# A NOTE ON MIS-SPECIFIED ESTIMATING FUNCTIONS

Grace Y. Yi and Nancy Reid

*University of Waterloo and University of Toronto*

*Abstract:* We consider the use of estimating functions that are not unbiased. Typically, to result in consistent estimators, unbiasedness of estimating functions is a pre-requisite. However, it is sometimes 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 and phrases:* Binary data, bridge function, Godambe information, missing at random, misclassified.

## 1. Introduction

We consider estimation of a vector parameter $\theta$, based on a sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of independent and identically distributed random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ defined on $\mathcal{Y} \subset \mathbb{R}^d$, drawn from the family of densities $\{f(\mathbf{y}; \theta) : \theta \in \Theta\}$, where $\Theta$ is a subset of $\mathbb{R}^p$, and $d$ and $p$ are the dimensions of $\mathbf{Y}_i$ and $\theta$, respectively. We distinguish random variables from their realizations by using upper case and lower case letters, respectively, and use the notation $\mathbf{Y}$ for a random variable with the same distribution as any $\mathbf{Y}_i$. 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 covariance matrix given by the inverse of the Godambe information $J^{-1}(\mathbf{g}) = \{E_\theta(\partial \mathbf{g}/\partial \theta^T)\}^{-1} E_\theta(\mathbf{g}\mathbf{g}^T) \{E_\theta(\partial \mathbf{g}^T/\partial \theta)\}^{-1}$ (Godambe (1960)). Yanagimoto and Yamamoto (1991) give a number of examples illustrating the role of unbiasedness of estimating equations, and relate it to conditional likelihood inference in the context of exponential families.

In some practical contexts, however, there is a natural choice of working estimating function that is not unbiased. We use the notation $\mathbf{h}(\mathbf{y}; \theta)$ for the vector of biased estimating functions, i.e., we assume $E_\theta\{\mathbf{h}(\mathbf{Y}_i; \theta)\} \neq \mathbf{0}$. 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

$$\widetilde{\mathbf{H}}_n(\theta) = n^{-1} \sum_{i=1}^n \mathbf{h}(\mathbf{y}_i; \theta) - E_\theta\{\mathbf{h}(\mathbf{Y}_i; \theta)\}; \tag{1.1}$$

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 the correction of estimating functions derived from the method of moments.

In this paper we consider a different, but related, approach to deriving a consistent estimate of $\theta$ from a set of biased estimating functions. Assume that the equation

$$\mathbf{H}_n(\theta) = n^{-1} \sum_{i=1}^n \mathbf{h}(\mathbf{y}_i; \theta) = \mathbf{0} \tag{1.2}$$

has a root $\widehat{\theta}_n^* \in \Theta$ for any given random sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$, and that for any $\theta \in \Theta$, there exists a $\theta^* \in \Theta$ for which

$$E_\theta\{\mathbf{h}(\mathbf{Y}; \theta^*)\} = \mathbf{0}. \tag{1.3}$$

Equation (1.3) defines $\theta^*$ as a function of $\theta$, say, $\theta^* = \widetilde{\mathbf{k}}(\theta)$ for some $p$-vector of functions $\widetilde{\mathbf{k}}(\cdot)$. Assuming the inverse functions

$$\theta = \mathbf{k}(\theta^*) \tag{1.4}$$

exist, then we use this to define a new estimator of $\theta$ as

$$\widehat{\theta}_n = \mathbf{k}(\widehat{\theta}_n^*). \tag{1.5}$$

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, $i = 1, \ldots, n$. Assume that $E(Y_{ij}) = \mu, j = 1, 2$, and $\mathrm{corr}(Y_{i1}, Y_{i2}) = \rho$ for $i = 1, \ldots, n$. Let $\theta = (\mu, \rho)^T$ denote the parameter of interest. Let $R_{ij} = 1$ if $Y_{ij}$ is observed, and 0 otherwise, and set $\lambda_{ij} = \mathrm{Pr}(R_{ij} = 1|Y_{i1}, Y_{i2})$, and $\lambda_{i12} = \mathrm{Pr}(R_{i1} = 1, R_{i2} = 1|Y_{i1}, Y_{i2})$. Assume

$$\mathrm{logit}\, \lambda_{ij} = \alpha_0 + \alpha_1 Y_{ij}, \quad \mathrm{logit}\, \lambda_{i12} = \gamma_0 + \gamma_1(Y_{i1} + Y_{i2}).$$

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

$$\hat{\mu}_n = \frac{\sum_{i=1}^n (Y_{i1} + Y_{i2})}{2n}, \qquad \hat{\rho}_n = \frac{\sum_{i=1}^n Y_{i1}Y_{i2} - n\hat{\mu}_n^2}{n\hat{\mu}_n(1 - \hat{\mu}_n)}. \tag{1.6}$$

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

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

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

$$\hat{\mu}_n^* = \frac{\sum_{i=1}^n (R_{i1}Y_{i1} + R_{i2}Y_{i2})}{\sum_{i=1}^n (R_{i1} + R_{i2})}, \tag{1.7}$$

$$\hat{\rho}_n^* = \frac{\sum_{i=1}^n R_{i1}R_{i2}Y_{i1}Y_{i2} - \hat{\mu}_n^* \sum_{i=1}^n R_{i1}R_{i2}}{\hat{\mu}_n^*(1 - \hat{\mu}_n^*) \sum_{i=1}^n R_{i1}R_{i2}}. \tag{1.8}$$

To find $\theta^*$ we use (1.3) to compute

$$E_\theta \begin{Bmatrix} h_\mu(\mathbf{Y}_i; \theta^*) \\ h_\rho(\mathbf{Y}_i; \theta^*) \end{Bmatrix} = \begin{bmatrix} E_\theta(R_{i1}Y_{i1} + R_{i2}Y_{i2}) - \mu^* E_\theta(R_{i1} + R_{i2}) \\ E_\theta(R_{i1}R_{i2}Y_{i1}Y_{i2}) - \{\rho^*\mu^*(1-\mu^*) + \mu^{*2}\}E_\theta(R_{i1}R_{i2}) \end{bmatrix} = \mathbf{0}. \tag{1.9}$$

