

Integrated Likelihood Inference in Poisson Distributions

Timothy Ruel

Department of Statistics and Data Science, Northwestern University

September 15, 2024

Abstract

The text of your abstract. 200 or fewer words.

Keywords: Directly standardized rate, Integrated likelihood ratio statistic, Maximum integrated likelihood estimator, Profile likelihood, Weighted sum, Zero score expectation parameter

1 Introduction

Consider a vector $\theta = (\theta_1, \dots, \theta_n)$ in which each component represents the mean of a distinct Poisson process. The purpose of this paper is to discuss the task of conducting likelihood-based inference for a real-valued parameter of interest $\psi = \tau(\theta)$. In particular, we will examine the utility of the integrated likelihood function as a tool for obtaining interval and point estimates for ψ , using the performance of the more easily calculated profile likelihood as a benchmark.

We may obtain a sample of values from each Poisson process through repeated measurements of the number of events it generates over a fixed period of time. Suppose we have done so, and let X_{ij} represent the j th count from the i th sample, so that $X_{ij} \sim \text{Poisson}(\theta_i)$ for $i = 1, \dots, n$ and $j = 1, \dots, m_i$. The probability mass function (pmf) for a single observation $X_{ij} = x_{ij}$ is

$$p(x_{ij}; \theta_i) = \frac{e^{-\theta_i} \theta_i^{x_{ij}}}{x_{ij}!}, \quad x_{ij} = 0, 1, 2, \dots; \quad \theta_i > 0. \quad (1)$$

Denote the sample of counts from the i th process by the vector $X_{i\bullet} = (X_{i1}, \dots, X_{im_i})$, its associated mean by $\bar{X}_{i\bullet} = \frac{1}{m_i} \sum_{j=1}^{m_i} X_{ij}$, and assume that all of the counts both within and between samples are measured independently. The likelihood function for an individual component θ_i based on the data $X_{i\bullet} = x_{i\bullet}$ is then equal to the product of the individual

probabilities of the observed counts, i.e.

$$\begin{aligned}
L(\theta_i; x_{i\bullet}) &= \prod_{j=1}^{m_i} p(x_{ij}; \theta_i) \\
&= \prod_{j=1}^{m_i} \frac{e^{-\theta_i} \theta_i^{x_{ij}}}{x_{ij}!} \\
&= \left(\prod_{j=1}^{m_i} e^{-\theta_i} \right) \left(\prod_{j=1}^{m_i} \theta_i^{x_{ij}} \right) \left(\prod_{j=1}^{m_i} x_{ij}! \right)^{-1} \\
&= \left(e^{-\sum_{j=1}^{m_i} \theta_i} \right) \left(\theta_i^{\sum_{j=1}^{m_i} x_{ij}} \right) \left(\prod_{j=1}^{m_i} x_{ij}! \right)^{-1} \\
&= e^{-m_i \theta_i} \theta_i^{m_i \bar{x}_{i\bullet}} \left(\prod_{j=1}^{m_i} x_{ij}! \right)^{-1}.
\end{aligned} \tag{2}$$

Since L is only useful to the extent that it informs our understanding of the value of θ_i , we are free to replace it with any other function differing from it by just a (nonzero) multiplicative term that is constant with respect to θ_i , provided that the result still satisfies the necessary regularity conditions, as this will not change any conclusions regarding θ_i that we draw from it. Hence, we may safely discard the term in parentheses on the final line of Equation 2 as it does not depend on θ_i and instead simply write

$$L(\theta_i; x_{i\bullet}) = e^{-m_i \theta_i} \theta_i^{m_i \bar{x}_{i\bullet}}. \tag{3}$$

It will generally be more convenient to work with the log-likelihood function, which is given by

$$\begin{aligned}
\ell(\theta_i; x_{i\bullet}) &= \log L(\theta_i; x_{i\bullet}) \\
&= \log \left(e^{-m_i \theta_i} \theta_i^{m_i \bar{x}_{i\bullet}} \right) \\
&= -m_i \theta_i + m_i \bar{x}_{i\bullet} \log \theta_i \\
&= m_i (\bar{x}_{i\bullet} \log \theta_i - \theta_i).
\end{aligned} \tag{4}$$

The sum of the log-likelihood functions for each component of θ then forms the basis of

the log-likelihood function for θ itself:

$$\begin{aligned}
\ell(\theta; x_{1\bullet}, \dots, x_{n\bullet}) &= \log L(\theta; x_{1\bullet}, \dots, x_{n\bullet}) \\
&= \log \left(\prod_{i=1}^n L(\theta_i; x_{i\bullet}) \right) \\
&= \sum_{i=1}^n \log L(\theta_i; x_{i\bullet}) \\
&= \sum_{i=1}^n \ell(\theta_i; x_{i\bullet}) \\
&= \sum_{i=1}^n m_i(\bar{x}_{i\bullet} \log \theta_i - \theta_i).
\end{aligned} \tag{5}$$

