Ann Inst Stat Math (2011) 63:535–558
DOI 10.1007/s10463-009-0243-3

Estimating functions for repeated measures with incidental parameters

Martin Crowder

Received: 31 March 2008 / Revised: 10 November 2008 / Published online: 4 June 2009

Abstract Repeated measures, or longitudinal data, are considered. The statistical characteristics for each individual case are supposed to be governed by a structural parameter, common to all, and an incidental parameter, specific to the individual. Introducing this terminology, Neyman and Scott studied the properties of estimators in a likelihood framework. In this paper the model specification is taken to be more limited, not sufficient to construct a proper likelihood function. The proposal here is to seek an estimating function, based on the data and the structural parameter alone, whose maximum has an identifiable limit as the sample size grows. Then a transformation of the maximum is sought so that the modified version is a consistent estimator. Some examples are worked through and asymptotic distributions of the resulting consistent estimators are outlined to enable tests and confidence regions to be derived. Relative efficiency of competing estimators is also considered.

Keywords Estimating functions · Incidental parameters · Neyman–Scott problem · Repeated measures · Structural parameters

1 Framework

We suppose the data to comprise independent observation vectors $y_i = (y_{i1}, \ldots, y_{ir_i})$ for $i = 1, \ldots, n$; thus, there are n cases, or individuals, with r_i values recorded for the ith. For example, this is typical of repeated-measures, or longitudinal data, where the y_{ij} are recorded values of the same variable on the same individual at different times (or at different points along some other dimension); see, e.g. Crowder and Hand

M. Crowder (✉)
Institute for Mathematical Sciences, Imperial College London,
53 Prince's Gate, Exhibition Road, London SW7 2PG, UK
e-mail: m.crowder@imperial.ac.uk

(1990) and Hand and Crowder (1996). Another example is provided by clustered data, where the variable is recorded for different individuals in a homogeneous group, usually at the same time. We suppose that the joint probability distribution of y_i depends on parameter vectors ϕ and ψ_i, ϕ being common to all cases and ψ_i specific to the ith. When the joint probability or density functions, $p(y_i; \phi, \psi_i)$, are known likelihood methods can be brought to bear. Here, however, we do not assume such full knowledge; the aim is to base estimation on more limited model specification, often given as an expression for the mean of y_{ij}, and perhaps its variance, in terms of ϕ and ψ_i.

In the terminology of Neyman and Scott (1948) ϕ is the structural parameter and the ψ_i are incidental parameters. They gave examples in which ϕ is consistently estimable by maximum likelihood and others in which it is not; it goes almost without saying that ψ_i, which appears in the distributions of only a finite number of observed quantities, is not normally consistently estimable. Neyman and Scott went on to set out some conditions to be satisfied by suitable estimating functions that would ensure consistency of ϕ-estimation. Lancaster (2000) reviewed progress on the problem since Neyman and Scott's original paper. In essence, this is a classic problem that has remained unsolved for over half a century. In this paper we propose a general approach to estimation of ϕ when the likelihood is not fully specified. We also consider the case where the incidental parameters, the ψ_i, are of interest in their own right, rather than just appearing as nuisance parameters.

Most of the literature on eliminating nuisance parameters (the ψ_i here) is likelihood-based: marginal, conditional and profile likelihoods, pivotal functions, together with a variety of refinements and approximations, form the core material in this field. Much of the modern theory can be found in the book by Barndorff-Nielsen and Cox (1994). The companion papers of Reid (1995) and Liang and Zeger (1995) give excellent reviews of methods for eliminating nuisance parameters, the first via likelihood-based conditioning and the second via estimating functions. However, they are mainly concerned with nuisance parameters of finite dimension.

The theoretical results given below in Sect. 2 are expressed formally in terms of asymptotic theory, but such analyses are regarded here just as a way of obtaining usable approximations. In particular, they include a method for converting a convergent ϕ-estimator into a consistent one, albeit one that is not necessarily efficient or optimal. In Sect. 3 some examples and applications are worked through, in Sect. 4 some notes on asymptotic inference for ϕ are made, and some discussion is given in Sect. 5.

2 Point estimation

We take as the first priority estimation of the structural parameter ϕ. Once that is achieved the ψ_i can be addressed if required. It will usually be the case in practice that some ψ_i can be estimated better than others because there is more information on some individuals than on others.

The focus here is on constructing an estimating function $h_n(\phi)$, a function of $(y_1, \ldots, y_n)$ and ϕ, whose maximisation yields a convergent ϕ-sequence, $\hat{\phi}_n$. We then seek to transform $\hat{\phi}_n$ into a consistent estimator for ϕ_0, rather than modifying the estimating function as in the vast majority of published work. Indeed, in the literature it often appears to be implicitly assumed that an unbiased estimating equation will yield a consistent estimator; this is not necessarily so, even for so called ‘regular likelihood’ situations with a finite-dimensional parameter (e.g. Crowder 1986).

In what follows estimating functions are often derived in some way from a likelihood-like function. Their use enables consideration of $\hat{\phi}_n$ to be confined to the parameter space of ϕ, assumed here to be compact. We then avoid having to deal with the parameter space of $\psi^{(n)} = (\psi_1, \ldots, \psi_n)$, whose dimension grows linearly with n. Portnoy (1988) obtained asymptotic results for the case where the number of parameters grows with n, but in a different way to that considered here: in his case, the whole parameter set is involved in each observation, the model's becoming more parameter-rich as n grows.

In the absence of a likelihood, and with limited model specifications, the scope for constructing suitable estimating functions is restricted. This means that we might have to settle for one that produces an inconsistent estimator, $\hat{\phi}_n$. One of the main themes here is to show how $\hat{\phi}_n$ can often be simply corrected to yield an estimator that is consistent.

2.1 Methods for eliminating the ψ_i

The general approach here is to base the estimating function on a suitable likelihood. Roughly speaking, suitability is judged here as being relevant to the application and of form simple enough for its performance to be assessed in terms of the limited model specifications made. The likelihood selected is not presumed to arise from the true stochastic mechanism underlying the observations, which is regarded as unknown: it is just treated as a vehicle for inference. Where applicable, any of the standard methods for eliminating nuisance parameters, here the ψ_i, from a likelihood can be employed.

The simplest way of getting rid of the individual ψ_i is just to replace them by a single ψ, either a specified value or as a parameter to be estimated. Less crude is to form a profile likelihood or, when available, a conditional or marginal likelihood.

Another method is to integrate the nuisance parameters out from a likelihood with respect to some weight function. In the Bayesian approach this weight function would be a prior distribution; in the case of random effects it would be the frequency distribution from which they are sampled. Berger et al. (1999) advocated this method: they listed various good properties, including its being ‘safer’ than profile likelihood, for instance, in that allowance is made for the uncertainty in the nuisance parameters. However, in our case, the integration often destroys the required simplicity: see Example 3 below.

Kalbfleisch and Sprott (1970, Sect. 4.2), independently of Andersen (1967), proposed eliminating nuisance parameters by finding a function $g(y_i, \phi)$ whose distribution is independent of ψ_i, so $g(y_i, \phi)$ is pivotal for ψ_i. They suggested that the corresponding ‘likelihood function’, based on the random variables $g(y_i, \phi)$, might be used for inference about ϕ. Cox (1993) investigated the proposal further, focussing on the question of whether the resulting estimating equation is unbiased. For us, the drawback is that, to define a pivotal function, one strictly needs a full probability framework.

2.2 Consistency

The first aspect that needs to be addressed is consistency of estimation. Once that is established the (asymptotic) distribution of estimators is easier to deal with: see Sect. 4.

The notation $m_n(\phi) = \mathrm{E}_0\{h_n(\phi)\}$ will be used, where E_0 is the expectation taken with respect to the true parameter $(\phi_0, \psi_0^{(n)})$; likewise, P_0 will denote probability evaluated under $(\phi_0, \psi_0^{(n)})$. In general, m_n will be a function of $(\phi_0, \psi_0^{(n)})$ as well as of ϕ; for tidiness these extra arguments will be omitted. A scaling sequence, $s_n \to \infty$, is also needed so that $s_n^{-1} m_n(\phi) = \mathrm{O}(1)$ as $n \to \infty$.

For the following result some conditions on $h_n(\phi)$ and $m_n(\phi)$ are required, involving continuity and convergence. The literature contains a wide variety of such conditions: the ones used here are framed to be as transparent as possible while suiting the present purpose, in particular, in having to accommodate an infinite-dimensional parameter.

Condition C1 below comprises routine restrictions on the behaviour of the estimating function $h_n(\phi)$. C1(iii) is a technical requirement necessary for a non-denumerable ϕ-space: a version of it appears in the classic Wald (1949) paper. The crux of the matter is separability of $s_n^{-1} h_n(\phi)$ as a process in ϕ (Doob 1953, Sect. II.2); in the present case it is justified by our assumed continuity of $s_n^{-1} h_n(\phi)$ in ϕ (Loeve 1963, Sect. 35.2A).

Condition C2 is more specifically focused on the case of incidental parameters. In particular, we have to deal with the increasing set of arguments of $m_n(\phi)$ as n increases. So, in C2, G is used to denote either the infinite sequence $(\psi_1, \psi_2, \ldots)$ or the probability distribution from which they are drawn, if that be the case. The latter interpretation is the more relevant in many applications in which the ψ_i are so called random effects, latent variables or frailties. This was the point of view adopted by Kiefer and Wolfowitz (1956) in their follow-up to Neyman and Scott (1948) work.

Condition C1 (behaviour of $s_n^{-1} h_n$)

(i) Continuity: $s_n^{-1} h_n(\phi)$ is continuous in ϕ, uniformly in n.
(ii) Convergence: $s_n^{-1} \mid h_n(\phi) - m_n(\phi) \mid \to 0$ in $(\phi_0, \psi_0^{(n)})$-probability for each ϕ as $n \to \infty$.
(iii) $\mathrm{E}_0\{\sup_{|\phi' - \phi| < \rho} s_n^{-1} h_n(\phi')\}$ exists for each ϕ and $\rho > 0$.

