INFERENCE FOR IMPUTATION ESTIMATORS

By JAMES M. ROBINS
Department of Epidemiology and Biostatistics, Harvard School of Public Health
677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
e-mail: robins@hsph.harvard.edu
and NAISYIN WANG
Department of Statistics, Texas A & M University
College Station, Texas 77843, U.S.A.
e-mail: nwang@picard.tamu.edu

November 22, 1999

Summary

We derive an estimator of the asymptotic variance of both single and multiple imputation estimators. We assume a parametric imputation model but allow for non- and semiparametric analysis models. Our variance estimator, in contrast to the estimator proposed by Rubin (1987), is consistent even when the imputation and analysis models are misspecified and incompatible with one another.

Some Key Words: Asymptotic Variance; Influence Function; Missing Data.

1. Introduction

In both observational and randomised studies, data are often missing either by chance or design. In recent years, the parametric multiple imputation method proposed by Rubin (1978, 1987) has become one of the most popular methods for handling missing data. His original goal was to impute m completed datasets for public usage in the context of public surveys in which a response rate of less than 60 percent for any variable was rare. Rubin (1987, 1996) envisaged the imputer as a trained statistician familiar with state-of-the-art missing data methods, knowledgeable about the reasons for non-response, and with possible access to additional confidential information, such as exact addresses and neighbourhood relationships. The user was conceptualised as a non-statistician who would only have access to standard complete data analysis models and software. As a result of its ease of implementation, Rubin's method is now also being used in settings where the fraction of missing data is large and the user and imputer are the same individual who chooses multiple imputation because of its convenience. Examples include the papers by Little & Yao (1996), Paik (1997), Taylor et al. (1990), Tu, Meng & Pagano (1993) and Clayton et al. (1998).

Unfortunately, Fay (1992, 1994, 1996), Meng (1994), Rubin (1996) and Clayton et al. (1998) have shown that, in certain settings, the variance estimator $\widehat{\Sigma}_{Rubin}$ proposed by Rubin will be inconsistent with upward bias, resulting in conservative confidence intervals whose expected length is longer, and occasionally much longer, than necessary. In § 4 we derive a general formula for the large sample bias of $\widehat{\Sigma}_{Rubin}$ which not only confirms the findings of the above authors but also indicates there are other scenarios under which $\widehat{\Sigma}_{Rubin}$ is downwardly biased, resulting in anti-conservative confidence intervals whose actual coverage rates are less than nominal.

The purpose of this paper is to provide a variance estimator which overcomes the deficiencies of Rubin's. The price we pay for the better performance of our variance estimator is a slight increase in computational complexity. However, with small modifications to existing complete data software, we show that this increased computational burden can be made invisible to the user.

Wang & Robins (1998) have recently proposed a variance estimator for imputation estimators under the assumption that the imputation and analysis model are the same correctly specified parametric model. This paper extends their results by allowing (i) for misspecification and incompatibility of the models and (ii) for non- or semiparametric analysis procedures. One final point is that the method we propose does not require multiple imputations. It works perfectly well if one decides to fill in the missing data with a single imputation, although this may not be the most efficient choice except when computational and data storage resources severely limit one's options.

2. Fully parametric probability model imputation

In this section, we suppose the imputer uses a fully parametric probability model. Let Y denote the complete data, which may not be not fully observed. Rather, we assume we observe n independent and identically distributed copies (Y_R^i, R^i) of (Y_R, R) where $Y_R = c_R(Y)$ is a known function, that is, a coarsening $c_R(\cdot)$ of Y depending on R, where R indexes which part of Y is observed. Missing data is a special case of coarsened data in which each univariate component Y_k of $Y = (Y_1, \ldots, Y_p)'$ is either observed exactly or not at all. With missing data, we let $R = (R_1, \ldots, R_p)'$ be a vector of 0's and 1's satisfying $R_k = 1$ only if Y_k is observed. We assume the imputer models the joint density of (Y, R) as belonging to a parametric family of densities, $\{f(Y, R; \psi); \psi \in \Psi \subset \mathcal{R}^q\}$. Note that this model allows for non-ignorable missing data mechanisms (Rubin, 1987, p. 203).

To avoid extraneous issues, we assume that (i) the observed data maximum likelihood estimator $\widehat{\psi}$ is the unique solution to the observed data score equation, $\sum S_{obs}^{i}(\psi) = 0$, where $S_{obs}(\psi) = E_{\psi}\{S(\psi) \mid Y_R, R\}$, $S(\psi) = s(Y, R; \psi) = \partial \log f(Y, R; \psi)/\partial\psi$, (ii) $\widehat{\psi}$ converges in probability to a limit ψ^*, and (iii) $\widehat{\psi}$ is an asymptotically linear estimator of ψ^* with influence function

$$D(\psi^*) = I_{obs}^{-1} S_{obs}(\psi^*), \tag{1}$$

where

$$I_{obs} = -E\{\partial S_{obs}(\psi)/\partial\psi'\}_{\psi=\psi^*}.$$

That is,

$$n^{1/2}\left(\widehat{\psi} - \psi^*\right) = n^{-1/2}\sum_i D^i(\psi^*) + o_p(1), \tag{2}$$

so that $n^{1/2}(\widehat{\psi} - \psi^*)$ is asymptotically normal with mean zero and covariance matrix

$$\Lambda(\psi^*) = E\left\{D(\psi^*)^{\otimes 2}\right\} = I_{obs}^{-1} E\left\{S_{obs}^{\otimes 2}(\psi^*)\right\} I_{obs}^{-1}, \tag{3}$$

where $A^{\otimes 2} = AA'$. We do not assume that the model $f(Y, R; \psi)$ is correctly specified. That is, there may be no value of ψ for which $f(Y, R; \psi)$ is the true joint density $f_0(Y, R)$ of (Y, R).

