*Biostatistics* (2001), **2**, 3, *pp.* 295–307
*Printed in Great Britain*

# Testing for bias in weighted estimating equations

STUART LIPSITZ*
*Department of Biostatistics, Dana-Farber Cancer Institute, 44 Binney Street, Boston MA 02115, USA*
*Department of Biometry and Epidemiology, Medical University of South Carolina, USA*
Lipsitzs@musc.edu

MICHAEL PARZEN
*Graduate School of Business, University of Chicago, USA*

GEERT MOLENBERGHS
*Center for Statistics, Limburgs Universitair Centrum, Belgium*

JOSEPH IBRAHIM
*Department of Biostatistics, Dana-Farber Cancer Institute, 44 Binney Street, Boston MA 02115, USA*

## SUMMARY

It is very common in regression analysis to encounter incompletely observed covariate information. A recent approach to analyse such data is weighted estimating equations (Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994), *JASA*, **89**, 846–866, and Zhao, L. P., Lipsitz, S. R. and Lew, D. (1996), *Biometrics*, **52**, 1165–1182). With weighted estimating equations, the contribution to the estimating equation from a complete observation is weighted by the inverse of the probability of being observed. We propose a test statistic to assess if the weighted estimating equations produce biased estimates. Our test statistic is similar to the test statistic proposed by DuMouchel and Duncan (1983) for weighted least squares estimates for sample survey data. The method is illustrated using data from a randomized clinical trial on chemotherapy for multiple myeloma.

*Keywords*: Estimating equations; Generalized linear model; Missing at random; Missing covariate data.

## 1. INTRODUCTION

The regression modeller of, for example, clinical trials or sample surveys, is often confronted with incompletely observed covariate information. Precisely, we consider a regression analysis of an outcome $y$ on a vector $x = (x_1, \dots, x_K)'$ of $K$ covariates which are always observed, and a covariate $z$, which can be missing for some subjects. Our interest centres on estimating the regression parameters, say $\beta$. In our example in Section 4, we analyse a randomized clinical trial in multiple myeloma (Kalish, 1992). The outcome variable is survival time, and there are eight covariates, one of which has missing values. The main covariate of interest is the effect of the new chemotherapy versus the standard therapy. Besides treatment, the other seven covariates of interest are: bone fractures at diagnosis (yes/no), logarithm of the blood urea nitrogen (LOGBUN), hemoglobin (HGB), platelet count (PLATELET), logarithm of the white blood cell count (LOGWBC), the logarithm of plasma cells in bone marrow (LOGPBM), and serum calcium (SCALC). Patients with high values of LOGBUN, HGB, LOGPBM, and SCALC, and

*To whom correspondence should be addressed.

low values of PLATELET and LOGWBC are expected to have shorter survival. Even though survival as an outcome typically calls for time-to-event type methods, we use Poisson regression to estimate the regression parameters for an exponentially distributed survival time.

The only covariate with missing values is $z$ = bone fractures at diagnosis, which is missing for 84 (37.5%) of the 224 cases. This covariate is a very important predictor since multiple myeloma is a haematologic cancer which attacks the bone marrow, so a patient with bone fractures at diagnosis may have more advanced cancer, and thus shorter survival. With such a large fraction of missing data, a complete case analysis using only the 140 subjects with no missing data could give highly inefficient and/or biased estimates. Nevertheless, such a complete case analysis is still one of the most commonly encountered modes of analysis in practice.

A recent approach that can reduce the bias of the complete case estimate is weighted estimating equations (WEE) (Robins *et al.*, 1994; Zhao *et al.*, 1996). With WEE, the contribution to the regression estimating equation from a complete observation $(y, x, z)$ is weighted by the inverse of the response probability, i.e. the probability that $z$ was observed. It has been shown that WEE are applicable to regression analysis when missing covariates are missing at random (MAR); (Robins *et al.*, 1994), i.e. when the probability that $z$ is missing depends on $(y, x)$ but not $z$. To use WEE, one must pose and estimate a binary regression model for the probability of $z$ being observed as a function of $(y, x)$. A logistic, probit, or complementary log–log binary regression model can be used, or the parametric assumptions can be lessened by using generalized additive models (Hastie and Tibshirani, 1990). Of course, the quality of the inference will depend on the correctness of the posited model of the probability of being observed.

WEE are sometimes preferred to maximum likelihood since one does not need to specify the full joint distribution of $(y, x, z)$, which could be complicated when $(y, x, z)$ are mixed discrete and non-normal continuous. If the model for the probability of $z$ being observed is correctly specified, then the WEE will produce a consistent estimate. However, even if the missing data model is mis-specified, the WEE estimate could have little or no bias. For example, if missingness depends on all of the covariates $x$ and $z$, but not $y$, and the (wrongly) posed missing data model contains any subset of $x$ and $z$, one can show that the WEE estimate will be consistent for $\beta$. Thus, although one can use goodness-of-fit statistics for the fit of the binary regression model for missingness, one would still want a direct test of whether the WEE is asymptotically unbiased. We propose a test statistic that gives the investigator an idea if the WEE produce biased regression parameter estimates. This statistic is similar to a statistic proposed by DuMouchel and Duncan (1983) for testing if unweighted or weighted least squares should be used in sample surveys with unequal probability of being selected into the sample, where the weights are the inverse probability of selection.

