NORTHWESTERN UNIVERSITY

Approximation of the Integrated Likelihood Function for a Scalar Parameter of Interest

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Statistics

By

Timothy Ruel

EVANSTON, ILLINOIS

June 2025

ABSTRACT

Approximation of the Integrated Likelihood Function for a Scalar Parameter of Interest

Timothy Ruel

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

CHAPTER 2

The Likelihood Function

2.1. 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

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{x}_n) = p(\mathbf{x}_n; \boldsymbol{\theta}), \ \ \boldsymbol{\theta} \in \Theta. \tag{2.1.1}$$

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 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 2.1.1. 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.

2.2. 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:

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \mathbf{x}_n) = \log L(\boldsymbol{\theta}; \mathbf{x}_n). \tag{2.2.1}$$

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

---

[1]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.

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

$$\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}_n) = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell(\boldsymbol{\theta}; \mathbf{x}_n) \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{\theta}_d} \ell(\boldsymbol{\theta}; \mathbf{x}_n) \end{pmatrix} = \begin{pmatrix} \mathcal{S}_1(\boldsymbol{\theta}; \mathbf{x}_n) \\ \vdots \\ \mathcal{S}_d(\boldsymbol{\theta}; \mathbf{x}_n) \end{pmatrix}, \tag{2.2.2}$$

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

$$\mathcal{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) = -\mathbf{H}_{\boldsymbol{\theta}}\Big(\ell(\boldsymbol{\theta}; \mathbf{x}_n)\Big) = -\mathbf{J}_{\boldsymbol{\theta}}\Big(\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n)\Big)^{\top} = -\begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_d} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_d \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_d \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_d^2} \end{pmatrix}. \tag{2.2.3}$$

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

$$\mathscr{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) = \text{Var}_{\boldsymbol{\theta}}[\mathcal{S}(\boldsymbol{\theta}; \mathbf{X}_n)]. \tag{2.2.4}$$

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

$$\mathscr{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}[\mathcal{S}(\boldsymbol{\theta}; \mathbf{X}_n)] \qquad \text{(by Eq. 2.2.4)}$$

$$= \mathrm{Cov}_{\boldsymbol{\theta}}\left[\left(\tfrac{\partial \ell}{\partial \theta_1}, \cdots, \tfrac{\partial \ell}{\partial \theta_d} \right)^T \right] \qquad \text{(by Eq. 2.2.2)}$$

$$= \begin{pmatrix} \mathrm{Var}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_1} \right) & \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_1}, \tfrac{\partial \ell}{\partial \theta_2} \right) & \cdots & \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_1}, \tfrac{\partial \ell}{\partial \theta_d} \right) \\ \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_2}, \tfrac{\partial \ell}{\partial \theta_1} \right) & \mathrm{Var}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_2} \right) & \cdots & \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_2}, \tfrac{\partial \ell}{\partial \theta_d} \right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_d}, \tfrac{\partial \ell}{\partial \theta_1} \right) & \mathrm{Cov}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_d}, \tfrac{\partial \ell}{\partial \theta_2} \right) & \cdots & \mathrm{Var}_{\boldsymbol{\theta}}\left(\tfrac{\partial \ell}{\partial \theta_d} \right) \end{pmatrix}. \qquad (2.2.5)$$

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,

$$\mathscr{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathscr{I}_{X_i}(\boldsymbol{\theta}). \qquad (2.2.6)$$

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:

$$\mathscr{I}_{\mathbf{X}_n}(\boldsymbol{\theta}) = n\mathscr{I}_X(\boldsymbol{\theta}). \qquad (2.2.7)$$

The above two equations hold true for the observed information under the same conditions as well.

2.3. Maximum Likelihood Estimation

2.3.1. 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 $\mathbf{x}_n$. Hence, 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.,

$$L(\hat{\boldsymbol{\theta}}; \mathbf{x}_n) = \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x}_n). \tag{2.3.1}$$

We call such a value a *maximum likelihood estimate* of $\boldsymbol{\theta}_0$. This definition of $\hat{\boldsymbol{\theta}}$ 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 $\hat{\boldsymbol{\theta}}(\mathbf{x}_n)$.

2.3.2. 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 second-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 $i, j = 1, ..., d$. 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 $\boldsymbol{\theta}_0$.

2.3.3. 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 is critical to 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.

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.

2.4. 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. This means 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

$$
\begin{aligned}
E_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta; X)\right] &= \int_\mathbb{R}\left[\frac{\partial}{\partial\theta}\log p(x; \theta)\right]p(x; \theta)dx \\
&= \int_\mathbb{R}\frac{\frac{\partial}{\partial\theta}p(x; \theta)}{p(x; \theta)}p(x; \theta)dx \\
&= \int_\mathbb{R}\frac{\partial}{\partial\theta}p(x; \theta)dx \\
&= \frac{d}{d\theta}\int_\mathbb{R}p(x; \theta)dx \\
&= \frac{d}{d\theta}1 \\
&= 0.
\end{aligned}
\tag{2.4.1}
$$

Equation 2.4.1 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

$$
\begin{aligned}
\mathscr{I}_X(\theta) &= \text{Var}_\theta[\mathcal{S}(\theta; X)] \\
&= \text{Var}_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta; X)\right] \\
&= \text{Var}_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta; X)\right] + \left(\text{E}_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta; X)\right]\right)^2 \quad \text{(by the first Bartlett identity)} \\
&= \text{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ell(\theta; X)\right)^2\right].
\end{aligned}
\tag{2.4.2}
$$

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

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

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,

$$
\mathrm{E}_\theta\left[\frac{\partial^2}{\partial \theta^2} \ell(\theta; X)\right] + \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial \theta} \ell(\theta; X)\right)^2\right] = 0. \tag{2.4.3}
$$

Equation 2.4.2 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 2.2.4, Equation 2.4.2, and Equation 2.4.3 imply that the following definitions for its expected information regarding its parameter θ are all equivalent:

$$\mathscr{I}_X(\theta) = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta;X)\right] = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ell(\theta;X)\right)^2\right] = \mathrm{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta;X)\right] = \mathrm{E}_\theta[\mathscr{J}_X(\theta)]. \qquad (2.4.4)$$

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. 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.

CHAPTER 3

Pseudolikelihood Functions

3.1. 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_0, the true value of ψ. We will examine some of the issues posed by high-dimensional nuisance parameters in greater detail in the next chapter.

---

[1]Nuisance parameters are not always uniquely defined. In fact, depending on the choice of parameter of interest, there may be multiple or even infinite ways to define one.

3.1.1. 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

$$\psi = (\theta_{I_1}, ..., \theta_{I_p}). \tag{3.1.1}$$

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

$$\lambda = (\theta_{J_1}, ..., \theta_{J_q}). \tag{3.1.2}$$

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

3.1.2. 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

$$\psi = \varphi(\boldsymbol{\theta}). \tag{3.1.3}$$

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.

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

$$\varphi(\boldsymbol{\theta}) = (\theta_{I_1}, ..., \theta_{I_p}) = \psi, \tag{3.1.4}$$

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.

3.1.3. Nuisance Parameter Elimination

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 $\boldsymbol{\theta} \in \Theta$. If we let $\Theta(\psi) = \{\boldsymbol{\theta} \in \Theta : \varphi(\boldsymbol{\theta}) = \psi\}$, then associated with each $\psi \in \Psi$ is the set of likelihoods $\mathcal{L}_\psi = \{L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta(\psi)\}$.

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. None of these summaries come without a cost though, meaning some information about ψ_0 is almost certainly sacrificed whenever a nuisance parameter is eliminated from a likelihood. One measure of a good pseudolikelihood, therefore, is how well it is able to retain information about ψ_0 without becoming overly complex in its computation.

In the previous chapter, we introduced the Bartlett identities as being a set of equations relating the derivatives of the log-likelihodd to one another. They can also be used to understand 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 being 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. A pseudolikelihood that satisfies the first Bartlett identity is called *score-unbiased*; one that satisfies the second is called *information-unbiased*. Historically, more attention has been given to constructing pseudo-likelihoods that are score-unbiased, at least asymptotically (Cf. Kalbfleisch and Sprott, 1970; De Bin et al., 2015; Schumann et al., 2021, 2023).

3.2. The Marginal Likelihood

Suppose we observe some data $\mathbf{X} = \mathbf{x}$ such that the pair of statistics $(\mathbf{T}(\mathbf{X}), \mathbf{S}(\mathbf{X}))$ is sufficient for $\boldsymbol{\theta} = (\psi, \lambda)$. Then, abusing notation slightly, we can write

$$p_{\boldsymbol{\theta}}(\mathbf{X}) = p_{\boldsymbol{\theta}}(\mathbf{T}, \mathbf{S}). \tag{3.2.1}$$

Since the right-hand side is a joint density for $\mathbf{T}$ and $\mathbf{S}$, it can be factored into a product of a marginal and a conditional density as follows:

$$p_{\boldsymbol{\theta}}(\mathbf{T}, \mathbf{S}) = p_{\boldsymbol{\theta}}(\mathbf{T}|\mathbf{S})p_{\boldsymbol{\theta}}(\mathbf{S}). \tag{3.2.2}$$

If the right-hand term in this product doesn't depend on λ, i.e., $p_{\boldsymbol{\theta}}(\mathbf{S}) = p_{\psi}(\mathbf{S})$, then we have

$$p_{\boldsymbol{\theta}}(\mathbf{T}, \mathbf{S}) = p_{\boldsymbol{\theta}}(\mathbf{T}|\mathbf{S})p_{\psi}(\mathbf{S}). \tag{3.2.3}$$

In this case, a pseudolikelihood for ψ is simply

$$L_m(\psi) = p_{\psi}(\mathbf{S}). \tag{3.2.4}$$

$L_m(\psi)$ is called a *marginal likelihood* for ψ as it is based on the marginal distribution of $\mathbf{S} = \mathbf{S}(\mathbf{X})$. The conditional part of the density, $p_{\boldsymbol{\theta}}(\mathbf{T}|\mathbf{S})$, does still depend on ψ, however, and so by choosing to base

our inferences regarding ψ_0 solely on $L_m(\psi)$, we have ignored some relevant information for ψ_0 that was contained in the data.

3.3. The Conditional Likelihood

If instead it is the left-hand term in Equation 3.2.2 that doesn't depend on λ, i.e., $p_{\boldsymbol{\theta}}(\mathbf{T}|\mathbf{S}) = p_\psi(\mathbf{T}|\mathbf{S})$, then we have

$$p_{\boldsymbol{\theta}}(\mathbf{T}, \mathbf{S}) = p_\psi(\mathbf{T}|\mathbf{S})p_\theta(\mathbf{S}). \tag{3.3.1}$$

Here, a pseudolikelihood for ψ may be given by

$$L_C(\psi) = p_\psi(\mathbf{T}|\mathbf{S}). \tag{3.3.2}$$

$L_C(\psi)$ is called a *conditional likelihood* for ψ as it is based on the conditional distribution of $\mathbf{T}$ given $\mathbf{S}$. As with the marginal likelihood, some information has been disregarded through our omission of $p_\theta(\mathbf{S})$ in our inference for ψ_0.

Thus, the use of either a marginal or a conditional likelihood as a pseudolikelihood for a parameter of interest is theoretically only justified when the benefits of eliminating the nuisance parameter through marginalization or conditioning outweigh the corresponding loss in information. In practice, however, the real limiting constraint of using a marginal or conditional likelihood tends not to be the lack of clarity they confer to our inferences, but rather their lack of existence in the first place. The ability to factor a density as in Equation 3.2.3 or Equation 3.3.1 is a rather strong condition, and there are plenty of models for which it is impossible to construct a marginal or conditional likelihood for its parameter of interest. When they do exist, it is typically worthwhile to use them.

3.4. The Profile Likelihood

Profile likelihoods are one of the most straightforward methods for eliminating a nuisance parameter λ from a likelihood function. The idea is to summarize the set $\mathcal{L}_\psi$ by its maximum value. Formally, the

profile likelihood is defined as

$$L_p(\psi) = \sup_{\boldsymbol{\theta} \in \Theta(\psi)} L(\boldsymbol{\theta}) = L(\hat{\boldsymbol{\theta}}_\psi),$$

where $\hat{\boldsymbol{\theta}}_\psi$ is the MLE of $\boldsymbol{\theta_0}$ for fixed ψ. If we define $\hat{\lambda}_\psi$ to be the conditional MLE for λ given ψ, meaning the value of λ that maximizes the likelihood for a particular value of ψ, then we must have $\hat{\boldsymbol{\theta}}_\psi = (\psi, \hat{\lambda}_\psi)$.