We now describe the procedure for estimation by imputation. Suppose, in the absence of missing data, the user would report an estimator $\widehat{\beta}_c$ that solves the complete data estimating equation

$$0 = \sum U^i(\beta), \tag{4}$$

where $U^i(\beta) = u\{Y^i; \beta\}$. For example, let $Y = (Z, X', W')'$ and $u(Y, \beta) = (Z - \beta' X)X$. Then $\widehat{\beta}_c$ is the ordinary least squares estimator of the regression of Z on X, ignoring the data on W. Throughout we suppose the user has available an off-the-shelf commercial software package that computes $\widehat{\beta}$ from n independent observations $Y^i, i = 1, \ldots, n$, and has some ability to do simple matrix calculations. In general, $\sum U^i(\beta)$ may be a complete data estimating equation under a non-, semi- or parametric statistical model that (i) may be misspecified and (ii) may have no connection at all and indeed may be incompatible with the imputer's model $f(Y, R; \psi)$.

In the presence of missing data, the imputer, for each subject i, imputes m completed data vectors $Y^{ij} = (Y^{ij}_{\overline{R}}, Y^i_R), j = 1, \ldots, m$, of Y. Each $Y^{ij}_{\overline{R}} \equiv Y^{ij}_{\overline{R}}(\widehat{\psi})$ is drawn independently from the conditional density $f(Y_{\overline{R}} \mid Y^i_R, R^i; \widehat{\psi})$ of $Y_{\overline{R}}$ given the observed data (Y^i_R, R^i) evaluated at the maximum likelihood estimator $\widehat{\psi}$. The user then reports the estimator $\widehat{\beta}$ solving

$$0 = \sum \overline{U}^i \left(\widehat{\psi}, \beta\right), \tag{5}$$

where $\overline{U}^i(\widehat{\psi}, \beta) = m^{-1} \sum_j U^{ij}(\widehat{\psi}, \beta)$ and $U^{ij}(\widehat{\psi}, \beta) = u\{Y^{ij}(\widehat{\psi}), \beta\}$. Note that, for a subject i without missing data, $Y^{ij}(\widehat{\psi}) \equiv Y^i$ and $U^{ij}(\widehat{\psi}, \beta) \equiv U^i(\beta)$, for all j. To compute $\widehat{\beta}$, the user simply inputs the mn observations $\{Y^{ij}(\widehat{\psi})\}$ as 'independent' observations to an off-the-shelf software package. We will assume that, with probability tending to 1, equation (5) has a unique solution $\widehat{\beta}$ which converges to a limit β^*.

The following theorem provides the asymptotic distribution of $n^{1/2}(\widehat{\beta} - \beta^*)$. In the theorem and elsewhere, for any $H = h\left(R, Y_R, Y^1_{\overline{R}}, Y^2_{\overline{R}}, \ldots, Y^m_{\overline{R}}\right)$, $E(H)$ denotes the expectation of H with respect to the density $f(R, Y_R, Y^1_{\overline{R}}, \ldots, Y^m_{\overline{R}}) = \prod_j f(Y^j_{\overline{R}} \mid Y_R, R; \psi^*) f_0(Y_R, R)$, where $f_0(Y_R, R)$ is the true marginal density of (Y_R, R).

THEOREM 1. *Under the regularity conditions given in the Appendix, $n^{1/2}(\widehat{\beta} - \beta^*)$ is asymptotically normal with mean zero and variance $\Sigma = \tau^{-1}\,\Omega\,(\tau')^{-1}$, where*

$$\tau = -E\left\{\partial \overline{U}\left(\psi^*, \beta\right)/\partial \beta'\right\}_{\beta=\beta^*},$$

$$\Omega = E\left\{\overline{U}\left(\psi^*, \beta^*\right)^{\otimes 2}\right\} + \kappa \Lambda\left(\psi^*\right)\kappa' + E\left[\kappa D\left(\psi^*\right)\overline{U}\left(\psi^*, \beta^*\right)' + \left\{\kappa D\left(\psi^*\right)\overline{U}\left(\psi^*, \beta^*\right)'\right\}'\right],$$

$$\kappa = E\left\{U\left(\psi^*, \beta^*\right) S_{mis}\left(\psi^*\right)'\right\}, \text{ and } S_{mis}\left(\psi^*\right) = \partial \log f\left(Y \mid Y_R, R; \psi\right)/\partial \psi \mid_{\psi=\psi^*}.$$

Note that, in Theorem 1, $\widehat{\beta}$ is centred around its probability limit β^*. As discussed further in § 6, Theorem 1 is totally agnostic as to bias in the sense that it is true regardless whether or not β^* equals the probability limit β_0 of $\widehat{\beta}_c$ that would be obtained in the absence of missing data. A non-zero difference $\beta^* - \beta_0$ implies that the imputation model $f(Y, R; \psi)$ is misspecified. The results in Theorem 1 suggest the following consistent estimator $\widehat{\Sigma}$ of Σ.

COROLLARY 1: *$\widehat{\Sigma} = \widehat{\tau}^{-1}\,\widehat{\Omega}\,(\widehat{\tau}')^{-1}$ is a consistent estimator of Σ, where*

$$\widehat{\tau} = \widehat{\tau}\left(\widehat{\psi}, \widehat{\beta}\right) = -n^{-1} \sum_{i=1}^{n} \partial \overline{U}^i\left(\widehat{\psi}, \beta\right)/\partial \beta' \mid_{\beta=\widehat{\beta}},$$

$$\widehat{\Omega} = \widehat{\Omega}\left(\widehat{\psi}, \widehat{\beta}\right) = \widehat{\Omega}_c + \widehat{\kappa}\widehat{\Lambda}\left(\widehat{\psi}\right)\widehat{\kappa}' + n^{-1} \sum_{i=1}^{n} \left[\widehat{\kappa}\widehat{D}^i\left(\widehat{\psi}\right)\overline{U}^i\left(\widehat{\psi}, \widehat{\beta}\right)' + \left\{\kappa \widehat{D}^i\left(\widehat{\psi}\right)\overline{U}^i\left(\widehat{\psi}, \widehat{\beta}\right)'\right\}'\right],$$