Our proposed test statistic is as follows. First, we estimate the probability of $z$ being observed given $(y, x)$, and then use WEE to obtain $\widehat{\beta}_{wee}$, in which only the complete cases are inversely weighted by the estimates probabilities of $z$ being observed. Although unrealistic, suppose we actually knew (say, from a follow-up sample) the missing $z$ values, and thus could obtain a consistent estimate of $\beta$, say $\widehat{\beta}$, using all of the data. To check for asymptotic unbiasedness of $\widehat{\beta}_{wee}$, we could use a quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$. This quadratic form will be asymptotically chi-square with $(K + 1)$ degrees of freedom under the null that $\widehat{\beta}_{wee}$ is asymptotically unbiased for $\beta$, or, equivalently, under the null that the 'WEE and the full data estimates converge in probability to the same parameter'. The quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$ is the test statistic that we would like to use. Unfortunately, we do not know the missing $z$ so we propose using the quadratic form $(\widehat{\beta}_{wee} - \widehat{\beta})$ under the constraint that the regression coefficient of $z$ equals 0. To obtain this new quadratic form, we fit the regression model of $y$ given $x$ (without $z$) using WEE, as well as all data. In this case, even though the estimated regression coefficients of $x$ are biased for the regression coefficients of $x$ for the model $E(y|x, z)$, the estimate using all of the data will converge to some fixed vector, say $\gamma$, and the weighted estimating equations will converge to the vector $\gamma_w$. As was the case for $\widehat{\beta}_{wee}$, if the model for the probability of $z$ being observed is correctly specified or, if missingness only depends on some function of $(x, z)$ and the probability of $z$ being observed is modelled as the wrong function of $(x, z)$, then WEE and the full data estimates converge in probability to the same parameter, i.e. $\gamma_w = \gamma$. In particular, our proposed test statistic, which is a quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$ under the constraint that the regression coefficient of $z$ equals 0, will be asymptotically chi-square with $K$ degrees of freedom under the null that the 'WEE and full data estimates converge in probability to the same parameter vector'. Note that this null hypothesis is more broad than a null that 'the probability of $z$ being observed is modelled correctly', because, as we have discussed above, we could reject the null 'the probability of $z$ being observed is modelled correctly', but still could have the null 'the WEE and full data estimates converge in probability to the same parameter vector' hold.

Section 2 introduces necessary notation and describes the weighted estimating equations. Section 3 describes the test statistic. Section 4 illustrates the methods with the example and, in Section 5, we give the results of simulations comparing the power of our proposed test statistic to the preferred, but unavailable statistic, which is the quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$.

## 2. Notation and model

Consider a regression problem involving $n$ independent subjects, $i = 1, \ldots, n$. The data collected on the $i$th subject are the outcome variable $y_i$, a vector $x_i = (x_{i1}, \ldots, x_{iK})^T$ of $K$ covariates that are always observed, and a covariate $z_i$ that is missing for some subjects. Since $z_i$ can be missing, we also define the indicator random variable $R_i$, which equals 1 if $z_i$ is observed and 0 if $z_i$ is missing. The distribution of $R_i$ given $(y_i, x_i, z_i)$, is Bernoulli with probability

$$\pi_i = \Pr(R_i = 1 | y_i, x_i, z_i),$$

and is referred to as the missing data mechanism. If missingness is non-ignorable, then $\pi_i$ depends on $z_i$. In this paper, we restrict attention to missing data mechanisms that are MAR, i.e. in which $\pi_i$ does not depend on $z_i$, but solely on $(y_i, x_i)$. If $\pi_i$ depends on none of $y_i$, $x_i$, and $z_i$, then the data are missing completely at random.

Let $\mu_i = E(y_i | x_i, z_i)$ denote the expectation of the outcome $y_i$ given the covariates. In most regression problems, it is of interest to estimate the regression coefficients $\beta' = (\beta_0, \beta_x', \beta_z)$, where $\beta_0$ and $\beta_z$ are scalars and $\beta_x$ is a vector, from the regression model

$$\mu_i = g(\beta_0 + x_i'\beta_x + z_i\beta_z), \tag{1}$$

where $g(\cdot)$ is a specified function, such as $g(a) = a$ for linear regression and $g(a) = \exp(a)/[1 + \exp(a)]$ for logistic regression.

With no missing covariate data, one can use a quasi-likelihood (McCullagh and Nelder, 1989) estimator of $\beta$, which is the solution to $u(\widehat{\beta}) = 0$, where

$$u(\beta) = \sum_{i=1}^{n} u_i(\beta) = \sum_{i=1}^{n} d_i v_i^{-1}(y_i - \mu_i). \tag{2}$$

Here, $d_i = \frac{\partial \mu_i}{\partial \beta}$, and $v_i = v_i(\beta) = \text{Var}(Y_i | x_i, z_i)$. Since $u_i(\beta)$ is a linear function of $(y_i - \mu_i)$, it has expection 0. Since $E[u(\beta)] = 0$, and we are solving $u(\widehat{\beta}) = 0$, using method of moment ideas, the estimate of $\beta$ from the quasi-likelihood converges to $\beta$, establishing consistency. Moreover, using a first-order Taylor series expansion,

$$(\widehat{\beta} - \beta) \approx \left[\sum_{i=1}^{n} d_i v_i^{-1} d_i'\right]^{-1} \sum_{i=1}^{n} d_i v_i^{-1}(y_i - \mu_i). \tag{3}$$

By application of the central limit theorem, the estimate of $\beta$ has an asymptotic normal distribution with mean $\beta$ and covariance matrix

$$\left[\sum_{i=1}^{n} d_i v_i^{-1} d_i'\right]^{-1}\left[\sum_{i=1}^{n} E[u_i(\beta)u_i(\beta)']\right]\left[\sum_{i=1}^{n} d_i v_i^{-1} d_i'\right]^{-1}, \tag{4}$$

which can be consistently estimated by

$$\left[\sum_{i=1}^{n} \widehat{d}_i \widehat{v}_i^{-1} \widehat{d}_i'\right]^{-1}\left[\sum_{i=1}^{n} \widehat{d}_i \widehat{d}_i' \widehat{v}_i^{-2}(y_i - \widehat{\mu}_i)^2\right]\left[\sum_{i=1}^{n} \widehat{d}_i \widehat{v}_i^{-1} \widehat{d}_i'\right]^{-1}, \tag{5}$$

where all quantities in (5) are evaluated at $\widehat{\beta}$. The solution to the estimating equation (2) cannot be solved in closed form. Fisher's method of scoring can be used to solve these nonlinear equations numerically.

With missing data, the most popular method of estimation is the complete case estimate, $\widehat{\beta}_{cc}$, which is the solution to the estimating equation $u_{cc}(\widehat{\beta}_{cc}) = 0$, where

$$u_{cc}(\beta) = \sum_{i=1}^{n} r_i d_i v_i^{-1}(y_i - \mu_i) = 0, \tag{6}$$

where $r_i$ is the realized value of $R_i$ and hence only complete cases contribute.

