December 19, 2008

A Note on Misspecified Estimating Functions

Grace Y. Yi
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1
yyi@uwaterloo.ca

and

Nancy Reid
Department of Statistics
University of Toronto
Toronto, Ontario
Canada M5S 3G3

Summary

We consider the use of estimating functions which are not unbiased. Typically, to result in consistent estimators, unbiasedness of estimating functions is a pre-requisite. However, it may sometimes be easier to find a useful estimating function that is biased, especially in the presence of missing data or misclassified observations. We show that the root of the estimating function can be modified to give a consistent and asymptotically normal estimator, and illustrate this on several examples with binary data. We compare this to the alternative approach of adjusting the estimating function, and show that it can be more efficient.

Key Words: Asymptotic normality; Consistency; Efficiency; Estimating function; Unbiasedness.

1 Introduction

We consider estimation of a vector parameter θ, based on a sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of independent and identically distributed random vectors on Ω, drawn from the family of densities $\{f(\mathbf{y};\theta) : \theta \in \Theta\}$, where Θ is a subset of a Euclidean space of dimension p. As an alternative to maximum likelihood estimation, we assume we have a $p \times 1$ vector of estimating functions $\mathbf{g}(\mathbf{y};\theta)$, and define an estimator $\widetilde{\theta}_n$ as the root of the set of p equations

$$\mathbf{G}_n(\widetilde{\theta}_n) = n^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{y}_i;\widetilde{\theta}_n) = \mathbf{0}.$$

Under regularity conditions on the model, and the condition that the estimating function is unbiased, $E_\theta\{\mathbf{g}(\mathbf{Y}_i;\theta)\} = \mathbf{0}$, the resulting estimator is consistent and asymptotically normal, with asymptotic variance given by the Godambe information $J(\mathbf{g}) = \{E_\theta(\partial\mathbf{g}/\partial\theta^T)\}^{-1}E_\theta(\mathbf{g}\mathbf{g}^T)\{E_\theta(\partial\mathbf{g}^T/\partial\theta)\}^{-1}$ (Godambe, 1960). Yanagimoto and Yamamoto (1991) give a number of examples illustrating the role of unbiasedness of estimating equations, and relating it to conditional likelihood inference in the context of exponential families.

In some practical contexts, however, there may be a natural choice of working estimating function that is not unbiased. The most direct approach to correcting a biased estimating function $\mathbf{h}(\mathbf{y};\theta)$ is to compute $E_\theta\{\mathbf{h}(\mathbf{Y};\theta)\}$ and construct a modified estimating function

$$\widetilde{\mathbf{H}}_n(\theta) = n^{-1}\sum_{i=1}^{n}\mathbf{h}(\mathbf{y}_i;\theta) - E_\theta\{\mathbf{h}(\mathbf{Y}_i;\theta)\}; \tag{1}$$

if $E_\theta\{\mathbf{h}(\mathbf{Y}_i;\theta)\}$ cannot be computed exactly then a suitable approximation might be available. For example, McCullagh and Tibshirani (1990) use a bootstrap estimate of the mean to correct the bias of score functions derived from the profile log-likelihood; Yanagimoto and Yamamoto (1991) illustrate correcting estimating functions derived from the method of moments.

In this paper we consider a different, but related, approach to deriving a consistent estimate of θ from a set of biased estimating functions. We use the notation $\mathbf{h}(\mathbf{y};\theta)$ for the vector of biased estimating functions; i.e. we assume $E_\theta\{\mathbf{h}(\mathbf{Y};\theta)\} \neq \mathbf{0}$. Assume that the equation

$$\mathbf{H}_n(\theta) = n^{-1}\sum_{i=1}^{n}\mathbf{h}(\mathbf{y}_i;\theta) = \mathbf{0}$$

has a root $\widehat{\theta}_n^* \in \Theta$ for any given random sample $\mathbf{y}_1, ..., \mathbf{y}_n$, and that $\theta^* \in \Theta$ exists, where θ^* is defined by

$$E_\theta\{\mathbf{h}(\mathbf{Y};\theta^*)\} = \mathbf{0}. \tag{2}$$

Equation (2) defines θ as a function of θ^*, say

$$\theta = \mathbf{k}(\theta^*) \tag{3}$$

for some p-vector of functions $\mathbf{k}(\cdot)$, and we use this to define a new estimator of θ as

$$\widehat{\theta}_n = \mathbf{k}(\widehat{\theta}^*_n). \tag{4}$$

As an illustration we consider a binary data problem with a simple missing data structure.

Example 1: binary pairs with missing data

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^T$ be a random sample of bivariate binary vectors, $i = 1, ..., n$. Assume that $E(Y_{ij}) = \mu, j = 1, 2$, and $\text{corr}(Y_{i1}, Y_{i2}) = \rho$ for $i = 1, ..., n$. Let $\theta = (\mu, \rho)^T$ denote the parameter of interest. Let $R_{ij} = 1$ if Y_{ij} is observed, and 0 otherwise, and define $\lambda_{ij} = P(R_{ij} = 1|Y_{i1}, Y_{i2})$, and $\lambda_{i12} = P(R_{i1} = 1, R_{i2} = 1|Y_{i1}, Y_{i2})$. Assume

$$\text{logit } \lambda_{ij} = \alpha_0 + \alpha_1 Y_{ij},$$

and

$$\text{logit } \lambda_{i12} = \gamma_0 + \gamma_1(Y_{i1} + Y_{i2}).$$

Let

$$u_\mu(\mathbf{Y}_i; \theta) = Y_{i1} + Y_{i2} - 2\mu$$

and

$$u_\rho(\mathbf{Y}_i; \theta) = Y_{i1}Y_{i2} - \rho\mu(1 - \mu) - \mu^2$$

be constructed based on the method of moments, and

$$\mathbf{u}(\mathbf{Y}_i; \theta) = \{u_\mu(\mathbf{Y}_i; \theta), u_\rho(\mathbf{Y}_i; \theta)\}^T.$$

If there is no missing data, $\sum_{i=1}^n \mathbf{u}(\mathbf{Y}_i; \theta)$ is unbiased for θ, yielding a consistent estimator for θ:

$$\hat{\mu}_n = \frac{\sum_{i=1}^n (Y_{i1} + Y_{i2})}{2n}, \qquad \hat{\rho}_n = \frac{\sum_{i=1}^n Y_{i1}Y_{i2} - n\hat{\mu}_n^2}{n\hat{\mu}_n(1 - \hat{\mu}_n)}. \tag{5}$$

Now if we naively apply these estimating functions to the observed data, we have

$$h_\mu(\mathbf{Y}_i; \theta) = R_{i1}Y_{i1} + R_{i2}Y_{i2} - (R_{i1} + R_{i2})\mu,$$

$$h_\rho(\mathbf{Y}_i; \theta) = R_{i1}R_{i2}\{Y_{i1}Y_{i2} - \rho\mu(1 - \mu) - \mu^2\},$$

and

$$\mathbf{h}(\mathbf{Y}_i; \theta) = \{h_\mu(\mathbf{Y}_i; \theta)h_\rho(\mathbf{Y}_i; \theta)\}^T.$$

Setting $\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i;\theta) = \mathbf{0}$ leads to

$$\hat{\mu}_n^* = \frac{\sum_{i=1}^n (R_{i1}Y_{i1} + R_{i2}Y_{i2})}{\sum_{i=1}^n (R_{i1} + R_{i2})}, \tag{6}$$

$$\hat{\rho}_n^* = \frac{\sum_{i=1}^n R_{i1}R_{i2}Y_{i1}Y_{i2} - \hat{\mu}_n^* \sum_{i=1}^n R_{i1}R_{i2}}{\hat{\mu}_n^*(1-\hat{\mu}_n^*)\sum_{i=1}^n R_{i1}R_{i2}}. \tag{7}$$

To find θ^* we use (2) to compute

$$E_\theta \begin{pmatrix} h_\mu(\mathbf{Y}_i;\theta^*) \\ h_\rho(\mathbf{Y}_i;\theta^*) \end{pmatrix} = \begin{pmatrix} E_\theta(R_{i1}Y_{i1} + R_{i2}Y_{i2}) - \mu^* E_\theta(R_{i1} + R_{i2}) \\ E_\theta(R_{i1}R_{i2}Y_{i1}Y_{i2}) - \{\rho^*\mu^*(1-\mu^*) + \mu^{*2}\}E_\theta(R_{i1}R_{i2}) \end{pmatrix} = \mathbf{0}. \tag{8}$$