Condition C2 (existence of unimodal limiting mean function $\bar{m}$)

(i) $s_n^{-1} m_n(\phi) \to \bar{m}(\phi, G)$ as $n \to \infty$ uniformly on the ϕ-space.
(ii) $\bar{m}(\phi, G)$ has a unique maximum point at $\phi_1 = \phi_1(G)$ such that, when $\mid \phi - \phi_1 \mid \geq \delta$, $\bar{m}(\phi, G) < \bar{m}(\phi_1, G) - \eta_\delta$ for some $\eta_\delta > 0$.

Proposition 1 *Under conditions C1 and C2, $\hat{\phi}_n \to_p \phi_1$ (in probability) as $n \to \infty$.*

Proof This is given in Sect. 6. □

If $\phi_1 = \phi_0$ the proposition gives a straightforward consistency result for ϕ_0. If $\phi_1 \neq \phi_0$ we have inconsistency. However, in the latter case it might be possible to identify ϕ_1 explicitly enough in terms of ϕ_0 to see how to correct $\hat{\phi}_n$ to obtain a consistent estimator for ϕ_0. Given ϕ_1 as a 1–1 function of ϕ_0, say $\phi_1 = f_1(\phi_0)$, a consistent estimator of ϕ_0 can be obtained by equating $f_1(\phi_0)$ to $\hat{\phi}_n$, so $\hat{\phi}_0 = f_1^{-1}(\hat{\phi}_n)$; see the examples below.

The approach here is alternative to the usual one, in which a modification of the estimating function, $h_n(\phi)$, is sought that will yield better asymptotic properties. For example, in the case of binary matched pairs, the mle of the odds ratio, ϕ, tends in probability to ϕ_0^2 as the number of pairs increases (Andersen 1973; Breslow 1981). Barndorff-Nielsen (1983) replaced the likelihood by a modified profile likelihood, which produces an estimator with probability limit $\{(5\phi_0+1)/(\phi_0+5)\}^2$; he observed that, for a range of ϕ-values, this limit is closer to the target value, ϕ_0, than the former ϕ_0^2. However, in the spirit of the present paper, we could simply correct the raw mle by taking its square root. Reid (1995) mentioned another example, this time involving unmatched pairs: y_{i1} and y_{i2} are independent and exponentially distributed with respective means $\phi\psi_i$ and ϕ/ψ_i. The mle of ϕ converges to $\phi_0(\pi/4)$, whereas the estimator based on Cox and Reid (1987) approximate conditional likelihood gets closer, with limit $\phi_0(\pi/3)$. In our approach, either estimator can be converted to consistency by simply rescaling. Firth (1993) was concerned with a rather more refined adjustment, namely that of removing the $O(n^{-1})$ asymptotic bias from a maximum likelihood estimator that is already consistent.

An examination of its proof shows that Proposition 1 can be easily generalised to cases where ϕ_1 is not a unique maximum point of $\bar{m}(\phi, G)$, e.g. where there is a lack of identifiability such as when a likelihood surface has a ridge. Then the small open sphere surrounding ϕ_1 into which $\hat{\phi}_n$ eventually migrates is replaced by an open set containing all such ϕ_1-points. This is an approach exploited in Crowder (1986).

2.3 Consistency of individual estimators of ϕ and ψ_i

As previously mentioned, the individual parameters, the ψ_i, will sometimes be of interest themselves. This would arise when they are to be used to make decisions about individuals. For example, in a medical context ψ_i might be associated with the state of the ith patient, which is not directly observable. In that case ψ_i is a latent characteristic or frailty that one needs to estimate by monitoring the patient, *i.e.* obtaining a sequence $y_{i1}, y_{i2}, \ldots$ of repeated measurements on him or her. In such situations we expect r_i, the number of observations on the ith case, to be large enough to give a decent estimate of ψ_i, at least for some individuals. In purely formal terms, we may consider consistent estimation of ψ_i, entailing $r_i \to \infty$. Though not explicitly ruled out in their 1948 paper, this perhaps goes a little beyond the Neyman–Scott framework.

Consistency of $\hat{\psi}^{(n)} = (\hat{\psi}_1, \ldots, \hat{\psi}_n)$, i.e. of its n components simultaneously, is problematic, requiring something like $\max_{i=1,\ldots,n} \mid \hat{\psi}_i - \psi_{i0} \mid \to_p 0$ as $n \to \infty$. It will often be the case in practice that consistency only obtains for some ψ_i, i.e. those for which y_i is sufficiently informative. A fairly obvious result for an individual $\hat{\psi}_i$ can be framed as follows; in this, $\hat{\psi}_{i\phi}$ represents the ψ_i-estimator for given ϕ.

Proposition 2 *Suppose that* (i) $\hat{\phi}_n$ *is consistent for* ϕ_0, (ii) $\hat{\psi}_{i\phi_0}$ *is consistent for* ψ_{i0}, *and* (iii) $\hat{\psi}_{i\phi}$ *is continuous in* ϕ. *Then* $\hat{\psi}_{i\hat{\phi}_n}$ *is consistent for* ψ_{i0}.

Proof This is given in Sect. 6. □

If ψ_i can be estimated consistently from y_i, the data on the ith individual, it is likely that the same is true for ϕ. Let $(\tilde{\phi}_i, \tilde{\psi}_i)$ be the estimator from y_i alone. A question arises as to the relative merits of the individual estimators, the $\tilde{\phi}_i$, and the overall one, $\hat{\phi}_n$: perhaps some sort of average of the $\tilde{\phi}_i$ might be preferred to $\hat{\phi}_n$. In fact, the following lemma shows that $\hat{\phi}_n$ can be expressed as the standard 'optimal linear combination' of the $\tilde{\phi}_i$. The estimating function whose maximisation produces $(\tilde{\phi}_i, \tilde{\psi}_i)$ from y_i $(i = 1, \ldots, n)$ will be denoted by $h_{ni} = h_{ni}(\phi, \psi_i)$, and then $h_n(\phi, \psi^{(n)}) = \sum_{i=1}^n h_{ni}(\phi, \psi_i)$.

Proposition 3 *When $\hat{\phi}_n$ and the $\tilde{\phi}_i$ $(i = 1, \ldots, n)$ are all consistent for ϕ_0, $\hat{\phi}_n$ can be expressed asymptotically as an optimally-weighted average of the $\tilde{\phi}_i$:*

$$\hat{\phi}_n \sim \left(\sum_{i=1}^n w_i\right)^{-1} \left(\sum_{i=1}^n w_i \tilde{\phi}_i\right), \quad w_i = \frac{\partial^2 h_{ni}}{\partial \phi^2} - \frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i} \left(\frac{\partial^2 h_{ni}}{\partial \psi_i^2}\right)^{-1} \frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi},$$

in which the derivatives of h_{ni} are evaluated at (ϕ_0, ψ_0).

Proof An outline is given in Sect. 6 in which the formal assumptions are indicated. □

It follows from the proposition that, if the $\tilde{\phi}_i$ are consistent for ϕ_0, the weighted average is (approximately) the overall estimator, $\hat{\phi}_n$, and this is also consistent. It might be preferable in some practical situations to proceed in this way: the $(n + 1)$th case produces w_{n+1} and $\tilde{\phi}_{n+1}$, and then the sums, $\sum_{i=1}^n w_i$ and $\sum_{i=1}^n w_i \tilde{\phi}_i$, can be updated accordingly.

3 Examples and applications

The following examples are meant to bring out various aspects of the approach proposed in this paper. We use the notation $r_+ = \sum_{i=1}^n r_i$, and $\bar{y}_i$ and s_i^2 will denote the sample mean and variance of $(y_{i1}, \ldots, y_{ir_i})$, respectively. The routine conditions in C1 will normally hold for these examples with $s_n = r_+$ and $r_+^{-2} \sum_{i=1}^n r_i^2 \to 0$ as $n \to \infty$; some restrictions on the values of covariates might also be necessary. We concentrate on C2 in these examples.

Example 1 Neyman and Scott (1948) gave a likelihood-based analysis of a variety of models. To illustrate the approach here, albeit in a full likelihood context, we will just consider briefly their first two examples.

In the first example the observations y_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, r_i)$ are independent, y_{ij} having distribution $\mathrm{N}(\phi, \psi_i^2)$. The overall log-likelihood function is

$$l_n(\phi, \psi^{(n)}) = -\frac{1}{2} \sum_{i=1}^n \left\{ r_i \log(2\pi \psi_i^2) + \sum_{j=1}^{r_i} (y_{ij} - \phi)^2 / \psi_i^2 \right\}.$$

With the ψ_i replaced by a single specified value, ψ, we obtain an estimating function, say $h_{n1}(\phi)$, whose maximisation yields the estimator $\hat{\phi}_{n1} = r_+^{-1}\sum_{i=1}^n r_i\bar{y}_i$. Since $\mathrm{E}_0(\hat{\phi}_{n1}) = \phi_0$, $\hat{\phi}_{n1}$ is consistent for ϕ_0, assuming that $\mathrm{var}_0(\hat{\phi}_{n1}) = r_+^{-2}\sum_{i=1}^n r_i\psi_{i0}^2$ tends to 0 as $n \to \infty$. As is often the case, this conclusion holds without assuming that the y_{ij} are normally distributed; only the forms of the mean and variance of the y_{ij} enter the calculation.

