APPROXIMATE MARGINAL INFERENCE IN MODELS WITH STRATUM NUISANCE PARAMETERS, WITH APPLICATIONS TO FISHERY DATA

JARED TOBIN

Approximate marginal inference in models with stratum nuisance parameters, with applications to fishery data

A practicum submitted to the
School of Graduate Studies
in partial fulfilment of the
requirements for the degree of
Master of Applied Statistics

Department of Mathematics and Statistics
Memorial University of Newfoundland

December 2010

St. John's Canada

Abstract

The profile likelihood is commonly used in cases where the maximum likelihood estimator for a shape or dispersion parameter depends on knowledge of the mean. We demonstrate that, in stratified models with many mean parameters, the maximum profile likelihood estimator for a common shape parameter can be severely biased or even inconsistent when the sample size per stratum is low. We note a 'marginal' likelihood function that eliminates these problematic mean parameters, but is usually intractable or even impossible to calculate in practice. We discuss approximations to this marginal likelihood - notably the modified profile likelihood of Barndorff-Nielsen [5], the adjusted profile likelihood of Cox & Reid [16], and quasi-likelihood variants - and demonstrate that estimators based on these functions have better bias properties than those based on the full likelihood. We apply these estimators to a stratified negative binomial model and achieve accurate estimates for the negative binomial dispersion parameter k in a simulation experiment. Finally, we provide an application of our methods to fishery data.

Acknowledgements

To thank my supervisor Dr. Noel Cadigan properly, I would probably need to devote more text to the acknowledgements section than I did to Chapter 3. For the sake of brevity, let's leave it that I am utterly indebted to him for the exemplary guidance, assistance, and financial support he has provided me during my graduate studies. He continues to this day to offer me opportunities that drive my professional & academic career forward. My most sincere thanks.

I acknowledge the financial support of the Department of Mathematics and Statistics, School of Graduate Studies, and Drs. Noel Cadigan & Gary Sneddon in the form of Graduate Assistantships, and Fisheries & Oceans Canada for providing me a student research assistant position for much of my program. I would like to thank Ms. Lynn Bryant, my higher-up at Transportation & Works, Government of Newfoundland & Labrador for her patience (and leave signatures, and journey authorization lobbies..) while I finished this practicum. A special thanks to my family and friends, who I assume still remember for-the-most-part what I look like.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The nuisance parameter problem

Maximum likelihood methods are widely used for parameter estimation in statistical models. The *likelihood function* treats the joint probability of a sample as a function of its parameters, and the maximum likelihood estimates of these parameters are the values that maximize the likelihood function.

Consider a general model for a response Y with H distinct strata, so that Y_h - the (random) response in the h^{th} stratum, $h = 1, \ldots, H$ - is assumed to follow a probability distribution with mean parameter ψ_h and scalar shape parameter θ, so that the shape parameter θ is common to all strata. This is a *strata mean model* (SMM), a simple class of models that is common and applicable to problems in science, engineering, medicine, and social research. As we will demonstrate, maximum likelihood estimators may be inadequate for estimating θ in an SMM when small sample sizes are used.

The nature of the problem is illustrated by the well-known example of estimating σ^2 from a normally-distributed sample; the normal density is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

and for a sample $\mathbf{x}$ of size n the likelihood function is

$$\begin{aligned} L(\mu, \sigma^2|\mathbf{x}) &= \prod_{i=1}^{n} f_X(x) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}. \end{aligned}$$

Typically the log-likelihood $l = \log L$ is more convenient to work with algebraically. The *score function* with respect to σ is[1]

$$\frac{\partial l(\mu, \sigma^2)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_i (x_i - \mu)^2$$

and the maximum likelihood estimator (MLE) $\hat{\sigma}^2$ satisfies $\partial l / \partial \sigma = 0$. Solving for the MLE yields

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

and the second derivative of the likelihood with respect to σ is

$$\frac{\partial^2 l(\mu, \sigma^2)}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_i (x_i - \mu)^2$$

which is negative when evaluated at $\hat{\sigma}^2$, indicating a maximum. As μ is unknown, the MLE $\hat{\sigma}^2$ is not useful. The usual thing that is done is to maximize the *profile likelihood*[2] function, a 'likelihood-like' or *pseudo-likelihood* function, defined as

$$l^{(P)}(\sigma^2|\hat{\mu}_\sigma) = l(\mu, \sigma^2)|_{\mu=\hat{\mu}_\sigma} \tag{1.1.1}$$

for this case, where $\hat{\mu}_\sigma$ is the MLE of μ with σ^2 held constant. Trivially, $\hat{\mu}_\sigma = \bar{x}$, and the value of σ^2 that maximizes the profile likelihood is $n^{-1} \sum_i (x_i - \bar{x})^2$. This estimator is consistent in that, for any $\epsilon > 0$, $\lim_{n\to\infty} P(|s^2 - \sigma^2| > \epsilon) = 0$ [21], but it is biased in that it has expectation $[(n-1)/n]\sigma^2 \neq \sigma^2$. Estimates can become notably inaccurate when a small sample size (i.e. $n = 2$) is used.[3] It is well known, on the other hand, that the sample variance $s^2 = (n-1)^{-1} \sum_i (x_i - \bar{x})^2$ is unbiased. Intuitively, maximizing the profile likelihood to estimate σ^2 does not account for the

[1] We will usually omit the condition on the sample $\mathbf{x}$ from the likelihood if no confusion is possible.

[2] Technically the profile log-likelihood function, but as the difference between a likelihood and log-likelihood function is unimportant for parameter estimation, we will usually not distinguish between the two.

[3] A convenient way to quantify a general error like this is by using 'big O' or 'big O in probability' notation; if $f(n)$ and $g(n)$ are functions, then $f(n) = O(g(n))$ means that f is asymptotically bounded by g. Formally, for some scalar

degrees of freedom lost by estimating the mean parameter μ from the same data. The problem grows worse in models with many mean parameters (known as *nuisance parameters* when they are not of immediate interest), as noted by Neyman & Scott as early as 1948 [33]. Continuing along the lines of the example above, if Y_{h1} and Y_{h2} are iid $N(\mu_h, \sigma^2)$ random variables for $h = 1, \ldots, H$, the log-likelihood is[4]

$$\begin{aligned} l(\mu_h, \sigma^2 | \mathbf{y}) &= \sum_{h,i} \log f(y_{hi}) \\ &= \sum_h \left\{ -\log 2\pi\sigma^2 - \frac{\left[(y_{h1} - \mu_h)^2 + (y_{h2} - \mu_h)^2\right]}{2\sigma^2} \right\}. \end{aligned}$$

The score function for μ_h is thus

$$\frac{\partial l(\mu_h, \sigma^2)}{\partial \mu_h} = \frac{1}{\sigma^2} \sum_h \left[(y_{h1} + y_{h2}) - 2\mu_h\right]$$

and setting it equal to zero yields the MLE for μ_h, $\tilde{\mu}_h = (y_{h1} + y_{h2})/2$. The profile score function for σ^2 is

$$\frac{\partial l^{(P)}(\sigma^2 | \tilde{\mu}_h)}{\partial \sigma} = \sum_h \left\{ -\frac{2}{\sigma} + \frac{1}{\sigma^3} \left[(y_{h1} - \tilde{\mu}_h)^2 + (y_{h2} - \tilde{\mu}_h)^2\right] \right\}$$

and setting it to zero yields the maximum profile likelihood estimator for σ^2

$$\hat{\sigma}^2 = \frac{1}{H} \sum_h \frac{\left[(y_{h1} - \hat{\mu}_h)^2 + (y_{h2} - \hat{\mu}_h)^2\right]}{2}.$$

Let $S_h^2 = [(Y_{h1} - \hat{\mu}_h)^2 + (Y_{h2} - \hat{\mu}_h)^2]/2$ and note that $S_h^2 = (Y_{h1} - Y_{h2})^2/4$ and $\sigma^2 = H^{-1} \sum_h S_h^2$. Then

$$\begin{aligned} E[S_h^2] &= \frac{1}{4} E(Y_{h1} - Y_{h2})^2 \\ &= \frac{1}{4} E(Y_{h1}^2 - 2Y_{h1}Y_{h2} + Y_{h2}^2) \\ &= \frac{1}{4} (\operatorname{var} Y_{h1} + \mu_h^2 + \operatorname{var} Y_{h2} + \mu_h^2 - 2E[Y_{h1}Y_{h2}]) \end{aligned}$$

M, $f(n)$ is $O(g(n))$ if

$$\lim_{n \to \infty} f(n) \leq M g(n).$$

The notation $f(n) = O_p(g(n))$ means that f is asymptotically bounded by g *in probability*. $f(n)$ is $O_p(g(n))$ if

$$\lim_{n \to \infty} P(f(n) \leq M g(n)) = 1.$$

For example, the maximum profile likelihood estimator is biased of order $O_p(n^{-1})$.

[4] Where $\sum_{h,i} = \sum_{h=1}^{H} \sum_{i=1}^{2}$

and since Y_{h1}, Y_{h2} are independent, $E[Y_{h1}Y_{h2}] = \mu_h^2$ so that $E[S_h^2] = \sigma^2/2$. Finally,

$$\begin{aligned} E[\hat{\sigma}^2] &= \frac{1}{H}\sum_h E[S_h^2] \\ &= \frac{1}{H}\sum_h \frac{\sigma^2}{2} \\ &= \frac{\sigma^2}{2} \end{aligned}$$

so that not only is $\hat{\sigma}^2$ biased, but also, since $E[\hat{\sigma}^2]$ does not depend on H, it is inconsistent as the sample size goes to infinity via H. This result holds for any finite per-stratum sample size n_h [33] and demonstrates that estimates based on the profile likelihood may not be accurate in stratified models with small per-stratum sample sizes. This is easy to fix in the normal model (i.e. by using the stratum sample variance), but no solution is obvious for the general (ψ_h, θ) model where the maximum profile likelihood estimator of θ may not exist in closed form.

The problem is also not limited to the class of SMMs; equivalent problems can occur in generalized linear models (GLMs) and quasi-GLMs for fixed mean or dispersion effects [26]. Random and mixed-effects models are also typically heavily parameterized, and so estimating a parameter of interest via the profile likelihood may introduce similar bias problems [35, 27]. In some applications, i.e. estimating confidence intervals, the interest is on estimating θ, and the underlying explanatory model for the response Y (i.e. a GLM or quasi-GLM) is not important. Estimating a traditional confidence interval for μ in the normal model, for example, does not require an underlying explanatory model, but only estimates of μ and σ^2. If the profile likelihood estimator $\hat{\sigma}^2$ is used, the estimated confidence interval may not be very accurate for small sample sizes. This may be a significant problem if μ is small, due to the magnitude of differences (in percentage terms) on that scale.

1.2 Proposed solutions

Numerous methods have been suggested to deal with the nuisance parameters problem. Non-likelihood methods, such as the method of moments, may better account for the presence of nuisance parameters in the model [12]. The method of moments estimator for σ^2 is the unbiased s^2, for example. In general, however, the moments may involve multiple parameters, and it is not necessarily clear how to construct a moment-based estimator for each parameter. Maximum likelihood esti-

mators have desirable asymptotic properties when nuisance parameters are not an issue; they are consistent, asymptotically normally distributed, and asymptotically efficient [21]. As a result, much work has been done to 'adjust' the likelihood for the presence of nuisance parameters. The goal is to eliminate the nuisance parameter effects on bias and consistency, while also preserving the MLE's ideal asymptotic properties.

In exceptional cases the nuisance parameters can be removed from the likelihood via integration or by conditioning on sufficient or ancillary statistics, but this is not tractable in most circumstances. The resulting *marginal likelihood* can be interpreted as a likelihood function solely for the interest parameter θ and is free of nuisance parameter effects [22, 23]. Much work has gone into *asymptotic approximations* to such a marginal likelihood. Barndorff-Nielsen wrote two seminal papers on this topic in the early 1980's. He developed a second-order asymptotic approximation to the distribution of the MLE in [4], and then used it to approximate the marginal likelihood in [5]. Many other approximations have since been suggested, the most influential being the one developed by Cox & Reid in [16]. Other authors that have done work on this topic are Lee & Nelder [26, 27, 24], McCullagh & Tibshirani [31], Severini [43], Pace & Salvan [34], and Sartori [42, 41]. The most important contributions from this work have been the Barndorff-Nielsen modified profile likelihood and Cox-Reid adjusted profile likelihood, both which seek to approximate the marginal likelihood by adjusting the profile likelihood.

Authors have also used asymptotic approximations to the standard (i.e. non-marginal) likelihood in order to estimate θ [14, 39, 40, 12]. The *extended quasi-likelihood* function of Nelder & Pregibon approximates the likelihood function and can be maximized to achieve an estimate of θ. Several authors have used multiple asymptotic approximations in order to form a 'marginal extended quasi-likelihood' function, that itself approximates the marginal likelihood [26, 12] as per above. Other authors have used still more complicated pseudo-likelihood functions [39, 40].

1.3 Scope and contributions of the practicum

In this practicum we seek to more thoroughly understand and resolve the nuisance parameter problem in stratified models, and extend on the work of Cadigan & Tobin [12]. We proceed by setting up the parameter estimation framework in **Chapter 2**. We first develop the method of maximum

likelihood, followed by the methods of maximum extended quasi-likelihood and maximum double-extended quasi-likelihood to form our estimation framework in section 2.2. We then more rigorously explore the theoretical properties of the profile likelihood function in section 2.3 and conclude that it (and the profile extended and double extended quasi-likelihood functions) may be unsuitable for use in the SMM.

In section 2.4 we describe the notion of marginal and conditional inference with respect to θ, and then introduce the Barndorff-Nielsen modified profile likelihood as a highly accurate asymptotic approximation to the marginal or conditional likelihood function. We explore the theoretical properties of the modified profile likelihood and show that it is more appropriate for use in the SMM. We then introduced the Cox-Reid adjusted profile likelihood, as well as Lee & Nelder's approximation to it for quasi-likelihood models.

In **Chapter 3** we develop a specific negative binomial strata mean model, which has been recommended for use in fields such as ecology, genetics, and epidemiology [20, 30, 38, 10]. We derive the estimators developed in **Chapter 2** for this negative binomial SMM, and then compare their performance measures empirically across a range of simulated stratification conditions in section 3.3. In this chapter we extend the work of Cadigan & Tobin by providing a more detailed insight on some of the estimators used in that work.

In **Chapter 4** we apply the results of **Chapter 2** and **Chapter 3** to count data for Atlantic cod caught in stratified random bottom trawl surveys off the south coast of Newfoundland & Labrador. We model these data with the negative binomial SMM and estimate the negative binomial dispersion parameter using an adjusted profile likelihood estimator. We then use our improved estimator to estimate confidence intervals for average trawlable abundance, an important figure used in fisheries research and stock assessment.

Chapter 2

Parameter estimation in the strata mean model

2.1 Content and notation

This chapter summarizes likelihood and asymptotic methods that have been recommended for parameter estimation in the literature. We describe the likelihood and hierarchical likelihood functions for the SMM, and then construct the extended and double-extended quasi-likelihood functions as approximations to these. We provide a summary of theoretical results regarding the bias of maximum profile likelihood and maximum profile extended quasi-likelihood estimators in the SMM, and then develop the Barndorff-Nielsen, Cox-Reid, and Lee-Nelder adjustments to correct for it.

Throughout this chapter we use the following notation: Y will denote an arbitrary random variable, while Y_h will denote a random variable in the h^{th} stratum from a total collection of H strata. All random variables will have parameters (ψ, θ) or (ψ_h, θ) when appropriate, where ψ is a nuisance parameter and θ is the scalar parameter of interest. This vector of parameters will be denoted as λ. We will often discuss results in the context of $H = 1$ and $n_h = 1$ to conserve notation, and when a point is relevant for the $H > 1$ case, we will call attention to it. Arbitrary density or mass functions for Y will be denoted by f. Other necessary notation will be introduced as is required.

2.2 Review of likelihood and quasi-likelihood methods

In this section we will discuss maximum likelihood, maximum extended quasi-likelihood, and maximum double extended quasi-likelihood. We will deal with the nuisance parameter concept in section 2.3.

2.2.1 Maximum likelihood

As detailed in **Chapter 1**, the likelihood function for a given sample $\mathbf{y} = (y_1, \ldots, y_n)$ is defined as

$$L(\psi, \theta|\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\psi, \theta) \tag{2.2.1}$$

and the log-likelihood function is the logarithm of L. In the H strata case where $H > 1$ we have, for $\mathbf{y} = (y_{11}, \cdots, y_{Hn_H})$,

$$l(\psi_1, \ldots, \psi_H, \theta|\mathbf{y}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \log f(y_{hi}|\psi_h, \theta) \tag{2.2.2}$$

which can also be written as

$$l(\psi_1, \ldots, \psi_H, \theta|\mathbf{y}) = \sum_{h} l_h(\psi_h, \theta) \tag{2.2.3}$$

so that the full log-likelihood is the sum of the stratum-specific log-likelihoods. The values of the arguments ψ and θ that maximize L are called the maximum likelihood estimators (MLEs) for the parameters ψ and θ, denoted by $\hat{\psi}_{ml}$ and $\hat{\theta}_{ml}$. The *score functions* are the components of the gradient of the log-likelihood; that is,

$$\frac{\partial l(\psi, \theta)}{\partial \lambda} = \left(\frac{\partial l(\psi, \theta)}{\partial \psi}, \frac{\partial l(\psi, \theta)}{\partial \theta} \right)$$

and, unless necessary, we will usually denote these in the above Leibniz notation rather than assign them specific names. Usually we will be interested in the score function for θ, $\partial l/\partial \theta$, and so we will refer to this as 'the' score unless noted otherwise.

An important property of the score functions is that they have expectation equal to zero. Assuming the standard regularity conditions[1] allowing for the interchange of integration and differentiation,

[1] See, for example, [21].

and using $\mathbf{y} = y$ without loss of generality (WLOG), we have

$$\begin{aligned} E_\theta\left[\frac{\partial l(\psi,\theta)}{\partial\theta}\right] &= \int_{\mathbb{R}} \frac{1}{f(y|\psi,\theta)} \frac{\partial f(y|\psi,\theta)}{\partial\theta} f(y|\psi,\theta)dy \\ &= \int_{\mathbb{R}} \frac{\partial f(y|\psi,\theta)}{\partial\theta} dy \\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f(y|\psi,\theta)dy \\ &= 0 \end{aligned} \tag{2.2.4}$$

where $\int_{\mathbb{R}}$ denotes integration over the real line. This property is important because in most practical problems the MLEs are interior points on the likelihood surface and thus satisfy the ML estimating equations $\partial l/\partial\lambda|_{\lambda=\hat{\lambda}} = (0,0)$. The property of the score having expectation zero is known as *score unbiasedness* in the literature [18]. Intuitively, it can be thought of as an unbiasedness property for the ML estimating equations.

The *observed information* for θ, or just *information*, is the negative derivative of the score, and thus the *observed information matrix* is the negative of the Hessian matrix. We will denote the elements of the information matrix as j, with subscripts to indicate their place in the information matrix. For example, $j_{\psi\psi} = -\partial^2 l/\partial\psi^2$ while $j_{\psi\theta} = -\partial^2 l/\partial\psi\partial\theta$. The *Fisher information matrix* is defined as $E[(\partial l/\partial\lambda)(\partial l/\partial\lambda^T)]$, and the elements are denoted by i and subscripted in the same fashion as per the observed information matrix, i.e. $i_{\theta\theta} = E[(\partial l/\partial\theta)^2]$. Again, as we are primarily interested in θ, we will refer to $i_{\theta\theta}$ as 'the' Fisher information. An important property of the Fisher information follows from the zero expectation of the score. Note the identity

$$\frac{\partial^2 l(\psi,\theta|y)}{\partial\theta^2} = \frac{1}{f(y|\psi,\theta)}\frac{\partial^2 f(y|\psi,\theta)}{\partial\theta^2} - \left[\frac{1}{f(y|\psi,\theta)}\frac{\partial f(y|\psi,\theta)}{\partial\theta}\right]^2 \tag{2.2.5}$$

from [28]. Taking expectations across both sides, we have (assuming sufficient regularity)

$$\begin{aligned} E\left[\frac{1}{f(y|\psi,\theta)}\frac{\partial^2 f(y|\psi,\theta)}{\partial\theta^2}\right] &= \int_{\mathbb{R}} \frac{\partial^2 f(y|\psi,\theta)}{\partial\theta^2} dy \\ &= \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}} f(y|\psi,\theta)dy \\ &= \frac{\partial^2}{\partial\theta^2} 1 \\ &= 0 \end{aligned}$$

and since score unbiasedness implies $\text{var}\,(\partial l/\partial\theta) = E[(\partial l/\partial\theta)^2]$, we have the identity

$$-E\left[\frac{\partial^2 l(\psi,\theta|y)}{\partial\theta^2}\right] = \text{var}\,\frac{\partial l(\psi,\theta|y)}{\partial\theta} \tag{2.2.6}$$

so that $i_{\theta\theta} = -E[\partial^2 l/\partial\theta^2] = E[j_{\theta\theta}]$. For this reason the Fisher information is also called the *expected information*. This identity holds under score unbiasedness and the property is often referred to as *information unbiasedness* in the literature [18].

Inference about the parameters can be performed using the *likelihood ratio statistic* Λ defined as $L(\lambda_0)/L(\hat{\lambda})$ for λ_0 a hypothesized parameter value and $\hat{\lambda}$ the MLE of λ. The notation Λ_ψ is used to denote the likelihood ratio statistic with θ held constant, i.e. $\Lambda_\psi = L(\psi_0|\theta)/L(\hat{\psi}|\theta)$ for inference about ψ. It can be shown that $-2\log\Lambda_\psi \sim \chi^2_1$ asymptotically, and thus p-values can be calculated for inference using the chi-squared distribution [21]. Predictably, there are issues with inference about θ using a likelihood ratio test (LRT) in the presence of nuisance parameters [44], but we will focus solely on point estimation of θ in this practicum.

2.2.2 Quasi-likelihood (QL) and extended quasi-likelihood (EQL)

Quasi-likelihood (QL) arises from an extension to the standard generalized linear model (GLM) framework. Briefly, a GLM extends normal-theory linear model framework to allow the use of any probability distribution from the one-parameter exponential family. For a response Y, a GLM consists of a *random component* (a probability model - and thus a likelihood - for Y from the one-parameter exponential family), a *linear predictor* $\eta = X\beta$ where X is a design or model matrix and β is a vector of parameters, and a monotonic, differentiable *link function* g such that $E[Y] = \psi = g^{-1}(\eta)$. Methods to fit GLMs are standard in all modern statistical software.

All members of the one-parameter exponential family have log-likelihood functions of the form

$$l(\zeta|y) = y\zeta - b(\zeta) + k(y) \tag{2.2.7}$$

where ζ, a function of ψ, is called the *canonical parameter* such that $\partial\zeta/\partial\psi = (\text{var}\,Y)^{-1} = [V(\psi)]^{-1}$, and b is a function such that $\partial b/\partial\zeta = \psi$ and $\partial^2 b/\partial\zeta^2 = V(\psi)$. The function $V(\psi)$ is called the *variance function*. Necessarily, all members must also have score functions with respect to ψ of the

same form, namely

$$\begin{aligned}\frac{\partial l(\zeta)}{\partial \psi} &= \frac{\partial}{\partial \psi}\{y\zeta - b(\zeta) + k(y)\} \\ &= y\frac{\partial \zeta}{\partial \psi} - \frac{\partial b(\zeta)}{\partial \zeta}\frac{\partial \zeta}{\partial \psi} \\ &= \frac{\partial \zeta}{\partial \psi}(y - \psi) \\ &= \frac{y - \psi}{V(\psi)}.\end{aligned}$$

Based on this common score function, Wedderburn developed the quasi-likelihood function in [45]. He defined the quasi-likelihood function $q(\psi|y)$ for a single observation y as

$$\frac{\partial q(\psi|y)}{\partial \psi} = \frac{y - \psi}{V(\psi)}$$

so that

$$q(\psi|y) = \int_y^{\psi} \frac{y - u}{V(u)} du \tag{2.2.8}$$

or, for a sample $\mathbf{y}$ of size n,

$$q(\psi|\mathbf{y}) = \sum_{i=1}^{n} \int_{y_i}^{\psi} \frac{y_i - u}{V(u)} du.$$

The quasi-likelihood function shares some of the important properties of a true likelihood without generally corresponding to an actual probability distribution.[2] The quasi-likelihood function is both score and information unbiased [45], and the maximum quasi-likelihood estimator is asymptotically normally distributed [32]. By replacing the likelihood function in the random component of a GLM with a quasi-likelihood function, Wedderburn extended the applicability of GLMs in much the same way that GLMs extend conventional linear models. This model is technically called a *quasi*-GLM, but due to its similarity to the original framework the name 'GLM' is freely applied to it as well. The method of quasi-likelihood also extends the GLM framework by allowing the modelling of 'extra' dispersion via a scalar multiplicative parameter ϕ in the variance function (i.e. by specifying $\operatorname{var} Y = \phi V(\psi)$). We will not discuss this for strict quasi-likelihood, but will come back to it in subsection 2.2.3.

