

PROFILE-SCORE ADJUSTMENTS FOR INCIDENTAL-PARAMETER PROBLEMS

GEERT DHAENE

KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
geert.dhaene@kuleuven.be

KOEN JOCHMANS

Sciences Po, 28 rue des Saints-Pères, 75007 Paris, France
koen.jochmans@sciencespo.fr

This version: September 12, 2017

We propose a scheme of iterative adjustments to the profile score to deal with incidental-parameter bias in models for stratified data with few observations on a large number of strata. The first-order adjustment is based on a calculation of the profile-score bias and evaluation of this bias at maximum-likelihood estimates of the incidental parameters. If the bias does not depend on the incidental parameters, the first-order adjusted profile score is fully recentered, solving the incidental-parameter problem. Otherwise, it is approximately recentered, alleviating the incidental-parameter problem. In the latter case, the adjustment can be iterated to give higher-order adjustments, possibly until convergence. The adjustments are generally applicable (e.g., not requiring parameter orthogonality) and lead to estimates that generally improve on maximum likelihood. We examine a range of nonlinear models with covariates. In many of them, we obtain an adjusted profile score that is exactly unbiased. In the others, we obtain approximate bias adjustments that yield much improved estimates, relative to maximum likelihood, even when there are only two observations per stratum.

Keywords: adjusted profile score, bias reduction, incidental parameters.

INTRODUCTION

Consider inference about a finite-dimensional parameter ψ based on m observations from each of n independent strata in the presence of incidental parameters λ_i ($i = 1, 2, \dots, n$). It is well known that maximum likelihood does not, in general, yield consistent point estimates of ψ as the number of strata, n , increases while their size, m , is kept fixed; see [Neyman and Scott \(1948\)](#). Only in some cases is it possible to separate inference about ψ from inference about the λ_i by means of a conditional or marginal likelihood; see [Andersen \(1970\)](#) and [Lancaster \(2000\)](#) for examples. Estimation of ψ may, alternatively, be based on an integrated likelihood, as discussed in [Kalbfleisch and Sprott \(1970\)](#), but the choice of prior density on λ_i may be hard to justify.

An alternative route is to work with either a modified likelihood ([Barndorff-Nielsen 1983](#)) or an approximate conditional likelihood ([Cox and Reid 1987](#)). In a rectangular-array embedding ([Li et al. 2003](#)), [Sartori \(2003\)](#) showed that the modified likelihood generally leads to superior inference. [Lancaster \(2002\)](#) found similar improvements for approximate conditional likelihoods. [Arellano and Bonhomme \(2009\)](#) extended these to situations where an information-orthogonalizing reparameterization is not possible. [Hahn and Newey \(2004\)](#) developed bias corrections with similar gains.

In this paper, we study estimation of ψ based on adjustments to the profile score, as in [McCullagh and Tibshirani \(1990\)](#). The basic (or first-order) adjustment is to calculate the bias of the profile score, evaluate it at maximum-likelihood estimates of the λ_i , and subsequently recenter the profile score. When the bias is free of incidental parameters, the so-adjusted profile score is fully recentered and the corresponding estimating equation has a root that is consistent for ψ under Neyman-Scott asymptotics. We find this to be the case in a number of relevant models. When the bias is not free of incidental parameters, the first-order adjusted profile

score leads to point estimates whose bias is $O(m^{-2})$, as opposed to the standard $O(m^{-1})$. We show that the adjustment can be iterated, possibly until convergence, generating higher-order adjustments. Assuming sufficient regularity, at each iteration the order of the bias is reduced, and the fully iterated bias adjustment may yield consistent estimates under Neyman-Scott asymptotics. We study the adjustments analytically and numerically in a range of nonlinear models. Invariably, the profile score adjustments are found to be very effective, either leading to consistent estimates under Neyman-Scott asymptotics or, else, to estimates with much smaller bias than maximum likelihood, even for $m = 2$. In the examples examined, we also find that, when a conditional or marginal likelihood exists, its score function coincides with the fully iterated adjusted profile score.

Working with the profile score, and adjustments thereof, has some further attractive features relating to problem diagnosis, invariance, generality, simplicity, and higher-order properties. Calculation of the bias of the profile score serves as a simple diagnostic tool, since it will reveal if the presence of incidental parameters does or does not lead to an incidental-parameter problem for the model at hand. Further, the profile-score adjustments are invariant under interest-respecting reparameterizations, and they can always be computed. In particular, they do not rely on the existence of an information-orthogonal parameterization or a minimal sufficient statistic for λ_i that is independent of ψ . It is well known that both of these do not always exist (Severini 2000). The adjustments are computationally simple in that they only require calculating certain expectations, which can always be obtained by simulation. Sample-space derivatives are not needed. The difficulty to compute these has been a major reason for the development of approximations to the modified profile likelihood such as those of Cox and Reid (1987) and Severini (1998), for example. Finally, the iterative procedure leads to higher-order improvements that the modified profile likelihood does not attain in general. The fully iterated adjustment yields estimators whose bias, in principle, shrinks exponentially fast in m .

1. ADJUSTING THE PROFILE SCORE

1.1. Bias of the profile score

Suppose we are given a rectangular-array data set $\{y_{ij}; i = 1, \dots, n; j = 1, \dots, m\}$ with n strata and m observations for each stratum. The observations y_{ij} are sampled from a probability density (or mass) function $f(y_{ij}; \psi, \lambda_i)$, where ψ and λ_i are finite-dimensional parameters. Although suppressed in the notation, the density may also depend on covariates, in which case the expectations taken below are conditional expectations, given the covariates. The parameter of interest is ψ , with $\lambda = (\lambda_1, \dots, \lambda_n)$ being treated as a nuisance parameter. For simplicity of the exposition, in this section we assume that the observations y_{ij} are independent across i and j . Some examples with dependent data will be discussed in the next section.

The profile log-likelihood and score functions for ψ , together with their i th contributions, are

$$\begin{aligned} l(\psi) &= \sum_{i=1}^n l_i(\psi), & l_i(\psi) &= \sum_{j=1}^m \log f(y_{ij}; \psi, \hat{\lambda}_i(\psi)), \\ s(\psi) &= \sum_{i=1}^n s_i(\psi), & s_i(\psi) &= \sum_{j=1}^m \nabla_{\psi} \log f(y_{ij}; \psi, \hat{\lambda}_i(\psi)), \end{aligned}$$

where $\hat{\lambda}_i(\psi) = \arg \max_{\lambda_i} \sum_{j=1}^m \log f(y_{ij}; \psi, \lambda_i)$ is the maximum likelihood estimator of λ_i for fixed ψ . Let $\hat{\psi} = \arg \max_{\psi} l(\psi)$ be the maximum likelihood estimator of ψ and assume sufficient regularity to ensure that $\hat{\psi}$ satisfies $s(\hat{\psi}) = 0$. Neyman and Scott (1948) showed that $\hat{\psi}$ is not, in general, a consistent estimator of the true value of ψ as $n \rightarrow \infty$ while m remains fixed. This is the incidental-parameter problem.

When $\widehat{\psi}$ is inconsistent, the inconsistency is due to a bias in the profile score function. One may view $s(\psi)$ as an approximation to the infeasible profile score function

$$s^{\text{in}}(\psi) = \sum_{i=1}^n s_i^{\text{in}}(\psi), \quad s_i^{\text{in}}(\psi) = \sum_{j=1}^m \nabla_{\psi} \log f(y_{ij}; \psi, \lambda_i(\psi)),$$

where $\lambda_i(\psi) = \arg \max_{\lambda_i^*} E_{\psi, \lambda_i} \sum_{j=1}^m \log f(y_{ij}; \psi, \lambda_i^*)$ and $E_{\psi, \lambda_i}(\cdot)$ denotes the expectation under the density $f(\cdot; \psi, \lambda_i)$. So $s(\psi)$ differs from $s^{\text{in}}(\psi)$ in that $s(\psi)$ uses $\widehat{\lambda}(\psi) = (\widehat{\lambda}_1(\psi), \dots, \widehat{\lambda}_n(\psi))$ whereas $s^{\text{in}}(\psi)$ uses the infeasible $\lambda(\psi) = (\lambda_1(\psi), \dots, \lambda_n(\psi))$, a difference that often introduces a bias. While $s^{\text{in}}(\psi)$ is unbiased, i.e., $E_{\psi, \lambda} s^{\text{in}}(\psi) = 0$ (see, e.g., [Pace and Salvan 2006](#)), it is often the case that

$$E_{\psi, \lambda} s(\psi) \neq 0,$$

causing $\widehat{\psi}$ to be inconsistent for fixed m and the limit distribution of $\widehat{\psi}$ to be incorrectly centered unless $m/n \rightarrow \infty$; see, e.g., [Li et al. \(2003\)](#). Typically, when the profile score is biased, its bias is $O(n)$ and the inconsistency of $\widehat{\psi}$, as $n \rightarrow \infty$ with fixed m , is $O(m^{-1})$.

1.2. Iterated bias adjustment

Our approach is to bias-adjust $s(\psi)$ and, therefore, requires calculating $E_{\psi, \lambda} s(\psi)$ analytically or numerically for given ψ and λ . Three cases can arise:

- (a) $E_{\psi, \lambda} s(\psi) = 0$;
- (b) $E_{\psi, \lambda} s(\psi) \neq 0$ but $E_{\psi, \lambda} s(\psi)$ is free of λ ;
- (c) $E_{\psi, \lambda} s(\psi) \neq 0$ and $E_{\psi, \lambda} s(\psi)$ depends on λ .

In Case (a), $s(\psi) = 0$ is an unbiased estimating equation and $\widehat{\psi}$ is consistent as $n \rightarrow \infty$ for fixed m . The interesting point here is that a simple calculation, that of $E_{\psi, \lambda} s(\psi)$, will reveal so.

In Case (b), $\widehat{\psi}$ is inconsistent for fixed m , but an unbiased estimating equation is readily obtained. The adjusted profile score

$$s_a(\psi) = s(\psi) - E_{\psi, \lambda} s(\psi)$$

is unbiased by construction and feasible because it does not depend on λ . [Neyman and Scott \(1948\)](#) already noted that, when $E_{\psi, \lambda} s(\psi)$ is free of λ , a fixed- m consistent estimator can be obtained by centering the profile score.