To form the profile log-likelihood for ϕ we need the mle of ψ_i^2 for given ϕ:

$$\hat{\psi}_{i\phi}^2 = r_i^{-1}\sum_{j=1}^{r_i}(y_{ij} - \phi)^2.$$

The resulting estimating function is

$$h_{n2}(\phi) = -\frac{1}{2}\sum_{i=1}^n r_i\left[1 + \log(2\pi/r_i) + \log\left\{\sum_{j=1}^{r_i}(y_{ij} - \phi)^2\right\}\right],$$

and the associated mean function, $m_{n2}(\phi) = \mathrm{E}_0\{h_{n2}(\phi)\}$, satisfies

$$m_{n2}(\phi) - m_{n2}(\phi_0) = \frac{1}{2}\sum_{i=1}^n r_i\mathrm{E}_0\left[\log\left\{\sum_{j=1}^{r_i}(y_{ij} - \phi_0)^2\right\} - \log\left\{\sum_{j=1}^{r_i}(y_{ij} - \phi)^2\right\}\right].$$

Under $(\phi_0, \psi_0^{(n)})$, $\sum_{j=1}^{r_i}(y_{ij} - \phi_0)^2$ and $\sum_{j=1}^{r_i}(y_{ij} - \phi)^2$ have respective distributions $\psi_{i0}^2\chi_{r_i}^2$ and $\psi_{i0}^2\chi_{r_i}^2\{r_i(\phi - \phi_0)^2\}$ (central and non-central Chi-squares). That $m_{n2}(\phi) - m_{n2}(\phi_0) < 0$ for $\phi \neq \phi_0$ follows from the fact that $\chi_r^2(\delta^2)$ is stochastically larger than χ_r^2. Hence, C2(ii) holds with $\phi_1 = \phi_0$, and $\hat{\phi}_{n2}$ is consistent for ϕ_0, though not an explicit function of the y_{ij}, unlike $\hat{\phi}_{n1}$.

Example 2 This is Neyman and Scott's second example: it is the complement of their first in that it is now the mean that is the incidental parameter. So, the observations are independent, y_{ij} having distribution $\mathrm{N}(\psi_i, \phi^2)$. The overall log-likelihood function is

$$l_n(\phi, \psi^{(n)}) = -\frac{1}{2}r_+\log(2\pi\phi^2) - \frac{1}{2}\phi^{-2}\sum_{i=1}^n\sum_{j=1}^{r_i}(y_{ij} - \psi_i)^2.$$

The estimating function with a single specified value, ψ, in place of the ψ_i yields

$$\hat{\phi}_{n1}^2 = r_+^{-1}\sum_{i=1}^n\sum_{j=1}^{r_i}(y_{ij} - \psi)^2,$$

which has mean $\mathrm{E}_0(\hat{\phi}_{n1}^2) = \phi_0^2 + \eta_n(\psi)$, where $\eta_n(\psi) = r_+^{-1}\sum_{i=1}^n r_i(\psi_{i0} - \psi)^2$. Hence, in order to define a bias-corrected estimator, as $\hat{\phi}_{n1}^2 - \eta_n(\psi)$, the value of

$\eta_n(\psi)$ is needed. Typically, the ψ_{i0} are random effects, independently sampled from distribution G, say with mean μ_ψ and variance σ_ψ^2. In this case the basic identities

$$\mathrm{E}(y_{ij}) = \mu_\psi, \quad \mathrm{var}(y_{ij}) = \phi^2 + \sigma_\psi^2, \quad \mathrm{E}(y_{ij} y_{ik}) = \sigma_\psi^2 + \mu_\psi^2 \ (j \neq k)$$

follow from the mean and variance specifications, and they can be used to construct an unbiased estimator of $\eta_n(\psi)$ as

$$r_+^{-1} \sum_{i=1}^{n} r_i (r_i - 1)^{-1} \left\{ \bar{y}_i^2 + (r_i - 1)(\bar{y}_i - \psi)^2 - r_i^{-1} \sum_{j=1}^{r_i} y_{ij}^2 \right\}.$$

The resulting bias-corrected estimator for ϕ_0^2 then reduces to

$$\hat{\phi}_0^2 = r_+^{-1} \sum_{i=1}^{n} r_i (r_i - 1)^{-1} \left\{ \sum_{j=1}^{r_i} y_{ij}^2 - r_i \bar{y}_i^2 \right\};$$

$\hat{\phi}_0^2$ is consistent for ϕ_0^2 if, for example, the y_{ij} have finite fourth moment.

A profile log-likelihood for ϕ can be constructed from $\hat{\psi}_{i\phi} = \bar{y}_i$ as

$$h_{n2}(\phi) = -\frac{1}{2} \left\{ r_+ \log(2\pi\phi^2) + \sum_{i=1}^{n} (r_i - 1) s_i^2 / \phi^2 \right\}.$$

This gives rise to the estimator

$$\hat{\phi}_{n2}^2 = r_+^{-1} \sum_{i=1}^{n} (r_i - 1) s_i^2,$$

which has mean $\mathrm{E}_0(\hat{\phi}_{n2}^2) = (1 - n/r_+)\phi_0^2$, so the bias-corrected version is $(1 - n/r_+)^{-1} \hat{\phi}_{n2}^2$. This simple result fits into the general framework here as follows. The mean function of $h_{n2}(\phi)$ satisfies

$$m_{n2}(\phi) - m_{n2}(\phi_0) = \frac{1}{2} \left\{ r_+ \log \rho - (r_+ - n)(\rho - 1) \right\},$$

in which $\rho = \phi_0^2/\phi^2$. This expression takes its maximum value, $-\frac{1}{2}\{n + r_+ \log (1 - n/r_+)\}$, at $\rho = \rho_{\max} = r_+/(r_+ - n)$. But, $\log(1 - x) < -x$ for $0 < x < 1$, so the maximum value exceeds $-\frac{1}{2}\{n + r_+(-n/r_+)\} = 0$. Thus, condition C2(ii), essentially that $m_{n2}(\phi) < m_{n2}(\phi_0)$ for $\phi \neq \phi_0$, fails here. However, we can identify ϕ_1 from $\phi_0^2/\phi_1^2 = \rho_{\max}$ as $\phi_1^2 = \phi_0^2(1 - n/r_+)$. It follows that $(1 - n/r_+)^{-1} \hat{\phi}_{n2}^2$ is consistent for ϕ_0^2. The correction factor, $(1 - n/r_+)^{-1}$, does not depend on G in this case, so the method is successful in this respect.

A maximum conditional likelihood estimator (mcle) can be constructed by conditioning on the $\bar{y}_i$, which are sufficient for the ψ_i (for known ϕ). The mcle is found to be identical to the bias-corrected version of $\hat{\phi}^2_{n2}$.

Example 3 Suppose that the y_{ij} are independent event-counts with means $\lambda_{ij} = \psi_i e^{x_{ij}\phi}$. For simplicity, the covariates, x_{ij}, are taken to be scalar, with $\dim(\phi) = 1$. The literature contains many examples of such data. In Gesch et al. (2002) y_{ij} is the number of misdemeanors of a young offender in custody during period j, the x_{ij} are binary, with $x_{ij} = 0$ representing the administration of a placebo and $x_{ij} = 1$ that of a nutritional supplement. The focus there was on the parameter ϕ, representing the effect of the supplement, and the ψ_i were treated as random effects to be integrated out over an assumed frequency distribution. However, the ψ_i themselves would clearly also be of potential interest in assessing individuals' levels of behaviour.

In order to construct a suitable estimating function let us begin with a simple Poisson model, for which the log-likelihood function is

$$l_n(\phi, \psi^{(n)}) = \sum_{i=1}^{n}\sum_{j=1}^{r_i}(-\lambda_{ij} + y_{ij}\log\lambda_{ij} - \log y_{ij}!).$$

The derived estimating function, with a single specified ψ replacing the ψ_i, is given by

$$h_{n1}(\phi) = \sum_{i=1}^{n}\sum_{j=1}^{r_i}\left\{-\psi e^{x_{ij}\phi} + y_{ij}\log(\psi e^{x_{ij}\phi}) - \log y_{ij}!\right\},$$

for which the corresponding mean function satisfies

$$m_{n1}(\phi) - m_{n1}(\phi_0) = \sum_{i=1}^{n}\left[-\psi\left\{\tau_{i0}(\phi) - \tau_{i0}(\phi_0)\right\} + \psi_{i0}\tau_{i1}(\phi_0)(\phi - \phi_0)\right],$$

$$m'_{n1}(\phi) = \sum_{i=1}^{n}\left\{-\psi\tau_{i1}(\phi) + \psi_{i0}\tau_{i1}(\phi_0)\right\}, \quad m''_{n1}(\phi) = -\psi\sum_{i=1}^{n}\tau_{i2}(\phi),$$

where $\tau_{ik}(\phi) = \sum_{j=1}^{r_i} x^k_{ij} e^{x_{ij}\phi}$ for $k = 0, 1, 2$. Hence, even though $m''_{n1}(\phi) < 0$ for all ϕ, by taking $\psi > 0$, so that a unique maximum of $m_{n1}(\phi)$ exists, it is a non-explicit function of ϕ_0; moreover, it depends on the ψ_{i0}. So, the crude, single-ψ method does not produce a useful result in this case.

The profile log-likelihood, which we adopt as a second estimating function, is based on $\hat{\psi}_{i\phi} = r_i\bar{y}_i/\tau_{i0}(\phi)$. Thus,

$$h_{n2}(\phi) = \sum_{i=1}^{n}\sum_{j=1}^{r_i}\left\{-\hat{\psi}_{i\phi}e^{x_{ij}\phi} + y_{ij}\log(\hat{\psi}_{i\phi}e^{x_{ij}\phi}) - \log y_{ij}!\right\},$$

and the associated mean function satisfies

$$m_{n2}(\phi) - m_{n2}(\phi_0) = \sum_{i=1}^{n} \psi_{i0} g_i(\phi),$$

where

$$g_i(\phi) = -\tau_{i0}(\phi_0)\{\log \tau_{i0}(\phi) - \log \tau_{i0}(\phi_0)\} + \tau_{i1}(\phi_0)(\phi - \phi_0).$$

Now,

$$g_i'(\phi) = -\tau_{i0}(\phi_0)\left\{\frac{\tau_{i1}(\phi)}{\tau_{i0}(\phi)} - \frac{\tau_{i1}(\phi_0)}{\tau_{i0}(\phi_0)}\right\} \quad \text{and}$$
$$g_i''(\phi) = -\frac{\tau_{i0}(\phi_0)}{\tau_{i0}(\phi)^2}\left\{\tau_{i0}(\phi)\tau_{i2}(\phi) - \tau_{i1}(\phi)^2\right\},$$

so $g_i'(\phi_0) = 0$ and $g_i''(\phi) < 0$ for all ϕ, by the Cauchy-Schwartz Inequality. Hence, ϕ_0 is the global maximum of $g_i(\phi)$, and therefore of $m_{n2}(\phi)$ too. It follows now that $\phi_1 = \phi_0$ in C2, so $\hat{\phi}_{n2}$ is consistent for ϕ_0.