Note that

$$E_\theta(R_{ij}Y_{ij}) = E_Y E_{R|Y}(R_{ij}Y_{ij}) = E_Y(Y_{ij}\lambda_{ij}) = \frac{\exp(\alpha_0+\alpha_1)}{1+\exp(\alpha_0+\alpha_1)}\mu.$$

Similarly,

$$\begin{aligned}
E_\theta(R_{ij}) &= \frac{\exp(\alpha_0+\alpha_1)}{1+\exp(\alpha_0+\alpha_1)}\mu + \frac{\exp(\alpha_0)}{1+\exp(\alpha_0)}(1-\mu), \\
E_\theta(R_{i1}R_{i2}Y_{i1}Y_{i2}) &= \frac{e^{\gamma_0+2\gamma_1}}{1+e^{\gamma_0+2\gamma_1}}\{\rho\mu(1-\mu)+\mu^2\}, \quad \text{and} \\
E_\theta(R_{i1}R_{i2}) &= \frac{\exp(\gamma_0+2\gamma_1)}{1+\exp(\gamma_0+2\gamma_1)}\{\rho\mu(1-\mu)+\mu^2\} + \frac{2\exp(\gamma_0+\gamma_1)}{1+\exp(\gamma_0+\gamma_1)}\{\mu - \rho\mu(1-\mu) - \mu^2\} \\
&\quad + \frac{\exp(\gamma_0)}{1+\exp(\gamma_0)}(1 - 2\mu + \rho\mu - \rho\mu^2 + \mu^2).
\end{aligned}$$

Therefore, the first equation of (8) gives

$$\mu = \frac{\mu^* \exp(\alpha_0)/\{1+\exp(\alpha_0)\}}{(1-\mu^*)\cdot\exp(\alpha_0+\alpha_1)/\{1+\exp(\alpha_0+\alpha_1)\} + \mu^*\exp(\alpha_0)/\{1+\exp(\alpha_0)\}}$$

with the same relationship between $\hat{\mu}_n$ and $\hat{\mu}_n^*$. It is easily shown that μ is equal to, less than, or greater than μ^* as $\alpha_1 = 0$, $\alpha_1 > 0$ and $\alpha_1 < 0$. In this model if $\alpha_1 = 0$ the data is missing completely at random, and the estimator based on the observed data is consistent, as has been noted in the literature; see, for example, Fitzmaurice, Molenberghs and Lipsitz (1995). However, if the missing data is not missing completely at random, then the moment estimator based only on the observed data either inflates or attenuates the true parameter, depending on how the response affects missingness. This result provides an interesting and transparent characterization of the asymptotic bias induced by ignoring missing values. Applying the second equation of (8), we obtain the relationship between ρ and ρ^*. If $\gamma_1 = 0, \alpha_1 = 0$, then $\rho = \rho^*$, showing that using the available data can still produce a consistent estimator of the correlation under missing completely at random mechanisms.

In this example we get the same estimators for μ and ρ by using (8) to compute $E_\theta\{\mathbf{h}(\mathbf{Y};\theta)\}$ and constructing $\widetilde{\mathbf{H}}_n(\theta)$, as at (1). In Appendix A we show that this will be the case whenever the estimating equation $\mathbf{h}(\mathbf{y};\theta)$ has a structure that is linear in functions of $\mathbf{y}$ and θ, and give a simple example where this does not hold.

In Section 2 we give results on asymptotic consistency and normality for the estimator $\widehat{\theta}_n^*$, and hence for $\widehat{\theta}_n$. The results generalize the discussion in White (1982), which studies model misspecification under the likelihood formulation. It is closely related to the results of Jiang, Turnbull and Clark (1999), who used methods very similar to those in this paper in the context of semiparametric Poisson models. Their biased estimating equations are the score equations from a likelihood function obtained from a working model that is subject to misspecification, and their bridge function, $s_0(\cdot)$ is the inverse of $\mathbf{k}(\theta^*)$.

In Section 3 we illustrate the approach with a series of examples of biased estimating equations for binary data models, where the bias is caused by missing data or misclassified data. In Section 4 we outline a brief comparison for the estimators obtained from (1) and (4), and Section 5 provides a brief discussion.

2 Asymptotic Results

Theorem 1: Suppose $\mathbf{h}\{\mathbf{y};\theta) = (h_1(\mathbf{y};\theta), ..., h_p(\mathbf{y};\theta)\}^T$ is a vector of functions defined on $\Omega \times \Theta$ such that $h_j(\mathbf{y};\theta)$ is a continuous function of θ for each $\mathbf{y}$ and a measurable function of $\mathbf{y}$ for each θ, $j = 1, ..., p$. Assume that Θ is a convex compact set and the true distribution of $\mathbf{Y}$ is $F(\mathbf{y};\theta)$, with density $f(\mathbf{y};\theta)$. Assume $|h_j(\mathbf{y};\theta)| \leq m_j(\mathbf{y})$ for all $\mathbf{y}$ and θ where $m_j(\cdot)$ is integrable with respect to F, $j = 1, ..., p$. Let $\mathbf{H}(\theta) = E_\theta\{\mathbf{h}(\mathbf{Y};\theta)\}$. If $\mathbf{H}(\theta) = \mathbf{0}$ has a unique solution θ^* and $\mathbf{H}_n(\theta) = 0$ has a solution $\widehat{\theta}_n^*$, then

$$\widehat{\theta}_n^* \rightarrow_p \theta^* \quad \text{as} \quad n \rightarrow \infty$$

for almost every sequence $\mathbf{Y}_1, \mathbf{Y}_2, ...$ which is a random sample from F.
Proof: Given j, by Theorem 2 of Jennrich (1969), we have, for almost every sequence $\{\mathbf{Y}_n\}$,

$$n^{-1}\sum_{i=1}^{n} h_j(\mathbf{Y}_i;\theta) \rightarrow \int h_j(\mathbf{y};\theta)dF(\mathbf{y})$$

uniformly for all $\theta \in \Theta$, thus,

$$\sup_{\theta\in\Theta} d\{\mathbf{H}_n(\theta), \mathbf{H}(\theta)\} \rightarrow_p 0 \tag{9}$$

where $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$ is the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. The set $\{\theta : d(\theta, \theta^*) \geq \epsilon\} = \Theta - \{\theta : d(\theta, \theta^*) < \epsilon\}$ is a compact subset of Θ for any $\epsilon > 0$. As $h_j(\mathbf{y};\theta)$ is continuous

in θ for each $\mathbf{y}$, $j = 1, ..., p$, we conclude that $||\mathbf{H}(\theta)||$ is a continuous function of θ. Therefore, there exists $\theta_1 \in \{\theta : d(\theta, \theta^*) \geq \epsilon\}$ such that

$$\inf_{\theta: d(\theta,\theta^*)\geq\epsilon} ||\mathbf{H}(\theta)|| = ||\mathbf{H}(\theta_1)||.$$

As θ^* is the unique solution of $\mathbf{H}(\theta) = \mathbf{0}$, and $\theta_1 \neq \theta^*$, we have $||\mathbf{H}(\theta_1)|| > 0$, i.e., $\inf_{\theta: d(\theta,\theta^*)\geq\epsilon} ||\mathbf{H}(\theta)|| > 0$. Furthermore, $\mathbf{H}_n(\widehat{\theta}_n^*) = \mathbf{0}$ gives

$$\inf_{\theta: d(\theta,\theta^*)\geq\epsilon} ||\mathbf{H}(\theta)|| > 0 = ||\mathbf{H}_n(\widehat{\theta}_n^*)||. \tag{10}$$

By (9) and (10), we conclude, applying Theorem 5.9 of van der Vaart (1998, p.46),

$$\widehat{\theta}_n^* \to_p \theta^* \quad \text{as } n \to \infty.$$

This theorem characterizes the convergence of the estimator $\widehat{\theta}_n^*$ obtained from estimating functions that are not necessarily unbiased. The difference $\theta^* - \theta$ is the asymptotic bias of using estimating functions that are not unbiased to perform estimation of θ. In particular, if $\mathbf{h}(\mathbf{Y}; \theta)$ is unbiased, then $\theta^* = \theta$ and $\widehat{\theta}_n^*$ is consistent for θ. If $\mathbf{k}(.)$ is continuous, then $\mathbf{k}(\widehat{\theta}_n^*)$ converges to $\mathbf{k}(\theta^*)$ in probability and the adjusted estimator $\widehat{\theta}_n$ is consistent for θ.