$$\widehat{\Omega}_c = \widehat{\Omega}_c\left(\widehat{\psi}, \widehat{\beta}\right) = n^{-1} \sum_{i=1}^{n} \overline{U}^i\left(\widehat{\psi}, \widehat{\beta}\right)^{\otimes 2}, \ \widehat{\kappa} = \widehat{\kappa}\left(\widehat{\psi}, \widehat{\beta}\right) = (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \overline{U}^{ij}\left(\widehat{\psi}, \widehat{\beta}\right)\left\{S^{ij}_{mis}\left(\widehat{\psi}\right)\right\}',$$

$$S^{ij}_{mis}\left(\widehat{\psi}\right) = \partial \log f\left\{Y^{ij}\left(\widehat{\psi}\right) \mid Y^i_R, R^i; \psi\right\}/\partial\psi \mid_{\psi=\widehat{\psi}}\ ,\ \widehat{\Lambda}\left(\widehat{\psi}\right) = n^{-1}\sum_{i=1}^{n}\widehat{D}^i\left(\widehat{\psi}\right)^{\otimes 2},$$
$$\widehat{D}^i\left(\widehat{\psi}\right) = -\left\{n^{-1}\sum_{i=1}^{n}\partial S^i_{obs}\left(\widehat{\psi}\right)/\partial\psi'\right\}^{-1} S^i_{obs}\left(\widehat{\psi}\right).$$

Note that, for a subject i without missing data, $S^{ij}_{mis}(\psi) \equiv 0$, for all j. Our results are completely non-parametric in the sense that our variance estimator is consistent whatever be the true but unknown joint distribution of the observed data (Y_R, R). Wang & Robins (1998) provided an alternative consistent estimator for the asymptotic variance of $\widehat{\beta}$ in the special case in which (i) the model $f(Y, R; \psi)$ is correctly specified and (ii) $U(\beta) = S(\psi)$ and $\beta = \psi$. When either (i) or (ii) is false, the variance estimator of Corollary 1 must be used, even when $\hat{\beta}$ is asymptotically unbiased in the sense that $\beta^* = \beta_0$.

It follows from the Corollary that a $(1-\alpha)$ large sample confidence interval for $c'\beta^*$ for a given constant vector c is $c'\widehat{\beta} \pm z_{(1-\alpha/2)}n^{-1/2}\{c'\widehat{\Sigma}c\}^{1/2}$, where z_α is the α-quantile of the standard normal distribution. The asymptotic coverage of this interval is $1-\alpha$, so it is valid without being conservative. In the absence of bias, this interval is also a $(1-\alpha)$ large sample confidence interval for $c'\beta_0$.

3. Software needed to compute the variance estimator $\widehat{\Sigma}$

We now discuss the software needed to compute $\widehat{\Sigma} = \widehat{\tau}^{-1}\,\widehat{\Omega}\,(\widehat{\tau}')^{-1}$. When the user inputs the mn observations $\{Y^{ij}(\widehat{\psi})\}$, as 'independent' observations, $\widehat{\tau}^{-1}/(nm)$ will most often be output by the user's software package as a component of the 'variance' of $\widehat{\beta}$. For example, if, as in generalised estimating equation models, the software outputs a 'robust' sandwich variance estimator, then $\widehat{\tau}^{-1}/(nm)$ will be the 'outside' of the sandwich.

Thus it remains to compute $\widehat{\Omega}$. In order to do so, one needs the $2q + q^*$ variables $\{S^{ij}_{mis}(\widehat{\psi})', \widehat{D}^i(\widehat{\psi})', U^{ij}(\widehat{\psi}, \widehat{\beta})'\}$, where q is the dimension of ψ and q^* is the dimension of β. The additional variables $\{S^{ij}_{mis}(\widehat{\psi})', \widehat{D}^i(\widehat{\psi})'\}$ can be appended by the imputer at the end of each of the nm rows $Y^{ij}(\widehat{\psi})$ of the observation matrix. These additional variables do not depend on the specific analysis chosen by the user, and are offered by the imputer as the essential information about his/her model which is needed to calculate $\widehat{\Sigma}$. When the assumed imputation model is simple, S^{ij}_{mis} and $\widehat{D}^i$ can often be obtained analytically. In any case, they are readily available from an estimate of S^i_{obs}, which can always be obtained by a numerical or Monte Carlo approximation to $E_{\widehat{\psi}}\{s(Y, R; \widehat{\psi}) \mid Y^i_R, R^i\}$ since $S^{ij}_{mis}(\psi) = S^{ij}(\psi) - S^i_{obs}(\psi)$ (Meilijson, 1989). Furthermore, it will be computationally convenient for the imputer to

supply separately the matrix $\widehat{\Lambda}(\widehat{\psi})$, rather than to require the user to calculate it from the $\widehat{D}^i(\widehat{\psi})$.

The q^* variables $U^{ij}(\widehat{\psi},\widehat{\beta})$ are supplied by the user and depend on the user's choice of analysis procedure. Unfortunately, most off-the-shelf software packages will not, upon convergence, output a dataset containing the nm individual contributions, $U^{ij}(\widehat{\psi},\widehat{\beta})$, to the estimating equation $0 = \sum\sum U^{ij}(\widehat{\psi},\widehat{\beta})$. However, the users can often calculate the $U^{ij}(\widehat{\psi},\widehat{\beta})$ for themselves with a few lines of additional programming. Specifically, in the aforementioned linear regression example, most software packages will output a dataset of predicted values $\widehat{\beta}'X^{ij}$ corresponding to each of the nm observations $Y^{ij}(\widehat{\psi}) = Y^{ij} = (Z^{ij}, X^{ij\prime}, W^{ij\prime})'$. It is then straightforward to compute $U^{ij}(\widehat{\psi},\widehat{\beta}) = (Z^{ij} - \widehat{\beta}'X^{ij})X^{ij}$. Finally, the simple algebra given in Corollary 1 can then be used to compute $\widehat{\Sigma} = \widehat{\tau}^{-1}\,\widehat{\Omega}\,(\widehat{\tau}')^{-1}$ for the dataset $\{Y^{ij}(\widehat{\psi})', \widehat{D}^i(\widehat{\psi})', S^{ij}_{mis}(\widehat{\psi})', U^{ij}(\widehat{\psi},\widehat{\beta})'\}$. We hope that, in the future, software developers will create packages that not only output the contributions $U^{ij}(\widehat{\psi},\widehat{\beta})$ but also include a small program to compute $\widehat{\Omega}$.