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,

$$E_\theta(R_{ij}) = \frac{\exp(\alpha_0 + \alpha_1)}{1 + \exp(\alpha_0 + \alpha_1)}\mu + \frac{\exp(\alpha_0)}{1 + \exp(\alpha_0)}(1 - \mu),$$

$$E_\theta(R_{i1}R_{i2}Y_{i1}Y_{i2}) = \frac{e^{\gamma_0 + 2\gamma_1}}{1 + e^{\gamma_0 + 2\gamma_1}}\{\rho\mu(1-\mu) + \mu^2\}, \quad \text{and}$$

$$E_\theta(R_{i1}R_{i2}) = \frac{\exp(\gamma_0 + 2\gamma_1)}{1 + \exp(\gamma_0 + 2\gamma_1)}\{\rho\mu(1-\mu) + \mu^2\}$$

$$+ \frac{2\exp(\gamma_0 + \gamma_1)}{1 + \exp(\gamma_0 + \gamma_1)}\{\mu - \rho\mu(1-\mu) - \mu^2\}$$

$$+ \frac{\exp(\gamma_0)}{1 + \exp(\gamma_0)}(1 - 2\mu + \rho\mu(1-\mu) + \mu^2).$$

Therefore, the first equation of (1.9) gives

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

with the same relationship between $\hat{\mu}_n$ and $\hat{\mu}_n^*$. It is easily shown that $\mu$ is equal to, less than, or greater than $\mu^*$ as $\alpha_1 = 0$, $\alpha_1 > 0$, and $\alpha_1 < 0$. If the data are missing completely at random, then $\alpha_1 = 0$ 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 are 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 (1.9), we obtain the relationship between $\rho$ and $\rho^*$. 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 $\mu$ and $\rho$ by using (1.9) to compute $E_\theta\{\mathbf{h}(\mathbf{Y}; \theta)\}$ and constructing $\widetilde{\mathbf{H}}_n(\theta)$, as at (1.1). In Appendix A we show that this is the case whenever the estimating equation $\mathbf{h}(\mathbf{y}; \theta)$ has a structure that is linear in functions of $\mathbf{y}$ and $\theta$, and we 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), that 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, $s_0(\cdot)$ from their Proposition 1, is our $\widetilde{\mathbf{k}}(\cdot)$.

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 comparison for the estimators obtained from (1.1) and (1.5), and Section 5 provides a brief discussion.

## 2. Asymptotic Results

**Theorem 1.** *Suppose* $\mathbf{h}(\mathbf{y}; \theta) = \{h_1(\mathbf{y}; \theta), \ldots, h_p(\mathbf{y}; \theta)\}^T$ *is a vector of functions defined on* $\mathcal{Y} \times \Theta$ *such that* $h_j(\mathbf{y}; \theta)$ *is a continuous function of* $\theta$ *for each* $\mathbf{y}$ *and a measurable function of* $\mathbf{y}$ *for each* $\theta$, $j = 1, \ldots, p$. *Assume that* $\Theta$ *is a*

*convex compact set and the true distribution of $\mathbf{Y}$ is $F = F(\mathbf{y}; \theta_0)$, with density $f(\mathbf{y}; \theta_0)$ for some $\theta_0 \in \Theta$. Assume $|h_j(\mathbf{y}; \theta)| \leq m_j(\mathbf{y})$ for all $\mathbf{y}$ and $\theta$ where $m_j(\cdot)$ is integrable with respect to $F$, $j = 1, \ldots, p$. Let $\mathbf{H}(\theta) = E_{\theta_0}\{\mathbf{h}(\mathbf{Y}; \theta)\}$, and $\mathbf{H}_n(\theta) = n^{-1}\sum_{i=1}^{n} \mathbf{h}(\mathbf{y}_i; \theta)$. If $\mathbf{H}(\theta) = \mathbf{0}$ has a unique solution $\theta_0^*$ and $\mathbf{H}_n(\theta) = \mathbf{0}$ has a solution $\widehat{\theta}_n^*$, then*

$$\widehat{\theta}_n^* \to_p \theta_0^* \quad \text{as} \quad n \to \infty$$

*for almost every sequence $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ which is a random sample from $F$.*

**Proof.** For any $\mathbf{y} \in \mathcal{Y}$, since $h_j(\mathbf{y}; \theta)$ is continuous for $\theta \in \Theta$ and $\Theta$ is compact, by the Heine-Cantor Theorem $h_j(\mathbf{y}; \theta)$ is uniformly continuous in $\theta$. It then follows that $||\mathbf{H}(\theta)||$ is continuous, where $||\cdot||$ is Euclidean norm.

Let $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$ be the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. Since the set $\{\theta : d(\theta, \theta_0^*) \geq \eta\} = \Theta - \{\theta : d(\theta, \theta_0^*) < \eta\}$ is a compact subset of $\Theta$ for any $\eta > 0$ there exists $\theta_1 \in \{\theta : d(\theta, \theta_0^*) \geq \eta\}$ such that

$$\inf_{\theta : d(\theta, \theta_0^*) \geq \eta} ||\mathbf{H}(\theta)|| = ||\mathbf{H}(\theta_1)||.$$

As $\theta_0^*$ is the unique solution of $\mathbf{H}(\theta) = \mathbf{0}$, and $\theta_1 \neq \theta_0^*$, we have $||\mathbf{H}(\theta_1)|| > 0$, i.e., $\inf_{\theta : d(\theta, \theta_0^*) \geq \eta} ||\mathbf{H}(\theta)|| > 0$. Furthermore, $\mathbf{H}_n(\widehat{\theta}_n^*) = \mathbf{0}$ gives

$$\inf_{\theta : d(\theta, \theta_0^*) \geq \eta} ||\mathbf{H}(\theta)|| > 0 = ||\mathbf{H}_n(\widehat{\theta}_n^*)||. \tag{2.1}$$