Much of the allure of a profile likelihood can be traced to its ease of computation. In the event of a regular model, finding the value of $\hat{\lambda}_\psi$ corresponding to a particular ψ is equivalent to solving a convex optimization problem. Either an analytical solution to this problem will exist or a numerical one can be obtained using modern computational tools. In both cases, $\hat{\lambda}_\psi$ can be found without much trouble, and from there it's just a matter of setting $\lambda = \hat{\lambda}_\psi$ inside $L(\psi, \lambda)$ to obtain $L_p(\psi)$.

The method is not without its drawbacks however. As noted by Kalbfleisch & Sprott (1970), a major disadvantage to profile likelihoods are their inability to take into account the dimensionality of the nuisance parameter. By effectively always assuming that $\lambda = \hat{\lambda}_\psi$, they fail to incorporate any uncertainty we might have in its value into the resulting pseudolikelihood, leading to overly confident estimates of ψ. This effect is especially pronounced when the dimension of Λ is large. Berger et al. (1999) also bring up the scenario in which the likelihood has a sharp ridge running in one direction as being one in which the profile likelihood will perform poorly. In such a situation, the profile of the likelihood along the ridge would not be representative of the shape of the likelihood elsewhere, and yet it is exactly that profile which will be obtained through maximization.

Nevertheless, a profile likelihood can still be a useful tool for conducting inference regarding ψ_0. At worst, it offers a baseline of sorts to assess the quality of other pseudolikelihoods. There is no point in using another pseudolikelihood that is harder to compute if it offers the same amount of information about ψ_0 - it must justify this increase in complexity with a corresponding increase in features that lend themselves to a greater degree of accuracy in our inference, such as higher level of peakedness around the value of ψ_0.

3.5. The Integrated Likelihood

The *integrated likelihood* for ψ seeks to summarize $\mathcal{L}_\psi$ by its average value with respect to some weight function π over $\Theta(\psi)$. From a theoretical standpoint, this is preferable to maximization as it incorporates (or at least is capable of incorporating) the uncertainty we have in the value of the nuisance parameter into the resulting pseudolikelihood in a very natural way. Formally, the integrated likelihood function is defined as

$$\bar{L}(\psi) = \int_\Lambda L(\psi, \lambda)\pi(\lambda|\psi)d\lambda, \tag{3.5.1}$$

where $\pi(\lambda|\psi)$ is a nonnegative function on Λ. $\pi(\lambda|\psi)$ is sometimes called a conditional prior density for λ given ψ, though it need not satisfy the requirements of a genuine density function.

Note the similarity in form between the integral in Equation 3.5.1 and the expression for the normalizing constant of a posterior distribution:

$$\int_\Theta L(\theta; X)\pi(\theta)d\theta.$$

This similarity lends weight to the idea that Bayesian techniques used to obtain empirical approximations to posterior distributions, such as Markov Chain Monte Carlo, could also be used to approximate an integrated likelihood function, with the result being useful for Bayesian and frequentist inference alike.

In general, the selection of the weight function plays an important role in the properties of the resulting integrated likelihood. The obvious first choice to make for it, and therefore the one that could be considered the "default", is simply $\pi(\lambda|\psi) \propto 1$, i.e., a uniform density. This yields what is known as the uniform-integrated likelihood:

$$\bar{L}^U(\psi) = \int_\Lambda L(\psi, \lambda)d\lambda. \tag{3.5.2}$$

In the next chapter, we will discuss a re-parameterization of the nuisance parameter developed by Severini (2007) that makes the integrated likelihood relatively insensitive to the exact weight function chosen. Using this new parameterization, we have great flexibility in choosing our weight function; as long as it does not depend on the parameter of interest, the integrated likelihood that is produced will enjoy many desirable frequency properties.

CHAPTER 4

Properties of the Integrated Likelihood

4.1. Asymptotic Theory

4.1.1. One-Index Asymptotics

The one-index asymptotics framework describes the behavior of likelihood-based statistics as the sample size (n) grows to infinity. The aim of this section is to present an overview of some of the theory's classic results. This provides us with a readily available baseline against which to compare the results of the following section discussing the two-index asymptotics framework, in which the full model parameter is partitioned into a parameter of interest and a nuisance parameter, and the dimension of the nuisance parameter is allowed to increase with the sample size.

Let $\boldsymbol{\theta}$ be the parameter of a regular model with true value $\boldsymbol{\theta}_0$. As a partial justification for our use of the MLE in estimating $\boldsymbol{\theta}_0$, we will start by showing that with probability tending to 1 as n tends toward infinity, the likelihood for $\boldsymbol{\theta}$ is strictly larger at $\boldsymbol{\theta}_0$ than for any other $\boldsymbol{\theta} \in \Theta$. **RC1** implies

$$L(\boldsymbol{\theta}; \mathbf{x}_n) = p(\mathbf{x}_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta}) \tag{4.1.1}$$

and

$$\ell(\boldsymbol{\theta}; \mathbf{x}_n) = \log p(\mathbf{x}_n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta}). \tag{4.1.2}$$

It follows that

$$L(\boldsymbol{\theta}; \mathbf{x}_n) < L(\boldsymbol{\theta}_0; \mathbf{x}_n) \iff \ell(\boldsymbol{\theta}; \mathbf{x}_n) < \ell(\boldsymbol{\theta}_0; \mathbf{x}_n)$$

$$\iff \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta}) - \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta}_0) < 0$$

$$\iff \sum_{i=1}^{n} \left[\log p(x_i; \boldsymbol{\theta}) - \log p(x_i; \boldsymbol{\theta}_0) \right] < 0 \qquad (4.1.3)$$

$$\iff \sum_{i=1}^{n} \log \frac{p(x_i; \boldsymbol{\theta})}{p(x_i; \boldsymbol{\theta}_0)} < 0$$

$$\iff \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i; \boldsymbol{\theta})}{p(x_i; \boldsymbol{\theta}_0)} < 0.$$

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,

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i; \boldsymbol{\theta})}{p(X_i; \boldsymbol{\theta}_0)} \to \mathrm{E}_{\boldsymbol{\theta}_0} \left[\log \frac{p(X; \boldsymbol{\theta})}{p(X; \boldsymbol{\theta}_0)} \right] \qquad (4.1.4)$$

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

$$\mathrm{E}_{\boldsymbol{\theta}_0} \left[\frac{p(X; \boldsymbol{\theta})}{p(X; \boldsymbol{\theta}_0)} \right] = \int_{\mathcal{X}} \left[\frac{p(x; \boldsymbol{\theta})}{p(x; \boldsymbol{\theta}_0)} \right] p(x; \boldsymbol{\theta}_0) dx = \int_{\mathcal{X}} p(x; \boldsymbol{\theta}) dx = 1. \qquad (4.1.5)$$

Since $\log(x)$ is a strictly concave function, it follows from Jensen's inequality (see Appendix A) and Equation 4.1.5

$$\mathrm{E}_{\boldsymbol{\theta}_0} \left[\log \frac{p(X; \boldsymbol{\theta})}{p(X; \boldsymbol{\theta}_0)} \right] < \log \mathrm{E}_{\boldsymbol{\theta}_0} \left[\frac{p(X; \boldsymbol{\theta})}{p(X; \boldsymbol{\theta}_0)} \right] = \log 1 = 0. \qquad (4.1.6)$$

Hence, the quantity on the left-hand side of Equation 4.1.4 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 4.1.3, it follows that

$$\lim_{n \to \infty} P_{\boldsymbol{\theta}_0} \left[L(\boldsymbol{\theta}; \mathbf{x}_n) < L(\boldsymbol{\theta}_0; \mathbf{x}_n) \right] = \lim_{n \to \infty} P_{\boldsymbol{\theta}_0} \left[\frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i; \boldsymbol{\theta})}{p(X_i; \boldsymbol{\theta}_0)} < 0 \right] = 1, \qquad (4.1.7)$$

which proves the claim.

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

$$\nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{4.1.8}$$

whenever it exists. 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 now turn our attention to the asymptotic distribution of the MLE. Similarly to the log-likelihood function, **RC1** implies the score function is equal to

$$\begin{aligned}
\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n) &= \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}_n) \\
&= \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; x_i) \\
&= \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; x_i) \\
&= \sum_{i=1}^{n} \mathcal{S}(\boldsymbol{\theta}; x_i).
\end{aligned} \tag{4.1.9}$$

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 $\boldsymbol{\theta} \in N_{\boldsymbol{\theta}_0}$, $\mathrm{LS}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \subset N_{\boldsymbol{\theta}_0}$. Then by Taylor's theorem with the Lagrange form of the remainder, there exists $\tilde{\boldsymbol{\theta}}_j \in \mathrm{LS}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ such that the j-th component of $\mathcal{S}(\boldsymbol{\theta}; x_i)$ may be expanded as

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

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

$$
\begin{pmatrix} \mathcal{S}_1(\boldsymbol{\theta}; x_i) \\ \vdots \\ \mathcal{S}_d(\boldsymbol{\theta}; x_i) \end{pmatrix} = \begin{pmatrix} \mathcal{S}_1(\boldsymbol{\theta}_0; x_i) \\ \vdots \\ \mathcal{S}_d(\boldsymbol{\theta}_0; x_i) \end{pmatrix} + \left[\begin{pmatrix} \nabla^{\top}\mathcal{S}_1(\boldsymbol{\theta}_0; x_i) \\ \vdots \\ \nabla^{\top}\mathcal{S}_d(\boldsymbol{\theta}_0; x_i) \end{pmatrix} + M(x_i)O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)\mathbf{1}_{d\times d} \right](\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\top}. \quad (4.1.10)
$$

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 2.2.3, this is just the negative transpose of the observed information matrix also evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\mathcal{J}(\boldsymbol{\theta}_0)$. 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 4.1.10 becomes

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

$$(4.1.11)$$

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

$$
\begin{aligned}
\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n) &= \sum_{i=1}^{n} \mathcal{S}(\boldsymbol{\theta}; x_i) \\
&= \sum_{i=1}^{n} \left[\mathcal{S}(\boldsymbol{\theta}_0; x_i) - \left[\mathcal{J}_{X_i}(\boldsymbol{\theta}_0) + M(x_i) O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||) \mathbf{1}_{d \times d} \right] (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \right] \\
&= \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) - \left[\mathcal{J}_{\mathbf{X}_n}(\boldsymbol{\theta}_0) + \left\{ \sum_{i=1}^{n} M(x_i) \right\} O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||) \mathbf{1}_{d \times d} \right] (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \\
&= \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) - \left[\frac{1}{n} \mathcal{J}_{\mathbf{X}_n}(\boldsymbol{\theta}_0) + \left\{ \frac{1}{n} \sum_{i=1}^{n} M(x_i) \right\} O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||) \mathbf{1}_{d \times d} \right] n(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top.
\end{aligned}
\tag{4.1.12}
$$

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

$$
\frac{1}{\sqrt{n}} \mathcal{S}(\boldsymbol{\theta}; \mathbf{x}_n) = \frac{1}{\sqrt{n}} \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) - \left[\frac{1}{n} \mathcal{J}_{\mathbf{X}_n}(\boldsymbol{\theta}_0) + \left\{ \frac{1}{n} \sum_{i=1}^{n} M(x_i) \right\} O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||) \mathbf{1}_{d \times d} \right] \sqrt{n} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top. \tag{4.1.13}
$$

We previously established 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 4.1.13 gives

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

and therefore,

$$
\frac{1}{\sqrt{n}} \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) = \left[\frac{1}{n} \mathcal{J}_{\mathbf{X}_n}(\boldsymbol{\theta}_0) + \left\{ \frac{1}{n} \sum_{i=1}^{n} M(x_i) \right\} O_p(||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||) \mathbf{1}_{d \times d} \right] \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top. \tag{4.1.14}
$$

For the terms in the square brackets in the line above, we have the following observations:

1) By Equation 3.5.4, $\mathrm{E}_{\boldsymbol{\theta}_0} [\mathcal{J}_X(\boldsymbol{\theta}_0)] = \mathscr{I}_X(\boldsymbol{\theta}_0)$, and thus $\frac{1}{n} \mathcal{J}_{\mathbf{X}_n}(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{J}_{X_i}(\boldsymbol{\theta}_0)$ is converging in probability to $\mathscr{I}_X(\boldsymbol{\theta}_0)$ as $n \to \infty$ by the Weak Law of Large Numbers, i.e., $\mathscr{I}_X(\boldsymbol{\theta}_0) = \frac{1}{n} \mathcal{J}_{\mathbf{X}_n}(\boldsymbol{\theta}_0) + o_p(1)$.

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 $\mathscr{I}_X(\boldsymbol{\theta}_0)$ as $n \to \infty$. This allows us to rewrite Equation 4.1.14 as

$$\frac{1}{\sqrt{n}} \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) = \left[\mathscr{I}_X(\boldsymbol{\theta}_0) + o_p(1) \right] \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top. \tag{4.1.15}$$

This is useful because if we know the distribution to which the term on the left-hand side is converging as $n \to \infty$, we can deduce the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$ using Slutsky's theorem.