The quasi-likelihood function can be used to estimate ψ in the SMM class of models. By construction the MQL estimator for ψ is always the sample mean; it is the solution to $\partial q/\partial \psi = 0$ and

[2] The exception being that if Y follows a distribution from the one-parameter exponential family, the corresponding log-likelihood and quasi-likelihood functions are obviously equivalent.

it is trivial to demonstrate that this is $\bar{y}$. Inference about ψ can be performed via the *deviance*, defined for a single observation y as

$$\begin{aligned} D(\psi|\theta, y) &= -2[q(\psi|\theta, y) - q(y|\theta, y)] \\ &= -2\int_y^{\psi} \frac{y-u}{V(u)} du \end{aligned} \tag{2.2.9}$$

and thus for an entire sample $\mathbf{y}$ as $D(\psi) = \sum_i D_i(\psi)$ for $D_i(\psi) = -2[q(\psi|\theta, y_i) - q(y_i|\theta, y_i)]$. The deviance performs an analogous function as the likelihood ratio statistic in likelihood inference and, similarly, D has an asymptotic chi-squared distribution when used for inference about ψ [32]. As the quasi-likelihood is defined in terms of the score of a one-parameter (i.e. mean parameter) exponential family, however, it cannot be be used to estimate the shape parameter θ, nor can the deviance statistic be used for inference about it.

Nelder and Pregibon developed the extended quasi-likelihood (EQL) function in [32] to allow for the estimation of 'nonlinear' parameters (i.e. those not appearing in the linear predictor η). They defined the EQL function q^+ for a single observation y as

$$q^+(\psi, \theta|y) = -\frac{1}{2}\log\{2\pi\phi V(y)\} - \frac{1}{2}D(\psi|\theta, y)/\phi \tag{2.2.10}$$

so that θ also enters via $V(y)$, the variance function with y replacing ψ in $V(\psi)$. $\exp\{q^+\}$ is a type of *saddlepoint approximation*[3], a very important asymptotic approximation that is here applied to the general exponential family likelihood [32]. The maximum EQL estimator for ψ remains unchanged, but now a viable 'quasi-score' equation $\partial q^+/\partial\theta$ can be solved to yield a maximum EQL estimator for θ, denoted $\hat{\theta}_{eql}$.

2.2.3 *h*-likelihood and double EQL (DEQL)

The framework of generalized linear mixed models (GLMMs) extends the GLM framework so that the linear predictor η can contain *random effects* in addition to fixed effects. That is, $\eta = X\beta + Z\gamma$ for Z a model matrix corresponding to the random effects γ, which are assumed to be normally distributed. In a similar way in that GLMs extend classical linear models, hierarchical generalized linear models (HGLMs) developed by Lee & Nelder in [25] and [27] extend the scope of GLMMs by allowing the specification of non-normal random effects. In essence, specifying a HGLM involves

[3] See Appendix A.

specifying two individual GLMs; one for the conditional response $Y|\gamma$ and another for the random effects γ. The family of *conjugate HGLMs* is the simplest of HGLMs, in which the distributions of the fixed and random effects have the same relationship as do a Bayesian posterior and conjugate prior [25]. For example, where the first distribution is that of the fixed effects and the second is that of the random effects, the conjugate HGLMs include the binomial-beta, gamma-inverse gamma, and Poisson-gamma models [27, 24]. HGLMs can be generalized to handle two or more different random effects [25], but we will limit our discussion to a one-random effect conjugate model.

The two distributions are linked by the concept of a *h-likelihood*, or *hierarchical likelihood* (HL) function, which has a somewhat complicated formation. Given random effects γ, the log-likelihood for $Y|\gamma$ has kernel[4]

$$l_0(\zeta(\psi);\, y|\gamma) = y\zeta(\psi) - b[\zeta(\psi)] \tag{2.2.11}$$

where $E[Y|\gamma] = \psi$, $\zeta(\psi)$ is the canonical parameter, and b is the function such that $\partial b[\zeta]/\partial\zeta = \psi$ and $\partial^2\zeta/\partial\zeta^2 = V(\psi)$. In a conjugate model, the distribution of γ depends on that of $Y|\gamma$ and must be chosen appropriately, as listed above. For $v = \zeta(\gamma)$, the chosen distribution has likelihood with kernel of the form[5]

$$l_1(\theta|v) = [\theta a(\theta)v - b[v]]\,/\theta \tag{2.2.12}$$

for θ a dispersion parameter as described in subsection 2.2.2. Let $\nu = \theta a(\theta)$. For any conjugate model, $E[\gamma] = \nu$ [27], and thus l_1 can be viewed as the likelihood of 'quasi-data' ν with quasi-parameter γ (and equivalently, v), where ν satisfies the relationships $E[\nu] = \gamma$ and $\operatorname{var}\nu = \theta V(\gamma)$ for $\partial b[v]/\partial v = \gamma$ and $\partial^2 b[v]/\partial v^2 = V(\gamma)$. The h-likelihood function is then defined as the sum

$$h(\zeta(\psi), \theta;\, y|\gamma, \zeta(\gamma)) = l_0(\zeta(\psi)|\, y|\gamma) + l_1(\theta|\zeta(\gamma)). \tag{2.2.13}$$

Lee & Nelder also developed a quasi-GLM analog for HGLMs in [27] by allowing the specification of a HGLM based on the mean and variance of the individual conditional response and random effects components, instead of on full likelihoods. These models are called hierarchical quasi-GLMs (HQGLM), and allow for the specification of a broader class of models in the same fashion that quasi-likelihood does for GLMs. This has the effect that each of the likelihoods in h are merely

[4]The *kernel* of a log-likelihood function is that part of the function that depends on the parameters.

[5]Note that the ζ in the transformation $V = \zeta(\gamma)$ is the same ζ as $\zeta(\psi)$ appearing as the canonical parameter in l_0. This restriction not required in general, but is necessary for conjugate models.

replaced by appropriate EQL functions. A HQGLM is specified by the variance and link functions V_0, g_0, V_1, g_1 where the 0 subscripts correspond to the quasi-GLM for $Y|\gamma$ and the 1 subscripts correspond to the quasi-GLM for γ. Then we have $E[Y|\gamma] = \psi$ and $\operatorname{var} Y|\gamma = V_0(\psi)$ in terms of the 0 subscripts, and if $E[\gamma] = \nu_1$, $v_1 = g_1(\gamma)$, and $\operatorname{var} \gamma = V(\nu_1)$ then we can think of $E[\nu_1] = \gamma$ and $\operatorname{var} \nu_1 = V_1(\gamma)$ as satisfying GLM relationships for quasi-data ν_1. For single observations y and γ, the deviance corresponding to $Y|\gamma$ is

$$D_0(\psi) = -2\int_y^{\psi} \frac{y-u}{V_0(u)}du$$

and the deviance corresponding to γ is

$$D_1(\nu) = -2\int_{\gamma}^{\nu} \frac{\nu-u}{V_1(u)}du.$$

The corresponding EQL functions are thus

$$q_0^+(\zeta(\psi);\, y|\gamma) = -\frac{1}{2}\log\{2\pi V_0(y)\} - \frac{1}{2}D_0(\psi)$$

and

$$q_1^+(\nu_1|v_1) = -\frac{1}{2}\log\{2\pi V_1(\nu_1)\} - \frac{1}{2\theta}D_1(\nu) + \log\left|\frac{\partial\zeta(\gamma)}{\partial v_1}\right|.$$

The last term in q_1^+ is a standard Jacobian adjustment required due to the change in scale resulting from use of the link function g_1, and it disappears in conjugate models [27]. Combining these functions akin to the h-likelihood case, we form the double extended quasi-likelihood (DEQL) function

$$Q(\zeta(\psi), \theta) = q_0^+(\zeta(\psi);\, y|\gamma) + q_1^+(\theta|v_1) \tag{2.2.14}$$

and it can be maximized to yield parameter estimates for ψ or θ in the SMM. As a composite of EQL functions, the DEQL function approximates the h-likelihood in the same fashion.

2.3 The profile likelihood and variants thereof

We noted in **Chapter 1** that in the presence of ψ, the maximum profile likelihood estimator of θ is biased and even inconsistent in the case where $H > 1$ and n_h is finite while $n = \sum_h n_h \to \infty$ via H. We begin this section with a more general and thorough review of the profile likelihood function and the properties of the maximum profile likelihood estimator.

As mentioned in section 1.1, for the random variable Y with parameter (ψ, θ) the *profile likelihood function* for θ is the function $l^{(P)}$, defined as

$$l^{(P)}(\theta|y, \hat{\psi}_\theta) = \sup_{\psi} l(\psi, \theta|y) \tag{2.3.1}$$

where $\hat{\psi}_\theta$ is the MLE for ψ with θ held fixed. The profile likelihood is useful in that it is trivially easy to specify from the likelihood function. The value of θ that maximizes $l^{(P)}$ is the maximum profile likelihood estimator for θ.

Recall that the profile likelihood is a pseudo-likelihood in that, as per the quasi-likelihood, it does not in general correspond to any particular density function. Unlike a likelihood (or even quasi-likelihood) however, the profile likelihood is not score unbiased. Notice that

$$E\left[\frac{\partial l^{(P)}(\theta|y, \hat{\psi}_\theta)}{\partial \theta}\right] = \int_{\mathbb{R}} \left[\frac{1}{f(y^*|\theta, \hat{\psi}^*_\theta)} \frac{\partial f(y^*|\theta, \hat{\psi}^*_\theta)}{\partial \theta}\right] f(y^*|\theta, \hat{\psi}_\theta) dy^* \tag{2.3.2}$$

where a * denotes a variable that is being integrated over. Then $\hat{\psi}^*_\theta$ is a function of the variable being integrated over in the kernel of the expectation, whereas it is a particular value in the parameter space of ψ in the density of Y. The required cancellation of terms (recall equation 2.2.4) leading to the expectation equalling zero does not occur in general [31]. Similarly, the profile likelihood is not information unbiased; for the second term in equation 2.2.5,

$$E\left[\frac{1}{f(y|\theta, \hat{\psi}_\theta)} \frac{\partial^2 f(y|\theta, \hat{\psi}_\theta)}{\partial \theta^2}\right] = \int_{\mathbb{R}} \left[\frac{1}{f(y^*|\theta, \hat{\psi}^*_\theta)} \frac{\partial^2 f(y^*|\theta, \hat{\psi}^*_\theta)}{\partial \theta^2}\right] f(y^*|\theta, \hat{\psi}_\theta) dy^*$$

so that the required cancellation does not occur to allow the profile Fisher information $i^{(P)}_{\theta\theta}$ to equal $-E[\partial l^{(P)}/\partial \theta]^2$.

McCullagh and Tibshirani derived an expression for the profile score bias in [31]. The Taylor expansion for the profile likelihood about the true parameter ψ is

$$l^{(P)}(\theta|\hat{\psi}_\theta) = l(\psi, \theta) + (\hat{\psi}_\theta - \psi)\frac{\partial l(\psi, \theta)}{\partial \psi} + \frac{(\hat{\psi}_\theta - \psi)^2}{2}\frac{\partial^2 l(\psi, \theta)}{\partial \psi^2} + \dots$$

and taking the derivative with respect to θ yields

$$\frac{\partial l^{(P)}(\theta|\hat{\psi}_\theta)}{\partial \theta} = \frac{\partial l(\psi, \theta)}{\partial \theta} + (\hat{\psi}_\theta - \psi)\frac{\partial^2 l(\psi, \theta)}{\partial \theta \partial \psi} + \frac{(\hat{\psi}_\theta - \psi)^2}{2}\frac{\partial^3 l(\psi, \theta)}{\partial \theta \partial \psi^2} + \dots \tag{2.3.3}$$

which can be reduced to [41]

$$\frac{\partial l^{(P)}(\theta|\hat{\psi}_\theta)}{\partial \theta} = \frac{\partial l(\psi, \theta)}{\partial \theta} - i_{\theta\psi} i^{-1}_{\psi\psi} \frac{\partial l(\psi, \theta)}{\partial \psi} + B^{(P)} + R^{(P)} \tag{2.3.4}$$

where $B^{(P)} = O(1)$ and $R^{(P)} = O(n^{-1/2})$ for a sample $\mathbf{y}$ of size n. The sum of the first two terms, $\partial l/\partial\theta - i_{\theta\psi}i_{\psi\psi}^{-1}\partial l/\partial\psi$, constitute an important quantity in that it is the leading term in the score function for θ when ψ is fixed [41]. This term is called the *partial score* for θ, and its variance

$$\text{var}\left[\frac{\partial l(\psi,\theta)}{\partial\psi} - i_{\theta\psi}i_{\psi\psi}^{-1}\frac{\partial l(\psi,\theta)}{\partial\psi}\right] = i_{\theta\theta} - i_{\theta\psi}i_{\psi\psi}^{-1}i_{\psi\theta} \tag{2.3.5}$$

is called the *partial information* for θ, denoted $i_{\theta\theta|\psi}$. These are the conditional score and expected information functions for θ given knowledge of ψ.

Finally, taking expectations of (2.3.4) we have [31, 18, 41]

$$E\left[\frac{\partial l^{(P)}(\theta|\hat{\psi}_\theta)}{\partial\theta}\right] = 0 - \rho + O(n^{-1}) \tag{2.3.6}$$

where $\rho = E[B^{(P)}] = O(1)$, and thus the profile score bias is of order $O(1)$. The profile information bias is also of order $O(1)$, and a proof of this for full exponential families is given in [31], while [18] gives a more general proof.

For the SMM likelihood (equation 2.2.2),

$$\begin{aligned} E\left[\frac{\partial l^{(P)}(\theta|\hat{\psi}_{1\theta},\ldots,\hat{\psi}_{H\theta})}{\partial\theta}\right] &= \sum_h E\left[\frac{\partial l_h^{(P)}(\theta|\hat{\psi}_{h\theta})}{\partial\theta}\right] \\ &= -\sum_h\left[\rho_h + O(n_h^{-1})\right] \\ &= -\sum_h\left[O(1) + O(n_h^{-1})\right] \end{aligned} \tag{2.3.7}$$

so that for the full model the profile score bias is of order $O(H)$. As the maximum profile likelihood estimator for θ is the solution to $\partial l^{(P)}/\partial\theta = 0$, this has the effect that the maximum profile likelihood estimating equation itself is 'biased' in some sense, relative to a true ML estimating equation.

The profile score bias has implications for the asymptotic properties of the maximum profile likelihood estimator $\hat{\theta}$. For the SMM $\hat{\theta}$ is inconsistent when $H \to \infty$ while n_h remains finite. Sartori considered the asymptotic qualities of $\hat{\theta}_{(P)}$ in [41] under a more complete two-index asymptotic setting, in which both n_h and H go to infinity, but at possibly different rates. In this scenario, and as alluded to in section 1.1, $\hat{\theta}_{(P)}$ is always consistent [41], independently of how quickly each variable increases without bound. However, the speed at which each variable approaches infinity affects the speed of convergence of $\hat{\theta}_{(P)} \to \theta$. Assuming WLOG that n_h is the same for all h, an important

condition in the asymptotic setting for the SMM is

$$n_h^{-1}H = O(1) \tag{2.3.8}$$

or the condition that n_h increase without bound at a faster rate than H does. Under this condition, an expansion for the maximum profile likelihood estimator is [41]

$$\hat{\theta}_{(P)} = \theta + O_p(n^{-1/2}) \tag{2.3.9}$$

so that the difference $\hat{\theta}_{(P)} - \theta$ is of order $O_p(n^{-1/2})$. If H increases faster than n_h so that $n_h^{-1}H \neq O(1)$, then $\hat{\theta}_{(P)} - \theta = O_p(n_h^{-1})$, akin to the bias in the one-index asymptotic example given in section 1.1. Condition 2.3.8 is also sufficient for the standardized profile score statistic

$$Z = i_{\theta\theta|\psi}^{-1/2} \frac{\partial l^{(P)}(\psi, \theta)}{\partial \theta}, \tag{2.3.10}$$

where $i_{\theta\theta|\psi}$ denotes the partial information for θ, to have an asymptotic $N(0,1)$ distribution [41].

Pace and Salvan noted in [34] that, in general[6], the profile likelihood effectively places excess weight on information about θ. They noted that the profile expected information $i_{\theta\theta}^{(P)}$ is first-order equivalent to the unconditional expected information $i_{\theta\theta}$ as calculated from the likelihood, whereas the partial information for θ given ψ $i_{\theta\theta|\psi}$ is, from equation 2.3.5, $i_{\theta\theta} - i_{\theta\psi}i_{\psi\psi}^{-1}i_{\psi\theta}$. Since $i_{\theta\theta}^{(P)}$ is first-order equivalent to $i_{\theta\theta}$, $i_{\theta\theta|\psi} < i_{\theta\theta}^{(P)}$ in an asymptotic sense. The maximum profile likelihood estimator for θ does not properly take into account the sampling variability of $\hat{\psi}_\theta$ [34], and thus the intuitive 'degrees of freedom' adjustment to $\hat{\theta}_{(P)}$ is missing.

To summarize, we have that

- The maximum profile likelihood estimator for θ is biased and inconsistent in the one-index asymptotic setting where the number of strata increases without bound, independent of the size of $n_h \,\forall\, h$ so long as n_h is finite. That is, even though $n \to \infty$ as $H \to \infty$, we have that, for any $\epsilon > 0$, $\lim_{n\to\infty} P(|\hat{\theta}_{(P)} - \theta| < \epsilon) = 0$.
- Both the profile score function and profile expected information for any stratum (given any $\hat{\psi}_{h\theta}$) are biased of order $O(1)$. For the complete model with H strata, the biases are of order $O(H)$.

[6] The exceptional case will be discussed in subsection 2.4.2

- In the two-index asymptotic setting, the maximum profile likelihood estimator for θ is consistent. However, it converges to θ more quickly when $n_h \to \infty$ more rapidly than H does. In this case, the difference $\hat{\theta}_{(P)} - \theta$ is of order $O_p(n^{-1/2})$. If $H \to \infty$ quicker than n_h, the same difference is of order $O_p(n_h^{-1})$.

- $n_h \to \infty$ faster than H is also a sufficient condition for the profile score bias to be asymptotically negligible, and $i_{\theta\theta|\psi}^{-1/2}\partial l^{(P)}(\psi,\theta)/\partial\theta \dot{\sim} N(0,1)$.

Some research has been done on so-called *'profile' quasi-likelihood* functions, which are not in general constructed in the same fashion as the profile likelihood. Barndorff-Nielsen [7] and Adimara & Ventura [2] discussed the construction of a profile quasi-likelihood function based on a multiplicative adjustment to the profile quasi-score function and noted that the resulting quasi-score and quasi-information biases were both - like the profile likelihood - of order $O(1)$. Lin and Zhang constructed a profile quasi-likelihood in [29] and similarly described a bound on their estimator for θ as $\hat{\theta} - \theta = O_p(n^{-1/2})$ in a one-index asymptotic setting. The construction of these functions seems similar to the *profile extended quasi-likelihood* function, defined analogously for the EQL as the profile likelihood is to the likelihood, and we suspect that the EQL approximately has the same bias properties. This is due to the fact that the exponentiated EQL function is a saddlepoint approximation to an exponential family log-likelihood and that the profile likelihood for an exponential family has information and score bias of order $O(1)$.

We are not aware of any research on the asymptotic score and information bias properties of the profile double-extended quasi-likelihood function, defined as

$$Q^{(P)}(\theta) = \sup_{\psi,\gamma} Q(\psi,\theta,\gamma). \tag{2.3.11}$$

for Q as in equation 2.2.14. Note that as the DEQL function includes random effects, the profile DEQL for θ depends on maximum likelihood estimates for both ψ and γ. An analysis of its asymptotic properties is beyond the scope of this practicum. For conjugate HGLMs we speculate that the bias properties are similar to the EQL function as $Q^{(P)}$ is the sum of two EQL functions.

The one- and two-index asymptotic analyses yield clues about the properties of the maximum profile estimators in a SMM with different combinations of H and n_h. The most severe bias is to be expected in the high-dimensional case where n_h is less than H on average. This occurs in practice,

and the MLE of θ may not be reliable in this setting. The negative binomial SMM based on DFO research survey data that we discuss in **Chapter 4** is a high-dimensional example.

2.4 Asymptotic methods for bias correction

Let (t_1, t_2) be jointly sufficient for (ψ, θ) and a be ancillary.[7] If we could factorize the likelihood as

$$L(\psi, \theta) \approx L(\theta; t_2|a)L(\psi, \theta; t_1|t_2, a) \tag{2.4.1}$$

or

$$L(\psi, \theta) \approx L(\theta; t_2|t_1, a)L(\psi, \theta; t_1|a) \tag{2.4.2}$$

then we could maximize $L(\theta; t_2|a)$, called the *marginal likelihood*, or $L(\theta; t_2|t_1, a)$, the *conditional likelihood*, to obtain an estimate of θ. Both $L(\theta; t_2|a)$ and $L(\theta; t_2|t_1, a)$ - special cases of the *partial likelihood* developed by Cox in [17] - condition on statistics that contain all of the information about θ and negligible information about ψ, and so seek to eliminate the effect of ψ when estimating θ. It is difficult to find such a factorization in general however, or even show that one exists [23, 37]; as a result, much work has been done on *approximating* the marginal or conditional likelihood.

Recent software development (i.e. AD Model Builder, `http://admb-project.org`) has made it possible to directly remove nuisance parameters by integrating them out of the likelihood. That is, it may be possible to estimate θ by using

$$L(\theta) = \int_{\psi} L(\psi, \theta) d\psi$$

called the *integrated likelihood* a similarly 'marginal' likelihood, provided the integral exists. We focus on approximate methods; a treatise on exact marginal inference using the integrated likelihood is beyond the scope of this practicum, but remains a promising research direction.

The most influential approximation to the marginal likelihood is based on an approximation to the asymptotic distribution to the MLE. It is well known that the asymptotic distribution of the MLE $\hat{\theta}$ is $N(0, i_{\theta\theta}^{-1})$ [21], and this is a first-order approximation (i.e. $O(n^{-1})$) for finite sample sizes. Barndorff-Nielsen developed a higher-order approximation in [4] based on a saddlepoint approximation to

[7] A *sufficient statistic* for a parameter contains the maximum information about the value of that parameter. An *ancillary statistic* for a parameter has a sampling distribution that does not depend on that parameter.

the distribution of the minimal sufficient statistic in full exponential families. Given an ancillary[8] statistic a, the conditional distribution of $\hat{\theta}$ can be approximated to order $O(n^{-3/2})$ by

$$p(\hat{\theta};\, \theta|a) \approx c(a) \left|j(\hat{\theta})\right|^{1/2} \Lambda_{\hat{\theta}} \tag{2.4.3}$$

where $j(\hat{\theta})$ is the profile observed information matrix, $\Lambda_{\hat{\theta}}$ is the likelihood ratio statistic $L(\theta)/L(\hat{\theta})$, and $c(a)$ is a normalizing constant dependent on a so that $\int_{\mathbb{R}} p(\hat{\theta}|a) = 1$. Denoting the right side of 2.4.3 as p^*, it is known as Barndorff-Nielsen's p^* *formula*.