### 2.1 *Weighted estimating equations*

Suppose now that $\pi_i$, the probability of $z_i$ being observed, depends on the observed outcome $y_i$ and the covariates $x_i$, and that the dependence is specified up to a known probability function indexed by a finite number of unknown parameters. Specifically, we consider a logistic regression for the probability of being observed,

$$\pi_i = \pi_i(\alpha) = \frac{\exp(\alpha' m_i)}{1 + \exp(\alpha' m_i)}, \tag{7}$$

where $\alpha$ is a vector of unknown parameters and $m_i$ is a function of $(y_i, x_i')'$. We could have $m_i = (y_i, x_i')'$, but $m_i$ could also include interactions among the elements of $(y_i, x_i')'$ whilst preserving the MAR nature of the mechanism.

In order to obtain a consistent estimate of the regression parameters under a MAR mechanism, we can use the WEE proposed by Robins *et al.* (1994) and Zhao *et al.* (1996). In the WEE, we replace $r_i$ in the complete case estimating equation (6) with $r_i/\pi_i$. In particular, the WEE are $u_{wee}(\widehat{\beta}_{wee}) = 0$, where

$$u_{wee}(\beta) = \sum_{i=1}^{n} \frac{r_i}{\pi_i} u_i(\beta) = \sum_{i=1}^{n} \frac{r_i}{\pi_i} d_i v_i^{-1}(y_i - \mu_i). \tag{8}$$

We provide a brief sketch of the argument for why $\widehat{\beta}_{wee}$ is consistent. The estimating equation given by (8) is unbiased for 0 at the true $\beta$ if $\pi_i$ is correctly specified because

$$\begin{aligned} E\left[\left(\frac{R_i}{\pi_i}\right) d_i v_i^{-1}(y_i - \mu_i)\right] &= E_{x_i,z_i}\left(E_{y_i|x_i,z_i}\left\{d_i v_i^{-1}(y_i - \mu_i)\left[E_{R_i|y_i,x_i,z_i}\left(\frac{R_i}{\pi_i}\right)\right]\right\}\right) \\ &= E_{x_i,z_i}(E_{y_i|x_i,z_i}\{d_i v_i^{-1}(y_i - \mu_i)\}) \\ &= E_{x_i,z_i}(0) = 0. \end{aligned} \tag{9}$$

As before, since the estimating equation is unbiased for 0, $\widehat{\beta}_{wee}$ defines a consistent estimator for $\beta$. If $\pi_i$ is either known or is consistently estimated, a consistent estimate of $\beta$ is obtained. Under a MAR mechanism, $\pi_i$ can be estimated independently from the model for $\mu_i$, and hence there is no need to fully specify the joint distribution of $(y_i, x_i, z_i, r_i)$. In most applications, $\pi_i(\alpha)$ is unknown and needs to be estimated and substituted in (8). We estimate $\pi_i$ using ordinary logistic regression with outcome $R_i$ and covariates $m_i$ given in (7). The ordinary logistic regression estimating equations for $\alpha$ are given by $u_\alpha(\widehat{\alpha}) = 0$, where

$$u_\alpha(\alpha) = \sum_{i=1}^{n} u_{\alpha i}(\alpha) = \sum_{i=1}^{n} m_i'[r_i - \pi_i(\alpha)]. \tag{10}$$

We can put the estimating equations for $(\beta, \alpha)$ together to get

$$S_{wee}(\beta, \alpha) = \sum_{i=1}^{n} \begin{bmatrix} u_{wee,i}(\beta) \\ u_{\alpha,i}(\alpha) \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} (r_i/\pi_i) d_i v_i^{-1}(y_i - \mu_i) \\ m_i'[r_i - \pi_i(\alpha)] \end{bmatrix}. \tag{11}$$

If we have the correct models for $\mu_i$ and $\pi_i$ (Zhao *et al.*, 1996), then the estimates of both $\beta$ and $\alpha$ are consistent. Using a Taylor series expansion, $(\widehat{\beta}_{wee}, \widehat{\alpha})$ also have an asymptotic normal distribution with covariance matrix that is consistently estimated by

$$\widehat{\text{Var}} \begin{bmatrix} \widehat{\beta}_{wee} \\ \widehat{\alpha} \end{bmatrix} = \left[ \sum_{i=1}^{n} \widehat{\text{A}}_i \right]^{-1} \left[ \sum_{i=1}^{n} \widehat{\text{B}}_i \right] \left[ \sum_{i=1}^{n} \widehat{\text{A}}_i' \right]^{-1}, \tag{12}$$

where

$$A_i = \begin{bmatrix} (R_i/\pi_i) v_i^{-1} d_i d_i' & (R_i/\pi_i)(1 - \pi_i) v_i^{-1}(y_i - \mu_i) d_i m_i' \\ 0 & \pi_i(1 - \pi_i) m_i m_i' \end{bmatrix} \tag{13}$$

and

$$B_i = \begin{bmatrix} (R_i/\pi_i^{-2})[v_i^{-1}(y_i - \mu_i)]^2 d_i d_i' & (R_i/\pi_i) w_i^{-1}(y_i - \mu_i)(R_i - \pi_i) d_i m_i' \\ (R_i/\pi_i) v_i^{-1}(y_i - \mu_i)(R_i - \pi_i) m_i d_i' & (R_i - \pi_i)^2 m_i m_i', \end{bmatrix}, \tag{14}$$

where 0 is a matrix of zeros with appropriate dimensions, and (12) is evaluated at $(\widehat{\beta}_{wee}, \widehat{\alpha})$.

