Summer Seminar: Philosophy of Statistics
Lecture Notes 7: Estimation II: Methods of Estimation

Aris Spanos [Summer 2019]

1 Introduction

In chapter 12 we discussed estimators and their properties. The essential finite sample and asymptotic properties of estimators are listed in table 12.1.

Table 12.1: Properties of Estimators

Finite sample $(1<n<\infty)$	**Asymptotic** $(n \rightarrow \infty)$
1. Unbiasedness,	5. Consistency (weak, strong)
2. Relative Efficiency,	6. Asymptotic Normality
3. Full Efficiency,	7. Asymptotic Unbiasedness
4. Sufficiency,	8. Asymptotic Efficiency

The primary aim of this chapter is to discuss four estimation methods (table 12.2) as general procedures for deriving estimators with good properties. The comparison between them revolves around how optimal are the estimators they give rise to.

Table 12.2: Methods of Estimation

1. The method of Maximum Likelihood
2. The Least Squares method
3. The Moment Matching principle
4. The Parametric Method of Moments

Historically the Least Squares method was the first to be developed in the early 1800s by Adrien-Marie Legendre (1752–1833), a French mathematician, and Gauss (1777–1855), a German mathematician, as a *curve-fitting method* in the context of *the theory of errors*; see Stigler (1986), Gorroochurn (2016). The Moment Matching principle arose in the 19th century as a result of a confusion between the probability moments associated with distribution functions and sample moments as functions of the data. The first to point out this confusion was Fisher (1922a). The Parametric Method of Moments (PMM) is an anachronistic variation on the Karl Pearson's Method of Moments he developed in the late 19th century. The PMM method provided the backbone of Karl Pearson's approach to statistical modeling. Pearson's approach commenced *from the data to the best descriptive model* in the form of a frequency curve from Pearson's family of distributions; see Appendix 12.A. The only estimation method that was developed in the context of modern model-based frequentist inference is the Maximum Likelihood method proposed by Fisher (1921); see Stigler (2005). In contrast to Pearson's method of moments, the PPM is a model-based procedure, where the statistical model is prespecified.

A bird's eye view of the chapter. In section 2 we discuss the Maximum Likelihood (ML) method as a prelude to the other estimation methods to be used for comparison purposes. Section 3 introduces the least-squares method, first as a mathematical approximation method and then as a proper estimation method.Section 4 discusses the moment matching principle where the unknown parameters are estimated by equating the distribution with the sample moments. Section 5 discusses briefly Pearson's method of moments with a view to contrast it with the parametric method of moments, an adaptation of the original method for the current model-based approach to statistical inference.

2 The Maximum Likelihood Method

2.1 The Likelihood function

In contrast to the other methods of estimation, Maximum Likelihood (ML) was specifically developed for the modern model-based approach to statistical inference as framed by Fisher (1912; 1922a; 1925b). This approach turns the Karl Pearson procedure from data to histograms and frequency curves (Appendix 12.A), on its head by viewing the data $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$ as a typical realization of the sample $\mathbf{X}{:=}(X_1, X_2, ..., X_n)$ from a prespecified stochastic generating mechanism, we call a *statistical model*:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta{\subset}\mathbb{R}^m\},\ \mathbf{x}{\in}\mathbb{R}^n_X,\ m{<}n. \tag{1}$$

The probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ are encapsulated by the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}^n_X$, *the joint distribution of the random variables making up the sample*.

The cornerstone of the ML method is the concept of the *likelihood function* (Fisher, 1921), defined by:

$$L(\boldsymbol{\theta};\mathbf{x}_0) \propto f(\mathbf{x}_0;\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\Theta,$$

where $\propto$ reads 'proportional to'. In light of viewing the statistical model as the stochastic mechanism that generated $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$, it seems intuitively obvious to evaluate $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}^n_X$, at $\mathbf{X}{=}\mathbf{x}_0$, and pose the reverse question:

► how likely does $f(\mathbf{x}_0;\boldsymbol{\theta})$ render the different values of $\boldsymbol{\theta}$ in Θ to have been the 'true' value $\boldsymbol{\theta}^*$? Recall that '$\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$' is a shorthand for saying that 'data $\mathbf{x}_0$ constitute a typical realization of the sample $\mathbf{X}$ with distribution $f(\mathbf{x};\boldsymbol{\theta}^*),\ \mathbf{x}{\in}\mathbb{R}^n_X$', and the primary objective of an estimator $\widehat{\boldsymbol{\theta}}_n(\mathbf{X})$ of $\boldsymbol{\theta}$ is to pinpoint $\boldsymbol{\theta}^*$. Hence, the likelihood function yields the *likelihood* (proportional to the probability) of getting $\mathbf{x}_0$ under different values of $\boldsymbol{\theta}$.

Note that the proportionality is important for both the interpretation of likelihood values to different θ, as well as for mathematical purposes because $L(\boldsymbol{\theta};\mathbf{x}_0)$ is interpreted as a function of $\boldsymbol{\theta}\in\Theta$ but $f(\mathbf{x};\boldsymbol{\theta})$ is a function of $\mathbf{x}\in\mathbb{R}^n_X$. In practice, Θ has considerably lower dimension than $\mathbb{R}^n_X$. Hence, the LF does NOT assign probabilities to $\boldsymbol{\theta}$, but reflects the relative likelihoods for different values of $\boldsymbol{\theta}\in\Theta$ stemming from data $\mathbf{x}_0$ when viewed through the prism of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}^n_X$. Collecting all these pieces together, the frequentist approach to inference is summarized in table 12.3.

Table 12.3: The frequentist approach to statistical inference

$$\begin{tikzcd}
\boxed{\begin{array}{c}\text{Statistical model}\\ \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\end{array}} \arrow[r, Rightarrow] & \boxed{\begin{array}{c}\text{Distribution of the sample}\\ f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X\end{array}} \arrow[d, Rightarrow] \\
\boxed{\textbf{Data}\text{: } \mathbf{x}_0\mathbf{:=}(x_1,x_2,...,x_n)} \arrow[u] \arrow[r] & \boxed{\begin{array}{c}\textbf{Likelihood function}\\ L(\boldsymbol{\theta};\mathbf{x}_0),\ \boldsymbol{\theta}\in\Theta\end{array}}
\end{tikzcd}$$

The fact that the maximum likelihood method is tailor-made for the modern approach to model-based statistical inference can be seen from table 12.3, where the distribution of the sample is defined so as to encapsulate all relevant information contained in the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}^n_X$.

Example 12.1. Consider the *simple Bernoulli model*, as specified in table 12.4.

Table 12.4: The simple Bernoulli model

Statistical GM: $X_t=\theta+u_t,\ t\in\mathbb{N}:=(1,2,...,n,...)$

[1]	Bernoulli:	$X_t \backsim \mathsf{Ber}(.,.),\ x_t=\{0,1\}$,
[2]	Constant mean:	$E(X_t)=\theta$, $0\leq\theta\leq1$, for all $t\in\mathbb{N}$,
[3]	Constant variance:	$Var(X_t)=\theta(1-\theta)$, for all $t\in\mathbb{N}$,
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$ - independent process.

Assumptions [1]-[4] imply that $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, takes the form:

$$\begin{aligned} f(\mathbf{x};\theta) &\overset{[4]}{=}\prod\nolimits_{k=1}^n f_k(x_k;\theta_k)\overset{[2]\text{-}[4]}{=}\prod\nolimits_{k=1}^n f(x_k;\theta)\overset{[1]\text{-}[4]}{=}\prod\nolimits_{k=1}^n \theta^{X_k}(1-\theta)^{1-X_k}= \\ &=\theta^{\sum_{k=1}^n X_k}(1-\theta)^{\sum_{k=1}^n(1-X_k)},\ \mathbf{x}\in\{0,1\}^n, \end{aligned} \tag{2}$$

where the reduction in (2) follows from the cumulative imposition of the assumptions [1]-[4]. Hence, the Likelihood Function (LF) takes the form:

$$L(\theta;\mathbf{x}_0)\propto\theta^{\sum_{k=1}^n x_k}(1-\theta)^{\sum_{k=1}^n(1-x_k)}=\theta^y(1-\theta)^{(n-y)},\ \theta\in[0,1], \tag{3}$$

where $Y=(\sum_{k=1}^n X_k)$.

Hence, the distribution of the sample is:

$$f(\mathbf{x};\boldsymbol{\theta})=\theta^{Y}(1-\theta)^{n-Y},\ \ \mathbf{x}\in\{0,1\}^{n},$$

and the Likelihood Function (LF) is:

$$L(\theta;\mathbf{x}_0)\propto\theta^{y}(1-\theta)^{(n-y)},\ \ \forall\theta\in[0,1]. \tag{4}$$

Note that $f(\mathbf{x};\theta)$, $\mathbf{x}\in\{0,1\}^{n}$ is a discrete density function of Y, but the LF, $L(\theta;\mathbf{x}_0)$, $\theta\in[0,1]$, is a continuous function of $\theta\in[0,1]$. In general a crucial distinction is:

$$f(\mathbf{x};\boldsymbol{\theta}),\ \ \mathbf{x}\in\mathbb{R}^{n}_{X}\ \text{vs.}\ L(\boldsymbol{\theta};\mathbf{x}_0),\ \ \boldsymbol{\theta}\in\Theta.$$

In the simple Bernoulli model, Y is Binomially distributed:

$$Y=\textstyle\sum_{k=1}^{n}X_k\backsim\mathsf{Bin}\left(n\theta,\ n\theta(1-\theta);n)\right), \tag{5}$$

Example 12.1 (continued). The distribution $f(y;\theta)$, $y=1,2,..,n$, is shown in figure 4 for $n=100$, $\theta=.56$, is a one-dimensional representation of $f(\mathbf{x};\theta)$, $\mathbf{x}\in\{0,1\}^{n}$ using $f(\mathbf{x};\theta)=\theta^{Y}(1-\theta)^{n-Y}$, $y=0,1,2,...,n$. This discrete distribution in fig. 4 should be contrasted with the Likelihood Function (LF) $L(\theta;\mathbf{x}_0)=\theta^{y}(1-\theta)^{n-y}$, $\theta\in[0,1]$, (figure 5) which is a continuous and differentiable function of θ.

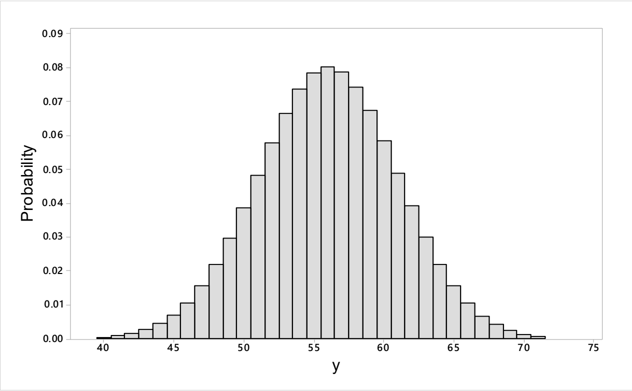

Fig. 4: $Y\backsim\mathsf{Bin}(n\theta,\ n\theta(1-\theta))$, $n=100$, $\theta=.56$

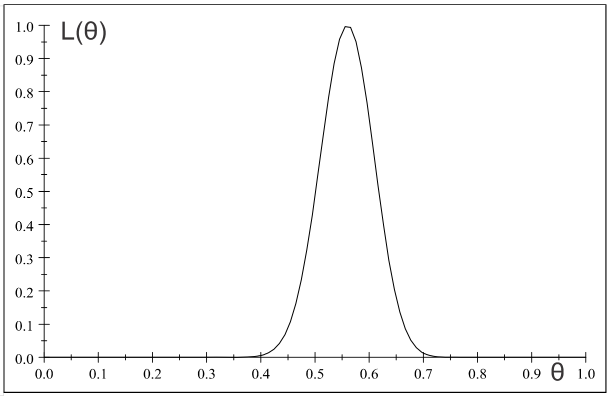

Fig. 5: $L(\theta;\mathbf{x}_0)$, $\theta\in[0,1]$, $Y=.56$

This brings out an important feature of the likelihood function that pertains to the scaling on the vertical axis. This scaling is arbitrary since one can define the Likelihood Function (LF), equivalently as:

$$L(\boldsymbol{\theta};\mathbf{x}_0)=c(\mathbf{x}_0)\cdot f(\mathbf{x}_0;\boldsymbol{\theta}),\ \ \boldsymbol{\theta}\in\Theta, \tag{6}$$

where $c(\mathbf{x}_0)$ depends only on the data $\mathbf{x}_0$ and *not* on $\boldsymbol{\theta}$. Indeed, the likelihood function graph in figure 5 has been scaled using $c(\mathbf{x}_0)=[1/L(\widehat{\theta};\mathbf{x}_0)]$, where $L(\widehat{\theta};\mathbf{x}_0)$ denotes the estimated likelihood with $\widehat{\theta}=.56$, being the Maximum Likelihood (ML) estimate of θ; see Lecture Notes 1. This renders the graph of the likelihood function easier to read as well as compare the likelihood values for different θ's. To get some idea of comparing likelihood values for different values of θ consider the following example.