2.4.1 Modified profile likelihood (MPL) and the Barndorff-Nielsen adjustment

Barndorff-Nielsen constructed a modification to the profile likelihood function in [5] by using the p^* formula in equation 2.4.3. Let $t = (t_1, t_2)$ be a minimal sufficient statistic and u be a statistic so that both $(\hat{\psi}, u)$ and $(\hat{\psi}, \hat{\theta})$ are one-to-one transformations of t with the distribution of u depending only on θ. The marginal density of u is then [5]

$$\begin{aligned} f(u;\, \theta) &= \frac{f(\hat{\psi}, u; \psi, \theta)}{f(\hat{\psi}; \psi, \theta|u)} \\ &= \left\{ f(\hat{\psi}, \hat{\theta}; \psi, \theta) \left|\frac{\partial(\hat{\psi}, \hat{\theta})}{\partial(\hat{\psi}, u)}\right| \right\} \Big/ \left\{ f(\hat{\psi}_\theta;\, \psi, \theta|u) \left|\frac{\partial \hat{\psi}_\theta}{\partial \hat{\psi}}\right| \right\}. \end{aligned} \tag{2.4.4}$$

The terms $|\partial(\hat{\psi}, \hat{\theta})/\partial(\hat{\psi}, u)|$ and $|\partial \hat{\psi}_\theta / \partial \hat{\psi}|$ are complicated and are described in [5]. From [46], $|\partial \hat{\psi}_\theta / \partial \hat{\psi}|^{-1}$ has the form

$$\left|\frac{\partial \hat{\psi}}{\partial \hat{\psi}_\theta}\right| = \left|j_{\psi\psi}(\hat{\lambda}_\theta)\right| \left|l_{\psi;\hat{\psi}}(\hat{\lambda}_\theta|\hat{\psi}, u)\right|^{-1}$$

where $\hat{\lambda}_\theta = (\theta, \hat{\psi}_\theta)$ and where the term $|l_{\psi;\hat{\psi}}(\hat{\lambda}_\theta|\hat{\psi}, u)|$ is the determinant of a *sample space derivative*, defined as the matrix $\partial^2 l(\lambda|\hat{\lambda}, u)/\partial\psi\partial\hat{\psi}^T$. Often this expression itself is intractable, and the sample space derivative must be approximated or eliminated altogether. The term $|\partial(\hat{\psi}, \hat{\theta})/\partial(\hat{\psi}, u)|$, on the other hand, does not depend on θ. Using the p^* formula to approximate both $f(\hat{\psi}, \hat{\theta}|\psi, \theta)$ and

[8] Or approximately ancillary, by which $\forall\, \delta > 0\ p(a|\theta + \delta n^{-1/2}) = p(a|\theta)\{1 + O(n^{-1})\}$ [37].

$f(\hat{\psi}_\theta;\, \psi, \theta|u)$ in equation 2.4.4, we have the kernel

$$\begin{aligned} L(\theta;u) &\propto \frac{\left|j(\hat{\psi},\hat{\theta})\right|^{1/2}}{\left|j(\hat{\psi}_\theta,\theta)\right|^{1/2}} \frac{L(\psi,\theta)}{L(\hat{\psi},\hat{\theta})} \frac{L(\hat{\psi}_\theta,\theta)}{L(\psi,\theta)} \left|j(\hat{\psi}_\theta,\theta)\right| \left|l_{\psi;\hat{\psi}}(\hat{\psi}_\theta,\theta)\right|^{-1} \\ &\propto L(\hat{\psi}_\theta,\theta) \left|j(\hat{\psi}_\theta,\theta)\right|^{1/2} \left|l_{\psi;\hat{\psi}}(\hat{\psi}_\theta,\theta)\right|^{-1} \\ &= L^{(P)}(\theta|\hat{\psi}_\theta) \left|j_{\psi\psi}(\theta,\hat{\psi}_\theta)\right|^{1/2} \left|l_{\psi;\hat{\psi}}(\hat{\psi}_\theta,\theta)\right|^{-1} \end{aligned}$$

and, taking the logarithm, we get

$$l(\theta;u) \approx l^{(P)}(\theta|\hat{\psi}_\theta) + \frac{1}{2}\log\left|j_{\psi\psi}(\theta,\hat{\psi}_\theta)\right| - \log\left|l_{\psi;\hat{\psi}}(\theta,\hat{\psi}_\theta)\right|. \tag{2.4.5}$$

This approximation, denoted $l^{(M)}(\theta)$, is called the *modified profile likelihood*, and as it is based on the saddlepoint approximation, it is a highly accurate approximation to the 'exact' marginal likelihood $l(\theta|u)$ which is in general not tractable. The term involving the profile observed information corrects for the emphasis placed on information about θ due to the profile likelihood; the term involving the sample space derivative $l_{\psi;\hat{\psi}}$ corrects for non-orthogonality[9] of the parameters ψ and θ. The modified profile likelihood also has the property that it is invariant to parameter transformations that preserve θ [5]. The value of θ that maximizes $l^{(M)}$ is the maximum modified profile likelihood (MPL) estimator for θ, denoted $\hat{\theta}_{(M)}$.

Denoting the last two terms in equation 2.4.5 as $M(\theta)$, the modified profile likelihood can be written as

$$l^{(M)}(\theta) = l^{(P)}(\theta|\hat{\psi}_\theta) + M(\theta) \tag{2.4.6}$$

to make the additive adjustment to the profile log-likelihood explicit. We can then expand $l^{(M)}$ about ψ in the same way as was done for the profile likelihood in equation 2.3.4. Proceeding identically, we have

$$\frac{\partial l^{(M)}(\theta|\hat{\psi}_\theta)}{\partial\theta} = \frac{\partial l^{(P)}(\theta|\hat{\psi}_\theta)}{\partial\theta} + B^{(P)} + R^{(P)} + B_1 + R_1 \tag{2.4.7}$$

so that

$$E\left[\frac{\partial l^{(M)}(\theta|\hat{\psi}_\theta)}{\partial\theta}\right] = -\rho + O(n^{-1}) + E[B_1] + E[R_1].$$

[9] We will discuss this point later, in subsection 2.4.2.

Diciccio et al showed in [18] that $E[B_1] = \rho$ and $E[R_1] = O(n^{-1})$, so that the modified profile score bias is

$$E\left[\frac{\partial l^{(M)}(\theta|\tilde{\psi}_\theta)}{\partial\theta}\right] = -\rho + \rho + O(n^{-1})$$
$$= O(n^{-1}) \tag{2.4.8}$$

as opposed to the profile score bias of order $O(1)$. Similarly, the modified profile information bias is of order $O(n^{-1})$ [18]. From equations 2.3.7 and 2.4.8, we have that in the SMM full model with H strata, the modified profile score bias is of order

$$E\left[\frac{\partial l^{(M)}(\theta)}{\partial\theta}\right] = \sum_h E\left[\frac{\partial l_h^M(\theta)}{\partial\theta}\right]$$
$$= \sum_h O(n_h^{-1})$$
$$= O(n_h^{-1}). \tag{2.4.9}$$

Sartori also studied the properties of the modified profile likelihood in the two-index asymptotic setting in [41]. The results about consistency for $\hat{\theta}$ in the two-index asymptotic setting obviously also hold for $\hat{\theta}_{(M)}$, but the rates of convergence to θ differ. Analogous to condition 2.3.8, another important condition is that

$$n_h^{-3}H = O(1) \tag{2.4.10}$$

in that n_h increases without bound faster than $H^{1/3}$, which is weaker than condition 2.3.8. Under this condition, the expansion for $\hat{\theta}_{(M)}$ is [41]

$$\hat{\theta}_{(M)} = \theta + O_p(n^{-1/2}) \tag{2.4.11}$$

which is the same as expansion 2.3.9. If condition 2.4.10 does not hold then the difference $\hat{\theta}_{(M)} - \theta$ is of order $O_p(n_h^{-2})$, as opposed to $O_p(n_h^{-1})$ in the profile likelihood case. Condition 2.4.10 is also sufficient for the standardized score statistic based on the modified profile likelihood to be asymptotically normal; i.e., $i_{\theta\theta|\psi}^{-1/2}\partial l^{(M)}(\theta)/\partial\theta \dot{\sim} N(0,1)$ [41]. Sartori's results in the two-index asymptotic setting demonstrate that the difference between either the maximum profile likelihood estimator or the maximum MPL estimator and the true parameter θ is bound in probability by order $O_p(n^{-1/2})$ in the respective 'best' cases. As $\hat{\theta}_{(M)}$ achieves this bound for $n_h^{-3}H = O(1)$ whereas $\hat{\theta}_{(P)}$ achieves it

for $n_h^{-1}H = O(1)$, the maximum MPL estimator achieves it under weaker conditions. In cases where these conditions do not hold, the difference $\hat{\theta}_{(M)} - \theta$ also enjoys a smaller bound in probability (of order $O_p(n_h^{-2})$) than $\hat{\theta}_{(P)} - \theta$ (of order $O_p(n_h^{-1})$). The analysis in the two-index setting suggests that the maximum MPL estimator is likely to provide better estimates than the profile likelihood in situations where $H > n_h$ on average.

Reid noted in [37] that the use of conditioning on an ancillary statistic accounts for the 'appropriate' degrees of freedom adjustment present in MPL estimators, as it reduces the dimension of the data Y from n to $n - H$. Determining an ancillary (or approximately ancillary) statistic is not trivial in a general model, however, and the calculation of $l_{\psi;\hat{\psi}}$ may be difficult to the point that the modified profile likelihood is prohibitively complicated to use in practice [43]. As such, numerous approximations have been made that affect the modification term $M(\theta)$ in equation 2.4.6 so as to produce a general set of 'adjusted' profile likelihoods that reduce the profile score bias [18]. McCullagh and Tibshirani developed a parametric bootstrap method to calculate an adjustment term in [31] based on the desired outcome of profile score and information unbiasedness. Pace and Salvan discussed a profile likelihood based on the existence of a least favorable curve in the parameter space in [34], similar to the profile extended quasi-likelihood function developed in [29] that was mentioned in section 2.3. Barndorff-Nielsen and Severini discussed approximations to $M(\theta)$ in [6] and [43] respectively. We will denote $M(\theta)$ as defined above as the *Barndorff-Nielsen adjustment* to the profile likelihood.

2.4.2 Adjusted profile likelihood (APL) and the Cox-Reid adjustment

The Cox-Reid *adjusted profile likelihood* (APL) developed in [16] is a popular modification that arises as an approximation to the modified profile likelihood $l^{(M)}$. It is based on the concept of *orthogonal parameters* in an information-geometric sense. In the SMM model, for example, the chosen probability measure f forms a differentiable manifold with coordinate system (ψ, θ), equipped with the Fisher information as a Riemannian metric [3]. In this context, the parameters ψ and θ are said to be orthogonal if the Fisher information $i_{\psi\theta}$ is equal to 0.

Recall from section 2.3 that the profile likelihood in general places excess weight on information about θ; the partial information $i_{\theta\theta|\psi}$ is $i_{\theta\theta} - i_{\theta\psi}i_{\psi\psi}^{-1}i_{\psi\theta}$, and as $i_{\theta\theta}^{(P)}$ is first-order equivalent to the unconditional expected information $i_{\theta\theta}$, $i_{\theta\theta|\psi} < i_{\theta\theta}^{(P)}$ in an asymptotic sense. If ψ and θ are

orthogonal, however, then $i_{\psi\theta} = 0$ and thus $i_{\theta\theta|\psi} = i_{\theta\theta}$. Cox and Reid showed that an orthogonal transformation of (ψ, θ) can always be constructed if θ is a scalar parameter [16], and so assuming (ψ, θ) are orthogonal, the Cox-Reid APL is

$$l^{(A)}(\theta) = l^{(P)}(\theta) - \frac{1}{2}\log\left|j_{\psi\psi}(\theta, \hat{\psi}_\theta)\right| \tag{2.4.12}$$

and we will call the term $-2^{-1}\log|j_{\psi\psi}(\theta, \hat{\psi}_\theta)|$ the *Cox-Reid adjustment*, denoted $A(\theta)$. The adjusted profile likelihood is a generalization of the method of restricted maximum likelihood (REML, see [35]) for GLMMs [26]. It is a special case of the modified profile likelihood; when ψ and θ are orthogonal, the two are equal [16].

It is interesting to view the APL by the order of its approximation to the 'exact' marginal likelihood $L(\theta)$. Whereas the modified profile likelihood is calculated using saddlepoint approximations in the form of the p^* formula, the adjusted profile likelihood is a lower-order Laplace approximation. Assuming ψ is scalar, applying equation A.1.3 from Appendix A to the marginal likelihood $L(\theta) = \int L(\psi, \theta)d\psi$ yields

$$\begin{aligned} L(\theta) &= \int_{\mathbb{R}} L(\psi, \theta)d\psi \\ &\approx \exp\left\{l^{(P)}(\theta|\hat{\psi}_\theta)\right\}\left\{-\frac{2\pi}{\frac{\partial^2 l(\psi,\theta)}{\partial\psi^2}\big|_{\psi=\hat{\psi}_\theta}}\right\} \\ &= L^{(P)}(\theta|\hat{\psi}_\theta)\left[j_{\psi\psi}(\hat{\psi}_\theta)^{-1/2}\right] \end{aligned} \tag{2.4.13}$$

which agrees with equation 2.4.12 for scalar ψ. As it is a special case of the MPL, its asymptotic properties are identical as long as ψ and θ are approximately orthogonal; if not, the MPL has preferable asymptotic properties [37].

A very important property was proved by Barndorff-Nielsen in [8], in that (ψ, ϕ) are orthogonal in joint quasi-GLMs for mean parameters ψ and dispersion parameters ϕ. Since θ is always a scalar in the SMM class of models, if a joint quasi-GLM can be reparameterized so that θ takes the 'usual' place of the dispersion parameter[10] ϕ, the orthogonality property holds for (ψ, θ) as well. As such, the Cox-Reid adjustment is particularly useful with respect to pseudo-likelihoods in the SMM class of models. It can be applied to more general pseudo-likelihood functions than the profile likelihood, such as the profile EQL and DEQL functions, by using the appropriate observed

[10] We do this in subsection 3.2.3.

pseudo-information. Lee & Nelder applied the Cox-Reid adjustment to the EQL function in [26] and to the DEQL function in [27] and [24]. The observed information is often multiplied by a $1/2\pi$ term to improve the approximation in the pseudo-likelihood case [26, 24].

2.4.3 The Lee-Nelder adjustment for quasi-likelihood functions

Lee & Nelder proposed a very simple adjustment for fixed-effects quasi-likelihood models that approximates the Cox-Reid adjustment in [26]. They noted that in quasi-GLMs with H fixed effects, the Cox-Reid adjustment (with the 2π factor) can be written as

$$A(k) = -\frac{1}{2}\log\left|\frac{\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}}{2\pi}\right| \tag{2.4.14}$$

where $\mathbf{X}$ is the model (design) matrix and $\hat{\mathbf{W}}$ is an $n \times n$ diagonal matrix with j^{th} diagonal element

$$[V(\hat{\psi}_j)]^{-1}\frac{\partial\psi_j}{\partial\eta}\Big|_{\psi_j=\hat{\psi}_j}$$

for η the linear predictor $\mathbf{X}\beta$. In this framework, the quasi-likelihood hat matrix is

$$\mathbf{\Lambda} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{1/2} \tag{2.4.15}$$

where $\mathbf{V}$ is an $n \times n$ diagonal matrix with j^{th} diagonal element equal to $V(\psi_j)$. Recalling the EQL function from equation 2.2.10, Lee & Nelder suggested optimizing

$$q^{+}_{(B)} = -\frac{1}{2}\log\{2\pi V(y)\} - \frac{1}{2}D^*(\psi|\theta, y). \tag{2.4.16}$$

Where the adjusted deviance $D^*(\psi|\theta, y)$ is $D(\psi|\theta, y)/(1 - \kappa_{hi})$ and κ_{hi} are the diagonal elements of $\mathbf{\Lambda}$. They noted that optimizing equation 2.4.16 was simpler than optimizing the Cox-Reid adjusted profile likelihood and provided esssentially identical results.

Next, in **Chapter 3**, we apply the results of this chapter to a specific strata mean model. In section 3.1 we develop a negative binomial SMM, and in section 3.2 we describe the ML, EQL, DEQL, AML, CREQL, and LNEQL estimators for the negative binomial dispersion parameter. We then perform a simulation study in section 3.3 to measure the various estimators' empirical bias and mean squared error properties. We use the results of this simulation study in **Chapter 4** to select an appropriate estimator for the negative binomial dispersion parameter in a highly-stratified survey application.

Chapter 3

The negative binomial strata mean model and dispersion parameter estimators

3.1 The negative binomial model

As mentioned in **Chapter 1**, a useful form for a strata-mean model with parameter (ψ, θ) is the negative binomial SMM. Specifically, for the response Y we have $Y \sim \text{negbin}(\mu, k)$, where the negative binomial mass function is

$$P(Y = y) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k. \tag{3.1.1}$$

for $y = 0, 1, 2, \ldots$, $\mu > 0$, and $k > 0$. The negative binomial distribution has been suggested as an appropriate model for count data and finds widespread use in ecology, genetics, and epidemiology [20, 30, 38, 10].

The negative binomial distribution is a generalization of the Poisson distribution with parameter μ, which itself is often suggested for modelling count data. The Poisson mass function is

$$P(Y = y \,|\, \mu) = \frac{e^{-\mu}\mu^y}{y!} \tag{3.1.2}$$

where $y = 0, 1, 2, \ldots$ and $\mu > 0$. The Poisson distribution is a member of the exponential family of distributions; the mass function can be written as

$$P(Y = y \mid \mu) = \exp\{y\zeta - b[\zeta] + k(y)\} \tag{3.1.3}$$

where $\zeta = \log \mu$, $b[\zeta] = \exp\{\zeta\}$, and $k(y) = -\log y!$. A Poisson random variable has the property that its mean and variance are equal; the Poisson mean is

$$\begin{aligned} E[Y] &= \sum_{j=0}^{\infty} \frac{j\mu^j e^{-\mu}}{j!} \\ &= 0 + \sum_{j=1}^{\infty} \frac{j\mu^j e^{-\mu}}{j!} \\ &= \sum_{j=1}^{\infty} \frac{\mu^j e^{-\mu}}{(j-1)!} \\ &= \sum_{j=0}^{\infty} \frac{\mu^{j+1} e^{-\mu}}{j!} \\ &= \mu \sum_{j=0}^{\infty} \frac{\mu^j e^{-\mu}}{j!} \\ &= \mu \end{aligned}$$

and its variance is

$$\begin{aligned} \operatorname{var} Y &= EY^2 - [EY]^2 \\ &= \sum_{j=0}^{\infty} \frac{j^2 \mu^j e^{-\mu}}{j!} - \mu^2 \\ &= \sum_{j=1}^{\infty} \frac{j\mu^j e^{-\mu}}{(j-1)!} - \mu^2 \\ &= \sum_{j=0}^{\infty} \frac{(j+1)\mu^{j+1} e^{-\mu}}{j!} - \mu^2 \\ &= \mu \left[\sum_{j=0}^{\infty} \frac{j\mu^j e^{-\mu}}{j!} + \sum_{j=0}^{\infty} \frac{\mu^j e^{-\mu}}{j!} \right] - \mu^2 \\ &= \mu[\mu + 1] - \mu^2 \\ &= \mu. \end{aligned}$$

For many sets of count data observed in practice, however, the sample variance exceeds the sample mean, which is a property known as *overdispersion*. In these cases, the Poisson model is insufficient. One generating mechanism for the negative binomial distribution is when the Poisson mean parameter μ is modelled as a gamma random variable. The gamma density function for a random variable U with parameters $\alpha > 0$ and $\beta > 0$ is

$$f_U(u|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} u^\alpha e^{\frac{-u}{\beta}} \tag{3.1.4}$$

and has mean

$$\begin{aligned} E[U] &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} u u^\alpha e^{\frac{-u}{\beta}} du \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty u^{\alpha+1} e^{\frac{-u}{\beta}} du \\ &= \frac{\Gamma(\alpha+1)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} u^{\alpha+1} e^{\frac{-u}{\beta}} du \\ &= \alpha\beta \end{aligned}$$

and variance

$$\begin{aligned} \operatorname{var} U &= EU^2 - [EU]^2 \\ &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} u^2 u^\alpha e^{\frac{-u}{\beta}} du - \alpha^2\beta^2 \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty u^{\alpha+2} e^{\frac{-u}{\beta}} du - \alpha^2\beta^2 \\ &= \frac{\Gamma(\alpha+2)\beta^{\alpha+2}}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty \frac{1}{\Gamma(\alpha+2)\beta^{\alpha+2}} u^{\alpha+2} e^{\frac{-u}{\beta}} du - \alpha^2\beta^2 \\ &= (\alpha+1)\alpha\beta^2 - \alpha^2\beta^2 \\ &= \alpha^2\beta^2 + \alpha\beta^2 - \alpha^2\beta^2 \\ &= \alpha\beta^2. \end{aligned}$$

Let $U \sim \text{gamma}(k, k^{-1}\mu)$ be an unobserved random variable for some $k > 0$ such that $E[U] = \mu$ and $\operatorname{var} U = k^{-1}\mu^2$. Then we have a hierarchical model

$$Y|U \sim \text{Poisson}(\mu)$$

$$U \sim \text{gamma}(k, k^{-1}\mu).$$

The unconditional mass function for Y is,

$$\begin{aligned}
P(Y=y) &= \int_U P(Y=y|U=u) f_U(u) du \\
&= \int_0^\infty \frac{u^y e^{-u}}{y!} \frac{1}{\Gamma(k)} \left(\frac{k}{\mu}\right)^k u^{k-1} e^{-u\left(\frac{k}{\mu}\right)} du \\
&= \frac{1}{y!\Gamma(k)} \left(\frac{k}{\mu}\right)^k \int_0^\infty u^y e^{-u} u^{k-1} e^{-u\frac{k}{\mu}} du \\
&= \frac{1}{y!\Gamma(k)} \left(\frac{k}{\mu}\right)^k \int_0^\infty u^{y+k-1} e^{-u\left(\frac{\mu+k}{\mu}\right)} du \\
&= \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{k}{\mu}\right)^k \left(\frac{\mu}{\mu+k}\right)^{y+k} \\
&= \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k
\end{aligned}$$

and since $y! = \Gamma(y+1)$ for $y = 0, 1, 2, \ldots$ we have (for purely aesthetic reasons)

$$P(Y=y) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k$$

which is the negative binomial mass function for Y given in equation 3.1.1. The negative binomial distribution is sometimes referred to as the *Poisson-gamma* distribution. The hierarchical model is useful for calculating the moments of Y as we can use well-known conditional identities. The mean is

$$\begin{aligned}
E[Y] &= E\left\{E[Y|U]\right\} \\
&= E[U] \\
&= \mu
\end{aligned} \tag{3.1.5}$$

and the variance is

$$\begin{aligned}
\operatorname{var} Y &= E[\operatorname{var}(Y|U)] + \operatorname{var}(E[Y|U]) \\
&= E[U] + \operatorname{var} U \\
&= \mu + k^{-1}\mu^2.
\end{aligned} \tag{3.1.6}$$

As $\operatorname{var} Y > E[Y]\ \forall\ k > 0$, the negative binomial distribution is more appropriate than the Poisson distribution for modelling overdispersed count data. The two distributions are equivalent in the

limiting case where $k \to \infty$, since

$$\begin{aligned}\lim_{k\to\infty} P(Y=y) &= \lim_{k\to\infty} \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k \\ &= \frac{\mu^y}{y!} \lim_{k\to\infty} \frac{\Gamma(y+k)}{\Gamma(k)(\mu+k)^y} \left(\frac{1}{1+k^{-1}\mu}\right)^k\end{aligned}$$

and as $\Gamma(y+k)/\Gamma(k) = \prod_{j=0}^{y-1}(j+k)$,

$$\begin{aligned}&= \frac{\mu^y}{y!} \lim_{k\to\infty} \frac{\prod_{j=0}^{y-1}(1+k^{-1}j)}{(1+k^{-1}\mu)^y} \left(\frac{1}{1+k^{-1}\mu}\right)^k \\ &= \frac{\mu^y e^{-\mu}}{y!}\end{aligned}$$

which is the Poisson mass function. As a mass function uniquely defines the probability distribution of a random variable, Y is distributed as Poisson(μ) when $k \to \infty$. As noted in [36] it is possible to relax the parameter space of k somewhat so that k can take some negative values while $P(Y=y)$ remains a valid probability mass function. For $k < -\mu$ this has the effect that $\operatorname{var} Y < \mu$, and thus this 'relaxed' negative binomial model is also capable of modeling *underdispersed* count data. If k is known, then the negative binomial distribution is a member of the one-parameter exponential family of distributions. Notice that

$$\begin{aligned}P(Y=y) &= \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k \\ &= \exp\left\{y\log\left(\frac{\mu}{\mu+k}\right) - \left[-k\log\left(\frac{k}{\mu+k}\right)\right] + \log\frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\right\} \\ &= \exp\left\{y\zeta - b(\zeta) + c(y)\right\}\end{aligned}$$

where the canonical parameter is $\zeta = \log[\mu/(\mu+k)]$ and $b(\zeta) = -k\log(1-e^{\zeta}) = -k\,\log[k/(\mu+k)]$.