Under a Poisson model, for known ϕ, the $\bar{y}_i$ are sufficient for the ψ_i, so a conditional likelihood can be formulated. However, this just produces the same estimator as the profile likelihood just given.

To apply the integrated likelihood method, a natural (conjugate) choice of weight function would be a gamma density for the ψ_i:

$$p(\psi_i) = \Gamma(\tau)^{-1}\gamma^{\tau}\psi_i^{\tau-1}\mathrm{e}^{-\gamma\psi_i},$$

in which γ and τ are taken as given. The resulting integrated log-likelihood function, based on a negative binomial probability function for the event counts, is

$$il_n(\phi) = \sum_{i=1}^{n}\big[\log\Gamma(\tau + y_{ij}) - \log\left\{y_{ij}!\Gamma(\tau)\right\} + \tau\log\gamma + x_{ij}y_{ij}\phi$$
$$-(\tau + y_{ij})\log(\gamma + \mathrm{e}^{x_{ij}\phi})\big].$$

But now it is difficult to proceed as before with this form armed only with a basic specification of the mean of y_{ij}. However, $il_n'(\phi)$ is linear in y_{ij}, so it is easy to evaluate

$$\mathrm{E}_0\{il_n'(\phi)\} = \sum_{i=1}^{n} x_{ij}(\gamma + \mathrm{e}^{x_{ij}\phi})^{-1}(\gamma\psi_{i0}\mathrm{e}^{x_{ij}\phi_0} - \tau\mathrm{e}^{x_{ij}\phi}).$$

This expression is not zero at $\phi = \phi_0$, so the estimating equation, $il_n'(\phi) = 0$, is biased, suggesting that the associated estimator is inconsistent. The bias could be removed, by subtracting the mean from $il_n(\phi)$, but the result would then be a function of ψ_{i0} as well as of ϕ and ϕ_0. Alternatively, a value, ϕ_1, for which $\mathrm{E}_0\{il_n'(\phi_1)\} = 0$ could be sought numerically as a function of ϕ_0. However, the previous approach, based on the profile likelihood, is much more straightforward.

Example 4 Suppose that the y_{ij} are independent, y_{ij} having some unspecified continuous distribution on (ψ_i, ∞) with mean $\psi_i + \phi$. This could serve as a model for clusters of survival times where the actual zero-time is unrecorded for each cluster. As a basis for constructing an estimating function we will use the exponential distribution. The corresponding (non-regular) log-likelihood function is

$$l_n(\phi, \psi^{(n)}) = \sum_{i=1}^{n} \big\{ - r_i \log \phi - \phi^{-1} r_i (\bar{y}_i - \psi_i) + \log \mathrm{I}(y_{i,\min} \geq \psi_i) \big\},$$

where $y_{i,\min} = \min(y_{i1}, \ldots, y_{ir_i})$ and I(.) is the indicator function.

It is difficult to specify an appropriate single value for ψ here in general. In the special case where the y_{ij} are non-negative, $\psi = 0$ can be tried. This gives the estimating function

$$h_{n1}(\phi) = -r_+ (\log \phi + \phi^{-1} \bar{y}),$$

where $\bar{y} = r_+^{-1} \sum_{i=1}^{n} r_i \bar{y}_i$ is the overall sample mean. The resulting estimator is $\hat{\phi}_{n1} = \bar{y}$, which has mean $\mathrm{E}_0(\hat{\phi}_{n1}) = \phi_0 + \eta_n$, where $\eta_n = r_+^{-1} \sum_{i=1}^{n} r_i \psi_{i0}$. Thus, the bias-corrected version, $\hat{\phi}_{n1} - \eta_n$, depends on the ψ_{i0}. We can try to construct an estimate of η_n, based on the $y_{i,\min}$, as follows. Assume that $\mathrm{E}_0(y_{i,\min}) = \psi_{i0} + \phi_0 \nu(r_i)$ for some function $\nu(.)$: this will be the case for any distribution on (ψ_{i0}, ∞) in which ϕ_0 is a scale parameter. Then,

$$\mathrm{E}_0 \left(r_+^{-1} \sum_{i=1}^{n} r_i y_{i,\min} \right) = \eta_n + \phi_0 b_n,$$

where $b_n = r_+^{-1} \sum_{i=1}^{n} r_i \nu(r_i)$, which leads to a bias-corrected estimator for ϕ_0 based on

$$\mathrm{E}_0 \left\{ r_+^{-1} \sum_{i=1}^{n} r_i (\bar{y}_i - y_{i,\min}) \right\} = (1 - b_n) \phi_0.$$

So, the lack of knowledge about η_n has been transferred, via estimation, to a lack of knowledge about $\nu(.)$. In general, as r_i increases from 1 to ∞, $\nu(r_i)$ decreases from 1 to 0; for example, if the y_{ij} are actually exponentially-distributed, $\nu(r_i) = r_i^{-1}$ and then $b_n = n/r_+$. In pragmatic vein, one could try to gain some idea of the form of $\nu(.)$ by plotting the points $(r_i, \bar{y}_i - y_{i,\min})$: these will lie around the curve $(r, \phi_0\{1 - \nu(r)\})$, which increases monotonically from (1,0) to (∞, ϕ_0).

For a profile log-likelihood we have $\hat{\psi}_{i\phi} = y_{i,\min}$, so an alternative estimating function is

$$h_{n2}(\phi) = -r_+ \log \phi - \phi^{-1} \sum_{i=1}^{n} r_i (\bar{y}_i - y_{i,\min}).$$

Maximisation of $h_{n2}(\phi)$ yields $\hat{\phi}_{n2} = r_+^{-1} \sum_{i=1}^n r_i (\bar{y}_i - y_{i,\min})$, similar to the previous estimator.

A maximum conditional likelihood estimator can be derived by noting that under the exponential model, for known ϕ, $y_{i,\min}$ is sufficient for ψ_i and has density function $r_i \phi^{-1} \mathrm{e}^{-r_i(y-\psi_i)/\phi}$ on (ψ_i, ∞). The resulting conditional log-likelihood function is

$$cl_n(\phi) = -\sum_{i=1}^n \left\{ \log r_i + (r_i - 1) \log \phi + r_i \phi^{-1} (\bar{y}_i - y_{i,\min}) \right\}.$$

But this just produces $\hat{\phi}_{nc} = (r_+ - n)^{-1} \sum_{i=1}^n r_i (\bar{y}_i - y_{i,\min})$, essentially the same estimator again.

Example 5 Suppose that y_{i1} and y_{i2} are independent, non-negative variates with means $\phi\psi_i$ and ϕ/ψ_i, respectively; this is an example from Reid (1995). The log-likelihood based on exponential distributions for y_{i1} and y_{i2} is

$$l_n(\phi, \psi^{(n)}) = -2n \log \phi - \phi^{-1} \sum_{i=1}^n (y_{i1}/\psi_i + y_{i2}\psi_i).$$

The estimating function with a single specified ψ yields $\hat{\phi}_{n1} = \frac{1}{2}(\bar{y}_1/\psi + \bar{y}_2\psi)$, in terms of the sample means, and

$$\mathrm{E}_0(\hat{\phi}_{n1}) = \frac{1}{2}\phi_0 n^{-1} \sum_{i=1}^n (\psi_{i0}/\psi + \psi/\psi_{i0}).$$

The estimator-bias depends on the ψ_{i0}, but we can derive $\psi_{i0}/\psi + \psi/\psi_{i0} \geq 2$, so $\mathrm{E}_0(\hat{\phi}_{n1}) \geq \phi_0$. Treating ψ as a parameter to be estimated produces estimators $\hat{\psi}_n = \bar{y}_1/\bar{y}_2$ and $\hat{\phi}_{n2} = \sqrt{\bar{y}_1\bar{y}_2}$; thus, $\hat{\phi}_{n2} < \hat{\phi}_{n1}$ and

$$\mathrm{E}_0(\hat{\phi}_{n2}^2) = \phi_0^2 \left(n^{-1} \sum_{i=1}^n \psi_{i0} \right) \left(n^{-1} \sum_{i=1}^n \psi_{i0}^{-1} \right) \geq \phi_0^2.$$

A profile log-likelihood follows from $\hat{\psi}_{i\phi}^2 = y_{i1}/y_{i2}$ as

$$h_{n3}(\phi) = -2n \log \phi - 2\phi^{-1} \sum_{i=1}^n \sqrt{y_{i1}y_{i2}},$$

and maximisation of $h_{n3}(\phi)$ produces $\hat{\phi}_{n3} = n^{-1} \sum_{i=1}^n \sqrt{y_{i1}y_{i2}}$. Reid (1995) pointed out that, when the y-distributions actually are exponential, $\hat{\phi}_{n3} \to \mathrm{E}_0(\hat{\phi}_{n3}) = \frac{1}{4}\pi\phi_0$. In our case, taking $\phi\psi_i$ and ϕ/ψ_i as scale parameters for y_{i1} and y_{i2}, $\mathrm{E}_0(\hat{\phi}_{n3}) = \phi_0\eta_1\eta_2$, where $\eta_k = \mathrm{E}_0(\sqrt{y_{ik}/\phi_0})$ $(k = 1, 2)$. Although $\eta_1\eta_2$ will not generally be known,

$$\eta_1\eta_2 < \phi_0^{-1}\sqrt{\mathrm{E}_0(y_{i1})\mathrm{E}_0(y_{i2})} = 1,$$

so $\mathrm{E}_0(\hat{\phi}_{n3}) < \phi_0$.