If the model for $\pi_i$ is correctly specified, then $\widehat{\beta}_{wee}$ is consistent for $\beta$. If the model for $\pi_i$ is underspecified, the estimate of the regression coefficients of interest, $\beta$, could be biased. Note, however, that $\widehat{\beta}_{wee}$ will sometimes be consistent even if $\pi_i$ is mis-modelled. The weighted estimate will be unbiased when $\pi_i$ truly depends on a function of the covariates $(x_i, z_i)$, say $\pi_i(x_i, z_i)$, but one models $\pi_i$ as the wrong function of $(x_i, z_i)$, say $\pi_i^*(x_i, z_i)$. Thus, even if the missing data are non-ignorably missing ($\pi_i$ depends on $z_i$), but missingness does not depend on $y_i$, we can mis-model $\pi_i$, and a still get consistent estimates using WEE. The following equation shows that the WEE with the wrong 'weights' $\pi_i^*(x_i, z_i)$ is still unbiased for 0,

$$\begin{aligned} E\left[\left(\frac{R_i}{\pi_i^*(x_i, z_i)}\right) d_i v_i^{-1}(y_i - \mu_i)\right] &= E_{x_i,z_i}\left[E_{y_i|x_i,z_i}\left(d_i v_i^{-1}(y_i - \mu_i)\left[E_{R_i|y_i,x_i,z_i}\left(\frac{R_i}{\pi_i^*(x_i, z_i)}\right)\right]\right)\right] \\ &= E_{x_i,z_i}\left[E_{y_i|x_i,z_i}\left(d_i v_i^{-1}(y_i - \mu_i)\left[\frac{\pi_i(x_i, z_i)}{\pi_i^*(x_i, z_i)}\right]\right)\right] \\ &= E_{x_i,z_i}\left[\left(\frac{\pi_i(x_i, z_i)}{\pi_i^*(x_i, z_i)}\right) E_{y_i|x_i,z_i}(d_i v_i^{-1}(y_i - \mu_i))\right] \\ &= E_{x_i,z_i}\left[\left(\frac{\pi_i(x_i, z_i)}{\pi_i^*(x_i, z_i)}\right) 0\right] = 0. \end{aligned} \tag{15}$$

As before, since the estimating equation is unbiased for 0, $\widehat{\beta}_{wee}$ defines a consistent estimator for $\beta$. Note that, if we set $\pi_i^*(x_i, z_i) = 1$ in (15), we get the complete case estimate described in (6). Thus, (15) can be used to show that the complete case estimate is asymptotically unbiased as long as the true $\pi_i$ is a function of $(x_i, z_i)$ but not $y_i$.

Even if $\pi_i$ depends on both $x_i$ and $y_i$, but we mis-model $\pi_i$, we still may get small bias in the weighted estimate. However, in order to ensure minimal bias, one should go about finding the best fit for $\pi_i$ as one usually does for logistic regression. Tests for interactions can be performed, and stepwise logistic regression can be used to obtain a model for $\pi_i$. Thus, although one can use goodness-of-fit statistics to obtain the best fit of the binary regression model for missingness, one would still want a direct test of whether the WEE is asymptotically unbiased. Further, our null of interest is that WEE produces asymptotically unbiased estimates, and one can use the test statistics proposed in the following section to assess this. Note that this null hypothesis is broader than a null that $\pi_i$ is correctly specified because, as we have discussed above, if missingness only depends on $(x_i, z_i)$, and we mis-model $\pi_i$, we could reject the null 'that $\pi_i$ is correctly specified', but still have the null 'WEE produces asymptotically unbiased estimates' hold.

## 3. TEST STATISTIC TO ASSESS IF WEE IS ASYMPTOTICALLY UNBIASED

Suppose $\widehat{\beta}_{wee}$ converges in probability to the parameter vector $\beta_w$. If the model for $\pi_i$ is correctly specified, then $\beta_w = \beta$, or, if missingness only depends on some function of $(x_i, z_i)$ and $\pi_i$ is modelled as the wrong function of $x_i$, we still have $\beta_w = \beta$. Although unrealistic, if we knew the true value of $\beta$, then we could test the null hypothesis that $\beta_w = \beta$ using a quadratic form in $(\widehat{\beta}_{wee} - \beta)$. Although again unrealistic, suppose we actually knew (say, from a follow-up sample) the values of the missing $z_i$, and thus could obtain a consistent estimate of $\beta$, say $\widehat{\beta}$, using all of the data. In this case, we could test the null hypothesis that $\beta_w = \beta$ using a quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$, which will be asymptotically chi-square with $(K + 1)$ degrees of freedom under the null that $\widehat{\beta}_{wee}$ is asymptotically unbiased for $\beta$. This quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$ is the test statistic that we would like to use, and is similar in spirit to the test statistic proposed by DuMouchel and Duncan (1983) for sample survey data. However, since we do not know the missing $z_i$, we cannot use it. Instead, we attempt to form a test statistic as 'close' as possible to a statistic based on $(\widehat{\beta}_{wee} - \widehat{\beta})$: our statistic is a quadratic form of $(\widehat{\beta}_{wee} - \widehat{\beta})$, under the restriction that $\beta_z = 0$. Our general null hypothesis is that the 'WEE estimates and full data estimates converge in probability to the same parameter'.

Thus, suppose we set $\beta_z = 0$, which is equivalent to dropping $z_i$ from the regression model, and we fit the regression model of $y_i$ given $x_i$ (without $z_i$) using both the WEE, and all of the data. Even though the estimated regression coefficients of $x_i$ when setting $\beta_z = 0$ are biased for the regression coefficients $\beta_x$ of $x_i$ for the model $E(y_i|x_i, z_i)$, the estimate of $\beta_x$ using all of the data will converge to some fixed vector, say $\gamma$, and the WEE will converge to the vector $\gamma_w$. Under the null 'the WEE and full data estimates converge in probability to the same parameter', $H_o : \gamma_w = \gamma$. Thus, our statistic, a quadratic form of $(\widehat{\beta}_{wee} - \widehat{\beta})$ under the restriction that $\beta_z = 0$, will be asymptotically chi-square with $K$ degrees of freedom under the null that the 'WEE and full data estimates converge in probability to the same parameter'.