In Case (c), consider the first-order adjusted profile score

$$s_a^{(1)}(\psi) = s(\psi) - E_{\psi, \widehat{\lambda}(\psi)} s(\psi).$$

[McCullagh and Tibshirani \(1990\)](#) suggested this approximate centering of the profile score in the generic context where nuisance parameters are profiled out of the likelihood. Under regularity conditions, the first-order adjusted profile score reduces the large- m asymptotic bias of the profile score by a factor $O(m^{-1})$. The profile score bias is

$$E_{\psi, \lambda} s(\psi) = E_{\psi, \lambda(\psi)} s(\psi) = \sum_{i=1}^n E_{\psi, \lambda_i(\psi)} s_i(\psi),$$

which we are approximating by

$$E_{\psi, \widehat{\lambda}(\psi)} s(\psi) = \sum_{i=1}^n E_{\psi, \widehat{\lambda}_i(\psi)} s_i(\psi).$$

Thus, in each $E_{\psi, \lambda_i(\psi)} s_i(\psi)$ the infeasible $\lambda_i(\psi)$ is approximated by $\widehat{\lambda}_i(\psi)$. This introduces a relative bias which is $O(m^{-1})$ as $m \rightarrow \infty$ (see, e.g., DiCiccio et al. 1996), i.e., $E_{\psi, \lambda_i} E_{\psi, \widehat{\lambda}_i(\psi)} s_i(\psi) = (1 + O(m^{-1})) E_{\psi, \lambda_i} s_i(\psi)$ and, on summing over i ,

$$E_{\psi, \lambda} E_{\psi, \widehat{\lambda}(\psi)} s(\psi) = (1 + O(m^{-1})) E_{\psi, \lambda} s(\psi).$$

Therefore, relative to profile score, the first-order adjusted profile score reduces the bias from $E_{\psi, \lambda} s(\psi) = O(n)$ to $E_{\psi, \lambda} s_a^{(1)}(\psi) = O(n/m)$. The adjustment removes the first-order asymptotic bias from $s(\psi)$; see also Sartori (2003).

In Case (c), the adjustment can be iterated, each iteration giving a further asymptotic improvement. The bias $E_{\psi, \lambda} s_a^{(1)}(\psi)$ can be approximated by $E_{\psi, \widehat{\lambda}(\psi)} s_a^{(1)}(\psi)$, again with relative bias $O(m^{-1})$, leading to the second-order adjusted profile score

$$\begin{aligned} s_a^{(2)}(\psi) &= s_a^{(1)}(\psi) - E_{\psi, \widehat{\lambda}(\psi)} s_a^{(1)}(\psi) \\ &= s(\psi) - 2E_{\psi, \widehat{\lambda}(\psi)} s(\psi) + E_{\psi, \widehat{\lambda}(\psi)} \left(E_{\psi, \widehat{\lambda}(\psi), \widehat{\lambda}(\psi)} s(\psi) \right), \end{aligned}$$

with bias $E_{\psi, \lambda} s_a^{(2)}(\psi) = O(n/m^2)$. Here, $\widehat{\lambda}(\psi, \widehat{\lambda}(\psi))$ is the maximum likelihood estimator of λ , for fixed ψ , based on a data set $\{y_{ij}^*; i = 1, \dots, n; j = 1, \dots, m\}$ where y_{ij}^* is sampled from $f(\cdot; \psi, \widehat{\lambda}_i(\psi))$. The structure of the iterated adjustments is now apparent. Defining the p -fold iteration of $E_{\psi, \widehat{\lambda}(\psi)}(\cdot)$ by the recursion

$$\begin{aligned} E_{\psi, \widehat{\lambda}(\psi)}^{(0)}(\cdot) &= (\cdot), \\ E_{\psi, \widehat{\lambda}(\psi)}^{(p)}(\cdot) &= E_{\psi, \widehat{\lambda}(\psi)} \left(E_{\psi, \widehat{\lambda}(\psi), \widehat{\lambda}(\psi)}^{(p-1)}(\cdot) \right), \quad p = 1, 2, \dots, \end{aligned}$$

the k th order adjusted profile score is

$$\begin{aligned} s_a^{(k)}(\psi) &= s_a^{(k-1)}(\psi) - E_{\psi, \widehat{\lambda}(\psi)} s_a^{(k-1)}(\psi) \\ &= s(\psi) - \sum_{p=1}^k \binom{k}{p} (-1)^{p-1} E_{\psi, \widehat{\lambda}(\psi)}^{(p)} s(\psi), \end{aligned}$$

with bias $O(n/m^k)$, given sufficient regularity. The associated estimator, $\widehat{\psi}_a^{(k)}$, is defined as the solution to $s_a^{(k)}(\psi) = 0$. The adjustment may be iterated until convergence. Upon existence of the limits, we define the (fully iterated) adjusted profile score as $s_a(\psi) = \lim_{k \rightarrow \infty} s_a^{(k)}(\psi)$ and the (fully) adjusted-score estimator as $\widehat{\psi}_a = \lim_{k \rightarrow \infty} \widehat{\psi}_a^{(k)}$.

1.3. Further discussion

In Case (c), the rationale for iterating the adjustment, possibly until convergence, is based on informal large- m arguments. Our focus in this paper is not on the large- m asymptotics per se and on providing detailed regularity conditions for them. Instead, we examine the profile-score adjustments for fixed m through many examples. Our main finding is that the fully iterated adjustment may have very good properties even when m is very small. In particular, it may occur that $\widehat{\psi}_a$ is fixed- m consistent whereas $\widehat{\psi}$ and $\widehat{\psi}_a^{(k)}$ (for any finite k) are not. Also, in cases where $\widehat{\psi}_a$ is not fixed- m consistent, it may still reduce the asymptotic bias of $\widehat{\psi}$ considerably. Such situations arise when ψ is not point identified for a given m . Obviously, when point identification fails, $s_a(\psi)$ cannot deliver an unbiased estimating equation, i.e., it must be the case that $E_{\psi, \lambda} s_a(\psi)$ is generically nonzero or that $s_a(\psi)$ vanishes in the neighborhood of the true value of ψ . Yet, in either case, $\widehat{\psi}_a$ may still be well defined (i.e., the required limit exists) and have far better properties than $\widehat{\psi}$. We will discuss examples to illustrate these points.

Approximate standard errors of $\widehat{\psi}_a$ follow from common arguments (and similarly for those of $\widehat{\psi}_a^{(k)}$). Under mild conditions, as $n \rightarrow \infty$ with m fixed, $\widehat{\psi}_a$ converges to some ψ_* (equal to ψ if $\widehat{\psi}_a$ is fixed- m consistent and depending on m otherwise) and is asymptotically normally distributed, i.e.,

$$\sqrt{n}(\widehat{\psi}_a - \psi_*) \xrightarrow{d} N(0, H^{-1} J H^{-1})$$

where $H = \text{plim}_{n \rightarrow \infty} \nabla_{\psi} s_a(\psi_*)/n$ and $J = \text{plim}_{n \rightarrow \infty} \sum_i s_{ai}(\psi_*) s_{ai}(\psi_*)^{\top}/n$, with plug-in estimates $\widehat{H} = \nabla_{\psi} s_a(\widehat{\psi}_a)$ and $\widehat{J} = \sum_i s_{ai}(\widehat{\psi}_a) s_{ai}(\widehat{\psi}_a)^{\top}/n$, expressions that allow for conditional dependency of the data across j , given the λ_i .

To compute $\widehat{\psi}_a$ an expression of $s_a(\psi)$ or $s_a^{(k)}(\psi)$ is needed, which in turn requires an expression of the expectations $E_{\psi, \widehat{\lambda}(\psi)}^{(p)} s(\psi)$. When the model is tractable enough, the expectations may be obtained in closed or semi-closed form, so that $\widehat{\psi}_a$ becomes computationally feasible. This is the case in several nonlinear regression models, as we discuss below. When the expectations are not available in a suitable form, it is more difficult to compute $\widehat{\psi}_a$. One may, however, approximate $\widehat{\psi}_a$ by $\widehat{\psi}_a^{(k)}$ for some chosen k , and approximate the required expectations by simulations or other numerical methods. For further details, we refer to a numerical example relating to the negative binomial regression model discussed in Sections 2 and 3.

The classification (a)–(c) helps to discern if there is an incidental-parameter problem for ψ and, if so, how it may be solved or mitigated. When ψ is multidimensional, there may be an incidental-parameter problem only for a subvector of ψ . Let ψ and $s(\psi)$ be conformably partitioned as

$$\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \quad s(\psi) = \begin{pmatrix} s_{\psi_1}(\psi) \\ s_{\psi_2}(\psi) \end{pmatrix}.$$

Then, if for any ψ'_2 ,

$$\begin{aligned} E_{\psi, \lambda} s_{\psi_1}(\psi') &= 0 & \text{for } \psi' &= \begin{pmatrix} \psi_1 \\ \psi'_2 \end{pmatrix}, \\ E_{\psi, \lambda} s_{\psi_2}(\psi) &\neq 0, \end{aligned}$$

there is an incidental-parameter problem for ψ_2 but it does not carry over to ψ_1 . Note that $E_{\psi, \lambda} s_{\psi_1}(\psi) = 0$ is necessary but not sufficient to prevent the incidental-parameter problem for ψ_2 to carry over to ψ_1 .

2. EXAMPLES

Many of our examples are conditional models of a variable y_{ij} given a vector of covariates x_{ij} . Accordingly, expectations are taken conditionally on the covariates.