The alternative estimators together yield, asymptotically,

$$\hat{\phi}_{n3} < \phi_0 \leq \hat{\phi}_{n2} < \hat{\phi}_{n1}.$$

However, none of these general methods produces the rather obvious estimator $\hat{\phi}_n = n^{-1}\sum_{i=1}^n y_{i1}y_{i2}$, which is unbiased for ϕ_0^2. So, according to the proposal here, we take $\hat{\phi}_n^{1/2}$ as our consistent estimator.

Example 6 A typical example of longitudinal data occurs when we have many short time series, e.g. in monitoring hospital patients. Suppose that the ith individual gives rise to an AR(1) process:

$$y_{ij} = \phi y_{i,j-1} + \psi_i e_{ij}.$$

The e_{ij} represent 'white noise': they are assumed to be independent innovations with mean 0 and variance 1. The regression parameter, ϕ, is homogeneous between series, but the variability, or volatility, represented by ψ_i, varies between them. Taking a normal distribution for the noise gives a log-likelihood, conditional on the initial levels, the y_{i0}, as

$$l_n(\phi, \psi^{(n)}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{r_i}\big\{\log(2\pi\psi_i^2) + (y_{ij} - \phi y_{i,j-1})^2/\psi_i^2\big\}.$$

The log-likelihood, with ψ_i replaced by a single ψ, is

$$h_{n1}(\phi) = -\frac{1}{2}\left\{r_+\log(2\pi\psi^2) + \psi^{-2}\sum_{i=1}^{n}\sum_{j=1}^{r_i}(y_{ij} - \phi y_{i,j-1})^2\right\},$$

so $\hat{\phi}_{n1}$ is just the least-squares estimator. The resulting mean function is given by

$$m_{n1}(\phi) = -\frac{1}{2}\left\{r_+\log(2\pi\psi^2) + \psi^{-2}\sum_{i=1}^{n} r_i\psi_{i0}^2 + \psi^{-2}(\phi_0 - \phi)^2\sum_{i=1}^{n}\sum_{j=1}^{r_i}\mathrm{E}_0(y_{i,j-1}^2)\right\},$$

where we have used $\mathrm{E}_0(e_{ij}y_{i,j-1}) = 0$. This quadratic in ϕ has a unique maximum at ϕ_0, and so $\hat{\phi}_{n1}$ is consistent for ϕ_0.

The profile log-likelihood for ϕ provides an alternative estimating function:

$$h_{n2}(\phi) = -\frac{1}{2}\sum_{i=1}^{n} r_i\big\{1 + \log(2\pi\hat{\psi}_{i\phi}^2)\big\},$$

in which $\hat{\psi}_{i\phi}^2 = r_i^{-1}\sum_{j=1}^{r_i}(y_{ij} - \phi y_{i,j-1})^2$. For this, the mean function satisfies

$$m_{n2}(\phi) - m_{n2}(\phi_0) = -\frac{1}{2}\sum_{i=1}^{n} r_i \left\{\mathrm{E}_0(\log \hat{\psi}_{i\phi}^2) - \mathrm{E}_0(\log \hat{\psi}_{i\phi_0}^2)\right\}.$$

As in Example 1, under the assumption of a normal distribution for the e_{ij}, the expression involves central and non-central Chi-squares, and is therefore negative for $\phi \neq \phi_0$. In fact, this assessment will hold provided only that the e_{ij}^2 have a density monotone decreasing from its maximum at zero. Hence, $m_{n2}(\phi)$ has a maximum at ϕ_0 and so the associated estimator, $\hat{\phi}_{n2}$ is consistent for ϕ_0.

Example 7 Cox (1993) gave an example in which, in our notation, $r_i = 2$, and y_{i1} and y_{i2} are independently normal with respective means ψ_i and $\phi\psi_i$, and unit variances; this is also Example 3 of Morton (1981). Cox examined two possible pivotal functions:

$$g_1(y_i, \phi) = (y_{i2} - \phi y_{i1})/(1+\phi^2)^{1/2} \quad \text{and} \quad g_2(y_i, \phi) = y_{i2} - \phi y_{i1};$$

$g_1(y_i, \phi)$ is $\mathrm{N}(0, 1)$ and $g_2(y_i, \phi)$ is $\mathrm{N}(0, 1+\phi^2)$. The corresponding 'log-likelihoods' are

$$h_{n1}(\phi) = -\frac{1}{2}\sum_{i=1}^{n}(y_{i2} - \phi y_{i1})^2/(1+\phi^2) \quad \text{and} \quad h_{n2}(\phi) = h_{n1}(\phi) - \frac{n}{2}\log(1+\phi^2).$$

Now, maximisation of $h_{n1}(\phi)$ with respect to ϕ looks like weighted least-squares, which is well-known to give rise to inconsistency; on the other hand, maximisation of $h_{n2}(\phi)$ looks like Gaussian estimation, which corrects weighted least-squares by adding the log-variance term, thus eliminating the bias in the estimating equation (e.g. Crowder 1986, Example 2.2). However, because of the structure here, in particular because y_{i1} is not a fixed covariate, the opposite is true: it turns out that it is $h_{n1}(\phi)$, rather than $h_{n2}(\phi)$, that gives an unbiased estimating equation.

It is straightforward to show that $h_{n1}(\phi)$ gives a consistent estimator. Its mean function satisfies

$$\begin{aligned} m_{n1}(\phi) &= -\frac{n}{2}\left\{1 + \eta_n(\phi - \phi_0)^2/(1+\phi^2)\right\}, \\ m'_{n1}(\phi) &= -n\eta_n(\phi - \phi_0)(1+\phi\phi_0)/(1+\phi^2)^2, \end{aligned}$$

where $\eta_n = n^{-1}\sum_{i=1}^{n} \psi_{i0}^2$. Thus, $m'_{n1}(\phi_0) = 0$, so $\phi_1 = \phi_0$ in C2; the other root, $\phi = -1/\phi_0$, of $m'_{n1}(\phi) = 0$ can be ignored since $m_{n1}(-1/\phi_0) < m_{n1}(\phi_0)$.

As Cox (1993) pointed out, $h_{n2}(\phi)$ does not yield an unbiased estimating equation. However, our primary interest in this example is to take $h_{n2}(\phi)$ and show how it can lead to a consistent estimator via Proposition 1. The mean function for $h_{n2}(\phi)$ and its derivative are given by

$$\begin{aligned} m_{n2}(\phi) &= -\frac{1}{2}n\left\{1 + \eta_n(\phi - \phi_0)^2(1+\phi^2)^{-1} + \log(1+\phi^2)\right\}, \\ m'_{n2}(\phi) &= -n(1+\phi^2)^{-2}\left\{\eta_n(1+\phi\phi_0)(\phi - \phi_0) + \phi(1+\phi^2)\right\}. \end{aligned}$$

Thus, $m'_{n2}(0) = n\eta_n\phi_0$ and $m'_{n2}(\phi_0) = -n\phi_0(1+\phi_0^2)^{-1}$, which shows that, whether ϕ_0 is positive or negative, $m_{n2}(\phi)$ has a maximum between 0 and ϕ_0. Assume that $\eta_n \to \eta$ as $n \to \infty$, where η is known or consistently estimated, e.g. by $n^{-1}\sum_{i=1}^{n}(y_{i1}^2 - 1)$. Then,

$$n^{-1}m_{n2}(\phi) \to \bar{m}_2(\phi) = -\frac{1}{2}\left\{1 + \eta(\phi-\phi_0)^2(1+\phi^2)^{-1} + \log(1+\phi^2)\right\}.$$

Like $m_{n2}(\phi)$, the function $\bar{m}_2(\phi)$ has a maximum between 0 and ϕ_0. It could also have a second maximum, depending on the number of real roots of the cubic equation $\bar{m}'_2(\phi) = 0$; in either case ϕ_1 is defined as the dominant maximum. Then a consistent estimator for ϕ_0 can be identified as follows: first compute $\hat{\phi}_{n2}$ that maximises $h_{n2}(\phi)$; then solve the equation $f_1(\phi_0) = \hat{\phi}_{n2}$, that is, $\bar{m}'_2(\hat{\phi}_{n2}) = 0$, for ϕ_0. This equation is quadratic,

$$\eta_n(\hat{\phi}_{n2} - \phi_0)(1 + \hat{\phi}_{n2}\phi_0) + \hat{\phi}_{n2}(1 + \hat{\phi}_{n2}^2) = 0,$$

and the ϕ_0-root that maximises $m_{n2}(\hat{\phi}_{n2})$ is selected.

Replacing ψ_i by a single ψ in the log-likelihood for this model gives

$$h_{n3}(\phi) = -n\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left\{(y_{i1}-\psi)^2 + (y_{i2}-\phi\psi)^2\right\}.$$

Treating ψ also as a parameter to be estimated, maximisation of h_{n3} produces $\hat{\phi}_{n3} = \bar{y}_2/\bar{y}_1$ and $\hat{\psi}_n = \bar{y}_1$, in terms of the sample means. Assuming that $\tau = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\psi_{i0}$ exists, $\bar{y}_1 \to_p \tau$ and $\bar{y}_2 \to_p \phi_0\tau$, and so $\hat{\phi}_{n3}$ is consistent for ϕ_0.

As is commonly the case, the asymptotic results depend mainly on the specifications for the first two moments of (y_{i1}, y_{i2}), not on the normal distributional assumption.

4 Inference for $\boldsymbol{\phi_0}$ based on the asymptotic distribution of $\boldsymbol{\hat{\phi}_0}$

Routine methods for obtaining confidence regions for and performing tests on ϕ_0 are based on the asymptotic distribution of $\hat{\phi}_0$, which is addressed in Sect. 4.1. In the subsequent sections efficiency comparisons and practical issues are discussed.

4.1 Asymptotic distribution of $\hat{\phi}_0$

We make standard assumptions; I here denotes the unit matrix.

Condition C3

(i) For some sequence $\{v_n\}$ of positive-definite matrices, $v_n^{-1/2}h'_n(\phi_1) \to_d \mathrm{N}(0, I)$.
(ii) For some sequence $\{w_n\}$ of non-singular matrices, $w_n^{-1}h''_n(\phi_1) \to_p -I$.