For stratified responses, with stratum means μ_h, the negative binomial mass function is

$$P(Y_h = y_h) = \frac{\Gamma(y_h+k)}{\Gamma(y_h+1)\Gamma(k)} \left(\frac{\mu_h}{\mu_h+k}\right)^{y_h} \left(\frac{k}{\mu_h+k}\right)^k. \tag{3.1.7}$$

and the variance for the response corresponding to the h^{th} fixed effect is $\operatorname{var} Y_h = \mu_h + k^{-1}\mu_h^2$.

3.2 Estimators for k

In this section we will derive the estimators described in **Chapter 2** specifically for the negative binomial SMM with a sample $\mathbf{y}$ consisting of n_h observations for each of $h = 1, \ldots, H$ strata, so that

the total size of $\mathbf{y}$ is $n = \sum_h n_h$. Note that in all the following cases, the corresponding estimator for μ_h is equal to the sample mean $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$ in the h^{th} effect. We will provide details on this in some cases.

3.2.1 ML for k

From equations 2.2.1 and 3.1.7, the log-likelihood function for μ_h and k is

$$\begin{aligned} l(\mu_1, \ldots, \mu_H, k|\mathbf{y}) &= \sum_{h,i} \log P(y_{hi} = y_{hi}|\mu_h, k) \\ &= \sum_{h,i} \left\{ y_{hi} \log \mu_h + k \log k - (y_{hi} + k) \log(\mu_h + k) + \log \frac{\Gamma(y_{hi} + k)}{\Gamma(y_h + 1)\Gamma(k)} \right\} \end{aligned} \tag{3.2.1}$$

and it has kernel

$$\sum_{h,i} \left\{ y_{hi} \log \mu_h + k \log k - (y_{hi} + k) \log(\mu_h + k) + \log \frac{\Gamma(y_{hi} + k)}{\Gamma(k)} \right\}.$$

The score for μ_h is

$$\frac{\partial l(\mu_h, k)}{\partial \mu_h} = \sum_{h,i} \left\{ \frac{y_{hi}}{\mu_h} - \frac{y_{hi} + k}{\mu_h + k} \right\} \tag{3.2.2}$$

and setting $\partial l / \partial \mu_h = 0$ yields $\hat{\mu}_h = \bar{y}_h$. The profile likelihood for k is thus

$$l^{(P)}(k|\bar{y}_1, \ldots, \bar{y}_H) = \sum_{h,i} \left\{ y_{hi} \log \bar{y}_h + k \log k - (y_{hi} + k) \log(\bar{y}_h + k) + \log \frac{\Gamma(y_{hi} + k)}{\Gamma(k)} \right\} \tag{3.2.3}$$

and the profile score for k is

$$\begin{aligned} \frac{\partial l^{(P)}(k)}{\partial k} &= \sum_{h,i} \left\{ \log k + 1 - \log(\bar{y}_h + k) - \frac{y_{hi} + k}{\bar{y}_h + k} + \Psi(y_{hi} + k) - \Psi(k) \right\} \\ &= \sum_h \left\{ n_h \log \frac{k}{\bar{y}_h + k} + n_h \right\} - \sum_h \frac{n_h(\bar{y}_h + k)}{\bar{y}_h + k} + \sum_{h,i} \left[\Psi(y_{hi} + k) - \Psi(k) \right] \\ &= \sum_h \left\{ n_h \log \frac{k}{\bar{y}_h + k} + n_h - n_h \right\} + \sum_{h,i} \left[\Psi(y_{hi} + k) - \Psi(k) \right] \\ &= \sum_h \left\{ n_h \log \frac{k}{\bar{y}_h + k} \right\} + \sum_{h,i} \left[\Psi(y_{hi} + k) - \Psi(k) \right]. \end{aligned} \tag{3.2.4}$$

Ψ denotes the digamma function, defined as

$$\Psi(x) = \frac{\partial}{\partial x} \log \Gamma(x)$$

which is usually implemented in numerical optimization packages. The digamma function can also be written in terms of elementary functions; notice that

$$\begin{aligned}\sum_{h,i}[\Psi(y_{hi}+k)-\Psi(k)] &= \sum_{h,i}\frac{\partial}{\partial k}\log\frac{\Gamma(y_{hi}+k)}{\Gamma(k)}\\ &= \sum_{h,i}\frac{\partial}{\partial k}\log\prod_{j=0}^{y_{hi}-1}(j+k)\\ &= \sum_{h,i}\frac{\partial}{\partial k}\sum_{j=0}^{y_{hi}-1}\log(j+k)\\ &= \sum_{h,i}\sum_{j=0}^{y_{hi}-1}(j+k)^{-1}\end{aligned}$$

which can be replaced in the profile score function in equation 3.2.4. The k-root of equation 3.2.4 is the maximum profile likelihood estimator of k, denoted $\hat{k}_{ml}$.

3.2.2 EQL for k

From equations 2.2.10 and 3.2.2, the deviance $D(\mu_h|k,y_{hi})$ for a single observation y_{hi} is

$$\begin{aligned}D(\mu_h|k,y_{hi}) &= -2\int_{y_{hi}}^{\mu_h}\frac{y_{hi}-u}{u+k^{-1}u^2}du\\ &= -2\left[y\log u-(y_{hi}+k)\log\frac{u+k}{k}\right]\Big|_{y_{hi}}^{\mu_h}\\ &= -2\left[y_{hi}\log\frac{\mu_h}{y_{hi}}-(y_{hi}+k)\log\frac{\mu_h+k}{y_{hi}+k}\right]\end{aligned}\tag{3.2.5}$$

and the extended quasi-likelihood function for μ_h and k is thus

$$q^+(\mu_1,\ldots,\mu_H,k) = -\frac{1}{2}\sum_{h,i}\left\{\log[2\pi V(y_{hi})]-2\left[y_{hi}\log\frac{\mu_h}{y_{hi}}-(y_{hi}+k)\log\frac{\mu_h+k}{y_{hi}+k}\right]\right\}\tag{3.2.6}$$

where $V(y_{hi}) = y_{hi}+k^{-1}y_{hi}^2$. As Y_{hi} can equal 0, this version is problematic for the negative binomial SMM since the function will not exist for any $y_{hi}=0$ in $\log[2\pi V(y_{hi})]$.

$V(y_{hi})$ arises in the function due to the nature of the EQL as a saddlepoint approximation. For the binomial, negative binomial, and Poisson distributions, the saddlepoint approximation has the form of the log-likelihood with Stirling's approximation

$$x! \approx \sqrt{2\pi x}x^x e^{-x}\tag{3.2.7}$$

in place of any $\Gamma(x+1)$.[1] In particular,

$$\begin{aligned}
\log \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} &\approx \log \frac{(y+k-1)!}{y!(k-1)!} \; \forall k \in \{1,2,\ldots\} \\
&= \log \frac{\sqrt{2\pi(y+k-1)}(y+k-1)^{y+k-1}e^{-(y+k-1)}}{\sqrt{2\pi(k-1)}(k-1)^{k-1}e^{-(k-1)}\sqrt{2\pi y}y^y e^{-y}} \\
&= \log \frac{\left[(y+k-1)^{2(y+k)-1}\right]^{1/2}}{\left[(k-1)^{2k-1}\right]^{1/2}(2\pi)^{1/2}\left[y^{2y+1}\right]^{1/2}} \\
&= \frac{1}{2}\log \frac{(y+k-1)^{2(y+k)}(k-1)}{(k-1)^{2k}(2\pi)y^{2y+1}(y+k-1)} \\
&= \frac{1}{2}\log \left(\frac{y+k-1}{y}\right)^{2y}\left(\frac{y+k-1}{k-1}\right)^{2k}\frac{(2\pi)^{-1}(k-1)}{y(y+k-1)} \\
&= \Big[(y+k)\log(y+k-1) - y\log y - k\log k\Big] \\
&\quad - \frac{1}{2}\log 2\pi - \frac{1}{2}\log(y+k-1) - \frac{1}{2}\log y + \frac{1}{2}\log(k-1)
\end{aligned}$$

and substituting this into $l(\mu, k)$ yields

$$\begin{aligned}
l(\mu,k) \approx & y\log\mu - y\log(\mu+k) + k\log k - k\log(\mu+k) + \Big[(y+k)\log(y+k-1) - y\log y - k\log k\Big] \\
& - \frac{1}{2}\log 2\pi - \frac{1}{2}\log(y+k-1) - \frac{1}{2}\log y + \frac{1}{2}\log(k-1) \\
\approx & \Big[y\log\frac{\mu}{y} - (y+k)\log\frac{\mu+k}{y+k}\Big] - \frac{1}{2}\log 2\pi - \frac{1}{2}\log(y+k) - \frac{1}{2}\log y + \frac{1}{2}\log(k) \\
= & -\frac{1}{2}\Big\{-2\Big[y\log\frac{\mu}{y} - (y+k)\log\frac{\mu+k}{y+k}\Big] + \log 2\pi\frac{y(y+k)}{k}\Big\} \\
= & -\frac{1}{2}\left\{-2D(\mu|k,y) + \log 2\pi V(y)\right\} \\
= & \; q^+(\mu,k)
\end{aligned}$$

as expected. However, this derivation is crude; the Stirling's approximation used in equation 3.2.7 is both poor for 0! and assumes integer k. A more appropriate approximation for the gamma function is [24]

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}}x^x e^{-x} \tag{3.2.8}$$

which is valid for real k and gives the approximation .922 for 0!. We can go a step further and use an even more exact Stirling's approximation; note that the *Stirling's series* is defined for the

[1] See Appendix A.

log-gamma function as

$$\log\Gamma(x) = \left(x - \frac{1}{2}\right)\log x + \frac{1}{2}\log 2\pi - x + \sum_{n=1}^{\infty} \frac{B_{2n}}{2n(2n-1)x^{2n-1}} \tag{3.2.9}$$

where B_n is the n^{th} Bernoulli number. The Stirling's approximation with one additional term is

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}} x^x e^{-x} e^{1/12x} \tag{3.2.10}$$

which gives the approximation 1.002 for 0!.[2] Using this Stirling's approximation, we have that

$$\begin{aligned}\log\frac{\Gamma(y+k)}{\Gamma(k)} &\approx \log\frac{k^{1/2}}{(y+k)^{1/2}}\frac{(y+k)^{y+k}}{k^k}\frac{e^{-(y+k)}}{e^{-k}}\frac{e^{1/12(y+k)}}{e^{1/12k}} \\ &\approx y\log(y+k) + \left(k-\frac{1}{2}\right)\log(y+k) - \left(k-\frac{1}{2}\right)\log k - \frac{y}{12k(y+k)}\end{aligned}$$

and so, substituting into the log-likelihood, the kernel of the improved EQL is

$$q^+(\mu,k) = \left[y\log\mu + (y+k)\log\frac{y+k}{\mu+k}\right] - \frac{1}{2}\log(y+k) + \frac{1}{2}\log k - \frac{y}{12k(y+k)}. \tag{3.2.11}$$

In terms of the SMM, and substituting $\bar{y}_h$ in for μ, we have the kernel

$$q^+_{(P)}(k|\bar{y}_h) = \sum_{h,i}\left\{\left[y_{hi}\log\bar{y}_h + (y_{hi}+k)\log\frac{y_{hi}+k}{\bar{y}_h+k}\right] - \frac{1}{2}\log(y_{hi}+k) + \frac{1}{2}\log k - \frac{y_{hi}}{12k(y_{hi}+k)}\right\} \tag{3.2.12}$$

the profile extended quasi-likelihood function for k. Differentiating with respect to k yields the profile extended quasi-score function

$$\frac{\partial q^+_{(P)}(k|\bar{y}_h)}{\partial k} = \sum_{h,i}\left\{\log\frac{y_{hi}+k}{\bar{y}_h+k} - \frac{1}{2(y_{hi}+k)} + \frac{1}{2k} - \frac{1}{12(y_{hi}+k)^2} + \frac{1}{12k^2}\right\} \tag{3.2.13}$$

and the k-root of equation 3.2.13 is the maximum profile extended quasi-likelihood estimator of k, denoted $\hat{k}_{eql}$.

3.2.3 DEQL for k

The HQGLM and DEQL construction is complex, but the DEQL function was recommended by Saha & Paul in [39] and Saha in [40] for estimating k in the negative binomial model. Cadigan & Tobin considered the DEQL function as given in [40] in [12], denoting the corresponding estimator

[2] The next term is $1/360x^3$ and provides the approximation 0.9995 for 0!, so the gain in accuracy drops off quickly.

for k by $\hat{k}_{deql}$. A closer analysis of the HQGLM construction, however, shows that it is *not useful* for estimating k, and we demonstrate this in this section.

Recall from subsection 2.2.3 that conjugate HGLMs and are specified using i) a distribution for the response, conditional on random effects, and ii) a distribution for the random effects taking the form of a conjugate prior. Using the hierarchical model developed in section 3.1, we have (using a slightly different parametrization)

$$Y|U \sim \text{Poisson}(\mu_0)$$

$$U \sim \text{gamma}(k, k^{-1})$$

where $\mu_0 = U\mu$ for the unobserved random variables U. Recall from equation 3.1.3 that the Poisson mass function can be written as

$$P(Y = y|\zeta) = \exp\{y\zeta - b[\zeta] + k(y)\}$$

where $\zeta = \log \mu_0$, $b[\zeta] = \exp\{\zeta\}$, and $k(y) = -\log y!$. Using this and $E[Y|U] = \mu_0$, $\text{var}[Y|U] = \mu_0$, we have a quasi-GLM for $Y|U$ with a canonical log link. The quasi-likelihood for a single observation y is thus

$$\begin{aligned} q_0(\mu_0; y|u) &= \int_y^{\mu_0} \frac{y-s}{V(s)} ds \\ &= [y \log s - s] \,|_y^{\mu_0} \\ &= y \log \frac{\mu_0}{y} - (\mu_0 - y) \end{aligned}$$

and $D_0(\mu_0) = -2q_0(\mu_0)$ is the deviance. The extended quasi-likelihood for a single observation y, adjusted[3] for the presence of 0 in the support of $Y|U$, is then

$$\begin{aligned} q_0^+(\mu_0|\, y|u) &= -\frac{1}{2}\{\log 2\pi V(y) + D_0(\mu_0)\} \\ &= -\frac{1}{2}\left\{\log 2\pi(y + 1/6) - 2\left[y \log \frac{\mu_0}{y} - (\mu_0 - y)\right]\right\}. \end{aligned} \tag{3.2.14}$$

For $U \sim$ gamma (k, k^{-1}), we have $E[U] = 1$ and $\text{var}\, U = kk^{-2} = k^{-1}V(u)$, where $V(u) = 1$. The quasi-data is the observation 1 with quasi-parameter u and thus $\zeta(u) = \log u = v$. We then

[3]Here we use the Stirling's approximation $y! \approx \sqrt{2\pi(y + 1/6)}y^y e^{-y}$ as y is integer-valued. This approximation yields 1.023 for 0! [32].

have a quasi-GLM for u with conjugate log link and conjugate variance function $V_1 = V_0$ so that the quasi-likelihood function for a single observation of quasi-data is

$$\begin{aligned} q_1(v, k|1) &= \int_1^u \frac{1-t}{k^{-1}V(t)}dt \\ &= \frac{1}{k^{-1}} \int_1^u (t^{-1} - 1)dt \\ &= \frac{1}{k^{-1}} \left[1 \log \frac{u}{1} - (u-1)\right] \\ &= \frac{1}{k^{-1}} \left[v - \exp\{v\} + 1\right] \end{aligned}$$

and $D_1(v, k) = -2q_1(v, k)$ is the deviance. The extended quasi-likelihood for a single observation of quasi-data is then

$$\begin{aligned} q_1^+(v, k|1) &= -\frac{1}{2}\left\{\log 2\pi k^{-1} V(u) + D_1(v, k)\right\} \\ &= -\frac{1}{2}\left\{\log 2\pi k^{-1} - 2k\left[v - \exp\{v\} + 1\right]\right\} \end{aligned} \tag{3.2.15}$$

and note that we do not use the adjustment of $V(u)$ at the origin since $V(u) = 1 > 0$. As this is a conjugate model no Jacobian adjustment is required, and thus the double extended quasi-likelihood function is

$$\begin{aligned} Q(\mu_0, k, v|y) &= q_0^+(\mu_0) + q_1^+(v, k) \\ &= -\log 2\pi - \frac{1}{2}\log(y + 1/6) + y\left[\log \frac{\mu_0}{y} + 1\right] - \mu_0 \\ &\quad + \frac{1}{2}\log k + k[v - e^v + 1]. \end{aligned}$$

For the SMM with $H > 1$ we have the relevant indices Y_{hi}, μ_h, and v_{hi}, and thus the DEQL function for the full model is

$$\begin{aligned} Q(\mu_{01}, \ldots, \mu_{0H}, k, v_{hi}|\mathbf{y}) = \sum_{h,i} \Big\{ &-\log 2\pi - \frac{1}{2}\log(y_{hi} + 1/6) + y_{hi}\left[\log \frac{\mu_{0h}}{y_{hi}} + 1\right] - \mu_{0h} \\ &+ \frac{1}{2}\log k + k[v_{hi} - e^{v_{hi}} + 1]\Big\} \end{aligned} \tag{3.2.16}$$

and we can write this in terms of μ_h and u_{hi} as

$$Q(\mu_1,\ldots,\mu_H,k,u_{hi}|\mathbf{y}) = \sum_{hi}\Big\{[-(\mu_h+k)u_{hi}+(y_{hi}+k)\log u_{hi}+y_{hi}\log\mu_h] \\ -\frac{1}{2}\log(y_{hi}+1/6)+(y_{hi}+k)+\frac{1}{2}\log k - y_{hi}\log y_{hi}-\log 2\pi\Big\}. \tag{3.2.17}$$

The DEQL estimator for μ_h is not immediately obvious. The DEQ-score for v_{hi} is

$$\begin{aligned}\frac{\partial Q(\mu_{0h},k,v_{hi})}{\partial v_{hi}} &= y_{hi}\frac{\partial}{\partial v_{hi}}\log\mu_{0h} - \frac{\partial}{\partial v_{hi}}\mu_{0h} + k[1-e^{v_{hi}}] \\ &= y_{hi}\frac{\partial}{\partial v_{hi}}\log\mu_h e^{v_{hi}} - \frac{\partial}{\partial v_{hi}}\mu_h e^{v_{hi}} + k[1-e^{v_{hi}}] \\ &= y_{hi} - \mu_h e^{v_{hi}} + k - k e^{v_{hi}} \\ &= [(y_{hi}+k) - e^{v_{hi}}(\mu_h+k)]\end{aligned}$$

so that, setting $\partial Q/\partial v_{hi}=0$, we get

$$\hat{v}_{hi} = \log\frac{y_{hi}+k}{\mu_h+k}$$

the maximum DEQL estimator for the random effects v_{hi}. The profile DEQL for μ_h is then

$$Q^{(P)}(\mu_h,k|\mathbf{y},\hat{v}_{hi}) = \sum_{h,i}\Big\{-\log 2\pi - \frac{1}{2}\log(y_{hi}+1/6)+\frac{1}{2}\log k - y_{hi}\log y_{hi} \\ + y_{hi}\log\mu_h + (y_{hi}+k)\log\frac{y_{hi}+k}{\mu_h+k}\Big\}$$

and so the profile DEQ-score for μ_h is

$$\frac{\partial Q^{(P)}(\mu_h,k)}{\partial\mu_h} = \sum_i\left\{\frac{y_{hi}}{\mu_h} - \frac{y_{hi}+k}{\mu_h+k}\right\}$$

so that setting $\partial Q^{(P)}/\partial\mu_h=0$ yields $\hat{\mu}_h^{(P)} = \bar{y}_{h\cdot}$ the maximum profile DEQL estimator for μ_h.

Now, we demonstrate that the HQGLM setup is unnecessary for the negative binomial model. Note that as we've used a full distributional assumption for $Y|U$ and U, we can specify the full h-likelihood in terms of u_{hi} by using the Poisson and gamma likelihoods as

$$\begin{aligned}h(\mu_h,k,u_{hi}) &= l_0(\mu_{0h}|u_{hi};y_{hi}) + l_1(k;u_{hi}) \\ &\propto \sum_{h,i}\{-(\mu_h+k)u_{hi}+(y_{hi}+k)\log u_{hi}+k\log k+y_{hi}\log\mu_h-\log\Gamma(y_{hi}+1)\Gamma(k)\}\end{aligned} \tag{3.2.18}$$

and by comparing with equation 3.2.17, the DEQL kernel is just the h-likelihood kernel using a Stirling's approximation for $\log\{k!\Gamma(y_{hi}+1)\}$ with integer k so that $D \approx h$. For this Poisson-gamma HGLM we can integrate out the random effects u_{hi} to achieve a marginal h-likelihood in terms of the random effects, rather than using a profile h-likelihood. Denoting this marginal h-likelihood by h_u, we have

$$\begin{aligned} h_u &= \log \int_0^\infty \exp\{h(\mu_h, k, u_{hi})\}\, du \\ &= \log \prod_{h,i} \frac{k^k \mu_h^{y_{hi}}}{\Gamma(y_{hi}+1)\Gamma(k)} \int_0^\infty e^{-(\mu_h+k)u_{hi}+(y_{hi}+k)\log u_{hi}} du_{hi} \\ &= \log \prod_{h,i} \frac{k^k \mu_h^{y_{hi}}}{\Gamma(y_{hi}+1)\Gamma(k)} \left[(\mu_h+k)^{-(y_{hi}+k)}\Gamma(y_{hi}+k)\right] \\ &= \sum_{h,i} \left\{ y_{hi} \log \frac{\mu_h}{\mu_h+k} + k \log \frac{k}{\mu_h+k} + \log \frac{\Gamma(y_{hi}+k)}{\Gamma(y_{hi}+1)\Gamma(k)} \right\}. \end{aligned}$$

That is, the marginal h-likelihood with respect to the random effects u_{hi} is the negative binomial log-likelihood $l(\mu_h, k)$, and so h cannot possibly contain any additional information about k. As Q approximates h and differs only by means of Stirling's approximations, the marginal DEQL function with respect to the random effects is analogously the EQL function of equation 3.2.12. Note that equation 3.2.12 is the kernel of the DEQL function given in [39] and [40].