By definition, $\text{Var}_{\boldsymbol{\theta}_0}[\mathcal{S}(\boldsymbol{\theta}_0; x_i)] = \mathscr{I}_X(\boldsymbol{\theta}_0)$. From Equation 4.1.9, we have that $\frac{1}{n}\mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}(\boldsymbol{\theta}_0; x_i)$. It follows from the Central Limit Theorem that

$$\sqrt{n}\left(\frac{1}{n} \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) - \text{E}_{\boldsymbol{\theta}_0}\left[\mathcal{S}(\boldsymbol{\theta}_0; x_i) \right] \right) \xrightarrow{d} \text{N}\left(\mathbf{0}, \mathscr{I}_X(\boldsymbol{\theta}_0) \right) \text{ as } n \to \infty. \tag{4.1.16}$$

But by the first Bartlett identity, $\text{E}_{\boldsymbol{\theta}_0}\left[\mathcal{S}(\boldsymbol{\theta}_0; x_i) \right] = \mathbf{0}$, and therefore

$$\frac{1}{\sqrt{n}} \mathcal{S}(\boldsymbol{\theta}_0; \mathbf{x}_n) \xrightarrow{d} \text{N}\left(\mathbf{0}, \mathscr{I}_X(\boldsymbol{\theta}_0) \right) \text{ as } n \to \infty. \tag{4.1.17}$$

Combining the results of Equation 4.1.15 and Equation 4.1.17, we see that

$$\left[\mathscr{I}_X(\boldsymbol{\theta}_0) + o_p(1) \right] \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \xrightarrow{d} \text{N}\left(\mathbf{0}, \mathscr{I}_X(\boldsymbol{\theta}_0) \right) \text{ as } n \to \infty. \tag{4.1.18}$$

Since the term in square brackets is converging in probability to $\mathscr{I}(\boldsymbol{\theta}_0)$, which by **RC8** is a positive definite matrix, a minor extension of Slutsky's Theorem (see Appendix A) allows us to deduce that the asymptotic distribution of the MLE for the true parameter value of a regular model is given by

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \xrightarrow{d} \mathscr{I}_X(\boldsymbol{\theta}_0)^{-1} \text{N}\left(\mathbf{0}, \mathscr{I}_X(\boldsymbol{\theta}_0) \right)$$
$$\xrightarrow{d} \text{N}\left(\mathbf{0}, \mathscr{I}_X(\boldsymbol{\theta}_0)^{-1} \right) \text{ as } n \to \infty. \tag{4.1.19}$$

4.1.2. 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 et al. (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 et al. (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

$$X_{ij} \sim p_{ij}(x_{ij}; \psi, \lambda_i), \tag{4.1.20}$$

where $i = 1, ..., q$ and $j = 1, ..., m_i$, 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 et al. (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 4.1.20 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

$$L^{(i)}(\psi, \lambda_i) = p_i(\mathbf{x}_i; \psi, \lambda_i), \tag{4.1.21}$$

and

$$\ell^{(i)}(\psi, \lambda_i) = \log L^{(i)}(\psi, \lambda_i), \tag{4.1.22}$$

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

$$\bar{L}^{(i)}(\psi) = \int_\Lambda L^{(i)}(\psi, \lambda_i) g(\lambda_i; \psi) d\lambda_i. \tag{4.1.23}$$

From here, we proceed by using Laplace's method as described by Tierney & 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 4.1.23 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

$$\hat{\bar{L}}^{(i)}(\psi) = f(\hat{\lambda}_{i\psi})\sqrt{\frac{2\pi}{m}}\sigma\exp[-mh(\hat{\lambda}_{i\psi})], \tag{4.1.24}$$

where

$$\sigma = \left[\left.\frac{\partial^2 h}{\partial\lambda_i^2}\right|_{\lambda_i=\hat{\lambda}_{i\psi}}\right]^{-1/2}$$

$$= \left[-\frac{1}{m}\left.\frac{\partial^2 \ell^{(i)}(\psi,\lambda_i)}{\partial\lambda_i^2}\right|_{\lambda_i=\hat{\lambda}_{i\psi}}\right]^{-1/2}$$

$$= \left[\frac{1}{m}\mathcal{I}(\hat{\lambda}_{i\psi})\right]^{-1/2}.$$

Here, $\mathcal{I}(\hat{\lambda}_{i\psi})$ 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 4.1.24, we arrive at

$$\begin{aligned}
\hat{\bar{L}}^{(i)}(\psi) &= f(\hat{\lambda}_{i\psi})\sqrt{\frac{2\pi}{m}}\sigma\exp[-mh(\hat{\lambda}_{i\psi})] \\
&= g(\hat{\lambda}_{i\psi};\psi)\sqrt{\frac{2\pi}{m}}\left[\frac{1}{m}\mathcal{I}(\hat{\lambda}_{i\psi})\right]^{-1/2}\exp\left\{-m\cdot-\frac{1}{m}\ell^{(i)}(\psi,\hat{\lambda}_{i\psi})\right\} \\
&= \frac{\sqrt{2\pi}}{m}L^{(i)}(\psi,\hat{\lambda}_{i\psi})g(\hat{\lambda}_{i\psi};\psi)[\mathcal{I}(\hat{\lambda}_{i\psi})]^{-1/2} \\
&= \frac{\sqrt{2\pi}}{m}L_P^{(i)}(\psi)g(\hat{\lambda}_{i\psi};\psi)[\mathcal{I}(\hat{\lambda}_{i\psi})]^{-1/2},
\end{aligned} \tag{4.1.25}$$

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

$$\bar{L}^{(i)}(\psi) = \hat{\bar{L}}^{(i)}(\psi)\left\{1 + O\left(\frac{1}{m}\right)\right\} \quad \text{as} \quad m \to \infty. \tag{4.1.26}$$

Let

$$\bar{\ell}^{(i)}(\psi) = \log\bar{L}^{(i)}(\psi) \tag{4.1.27}$$

denote the integrated log-likelihood for ψ. Putting the results in Equation 4.1.25, Equation 4.1.26, and Equation 4.1.27 together, we have

$$\bar{\ell}^{(i)}(\psi) = \log \bar{L}^{(i)}(\psi) \qquad \text{(by Equation 4.2.8)}$$

$$= \log\left(\hat{\bar{L}}^{(i)}(\psi)\left\{1 + O\left(\frac{1}{m}\right)\right\}\right) \qquad \text{(by Equation 4.2.7)}$$

$$= \log \hat{\bar{L}}^{(i)}(\psi) + \log\left\{1 + O\left(\frac{1}{m}\right)\right\} \qquad \text{(by Equation 4.2.6)}$$

$$= \log\left\{\frac{\sqrt{2\pi}}{m} L_P^{(i)}(\psi) g(\hat{\lambda}_{i\psi}; \psi)[\mathcal{J}(\hat{\lambda}_{i\psi})]^{-1/2}\right\} + O\left(\frac{1}{m}\right)$$

$$= \frac{1}{2}\log(2\pi) - \log(m) + \log L_P^{(i)}(\psi) + \log g(\hat{\lambda}_{i\psi}; \psi) - \frac{1}{2}\log \mathcal{J}(\hat{\lambda}_{i\psi}) + O\left(\frac{1}{m}\right)$$

$$= \ell_P^{(i)}(\psi) + \log g(\hat{\lambda}_{i\psi}; \psi) - \frac{1}{2}\log \mathcal{J}(\hat{\lambda}_{i\psi}) + \frac{1}{2}\log(2\pi) - \log(m) + O\left(\frac{1}{m}\right) \quad \text{as} \quad m \to \infty,$$

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:

$$\hat{\bar{\ell}}^{(i)}(\psi) = \ell_P^{(i)}(\psi) + \log g(\hat{\lambda}_{i\psi}; \psi) - \frac{1}{2}\log \mathcal{J}(\hat{\lambda}_{i\psi}). \qquad (4.1.28)$$

The error in this approximation is given by

$$\bar{\ell}^{(i)}(\psi) = \hat{\bar{\ell}}^{(i)}(\psi) + O\left(\frac{1}{m}\right). \qquad (4.1.29)$$

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

$$L(\psi, \lambda) = \prod_{i=1}^{q} L^{(i)}(\psi, \lambda_i) \qquad (4.1.30)$$

and

$$\ell(\psi, \lambda) = \sum_{i=1}^{q} \ell^{(i)}(\psi, \lambda_i), \qquad (4.1.31)$$

respectively. If we define the weight function

$$G(\lambda; \psi) \equiv \prod_{i=1}^{q} g(\lambda_i; \psi) \tag{4.1.32}$$

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

$$\begin{aligned}
\bar{L}(\psi) &= \int_{\Lambda^q} L(\psi, \lambda) G(\lambda; \psi) d\lambda \\
&= \int_{\Lambda} \cdots \int_{\Lambda} \left[\prod_{i=1}^{q} L^{(i)}(\psi, \lambda_i) \right] \left[\prod_{i=1}^{q} g(\lambda_i; \psi) \right] d\lambda_1 \cdots d\lambda_q \\
&= \prod_{i=1}^{q} \int_{\Lambda} L^{(i)}(\psi, \lambda_i) g(\lambda_i; \psi) d\lambda_i \\
&= \prod_{i=1}^{q} \bar{L}^{(i)}(\psi).
\end{aligned} \tag{4.1.33}$$

Let $\bar{\ell}(\psi) = \log \bar{L}(\psi)$ denote the integrated log-likelihood function for ψ. Taking the logarithm of both sides in Equation 4.1.33, we have

$$\bar{\ell}(\psi) = \sum_{i=1}^{q} \bar{\ell}^{(i)}(\psi). \tag{4.1.34}$$

Plugging in our approximation to $\bar{\ell}^{(i)}(\psi)$ and its error term in Equation 4.1.28 and Equation 4.1.29, respectively, yields

$$\begin{aligned}
\bar{\ell}(\psi) &= \sum_{i=1}^{q} \left[\ell_P^{(i)}(\psi) + \log g(\hat{\lambda}_{i\psi}; \psi) - \log \mathcal{I}(\hat{\lambda}_{i\psi}) + O\left(\frac{1}{m}\right) \right] \\
&= \ell_P(\psi) + \sum_{i=1}^{q} \log g(\hat{\lambda}_{i\psi}; \psi) - \frac{1}{2} \sum_{i=1}^{q} \log \mathcal{I}(\hat{\lambda}_{i\psi}) + O\left(\frac{q}{m}\right) \quad \text{as} \quad m \to \infty.
\end{aligned} \tag{4.1.35}$$

4.2. Frequentist Properties

4.3. Dependence on Nuisance Parameter Weight Function

4.4. Parameterization Invariance

4.4.1. The Zero-Score Expectation Parameter

Severini ([2007](#)) considered the problem of selecting a weight function $\pi(\lambda|\psi)$ such that when the likelihood function is integrated with respect to this density over the nuisance parameter space, the result is useful for non-Bayesian inference. To do this, he outlined four properties (see Appendix **??**) that an integrated likelihood must satisfy if it is to be of any use and went on to show that such a function could be obtained by doing the following:

1) Find a reparameterization $(\psi, \lambda) \mapsto (\psi, \phi)$ of the model such that the new nuisance parameter ϕ is unrelated to ψ in the sense that $\hat{\phi}_\psi = \hat{\phi}$; that is, the conditional maximum likelihood estimate of ϕ given ψ is simply equal to the unrestricted maximum likelihood estimate of ϕ.

2) Select a prior density $\pi(\lambda|\psi)$ such that when the model undergoes the above reparameterization, the resulting prior density $\pi(\phi)$ does not depend on ψ.