Normally, v_n can be taken as $\mathrm{var}_0\{h_n'(\phi_1)\}$ and w_n as $\mathrm{E}_0\{-h_n''(\phi_1)\}$.

The estimating equation $h_n'(\phi) = 0$ can be expanded about ϕ_1, rather than the more usual ϕ_0, to obtain

$$0 = h_n'(\hat{\phi}_n) = h_n'(\phi_1) + h_n''(\phi_*)(\hat{\phi}_n - \phi_1),$$

where ϕ_* lies between ϕ_1 and $\hat{\phi}_n$. Now, $\hat{\phi}_n \rightarrow_p \phi_1$ under C1 and C2, and then it follows from C3 that

$$\hat{\phi}_n - \phi_1 \sim w_n^{-1} h_n'(\phi_1) \sim_d \mathrm{N}\big\{0, w_n^{-1} v_n (w_n^{-1})^{\mathrm{T}}\big\}.$$

For the consistency-corrected estimator, $\hat{\phi}_0 = f_1^{-1}(\hat{\phi}_n)$, we have $\hat{\phi}_0 \sim_d \mathrm{N}(\phi_0, c_n)$, where $c_n = u w_n^{-1} v_n (w_n^{-1})^{\mathrm{T}} u^{\mathrm{T}}$ with $u = f_1'(\phi_0)^{-1}$; in application, ϕ_0 and ϕ_1 are replaced by their estimates in the matrix expressions. So, $c_n^{-1/2}(\hat{\phi}_0 - \phi_0)$ is an asymptotically-pivotal function from which approximate confidence regions and tests for ϕ_0 can be generated in the usual way.

A more direct route is to invert the function $h_n'(\phi)$ (e.g. Boos 1980). Thus, for a given set B,

$$\mathrm{P}_0\{v_n^{-1/2} h_n'(\phi_1) \in B\} = \mathrm{P}\{\phi_1 \in g_n(B)\} = \mathrm{P}_0\big[\phi_0 \in f_1^{-1}\{g_n(B)\}\big],$$

where g_n is the inverse function to $v_n^{-1/2} h_n'(\phi)$; here, $f_1^{-1}\{g_n(B)\}$ is a random subset of the ϕ-space. An approximate confidence region is obtained for ϕ_0 on applying the asymptotic normal distribution of $v_n^{-1/2} h_n'(\phi_1)$. A typical choice for B would be an origin-centred sphere of given probability content under $\mathrm{N}(0, I)$. For example, suppose that $\dim(\phi) = 1$. Then, $v_n^{-1/2} h_n'(\phi_1)$ is asymptotically $\mathrm{N}(0, 1)$ under $(\phi_0, \psi_0^{(n)})$, and so $\mathrm{P}_0\{v_n^{-1/2} h_n'(\phi_1) \leq x\} \sim \Phi(x)$, where Φ is the standard normal distribution function. Hence, assuming that $f_1(.)$ and $h_n'(.)$ are monotone increasing functions,

$$\mathrm{P}_0\big[\phi_0 \leq f_1^{-1}\{g_n(x)\}\big] = \mathrm{P}_0\{\phi_1 \leq g_n(x)\} = \mathrm{P}_0\{v_n^{-1/2} h_n'(\phi_1) \leq x\} \sim \Phi(x).$$

So, for example, an approximate 95% confidence interval for ϕ_0 can be obtained as $[f_1^{-1}\{g_n(-1.96)\}, f_1^{-1}\{g_n(1.96)\}]$.

4.2 Efficiency comparisons

Absolute efficiency is difficult to determine in the absence of a likelihood; likewise, it is not much use to know that the optimal estimating function is the likelihood when we do not have a likelihood. However, the relative efficiency of competing consistent estimators, when more than one is available, can be assessed by comparing their asymptotic variances.

We will use Example 7, which has three competing estimators, for illustration. Recall that, conditionally on ψ_i, y_{1i} and y_{2i} are, respectively, distributed as $\mathrm{N}(\psi_i, 1)$ and $\mathrm{N}(\phi\psi_i, 1)$. In order to obtain simulated results, in addition to ones based on the

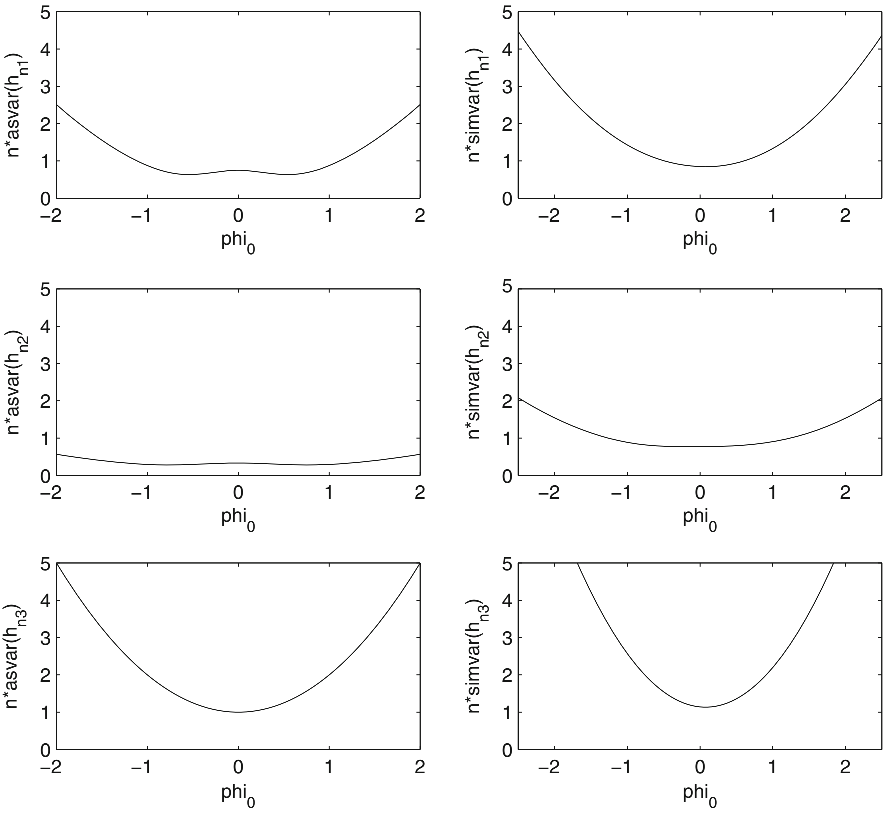

Fig. 1 The left-hand column shows $n * \mathrm{asvar}(h_{nj})$ $(j = 1, 2, 3)$ plotted against ϕ_0: 'asvar' means the asymptotic variance of the estimator associated with h_{nj}, for the three estimating functions of Example 7, calculated by the asymptotic formula given in Sect. 4.1. The right-hand column shows the corresponding sample variances, $n * \mathrm{simvar}(h_{nj})$ from 1,000 simulated samples of size 50

asymptotic formulae, we take the ψ_i to be independently sampled from $N(1, 1)$; a non-zero mean for ψ_i avoids identifiability problems with ϕ.

The algebraic forms of the three sets of w_n and v_n were derived and coded in Matlab, and the left-hand column of Fig. 1 gives plots of them, scaled by n, against ϕ_0. It appears that h_{n2} gives slightly smaller asymptotic variances than h_{n1} and that they both dominate h_{n3}. The asymptotic variances tend to be smaller in the vicinity of $\phi = 0$ though h_{n1} has a slight hump there; the added log-term in h_{n2} appears to give a slight advantage.

The right-hand column in Fig. 1 gives the corresponding plots for simulated data. The simulations comprised 1,000 samples, each of $n = 50$ pairs (y_1, y_2). For each sample the three estimators are computed as follows. The equation $h'_{n1}(\phi) = 0$ is a quadratic, yielding $\hat{\phi}_{n1} = r \pm \sqrt{r^2 + 1}$, where $r = \sum(y_{i2}^2 - y_{i1}^2)/\sum y_{i1}y_{i2}$; there is thus one positive and one negative root. However, it is easy to choose between them, for instance by noting that $\bar{y}_1\bar{y}_2$ will most probably have the sign of ϕ_0.

Computation of $\hat{\phi}_{n2}$ is more involved, as described earlier. The function h_{n3} produces the very simple estimator $\hat{\phi}_{n3} = \bar{y}_2/\bar{y}_1$. The plotted variances are computed as the average mean-square differences between $\hat{\phi}_0$ and ϕ_0, again scaled by n. The values are a little larger than the asymptotic ones but follow similar patterns.

In summary, it seems that h_{n2}, the function that gives a biased estimating function and an inconsistent sequence requiring transformation, is the one that yields the best estimator.

4.3 Some practical issues

Unfortunately, with the type of limited model specification assumed here, analytic evaluation of v_n and w_n will often not be possible. For instance, in Example 2 we have

$$h'_{n2}(\phi) = -r_+\phi^{-1} + \phi^{-3}\sum_{i=1}^{n}(r_i - 1)s_i^2, \quad h''_{n2}(\phi) = r_+\phi^{-2} - 3\phi^{-4}\sum_{i=1}^{n}(r_i - 1)s_i^2.$$

For w_n we need $\mathrm{E}_0(s_i^2)$ and, with the specifications made for $\mathrm{E}_0(y_{ij})$ and $\mathrm{var}_0(y_{ij})$, we have

$$w_n = -r_+\phi_1^{-2} + 3\phi_1^{-4}(r_+ - n)\phi_0^2 = 2\phi_0^{-2}r_+^2/(r_+ - n),$$

using $\phi_1^2 = (1 - n/r_+)\phi_0^2$. For v_n we need $\mathrm{var}_0(s_i^2)$, which cannot be evaluated solely from the specifications made for the mean and variance of y_{ij}. Under the normal distribution,

$$v_n = \phi_1^{-6}\sum_{i=1}^{n}2(r_i - 1)\phi_0^4 = w_n;$$

then, since $u = (1 - n/r_+)^{-1/2}$, $c_n = \frac{1}{2}\phi_0^2/r_+$. But this evaluation goes beyond the basic model specifications.