3.2.4 AML for k

One benefit of the hierarchical construction is that, due to the parameterization used (recall $\operatorname{var} U = k^{-1}V(u)$) and the result mentioned in subsection 2.4.2, μ_h and k are orthogonal. Alternatively, Saha & Paul proved the orthogonality of μ and k analytically for the *iid* negative binomial case in [39]. The Barndorff-Nielsen modified profile likelihood and the Cox-Reid adjusted profile likelihood are therefore equivalent in the NB SMM. From equation 2.4.12 in subsection 2.4.2, the Cox-Reid adjustment to the profile likelihood is

$$M(k) = -\frac{1}{2} \log \left| j_{\mu_h \mu_h}(k, \hat{\mu}_{h|k}) \right|.$$

The score for μ_h is the gradient $(\partial l/\partial \mu_h)$, $h = 1, \ldots, H$ and it has elements

$$\begin{aligned}\frac{\partial l(\mu_h, k)}{\partial \mu_h} &= \sum_i \left\{ \frac{y_{hi}}{\mu_h} - \frac{y_{hi} + k}{\mu_h + k} \right\} \\ &= n_h \left[\frac{\bar{y}_h}{\mu_h} - \frac{\bar{y}_h + k}{\mu_h + k} \right].\end{aligned}$$

As $\partial^2 l/\partial \mu_h \mu_j = 0 \; \forall \; h \neq j$, the μ_h-μ_h block of the Hessian's only nonzero elements are the diagonals

$$\frac{\partial^2 l(\mu_h, k)}{\partial \mu_h^2} = -n_h \left[\frac{\bar{y}_h}{\mu_h^2} + \frac{\bar{y}_h + k}{(\mu_h + k)^2} \right] \tag{3.2.19}$$

so that the μ_h-μ_h block of the profile observed information matrix has diagonals

$$\begin{aligned}-\frac{\partial^2 l(\mu_h, k)}{\partial \mu_h^2}\Big|_{\mu_h = \bar{y}_h} &= -\frac{n_h \left[-(\bar{y}_h + k) + \bar{y}_h \right]}{\bar{y}_h (\bar{y}_h + k)} \\ &= \frac{n_h}{V(\bar{y}_h)}\end{aligned} \tag{3.2.20}$$

where $V(\bar{y}_h) = \bar{y}_h + k^{-1}\bar{y}_h$. Defining $\hat{v}(\bar{y}_h) = n_h^{-1} V(\bar{y}_h)$, we then have

$$M(k) = -\frac{1}{2} \log \left| j_{\mu_h \mu_h}(\hat{\mu}_{h|k}) \right| = \frac{1}{2} \sum_h \log \hat{v}(\bar{y}_h). \tag{3.2.21}$$

The Cox-Reid adjusted profile likelihood is then

$$\begin{aligned}l^{(A)}(k) &= l^{(P)}(k|\bar{y}_h) + M(k) \\ &= l^{(P)}(k|\bar{y}_h) + \frac{1}{2} \sum_h \log \hat{v}(\bar{y}_h).\end{aligned} \tag{3.2.22}$$

and it is worth noting that this is equivalent to the modified profile likelihood as derived in [12]. The adjusted profile score is

$$\frac{\partial l^{(A)}(k)}{\partial k} = \sum_h \left\{ n_h \log \frac{k}{\bar{y}_h + k} - \frac{\bar{y}_h}{2k(\bar{y}_h + k)} \right\} + \sum_{h,i} \left[\Psi(y_{hi} + k) - \Psi(k) \right] \tag{3.2.23}$$

and the maximum adjusted profile likelihood estimator for k, $\hat{k}_{aml}$ is the k-root of equation 3.2.23.

3.2.5 CREQL for k

As mentioned in subsection 2.4.2, Lee & Nelder applied the Cox-Reid adjustment to an EQL function in [26] in order to estimate dispersion parameters in a joint quasi-GLM model. The extended quasi-score for μ_h is

$$\frac{\partial q^+(\mu_h, k)}{\partial \mu_h} = \sum_i \left\{ \frac{y_{hi}}{\mu_h} - \frac{y_{hi} + k}{\mu_h + k} \right\}$$

which is identical to the score function for the likelihood as given in 3.2.2. The Cox-Reid adjustment for the extended quasi-likelihood function is identical to that of the likelihood function, and we write the Cox-Reid adjusted extended quasi-likelihood function as

$$\begin{aligned} q^+_{(A)}(k) &= q^+_{(P)}(k) + M(k) \\ &= q^+_{(P)}(k) + \frac{1}{2}\sum_h \log \hat{v}(\bar{y}_h) \end{aligned} \tag{3.2.24}$$

with the Cox-Reid adjusted profile quasi-score

$$\frac{\partial q^+_{(A)}(k)}{\partial k} = \sum_{h,i}\left\{\log\frac{y_{hi}+k}{\bar{y}_h+k} - \frac{1}{2(y_{hi}+k)} + \frac{1}{2k} - \frac{1}{12(y_{hi}+k)^2} + \frac{1}{12k^2}\right\} - \sum_h\left\{\frac{\bar{y}_h}{2k(\bar{y}_h+k)}\right\}. \tag{3.2.25}$$

The Cox-Reid adjusted EQL estimator for k, denoted $\hat{k}_{CReql}$, is the k-root of equation 3.2.25.

3.2.6 LNEQL for k

From subsection 2.4.3, the Lee-Nelder adjustment is achieved by finding the diagonals of the quasi-likelihood hat matrix

$$\mathbf{\Lambda} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{1/2}.$$

For the NB SMM it is straightforward to show that $(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1} = \text{diag}\{[n_jV(\mu_j)]^{-1}\}$ and thus $\mathbf{\Lambda} = \text{diag}\{n_j^{-1}\}$ so that $\kappa_{hi} = n_h^{-1}$. The deviance adjustment is thus $1/(1-n_h^{-1}) = n_h/(n_h-1)$, and applying it to the profile extended quasi-score function yields

$$\frac{\partial q^+_{(B)}(k)}{\partial k} = \sum_{h,i}\left\{\frac{n_h}{n_h-1}\left[\log\frac{y_{hi}+k}{\bar{y}_h+k}\right] - \frac{1}{2(y_{hi}+k)} + \frac{1}{2k} - \frac{1}{12(y_{hi}+k)^2} + \frac{1}{12k^2}\right\}. \tag{3.2.26}$$

The Lee-Nelder adjusted EQL estimator for k, denoted $\hat{k}_{LNeql}$, is the k-root of equation 3.2.26.

3.2.7 Comments

We seek to estimate k in the presence of the nuisance mean parameters μ_h in the negative binomial model, and have amassed a collection of estimators to do so. Based on the discussion in **Chapter 1** and **Chapter 2**, we can expect the maximum profile likelihood estimator to perform poorly in situations with small sample sizes, particularly when $H > n_h$ on average in the SMM sense. As a

highly accurate saddlepoint approximation to the likelihood, we can expect the profile EQL function to perform similarly, and mentioned rationale for this in section 2.3.

Since (μ_h, k) are orthogonal, the adjusted profile likelihood of Cox & Reid is equivalent to the modified profile likelihood of Barndorff-Nielsen for the NB SMM. The modified profile likelihood is a saddlepoint-based approximation to the marginal or conditional likelihoods of equations 2.4.1 and 2.4.2, and we discussed the expected improvement in estimator performance in subsection 2.4.1. The Cox-Reid adjusted EQL function uses a combination of saddlepoint approximations for the orthogonal case; the EQL approximates the likelihood, and then the Cox-Reid adjustment approximates the marginal likelihood. The use of saddlepoint approximations suggest that this estimator will also perform well.

The Lee & Nelder adjustment, as a model-specific adjustment to the deviance in a quasi-GLM, may not perform as well. As discussed in section subsection 2.4.2, the Cox-Reid adjustment for the (ψ, θ) model with scalar θ is, in general, a Laplace approximation to the marginal likelihood. Since the Lee-Nelder adjustment is approximating a lower-order Laplace approximation, it may not perform as well as the Cox-Reid adjusted estimators.

3.3 Comparison of estimators

3.3.1 Experimental design & methodology

To empirically compare estimator performance we used the experimental design outlined by Cadigan & Tobin in [12]. This design contains a range of stratifications in order to compare performance in the presence of nuisance parameters when both $n_h < H$ and $n_h > H$. The full factorial design has factors: i) sample size per stratum, $n_h = 2$ or 10; ii) number of strata, $H = 5$ or 51; iii) average stratum mean $\mu = H^{-1} \sum_h \mu_h$, $\mu = 5$ or 50, and iv) $k = 0.5$, 1, or 5. Individual strata means were set uniformly over the interval $\mu \pm \mu/2$. The combinations of H and n_h determine the degree of 'stratification' in the design; when $n_h = 2$ and $H = 5$ there are 10 observations for five strata (or a strata-to-observations proportion of 50%), and when $n_h = 10$ and $H = 5$ there are 50 observations for five strata (a proportion of 10%). These proportions remain the same for $H = 51$. We refer to the combinations including $n_h = 10$ as having 'low stratification' and those including $n_h = 2$ as having 'high stratification'.

Data were simulated using the *rnegbin()* function in the R package *MASS*. One thousand data sets were generated for each of the 24 factor combinations in the design, and various estimates of k (summarized below) were computed for each simulated data set. Results were summarized using the bias and mean squared error from the simulations. All estimates were performed using the appropriate score and R's *uniroot()* function which searches an interval for a root. We defined the left and right endpoints of the interval to be 0.00001 and $10k$ respectively; if the signs of the score function evaluated at these endpoints were the same, then a value of NA was returned for the estimate. The proportion of estimators receiving a value of NA were recorded so as to measure the average reliability of each estimator.

Cadigan & Tobin found that their adjusted double extended quasi-likelihood, or ADEQL, estimator performed well in the class of estimators they compared [12]. In lieu of the discussion in subsection 3.2.3, this estimator is actually a variant of the EQL function. They noted the estimating equation $\partial q^+ / \partial k = 0$ can be written as

$$\sum_{h,i} \left\{ 2k \log \frac{\bar{y}_h + k}{y_{hi} + k} + \frac{k}{y_{hi} + k} - \frac{y_{hi}(y_{hi} + 2k)}{6k(y_{hi} + k)^2} \right\} - n = 0$$

and, with motivation stemming from the method of moments estimator as discussed in [12], they proposed the adjusted estimating equation

$$\sum_{h,i} \left\{ 2k \log \frac{\bar{y}_h + k}{y_{hi} + k} + \frac{k}{y_{hi} + k} - \frac{y_{hi}(y_{hi} + 2k)}{6k(y_{hi} + k)^2} \right\} - (n - H) = 0 \tag{3.3.1}$$

and the Cadigan-Tobin adjusted EQL estimator, $\hat{k}_{CTeql}$, solves equation 3.3.1. This is not an asymptotic method, but for purposes of comparison to their empircal work, we include it in our simulations here. Our collection of estimators is then the ML, EQL, AML, CREQL, LNEQL, and CTEQL estimators.

3.3.2 Simulations and results

We ranked the estimators according to their absolute bias, mean squared error, and proportion converging in 1000 data sets (see Tables 3.5 and 3.6) and then averaged those ranks across estimators in order to achieve an overall performance ranking.

The results of the simulations conform well to what we expect theoretically, based on the discussion in **Chapter 2**. They clearly separated into two groups; the maximum profile likelihood

estimator $\hat{k}_{ml}$ and maximum profile EQL estimator $\hat{k}_{eql}$ performed almost identically, while the maximum AML, CREQL, LNEQL, and CTEQL estimators also had similar performance.

A good way to visualize the results of these simulations is via conditional inference trees.[4] We display conditional inference trees for absolute percentage bias and MSE in Figures 3.1 and 3.2. The conditional inference tree estimates a regression relationship by binary recursive partitioning; from top to bottom, the conditional inference tree selects the input with strongest association to the average absolute percentage bias, implements a binary split of that input, and then recursively repeats until all significant inputs have been exhausted.

Figure 3.1 shows that n_h, which determines the degree of stratification, has the strongest association with bias. For the highly-stratified case with $n_h = 2$, the estimator (i.e. AML, EQL, etc.) has the next highest association. Notice that the estimators clump into obvious groups based on the approximation to the marginal likelihood. In relatively unstratified conditions where $n_h = 10$, the estimator choice is not significantly associated with bias. Similarly, referring to Figure 3.2, n_h has the highest association to MSE. However, in both high and low stratifications, k is the next most highly associated variable. When $n_h = 2$, the estimators clump into the same two groups based on the approximation to the marginal likelihood for low and medium values of k.

$\hat{k}_{ml}$ and $\hat{k}_{eql}$ performed exceptionally poorly for highly-stratified models. As expected, when $n_h < H$ the two estimators were notably biased, with biases from 105% to upwards of 431% for large k. What is interesting is that, in the highly-stratified cases, the biases did not seem to get significantly worse as the number of strata increased for low and medium levels of k. If $n_h < H$, it did not seem to make much of a difference if H was 5 or 51, as evidenced by the absolute bias performance of $\hat{k}_{ml}$ and $\hat{k}_{eql}$ for the $k = 0.5$ or 1 and $H = 5$ or 51 factor combinations (see Table 3.2). This is consistent with the results of the two-index asymptotic analysis in section 2.3, which states that the bias is of order $O_p(n_h^{-1})$ and thus depends only on n_h if $n_h^{-1}H = O(1)$. Since the EQL function approximates the NB likelihood, $\hat{k}_{eql}$ does not escape the nuisance parameter problem. For increasing levels of k the AML and CREQL estimators performed more similarly, and when $k = 5$ they performed identically as measured by MSE. $\hat{k}_{ml}$ and $\hat{k}_{eql}$ had poor convergence in the highly-stratified case when k was large, failing to converge upwards of 68% of the time when $H = 51$.

On the other hand, the AML, CREQL, and LNEQL functions approximate the marginal likeli-

[4] The *ctree()* function in the R package 'party'.

hood and thus performed considerably better. Again, the results were consistent with theory. Due to the accuracy of EQL as an approximation to the likelihood, $\hat{k}_{aml}$ and $\hat{k}_{CReql}$ performed almost identically. $\hat{k}_{aml}$ and $\hat{k}_{CReql}$ both outperformed $\hat{k}_{LNeql}$ on average (see Table 3.5), although the LNEQL estimator was less biased for some individual factor combinations when k was high. As the Lee-Nelder adjustment itself approximates the Cox-Reid adjustment, this makes intuitive sense. $\hat{k}_{LNeql}$ had a substantially higher average MSE than either the AML or CREQL estimators, but it was far better than the average MSEs of $\hat{k}_{ml}$ and $\hat{k}_{eql}$.

The Cadigan-Tobin EQL estimator ranked best in terms of average absolute bias, but ranked slightly poorer than either $\hat{k}_{aml}$ or $\hat{k}_{CReql}$ in terms of average MSE. One point of note however is that Cadigan & Tobin found that the CTEQL estimator outperformed the AML estimator in their study, while here the AML performed best. We attribute this difference to the difference in methodology used in both studies; if an estimator failed to converge in [12], it was automatically assigned the maximum bound on the estimate of $10k$. In this study we record an NA for nonconverging estimates, which we feel is a more appropriate procedure. Cadigan & Tobin also found that their AEQL estimator - functionally equivalent to our LNEQL estimator, but it uses a cruder Stirling's approximation in the derivation - outperformed $\hat{k}_{aml}$ in their study. As the LNEQL estimator is 'doubly' approximating the adjusted profile likelihood, this result does not make intuitive sense. In this study our results seem to match up with intuition; the AML, which closely approximates the marginal likelihood, performs best, while further approximations (CREQL, LNEQL), perform slightly worse. Our overall conclusions are the same as in [12]; when the sample size per stratum is small, the estimators that do not adjust for the presence of nuisance parameters perform poorly. This practicum adds the observation that the estimators that better approximate the marginal likelihood perform better on average; of these estimators we find the intuitive ordinal ranking of AML, CREQL, LNEQL. The Cadigan-Tobin estimator $\hat{k}_{CTeql}$ also performs extremely well, ranking second overall in this study.

In the next chapter we apply these methods to a highly stratified negative binomial model for real data and discuss confidence intervals that require an estimate of k. The theoretical results of **Chapter 2** and the simulation study in this chapter suggest that the MLE for k may perform poorly for this model, which has implications for the accuracy of the confidence intervals we discuss.

k	H	n_h	μ	ML	AML	EQL	CREQL	LNEQL	CTEQL
0.5	5	2	5	234.30	71.57	237.23	75.51	76.33	33.31
0.5	5	2	51	162.44	54.67	164.88	58.03	43.19	50.15
0.5	5	10	5	10.15	6.02	15.73	11.69	9.51	7.81
0.5	5	10	51	6.53	4.10	10.80	8.44	5.96	6.49
0.5	51	2	5	234.15	56.41	236.26	60.22	32.19	9.06
0.5	51	2	51	110.33	39.67	112.60	43.01	8.49	17.87
0.5	51	10	5	7.56	3.78	13.15	9.44	7.15	5.63
0.5	51	10	51	4.61	2.32	8.92	6.69	4.17	4.71
1	5	2	5	229.21	12.57	229.95	13.92	60.55	5.76
1	5	2	51	169.05	13.45	169.55	14.43	53.70	54.35
1	5	10	5	10.25	2.18	11.75	3.75	4.86	2.48
1	5	10	51	7.17	2.25	8.15	3.28	3.08	3.41
1	51	2	5	249.52	5.50	249.96	6.81	17.83	-14.77
1	51	2	51	105.39	4.27	105.82	5.25	1.79	7.49
1	51	10	5	7.51	0.15	9.00	1.71	2.38	0.27
1	51	10	51	4.92	0.31	5.90	1.34	0.96	1.30
5	5	2	5	169.40	-67.62	169.43	-67.47	32.71	-55.35
5	5	2	51	189.13	-64.17	189.13	-64.09	58.15	30.77
5	5	10	5	29.62	-23.45	29.66	-23.39	14.04	-2.72
5	5	10	51	8.62	-15.92	8.64	-15.90	2.66	2.03
5	51	2	5	431.07	-67.94	431.08	-67.80	34.29	-59.60
5	51	2	51	140.25	-64.42	140.25	-64.34	4.74	-5.69
5	51	10	5	13.45	-25.84	13.50	-25.78	1.81	-9.32
5	51	10	51	5.73	-17.19	5.75	-17.17	-0.01	-0.55

Table 3.1: Average percentage bias by estimator and factor combination. Factor combinations are listed in the leftmost columns. Factors and levels are, in order from left to right: $k = 0.5$, 1, or 5; $H = 5$ or 51; $n_h = 2$ or 10; and $\mu = 5$ or 50.

k	H	n_h	μ	ML	AML	EQL	CREQL	LNEQL	CTEQL
0.5	5	2	5	235.58	77.23	237.93	79.26	96.04	58.62
0.5	5	2	51	164.90	61.60	166.56	62.99	66.00	67.07
0.5	5	10	5	16.80	14.62	19.10	16.18	15.26	14.40
0.5	5	10	51	12.10	11.11	13.51	12.00	11.08	11.27
0.5	51	2	5	234.15	56.41	236.26	60.22	33.61	15.13
0.5	51	2	51	110.33	39.68	112.60	43.02	14.79	19.89
0.5	51	10	5	8.01	5.31	13.16	9.53	7.49	6.29
0.5	51	10	51	5.25	3.80	8.97	6.83	4.77	5.17
1	5	2	5	232.27	29.63	232.65	29.47	88.75	45.36
1	5	2	51	172.12	28.65	172.36	28.57	79.97	75.35
1	5	10	5	17.59	13.97	17.90	13.82	15.05	14.29
1	5	10	51	12.57	10.53	12.79	10.51	10.95	10.98
1	51	2	5	249.52	10.19	249.96	10.55	26.05	17.84
1	51	2	51	105.39	9.10	105.82	9.36	15.06	15.35
1	51	10	5	8.14	4.38	9.36	4.49	4.96	4.46
1	51	10	51	5.52	3.25	6.28	3.36	3.40	3.48
5	5	2	5	184.36	67.62	184.35	67.47	89.52	55.45
5	5	2	51	192.49	64.17	192.49	64.09	88.03	62.95
5	5	10	5	40.68	24.81	40.68	24.76	31.23	22.44
5	5	10	51	14.75	16.59	14.75	16.58	12.87	12.66
5	51	2	5	431.07	67.94	431.08	67.80	53.31	59.60
5	51	2	51	140.25	64.42	140.25	64.34	18.51	15.24
5	51	10	5	14.53	25.84	14.56	25.78	7.96	10.32
5	51	10	51	6.41	17.19	6.42	17.17	3.86	3.85

Table 3.2: Average absolute percentage bias by estimator and factor combination. Factor combinations are listed in the leftmost columns. Factors and levels are, in order from left to right: $k = 0.5$, 1, or 5; $H = 5$ or 51; $n_h = 2$ or 10; and $\mu = 5$ or 50.

k	H	n_h	μ	ML	AML	EQL	CREQL	LNEQL	CTEQL
0.5	5	2	5	4.94	0.51	4.98	0.53	1.45	0.46
0.5	5	2	51	2.79	0.35	2.81	0.36	0.69	0.68
0.5	5	10	5	0.02	0.02	0.03	0.02	0.02	0.02
0.5	5	10	51	0.01	0.01	0.02	0.01	0.01	0.01
0.5	51	2	5	3.22	0.18	3.26	0.20	0.09	0.02
0.5	51	2	51	0.68	0.09	0.70	0.11	0.02	0.03
0.5	51	10	5	0.00	0.00	0.01	0.01	0.00	0.00
0.5	51	10	51	0.00	0.00	0.00	0.00	0.00	0.00
1	5	2	5	10.21	0.14	10.23	0.14	2.31	0.43
1	5	2	51	6.11	0.13	6.12	0.13	2.23	1.82
1	5	10	5	0.06	0.03	0.06	0.03	0.04	0.03
1	5	10	51	0.03	0.02	0.03	0.02	0.02	0.02
1	51	2	5	7.84	0.02	7.85	0.02	0.14	0.04
1	51	2	51	1.30	0.01	1.31	0.01	0.04	0.04
1	51	10	5	0.01	0.00	0.01	0.00	0.00	0.00
1	51	10	51	0.00	0.00	0.01	0.00	0.00	0.00
5	5	2	5	38.02	2.30	38.02	2.29	11.66	1.72
5	5	2	51	37.57	2.07	37.57	2.06	12.16	5.23
5	5	10	5	2.66	0.39	2.66	0.39	1.34	0.47
5	5	10	51	0.19	0.18	0.19	0.18	0.13	0.13
5	51	2	5	117.20	2.31	117.20	2.30	5.08	1.79
5	51	2	51	12.08	2.08	12.08	2.07	0.31	0.18
5	51	10	5	0.16	0.34	0.16	0.34	0.05	0.07
5	51	10	51	0.03	0.15	0.03	0.15	0.01	0.01

Table 3.3: Average mean squared error (MSE) by estimator and factor combination. Factor combinations are listed in the leftmost columns. Factors and levels are, in order from left to right: $k = 0.5$, 1, or 5; $H = 5$ or 51; $n_h = 2$ or 10; and $\mu = 5$ or 50.

k	H	n_h	μ	ML	AML	EQL	CREQL	LNEQL	CTEQL
0.5	5	2	5	0.2		0.2		0.07	
0.5	5	2	51	0.04		0.04		0.01	
0.5	5	10	5						
0.5	5	10	51						
0.5	51	2	5						
0.5	51	2	51						
0.5	51	10	5						
0.5	51	10	51						
1	5	2	5	0.28		0.28		0.09	
1	5	2	51	0.06		0.06		0.02	
1	5	10	5						
1	5	10	51						
1	51	2	5	0.01		0.01			
1	51	2	51						
1	51	10	5						
1	51	10	51						
5	5	2	5	0.63		0.63		0.29	
5	5	2	51	0.14		0.14		0.04	
5	5	10	5						
5	5	10	51						
5	51	2	5	0.68		0.68			
5	51	2	51						
5	51	10	5						
5	51	10	51						

Table 3.4: Proportion of estimators failing to converge in 1000 data sets, by estimator and factor combination. Zero proportions omitted for readability. Factor combinations are listed in the leftmost columns. Factors and levels are, in order from left to right: $k = 0.5$, 1, or 5; $H = 5$ or 51; $n_h = 2$ or 10; and $\mu = 5$ or 50.

	ML	AML	EQL	CREQL	LNEQL	CTEQL
Avg. absolute percentage bias	109.00	30.00	110.00	31.00	33.00	26.00
Avg. MSE	10.21	0.47	10.22	0.47	1.58	0.55
Avg. proportion non-converging	0.09	0.00	0.09	0.00	0.02	0.00

Table 3.5: Average perfomance measures across all factor combinations, by estimator.

	ML	AML	EQL	CREQL	LNEQL	CTEQL
Avg. absolute percentage bias	5	2	6	3	4	1
Avg. MSE	5	1	6	2	4	3
Avg. proportion non-converging	5.5	2	5.5	2	4	2
Overall rank	5	1	6	3	4	2

Table 3.6: Ranks of estimator by criterion. Overall rank is calculated as the ranked average of all other ranks.