An integrated likelihood function for ψ that possesses the desired properties will then be given by

$$\bar{L}(\psi) = \int_\Phi \tilde{L}(\psi, \phi)\pi(\phi)d\phi, \tag{4.4.1}$$

where $\tilde{L}(\psi, \phi)$ is the likelihood function for the model after it has been reparameterized in terms of ϕ. The exact choice of prior density for ϕ is not particularly important; the only restriction placed upon it is that it must not depend on ψ. Hence, the crux of the matter really lies in completing the first step. The approach taken by Severini ([2007](#)) is to define a new nuisance parameter ϕ as the solution to the equation

$$\mathrm{E}_{(\psi_0, \lambda_0)}\left[\nabla_\lambda \ell(\psi, \lambda; \mathbf{X}_n)\right]\bigg|_{(\psi_0, \lambda_0)=(\hat{\psi}, \phi)} = \mathbf{0}, \tag{4.4.2}$$

where ψ_0 and λ_0 denote the true values of ψ and λ, and $\widehat{\psi}$ is the unrestricted MLE for ψ_0. The expectation here is being taken with respect to the data $\mathbf{X}_n = \mathbf{x}_n$ and not the parameters themselves. Equation 4.4.2 can thus be rewritten as

$$I(\psi, \lambda, \widehat{\psi}, \phi) = \mathbf{0},$$

where

$$I(\psi, \lambda, \psi_0, \lambda_0) = \int_{\mathbb{R}^n} [\nabla_\lambda \ell(\psi, \lambda; \mathbf{x}_n)] p(\mathbf{x}_n; \psi_0, \lambda_0)) d\mathbf{x}_n.$$

Assuming I is invertible, for a particular value of $(\psi, \lambda, \widehat{\psi})$, there will be a unique value of ϕ that solves Equation 4.4.2. ϕ is called the *zero-score expectation* (ZSE) parameter because it is defined as the value that makes the expectation of the score function (in terms of λ, not the full parameter) with respect to $p(\mathbf{x}_n; \psi_0, \lambda_0)$ evaluated at the point $(\psi_0, \lambda_0) = (\widehat{\psi}, \phi)$ equal to zero. This means that ϕ is really a function of $(\psi, \lambda, \widehat{\psi})$, i.e., $\phi = \phi(\psi, \lambda, \widehat{\psi})$. This in turn implies that ϕ is a function of the data through $\widehat{\psi}$. Normally we try to 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 ϕ, the corresponding value of λ can be found by

$$\lambda(\psi, \phi) = \operatorname*{argmax}_{\lambda \in \Lambda} \mathrm{E}_{(\widehat{\psi}, \phi)}[\ell(\psi, \lambda; \mathbf{X}_n)]. \tag{4.4.3}$$

For a certain choice of prior density $\pi(\phi)$, this allows us to write Equation 4.4.1 in terms of $L(\psi, \lambda)$:

$$\bar{L}(\psi) = \int_\Phi L(\psi, \lambda(\psi, \phi)) \pi(\phi) d\phi. \tag{4.4.4}$$

CHAPTER 5

Methodology

5.1. Explicit Parameters of Interest

5.2. Implicit Parameters of Interest

The procedure described in the previous section is based on the assumption that λ is an explicit nuisance parameter so that taking partial derivatives of ℓ with respect to its components is a well-defined operation. However, Severini (2018) proved that reparameterizing the model in terms of the ZSE parameter yields the same nice properties in the subsequent integrated likelihood when ψ and λ are implicit as well. In this section, we consider an approach to approximating the integrated likelihood function that has been adapted from this method.

Consider a model with parameter $\theta \in \Theta$ and implicit parameter of interest $\psi = \varphi(\theta)$ for some function $\varphi : \Theta \to \mathbb{R}$. Equation 4.4.4 tells us that we can calculate the integrated likelihood for ψ if we know the value of θ corresponding to a given value of the ZSE parameter, and this will be true for models with both explicit and implict parameters. Let $\hat{\psi} = \varphi(\hat{\theta})$ denote the unrestricted MLE for ψ and define the set

$$\Omega_{\psi} = \Big\{ \omega \in \Theta : \varphi(\omega) = \psi \Big\}. \tag{5.2.1}$$

Then an element of $\Omega_{\hat{\psi}}$ for a model without an explicit nuisance parameter is functionally equivalent to a value $(\hat{\psi}, \phi)$ for a model with an explicit nuisance parameter.

Generalizing the result in Equation 4.4.3, for a given element $\omega \in \Omega_{\hat{\psi}}$, the corresponding value of θ is that which maximizes $\mathrm{E}_{\omega}\big[\ell(\theta; \mathbf{X}_n)\big]$ subject to the restriction that $\varphi(\theta) = \psi$. This allows us to define a function $T_{\psi} : \Omega_{\hat{\psi}} \to \Theta(\psi)$ such that $T_{\psi}(\omega) = \underset{\theta \in \Theta(\psi)}{\mathrm{argmax}} \, \mathrm{E}_{\omega}\big[\ell(\theta; \mathbf{X}_n)\big]$. The integrated likelihood for ψ is

then given by

$$\bar{L}(\psi) = \int_{\Omega_{\hat{\psi}}} L(T_\psi(\omega))\pi(\omega)d\omega, \tag{5.2.2}$$

where $\pi(\omega)$ is a density defined on $\Omega_{\hat{\psi}}$.

Equation 5.2.2 can also be written as

$$\bar{L}(\psi) = \mathrm{E}_\omega\Big[L(T_\psi(W))\Big], \tag{5.2.3}$$

where W represents a random variable with density $\pi(\omega)$. We can further define a function $Q : \mathcal{U} \to \Omega_{\hat{\psi}}$ for some set $\mathcal{U}$ such that if U is a random variable taking values in $\mathcal{U}$, then $Q(U)$ will be a random variable in $\Omega_{\hat{\psi}}$ with a distribution that is completely determined by our choice of U. Therefore,

$$\bar{L}(\psi) = \mathrm{E}_\omega\Big[L(T_\psi(Q(U)))\Big] \tag{5.2.4}$$

is an integrated likelihood for ψ with a weight function corresponding to the density of $Q(U)$.

Define

$$\tilde{L}(u;\psi) = L(T_\psi(Q(u)), \ u \in \mathcal{U}. \tag{5.2.5}$$

Then we simply have $\tilde{L}(u;\psi) = L(\boldsymbol{\theta})$, with $\boldsymbol{\theta}$ taken to be $T_\psi(Q(u))$. $\tilde{L}(u;\psi)$ will be a genuine likelihood for the parameter (u,ψ) provided that there always exists a value (u,ψ) such that $T_\psi(Q(u)) = \boldsymbol{\theta}_0$ for any possible value of $\boldsymbol{\theta}_0$. A sufficient condition for this to occur is that for any $\psi \in \Psi$, $T_\psi(Q(\mathcal{U})) = \Omega_\psi$.

We can then write the integrated likelihood for ψ in terms of $\tilde{L}(u;\psi)$ as follows:

$$\bar{L}(\psi) = \int_{\mathcal{U}} \tilde{L}(u;\psi)\breve{\pi}(u)du, \tag{5.2.6}$$

where $\breve{\pi}(u)$ is a density on $\mathcal{U}$ of our choosing. If $\breve{\pi}(u)$ doesn't depend on ψ, the integrated likelihood given by Equation 5.2.6 will have the properties we desire.

We can further rewrite Equation 5.2.6 as

$$\bar{L}(\psi) = \int_\Theta \frac{\tilde{L}(u;\psi)}{\breve{L}(u)}\breve{L}(u)\breve{\pi}(u)du, \tag{5.2.7}$$

where $\check{L}(u)$ represents an arbitrary likelihood function for the "parameter" u. From a Bayesian point of view, the quantity $\check{L}(u)\check{\pi}(u)$ can then be thought of as a posterior density for u up to some proportionality constant. If $\check{\pi}(u)$ is chosen to be a conjugate prior for $\check{L}(u)$ such that the posterior density $\check{L}(u)\check{\pi}(u)$ has the same form as $\check{L}$ itself, and $\check{L}$ is chosen to be a known distribution, then random samples can be drawn directly from this posterior density using modern statistical computing packages. Alternatively, it may also be possible to obtain random samples from the posterior density through Monte Carlo methods such as importance sampling or MCMC.

In either case, once random variates $u_1, ..., u_R$ have been sampled from $\check{L}(u)\check{\pi}(u)$, a simple empirical estimate to $\bar{L}(\psi)$ at a particular value of ψ is given by

$$\hat{\bar{L}}(\psi) = \frac{1}{R} \sum_{i=1}^{R} \frac{\tilde{L}(u_i; \psi)}{\check{L}(u_i)}. \tag{5.2.8}$$

Repeating this procedure for every value of $\psi \in \Psi$ will give an overall shape for $\hat{\bar{L}}(\psi)$, which can be used to conduct inference for ψ_0 without interference from a nuisance parameter.

5.3. Laplace's Method

5.4. Monte Carlo Methods

5.4.1. Simple Monte Carlo

5.4.2. Importance Sampling

To obtain an integrated likelihood for ψ alone, we can use the procedure described in the previous chapter to find an approximation to the integral in Equation 5.2.7 where $\tilde{L}(u; \psi)$ is the likelihood function reparameterized in terms of the ZSE parameter, and $\check{L}(u)$ and $\check{\pi}(u)$ are chosen such that $\check{\pi}$ is a conjugate prior for $\check{L}$.[1] In this case, a natural choice exists due to the fact that the Dirichlet distribution is a conjugate prior for the multinomial distribution. Since our original likelihood function L is based on a multinomial distribution, we can simply set $\check{L}(u) := L(u)$ and $\check{\pi}(u) \sim \text{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_d)$.

---

[1]This procedure is an adapted version of another algorithm developed by Severini (2022) for approximating the integrated likelihood for the entropy of a multinomial distribution.

Then the posterior distribution for u based on data $\mathbf{n} = (n_1, ..., n_d)$ is given by $L(u)\breve{\pi}(u) \sim \text{Dir}(\mathbf{n} + \boldsymbol{\alpha})$. Consequently, we will take $\breve{\pi}(u)$ to be the symmetric Dirichlet distribution on the probability simplex in $\mathbb{R}^d$ with $\alpha_j = 1$ for all j so that random variates of u can be sampled from a $\text{Dir}(\mathbf{n} + 1)$ distribution.

From Equation 5.2.5, finding $\widetilde{L}(u; \psi)$ is a matter of defining two functions, Q and T_ψ such that $\widetilde{L}(u; \psi) = L(T_\psi(Q(u)))$. Q maps a random variate u sampled from the above posterior to an element ω in $\Omega_{\hat{\psi}}$, and T_ψ then maps ω to an element $\boldsymbol{\theta}$ in $\Theta(\psi)$. Since in this situation, these quantities u, ω, and $\boldsymbol{\theta}$ are all members of the probability simplex, the maps Q and T_ψ can be defined as returning the elements in their respective output spaces that are closest to the input element they have each received. That is, Q returns the element ω that minimizes the distance to an input u, subject to the constraints that the sum of the components of ω equal 1 and $\varphi(\omega) = \hat{\psi}$. Similarly, T_ψ returns the element $\boldsymbol{\theta}$ that minimizes the distance to an input ω, subject to the constraints that the sum of the components of $\boldsymbol{\theta}$ equal 1 and $\varphi(\boldsymbol{\theta}) = \psi$.

From here, we sample R random variates from the appropriate Dirichlet distribution, calculate $\widetilde{L}(u; \psi)$ and $L(u)$ for each one, and use Equation 5.2.8 to approximate the integrated likelihood at a particular value of ψ. We then repeat this process over a finely spaced sequence of the possible values of ψ in order to get a shape of the overall integrated likelihood. See Appendix **??** for a graph comparing the plot of one such integrated likelihood to the profile likelihood, for observed data $(n_1, ..., n_6) = (1, 1, 4, 7, 10)$ and 250 samples of the appropriate Dirichlet distribution drawn for each value of ψ.[2]

---

[2]The samples were obtained using the 'LaplacesDemon' R package. The distance minimizations needed for Q and T_ψ were computed numerically using the 'nloptr' R package.

CHAPTER 6

Multinomial Parameters of Interest

6.1. Introduction

This chapter concerns parameters of interest derived from a multinomial distribution using the integrated likelihood function. The parameters of interest considered are various measures of diversity, such as Shannon entropy and Simpson's diversity index, as well as a shared effect coefficient in a multinomial generalized linear model. We begin with a brief review of the multinomial distribution and its relevant properties.

A random vector $\mathbf{Y} = (Y_1, ..., Y_J)$ follows a categorical distribution with parameter

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_J) \in \Delta_{J-1}$$

if $\sum_{j=1}^{J} Y_j = 1$, $Y_j \in \{0, 1\}$, and $\mathbb{P}(Y_j = 1) = \theta_j$ for all j, where

$$\Delta_{J-1} \equiv \left\{ \boldsymbol{\theta} \in [0, 1]^J : \sum_{j=1}^{J} \theta_j = 1 \right\}$$

denotes the $(J-1)$-dimensional probability simplex. A multinomial random vector arises as the sum of n i.i.d. categorical random vectors:

$$\sum_{i=1}^{n} \mathbf{Y}_i \sim \text{Multinomial}(n, \theta), \;\; \mathbf{Y}_i \overset{\text{i.i.d.}}{\sim} \text{Categorical}(\theta), \;\; i = 1, ..., n.$$

Consequently, the categorical distribution is a special case of the multinomial distribution with $n = 1$, mirroring the relationship between the Bernoulli and binomial distributions.[1]

---

[1] The categorical distribution is also known as the *multinoulli* distribution.

