ABSTRACT

Proposal for an Adjusted Numerical Approximation to the Integrated Likelihood Function

Timothy Ruel

The primary focus of this prospectus is to motivate and explain an adapted version of numerical likelihood integration as a method for eliminating nuisance parameters from a statistical model.

Table of Contents

CHAPTER 1

Introduction and Motivation

CHAPTER 2

Experiments and Statistical Models

The acquisition of knowledge regarding a population of interest has long been the impetus for the field of statistical inference. In all but the most basic of circumstances, limiting constraints such as time, accessibility, and cost make perfect knowledge of a population essentially impossible to obtain. It therefore becomes necessary to infer characteristics of the population based on a random and representative sample of observations drawn from it. The procedures by which these samples might be procured are themselves far from trivial, and indeed an entire branch of statistics has been dedicated to their study. However, we are primarily concerned in this paper with what occurs after the sample has been taken, and so we will generally take it for granted that a suitably representative sample of the population already exists.

Suppose $(x_1, ..., x_n)$ is one such sample. What information can we then glean from this sample about the population from which it has been drawn? Where is its point of central tendency located? Are its values clustered tightly around this point, or are they more diffuse? Are they distributed symmetrically or skewed to one side or the other? As the questions increase in complexity, so do the techniques required to answer them. Unfortunately the natural chaos of the real world all but guarantees there will never be an instrument capable of completely capturing the intricacies of a population whose properties we wish to infer. Hence, some amount of idealization will always be required in order to proceed.

This idealization typically comes in the form of additional assumptions that we impose on the population with the goal of sacrificing what we hope is only a small amount of accuracy in exchange for a large reduction in complexity. These assumptions are essentially never "true" in the sense that they are not a flawless representation of reality, but they may nevertheless serve as convenient approximations that are capable of producing answers with degrees of accuracy high enough to be useful in their own right. Taken as a whole, they form the basis for a statistical model.

The traditional framework for a statistical model begins by assuming that there exists an unknown probability distribution P over the population of interest that generates the data we observe from it. We choose to model this observed data as being the realized outcomes ("realizations") of some random variable X that is distributed according to P. Let $\mathcal{P}$ denote the set of all such distributions that we are willing to consider as candidates for the true distribution P. Out of necessity, we will proceed as though our choice of $\mathcal{P}$ always contains P though in reality there is nothing specifically requiring it.

The next assumption we make is that $\mathcal{P}$ is *parameterized*. That is, there exists a *parameter* $\boldsymbol{\theta}$ which indexes $\mathcal{P}$, acting as a label that allows us to differentiate between the distributions it contains. For a particular value of the parameter $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_1$, we can refer to its corresponding distribution in $\mathcal{P}$ with the notation $P_{\boldsymbol{\theta}_1}$, and therefore $\mathcal{P}$ itself may be written as $\mathcal{P} = \{P_{\boldsymbol{\theta}} | \, \boldsymbol{\theta} \in \Theta\}$. Θ is called the *parameter space* and represents the set of all possible values $\boldsymbol{\theta}$ can take on.

We will restrict our attention in this paper to distributions that are absolutely continuous with respect to some σ-finite measure μ, so that they admit a probability density function by the Radon-Nikodym Theorem. Let $p_{\boldsymbol{\theta}}$ denote the density function associated with the distribution $P_{\boldsymbol{\theta}}$. This one-to-one correspondence between distribution and density allows us to define $\mathcal{P}$ as $\mathcal{P} = \{p_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \boldsymbol{\theta}\}$. Going forward, we will use the notation $p_{\boldsymbol{\theta}}(x)$ and $p(x; \boldsymbol{\theta})$ interchangeably to refer to a density function with parameter $\boldsymbol{\theta}$. We will also simply write $\int p_{\boldsymbol{\theta}}(x)dx$ instead of $\int p_{\boldsymbol{\theta}} d\mu$ or $\int p_{\boldsymbol{\theta}}(x)d\mu(x)$ with the understanding that $p_{\boldsymbol{\theta}}$ is always defined with respect to some dominating measure μ.

In general, a model's parameterization is not unique, and for a given parameter $\boldsymbol{\theta}$, we are free to choose any invertible function of $\boldsymbol{\theta}$ as a new parameter. Once we have made our choice of parameterization, we will assume that $\boldsymbol{\theta}$ does contain a singular true parameter value, which we will denote by $\boldsymbol{\theta}_0$. The conventional interpretation of $\boldsymbol{\theta}_0$ is as a fixed but unknown constant that represents the value of the parameter corresponding to the true density function $p_{\boldsymbol{\theta}_0}$ in $\mathcal{P}$. Conversely, $\boldsymbol{\theta}$ represents an arbitrary parameter value that is allowed to range over all possible elements of Θ, including $\boldsymbol{\theta}_0$. In other words, $\boldsymbol{\theta}$ acts like a tuning dial for the population - rotate the dial and certain behaviors of the population (e.g. its location, scale, or shape) will change. Making inferences regarding $\boldsymbol{\theta}_0$ is like trying to figure out the particular value ($\boldsymbol{\theta}_0$) to which a population's dial ($\boldsymbol{\theta}$) has been set. Note that it is possible for the value

of $\boldsymbol{\theta}_0$ itself to change over time as well, depending on the population. In such cases, any estimate of $\boldsymbol{\theta}_0$ based on a cross-sectional sample drawn from the population is best thought of as an estimate of the true parameter value during the particular time in which the sample was collected.

Crucially, it must always be possible to identify the parameter in our model on the basis of the data we observe. A model is considered *identifiable* if having perfect knowledge of the population would enable us to determine $\boldsymbol{\theta}_0$ with absolute certainty. This is equivalent to requiring that for some observed data x and any two parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\theta}$, if $p_{\boldsymbol{\theta}_1}(x) = p_{\boldsymbol{\theta}_2}(x)$, then it must follow that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. A model that is not identifiable could potentially have two or more distinct parameter values that give rise to the same probability distribution. For example, suppose Y is distributed uniformly on the interval $(0, \alpha + \beta)$, where $\alpha, \beta > 0$. If we use $\boldsymbol{\theta} = (\alpha, \beta)$ as a parameter for the distribution of Y, then $\boldsymbol{\theta}$ is unidentifiable since, for instance, the case where $\boldsymbol{\theta}_1 = (0, 1)$ and $\boldsymbol{\theta}_1 = (1, 0)$ implies that $p_{\boldsymbol{\theta}_1}(y) = p_{\boldsymbol{\theta}_2}(y)$ despite the fact that $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. This is clearly an undesirable property for a model to possess, and so we will consider only identifiable models in this paper as a means of avoiding it.

Finally, we must make a choice regarding the dimension of the parameter space Θ when formulating our models. *Parametric* models are defined as having finite-dimensional parameter spaces. Any model that is not parametric is either *semi-parametric* or *nonparametric*. In this paper, we will consider only parametric models whose parameter spaces are subsets of the d-dimensional real coordinate space, i.e., $\Theta \subseteq \mathbb{R}^d$, where $d \in \mathbb{Z}^+$.

CHAPTER 3

The Likelihood Function

3.1. Definition

Upon choosing a statistical model that we think best characterizes our population of interest, the obvious next step is to identify the true distribution in $\mathcal{P}$ or at the very least, the one that best approximates the truth. This is equivalent to making inferences about $\boldsymbol{\theta}_0$ in the case where the model is parametric and identifiable. That is, given the particular form(s) we have chosen for the distributions in $\mathcal{P}$, the only unknown remaining is the value of $\boldsymbol{\theta}_0$ itself. Since this value is ultimately what controls the mechanism generating any sample of data $\mathbf{x}_n = (x_1, ..., x_n)$ that we might observe from the population, it stands to reason that information regarding $\boldsymbol{\theta}_0$ can be inferred from the specific values of $x_1, ..., x_n$ that we obtain. To make this notion more rigorous, we require some method of analyzing the joint probability of our sample as a function of our parameter $\boldsymbol{\theta}$.

Given some observed data $\mathbf{x}_n$, the *likelihood function* for $\boldsymbol{\theta}$ is defined as

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{x}_n) = p(\mathbf{x}_n; \boldsymbol{\theta}), \;\; \boldsymbol{\theta} \in \Theta. \tag{3.1.1}$$

In other words, the value of the likelihood function evaluated at a particular $\boldsymbol{\theta} \in \Theta$ is simply equal to the output of the model's density function evaluated at the same inputs. However, while $p(\mathbf{x}_n; \boldsymbol{\theta})$ is viewed primarily as a function of $\mathbf{x}_n$ for fixed $\boldsymbol{\theta}$, the reverse is actually true for $L(\boldsymbol{\theta}; \mathbf{x}_n)$. Indeed, we regard the likelihood as being a function of the parameter $\boldsymbol{\theta}$ for fixed $\mathbf{x}_n$. The reversal of the order of the arguments $\boldsymbol{\theta}$ and x is a reflection of this difference in perspectives.

When X is discrete, we may interpret $L(\boldsymbol{\theta}; x)$ as the probability that $X = x$ given that $\boldsymbol{\theta}$ is the true parameter value.[1] Crucially, this is *not* equivalent to the inverse probability that $\boldsymbol{\theta}$ is the true parameter

[1]It is important to note here that whichever value of $\boldsymbol{\theta}$ we choose to plug into $L(\boldsymbol{\theta}; x)$ is the value that we are currently "pretending" is the true one, regardless of whether or not it actually equals $\boldsymbol{\theta}_0$ in reality.

value given $X = x$. The likelihood does not directly tell us anything about the probability that $\boldsymbol{\theta}$ assumes any particular value at all. Though intuitively appealing, this interpretation constitutes a fundamental misunderstanding of what a likelihood function is, and great care must be taken to avoid it.

When X is continuous, the likelihood for $\boldsymbol{\theta}$ may still be defined as it is in Equation 3.1.1. However, we must forfeit our previous interpretation of $L(\boldsymbol{\theta})$ as a probability since the probability that X takes on any particular value is now 0. We may however still think of the likelihood as being proportional to the probability that X takes on a value "close" to x, meaning that that X is within a tiny ball centered at x. Specifically, for two different observations x_1 and x_2, if $L(\boldsymbol{\theta}; x_1) = c \cdot L(\boldsymbol{\theta}; x_2)$, where $c > 1$, then under this model we may conclude X is c times more likely to assume a value closer to x_1 than x_2 given that $\boldsymbol{\theta}$ is the true value of the parameter.

As in the discrete case, we must also be careful when X is continuous to avoid using $L(\boldsymbol{\theta}; \mathbf{x}_n)$ to make probabilistic assertions regarding $\boldsymbol{\theta}$. Despite our use of probability in its definition, the likelihood itself is *not* a probability density function for the parameter $\boldsymbol{\theta}$ and is subject to neither the same rules nor interpretations as one.