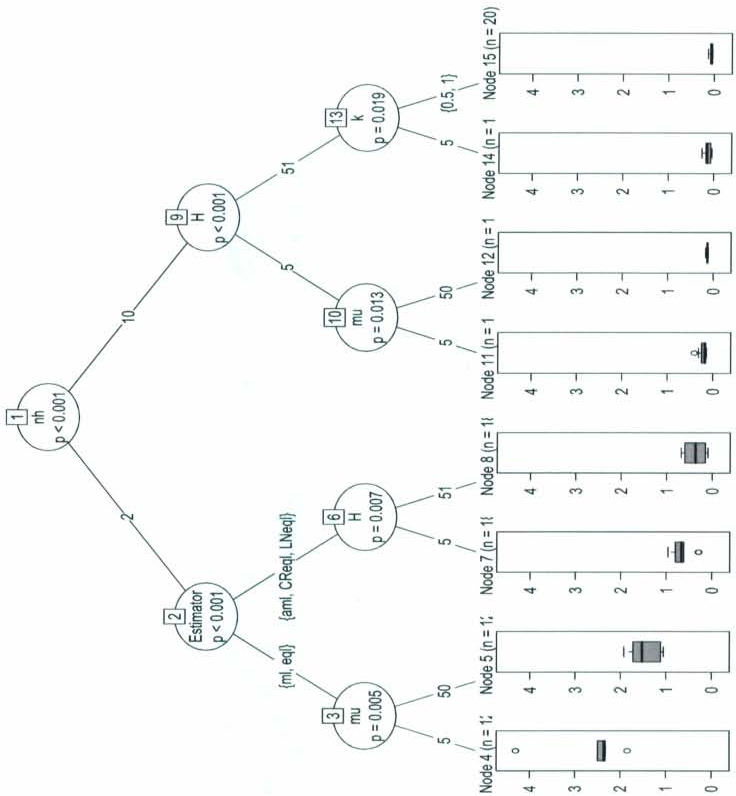

Figure 3.1: Conditional inference tree for average absolute percentage bias.

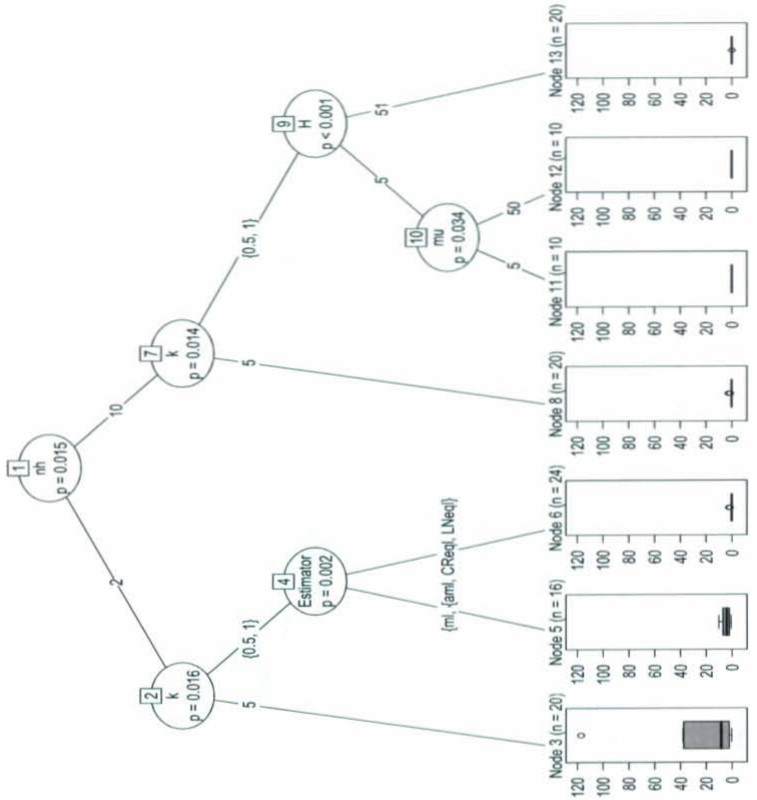

Figure 3.2: Conditional inference tree for mean squared error.

Chapter 4

Application to 3Ps Atlantic cod data

4.1 Background

The negative binomial SMM has been used in fisheries research and stock assessment [20, 11, 12, 13]. In Newfoundland & Labrador the stock of Atlantic cod (*Gadus morhua*, see 4.1, commonly referred to as 'northern cod') in NAFO[1] subdivision 3Ps, off the south coast of the island of Newfoundland (see Figure 4.2), currently supports the largest cod fishery off eastern Canada.

In this chapter we use the negative binomial SMM to model catches of 3Ps Atlantic cod in research trawls. We briefly describe a method for constructing negative binomial confidence intervals based on the SMM, and then use the adjusted profile likelihood estimator $\hat{k}_{amt}$ to construct the confidence intervals for inference on mean trawlable abundance.

4.1.1 Survey design and abundance estimation

The Canadian Department of Fisheries & Oceans (DFO) conducts two research (trawl) surveys per year in the waters off the coast of Newfoundland & Labrador. The 'Spring survey' runs between

[1] North Atlantic Fisheries Organization

April and June and covers divisions 3LNOP[2], so that subdivision 3Ps is only surveyed once per year. Survey data are particularly important for subdivision 3Ps; it currently has the highest total allowable catches (TACs) of any NAFO division, and survey results are scrutenized closely by both fishery managers and stakeholders.

Surveys follow a stratified random sampling scheme; each NAFO division is divided into a certain number of strata that are largely determined by ocean depth (see Figure 4.3). Stratified simple random sampling is used to determine sampling locations at approximately the same time each year. An observation from the sample consists of the number of cod caught in one standardized research trawl tow, in which a specific fishing gear is towed by a research vessel at a constant speed over a fixed distance. The sampling unit is the area over the bottom covered by one standardized tow, and a minimum of two tows are made from every index stratum in the design.

Of primary importance in stock assessment is the quantification of abundance, the total number of fish in a stock, or equivalently, the average fish density over the stock area. Abundance estimators based solely on the stratified sampling design (called *design-based* estimators) are thought not to be sufficient for estimating abundance in trawl surveys [13]. One important problem with the design-based approach is due to the *measurement error* inherent in trawl sampling, in that not all fish available for catching at a tow site are actually caught (due to net avoidance, trawl mesh size, etc.). In the presence of measurement error, design-based estimators - which do not take into account the fraction of fish that tends to be caught - would be imprecise even if all tow sites were actually sampled. A probabilistic model and thus a *model-based* estimator for *trawlable* abundance are desirable.

4.1.2 A probabilistic model for trawl catches

Let the stock area be divided into N equally-sized sampling units with the j^{th} unit, $j = 1, \ldots, N$ containing λ_j fish, so that the total number of fish in the stock is $\lambda = \sum_{j=1}^{N} \lambda_j$. Recall that each sampling unit corresponds to the area over the bottom covered by a standardized tow with a fixed speed, distance, and equivalently, duration. For any such tow (with duration T) corresponding to the j^{th} sample unit, the number of fish that have arrived at the trawl at time t, $t \leq T$ is described

[2]This abbreviation is used to indicate divisions 3L, 3N, 3O, and 3P.

Figure 4.1: *Gadus Morhua.* Photo: Hans-Petter Fjeld (CC-BY-SA)

by the Poisson process

$$\{N^j(t); 0 < t \leq T, \lambda_j/T\} \tag{4.1.1}$$

with rate λ_j/T. Due to the mechanisms resulting in measurement error, not all of these fish are actually caught; the unknown fraction of λ_j that is actually caught is commonly denoted as q, which is typically assumed to be constant throughout sampling units[3]. There are only two outcomes for fish arriving in the trawl - they may be successfully caught, or they may escape. Thinning the arrival process in equation 4.1.1 by incorporating q yields the independent Poisson processes

$$\{N_1^j(t); 0 < t \leq T, q\lambda_j/T\} \tag{4.1.2}$$

$$\{N_2^j(t); 0 < t \leq T, (1-q)\lambda_j/T\} \tag{4.1.3}$$

for the number of fish caught and escaped at time t respectively. Our model for trawlable catches is the catch process $\{N_1^j(t)\}$; the *expected catch* in the j^{th} unit is

$$\begin{aligned} \mu_j &= \int_0^T \frac{q\lambda_j}{T} dt \\ &= q\lambda_j \end{aligned}$$

[3]In practice q is often dependent on the age and length of fish arriving in the trawl, but we will not consider this possibility here.

The *average trawlable stock size* is

$$\mu = \frac{1}{N}\sum_j \mu_j$$
$$= q\lambda$$

and μ is typically the measure of trawlable abundance that is of interest.

DFO research surveys use stratified simple random sampling, and so the stratified sample mean is used to estimate μ. As per the notation we have used throughout the practicum we have H strata, but with N_h possible sampling units in the h^{th} stratum so that we have a total of $\sum_h N_h = N$ possible sampling units. Let $W_h = N_h/N$ be the weight, or *proportional size*, of the h^{th} stratum. A random sample of n_h units is selected for each stratum and, as before, we have $\sum_h n_h = n$ as the size of the sample. The sample consists of the n observations $\{Y_{hi};\ h = 1, \ldots, H,\ i = 1, \ldots, n_h\}$, and we define the stratum sample mean $\bar{y}_h$ and stratum sample variance s_h^2 in obvious fashion as estimators for the true stratum mean μ_h and true stratum variance σ_h^2. The *stratified sample mean* is

$$\hat{\mu} = \sum_h W_h \bar{y}_h. \tag{4.1.4}$$

and it is unbiased for μ without needing any assumption about Y_{hi} [15].

In practice, extra-Poisson variation exists between tows for a variety of reasons; there is often local & random variation in stock densities, random variation in trawl catchability due to trawl configuration, random variation in ocean conditions (i.e. currents), etc. that cause overdispersion and thus skewed catch distributions relative to the idealized Poisson model. The negative binomial distribution is a natural choice for modelling Y_{hi}, so we have $Y_{hi} \sim \text{negbin}(\mu_h, k)$ as per the NB SMM.

4.2 Inference about μ

4.2.1 Design-based inference

The design-based variance for $\hat{\mu}$ is [15]

$$\text{var}_D(\hat{\mu}) = \frac{1}{N}\sum_h W_h \left(\frac{N_h - n_h}{n_h}\right)\sigma_h^2 \tag{4.2.1}$$

and an estimator based on the stratum sample variance is

$$\hat{\text{var}}_D(\hat{\mu}) = \frac{1}{N}\sum_h W_h\left(\frac{N_h - n_h}{n_h}\right)s_h^2 \tag{4.2.2}$$

which we will denote $v_{st}(\hat{\mu})$. We have that $\hat{\mu}$ is normally distributed by the central limit theorem, and since $E[\hat{\mu}] = \mu$, the statistic

$$T = \frac{\hat{\mu} - \mu}{\sqrt{\text{var}_D(\hat{\mu})}} \sim N(0,1) \tag{4.2.3}$$

for sufficiently large sample sizes. If we could calculate $\text{var}_D(\hat{\mu})$ exactly then a $(1-\alpha)\%$ confidence interval for μ can be calculated as

$$\begin{aligned} 1-\alpha &= P\left(-z_{\alpha/2} < T < z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2} < \frac{\hat{\mu} - \mu}{\sqrt{\text{var}_D\left(\hat{\mu}\right)}} < z_{\alpha/2}\right) \\ &= P\left(\hat{\mu} - z_{\alpha/2}\sqrt{\text{var}_D\left(\hat{\mu}\right)} < \mu < \hat{\mu} + z_{\alpha/2}\sqrt{\text{var}_D\left(\hat{\mu}\right)}\right) \end{aligned}$$

so that $\hat{\mu} \pm z_{\alpha/2}\sqrt{\text{var}_D\left(\hat{\mu}\right)}$ would contain μ in $(1-\alpha)\%$ of samples, where $P(T > z_{\alpha/2}) = \alpha/2$. In practice, we cannot calculate $\text{var}_D(\hat{\mu})$ exactly. We can however estimate var_D by $v_{st}(\hat{\mu})$, and use the statistic

$$T^* = \frac{\hat{\mu} - \mu}{\sqrt{v_{st}(\hat{\mu})}}. \tag{4.2.4}$$

Notice that since $T \xrightarrow{D} N(0,1)$ and $R = \hat{\text{var}}_D(\hat{\mu})/\text{var}_D(\hat{\mu}) \xrightarrow{P} 1$, we have

$$T^* = \frac{T}{R} \xrightarrow{D} N(0,1). \tag{4.2.5}$$

For practical sample sizes, T^* better follows a t-distribution with ν degrees of freedom[4]. This yields the 'standard' design-based $(1-\alpha)\%$ confidence interval for μ,

$$\hat{\mu} \pm t_{\nu,\alpha/2}\sqrt{v_{st}(\hat{\mu})} \tag{4.2.6}$$

which is the interval estimator used in current DFO stock assessment reports (SARs) [1].

[4]Satterthwaite's approximation is used in practice for the degrees of freedom calculation. That is,

$$\nu = \frac{\sum_h (g_h s_h^2)^2}{\sum_h g_h^2 s_h^4/(n_h - 1)}$$

where $g_h = N_h(N_h - n_h)/n_h$. [15]

The problem with using T^* as a basis for inference is that it may not approximate a t-distribution well in practice due to the highly skewed nature of survey catch distributions. Standard t-statistic confidence intervals may have poor coverage or include unfeasible negative values [13]. In the presence of measurement error, $\text{var}_D(\hat{\mu})$ also underestimates the total variance and may further distort confidence interval coverage, although this is usually not an important problem when the sampling fraction is small (as it is in fisheries surveys) [13].

4.2.2 Model-based inference

Incorporating the negative binomial model for catches, we have $Y_{hi} \sim \text{negbin}(\mu_h, k)$ and can directly calculate $\text{var}\,(\hat{\mu})$ as follows:

$$\begin{aligned}
\text{var}\,(\hat{\mu}) &= \text{var}\left(\sum_h W_h \bar{y}_h\right) \\
&= \sum_h W_h^2 \text{var}\left(\frac{1}{n_h}\sum_{i=1}^{n_h} y_{h_i}\right) \\
&= \sum_h \frac{W_h^2}{n_h^2}\sum_i \left(\mu_h + k^{-1}\mu_h^2\right) \\
&= \sum_h \frac{W_h^2}{n_h^2} n_h \left(\mu_h + k^{-1}\mu_h^2\right) \\
&= \sum_h \frac{W_h^2}{n_h} V(\mu_h)
\end{aligned}$$

where $V(\mu_h)$ denotes the variance function $\mu_h + k^{-1}\mu_h^2$. An estimator for $\text{var}\,(\hat{\mu})$ is

$$\hat{\text{var}}\,(\hat{\mu}) = \sum_h \frac{W_h^2}{n_h} \frac{n_h \hat{k}}{n_h \hat{k} + 1}\left(\bar{y}_h + \hat{k}^{-1}\bar{y}_h{}^2\right) \tag{4.2.7}$$

for an estimate $\hat{k}$ of k. The term $n_h\hat{k}/(n_h\hat{k}+1)$ corrects for bias; if we were to simply substitute $\bar{y}_h$ in for μ_h to estimate $V(\mu_h)$ then $V(\bar{y}_h)$ has expectation

$$\begin{aligned} E[V(\bar{y}_h)] &= \mu_h + \hat{k}^{-1}E[\bar{Y}_h^2] \\ &= \mu_h + \hat{k}^{-1}\mu_h^2 + \left[\frac{\mu_h + \hat{k}^{-1}\mu_h^2}{n_h\hat{k}}\right] \\ &= V(\mu_h) + n_h\hat{k}V(\mu_h) \\ &= \left[\frac{n_h\hat{k}+1}{n_h\hat{k}}\right]V(\mu_h) \end{aligned}$$

and so $[n_h\hat{k}/(n_h\hat{k}+1)]V(\bar{y}_h)$ is unbiased for $V(\mu_h)$. If k were known exactly, then $\hat{\text{var}}\,(\hat{\mu})$ would be unbiased for $\text{var}\,(\hat{\mu})$.

The estimator $\hat{var}(\hat{\mu})$ accounts for both modes of variation present - that due to the sample design, and that due to measurement error - and can be calculated from the available survey data [13]. We could then use the statistic

$$T_{NB}^{*} = \frac{\hat{\mu}-\mu}{\sqrt{\hat{\text{var}}\,(\hat{\mu})}}$$

to derive confidence intervals for μ, but due to the skewed nature of trawl survey data, this may perform no better than T^{*}.

Cadigan proposed a method in [13] for calculating negative binomial confidence intervals that can partially account for the presence of skewness. If we have approximately proportional allocation in the sample so that $n_h/N_h \approx n/N$, then

$$\begin{aligned} \text{var}\,(\hat{\mu}) &= \sum_h \frac{W_h^2}{n_h}\left(\mu_h + k^{-1}\mu_h^2\right) \\ &= \sum_h \left(\frac{N_h}{N}\right)^2 \frac{1}{n_h}\left(\mu_h + k^{-1}\mu_h^2\right) \\ &\approx \sum_h \left(\frac{N_h}{N}\frac{N}{n}\right)\frac{1}{N}\left(\mu_h + k^{-1}\mu_h^2\right) \\ &= \frac{1}{n}\sum_h W_h\left(\mu_h + k^{-1}\mu_h^2\right) \\ &= \frac{1}{n}\left[\sum_h W_h\mu_h + k^{-1}\sum_h W_h\mu_h^2\right] \end{aligned}$$

and since $(\mu_h - \mu)^2 = \mu_h^2 - 2\mu_h\mu + \mu^2$, we can replace μ_h^2 in the above expression by $(\mu_h - \mu)^2 + 2\mu_h\mu - \mu^2$ so that

$$\begin{aligned}
\text{var}\,(\hat{\mu}) &\approx \frac{1}{n}\left[\mu + k^{-1}\left(\sum_h W_h(\mu_h - \mu)^2 + 2\mu\sum_h W_h\mu_h - \sum_h W_h\mu^2\right)\right] \\
&= \frac{1}{n}\left[\mu + k^{-1}\left(\sum_h W_h(\mu_h - \mu)^2 + \mu^2\right)\right] \\
&= \frac{1}{n}\left(\mu + k^{-1}\mu^2\right) + \frac{k^{-1}}{n}\sum_h W_h(\mu_h - \mu)^2.
\end{aligned}$$

If $\sum_h W_h(\mu_h - \mu)^2$ is proportional to μ^2 such that

$$\sum_h W_h(\mu_h - \mu)^2 \approx \beta\mu^2$$

for some $\beta > 0$, then, finally,

$$\begin{aligned}
\text{var}\,(\hat{\mu}) &\approx \frac{1}{n}(\mu + k^{-1}\mu^2) + \frac{k^{-1}}{n}\beta\mu^2 \\
&= \frac{1}{n}\left[\mu + k^{-1}\mu^2(1+\beta)\right] \\
&= \frac{1}{n}\left[\mu + k_p^{-1}\mu^2\right]
\end{aligned}$$

where $k_p = k/(1+\beta)$. Cadigan noted that empirical evidence seemed to suggest between-strata variation in μ_h was frequently proportional to μ^2, with $\beta > 1$ in practice. Thus, the major mode of variation in $\hat{\mu}$ is the term $n^{-1}(\mu + k_p^{-1}\mu^2)$, and we can use this approximation to estimate confidence intervals for μ. We define

$$Z_{NB} = \frac{\hat{\mu} - \mu}{n^{-1/2}\sqrt{\mu + \hat{k}_p^{-1}\mu^2}} \sim N(0,1) \tag{4.2.8}$$

and use it to estimate confidence intervals similarly. Confidence intervals for μ based on Z_{NB} are sometimes called score intervals, and Cadigan noted that these were more accurate than those based on T^* for iid Poisson and negative binomial data [13]. Confidence intervals based on Z_{NB} also do not cover negative values, while this can be a problem for those based on T^* (see Figure 4.14). Intervals based on Z_{NB} are also asymmetric and better reflect the skewed nature of trawl survey data. Cadigan noted that symmetric intervals, such as those based on T^*, can have poor one-sided coverage properties. Using Z_{NB}, a $(1-2\alpha)\%$ confidence interval for μ is

$$\frac{a}{2} \pm \sqrt{\frac{a^2}{4} + b}$$

where

$$a = \frac{2\hat{\mu} + z_\alpha^2/n}{1 - z_\alpha^2/(n\hat{k}_p)}$$

and

$$b = \frac{-\hat{\mu}^2}{1 - z_\alpha^2/(n\hat{k}_p)}.$$

Cadigan gave a moment-based estimator for k_p as [13]

$$\hat{k}_p^{-1} = \frac{n\hat{\text{var}}(\hat{\mu}) - \hat{\mu}}{\hat{\mu}^2 - \hat{\text{var}}(\hat{\mu})}$$

and it requires an estimate of k in $\text{var}(\hat{\mu})$. Here, we can use an appropriately adjusted estimator as discussed in **Chapter 3**, for example, the adjusted profile likelihood estimator $\hat{k}_{aml}$.

4.3 Estimates of trawlable abundance for 3Ps Atlantic cod, 1996-2007

We used DFO research data from 1996-2007 to estimate both point and interval estimates for trawlable abundance of Atlantic cod by year. Note that we do not use data for 2006 as the survey was not completed that year. Survey tow locations and catches by year are plotted in Figures 4.4, 4.5, and 4.6, and summary statistics for the 3Ps data are shown in Table 4.1. The data are heavily stratified; most years have 45 strata, while the maximum number of observations sampled from any given stratum is between 11 and 13. In almost all years, over 50% of the strata sampled contained only two observations.

Stratum sample means are plotted against stratum sample variances across years in Figures 4.7, 4.8, and 4.9. The same plots on a log scale are shown in Figures 4.10, 4.11, and 4.12. Since $\text{var}\, Y_{hi}$ is $O(\mu_h^2)$, the relationship on the log scale should be an approximate line segment with slope equal to 2. The quadratic relationship is clearly visible, so the negative binomial specification seems appropriate.

We calculated normal, t, and negative binomial 95% confidence intervals for mean trawlable abundance by year. The calculated values are shown in Table 4.4, and detailed figures required for their calculations are shown in Table 4.2. Figures for these intervals can also be found at the end of this chapter. The normal intervals, being symmetric about $\hat{\mu}$, tend to be relatively conservative in

the lower endpoints (i.e., lower endpoints were small) while being relatively restrictive in the upper endpoints (i.e., upper endpoints were small). The t intervals behave similarly, but are even more conservative than the normal counterparts. The t intervals also included negative lower endpoints, which is impossible for the survey index. The negative binomial intervals used $\hat{k}_{aml}$ and were strictly positive. They were usually more conservative on upper endpoints than the other intervals, although the t upper endpoints occasionally exceeded them.

We also calculated another set of 95% negative binomial intervals using $\hat{k}_{ml}$. A table of these values is found in Table 4.5, and a plot is found in Figure 4.15. The intervals calculated using $\hat{k}_{ml}$ had smaller upper endpoints compared to those calculated using $\hat{k}_{aml}$, and the results of this practicum indicate that $\hat{k}_{aml}$ is likely the more accurate estimator in this scenario.

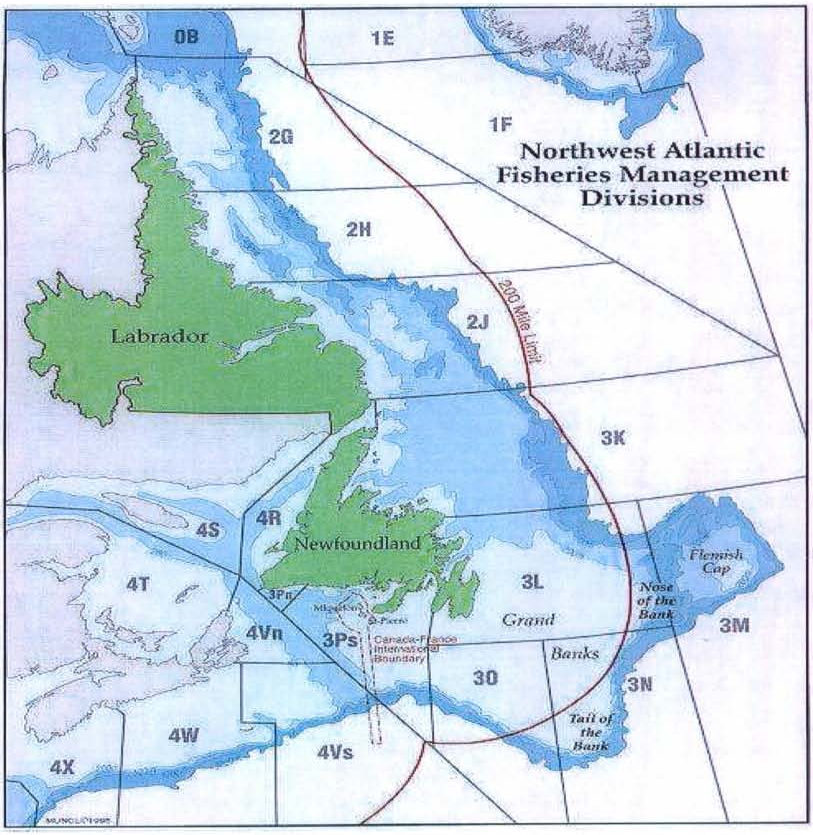

Figure 4.2: NAFO Divisions. Divisions 3LNOP are covered in DFO's Spring survey. Division 3P is divided into subdivisions 3Ps and 3Pn, both visible off Newfoundland's south coast.