Example 12.1 (continued). For the *simple Bernoulli model*, with n=100, θ=.56, let us compare the likelihood of two values of $\theta=\mathbb{P}(X=1)$ within the interval $[0,1]$, θ_1=.45 and θ_2=.62; see fig. 5. The values of the likelihood function are:

$$L(.45;\mathbf{x}_0)=(.45)^{56}(1-.45)^{44}=1.431\,7\times 10^{-31},$$
$$L(.62;\mathbf{x}_0)=(.62)^{56}(1-.62)^{44}=7.663\,2\times 10^{-31},$$

which are tiny, and thus highly vulnerable to approximation errors. Having said that, due to the presence of the arbitrary constant $c(\mathbf{x}_0)$ in (**??**), the LF can be scaled to avoid such problems. An obvious way to scale the LF is to divie by the estimated LF:

$$L(\widehat{\theta};\mathbf{x}_0)=(.56)^{56}(1-.56)^{44}=1.623\,5\times10^{-30},$$

which is also a tiny number. The scaled likelihood function $\frac{L(\theta;\mathbf{x}_0)}{L(\widehat{\theta};\mathbf{x}_0)}$, however, takes values between zero and one:

$$\tfrac{L(.45;\mathbf{x}_0)}{L(\widehat{\theta};\mathbf{x}_0)}=\tfrac{(.45)^{56}(1-.45)^{44}}{(.56)^{56}(1-.56)^{44}}=.0882,\;\; \tfrac{L(.62;\mathbf{x}_0)}{L(\widehat{\theta};\mathbf{x}_0)}=\tfrac{(.62)^{56}(1-.62)^{44}}{(.56)^{56}(1-.56)^{44}}=.472,$$

which renders the comparison of the two easier. CAUTION, however, is advised to avoid misconstruing the scaled likelihood function as assigning probabilities to $\theta\in[0,1]$, just because of the particular scaling used.

In light of the arbitrariness of the scaling factor $c(\mathbf{x}_0)$, the only meaningful measure of relative likelihood for two values of θ comes in the form of the ratio:

$$\tfrac{L(.62;\mathbf{x}_0)}{L(.45;\mathbf{x}_0)}=\tfrac{c(\mathbf{x}_0)(.62)^{56}(1-.62)^{44}}{c(\mathbf{x}_0)(.45)^{56}(1-.45)^{44}}=\tfrac{(.62)^{56}(1-.62)^{44}}{(.45)^{56}(1-.45)^{44}}=5.353,$$

since the scaling factor *cancels out*, being the same for all values $\theta\in[0,1]$. This renders the value θ=.62 more than 5 times likelier than θ=.45. Does that mean that $\mathbf{x}_0$ this provides evidence that θ=.62 is close to the θ^*, the true θ?

Not necessarily! This is because, by definition, the values of the likelihood function $L(\theta;\mathbf{x}_0)$ are dominated by the Maximum Likelihood (ML) estimate $\widehat{\theta}$=.56. Moreover, in point estimation there is no warranted inferential claim that $\widehat{\theta}$=.56 is approximately equal to θ^* due to the sampling variability associated with the ML estimator:

$$\widehat{\theta}(\mathbf{X})=\overline{X}=\tfrac{1}{n}\textstyle\sum_{i=1}^{n}X_i \backsim \mathsf{Bin}(\theta,\tfrac{\theta(1-\theta)}{n}),$$

where $\mathsf{Bin}(\theta,\frac{\theta(1-\theta)}{n})$ denotes a 'scaled' Binomial distribution with mean θ and variance $\frac{\theta(1-\theta)}{n}$; see (5). This suggests that for a particular sample realization $\mathbf{x}_0$ there is no reason to presume that $\widehat{\theta}(\mathbf{X})\simeq\theta^*$, since for an unbiased estimator $\widehat{\theta}(\mathbf{X})$ of θ^* is only its mean that has such property: $E(\widehat{\theta}(\mathbf{X}))=\theta^*$. That is, if one were to use the long-run metaphor to visualize the sampling distribution of $\widehat{\theta}(\mathbf{X})$, one would have to draw N (say $N{=}10000$) sample realizations $\mathbf{x}_i$, $i{=}1,2,...N$ and construct the empirical sampling distribution of $\widehat{\theta}(\mathbf{X})$ and evaluate its mean to be able to claim that $\widehat{E}(\widehat{\theta}(\mathbf{X}))\simeq\theta^*$.

In contrast to a point estimator, both confidence intervals and hypothesis testing account for this sampling variability by using statistics of the form:

$$\widehat{\theta}(\mathbf{X})\pm c_{\frac{\alpha}{2}}\tfrac{\sqrt{\widehat{\theta}(\mathbf{X})(1-\widehat{\theta}(\mathbf{X}))}}{\sqrt{n}},\ \ \tfrac{\sqrt{n}(\widehat{\theta}(\mathbf{X})-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}.$$

Maximum Likelihood method and learning from data. In the case of a simple statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ whose distribution of the sample $f(\mathbf{x};\theta)$ is a one-to-one function of θ, it can be shown that, under certain regularity conditions, $\ln L_n(\theta;\mathbf{x}){=}\sum_{t=1}^{n}\ln f(x_t;\theta)$ attains its maximum at the true value θ^* in the sense that:

$$\mathbb{P}(\lim_{n\rightarrow\infty}\left[\ln(\tfrac{\frac{1}{n}L_n(\theta^*;\mathbf{x})}{\frac{1}{n}L_n(\theta;\mathbf{x})})\right]>0){=}1,\ \forall\theta{\in}\left[\Theta{-}\{\theta^*\}\right]. \tag{7}$$

This result follows directly from applying the SLLN (chapter 9) to $\frac{1}{n}\sum_{t=1}^{n}\ln f(x_t;\theta)$. The result in (7) can be extended to statistical models beyond IID samples.

2.2 Maximum Likelihood estimators

In order to be able to derive results of some generality we confine the discussion to *regular statistical models* which satisfy the restrictions **R1-R4** (chapter 11) that ensure the existence of the Fisher information. The extend to which these regularity conditions restrict the probability models; see Gourieroux and Monfort (1995) for more details.

For simplicity of exposition and notational convenience, let us consider the case where θ is a scalar. Estimating by maximum likelihood amounts to finding that particular value $\widehat{\theta}_{ML}{=}h(\mathbf{x})$ that maximizes the likelihood function:

$$L(\widehat{\theta}_{ML};\mathbf{x}_0){=}\max_{\theta\in\Theta}L(\theta;\mathbf{x}_0)\Longleftrightarrow\widehat{\theta}_{ML}{=}\arg[\max_{\theta\in\Theta}L(\theta;\mathbf{x}_0)], \tag{8}$$

but then turn it into a statistic (a function of $\mathbf{X}$). That is, $\widehat{\theta}_{ML}(\mathbf{X}){=}h(\mathbf{X})$ is the *Maximum Likelihood Estimator* (MLE) of θ and $\widehat{\theta}_{ML}(\mathbf{x}_0){=}h(\mathbf{x}_0)$ is the ML *estimate*. There are several things to note about MLE in (8):

(a) the MLE $\widehat{\theta}_{ML}(\mathbf{X})$ may *not exist*,

(b) the MLE $\widehat{\theta}_{ML}(\mathbf{X})$ may *not* be *unique*,

(c) the MLE may *not* have a *closed form* expression $\widehat{\theta}_{ML}(\mathbf{X}){=}h(\mathbf{X})$.

Example 12.3. Consider the *simple Uniform model*:

$$X_k\backsim\mathsf{UIID}(\theta{-}\tfrac{1}{2},\theta{+}\tfrac{1}{2}),\ \theta{\in}\mathbb{R},\ t{=}1,2,...,n,...,$$

whose density function is $f(x;\theta){=}1,\ x{\in}[\theta{-}\frac{1}{2},\ \theta{+}\frac{1}{2}]$, and:

$$E(X){=}\int_{\theta-.5}^{\theta+.5}xdx{=}\theta,\ Var(X){=}\int_{\theta-.5}^{\theta+.5}(x{-}\theta)^2dx{=}\tfrac{1}{12}.$$

These assumptions imply that the distribution of the sample is:

$$f(\mathbf{x};\theta)\ \ ={\textstyle\prod_{k=1}^{n}}1{=}1,\ \mathbf{x}{\in}[\theta{-}\tfrac{1}{2},\ \theta{+}\tfrac{1}{2}]^n,$$

Given that $\left[\theta-\frac{1}{2}\leq(x_1,x_2,...x_n)\leq\theta+\frac{1}{2}\right]$ it implies that the set of possible values of θ is:

$$\left(x_{[n]}-\tfrac{1}{2}\leq\theta\leq x_{[1]}+\tfrac{1}{2}\right), \tag{9}$$

where $x_{[1]}{=}\min(x_1,x_2,...,x_n)$ and $x_{[n]}{=}\max(x_1,x_2,...,x_n)$. The *likelihood function* is:

$$L(\theta;\mathbf{x}){=}1\ \ \text{if }\theta-\tfrac{1}{2}\leq x_{[1]}\text{ and }x_{[n]}\leq\theta+\tfrac{1}{2}, \tag{10}$$

and thus the MLE of θ is non-unique since it could be any value in (9). Despite its non-uniqueness, the preferred ML estimator is the *midrange* of X, $\widehat{\theta}_{ML}(\mathbf{X}){=}\frac{X_{[n]}+X_{[1]}}{2}$, because it is unbiased and consistent since:

$$E(\widehat{\theta}_{ML}(\mathbf{X})){=}\theta,\ \ \ Var(\widehat{\theta}_{ML}(\mathbf{X})){=}\tfrac{1}{2(n+1)(n+2)}.$$

Two things are worth noting about this example. First, the source of the non-uniqueness of the MLE is that fact that the above statistical model is non-regular since the support of $f(\mathbf{x};\theta)$ depends on θ; it violates condition R2 in table 11.4. Hence, the Cramer-Rao lower bound cannot be used to evaluate the full efficiency of estimators. Second, the midrange estimator is relatively more efficient than the sample mean $\widehat{\theta}_n(\mathbf{X}){=}\frac{1}{n}\sum_{k=1}^{n}X_k$ since $E(\widehat{\theta}_n(\mathbf{X}))$ and $Var(\widehat{\theta}_n(\mathbf{X})){=}\frac{1}{12n}{>}Var(\widehat{\theta}_{ML}(\mathbf{X}))$, for any $n{>}1$.

Despite the few pathological cases for which existence and uniqueness of the MLE $\widehat{\theta}$ is not guaranteed (Gourieroux and Monfort, 1995), in practice $\widehat{\theta}_{ML}(\mathbf{X})$ exists and is unique in the overwhelming number of cases of interest. In order to reduce the pathological cases for which $\widehat{\theta}_{ML}(\mathbf{X})$ may not exist we often restrict our discussion to cases where two additional restrictions to **R1-R4** in table 11.4 are imposed on $\mathcal{M}_\theta(\mathbf{x})$ (table 12.5).

Table 12.5: Regularity for $\mathcal{M}_\theta(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta\},\ \mathbf{x}{\in}\mathbb{R}_X^n$	
(**R5**)	$L(.;\mathbf{x}_0)$: $\Theta\rightarrow[0,\infty)$, is *continuous* at all points $\boldsymbol{\theta}{\in}\Theta$.
(**R6)**	For all values $\boldsymbol{\theta}_1{\neq}\boldsymbol{\theta}_2$ in Θ, $f(\mathbf{x};\boldsymbol{\theta}_1){\neq}f(\mathbf{x};\boldsymbol{\theta}_2),\ \mathbf{x}{\in}\mathbb{R}_X^n$.

Condition (**R5**) ensures that $L(\boldsymbol{\theta};\mathbf{x})$ is smooth enough to locate its maximum, and (**R6**) ensures that $\boldsymbol{\theta}$ is *identifiable* and thus unique. When the LF is also differentiable, one can locate the maximum by solving the first-order conditions:

$$\tfrac{dL(\theta;\mathbf{x})}{d\theta}{=}g(\widehat{\theta}_{ML}){=}0,\text{ given that }\left.\tfrac{d^2L(\theta;\mathbf{x})}{d^2\theta}\right|_{\theta=\widehat{\theta}_{ML}}<0.$$

In practice, it is often easier to maximize the log likelihood function instead, because they have the same maximum (the logarithm is a monotonic transformation):

$$\tfrac{d\ln L(\theta;\mathbf{x})}{d\theta}=\ell(\widehat{\theta}_{ML})=\left(\tfrac{1}{L}\right)\tfrac{dL(\theta;\mathbf{x})}{d\theta}=\left(\tfrac{1}{L}\right)g(\widehat{\theta}_{ML})=0,\ \text{ given } L\neq 0.$$

Example 12.4. For the *simple Bernoulli model* (table 12.4), the log-likelihood function is:

$$\ln L(\mathbf{x};\theta)=(\textstyle\sum_{i=1}^{n}X_i)\ln\theta+(\sum_{i=1}^{n}[1-X_i])\ln(1-\theta)=Y\ln\theta+(n-Y)\ln(1-\theta), \qquad (11)$$

where $Y=\sum_{k=1}^{n}X_k$. Solving the first order condition:

$$\tfrac{d\ln L(\mathbf{x};\theta)}{d\theta}=(\tfrac{1}{\theta})Y-(\tfrac{1}{1-\theta})(n-Y)=0 \Rightarrow Y(1-\theta)=\theta(n-Y) \Rightarrow n\theta=Y,$$

for θ yields the MLE: $\widehat{\theta}_{ML}=\frac{1}{n}\sum_{k=1}^{n}X_k$ of θ, which is just the sample mean. To ensure that $\widehat{\theta}_{ML}$ is a maximum of $\ln L(\mathbf{x};\theta)$, we need to check that $\frac{d^2\ln L(\mathbf{x};\theta)}{d\theta^2}\Big|_{\theta=\widehat{\theta}_{ML}}<0$. Note that when $\frac{d^2\ln L(\mathbf{x};\theta)}{d\theta^2}\Big|_{\theta=\widehat{\theta}_{ML}}>0$, $\widehat{\theta}_{ML}$ is a minimum. The second order conditions confirm that $\widehat{\theta}_{ML}$ is a maximum since:

$$\tfrac{d^2\ln L(\mathbf{x};\theta)}{d\theta^2}\Big|_{\theta=\widehat{\theta}_{ML}}=-(\tfrac{1}{\theta^2})Y-(\tfrac{1}{1-\theta})^2(n-Y)=-\tfrac{\left(n\theta^2+2Y\theta-2Y\theta^2-Y\right)}{\theta^2(\theta-1)^2}\Big|_{\theta=\widehat{\theta}_{ML}}=-\tfrac{n^3}{Y(n-Y)}<0,$$

because both the numerator (n^3) and denominator ($Y(n-Y),\ n>Y$) are positive.