Let $(N_1, ..., N_J) \sim \text{Multinomial}(n, \boldsymbol{\theta})$, where N_j denotes the number of observations falling in category j and $n = \sum_{j=1}^{J} N_j$. Each marginal count satisfies $N_j \sim \text{Binomial}(n, \theta_j)$, with $\mathbb{E}_{\boldsymbol{\theta}}(N_j) = n\theta_j$. Since $\boldsymbol{\theta}$ represents category probabilities, the natural parameter space is the probability simplex $\Theta = \Delta_{J-1} \subset \mathbb{R}^J$. The likelihood and log-likelihood functions are given by

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} \theta_j^{N_j}$$

and

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{J} N_j \log \theta_j,$$

respectively. Maximization of $\ell(\boldsymbol{\theta})$ subject to $\boldsymbol{\theta} \in \Delta_{J-1}$ yields the familiar maximum likelihood estimator

$$\hat{\theta}_j = \frac{N_j}{n}, \;\; j = 1, ..., J.$$

6.2. Shannon Entropy

The Shannon entropy of a discrete random variable X with support $\mathcal{X}$ and probability mass function p is given by

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$

with the convention that $0 \log 0 = 0$. In general, $H(X) \geq 0$, with equality if and only if $p(x^*) = 1$ for some $x^* \in \mathcal{X}$ and $p(x) = 0$ for all $x \neq x^*$. If $\mathcal{X}$ is finite with cardinality $J = |\mathcal{X}|$, then $H(X) \leq J$, with equality if and only if $p(x) = 1/J$ for all $x \in \mathcal{X}$.

Thus, Shannon entropy is minimized when all probability mass is placed on a single outcome and maximized when it is distributed unifomrly across all possible outcomes. This property makes it a natural measure of the relative concentration or dispersion of probability mass over a finite set of categories. In applied settings, entropy is commonly used as an index of diversity, for example to quantify species heterogeneity in ecological communities.

In the present setting, our parameter of interest is the entropy of a single draw from a multinomial model. Although the observed data consist of aggregated counts $(N_1, ..., N_J)$ arising from n independent

realizations of a categorical random vector, the entropy itself is a property of the underlying categorical distribution with probability vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_J)$. Equivalently, the multinomial model describes the distribution of the *aggregate* $(N_1, ..., N_J)$, whereas the entropy we are interested in is a property of the *single-trial* distribution. Let $X \in \{1, ..., J\}$ denote the outcome of one trial, with $\mathbb{P}_\theta(X = j) = \theta_j$. Then

$$H(X) = -\sum_{j=1}^{J} \theta_j \log \theta_j,$$

and so we may define the estimand

$$\psi(\theta) := H(X) = -\sum_{j=1}^{J} \theta_j \log \theta_j.$$

Hence, while the data enter the model through the multinomial likelihood for θ, the parameter of interest is the entropy of the corresponding categorical distribution.

6.3. Simpson's Diversity Index

6.4. Shared Effect

6.5. Discussion

CHAPTER 7

Count Dyads

This chapter examines the utility of the integrated likelihood function as a tool for estimating parameters of interest in dyads with models of data generated from count distributions. In each dyad we investigate, we will compare and contrast the performances of the point and interval estimators produced by the integrated likelihood (or more commonly, our approximations to the integrated likelihood) for the dyad's parameter of interest to the analogous estimators produced by the profile likelihood function. Our use of the profile likelihood stems from its conceptual simplicity and ease of computation, making it a convenient choice for an estimation benchmark.

7.1. Weighted Sums of Rate Parameters

To motivate our discussion, let us imagine a hypothetical scenario to which we will return at various points throughout this section for context. Suppose there has been an outbreak of disease among a group of hospitals, and we have been tasked with assessing its severity. A basic question to ask is, "What has been the overall rate of infection across all affected hospitals since the start of the outbreak?" A seemingly straightforward answer is the simple ratio of total infections to total patient-days reported across all the hospitals since the start of the outbreak.[1] We must tread with some caution here, however. The number of infections is clearly random, but what about the patient-days? The properties of our estimator hinge upon the answer so it would be prudent to give it some consideration.

To interpret the number of patient-days as a random variable is to declare that what really matters is not the rate of infection among hospitals in this outbreak in particular but rather the more abstract notion of the disease's "true" rate of infection in general. This would mean that both the numerator and denominator in our ratio are random variables, transforming it from an "estimator which happens to be

---

[1]A patient-day is a single day spent by a patient in a hospital.

a ratio" into an actual "ratio estimator". Conversely, to treat them as fixed parameters is to condition on their exact values reported by each hospital in the group, thereby restricting our attention to the rate of infection among these hospitals only. Neither choice is inherently right or wrong - it is up to us to decide where our interest lies - but once we have decided, we must honor the consequences of that choice in our analysis. Failure to do so would lead to bias in our estimates. We will begin with the assumption that the point of our analysis is to gauge the risk involved with this outbreak specifically, so until such time as it is explicitly noted otherwise, it is safe to regard the patient-days that each hospital reports as being fixed parameters.

The overall rate of infection among only the hospitals in our group is therefore what constitutes our parameter of interest. As usual we will call this value ψ and its estimator $\hat{\psi}$. In practice, especially for an example such as this, the line between ψ and $\hat{\psi}$ often gets blurred, with many treating the observed rate as though it is the same as the true underlying rate. From the perspective of assessing impact, this is understandable as the observed rate is by definition what the hospitals actually experience. However, for the sake of avoiding conceptual ambiguity in our language and intent, we will continue to treat them as two distinct entities. If nothing else, it serves as a friendly reminder of the importance of maintaining proper separation between estimand and estimator.

Assume there are J hospitals in the group, and let Y_j and t_j represent the number of infections and patient-days, respectively, reported by hospital j since the start of the outbreak. Then our estimator may be written as

$$
\begin{aligned}
\hat{\psi} &= \frac{\sum_{j=1}^{J} Y_j}{\sum_{j=1}^{J} t_j} \\
&= \frac{\sum_{j=1}^{J} t_j (Y_j / t_j)}{\sum_{j=1}^{J} t_j} \\
&= \frac{\sum_{j=1}^{J} t_j \hat{\theta}_j}{\sum_{j=1}^{J} t_j} \\
&= \sum_{j=1}^{J} t_j' \hat{\theta}_j,
\end{aligned}
\tag{7.1.1}
$$

where $t'_j \equiv \frac{t_j}{\sum_{j=1}^{J} t_j}$ and $\hat{\theta}_j \equiv Y_j/t_j$ is the observed rate of infection at hospital j. If θ_j represents its associated estimand, we can similarly rewrite our expression for ψ in a similar form as

$$\psi = \sum_{j=1}^{J} t'_j \theta_j. \tag{7.1.2}$$

Thus, the overall rate of infection can be decomposed as the weighted *average* of the individual hospitals' rates of infection. Note that the term "average" is proper here since $\sum_{j=1}^{J} t'_j = 1$. This follows directly from our requirement that each weight t'_j be proportional to the number of patient-days contributed by hospital j to the total number of patient-days across all the hospitals since the outbreak began. We made this restriction as a consequence of the question that motivated our analysis, but it was entirely arbitrary. There is nothing actually requiring the weights to sum to one, or indeed that they depend on the patient-days at all. We could just as easily have asked about an aggregated infection rate relative to some other hospital-specific metric (e.g. number of available beds, population of surrounding region, etc.) that would in turn inform the precise values of the weights we choose for our parameter of interest.

And of course the utility of a weighted sum as a parameter of interest is not unique to our example of evaluating the severity of a disease outbreak in a group of hospitals either. In general, we may consider any set of J independent count-generating processes with distinct rates (countable events per unit time) of $\theta_1, ..., \theta_J$ and corresponding weights $\alpha_1, ..., \alpha_J$ as inducing a parameter of interest of the form

$$\psi = \sum_{j=1}^{J} \alpha_j \theta_j. \tag{7.1.3}$$

It is not necessary to assume that the weights are normalized, though we will still require that $\alpha_j > 0$ for all j in order to retain the interpretation of ψ as being a weighted *sum* of individual rates.

In the following sections, we will assess the ability of the integrated likelihood to produce accurate estimates for ψ under three separate model frameworks. The models we will consider are (i) a naive one in which we estimate the rate in each group using only the group's observed count and exposure time, (ii) a fixed effects regression model in which we assume there is an underlying covariate structure influencing

the observed counts in each group, and (iii) a hierarchical mixed effects regression model in which we assume some or all of the intercepts and slopes from the previous model are random effects instead of fixed. We will also take into account the robustness of the pseudolikelihood-derived estimators to model misspecification when evaluating their performance.

7.1.1. Naive Rates

We begin with a model characterized by its sole use of a process's observed count of events over some recorded interval of time to estimate its associated rate. This is perhaps the simplest nontrivial model one can employ for our parameter of interest, so we will refer to it as the *naive rates model*.

Let Y_j denote the count of events that we record from the j-th process during an exposure time t_j (the patient-days in our outbreak example). We will proceed for the moment under the assumption that the mean of the random variable underlying each count-generating process is roughly equal to its variance, meaning there is no risk of overdispersion. It is therefore safe to assume that the counts follow a Poisson distribution, i.e.

$$Y_j \sim \text{Poisson}(t_j \theta_j).$$

The log-likelihood for θ then becomes

$$\ell(\theta; Y) = \sum_{j=1}^{J} \left(Y_j \log \theta_j - t_j \theta_j \right),$$

where we have discarded any additive terms not depending on θ. The unconstrained MLE for θ_j is given by $\hat{\theta}_j = Y_j / t_j$, with the corresponding MLE for ψ being obtained by plugging these estimates into Equation 7.1.3, yielding

$$\hat{\psi} = \sum_{j=1}^{J} \alpha_j \hat{\theta}_j. \tag{7.1.4}$$

Since ψ is a function of the full model parameter, we can employ the ZSE parameterization method for an implicit nuisance parameter introduced in Severini (2018) to approximate the integrated likelihood. Recall that elements of the subspace $\Omega_{\hat{\psi}} = \{\omega \in \Theta : \alpha^\top \omega = \hat{\psi}\}$ play the role of the ZSE parameter in models for which the nuisance parameter is implicit. Let $Q(\theta; \omega)$ represent the expected value of $\ell(\theta; Y)$

under the probability distribution indexed by an element $\omega \in \Omega_{\hat{\psi}}$. That is,

$$Q(\theta; \omega) := \mathrm{E}_\omega[\ell(\theta; Y)]$$

$$= \mathrm{E}_\omega\left[\sum_{g=1}^G (Y_g \log \theta_g - t_g \theta_g)\right]$$

$$= \sum_{g=1}^G \left(\mathrm{E}_{\omega_g}[Y_g] \log \theta_g - t_g \theta_g\right)$$

$$= \sum_{g=1}^G (t_g \omega_g \log \theta_g - t_g \theta_g).$$

For a given value of ψ, the ZSE-constrained optimizer is

$$\tilde{\theta}(\psi; \omega) := \arg\max_\theta Q(\theta; \omega) \text{ s.t. } \alpha^\top \theta = \psi.$$

Evaluating $L(\theta; Y) := \exp(\ell(\theta; Y))$ at $\theta = \tilde{\theta}(\psi; \omega)$ gives us a mapping from ψ to its associated likelihood value under the distribution indexed by ω. If we integrate this mapping over $\Omega_{\hat{\psi}}$ with respect to some weight function $\pi(\omega)$ that doesn't depend on ψ, we can obtain the integrated likelihood function for ψ based on the ZSE parameterization, i.e.

$$\bar{L}(\psi) = \int_{\Omega_{\hat{\psi}}} L(\tilde{\theta}(\psi; \omega)) \pi(\omega) d\omega.$$

Unsurprisingly, no closed-form solution exists for this integral. Instead, we can use Monte Carlo integration to approximate it with the following estimator:

$$\hat{\bar{L}}(\psi) = \frac{1}{R} \sum_{j=1}^R L(\tilde{\theta}(\omega^{(j)}; \psi); Y),$$

where the choice of $\pi(\omega)$ is now understood to be defined implicitly as the density function admitted by the probability measure $d\Pi(\omega)$ that governs how each $\omega^{(j)}$ is drawn from $\Omega_{\hat{\psi}}$. As long as the selection does not depend on the value of ψ, we are essentially free to choose any (non-degenerate) probability distribution we wish, and the form of $\pi(\omega)$ will be adjusted accordingly.

Hence, under a simple MC approximation scheme, the functions $L(\tilde{\theta}(\omega^{(j)}; \psi); Y)$ for $j = 1, ..., R$ are individual "branches" of the integrated likelihood that we average to obtain an estimate of its true underlying value at a particular value of ψ. Let

$$b_\omega(\psi) := \log L(\tilde{\theta}(\psi; \omega^{(j)}); Y)$$

denote the branch curve generated by the value $\omega \in \Omega_{\hat{\psi}}$, and let its mode be given by

$$\psi_\omega^* := \arg\max_\psi b_\omega(\psi).$$

For an arbitrary dyad, two distinct values $\omega^{(1)}, \omega^{(2)} \in \Omega_{\hat{\psi}}$ will in general produce branch curves with modes $\psi_{\omega^{(1)}}^*$ and $\psi_{\omega^{(2)}}^*$, respectively, that are also distinct. In other words, there is no reason to expect that branch curves corresponding to different values drawn from $\Omega_{\hat{\psi}}$ will be maximized at the same location. However, if our dyad is such that the model belongs to the exponential family and our parameter of interest ψ is a linear function of the full model parameter, then this is no longer case. Indeed, it can be shown that under these two conditions, the branch curves of the integrated likelihood function for ψ will all have the same mode regardless of the values of ω that generated them, and furthermore that mode will simply be the unconstrained MLE for ψ, given by $\hat{\psi} = \alpha^\top \hat{\theta}$ (see appendix:A for a proof). Therefore our MC approximation to the integrated likelihood for ψ, being nothing more a simple average of these branches at each value of ψ, must also be maximized at $\hat{\psi}$.