Given $j$, by Theorem 2 of Jennrich (1969), we have, for almost every sequence $\{\mathbf{Y}_n\}$, as $n \to \infty$,

$$n^{-1}\sum_{i=1}^{n} h_j(\mathbf{Y}_i; \theta) \to \int h_j(\mathbf{y}; \theta) dF(\mathbf{y}; \theta_0)$$

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

$$\sup_{\theta \in \Theta} d\{\mathbf{H}_n(\theta), \mathbf{H}(\theta)\} \to_p 0, \tag{2.2}$$

By (2.2) and (2.1), we conclude, applying Theorem 5.9 of van der Vaart (1998, p.46), $\widehat{\theta}_n^* \to_p \theta_0^*$ 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 $\theta_0^* - \theta_0$ is the asymptotic bias of using estimating functions that are not unbiased to perform estimation of $\theta$. In particular, if $\mathbf{h}(\mathbf{Y}; \theta)$ is unbiased, then $\theta_0^* = \theta_0$ and $\widehat{\theta}_n^*$ is consistent for $\theta$. If $\mathbf{k}(\cdot)$ 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 $\theta$.

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

$$\mathbf{A}_n(\theta) = n^{-1} \sum_{i=1}^n (\partial/\partial\theta)\{\mathbf{h}(\mathbf{Y}_i; \theta)\}^T, \qquad \mathbf{A}(\theta) = E_{\theta_0}\{\mathbf{A}_n(\theta)\},$$

$$\mathbf{B}_n(\theta) = n^{-1} \sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \theta)\{\mathbf{h}(\mathbf{Y}_i; \theta)\}^T, \quad \mathbf{B}(\theta) = E_{\theta_0}\{\mathbf{B}_n(\theta)\},$$

$$\mathbf{C}_n(\theta) = \mathbf{A}_n^{-1}(\theta)\mathbf{B}_n(\theta)\mathbf{A}_n^{-T}(\theta), \quad \text{and} \quad \mathbf{C}(\theta) = \mathbf{A}^{-1}(\theta)\mathbf{B}(\theta)\mathbf{A}^{-T}(\theta).$$

**Theorem 2**. *Suppose the conditions in Theorem 1 are satisfied, and $h_j(\mathbf{y}; \theta)$ is a continuously differentiable function of $\theta$ for each $\mathbf{y}$, $j = 1, \ldots, p$. Assume that $\mathbf{A}(\theta_0^*)$ is nonsingular, then under some regularity conditions on $h_j$ and the model $F$, we have: as $n \to \infty$,*

(i)   $\sqrt{n}(\widehat{\theta}_n^* - \theta_0^*) \to_d N\{\mathbf{0}, \mathbf{C}(\theta_0^*)\};$

(ii)  $\mathbf{C}_n(\widehat{\theta}_n^*) \to_p \mathbf{C}(\theta_0^*)$, *and assuming* $\mathbf{k}(\cdot)$ *defined at (1.4) exists and is differentiable,*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d N\left\{\mathbf{0}, \left(\frac{\partial \mathbf{k}^T(\theta_0^*)}{\partial\theta}\right)\mathbf{C}(\theta_0^*)\left(\frac{\partial \mathbf{k}(\theta_0^*)}{\partial\theta^T}\right)\right\}. \qquad (2.3)$$

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

$$\sum_{i=1}^n h_j(\mathbf{y}_i; \widehat{\theta}_n^*) = \sum_{i=1}^n h_j(\mathbf{y}_i; \theta_0^*) + \frac{\partial}{\partial\theta^T}\left\{\sum_{i=1}^n h_j(\mathbf{y}_i; \bar{\theta}_{jn})\right\}(\widehat{\theta}_n^* - \theta_0^*),$$

where $\bar{\theta}_{jn}$ lies on the "segment" joining $\widehat{\theta}_n^*$ and $\theta_0^*$. Stacking these $p$ expansions, we obtain, in matrix form,

$$\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, \ldots, \bar{\theta}_{pn})\sqrt{n}(\widehat{\theta}_n^* - \theta_0^*) = -n^{-1/2}\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \theta_0^*) + n^{-1/2}\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \widehat{\theta}_n^*),$$

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

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

$$\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, \ldots, \bar{\theta}_{pn})\sqrt{n}(\widehat{\theta}_n^* - \theta_0^*) = -n^{-1/2}\sum_{i=1}^n \mathbf{h}(\mathbf{Y}_i; \theta_0^*). \qquad (2.4)$$

As $E_{\theta_0}\{\mathbf{h}(\mathbf{Y}_i; \theta_0^*)\} = \mathbf{H}(\theta_0^*) = \mathbf{0}$, and $\mathrm{cov}_{\theta_0}\{\mathbf{h}(\mathbf{Y}_i; \theta_0^*)\} = E_{\theta_0}[\mathbf{h}(\mathbf{Y}_i; \theta_0^*)\{\mathbf{h}(\mathbf{Y}_i; \theta_0^*)\}^T]$ $= \mathbf{B}(\theta_0^*)$, by the Central Limit Theorem, we conclude

$$n^{-1/2} \sum_{i=1}^{n} \mathbf{h}(\mathbf{Y}_i; \theta_0^*) \to_d N\{\mathbf{0}, \mathbf{B}(\mathbf{\theta_0^*})\}. \qquad (2.5)$$

Note that for each $j = 1, \ldots, p$, $\bar{\theta}_{jn} \to_p \theta_0^*$ as $n \to \infty$. Therefore,

$$n^{-1} \sum_{i=1}^{n} \frac{\partial h_j(\mathbf{Y}_i; \bar{\theta}_{jn})}{\partial \theta} \to_p E_{\theta_0} \left\{ \frac{\partial h_j(\mathbf{Y}_i; \theta_0^*)}{\partial \theta} \right\}, \qquad (2.6)$$

and hence

$$\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, \ldots, \bar{\theta}_{pn}) \to_p \mathbf{A}(\theta_0^*) \quad \text{as } n \to \infty. \qquad (2.7)$$