Clayton et al. (1998) derive an alternative analytic estimator for the asymptotic variance of imputation estimators. However, their variance estimator is not expressed in a form that lends itself to the creation of software that makes the calculation of the estimator essentially invisible to the user.

4. Large sample bias of $\widehat{\Sigma}_{Rubin}$

In Theorem 2 below, we characterise the large sample bias of Rubin's variance estimator $\widehat{\Sigma}_{Rubin}$. Recall that $\widehat{\beta}_c$ is the solution to (4) were there no missing data. In that case, $n^{1/2}(\widehat{\beta}_c - \beta_0)$ is asymptotically normal with mean zero and variance that can be consistently estimated by the sandwich variance formula $\widehat{V}(\widehat{\psi},\widehat{\beta}_c) = \widehat{\tau}(\widehat{\psi},\widehat{\beta}_c)^{-1}\,\widehat{\Omega}_c(\widehat{\psi},\widehat{\beta}_c)\,\{\widehat{\tau}(\widehat{\psi},\widehat{\beta}_c)'\}^{-1}$.

Rubin (1987, p. 76) proposed an estimator $\widehat{\Sigma}_{Rubin}$ for the variance of the multiple imputation estimator $\widehat{\beta}_{Rubin} = m^{-1}\sum_j \widehat{\beta}_{cj}$, where $\widehat{\beta}_{cj}$ is the value of $\widehat{\beta}_c$ based on the jth completed dataset, and the imputed values for the jth completed dataset are drawn from $f(Y_{\overline{R}} \mid Y_R, R; \psi_j)$, with ψ_j drawn from the posterior distribution of ψ under a Bayesian model. Specifically, $\widehat{\Sigma}_{Rubin} = \widehat{V}_\bullet + (1 + m^{-1})\overline{B}$, where $\widehat{V}_\bullet = m^{-1}\sum \widehat{V}_j$, $\widehat{V}_j = \widehat{V}(\psi_j, \widehat{\beta}_{cj})$ and $\overline{B} = n(m-1)^{-1}\sum_j(\widehat{\beta}_{cj} - \widehat{\beta}_{Rubin})^{\otimes 2}$. The following theorem restricts consideration to the situation in which the number of imputations m is infinite. Meng (1994) and Rubin (1996) also restricted their theoretical calculations to the infinite-m case. By arguments analogous to those in Wang & Robins (1998), it can be shown that, when the number of

imputations m is infinite, the estimator $\widehat{\beta}_{Rubin}$ and the estimator $\widehat{\beta}$ solving (5) are asymptotically equivalent with asymptotic mean β^* and the asymptotic variance, Σ, specified in Theorem 1. In contrast, for finite m, $\widehat{\beta}$ and $\widehat{\beta}_{Rubin}$ are asymptotically normal with the same mean β^*; however, the asymptotic variance Σ of $\widehat{\beta}$ is strictly smaller than that of $\widehat{\beta}_{Rubin}$. In the infinite-m case, the large sample bias of Rubin's variance estimator $\widehat{\Sigma}_{Rubin}$ is defined to be $\Sigma_{Rubin} - \Sigma$, where Σ_{Rubin} is the limit of $\widehat{\Sigma}_{Rubin}$, as n, m go to infinity.

THEOREM 2. *Under the regularity conditions given in the Appendix, in the infinite-m case,* $\Sigma_{Rubin} - \Sigma = \tau^{-1}\left[G + G' + \kappa I_{obs}^{-1}\left\{I_{obs} - E\left(S_{obs}^{\otimes 2}\right)\right\} I_{obs}^{-1}\kappa'\right](\tau')^{-1}$, *where* $G = E\left[\text{var}\left\{U\left(\psi^*, \beta^*\right) \mid R, Y_R\right\}\right] - \kappa E\left\{D\left(\psi^*\right) U\left(\psi^*, \beta^*\right)'\right\}$, *and* $E\left(S_{obs}^{\otimes 2}\right) = E\left\{S_{obs}^{\otimes 2}\left(\psi^*\right)\right\}$.

Remark: Note that, if the imputation model is correctly specified then (i) I_{obs} will equal $E(S_{obs}^{\otimes 2})$ and the third term in the square braces is zero, and (ii) if the user chooses $\beta = \psi$ and $U\left(\beta\right) = S\left(\psi\right)$, so that the user computes the maximum likelihood estimator under the imputer's model, then G is zero and Rubin's variance estimator is without large sample bias (Meng, 1994; Wang & Robins, 1998).

***Example* 1.** We consider a simple hypothetical example which illustrates that Rubin's variance estimator may be either upwardly or downwardly biased even when $\hat{\beta}$ is asymptotically unbiased in the sense that $\beta^* = \beta_0$. Suppose there are two classes at a daycare centre. The relevant variables are the classroom indicator A, with 0 denoting infants and 1 denoting toddlers, the age X of the children, the child's score Z on a test of gross motor skills and the indicator variable R, that takes the value 1 if Z is observed and is 0 otherwise. A fraction π of the toddlers are missing Z because of illness on the day the test was given. There is no missing data among the infants.