To avoid the misleading impression that the Maximum Likelihood estimator for simple statistical models can always be derived using differentiation, compare example 12.4 with the following.

Example 12.5. Consider the *simple Laplace model* (table 12.6) whose density function is:

$$f(x;\theta)=\tfrac{1}{2}\exp\{-|x-\theta|\},\ \theta\in\mathbb{R},\ x\in\mathbb{R}.$$

Table 12.6: The simple Laplace model

Statistical GM: $X_t=\theta+u_t,\ t\in\mathbb{N}:=(1,2,...,n,...)$		
[1]	Laplace:	$X_t\backsim\mathsf{Lap}(.,.)$,
[2]	Constant mean:	$E(X_t)=\theta$, for all $t\in\mathbb{N}$,
[3]	Constant variance:	$Var(X_t)=2$, for all $t\in\mathbb{N}$,
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$ - independent process.

The distribution of the sample takes the form:

$$f(\mathbf{x};\theta)\ =\prod_{t=1}^{n}\tfrac{1}{2}\exp\{-|x_t-\theta|\}=(\tfrac{1}{2})^n\exp\{-\textstyle\sum_{t=1}^{n}|x_t-\theta|\},\ \mathbf{x}\in\mathbb{R}^n,$$

and thus the log-likelihood function is:

$$\ln L(\theta;\mathbf{x}){=}\text{const}{-}n\ln(2){-}\textstyle\sum_{t=1}^{n}|x_t{-}\theta|,\ \ \theta{\in}\mathbb{R}.$$

Since $\ln L(\theta;\mathbf{x})$ is *non-differentiable* one needs to use alternative methods to derive the maximum of this function. In this case maximizing $\ln L(\theta;\mathbf{x})$ with respect to θ is equivalent to minimizing the function:

$$\ell(\theta){=}\textstyle\sum_{t=1}^{n}|x_t-\theta|,$$

which (in the case of n odd) gives rise to the sample median:

$$\widehat{\theta}_{ML}{=}\text{median}(X_1,X_2,,...,X_n).$$

2.3 The Score function

The quantity $\frac{d}{d\theta}\ln L(\theta;\mathbf{x})$ has been encountered in chapter 11 in relation to full efficiency, but at that point we used the log of the distribution of the sample $\ln f(\mathbf{x};\theta)$ instead of $\ln L(\theta;\mathbf{x})$ to define the *Fisher information*:

$$\mathcal{I}_n(\theta){:=}E\left\{\left(\tfrac{\partial\ln f(\mathbf{x};\theta)}{\partial\theta}\right)^2\right\}. \tag{12}$$

In terms of the log-likelihood function the *Cramer-Rao* (C-R) *lower bound* takes the form:

$$Var(\widehat{\theta})\geq\left[E\left\{\left(\tfrac{\partial\ln L(\theta;\mathbf{x})}{\partial\theta}\right)^2\right\}\right]^{-1}, \tag{13}$$

for *any unbiased estimator* $\widehat{\theta}$ of θ.

A SHORT DIGRESSION. From a mathematical perspective:

$$E\left\{\left(\tfrac{\partial\ln f(\mathbf{x};\theta)}{\partial\theta}\right)^2\right\}{=}E\left\{\left(\tfrac{\partial\ln L(\theta;\mathbf{x})}{\partial\theta}\right)^2\right\},$$

but the question is which choice between $\ln f(\mathbf{x};\theta)$ and $\ln L(\theta;\mathbf{x})$ provides a correct way to express the C-R bound in a probabilistically meaningful way. It turns out that neither of these concepts is entirely correct for that. Using $\ln L(\theta;\mathbf{x})$ renders taking the derivative with respect to θ meaningful since it is a function of $\theta{\in}\Theta$, in contrast to $f(\mathbf{x};\theta)$ that is a function of $\mathbf{x}{\in}\mathbb{R}_X^n$ with θ assumed fixed at a particular value. On the other hand, the expectation $E(.)$ is always with respect to $\mathbf{x}{\in}\mathbb{R}_X^n$ and that makes sense only with respect to $f(\mathbf{x};\theta)$. Hence, what is implicitly assumed in the derivation of the C-R bound is a more general real-valued function with two arguments:

$$g(.,.)\text{: }(\mathbb{R}_X^n\times\Theta)\rightarrow\mathbb{R},$$

such that: (i) for a given $\mathbf{x}{=}\mathbf{x}_0$, $g(\mathbf{x}_0;\theta)\propto L(\theta;\mathbf{x}_0),\ \theta{\in}\Theta$, and (ii) for a fixed θ, say $\theta{=}\theta^*,\ g(\mathbf{x};\theta){=}f(\mathbf{x};\theta^*),\ \mathbf{x}{\in}\mathbb{R}_X^n$.

The first derivative of the log-likelihood function, when interpreted as a function of the sample $\mathbf{X}$, defines:

$$\textbf{the score function}\text{: } \mathfrak{s}(\theta;\mathbf{x}){=}\tfrac{d}{d\theta}\ln L(\theta;\mathbf{x}),\ \forall\mathbf{x}{\in}\mathbb{R}_X^n,$$

that satisfies the properties in table 12.7.

Table 12.7: Score function: Properties

(**Sc1**)	$E[\mathfrak{s}(\theta;\mathbf{X})]{=}0,$
(**Sc2)**	$Var[\mathfrak{s}(\theta;\mathbf{X})]{=}E[s(\theta;\mathbf{X})]^2{=}E\left(-\frac{d^2}{d\theta^2}\ln L(\theta;\mathbf{X})\right){:=}\mathcal{I}_n(\theta).$

That is, the Fisher information is the variance of the score function. As shown in the previous chapter, an unbiased estimator $\widehat{\theta}_n(\mathbf{X})$ of θ achieves the Cramer-Rao (C-R) lower bound if and only if $(\widehat{\theta}_n(\mathbf{X}){-}\theta)$ can be expressed in the form:

$$(\widehat{\theta}_n(\mathbf{X})-\theta){=}h(\theta){\cdot}\mathfrak{s}(\theta;\mathbf{X}),\text{ for some function } h(\theta).$$

Example 12.6. In the case of the *Bernoulli model* the score function is:

$$\mathfrak{s}(\theta;\mathbf{X}){:=}\tfrac{d}{d\theta}\ln L(\theta;\mathbf{X}){=}\left(\tfrac{1}{\theta}Y-(\tfrac{1}{1-\theta})(n{-}Y)\right)\Rightarrow$$

$$\left[\tfrac{\theta(1-\theta)}{n}\right]\mathfrak{s}(\theta;\mathbf{X}){=}\tfrac{1}{n}\left(Y{-}n\theta\right){=}(\widehat{\theta}_{ML}{-}\theta)\ \Rightarrow\ (\widehat{\theta}_{ML}{-}\theta){=}\left[\tfrac{\theta(1-\theta)}{n}\right]\mathfrak{s}(\theta;\mathbf{X}),$$

which implies that $\widehat{\theta}_{ML}{=}\frac{1}{n}\sum_{i=1}^n X_i$ achieves the C-R lower bound:

$$Var(\widehat{\theta}_{ML}){=}\textsf{C-R}(\theta){=}\tfrac{\theta(1-\theta)}{n},$$

confirming the result in example 11.15.

Example 12.7. Consider the simple *Exponential model* in table 12.8.

Table 12.8: The simple Exponential model

Statistical GM:		$X_t{=}\theta+u_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$
[1]	Exponential:	$X_t \backsim \textsf{Exp}(.,.),\ x_t{\in}\mathbb{R}_+,$
[2]	Constant mean:	$E(X_t){=}\theta,\ \theta{\in}\mathbb{R},\ \forall t{\in}\mathbb{N},$
[3]	Constant variance:	$Var(X_t){=}\theta^2,\ \forall t{\in}\mathbb{N},$
[4]	Independence:	$\{X_t,\ t{\in}\mathbb{N}\}$-independent process.

Assumptions [1]-[4] imply that $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}_X^n$, takes the form:

$$\begin{aligned} f(\mathbf{x};\theta)\ &\overset{[4]}{=}\prod_{k=1}^n f_k(x_k;\theta_k)\overset{[2]\text{-}[4]}{=}\prod_{k=1}^n f(x_t;\theta){=}\\ &\overset{[1]\text{-}[4]}{=}\prod_{k=1}^n \tfrac{1}{\theta}\exp\left\{-\tfrac{x_k}{\theta}\right\}=\left(\tfrac{1}{\theta}\right)^n\exp\left\{-\tfrac{1}{\theta}\sum_{k=1}^n x_k\right\},\ \mathbf{x}{\in}\mathbb{R}_+^n,\end{aligned}$$

and thus the log-likelihood function is:

$$\ln L(\theta;\mathbf{x}){=}-n\ln(\theta)-\tfrac{1}{\theta}\textstyle\sum_{k=1}^{n}x_k.$$

$$\tfrac{d}{d\theta}\ln L(\theta;\mathbf{x}){=}-\tfrac{n}{\theta}{+}\tfrac{1}{\theta^2}\textstyle\sum_{k=1}^{n}x_k{=}0\Rightarrow\ \ \widehat{\theta}_{ML}{=}\tfrac{1}{n}\textstyle\sum_{k=1}^{n}X_k.$$

The second-order condition:

$$\tfrac{d^2}{d\theta^2}\ln L(\theta;\mathbf{x})\Big|_{\theta=\widehat{\theta}_{ML}}=\tfrac{n}{\theta^2}{-}\tfrac{2}{\theta^3}\textstyle\sum_{k=1}^{n}x_k\Big|_{\theta=\widehat{\theta}_{ML}}=-\tfrac{n}{\widehat{\theta}_{ML}^2}{<}0,$$

ensures that $\ln L(\widehat{\theta};\mathbf{x})$ is a maximum and not a minimum or a point of inflection. Using the second derivative of the log-likelihood function we can derive the Fisher information:

$$\mathcal{I}_n(\theta){:=}E\left(-\tfrac{d^2}{d\theta^2}\ln L(\theta;\mathbf{x})\right){=}\tfrac{n}{\theta^2}.$$

The above results suggest that the ML estimator $\widehat{\theta}_{ML}{=}\frac{1}{n}\sum_{k=1}^{n}X_k$ is both unbiased and fully efficient (verify!).

2.4 Two-parameter statistical model

In the case where $\boldsymbol{\theta}$ contains more than one parameter, say $\boldsymbol{\theta}{:=}(\theta_1,\theta_2)$, the first-order conditions for the MLEs take the form of a system of equations:

$$\tfrac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_1}{=}0,\ \tfrac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_2}{=}0,$$

which need to be solved simultaneously in order to derive the MLEs $\widehat{\boldsymbol{\theta}}_{ML}(\mathbf{X})$. Moreover, the second order conditions for a maximum are more involved that the one-parameter case since they involve three restrictions:

$$\text{(i)}\det\begin{pmatrix}\frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_1^2} & \frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_1\partial\theta_2}\\ \frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_2\partial\theta_1} & \frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_2^2}\end{pmatrix}\Bigg|_{\widehat{\boldsymbol{\theta}}_{ML}}{>}0,\ \text{(ii)}\frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_1^2}\Big|_{\widehat{\boldsymbol{\theta}}_{ML}}{<}0\ \text{and(iii)}\ \frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_2^2}\Big|_{\widehat{\boldsymbol{\theta}}_{ML}}{<}0.$$

Note that when (ii) and (iii) are positive then the optimum is a minimum.

The Fisher information matrix is defined by:

$$\mathcal{I}_n(\boldsymbol{\theta}){=}E\left(\tfrac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\boldsymbol{\theta}}\tfrac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\boldsymbol{\theta}^\top}\right){=}E\left(-\tfrac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right){=}Cov\left(\tfrac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\boldsymbol{\theta}}\right).$$

Example 12.8. Consider the *simple Normal model* in table 12.9.

Table 12.9: The simple Normal model

Statistical GM:		$X_t{=}\mu+u_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$
[1]	Normal:	$X_t\backsim\mathsf{N}(.,.),\ x_t{\in}\mathbb{R},$
[2]	Constant mean:	$E(X_t){=}\mu,\ \mu{\in}\mathbb{R},\ \forall t{\in}\mathbb{N},$
[3]	Constant variance:	$Var(X_t){=}\sigma^2,\ \forall t{\in}\mathbb{N},$
[4]	Independence:	$\{X_t,\ t{\in}\mathbb{N}\}$-independent process.