2.1. Case (a) models

Poisson regression. Consider Poisson random variables y_{ij} with mean $\mu_{ij} = \lambda_i \exp(x_{ij}^{\top} \psi)$. The probability mass function is $f(y_{ij}; \psi, \lambda_i) = \mu_{ij}^{y_{ij}} \exp(-\mu_{ij})/y_{ij}!$. For fixed ψ , the maximum likelihood estimator of λ_i is $\widehat{\lambda}_i(\psi) = \sum_j y_{ij} / \sum_j \exp(x_{ij}^{\top} \psi)$. The profile log-likelihood and score are

$$\begin{aligned} l(\psi) &= \sum_{i,j} y_{ij} \left(-\log \sum_j \exp(x_{ij}^{\top} \psi) + x_{ij}^{\top} \psi \right), \\ s(\psi) &= \sum_{i,j} y_{ij} \left(-\frac{\sum_j \exp(x_{ij}^{\top} \psi) x_{ij}}{\sum_j \exp(x_{ij}^{\top} \psi)} + x_{ij} \right). \end{aligned}$$

(We omit additive constants from $l(\psi)$ in all examples.) Taking expectations gives $E_{\psi,\lambda}s(\psi) = 0$. There is no incidental-parameter problem in this model and, accordingly, the adjustment leaves the profile score unaltered.

Blundell et al. (1999) gave a closely related derivation. Further, Lancaster (2002) showed that ψ and λ are likelihood orthogonal after an interest-respecting reparametrization (i.e., the unprofiled likelihood is separable), which is another way of showing that maximum likelihood is consistent. The modifications of Barndorff-Nielsen (1983) and Cox and Reid (1987) to the profile likelihood also leave it unchanged. The profile likelihood is equal to the conditional likelihood given $\sum_j y_{ij}$, $i = 1, \dots, n$, which is a sufficient statistic for λ . Hence, from Hahn (1997) it follows that maximum likelihood attains the semiparametric efficiency bound.

Exponential regression. Let y_{ij} be exponentially distributed with scale $\mu_{ij} = \lambda_i \exp(x_{ij}^\top \psi)$, i.e., $f(y_{ij}; \psi, \lambda_i) = \mu_{ij}^{-1} \exp(-y_{ij}/\mu_{ij})$. Then $\hat{\lambda}_i(\psi) = m^{-1} \sum_j y_{ij} \exp(-x_{ij}^\top \psi)$ and the profile log-likelihood and score are

$$l(\psi) = \sum_{i,j} \left(-\log \sum_j y_{it} \exp(-x_{ij}^\top \psi) - x_{ij}^\top \psi \right),$$

$$s(\psi) = \sum_{i,j} \left(\frac{\sum_j y_{ij} \exp(-x_{ij}^\top \psi) x_{ij}}{\sum_j y_{ij} \exp(-x_{ij}^\top \psi)} - x_{ij} \right).$$

Now write

$$\frac{\sum_j y_{ij} \exp(-x_{ij}^\top \psi) x_{ij}}{\sum_j y_{ij} \exp(-x_{ij}^\top \psi)} = \frac{\sum_j z_{ij} x_{ij}}{\sum_j z_{ij}},$$

$z_{ij} = y_{ij}/\mu_{ij}$ being independent unit-exponential random variables. Because the z_{ij} are identically distributed,

$$E \left(\frac{\sum_j z_{ij} x_{ij}}{\sum_j z_{ij}} \right) = \sum_j E \left(\frac{z_{ij}}{\sum_j z_{ij}} \right) x_{ij} = \frac{1}{m} \sum_j x_{ij}.$$

Hence $E_{\psi,\lambda}s(\psi) = 0$. There is no incidental-parameter problem. Again, the profile likelihood coincides with the modified profile likelihood of Barndorff-Nielsen (1983). It is also identical to the marginal likelihood of the ratios y_{ij}/y_{i1} , which is free of λ .

2.2. Case (b) models

Many normal means. This is the classic Neyman and Scott (1948) example of the incidental-parameter problem. The goal is to infer the variance ψ from independent observations $y_{ij} \sim \mathcal{N}(\lambda_i, \psi)$. The profile score is

$$s(\psi) = -\frac{nm}{2\psi} + \frac{1}{2\psi^2} \sum_{i,j} (y_{ij} - \bar{y}_i)^2,$$

with bias $E_{\psi,\lambda}s(\psi) = -n(2\psi)^{-1}$, free of λ , and $\hat{\psi} = (nm)^{-1} \sum_{i,j} (y_{ij} - \bar{y}_i)^2$ converges to $(1 - m^{-1})\psi$. The solution of the adjusted profile score equation $s_a(\psi) = 0$ is $\hat{\psi}/(1 - m^{-1})$, which is consistent for fixed m . Numerous other approaches lead to the same estimator.

A regression version of this model has $y_{ij} \sim \mathcal{N}(\lambda_i + x_{ij}^\top \psi_1, \psi_2)$, with ψ consisting of ψ_1 and ψ_2 . Here, the bias of the profile score for ψ_1 is zero and for ψ_2 it is $-n(2\psi_2)^{-1}$, as in the no-covariate case. Again, solving the adjusted profile score equation yields the standard solution: (i) the maximum likelihood estimator of ψ_1

(which is least-squares applied to the pooled group-wise demeaned data) is unaltered; (ii) a one-degree-of-freedom correction is applied to the maximum likelihood estimator of ψ_2 .

Autoregression. Since [Nickell \(1981\)](#), autoregressive models have become another classic instance of the incidental-parameter problem. Suppose that

$$y_{ij} = \lambda_i + \psi_1 y_{ij-1} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \psi_2),$$

where we observe y_{ij} for $j = 0, 1, \dots, m$ and leave the initial observations, y_{i0} , unrestricted (i.e., we condition on them). The profile score is

$$s(\psi) = \begin{pmatrix} -\psi_2^{-1} \sum_i (y_i - \psi_1 y_{i-})^\top M y_{i-} \\ -(2\psi_2)^{-1} (nm - \psi_2^{-1} \sum_i (y_i - \psi_1 y_{i-})^\top M (y_i - \psi_1 y_{i-})) \end{pmatrix},$$

where $y_i = (y_{i1}, y_{i2}, \dots, y_{im})^\top$, $y_{i-} = (y_{i0}, y_{i1}, \dots, y_{i,m-1})^\top$, $M = I - m^{-1} \iota \iota^\top$, I is the $m \times m$ identity matrix, and ι is an m -vector of ones. Using backward substitution, it follows that

$$E_{\psi, \lambda} s(\psi) = \begin{pmatrix} -n(m-1)^{-1} \sum_{j=1}^{m-1} (m-j) \psi_1^{j-1} \\ -n(2\psi_2)^{-1} \end{pmatrix},$$

which is free of λ . In this model, the adjusted score equation typically has more than one root, so the correct root has to be selected, as discussed in [Dhaene and Jochmans \(2016\)](#).

[Lancaster \(2002\)](#) showed that λ and ψ can be orthogonalized and derived a [Cox and Reid \(1987\)](#) conditional profile log-likelihood, whose score function coincides with the adjusted profile score. When the model is extended to include covariates or more than one autoregressive term, the bias of the profile score still admits a closed form and remains free of incidental parameters. In contrast, an orthogonalizing reparameterization of λ and ψ no longer exists (see [Dhaene and Jochmans 2016](#)).

Weibull regression. Suppose that y_{ij} is Weibull distributed with survival function $\exp(-(y_{ij}/\mu_{ij})^{\psi_2})$, where $\mu_{ij} = \lambda_i \exp(x_{ij}^\top \psi_1)$. Then $\hat{\lambda}_i(\psi) = (m^{-1} \sum_j w_{ij}(\psi))^{1/\psi_2}$, where $w_{ij}(\psi) = (y_{ij} \exp(-x_{ij}^\top \psi_1))^{\psi_2}$. The profile log-likelihood and score are

$$l(\psi) = \sum_{i,j} \left(\log \psi_2 + \psi_2 \log y_{ij} - \log \sum_j w_{ij}(\psi) - \psi_2 x_{ij}^\top \psi_1 \right),$$

$$s(\psi) = \sum_{i,j} \begin{pmatrix} \psi_2 \sum_j w_{ij}(\psi) x_{ij} / \sum_j w_{ij}(\psi) - \psi_2 x_{ij} \\ \psi_2^{-1} + \log y_{ij} - \psi_2^{-1} \sum_j w_{ij}(\psi) \log w_{ij}(\psi) / \sum_j w_{ij}(\psi) - x_{ij}^\top \psi_1 \end{pmatrix}.$$

A calculation summarized in the Appendix gives the bias of the profile score as

$$E_{\psi, \lambda} s(\psi) = \begin{pmatrix} 0 \\ n\psi_2^{-1} \end{pmatrix},$$

which is free of λ . The solutions of $s(\psi) = 0$ and $s_a(\psi) = 0$ differ for both ψ_1 and ψ_2 . However, there is an incidental-parameter problem only for ψ_2 because, as shown in the Appendix, the first component of $s(\psi')$, with $\psi' = (\psi_1, \psi_2')$, has zero expectation for any ψ_2' .

Several other approaches turn out to be equivalent to adjusting the profile score. [Lancaster \(2000\)](#) showed that λ and ψ can be orthogonalized. Integrating the reparameterized λ from the likelihood using a uniform prior gives a [Cox and Reid \(1987\)](#) conditional profile likelihood. [Chamberlain \(1985\)](#) suggested to use the marginal likelihood of the ratios y_{ij}/y_{i1} , which is free of λ . The conditional profile log-likelihood and the marginal log-likelihood are identical, and their score functions are identical to the adjusted profile score function.