Next we establish the asymptotic normality of the estimator $\widehat{\theta}_n^*$ and hence of $\widehat{\theta}_n$. Let

$$\begin{aligned}
\mathbf{A}_n(\theta) &= n^{-1} \sum_{i=1}^{n} (\partial/\partial\theta^T)\mathbf{h}(\mathbf{Y}_i; \theta), \qquad \mathbf{A}(\theta) = E_\theta\{\mathbf{A}_n(\theta)\}, \\
\mathbf{B}_n(\theta) &= n^{-1} \sum_{i=1}^{n} \mathbf{h}(\mathbf{Y}_i; \theta)\{\mathbf{h}(\mathbf{Y}_i; \theta)\}^T, \quad \mathbf{B}(\theta) = E_\theta\{\mathbf{B}_n(\theta)\}, \\
\mathbf{C}_n(\theta) &= \{\mathbf{A}_n^{-1}(\theta)\}\mathbf{B}_n(\theta)\{\mathbf{A}_n^{-1}(\theta)\}^T, \quad \text{and} \\
\mathbf{C}(\theta) &= \{\mathbf{A}^{-1}(\theta)\}\mathbf{B}(\theta)\{\mathbf{A}^{-1}(\theta)\}^T.
\end{aligned}$$

Theorem 2: Suppose the conditions in Theorem 1 are satisfied, and $h_j(\mathbf{y}; \theta)$ is a continuously differentiable function of θ for each $\mathbf{y}$, $j = 1, ..., p$. Assume that $\mathbf{A}(\theta^*)$ is nonsingular, then under some regularity conditions on h_j and the model F, we have: as $n \to \infty$, (i) $\sqrt{n}(\widehat{\theta}_n^* - \theta^*) \to_d N\{\mathbf{0}, \mathbf{C}(\theta^*)\}$; (ii) $\mathbf{C}_n(\widehat{\theta}_n^*) \to_p \mathbf{C}(\theta^*)$, and assuming $\mathbf{k}(\cdot)$ defined at (2) is differentiable,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \to_d N\left\{\mathbf{0}, \left(\frac{\partial \mathbf{k}^T(\theta^*)}{\partial\theta}\right)\mathbf{C}(\theta^*)\left(\frac{\partial \mathbf{k}(\theta^*)}{\partial\theta^T}\right)\right\}. \tag{11}$$

Proof: For each $j = 1, ..., p$, applying Lemma 3 of Jennrich (1969) to $\sum_{i=1}^{n} h_j(\mathbf{y}_i; \widehat{\theta}_n^*)$, we obtain

$$\sum_{i=1}^{n} h_j(\mathbf{y}_i; \widehat{\theta}_n^*) = \sum_{i=1}^{n} h_j(\mathbf{y}_i; \theta^*) + \frac{\partial}{\partial\theta^T}\left\{\sum_{i=1}^{n} h_j(\mathbf{y}_i; \bar{\theta}_{jn})\right\}(\widehat{\theta}_n^* - \theta^*)$$

where $\bar{\theta}_{jn}$ lies on the "segment" joining $\widehat{\theta}_n^*$ and θ^*. Stacking these p expansions together, we obtain an expression in a matrix form

$$\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, ..., \bar{\theta}_{pn})\sqrt{n}(\widehat{\theta}_n^* - \theta^*) = -n^{-1/2}\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \theta^*) + n^{-1/2}\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \widehat{\theta}_n^*),$$

where $\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, ..., \bar{\theta}_{pn}) = n^{-1}\{\sum_{i=1}^n \partial h_1(\mathbf{Y}_i; \bar{\theta}_{1n})/\partial\theta^T, \ldots \sum_{i=1}^n \partial h_p(\mathbf{Y}_i; \bar{\theta}_{pn})/\partial\theta^T\}^T$.

By $H_n(\widehat{\theta}_n^*) = \mathbf{0}$, we obtain

$$\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, ..., \bar{\theta}_{pn})\sqrt{n}(\widehat{\theta}_n^* - \theta^*) = -n^{-1/2}\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \theta^*). \tag{12}$$

As

$$E_\theta\{\mathbf{h}(\mathbf{Y}_i; \theta^*)\} = \mathbf{H}(\theta^*) = \mathbf{0},$$

and

$$\mathrm{cov}_\theta\{\mathbf{h}(\mathbf{Y}_i; \theta^*)\} = E_\theta[\mathbf{h}(\mathbf{Y}_i; \theta^*)\{\mathbf{h}(\mathbf{Y}_i; \theta^*)\}^T] = \mathbf{B}(\theta^*),$$

by the Central Limit Theorem, we conclude

$$n^{-1/2}\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \theta^*) \to_d N\{\mathbf{0}, \mathbf{B}(\theta^*)\}. \tag{13}$$

Note that for each $j = 1, ..., p$, $\bar{\theta}_{jn} \to_p \theta^*$ as $n \to \infty$, therefore,

$$n^{-1}\sum_{i=1}^n \partial h_j(\mathbf{Y}_i; \bar{\theta}_{jn})/\partial\theta^T \to_p E\{\partial h_j(\mathbf{Y}_i; \theta^*)/\partial\theta^T\}, \tag{14}$$

and hence

$$\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, ..., \bar{\theta}_{pn}) \to_p \mathbf{A}(\theta^*) \quad \text{as } n \to \infty. \tag{15}$$

Assuming that $\mathbf{A}(\theta^*)$ is nonsingular, we have that $\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, ..., \bar{\theta}_{pn})$ is nonsingular for sufficiently large n (in probability). Therefore, (12) leads to

$$\sqrt{n}(\widehat{\theta}_n^* - \theta^*) = -\{\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, ..., \bar{\theta}_{pn})\}^{-1} n^{-1/2}\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \theta^*).$$

By (13) and (15),

$$\sqrt{n}(\widehat{\theta}_n^* - \theta^*) \to_d N[\mathbf{0}, \{\mathbf{A}^{-1}(\theta^*)\}\mathbf{B}(\theta^*)\{\mathbf{A}^{-1}(\theta^*)\}^T],$$

which is conclusion (i). Conclusion (ii) is straightforward as $\{\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, ..., \bar{\theta}_{pn})\}^{-1} \to_p \mathbf{A}_n^{-1}(\theta^*)$ and $\mathbf{B}_n(\widehat{\theta}_n^*) \to_p \mathbf{B}(\theta^*)$. The asymptotic normality of $\widehat{\theta}_n$ follows from an application of the delta method.

The regularity conditions for Theorems 1 and 2 are similar to those outlined in Ch. 5 of Van der Waart; see in particular the discussion following his Theorems 5.9 and 5.21. The compactness assumption on Θ is the simplest way to ensure consistency, but may be relaxed to conditions similar to those discussed in Walker (1969) or Huber (1967). For asymptotic normality, assumptions on the existence of first and second moments of $\mathbf{h}$ and $\partial\mathbf{h}/\partial\theta$ are needed, as well as an assumption on the model and the estimating equation that ensures differentiation with respect to θ and expectation can be exchanged.

3 Applications to inference for binary data

In this section we look at several examples related to binary data, where biased estimating equations arise from ignoring various complexities of the data. We show that the method of adjusting the estimator based on (4) can be simpler than correcting the bias of the estimating function and can also lead to insight about the effect of ignoring the complexities.

First we illustrate the method with a somewhat artificial example related to Example 1.
Example 2: complete binary data. Suppose as in Example 1 that $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^T$ is a random sample of bivariate binary vectors, $i = 1, ..., n$ with $E(Y_{ij}) = \mu, j = 1, 2$, and $corr(Y_{i1}, Y_{i2}) = \rho$ for $i = 1, ..., n$, and $\theta = (\mu, \rho)^T$.