Assumptions [1]-[4] imply that $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, takes the form:

$$f(\mathbf{x};\boldsymbol{\theta}) \overset{[4]}{=} \prod\nolimits_{k=1}^n f_k(x_k;\boldsymbol{\theta}_k) \overset{[2]\text{-}[4]}{=} \prod\nolimits_{k=1}^n f(x_t;\boldsymbol{\theta})=$$

$$\overset{[1]\text{-}[4]}{=} \prod\nolimits_{k=1}^n \tfrac{1}{\sqrt{2\pi\sigma^2}}\exp(-\tfrac{(x_k-\mu)^2}{2\sigma^2})=(\tfrac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left\{\tfrac{1}{2\sigma^2}\sum\nolimits_{k=1}^n(x_k-\mu)^2\right\},\ \mathbf{x}\in\mathbb{R}^n, \tag{14}$$

Hence, the log-likelihood function is:

$$\ln L(\mu,\sigma^2;\mathbf{x})=\text{const.}-\tfrac{n}{2}\ln\sigma^2-\tfrac{1}{2\sigma^2}\sum\nolimits_{k=1}^n(x_k-\mu)^2.$$

Hence, we can derive the MLEs of μ and σ^2 via the first-order conditions:

$$\frac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\mu}=-\tfrac{1}{2\sigma^2}(-2)\sum_{k=1}^n(x_k-\mu)=0,\quad \frac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\sigma^2}=-\tfrac{n}{2\sigma^2}+\tfrac{1}{2\sigma^4}\sum_{k=1}^n(x_k-\mu)^2=0.$$

Solving these for μ and σ^2 yields:

$$\widehat{\mu}_{ML}=\tfrac{1}{n}\sum\nolimits_{k=1}^n X_k,\quad \widehat{\sigma}^2_{ML}=\tfrac{1}{n}\sum\nolimits_{k=1}^n(X_k-\widehat{\mu}_{ML})^2.$$

Again, the MLEs coincide with the estimators suggested by the other three methods.

$\ln L(\widehat{\boldsymbol{\theta}};\mathbf{x})$ for $\widehat{\boldsymbol{\theta}}:=(\widehat{\mu},\hat{\sigma}^2)$ is indeed a maximum since the second derivatives at $\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{ML}$ take the following signs:

$$\left.\frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\mu^2}\right|_{\widehat{\boldsymbol{\theta}}_{ML}}=-\left(\tfrac{n}{\sigma^2}\right)\Big|_{\widehat{\boldsymbol{\theta}}_{ML}}=-\tfrac{n}{\hat{\sigma}^2}<0,\quad \left.\frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\sigma^2\partial\mu}\right|_{\widehat{\boldsymbol{\theta}}_{ML}}=-\tfrac{1}{\sigma^4}\left.\sum_{k=1}^n(x_k-\mu)\right|_{\widehat{\boldsymbol{\theta}}_{ML}}=0$$

$$\left.\frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\sigma^4}\right|_{\widehat{\boldsymbol{\theta}}_{ML}}=\tfrac{n}{2\sigma^4}-\tfrac{1}{\sigma^6}\left.\sum_{k=1}^n(x_k-\mu)^2\right|_{\widehat{\boldsymbol{\theta}}_{ML}}=-\tfrac{n^2}{\hat{\sigma}^6}<0,$$

$$\left.\left(\frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\mu^2}\right)\left(\frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\sigma^4}\right)-\left(\frac{\partial^2\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\sigma^2\partial\mu}\right)\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{ML}}>0.$$

The second derivatives and their expected values for the simple Normal model were derived in section 11.6 and yielded the following Fisher Information matrix and the C-R lower bounds for any unbiased estimators of μ and σ^2:

$$\mathcal{I}_n(\boldsymbol{\theta})=\begin{pmatrix}\frac{n}{\sigma^2}, & 0\\ 0, & \frac{n}{2\sigma^4}\end{pmatrix},\ \text{(a) C-R}(\mu)=\tfrac{\sigma^2}{n},\ \text{(b) C-R}(\sigma^2)=\tfrac{2\sigma^4}{n}.$$

In addition, the sampling distributions of the MLEs take the form (section 11.6):

$$\text{(i) }\widehat{\mu}_{ML}\backsim\mathsf{N}(\mu,\tfrac{\sigma^2}{n}),\quad \text{(ii) }(\tfrac{n\widehat{\sigma}^2_{ML}}{\sigma^2})\backsim\chi^2(n-1). \tag{15}$$

Hence, $\widehat{\mu}_{ML}$ is an unbiased, fully efficient, sufficient, consistent, asymptotically Normal, asymptotically efficient estimator of μ, but $\widehat{\sigma}^2_{ML}$ is biased, sufficient, consistent, asymptotically Normal and asymptotically efficient.

Observed information matrix. At this point it is important to digress for a few seconds in order to introduce a concept sometimes used in place of the Fisher information matrix, the *observed information matrix*:

$$\mathcal{J}_n(\boldsymbol{\theta}){=}-\left(\tfrac{\partial^2 \ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right).$$

In the case of the simple Normal model this matrix takes the form:

$$\mathcal{J}_n(\boldsymbol{\theta}){=}\left(\begin{array}{cc} \frac{n}{\sigma^2}, & \frac{1}{\sigma^4}\sum_{k=1}^n(x_k{-}\mu) \\ \frac{1}{\sigma^4}\sum_{k=1}^n(x_k{-}\mu), & -\frac{n}{2\sigma^4}+\frac{1}{\sigma^6}\sum_{k=1}^n(x_k{-}\mu)^2 \end{array}\right).$$

As we can see $\mathcal{J}_n(\boldsymbol{\theta})$ is much easier to evaluate because no expectations need to be taken. Efron and Hinkley (1978) argued that $\mathcal{J}_n(\boldsymbol{\theta})$ should be used in preference to $\mathcal{I}_n(\boldsymbol{\theta})$ when using a Normal approximation for the distribution of a ML estimator because it provides a better approximation to the finite sampling distribution.

Before the reader jumps to the erroneous conclusion that all ML estimators have closed form expressions $\widehat{\theta}_{ML}(\mathbf{X}){=}h(\mathbf{X})$ that usually coincide with the sample moments, let us consider the following example.

Example 12.9. Consider the simple *Gamma model* (table 12.10), with a density function:

$$f(x;\boldsymbol{\theta}){=}\tfrac{\beta^{-1}}{\Gamma[\alpha]}\left(\tfrac{x}{\beta}\right)^{\alpha-1}\exp\left\{-\left(\tfrac{x}{\beta}\right)\right\},\ \boldsymbol{\theta}{:=}(\alpha,\beta){\in}\mathbb{R}_+^2,\ x{\in}\mathbb{R}_+.$$

where $\Gamma[\beta]$ is the Gamma function (see Appendix 3.A).

Table 12.10: The simple Gamma model

Statistical GM:		$X_t{=}\alpha\beta+u_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$
[1]	Gamma:	$X_t \backsim \mathsf{G}(.,.),\ x_t{\in}\mathbb{R}_+,$
[2]	Constant mean:	$E(X_t){=}\alpha\beta,\ (\alpha,\beta){\in}\mathbb{R}_+^2,\ \forall t{\in}\mathbb{N},$
[3]	Constant variance:	$Var(X_t){=}\alpha\beta^2,\ \forall t{\in}\mathbb{N},$
[4]	Independence:	$\{X_t,\ t{\in}\mathbb{N}\}$-independent process.

Assumptions [1]-[4] imply that $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}_X^n$, takes the form:

$$f(\mathbf{x};\boldsymbol{\theta})\ \overset{[4]}{=}\prod_{k=1}^n f_k(x_k;\boldsymbol{\theta}_k)\overset{[2]\text{-}[4]}{=}\prod_{k=1}^n f(x_t;\boldsymbol{\theta}){=}$$

$$\overset{[1]\text{-}[4]}{=}\prod_{k=1}^n\left(\tfrac{\beta^{-\alpha}x_k^{\alpha-1}}{\Gamma[\alpha]}\right)e^{\{-(\frac{x_k}{\beta})\}}{=}\left(\tfrac{\beta^{-\alpha}}{\Gamma[\alpha]}\right)^n\prod_{k=1}^n(x_k^{\alpha-1})\exp\left\{\text{-}\sum_{k=1}^n\tfrac{x_k}{\beta}\right\},\ \mathbf{x}{\in}\mathbb{R}_+^n.$$

The log-likelihood function, with $\boldsymbol{\theta}{:=}(\alpha,\beta)$, takes the form:

$$\ln L(\boldsymbol{\theta};\mathbf{x}){=}\ \ \text{const}{-}n\ln\Gamma[\alpha]{-}n\alpha\ln\beta+(\alpha{-}1)\sum_{k=1}^n\ln x_k{-}\sum_{k=1}^n\tfrac{x_k}{\beta},$$

The first order conditions yield:

$$\frac{\partial \ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\beta}{=}-\frac{n\alpha}{\beta}{+}\frac{1}{\beta^2}\sum_{k=1}^{n}x_k{=}0,\quad \frac{\partial \ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\alpha}{=}-n\psi'[\alpha]{-}n\ln\beta{+}\sum_{k=1}^{n}\ln x_k{=}0,$$

where $\psi'[z]{:=}\frac{d}{dz}\ln\Gamma[z]$, is known as the digamma function (see Abramowitz and Stegum, 1970). Solving the first equation yields: $\widehat{\beta}_{ML}{=}\frac{\overline{X}_n}{\widehat{\alpha}}$, where $\overline{X}_n{=}\frac{1}{n}\sum_{k=1}^{n}X_k$. Substituting this into the second equation yields:

$$\ell(\alpha){=}-n\psi'[\alpha]-n\ln(\tfrac{\overline{X}_n}{\widehat{\alpha}})+\sum_{k=1}^{n}\ln X_k{=}0, \tag{16}$$

which cannot be solved explicitly for $\widehat{\alpha}$. It can, however, be solved numerically.

2.4.1 Numerical evaluation

As in the case of the simple Gamma model discussed above, solving the first-order conditions for MLEs one will need to use numerical methods because no closed form expression $\widehat{\theta}{=}h(\mathbf{X})$ can be derived from 16. In its simplest form the numerical evaluation amounts to solving numerically the score function equation, $\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}{=}0$, being a non-linear function of θ. There are several numerical algorithms which can be used to solve this problem which are appropriate for different circumstances. One of the simplest and most widely used algorithms is the *Newton Raphson* which we can describe briefly as follows in the case of finding the value of θ in Θ that minimizes the function $\ell(\theta){=}-(\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta})$ by ensuring that $\frac{d\ell(\theta)}{d\theta}{:=}\ell'(\theta){\simeq}0$. Note that maximizing $h(\theta)$ is equivalent to minimizing $-h(\theta)$.

Step 1. Choose an initial (tentative) best guess 'value': θ_0.

Step 2. The Newton-Raphson algorithm improves this value θ_0 by choosing:

$$\theta_1{=}\theta_0-[\ell'(\theta_0)]^{-1}\,\ell(\theta_0),\ \text{where}\ \ell'(\theta_0){=}\frac{d\ell(\theta_0)}{d\theta}.$$

This is based on taking a first-order Taylor approximation:

$$\ell(\theta_1)\simeq\ell(\theta_0)+(\theta_1-\theta_0)\ell'(\theta_0),$$

setting it equal to zero, $\ell(\theta_1){=}0$, and solving it for θ_1. This provides a quadratic approximation of the function $\ell(\theta)$.

Step 3. Continue iterating using the algorithm:

$$\widehat{\theta}_{k+1}{=}\widehat{\theta}_k-\left[\ell'(\widehat{\theta}_k)\right]^{-1}\ell(\widehat{\theta}_k),\ k{=}1,2,...,N+1,$$

until the difference between $\widehat{\theta}_{k+1}$ and $\widehat{\theta}_k$ is less than a pre-assigned small value ϵ, say $\epsilon{=}.00001$, i.e.

$$\left|\widehat{\theta}_{N+1}-\widehat{\theta}_N\right|<\epsilon.$$

NOTE that $\left[-\ell'(\widehat{\theta}_k)\right]$ is the observed information (matrix) encountered above.

Step 4. The MLE is chosen to be the value $\widehat{\theta}_{N+1}$ for which: $\ell'(\widehat{\theta}_{N+1})\simeq 0$.

A related numerical algorithm, known as the *method of scoring*, replaces $\ell'(\widehat{\theta}_k)$ with the Fisher information $\mathcal{I}_n(\theta)$, the justification being the convergence result:

$$\tfrac{1}{n}\ell'(\widehat{\theta}_k)\overset{a.s.}{\to}\mathcal{I}_n(\theta),$$

yielding the sequential iteration scheme:

$$\widehat{\theta}_{k+1}{=}\widehat{\theta}_k-\tfrac{1}{n}\left[\mathcal{I}_n(\widehat{\theta}_k)\right]^{-1}\ell(\widehat{\theta}_k),\ k{=}1,2,...,N+1.$$

IMPORTANT: It turns out that all one needs to do in order to achieve asymptotically efficient estimators is to use any one of the above iteration schemes for one iteration! One iteration is sufficient for asymptotic efficiency. For an extensive discussion of such numerical algorithms used in econometrics, see Gourieroux and Monfort (1995), Hendry (1995) and Davidson and McKinnon (2004).