We can derive the maximum likelihood estimate (MLE) for θ_i by differentiating Equation 4 with respect to θ_i , setting the result equal to 0, and solving for θ_i . This gives the nice result that the MLE is simply equal to the mean of the sample of data $X_{i\bullet}$. That is,

$$\hat{\theta}_i = \bar{X}_{i\bullet}. \tag{6}$$

Similarly, the MLE for the full parameter θ is just the vector of MLEs for its individual components:

$$\hat{\theta} \equiv (\hat{\theta}_1, \dots, \hat{\theta}_n) = (\bar{X}_{1\bullet}, \dots, \bar{X}_{n\bullet}). \tag{7}$$

2 Pseudolikelihoods

Let $\Theta \subseteq \mathbb{R}_+^n$ represent the space of possible values for θ and suppose we have a real-valued *parameter of interest* $\psi = \tau(\theta)$, where $\tau : \Theta \rightarrow \Psi$ is a known function with at least two continuous derivatives. Though it is not strictly necessary, in order to align with the tendency of researchers to focus on one-dimensional summaries of vector quantities we will assume for our purposes that ψ is a scalar, i.e. $\Psi \subseteq \mathbb{R}$.

This reduced dimension of Ψ relative to Θ implies the existence of a *nuisance parameter* $\lambda \in \Lambda \subseteq \mathbb{R}^{n-1}$. As its name suggests, λ tends to obfuscate or outright preclude inference

regarding ψ and typically must be eliminated from the likelihood before proceeding. The product of this elimination is called a *pseudolikelihood function*. Any function of the data and ψ alone could theoretically be considered a pseudolikelihood, though course in practice some are more useful than others.

If we let $\Theta_\psi = \{\theta \in \Theta : \tau(\theta) = \psi\}$, then associated with each $\psi \in \Psi$ is the set of likelihood values $\mathcal{L}_\psi = \{L(\theta) : \theta \in \Theta_\psi\}$. For a given value of ψ , there may exist multiple corresponding values of λ .

We can construct pseudolikelihoods for ψ through clever choices by which to summarize \mathcal{L}_ψ over all possible values of λ . Among the most popular methods of summary are profiling (i.e. maximization), conditioning, and integration, each with respect to the nuisance parameter. These summaries do come at a cost, however; eliminating a model's nuisance parameter from its likelihood almost always sacrifices some information about its parameter of interest as well. One measure of a good pseudolikelihood, therefore, is the balance it strikes between the amount of information it retains about ψ and the ease with which it can be computed.

2.1 The Profile Likelihood

The most straightforward method we can use to construct a pseudolikelihood (or equivalently, a pseudo-log-likelihood) function for ψ is usually to find the maximum of $\ell(\theta)$ over all possible values of θ for each value of ψ . This yields what is known as the *profile* log-likelihood function, formally defined as

$$\ell_p(\psi) = \sup_{\theta \in \Theta : \tau(\theta) = \psi} \ell(\theta), \quad \psi \in \Psi. \quad (8)$$

In the case where an explicit nuisance parameter λ exists so that θ may be written as $\theta = (\psi, \lambda)$, Equation 8 is equivalent to replacing λ with $\hat{\lambda}_\psi$, its conditional MLE given ψ :

$$\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi). \quad (9)$$

Historically, the efficiency with which the profile is capable of producing accurate estimates of ψ relative to its ease of computation has made it the method of choice for statisticians when performing likelihood-based inference regarding a parameter of interest. Examples of profile-based statistics are the MLE for ψ , i.e.,

$$\hat{\psi} = \arg \sup_{\psi \in \Psi} \ell_p(\psi), \quad (10)$$

and the signed likelihood ratio statistic for ψ , given by

$$R_\psi = \text{sgn}(\hat{\psi} - \psi)(2(\ell_p(\hat{\psi}) - \ell_p(\psi)))^{\frac{1}{2}}. \quad (11)$$

2.2 The Integrated Likelihood

The *integrated likelihood* for ψ seeks to summarize \mathcal{L}_ψ by its average value with respect to some weight function π over the space Θ_ψ . From a theoretical standpoint, this is preferable to the maximization procedure found in the profile likelihood as it naturally incorporates our uncertainty regarding the nuisance parameter's true value into the resulting pseudo-likelihood. The general form of an integrated likelihood function is given

$$\bar{L}(\psi) = \int_{\Theta_\psi} L(\theta) \pi(\theta; \psi) d\theta. \quad (12)$$

It is up to the researcher to choose the weight function $\pi(\cdot; \psi)$, which plays an important role in the properties of the resulting integrated likelihood. Severini (2007) developed a method for re-parameterizing λ 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.