As shown in Example 1 at (5), consistent estimators are available for μ and ρ from a simple method of moments approach. If we deliberately misspecify estimating functions by switching the meaning of moments, considering for example

$$h_\mu(\mathbf{Y}_i; \theta) = Y_{i1}Y_{i2} - \mu, \quad h_\rho(\mathbf{Y}_i; \theta) = Y_{i1} + Y_{i2} - \rho,$$

the resulting estimator is

$$\hat{\mu}_n^* = \frac{\sum_{i=1}^n Y_{i1}Y_{i2}}{n}, \qquad \hat{\rho}_n^* = \frac{\sum_{i=1}^n (Y_{i1} + Y_{i2})}{n}. \tag{16}$$

Obviously, $\widehat{\theta} = (\hat{\mu}^*, \hat{\rho}^*)^T$ is not a consistent estimator for θ. Applying the adjustment function (2) to $\mathbf{h}(\mathbf{y}_i; \theta)$:

$$\begin{pmatrix} E_\theta(Y_{i1}Y_{i2}) - \mu^* \\ E_\theta(Y_{i1} + Y_{i2}) - \rho^* \end{pmatrix} = \begin{pmatrix} \rho\mu(1-\mu) + \mu^2 - \mu^* \\ 2\mu - \rho^* \end{pmatrix} = \mathbf{0},$$

we obtain

$$\mu = \frac{1}{2}\rho^*, \quad \rho = \frac{4\mu^* - \rho^{*2}}{\rho^*(2 - \rho^*)}, \tag{17}$$

which gives the adjusted estimator

$$\hat{\mu}_n = \frac{\sum_{i=1}^n (Y_{i1} + Y_{i2})}{2n}, \quad \hat{\rho}_n = \frac{4n\sum_{i=1}^n Y_{i1}Y_{i2} - \{\sum_{i=1}^n (Y_{i1} + Y_{i2})\}^2}{\sum_{i=1}^n (Y_{i1} + Y_{i2})\{2n - \sum_{i=1}^n (Y_{i1} + Y_{i2})\}} \tag{18}$$

which is identical to (5).

We may alternatively consider another set of misspecified functions

$$h_\mu(\mathbf{Y}_i;\theta) = Y_{i1} - \mu, \quad h_\rho(\mathbf{Y}_i;\theta) = Y_{i1}Y_{i2} - \rho,$$

which produces

$$\hat{\mu}_n^* = \frac{\sum_{i=1}^n Y_{i1}}{n}, \quad \hat{\rho}_n^* = \frac{\sum_{i=1}^n Y_{i1}Y_{i2}}{n}. \tag{19}$$

Note here that $\hat{\mu}_n^*$ is a consistent estimator for μ, but $\hat{\rho}_n^*$ is not consistent for ρ. We now apply the adjustment function (2) to $\mathbf{h}(\mathbf{y}_i;\theta^*)$:

$$\begin{pmatrix} E_\theta(Y_{i1}) - \mu^* \\ E_\theta(Y_{i1}Y_{i2}) - \rho^* \end{pmatrix} = \begin{pmatrix} \mu - \mu^* \\ \rho\mu(1-\mu) + \mu^2 - \rho^* \end{pmatrix} = \mathbf{0},$$

yielding

$$\mu = \mu^*, \quad \rho = \frac{\rho^* - \mu^{*2}}{\mu^*(1-\mu^*)}. \tag{20}$$

Applying (20) to (19), we obtain an adjusted estimator

$$\hat{\mu}_n = \frac{\sum_{i=1}^n Y_{i1}}{n}, \quad \hat{\rho}_n = \frac{n\sum_{i=1}^n Y_{i1}Y_{i2} - (\sum_{i=1}^n Y_{i1})^2}{(\sum_{i=1}^n Y_{i1})(n - \sum_{i=1}^n Y_{i1})},$$

which is consistent for θ, although clearly less efficient than (18).

As in Example 1, suppose now that there is missing data, and R_{ij} records whether or not Y_{ij} is missing, for $j = 1, 2$ and $i = 1, \ldots, n$. Using these misspecified estimating equations for the observed data gives

$$h_\mu(\mathbf{Y}_i;\theta) = R_{i1}Y_{i1} - \mu, \quad h_\rho(\mathbf{Y}_i;\theta) = R_{i1}R_{i2}Y_{i1}Y_{i2} - \rho.$$

Then the resulting estimator is

$$\hat{\mu}_n^* = \frac{\sum_{i=1}^n R_{i1}R_{i2}Y_{i1}Y_{i2}}{n}, \quad \hat{\rho}_n^* = \frac{\sum_{i=1}^n R_{i1}Y_{i1}}{n}. \tag{21}$$

Adjusting it as before gives

$$E_\theta(R_{i1}Y_{i1}) = \mu^*, \quad E_\theta(R_{i1}R_{i2}Y_{i1}Y_{i2}) = \rho^*,$$

which leads to

$$\begin{aligned} \mu &= \{1 + \exp(-\alpha_0 - \alpha_1)\}\mu^*, \\ \rho &= \frac{\{1 + \exp(-\gamma_0 - 2\gamma_1)\}\rho^* - \{1 + \exp(-\alpha_0 - \alpha_1)\}^2\mu^{*2}}{\{1 + \exp(-\alpha_0 - \alpha_1)\}\mu^*[1 - \{1 + \exp(-\alpha_0 - \alpha_1)\}\mu^*]}. \end{aligned} \tag{22}$$

Combining (22) with (21) gives a consistent estimator for θ.

We now consider extension to a regression setting, assuming Y_{ij} is a binary response for subject i at time point j, $j = 1, ..., m$, with an associated covariate vector $\mathbf{x}_{ij}$, and model the mean vector as a logistic regression:

$$\text{logit } \mu_{ij} = \theta^T \mathbf{x}_{ij}, \tag{23}$$

where $\mu_{ij} = E(Y_{ij}|\mathbf{x}_i)$ with $\mathbf{x}_i = (\mathbf{x}_{i1}^T, ..., \mathbf{x}_{im}^T)^T$. The score equations for θ, assuming independence of the observations in both i and j, are

$$\sum_{i=1}^{n} U_i(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{x}_{ij} \left\{ y_{ij} - \frac{\exp(\theta^T \mathbf{x}_{ij})}{1 + \exp(\theta^T \mathbf{x}_{ij})} \right\}; \tag{24}$$

these are also the generalized estimating equations (Liang and Zeger, 1986), under a working model of independence. Denote by $\widehat{\theta}_U$ the estimator based on (24).

For computing $\mathbf{k}(\theta^*)$ in settings with missing or misclassified data, discussed below, we will use the alternative unbiased estimating equation

$$\sum_{i=1}^{n} \mathbf{g}(\mathbf{y}_i; \theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{x}_{ij} [y_{ij} \{1 + \exp(\theta^T \mathbf{x}_{ij})\} - \exp(\theta^T \mathbf{x}_{ij})], \tag{25}$$

and denote by $\widehat{\theta}_g$ the estimator based on (25). In the special case that a single covariate $x_{ij} = 0$ or 1, both $\widehat{\theta}_U$ and $\widehat{\theta}_g$ are given by

$$\exp(\widehat{\theta}_U) = \exp(\widehat{\theta}_g) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}(1 - y_{ij})},$$

although if $x_{ij} = \pm 1$ with equal frequencies, then

$$\exp(\hat{\theta}_U) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (1 + x_{ij}) y_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{m} (1 - x_{ij})(1 - y_{ij})}{\sum_{i=1}^{n} \sum_{j=1}^{m} (1 + x_{ij})(1 - y_{ij}) + \sum_{i=1}^{n} \sum_{j=1}^{m} (1 - x_{ij}) y_{ij}},$$

whereas

$$\exp(\hat{\theta}_g) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (1 - x_{ij})(1 - y_{ij})}{\sum_{i=1}^{n} \sum_{j=1}^{m} (1 + x_{ij})(1 - y_{ij})}.$$

Example 3: binary data with misclassification. Suppose now that we have some misclassification of the binary responses, so that the observed data is S_{ij}, where

$$\begin{aligned} \Pr(S_{ij} = 1 \mid Y_{ij} = 0) &= p_1 \\ \Pr(S_{ij} = 0 \mid Y_{ij} = 1) &= p_0, \end{aligned}$$

but that we ignore the misclassification error, and use the estimating function (25) based on S_{ij}:

$$\sum_{i=1}^{n} h(\mathbf{s}_i;\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{x}_{ij}[s_{ij}\{1+\exp(\theta^T\mathbf{x}_{ij})\} - \exp(\theta^T\mathbf{x}_{ij})]. \tag{26}$$

The linear structure of (26) simplifies the calculation of $E_\beta\{h(\mathbf{s}_i;\theta^*)\}$:

$$\begin{aligned} E_\theta\{h(\mathbf{S}_i;\theta^*)\} &= \sum_{j=1}^{m} \mathbf{x}_{ij}[(1-p_0)\{1+\exp(\theta^{*T}\mathbf{x}_{ij})\}\frac{\exp(\theta^T\mathbf{x}_{ij})}{1+\exp(\theta^T\mathbf{x}_{ij})} \\ &\quad + p_1\{1+\exp(\theta^{*T}\mathbf{x}_{ij})\}\frac{1}{1+\exp(\theta^T\mathbf{x}_{ij})} - \exp(\theta^{*T}\mathbf{x}_{ij})], \end{aligned} \tag{27}$$

where we have assumed that $\Pr(S_{ij}=s \mid Y_{ij}=y,\mathbf{x}_i) = \Pr(S_{ij}=s \mid Y_{ij}=y)$. The solution of θ as a function of θ^* obtained by setting (27) to zero defines $\widehat{\theta}_n$ as a function of $\widehat{\theta}_n^*$, the root of (26).