Example 12.10. Consider the simple *Logistic (one parameter) model* (table 12.11), with a density function:

$$f(x;\theta){=}\tfrac{\exp(-(x-\theta))}{[1+\exp(-(x-\theta))]^2},\ \theta{\in}\mathbb{R},\ x{\in}\mathbb{R}.$$

Table 12.11: The simple (one-parameter) **Logistic model**

Statistical GM:		$X_t{=}\theta+u_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$
[1]	Logistic:	$X_t\backsim\mathsf{Log}(.),\ x_t{\in}\mathbb{R},$
[2]	Constant mean:	$E(X_t){=}\theta,\ \theta{\in}\mathbb{R},,\ \forall t{\in}\mathbb{N},$
[3]	Constant variance:	$Var(X_t){=}\frac{\pi^2}{3},\ \forall t{\in}\mathbb{N},$
[4]	Independence:	$\{X_t,\ t{\in}\mathbb{N}\}$-independent process.

Assumptions [1]-4] imply that $\ln L(\theta;\mathbf{x})$ and the first-order conditions are:

$$\ln L(\theta;\mathbf{x}){=}\text{-}\sum_{k=1}^{n}(x_k{-}\theta){-}2\sum_{k=1}^{n}\ln\left[1{+}e^{\text{-}(x_k-\theta)}\right],\quad \tfrac{d\ln L(\theta;\mathbf{x})}{d\theta}{=}n{-}2\sum_{k=1}^{n}\tfrac{\exp(-(x_k-\theta))}{[1+\exp(-(x_k-\theta))]}{=}0.$$

The MLEs of θ can be derived using the Newton-Raphson algorithm with:

$$\ell'(\theta){=}-2\sum_{k=1}^{n}\tfrac{\exp(x_k-\theta)}{[1+\exp(x_k-\theta)]^2},$$

and $\overline{X}_n$ as initial value for θ. For comparison purposes NOTE that:

$$\sqrt{n}(\overline{X}_n-\theta)\underset{n\to\infty}{\backsim}\mathsf{N}(0,\tfrac{\pi^2}{3}),\ \tfrac{\pi^2}{3}{=}3.2899,\quad\text{and}\quad\sqrt{n}(\widehat{\theta}_{ML}-\theta)\underset{n\to\infty}{\backsim}\mathsf{N}(0,3).$$

2.5 Properties of Maximum Likelihood Estimators

2.5.1 Finite sample properties

Maximum likelihood estimators are not unbiased in general, but instead, they are invariant with respect to well-behaved functional parameterizations, and the two properties are incompatible.

(1) Parameterization invariance

For $\phi{=}g(\theta)$ a well-behaved (Borel) function of θ,the MLE of ϕ is given by:

$$\widehat{\phi}_{ML}{=}g(\widehat{\theta}_{ML}).$$

This property is particularly useful because the *substantive* (structural) parameters of interest $\boldsymbol{\varphi}$ do not often coincide with the statistical parameters $\boldsymbol{\theta}$, and this property enables us to derive the MLEs of the former.

In view of the fact that in general: $E(\hat{\phi}_{ML}) \neq g(E(\widehat{\theta}_{ML}))$,

one can think of the bias in certain MLEs as the price to pay for the invariance property. That is, if $E(\widehat{\theta}_{ML}){=}\theta$, $E(\hat{\phi}_{ML}){\neq}\phi$ in general.

Fisher in his classic (1922a) paper emphasized the crucial importance of the parameterization invariance property, and used it to bring out a major weakness for *unbiasedness*: "... lack of bias, which ... is not invariant for functional transformation of parameters has never had the least interest for me." (Bennett, 1990, p. 196). Indeed, Fisher used this invariance property to question the claim by Bayesians that a Uniform prior is 'uninformative' about the unknown parameters; see chapter 10.

Example 12.11. For the simple *Normal model* (table 12.9) $\widehat{\mu}_{ML}$ is an unbiased estimator of μ. Assuming that the parameter of interest is μ^2, is $\widehat{\mu}^2_{ML}$ an unbiased estimator? The answer is no since:

$$\begin{aligned} E(\widehat{\mu}^2_{ML}) &= E\left(\tfrac{1}{n}\textstyle\sum_{k=1}^n X_k\right)^2 = (\tfrac{1}{n})^2\left[\textstyle\sum_{k=1}^n E(X_k^2) + \sum_{i\neq j}^n E(X_iX_j)\right] = \\ &\overset{[4]}{=} (\tfrac{1}{n})^2\left[n(\mu^2{+}\sigma^2){+}n(n{-}1)\mu^2\right] = \tfrac{1}{n^2}(n\,(\sigma^2{+}n\mu^2)) = \mu^2{+}\tfrac{\sigma^2}{n}, \end{aligned}$$

since $E(X_k^2){=}\mu^2{+}\sigma^2$ and $E(X_iX_j)\overset{[4]}{=}E(X_i){\cdot}E(X_j){=}\mu^2$.

Example 12.12: Bivariate Bernoulli model. The basic statistical model underlying cross-classified binary data is the simple *bivariate Bernoulli* model:

$$\mathbf{Z}_i \backsim \mathsf{BerIID}(\boldsymbol{\mu}(\boldsymbol{\theta}),\ \boldsymbol{\Sigma}(\boldsymbol{\theta})),\ i{=}1,2,...,n,..., \tag{17}$$

where $\mathbf{Z}_i{:=}(X_i, Y_i)^\top$, and the parameters $\boldsymbol{\theta}{:=}(\theta_1, \theta_2, \theta_3)$ define the mean and covariance:

$$\boldsymbol{\mu}(\boldsymbol{\theta}){=}(\theta_1, \theta_2)^\top,\ \boldsymbol{\Sigma}(\boldsymbol{\theta}){=}[\sigma_{ij}]_{i,j=1}^2,\ \sigma_{11}{=}\theta_1(1{-}\theta_1),\ \sigma_{22}{=}\theta_2(1{-}\theta_2),\ \sigma_{12}{=}\theta_3.$$

As shown in chapter 6, the bivariate density function is not specified in terms of the first two moments, but $f(x, y; \boldsymbol{\theta}){=}\mathbb{P}(X{=}i, Y{=}j),\ i,j{=}0,1$ (table 12.12). In this section we use the notation $\pi_{ij}=\mathbb{P}(X=i-1, Y=j-1),\ i,j=1,2,$ that can be easily extended from the 2×2 Bernoulli to a multinomial distribution for $I\times J$ contingency tables with discrete values other than 0 and 1 (Bishop et al., 1975):

$$f(x,y;\boldsymbol{\theta})=\pi_{11}^{(1-x)(1-y)}\pi_{21}^{x(1-y)}\pi_{12}^{(1-x)y}\pi_{22}^{xy},\ x=0,1,\ y=0,1,$$
$$\theta_1=\pi_{21}+\pi_{22}=\pi_{+2},\ \theta_2=\pi_{12}+\pi_{22}=\pi_{+2},\ \theta_3=\pi_{22}-\pi_{+2}\pi_{+2}. \tag{18}$$

Table 12.12: 2×2 table

$x\diagdown y$	0	1	total
0	π_{11}	π_{12}	π_{1+}
1	π_{21}	π_{22}	π_{2+}
total	π_{+1}	π_{+2}	1

Note that the specification in table 12.12 imposes the assumptions in (17) and obviates the relevant ordering $i=1,2,...,n,$ that often pertains to individual units (individuals, cities, etc.). The distribution of the sample is:

$$f(\mathbf{x},\mathbf{y};\boldsymbol{\theta})=\prod_{i=1}^{n}\pi_{11}^{(1-y_i)(1-x_i)}\pi_{21}^{(1-y_i)x_i}\pi_{12}^{y_i(1-x_i)}\pi_{22}^{x_iy_i},\ \text{s.t.}\ \sum_{i=1}^{2}\sum_{j=1}^{2}\pi_{i,j}=1,\ x=0,1,\ y=0,1,$$

where 's.t.' denotes 'subject to', giving rise to the log-likelihood function:

$$\ln L(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\propto n_{11}\ln\pi_{11}+n_{21}\ln\pi_{21}+n_{12}\ln\pi_{12}+n_{22}\ln\pi_{22}+\lambda(\textstyle\sum_{j=1}^{2}\sum_{i=1}^{2}\pi_{i,j}-1),$$

where $n_{ij},\ i,j=1,2,$ denote the observed frequencies corresponding to $\pi_{ij},\ i,j=1,2$:

$$n_{11}=\textstyle\sum_{i=1}^{n}(1-x_i)(1-y_i),\ n_{21}=\sum_{i=1}^{n}x_i(1-y_i),\ n_{12}=\sum_{i=1}^{n}(1-x_i)y_i,\ n_{22}=\sum_{i=1}^{n}x_iy_i,$$

Solving the first order conditions with respect to $\pi_{ij},\ i,j=1,2$ and λ:

$$\tfrac{\partial L}{\partial\pi_{ij}}=\tfrac{n_{ij}}{\pi_{ij}}-\lambda=0,\ \tfrac{\partial L}{\partial\lambda}=(\textstyle\sum_{j=1}^{2}\sum_{i=1}^{2}\pi_{i,j}-1)=0,\ i,j=1,2. \tag{19}$$

yields the MLEs $\widehat{\pi}_{ij}$ of π_{ij}, $i,j=1,2,$ that coincide with the relative frequencies:

$$\lambda=\tfrac{n_{ij}}{\pi_{ij}}\Longrightarrow\tfrac{1}{\lambda}\textstyle\sum_{j=1}^{2}\sum_{i=1}^{2}n_{ij}=1\Longrightarrow\lambda=\sum_{j=1}^{2}\sum_{i=1}^{2}n_{ij}=n\Longrightarrow\widehat{\pi}_{ij}=\tfrac{n_{ij}}{n},\ i,j=1,2.$$

Using the parameterization invariance property, one can derive the ML estimators of $\boldsymbol{\phi}:=(\theta_1,\theta_2,\theta_3)$ as well as the cross-product ratio ψ_{12} (chapter 6):

$$\widehat{\theta}_1=\widehat{\pi}_{21}+\widehat{\pi}_{22},\ \widehat{\theta}_2=\widehat{\pi}_{12}+\widehat{\pi}_{22},\ \theta_3=\widehat{\pi}_{22}-\widehat{\theta}_1\cdot\widehat{\theta}_2,\ \widehat{\psi}_{12}=\tfrac{\widehat{\pi}_{11}\cdot\widehat{\pi}_{22}}{\widehat{\pi}_{21}\cdot\widehat{\pi}_{12}}.$$

Example 12.13: **One-way Analysis of Variance (ANOVA)**. This is a heterogeneous extension of the simple Normal model, known as the one factor (or one-way) model, specified in terms of the statistical GM:

$$X_{ij}=\mu_i+u_{ij},\ u_{ij}\backsim\mathsf{NIID}\left(0,\sigma^2\right),\ j=1,2,...,n_i,\ i=1,2,\dots,p. \tag{20}$$

In terms of the observable i-heterogenous process: $X_{ij}\backsim\mathsf{NI}(\mu_i,\sigma^2)$, $i\in\mathbb{N}$, $t\in\mathbb{J}$. The data come in the form of $\mathbf{X}_0:=\{x_{ij},\ j=1,2,...,n_i,\ i=1,2,...,p\}$, and the log-likelihood is:

$$\ln L(\boldsymbol{\mu},\sigma^2;\mathbf{x}) \;\; =-\tfrac{N}{2}\left[\ln(2\pi)\right]-\tfrac{N}{2}\ln(\sigma^2)-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2.$$

where $N=\sum_{i=1}^{p}n_i$, $\boldsymbol{\mu}:=(\mu_1,\mu_2,...,\mu_p)$. The first order conditions yield the MLEs:

$$\widehat{\mu}_i=\tfrac{1}{n_i}\textstyle\sum_{j=1}^{n_i}X_{ij}\backsim\mathsf{N}\left(\mu_i,\tfrac{\sigma^2}{n_i}\right),\ i=1,2,\ldots,p,$$

$$\widehat{\sigma}^2=\tfrac{1}{N}\textstyle\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu}_i)^2,\ \tfrac{n\widehat{\sigma}^2}{\widehat{\sigma}^2}\backsim\chi^2\left(N-p\right).$$

In practice, the ANOVA model in (20) is often reparameterized using:

$$\mu_i=\mu+\alpha_i,\ \mu=\textstyle\sum_{i=1}^{p}n_i\mu_i,\quad \sum_{i=1}^{p}n_i\alpha_i=0.$$

Since $\sum_{i=1}^{p}\sum_{j=1}^{n_i}\mu\alpha_i=\mu\sum_{i=1}^{p}n_i\alpha_i$, the restriction $\sum_{i=1}^{p}n_i\alpha_i=0$ gives rise to the reparameterized model:

$$X_{ij}=\mu+\alpha_i+u_{ij},\ u_{ij}\backsim\mathsf{NIID}\left(0,\sigma^2\right),\ j=1,2,...,n_i,\ i=1,2,\ldots,p,$$

where μ and α_i, $i=1,2,\ldots,p$, are orthogonal in the sense that $\sum_{i=1}^{p}\sum_{j=1}^{n_i}\mu\alpha_i=0$. The parameterization invariance of the MLEs, implies that:

$$\widehat{\mu}=\tfrac{1}{N}\textstyle\sum_{i=1}^{p}\sum_{j=1}^{n_i}X_{ij}=\sum_{j=1}^{n_i}(\tfrac{n_i}{N})X_{ij},\quad \widehat{\alpha}_i=\widehat{\mu}_i-\widehat{\mu},\ i=1,2,\ldots,p,$$

are the MLEs of the orthogonal parameters $\boldsymbol{\phi}:=\left(\mu,\alpha_i,\ i=1,2,\ldots,p\right)$. It is important to emphasize that $\widehat{\mu}$ is a *weighted* average of X_{ij}, with weights $(\frac{n_i}{N})$, $i=1,2,\ldots,p$. In the *balanced case* where $n_1=n_2=\cdots=n_p=n$:

$$N=pn \implies \widehat{\mu}=\tfrac{1}{p}\textstyle\sum_{i=1}^{p}\widehat{\mu}_i.$$

The one-way ANOVA was introduced by Fisher (1921) and popularized after his first book, Fisher (1925a). The term ANOVA stems from the fact that the above model gives rise to a decomposition of the Total Sum of Squares (TSS) into the Explained (ESS) and the Residual Sum of Squares (RSS), which are all chi-square distributed with different degress of freedom (df), as shown in table 12.13, where $\mathsf{F}(p-1,N-p)$ denotes the F-distribution with $(p-1)$ and $(N-p)$ df.