3.2. Transformations

There are a few useful transformations of the likelihood function that we will define here for use in future sections. The first is the *log-likelihood function*, which is defined as the natural logarithm of the likelihood function:

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \mathbf{x}_n) = \log L(\boldsymbol{\theta}; \mathbf{x}_n). \tag{3.2.1}$$

In practice, we will typically eschew direct analysis of the likelihood in favor of the log-likelihood due to the nice mathematical properties logarithms possess. Chief among these properties is the ability to turn products into sums (i.e. $\log(ab) = \log(a) + \log(b)$ for $a, b > 0$). Sums tend to be easier to differentiate than products, making this is a particularly useful feature for likelihood functions, which are often expressed as the product of marginal density functions when the observations are independent.

The other key property of logarithms that makes the log-likelihood so useful is that they are strictly increasing functions of their arguments (i.e. $\log x > \log y$ for $x > y > 0$). This monotonicity ensures that the locations of a function's extrema are preserved when the function is passed to the argument of a logarithm. For example, for a positive function f with a global maximum, $\underset{x}{\operatorname{argmax}} f(x) = \underset{x}{\operatorname{argmax}} \log f(x)$.

In the general case in which $\boldsymbol{\theta}$ is a d-dimensional vector, where d is an integer greater than 1, it follows that the first derivative of the log-likelihood with respect to $\boldsymbol{\theta}$ will also be a d-dimensional vector, the second derivative will be a $d \times d$ matrix, the third derivative will be a $d \times d \times d$ array, and so forth. To emphasize the multidimensional nature of these results, we will use notation typically associated with partial derivatives involving functions of more than one variable (e.g. $\nabla, \mathbf{J}, \mathbf{H}$, etc.) along with subscripts that indicate the variable with respect to which the partial derivatives are being taken. See Appendix A for a review of this notation.

The gradient of ℓ with respect to $\boldsymbol{\theta}$ appears frequently enough in the analysis of likelihood functions that it has earned its own name - the *score function*, or just the *score*. Formally, it is defined as

$$\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}_n) = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell(\boldsymbol{\theta}; \mathbf{x}_n) \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{\theta}_d} \ell(\boldsymbol{\theta}; \mathbf{x}_n) \end{pmatrix} = \begin{pmatrix} \mathcal{S}_1(\boldsymbol{\theta}; \mathbf{x}_n) \\ \vdots \\ \mathcal{S}_d(\boldsymbol{\theta}; \mathbf{x}_n) \end{pmatrix}, \tag{3.2.2}$$

where we think of each component as being a function $S_j : \Theta \to \mathbb{R}$.

Similarly, the Hessian matrix of the log-likelihood function with respect to $\boldsymbol{\theta}$ (i.e. the transpose of the Jacobian matrix of the score) multiplied by -1 is called the *observed information*, or just the *information*, and is denoted by

$$\mathcal{J}(\boldsymbol{\theta}) = -\mathbf{H}_{\boldsymbol{\theta}}\Big(\ell(\boldsymbol{\theta}; \mathbf{x}_n)\Big) = -\mathbf{J}_{\boldsymbol{\theta}}\Big(\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n)\Big)^{\top} = - \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_d} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_d \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_d \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_d^2} \end{pmatrix}. \tag{3.2.3}$$

The use of the term "information" here derives from the fact that the second partial derivatives of ℓ with respect to the components of $\boldsymbol{\theta}$ are all related to the curvature of ℓ near its maximum - the sharper the curve, the less uncertainty and therefore more information we have about $\boldsymbol{\theta}$.

Recall that $L(\boldsymbol{\theta}; \mathbf{x}_n)$ is defined as a function of $\boldsymbol{\theta}$ for a fixed sample of observations $\mathbf{x}_n = (x_1, ..., x_n)$, where we think of each x_i as being a realization of a random variable X_i. We may therefore interpret $L(\boldsymbol{\theta}; \mathbf{x}_n)$ as a random variable in the following sense: for a given $\boldsymbol{\theta}$, the value of $L(\boldsymbol{\theta}; \mathbf{x}_n)$ depends entirely on the values of $X_1, ..., X_n$ that we happened to observe, and so $L(\boldsymbol{\theta}; \mathbf{X}_n)$ is itself a random variable with respect to the joint probability distribution of $\mathbf{X}_n = (X_1, ..., X_n)$. The same is also true for any function or estimate based on the likelihood, as they ultimately will all depend on the data through it as well. Going forward, we will use capital letters inside these functions when we want to emphasize this interpretation. For example, $\mathcal{S}(\boldsymbol{\theta}; \mathbf{X}_n)$ is a random variable for which we have observed the value $\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n)$.

The random nature of these likelihood-based quantities further implies that finding their expectations and variances with respect to $p_{\boldsymbol{\theta}}(\mathbf{x}_n)$ is a well-defined, nontrivial task. The variance of the score function will be of particular importance, as it also relates to the amount of information pertaining to $\boldsymbol{\theta}_0$ that is contained within the log-likelihood function of our model. Properly known as the *Fisher information* or the *expected information*, it is defined as

$$\mathcal{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}[\mathcal{S}(\boldsymbol{\theta}; \mathbf{X}_n)]. \tag{3.2.4}$$

Since we are working in the more general framework in which $\mathcal{S}(\boldsymbol{\theta})$ is a $d \times 1$ random vector, it would be more accurate to speak of the *Fisher information matrix*, which is equal to the variance-covariance

matrix of $\mathcal{S}(\boldsymbol{\theta})$. Hence, we have

$$\begin{aligned}
\mathscr{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) &= \mathrm{Var}_{\boldsymbol{\theta}}[\mathcal{S}(\boldsymbol{\theta}; \mathbf{X}_n)] && \text{(by Eq. 2.2.4)} \\
&= \mathrm{Cov}_{\boldsymbol{\theta}}\left[\left(\tfrac{\partial \ell}{\partial \theta_1}, \ldots, \tfrac{\partial \ell}{\partial \theta_d} \right)^T \right] && \text{(by Eq. 2.2.2)} \\
&= \begin{pmatrix}
\mathrm{Var}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_1}\right) & \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_1}, \tfrac{\partial \ell}{\partial \theta_2}\right) & \cdots & \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_1}, \tfrac{\partial \ell}{\partial \theta_d}\right) \\
\mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_2}, \tfrac{\partial \ell}{\partial \theta_1}\right) & \mathrm{Var}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_2}\right) & \cdots & \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_2}, \tfrac{\partial \ell}{\partial \theta_d}\right) \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_d}, \tfrac{\partial \ell}{\partial \theta_1}\right) & \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_d}, \tfrac{\partial \ell}{\partial \theta_2}\right) & \cdots & \mathrm{Var}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_d}\right)
\end{pmatrix}. && (3.2.5)
\end{aligned}$$

Note that if the observations are independent, the Fisher information of the whole sample is equal to the sum of the Fisher information values for each of the observations individually. That is,

$$\mathscr{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathscr{I}_{X_i}(\boldsymbol{\theta}). \tag{3.2.6}$$

If the observations are also identically distributed according to the distribution of some random variable X, then $\mathscr{I}_{X_i}(\boldsymbol{\theta}) = \mathscr{I}_X(\boldsymbol{\theta})$ for all i, and so the Fisher information for the entire sample is simply equal to the Fisher information for a single observation of X multiplied by a factor of n:

$$\mathscr{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) = n\mathscr{I}_X(\boldsymbol{\theta}). \tag{3.2.7}$$

3.3. Maximum Likelihood Estimation

3.3.1. Motivation

Maximum likelihood estimation is one of the most powerful and widespread techniques for obtaining point estimates of model parameters. The original intuition behind the method derives from the observation that when faced with a choice between two possible values of a parameter, the sensible choice is the one which makes the data we actually did observe more probable to have been observed. We have already defined the likelihood function as a means of capturing this probability, which makes expressing this decision rule in terms of it very easy - we simply choose for our estimate the option that produces

the higher value of the likelihood function. That is, if $L(\boldsymbol{\theta}_1; \mathbf{x}_n) > L(\boldsymbol{\theta}_2; \mathbf{x}_n)$, then under the preceding logic, $\boldsymbol{\theta}_1$ is the better estimate of the true parameter value.

This can be extended to include as many candidate parameter values as we would like. For n potential estimates of $\boldsymbol{\theta}_0$, the best is the one that corresponds to the highest value of the likelihood function. Following this line of reasoning to its natural conclusion, a sensible choice for an estimate of $\boldsymbol{\theta}_0$ is any value that maximizes the likelihood function based on an observed dataset $\mathbf{x}_n$.

To help make this argument rigorous, we can show that with probability tending to 1 as the sample size tends toward infinity, the likelihood for a regular model will be strictly larger at $\boldsymbol{\theta}_0$ than for any other $\boldsymbol{\theta} \in \Theta$. We start by observing that **RC1** implies

$$L(\boldsymbol{\theta}; \mathbf{x}_n) = p(\mathbf{x}_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta}) \tag{3.3.1}$$

and

$$\ell(\boldsymbol{\theta}; \mathbf{x}_n) = \log p(\mathbf{x}_n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta}). \tag{3.3.2}$$

It follows that

$$\begin{aligned}
L(\boldsymbol{\theta}; \mathbf{x}_n) < L(\boldsymbol{\theta}_0; \mathbf{x}_n) &\iff \ell(\boldsymbol{\theta}; \mathbf{x}_n) < \ell(\boldsymbol{\theta}_0; \mathbf{x}_n) \\
&\iff \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta}) - \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta}_0) < 0 \\
&\iff \sum_{i=1}^{n} \left[\log p(x_i; \boldsymbol{\theta}) - \log p(x_i; \boldsymbol{\theta}_0) \right] < 0 \\
&\iff \sum_{i=1}^{n} \log \frac{p(x_i; \boldsymbol{\theta})}{p(x_i; \boldsymbol{\theta}_0)} < 0 \\
&\iff \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i; \boldsymbol{\theta})}{p(x_i; \boldsymbol{\theta}_0)} < 0.
\end{aligned} \tag{3.3.3}$$

Note that **RC3** guarantees that the ratio $p(x;\boldsymbol{\theta})/p(x;\boldsymbol{\theta}_0)$ is well-defined and finite for all $x \in \mathcal{X}$, the region of common support. Then by the Weak Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(X_i;\boldsymbol{\theta})}{p(X_i;\boldsymbol{\theta}_0)} \to \mathrm{E}_{\boldsymbol{\theta}_0}\left[\log\frac{p(X;\boldsymbol{\theta})}{p(X;\boldsymbol{\theta}_0)}\right] \tag{3.3.4}$$

in probability as $n \to \infty$. Furthermore,

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left[\frac{p(X;\boldsymbol{\theta})}{p(X;\boldsymbol{\theta}_0)}\right] = \int_{\mathcal{X}}\left[\frac{p(x;\boldsymbol{\theta})}{p(x;\boldsymbol{\theta}_0)}\right]p(x;\boldsymbol{\theta}_0)dx = \int_{\mathcal{X}} p(x;\boldsymbol{\theta})dx = 1. \tag{3.3.5}$$

Since $\log(x)$ is a strictly concave function, it follows from Jensen's inequality (see Appendix A) and Equation 3.3.5

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left[\log\frac{p(X;\boldsymbol{\theta})}{p(X;\boldsymbol{\theta}_0)}\right] < \log\mathrm{E}_{\boldsymbol{\theta}_0}\left[\frac{p(X;\boldsymbol{\theta})}{p(X;\boldsymbol{\theta}_0)}\right] = \log 1 = 0. \tag{3.3.6}$$

Hence, the quantity on the left-hand side of Equation 3.3.4 is converging in probability to a constant that is less than 0 as n tends to infinity. From this and the equivalence we established in Equation 3.3.3, it follows that

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}_0}\left[L(\boldsymbol{\theta};\mathbf{x}_n) < L(\boldsymbol{\theta}_0;\mathbf{x}_n)\right] = \lim_{n\to\infty} P_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(X_i;\boldsymbol{\theta})}{p(X_i;\boldsymbol{\theta}_0)} < 0\right] = 1, \tag{3.3.7}$$

which proves the claim.