Assuming that $\mathbf{A}(\theta_0^*)$ is nonsingular, we have that $\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, \ldots, \bar{\theta}_{pn})$ is nonsingular for sufficiently large $n$ (in probability). Therefore, (2.4) leads to

$$\sqrt{n}(\widehat{\theta}_n^* - \theta_0^*) = -\{\mathbf{A}_n^*(\bar{\theta}_{1n}, \bar{\theta}_{2n}, \ldots, \bar{\theta}_{pn})\}^{-1} n^{-1/2} \sum_{i=1}^{n} \mathbf{h}(\mathbf{Y}_i; \theta_0^*).$$

By (2.5) and (2.7),

$$\sqrt{n}(\widehat{\theta}_n^* - \theta_0^*) \to_d N[\mathbf{0}, \mathbf{A}^{-1}(\theta_0^*)\mathbf{B}(\theta_0^*)\mathbf{A}^{-T}(\theta_0^*)],$$

which is conclusion (i). Conclusion (ii) is straightforward as $\mathbf{B}_n(\widehat{\theta}_n^*) \to_p \mathbf{B}(\theta_0^*)$ and $\{\mathbf{A}_n^*(\widehat{\theta}_n^*)\}^{-1} \to_p \mathbf{A}^{-1}(\theta_0^*)$. The asymptotic normality of $\widehat{\theta}_n$ follows from an application of the delta method.

Result (2.3) provides means to conduct inference on $\theta$, such as constructing confidence intervals or testing hypotheses. In doing so, one may replace the relevant quantities with their empirical counterparts to obtain a consistent estimate for the asymptotic covariance matrix. The regularity conditions for Theorems 1 and 2 are similar to those outlined in Ch. 5 of Van der Waart (1998); see in particular the discussion following his Theorems 5.9 and 5.21. These conditions are sufficient to ensure the convergence in Theorems 1 and 2, but they are not necessarily the weakest conditions. The compactness assumption on $\Theta$ 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 $\partial \mathbf{h}/\partial \theta^T$ are needed, as well as an assumption on the model and the estimating equation that ensures differentiation with respect to $\theta$ and expectation can be exchanged.

## 3. Applications to Inference from 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 (1.5) can be simpler than correcting the bias of the estimating function, and can also lead to insight about the effect of ignoring the complexities.

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, $i = 1, \ldots, n$, with $E(Y_{ij}) = \mu, j = 1, 2$, $corr(Y_{i1}, Y_{i2}) = \rho$ for $i = 1, \ldots, n$, and $\theta = (\mu, \rho)^T$.

As shown in Example 1 at (1.6), consistent estimators are available for $\mu$ and $\rho$ 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

$$\hat{\mu}_n^* = \frac{1}{n}\sum_{i=1}^n Y_{i1}Y_{i2}, \qquad \hat{\rho}_n^* = \frac{1}{n}\sum_{i=1}^n (Y_{i1} + Y_{i2}). \tag{3.1}$$

Obviously, $\widehat{\theta} = (\hat{\mu}^*, \hat{\rho}^*)^T$ is not a consistent estimator for $\theta$. Applying the adjustment function (1.3) 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

$$\mu = \frac{1}{2}\rho^*, \quad \rho = \frac{4\mu^* - \rho^{*2}}{\rho^*(2 - \rho^*)}, \tag{3.2}$$

which gives the adjusted estimator

$$\hat{\mu}_n = \frac{1}{2n}\sum_{i=1}^n (Y_{i1} + Y_{i2}), \quad \hat{\rho}_n = \frac{4n\sum_{i=1}^n Y_{i1}Y_{i2} - \{\sum_{i=1}^n (Y_{i1} + Y_{i2})\}^2}{\sum_{i=1}^n (Y_{i1} + Y_{i2})\{2n - \sum_{i=1}^n (Y_{i1} + Y_{i2})\}} \tag{3.3}$$

that is identical to (1.6).

We consider another set of misspecified functions

$$h_\mu(\mathbf{Y}_i; \theta) = Y_{i1} - \mu, \quad h_\rho(\mathbf{Y}_i; \theta) = Y_{i1}Y_{i2} - \rho, \tag{3.4}$$

which produces

$$\hat{\mu}_n^* = \frac{1}{n}\sum_{i=1}^n Y_{i1}, \quad \hat{\rho}_n^* = \frac{1}{n}\sum_{i=1}^n Y_{i1}Y_{i2}. \tag{3.5}$$

Note here that $\hat{\mu}_n^*$ is a consistent estimator for $\mu$, but $\hat{\rho}_n^*$ is not consistent for $\rho$. We now apply the adjustment function (1.3) 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

$$\mu = \mu^*, \quad \rho = \frac{\rho^* - \mu^{*2}}{\mu^*(1-\mu^*)}. \tag{3.6}$$

Applying (3.6) to (3.5), we obtain an adjusted estimator

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

that is consistent for $\theta$, although clearly less efficient than (3.3).

As in Example 1, suppose now that there are missing data, and $R_{ij}$ records whether or not $Y_{ij}$ is missing, for $j = 1, 2$ and $i = 1, \ldots, n$. Using misspecified estimating equations (3.4) 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

$$\hat{\mu}_n^* = \frac{1}{n}\sum_{i=1}^n R_{i1}Y_{i1}, \quad \hat{\rho}_n^* = \frac{1}{n}\sum_{i=1}^n R_{i1}R_{i2}Y_{i1}Y_{i2}. \tag{3.7}$$

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

$$\begin{aligned} \mu &= \{1 + \exp(-\alpha_0 - \alpha_1)\}\mu^*, \\ \rho &= \frac{\{1 + \exp(-\gamma_0 - 2\gamma_1)\}\rho^* - \{1 + \exp(-\alpha_0 - \alpha_1)\}^2\mu^{*2}}{\{1 + \exp(-\alpha_0 - \alpha_1)\}\mu^*[1 - \{1 + \exp(-\alpha_0 - \alpha_1)\}\mu^*]}. \end{aligned} \tag{3.8}$$

Combining (3.8) with (3.7) gives a consistent estimator for $\theta$.