Table 12.13: One-way Analysis of Variance (ANOVA)

Source of variation	Sum of Squares	df	Mean Square	$\mathsf{F}(p-1,N-p)$
Between groups	$\mathsf{ESS}=\sum_{i=1}^{p}n_i(\widehat{\mu}_i-\widehat{\mu})^2$	$(p-1)$	$\frac{1}{(p-1)}\sum_{i=1}^{p}n_i(\widehat{\mu}_i-\widehat{\mu})^2$	$\frac{\mathsf{ESS}/(p-1)}{\mathsf{RSS}/(N-p)}$
Within groups	$\mathsf{RSS}=\sum\limits_{i=1}^{p}\sum\limits_{j=1}^{n_i}(x_{ij}-\widehat{\mu}_i)^2$	$(N-p)$	$\frac{1}{(N-p)}\sum\limits_{i=1}^{p}\sum\limits_{j=1}^{n_i}(x_{ij}-\widehat{\mu}_i)^2$	
TSS	$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\widehat{\mu})^2$	$(N-1)$		

(2) Unbiasedness and Full efficiency

In a regular statistical model (see table 11.4), when there exists an unbiased estimator that also attains the Cramer-Rao Lower bound, say $\widehat{\theta}_U$, then it coincides with the maximum likelihood estimator, i.e. $\widehat{\theta}_U=\widehat{\theta}_{ML}$.

Example 12.14. Consider the simple *Poisson model* in table 12.14, whose density function is:

$$f(x;\theta)=(\tfrac{e^{-\theta}\theta^{x}}{x!}),\ \theta>0, x\in\mathbb{N}_0=\{0,1,2,....\}.$$

Given that $\theta=E(X_t)$, an obvious unbiased estimator of θ is $\widehat{\theta}_U=\frac{1}{n}\sum_{k=1}^{n}X_k$ since:

$$E(\widehat{\theta}_U)=\theta \text{ and } Var(\widehat{\theta}_U)=\tfrac{\theta}{n}.$$

Is $\widehat{\theta}_U$ also fully efficient: Assumptions [1]-[4] imply that $L(\theta;\mathbf{x})$ and $\frac{d\ln L(\theta;\mathbf{x})}{d\theta}=0$ are:

$$L(\theta;\mathbf{x})=\prod_{k=1}^{n}\theta^{x_k}e^{-\theta}(1/(x_k)!)=\theta^{\sum_{k=1}^{n}x_k}e^{-n\theta}\prod_{k=1}^{n}(1/(x_k)!)\Rightarrow$$
$$\Rightarrow\ln L(\theta;\mathbf{x})=(\textstyle\sum_{k=1}^{n}x_k)\ln\theta-n\theta-\sum_{k=1}^{n}\ln(x_k)\Rightarrow$$
$$\Rightarrow\frac{d\ln L(\theta;\mathbf{x})}{d\theta}=\left(-n+\tfrac{1}{\theta}\textstyle\sum_{k=1}^{n}X_k\right)\Rightarrow\frac{d^2\ln L(\mathbf{x};\theta)}{d\theta^2}=-\left(\tfrac{1}{\theta^2}\textstyle\sum_{k=1}^{n}X_k\right)\Rightarrow$$
$$\Rightarrow\mathcal{I}_n(\boldsymbol{\theta})=E(-\tfrac{d^2\ln L(\mathbf{x};\theta)}{d\theta^2})=\tfrac{1}{\theta^2}\textstyle\sum_{k=1}^{n}E(X_k)=\tfrac{n\theta}{\theta^2}=\tfrac{n}{\theta}.$$

Give that the Cramer-Rao lower bound is $\mathsf{C\text{-}R}(\theta)=\frac{\theta}{n}$, we can deduce that $\widehat{\theta}_U$ is fully efficient and thus it coincides with the ML estimator since:

$$\tfrac{d\ln L(\theta;\mathbf{x})}{d\theta}=(-n+\tfrac{1}{\theta}\textstyle\sum_{k=1}^{n}X_k)=0\Rightarrow\widehat{\theta}_{ML}=\tfrac{1}{n}\sum_{k=1}^{n}X_k=\widehat{\theta}_U.$$

Table 12.14: The simple Poisson model

Statistical GM:		$X_t=\theta+u_t,\ t\in\mathbb{N}:=(1,2,...,n,...)$
[1]	Logistic:	$X_t\backsim\mathsf{Poisson}(.),\ x_t\in\mathbb{N}_0,$
[2]	Constant mean:	$E(X_t)=\theta,\ \theta\in\mathbb{R},\ \forall t\in\mathbb{N},$
[3]	Constant variance:	$Var(X_t)=\theta,\ \forall t\in\mathbb{N},$
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$-independent process.

(3) Sufficiency

The notion of a sufficient statistic is operationalized using the *Factorization theorem*. A statistic $S(\mathbf{X})$ is said to be a *sufficient statistic* for θ if and only if there exist functions $g(S(\mathbf{X});\theta)$ and $v(\mathbf{X})$ such that:

$$f(\mathbf{x};\theta)=g(S(\mathbf{x});\theta)\cdot v(\mathbf{x}),\ \forall\mathbf{x}\in\mathbb{R}_X^n. \tag{21}$$

The result in (21) suggests that if there exists a sufficient statistic $h(\mathbf{X})$, and the MLE $\widehat{\theta}_{ML}(\mathbf{X})$ exists and is unique, then $\widehat{\theta}_{ML}(\mathbf{X})=h(S(\mathbf{X}))$ because:

$$L(\mathbf{x}_0;\theta)=[c(\mathbf{x}_0)\cdot v(\mathbf{x}_0)]\cdot g(S(\mathbf{x}_0);\theta)\propto g(S(\mathbf{x}_0);\theta),\ \forall\theta\in\Theta\Rightarrow \tag{22}$$

$$\tfrac{dL(\mathbf{x};\theta)}{d\theta}{=}\tfrac{dg(S(\mathbf{x});\theta)}{d\theta} \Rightarrow \widehat{\theta}_{ML}{=}h(S(\mathbf{X})),$$

ensuring that $\widehat{\theta}_{ML}{=}h(S(\mathbf{X}))$ depends on $\mathbf{X}$ only through the sufficient statistic.

(4) Full Efficiency

Recalling from chapter 11 that an estimator $\widehat{\theta}_n(\mathbf{X})$ is fully efficient iff:

$$(\widehat{\theta}_n(\mathbf{X})-\theta){=}h(\theta)\left[\tfrac{d\ln L(\mathbf{x};\theta)}{d\theta}\right], \tag{23}$$

for some function $h(\theta)$, implies that $L(\mathbf{x}_0;\theta)$ has the form in (22), and thus if a fully efficient estimator $\widehat{\theta}_n(\mathbf{X})$ exists, $\widehat{\theta}_n(\mathbf{X}){=}\widehat{\theta}_{ML}(\mathbf{X})$. This suggests that the existence of a sufficient statistic is weaker than that of a fully efficient estimator.

2.5.2 Asymptotic properties (IID sample)

Let us consider the asymptotic properties of MLEs in the simple IID sample case where:

$$\mathcal{I}_n(\theta){=}n\mathcal{I}(\theta),\ \mathcal{I}(\theta){=}E\left(\tfrac{d\ln f(x;\theta)}{d\theta}\right)^2>0, \tag{24}$$

where $\mathcal{I}(\theta)$ is known as Fisher's information for one observation. In addition to **R1-R6,** we will need the two conditions in table 12.15.

Table 12.15: Regularity conditions for $\ln L(\theta;\mathbf{x}),\ \forall\theta{\in}\Theta$.

(**R7**)	$E(\ln f(x;\theta))$ exists,
(**R8**)	$\frac{1}{n}\ln L(\theta;\mathbf{x})\overset{a.s.}{\to}E(\ln f(x;\theta)),\ \forall\theta{\in}\Theta,$
(**R9**)	$\ln L(\theta;\mathbf{x})$ is twice differentiable in an open interval around θ.

(5) Consistency

(a) **Weak Consistency**. Under these regularity conditions, MLEs are weakly consistent, i.e. for some $\varepsilon>0$:

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\widehat{\theta}_{ML}{-}\theta^*\right|<\varepsilon\right){=}1,\text{ and denoted by: }\widehat{\theta}_{ML}\overset{\mathbb{P}}{\to}\theta.$$

(b) **Strong Consistency**. Under these regularity conditions, MLEs are strongly consistent:

$$\mathbb{P}(\lim_{n\to\infty}\widehat{\theta}_{ML}{=}\theta^*){=}1,\text{ and denoted by: }\widehat{\theta}_{ML}\overset{a.s.}{\to}\theta.$$

See chapter 9 for a discussion between these two different modes of convergence.

(6) Asymptotic Normality

Under the regularity conditions **(R1)-(R9)**, MLEs are asymptotically Normal:

$$\sqrt{n}(\widehat{\theta}_{ML}-\theta^*)\underset{n\to\infty}{\backsim}\mathsf{N}(0,\ V_\infty(\theta)),\tag{25}$$

where $V_\infty(\theta)$ denotes the asymptotic variance of $\widehat{\theta}_{ML}$.

(7) Asymptotic Unbiasedness

The asymptotic Normality for MLEs also implies asymptotic unbiasedness:

$$\lim_{n\to\infty}E(\widehat{\theta}_{ML})=\theta^*.$$

(8) Asymptotic (full) Efficiency

Under the same regularity conditions the asymptotic variance of maximum likelihood estimators achieves the asymptotic Cramer-Rao lower bound, which in view of (24) is:

$$V_\infty(\widehat{\theta}_{ML})=\mathcal{I}^{-1}(\theta).$$

Example 12.15. For *the simple Bernoulli model* (table 12.4):

$$\sqrt{n}(\widehat{\theta}_{ML}-\theta)\underset{n\to\infty}{\backsim}\mathsf{N}(0,\theta(1-\theta)).$$

Example 12.16. For *the simple Exponential model* (table 12.8):

$$\sqrt{n}(\widehat{\theta}_{ML}-\theta)\underset{n\to\infty}{\backsim}\mathsf{N}(0,\theta^2).$$

Example 12.17. For *the simple Logistic model* (table 12.11):

$$\sqrt{n}(\widehat{\theta}_{ML}-\theta)\underset{n\to\infty}{\backsim}\mathsf{N}(0,3).$$

Example 12.18. For *the simple Normal model* (table 12.9):

$$\sqrt{n}(\widehat{\mu}_{ML}-\mu)\underset{n\to\infty}{\backsim}\mathsf{N}(0,\sigma^2),\quad \sqrt{n}(\widehat{\sigma}^2_{ML}-\sigma^2)\underset{n\to\infty}{\backsim}\mathsf{N}(0,2\sigma^4).$$

2.5.3 Asymptotic properties (Independent (I) but non-ID sample)

The above asymptotic properties need to be modified somewhat in the case where the sample is independent but non-identically distributed. In this case the relationship between the individual observation Fisher information $\mathcal{I}(\theta)$ and the sample Fisher information $\mathcal{I}_n(\theta)$ take the form:

$$\mathcal{I}_n(\theta)\overset{\text{I}}{=}\textstyle\sum_{k=1}^n\mathcal{I}_k(\theta),\ \mathcal{I}_k(\theta)=E\left(\left[\frac{d\ln f(x_k;\theta)}{d\theta}\right]^2\right).\tag{26}$$

For the above properties to hold we need to impose certain restrictions on the asymptotic behavior of $\mathcal{I}_n(\theta)$ (see Spanos, 1986, ch. 10) as given in table 12.16.

Table 12.16: Regularity conditions for $\mathcal{I}_n(\theta)$.

(**a**)	$\lim\limits_{n\to\infty}\mathcal{I}_n(\theta)=\infty$,
(**b**)	There exists a sequence $\{c_n\}_{n=1}^{\infty}$ such that $\lim\limits_{n\to\infty}\left(\frac{1}{c_n^2}\mathcal{I}_n(\theta)\right)=\mathcal{I}_\infty(\theta)>0$.

The first condition ensures consistency, and the second ensures asymptotic Normality.