Of course, by definition $\hat{\psi}$ is also the maximizer of the profile likelihood, meaning that the integrated likelihood and profile likelihood will produce identical maximum likelihood estimates for ψ in any dyad satisfying these criteria. We can further use a quadratic expansion of an arbitrary branch curve to show that the branch's curvature around its peak at $\hat{\psi}$ will be equal to that of the profile likelihood up to a second-order approximation (again, see appendix:A for a proof). It follows that the behavior of our MC approximation to the integrated likelihood for ψ, itself being nothing more than a simple average of these branches, will be virtually indistinguishable from that of the profile likelihood near $\hat{\psi}$ since each branch contributes a quadratic with the same peak and very similar curvature. Thus it is clear that for any

dyad satisfying these criteria, it does not matter whether we use the integrated or the profile likelihood to conduct inference regarding its parameter of interest - our conclusions will be identical. However, the profile likelihood is in general both conceptually and computationally easier to implement, making it the better choice in such cases.

7.1.2. Fixed Effects Regression

In the previous section, we relied solely on the observed counts and exposure times to estimate ψ. In practice, however, additional external factors often influence the counts we observe from each process. For instance, in the disease-outbreak example, variations in patient age distributions or other comorbidities across hospitals could affect the number of reported infections. To account for such influences, we turn to the standard framework of linear regression, which formally relates covariates to systematic variation in the observed counts.

Let Y_{ij} denote the i-th count from process j observed over an exposure time t_{ij}. Unlike the naive rates model in which only the total count per process was relevant, the presence of covariates that vary by observation requires us to retain each individual Y_{ij} and t_{ij}. If we were to ignore the covariates and sum over the index i, we would recover the structure of the naive model. Suppose that for each count we also observe a vector of covariates X_{ij} that may affect the rate at which process j generates events. We model these effects using a *fixed effects regression* framework, in which the coefficients associated with each covariate are treated as constant but unknown parameters to be estimated from the data. Some covariates may have *process-specific* effects (their impact varies across j), while others may have *process-agnostic* effects (affecting Y_{ij} uniformly across all processes $j = 1, ..., J$). We partition the covariate vector as

$$X_{ij} = \begin{pmatrix} X_{ij}^{(s)} \\ X_{ij}^{(a)} \end{pmatrix} \tag{7.1.5}$$

where $X_{ij}^{(s)} \in \mathbb{R}^{p \times 1}$ and $X_{ij}^{(a)} \in \mathbb{R}^{q \times 1}$ collect the process-specific and process-agnostic covariates, respectively.

Because covariates can influence the number of events recorded in each observation, the underlying rate

$$\theta_{ij} \equiv \theta_j(X_{ij}) \tag{7.1.6}$$

may vary with i. Each observed count Y_{ij} then has a conditional mean

$$\mu_{ij} \equiv \mathrm{E}(Y_{ij}|X_{ij}) = t_{ij}\theta_{ij}. \tag{7.1.7}$$

We define the *marginal rate* of process j as the expected rate averaged over the distribution of covariates, i.e.,

$$\bar{\theta}_j \equiv \mathrm{E}_{X_j}[\theta_j(X_j)] = \int \theta_j(X_j)dF_{X_j}(X_j), \tag{7.1.8}$$

where X_j represents a generic covariate vector associated with process j, drawn according to the joint distribution F_{X_j}. These marginal rates generalize the naive rates from the previous model by averaging over covariates rather than relying on raw totals. Our parameter of interest for this model may then be written as

$$\psi = \sum_{j=1}^{J} \alpha_j \bar{\theta}_j, \tag{7.1.9}$$

for some positive weights $\alpha_1, ..., \alpha_J$.

Up to this point, we have refrained from placing any distributional assumption on the observed counts. In this chapter, we focus on two commonly used distributions for count data: the Poisson and the negative binomial. Since both distributions belong to the exponential family, our model fits naturally within the framework of generalized linear models (GLMs), which provide a formal connection between the covariate-dependent rates $\theta_j(X_{ij})$ and the observed counts. In the following subsections, we examine each distribution in turn and assess how well an integrated likelihood function can produce estimators of weighted sums of the marginal rates $\bar{\theta}_j$ under these standard count models.

7.1.2.1. The Poisson GLM

Suppose the conditional distribution of Y_{ij} given X_{ij} is $Y_{ij}|X_{ij} \sim \mathrm{Poisson}(\mu_{ij})$. Then the conditional probability mass function of Y_{ij} can be expressed in canonical exponential-family form as

$$\mathrm{P}(Y_{ij} = y | X_{ij}) = \frac{e^{-\mu_{ij}} \mu_{ij}^y}{y!}$$

$$= \exp\left[y \log(\mu_{ij}) - \mu_{ij} - \log(y!)\right]$$

$$= \exp\left[y \log(t_{ij}\theta_{ij}) - t_{ij}\theta_{ij} - \log(y!)\right]$$

$$= \exp\left[y \log(\theta_{ij}) - t_{ij}\theta_{ij} + y \log(t_{ij}) - \log(y!)\right]$$

$$= \exp\left[y\eta_{ij} - t_{ij}\exp(\eta_{ij}) + y \log(t_{ij}) - \log(y!)\right],$$

where $\eta_{ij} = \log(\theta_{ij})$ is the natural parameter of the model.

As a transformation of the rate, η_{ij} governs the mean of the distribution and thus captures how the expected count depends on the covariates. Following from the bedrock regression principle that systematic variation in the mean response should be explained linearly in terms of the predictors, we formalize this dependence by defining the *linear predictor* for observation i from process j as $X_{ij}^{\top}\beta_j$, a linear combination of the covariates with a vector of unknown coefficients β_j.

In a GLM, the systematic component represented by $X_{ij}^{\top}\beta_j$ must be connected to its random component μ_{ij}, the conditional mean of the response. This is done via a *link function* $g(\cdot)$, which specifies the relationship

$$g(\mu_{ij}) = X_{ij}^{\top}\beta_j.$$

The *canonical link function* is the special choice of $g(\cdot)$ that sets the natural parameter equal to the linear predictor, i.e.,

$$\eta_{ij} = X_{ij}^{\top}\beta_j. \tag{7.1.10}$$

For the Poisson distribution, we have already shown that the natural parameter is related to the rate through $\eta_{ij} = \log(\theta_{ij})$. Because the conditional mean of each Poisson process includes the exposure time,

$\mu_{ij} = t_{ij}\theta_{ij}$, the model can be equivalently written as

$$
\begin{aligned}
\log(\mu_{ij}) &= \log(t_{ij}\theta_{ij}) \\
&= \log(t_{ij}) + \log(\theta_{ij}) \\
&= \log(t_{ij}) + \eta_{ij} \\
&= \log(t_{ij}) + X_{ij}^{\top}\beta_j, \quad \text{(using the canonical link)}
\end{aligned}
\tag{7.1.11}
$$

where $\log(t_{ij})$ acts as a known offset. It follows that the canonical link for a Poisson GLM is the *log link* $g(\mu_{ij}) = \log(\mu_{ij}/t_{ij})$, under which additive effects of X_{ij} correspond to multiplicative effects on μ_{ij}.

To simplify notation, we include a leading 1 in the process-specific covariate vector $X_{ij}^{(s)}$ that corresponds to the intercept term for process j. We then partition the coefficient vector β_j as

$$
\beta_j = \begin{pmatrix} \beta_j^{(s)} \\ \beta^{(a)} \end{pmatrix} \in \mathbb{R}^{(p+q)\times 1},
\tag{7.1.12}
$$

where $\beta_j^{(s)} \in \mathbb{R}^{p\times 1}$ contains all the coefficients specific to process j (including the intercept) and $\beta^{(a)} \in \mathbb{R}^{q\times 1}$ contains the shared, process-agnostic slopes. The linear predictor for observation i from process j can then be written as

$$
\eta_{ij} = \left(X_{ij}^{(s)}\right)^{\top}\beta_j^{(s)} + \left(X_{ij}^{(a)}\right)^{\top}\beta^{(a)}.
\tag{7.1.13}
$$

Stacking the observations for process j, we can express the model in matrix form as

$$
\eta_j = X_j^{(s)}\beta_j^{(s)} + X_j^{(a)}\beta^{(a)},
\tag{7.1.14}
$$

where $X_j^{(s)} \in \mathbb{R}^{n_j\times p}$ and $X_j^{(a)} \in \mathbb{R}^{n_j\times q}$ are the design matrices of process-specific and process-agnostic covariates, respectively, and $\eta_j = (\eta_{1j},\dots,\eta_{n_jj})^{\top}$ is the vector of linear predictors for process j. If we concatenate the design matrices so that $X_j = [X_j^{(s)}\, X_j^{(a)}] \in \mathbb{R}^{n_j\times(p+q)}$, we can further simplify Equation 7.1.30 as

$$
\eta_j = X_j\beta_j.
\tag{7.1.15}
$$

Collecting all coefficients across all processes gives the global parameter vector

$$\beta = \begin{pmatrix} \beta_1^{(s)} \\ \vdots \\ \beta_J^{(s)} \\ \beta^{(a)} \end{pmatrix} \in \mathbb{R}^{(pJ+q)\times 1},$$

and we view $\beta_j = (\beta_j^{(s)\top}, \beta^{(a)\top})^\top$ as the subvector of β relevant to process j.

The log-likelihood for an individual observation Y_{ij} follows directly from the exponential-family representation of the Poisson distribution:

$$\ell_{ij}(\beta_j; Y_{ij}) = Y_{ij} X_{ij}^\top \beta_j - t_{ij} \exp(X_{ij}^\top \beta_j)), \tag{7.1.16}$$

where we have discarded any additive terms not depending on the parameters. Summing over all observations within process j gives the process-level log-likelihood,

$$\begin{aligned}
\ell_j(\beta_j; Y_j) &= \sum_{i=1}^{n_j} \ell_{ij}(\beta_j; Y_{ij}) \\
&= \sum_{i=1}^{n_j} \left[Y_{ij} X_{ij}^\top \beta_j - t_{ij} \exp(X_{ij}^\top \beta_j)) \right] \\
&= Y_j^\top X_j \beta_j - t_j^\top \exp(X_j \beta_j),
\end{aligned} \tag{7.1.17}$$

where $Y_j \in \mathbb{R}^{n_j \times 1}$ and $t_j \in \mathbb{R}^{n_j \times 1}$ collect the observed counts and exposure times, respectively, for process j, and $\exp(X_j \beta_j)$ is understood to mean the exponential function applied component-wise to the vector $X_j \beta_j$.

Finally, aggregating across all processes yields the full log-likelihood,

$$
\begin{aligned}
\ell(\beta; Y) &= \sum_{j=1}^{J} \ell_j(\beta_j; Y_j) \\
&= \sum_{j=1}^{J} \left[Y_j^\top X_j \beta_j - t_j^\top \exp(X_j \beta_j) \right] \\
&= Y^\top X \beta - t^\top \exp(X\beta),
\end{aligned}
\tag{7.1.18}
$$

where $n = \sum_{j=1}^{J} n_j$, $Y = (Y_1^\top, ..., Y_J^\top)^\top \in \mathbb{R}^{n \times 1}$ is the stacked response vector, $t = (t_1^\top, ..., t_J^\top)^\top \in \mathbb{R}^{n \times 1}$ is the stacked vector of exposure times, and $X = \mathrm{blockdiag}(X_1, ..., X_j) \in \mathbb{R}^{n \times (pJ+q)}$ is the block-diagonal design matrix with process-specific blocks, augmented with shared covariates $X_j^{(a)}$ repeated across processes as necessary.

Let $\hat{\beta}_j$ denote the MLE of the coefficient vector β_j. Because the exponential term in the log-likelihood in Equation 7.1.34 makes it a nonlinear function of each β_j, the score equations for this model cannot be solved algebraically, and thus no closed-form expressions for $\hat{\beta}_j$ exists. Instead, $\hat{\beta}_j$ can be obtained numerically using the `glm()` function from the `stats` package in R, which implements the standard iteratively reweighted least squares (IRLS) algorithm for approximating the MLEs in the GLM framework.