In formally explaining our test statistic, since $y_i$ and $x_i$ are observed for all individuals, we can use the full data and fit the model

$$E(Y_i|x_i, \gamma) = \mu_{i*}(\gamma) = g(\gamma_0 + \gamma_1 x_i). \tag{16}$$

Note that, assuming that (1) is the correct model for $E(Y_i|x_i, z_i)$, it is very unlikely that (16) is the true model for $E(Y_i|x_i)$. This is because it is very unlikely that $E(Y_i|x_i)$ and $E(Y_i|x_i, z_i)$ will both have the same link function $g(\cdot)$, except in the linear case, when $x$ and $z$ are orthogonal. To estimate $\gamma = [\gamma_0, \gamma_1']'$, one can use the full data estimating equation (2), in which we replace $\mu_i$ with $\mu_{i*}$, $d_i$ with $d_{i*} = \frac{\mathrm{d}\mu_{i*}}{\mathrm{d}\gamma}$ and $v_i$ with $v_{i*} = v_i(\gamma) = \mathrm{Var}(Y_i|x_i, \gamma)$. The resulting estimate $\widehat{\gamma}$ has similar asymptotic properties as the estimate in Section 2, i.e. $\widehat{\gamma}$ is asymptotically normal with mean $\gamma$ and asymptotic variance given by (4). Also, $\widehat{\gamma}$ is consistent for $\gamma$. Following White (1982), $\gamma$ is the quantity that minimizes the Kullback–Leibler distance between the true model and the chosen parametric family. For the purposes of our proposed methods, it is not important that (16) is not the correct model.

Next, suppose that we use the WEE to fit model (16). In particular, suppose we solve $S_{wee}(\widehat{\gamma}_{wee}, \widehat{\alpha}) = 0$, where

$$S_{wee}(\gamma, \alpha) = \sum_{i=1}^{n} \begin{bmatrix} u_i(\gamma) \\ u_i(\alpha) \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} (r_i/\pi_i) d_{i*} v_{i*}^{-1} (y_i - \mu_{i*}) \\ m_i[r_i - \pi_i(\alpha)] \end{bmatrix}, \tag{17}$$

where $d_{i*} = \frac{\mathrm{d}\mu_{i*}}{\mathrm{d}\gamma}$, and $v_{i*} = v_{i*}(\gamma) = \mathrm{Var}(Y_i|x_i)$.

Using results similar to (9), one can show that, if the model for $\pi_i$ is correctly specified, then the solution to (17), $\widehat{\gamma}_{wee}$, is consistent for $\gamma$, the same quantity as the full data estimator $\widehat{\gamma}$ converges to. In particular, under the null 'WEE estimates and all data estimates converge in probability to the same parameter', then $(\widehat{\gamma}_{wee} - \widehat{\gamma})$ is asymptotically normal with mean 0, and covariance matrix equal to

$$\mathrm{Var}(\widehat{\gamma}_{wee} - \widehat{\gamma}) = \mathrm{Var}(\widehat{\gamma}_{wee}) + \mathrm{Var}(\widehat{\gamma}) - 2\mathrm{Cov}(\widehat{\gamma}_{wee}, \widehat{\gamma}).$$

We can consistently estimate $\mathrm{Var}(\widehat{\gamma})$ with (5) and $\mathrm{Var}(\widehat{\gamma}_{wee})$ with (12), and using a Taylor series expansion, one can show that $\mathrm{Cov}(\widehat{\gamma}_{wee}, \widehat{\gamma})$ can be consistently estimated with the upper $(p + 1) \times (p + 1)$ block of

$$\left[\sum_{i=1}^{n} \widehat{\mathrm{A}}_i\right]^{-1} \left[\sum_{i=1}^{n} \widehat{\mathrm{C}}_i\right] \left[\sum_{i=1}^{n} \widehat{\mathrm{A}}_{i2}'\right]^{-1}, \tag{18}$$

where $\widehat{\mathrm{A}}_{i2} = \widehat{d}_{i*} \widehat{v}_{i*}^{-1} \widehat{d}_{i*}'$,

$$A_i = \begin{bmatrix} (R_i/\widehat{\pi}_i) \widetilde{v}_{i*}^{-1} \widetilde{d}_{i*} \widetilde{d}_{i*}' & (R_i/\widehat{\pi}_i)(1 - \widetilde{\pi}_i) \widetilde{v}_{i*}^{-1} (y_i - \widetilde{\mu}_{i*}) \widetilde{d}_{i*} m_i' \\ 0 & \widetilde{\pi}_i (1 - \widetilde{\pi}_i) m_i m_i' \end{bmatrix},$$

$$C_i = \left[(R_i/\widetilde{\pi}_i^{-2}) \widehat{v}_{i*}^{-1} (y_i - \widehat{\mu}_{i*})(R_i - \widetilde{\pi}_i) m_i \widehat{d}_{i*}', \quad (R_i/\widetilde{\pi}_i^{-2}) \widehat{v}_{i*}^{-1} (y_i - \widehat{\mu}_{i*})(y_i - \widetilde{\mu}_{i*}) \widetilde{v}_{i*}^{-1} \widehat{d}_{i*} \widetilde{d}_{i*}'\right],$$

and $\widehat{d}_i = d_i(\widehat{\gamma})$, $\widehat{v}_i = v_i(\widehat{\gamma})$, $\widetilde{d}_i = d_i(\widehat{\gamma}_{wee})$, $\widetilde{v}_i = v_i(\widehat{\gamma}_{wee})$.

To test if $H_o : \gamma_w = \gamma$, we propose the global Wald statistic,

$$G = (\widehat{\gamma}_{wee} - \widehat{\gamma})' [\widehat{\mathrm{Var}}(\widehat{\gamma}_{wee} - \widehat{\gamma})]^{-1} (\widehat{\gamma}_{wee} - \widehat{\gamma}), \tag{19}$$

which is approximately chi-square with $(p + 1)$ degrees of freedom (the dimension of $\gamma$) under the null.