For the special case that $x_{ij}=0,1$, we get

$$\exp(\widehat{\theta}_n^*) = \frac{\sum_{i=1}^n\sum_{j=1}^m x_{ij}s_{ij}}{\sum_{i=1}^n\sum_{j=1}^m x_{ij}(1-s_{ij})}$$

and

$$\exp(\widehat{\theta}_n) = \frac{(1-p_1)\exp(\widehat{\theta}_n^*) - p_1}{1-p_0-p_0\exp(\widehat{\theta}_n^*)},$$

which will be different from $\exp(\hat{\theta}_n^*)$ unless $p_0=p_1=0$.

Because in this case $h(\mathbf{s}_i;\theta)$ has a simple linear structure, we can also construct an unbiased estimating equation from (26) using (27):

$$\begin{aligned} \widetilde{\mathbf{H}}_n(\theta) &= n^{-1}\sum_{i=1}^{n} h(\mathbf{s}_i;\theta) - E_\theta\{h(\mathbf{S}_i;\theta)\} \\ &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{x}_{ij}[\{1+\exp(\theta^T\mathbf{x}_{ij})\}s_{ij} - (1-p_0)\exp(\theta^T\mathbf{x}_{ij}) - p_1] \end{aligned}$$

which leads in the special case that $x_{ij}=0,1$ to

$$\exp(\widetilde{\theta}_n) = \frac{\sum_{i=1}^n\sum_{j=1}^m x_{ij}(p_1-s_{ij})}{\sum_{i=1}^n\sum_{j=1}^m x_{ij}(p_0-1+s_{ij})},$$

which is identical to $\exp(\widehat{\theta}_n)$.

Example 4: missing data. We now assume that there are some missing observations, and $\lambda_{ij} = \Pr(R_{ij}=1 \mid \mathbf{y}_i,\mathbf{x}_i)$, where logit $\lambda_{ij} = \alpha_0 + \alpha_1 y_{ij}$. Suppose we use the estimating equations (25), which are unbiased for complete data, for the observed data:

$$h(\mathbf{y}_i;\theta) = \sum_{j=1}^{m} r_{ij}\mathbf{x}_{ij}[\{1+\exp(\theta^T\mathbf{x}_{ij})\}y_{ij} - \exp(\theta^T\mathbf{x}_{ij})]. \tag{28}$$

$\sum_{i=1}^n h(\mathbf{y}_i;\widehat{\theta}_n^*)=\mathbf{0}$ defines the estimator $\widehat{\beta}_n^*$. Using calculations similar to those in Example 1 we obtain

$$E_\theta\{h(\mathbf{Y}_i;\theta^*)\}=\sum_{j=1}^m x_{ij}\left\{\frac{\exp(\alpha_0+\alpha_1)}{1+\exp(\alpha_0+\alpha_1)}\frac{\exp(\theta^T\mathbf{x}_{ij})}{1+\exp(\theta^T\mathbf{x}_{ij})}-\frac{\exp(\alpha_0)}{1+\exp(\alpha_0)}\frac{\exp(\theta^{*T}\mathbf{x}_{ij})}{1+\exp(\theta^T\mathbf{x}_{ij})}\right\}. \quad (29)$$

The naive estimator has the explicit expression, in the special case that $x_{ij}=0,1$

$$\exp(\hat{\theta}_n^*)=\frac{\sum_{i=1}^n\sum_{j=1}^m x_{ij}r_{ij}y_{ij}}{\sum_{i=1}^n\sum_{j=1}^m x_{ij}r_{ij}(1-y_{ij})}$$

and the adjusted version leads to

$$\exp(\hat{\theta}_n)=\frac{1+\exp(\alpha_0+\alpha_1)}{\exp(\alpha_1)+\exp(\alpha_0+\alpha_1)}\exp(\hat{\theta}_n^*)$$

indicating as in Example 1 attenuation or enhancement of the true effect as α_1 is greater than or less than 0.

Another way to obtain an unbiased estimating equation is to introduce λ_{ij} as a weight in (28), leading to the inverse probability weighted generalized estimating equations of Robins et al. (1995) and Fitzmaurice et al. (1995). These are

$$g(\mathbf{y}_i;\theta)=\sum_{j=1}^m\frac{1}{\lambda_{ij}}r_{ij}\mathbf{x}_{ij}[\{1+\exp(\theta^T\mathbf{x}_{ij})\}y_{ij}-\exp(\theta^T\mathbf{x}_{ij})],$$

and in the case of binary x's have the solution

$$\exp(\tilde{\theta}_g)=\frac{\sum_{i=1}^n\sum_{j=1}^m x_{ij}r_{ij}y_{ij}\exp(\alpha_0+\alpha_1y_{ij})/\{1+\exp(\alpha_0+\alpha_1y_{ij})\}}{\sum_{i=1}^n\sum_{j=1}^m x_{ij}r_{ij}(1-y_{ij})\exp(\alpha_0+\alpha_1y_{ij})/\{1+\exp(\alpha_0+\alpha_1y_{ij})\}}$$

which may be compared with the adjusted version

$$\exp(\hat{\theta}_n)=\frac{1+\exp(\alpha_0+\alpha_1)}{\exp(\alpha_1)+\exp(\alpha_0+\alpha_1)}\frac{\sum_{i=1}^n\sum_{j=1}^m x_{ij}r_{ij}y_{ij}}{\sum_{i=1}^n\sum_{j=1}^m x_{ij}r_{ij}(1-y_{ij})}.$$

If we try to obtain an unbiased estimating equation from $h(\mathbf{y}_i;\theta)$ using (1) the resulting expression involves a quadratic function of $\exp(\widetilde{\theta}_n)$ which is quite cumbersome.

Example 5: covariate misclassification. Now suppose that we have a single binary covariate x_{ij}, but misclassified, so that we observe w_{ij} with

$$P(w_{ij}=1|x_{ij}=0)=p_1 \quad \text{and} \quad P(w_{ij}=0|x_{ij}=1)=p_0.$$

We further assume $x_{ij}=1$ with probability π and 0 with probability $1-\pi$. The estimating function based on (25) is easier to work with than the GEE version (24), so we assume that we start with a naive estimating function

$$h(\mathbf{y}_i;\theta)=\sum_{j=1}^m w_{ij}[\{1+\exp(\theta w_{ij})\}y_{ij}-\exp(\theta w_{ij})].$$

We then have

$$E_\theta\{h(\mathbf{Y}_i;\theta^*)\} = m\left\{(1-p_0)\pi\frac{\exp(\theta)-\exp(\theta^*)}{1+\exp(\theta)} + p_1(1-\pi)\frac{1-\exp(\theta^*)}{2}\right\},$$

which, by (3), leads to

$$\exp(\theta^*) = \frac{\{2(1-p_0)\pi + p_1(1-\pi)\}\exp(\theta) + p_1(1-\pi)}{\{2(1-p_0)\pi + p_1(1-\pi)\} + p_1(1-\pi)\exp(\theta)}. \tag{30}$$

This relationship reveals that in special situations, such as $p_0 \neq 1$ but $p_1 = 0$ or $\pi = 1$, we have $\theta^* = \theta$. In general situations with $0 < p_1 \leq 1$ and $0 \leq \pi < 1$, we have $\theta^* \geq \theta$ if and only if $\theta \leq 0$.

