Damião Nóbrega Da Silva, Chris Skinner, Jae Kwang Kim

Using binary paradata to correct for measurement error in survey data analysis

Article (Accepted version) (Refereed)

Original citation:
Da Silva, Damião Nóbrega, Skinner, Chris J. and Kim, Jae Kwang (2016) *Using binary paradata to correct for measurement error in survey data analysis.* Journal of the American Statistical Association, 111 (514). pp. 526-537. ISSN 0162-1459

DOI: 10.1080/01621459.2015.1130632

Using Binary Paradata to Correct for Measurement Error in Survey Data Analysis

Damião Nóbrega Da Silva

Departamento de Estatística, Universidade Federal do Rio Grande do Norte,

Natal, RN, Brazil, 59078-970

email: damiao@ccet.ufrn.br

Chris Skinner

London School of Economics and Political Science, UK, WC2A 2AE

email: c.j.skinner@lse.ac.uk

Jae Kwang Kim

Department of Statistics, Iowa State University, Ames, IA 50011

email: jkim@iasate.edu

November 16, 2015

Author's Footnote:

Damião Nóbrega Da Silva is Associate Professor, Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil, 59078-970. Chris Skinner is Professor, London School of Economics and Political Science, UK, WC2A 2AE. Jae Kwang Kim is Professor, Department of Statistics, Iowa State University, Ames, IA 50011. The first author acknowledges the development of this research on a postdoctoral research fellowship at the Southampton Statistical Sciences Research Institute and the use of the IRIDIS High Performance Computing Facility both at the University of Southampton, UK.

Abstract

Paradata refers here to data at unit level on an observed auxiliary variable, not usually of direct scientific interest, which may be informative about the quality of the survey data for the unit. There is increasing interest among survey researchers in how to use such data. Its use to reduce bias from nonresponse has received more attention so far than its use to correct for measurement error. This paper considers the latter with a focus on binary paradata indicating the presence of measurement error. A motivating application concerns inference about a regression model, where earnings is a covariate measured with error and whether a respondent refers to pay records is the paradata variable. We specify a parametric model allowing for either normally or t-distributed measurement errors and discuss the assumptions required to identify the regression coefficients. We propose two estimation approaches which take account of complex survey designs: pseudo-maximum likelihood estimation and parametric fractional imputation. These approaches are assessed in a simulation study and are applied to a regression of a measure of deprivation given earnings and other covariates using British Household Panel Survey data. It is found that the proposed approach to correcting for measurement error reduces bias and improves on the precision of a simple approach based on accurate observations. We outline briefly possible extensions to uses of this approach at earlier stages in the survey process. Supplemental materials are available online.

KEY WORDS: Auxiliary survey information; Complex sampling; Fractional imputation; Pseudo maximum likelihood.

1. INTRODUCTION

Survey researchers have shown increasing interest in potential uses of paradata, taken here to refer to variables recorded in a survey which are not of direct interest for analysis but may be informative about data quality. Kreuter (2013, sec. 1.4) notes that there has been a particular focus on the use of such information to reduce nonresponse bias, whereas questions about measurement error have received rather less attention. In this paper we consider the latter and specifically the question of how paradata might be used to correct regression analyses of survey data for bias induced by measurement error.

Examples of paradata which may be related to measurement error include those obtained automatically from computerized survey systems, such as times to respond to questions (Olson and Parkhurst 2013). Other examples are obtained from interviewer observations, such as whether an interviewer feels that a respondent's answers are accurate (Barrett et al. 2006) or whether the respondent answered a question with an expression of uncertainty (Mathiowetz 1998). In each case, the paradata variable is binary and might be interpreted as indicating the presence of measurement error. Thus, if the paradata variable is denoted a_i, it may take the value 1 if the observed value u_i^* of a variable of interest for unit i is accurate and 0 if inaccurate. Assuming that the accurate observation is without measurement error, we write

$$u_i^* = \begin{cases} u_i, & a_i = 1, \\ u_i + \tau\epsilon_i, & a_i = 0, \end{cases} \tag{1}$$

where u_i is the true value of the variable of interest, $\tau\epsilon_i$ denotes measurement error for the inaccurate measurement and it will be convenient to introduce the constant $\tau > 0$ as a scale parameter. Schouten and Calinescu (2013) discuss a related idea of a 'measurement profile', which indicates the existence of a specified form of survey measurement error.