3 Application to Poisson Models

We now turn our attention to the task of using the ZSE parameterization to construct an integrated likelihood that can be used to make inferences regarding a parameter of interest derived from the Poisson model described in the introduction. We will

4 Estimating the Weighted Sum of Poisson Means

Consider the weighted sum

$$Y = \sum_{i=1}^n w_i X_i,$$

where each w_i is a known constant greater than zero. Suppose we take for our parameter of interest the expected value of this weighted sum, so that

$$\psi \equiv E(Y) = \sum_{i=1}^n w_i \theta_i.$$

4.1 Examples

5 Zero-Inflated Poisson Regression

A sample of count data is called *zero-inflated* when it contains an excess amount of zero-valued observations. A common tactic to account for this excess is to model the data using a mixture of two processes, one that generates zeros and another that generates counts, some of which may also be zeros. When this count-generating process follows a Poisson distribution, we call the resulting mixture a zero-inflated Poisson (ZIP) model.

Let $U \sim \text{Bernoulli}(1 - \pi)$ and $V \sim \text{Poisson}(\mu)$. Suppose U and V are independent and let $W = UV$. Then $W \sim \text{ZIP}(\mu, \pi)$. Note that $W = 0$ when either $U = 0$ or $V = 0$ so that

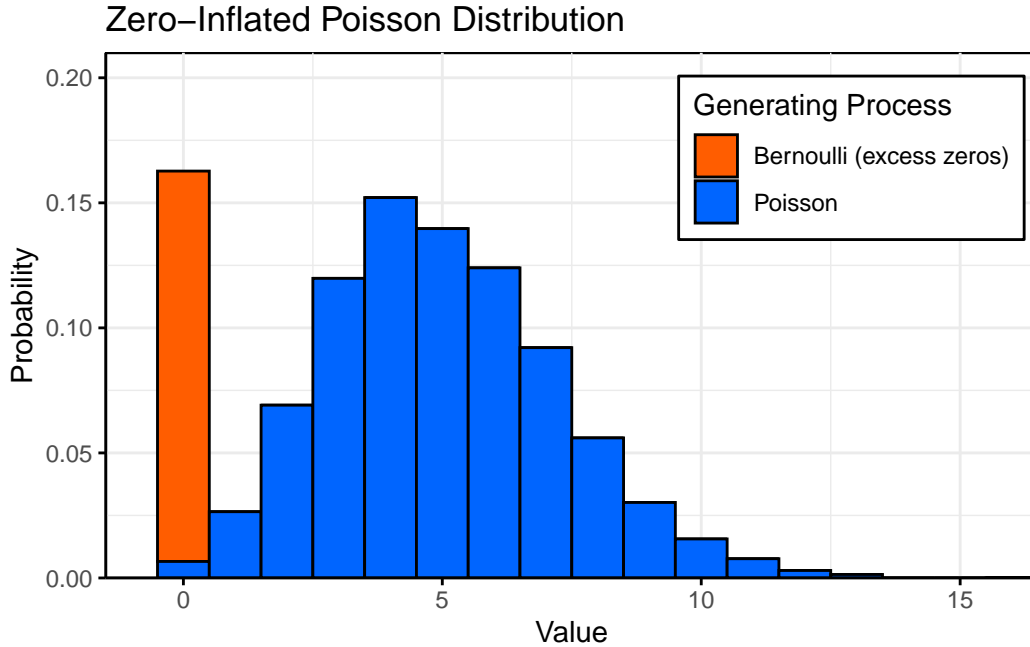
$$\begin{aligned}
 \mathbb{P}(W = 0) &= \mathbb{P}(U = 0 \cup V = 0) \\
 &= \mathbb{P}(U = 0) + \mathbb{P}(V = 0) - \mathbb{P}(U = 0 \cap V = 0) \\
 &= \mathbb{P}(U = 0) + \mathbb{P}(V = 0) - \mathbb{P}(U = 0)\mathbb{P}(V = 0) \\
 &= \pi + e^{-\mu} - \pi e^{-\mu} \\
 &= \pi + (1 - \pi)e^{-\mu}.
 \end{aligned}$$

In order for W to take on a value $w > 0$, we must have $U = 1$ and $V = w$. That is,

$$\begin{aligned}
 \mathbb{P}(W = w) &= \mathbb{P}(U = 1 \cap V = w) \\
 &= \mathbb{P}(U = 1)\mathbb{P}(V = w) \\
 &= (1 - \pi) \frac{e^{-\mu} \mu^w}{w!}, \quad w = 1, 2, \dots
 \end{aligned}$$

Thus, the full probability mass function for a ZIP random variable is given by

$$\mathbb{P}(W = w) = \begin{cases} \pi + (1 - \pi)e^{-\mu}, & w = 0 \\ (1 - \pi) \frac{e^{-\mu} \mu^w}{w!}, & w = 1, 2, \dots \end{cases}$$



References