has to be related to covariates whose values are measured on the same units where values \( y_i \) are measured. Therefore the data pertaining to the \( i \)-th unit are

\[
(y_i, x_{i1}, \ldots, x_{ip}).
\]

The simplest way to relate \( \mu_i \) to \( (x_{i1}, \ldots, x_{ip}) \) is through the linear model

\[
\mu_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p
\]

where \( \beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p \).
2.6 Example: linear model with normal errors

The model parametric statistical model thus defined is known as multiple regression model with normal errors, in brief normal linear regression model.

It is customary to express this model also in the form

\[ Y = X\beta + \epsilon, \]

where \( Y \) is the \( n \times 1 \) random variable generating \( y \), \( X \) is the \( n \times p \) matrix containing the fixed (non-stochastic) values of the covariates, \( \epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n) \) is a vector of i.i.d. Gaussian errors.

Even more succinctly \( y \) is modeled as an observation of \( Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n) \). The parameter space of the model is \( \Theta = \mathbb{IR}^p \times (0, +\infty) \) and the likelihood function is

\[
L(\beta, \sigma^2) = (\sigma^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma^2}(y - X\beta)^\top(y - X\beta)\right\}.
\]
2.6 Example: linear model with normal errors

The log likelihood function is therefore

\[
l(\beta, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta)\] .

For fixed \( \sigma^2 \) the (log) likelihood is maximized by a value of \( \beta \) that renders the shortest possible the distance between \( y \) and \( X\beta \). This value of \( \beta \) does not depend on \( \sigma^2 \). Geometrically, we need to find the orthogonal projection of \( y \) upon the vector space having the columns of \( X \) as its basis. Orthogonality between the "residual" \( y - X\hat{\beta} \) and the basis \( X \) means in terms of scalar products

\[
X^\top (y - X\hat{\beta}) = 0 .
\]

Therefore

\[
X^\top y = X^\top X \hat{\beta} .
\]

When the columns of \( X \) are linearly independent, i.e. \( X \) has rank \( p < n \), one has

\[
\hat{\beta} = (X^\top X)^{-1} X^\top y .
\]
2.6 Example: linear model with normal errors

A simple maximization of $l(\hat{\beta}, \sigma^2)$ gives

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\top(y - X\hat{\beta}).$$

Inference on the linear model with normal error is greatly simplified by the fact that the exact sampling distribution of the maximum likelihood estimator is easily described:

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X^\top X)^{-1})$$

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi^2_{n-p}$$

$\hat{\beta}$ and $\hat{\sigma}^2$ are independent.