The naive estimator is, from solving $\sum_{i=1}^n h(\mathbf{y}_i;\theta) = 0$, given by

$$\exp(\widehat{\theta}_n^*) = \frac{\sum_{i=1}^n\sum_{j=1}^m w_{ij}y_{ij}}{\sum_{i=1}^n\sum_{j=1}^m w_{ij}(1-y_{ij})}.$$

Therefore, the adjusted estimator is

$$\exp(\hat{\theta}_n) = \frac{2\{(1-p_0)\pi + p_1(1-\pi)\}\sum_{i=1}^n\sum_{j=1}^m w_{ij}y_{ij} - p_1(1-\pi)\sum_{i=1}^n\sum_{j=1}^m w_{ij}}{2\{(1-p_0)\pi + p_1(1-\pi)\}\sum_{i=1}^n\sum_{j=1}^m w_{ij}(1-y_{ij}) - p_1(1-\pi)\sum_{i=1}^n\sum_{j=1}^m w_{ij}}.$$

We compare this approach to that of correcting $h(\cdot)$ for its bias, which leads to the estimating equation

$$n^{-1}\sum_{i=1}^n\sum_{j=1}^m w_{ij}[\{1+\exp(\theta w_{ij})\}y_{ij} - \exp(\theta w_{ij})] - (m/2)[p_1(1-\pi)\{1-\exp(\theta)\}] = 0$$

and gives the consistent estimator of θ:

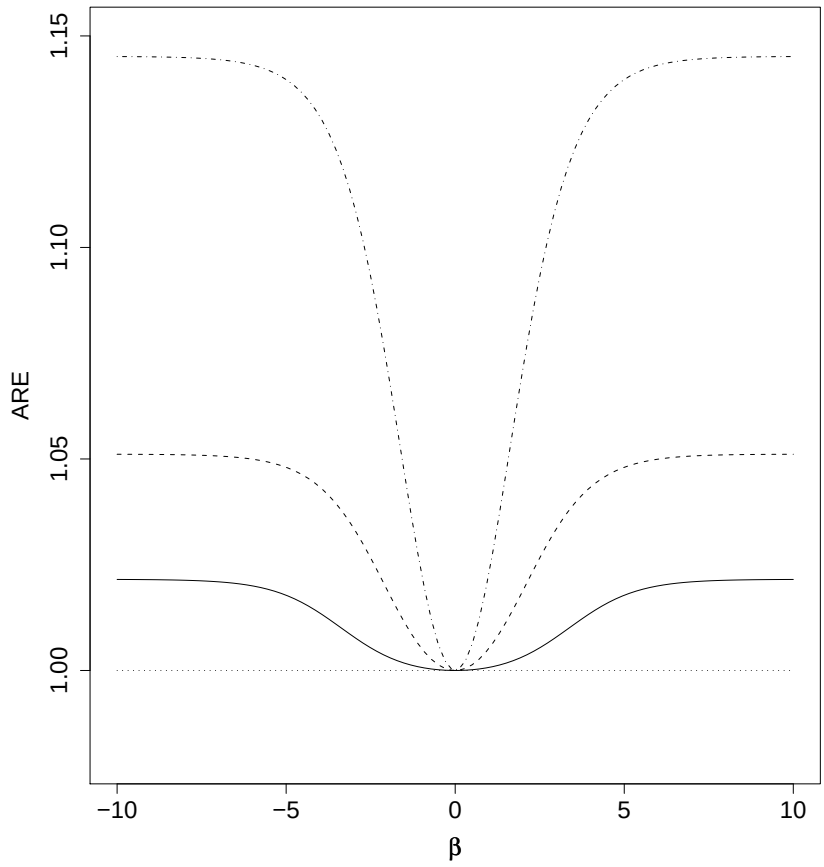
$$\exp(\widetilde{\theta}_n) = \frac{2\sum_{i=1}^n\sum_{j=1}^m w_{ij}y_{ij} - mnp_1(1-\pi)}{2\sum_{i=1}^n\sum_{j=1}^m w_{ij}(1-y_{ij}) - mnp_1(1-\pi)}.$$

Figure 1 shows the asymptotic relative efficiency of $\hat{\theta}_n$ and $\tilde{\theta}_n$, for three different choices of the probabilities of misclassification.

4 Comparison of Estimators

In this section we compare the estimators obtained from the two approaches described in Section 1: modifying the estimating equation by subtracting the bias, or modifying the point estimator using the relationship between θ^* and θ. As shown in Appendix A, in special cases these two methods may lead to the same estimators; however in general, the two estimators and their asymptotic variances are different.

Figure 1: Asymptotic relative efficiency of $\hat{\theta}_n$ relative to $\tilde{\theta}_n$, based on the expressions given in Theorem 2, presented as a function of θ for three choices of miisclassification probabilities: (i) $p_0 = 0.05$, $p_1 = 0.30$ (solid), (ii) $p_0 = 0.05, p_1 = 0.30$ (dashed), (iii) $p_0 = p_1 = 0.45$ (dashed-dotted). The x_{ij}s are equal to 0 or 1 with probability 1/2.



For ease of notation we consider the case that θ is a scalar, which is still instructive. Assume the subsequent quantities such as inverses and derivariatives all exist. Applying Taylor series expansions to $H_n(\widehat{\theta}^*)$ and $\widetilde{H}_n(\widehat{\theta})$ leads to

$$\widehat{\theta}_n = \theta - k'(\theta^*)\frac{H_n(\theta^*)}{H_n'(\theta^*)} + O_p(1/n),$$

and

$$\widetilde{\theta}_n = \theta - \frac{\widetilde{H}_n(\theta)}{\widetilde{H}_n'(\theta)} + O_p(1/n).$$

Now we examine approximations to the denominators $H_n'(\theta^*)$ and $\widetilde{H}_n'(\theta)$. Let $u(\theta)$ be the

score function and $\partial_\theta h(Y;\theta)$ denote the derivative of $h(Y;\theta)$ with respect to θ. Then

$$\begin{aligned} H_n'(\theta^*) &= E_\theta[H_n'(\theta^*)] + O_p(1/\sqrt{n}) \\ &= E_\theta[\partial_{\theta^*} h(Y;\theta^*)] + O_p(1/\sqrt{n}), \end{aligned}$$

and

$$\begin{aligned} \widetilde{H}_n'(\theta) &= E_\theta[\widetilde{H}_n'(\theta)] + O_p(1/\sqrt{n}) \\ &= E_\theta[\partial_\theta \widetilde{h}(Y;\theta)] + O_p(1/\sqrt{n}) \\ &= -E_\theta[u(\theta)\widetilde{h}(Y;\theta)] + O_p(1/\sqrt{n}) \text{ by unbiasedness of } \widetilde{h}(Y;\theta) \\ &= -E_\theta[u(\theta)h(Y;\theta)] - E_\theta[u(\theta)]E_\theta[h(Y;\theta)] + O_p(1/\sqrt{n}) \\ &= -E_\theta[u(\theta)h(Y;\theta)] + O_p(1/\sqrt{n}) \text{ by unbiasedness of } u(\theta), \end{aligned}$$

leading to

$$\widehat{\theta}_n - \widetilde{\theta}_n = -k'(\theta^*)\frac{H_n(\theta^*)}{E_\theta[\partial\theta^* h(Y;\theta^*)]} + \frac{\widetilde{H}_n(\theta^*)}{E_\theta[u(\theta)h(Y;\theta)]} + O_p(1/n). \tag{31}$$

By the definition of θ^*, we have

$$\int h(y;\theta^*) f(y;k(\theta^*))dy = 0. \tag{32}$$

Differentiating (32) with respect to θ^*, we obtain

$$\int \partial_{\theta^*} h(y;\theta^*) f(y;\theta)dy + \int h(y;\theta^*)\partial_\theta f(y;\theta)k'(\theta^*)dy = 0,$$

and hence,

$$k'(\theta^*) = -\frac{E_\theta[\partial_{\theta^*} h(Y;\theta^*)]}{E_\theta[u(\theta)h(Y;\theta^*)]}. \tag{33}$$

Expanding $h(y;\theta^*)$ in (32) around θ, we obtain

$$\int \{h(y;\theta) + (\theta^* - \theta)\partial_\theta h(y;\theta) + o(\theta^* - \theta)\} f(y;\theta)dy = 0$$

leading to

$$H(\theta) = -(\theta^* - \theta)E_\theta[\partial_\theta h(Y;\theta)] + o(\theta^* - \theta). \tag{34}$$