Let $\hat{\boldsymbol{\theta}} \in \Theta$ be any parameter value that renders the likelihood at the observed $\mathbf{X}_n = \mathbf{x}_n$ as large as possible, i.e.,

$$L(\hat{\boldsymbol{\theta}}, \mathbf{x}_n) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}). \tag{3.3.8}$$

We call such a value a *maximum likelihood estimate* of θ_0. Note that this definition of $\hat{\theta}$ as a maximizer of $L(\hat{\boldsymbol{\theta}}, \mathbf{x}_n)$ necessarily makes it a function of the observed data. When this function is measurable, then we can further define the *maximum likelihood estimator* (MLE) of $\boldsymbol{\theta}_0$ as the statistic $\hat{\boldsymbol{\theta}}(\mathbf{X}_n)$ for which we observe the value $\hat{\boldsymbol{\theta}}(\mathbf{x}_n)$.

3.4. Regularity Conditions

As a consequence of the random variable interpretation of likelihood-based quantities, a natural line of inquiry to investigate is how the behavior of these random variables changes as the sample size n increases. Of particular interest is the distribution to which the MLE converges, if any, as n tends toward infinity. To that end, it will be useful to establish some *regularity conditions* for our models. We can think of these conditions as being assumptions similar to those we discussed in the introduction to this paper that, when satisfied, endow our models with certain properties that enable us, among other things, to determine the aforementioned distribution.

For our purposes, we will call a model *regular* if it satisfies the following conditions:

RC1) Any observations $x_1, ..., x_n$ belonging to a sample that has been drawn from the model's sample space are independent and identically distributed (i.i.d.) realizations of a random variable X with density function $p_{\boldsymbol{\theta}}(x)$.

RC2) $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2} \implies \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$.

RC3) The distributions in $\mathcal{P}$ have a common support $\mathcal{X} = \{x : p_{\boldsymbol{\theta}}(x) > 0\} \subseteq \mathbb{R}$ not depending on $\boldsymbol{\theta}$.

RC4) There exists an open set $\Theta^* \subseteq \Theta$ of which $\boldsymbol{\theta}_0$ is an interior point.

RC5) $p(x; \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$.

RC6) There exists a random function $M(x)$ (that does not depend on $\boldsymbol{\theta}$) satisfying $\mathrm{E}[M(X)] < \infty$ such that each third partial derivative of $\ell(\boldsymbol{\theta}; x)$ is bounded in absolute value by $M(x)$ uniformly in some neighborhood of $\boldsymbol{\theta}_0$.

RC7) The integral $\int_{\mathcal{X}} p(x; \boldsymbol{\theta}) dx$ can be differentiated twice under the integral sign with respect to the components of $\boldsymbol{\theta} \in \Theta^*$.

RC8) $\mathscr{I}_X(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in \Theta$.

RC9) Θ is a compact and convex subset of $\mathbb{R}^d$.

While not strictly necessary, **RC1** is often assumed as a matter of convenience since it tends to simplify calculations greatly. We include it here for that purpose and to frame our discussion in the context of a standard case in which likelihood theory holds. Of course, it is possible to construct models

lacking i.i.d. observations yet still possessing real world applications for which the results discussed in this paper hold.

The implication in **RC2** is simply the identifiability property we mentioned in Chapter 2. We repeat it here as it is necessary to guarantee the consistency of the MLE, i.e., that it converges in probability to $\boldsymbol{\theta}_0$ as $n \to \infty$.

RC3 requires that the distributions in $\mathcal{P}$ be supported on a common subset of the real line, and the definition of this subset cannot depend on $\boldsymbol{\theta}$. This is to prevent situations in which, for example, the event $\{X_i \leq x_i\}$ occurs with positive probability when $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ but not $\boldsymbol{\theta} = \boldsymbol{\theta}_2$.

RC4 guarantees the existence of an open subset Θ^* of Θ containing $\boldsymbol{\theta}_0$ as an interior point. The fact that $\boldsymbol{\theta}_0$ is an interior point of Θ^* further implies that it is possible to find a neighborhood of $\boldsymbol{\theta}_0$ that is contained in Θ^*. **RC5** then goes on to assert the existence and continuity of the first two partial derivatives with respect to the components of $\boldsymbol{\theta}$ of $p(x; \boldsymbol{\theta})$ in this neighborhood. This is a necessary requirement for defining a first-order Taylor series expansion of the score function around $\boldsymbol{\theta}_0$.

Another way of stating **RC6** is that for all $\boldsymbol{\theta}$ in a neighborhood $N_{\boldsymbol{\theta}_0}$, there exists a random function $M(x)$ with finite expectation such that

$$\sup_{\boldsymbol{\theta} \in N_{\boldsymbol{\theta}_0}} \left| \frac{\partial^3 \ell(\boldsymbol{\theta}; x)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| \leq M(x)$$

for all integers $1 \leq i, j, k \leq d$. Equivalently, we could say that the entries of the Hessian matrix for each component of the score function are all bounded by $M(x)$ as well. This ensures the remainder terms in a second-order Taylor series expansion of the score function around $\boldsymbol{\theta_0}$ become negligible as the sample size increases to infinity.

RC7 grants us the ability to freely interchange integration and second-order partial differentiation with respect to the components of $\boldsymbol{\theta}$, i.e.,

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \int_{\mathcal{X}} p(x; \boldsymbol{\theta}) dx = \int_{\mathcal{X}} \frac{\partial^2}{\partial \theta_i \partial \theta_j} p(x; \boldsymbol{\theta}) dx$$

for all $\boldsymbol{\theta} \in \Theta^*$ and $i, j = 1, ..., d$. Note that this implies first-order partial derivatives can be passed under the integral sign as well. This will prove useful in our discussion of the Bartlett identities in Section 3.5.

Finally, **RC8-9** play an important role in ensuring the existence and uniqueness of the maximum likelihood estimator of $\boldsymbol{\theta}_0$.

3.4.1. Properties

In general, there is no guarantee that an MLE for a model's parameter will exist, and even if it does, it will not necessarily be unique. However, since maximum likelihood estimation plays such an important role in our discussion in Chapters 5 and 6, it would come as a great convenience if we possessed the ability to speak freely of *the* MLE of a model's parameter without having to clarify *which* MLE we mean or whether one even exists at all. Hence, some discussion of the conditions under which the MLE of a model's parameter exists and is unique is warranted.

The extreme value theorem (see Appendix A) implies that sufficient conditions for the existence of a model's MLE are that Θ is compact, and $\ell(\boldsymbol{\theta}; x)$ is continuous on Θ. The former is satisfied directly by **RC9** and the latter is implied through our assumption in **RC5** of the differentiability of $p(x; \boldsymbol{\theta})$ in $\boldsymbol{\theta}$. Therefore, at least one MLE will always exist for the true parameter value of a regular model. These are only sufficient conditions, however, and not necessary; MLEs may exist for parameters of non-regular models as wel.

Similarly, when the MLE does exist, a sufficient condition for its uniqueness is that Θ is convex, and $\ell(\boldsymbol{\theta}; x)$ is strictly concave on Θ, as this ensures that it has exactly one global maximum, which it attains at the $\hat{\boldsymbol{\theta}}$. **RC9** directly satisfies the compactness criterion. Our assumption in **RC8** that the Fisher information matrix is positive definite forces the Hessian matrix of $\ell(\boldsymbol{\theta}; x)$ to be negative definite. This in turn implies that $\ell(\boldsymbol{\theta}; x)$ is strictly concave, so the requirement is met. Hence, a regular model will always have a unique MLE for its parameter.

As a global maximizer of the log-likelihood function, the MLE $\hat{\boldsymbol{\theta}}$ must be a root of the log-likelihood function, i.e., it must satisfy the *likelihood equation*,

$$\nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{3.4.1}$$

whenever it exists. It is important to note that for an arbitrary model there may be other roots as well, even when the MLE doesn't exist. Assuming **RC2** and **RC5-8**, it can be shown that there will always be at least one sequence of roots $\hat{\boldsymbol{\theta}}_n$ of its log-likelihood such that $\hat{\boldsymbol{\theta}}_n$ tends to $\boldsymbol{\theta}_0$ in probability as $n \to \infty$ (Cf. Cramér 1945). The MLE will not necessarily be a part of this sequence though, even if it exists. Adding **RC9** is enough to ensure the MLE must be the unique solution to the likelihood equation, however, and therefore this sequence of roots will also be unique and for a given sample $\mathbf{x}_n$, the corresponding root $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}(\mathbf{x}_n)$ will be the unique MLE of $\boldsymbol{\theta}_0$. It follows that the MLE is a consistent estimator of $\boldsymbol{\theta}_0$ for regular models. We will explore the asymptotic properties of the MLE in more detail in Section 3.6.

One last useful property of the maximum likelihood estimator is its functional invariance. If $\hat{\boldsymbol{\theta}}$ is an MLE of $\boldsymbol{\theta}_0$, then any function $h(\boldsymbol{\theta}_0)$ will have $h(\hat{\boldsymbol{\theta}})$ as its MLE. Hence, it is straightforward to find the MLE of a model that has undergone a reparameterization given that we know the MLE of the original parameter.

3.5. The Bartlett Identities

The Bartlett identities are a set of equations relating to the expectations of the derivatives of a log-likelihood function to one another. In general, there is no guarantee that an arbitrary function of a random variable X and its parameter $\boldsymbol{\theta}$ will satisfy the Bartlett identities. It is guaranteed, however, that the log-likelihood function associated with X and $\boldsymbol{\theta}$ will satisfy them, provided that the model is regular. Thus, we can think of any function that does satisfy the Bartlett identities (or at least some of them) as resembling that of a genuine log-likelihood.