Gamma regression. Here, y_{ij} is gamma distributed with scale $\mu_{ij} = \lambda_i \exp(x_{ij}^\top \psi_1)$ and shape parameter ψ_2 . The density is $f(y_{ij}; \psi, \lambda_i) = y_{ij}^{\psi_2-1} \mu_{ij}^{-\psi_2} \exp(-y_{ij}/\mu_{ij})/\Gamma(\psi_2)$ and the maximum likelihood estimator of λ_i for fixed ψ is $\hat{\lambda}_i(\psi) = \psi_2^{-1} m^{-1} \sum_j y_{ij} \exp(-x_{ij}^\top \psi_1)$. The profile log-likelihood and score are

$$l(\psi) = \sum_{i,j} \left(-\log \Gamma(\psi_2) + \psi_2 \log(m\psi_2 y_{ij}) - \psi_2 - \psi_2 \log \sum_j y_{ij} \exp(-x_{ij}^\top \psi_1) - \psi_2 x_{ij}^\top \psi_1 \right),$$

$$s(\psi) = \sum_{i,j} \left(\begin{array}{c} \psi_2 \sum_j y_{ij} \exp(-x_{ij}^\top \psi_1) x_{ij} / \sum_j y_{ij} \exp(-x_{ij}^\top \psi_1) - \psi_2 x_{ij} \\ -\text{psi}(\psi_2) + \log(m\psi_2 y_{ij}) - \log \sum_j y_{ij} \exp(-x_{ij}^\top \psi_1) - x_{ij}^\top \psi_1 \end{array} \right),$$

where $\text{psi}(z) = \nabla_z \log \Gamma(z)$, the digamma function. A calculation, given in the Appendix, yields

$$E_{\psi, \lambda} s(\psi) = \begin{pmatrix} 0 \\ nm(\log(m\psi_2) - \text{psi}(m\psi_2)) \end{pmatrix},$$

which is free of λ . In this model, again, there is an incidental-parameter problem only for ψ_2 . The solutions of $s(\psi) = 0$ and $s_a(\psi) = 0$ coincide for ψ_1 but differ for ψ_2 .

The adjusted profile score function is equal to the score function of the marginal log-likelihood of the ratios y_{ij}/y_{i1} , which is free of λ (Chamberlain 1985).

Inverse Gaussian regression. Suppose y_{ij} has the inverse Gaussian distribution with mean $\mu_{ij} = \lambda_i \exp(x_{ij}^\top \psi_1)$ and variance μ_{ij}^3/ψ_2 . The density is $f(y_{ij}; \psi, \lambda_i) = \sqrt{\psi_2/(2\pi y_{ij}^3)} \exp(-\psi_2(2y_{ij})^{-1}(y_{ij}\mu_{ij}^{-1}-1)^2)$. For fixed ψ , the maximum likelihood estimator of λ_i is $\hat{\lambda}_i(\psi) = \sum_j y_{ij} \exp(-2x_{ij}^\top \psi_1) / \sum_j \exp(-x_{ij}^\top \psi_1)$. The profile log-likelihood and score are

$$l(\psi) = \sum_{i,j} (2^{-1} \log \psi_2 - \psi_2 (2y_{ij})^{-1} (y_{ij} \hat{\mu}_{ij}^{-1} - 1)^2),$$

$$s(\psi) = \sum_{i,j} \left(\begin{array}{c} \psi_2 (y_{ij} \hat{\mu}_{ij}^{-2} - \hat{\mu}_{ij}^{-1}) x_{ij} \\ (2\psi_2)^{-1} - (2y_{ij})^{-1} (y_{ij} \hat{\mu}_{ij}^{-1} - 1)^2 \end{array} \right),$$

where $\hat{\mu}_{ij} = \hat{\lambda}_i(\psi) \exp(x_{ij}^\top \psi_1)$. The profile score bias is

$$E_{\psi, \lambda} s(\psi) = \begin{pmatrix} 0 \\ n(2\psi_2)^{-1} \end{pmatrix},$$

as shown in the Appendix. Here, again, there is an incidental-parameter problem only for ψ_2 , and the solutions of $s(\psi) = 0$ and $s_a(\psi) = 0$ coincide for ψ_1 but differ for ψ_2 .

Exponential matched pairs. In this example, taken from Cox and Reid (1992), there are n pairs $y_i = (y_{i1}, y_{i2})$ of independent exponentially distributed random variables with means $Ey_{i1} = \psi/\lambda_i$ and $Ey_{i2} = \psi\lambda_i$. The maximum likelihood estimate of ψ is $\hat{\psi} = n^{-1} \sum_i \sqrt{y_{i1}y_{i2}}$ and converges to $\psi\pi/4$. Cox and Reid (1992) derived the conditional profile likelihood of Cox and Reid (1987). Its maximizer is $4\hat{\psi}/3$ and converges to $\psi\pi/3$. The profile score is

$$s(\psi) = -2n\psi^{-1} + 2\psi^{-2} \sum_i \sqrt{y_{i1}y_{i2}},$$

with bias $E_{\psi, \lambda} s(\psi) = -n\psi^{-1} (2 - \pi/2)$, free of λ . The adjusted-score estimator is $\hat{\psi}_a = 4\hat{\psi}/\pi$ and converges to ψ .

2.3. Case (c) models

Binary matched pairs. Consider n pairs $y_i = (y_{i1}, y_{i2})$ of independent binary variables with success probabilities $\Pr(y_{i1} = 1) = (1 + e^{-\lambda_i})^{-1}$ and $\Pr(y_{i2} = 1) = (1 + e^{-\lambda_i - \psi})^{-1}$; cf. Cox (1958). The parameter ψ is the log odds ratio and it is well known that $\text{plim}_{n \rightarrow \infty} \widehat{\psi} = 2\psi$. The classic solution to this incidental-parameter problem is to use the conditional likelihood given $y_{i1} + y_{i2}$, $i = 1, \dots, n$ (Rasch 1961, Andersen 1970). The conditional maximum likelihood estimator is consistent and semi-parametrically efficient (Hahn 1997). Here, we show that the adjusted profile score coincides with the score of the conditional log-likelihood.

For pairs of the form $y_i = (0, 0)$ or $y_i = (1, 1)$, the maximum likelihood estimator of λ_i for any fixed ψ is $\widehat{\lambda}_i(\psi) = -\infty$ and $\widehat{\lambda}_i(\psi) = +\infty$, respectively, so $l_i(\psi) = s_i(\psi) = 0$ for such pairs. For pairs of the form $y_i = (0, 1)$ or $y_i = (1, 0)$, $\widehat{\lambda}_i(\psi) = -\psi/2$. The profile log-likelihood and score are

$$l(\psi) = -2n_{01} \log(1 + e^{-\psi/2}) - 2n_{10} \log(1 + e^{\psi/2}),$$

$$s(\psi) = \frac{n_{01}}{1 + e^{\psi/2}} - \frac{n_{10}}{1 + e^{-\psi/2}},$$

where n_{01} and n_{10} are the number of $(0, 1)$ and $(1, 0)$ pairs, respectively, with expected values

$$E_{\psi, \lambda} n_{01} = \sum_i (1 + e^{\lambda_i})^{-1} (1 + e^{-\lambda_i - \psi})^{-1},$$

$$E_{\psi, \lambda} n_{10} = \sum_i (1 + e^{-\lambda_i})^{-1} (1 + e^{\lambda_i + \psi})^{-1} = e^{-\psi} E_{\psi, \lambda} n_{01}.$$

Defining $a_\psi = (1 - e^{-\psi/2})(1 + e^{\psi/2})^{-1}$, the bias of the profile score is

$$E_{\psi, \lambda} s(\psi) = a_\psi E_{\psi, \lambda} n_{01},$$

which depends on λ via $E_{\psi, \lambda} n_{01}$. Now consider the sequence $s_a^{(k)}(\psi)$ of finite-order adjusted profile scores. We have

$$E_{\psi, \widehat{\lambda}(\psi)} n_{01} = (n_{01} + n_{10})(1 + e^{-\psi/2})^{-2},$$

$$E_{\psi, \lambda} \left(E_{\psi, \widehat{\lambda}(\psi)} n_{01} \right) = b_\psi E_{\psi, \lambda} n_{01},$$

where $b_\psi = (1 + e^{-\psi})(1 + e^{-\psi/2})^{-2}$. Hence, for $k = 1, 2, \dots$,

$$E_{\psi, \widehat{\lambda}(\psi)}^{(k)} s(\psi) = a_\psi b_\psi^{k-1} E_{\psi, \widehat{\lambda}(\psi)} n_{01},$$

$$E_{\psi, \lambda} \left(E_{\psi, \widehat{\lambda}(\psi)}^{(k)} s(\psi) \right) = a_\psi b_\psi^k E_{\psi, \lambda} n_{01},$$

and we obtain

$$s_a^{(k)}(\psi) = s(\psi) - \sum_{p=1}^k \binom{k}{p} (-1)^{p-1} a_\psi b_\psi^{p-1} E_{\psi, \widehat{\lambda}(\psi)} n_{01}$$

$$= s(\psi) - (1 - (1 - b_\psi)^k) a_\psi b_\psi^{-1} E_{\psi, \widehat{\lambda}(\psi)} n_{01},$$

with bias

$$E_{\psi, \lambda} s_a^{(k)} = a_\psi (1 - b_\psi)^k E_{\psi, \lambda} n_{01}.$$

For all k , the bias has the same sign as ψ and, since $0 < b_\psi < 1$, decays monotonically to zero at a geometric rate in k . Letting $k \rightarrow \infty$, we find

$$s_a(\psi) = \frac{n_{01}}{1 + e^\psi} - \frac{n_{10}}{1 + e^{-\psi}} = s_c(\psi),$$

where $s_c(\psi)$ is the score function of the conditional log-likelihood. Thus, fully iterating the bias adjustment

of the profile score leads to the conditional likelihood. Accordingly, $\widehat{\psi}_a = \log(n_{01}/n_{10})$, the conditional maximum likelihood estimator.

It may be remarked that, in this model, the fully iterated adjustment can also be obtained without iterating, essentially as in Case (b). First rescale $s(\psi)$ as

$$q(\psi) = \frac{s(\psi)}{n_{01} + n_{10}} = \frac{n_{01}}{n_{01} + n_{10}} - \frac{1}{1 + e^{-\psi/2}},$$

assuming $n_{01} + n_{10} > 0$. As $n \rightarrow \infty$, $q(\psi)$ converges in probability to

$$q_\infty(\psi) = \frac{1}{1 + e^{-\psi}} - \frac{1}{1 + e^{-\psi/2}}$$

for any sequence $\lambda_1, \lambda_2, \dots$ for which $E_{\psi, \lambda} n_{01}/n$ converges. Since $q_\infty(\psi)$ is free of λ , $q_a(\psi) = q(\psi) - q_\infty(\psi)$ is a bias-adjusted version of $q(\psi)$. This yields $q_a(\psi) = s_c(\psi)/(n_{01} + n_{10})$, again leading to the conditional maximum likelihood estimator.