From Equation 7.1.26, the corresponding estimated linear predictors for process j are given by

$$
\hat{\eta}_{ij} = X_{ij}^\top \hat{\beta}_j.
\tag{7.1.19}
$$

Since $\eta_{ij} = \log(\theta_{ij})$, the estimated observation-level rate is

$$
\hat{\theta}_{ij} = \exp(X_{ij}^\top \hat{\beta}_j).
\tag{7.1.20}
$$

An estimate of the marginal rate $\bar{\theta}_j$ defined in Equation 7.1.8 can then be obtained by averaging these observation-level estimates over the index i:

$$
\hat{\theta}_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \exp(X_{ij}^\top \hat{\beta}_j) = \frac{1}{n_j} \mathbf{1}_{n_j}^\top \exp(X_j \hat{\beta}_j).
\tag{7.1.21}
$$

Finally, substituting these into Equation 7.1.9 yields the MLE for ψ:

$$\hat{\psi} = \sum_{j=1}^{J} \alpha_j \hat{\theta}_{\bullet j} = \alpha^\top \hat{\theta}_\bullet, \tag{7.1.22}$$

where

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_J \end{pmatrix} \quad \text{and} \quad \hat{\theta}_\bullet = \begin{pmatrix} \hat{\theta}_{\bullet 1} \\ \vdots \\ \hat{\theta}_{\bullet J} \end{pmatrix}.$$

If we let

$$\gamma = \begin{pmatrix} \frac{\alpha_1}{n_1} \mathbf{1}_{n_1} \\ \vdots \\ \frac{\alpha_J}{n_J} \mathbf{1}_{n_J} \end{pmatrix} \in \mathbb{R}^{n \times 1}, \tag{7.1.23}$$

we can rewrite our expression for $\hat{\psi}$ directly in terms of the MLE for β:

$$\hat{\psi} = \gamma^\top \exp(X\hat{\beta}). \tag{7.1.24}$$

This allows us to write our manifold of parameter vectors that produce the same estimate of ψ with the compact notation

$$\Omega_{\hat{\psi}} = \left\{ \omega \in \mathbb{R}^{(pJ+q) \times 1} : \gamma^\top \exp(X\omega) = \gamma^\top \exp(X\hat{\beta}) \right\}. \tag{7.1.25}$$

To estimate the value of the ZSE-parameterized integrated likelihood at a specific value of ψ, say ψ_1, using the method in Chapter 5, we use the following procedure.

7.1.2.1 Integrated Likelihood Calculation Procedure

1. Draw a random variate $\hat{\omega} \in \Omega_{\hat{\psi}}$ according to a distribution not depending on ψ. The method we have chosen for our simulations is:

    (i) Draw a random vector $u = (u_1, \ldots, u_m)$, where $m = pJ + q$ and each $u_k \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1)$, $k = 1, \ldots, m$.

    (ii) Pass u as the initial guess to an `auglag()` call from the `nloptr` R package that minimizes a dummy function $f(\hat{\omega}) \equiv 0$ subject to the constraint

    $$h(\hat{\omega}) \equiv \gamma^\top \exp(X\hat{\omega}) - \hat{\psi} = 0.$$

2. Solve

    $$\tilde{\beta} = \underset{\beta \in \mathbb{R}^m}{\arg\max} \; \left(t \odot \exp(X\hat{\omega})\right)^\top X\beta - t^\top \exp(X\beta) \quad \text{subject to } \gamma^\top \exp(X\hat{\omega}) = \psi.$$

3. Define the branch evaluated at ψ_1 as

    $$\ell_b^{(1)}(\psi_1) = Y^\top X\tilde{\beta} - t^\top \exp(X\tilde{\beta}).$$

4. Repeat steps (1) - (3) $R - 1$ times to obtain $\ell_b^{(1)}(\psi_1), \ldots, \ell_b^{(R)}(\psi_1)$, then calculate

    $$\hat{\tilde{L}}(\psi_1) = \frac{1}{R} \sum_{r=1}^{R} \exp\left(\ell_b^{(r)}(\psi_1)\right).$$

7.1.2.1 Profile Likelihood Calculation Procedure

1. Solve

    $$\tilde{\beta} = \underset{\beta \in \mathbb{R}^m}{\arg\max} \; \left(t \odot \exp(X\hat{\beta})\right)^\top X\beta - t^\top \exp(X\beta) \quad \text{subject to } \gamma^\top \exp(X\hat{\beta}) = \psi.$$

2. Obtain the profile likelihood evaluated at ψ_1 by calculating

    $$L_p(\psi_1) = \exp\left(Y^\top X\tilde{\beta} - t^\top \exp(X\tilde{\beta})\right).$$

7.1.2.2. The Negative Binomial GLM

Suppose the conditional distribution of Y_{ij} given X_{ij} is $Y_{ij}|X_{ij} \sim \mathrm{NB}(\mu_{ij}, r_j)$. We may continue to interpret μ_{ij} as the expected value of Y_{ij} conditional on X_{ij}, while r_j is best interpreted as a dispersion parameter that measures the excess variation across counts within process j that a Poisson random variable would fail to capture. Then the conditional probability mass function of Y_{ij} can be expressed in canonical exponential-family form as

$$\mathrm{P}(Y_{ij} = y|X_{ij}) = \frac{\Gamma(y + r_j)}{\Gamma(r_j)y!} \left(\frac{r_j}{r_j + \mu_{ij}} \right)^{r_j} \left(\frac{\mu_{ij}}{r_j + \mu_{ij}} \right)^{y}, \ \ y = 0, 1, 2, ...$$

where $\eta_{ij} = \log(\theta_{ij})$ is the natural parameter of the model.

As a transformation of the rate, η_{ij} governs the mean of the distribution and thus captures how the expected count depends on the covariates. Following from the bedrock regression principle that systematic variation in the mean response should be explained linearly in terms of the predictors, we formalize this dependence by defining the *linear predictor* for observation i from process j as $X_{ij}^{\top}\beta_j$, a linear combination of the covariates with a vector of unknown coefficients β_j.

In a GLM, the systematic component represented by $X_{ij}^{\top}\beta_j$ must be connected to its random component μ_{ij}, the conditional mean of the response. This is done via a *link function* $g(\cdot)$, which specifies the relationship

$$g(\mu_{ij}) = X_{ij}^{\top}\beta_j.$$

The *canonical link function* is the special choice of $g(\cdot)$ that sets the natural parameter equal to the linear predictor, i.e.,

$$\eta_{ij} = X_{ij}^{\top}\beta_j. \tag{7.1.26}$$

For the Poisson distribution, we have already shown that the natural parameter is related to the rate through $\eta_{ij} = \log(\theta_{ij})$. Because the conditional mean of each Poisson process includes the exposure time,

$\mu_{ij} = t_{ij}\theta_{ij}$, the model can be equivalently written as

$$
\begin{aligned}
\log(\mu_{ij}) &= \log(t_{ij}\theta_{ij}) \\
&= \log(t_{ij}) + \log(\theta_{ij}) \\
&= \log(t_{ij}) + \eta_{ij} \\
&= \log(t_{ij}) + X_{ij}^\top \beta_j, \quad \text{(using the canonical link)}
\end{aligned}
\tag{7.1.27}
$$

where $\log(t_{ij})$ acts as a known offset. It follows that the canonical link for a Poisson GLM is the *log link* $g(\mu_{ij}) = \log(\mu_{ij}/t_{ij})$, under which additive effects of X_{ij} correspond to multiplicative effects on μ_{ij}.

To simplify notation, we include a leading 1 in the process-specific covariate vector $X_{ij}^{(s)}$ that corresponds to the intercept term for process j. We then partition the coefficient vector β_j as

$$
\beta_j = \begin{pmatrix} \beta_j^{(s)} \\ \beta^{(a)} \end{pmatrix} \in \mathbb{R}^{(p+q)\times 1},
\tag{7.1.28}
$$

where $\beta_j^{(s)} \in \mathbb{R}^{p\times 1}$ contains all the coefficients specific to process j (including the intercept) and $\beta^{(a)} \in \mathbb{R}^{q\times 1}$ contains the shared, process-agnostic slopes. The linear predictor for observation i from process j can then be written as

$$
\eta_{ij} = \left(X_{ij}^{(s)}\right)^\top \beta_j^{(s)} + \left(X_{ij}^{(a)}\right)^\top \beta^{(a)}.
\tag{7.1.29}
$$

Stacking the observations for process j, we can express the model in matrix form as

$$
\eta_j = X_j^{(s)}\beta_j^{(s)} + X_j^{(a)}\beta^{(a)},
\tag{7.1.30}
$$

where $X_j^{(s)} \in \mathbb{R}^{n_j \times p}$ and $X_j^{(a)} \in \mathbb{R}^{n_j \times q}$ are the design matrices of process-specific and process-agnostic covariates, respectively, and $\eta_j = (\eta_{1j}, \dots, \eta_{n_j j})^\top$ is the vector of linear predictors for process j. If we concatenate the design matrices so that $X_j = [X_j^{(s)}\, X_j^{(a)}] \in \mathbb{R}^{n_j \times (p+q)}$, we can further simplify Equation 7.1.30 as

$$
\eta_j = X_j \beta_j.
\tag{7.1.31}
$$

Collecting all coefficients across all processes gives the global parameter vector

$$
\beta = \begin{pmatrix} \beta_1^{(s)} \\ \vdots \\ \beta_J^{(s)} \\ \beta^{(a)} \end{pmatrix} \in \mathbb{R}^{(pJ+q)\times 1},
$$

and we view $\beta_j = (\beta_j^{(s)\top}, \beta^{(a)\top})^\top$ as the subvector of β relevant to process j.

The log-likelihood for an individual observation Y_{ij} follows directly from the exponential-family representation of the Poisson distribution:

$$
\ell_{ij}(\beta_j; Y_{ij}) = Y_{ij}X_{ij}^\top\beta_j - t_{ij}\exp(X_{ij}^\top\beta_j)), \tag{7.1.32}
$$

where we have discarded any additive terms not depending on the parameters. Summing over all observations within process j gives the process-level log-likelihood,

$$
\begin{aligned}
\ell_j(\beta_j; Y_j) &= \sum_{i=1}^{n_j} \ell_{ij}(\beta_j; Y_{ij}) \\
&= \sum_{i=1}^{n_j} \left[Y_{ij}X_{ij}^\top\beta_j - t_{ij}\exp(X_{ij}^\top\beta_j)) \right] \\
&= Y_j^\top X_j\beta_j - t_j^\top\exp(X_j\beta_j),
\end{aligned} \tag{7.1.33}
$$

where $Y_j \in \mathbb{R}^{n_j\times 1}$ and $t_j \in \mathbb{R}^{n_j\times 1}$ collect the observed counts and exposure times, respectively, for process j, and $\exp(X_j\beta_j)$ is understood to mean the exponential function applied component-wise to the vector $X_j\beta_j$.

Finally, aggregating across all processes yields the full log-likelihood,

$$
\begin{aligned}
\ell(\beta; Y) &= \sum_{j=1}^{J} \ell_j(\beta_j; Y_j) \\
&= \sum_{j=1}^{J} \left[Y_j^\top X_j \beta_j - t_j^\top \exp(X_j \beta_j) \right] \\
&= Y^\top X \beta - t^\top \exp(X\beta),
\end{aligned}
\tag{7.1.34}
$$

where $n = \sum_{j=1}^{J} n_j$, $Y = (Y_1^\top, ..., Y_J^\top)^\top \in \mathbb{R}^{n \times 1}$ is the stacked response vector, $t = (t_1^\top, ..., t_J^\top)^\top \in \mathbb{R}^{n \times 1}$ is the stacked vector of exposure times, and $X = \mathrm{blockdiag}(X_1, ..., X_j) \in \mathbb{R}^{n \times (pJ+q)}$ is the block-diagonal design matrix with process-specific blocks, augmented with shared covariates $X_j^{(a)}$ repeated across processes as necessary.

Let $\hat{\beta}_j$ denote the MLE of the coefficient vector β_j. Because the exponential term in the log-likelihood in Equation 7.1.34 makes it a nonlinear function of each β_j, the score equations for this model cannot be solved algebraically, and thus no closed-form expressions for $\hat{\beta}_j$ exists. Instead, $\hat{\beta}_j$ can be obtained numerically using the `glm()` function from the `stats` package in R, which implements the standard iteratively reweighted least squares (IRLS) algorithm for approximating the MLEs in the GLM framework.