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

$$\text{logit } \mu_{ij} = \theta^T \mathbf{x}_{ij}, \tag{3.9}$$

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

$$\sum_{i=1}^{n} U_i(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{x}_{ij}\left\{ y_{ij} - \frac{\exp(\theta^T\mathbf{x}_{ij})}{1+\exp(\theta^T\mathbf{x}_{ij})}\right\}; \tag{3.10}$$

these are also the generalized estimating equations (GEE) (Liang and Zeger (1986)), under a working model of independence. Denote by $\widehat{\theta}_U$ the estimator based on (3.10).

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

$$\sum_{i=1}^{n} \mathbf{g}(\mathbf{y}_i;\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{x}_{ij}[Y_{ij}\{1+\exp(\theta^T\mathbf{x}_{ij})\} - \exp(\theta^T\mathbf{x}_{ij})], \tag{3.11}$$

and denote by $\widehat{\theta}_g$ the estimator based on (3.11). 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 response misclassification.** Suppose now that we have some misclassification of the binary responses, so that the observed responses are $s_{ij}$, where the model governing the random variables $S_{ij}$ is

$$\Pr(S_{ij} = 1 \mid Y_{ij} = 0) = p_1,$$
$$\Pr(S_{ij} = 0 \mid Y_{ij} = 1) = p_0.$$

Suppose we ignore the misclassification error, and use the estimating function (3.11) based on $s_{ij}$:

$$\sum_{i=1}^{n} h(\mathbf{s}_i;\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{x}_{ij}[s_{ij}\{1+\exp(\theta^T\mathbf{x}_{ij})\} - \exp(\theta^T\mathbf{x}_{ij})]. \tag{3.12}$$

The linear structure of (3.12) simplifies the calculation of $E_\theta\{h(\mathbf{S}_i; \theta^*)\}$:

$$E_\theta\{h(\mathbf{S}_i; \theta^*)\} = \sum_{j=1}^{m} \mathbf{x}_{ij}[(1 - p_0)\{1 + \exp(\theta^{*T}\mathbf{x}_{ij})\}\frac{\exp(\theta^T \mathbf{x}_{ij})}{1 + \exp(\theta^T \mathbf{x}_{ij})}$$

$$+ p_1\{1 + \exp(\theta^{*T}\mathbf{x}_{ij})\}\frac{1}{1 + \exp(\theta^T \mathbf{x}_{ij})} - \exp(\theta^{*T}\mathbf{x}_{ij})], \quad (3.13)$$

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 $\theta$ as a function of $\theta^*$ obtained by setting (3.13) to zero defines $\widehat{\theta}_n$ as a function of $\widehat{\theta}_n^*$, the root of (3.12).

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

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

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

which are different from $\exp(\widehat{\theta}_n^*)$ 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 (3.12) using (3.13):

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

$$= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{x}_{ij}[\{1 + \exp(\theta^T \mathbf{x}_{ij})\}s_{ij} - (1 - p_0)\exp(\theta^T \mathbf{x}_{ij}) - p_1],$$

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

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

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

**Example 4: missing responses.** 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 (3.11), which are unbiased for complete data, for the observed data:

$$h(\mathbf{y}_i; \theta) = \sum_{j=1}^{m} r_{ij}\mathbf{x}_{ij}[\{1 + \exp(\theta^T \mathbf{x}_{ij})\}y_{ij} - \exp(\theta^T \mathbf{x}_{ij})], \quad (3.14)$$

where $r_{ij}$ is a realization of the missing data indicator $R_{ij}$. $\sum_{i=1}^{n} h(\mathbf{y}_i; \widehat{\theta}_n^*) = \mathbf{0}$ defines the estimator $\widehat{\theta}_n^*$. Using calculations similar to those in Example 1 we obtain

$$E_\theta\{h(\mathbf{Y}_i; \theta^*)\} = \sum_{j=1}^{m} \mathbf{x}_{ij} \left\{ \frac{\exp(\alpha_0 + \alpha_1)}{1 + \exp(\alpha_0 + \alpha_1)} \frac{\exp(\theta^T \mathbf{x}_{ij})}{1 + \exp(\theta^T \mathbf{x}_{ij})} \right.$$
$$\left. - \frac{\exp(\alpha_0)}{1 + \exp(\alpha_0)} \frac{\exp(\theta^{*T} \mathbf{x}_{ij})}{1 + \exp(\theta^T \mathbf{x}_{ij})} \right\}. \tag{3.15}$$

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 $\alpha_1$ is greater than or less than 0.

Another way to obtain an unbiased estimating equation is to introduce $\lambda_{ij}$ as a weight in (3.14), leading to the inverse probability weighted generalized estimating equations of Robins, Rotnizky and Zhao (1995) and Fitzmaurice, Molenberghs and Lipsitz (1995). These are

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

and, in the case of binary $x$'s, have as the solution the estimator

$$\exp(\tilde{\theta}_g) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} r_{ij} Y_{ij} \{1 + \exp(-\alpha_0 - \alpha_1 Y_{ij})\}}{\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} r_{ij} (1 - Y_{ij}) \{1 + \exp(-\alpha_0 - \alpha_1 Y_{ij})\}}$$

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 (1.1), 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

$$\Pr(W_{ij} = 1|x_{ij} = 0) = p_1 \quad \text{and} \quad \Pr(W_{ij} = 0|x_{ij} = 1) = p_0,$$

where $p_0$ and $p_1$ have a different meaning than in Example 3. We further assume $x_{ij} = 1$ with probability $\pi$ and 0 with probability $1 - \pi$. The estimating function based on (3.11) is easier to work with than the GEE version (3.10), so we start with a naive estimating function replacing $x_{ij}$ by $w_{ij}$

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

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 (1.3) leads to

$$\exp(\theta^*) = \frac{\{2(1 - p_0)\pi + p_1(1 - \pi)\}\exp(\theta) + p_1(1 - \pi)}{\{2(1 - p_0)\pi + p_1(1 - \pi)\} + p_1(1 - \pi)\exp(\theta)}. \tag{3.16}$$