Now consider the following generalization. Suppose $\Pr(y_{i1} = 1) = G(\lambda_i)$ and $\Pr(y_{i2} = 1) = G(\lambda_i + \psi)$, where G is a distribution function with a density g that is symmetric about zero, unimodal, continuous, and non-zero everywhere. A conditional likelihood that is free of λ exists only when G is logistic, as above. The profile score and the adjusted profile score are

$$\begin{aligned} s(\psi) &= (n_{01} - Q(\psi/2)n_{10}) c_\psi, \\ s_a(\psi) &= (n_{01} - Q(\psi/2)^2 n_{10}) d_\psi, \end{aligned}$$

where $Q(z) = G(z)/G(-z)$, $c_\psi = g(\psi/2)/G(\psi/2)$, and $d_\psi = g(\psi/2)Q(-\psi/2)/(G(\psi/2)^2 + G(-\psi/2)^2)$; the derivation is given in the Appendix. It follows that $s_a(\psi)$ is unbiased if and only if

$$\frac{E_{\psi, \lambda} n_{01}}{E_{\psi, \lambda} n_{10}} = Q(\psi/2)^2,$$

regardless of λ . Therefore, unbiasedness requires that

$$\frac{E_{\psi, \lambda} n_{01}}{E_{\psi, \lambda} n_{10}} = \frac{\sum_i G(-\lambda_i)G(\lambda_i + \psi)}{\sum_i G(\lambda_i)G(-\lambda_i - \psi)}$$

be free of λ . Setting all λ_i equal gives the requirement

$$\frac{Q(\lambda_i + \psi)}{Q(\lambda_i)} = h(\psi)$$

for some function h . Setting $\lambda_i = 0$ gives $h(\psi) = Q(\psi)$ and hence

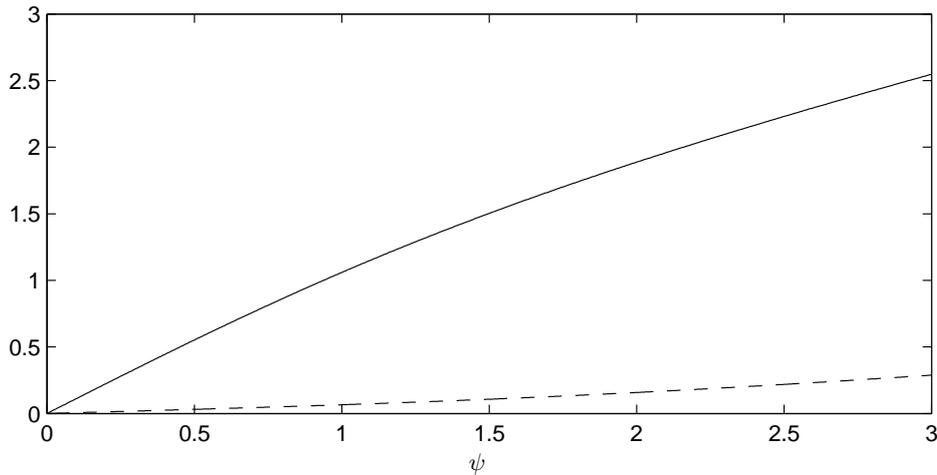
$$Q(\lambda_i + \psi) = Q(\lambda_i)Q(\psi),$$

whose solution is of the form $Q(z) = e^{\gamma z}$ and, therefore, $G(z) = (1 + e^{-\gamma z})^{-1}$ is logistic. When G is logistic, $s_a(\psi)$ is unbiased, as shown earlier. When G is not logistic, $E_{\psi, \lambda} n_{01}/E_{\psi, \lambda} n_{10}$ is not free of λ and, therefore, ψ is not point identified everywhere and an unbiased estimating equation does not exist; see also [Chamberlain \(1980, 2010\)](#).

When G is not logistic, the asymptotic bias of $\widehat{\psi}$ and $\widehat{\psi}_a$ can be signed. Suppose $\psi > 0$ (the case $\psi < 0$ follows by symmetry). Given $Q^{-1}(z) = G^{-1}(z/(1+z))$, we have

$$\begin{aligned} \widehat{\psi} &= 2G^{-1}\left(\frac{n_{01}/n_{10}}{1 + n_{01}/n_{10}}\right), \\ \widehat{\psi}_a &= 2G^{-1}\left(\frac{\sqrt{n_{01}/n_{10}}}{1 + \sqrt{n_{01}/n_{10}}}\right). \end{aligned}$$

Assume that $E_{\psi, \lambda} n_{01}/n$ converges, so that the probability limits of $\widehat{\psi}$ and $\widehat{\psi}_a$ exist. Now, since $Q(\lambda_i +$

Figure 1. Asymptotic biases in the probit model when $m = 2$ Asymptotic bias of $\hat{\psi}$ (solid) and $\hat{\psi}_a$ (dashed) when $\lambda_i \sim N(0, 1)$ and G is standard normal.

$\psi)/Q(\lambda_i) \geq Q(\psi/2)/Q(-\psi/2)$, with equality if and only if $\lambda_i = -\psi/2$,

$$\frac{E_{\psi, \lambda} n_{01}}{E_{\psi, \lambda} n_{10}} \geq Q(\psi/2)^2.$$

Therefore, $\text{plim}_{n \rightarrow \infty} \hat{\psi} > \text{plim}_{n \rightarrow \infty} \hat{\psi}_a \geq \psi$, with equality if and only if $\lambda_i = -\psi/2$ for almost all i . Hence, although $\hat{\psi}_a$ is generally inconsistent, it improves on maximum likelihood uniformly across the parameter space. Figure 1 illustrates the improvement for the case where $\lambda_1, \lambda_2, \dots$ are drawn independently from $N(0, 1)$ and G is the standard normal distribution function. The curves are the asymptotic biases of $\hat{\psi}$ (solid) and $\hat{\psi}_a$ (dashed). The asymptotic bias of $\hat{\psi}_a$ is small when most of the $|\lambda_i + \psi/2|$ are only moderately large, but is unbounded since $Q(\lambda_i + \psi)/Q(\lambda_i) \rightarrow \infty$ as $\lambda_i \rightarrow \pm\infty$.

Binary autoregressive pairs. Consider n independent pairs $y_i = (y_{i1}, y_{i2})$ of binary variables with success probabilities

$$\begin{aligned} \Pr(y_{i1} = 1) &= (1 + e^{-\lambda_i})^{-1}, \\ \Pr(y_{i2} = 1 | y_{i1}) &= (1 + e^{-\lambda_i - \psi y_{i1}})^{-1}. \end{aligned}$$

This is an autoregressive logit model with $y_{i0} = 0$ for all i . Point identification of ψ in this setting requires at least triplets of observations (Cox 1958; Honoré and Tamer 2006). Here, we examine the profile score adjustment when only pairs of data are available.

Pairs of the form $y_i = (0, 0)$ or $y_i = (1, 1)$ give $\hat{\lambda}_i(\psi) = -\infty$ and $\hat{\lambda}_i(\psi) = +\infty$, respectively, with $l_i(\psi) = s_i(\psi) = 0$ for such pairs. The pairs $y_i = (0, 1)$ give $\hat{\lambda}_i(\psi) = 0$, $l_i(\psi) = -\log 4$, and $s_i(\psi) = 0$. Only the pairs $y_i = (1, 0)$, for which $\hat{\lambda}_i(\psi) = -\psi/2$, contribute to the profile likelihood. Let n_{01} and n_{10} be the number of $(0, 1)$ and $(1, 0)$ pairs. The profile log-likelihood and score are

$$\begin{aligned} l(\psi) &= -2n_{10} \log(1 + e^{\psi/2}), \\ s(\psi) &= -n_{10}(1 + e^{-\psi/2})^{-1}, \end{aligned}$$

and the maximum likelihood estimator of ψ (assuming $n_{10} > 0$) is $-\infty$. The expectations of n_{01} and n_{10} are

$$E_{\psi, \lambda} n_{01} = \sum_i (1 + e^{\lambda_i})^{-1} (1 + e^{-\lambda_i})^{-1},$$

$$E_{\psi, \lambda} n_{10} = \sum_i (1 + e^{-\lambda_i})^{-1} (1 + e^{\lambda_i + \psi})^{-1}.$$

Evaluating these at $\lambda_i = \widehat{\lambda}_i(\psi)$, we obtain

$$E_{\psi, \widehat{\lambda}(\psi)} n_{01} = 4^{-1} n_{01} + (1 + e^{\psi/2})^{-1} (1 + e^{-\psi/2})^{-1} n_{10},$$

$$E_{\psi, \widehat{\lambda}(\psi)} n_{10} = 2^{-1} (1 + e^{\psi})^{-1} n_{01} + (1 + e^{\psi/2})^{-2} n_{10},$$

from which it follows that

$$E_{\psi, \widehat{\lambda}(\psi)}^{(k)} \begin{pmatrix} n_{01} \\ n_{10} \end{pmatrix} = B_{\psi}^k \begin{pmatrix} n_{01} \\ n_{10} \end{pmatrix}, \quad k = 1, 2, \dots,$$

where

$$B_{\psi} = \begin{pmatrix} 4^{-1} & (1 + e^{\psi/2})^{-1} (1 + e^{-\psi/2})^{-1} \\ 2^{-1} (1 + e^{\psi})^{-1} & (1 + e^{\psi/2})^{-2} \end{pmatrix}.$$

Hence, on writing the profile score as

$$s(\psi) = a_{\psi} \begin{pmatrix} n_{01} \\ n_{10} \end{pmatrix}, \quad a_{\psi} = \begin{pmatrix} 0 & -(1 + e^{-\psi/2})^{-1} \end{pmatrix},$$

the k th order adjusted profile score follows as

$$s_a^{(k)}(\psi) = a_{\psi} (I - B_{\psi})^k \begin{pmatrix} n_{01} \\ n_{10} \end{pmatrix}.$$