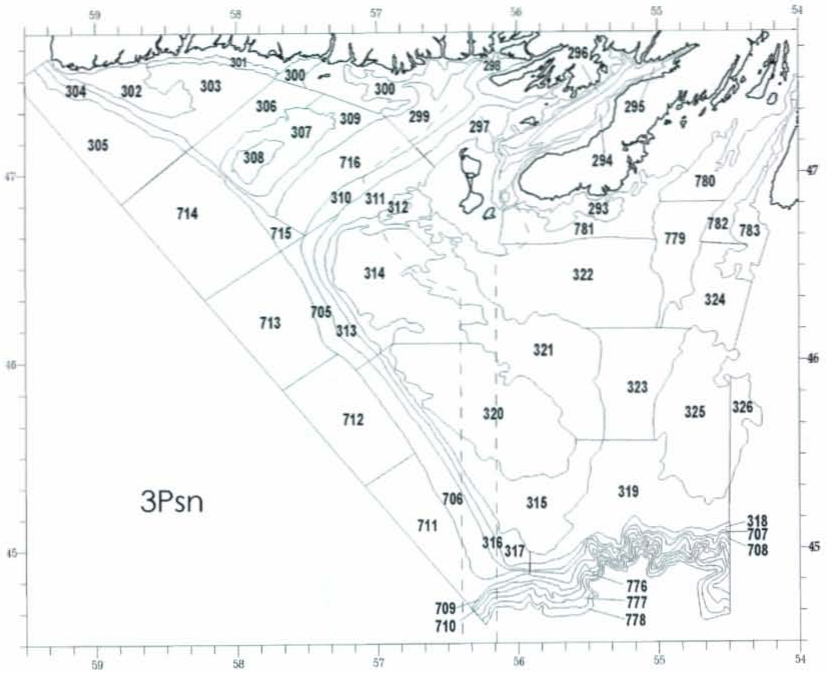

Figure 4.3: NAFO division 3P, with numbers indicating strata. Light grey lines indicate the strata borders, which are largely based on ocean depth. The variety of shapes and sizes of strata is evident; some are quite large (i.e. 322, 714) while many others are smaller. Note the many long, skinny strata occurring at the edge of the continental shelf.

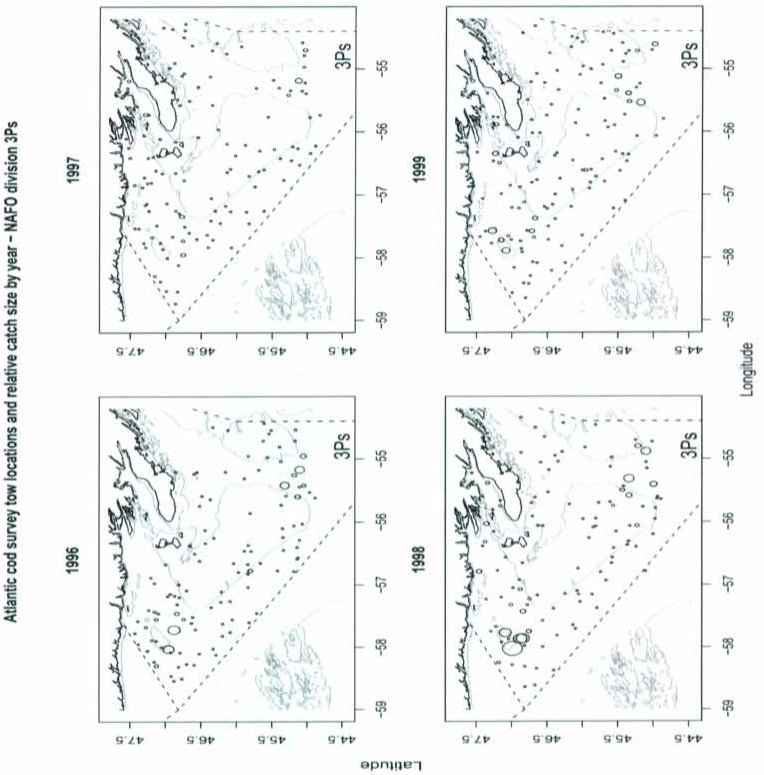

Figure 4.4: 3Ps survey catch locations for Atlantic cod, 1996-1999. Bubbles indicate a tow location, and the size of the bubble indicates the relative size of the catch.

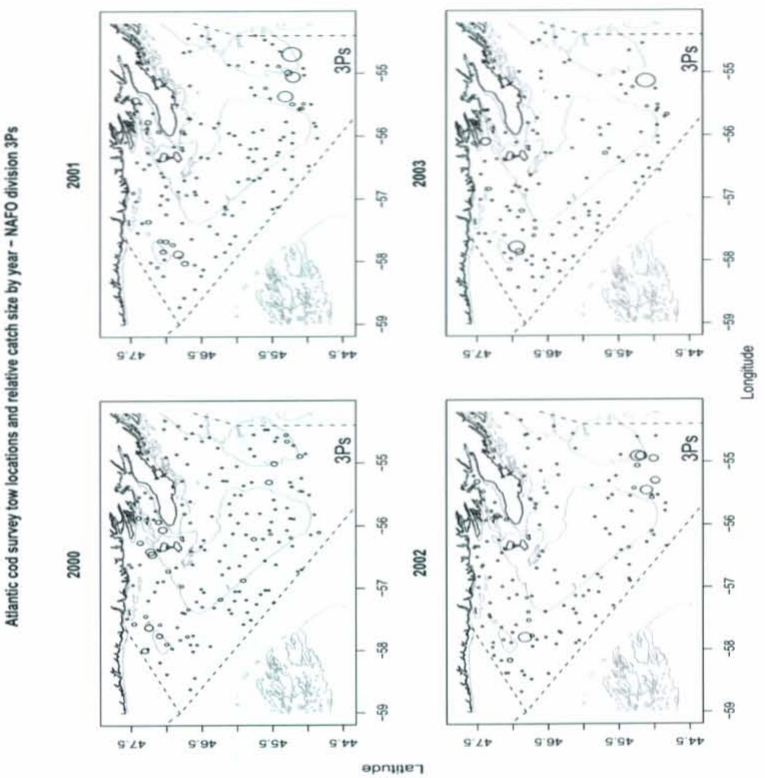

Figure 4.5: 3Ps survey catch locations for Atlantic cod, 2000-2003.

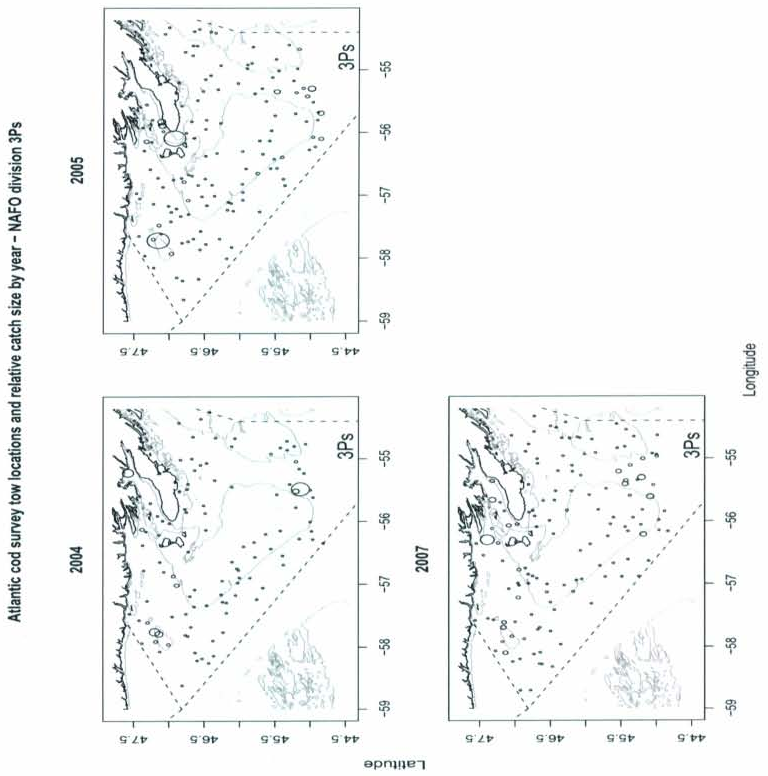

Figure 4.6: 3Ps survey catch locations for Atlantic cod, 2004-2007, excluding 2006 because the survey was not completed that year.

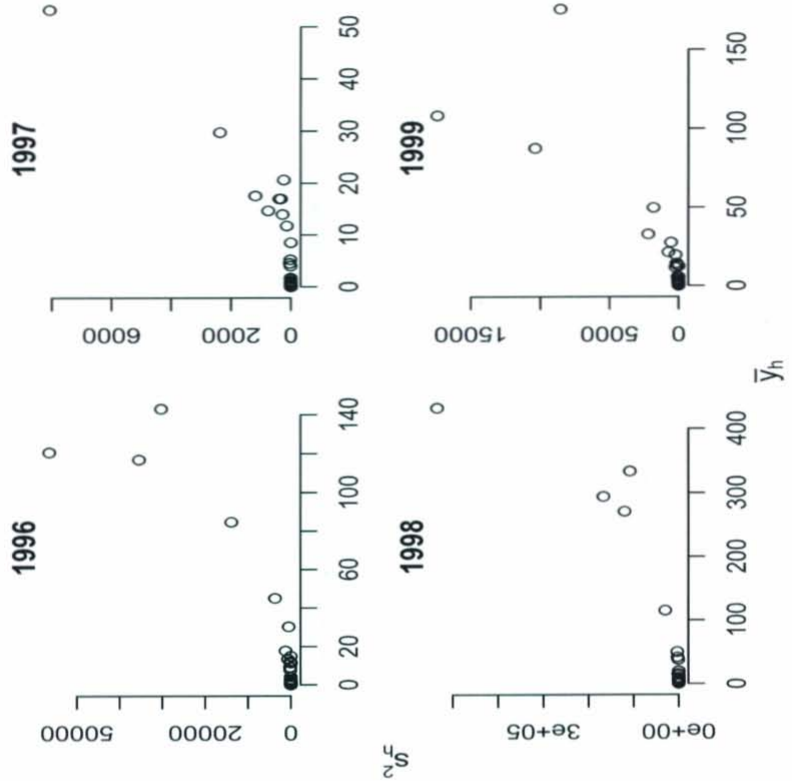

Figure 4.7: 3Ps strata sample means plotted against strata sample variances, 1996-1999. Note the approximate quadratic relationship, indicating that the negative binomial variance $\mu_h + k^{-1}\mu_h^2$ is appropriate.

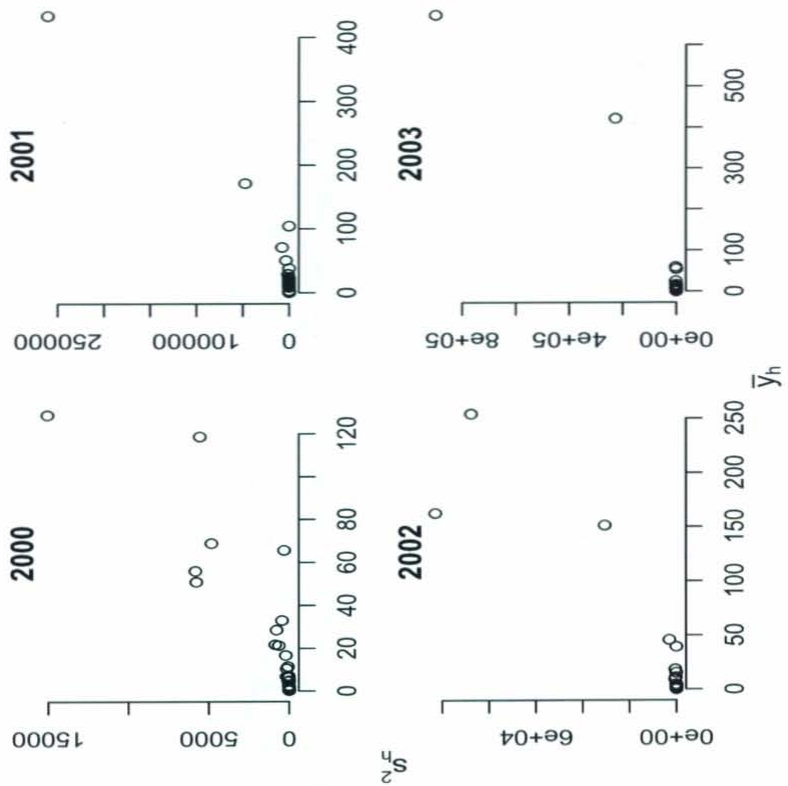

Figure 4.8: 3Ps strata sample means plotted against strata sample variances, 2000-2003.

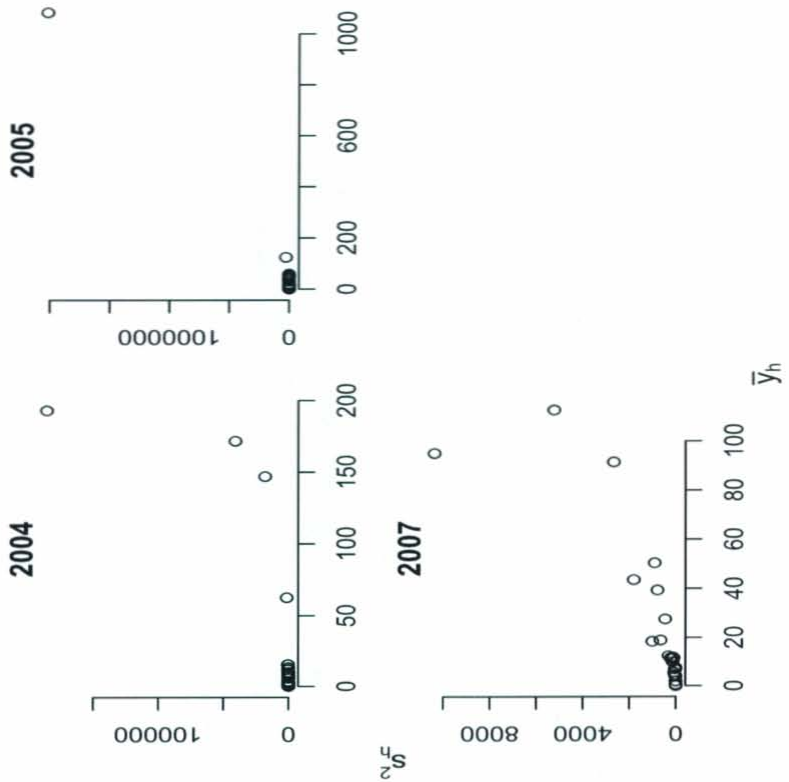

Figure 4.9: 3Ps strata sample means plotted against strata sample variances, 2004-2007.

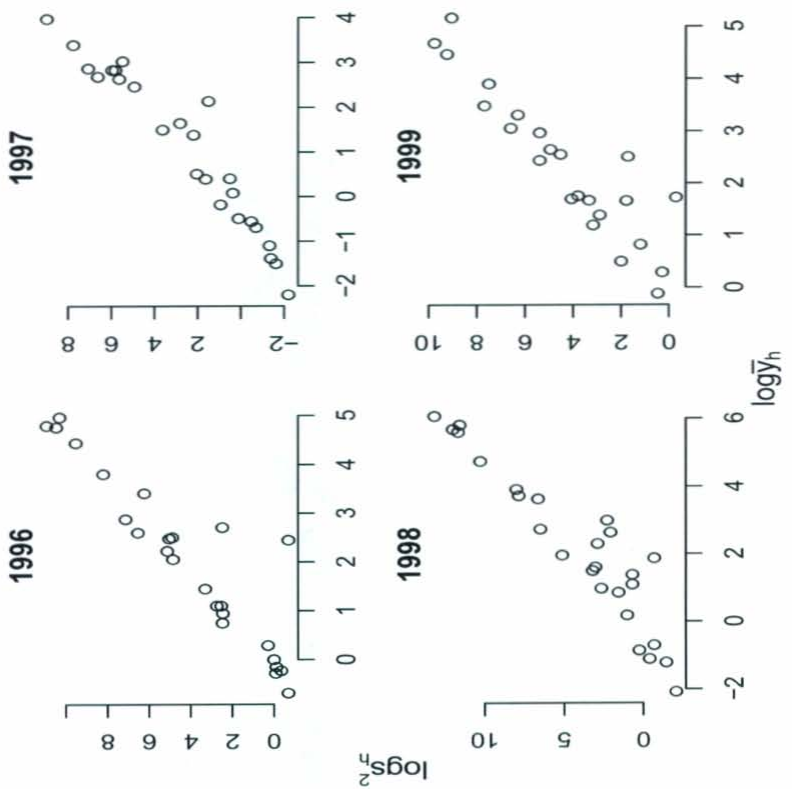

Figure 4.10: 3Ps log strata sample means plotted against log strata sample variances, 1996-1999. Note the linear relationship on the log scale, more clearly illustrating the quadratic relationship.

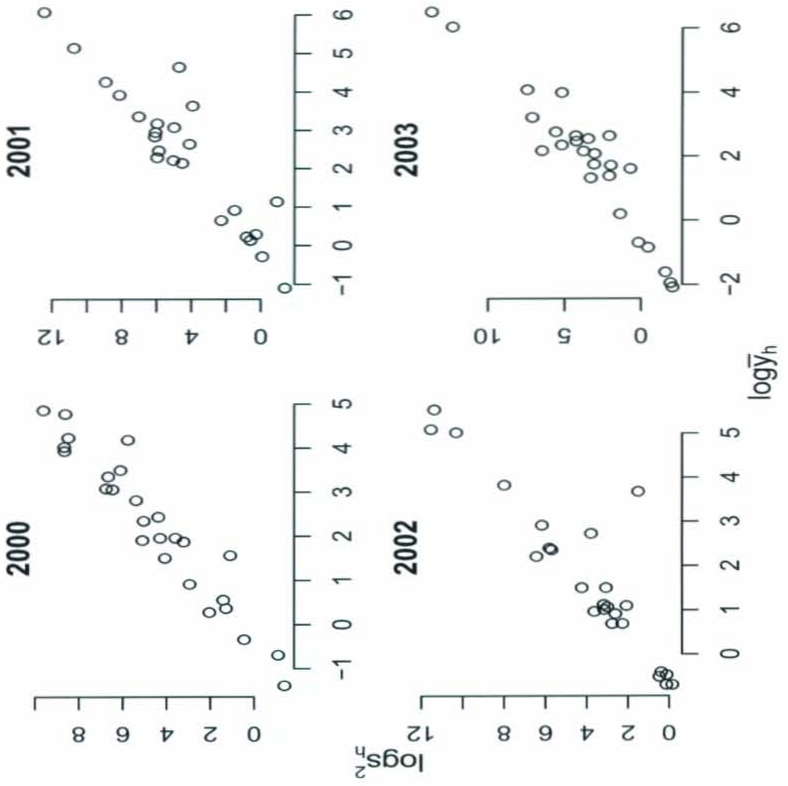

Figure 4.11: 3Ps log strata sample means plotted against log strata sample variances, 2000-2003.

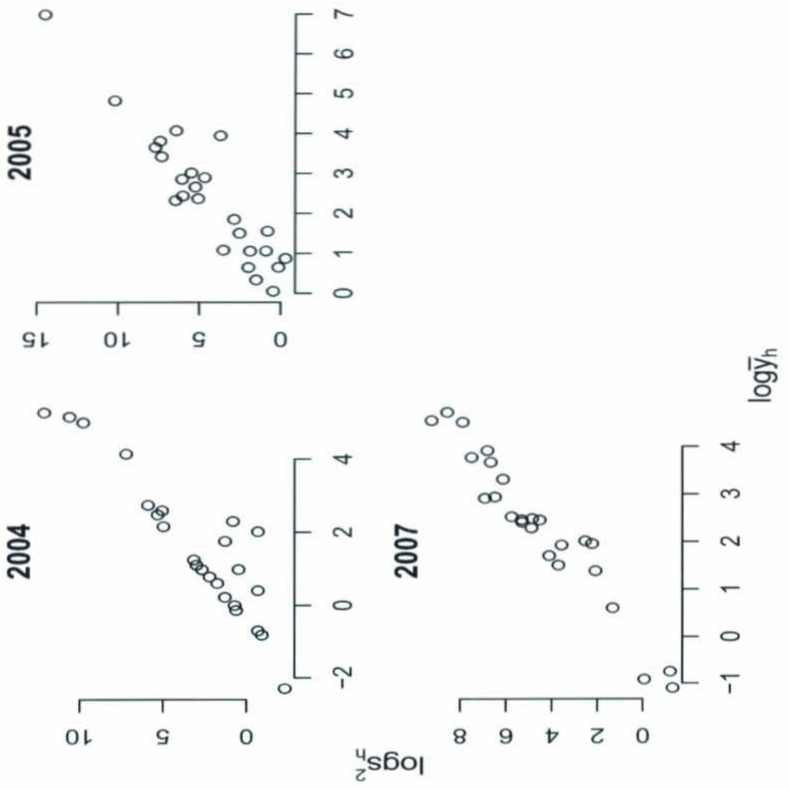

Figure 4.12: 3Ps log strata sample means plotted against log strata sample variances, 2004-2007.

Survey year	n	H	$\bar{y}$	s^2	CV	PH_2	$\max_h(n_h)$
1996	148	34	22.00	71.40	10.50	38	11
1997	157	44	9.00	21.50	5.70	54	11
1998	177	45	32.00	126.60	15.70	51	13
1999	175	46	17.70	41.90	5.60	54	12
2000	171	45	24.50	52.50	4.60	56	11
2001	173	45	35.50	141.10	15.80	53	12
2002	177	45	25.40	97.30	14.60	53	13
2003	176	45	24.90	128.40	26.60	53	13
2004	177	45	21.00	108.10	26.50	53	13
2005	178	45	38.70	218.90	32.00	53	13
2007	178	45	21.70	58.30	7.30	53	13

Table 4.1: Summary statistics for the 3Ps Atlantic cod survey data, 1996-2007. $\bar{y}$ and s^2 refer to the overall sample mean and variance respectively, and CV is the coefficient of variation. PH_2 is the percent of instances of $n_h = 2$ for $h = 1, \ldots, H$ out of H. $\max_h(n_h)$ is the maximum value of n_h across H strata.

Survey year	$\hat{\mu}$	$\hat{\text{var}}_D(\hat{\mu})$	ν	$\hat{k}_{aml}$	$\hat{\text{var}}(\hat{\mu})$	$\hat{k}_p$	a	b
1996	16.21	17.29	13.64	0.38	14.37	0.13	38.71	-311.06
1997	4.62	0.69	18.28	0.48	0.68	0.25	10.10	-23.30
1998	41.95	235.65	3.78	0.48	119.11	0.12	107.81	-2188.31
1999	21.07	15.00	10.42	0.81	11.76	0.34	45.74	-478.61
2000	23.66	11.46	10.46	0.62	14.16	0.28	51.15	-601.53
2001	44.77	186.42	7.33	0.75	155.61	0.11	120.72	-2593.45
2002	27.31	74.79	9.47	0.32	109.30	0.04	105.74	-1390.88
2003	25.34	112.38	3.02	0.59	78.78	0.06	84.38	-1031.12
2004	25.00	142.71	8.26	0.55	50.08	0.10	66.86	-816.66
2005	21.85	76.25	1.09	0.69	39.61	0.10	59.00	-631.21
2007	20.44	6.75	13.46	0.83	12.39	0.29	44.82	-454.86

Table 4.2: Table of values for the t and negative binomial confidence interval calculations. ν are the degrees of freedom via Satterthwaite's approximation.