The data are generated as follows. For $i = 1, \ldots, n$, $(R^i, Y^i) = (R^i, A^i, X^i, Z^i)$ is independent, identically distributed realisation of (R, A, X, Z); $\text{pr}(A = 0) = 2/3$, so there is a two-to-one infant to toddler ratio; $[X \mid A = 0]$ is Un$[0.1, 0.8]$ and $[X \mid A = 1]$ is Un$[0.8, 2.0]$; $Z \mid A, X \sim N\left(\beta X, \sigma^2 X^{\eta_A^*}\right)$; and $\text{pr}(R = 0 \mid X, A, Z) = A\pi$, so that missingness among the toddlers is completely at random.

The imputer specifies that the data in the toddlers are missing completely at random and that $Z \mid A, X \sim N\left(\beta X,\ \sigma^2 X^{\eta}\right)$, where η is regarded as known and (β, σ^2) are unknown parameters to be estimated. Thus, the imputer's model is correctly specified when $\eta = \eta_0^*$ $= \eta_1^*$. The imputer fits his/her model to the data on all the children.

The user's procedure is to fit the no-intercept regression model $Z = \beta X + \varepsilon$ by ordinary least squares through the origin. In this example, $\beta^* = \beta_0 = \beta$.

Scenario 1. In our first scenario, in order to estimate β for the toddlers, the user fits only the completed data in the toddlers, ignoring the data on the infants. Table 1 reports the infinite-m relative bias $(\Sigma_{Rubin} - \Sigma)/\Sigma$ as a function of π when $\eta = \eta_0^* = \eta_1^* = 0$, so that the imputer's and user's models are both correct. This example is similar in spirit to the example discussed in Meng (1994). As in Meng (1994), Rubin's variance estimator is upwardly biased.

Scenario 2. In our second scenario, in order to estimate β for all children, the user fits the completed data on both toddlers and infants by ordinary least squares through the origin. The imputer continues to assume the errors are homoscedastic, that is, $\eta = 0$. However, $\eta_0^* = \eta_1^* = 1$, $\eta_0^* = \eta_1^* = -1$ or $\eta_0^* = 2$ and $\eta_1^* = 1$. Note that, since $\eta^* \neq \eta$, the imputer's model is misspecified. Reading from Table 1, we observe that Rubin's variance estimator can suffer from substantial downward bias.

Scenario 3. Our final scenario differs from Scenario 2 only in that now $\eta = \eta_0^* = \eta_1^*$, with all equal to either -1 or 1, so that the imputer's model is again correct. Even when the imputer's model is correct, Rubin's variance estimator can still be downwardly biased when $\eta = \eta_0^* = \eta_1^* = -1$, although the magnitude of the bias is much smaller than that in Scenario 2.

We conducted a small simulation study under Scenario 2 to determine whether the large sample downward bias of $\widehat{\Sigma}_{Rubin}$ reported for Scenario 2 in Table 1 is also present in small to moderate size samples. In our simulation study, we chose $n = 150$, $\beta = 1$, $\sigma^2 = 1$, $\eta_0^* = 2$, $\eta_1^* = 1$, $\eta = 0$ and $\pi = 0.6$. The number of completed datasets m was either 5 or 20. Results for the estimators $\widehat{\beta}$ and $\widehat{\Sigma}$ and the nominal 95% interval $\widehat{\beta} \pm 1.96\widehat{\Sigma}^{1/2}$ are reported in the rows 'R-W'. Results based on $\widehat{\beta}_{Rubin}$ and $\widehat{\Sigma}_{Rubin}$ and the nominal 95% t-interval proposed by Rubin (1987, p. 77) are reported in the row 'Rubin'. To carry out Rubin's procedure, it was necessary to specify a prior distribution for the unknown parameters (β, σ^2) of the imputer's model. We chose independent flat priors on β and $\log \sigma^2$ as suggested by Rubin (1987, p. 166). Reading from Table 2, we observe that $\widehat{\Sigma}_{Rubin}$ has a large downward bias, underestimating the simulation variance of $\widehat{\beta}_{Rubin}$ by over 50%. As a consequence, Rubin's interval estimator undercovers. Note also that both $\hat{\beta}$ and $\hat{\beta}_{Rubin}$ are essentially unbiased for $\beta = 1$. Furthermore, as expected, when $m = 5$, $\hat{\beta}_{Rubin}$ is slightly less efficient than $\hat{\beta}$. Finally, because the user's model is mispecified, the variance estimator of Wang & Robins (1998) would also be biased since it fails to properly account for heteroscedacticity.

5. A conditional imputation model

So far, we have assumed that the imputation model was a fully parametric probability model for the joint distribution of (Y, R). However, since all imputations are drawn from the law of Y given Y_R and R, we can in fact specify a parametric model $f(Y \mid Y_R, R; \psi)$ for the law of Y given Y_R and R, and leave the rest of the joint distribution of (Y, R) unspecified. In this approach, which was suggested by Clayton et al. (1998), we do not even require that the model $f(Y \mid Y_R, R; \psi)$ be internally consistent in the sense that there exists some single probability law for (Y, R) that has the implied functional form $f(Y \mid Y_R, R; \psi)$, as (Y_R, R) varies. Indeed, all that is required is some method for constructing an estimator $\widehat{\psi}$ of ψ from the observed data so that we can simulate from $f(Y \mid Y_R, R; \widehat{\psi})$. Specifically, we shall assume that we have available an estimator $\widehat{\psi}$ obtained by solving some set of estimating equations based on the observed data $(R^i, Y_R^i, i = 1, \ldots, n)$ and that (i) $\widehat{\psi}$ converges to a limit ψ^* and (ii) there exists a zero mean finite variance influence function of $\widehat{\psi}$, $D(\psi^*)$. Then Theorem 1 and its corollary remain true, except that the influence function of $\widehat{\psi}$, $D(\psi^*)$, will no longer be obtained based on the right-hand sides of (1) and (2). Rather, if $\widehat{\psi}$ solves

$$0 = \sum_{i=1}^{n} M_i(\psi), \tag{6}$$

where $M(\psi) = m(Y_R, R, \psi)$, then by a Taylor expansion we obtain

$$D(\psi^*) = \left[-E\left\{\partial M(\psi)/\partial \psi'\right\}\right]^{-1}_{\psi=\psi^*} M(\psi^*), \; \widehat{D}\left(\widehat{\psi}\right) = \left\{-n^{-1}\textstyle\sum_i \partial M_i(\psi)/\partial \psi'\right\}^{-1}_{\psi=\widehat{\psi}} M\left(\widehat{\psi}\right).$$