The eigenvalues of $I - B_{\psi}$ are less than one in absolute value, so $s_a(\psi) = 0$ for every ψ . Unlike the case discussed in the previous example, here the lack of point identification results in $s_a(\psi)$ being uninformative about ψ . However, the equation $s_a^{(k)}(\psi) = 0$ has a unique solution, $\widehat{\psi}^{(k)}$, for every value of the ratio n_{01}/n_{10} , and this solution converges as $k \rightarrow \infty$. The limit solution, derived in the Appendix, is

$$\widehat{\psi}_a = g^{-1}(n_{01}/n_{10})$$

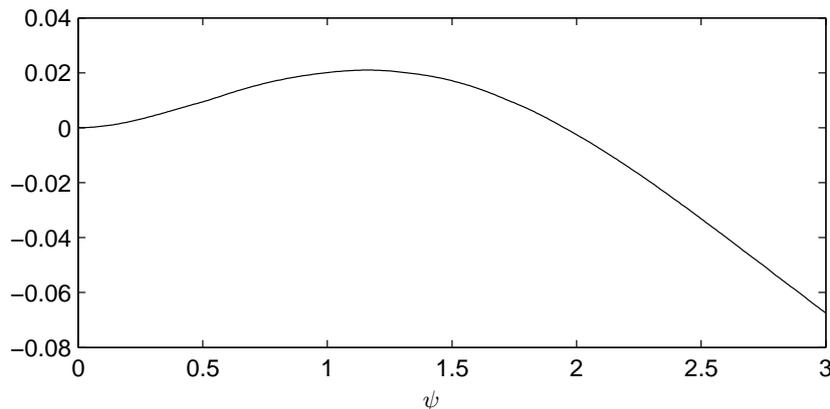
where

$$g(\psi) = u_{\psi} + \sqrt{u_{\psi}^2 + v_{\psi}},$$

$$u_{\psi} = (1 + e^{\psi})(4^{-1} - (1 + e^{\psi/2})^{-2}),$$

$$v_{\psi} = 2(1 + e^{\psi})(1 + e^{\psi/2})^{-1} (1 + e^{-\psi/2})^{-1}.$$

While inconsistent, $\widehat{\psi}_a$ improves rather drastically on maximum likelihood. Figure 2 shows its asymptotic bias for the case where $\lambda_1, \lambda_2, \dots$ are drawn independently from $N(0, 1)$. The bias is uniformly small over the range of values of ψ considered.

Figure 2. Asymptotic bias in the autoregressive logit model when $m = 2$ Asymptotic bias of $\widehat{\psi}_a$ when $\lambda_i \sim N(0, 1)$.

Negative binomial regression. Let y_{ij} be a negative binomial random variable with mean $\mu_{ij} = \lambda_i \exp(x_{ij}^\top \psi_1)$ and variance $\mu_{ij} + \mu_{ij}^2 / \psi_2$. The parameter $\psi_2^{-1} \geq 0$ is an overdispersion parameter, with $\psi_2 \rightarrow \infty$ yielding Poisson regression. The probability mass function is

$$f(y_{ij}; \psi, \lambda_i) = \frac{\Gamma(\psi_2 + y_{ij})}{\Gamma(\psi_2) \Gamma(y_{ij} + 1)} \left(\frac{\mu_{ij}}{\mu_{ij} + \psi_2} \right)^{y_{ij}} \left(\frac{\psi_2}{\mu_{ij} + \psi_2} \right)^{\psi_2}.$$

$\widehat{\lambda}_i(\psi)$ satisfies the equation $\sum_j g_{ij}(\psi, \widehat{\lambda}_i(\psi)) = 0$, where

$$g_{ij}(\psi, \lambda_i) = \frac{y_{ij} - \lambda_i \exp(x_{ij}^\top \psi_1)}{\psi_2 + \lambda_i \exp(x_{ij}^\top \psi_1)}, \quad \lambda_i \geq 0.$$

This equation is equivalent to a m th order polynomial equation and has a unique nonnegative root. The profile score is

$$s(\psi) = \sum_{i,j} \left(\begin{array}{c} \psi_2 g_{ij}(\psi, \widehat{\lambda}_i(\psi)) x_{ij}, \\ \text{psi}(\psi_2 + y_{ij}) - \text{psi}(\psi_2) + \log \psi_2 - \log(\psi_2 + \widehat{\lambda}_i(\psi) \exp(x_{ij}^\top \psi_1)) \end{array} \right),$$

with bias

$$E_{\psi, \lambda} s(\psi) = \sum_i \sum_{y_{i1}=0}^{\infty} \dots \sum_{y_{im}=0}^{\infty} s_i(\psi) \prod_j f(y_{ij}; \psi, \lambda_i).$$

For small m the bias can be computed directly. Both components of $E_{\psi, \lambda} s(\psi)$ are non-zero and depend on ψ , λ , and the covariate values.

Table 1 gives the result of a numerical computation, with the infinite sums in $E_{\psi, \lambda} s(\psi)$ truncated at 400, for the case $m = 2$, $\lambda_i = \psi_1 = \psi_2 = 1$, and $(x_{i1}, x_{i2}) = (0, \log 2)$. The overdispersion in this case is large, the means of y_{i1} and y_{i2} being 1 and 2, and the variances 2 and 6, respectively. We computed the finite-order adjusted profile score bias per observation, $E_{\psi, \lambda} s_a^{(k)}(\psi) / n$, and $\text{plim}_{n \rightarrow \infty} \widehat{\psi}^{(k)} = \arg \text{solve}_{\psi^*} \{E_{\psi, \lambda} s_a^{(k)}(\psi^*) = 0\}$ for $k = 0, \dots, 5$. The row $k = 0$ shows that the profile score bias is very small for ψ_1 and large for ψ_2 . Accordingly, the probability limits of the maximum likelihood estimates $\widehat{\psi}_1$ and $\widehat{\psi}_2$ are very close to ψ_1 and very different from ψ_2 , respectively. This is in line with the simulation results of Allison and Waterman (2002), who found in various designs with $m = 2$ that there is hardly indication of incidental parameter bias for ψ_1 , while the maximum likelihood estimator of ψ_2 was often infinite. The computation here suggests that the adjusted profile score is unbiased, and the adjusted score estimator consistent.

Table 1. Probability limits in the negative binomial regression when $m = 2$

k	$\text{plim}_{n \rightarrow \infty} E_{\psi, \lambda} s_a^{(k)}(\psi)/n$		$\text{plim}_{n \rightarrow \infty} \widehat{\psi}_a^{(k)}$	
	ψ_1	ψ_2	ψ_1	ψ_2
0 (MLE)	.00424	.26554	.9899	53.614
1	.00109	.05083	.9877	1.510
2	.00029	.01224	.9969	1.122
3	.00005	.00342	.9988	1.035
4	-.00003	.00097	.9992	1.010
5	-.00005	.00020	.9993	1.002

$$\lambda_i = \psi_1 = \psi_2 = 1, (x_{i1}, x_{i2}) = (0, \log 2)$$

3. SIMULATIONS

Table 2 presents the results of a simulation study for the nonlinear models considered above where the adjusted profile score was obtained in closed form. For simulations in the linear autoregression, see [Dhaene and Jochmans \(2016\)](#). In all designs, we set $n = 500$ and $m = 2$. The regression models were run with a single regressor, generated as $x_{ij} \sim N(0, 1)$, and $\lambda_i \sim U(.5, 1.5)$. The exponential pairs were generated with $\lambda_i \sim U(.5, 1.5)$, and the binary pairs with $\lambda_i \sim N(0, 1)$. We set ψ as indicated in the table and ran 100,000 Monte Carlo replications for each model. The table reports the mean and standard deviation (sd) of the maximum likelihood estimator, $\widehat{\psi}$, and the adjusted score estimator, $\widehat{\psi}_a$. All theoretical results derived above are confirmed. Whenever there is incidental-parameter bias, the adjusted profile score eliminates it, except in the probit binary-pair and the logit autoregressive-pair cases, where ψ is not point identified and a small bias remains, in line with the theory. The table also reports the ratio of the mean standard error estimate to the standard deviation (se/sd) and the coverage rate of the normal-theory approximate 95% confidence interval centered at the corresponding point estimate (95% c.i.). The se/sd ratio is very close to one (both for $\widehat{\psi}$ and $\widehat{\psi}_a$) and the confidence intervals centered at $\widehat{\psi}_a$ have coverage rates very close to 95%, again confirming the theory. Only for probit binary pairs is the coverage rate somewhat below 95%, in line with the non-negligible bias of $\widehat{\psi}_a$ in this case.