Consider the case where a random variable X has density function $p_\theta(x)$, where θ is a scalar. For a single observation $X = x$, the expectation of $\frac{\partial}{\partial \theta} \ell(\theta; X)$ gives

$$
\begin{aligned}
E_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta;X)\right] &= \int_{\mathbb{R}}\left[\frac{\partial}{\partial\theta}\log p(x;\theta)\right]p(x;\theta)dx \\
&= \int_{\mathbb{R}}\frac{\frac{\partial}{\partial\theta}p(x;\theta)}{p(x;\boldsymbol{\theta})}p(x;\theta)dx \\
&= \int_{\mathbb{R}}\frac{\partial}{\partial\theta}p(x;\theta)dx \\
&= \frac{d}{d\theta}\int_{\mathbb{R}}p(x;\theta)dx \\
&= \frac{d}{d\theta}1 \\
&= 0.
\end{aligned} \tag{3.5.1}
$$

Equation 3.5.1 is called the first Bartlett identity. In words, it states that the expectation of the first partial derivative of the log-likelihood function of a statistical model with respect to the parameter will always be 0. Since the score is defined as $\frac{\partial}{\partial\theta}\ell(\theta;x)$, any function that satisfies the first Bartlett identity is said to be *score-unbiased*.

For any model with a log-likelihood satisfying the first Bartlett identity, the expected information for its parameter θ may be rewritten as

$$
\begin{aligned}
\mathscr{I}_X(\theta) &= \mathrm{Var}_\theta[\mathcal{S}(\theta;X)] \\
&= \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta;X)\right] \\
&= \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta;X)\right] + \left(\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta;X)\right]\right)^2 \quad \text{(by the first Bartlett identity)} \\
&= \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ell(\theta;X)\right)^2\right].
\end{aligned} \tag{3.5.2}
$$

If we now consider the second partial derivative of $\ell(\theta; x)$ with respect to θ, we have

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\ell(\theta;x) &= \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\ell(\theta;x)\right] \\
&= \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\log p(x;\theta)\right] \\
&= \frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}p(x;\theta)}{p(x;\theta)}\right] \\
&= \frac{\left[\frac{\partial^2}{\partial\theta^2}p(x;\theta)\right]p(x;\theta) - \left[\frac{\partial}{\partial\theta}p(x;\theta)\right]\left[\frac{\partial}{\partial\theta}p(x;\theta)\right]}{\left[p(x;\theta)\right]^2} \\
&= \frac{\frac{\partial^2}{\partial\theta^2}p(x;\theta)}{p(x;\theta)} - \left[\frac{\frac{\partial}{\partial\theta}p(x;\theta)}{p(x;\theta)}\right]^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2}p(x;\theta)}{p(x;\theta)} - \left[\frac{\partial}{\partial\theta}\log p(x;\theta)\right]^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2}p(x;\theta)}{p(x;\theta)} - \left[\frac{\partial}{\partial\theta}\ell(\theta;x)\right]^2.
\end{aligned}$$

Rearranging terms and taking expectations yields

$$\begin{aligned}
\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta;X)\right] + \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ell(\theta;X)\right)^2\right] &= \mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}p(X;\theta)}{p(X;\theta)}\right] \\
&= \int_{\mathbb{R}}\left[\frac{\frac{\partial^2}{\partial\theta^2}p(x;\theta)}{p(x;\theta)}\right]p(x;\theta)dx \\
&= \int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}p(x;\theta)dx \\
&= \frac{d^2}{d\theta^2}\int_{\mathbb{R}}p(x;\theta)dx \\
&= \frac{d^2}{d\theta^2}1 \\
&= 0.
\end{aligned}$$

Therefore,

$$\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta;X)\right] + \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ell(\theta;X)\right)^2\right] = 0. \tag{3.5.3}$$

Equation 3.5.2 is called the second Bartlett identity. Any function that satisfies it is said to be *information-unbiased*. Regular models as we have defined them will automatically satisfy both the first and second Bartlett identities. Hence, for any regular model, the statements in Equation 3.2.4, Equation 3.5.2, and Equation 3.5.3 imply that the following definitions for its expected information regarding its parameter θ are all equivalent:

$$\mathscr{I}_X(\theta) = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta;X)\right] = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ell(\theta;X)\right)^2\right] = \mathrm{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta;X)\right] = \mathrm{E}_\theta[\mathcal{I}_X(\theta)]. \tag{3.5.4}$$

It is possible to derive further Bartlett identities by continuing in this manner for an arbitrary number of partial θ-derivatives of the log-likelihood function, provided that they exist. However, the first two are sufficient for our purposes of evaluating the validity of approximations to genuine likelihoods so we will not go further here. Note that while the above derivations were performed under the assumption that θ is a scalar, the Bartlett identities also hold in the case where $\boldsymbol{\theta}$ is a $d \times 1$ vector.

3.6. One-Index Asymptotics

The one-index asymptotics framework describes the behavior of likelihood-based statistics as the sample size (n) grows to infinity while the dimension of the nuisance parameter (q) remains fixed. The aim of this section is to present a basic overview of the theory's results so that we will have a readily available baseline against which to compare the results of the following section discussing the two-index asymptotics framework, in which q is allowed to increase with n.

Assume the regularity conditions of the previous section apply and let $\hat{\boldsymbol{\theta}}$ denote the MLE for $\boldsymbol{\theta}_0$. **RC1** implies the score function is equal to

$$\begin{aligned}
\mathcal{S}(\boldsymbol{\theta};\mathbf{x}_n) &= \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathbf{x}_n) \\
&= \nabla_{\boldsymbol{\theta}}\sum_{i=1}^{n}\ell(\boldsymbol{\theta};x_i) \\
&= \sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};x_i) \\
&= \sum_{i=1}^{n}\mathcal{S}(\boldsymbol{\theta};x_i).
\end{aligned} \tag{3.6.1}$$

where the last equality is true by . In other words, the score function for the parameter $\boldsymbol{\theta}$ based on data $x_1, ..., x_n$ can be written as the sum of independent contributions $\mathcal{S}(\boldsymbol{\theta}; x_i)$ $(i = 1, ..., n)$ where each $\mathcal{S}(\boldsymbol{\theta}; x_i)$ can be thought of as the score function for $\boldsymbol{\theta}$ based only on observation x_i. This implies that a Taylor series expansion of $\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n)$ will be equal to the sum of the Taylor series expansions of its individual contributions, plus a remainder term that depends on n. Since the observations are identically distributed, it suffices to consider the expansion for an arbitrary contribution, $\mathcal{S}(\boldsymbol{\theta}; x_i)$.

RC4-5 guarantee the existence of a neighborhood of $\boldsymbol{\theta}_0$ on which the first two partial derivatives of $\mathcal{S}(\boldsymbol{\theta}; x_i)$ with respect to $\boldsymbol{\theta}$ exist and are continuous. Without loss of generality, we may assume this neighborhood, call it $N_{\boldsymbol{\theta}_0}$, is convex so that it contains all of the line segments connecting any two of its points. In particular, for any $\boldsymbol{\theta} \in N_{\boldsymbol{\theta}_0}$, $\text{LS}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \subset N_{\boldsymbol{\theta}_0}$. Then by Taylor's theorem with the Lagrange form of the remainder, there exists $\bar{\boldsymbol{\theta}}_j \in \text{LS}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ such that the j-th component of $\mathcal{S}(\boldsymbol{\theta}; x_i)$ may be expanded as

$$\begin{aligned}
\mathcal{S}_j(\boldsymbol{\theta}; x_i) &= \mathcal{S}_j(\boldsymbol{\theta}_0; x_i) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}}\mathcal{S}_j(\boldsymbol{\theta}_0; x_i) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\mathbf{H}_{\boldsymbol{\theta}}\Big(\mathcal{S}_j(\bar{\boldsymbol{\theta}}_j; x_i)\Big)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \\
&= \mathcal{S}_j(\boldsymbol{\theta}_0; x_i) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)\Big[\nabla_{\boldsymbol{\theta}}\mathcal{S}_j(\boldsymbol{\theta}_0; x_i) + \frac{1}{2}\mathbf{H}_{\boldsymbol{\theta}}\Big(\mathcal{S}_j(\bar{\boldsymbol{\theta}}_j; x_i)\Big)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top\Big] \\
&= \mathcal{S}_j(\boldsymbol{\theta}_0; x_i) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)\Big[\nabla_{\boldsymbol{\theta}}\mathcal{S}_j(\boldsymbol{\theta}_0; x_i) + M(x_i)O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\Big] \qquad \text{(by } \mathbf{RC6}\text{)} \\
&= \mathcal{S}_j(\boldsymbol{\theta}_0; x_i) + \Big[\nabla_{\boldsymbol{\theta}}^\top\mathcal{S}_j(\boldsymbol{\theta}_0; x_i) + M(x_i)O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\Big](\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top.
\end{aligned}$$

When we stack each of these individual equations into a system of equations, we get

$$\begin{pmatrix} \mathcal{S}_1(\boldsymbol{\theta}; x_i) \\ \vdots \\ \mathcal{S}_d(\boldsymbol{\theta}; x_i) \end{pmatrix} = \begin{pmatrix} \mathcal{S}_1(\boldsymbol{\theta}_0; x_i) \\ \vdots \\ \mathcal{S}_d(\boldsymbol{\theta}_0; x_i) \end{pmatrix} + \left[\begin{pmatrix} \nabla^\top\mathcal{S}_1(\boldsymbol{\theta}_0; x_i) \\ \vdots \\ \nabla^\top\mathcal{S}_d(\boldsymbol{\theta}_0; x_i) \end{pmatrix} + M(x_i)O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d} \right](\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top. \tag{3.6.2}$$

The matrix of gradient vectors in the second term on the right-hand side of the above equation is simply the Jacobian of the score function evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, i.e., $\mathbf{J}_{\boldsymbol{\theta}}\Big(\mathcal{S}(\boldsymbol{\theta}_0; x_i)\Big)$. However, by Equation 3.2.3, this is just the negative transpose of the observed information matrix also evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\mathcal{I}(\boldsymbol{\theta}_0)$. Furthermore, since we have assumed ℓ is continuous in $\boldsymbol{\theta}$, we can freely swap the order of differentiation in all of its second partial derivatives with respect to $\boldsymbol{\theta}$. This implies the Jacobian of the score will be a

symmetric matrix, and so we simply have $\mathbf{J}_{\boldsymbol{\theta}}\Big(\mathcal{S}(\boldsymbol{\theta}_0; x_i)\Big) = -\mathcal{J}_i(\boldsymbol{\theta}_0)$. Hence, using more compact notation, Equation 3.6.2 becomes

$$\begin{aligned}\mathcal{S}(\boldsymbol{\theta}; x_i) &= \mathcal{S}(\boldsymbol{\theta}_0; x_i) + \Big[\mathbf{J}_{\boldsymbol{\theta}}\Big(\mathcal{S}(\boldsymbol{\theta}_0; x_i)\Big) + M(x_i)O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d}\Big](\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \\ &= \mathcal{S}(\boldsymbol{\theta}_0; x_i) - \Big[\mathcal{J}_i(\boldsymbol{\theta}_0) + M(x_i)O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d}\Big](\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top\end{aligned} \tag{3.6.3}$$