## 4. ANALYSIS OF MULTIPLE MYELOMA DATA

To illustrate our proposed approach we consider data on a subset of $n = 224$ patients from an Eastern Cooperative Oncology Group clinical trial, E2479 (Kalish, 1992). The main purpose of the trial was to evaluate whether Vincristine, BCNU, Melphalan, Cyclophosphamide, and Prednosone (VBMCP) should replace Melphalan plus Prednosone (MP) as a standard therapy for patients with previously untreated multiple myeloma. We are primarily interested in how treatment affects survival time, the time of entry into the study until death; we are also interested in how survival is predicted by seven other baseline characteristics. The other seven covariates have been described in Section 1. Patients with high values of LOGBUN, HGB, LOGPBM and SCALC, and low values of PLATELET and LOGWBC, are expected to have shorter survival. The covariate $z$ = bone fractures at diagnosis is missing for 84 (37.5%) of the 224 patients; all other covariates are completely observed. To use WEE, this FRAC covariate should be MAR. FRAC is determined by x-ray, and sicker patients at diagnosis were not always well enough to have an x-ray taken. Since multiple myeloma is a haematologic cancer which attacks the bone marrow, a patient with bone fractures at diagnosis may have more advanced cancer, and thus shorter survival. Thus, we would expect $\text{pr}(R_i = 1|z_i)$ to depend on $z_i$. However, conditional on the censoring time, and the covariates $x_i$, we expect the probability of being observed to be conditionally independent of the FRAC covariate.

For illustration, we assume the survival time is exponentially distributed. Let $T_i$ be the true failure time for subject $i$. We note that 10 (2.4%) of the 224 cases have their survival time censored. If we let $U_i$ be the censoring time, then we observe $X_{i0} = \min(T_i, U_i)$ and the censoring indicator $Y_i = I\{T_i \leqslant U_i\}$. Under non-informative censoring, the density of $(y_i, x_{i0}|x_i, z_i)$ is

$$p(y_i, x_{i0}|x_i, z_i, \beta) \propto \mathrm{e}^{-\lambda_i x_{i0}} \lambda_i^{y_i}, \tag{20}$$

where

$$\lambda_i = \exp[\beta_0 + \beta_1' x_i + \beta_2' z_i] \tag{21}$$

is the hazard and $x_i$ contains LOGBUN, HGB, PLATELET, LOGWBC, LOGPBM and SCAL. We note that the complete case estimator can be obtained as the solution to (2), by treating the censoring indicator $Y_i$ as a Poisson outcome random variable with mean

$$\mu_i = x_{i0}\lambda_i = \exp[\log(x_{i0}) + \beta_0 + \beta_1' x_i + \beta_2' z_i],$$

where $\log(x_{i0})$ is an offset.

To use the WEE, we must specify the logistic model for $\pi_i = \text{pr}(R_i = 1|y_i, x_i)$, ensuring that it is well fitting. Since a good fit for $\pi_i$ is indeed important, we do not worry about a high type I error rate; we would rather over-specify than under-specify, since under-specification could bias the estimates of $\beta$. Thus, one should fit the largest possible model for $\pi_i$ that appears reasonable (say, for a given model, one could use the jackknife, making sure that there are no 'outlying estimates' when dropping an observation). In this dataset, very few of the models with three-way interactions converged to a solution (some of the parameter estimates were converging to $\pm\infty$ using the Newton–Raphson algorithm). We suggest keeping any parameter in the model for $\pi_i$ that is significant at the 0.35 level. We can find the best-fitting model for $\pi_i$ using ordinary logistic regression, without having to worry about the specification of the model for $[y_i \mid x_i, z_i, \beta]$. We considered the main effects model in $x_i$ and log censoring time as the baseline model for $\pi_i$, and used a step-up logistic regression approach from the main effects. The best-fitting model for $\pi_i$ contained pairwise interactions between log(survival) and SCALC, LOGBUN and HGB, and HGB and SCALC. The results are shown in Table 1. The Hosmer–Lemeshow goodness-of-fit statistic has a value

Table 1. *Maximum likelihood estimates of the missingness model* $\mathrm{pr}(R_i = 1|x_i, y_i)$ *for myeloma data*

| Parameter | Estimate | Standard error | $Z$-value | $p$-value |
|---|---|---|---|---|
| INTERCEPT | 2.615 | 7.050 | 0.37 | 0.711 |
| LOGSURV | −1.776 | 0.657 | −2.70 | 0.007 |
| LOGBUN | −3.772 | 2.518 | −1.50 | 0.134 |
| HGB | 0.260 | 0.677 | 0.38 | 0.701 |
| PLATELET | −0.550 | 0.680 | −0.81 | 0.417 |
| LOGWBC | 0.295 | 0.318 | 0.93 | 0.353 |
| LOGPBM | −0.089 | 0.188 | −0.48 | 0.634 |
| SCALC | 0.651 | 0.460 | 1.41 | 0.158 |
| TRT | −0.107 | 0.252 | −0.42 | 0.673 |
| LOGSURV*SCALC | 0.165 | 0.064 | 2.59 | 0.010 |
| LOGBUN*HGB | 0.367 | 0.259 | 1.42 | 0.156 |
| HGB*SCALC | −0.103 | 0.047 | −2.17 | 0.030 |

of 6.753, which, when compared to a chi-square with eight degrees of freedom has a $p$-value of 0.5635. Thus, the Hosmer–Lemeshow statistic indicates a good fit for $\pi_i$.

Table 2 gives estimates and standard errors for the regression coefficients based on WEE (with $\widehat{\pi}_i$ from Table 1), and CC (complete case) estimation. We see that the only appreciable difference between WEE and CC is in the estimate of the SCALC effect, which is about 50% greater for WEE than CC, and is also significant for WEE and not CC. Now, we use our method fitting the Poisson regression model

$$\mu_{i*} = x_{i0}\lambda_{i*} = \exp[\log(x_{i0}) + \gamma_0 + \gamma_1' x_i],$$

using WEE, CC and the full data. Here we note that our method does apply, not only to WEE, but also to CC, merely by considering the CC estimates as a special form of WEE with $\pi_i = \pi$. For the WEE, our test has a value of 4.970, which, when compared to a chi-square with eight degrees of freedom has a $p$-value of 0.7608. For the CC, our test has a value of 7.907, which, when compared to a chi-square with eight degrees of freedom has a $p$-value of 0.4426. Although they are global test statistics, the statistics indicate that the WEE and CC estimates are not significantly different than the full data. Thus, we feel comfortable that $\widehat{\beta}_{wee}$ is not asymptotically biased. In the following section, we perform simulations comparing the power of our proposed test statistic to the preferred, but unavailable, test statistic, which is a quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$.