Asymptotic Normality under these conditions takes the form:

$$c_n(\widehat{\theta}_{ML}-\theta)\underset{n\to\infty}{\backsim}\mathsf{N}(0,\ \mathcal{I}_\infty(\theta)).$$

Example 12.19. Consider *a Poisson model* with separable heterogeneity:

$$X_k\backsim\mathsf{PI}(k\theta),\ f(x_k;\theta)=\tfrac{e^{-\theta k}(\theta k)^x}{x!},\ E(X_k)=Var(X_k)=k\theta,\ k\in\mathbb{N},\ \theta>0,\ x=\{0,1,2,....\}.$$

$$L(\theta;\mathbf{x})=\prod_{k=1}^n(k\theta)^{x_k}e^{-(k\theta)}(1/(x_k)!)=\prod_{k=1}^n\tfrac{(k)^{x_k}}{x_k!}\exp(\textstyle\sum_{k=1}^n x_k\ln\theta)-\theta\sum_{k=1}^n k\Rightarrow$$

$$\Rightarrow\ln L(\theta;\mathbf{x})=\text{const}+\textstyle\sum_{k=1}^n x_k\ln\theta-\theta a_n,\ \text{ where } a_n=\sum_{k=1}^n k=\frac{1}{2}(n(n+1)$$

$$\Rightarrow\tfrac{d\ln L(\theta;\mathbf{x})}{d\theta}=(\tfrac{1}{\theta}\textstyle\sum_{k=1}^n X_k-a_n)=0\Rightarrow\widehat{\theta}_{ML}=\frac{1}{a_n}\sum_{k=1}^n X_k,\ E(\widehat{\theta}_{ML})=\theta,\ Var(\widehat{\theta}_{ML})=\frac{\theta}{a_n}.$$

The question is whether, in addition to being unbiased, $\widehat{\theta}_{ML}$ is fully efficient:

$$\tfrac{d^2\ln L(\mathbf{x};\theta)}{d\theta^2}=-\ (\tfrac{\sum_{k=1}^n X_k}{\theta^2})\Rightarrow\mathcal{I}_n(\theta)=E(-\tfrac{d^2\ln L(\mathbf{x};\theta)}{d\theta^2})=\tfrac{\sum_{k=1}^n E(X_k)}{\theta^2}=\tfrac{\sum_{k=1}^n\theta k}{\theta^2}=\tfrac{\theta a_n}{\theta^2}=\tfrac{a_n}{\theta}.$$

Hence, the $\mathsf{C\text{-}R}(\theta)=\frac{\theta}{a_n}=Var(\widehat{\theta}_{ML})$, and thus $\widehat{\theta}_{ML}$ is fully efficient. In terms of asymptotic properties $\widehat{\theta}_{ML}$ is clearly consistent since $Var(\widehat{\theta}_{ML})\underset{n\to\infty}{\to}0$.

The asymptotic Normality is less obvious, but since $\frac{1}{a_n}\mathcal{I}_n(\theta)\underset{n\to\infty}{\to}\frac{1}{\theta}$, the scaling sequence is $\{\sqrt{a_n}\}_{n=1}^{\infty}$:

$$\sqrt{a_n}(\widehat{\theta}_{ML}-\theta)\underset{n\to\infty}{\backsim}\mathsf{N}(0,\tfrac{1}{\theta}).$$

This, however, is not a satisfactory result because the variance involves the unknown θ. A more general result that is often preferable is to use $\{\sqrt{\mathcal{I}_n(\theta)}\}_{n=1}^{\infty}$ as the scaling sequence:

$$\sqrt{\mathcal{I}_n(\theta)}(\widehat{\theta}_{ML}-\theta)=(\tfrac{Y_n-\theta a_n}{\sqrt{\theta a_n}})\underset{n\to\infty}{\backsim}\mathsf{N}(0,1),\ Y_n=\textstyle\sum_{k=1}^n X_k\backsim\mathsf{P}\left(\theta a_n\right).$$

Example 12.20. Consider *an Independent Normal model* with separable heterogeneity:

$$X_k\backsim\ \mathsf{NI}(k\mu,\ 1),\ f(x_k;\theta)=\tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{(x_k-k\mu)^2}{2}),\ k\in\mathbb{N},\ \mu\in\mathbb{R},\ x\in\mathbb{R}.$$

The distribution of the sample is:

$$\begin{aligned} f(\mathbf{x};\theta) &= \prod_{k=1}^{n}\tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{(x_k-k\mu)^2}{2}) = (\tfrac{1}{\sqrt{2\pi}})^n\exp(-\tfrac{1}{2}\sum_{k=1}^{n}(x_k-k\mu)^2) = \\ &= (\tfrac{1}{\sqrt{2\pi}})^n\exp(-\tfrac{1}{2}\sum_{k=1}^{n}x_k^2)\exp(\mu\sum_{k=1}^{n}kx_k-\tfrac{b_n\mu^2}{2}), \end{aligned}$$

since $(x_k-k\mu)^2=x_k^2+k^2\mu^2-2k\mu x_k,\ b_n=\sum_{k=1}^{n}k^2=\frac{n(n+1)(2n+1)}{6}$ and thus $\ln L(\mu;\mathbf{x})$ is:

$$\begin{aligned} &\ln L(\mu;\mathbf{x})=\text{const.}+\mu\sum_{k=1}^{n}kx_k-\tfrac{b_n\mu^2}{2}\Rightarrow \\ &\tfrac{d\ln L(\mu;\mathbf{x})}{d\mu}=\sum_{k=1}^{n}kx_k-\mu b_n=0\Rightarrow\widehat{\mu}_{ML}=\tfrac{1}{b_n}\sum_{k=1}^{n}kX_k,\Rightarrow \\ &E(\widehat{\mu}_{ML})=\mu,\ Var(\widehat{\mu}_{ML})=\tfrac{1}{b_n^2}\sum_{k=1}^{n}k^2Var(X_k)=\tfrac{b_n}{b_n^2}=\tfrac{1}{b_n}\Rightarrow \\ &\mathcal{I}_n(\mu)=E(-\tfrac{d^2\ln L(\mu;\mathbf{x})}{d\mu^2})=b_n\Rightarrow \text{C-R}\,(\mu)=\tfrac{1}{b_n}=Var(\widehat{\mu}_{ML}). \end{aligned}$$

These results imply that $\widehat{\mu}_{ML}$ is unbiased, fully efficient and consistent. In addition, since $\frac{1}{b_n}\mathcal{I}_n(\mu)\underset{n\to\infty}{\to}1$:

$$\sqrt{b_n}(\widehat{\mu}_{ML}-\mu)\underset{n\to\infty}{\backsim}\mathsf{N}(0,1).$$

Summary of optimal properties of MLEs. The *Maximum Likelihood method* yields estimators which, under certain regularity conditions, enjoy all the asymptotic optimal properties, *consistency*, *asymptotic Normality*, *unbiasedness* and *efficiency*, and in addition they satisfy excellent finite sample properties, such as *reparameterization invariance, sufficiency* as well as *unbiasedness-full efficiency* when they hold simultaneously.

2.6 The Maximum Likelihood method and its critics

The results relating to MLEs discussed above justify the wide acceptance of the maximum likelihood (ML) as the method of choice for estimation purposes in frequentist statistics. It turns out that there are good reasons for the ML method to be preferred for testing purposes as well (see chapter 14). Despite the wide acceptance of the ML method there are also critics who point to several examples where the method does not yield satisfactory results. Such examples range from cases where (a) the sample size is inappropriately small, (b) the regularity conditions do not hold, and (c) the postulated statistical model is problematic.

The criticism in (a) is completely misplaced because the modeler is looking for the famous 'free' lunch. As argued in chapter 1, if the sample size is too small to enable the modeler to test the model assumptions adequately, it is too small for inference purposes. The criticism of the ML method based on examples which do not satisfy the regularity conditions is also somewhat misplaced because when the modeler seeks methods with any generality the regularity conditions are inevitable. Without regularity conditions each estimation problem will be viewed as unique; no unifying principles are possible. Category (c) deserves more discussion because the assumed statistical models are ill-specified. From this category let us consider a widely discussed example.

Example 12.21: **Neyman and Scott** (1948) **model.** The statistical GM for this N-S model takes the form**:**

$$\mathbf{X}_t{=}\boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t,\;\; t{=}1,2,...,n,...,$$

where the underlying distribution Normal of the form:

$$\mathbf{X}_t{:=}\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix} \backsim \mathsf{NI}\left(\begin{pmatrix} \mu_t \\ \mu_t \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right),\;\; t{=}1,2,...,n,\ldots \tag{27}$$

NOTE that this model is not well-defined since it has an *incidental parameter problem*: the unknown parameters $(\mu_1,\mu_2,...,\mu_n)$ increase with the sample size n. Neyman and Scott attempted to sidestep this problem by declaring σ^2 the only parameter of interest and designating $(\mu_1,\mu_2,...,\mu_n)$ as *nuisance* parameters, which does not deal with the problem.

Let us ignore the incidental parameter problem and proceed to derive the distribution of the sample and the log-likelihood function:

$$f(\mathbf{x};\boldsymbol{\theta}) \;=\prod_{t=1}^{n}\prod_{i=1}^{2}\tfrac{1}{\sigma\sqrt{2\pi}}e^{\left\{-\frac{1}{2\sigma^2}(x_{it}-\mu_t)^2\right\}}{=}\prod_{t=1}^{n}\tfrac{1}{2\pi\sigma^2}e^{\left\{-\frac{1}{2\sigma^2}[(x_{1t}-\mu_t)^2+(x_{2t}-\mu_t)^2]\right\}}$$

$$\ln L(\boldsymbol{\theta};\mathbf{x}){=}-n\ln\sigma^2{-}\tfrac{1}{2\sigma^2}\textstyle\sum_{t=1}^{n}[(x_{1t}{-}\mu_t)^2+(x_{2t}{-}\mu_t)^2]. \tag{28}$$

In light of (28), the "MLEs" are then derived by solving the first-order conditions:

$$\begin{aligned} \tfrac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\mu_t} &{=}\tfrac{1}{\sigma^2}[(x_{1t}{-}\mu_t){+}(x_{2t}{-}\mu_t)]{=}0 \Rightarrow \widehat{\mu}_t{=}\tfrac{1}{2}(X_{1t}{+}X_{2t}),\;\; t{=}1,...,n, \\ \tfrac{\partial\ln L(\boldsymbol{\theta};\mathbf{x})}{\partial\sigma^2} &{=}{-}\tfrac{n}{\sigma^2}{+}\tfrac{1}{2\sigma^4}\textstyle\sum_{t=1}^{n}[(x_{1t}{-}\mu_t)^2+(x_{2t}{-}\mu_t)^2]{=}0 \Rightarrow \\ &\Rightarrow \widehat{\sigma}^2{=}\tfrac{1}{2n}\textstyle\sum_{t=1}^{n}[(X_{1t}{-}\widehat{\mu}_t)^2+(X_{2t}{-}\widehat{\mu}_t)^2]{=}\tfrac{1}{n}\sum_{t=1}^{n}\tfrac{(X_{1t}-X_{2t})^2}{4}, \end{aligned} \tag{29}$$

Critics of the ML method claim that ML yields *inconsistent* estimators since:

$$E(\widehat{\mu}_t){=}\mu_t,\; Var(\widehat{\mu}_t){=}\tfrac{1}{2}\sigma^2 \underset{n\to\infty}{\nrightarrow} 0,\;\; E(\widehat{\sigma}^2){=}\tfrac{1}{2}\sigma^2,\;\; \widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \tfrac{1}{2}\sigma^2{\neq}\sigma^2.$$

This, however, is a misplaced criticism since by definition $\sigma^2{=}E(X_{it}{-}\mu_t)^2$, and thus any attempt to find a consistent estimator of σ^2 calls for a consistent estimator of μ_t, but $\widehat{\mu}_t{=}\frac{1}{2}(X_{1t}{+}X_{2t})$ is *inconsistent*.

In light of that, the real question is not why the ML does not yield a consistent estimator of σ^2, but given that (27) is ill-specified:

► **Why would the ML method yield a consistent estimator of** σ^2?

Indeed, the fact that the ML method does *not* yield consistent estimators in such cases is an argument in its favor, not against it! A modeler should be skeptical of any method of estimation that yields consistent estimators in the context of (27).

The source of the problem is not the ML method but the statistical model in (27). Hence, one should focus on respecifing the ill-defined model with a view to find an optimal estimator σ^2, without the incidental parameter problem. This problem can be addressed by respcifying (27) to address the incidental parameter problem using the transformation:

$$Y_t = \tfrac{1}{\sqrt{2}}(X_{1t} - X_{2t}) \backsim \mathsf{NIID}\left(0, \sigma^2\right), \ t=1,2,...,n,..., \tag{30}$$

For the respecified model in (30) the MLE for σ^2 is: $\widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{t=1}^{n} Y_t^2$, which is unbiased, fully efficient and strongly consistent: $E(\widehat{\sigma}^2_{ML}) = \sigma^2$, $Var(\widehat{\sigma}^2_{ML}) = \frac{2\sigma^4}{n}$, $\widehat{\sigma}^2_{ML} \overset{a.s.}{\to} \sigma^2$.

The criticism in (c) relates to ill-specified models with suffering from the incidental parameter or contrived constraints that give rise to unnatural reparameterizations are imposed on the parameters at the outset; see Spanos (2010b; 2011a; 2012b; 2013a-d). Cautionary note: when the ML method does not gives rise to 'optimal' estimators, one should first take a closer look at the assumed statistical model to verify that it is well-specified before blaming the ML method.