The above represents the second-order Taylor expansion around $\boldsymbol{\theta}_0$ for an individual observation x_i's contribution to the score function. Summing over all of these contributions yields

$$\begin{aligned}\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n) &= \sum_{i=1}^{n} \mathcal{S}(\boldsymbol{\theta}; x_i) \\ &= \sum_{i=1}^{n} \left[\mathcal{S}(\boldsymbol{\theta}_0; x_i) - \Big[\mathcal{J}_i(\boldsymbol{\theta}_0) + M(x_i)O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d}\Big](\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top\right] \\ &= \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) - \left[\mathcal{J}_n(\boldsymbol{\theta}_0) + \left\{\sum_{i=1}^{n} M(x_i)\right\}O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d}\right](\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \\ &= \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) - \left[\frac{1}{n}\mathcal{J}_n(\boldsymbol{\theta}_0) + \left\{\frac{1}{n}\sum_{i=1}^{n} M(x_i)\right\}O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d}\right] n(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top.\end{aligned} \tag{3.6.4}$$

If we divide through by $\sqrt{n}$, we arrive at

$$\frac{1}{\sqrt{n}}\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n) = \frac{1}{\sqrt{n}}\mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) - \left[\frac{1}{n}\mathcal{J}_n(\boldsymbol{\theta}_0) + \left\{\frac{1}{n}\sum_{i=1}^{n} M(x_i)\right\}O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d}\right]\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top. \tag{3.6.5}$$

We established in Section 3.4.1 that regular models will always have a sequence of MLEs $\hat{\boldsymbol{\theta}}_n$ that converge in probability to $\boldsymbol{\theta}_0$ as $n \to \infty$, and that each $\hat{\boldsymbol{\theta}}_n$ in this sequence will satisfy $\mathcal{S}(\hat{\boldsymbol{\theta}}_n; \mathbf{x}_n) = \mathbf{0}$. Plugging $\hat{\boldsymbol{\theta}}_n$ in for $\boldsymbol{\theta}$ in Equation 3.6.5 gives

$$\frac{1}{\sqrt{n}}\mathcal{S}(\hat{\boldsymbol{\theta}}_n; \mathbf{x}_n) = \frac{1}{\sqrt{n}}\mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) - \left[\frac{1}{n}\mathcal{J}_n(\boldsymbol{\theta}_0) + \left\{\frac{1}{n}\sum_{i=1}^{n} M(x_i)\right\}O(||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d}\right]\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top$$

and therefore,

$$\frac{1}{\sqrt{n}}\mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) = \left[\frac{1}{n}\mathcal{J}_n(\boldsymbol{\theta}_0) + \left\{\frac{1}{n}\sum_{i=1}^{n} M(x_i)\right\}O_p(||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d}\right]\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top. \tag{3.6.6}$$

Note the following observations about the terms in the square brackets in the line above:

1) By Equation 3.5.4, $\mathrm{E}_{\boldsymbol{\theta}_0}[\mathcal{I}(\boldsymbol{\theta}_0)] = \mathscr{I}(\boldsymbol{\theta}_0)$, and thus $\frac{1}{n}\mathcal{I}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\boldsymbol{\theta}_0)$ is converging in probability to $\mathscr{I}(\boldsymbol{\theta}_0)$ as $n \to \infty$ by the Weak Law of Large Numbers, i.e., $\frac{1}{n}\mathcal{I}_n(\boldsymbol{\theta}_0) = \mathscr{I}(\boldsymbol{\theta}) + o_p(1)$.
2) $M(x_i)$ has finite expectation and does not depend on $\boldsymbol{\theta}$ by **RC6**, which implies through Markov's inequality that it is bounded in probability as $n \to \infty$, i.e., $M(x_i) = O_p(1)$. It follows that $\frac{1}{n}\sum_{i=1}^{n} M(x_i) = O_p(1)$ as well.
3) The fact that $\hat{\boldsymbol{\theta}}_n$ is converging in probability to $\boldsymbol{\theta}_0$ as $n \to \infty$ implies that $||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||$ is $o_p(1)$.

From these three facts we can conclude that the entire term inside the square brackets is converging in probability to $\mathscr{I}(\boldsymbol{\theta})$ as $n \to \infty$. This allow us to rewrite Equation 3.6.6 as

$$\frac{1}{\sqrt{n}}\mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) = \Big[\mathscr{I}(\boldsymbol{\theta}_0) + o_p(1)\Big]\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top. \tag{3.6.7}$$

CHAPTER 4

Pseudolikelihood Functions

4.1. Model Parameter Decomposition

It is often the case that we are not interested in estimating the full parameter $\boldsymbol{\theta} \in \boldsymbol{\theta} \subseteq \mathbb{R}^d$, but rather a different parameter ψ taking values in a set $\Psi \subseteq \mathbb{R}^p$, where $p < d$. In such an event, we refer to ψ as the *parameter of interest*. Crucially, as we will see, ψ can always be expressed as a function of $\boldsymbol{\theta}$.

Since ψ is of lower dimension than $\boldsymbol{\theta}$, it necessarily follows that there is another parameter λ, taking values in a set $\Lambda \subseteq \mathbb{R}^q$, where $p+q = d$, that is made up of whatever is "left over" from the full parameter $\boldsymbol{\theta}$. We refer to λ as the *nuisance parameter*, so named for its ability to complicate inference regarding the parameter of interest. Despite not being the object of study themselves, nuisance parameters are nevertheless capable of modifying the distributions of our observations and therefore must be accounted for when conducting inference or estimation regarding the parameter of interest.[1] The process by which this is accomplished is nontrivial and often represents a serious obstacle that must be overcome.

While not required, we will assume the parameter of interest ψ is always one-dimensional. That is, $\Psi \subseteq \mathbb{R}$ and consequently $\Lambda \subseteq \mathbb{R}^{d-1}$. This restriction reflects the common habit of researchers to focus on scalar-valued summaries of vector quantities. For example, suppose we observe data $Y = (y_1, ..., y_n)$, where each y_i is the outcome of some random variable $Y_i \sim N(\mu_i, \sigma_i^2)$, and we are interested in estimating the average of the population means, $\frac{1}{n}\sum_{i=1}^{n} \mu_i$. Rather than defining $\psi = (\mu_1, ..., \mu_n)$, we can instead define $\psi = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ directly, bypassing the need to estimate each μ_i individually before taking their average. This does carry the trade-off of increasing the dimension of the nuisance parameter, which must be dealt with before conducting inference or estimation on ψ. We will examine some of the issues posed by high-dimensional nuisance parameters in greater detail in the next chapter.

[1] Note that nuisance parameters are not always uniquely defined. Depending on the choice of parameter of interest, there may be multiple or even infinite ways to define a nuisance parameter.

4.1.1. Explicit Parameters

Parameters of interest and nuisance parameters can be broadly classified into two categories, explicit or implicit. For a given statistical model, both types of parameter must occupy the same category - it is not possible for ψ to be explicit and λ to be implicit, or vice versa.

Let us first consider the case in which ψ and λ are *explicit* parameters. This means that ψ is a sub-vector of $\boldsymbol{\theta}$, so that all the components of ψ are also components of $\boldsymbol{\theta}$. Then there exists a set $I = \{I_1, ..., I_p\} \subsetneq \{1, ..., d\}$ such that

$$\psi = (\boldsymbol{\theta}_{I_1}, ..., \boldsymbol{\theta}_{I_p}). \tag{4.1.1}$$

It immediately follows that λ is the sub-vector of all components of $\boldsymbol{\theta}$ that are not part of ψ. More precisely, if we let $J = \{J_1, ..., J_q\} \subsetneq \{1, ..., d\}$ such that $I \cup J = \{1, ..., d\}$ and $I \cap J = \emptyset$, then

$$\lambda = (\boldsymbol{\theta}_{J_1}, ..., \boldsymbol{\theta}_{J_q}). \tag{4.1.2}$$

$\boldsymbol{\theta}$ can therefore be decomposed as $\boldsymbol{\theta} = (\psi, \lambda)$ when ψ and λ are explicit, provided we shuffle the indices appropriately.

4.1.2. Implicit Parameters

Now let us consider the case in which ψ and λ are *implicit* parameters. This means there exists some function $\varphi : \boldsymbol{\theta} \to \Psi$ for which the parameter of interest can be written as

$$\psi = \varphi(\boldsymbol{\theta}). \tag{4.1.3}$$

As before, Ψ is still assumed to be a subset of $\mathbb{R}^p$ where p is less than d. This reduction in dimension again implies the existence of a nuisance parameter $\lambda \in \Lambda \subseteq \mathbb{R}^{k-m}$. However, unlike in the explicit case, a closed form expression for λ in terms of the original components of $\boldsymbol{\theta}$ need not exist. For this reason, implicit nuisance parameters are in general more difficult to eliminate compared to their explicit counterparts.

Note that when the parameter of interest and nuisance parameter are explicit, it is always possible to define a function φ such that

$$\varphi(\boldsymbol{\theta}) = (\boldsymbol{\theta}_{I_1}, ..., \boldsymbol{\theta}_{I_p}) \equiv \psi, \tag{4.1.4}$$

where $\{I_1, ..., I_p\}$ is defined as above. Hence, the first case is really just a special example of this more general one in which $\psi = \varphi(\boldsymbol{\theta})$. With this understanding in mind, we will use the notation $\psi = \varphi(\boldsymbol{\theta})$ to refer to the parameter of interest in general, only making the distinction between implicitness and explicitness when the difference is relevant to the situation.

4.2. Types of Pseudolikelihoods

The natural solution to the hindrance nuisance parameters pose to making inferences on the parameter of interest is to find a method for eliminating them from the likelihood function altogether. The result of this elimination is what is known as a pseudolikelihood function.

In general, a *pseudolikelihood function* for ψ is a function of the data and ψ only, having properties resembling that of a genuine likelihood function. Suppose $\psi = \varphi(\boldsymbol{\theta})$ for some function ϕ and parameter $\boldsymbol{\theta} \in \boldsymbol{\theta}$. If we let $\boldsymbol{\theta}(\psi) = \{\boldsymbol{\theta} \in \boldsymbol{\theta} \ : \ \varphi(\boldsymbol{\theta}) = \psi\}$, then associated with each $\psi \in \Psi$ is the set of likelihoods $\mathcal{L}_\psi = \{L(\boldsymbol{\theta}) \ : \ \boldsymbol{\theta} \in \boldsymbol{\theta}(\psi)\}$.