To make the above concrete, we shall use data given in Clayton et al. (1998), where the aforementioned conditional imputation approach was applied, to illustrate the performance of the proposed consistent variance estimator. This dataset, which was referred to as dataset 1 in Clayton et al. (1998), was made available to us by David Clayton. Clayton et al. (1998) also analysed this dataset using a conditional imputation approach. The complete data are $Y = (D_0, D_1, S_0, S_1, X)$, where, for $j = 0, 1$, D_j recorded a subject's dementia status at time j as diagnosed by a physician, S_j was a subject's mini-mental status exam score at time j and X=(1,sex,age$'$) $'$, with sex indicating male/female and age being a vector of five dummy variables encoding six age categories. Following Clayton et al. (1998), our analysis model is a linear logistic model for development of dementia between times 0 and 1 as a function of the regressors X, that is,

$$\mathrm{pr}(D_1 = 1 \mid X, D_0 = 0) = \mathrm{expit}(\beta' X) \;\text{ where } \mathrm{expit}(u) = \exp(u)/\{1 + \exp(u)\}.$$

Thus $U(\beta) = \{D_1 - \text{expit}(\beta' X)\}X(1 - D_0)$ is the usual score function for logistic regression restricted to subjects without dementia at time zero. In dataset 1, Y was not fully observed; of the ten thousand study subjects, the proportions of subjects missing both D_1 and D_0, missing D_1 alone and missing D_0 alone are, respectively, 55, 18 and 19 percent. The observed data were $R = (R_0, R_1), Y_R = (R_0 D_0, R_1 D_1, X, S)$, where $S = (S_0, S_1)'$ and $R_j = 1$ if D_j was observed and $R_j = 0$ otherwise.

We used Clayton et al.'s conditional imputation model $f(Y \mid R, Y_R; \psi)$ for the above three patterns of missing data; for $j = 0, 1$,

$$f\left(D_j \mid R_j = 0, R_{1-j} = 1, D_{1-j}, X, S; \psi\right) = g_j\left(D_j, X, S_j, \psi\right),$$

where $g_j(1, X, S_j, \psi) = \text{expit}(\psi' W_j)$ with $W_j = \{X', jX', (\text{age} \times \text{sex})', j(\text{age} \times \text{sex})', S_j, S_j^2\}'$, and age×sex encoding the gender-age interaction. In addition, for $k, \ell = 0, 1$,

$$f\left(D_0 = k, D_1 = \ell \mid R_0 = 0, R_1 = 0, X, S\right) = g_0\left(k, X, S_0, \psi\right) g_1\left(\ell, X, S_1, \psi\right).$$

We estimated ψ from the complete cases, with $R_0 = R_1 = 1$, by logistic regression, treating each subject's two outcomes D_0 and D_1 as independent. That is, in (6), we chose $M(\psi) = M_0(\psi) + M_1(\psi)$, where $M_j(\psi) = R_0 R_1\{D_j - \text{expit}(\psi' W_j)\}W_j$. We then imputed $m = 5$ completed datasets from $f(Y \mid R, Y_R; \widehat{\psi})$.

Following Clayton et al. (1998), we shall focus on the sex effect β_{sex}. The estimated standard error of $\widehat{\beta}_{sex}$ based on Corollary 1 was 0.1639. It is of some interest to compare this with the nonparametric bootstrap estimate of the standard error of $\widehat{\beta}_{sex}$ (Efron & Tibshirani, 1993, Ch. 6; Efron, 1994), since, like our standard error estimator, the bootstrap estimator is consistent for the asymptotic standard error of $n^{1/2}(\widehat{\beta}_{sex} - \beta^*_{sex})$ regardless of model misspecification or incompatibility.

The bootstrap estimate of the standard error of $\widehat{\beta}_{sex}$ based on 200 bootstrap resamples of the 10,000 observed data vectors was 0.1652, which, as predicted by our asymptotic theory, was similar to our analytic estimate of 0.1639. Indeed, the nonparametric bootstrap variance estimator could be an alternative to our analytic estimator. However, as pointed out by Rubin (1994, 1996), the nonparametric bootstrap estimator is much more computationally intensive, especially in handling a dataset as large as the current one, and this computational burden is on the users rather than on the imputer.

6. Discussion

We have derived an estimate of the asymptotic variance of the imputation estimator $\widehat{\beta}$ that is consistent even when the imputation analysis models are misspecified and incompatible with one another. It follows that in large samples the associated Wald interval estimator will cover the limit β^* of $\widehat{\beta}$ at its nominal rate. An important limitation of our approach as well as those of Rubin and Clayton et al. is that, if the parametric imputation model is misspecified, then the parameter β_0 that would be estimated in the absence of missing data may greatly differ from β^*. In that case, β^* may be of no scientific interest and our Wald intervals will fail to cover β_0 at the nominal rate. For this reason, one should consider, when possible, alternative estimators that are more robust than a parametric imputation estimator. For example, when the only source of missing data is by design and thus the non-response probabilities are known, the locally semiparametric efficient augmented inverse probability of response weighted estimators described by Robins & Wang (1998), Robins & Ritov (1997) and Robins, Rotnitzky & Zhao (1994) guarantee asymptotic unbiasedness, while often vastly improving upon the poor efficiency of the estimator of Horvitz & Thompson (1952).