## 5. A SIMULATION STUDY

We performed a simulation study based on the myeloma example discussed in Section 4, with survival time as response, and covariates TRT ($x_{i1}$), SCALC ($x_{i2}$) and FRAC ($z_i$). We are mainly interested in how the power of the quadratic form in $(\widehat{\gamma}_{wee} - \widehat{\gamma})$ (given in (19)) compares to the similar quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$, which is unavailable for a given dataset with $z_i$ missing, but easily obtained for a simulation in which we know the value of the missing $z_i$.

We formulate the true model from which we simulate by specifying each term on the right side of

$$\begin{aligned} &p(r_i, z_i, y_i, x_{i1}, x_{i2}, \alpha, \beta, \phi) \\ &\quad = p(r_i|y_i, x_i, z_i, \alpha)p(y_i|x_{i1}, x_{i2}, z_i, \beta)p(z_i|x_{i1}, x_{i2}, \phi_1)p(x_{i2}|x_{i1}, \phi_2)p(x_{i1}|\phi_3). \end{aligned} \tag{22}$$

Table 2. *Regression parameter* ($\beta$) *estimates for the myeloma data*

| Effect | Approach | $\hat{\beta}$ | SE | $Z$-statistic | $p$-value |
|---|---|---|---|---|---|
| INTERCEPT | WEE | −6.352 | 1.007 | −6.31 | 0.000 |
| | CC | −5.952 | 0.947 | −6.28 | 0.000 |
| LOGBUN | WEE | 0.318 | 0.260 | 1.22 | 0.222 |
| | CC | 0.273 | 0.232 | 1.18 | 0.238 |
| HGB | WEE | −0.027 | 0.025 | −1.10 | 0.271 |
| | CC | −0.036 | 0.026 | −1.40 | 0.160 |
| PLATELET | WEE | −1.457 | 0.762 | −1.91 | 0.056 |
| | CC | −1.400 | 0.600 | −2.33 | 0.024 |
| LOGWBC | WEE | 0.198 | 0.162 | 1.22 | 0.222 |
| | CC | 0.208 | 0.156 | 1.33 | 0.183 |
| LOGPBM | WEE | 0.285 | 0.086 | 3.32 | 0.001 |
| | CC | 0.279 | 0.085 | 3.30 | 0.001 |
| SCALC | WEE | 0.074 | 0.035 | 2.13 | 0.033 |
| | CC | 0.050 | 0.043 | 1.16 | 0.244 |
| TREATMENT | WEE | −0.070 | 0.107 | −0.66 | 0.512 |
| | CC | −0.045 | 0.112 | −0.41 | 0.685 |
| FRAC | WEE | −0.035 | 0.110 | −0.32 | 0.753 |
| | CC | −0.024 | 0.116 | −0.21 | 0.834 |

In the covariate distributions, we let $X_{i1}$, be a Bernoulli random variable with $\text{pr}(X_i = 1) = 0.5$. We let $p(x_{i2}|x_{i1}, \phi_2)$ be a Bernoulli distribution with the logit of the probability of success equal to

$$\text{logit}[\text{pr}(X_{i2} = 1|x_{i1}, \phi_2)] = -0.2 + 0.9x_{i1}.$$

Next, we let $p(z_i|x_{i1}, x_{i2}, \phi_1)$ be a Bernoulli distribution with the logit of the probability of success equal to

$$\text{logit}[\text{pr}(Z_i = 1|x_{i1}, x_{i2}, \phi_1)] = 0.5 - 0.5x_{i1} - 0.5x_{i2} + 0.5x_{i1}x_{i2}. \tag{23}$$

In each simulation, the exponential model for the survival time ($T_i$) had hazard

$$\lambda_i = \exp[-3 - x_{i1} + x_{i2} + z_i], \tag{24}$$

i.e. $\beta = (\beta_0, \beta_1, \beta_2, \beta_z) = (-3.8, -1, 1, 1)'$. We then sampled $T_i$ accordingly, without censoring any values. In all simulations, the true model for $\text{pr}(R_i = 1|y_i, x_{i1}, x_{i2})$ was

$$\begin{aligned}&\text{logit}\{\text{pr}(R_i = 1|t_i, x_{i1}, x_{i2}, \alpha)\}\\ &\quad = 0.5 - 0.5\log(t_i) + 0.5x_{i1} + 0.5x_{i2} - \theta[\log(t_i)x_{i1} + \log(t_i)x_{i2} + x_{i1}x_{i2}],\end{aligned} \tag{25}$$

where we varied $\theta$ from 0 to 1.2 in 0.1 intervals. At each value of $\theta$, we performed 1000 simulation replications, where each replication had a sample size of $n = 1500$. In the simulations, when fitting the WEE, we specified the model for $\pi_i$ as in (25), except we set $\theta = 0$. We then studied the power of our

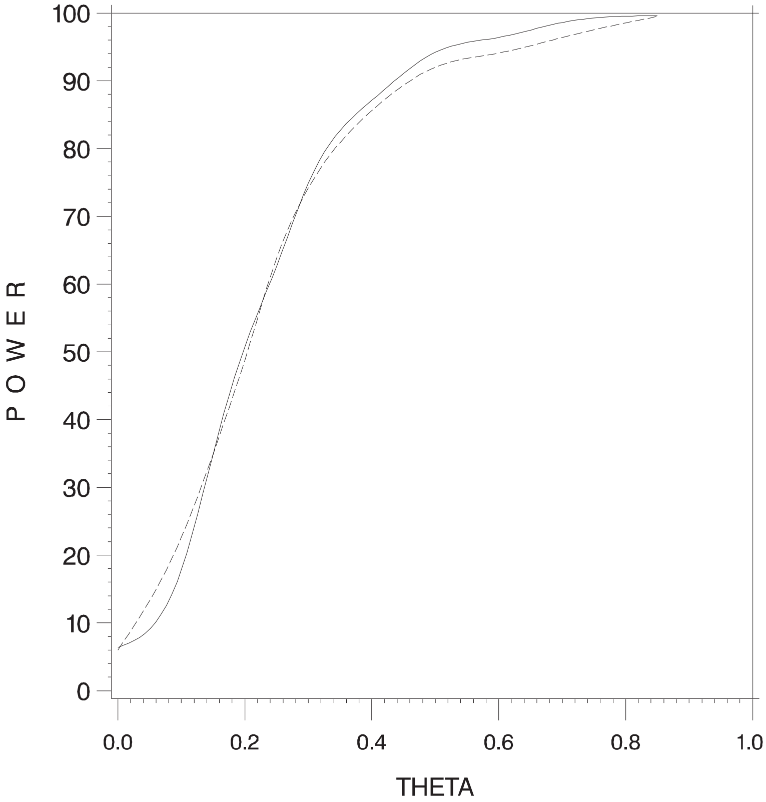

Fig. 1. Rejection probabability of our method (– – –: GAMMA) and the unavailable method (——: BETA) for testing $H_o : \beta_w = \beta$.