Any summary of the values in $\mathcal{L}_\psi$ that does not depend on λ theoretically constitutes a pseudolikehood function for ψ. There exist a variety of methods to obtain this summary but among the most popular are profiling (maximization), conditioning, and integration, each with respect to the nuisance parameter. We will explore each of these methods in more detail in the following sections.

4.2.1. The Profile Likelihood

The profile likelihood is the most straightforward method for eliminating a nuisance parameter from a likelihood function.

For example, suppose we are interested in estimating the mean of a random variable Y, where $Y \sim N(\mu, \sigma^2)$. The full model parameter is $\boldsymbol{\theta} = (\mu, \sigma^2)$ but since we are only interested in estimating the mean, the parameter of interest is $\psi = \mu$ and the nuisance parameter is $\lambda = \sigma^2$.

4.2.2. The Conditional Likelihood

4.2.3. The Marginal Likelihood

4.2.4. The Integrated Likelihood

4.3. The Bartlett Identities Revisited

The Bartlett identities offer an alternative way of characterizing the difference between likelihood and pseudolikelihood functions. A genuine likelihood function of $\boldsymbol{\theta}$ can be characterized as any nonnegative random function of $\boldsymbol{\theta}$ for which all of the Bartlett identities hold. Similarly, we can think of a pseudolikelihood function of $\boldsymbol{\theta}$ as behing any nonnegative random function of $\boldsymbol{\theta}$ for which at least one of the Bartlett identities does not hold. Hence, the identities act as a litmus test of sorts for determining the validity of a pseudolikelihood as an approximation to the genuine likelihood from which it originated - the more identities it does satisfy, the better the approximation.

CHAPTER 5

Approximating the Integrated Likelihood Function

5.1. The Zero-Score Expectation Parameter

Let $\psi = \varphi(\boldsymbol{\theta})$ and λ denote the parameter of interest and nuisance parameter, respectively, for some statistical model $(\mathcal{S}, \mathcal{P}_{\boldsymbol{\theta}})$. Then the general expression to obtain an integrated likelihood for ψ may be written as

$$\bar{L}(\psi) = \int_{\Lambda} L(\psi, \lambda)\pi(\lambda|\psi)d\lambda, \tag{5.1.1}$$

where $\pi(\lambda|\psi)$ is a conditional prior density for λ given ψ.

Severini (2007) considered the problem of selecting $\pi(\lambda|\psi)$ such that when the likelihood function is integrated with respect to this density, the result is useful for non-Bayesian inference. To do this, he outlined four properties (see Appendix B) that an integrated likelihood function must satisfy if it is to be of any use. He went on to prove that an integrated likelihood satisfying these properties could be obtained by first constructing a new nuisance parameter $\phi \in \Phi$ that is unrelated to the parameter of interest (in the sense that its maximum likelihood estimator remains roughly constant for all values of ψ) and then choosing a prior density $\pi(\phi)$ that is independent of ψ. Once chosen, the desired integrated likelihood function for ψ is given by

$$\bar{L}(\psi) = \int_{\Phi} \tilde{L}(\psi, \phi)\pi(\phi)d\phi, \tag{5.1.2}$$

where $\tilde{L}(\psi, \phi)$ is the likelihood function for the model after it has been reparameterized in terms of ϕ. It is important to note that the exact choice of prior density for ϕ is not particularly important; the only restriction we place upon it is that it must not depend on ψ.

Suppose that we have an explicit parameter of interest and nuisance parameter, so that $\boldsymbol{\theta} = (\psi, \lambda)$. Then Severini (2007) defines this new nuisance parameter ϕ as the solution to the equation

$$\mathbb{E}(\ell_\lambda(\psi, \lambda); \psi_0, \lambda_0)\Big|_{(\psi_0, \lambda_0) = (\widehat{\psi}, \phi)} = 0, \tag{5.1.3}$$

where $\ell_\lambda(\psi, \lambda) = \frac{\partial \ell(\psi, \lambda)}{\partial \lambda}$, ψ_0 and λ_0 denote the true values of ψ and λ, and $\widehat{\psi}$ is the MLE for ψ_0. In other words, for a particular value of $(\psi, \lambda, \widehat{\psi})$, we can find the corresponding value of ϕ by solving for it in Equation 5.1.3. ϕ is called the *zero-score expectation* (ZSE) parameter because it is defined as the value that makes the expectation of the score function (where the derivative is taken with respect to λ) evaluated at the point $(\widehat{\psi}, \phi)$ equal to zero. Note that ϕ is a function of the data through $\widehat{\psi}$. Normally we avoid creating such dependencies in our parameters as it renders them useless for the purpose of parameterizing a statistical model. However, from the perspective of the likelihood function, once the data have been collected they are considered fixed in place and there is no issue with using a quantity such as ϕ that depends on the data to parameterize it.

For a given value of $(\psi, \phi, \widehat{\psi})$, it is also possible to solve Equation 5.1.3 for λ. This allows us to write Equation 5.1.2 in terms of $L(\psi, \lambda)$:

$$\bar{L}(\psi) = \int_\Phi L(\psi, \lambda(\psi, \phi))\pi(\phi)d\phi. \tag{5.1.4}$$

Severini (2018) proved that reparameterizing the nuisance parameter in terms of the ZSE parameter yields the same desirable properties in the subsequent integrated likelihood when ψ and λ are implicit. Suppose $\psi = \varphi(\boldsymbol{\theta})$, for some function $\varphi : \boldsymbol{\theta} \to \Psi$, and consider the set of all values of $\boldsymbol{\theta}$ satisfying $\varphi(\boldsymbol{\theta}) = \widehat{\psi}$. Call this set $\Omega_{\widehat{\psi}}$ so that

$$\Omega_{\widehat{\psi}} = \left\{\omega \in \boldsymbol{\theta} : \varphi(\omega) = \widehat{\psi}\right\}. \tag{5.1.5}$$

Elements of $\Omega_{\widehat{\psi}}$ take the form $(\widehat{\psi}, \phi)$, where $\phi \in \Lambda$.

5.1.1. Weight Functions

5.2. Two-Index Asymptotics

Earlier we discussed the one-index asymptotics setting, in which the sample size (n) of the model diverged to infinity while the dimension of the nuisance parameter (q) remained fixed. Now we turn our attention to the two-index asymptotics setting which describes the behavior of likelihood and pseudolikelihood functions as n and q both tend to infinity, with q growing at least as fast as n. Under such a framework, De Bin, Sartori, and Severini (2015) showed that estimates for ψ based on a suitably constructed integrated likelihood function will outperform those coming from more traditional pseudolikelihoods, such as the profile likelihood. Such findings provide the motivation for our ensuing examination of two-index asymptotics theory, insofar as it relates to the performance of the integrated likelihood function as a method of inference regarding a parameter of interest.

To mirror the strategy used by Sartori (2003) and De Bin, Sartori, and Severini (2015), we will frame our discussion in terms of a stratified sample of data in which each stratum contributes one component to the overall nuisance parameter. Consider a model with parameter $\boldsymbol{\theta} = (\psi, \lambda)$ where ψ is the parameter of interest and $\lambda = (\lambda_1, ..., \lambda_q)$ is a q-dimensional nuisance parameter. For the sake of reducing complexity in our notation, we will only consider the case in which ψ and the individual components of λ are scalar parameters, though the results of this section should hold in the case where they are all vectors as well. Suppose that we have divided the model's population into q strata and collected a sample of size m_i from each stratum such that observation j from stratum i may be modeled as

$$X_{ij} \sim p_{ij}(x_{ij}; \psi, \lambda_i), \tag{5.2.1}$$

where $i = 1, ..., q$ and $j = 1, ..., n_i$, making the total sample size $n = \sum_{i=1}^{q} m_i$.[1] Hence, there is a one-to-one correspondence between the strata and the components of λ; assume that each $\lambda_i \in \Lambda$, where the space Λ is the same for all i, and they all have the same interpretation within their respective strata.

[1]It will be convenient to work under the restriction that the stratum sample sizes are all identical, meaning there exists some positive integer m such that $m_i = m$ for all i. However, as both Sartori (2003) and De Bin, Sartori, and Severini (2015) note, we could also assume a looser condition in which $m_i = K_i m$ where $0 < K_i < \infty$ without compromising our results.

Assume that all the regularity conditions set forth in Section 3.4 apply, except possibly for RC1) - it is not necessary to assume that the observations are i.i.d. here, and in fact it is perfectly acceptable for the p_{ij}'s in Equation 5.2.1 to differ from one another. We will also allow for the possibility of dependence among observations within a stratum, though not between them.

Let $\mathbf{x}_i = (x_{i1}, ..., x_{im})$ denote the sample of observations from stratum i, so that their joint density may be written as $p_i(\mathbf{x}_i; \psi, \lambda_i)$. Therefore, the likelihood and log-likelihood for the ith stratum are

$$L^{(i)}(\psi, \lambda_i) = p_i(\mathbf{x}_i; \psi, \lambda_i), \tag{5.2.2}$$

and

$$\ell^{(i)}(\psi, \lambda_i) = \log L^{(i)}(\psi, \lambda_i), \tag{5.2.3}$$

respectively. For a particular choice of weight function $g(\lambda_i; \psi)$, the integrated likelihood for ψ in stratum i is given by

$$\bar{L}^{(i)}(\psi) = \int_\Lambda L^{(i)}(\psi, \lambda_i) g(\lambda_i; \psi) d\lambda_i. \tag{5.2.4}$$

From here, we proceed by using Laplace's method as described by Tierney and Kadane (1986) (see Appendix B for a brief review) to obtain an analytic approximation to $\bar{L}^{(i)}(\psi)$. Setting $h(\lambda_i) = -\frac{1}{m}\ell^{(i)}(\psi, \lambda_i)$ and $f(\lambda_i) = g(\lambda_i; \psi)$, we may rewrite the integral in Equation 5.2.4 as

$$\bar{L}^{(i)}(\psi) = \int_\Lambda f(\lambda_i) \exp[-mh(\lambda_i)] d\lambda_i.$$

One consequence of the regularity conditions we have assumed is that $L^{(i)}(\psi, \lambda_i)$ is that an MLE for $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}$, exists and is unique. This further implies the existence and uniqueness of a conditional MLE for the true value of each stratum-specific nuisance parameter given ψ - denote this value for the ith stratum by $\hat{\lambda}_{i\psi}$. By definition this value maximizes $\ell^{(i)}(\psi, \lambda_i)$ as a function of λ_i, and so it also maximizes $-h(\lambda_i)$ since the two functions differ only by a multiplicative constant $\frac{1}{m}$.