Indeed, even when missingness is unplanned rather than by design, locally semiparametric efficient augmented inverse probability of response weighted estimators are still considerably more robust than parametric multiple imputation estimators. Specifically, if non-response is non-ignorable, consistency of a locally semiparametric efficient estimator only requires a correctly specified model for the non-response probabilities (Rotnitzky & Robins, 1997; Robins, 1997; Robins, Rotnitzky & Scharfstein, 1999). In contrast, for consistency, a parametric multiple imputation estimator additionally requires a correctly specified parametric model for the marginal distribution of the complete data. If non-response is ignorable, a locally semiparametric efficient estimator is doubly protected; that is, it is consistent if either a model for non-response or a parametric model for the complete data can be correctly specified. On the other hand, consistency of a parametric multiple imputation estimator requires correct specification of a parametric model for the complete data (Scharfstein, Rotnitzky & Robins, 1999).

Thus, to avoid bias, one should always be cautious in the use of parametric imputation models. However, in cases in which the variance of the 'inverse probability' weights is very large, the sampling distribution of a locally semiparametric efficient augmented inverse probability of response weighted estimator can be markedly skew and highly variable, and a parametric imputation estimator may be preferred. Since the specification of the parametric imputation model $f(Y, R; \psi)$ cannot be fully checked from the observed data, we would

recommend a sensitivity analysis in which the data are reanalysed under a number of different models $f(Y, R; \psi)$. Finally, a small note of caution: as with any procedure motivated by large sample theory, the performance of our variance estimator may degrade in small samples. When there is doubt, investigation by simulation would be warranted.

Acknowledgement

This research was supported by grants from the US National Institutes of Health. We thank the editor and referees, whose comments stimulated improvement of the paper.

Appendix

Assumptions and proofs

Along with the regularity assumptions mentioned in the text, we further assume that the following conditions hold for Theorem 1, for all (ψ, β) in a neighbourhood of (ψ^*, β^*).

(S1). Let $\lambda(\psi, \beta)$ be $E\{U(\psi, \beta)\}$; note the definition of the 'expectation operator' given in § 2. We also assume that $\partial\lambda(\psi, \beta)/\partial\psi'$ and $\partial\lambda(\psi, \beta)/\partial\beta'$ exist and are continuous in (ψ, β); in addition, the inverse of $(\partial/\partial\beta')\lambda(\psi, \beta)$ exists.

(S2). Both $\partial \log f(Y_{\overline{R}}|Y_R, R, \psi)/\partial\psi'$ and $\partial U(\psi, \beta)/\partial\beta'$ exist and are bounded in L^2.

(S3). Let $\mathcal{Z}_{n,\beta}(\psi_1, \psi_2) = n^{-1/2} \mid \sum \overline{U}^i(\psi_1, \beta) - \sum \overline{U}^i(\psi_2, \beta) - \lambda(\psi_1, \beta) + \lambda(\psi_2, \beta) \mid$. We assume that there exists a positive ι such that, for any ψ_1, ψ_2 in a neighbourhood of ψ^*, $\sup_{|\psi_1 - \psi_2| < \iota} \mathcal{Z}_{n,\beta}(\psi_1, \psi_2) \to 0$ uniformly in β as $n \to \infty$.

(S4). There exists a positive d such that $E\{U(\psi, \beta)^{2+d}\}$ is finite.

Sketch proof for Theorem 1. The proof of Theorem 1 mimics that in Wang & Robins (1998) with their score function S replaced by our U. Note that Wang & Robins assumed a correctly specified parametric distributional structure, but we do not. Their S^{mis} denotes the derivative of the 'true' log conditional density of $Y_{\overline{R}}$ given the observed data, while the S^{mis} here is the derivative of the 'assumed' log conditional density of Y given the observed data under the imputation model. By (S1) and (S2),

$$\left\{ \frac{\partial}{\partial\tilde{\psi}'} \lambda\left(\tilde{\psi}, \beta\right) \right\}\Big|_{\tilde{\psi}=\psi^*} = E\left\{ \int \overline{U}(\tilde{\psi}, \beta^*) \frac{\partial}{\partial\tilde{\psi}'} f(Y_{\overline{R}}|Y_R, R, \tilde{\psi}) dY_{\overline{R}} \Big|_{\tilde{\psi}=\psi^*} \right\}$$

and $\partial f(Y_{\overline{R}}|Y_R, R, \tilde{\psi})/\partial\tilde{\psi}' = S_{mis}(\tilde{\psi}) f(Y_{\overline{R}}|Y_R, R, \tilde{\psi})$. The rest of the proof follows closely the proof of Theorem 1 in Wang & Robins (1998).

Sketch proof for Theorem 2. Let $\nu_j = \psi_j - \hat{\psi}$. By an argument analogous to that in the Appendix of Wang & Robins (1998), $\lim nm^{-1} \sum(\nu_j - \bar{\nu})^{\otimes 2}$ converges to $\lim\{nE(\nu_1 - \bar{\nu})^{\otimes 2}\}$ $= I_{obs}^{-1}$. Denote $E_\psi\{U(\psi,\beta)|Y_R, R\}$ by $U_{obs}(\psi,\beta)$. Write $\overline{B}$ as $n(m-1)^{-1}\sum_j\{(\widehat{\beta}_{cj} - \beta^*) - (\widehat{\beta}_{Rubin} - \beta^*)\}^{\otimes 2}$. Straightforward derivations show that, as $m, n \to \infty$, $\overline{B}$ converges to

$$\tau^{-1}\left[\kappa \lim_{n\to\infty}\left\{nE\left(\nu_1 - \bar{\nu}\right)^{\otimes 2}\right\}\kappa' + E\left\{U(\psi^*,\beta^*) - U_{obs}(\psi^*,\beta^*)\right\}^{\otimes 2}\right](\tau')^{-1}, \tag{A1}$$

while $\widehat{V}_\bullet$ converges to

$$\tau^{-1}E\left[U_{obs}(\psi^*,\beta^*)^{\otimes 2} + \left\{U(\psi^*,\beta^*) - U_{obs}(\psi^*,\beta^*)\right\}^{\otimes 2}\right](\tau')^{-1}. \tag{A2}$$