Substituting (33) and (34) into (31) yields

$$\begin{aligned} \widehat{\theta}_n - \widetilde{\theta}_n &= \frac{H_n(\theta^*)}{E_\theta[u(\theta)h(Y;\theta^*)]} - \frac{H_n(\theta) - H(\theta)}{E_\theta[u(\theta)h(Y;\theta)]} + O_p(1/\sqrt{n}) \\ &= \frac{H_n(\theta^*)}{E_\theta[u(\theta)h(Y;\theta^*)]} - \frac{H_n(\theta) + (\theta^* - \theta)E_\theta[\partial_\theta h(Y;\theta)] + o(\theta^* - \theta)}{E_\theta[u(\theta)h(Y;\theta)]} + O_p(1/\sqrt{n}) \\ &= \left\{\frac{H_n(\theta^*)}{E_\theta[u(\theta)h(Y;\theta^*)]} - \frac{H_n(\theta)}{E_\theta[u(\theta)h(Y;\theta)]}\right\} - (\theta^* - \theta)\frac{E_\theta[\partial_\theta h(Y;\theta)]}{E_\theta[u(\theta)h(Y;\theta)]} + o(\theta^* - \theta) + O_p(1/\sqrt{n}). \end{aligned}$$

Further examining the term in braces by Taylor expansion, we have

$$H_n(\theta^*) \;=\; H_n(\theta) + (\theta^* - \theta)H_n'(\theta) + o(\theta^* - \theta)$$

and

$$\begin{aligned} E_\theta[u(\theta)h(Y;\theta^*)] &= \int u(\theta)h(y;\theta^*)f(y;\theta)dy \\ &= \int u(\theta)h(y;\theta)f(y;\theta)dy + \int u(\theta)[\partial_{\theta^*}h(y;\theta^*)|_{\theta^*=\theta}]f(y;\theta)dy(\theta^* - \theta) + o(\theta^* - \theta) \\ &= E_\theta[u(\theta)h(Y;\theta)] + (\theta^* - \theta)E_\theta[u(\theta)\partial_\theta h(Y;\theta)] + o(\theta^* - \theta). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{H_n(\theta^*)}{E_\theta[u(\theta)h(Y;\theta^*)]} &= \frac{H_n(\theta)}{E_\theta[u(\theta)h(Y;\theta)]}\left\{1 + (\theta^* - \theta)\frac{H_n'(\theta)}{H_n(\theta)} + o(\theta^* - \theta)\right\} \\ &\quad \cdot\left\{1 - (\theta^* - \theta)\frac{E_\theta[u(\theta)\partial_\theta h(Y;\theta)]}{E_\theta[u(\theta)h(Y;\theta)]} + o(\theta^* - \theta)\right\} \\ &= \frac{H_n(\theta)}{E_\theta[u(\theta)h(Y;\theta)]}\left\{1 + (\theta^* - \theta)\left(\frac{H_n'(\theta)}{H_n(\theta)} - \frac{E_\theta[u(\theta)\partial_\theta h(Y;\theta)]}{E_\theta[u(\theta)h(Y;\theta)]}\right) + o(\theta^* - \theta)\right\}. \end{aligned}$$

As a result, we obtain

$$\begin{aligned} \widehat{\theta}_n - \widetilde{\theta}_n &= (\theta^* - \theta)\left\{\frac{H_n'(\theta)}{E_\theta[u(\theta)h(Y;\theta)]} - \frac{H_n(\theta)E_\theta[u(\theta)\partial_\theta h(Y;\theta)]}{(E_\theta[u(\theta)h(Y;\theta)])^2} - \frac{E_\theta[\partial_\theta h(Y;\theta)]}{E_\theta[u(\theta)h(Y;\theta)]}\right\} \\ &\quad + o(\theta^* - \theta) + O_p(1/\sqrt{n}). \end{aligned} \tag{35}$$

Equation (35) characterizes the difference between estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ obtained from the two methods; this depends on function $h(Y;\theta)$ and its derivative, the correlations of these two functions with the score function, and the asymptotic bias $\theta^* - \theta$.

The theory of estimating functions summarized in the introduction gives the result that under regularity conditions, $\sqrt{n}(\widetilde{\theta}_n - \theta)$ asymptotically follows a normal distribution with mean zero and variance $\Gamma^{-1}(\theta)\Sigma(\theta)\{\Gamma^{-1}(\theta)\}^T$, where $\Gamma(\theta) = E_\theta\{\partial\widetilde{H}_n(\theta)/\partial\theta^T\}$, and $\Sigma(\theta) = E_\theta\{\widetilde{H}_n(\theta)\widetilde{H}_n^T(\theta)\}$. Consequently, the asymptotic relative efficiency between estimators $\widetilde{\theta}_n$ and $\widehat{\theta}_n$ can be obtained using Theorem 2. In general, neither estimator will outperform the other uniformly. Depending on the specification of the $h(\mathbf{y};\theta)$ function, one estimator may lead to smaller asymptotic variance than the other. In Appendix B, we give an example to illustrate this point.

5 Discussion

In this paper we investigate issues concerning misspecification of estimating functions and establish some asymptotic properties. This gives a means for developing consistent estimators by modifying estimators obtained from convenient estimating functions which may not be unbiased. This may be particularly useful in understanding the bias induced by missing or mismeasured data. Starting from a manageable estimating function, we can apply Theorem 1 to obtain a consistent estimator, and Theorem 2 to choose among alternatives.

For incomplete longitudinal data, Rotnitzky and Wypij (1994) provide an algorithm for determining $\mathbf{k}(\theta^*)$ when the responses and covariates follow a discrete distribution, and illustrate this under an assumed model for missing data. This could be used to check if $\mathbf{k}(\theta^*)$ is monotone, which is needed for the application of the delta method in Theorem 2. Their Figure 1 is consistent with the results of our Examples 1 and 5, showing positive or negative asymptotic bias in the naive estimator. As they point out, their method does not give a means of constructing a bias adjustment. The current development could also be used as a convenient tool for indirect likelihood inference, reviewed in Jiang and Turnbull (2004). The formulation of an indirect likelihood requires an intermediate statistic that has an asymptotically normal distribution, and our results provide a theoretical basis for this.

One interesting extension of this work concerns partial misspecification of models. It may be possible to develop a hybrid inference method by combining the development here with the pairwise likelihood techniques discussed in Cox and Reid (2004). More convenient and efficient inference procedures may be generated to preserve robustness of estimating functions and efficiency of likelihood-related formulation.

Acknowledgement

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix A

Suppose we start with a biased estimating function $\sum_{i=1}^n \mathbf{h}(\mathbf{y}_i;\theta)$ and create an unbiased estimating function in the usual way as at (1):

$$\widetilde{\mathbf{H}}_n(\theta) = n^{-1}\sum_{i=1}^n \widetilde{\mathbf{h}}(\mathbf{y}_i;\theta) = n^{-1}\sum_{i=1}^n \mathbf{h}(\mathbf{y}_i;\theta) - E_\theta\{\mathbf{h}(\mathbf{Y};\theta)\}.$$

Denote by $\widetilde{\theta}_n$ the root of $\widetilde{\mathbf{H}}_n(\theta) = \mathbf{0}$. We know under regularity conditons on $\widetilde{\mathbf{h}}$ and the underlying family of distributions that $\widetilde{\theta}_n$ is consistent for θ as $n \to \infty$.

If $\mathbf{h}(\mathbf{y};\theta) = \mathbf{h}_1(\theta)\mathbf{h}_2(\mathbf{y}) + \mathbf{h}_3(\theta)$, where $\mathbf{h}_1(\theta)$ is a $p \times p$ non-singular matrix, and $\mathbf{h}_2(\mathbf{y})$ and $\mathbf{h}_3(\theta)$ are $p \times 1$ vectors, then this is identical to the adjustment method outlined at (3) and (4), as

$$E_\theta\{\mathbf{h}(\mathbf{Y};\theta^*)\} = \mathbf{h}_1(\theta^*)E_\theta\{\mathbf{h}_2(\mathbf{Y})\} + \mathbf{h}_3(\theta^*)$$

showing that $E_\theta\{\mathbf{h}_2(\mathbf{Y})\} = -\{\mathbf{h}_1(\theta^*)\}^{-1}\mathbf{h}_3(\theta^*)$ and thus that

$$\mathbf{K}(\widehat{\theta}_n) = -\{\mathbf{h}_3(\widehat{\theta}_n^*)\}^{-1}\mathbf{h}_3(\widehat{\theta}_n^*),$$

where $\mathbf{K}(\theta) = E_\theta\{\mathbf{h}_2(\mathbf{Y})\}$. On the other hand, $n^{-1}\sum_{i=1}^n[\mathbf{h}_2(\mathbf{y}_i) - E_\theta\{\mathbf{h}_2(\mathbf{Y}_i)\} = \mathbf{0}$ is solved by $\widetilde{\theta}_n$, showing that the two estimators are identical, provided $\mathbf{K}(\cdot)$ is a vector of monotone functions.