The Laplace approximation to $\bar{L}^{(i)}(\psi)$ is then given by

$$\hat{\bar{L}}^{(i)}(\psi) = f(\hat{\lambda}_{i\psi})\sqrt{\frac{2\pi}{m}}\sigma \exp[-mh(\hat{\lambda}_{i\psi})], \tag{5.2.5}$$

where

$$\begin{aligned}\sigma &= \left[\left.\frac{\partial^2 h}{\partial \lambda_i^2}\right|_{\lambda_i=\hat{\lambda}_{i\psi}}\right]^{-1/2} \\ &= \left[-\frac{1}{m}\left.\frac{\partial^2 \ell^{(i)}(\psi,\lambda_i)}{\partial \lambda_i^2}\right|_{\lambda_i=\hat{\lambda}_{i\psi}}\right]^{-1/2} \\ &= \left[\frac{1}{m}\mathcal{J}(\hat{\lambda}_{i\psi})\right]^{-1/2}.\end{aligned}$$

Here, $\mathcal{J}(\hat{\lambda}_{i\psi})$ denotes the observed information function for λ_i only (i.e. the negative second partial derivative of the log-likelihood with respect to λ_i) evaluated at $\hat{\lambda}_{i\psi}$. Plugging in the appropriate quantities for f, h, and σ into Equation 5.2.5, we arrive at

$$\begin{aligned}\hat{\bar{L}}^{(i)}(\psi) &= f(\hat{\lambda}_{i\psi})\sqrt{\frac{2\pi}{m}}\sigma \exp[-mh(\hat{\lambda}_{i\psi})] \\ &= g(\hat{\lambda}_{i\psi};\psi)\sqrt{\frac{2\pi}{m}}\left[\frac{1}{m}\mathcal{J}(\hat{\lambda}_{i\psi})\right]^{-1/2}\exp\left\{-m\cdot -\frac{1}{m}\ell^{(i)}(\psi,\hat{\lambda}_{i\psi})\right\} \\ &= \frac{\sqrt{2\pi}}{m}L^{(i)}(\psi,\hat{\lambda}_{i\psi})g(\hat{\lambda}_{i\psi};\psi)[\mathcal{J}(\hat{\lambda}_{i\psi})]^{-1/2} \\ &= \frac{\sqrt{2\pi}}{m}L_P^{(i)}(\psi)g(\hat{\lambda}_{i\psi};\psi)[\mathcal{J}(\hat{\lambda}_{i\psi})]^{-1/2},\end{aligned} \tag{5.2.6}$$

where $L_P^{(i)}(\psi) = L^{(i)}(\psi,\hat{\lambda}_{i\psi})$ is the profile likelihood for ψ. The error in this approximation is

$$\bar{L}^{(i)}(\psi) = \hat{\bar{L}}^{(i)}(\psi)\left\{1 + O\left(\frac{1}{m}\right)\right\} \quad \text{as} \quad m \to \infty. \tag{5.2.7}$$

Let

$$\bar{\ell}^{(i)}(\psi) = \log \bar{L}^{(i)}(\psi) \tag{5.2.8}$$

denote the integrated log-likelihood for ψ. Putting the results in Equation 5.2.6, Equation 5.2.7, and Equation 5.2.8 together, we have

$$\begin{aligned}
\overline{\ell}^{(i)}(\psi) &= \log \overline{L}^{(i)}(\psi) && \text{(by Equation 5.2.8)} \\
&= \log\left(\hat{\bar{L}}^{(i)}(\psi)\left\{1 + O\left(\frac{1}{m}\right)\right\}\right) && \text{(by Equation 5.2.7)} \\
&= \log \hat{\bar{L}}^{(i)}(\psi) + \log\left\{1 + O\left(\frac{1}{m}\right)\right\} && \text{(by Equation 5.2.6)} \\
&= \log\left\{\frac{\sqrt{2\pi}}{m} L_P^{(i)}(\psi) g(\hat{\lambda}_{i\psi}; \psi)\left[\mathcal{J}(\hat{\lambda}_{i\psi})\right]^{-1/2}\right\} + O\left(\frac{1}{m}\right) \\
&= \frac{1}{2}\log(2\pi) - \log(m) + \log L_P^{(i)}(\psi) + \log g(\hat{\lambda}_{i\psi}; \psi) - \frac{1}{2}\log \mathcal{J}(\hat{\lambda}_{i\psi}) + O\left(\frac{1}{m}\right) \\
&= \ell_P^{(i)}(\psi) + \log g(\hat{\lambda}_{i\psi}; \psi) - \frac{1}{2}\log \mathcal{J}(\hat{\lambda}_{i\psi}) + \frac{1}{2}\log(2\pi) - \log(m) + O\left(\frac{1}{m}\right) \quad \text{as} \quad m \to \infty,
\end{aligned}$$

where $\ell_P^{(i)}(\psi) = \ell^{(i)}(\psi, \hat{\lambda}_{i\psi})$ is the profile log-likelihood for ψ. Since log-likelihoods are equivalent up to additive constants, we can discard the $\frac{1}{2}\log(2\pi)$ and $\log(m)$ terms in the final line above to arrive at our final approximation for the integrated log-likelihood in stratum i:

$$\hat{\bar{\ell}}^{(i)}(\psi) = \ell_P^{(i)}(\psi) + \log g(\hat{\lambda}_{i\psi}; \psi) - \frac{1}{2}\log \mathcal{J}(\hat{\lambda}_{i\psi}). \tag{5.2.9}$$

The error in this approximation is given by

$$\overline{\ell}^{(i)}(\psi) = \hat{\bar{\ell}}^{(i)}(\psi) + O\left(\frac{1}{m}\right). \tag{5.2.10}$$

Since the observations between the strata are independent, we may write the likelihood and log-likelihood functions for the entire model as

$$L(\psi, \lambda) = \prod_{i=1}^{q} L^{(i)}(\psi, \lambda_i) \tag{5.2.11}$$

and

$$\ell(\psi, \lambda) = \sum_{i=1}^{q} \ell^{(i)}(\psi, \lambda_i), \tag{5.2.12}$$

respectively. If we define the weight function

$$G(\lambda; \psi) \equiv \prod_{i=1}^{q} g(\lambda_i; \psi) \tag{5.2.13}$$

then the integrated likelihood function for ψ becomes separable. That is,

$$\begin{aligned} \bar{L}(\psi) &= \int_{\Lambda^q} L(\psi, \lambda) G(\lambda; \psi) d\lambda \\ &= \int_{\Lambda} \cdots \int_{\Lambda} \left[\prod_{i=1}^{q} L^{(i)}(\psi, \lambda_i) \right] \left[\prod_{i=1}^{q} g(\lambda_i; \psi) \right] d\lambda_1 \cdots d\lambda_q \\ &= \prod_{i=1}^{q} \int_{\Lambda} L^{(i)}(\psi, \lambda_i) g(\lambda_i; \psi) d\lambda_i \\ &= \prod_{i=1}^{q} \bar{L}^{(i)}(\psi). \end{aligned} \tag{5.2.14}$$

Let $\bar{\ell}(\psi) = \log \bar{L}(\psi)$ denote the integrated log-likelihood function for ψ. Taking the logarithm of both sides in Equation 5.2.14, we have

$$\bar{\ell}(\psi) = \sum_{i=1}^{q} \bar{\ell}^{(i)}(\psi). \tag{5.2.15}$$

Plugging in our approximation to $\bar{\ell}^{(i)}(\psi)$ and its error term in Equation 5.2.9 and Equation 5.2.10, respectively, yields

$$\begin{aligned} \bar{\ell}(\psi) &= \sum_{i=1}^{q} \left[\ell_P^{(i)}(\psi) + \log g(\hat{\lambda}_{i\psi}; \psi) - \log \mathcal{J}(\hat{\lambda}_{i\psi}) + O\left(\frac{1}{m}\right) \right] \\ &= \ell_P(\psi) + \sum_{i=1}^{q} \log g(\hat{\lambda}_{i\psi}; \psi) - \frac{1}{2} \sum_{i=1}^{q} \log \mathcal{J}(\hat{\lambda}_{i\psi}) + O\left(\frac{q}{m}\right) \quad \text{as} \quad m \to \infty. \end{aligned} \tag{5.2.16}$$

CHAPTER 6

Applications

6.1. Multinomial Distribution

6.2. Standardized Mean Difference

References

Barndorff-Nielsen, O. E., and David R. Cox. 1996. "Prediction and Asymptotics." *Bernoulli* 2 (4): 319–40. http://www.jstor.org/stable/3318417.

Basu, Debabrata. 1977. "On the Elimination of Nuisance Parameters." *Journal of the American Statistical Association* 72 (358): 355–66. http://www.jstor.org/stable/2286800.

Berger, James O., Brunero Liseo, and Robert L. Wolpert. 1999. "Integrated Likelihood Methods for Eliminating Nuisance Parameters." *Statistical Science* 14 (1): 1–22. http://www.jstor.org/stable/2676641.

Cramér, Harald. 1945. "Section 33.3. Asymptotic Properties of Maximum Likelihood Estimates." In *Mathematical Methods of Statistics*, 500–506. Princeton University Press.

De Bin, Riccardo, Nicola Sartori, and Thomas A. Severini. 2015. "Integrated likelihoods in models with stratum nuisance parameters." *Electronic Journal of Statistics* 9 (1): 1474–91. https://doi.org/10.1214/15-EJS1045.

Kalbfleisch, J. D., and D. A. Sprott. 1973. "Marginal and Conditional Likelihoods." *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 35 (3): 311–28. http://www.jstor.org/stable/25049882.

Liseo, Brunero. 1993. "Elimination of Nuisance Parameters with Reference Priors." *Biometrika* 80 (2): 295–304. http://www.jstor.org/stable/2337200.

Sartori, N. 2003. "Modified Profile Likelihoods in Models with Stratum Nuisance Parameters." *Biometrika* 90 (3): 533–49. http://www.jstor.org/stable/30042064.

Schumann, Martin, Thomas A. Severini, and Gautam Tripathi. 2021. "Integrated Likelihood Based Inference for Nonlinear Panel Data Models with Unobserved Effects." *Journal of Econometrics* 223 (1): 73–95. https://doi.org/10.1016/j.jeconom.2020.10.001.

———. 2023. "The Role of Score and Information Bias in Panel Data Likelihoods." *Journal of Econometrics* 235 (2): 1215–38. https://doi.org/10.1016/j.jeconom.2022.08.011.

Severini, Thomas A. 2000. *Likelihood Methods in Statistics*. Oxford University Press.

———. 2007. "Integrated Likelihood Functions for Non-Bayesian Inference." *Biometrika* 94 (3): 529–42. http://www.jstor.org/stable/20441394.

———. 2018. "Integrated Likelihoods for Functions of a Parameter." *Stat* 7 (1): e212. https://doi.org/10.1002/sta4.212.

———. 2022. "Integrated Likelihood Inference in Multinomial Distributions." *Metron*. https://doi.org/10.1007/s40300-022-00236-x.

Tierney, Luke, and Joseph B. Kadane. 1986. "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association* 81 (393): 82–86. http://www.jstor.org/stable/2287970.