There seems to be no easy answer to the problem, though some pragmatic suggestions can be made. The estimating functions considered here tend to be of the general form

$$h_n(\phi) = \sum_{i=1}^{n} u_i(y_i; \phi),$$

in which the y_i are independent vectors of lengths r_i. Consider the special case in which the u_i are identically distributed; this will be so when the y_i are identically distributed, entailing equal r_i and the u_i's being the same function for all i. Then, in $h'_n(\phi) = \sum_{i=1}^{n} \partial u_i/\partial\phi$, the summands are identically distributed and so their sample covariance matrix will provide a consistent estimator of $n^{-1}v_n$; the estimator $\hat{\phi}_n$ is inserted for ϕ_1 here. For instance, in Example 4, $\dim(\phi) = 1$ and $h_{n2}(\phi) = \sum u_i$, in which $u_i = -r_i\{\log\phi + \phi^{-1}(\bar{y}_i - y_{i,\min})\}$. Then, assuming that $r_i = r$ for all i, we can use the sample variance of the quantities $r\{\phi_1^{-1} - \phi_1^{-2}(\bar{y}_i - y_{i,\min})\}$ $(i = 1, \ldots, n)$, in which ϕ_1 is replaced by $\hat{\phi}_{n2}$. Likewise, the sample mean of the quantities $-\partial^2 u_i/\partial\phi^2$can be used to estimate $n^{-1}w_n$. When the r_i are not all equal it might be feasible to partition the cases into groups of equal r_i and aggregate the separate sample means and covariances; it might be desirable to weight the groups according to their sizes.

The r_i will tend to be equal when there is some fixed regime or design governing the acquisition of data on individuals. Otherwise, unequal r_i might result from variation in individual circumstances. For instance, it might be that more frequent monitoring of a patient is called for as a result of some unforeseen condition. Previously, it has been assumed that the r_i are given, i.e. the probability statements are all conditional on the r_i appearing in the current data. Consider now the case where the r_i are themselves regarded as random effects. Provided that there are no other individual effects on which the estimating function is conditioned, such as covariates, the u_i-derivatives can be regarded as independently sampled from common populations. Then their sample mean and covariance can be used to estimate v_n and w_n as described above. No doubt, there will be stronger justification for this approach in some situations than in others but, where the case can be made, it appears to solve the problem of assessing the asymptotic distribution of $\hat{\phi}_0$ and of asymptotic inference about ϕ_0.

5 Discussion

The strategy suggested here for constructing a consistent estimator for the structural parameter ϕ departs from the usual practice. Most standard methods attempt to modify a likelihood function so that it will produce a better estimator of the parameter of interest while reducing the effect of nuisance parameters. Here, because of limited model specification, we do not assume that a correct likelihood function is available, and so there is restricted choice for useful estimating functions. We apply any available estimating function without modification, even though it might be unsatisfactory as it stands, but then try to modify the estimator that it produces. This works best when ϕ_1 is simply related to ϕ_0 but the method can still be applied when the relation can only be realised numerically; this was the case in Example 7.

Although the model specification is assumed here not to be sufficient to formulate a correct likelihood function, we can use working likelihoods from which useful estimating functions can be derived. The examples in Sect. 3 are meant to illustrate the procedure in a variety of settings. In particular, some of them reveal how certain aspects of G, the sequence or distribution of the ψ_i, might be needed to establish the properties of the estimators. The charge can be levelled that sometimes the approach works and sometimes it does not. But this is to be expected: if there were a panacea it would probably have been discovered by now. So, the approach suggested here gets all the way in some cases, and gets a long way towards a solution in others. The examples presented in Sect. 3 illustrate this varying success rather than just being a specially-selected set of favourable showcases.

The framework here has been described as repeated measures but the formal results of Sect. 2 apply more generally. Thus, $y_i = (y_{i1}, \ldots, y_{ir_i})$ can have any multivariate joint distribution, the y_{ij} not necessarily being observations of the same attribute on an individual.

6 Proofs and details

6.1 Consistency

The proof of Proposition 1 is based on the classic method of Wald (1949), but to extend it to the present case a preliminary lemma is needed. We use $S(\phi, \rho) = \{\phi' :\mid \phi' - \phi \mid < \rho\}$ to denote an open sphere of radius ρ centred at ϕ.

Lemma 1 *Assume that Conditions C1 and C2 hold, and let* $\mid \phi - \phi_1 \mid \geq \delta$. *Given* $\epsilon > 0$, *one can find* $\rho(\phi, \epsilon) > 0$ *and* $n(\phi, \delta, \epsilon)$ *so that*

$$\mathrm{P}_0 \left\{ \sup_{|\phi' - \phi| < \rho} h_n(\phi') - h_n(\phi_1) \geq 0 \right\} < \epsilon \text{ foreach } n > n(\phi, \delta, \epsilon) \quad \text{and} \quad 0 < \rho < \rho(\phi, \epsilon).$$

Proof Let

$$a_n = \sup_{|\phi' - \phi| < \rho} h_n(\phi') - h_n(\phi), \quad b_n = h_n(\phi) - m_n(\phi), \quad c_n = m_n(\phi_1) - h_n(\phi_1),$$
$$d_n = s_n^{-1} \{m_n(\phi_1) - m_n(\phi)\}.$$

Then the probability in the statement of the lemma is equal to $\mathrm{P}_0(a_n + b_n + c_n \geq s_n d_n)$. Now,

$$d_n = \{s_n^{-1} m_n(\phi_1) - \bar{m}(\phi_1)\} - \{s_n^{-1} m_n(\phi) - \bar{m}(\phi)\} + \{\bar{m}(\phi_1) - \bar{m}(\phi)\}.$$

By C2, the first two terms are each less than $\frac{1}{4}\eta_\delta$ in modulus for $n > n(\delta)$, and the third term is greater than η_δ. Then, for $n > n(\delta)$, $d_n > \frac{1}{2}\eta_\delta > 0$ and so

$$\mathrm{P}_0\left(s_n^{-1} a_n \geq \frac{1}{3} d_n\right) < \mathrm{P}_0\left(s_n^{-1} a_n \geq \frac{1}{6}\eta_\delta\right) < \frac{6}{\eta_\delta} \mathrm{E}_0(s_n^{-1} a_n),$$

which tends to 0 as $\rho \to 0$ (by monotone convergence, resulting from Condition C1(i)). So, $\rho = \rho(\phi, \epsilon)$ can be found to ensure that $\mathrm{P}_0(s_n^{-1} a_n \geq \frac{1}{3} d_n) < \frac{1}{3}\epsilon$. For the other two quantities, b_n and c_n, it follows from C1(ii) that one can find $n(\phi, \epsilon)$ so that, for $n > n(\phi, \epsilon)$,

$$\mathrm{P}_0\left(s_n^{-1} b_n \geq \frac{1}{3} d_n\right) < \frac{1}{3}\epsilon \quad \text{and} \quad \mathrm{P}_0\left(s_n^{-1} c_n \geq \frac{1}{3} d_n\right) < \frac{1}{3}\epsilon.$$

Hence, taking $n(\phi, \delta, \epsilon) = \max\{n(\delta), n(\phi, \epsilon)\}$, for $n > n(\phi, \delta, \epsilon)$,

$$\begin{aligned}\mathrm{P}_0(a_n + b_n + c_n \geq s_n d_n) &< \mathrm{P}_0\left\{\left(s_n^{-1} a_n \geq \frac{1}{3} d_n\right) \cup \left(s_n^{-1} b_n \geq \frac{1}{3} d_n\right)\right.\\ &\qquad \left.\cup \left(s_n^{-1} c_n \geq \frac{1}{3} d_n\right)\right\}\\ &\leq \mathrm{P}_0\left(s_n^{-1} a_n \geq \frac{1}{3} d_n\right) + \mathrm{P}_0\left(s_n^{-1} b_n \geq \frac{1}{3} d_n\right)\\ &\quad + \mathrm{P}_0\left(s_n^{-1} c_n \geq \frac{1}{3} d_n\right) < \epsilon.\end{aligned}$$

□

Proof of Proposition 1 By monotone convergence

$$\mathrm{E}_0\left\{\sup_{|\phi'-\phi|<\rho} s_n^{-1} h_n(\phi')\right\} \to s_n^{-1} m_n(\phi) \text{ as } \rho \to 0.$$

But, for $\phi \neq \phi_1$, one can find n_1 (independent of ϕ) such that $s_n^{-1} m_n(\phi) < \bar{m}(\phi_1)$ for $n > n_1$, by Condition C2. Thus, we can define $\rho_\phi > 0$ such that, for $n > n_1$,

$$\mathrm{E}_0\left\{\sup_{|\phi'-\phi|<\rho_\phi} s_n^{-1} h_n(\phi')\right\} < \bar{m}(\phi_1).$$

The collection of sets $S(\phi, \rho_\phi)$ forms an open covering of S_ϕ, the compact ϕ-space. Thus, there is a finite sub-covering, say $S_1, \ldots, S_k$, of the compact subset $S_\phi - S(\phi_1, \delta)$ and

$$\sup_{|\phi-\phi_1|\geq\delta} h_n(\phi) = \max_{j=1,\ldots,k} \sup_{S_j} h_n(\phi).$$