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

$$\exp(\widehat{\theta}_n) = \frac{2\{(1 - p_0)\pi + p_1(1 - \pi)\}\sum_{ij} W_{ij}Y_{ij} - p_1(1 - \pi)\sum_{ij}^{m} W_{ij}}{2\{(1 - p_0)\pi + p_1(1 - \pi)\}\sum_{ij}^{m} W_{ij}(1 - Y_{ij}) - p_1(1 - \pi)\sum_{ij}^{m} W_{ij}}.$$

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

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

and gives the consistent estimator

$$\exp(\widetilde{\theta}_n) = \frac{2\sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}Y_{ij} - mnp_1(1 - \pi)}{2\sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}(1 - Y_{ij}) - mnp_1(1 - \pi)}.$$
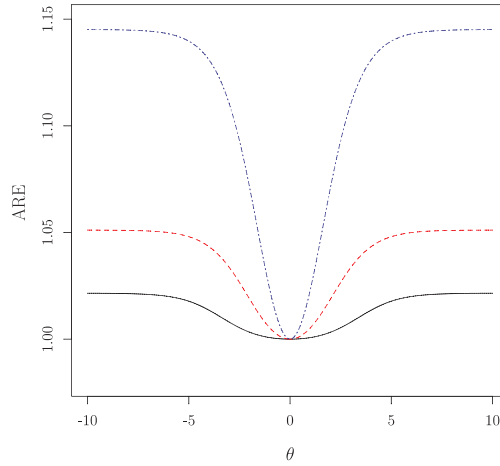
Figure 1. Asymptotic relative efficiency of $\hat{\theta}_n$ relative to $\tilde{\theta}_n$, based on the expressions given in Theorem 2, presented as a function of $\theta$ for three choices of misclassification probabilities: (i) $p_0 = 0.30$, $p_1 = 0.05$ (solid), (ii) $p_0 = 0.05, p_1 = 0.30$ (dashed), (iii) $p_0 = p_1 = 0.45$ (dashed-dotted). The $x_{ij}$s are equal to 0 or 1 with probability 1/2.

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 a biased estimating equation by subtracting its expectation, or modifying the point estimator using the relationship between $\theta^*$ and $\theta$. As shown in Appendix A, in special cases these two methods lead to the same estimators though in general, the two do not.

For ease of notation we consider the case that $\theta$ is a scalar. Assume subsequent quantities, such as inverses and derivatives, all exist. As before, $\widehat{\theta}_n^*$ is the estimator obtained from (1.2). Now let $\tilde{\theta}_n$ denote the estimator obtained from (1.1). Applying Taylor series expansions to $H_n(\widehat{\theta}_n^*)$ and $\widetilde{H}_n(\tilde{\theta}_n)$ around $\theta^*$ and $\theta$, respectively, and using (1.5) gives

$$\widehat{\theta}_n = \theta - k'(\theta^*)\frac{H_n(\theta^*)}{H_n'(\theta^*)} + O_p(\frac{1}{n}), \quad \text{and } \tilde{\theta}_n = \theta - \frac{\widetilde{H}_n(\theta)}{\widetilde{H}_n'(\theta)} + O_p(\frac{1}{n}).$$

Now we examine approximations to the denominators $H_n'(\theta^*)$ and $\widetilde{H}_n'(\theta)$ . Let $u(\theta)$ be the score function $(\partial/\partial\theta) \log f(y; \theta)$ and $\partial_\theta h(Y; \theta)$ denote the derivative of $h(Y; \theta)$ with respect to $\theta$. Then we find

$$H_n'(\theta^*) = E_\theta\{H_n'(\theta^*)\} + O_p(\frac{1}{\sqrt{n}}) = E_\theta\{\partial_{\theta^*} h(Y; \theta^*)\} + O_p(\frac{1}{\sqrt{n}}),$$

$$\widetilde{H}'_n(\theta) = E_\theta\{\widetilde{H}'_n(\theta)\} + O_p(\frac{1}{\sqrt{n}})$$

$$= E_\theta\{\partial_\theta \widetilde{h}(Y;\theta)\} + O_p(\frac{1}{\sqrt{n}})$$

$$= -E_\theta\{u(\theta)\widetilde{h}(Y;\theta)\} + O_p(\frac{1}{\sqrt{n}}) \text{ by unbiasedness of } \widetilde{h}(Y;\theta)$$

$$= -E_\theta\{u(\theta)h(Y;\theta)\} - E_\theta\{u(\theta)\}E_\theta\{h(Y;\theta)\} + O_p(\frac{1}{\sqrt{n}})$$

$$= -E_\theta\{u(\theta)h(Y;\theta)\} + O_p(\frac{1}{\sqrt{n}}) \text{ by unbiasedness of } u(\theta),$$

leading to

$$\widehat{\theta}_n - \widetilde{\theta}_n = -k'(\theta^*)\frac{H_n(\theta^*)}{E_\theta\{\partial_{\theta^*}h(Y;\theta^*)\}} + \frac{\widetilde{H}_n(\theta)}{E_\theta\{u(\theta)h(Y;\theta)\}} + O_p(\frac{1}{\sqrt{n}}). \qquad (4.1)$$

By the definition of $\theta^*$, we have

$$\int h(y;\theta^*)f(y;k(\theta^*))dy = 0. \qquad (4.2)$$

Differentiating (4.2) with respect to $\theta^*$, 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,

$$k'(\theta^*) = -\frac{E_\theta\{\partial_{\theta^*}h(Y;\theta^*)\}}{E_\theta\{u(\theta)h(Y;\theta^*)\}}. \qquad (4.3)$$

Expanding $h(y;\theta^*)$ in (4.2) around $\theta$, we obtain

$$\int \{h(y;\theta) + (\theta^* - \theta)\partial_\theta h(y;\theta) + o(\theta^* - \theta)\}f(y;\theta)dy = 0, \qquad (4.4)$$

leading to

$$H(\theta) = -(\theta^* - \theta)E_\theta\{\partial_\theta h(Y;\theta)\} + o(\theta^* - \theta). \qquad (4.5)$$