APPENDIX A

Chapter 3

A.1. Definitions and Notation

A.1.1. Open and Closed Balls

A.1.1.1. Open Ball. The *open ball of radius* $r > 0$ centered at a point $p \in \mathbb{R}^d$, is the set of all points $x \in \mathbb{R}^d$ such that the distance between p and x is less than r We denote this set using the notation

$$B_r(\mathbf{p}) = \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x} - \mathbf{p}|| < r\},$$

where $||\cdot||$ indicates the Euclidean norm, i.e., $||\mathbf{x}|| = \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}$ for $\mathbf{x} \in \mathbb{R}^d$.

A.1.1.2. Closed Ball. The *closed ball of radius* $r > 0$ centered at a point $p \in \mathbb{R}^d$, is the set of all points $x \in \mathbb{R}^d$ such that the distance between p and x is less than or equal to r. We denote this set using the notation

$$B_r[\mathbf{p}] = \{\mathbf{x} \in \mathbb{R}^d : \ ||\mathbf{x} - \mathbf{p}||\} \leq r.$$

A.1.2. Open and Closed Sets

A.1.2.1. Open Set. A subset A of $\mathbb{R}^d$ is called an *open set* of $\mathbb{R}^d$ if every point in A is the center of an open ball entirely contained in A. That is, A is open if and only if for any $a \in A$, there exists a radius $r > 0$ such that $B_r(a) \subseteq A$.

A.1.2.2. Closed Set. A subset B of $\mathbb{R}^d$ is called a *closed set* of $\mathbb{R}^d$ if its complement $B^c = \mathbb{R}^d \setminus B$ is open.

A.1.3. Neighborhoods

A set $N_{\mathbf{p}} \subseteq \mathbb{R}^d$ is called a *neighborhood* of a point $\mathbf{p} \in \mathbb{R}^d$ if it contains an open ball centered at $\mathbf{p}$, i.e., for some radius $r > 0$ there exists an open ball $B_r(\mathbf{p})$ such that $B_r(\mathbf{p}) \subseteq N_{\mathbf{p}}$.

A.1.4. Boundedness

A.1.4.1. Bounded Set. A set $S \subset \mathbb{R}^d$ is called *bounded* if there exists some radius $r > 0$ such that $B_r(\,) \subset S$.

A.1.4.2. Bounded Function. A function $f : X \to \mathbb{R}$ is called *bounded* if there exists a real number M such that $|f(x)| \leq M$ for all $x \in X$.

A.1.5. Compact Set

A subset of $\mathbb{R}^d$ is called *compact* if it is closed and bounded.

A.1.6. Interiority

A.1.6.1. Interior Point. A point $p \in S \subset \mathbb{R}^d$ is called an *interior point* of S if there exists some radius $r > 0$ such that $\subseteq B_r(\mathbb{p})$.

A.1.6.2. Interior of a Set. The *interior of a set* $S \subset \mathbb{R}^d$, denoted by $\operatorname{int} S$, is the set of all interior points of S.

A.1.7. Line Segments

For two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$, a third point $\bar{\mathbf{x}}$ is said to be on the *line segment* connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ if there exists $\omega \in [0, 1]$ such that $\bar{\mathbf{x}} = \omega\mathbf{x}_1 + (1 - \omega)\mathbf{x}_2$. We use the following notation to refer to such line segments:

$$\mathrm{LS}(\mathbf{x}_1, \mathbf{x}_2) = \{\omega\mathbf{x}_1 + (1 - \omega)\mathbf{x}_2 : \omega \in [0, 1]\}.$$

A.1.8. Convexity and Concavity

A.1.8.1. Convex Set. A set $S \subseteq \mathbb{R}^d$ is called *convex* if for any two points $\mathbf{x}_1, \mathbf{x}_2 \in S$, the line segment connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ is entirely contained within S, i.e.,

$$\mathrm{LS}(\mathbf{x}_1, \mathbf{x}_2) \subseteq S \text{ for all } \mathbf{x}_1, \mathbf{x}_2 \in S.$$

A.1.8.2. Convex Function. Let $f : X \to \mathbb{R}$, where X is a convex set. f is called a *convex function* if for all $t \in [0, 1]$ and all $x_1, x_2 \in X$,

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$

If it is possible to graph the function on the coordinate plane, this is equivalent to saying that the line segment between any two distinct points on the graph of the function lies above the graph.

f is called *strictly convex* if the equality is tightened, i.e.,

$$f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2).$$

A.1.8.3. Concave Function. Let $f : X \to \mathbb{R}$, where X is a convex set. f is called a *concave function* if for all $t \in [0, 1]$ and all $x_1, x_2 \in X$,

$$f(tx_1 + (1-t)x_2) \geq tf(x_1) + (1-t)f(x_2).$$

If it is possible to graph the function on the coordinate plane, this is equivalent to saying that the line segment between any two distinct points on the graph of the function lies below the graph.

f is called *strictly concave* if the equality is tightened, i.e.,

$$f(tx_1 + (1-t)x_2) > tf(x_1) + (1-t)f(x_2).$$

A.1.9. Positive Definiteness

A $d \times d$ symmetric real matrix M is called *positive definite* if $\mathbf{x}^\top M\mathbf{x} > 0$ for all non-zero $\mathbf{x} \in \mathbb{R}^d$.

A.1.10. Derivatives of Multivariable Functions

A.1.10.1. Gradient. The *gradient* of a multivariable scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is given by its $d \times 1$ vector of partial derivatives:

$$\nabla f(x_1, x_2, ..., x_d) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix}.$$

A.1.10.2. Jacobian Matrix. The *Jacobian matrix* of a multivariable vector-valued function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^k$ given by

$$\mathbf{f}(x_1, x_2, ..., x_d) = \begin{pmatrix} f_1(x_1, x_2, ..., x_d) \\ f_2(x_1, x_2, ..., x_d) \\ \vdots \\ f_k(x_1, x_2, ..., x_d) \end{pmatrix}.$$

is defined as its $k \times d$ matrix of partial derivatives:

$$\mathbf{J}(\mathbf{f}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & ... & \frac{\partial f_1}{\partial x_d} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & ... & \frac{\partial f_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_k}{\partial x_1} & \frac{\partial f_k}{\partial x_2} & ... & \frac{\partial f_k}{\partial x_d} \end{pmatrix} = \begin{pmatrix} \nabla^\top f_1 \\ \vdots \\ \nabla^\top f_k \end{pmatrix}$$

In the case where $k = 1$, the Jacobian matrix simply reduces to $\nabla^\top f$, the transpose of the gradient of $\mathbf{f}$.

A.1.10.3. Hessian Matrix. The Hessian matrix of a multivariable scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is given by its $d \times d$ matrix of second partial derivatives:

$$\mathbf{H}(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{pmatrix}$$

The *Hessian matrix* of f is also equal to the transpose of the Jacobian matrix of the gradient of f, i.e., $\mathbf{H}(f) = \mathbf{J}(\nabla f)^\top$. When the order of differentiation does not matter, which occurs if and only if all of f's second partial derivatives are continuous, both matrices become symmetric and we simply have $\mathbf{H}(f) = \mathbf{J}(\nabla f)$.

A twice differentiable function of several variables is strictly convex (concave) on a convex set if and only if its Hessian matrix is positive (negative) definite on the interior of the convex set.

A.2. Theorems

A.2.1. The Extreme Value Theorem

If K is a compact set and $f : K \to \mathbb{R}$ is a continuous function, then f is bounded and there exists $p, q \in K$ such that

$$f(p) = \sup_{x \in K} f(x)$$

and

$$f(q) = \inf_{x \in K} f(x).$$

A.2.2. The Mean-Value Theorem for Multivariable Vector-Valued Functions

Suppose the function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^k$ is continuously differentiable for all points $\mathbf{x} \in B_r(\mathbf{x_0})$. Then for $||\mathbf{t}|| < r$,

$$f(\mathbf{x_0} + \mathbf{t}) = f(\mathbf{x_0}) + \left[\int_0^1 \mathbf{J}\Big(f(\mathbf{x} + u\mathbf{t}) \Big) du \right] \mathbf{t}.$$

The term in the square brackets above denotes a $k \times d$ matrix whose (i, j)th entry is given by

$$\int_0^1 \frac{\partial f_i(\mathbf{x} + u\mathbf{t})}{\partial x_j} du.$$

A.2.3. Taylor's Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that is $(k+1)$-times continuously differentiable in a neighborhood $N_r(\mathbf{x}_0)$ of some point $\mathbf{x}_0 \in \mathbb{R}^d$ and suppose there exists M satisfying $|D^{\boldsymbol{\alpha}} f| \leq M$ for all $x \in N_r(x_0)$ and all $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| = k + 1$. Then

$$f(\mathbf{x}) = \sum_{0 \leq |\boldsymbol{\alpha}| \leq k} \frac{D^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{x}_0)^{\boldsymbol{\alpha}} + R_k(\mathbf{x}),$$

where the remainder term $R_k(\mathbf{x})$ satisfies

$$|R_k(\mathbf{x})| \leq \frac{M}{\boldsymbol{\alpha}!} |\mathbf{x} - \mathbf{x}_0|^{\boldsymbol{\alpha}}$$

for all $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| = k + 1$

A.2.4. Jensen's Inequality

For a real-valued random variable X with finite expectation and a strictly concave function φ,

$$\mathrm{E}[\varphi(X)] < \varphi(\mathrm{E}[X]).$$

APPENDIX B

Chapter 5

B.1. Desirable Properties of the Integrated Likelihood

B.1.1. Property 1

Suppose the likelihood function for a parameter $\boldsymbol{\theta}$ can be decomposed as the product $L(\boldsymbol{\theta}) = L_1(\psi)L_2(\lambda)$. Then the integrated likelihood for ψ should satisfy

$$\bar{L}(\psi) = L_1(\psi).$$

B.1.2. Property 2

B.1.3. Property 3

B.1.4. Property 4

B.2. Laplace's Method

Let $\boldsymbol{\theta}$ be a scalar parameter taking values in $\mathbb{R}$ and consider an integral of the form

$$I = \int_{-\infty}^{\infty} f(\boldsymbol{\theta}) \exp[-nh(\boldsymbol{\theta})] d\boldsymbol{\theta}.$$

Suppose the function $-h(\boldsymbol{\theta})$ is smooth, bounded, and unimodal so that it attains a maximum at a point $\hat{\boldsymbol{\theta}}$. Then Laplace's method states that an approximation for I is given by

$$\hat{I} = f(\hat{\boldsymbol{\theta}}) \sqrt{\frac{2\pi}{n}} \sigma \exp[-nh(\hat{\boldsymbol{\theta}})],$$

where

$$\sigma = \left[\left. \frac{\partial^2 h}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right]^{-1/2}.$$

It can further be shown that

$$I = \hat{I}\left\{1 + O\left(\frac{1}{n}\right)\right\},$$

where the term n may be interpreted as the sample size.