3 The Least-Squares method

3.1 The mathematical principle of least-squares

The *principle of least-squares* was originally proposed as a mathematical approximation procedure by Legendre in 1805; see Harter (1974-76). In its simplest form the problem involves the approximating of an *unknown* function $h(.)$: $\mathbb{R}_X \to \mathbb{R}_Y$:

$$y = h(x), \ (x, y) \in \left(\mathbb{R}_X \times \mathbb{R}_Y\right),$$

by selecting an *approximating* function, say linear: $g(x) = \alpha_0 + \alpha_1 x$, $(x, y) \in \left(\mathbb{R}_X \times \mathbb{R}_Y\right)$, and fitting $g(x)$ using data $\mathbf{z}_0 := \{(x_t, y_t), \ t=1,2,..,n\}$. This curve-fitting problem involves the approximation error: $\epsilon_t = h(x_t) - g(x_y)$, giving rise to the problem of how to use data $\mathbf{z}_0$ to get the best approximation by fitting:

$$y_t = \alpha_0 + \alpha_1 x_t + \epsilon_t, \ t=1,2,\ldots,n. \tag{31}$$

The earliest attempt to address this problem was made by Boscovitch in 1757 by proposing (Hald, 1998, 2007) the criterion:

$$\min_{\alpha_0,\alpha_1} \sum_{t=1}^{n} |\epsilon_t| \ \text{ subject to } \ \sum_{t=1}^{n} \epsilon_t = 0, \tag{32}$$

using a purely geometric argument about its merits. In 1789 Laplace proposed an analytic solution to the minimization problem in (32) that was rather laborious to implement. In 1805 Legendre offered a less laborious solution to the approximation problem by replacing $\sum_{t=1}^{n} |\epsilon_t|$ with $\sum_{t=1}^{n} \epsilon_t^2$, giving rise to the much easier minimization of the sum of squares (least-squares) of the errors:

$$\min_{\alpha_0,\alpha_1} \sum_{t=1}^{n} \epsilon_t^2.$$

In the case of (31), the principle of least squares amounts to minimizing:

$$\ell(a_0,\alpha_1){=}\textstyle\sum_{t=1}^{n}\left(y_t-\alpha_0-\alpha_1x_t\right)^2. \tag{33}$$

The first order conditions for a minimum, called the *normal equations***,** are:

$$\text{(i) } \tfrac{\partial\ell}{\partial\alpha_0}{=}(-2)\textstyle\sum_{t=1}^{n}(y_t{-}\alpha_0{-}\alpha_1x_t){=}0,\ \text{(ii) } \tfrac{\partial\ell}{\partial\alpha_1}{=}(-2)\textstyle\sum_{t=1}^{n}(y_t{-}\alpha_0{-}\alpha_1x_t)x_t{=}0.$$

Solving these two equations for (a_0,α_1) yields the Least-Square estimates:

$$\widehat{\alpha}_0{=}\overline{y}-\widehat{\alpha}_1\bar{x},\quad \widehat{\alpha}_1{=}\tfrac{\sum_{t=1}^{n}(y_t-\overline{y})(x_t-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}. \tag{34}$$

Example 12.22. The fitted line $\widehat{y}_t{=}\widehat{\alpha}_0{+}\widehat{\alpha}_1x_t$, through a scatter-plot of data $(n{=}200)$ in figure 12.1 is:

$$\widehat{y}_t{=}1.105+.809x_t. \tag{35}$$

In addition to (35), one could construct goodness-of-fit measures:

$$s^2{=}\tfrac{1}{n-2}\textstyle\sum_{t=1}^{n}\widehat{\epsilon}_t^2{=}.224,\ R^2{=}1{-}\left[\sum_{t=1}^{n}\widehat{\epsilon}_t^2/\sum_{t=1}^{n}(y_t{-}\overline{y})^2\right]{=}.77.8. \tag{36}$$

As it stands, however, (35)-(36) provides *no basis* for inductive inference. The fitted line in (35) cannot be used as a basis of any form of statistical inference because it has no inductive premises to provide measures for the uncertainty associated with $(\widehat{\alpha}_0,\widehat{\alpha}_1)$.

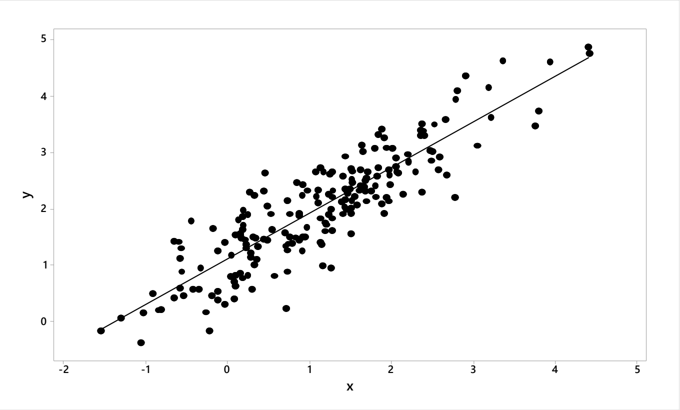

Fig. 12.1: Least-squares line fitting

The above mathematical approximation perspective to curve-fitting does not have any *probabilistic premises* stating the conditions under which the statistics $(\widehat{\alpha}_0,\widehat{\alpha}_1,s^2,R^2)$ are inferentially meaningful and reliable, as opposed to mathematically meaningful.

3.2 Least squares as a statistical method

It is interesting to note that Legendre's initial justification for the least-squares method was that for the simplest case where $g(x){=}\mu,\ (x,y){\in}(\mathbb{R}{\times}\mathbb{R})$:

$$Y_t=\mu+\epsilon_t,\ t{=}1,2,\ldots,n, \tag{37}$$

minimizing the sum of squares $\sum_{t=1}^{n}\epsilon_t^2$ yields:

$$\ell(\mu)=\sum_{t=1}^{n}(Y_t-\mu)^2 \;\;\Rightarrow\;\; \tfrac{d\ell}{d\mu}=(-2)\sum_{t=1}^{n}(Y_t-\mu)=0,$$

giving rise to the *arithmetic mean*: $\widehat{\mu}=\frac{1}{n}\sum_{t=1}^{n}Y_t$. At that time, the arithmetic mean was considered to be the gold standard for summarizing the information contained in the n data points $y_1, y_2, ..., y_n$, unaware that this presumes that $(Y_1, ..., Y_n)$ are IID.

The first probabilistic framing for least-squares was given by Gauss (1809). He introduced the Normal distribution by arguing that for a sequence of n independent random variables $Y_1, Y_2, ..., Y_n$, whose density function $f(y_t)$ satisfy certain regularity conditions, if $\bar{y}$ is the most probable combination for all values of $y_1, y_2, ..., y_n$ and each $n \geq 1$, then $f(y_t)$ is Normal; see Heyde and Seneta (1977), p. 63. This provided the missing probabilistic premises, and Gauss (1821) went on to prove an important result known today as the Gauss-Markov theorem.

Gauss-Markov theorem. Gauss supplemented the statistical GM (37) with the probabilistic assumptions:

$$\text{(i) } E(\epsilon_t)=0,\;\; \text{(ii) } E(\epsilon_t^2)=\sigma^2>0,\;\; \text{(iii) } E(\epsilon_t\epsilon_s)=0,\;\; t{\neq}s,\;\; t,s{=}1,2,\ldots,n,$$

and proved that under assumptions (i)-(iii) the least-squares estimator $\widehat{\mu}_{LS}=\frac{1}{n}\sum_{t=1}^{n}Y_t$ is Best (smallest variance) within the class of Linear and Unbiased Estimators (BLUE).

Proof. Any *linear* estimator of μ will be of the form $\widetilde{\mu}(\mathbf{w})=\sum_{t=1}^{n}w_tY_t$, where $\mathbf{w}{:=}(w_1, w_2, ..., w_n)$ denote constant weights. For $\widetilde{\mu}(\mathbf{w})$ to be unbiased it must be the case that $\sum_{t=1}^{n}w_t=1$, since $E(\widetilde{\mu}(\mathbf{w}))=\sum_{t=1}^{n}w_tE(Y_t)=\mu$. This implies that the problem of minimizing $Var(\widetilde{\mu}(\mathbf{w}))=\sigma^2\sum_{t=1}^{n}w_t^2$ can be transformed into a Largange multiplier problem:

$$\min_{\mathbf{w}}\mathcal{L}(\mathbf{w})=\left(\textstyle\sum_{t=1}^{n}w_t^2\right)-2\lambda\left(\textstyle\sum_{t=1}^{n}w_t-1\right),$$

whose first order conditions for a minimum yield:

$$\left.\begin{array}{l}\frac{\partial\mathcal{L}(\mathbf{w})}{w_t}=2w_t-2\lambda=0 \Rightarrow (w_t=\lambda)\\ \frac{\partial\mathcal{L}(\mathbf{w})}{\lambda}=-2\left(\sum_{t=1}^{n}w_t-1\right)=0\end{array}\right\} \Rightarrow \sum_{t=1}^{n}\lambda=1 \Rightarrow \lambda=\frac{1}{n},\;\; t=1,2,\ldots,n.$$

This proves that $\widehat{\mu}=\frac{1}{n}\sum_{t=1}^{n}Y_t$ is BLUE of μ. ■

The Gauss-Markov theorem is of very limited value in 'learning from data' because a BLUE estimator provides a very poor basis for inference, since the sampling distributions of $\widehat{\mu}_{LS}$ and $\widehat{\sigma}_{LS}^2=\frac{1}{n-1}\sum_{t=1}^{n}(Y_t-\widehat{\mu}_{LS})^2$ are unknown, and their first two moments involve unknown parameters that need to be estimated (table 12.21):

$$\widehat{\mu}_{LS}\overset{?}{\backsim}D_1(\mu,\tfrac{\sigma^2}{n}),\;\;\; \widehat{\sigma}_{LS}^2\overset{?}{\backsim}D_2(\left(\tfrac{n-1}{n}\right)\sigma^2,\tfrac{\mu_4-\mu_2^2}{n}-\tfrac{2(\mu_4-2\mu_2^2)}{n^2}+\tfrac{\mu_4-3\mu_2^2}{n^3})\;\;\; Cov(\widehat{\mu},\widehat{\sigma}^2)=\left(\tfrac{n-1}{n}\right)\mu_3.$$

In addition, the class of Linear and Unbiased estimators is unnecessarily narrow. For instance, in the case where the distribution of ϵ_t is Laplace (Appendix 3.A), the MLE of μ is the sample median $m(\mathbf{Y})=Y_{[\frac{n+1}{2}]}$ for n odd, and its variance is smaller than that of $\widehat{\mu}$; see Norton (1984). The Gauss-Markov theorem evades this problem because $m(\mathbf{x})$ is excluded from consideration for being a non-linear function of $\mathbf{Y}$; see chapter 14 for further discussion.

4 Summary and conclusions

1. Maximum Likelihood (ML) method. The method of ML is tailor-made for frequentist estimation because the likelihood function contains all the probabilistic information comprising the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}_X^n$, since it is defined as proportional to the distribution of the sample**:**

$$L(\boldsymbol{\theta};\mathbf{x}_0) \propto f(\mathbf{x}_0;\boldsymbol{\theta}),\ \text{for all } \boldsymbol{\theta}\in\Theta.$$

The property of sufficiency for MLEs often ensures optimal finite sample properties, and under certain regularity conditions, MLEs enjoy all optimal asymptotic properties. These optimal properties justify the wide acceptance of the ML as the method of choice for estimation purposes. The critics of the ML method often use problematic examples that range from cases where: (a) the sample size n is too small, (b) the regularity conditions do not hold and (c) the postulated model is *not* well-defined. The rule of thumb for how large n should be is: if n is too small to test the model assumptions using comprehensive misspecification testing, it it too small for inference purposes!

Does the superiority of the ML method implies that the other methods of estimation are redundant? The answer is that the other methods have something to contribute by supplementing and shedding additional light on the ML method.

2. Method of Least-Squares (LS). The LS procedure provides additional insight into the ML estimation of statistical models based on the Normal distribution. The additional insight stems from the geometry of fitting a line to a scatter-plot. Beyond that, the method of Least-Squares can be very misleading in practice. A closer look at this theorem reveals that its results are of very limited value for inference purposes! One needs to invoke asymptotics for inference purposes; chapters 9 and 14.

3. Moment Matching (MM) principle. This is not a fully fleshed method of estimation, but it can be used to provide additional intuition and insight into other estimation methods, including the ML method.

4. Parametric Method of Moments (PMM). This estimation method is clearly problematic because it does not utilize all the systematic information included in the statistical model. The same comments pertaining to invoking asymptotics for inference purposes apply to this method as well. Its real value is to provide respectable initial estimates in the context of numerical optimization for MLEs.

Additional references: Stuart et al. (1999), Pawitan (2001), Severini (2000).

Important concepts

Method of maximum likelihood, least squares method, moment matching principle, Pearson's method of moments, parametric method of moments, maximum likelihood estimator, regular statistical models, score function, Fisher information, Kullback-Leibler distance, parameterization invariance, Gauss-Markov theorem, sample moments and their sampling distributions.

Crucial distinctions

Pearson's vs. parametric method of moments, distribution of the sample vs. likelihood function, nuisance parameters vs. parameters of interest, least squares as mathematical approximation vs. statistical estimation procedure, Gauss-Markov theorem, parameters vs. estimators vs. estimates, sampling distributions of sample moments under IID vs. NIID.

Essential ideas

- The method of Maximum Likelihood is custom-made for parametric inference, and delivers the most optimal estimators for regular statistical models.
- The Least-squares method is an adaptation of a numerical approximation method that adds geometric intutition to estimation in certains cases, but very little else.
- The Moment Matching Principle is the result of a major confusion in statistics, initially brought out by Fisher (1922a), but in some cases (Normal, Bernoulli) it delivers good estimators.
- The Parametric Method of Moments is an anachronistic interpretation of Karl Pearson's method of moments, which was designed for a very different approach to statistical modeling and inference.
- Reliance on the asymptotic sampling distributions of sample moments without a distribution assumptions often gives rise to highly imprecise and potentially unreliable inferences. One is always better off assuming an explicit distribution and testing it rather than being agnostic.