Now, $|\ \hat{\phi}_n - \phi_1\ | < \delta$ is implied by $\sup_{|\phi-\phi_1|\geq\delta} h_n(\phi) < h_n(\phi_1)$, so

$$\begin{aligned}\mathrm{P}_0(|\ \hat{\phi}_n - \phi_1\ | < \delta) &> \mathrm{P}_0\left\{\sup_{|\phi-\phi_1|\geq\delta} h_n(\phi) < h_n(\phi_1)\right\}\\ &= \mathrm{P}_0\left\{\max_{j=1,\ldots,k} \sup_{S_j} h_n(\phi) < h_n(\phi_1)\right\}\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{P}_0\left[\bigcap_{j=1}^{k}\left\{\sup_{S_j} h_n(\phi) < h_n(\phi_1)\right\}\right] \\
&= 1-\mathrm{P}_0\left\{\bigcup_{j=1}^{k}\sup_{S_j} h_n(\phi) \geq h_n(\phi_1)\right\} \\
&\geq 1-\sum_{j=1}^{k}\mathrm{P}_0\left\{\sup_{S_j} h_n(\phi) - h_n(\phi_1) \geq 0\right\}.
\end{aligned}
$$

It now follows from Lemma 1 that, for $n > n(\delta, \epsilon)$, the last expression exceeds $1-k\epsilon$, which can be made arbitrarily close to 1 by choice of ϵ. □

Proof of Proposition 2 From

$$
\mid \hat{\psi}_{i\hat{\phi}_n} - \psi_{i0} \mid \leq \mid \hat{\psi}_{i\hat{\phi}_n} - \hat{\psi}_{i\phi_0} \mid + \mid \hat{\psi}_{i\phi_0} - \psi_{i0} \mid
$$

follows

$$
\begin{aligned}
\mathrm{P}_0(\mid \hat{\psi}_{i\hat{\phi}_n} - \psi_{i0} \mid > \delta) &\leq \mathrm{P}_0(\mid \hat{\psi}_{i\hat{\phi}_n} - \hat{\psi}_{i\phi_0} \mid + \mid \hat{\psi}_{i\phi_0} - \psi_{i0} \mid > \delta) \\
&\leq \mathrm{P}_0(\mid \hat{\psi}_{i\hat{\phi}_n} - \hat{\psi}_{i\phi_0} \mid > \delta/2) + \mathrm{P}_0(\mid \hat{\psi}_{i\phi_0} - \psi_{i0} \mid > \delta/2).
\end{aligned}
$$

Under the conditions assumed both probabilities tend to zero as $n \to \infty$. □

6.2 Relation between overall and individual estimators of ϕ

Outline proof of Proposition 3. The assumptions here, informally stated, are that h_n and the h_{ni} are twice continuously differentiable and that the orders of magnitude of the various random quantities are similar to those in regular likelihood theory, hence the $\mathrm{O}_p(1)$ terms. The overall estimator, $(\hat{\phi}_n, \hat{\psi}^{(n)})$ from the whole sample, satisfies the equations:

$$
\begin{aligned}
0 = \frac{\partial h_n}{\partial \phi} \mid_{(\hat{\phi}_n, \hat{\psi}^{(n)})} &= \frac{\partial h_n}{\partial \phi} + \left\{\frac{\partial^2 h_n}{\partial \phi^2}(\hat{\phi}_n - \phi_0) + \mathrm{O}_p(1)\right\} \\
&+ \sum_{i=1}^{n}\left\{\frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i}(\hat{\psi}_i - \psi_{i0}) + \mathrm{O}_p(1)\right\}, \qquad (1)
\end{aligned}
$$

where $\frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i}$ is the matrix with (j, k)th element $\frac{\partial^2 h_{ni}}{\partial \phi_j \partial \psi_{ik}}$ and the partial derivatives on the right-hand sides of these equations are all evaluated at (ϕ_0, ψ_0). Again, for $i = 1, \ldots, n$,

$$
\begin{aligned}
0 = \frac{\partial h_n}{\partial \psi_i} \mid_{(\hat{\phi}_n, \hat{\psi}^{(n)})} = \frac{\partial h_{ni}}{\partial \psi_i} \mid_{(\hat{\phi}_n, \hat{\psi}_i)} = \frac{\partial h_{ni}}{\partial \psi_i} + \frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi}(\hat{\phi}_n - \phi_0) + \frac{\partial^2 h_{ni}}{\partial \psi_i^2}(\hat{\psi}_i - \psi_{i0}) \\
+\mathrm{O}_p(1). \qquad (2)
\end{aligned}
$$

On the other hand, the estimator $(\tilde{\phi}_i, \tilde{\psi}_i)$ from y_i alone satisfies

$$0 = \frac{\partial h_{ni}}{\partial \phi} \mid_{(\tilde{\phi}_i, \tilde{\psi}_i)} = \frac{\partial h_{ni}}{\partial \phi} + \frac{\partial^2 h_{ni}}{\partial \phi^2}(\tilde{\phi}_i - \phi_0) + \frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i}(\tilde{\psi}_i - \psi_{i0}) + \mathrm{O}_p(1) \quad (3)$$

and

$$0 = \frac{\partial h_{ni}}{\partial \psi_i} \mid_{(\tilde{\phi}_i, \tilde{\psi}_i)} = \frac{\partial h_{ni}}{\partial \psi_i} + \frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi}(\tilde{\phi}_i - \phi_0) + \frac{\partial^2 h_{ni}}{\partial \psi_i^2}(\tilde{\psi}_i - \psi_{i0}) + \mathrm{O}_p(1). \quad (4)$$

Sum (3) over $i = 1, \ldots, n$ and subtract the result from (1a):

$$0 = \sum_{i=1}^{n} \left\{ \frac{\partial^2 h_{ni}}{\partial \phi^2}(\hat{\phi}_n - \tilde{\phi}_i) + \frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i}(\hat{\psi}_i - \tilde{\psi}_i) + \mathrm{O}_p(1) \right\}. \quad (5)$$

From (2) and (4)

$$0 = \frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi}(\hat{\phi}_n - \tilde{\phi}_i) + \frac{\partial^2 h_{ni}}{\partial \psi_i^2}(\hat{\psi}_i - \tilde{\psi}_i) + \mathrm{O}_p(1),$$

from which,

$$\hat{\psi}_i - \tilde{\psi}_i = -\left(\frac{\partial^2 h_{ni}}{\partial \psi_i^2} \right)^{-1} \left\{ \frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi}(\hat{\phi}_n - \tilde{\phi}_i) + \mathrm{O}_p(1) \right\}.$$

On substitution into (5) this yields

$$0 = \sum_{i=1}^{n} \left\{ w_i(\hat{\phi}_n - \tilde{\phi}_i) + \mathrm{O}_p(1) \right\}, \quad \text{where } w_i = \frac{\partial^2 h_{ni}}{\partial \phi^2} - \frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i} \left(\frac{\partial^2 h_{ni}}{\partial \psi_i^2} \right)^{-1} \frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi},$$

and the required expression follows. We assume that $w_i(\hat{\phi} - \tilde{\phi}) = \mathrm{O}(r_i \times r_i^{-1/2})$ and so dominates $\mathrm{O}_p(1)$ for large r_i. That the weights w_i provide the optimal linear combination follows from

$$\mathrm{var}(\tilde{\phi}_i, \tilde{\psi}_i) \sim -\begin{pmatrix} \frac{\partial^2 h_{ni}}{\partial \phi^2} & \frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i} \\ \frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi} & \frac{\partial^2 h_{ni}}{\partial \psi_i^2} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} w_i^{-1} & -w_i^{-1} \left(\frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i} \right) v_i^{-1} \\ -v_i^{-1} \left(\frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi} \right) w_i^{-1} & v_i^{-1} + v_i^{-1} \left(\frac{\partial^2 h_{ni}}{\partial \psi_i \partial \phi} \right) w_i^{-1} \left(\frac{\partial^2 h_{ni}}{\partial \phi \partial \psi_i} \right) v_i^{-1} \end{pmatrix},$$

where $v_i = \frac{\partial^2 h_{ni}}{\partial \psi_i^2}$. So, $\mathrm{var}(\tilde{\phi}_i) \sim w_i^{-1}$. □

References

Andersen, E. B. (1967). On partial sufficiency and partial ancillarity. *Skandinavian Aktuarial Journal, 50*, 137–152.

Andersen, E. B. (1973). *Conditional inference and models for measurement*. Copenhagen: Mentalhygienisk Forlag.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika, 70*, 343–365.

Barndorff-Nielsen, O. E., Cox, D. R. (1994). *Inference and asymptotics*. London: Chapman and Hall.

Berger, J. O., Liseo, B., Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science, 14*, 1–28.

Boos, D. D. (1980). A new method for constructing approximate confidence intervals from M estimates. *Journal of the American Statistical Association, 75*, 142–145.

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika, 68*, 73–84.

Cox, D. R. (1993). Unbiased estimating equations derived from statistics that are functions of a parameter. *Biometrika, 80*, 905–909.

Cox, D. R., Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society B, 49*, 1–39.

Crowder, M. J. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory, 3*, 305–330.

Crowder, M. J., Hand, D. J. (1990). *Analysis of repeated measures*. London: Chapman and Hall.

Doob, J. L. (1953). *Stochastic processes*. New York: Wiley.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*, 27–38.

Gesch, C. B., Hammond, S. M., Hampson, S. E., Eves, A., Crowder, M. J. (2002). Influence of supplementary vitamins, minerals and essential fatty acids on the antisocial behaviour of young adult prisoners. *British Journal of Psychiatry, 181*, 22–28.

Hand, D. J., Crowder, M. J. (1996). *Practical longitudinal data analysis*. London: Chapman and Hall.

Kalbfleisch, J. D., Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *Journal of the Royal Statistical Society B, 32*, 175–208.

Kiefer, J., Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27*, 887–906.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics, 95*, 391–413.

Liang, K.-Y., Zeger, S. L. (1995). Inference based on estimating functions in the presence of nuisance parameters (with discussion). *Statistical Science, 10*, 158–199.

Loeve, M. (1963). *Probability theory*. New York: Van Nostrand Reinhold.

Morton, R. (1981). Efficiency of estimating equations and the use of pivots. *Biometrika, 68*, 227–233.

Neyman, J., Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1–32.

Portnoy, S. (1988). Asymptotic behaviour of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics, 16*, 356–366.

Reid, N. (1995). The roles of conditioning in inference (with discussion). *Statistical Science, 10*, 138–199.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics, 20*, 595–601.