As an example to show that the methods lead to different estimators in nonlinear situations, suppose $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$ is a binary vector with independent components and $E(Y_{ij}) = \mu$. Let

$$h(\mathbf{y}_i;\mu) = \frac{\mu + y_{i2}}{1 + y_{i1}} - 1;$$

we have

$$E\{h(\mathbf{Y}_i;\mu)\} = 2\mu - \mu^2 - 1,$$

and hence

$$\widetilde{H}_n(\mu) = \frac{1}{n}\sum_{i=1}^n \frac{\mu + y_{i2}}{1 + y_{i1}} - (2\mu - \mu^2) = 0$$

has the solutions

$$\tilde{\mu}_n = (1/2)[2 - \frac{1}{n}\sum_{i=1}^n \frac{1}{1 + y_{i1}} \pm \sqrt{\{(\frac{1}{n}\sum_{i=1}^n \frac{1}{1 + y_{i1}})^2 - \frac{4}{n}\sum_{i=1}^n \frac{y_{i2}}{1 + y_{i1}}\}}],$$

where detailed examination indicates that for consistency we need to take the positive square root if $\mu \geq 2/3$ and otherwise the negative square root. Using the biased estimating equation we get the preliminary root

$$\hat{\mu}_n^* = \frac{n - \sum_{i=1}^n \sum_{j=1}^m y_{i2}/(1 + y_{i1})}{\sum_{i=1}^n \sum_{j=1}^m 1/(1 + y_{i1})},$$

and combining this with $E_\mu\{h(\mathbf{Y}_i;\mu^*)\} = (\mu^* + \mu)(1 - \mu/2) - 1$ gives

$$\hat{\mu}_n = (1/2)\{2 - \hat{\mu}_n^* \pm \sqrt{(\hat{\mu}_n^{*2} + 4\hat{\mu}_n^* - 4)}\}.$$

For example if the four pairs $(1,1)$, $(1,0)$, $(0,1)$ and $(0,0)$ have equal frequencies $n/4$ in the sample, then $\hat{\mu}_n$ is $2/3$ or $1/2$, whereas $\tilde{\mu}_n$ is $3/4$ or $1/2$.

Appendix B

We consider a simple case with independent binary variables Y_{i1} and Y_{i2}, $i = 1, 2, ..., n$. Let $\mu = E(Y_{i1}) = E(Y_{i2})$ be the parameter of interest. Consider an artificial, but instructive case in which the function $h(\mathbf{y}; \mu)$ is specified as

$$h(\mathbf{y}_i; \mu) = y_{i1} g_1(\mu) + y_{i2} g_2(\mu) + g_c(c)$$

for some functions $g_k(\cdot)(k = 1, 2, 3)$ and a constant c.

This function is not unbiased, as $E_\theta\{h(\mathbf{Y}; \mu)\} = \mu(g_1(\mu) + g_2(\mu)) + g_3(c)$. Thus μ^* is the point satisfying

$$\mu\{g_1(\mu^*) + g_2(\mu^*)\} + g_3(c) = 0. \tag{36}$$

It is easily seen that (3) is given by

$$\mu = k(\mu^*) = -\frac{g_3(c)}{g_1(\mu^*) + g_2(\mu^*)}. \tag{37}$$

To obtain the estimator $\widetilde{\mu}_n$, let

$$\widetilde{H}_n(\mu) = n^{-1} \sum_{i=1}^{n} h(\mathbf{y}_i; \mu) - \mu\{g_1(\mu) + g_2(\mu)\} - g_3(c).$$

Direct calculations lead to

$$\Gamma(\mu) = E_\mu\{\frac{\partial \widetilde{H}_n(\mu)}{\partial \mu}\} = \mu\{g_1'(\mu) + g_2'(\mu)\} - \{g_1(\mu) + g_2(\mu)\},$$

and

$$\Sigma(\mu) = E\{\widetilde{H}_n^2(\mu)\} = \mu(1 - \mu)\{g_1^2(\mu) + g_2^2(\mu)\}.$$

Therefore, $\text{avar}(\widetilde{\mu}_n) = \Gamma^{-2}(\mu)\Sigma(\mu)$.

Regarding the estimator $\widehat{\mu}_n$, direct calculations lead to

$$A(\mu) = E_\mu\{\frac{\partial h(\mathbf{Y}_i; \mu)}{\partial \mu}\} = \mu\{g_1'(\mu) + g_2'(\mu)\}$$

and

$$B(\mu) = E_\mu\{h^2(\mathbf{Y}_i; \mu)\} = \mu\{g_1^2(\mu) + g_2^2(\mu)\} + g_3^2(c) + 2\mu^2 g_1(\mu) g_2(\mu) + 2\mu\{g_1(\mu) + g_2(\mu)\} g_3(c).$$

Therefore,

$$\text{avar}(\widehat{\mu}_n) = \left\{\frac{\partial k(\mu^*)}{\partial \mu^*}\right\}^2 A^{-2}(\mu^*) B(\mu^*).$$

It is readily seen that by choice of the functions $g_k(\cdot)$ and constant c, we can make $\text{avar}(\widehat{\mu}_n)$ be smaller than $\text{avar}(\widetilde{\mu}_n)$, or vice versa. For example, with $g_1(t) = t$ and $g_2(t) = 1$, then

$$\text{avar}(\widetilde{\mu}_n) \quad = \quad \mu(1 - \mu)(1 + \mu^2)$$

and

$$\text{avar}(\widehat{\mu}_n) = \frac{\mu^5}{g_3^2(c)(g_3(c)+\mu)} + \frac{2\mu^4}{g_3(c)+\mu} - \frac{3\mu^3(g_3(c)+\mu)}{g_3^2(c)} + \frac{\mu^6}{g_3(c)(g_3(c)+\mu)}$$

in combination with (37). Given a value of μ, choosing a function $g_3(\cdot)$ and a consistant c satisfying

$$\begin{aligned} & (1-\mu+\mu^2-\mu^3)g_3^4(c) + (3+\mu-\mu^2-\mu^3-\mu^4)g_3^3(c) \\ & +(9\mu+\mu^2-\mu^3-\mu^4-\mu^5)g_3^2(c) + (9\mu^2-\mu^3-\mu^4)g_3(c) + (3\mu^3-\mu^5) \\ \geq\ & 0 \end{aligned}$$

results in $\text{avar}(\widehat{\mu}_n) \leq \text{avar}(\widetilde{\mu}_n)$. In particular, choosing a non-negative $g_3(c)$ leads to a more efficient estimator $\widehat{\mu}_n$ asymptotically.

References

[1] Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91, 729-737.

[2] Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society, Ser. B*, 57, 691-704.

[3] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31, 1208-1211.

[4] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, Berkeley: University of California Press, 221-233.

[5] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40, 633-643.

[6] Jiang, W., Turnbull, B. W., and Clark, L. C. (1999). Semiparametric regression models for repeated events with random effects and measurement error. *Journal of the American Statistical Association*, 94, 111-124.

[7] Jiang, W. and Turnbull, B. W. (2004). The indirect method: inference based on intermediate statistics - A synthesis and examples. *Statistical Science*, 19, 239-263.

[8] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

[9] McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society. Series B*, 52, 325-344.

[10] Robins, J.M., Rotnizky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

[11] Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, 50, 1163-1170.

[12] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[13] Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B*, 31, 80-88.

[14] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

[15] Yanagimoto, T. and Yamamoto, E. (1991). The role of unbiasedness in estimating equations. *Estimating Functions*. Ed. by V. P. Godambe. Oxford University Press, New York.