Hence, Σ_{Rubin} is the sum of (A1) and (A2). Furthermore, Ω, as defined in Theorem 1, converges to

$$E\left\{U_{obs}(\psi^*,\beta^*)^{\otimes 2}\right\} + \kappa\Lambda\left(\psi^*\right)\kappa' + \kappa E\left\{D\left(\psi^*\right)U\left(\psi^*,\beta^*\right)'\right\} + E\left\{D\left(\psi^*\right)U\left(\psi^*,\beta^*\right)'\right\}'\kappa'. \tag{A3}$$

Note that $E\left(E_{\psi^*}\left[\left\{U(\psi^*,\beta^*) - U_{obs}(\psi^*,\beta^*)\right\}^{\otimes 2}|Y_R,R\right]\right) = E\left[\mathrm{var}\left\{U(\psi^*,\beta^*)|Y_R,R\right\}\right]$. Replace $\Lambda(\psi^*)$ in (A3) by its second expression in (3) and recall that $\Sigma = \tau\,\Omega\,(\tau')^{-1}$. Theorem 2 now follows directly from a comparison of (A1), (A2) and (A3).

REFERENCES

Clayton, D., Dunn, G., Pickles, A. & Spiegelhalter, D. (1998). Analysis of longitudinal binary data from multiphase sampling (with Discussion). *J. R. Statist. Soc.* B **60,** 71-87.

Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Efron, B. (1994). Missing data, imputations, and the bootstrap (with Discussion). *J. Am. Statist. Assoc.* **89**, 463-79.

Fay, R. (1992). When are inferences from multiple imputation valid? In *Proc. Survey Res. Methodol. Sec. Am. Statist. Assoc.,* pp. 227-32.

Fay, R. (1994). Discussion of 'Multiple imputation inferences with uncongenial sources of input' by X.-L. Meng. *Statist. Sci.* **9,** 558-60.

Fay, R. (1996). Alternative paradigms for the analysis of imputed survey data. *J. Am. Statist. Assoc.* **91**, 490-8.

Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663-85.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Little, R.J.A. & Yao, L. (1996). Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* **52**, 1324-33.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. R. Statist. Soc.* B **51,** 127-38.

Meng, X.-L. (1994). Multiple imputation inferences with uncongenial sources of input (with Discussion). *Statist. Sci.* **9,** 538-73.

Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *J. Am. Statist. Assoc.* **92,** 1320-9.

Robins, J.M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statist. Med.* **16**, 39-56.

Robins, J.M. & Ritov, Y. (1997). A curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statist. Med.* **16**, 285-319.

Robins, J.M., Rotnitzky, A & Zho, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846-67.

Robins, J.M., Rotnitzky, A & Scharfstein, D.O. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Springer Lecture Notes in Statistics: Statistical Methods in Epidemiology,* Ed. M.E. Halloran, to appear. New York: Springer.

Robins, J.M. & Wang, N. (1998). Discussion of 'Analysis of longitudinal binary data from multiphase sampling' by D. Clayton, G. Dunn, A. Pickles and D. Spiegelhalter. *J. R. Statist. Soc.* B **60,** 91-3.

Rotnitzky, A. & Robins, J.M. (1997). Analysis of semiparametric regression models with non-ignorable non-response. *Statist. Med.* **16**, 81-102.

Rubin, D.B. (1978). Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. In *Proc. Survey Res. Methodol. Sec. Am. Statist. Assoc.*, pp. 20-

34.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D.B. (1994). Discussion of 'Missing data, imputations, and the bootstrap' by B. Efron. *J. Am. Statist. Assoc.* **89**, 475-8.

Rubin, D.B. (1996). Multiple imputation after 18 years. *J. Am. Statist. Assoc.* **91**, 473-90.

Scharfstein, D.O., Rotnitzky, A. & Robins J.M. (1999). Rejoinder to 'Adjusting for non-ignorable drop-out using semiparametric non-response models.' *J. Am. Statist. Assoc.* To appear.

Taylor, J.M., Munoz, A., Bass, S.M., Sah, A.J., Chmiel, J., Kingsley, L. et al. (1990). Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation. *Statist. Med.* **9,** 505-14.

Tu, X.M., Meng, X.-L. & Pagano, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *J. Am. Statist. Assoc.* **88,** 26-36.

Wang, N. & Robins, J.M. (1998). Large sample inference in parametric multiple imputation. *Biometrika* **85,** 935-48.

Table 1. *Theoretically calculated percent bias ratios,* $100 \times (\Sigma_{Rubin} - \Sigma)/\Sigma$.

	$\pi = 0.2$	$\pi = 0.4$	$\pi = 0.6$	$\pi = 0.8$
Scenario 1				
$\eta = \eta_0^* = \eta_1^* = 0$	7.88	16.4	25.7	35.8
Scenario 2				
$\eta = 0$, $\eta_0^* = 1$, $\eta_1^* = 1$	-13.8	-26.2	-36.5	-42.6
$\eta = 0$, $\eta_0^* = -1$, $\eta_1^* = -1$	44.4	79.0	99.4	98.0
$\eta = 0$, $\eta_0^* = 2$, $\eta_1^* = 1$	-16.6	-31.7	-45.1	-54.7
Scenario 3				
$\eta = \eta_0^* = \eta_1^* = 1$	10.7	21.8	33.4	45.4
$\eta = \eta_0^* = \eta_1^* = -1$	-2.18	-5.03	-8.61	-13.0

Table 2. *Results based on 1000 simulations for Scenario 2 with* $\pi = 0.6$, $n = 150$, $\beta = 1$, $\eta_0^* = 2$, $\eta_1^* = 1$, $\eta = 0$.

	Monte Carlo Mean	Monte Carlo Variance	Monte Carlo Average Estimated Variance	Empirical Coverage Probability
$m = 20$				
R-W	0.9904	0.01770	0.01636	0.941
Rubin	0.9900	0.01773	0.00830	0.811
$m = 5$				
R-W	0.9895	0.01773	0.01661	0.937
Rubin	0.9900	0.01806	0.00834	0.826

R-W: results from the method in this paper
Rubin: results from Rubin's method