Survey year	$\hat{\mu}$	$\hat{\text{var}}_D(\hat{\mu})$	$\hat{k}_{ml}$	$\hat{\text{var}}^*(\hat{\mu})$	$\hat{k}_p^*$	a^*	b^*
1996	16.21	17.29	0.41	13.48	0.13	38.23	-307.34
1997	4.62	0.69	0.56	0.62	0.28	10.01	-23.09
1998	41.95	235.65	0.55	109.67	0.13	105.36	-2144.07
1999	21.07	15.00	1.01	9.86	0.40	45.10	-472.42
2000	23.66	11.46	0.71	12.76	0.32	50.73	-596.97
2001	44.77	186.42	0.93	129.29	0.13	113.83	-2462.27
2002	27.31	74.79	0.34	102.98	0.04	99.72	-1314.49
2003	25.34	112.38	0.73	68.90	0.07	77.41	-950.11
2004	25.00	142.71	0.66	43.82	0.11	64.02	-784.20
2005	21.85	76.25	0.84	35.12	0.11	56.62	-607.14
2007	20.44	6.75	1.07	10.06	0.36	44.00	-447.11

Table 4.3: Table of values for the alternative negative binomial confidence interval calculations. The * superscripts indicate that they are calculated using the maximum likelihood estimator for k.

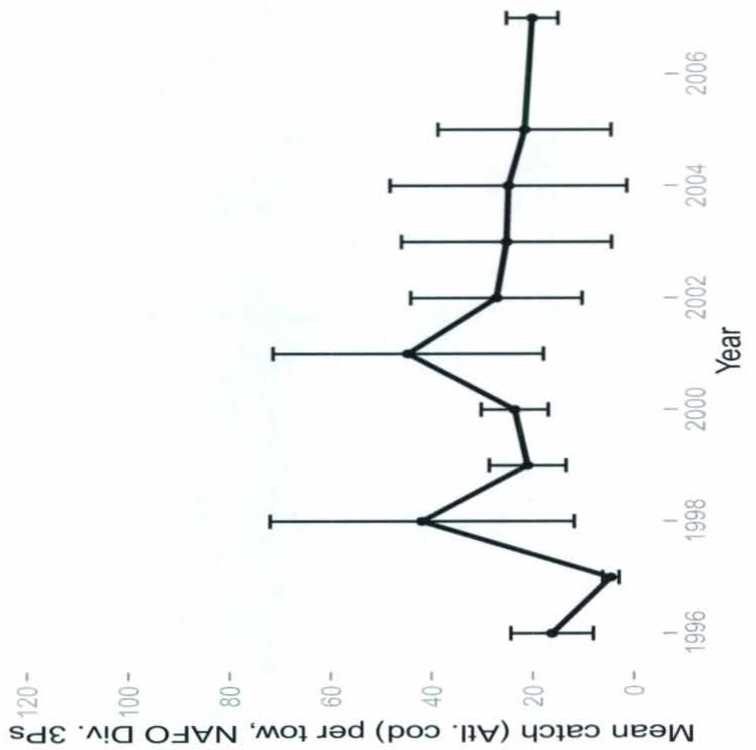

Figure 4.13: Time series of estimated average trawlable abundance $\hat{\mu}$ with the black segments indicating 95% normal confidence intervals, defined as $\hat{\mu} \pm z_{.025}\sqrt{\text{var}\,(\hat{\mu})}$. Notice the intervals are symmetric about the time series and can include negative values.

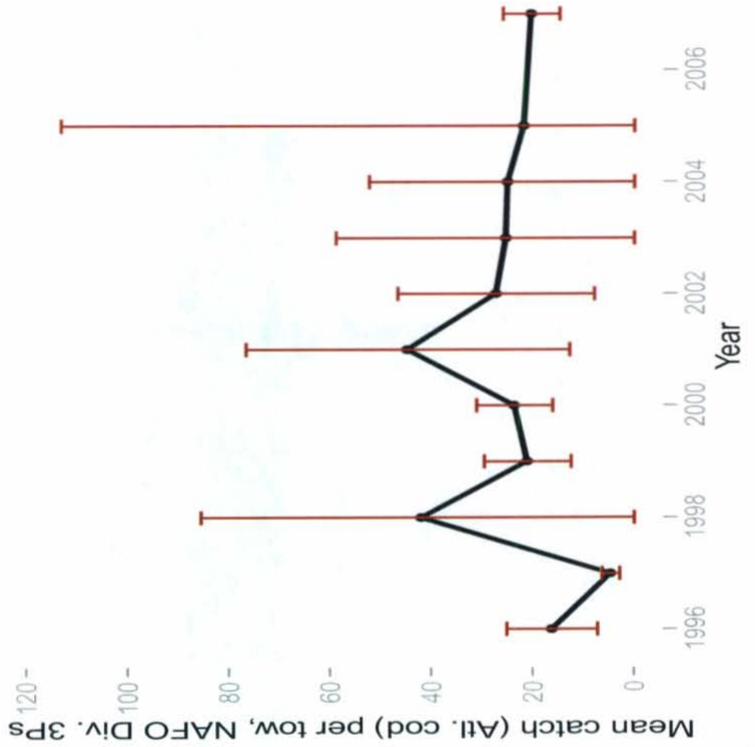

Figure 4.14: Time series of estimated average trawlable abundance $\hat{\mu}$ with the red segments indicating 95% Student's t confidence interval as defined in equation 4.2.6. The intervals are symmetric about the time series, but we have capped a lower limit at 0 for plotting purposes. Note that these intervals can (and do) take negative values otherwise.

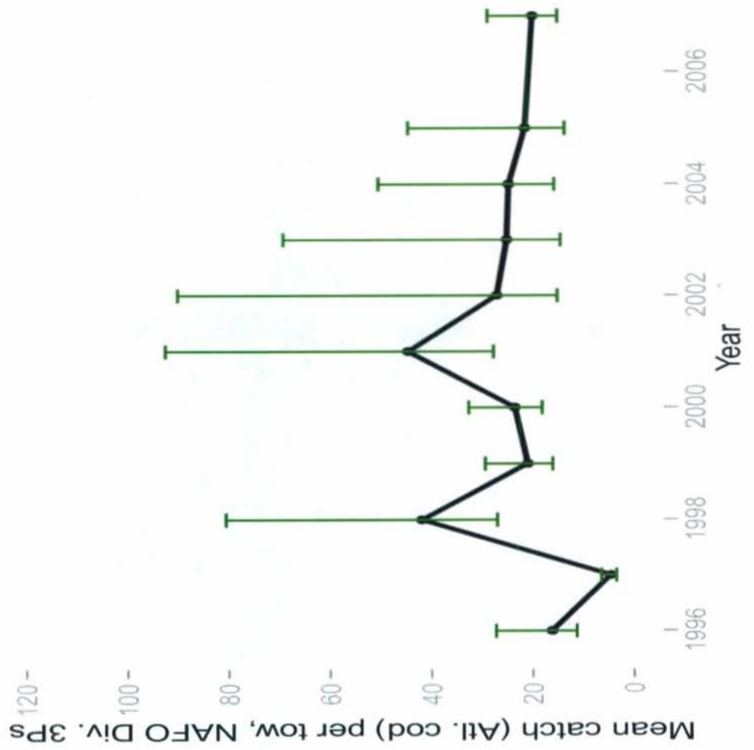

Figure 4.15: Time series of estimated average trawlable abundance $\hat{\mu}$ with the dark green segments indicating 95% negative binomial confidence intervals as defined in subsection 4.2.2. The intervals are not symmetric about the time series and cannot take negative values. $\tilde{k}_{aml}$ is used to estimate k.

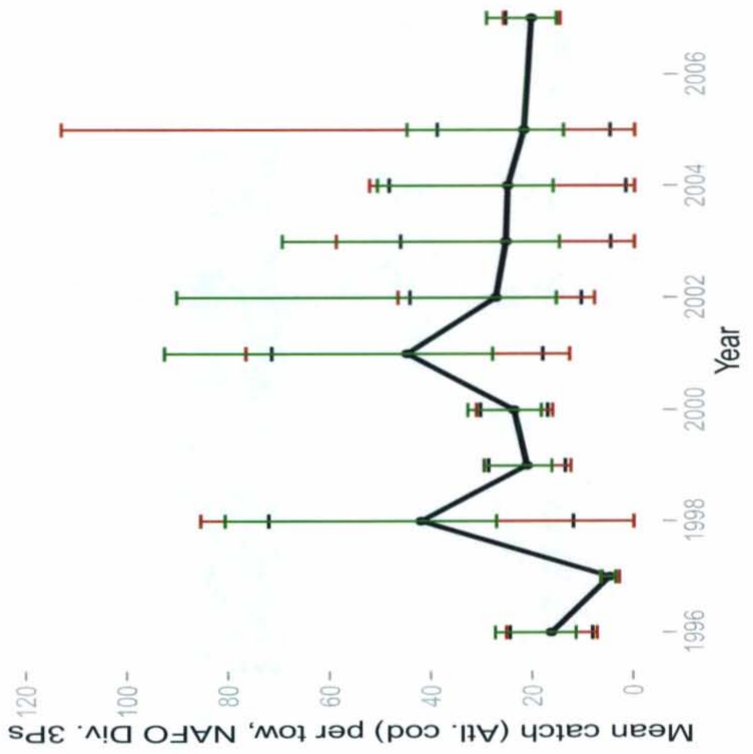

Figure 4.16: Time series of estimated average trawlable abundance $\hat{\mu}$ with various 95% confidence intervals. Black = normal, red = t, and dark green = negative binomial.

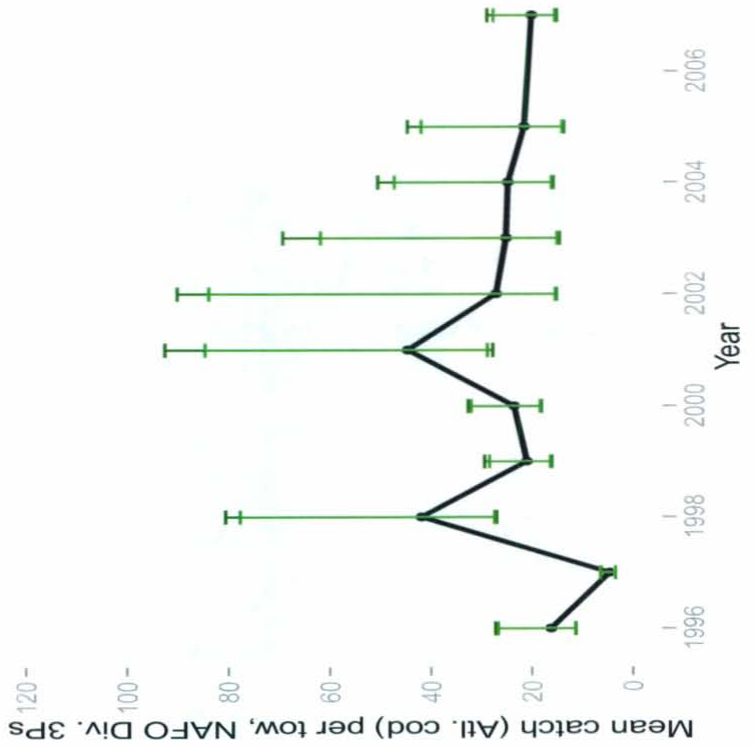

Figure 4.17: Time series of estimated average trawlable abundance $\hat{\mu}$ with 95% negative binomial confidence intervals. The dark green intervals use $\hat{k}_{aml}$ while the light green interval use $\hat{k}_{ml}$, which we expect to be biased in this highly-stratified model. Note that the NB intervals using $\hat{k}_{aml}$ are less pessimistic about the level of the time series.

Survey year	$\hat{\mu}$	z_-	z_+	t_-	t_+	nb_-	nb_+
1996	16.21	8.06	24.36	7.27	25.15	11.38	27.33
1997	4.62	2.99	6.25	2.88	6.36	3.56	6.54
1998	41.95	11.86	72.04	-1.67	85.57	27.12	80.70
1999	21.07	13.48	28.66	12.49	29.65	16.21	29.53
2000	23.66	17.03	30.29	16.16	31.16	18.33	32.82
2001	44.77	18.01	71.53	12.78	76.76	27.96	92.76
2002	27.31	10.36	44.26	7.89	46.73	15.39	90.35
2003	25.34	4.56	46.12	-8.27	58.95	14.82	69.56
2004	25.00	1.59	48.41	-2.40	52.40	16.08	50.78
2005	21.85	4.74	38.96	-69.63	113.33	14.04	44.96
2007	20.44	15.35	25.53	14.85	26.03	15.53	29.29

Table 4.4: Estimates of mean trawlable abundance and 95% confidence intervals by year. z, t, and nb refer to the the normal, t, and negative binomial intervals respectively. A $-$ subscript indicates the lower 95% CI endpoint and the $+$ subscript indicates the upper 95% CI endpoint. The normal intervals are too conservative in the lower endpoints and too tight in the upper endpoints. The t intervals behave similarly, and also include negative values.

Survey year	$\hat{\mu}$	nb_-	nb_+	nb_-^*	nb_+^*
1996	16.21	11.38	27.33	11.50	26.73
1997	4.62	3.56	6.54	3.60	6.41
1998	41.95	27.12	80.70	27.56	77.81
1999	21.07	16.21	29.53	16.54	28.55
2000	23.66	18.33	32.82	18.55	32.18
2001	44.77	27.96	92.76	29.04	84.80
2002	27.31	15.39	90.35	15.63	84.09
2003	25.34	14.82	69.56	15.30	62.11
2004	25.00	16.08	50.78	16.50	47.51
2005	21.85	14.04	44.96	14.37	42.24
2007	20.44	15.53	29.29	15.93	28.07

Table 4.5: Estimates of mean trawlable abundance and 95% negative binomial confidence intervals by year. The * superscript indicates estimates that were made using $\hat{k}_{ml}$. The others used $\hat{k}_{aml}$. The difference is noticeably apparent in the 2001-2003 index. In 2001 particularly the maximum profile likelihood estimator for k yields an upper limit of 84.8 for average trawlable abundance, while the maximum adjusted profile likelihood estimator yields 92.76.

Appendix A

Appendix A

A.1 Laplace approximation

Consider $L(\theta|\mathbf{y})$ a likelihood and l the log-likelihood. The Taylor expansion of L around any point θ_0 can be written in terms of l as

$$L(\theta|\mathbf{y}) = \exp\left\{l(\theta_0|\mathbf{y}) + (\theta-\theta_0)\frac{\partial l}{\partial\theta}\Big|_{\theta=\theta_0} + \frac{(\theta-\theta_0)^2}{2}\frac{\partial^2 l}{\partial\theta^2}\Big|_{\theta=\theta_0} + \dots\right\}.$$

Expanding around $\theta_0 = \hat{\theta}$, $l'(\hat{\theta}) = 0$ and L can be approximated as (using shorthand for derivatives)

$$L(\theta|\mathbf{y}) \approx \exp\left\{l(\hat{\theta}) + \frac{(\theta-\hat{\theta})^2}{2}l''(\hat{\theta})\right\}. \tag{A.1.1}$$

Notice also that the MLE $\hat{\theta}$ implies $l''(\hat{\theta}) < 0$. Integrating equation A.1.1 over the real line yields

$$\begin{aligned}\int_{-\infty}^{+\infty} L(\theta|\mathbf{y})d\theta &\approx \int_{-\infty}^{+\infty} \exp\left\{l(\hat{\theta}) + \frac{(\theta-\hat{\theta})^2}{2}l''(\hat{\theta})\right\}d\theta \\ &= \exp\left\{l(\hat{\theta})\right\}\int_{-\infty}^{+\infty}\exp\left\{\frac{(\theta-\hat{\theta})^2}{2l''(\hat{\theta})^{-1}}\right\}d\theta\end{aligned}$$

so that the integrand is the kernel of a $N(\hat{\theta}, -1/l''(\hat{\theta}))$ distribution. Knowing that the distribution integrates to one yields the required normalizing constant and thus

$$\int_{-\infty}^{+\infty} L(\theta|\mathbf{y})d\theta \approx \exp\left\{l(\hat{\theta})\right\}\left\{-\frac{2\pi}{l''(\hat{\theta})}\right\}^{\frac{1}{2}} \tag{A.1.2}$$

which is known as the *Laplace approximation* to $\int_{-\infty}^{+\infty} L(\theta|\mathbf{y})d\theta$. Note that the approximation holds well for sufficiently distant bounds (a, b), $a < b$ on the integral as a Gaussian function decreases rapidly as it departs from its mean value. Define $L(\theta|\mathbf{y})$ in terms of a positive function $m(\theta, t)$ by

$$L(\theta|\mathbf{y}) = \int_{-\infty}^{+\infty} m(\theta, t)dt.$$

Now define $k(\theta, t) = \log m(\theta, t)$. Applying the Laplace approximation to L for fixed θ yields

$$\begin{aligned} L(\theta|\mathbf{y}) &\approx \int_{\mathbb{R}} \exp\left\{k(\theta, \hat{t}(\theta)) + \frac{(t - \hat{t}(\theta))^2}{2} \frac{\partial^2 k}{\partial t^2}\Big|_{\hat{t}(\theta)}\right\} dt \\ &= \exp\left\{k(\theta, \hat{t}(\theta)\right\} \left\{-\frac{2\pi}{\frac{\partial^2 k}{\partial t^2}\Big|_{\hat{t}(\theta)}}\right\}^{\frac{1}{2}}. \end{aligned} \tag{A.1.3}$$

A.2 Saddlepoint approximation

Consider a density $f_X(x)$ with moment generating function (or *Laplace transform*) defined as

$$\phi_X(t) = \int_{-\infty}^{+\infty} e^{tx} f_X(x)dx$$

provided the integral is finite for $t \in (-\delta_0, \delta_0)$, $\delta_0 > 0$. Define the cumulant generating function $K_X(t) = \log \phi_X(t)$ and the characteristic function (or *Fourier transform*) as $\phi_X(it)$. The *inversion formula* is ([9])

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \phi_X(it) e^{itx} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp\left\{K_X(it) - itx\right\} dt. \end{aligned}$$

Define $t' = it$. Then

$$f_X(x) = \frac{1}{2\pi i} \int_{\tau-\infty}^{\tau+\infty} \exp\left\{K_X(it) - itx\right\} dt \tag{A.2.1}$$

for $\tau \in (-\delta_1, \delta_1)$, $\delta_1 > 0$. Taking a Taylor approximation of $K_X(t) - tx$ yields

$$K_X(t) - tx \approx K_X(\hat{t}(x)) - \hat{t}(x)x + \frac{(t - \hat{t}(x))^2}{2} K_X''|(\hat{t}(x))$$

and so substituting into equation A.2.1 and integrating yields [19]

$$f_X(x) \approx \exp\left\{K_X(\hat{t}(x)) - \hat{t}(x)x\right\} \left(\frac{1}{2\pi K_X''(\hat{t}(x))}\right)^{\frac{1}{2}} \tag{A.2.2}$$

the *saddlepoint approximation* to f.

For the Poisson, bionomial, and negative binomial distributions (with integer k), the saddlepoint approximation is found by using Stirling's approximation[1], given by

$$x! \approx \sqrt{2\pi} x^x e^{-x} \tag{A.2.3}$$

for all instances of $x!$ in the mass function [32]. The negative binomial distribution derived from the Poisson-gamma mixture allows non-integer k so that the gamma function terms $\Gamma(y+k)$ and $\Gamma(k)$ occur in the mass function, rather than factorials. In this case the approximation A.2.3 is crude, and the alternative Stirling's approximation

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}} x^x e^{-x} \tag{A.2.4}$$

for the gamma function yields a more accurate saddlepoint approximation. [24]

[1] Itself a saddlepoint approximation to $\Gamma(x+1)$ [19].

Bibliography

[1] DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2009/008. Stock assessment of subdivision 3ps cod. Technical report, DFO, 2009.

[2] G. Adimari and L. Ventura. Quasi-profile log likelihoods for unbiased estimating functions. *Annals of the Institute of Statistical Mathematics*, 54(2):235–244, 2002.

[3] S. I. Amari and H. Nagaoka. *Methods of information geometry*. Amer Mathematical Society, 2007.

[4] O. Barndorff-Nielsen. Conditionality resolutions. *Biometrika*, 67(2):293–310, 1980.

[5] O. Barndorff-Nielsen. On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343, 1983.

[6] O. E. Barndorff-Nielsen. Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 125–140, 1994.

[7] O. E. Barndorff-Nielsen. Quasi profile and directed likelihoods from estimating functions. *Annals of the Institute of Statistical Mathematics*, 47(3):461–464, 1995.

[8] O.E. Barndorff-Nielsen. *Information and exponential families in statistical theory*. New York: Wiley, 1978.

[9] P. Billingsley. *Probability and measure*. Wiley-India, 2008.

[10] M. Boyce and D. MacKenzie. Negative binomial models for abundance estimation of multiple closed populations. *Journal of Wildlife Management*, 65:498–509, 2001.

[11] N. G. Cadigan. *Statistical Inference about Fish Abundance: an Approach Based on Research Survey Data.* PhD thesis, University of Waterloo, 1999.

[12] N. G. Cadigan and J. Tobin. Estimating the negative binomial dispersion parameter with highly stratified surveys. *Journal of Statistical Planning and Inference*, 2010.

[13] N.G. Cadigan. Confidence intervals for trawlable abundance from stratified-random bottom trawl surveys. *Can. J. Fish. Aquat. Sci. (to appear)*, 2010.

[14] S. J. Clark and J. N. Perry. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics*, 45(1):309–316, 1989.

[15] W.G. Cochran. *Sampling Techniques.* Wiley, NY, 1977.

[16] D. R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1):1–39, 1987.

[17] D.R. Cox. Partial likelihood. *Biometrika*, 62:269, 1975.

[18] T. J. DiCiccio, M. A. Martin, S. E. Stern, and G. A. Young. Information bias and adjusted profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):189–203, 1996.

[19] C. Goutis and G. Casella. Explaining the saddlepoint approximation. *The American Statistician*, 53(3), 1999.

[20] D. R. Gunderson. *Surveys of fisheries resources.* John Wiley & Sons Inc, 1993.

[21] R. V. Hogg, J. W. McKean, and A. T. Craig. *Introduction to mathematical statistics.* Prentice Hall, 2005.

[22] Robert L. Wolpert James O. Berger, Brunero Liseo. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28, 1999.

[23] J. D. Kalbfleisch and D. A. Sprott. Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):175–208, 1970.

[24] Y. Lee and J. Nelder. Extended-reml estimators. *Journal of Applied Statistics*, 30(8):845–856, 2003.

[25] Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):619–678, 1996.

[26] Y. Lee and J. A. Nelder. Generalized linear models for the analysis of quality-improvement experiments. *Canadian journal of statistics*, 26(1):95–105, 1998.

[27] Y. Lee and J. A. Nelder. Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987, 2001.

[28] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Verlag, 1998.

[29] L. Lin and R. Zhang. Profile quasi-likelihood. *Statistics & Probability Letters*, 56(2):147–154, 2002.

[30] J. O. Lloyd-Smith. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PloS one*, 2(2):180, 2007.

[31] P. McCullagh and R. Tibshirani. A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 325–344, 1990.

[32] J. A. Nelder and D. Pregibon. An extended quasi-likelihood function. *Biometrika*, pages 221–232, 1987.

[33] J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

[34] L. Pace and A. Salvan. Adjustments of the profile likelihood from a new perspective. *Journal of Statistical Planning and Inference*, 136(10):3554–3564, 2006.

[35] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545, 1971.

[36] W. W. Piegorsch. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, 46(3):863–867, 1990.

[37] N. Reid. Asymptotics and the theory of inference. *Annals of Statistics*, 31(6):1695–1731, 2003.

[38] D. Robinson and G. Smyth. Small sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9:321–332, 2008.

[39] K. Saha and S. Paul. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61(1):179–185, 2005.

[40] K. K. Saha. Semiparametric estimation for the dispersion parameter in the analysis of over-or underdispersed count data. *Journal of Applied Statistics*, 35(12):1383–1397, 2008.

[41] N. Sartori. Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, 90(3):533, 2003.

[42] N. Sartori, R. Bellio, A. Salvan, and L. Pace. The directed modified profile likelihood in models with many nuisance parameters. *Biometrika*, 86:735–742, 1999.

[43] T. A. Severini. An approximation to the modified profile likelihood function. *Biometrika*, 85(2):403, 1998.

[44] T. A. Severini. Likelihood ratio statistics based on an integrated likelihood. *Biometrika*, 2010.

[45] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss–newton method. *Biometrika*, 61(3):439, 1974.

[46] J. Wu. *Asymptotic Likelihood Inference*. PhD thesis, University of Toronto, 1999.