Here $o(\theta^* - \theta)$ denotes the remainder in the Taylor's expansion in (4.4), which can be expressed as a polynomial in $\theta^* - \theta$. Substituting (4.3) and (4.5) into (4.1) yields

$$\widehat{\theta}_n - \widetilde{\theta}_n = \frac{H_n(\theta^*)}{E_\theta\{u(\theta)h(Y;\theta^*)\}} - \frac{H_n(\theta) - H(\theta)}{E_\theta\{u(\theta)h(Y;\theta)\}} + O_p(\frac{1}{\sqrt{n}})$$

$$= \frac{H_n(\theta^*)}{E_\theta\{u(\theta)h(Y;\theta^*)\}} - \frac{H_n(\theta) + (\theta^* - \theta)E_\theta\{\partial_\theta h(Y;\theta)\} + o(\theta^* - \theta)}{E_\theta\{u(\theta)h(Y;\theta)\}}$$

$$+ O_p(\frac{1}{\sqrt{n}})$$

$$= \left[ \frac{H_n(\theta^*)}{E_\theta\{u(\theta)h(Y;\theta^*)\}} - \frac{H_n(\theta)}{E_\theta\{u(\theta)h(Y;\theta)\}} \right] - (\theta^* - \theta) \frac{E_\theta\{\partial_\theta h(Y;\theta)\}}{E_\theta\{u(\theta)h(Y;\theta)\}}$$

$$+ o(\theta^* - \theta) + O_p(\frac{1}{\sqrt{n}}).$$

Further examining the terms in square brackets by Taylor expansion, we have

$$H_n(\theta^*) = H_n(\theta) + (\theta^* - \theta)H'_n(\theta) + o(\theta^* - \theta),$$

$$E_\theta\{u(\theta)h(Y;\theta^*)\} = \int u(\theta)h(y;\theta^*)f(y;\theta)dy$$

$$= \int u(\theta)h(y;\theta)f(y;\theta)dy$$

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

$$= E_\theta\{u(\theta)h(Y;\theta)\} + (\theta^* - \theta)E_\theta\{u(\theta)\partial_\theta h(Y;\theta)\} + o(\theta^* - \theta).$$

Therefore,

$$\frac{H_n(\theta^*)}{E_\theta\{u(\theta)h(Y;\theta^*)\}} = \frac{H_n(\theta)}{E_\theta\{u(\theta)h(Y;\theta)\}} \left\{ 1 + (\theta^* - \theta)\frac{H'_n(\theta)}{H_n(\theta)} + o(\theta^* - \theta) \right\}$$

$$\cdot \left[ 1 - (\theta^* - \theta)\frac{E_\theta\{u(\theta)\partial_\theta h(Y;\theta)\}}{E_\theta\{u(\theta)h(Y;\theta)\}} + o(\theta^* - \theta) \right]$$

$$= \frac{H_n(\theta)}{E_\theta\{u(\theta)h(Y;\theta)\}} \left( 1 + (\theta^* - \theta)\left[ \frac{H'_n(\theta)}{H_n(\theta)} - \frac{E_\theta\{u(\theta)\partial_\theta h(Y;\theta)\}}{E_\theta\{u(\theta)h(Y;\theta)\}} \right] \right.$$

$$\left. + o(\theta^* - \theta) \right).$$

As a result, we obtain

$$\widehat{\theta}_n - \widetilde{\theta}_n = (\theta^* - \theta)\left( \frac{H'_n(\theta)}{E_\theta\{u(\theta)h(Y;\theta)\}} - \frac{H_n(\theta)E_\theta\{u(\theta)\partial_\theta h(Y;\theta)\}}{[E_\theta\{u(\theta)h(Y;\theta)\}]^2} \right.$$

$$\left. - \frac{E_\theta\{\partial_\theta h(Y;\theta)\}}{E_\theta\{u(\theta)h(Y;\theta)\}} \right) + o(\theta^* - \theta) + O_p\left(\frac{1}{\sqrt{n}}\right). \qquad (4.6)$$

Equation (4.6) as a formal expansion shows that the difference between estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ depends on the estimating function $h(Y;\theta)$ and its derivative, the correlations of these two functions with the score function, and the asymptotic bias $\theta^* - \theta$. The term $o(\theta^* - \theta)$ is only useful if there is some measure by which the asymptotic bias is small. In problems with missing or misclassified data, as

in our examples, this bias will be determined by the missing or misclassification mechanism.

The theory of estimating functions summarized in the introduction gives the result that under regularity conditions, $\sqrt{n}(\widetilde{\theta}_n - \theta)$ is asymptotically normal with mean zero and $r$ co/matrix $\mathbf{\Gamma}^{-1}(\theta)\mathbf{\Sigma}(\theta)\{\mathbf{\Gamma}^{-1}(\theta)\}^T$, where $\mathbf{\Gamma}(\theta) = E_\theta\{\partial\widetilde{\mathbf{h}}(\mathbf{Y};\theta)/\partial\theta^T\}$, $\mathbf{\Sigma}(\theta) = E_\theta\{\widetilde{\mathbf{h}}(\mathbf{Y};\theta)\widetilde{\mathbf{h}}^T(\mathbf{Y};\theta)\}$, and $\widetilde{\mathbf{h}}(\mathbf{Y};\theta) = \mathbf{h}(\mathbf{Y};\theta) - E_\theta\{\mathbf{h}(\mathbf{Y};\theta)\}$. 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. 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 that 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 note, their method does not give a means for constructing a bias adjustment.

As pointed out by a reviewer, in our examples the parameters governing the response are all assumed to be of interest. If some components of the parameter $\theta$ in the estimating equation are considered to be nuisance parameters, then the extension of the results here using constrained estimators for the nuisance parameters should be relatively straightforward.