Table 3 presents simulation results for the negative binomial regression with $n = 500$ and $m = 2$. We computed the maximum likelihood estimate and the adjusted-score estimates $\widehat{\psi}_a^{(k)}$ for $k = 1, 2, 3$. The main goal is to explore the effect of using simulations to approximate $E_{\psi, \widehat{\lambda}(\psi)}^{(p)}$. To make the results comparable with Table 1, we used the same design, i.e., with $\lambda_i = \psi_1 = \psi_2 = 1$ and $(x_{i1}, x_{i2}) = (0, \log 2)$. For each $p = 1, \dots, k$, we approximated the p -fold iterated expectation $E_{\psi, \widehat{\lambda}(\psi)}^{(p)}$ by nested simulations with R replications in the outermost expectation and a single replication in all inner expectations, all replications being independent across p and across the levels of nesting. To make the iterative root finding of $s_a^{(k)}(\psi) = 0$ numerically stable, we kept the same basic stream of random numbers across all values of ψ . We set R equal to 1, 3, or 10. The results, based on 10,000 Monte Carlo replications, are in line with the predictions of Table 1 and, with regard to maximum likelihood, with the simulations of [Allison and Waterman \(2002\)](#). For ψ_1 , all estimates, including maximum likelihood, are nearly unbiased. For ψ_2 , the maximum likelihood was found to be infinite around 40% of the time. The adjusted-score estimates of ψ_2 , on the other hand, were always finite, with a mean that is close to the large- n probability limit given in Table 1, especially when R is not too small. As R increases, the standard deviations decrease (because the approximation of $s_a^{(k)}(\psi)$ becomes more accurate) and the se/sd ratio tends to become closer to one. The effect of R on the coverage rate of the

Table 2. Simulations for various nonlinear models with $n = 500$ and $m = 2$ (100,000 replications)

	ψ	mean		sd		se/sd		95% c.i.	
		$\hat{\psi}$	$\hat{\psi}_a$	$\hat{\psi}$	$\hat{\psi}_a$	$\hat{\psi}$	$\hat{\psi}_a$	$\hat{\psi}$	$\hat{\psi}_a$
Poisson regression	1	1.002	1.002	.051	.051	.991	.991	.948	.948
exponential regression	1	1.000	1.000	.055	.055	.993	.993	.948	.948
Weibull regression	1	1.000	1.000	.037	.037	.994	.994	.948	.948
gamma regression	1.5	2.531	1.504	.095	.056	.999	.998	.000	.949
	1	1.000	1.000	.042	.042	1.000	1.000	.950	.950
inverse Gaussian regression	1.5	2.758	1.507	.164	.084	.988	.989	.000	.948
	1	.999	.999	.051	.051	.989	.989	.942	.942
exponential matched pairs	1.5	3.020	1.510	.191	.096	.995	.995	.000	.949
	1	.785	1.000	.028	.035	.999	.999	.000	.948
binary matched pairs (logit)	1	2.012	1.006	.313	.156	.994	.994	.080	.951
binary matched pairs (probit)	1	2.074	1.075	.227	.126	.993	.991	.001	.929
binary autoregr. pairs (logit)	1	–	1.016	–	.249	–	.995	–	.948

regressions: $x_{ij} \sim N(0, 1)$, $\lambda_i \sim U(.5, 1.5)$; exponential pairs: $\lambda_i \sim U(.5, 1.5)$; binary pairs: $\lambda_i \sim N(0, 1)$.

confidence intervals is somewhat mixed. When there is little bias ($k = 3$), the coverage rates improve as R increases. When the bias is non-negligible ($k = 1$), as R increases the confidence intervals become narrower and therefore the bias shows up in deteriorating coverage rates. Finally, for fixed R , the standard deviations tend to increase with k . This is because the expectations $E_{\psi, \hat{\lambda}(\psi)}^{(p)}$ enter $s_a^{(k)}(\psi)$ with binomial coefficients, which grow in k .

Table 3. Simulations for the negative binomial regression with $n = 500$ and $m = 2$ (10,000 replications)

k	R	mean		sd		se/sd		95% c.i.	
		$\hat{\psi}_{1a}^{(k)}$	$\hat{\psi}_{2a}^{(k)}$	$\hat{\psi}_{1a}^{(k)}$	$\hat{\psi}_{2a}^{(k)}$	$\hat{\psi}_{1a}^{(k)}$	$\hat{\psi}_{2a}^{(k)}$	$\hat{\psi}_{1a}^{(k)}$	$\hat{\psi}_{2a}^{(k)}$
0 (MLE)	–	.992	–	.122	–	1.017	–	.954	.762
1	1	.990	1.552	.169	.297	1.003	.934	.948	.522
1	3	.988	1.544	.140	.237	1.019	.939	.954	.212
1	10	.990	1.541	.130	.216	1.012	.933	.954	.119
2	1	.998	1.179	.236	.347	1.027	.959	.950	.933
2	3	.996	1.155	.171	.233	1.016	.910	.951	.941
2	10	.998	1.149	.143	.184	1.010	.902	.954	.931
3	1	.999	1.142	.318	.538	1.465	2.216	.953	.867
3	3	.997	1.087	.211	.308	1.049	.939	.954	.913
3	10	.998	1.063	.158	.207	1.026	.895	.955	.928

$\lambda_i = \psi_1 = \psi_2 = 1$, $(x_{i1}, x_{i2}) = (0, \log 2)$.

CONCLUSION

In models with incidental parameters, the profile score is often biased, leading to inconsistent maximum-likelihood estimates under Neyman-Scott asymptotics. It is natural, then, to seek to remove this bias, as proposed by [Neyman and Scott \(1948\)](#) and [McCullagh and Tibshirani \(1990\)](#). Motivated by asymptotic (i.e., large m) arguments, we propose to iterate the bias adjustment. Given sufficient regularity, at each iteration

the bias of the profile score is reduced by a factor $O(m^{-1})$. Despite its asymptotic motivation, we often find that the bias adjustment also delivers consistency under Neyman-Scott asymptotics, either because the profile-score bias is free of incidental parameters, so that no iteration is needed, or because the fully iterated adjusted profile score exists, is unbiased, and not identically zero. Of course, point identification may fail under Neyman-Scott asymptotics and, therefore, it cannot be guaranteed that the iterated adjustment gives a consistent estimate, although in the models examined it invariably improves on maximum likelihood.

APPENDIX

Weibull regression. Write the profile score as

$$s(\psi) = \sum_{i,j} \left(\begin{array}{c} \psi_2 \sum_j z_{ij} x_{ij} / \sum_j z_{ij} - \psi_2 x_{ij} \\ \psi_2^{-1} + \psi_2^{-1} \log z_{ij} - \psi_2^{-1} \sum_j z_{ij} \log z_{ij} / \sum_j z_{ij} \end{array} \right)$$

with $z_{ij} = w_{ij}(\psi) \lambda_i^{-\psi_2} = (y_{it}/\mu_{it})^{\psi_2}$ being independent unit-exponential random variables. The first component of $s(\psi)$ has zero expectation, as in the exponential regression case. For the second component, write $\sum_j z_{ij} = z_{ij} + A$, where $A = \sum_{j' \neq j} z_{ij'}$ is independent of z_{ij} and is Erlang distributed with shape parameter $m-1$ and scale parameter 1. The density of A is $g_A(a) = a^{m-2} \exp(-a)/(m-2)!$, so

$$E \left(\frac{z_{ij} \log z_{ij}}{\sum_j z_{ij}} \right) = \int_0^\infty \int_0^\infty \frac{z \log z}{z+a} \exp(-z) \frac{a^{m-2} \exp(-a)}{(m-2)!} dz da = \frac{m-1-m\gamma}{m^2},$$

where γ is Euler's gamma. Setting $m=1$ gives $E \log z_{ij} = -\gamma$. Hence the second component of $s(\psi)$ has expectation $n\psi_2^{-1}$. Now let $\psi' = (\psi_1, \psi_2')$, with arbitrary $\psi_2' > 0$. The first component of $s(\psi')$ is

$$s_{\psi_1}(\psi') = \sum_{i,j} \psi_2' \left(\frac{\sum_j z'_{ij} x_{ij}}{\sum_j z'_{ij}} - x_{ij} \right)$$

with $z'_{ij} = z_{ij}^{\psi_2'/\psi_2}$ and z_{ij} as above. It follows that $E_{\psi, \lambda} s_1(\psi') = 0$.

Gamma regression. Write the profile score as

$$s(\psi) = \sum_{i,j} \left(\begin{array}{c} \psi_2 \sum_j z_{ij} x_{ij} / \sum_j z_{ij} - \psi_2 x_{ij} \\ -\text{psi}(\psi_2) + \log(m\psi_2) + \log z_{ij} - \log \sum_j z_{ij} \end{array} \right),$$

$z_{ij} = y_{ij}/\mu_{ij}$ being independent gamma distributed random variables with shape parameter ψ_2 and scale 1. By the same argument as in the exponential regression model, the first component of $s(\psi)$ has zero expectation. Given that $E \log z_{ij} = \text{psi}(\psi_2)$ and that $\sum_j z_{ij}$ is gamma distributed with shape parameter $m\psi_2$ and scale 1, the second component of $s(\psi)$ has expectation $nm(\log(m\psi_2) - \text{psi}(m\psi_2))$. This expectation is $O(n)$ uniformly in m because $m(\log(m\psi_2) - \text{psi}(m\psi_2)) = (2\psi_2)^{-1} + O(m^{-1})$ as $m \rightarrow \infty$.

Inverse Gaussian regression. Write $\hat{\mu}_{ij}$ as

$$\hat{\mu}_{ij} = \mu_{ij} \frac{\hat{\lambda}_i(\psi)}{\lambda_i} = \mu_{ij} \frac{\sum_j y_{ij} \lambda_i^{-2} \exp(-2x_{ij}^\top \psi_1)}{\sum_j \lambda_i^{-1} \exp(-x_{ij}^\top \psi_1)} = \mu_{ij} \frac{\sum_j y_{ij} \mu_{ij}^{-2}}{\sum_j \mu_{ij}^{-1}} = c^{-1} \mu_{ij} \sum_j z_{ij}$$

where $z_{ij} = y_{ij} \mu_{ij}^{-2}$ and $c = \sum_j \mu_{ij}^{-1}$. Denote the distribution of y_{ij} as $\mathcal{IG}(\mu_{ij}, \psi_2)$. Then, by properties of the inverse Gaussian distribution derived by Tweedie (1957),

$$z_{ij} \sim \mathcal{IG}(\mu_{ij}^{-1}, \mu_{ij}^{-2} \psi_2), \quad \sum_j z_{ij} \sim \mathcal{IG}(c, c^2 \psi_2), \quad \hat{\mu}_{ij} \sim \mathcal{IG}(\mu_{ij}, \mu_{ij} c \psi_2),$$

and

$$E y_{ij} = \mu_{ij}, \quad E y_{ij}^{-1} = \mu_{ij}^{-1} + \psi_2^{-1}, \quad E \widehat{\mu}_{ij}^{-1} = \mu_{ij}^{-1} + \mu_{ij}^{-1} c^{-1} \psi_2^{-1}.$$

Hence

$$E \sum_{i,j} (y_{ij}^{-1} - \widehat{\mu}_{ij}^{-1}) = n(m-1)\psi_2^{-1}. \quad (\text{A.1})$$