proposed test statistic as compared to the similar quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$. In our statistic, we fit the model dropping $z_i$, i.e.

$$\lambda_i = \exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2}),$$

and test if $\gamma = (\gamma_0, \gamma_1, \gamma_2)'$ estimated from all data, and $\gamma$ estimated from WEE are the same.

Figure 1 gives the results of the simulations. The line labelled ‘GAMMA’ is the proportion of two-sided $p$-values less than 0.05 when testing if $\gamma$ is equal using all data and WEE. The line labelled ‘BETA’ is the quadratic form for testing if $\beta$ is equal using all data and WEE. Both test statistics appear to have very similar power properties, and 95% confidence intervals for both power curves overlap. We have performed many simulations comparing the two statistics, including letting $Z_i$ be independent of the other covariates, letting $(x_{i1}, x_{i2}, z_i)$ be continuous, and setting the $\alpha$, $\beta$, and $\phi$ to many different values. In all of these simulations, the resulting power curves were very similar to Figure 1. These results are very encouraging for the use of our statistic. However, because of the broad range of possible data configurations and underlying probability distributions generating the data, it is difficult to draw definitive conclusions from simulations. We can only make general suggestions. In our simulation study, we have seen that our proposed test statistic performs very similarly to the preferred, but unavailable, statistic based on $(\widehat{\beta}_{wee} - \widehat{\beta})$.

## 6. DISCUSSION

We have developed a global goodness-of-fit test statistic to assess if WEE produces asymptotically unbiased estimates. The test statistic is relatively easy to calculate, and can be used in addition to stepwise regression and any other goodness-of-fit statistics for the missingness model. Ideally, one would like to compare the WEE estimate of $\beta$ to those obtained using the full dataset with no missing data. Unfortunately, the missing data makes this impossible. Thus, we feel that the comparison of the estimate of $\gamma$ (after setting $\beta_z = 0$ or, equivalently, dropping out $z$) using WEE and the full dataset is the next best thing. In our simulations, we have seen that quadratic forms based on $(\widehat{\beta}_{wee} - \widehat{\beta})$ or $(\widehat{\gamma}_{wee} - \widehat{\gamma})$ have very similar power properties; thus the quadratic form in $(\widehat{\gamma}_{wee} - \widehat{\gamma})$ appears to be a very good 'surrogate' for the quadratic form in $(\widehat{\beta}_{wee} - \widehat{\beta})$.

If one fits the fullest possible model for $\pi_i$ and still gets significant differences between $\widehat{\gamma}$ and $\widehat{\gamma}_{wee}$, then what should the researcher do? Here we briefly discuss the possibilities. One alternative is maximum likelihood, in which one must also specify the conditional distribution for $z_i$ given $x_i$, but not the conditional distribution for $r_i$ given $x_i$. One would then run into the problem of bias caused by the mis-specification of the distribution of $z_i$ given $(y_i, x_i)$. When using WEE, one has to specify the logistic regression model for $\pi_i$ correctly (except in the case where the true $\pi_i$ depends on $(x_i, z_i)$). When using maximum likelihood, we would have to correctly specify $p(z_i|x_i, \phi)$. Thus, besides needing to correctly specify the regression model for $y_i$, ML and WEE require us to correctly specify another distribution. As an even better extension, one could use the modified WEE of Sharfstein *et al.* (1999) in which the estimate of $\beta$ will be consistent if either $\pi_i$ or $p(z_i|x_i, \phi)$ is correctly specified.

## ACKNOWLEDGEMENTS

We are grateful for the support provided by grants CA 55576, CA 55670, CA 70101 and CA 74015. We gratefully acknowledge support from FWO-Vlaanderen (Belgium) Research Project 'Sensitivity Analysis for Incomplete and Coarse Data'.

## REFERENCES

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society,* Series B **39**, 1–38.

DUMOUCHEL, W. H. AND DUNCAN, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *JASA* **78**, 535–543.

HASTIE, T. J. AND TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

IBRAHIM, J. G., CHEN, M. AND LIPSITZ, S. R. (1999). Missing covariates in parametric regression models with ignorable missing data. *Biometrics,* to appear.

KALISH, L. A. (1992). Phase III multiple myeloma: evaluation of combination chemotherapy in previously untreated patients. *Technical Report # 726E*. Department of Biostatistics, Dana-Farber Cancer Institute.

LITTLE, R. J. A. AND RUBIN, D. B. (1987). *Statistical Analysis With Missing Data*. New York: Wiley.

MCCULLAGH, R. AND NELDER, J. A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

MOLENBERGHS, G., GOETGHEBEUR, E., LIPSITZ, S. R. AND KENWARD, M. G. (1999). Non-random missingness in categorical data: strengths and limitations. *The American Statistician* **53**, 110–118.

ROBINS, J. M., ROTNITZKY, A. AND ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *JASA* **89**, 846–866.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

SCHARFSTEIN, D. O., ROTNITZKY, A. AND ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *JASA* **94**, 1096–1120.

SHAH, B. V., BARNWELL, B. G. AND BIELER, G. S. (1996). *SUDAAN User's Manual: Release 7.0.* Research Triangle Park, NC: Research Triangle Institute.

WHITE, H. (1982). Maximum likelihood estimation under mis-specified models. *Econometrica* **50**, 1–26.

ZHAO, L. P., LIPSITZ, S. R. AND LEW, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165–1182.

[*Received November 15, 1999; revised May 30, 2000; accepted for publication August 10, 2000*]