The parameters governing the missing data or misclassification processes, on the other hand, are indeed nuisance parameters in this setting, and have been treated as known. This allows us to focus the discussion on the parameters of primary interest, and is useful for certain situations, especially for sensitivity analyses to assess the impact of different degrees of missingness or misclassification on estimation of the parameters of interest. If the values of these nuisance parameters are not known, they must be estimated, either from additional data sources or from a specification of a model for the missing data. For instance,

estimation of the misclassification parameters can be directly undertaken if there is a validation subsample, treated for example in Carroll et al. (2006). Another widely used approach to estimation of the missing data parameters is to fit a logistic regression to the missing data, which is appropriate under missing at random or missing completely at random mechanisms; see, e.g., Diggle and Kenward (1994). With data missing not at random, the issue of nonidentifiability may arise and sensitivity analyses can be a useful alternative to provide insight into the possible effects on estimation of the parameters of interest (Fitzmaurice, Molenberghs and Lipsitz (1995)). It would be interesting to incorporate a general theory of nuisance parameter estimation into the biased estimating equation framework but this is beyond the scope of this paper.

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. Another 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.

## Acknowledgements

## 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 as at (1.1):

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

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 $\theta$ 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 $\widetilde{\theta}_n$ is identical to the adjusted estimator outlined at (1.4) and (1.5), 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 $E_\theta\{\mathbf{h}_2(\mathbf{Y})\} = -\{\mathbf{h}_1(\theta^*)\}^{-1}\mathbf{h}_3(\theta^*)$. Since $0 = (1/n)\sum_{i=1}^n \mathbf{h}(\mathbf{y}_i; \widehat{\theta}_n^*) = \mathbf{h}_1(\widehat{\theta}_n^*)\sum_{i=1}^n h_2(y_i) + \mathbf{h}_3(\widehat{\theta}_n^*)$, $\mathbf{K}(\widehat{\theta}_n) = -\{\mathbf{h}_1(\widehat{\theta}_n^*)\}^{-1}\mathbf{h}_3(\widehat{\theta}_n^*)$, 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 $E_\mu\{h(\mathbf{Y}_i; \mu)\} = 2\mu - \mu^2 - 1$, 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

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

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

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

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

This example illustrates the discussion of relative efficiency at the end of Section 4. We consider a simple case with independent binary variables $Y_{i1}$ and $Y_{i2}$, $i = 1, \ldots, n$. Let $\mu = E(Y_{i1}) = E(Y_{i2})$ be the parameter of interest. Consider an artificial case in which the function $h(\mathbf{y}_i; \mu)$ is specified as $h(\mathbf{y}_i; \mu) = y_{i1}g_1(\mu) + y_{i2}g_2(\mu) + g_3(c)$ for some functions $g_k(\cdot)(k = 1, 2, 3)$ and a constant $c$. This function is not unbiased in general, as $E_\mu\{h(\mathbf{Y}; \mu)\} = \mu\{g_1(\mu) + g_2(\mu)\} + g_3(c)$. Thus $\mu^*$ is the point satisfying

$$\mu\{g_1(\mu^*) + g_2(\mu^*)\} + g_3(c) = 0. \tag{B.1}$$

It is easily seen that (1.4) is given by

$$\mu = k(\mu^*) = -\frac{g_3(c)}{g_1(\mu^*) + g_2(\mu^*)}. \tag{B.2}$$

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

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

Direct calculations lead to

$$\Gamma(\mu) = E_\mu\{\frac{\partial \widetilde{h}(\mathbf{Y}_i; \mu)}{\partial \mu}\} = -\{g_1(\mu) + g_2(\mu)\},$$

$$\Sigma(\mu) = E_\mu\{\widetilde{h}^2(\mathbf{Y}_i; \mu)\} = \mu(1 - \mu)\{g_1^2(\mu) + g_2^2(\mu)\}.$$

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

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

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

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

Therefore the asymptotic variance of $\widehat{\mu}_n$ is

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

It is readily seen that by choice of the functions $g_k(\cdot)$ and constant $c$, we can make this smaller or larger than the asymptotic variance of $\widetilde{\mu}_n$. For example, with $g_1(t) = t$ and $g_2(t) = 0$, then $\text{avar}(\widetilde{\mu}_n) = \mu(1 - \mu)$ and $\text{avar}(\widehat{\mu}_n) = -(\mu^3/g_3^2(c))[2g_3(c)\mu + \mu^3 - g_3(c)]$ in combination with (B.2). Given a value of $\mu$, choosing a function $g_3(\cdot)$ and a constant $c$ satisfying $g_3(c)(2\mu - 1) + \mu^3 > 0$ and $g_3^2(c)(1 - \mu) + g_3(c)(1 - 2\mu)\mu^2 - \mu^5 \geq 0$ results in $\text{avar}(\widehat{\mu}_n) \leq \text{avar}(\widetilde{\mu}_n)$. In particular, choosing $g_3(c) = -1$ leads to a more efficient estimator $\widehat{\mu}_n$ asymptotically if $\mu < (\sqrt{5} - 1)/2$.

## References

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, 2nd ed. Chapman & Hall/CRC, Boca Raton.

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

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.* **43**, 49-93.

Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *J. Roy. Statist. Soc. Ser. B* **57**, 691-704.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208-1211.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Symp. Math. Statist. and Probability*, University of California Press, Berkeley. 221-233.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40**, 633-643.

Jiang, W., Turnbull, B. W. and Clark, L. C. (1999). Semiparametric regression models for repeated events with random effects and measurement error. *J. Amer. Statist. Assoc.* **94**, 111-124.

Jiang, W. and Turnbull, B. W. (2004). The indirect method: inference based on intermediate statistics - A synthesis and examples. *Statist. Sci.* **19**, 239-263.

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

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **52**, 325-344.

Robins, J. M., Rotnizky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

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

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

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *J. Roy. Statist. Soc. Ser. B* **31**, 80-88.

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

Yanagimoto, T. and Yamamoto, E. (1991). The role of unbiasedness in estimating equations. In *Estimating Functions*. (Edited by by V. P. Godambe). Oxford University Press, Oxford.

Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1.

E-mail: yyi@uwaterloo.ca

Department of Statistics, University of Toronto, Sidney Smith Hall, 100 St. George St., Toronto, Ontario, M5S 3G3 Canada.

E-mail: reid@utstat.utoronto.ca