We now calculate the expectation of z_{ij}/S^2 , where $S = \sum_j z_{ij}$. The joint moment generating function of z_{ij} and S is

$$\begin{aligned} \text{MGF}_{z_{ij}, S}(t_1, t_2) &= E \exp(t_1 z_{ij} + t_2 S) \\ &= \exp\left(\psi_2 c - \mu_{ij}^{-1} \sqrt{\psi_2(\psi_2 - 2t_1 - 2t_2)} - (c - \mu_{ij}^{-1}) \sqrt{\psi_2(\psi_2 - 2t_2)}\right) \end{aligned}$$

Following [Cressie et al. \(1981\)](#), we obtain

$$E \frac{z_{ij}}{S^2} = \int_0^\infty t_2 \lim_{t_1 \rightarrow 0} \frac{\partial}{\partial t_1} \text{MGF}_{z_{ij}, S}(t_1, -t_2) dt_2 = \mu_{ij}^{-1} \frac{1 + c\psi_2}{c^3 \psi_2}.$$

Hence, from $y_{ij} \widehat{\mu}_{ij}^{-2} = c^2 z_{ij} S^{-2}$, we have

$$E(y_{ij} \widehat{\mu}_{ij}^{-2} - \widehat{\mu}_{ij}^{-1}) = 0. \quad (\text{A.2})$$

On writing the second component of $s(\psi)$ as

$$mn(2\psi_2)^{-1} - \sum_{i,j} 2^{-1}(y_{ij} \widehat{\mu}_{ij}^{-2} - \widehat{\mu}_{ij}^{-1} + y_{ij}^{-1} - \widehat{\mu}_{ij}^{-1}),$$

$E_{\psi, \lambda} s(\psi)$ follows from (A.1) and (A.2).

Binary matched pairs. Since g is symmetric about zero and unimodal, $\widehat{\lambda}_i(\psi) = -\psi/2$. The profile log-likelihood and score are

$$\begin{aligned} l(\psi) &= 2n_{01} \log G(\psi/2) + 2n_{10} \log G(-\psi/2), \\ s(\psi) &= n_{01} \frac{g(\psi/2)}{G(\psi/2)} - n_{10} \frac{g(\psi/2)}{G(-\psi/2)} \\ &= (n_{01} - Q(\psi/2)n_{10}) c_\psi. \end{aligned}$$

Given

$$\begin{aligned} E_{\psi, \lambda} n_{01} &= \sum_i G(-\lambda_i) G(\lambda_i + \psi), \\ E_{\psi, \lambda} n_{10} &= \sum_i G(\lambda_i) G(-\lambda_i - \psi), \end{aligned}$$

we have

$$\begin{aligned} E_{\psi, \widehat{\lambda}(\psi)} n_{01} &= (n_{01} + n_{10}) \alpha_{01}, & \alpha_{01} &= \alpha_{01}(\psi) = G(\psi/2)^2, \\ E_{\psi, \widehat{\lambda}(\psi)} n_{10} &= (n_{01} + n_{10}) \alpha_{10}, & \alpha_{10} &= \alpha_{10}(\psi) = G(-\psi/2)^2, \end{aligned}$$

and, for general k ,

$$\begin{aligned} E_{\psi, \widehat{\lambda}(\psi)}^{(k)} n_{01} &= \alpha_{01} E_{\psi, \widehat{\lambda}(\psi)}^{(k-1)} (n_{01} + n_{10}) = \alpha_{01} (\alpha_{01} + \alpha_{10})^{k-1} (n_{01} + n_{10}) \\ &= (\alpha_{01} + \alpha_{10})^{k-1} E_{\psi, \widehat{\lambda}(\psi)} n_{01}, \\ E_{\psi, \widehat{\lambda}(\psi)}^{(k)} n_{10} &= (\alpha_{01} + \alpha_{10})^{k-1} E_{\psi, \widehat{\lambda}(\psi)} n_{10}. \end{aligned}$$

Therefore, with $s(\psi)$ written as

$$s(\psi) = \beta_{01}n_{01} + \beta_{10}n_{10}, \quad \beta_{01} = \beta_{01}(\psi) = \frac{g(\psi/2)}{G(\psi/2)}, \quad \beta_{10} = \beta_{10}(\psi) = \frac{g(\psi/2)}{G(-\psi/2)},$$

we have

$$\begin{aligned} E_{\psi, \widehat{\lambda}(\psi)}^{(k)} s(\psi) &= (\alpha_{01} + \alpha_{10})^{k-1} (\beta_{01} E_{\psi, \widehat{\lambda}(\psi)} n_{01} - \beta_{10} E_{\psi, \widehat{\lambda}(\psi)} n_{10}) \\ &= (\alpha_{01} + \alpha_{10})^{k-1} (\alpha_{01} \beta_{01} - \alpha_{10} \beta_{10}) (n_{01} + n_{10}) \end{aligned}$$

and

$$\begin{aligned} s_a^{(k)}(\psi) &= s(\psi) - \sum_{p=1}^k \binom{k}{p} (-1)^{p-1} (\alpha_{01} + \alpha_{10})^{p-1} (\alpha_{01} \beta_{01} - \alpha_{10} \beta_{10}) (n_{01} + n_{10}) \\ &= s(\psi) - (1 - (1 - \alpha_{01} - \alpha_{10})^k) \left(\frac{\alpha_{01} \beta_{01} - \alpha_{10} \beta_{10}}{\alpha_{01} + \alpha_{10}} \right) (n_{01} + n_{10}). \end{aligned}$$

Given that $0 < \alpha_{01} + \alpha_{10} < 1$, we obtain

$$\begin{aligned} s_a(\psi) &= s(\psi) - \left(\frac{\alpha_{01} \beta_{01} - \alpha_{10} \beta_{10}}{\alpha_{01} + \alpha_{10}} \right) (n_{01} + n_{10}) \\ &= (n_{01} - Q(\psi/2)^2 n_{10}) d_\psi. \end{aligned}$$

Binary autoregressive pairs. Write B_ψ as

$$B_\psi = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

where $a = 4^{-1}$ and b, c, d are functions of ψ , and decompose $I - B_\psi$ as $P\Delta P^{-1}$, where

$$\begin{aligned} \Delta &= \begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{pmatrix}, \quad \delta_1 = 1 - (a + d - D)/2, \quad \delta_2 = 1 - (a + d + D)/2, \\ P &= \begin{pmatrix} (a - d - D)/(2c) & (a - d + D)/(2c) \\ 1 & 1 \end{pmatrix}, \\ P^{-1} &= \begin{pmatrix} -c/D & (a - d + D)/(2D) \\ c/D & -(a - d - D)/(2D) \end{pmatrix}, \end{aligned}$$

and $D = \sqrt{(a - d)^2 + 4bc}$. Note that $1 > \delta_1 > \delta_2 > 0$. For given k , $\widehat{\psi}_a^{(k)}$ solves

$$a_\psi (I - B_\psi)^k \begin{pmatrix} n_{01} \\ n_{10} \end{pmatrix} = 0,$$

where the first element of a_ψ is zero. Given $(I - B_\psi)^k = P\Delta^k P^{-1}$, this equation is equivalent to

$$\delta_1^k (g - n_{01}/n_{10}) + \delta_2^k (h + n_{01}/n_{10}) = 0$$

where $g = (a - d + D)/(2c)$ and $h = -(a - d - D)/(2c)$. As $k \rightarrow \infty$, the first term dominates. Hence, the limiting solution solves $g = n_{01}/n_{10}$, that is, $g(\psi) = n_{01}/n_{10}$, on writing g as a function of ψ .

REFERENCES

- Allison, P. D. and R. P. Waterman (2002). Fixed-effects negative binomial models. *Sociological Methodology* 32, 247–266.
- Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B* 32, 283–301.
- Arellano, M. and S. Bonhomme (2009). Robust priors in nonlinear panel data models. *Econometrica* 77, 489–536.

- Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70, 343–365.
- Blundell, R., R. Griffith, and F. Windmeijer (1999). Individual effects and dynamics in count data models. The Institute for Fiscal Studies, Working Paper No. W99/3.
- Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies* 47, 225–238.
- Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In J. J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*. Cambridge University Press.
- Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica* 78, 159–168.
- Cox, D. (1958). Two further applications of a model for binary regression. *Biometrika* 45, 562–565.
- Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B* 49, 1–39.
- Cox, D. R. and N. Reid (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* 79, 408–411.
- Cressie, N., A. S. Davis, J. L. Folks, and G. E. Policello II (1981). The moment-generating function and negative integer moments. *The American Statistician* 35, 148–150.
- Dhaene, G. and K. Jochmans (2016). Likelihood inference in an autoregression with fixed effects. *Econometric Theory* 32, 1178–1215.
- DiCiccio, T. J., M. A. Martin, S. E. Stern, and A. Young (1996). Information bias and adjusted profile likelihoods. *Journal of the Royal Statistical Society, Series B* 58, 189–203.
- Hahn, J. (1997). A note on the efficient semiparametric estimation of some exponential panel models. *Econometric Theory* 13, 583–588.
- Hahn, J. and W. K. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72, 1295–1319.
- Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica* 74, 611–629.
- Kalbfleisch, J. D. and D. A. Sprott (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society, Series B* 32, 175–208.
- Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics* 95, 391–413.
- Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies* 69, 647–666.
- Li, H., B. Lindsay, and R. Waterman (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society, Series B* 65, 191–208.
- McCullagh, P. and R. Tibshirani (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society, Series B* 52, 325–344.
- Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.
- Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica* 49, 1417–1426.
- Pace, L. and A. Salvani (2006). Adjustments of profile likelihood from a new perspective. *Journal of Statistical Planning and Inference* 136, 3554–3564.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 4*, pp. 321–333. University of California Press.

- Sartori, N. (2003). Modified profile likelihood in models with stratum nuisance parameters. *Biometrika* 90, 533–549.
- Severini, T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika* 85, 403–411.
- Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford Statistical Science Series. Oxford University Press.
- Tweedie, M. (1957). Properties of inverse gaussian distributions I. *Annals of Mathematical Statistics* 28, 362–377.