From Equation 7.1.26, the corresponding estimated linear predictors for process j are given by

$$
\hat{\eta}_{ij} = X_{ij}^\top \hat{\beta}_j.
\tag{7.1.35}
$$

Since $\eta_{ij} = \log(\theta_{ij})$, the estimated observation-level rate is

$$
\hat{\theta}_{ij} = \exp(X_{ij}^\top \hat{\beta}_j).
\tag{7.1.36}
$$

An estimate of the marginal rate $\bar{\theta}_j$ defined in Equation 7.1.8 can then be obtained by averaging these observation-level estimates over the index i:

$$
\hat{\theta}_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \exp(X_{ij}^\top \hat{\beta}_j) = \frac{1}{n_j} \mathbf{1}_{n_j}^\top \exp(X_j \hat{\beta}_j).
\tag{7.1.37}
$$

Finally, substituting these into Equation 7.1.9 yields the MLE for ψ:

$$\hat{\psi} = \sum_{j=1}^{J} \alpha_j \hat{\theta}_{\bullet j} = \alpha^\top \hat{\theta}_\bullet, \tag{7.1.38}$$

where

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_J \end{pmatrix} \quad \text{and} \quad \hat{\theta}_\bullet = \begin{pmatrix} \hat{\theta}_{\bullet 1} \\ \vdots \\ \hat{\theta}_{\bullet J} \end{pmatrix}.$$

If we let

$$\gamma = \begin{pmatrix} \frac{\alpha_1}{n_1} \mathbf{1}_{n_1} \\ \vdots \\ \frac{\alpha_J}{n_J} \mathbf{1}_{n_J} \end{pmatrix} \in \mathbb{R}^{n \times 1}, \tag{7.1.39}$$

we can rewrite our expression for $\hat{\psi}$ directly in terms of the MLE for β:

$$\hat{\psi} = \gamma^\top \exp(X\hat{\beta}). \tag{7.1.40}$$

This allows us to write our manifold of parameter vectors that produce the same estimate of ψ with the compact notation

$$\Omega_{\hat{\psi}} = \left\{ \omega \in \mathbb{R}^{(pJ+q) \times 1} : \gamma^\top \exp(X\omega) = \gamma^\top \exp(X\hat{\beta}) \right\}. \tag{7.1.41}$$

To estimate the value of the ZSE-parameterized integrated likelihood at a specific value of ψ, say ψ_1, using the method in Chapter 5, we use the following procedure.

7.1.2.1 Integrated Likelihood Calculation Procedure

1. Draw a random variate $\hat{\omega} \in \Omega_{\hat{\psi}}$ according to a distribution not depending on ψ. The method we have chosen for our simulations is:

   (i) Draw a random vector $u = (u_1, \ldots, u_m)$, where $m = pJ + q$ and each $u_k \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1)$, $k = 1, \ldots, m$.

   (ii) Pass u as the initial guess to an `auglag()` call from the `nloptr` R package that minimizes a dummy function $f(\hat{\omega}) \equiv 0$ subject to the constraint

$$h(\hat{\omega}) \equiv \gamma^\top \exp(X\hat{\omega}) - \hat{\psi} = 0.$$

2. Solve

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^m}{\arg\max} \; \left(t \odot \exp(X\hat{\omega})\right)^\top X\beta - t^\top \exp(X\beta) \quad \text{subject to } \gamma^\top \exp(X\hat{\omega}) = \psi.$$

3. Define the branch evaluated at ψ_1 as

$$\ell_b^{(1)}(\psi_1) = Y^\top X\tilde{\beta} - t^\top \exp(X\tilde{\beta}).$$

4. Repeat steps (1) - (3) $R - 1$ times to obtain $\ell_b^{(1)}(\psi_1), \ldots, \ell_b^{(R)}(\psi_1)$, then calculate

$$\hat{\tilde{L}}(\psi_1) = \frac{1}{R} \sum_{r=1}^{R} \exp\left(\ell_b^{(r)}(\psi_1)\right).$$

7.1.2.1 Profile Likelihood Calculation Procedure

1. Solve

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^m}{\arg\max} \; \left(t \odot \exp(X\hat{\beta})\right)^\top X\beta - t^\top \exp(X\beta) \quad \text{subject to } \gamma^\top \exp(X\hat{\beta}) = \psi.$$

2. Obtain the profile likelihood evaluated at ψ_1 by calculating

$$L_p(\psi_1) = \exp\left(Y^\top X\tilde{\beta} - t^\top \exp(X\tilde{\beta})\right).$$

7.1.3. Mixed Effects Regression

7.1.3.1. Fixed Intercepts and Random Slopes

7.1.3.2. Random Intercepts and Fixed Slopes

7.1.3.3. Random Intercepts and Random Slopes

7.2. Overdispersion

7.3. Zero-Inflation

CHAPTER 8

Discussion

References

Berger, J. O., Liseo, B., & Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, *14*(1), 1–22. http://www.jstor.org/stable/2676641

De Bin, R., Sartori, N., & Severini, T. A. (2015). Integrated likelihoods in models with stratum nuisance parameters. *Electronic Journal of Statistics*, *9*(1), 1474–1491. https://doi.org/10.1214/15-EJS1045

Kalbfleisch, J. D., & Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, *32*(2), 175–208. http://www.jstor.org/stable/2984524

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, *90*(3), 533–549. http://www.jstor.org/stable/30042064

Severini, T. A. (2007). Integrated likelihood functions for non-bayesian inference. *Biometrika*, *94*(3), 529–542. http://www.jstor.org/stable/20441394

Severini, T. A. (2018). Integrated likelihoods for functions of a parameter. *Stat*, *7*(1), e212. https://doi.org/10.1002/sta4.212

Severini, T. A. (2022). Integrated likelihood inference in multinomial distributions. *Metron*. https://doi.org/10.1007/s40300-022-00236-x

Tierney, L., & Kadane, J. B. (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 7

Proposition: $\psi_\omega^* = \hat{\psi}$ for all $\omega \in \Omega_{\hat{\psi}}$.

Proof: It suffices to show that an arbitrary branch curve $b_\omega(\psi)$ has a stationary point at $\psi = \hat{\psi}$, i.e. $b_\omega'(\hat{\psi}) = 0$. Since $\ell(\theta; Y)$ is concave in θ and the mapping $\psi \mapsto \tilde{\theta}(\psi; \omega)$ is smooth, such a point will be the unique maximizer of $b_\omega(\psi)$. Differentiating the branch yields

$$b_\omega'(\psi) = \nabla_\theta \ell(\tilde{\theta}(\psi; \omega); Y)^\top \frac{\partial \tilde{\theta}(\psi; \omega)}{\partial \psi}. \tag{A.0.1}$$

Thus, we need to be able to evaluate both the ZSE optimizer $\tilde{\theta}(\psi; \omega)$ and its partial derivative with respect to ψ at $\psi = \hat{\psi}$. Recall that to evaluate $\tilde{\theta}(\psi; \omega)$ at a point ψ for a given ω is to solve the constrained maximization problem

$$\arg\max_\theta \sum_{g=1}^G \left(t_g \omega_g \log \theta_g - t_g \theta_g \right) \text{ s.t. } \alpha^\top \theta = \psi.$$

We proceed by the method of Lagrange multipliers. For some multiplier λ (which may depend on ψ and ω), define the Lagrangian function

$$\mathcal{L}(\theta, \lambda) = \sum_{g=1}^G \left(t_g \omega_g \log \theta_g - t_g \theta_g \right) + \lambda \left[\psi - \sum_{g=1}^G \alpha_g \theta_g \right].$$

Per the Lagrange multiplier theorem, the solution to the maximization problem will satisfy

$$\left. \frac{\partial \mathcal{L}(\theta, \lambda)}{\theta_g} \right|_{\theta_g = \tilde{\theta}_g} = 0, \text{ for } g = 1, ..., G \tag{A.0.2}$$

and

$$\frac{\partial \mathcal{L}(\theta, \lambda)}{\lambda} = 0. \tag{A.0.3}$$

From Equation A.0.2, we have

$$\frac{t_g \omega_g}{\tilde{\theta}_g} - t_g - \lambda \alpha_g = 0.$$

This implies that the optimal solution $\tilde{\theta}$, viewed here as a function of λ, satisfies

$$\tilde{\theta}_g(\lambda) = \frac{t_g \omega_g}{t_g + \lambda \alpha_g}. \tag{A.0.4}$$

Paired with the constraint condition enforced by Equation A.0.3, this in turn implies that the optimal value of λ will satisfy

$$\sum_{g=1}^{G} \alpha_g \tilde{\theta}_g(\lambda) = \psi. \tag{A.0.5}$$

Note that $\lambda = 0 \iff \tilde{\theta} = \omega \iff \psi = \hat{\psi}$. It follows that $\tilde{\theta}(\hat{\psi}; \omega) = \omega$.

Turning our attention now to the derivative of the ZSE optimizer with respect to ψ, we have that

$$\frac{\partial \tilde{\theta}_g}{\partial \psi} = \frac{\partial \tilde{\theta}_g}{\partial \lambda} \frac{d\lambda}{d\psi}.$$

From Equation A.0.4, we can compute

$$\frac{\partial \tilde{\theta}_g}{\partial \lambda} = -\frac{t_g \omega_g \alpha_g}{(t_g + \lambda \alpha_g)^2}.$$

At $\lambda = 0$, this evaluates to

$$\frac{\partial \tilde{\theta}_g}{\partial \lambda}\bigg|_{\lambda=0} = -\frac{\omega_g \alpha_g}{t_g}.$$

Similarly, differentiating both sides of Equation A.0.5 with respect to ψ gives

$$\sum_{g=1}^{G} \alpha_g \frac{\partial \tilde{\theta}_g}{\partial \lambda} \frac{d\lambda}{d\psi} = 1 \implies \frac{d\lambda}{d\psi} = \frac{1}{\sum_{g=1}^{G} \alpha_g \frac{\partial \tilde{\theta}_g}{\partial \lambda}}.$$

At $\psi = \hat{\psi}$, this evaluates to

$$\frac{d\lambda}{d\psi}\bigg|_{\psi=\hat{\psi}} = \frac{1}{\sum_{g=1}^{G} \alpha_g \frac{\partial \tilde{\theta}_g}{\partial \lambda}\big|_{\lambda=0}} = -\frac{1}{\sum_{g=1}^{G} \alpha_g^2 \omega_g / t_g}.$$

Thus

$$\left.\frac{\partial\tilde{\theta}_g}{\partial\psi}\right|_{\psi=\hat{\psi}} = \left.\frac{\partial\tilde{\theta}_g}{\partial\lambda}\right|_{\lambda=0} \cdot \left.\frac{d\lambda}{d\psi}\right|_{\psi=\hat{\psi}}$$

$$= \left(-\frac{\omega_g\alpha_g}{t_g}\right)\left(-\frac{1}{\sum_{h=1}^{H}\alpha_h^2\omega_h/t_h}\right)$$

$$= \frac{\alpha_g\omega_g/t_g}{\sum_{h=1}^{H}\alpha_h^2\omega_h/t_h},$$

where we have changed the index of summation from g to h in the denominator of the final two lines to avoid confusion with the particular index g being considered in the numerator.

Returning to Equation A.0.1, we can now compute

$$b'_\omega(\hat{\psi}) = \left[\left.\nabla_\theta\ell(\tilde{\theta}(\psi;\omega);Y)\right|_{\psi=\hat{\psi}}\right]^\top \left[\left.\frac{\partial\tilde{\theta}(\psi;\omega)}{\partial\psi}\right|_{\psi=\hat{\psi}}\right]$$

$$= \sum_{g=1}^{G}\left(\left.\frac{\partial\ell}{\partial\theta_g}\right|_{\theta_g=\tilde{\theta}_g(\hat{\psi};\omega)}\right)\left(\left.\frac{\partial\tilde{\theta}_g}{\partial\psi}\right|_{\psi=\hat{\psi}}\right)$$

$$= \sum_{g=1}^{G}\left(\frac{Y_g}{\tilde{\theta}_g(\hat{\psi};\omega)} - t_g\right)\left(\frac{\alpha_g\omega_g/t_g}{\sum_{h=1}^{H}\alpha_h^2\omega_h/t_h}\right)$$

$$= \sum_{g=1}^{G}\left(\frac{Y_g}{\omega_g} - t_g\right)\left(\frac{\alpha_g\omega_g/t_g}{\sum_{h=1}^{H}\alpha_h^2\omega_h/t_h}\right)$$

$$= \frac{1}{\sum_{h=1}^{H}\alpha_h^2\omega_h/t_h}\sum_{g=1}^{G}\alpha_g\left(\frac{Y_g}{t_g} - \omega_g\right)$$

$$= \frac{1}{\sum_{h=1}^{H}\alpha_h^2\omega_h/t_h}\left(\sum_{g=1}^{G}\alpha_g\hat{\theta}_g - \sum_{g=1}^{G}\alpha_g\omega_g\right)$$

$$= \frac{1}{\sum_{h=1}^{H}\alpha_h^2\omega_h/t_h}(\hat{\psi} - \hat{\psi})$$

$$= 0.$$

$\therefore b'_\omega(\hat{\psi}) = 0.$

Since the choice of branch was arbitrary, this holds true for all $\omega \in \Omega_{\hat{\psi}}$.

$\therefore \psi^*_\omega = \hat{\psi}$ for all $\omega \in \Omega_{\hat{\psi}}$. $\blacksquare$