One variable which is frequently collected in household surveys, is often included as a regressor in regression analyses of survey data, but has long been known to be measured with error is earnings (Rodgers et al. 1993; Moore et al. 2000). A natural binary paradata variable to use in a face-to-face survey in this case is whether the interviewer observes the respondent referring to their pay records when responding. There is reason to expect an accurate response if their records are referred to but not otherwise. In this paper we shall consider an application with these choices of u_i^* and a_i.

We consider a regression analysis where the outcome variable is a measure of hardship experienced by the respondent. There is interest and possible welfare policy implications in whether different kinds of people experience different levels of hardship for a given level of earnings or income and, in particular, whether there is variation by age. Such questions can be addressed through regression analysis and we shall consider an analysis, based upon that of Berthoud et al. (2009), using data from the British Household Panel Survey to explore the impact of using such a paradata variable to adjust the estimated regression coefficients.

The methodological objective of this paper is to consider how to analyse survey data when observations on one of the variables of interest consist of pairs of values (u_i^*, a_i), where a_i is a binary paradata variable and the model in (1) holds. We also consider sensitivity analysis to departures from model (1) where $a_i = 1$ may not guarantee accurate measurement. We focus on the case of regression analysis where u_i is a covariate. Battistin et al. (2003) considered a similar problem where a covariate is a measure of household consumption which is subject to recall error and where paradata in the form of interview quality indicators, such as an interviewer's assessment of how well the respondent understood the questions, were available as predictors of the recall error. They developed ways of identifying the model through the use of data from a second survey and through other restrictions on the model. Our approach, restricted to binary paradata, builds on the approach of Da Silva and Skinner (2014), who refer to the binary paradata variable as an accuracy indicator. They developed a pseudo maximum likelihood approach to estimating the finite population distribution function of u_i, under some normality assumptions. This paper differs in two main ways. Firstly, the target of inference is different. We consider here the problem of fitting a linear regression model when u_i is one of the covariates, for which it turns out that the parameter identification issues are somewhat different. Both this problem and the one of estimating a distribution function are, nevertheless, well-known cases where measurement error can lead to estimation bias, even if ϵ_i has mean 0. The second main new feature of this paper is that we develop a fractional imputation approach (Kim 2011), which may be applied to problems where the pseudo maximum likelihood approach is not tractable, but the model is still expressed in a parametric form. We seek to develop an approach which can accommodate complex sampling schemes and the fractional imputation method is well-suited to this objective.

There is a literature on the use of multiple imputation to correct for covariate measurement error when data from a calibration study are available (Cole et al. 2006; He and Zaslavsky 2009; Guo et al. 2012). Blackwell et al. (2015) also discuss the use of multiple imputation for measurement error and specifically allow for the presence of an observed binary variable a_i governing the occurrence of measurement error, as in (1), but they assume additional information about the measurement error variance and do not allow for complex sampling.

The use of paradata for inference in measurement error models is analogous to the use of instrumental variable methods (Carroll et al. 2006, ch. 6) in the sense that both paradata and instrumental variables are auxiliary variables. They are completely different, however, in that instrumental variables are assumed to be independent of (or at least uncorrelated with)the measurement error, whereas we are interested in paradata variables precisely because they are related to measurement error.

Our setting is related to the literature on measurement error with unequal variances (e.g. Fuller 1987, sec. 3.1), since a key feature of our model is that error variances are either zero or non-zero, according to the value of a_i. However, that literature generally assumes that auxiliary information is available about the unequal measurement error variances, whereas we only assume that a_i is observed.

We set out our framework and broad estimation approaches in Sections 2 and 3. The fractional imputation approach is presented in Section 4, with variance estimation covered in Section 5. An initial investigation of the proposed approaches is conducted by simulation in Section 6. The application of the proposed approach to data from the British Household Panel Survey is given in Section 7 with some concluding remarks in Section 8. Our approach makes strong parametric modeling assumptions and we discuss the sensitivity of our methods to departures from these assumptions both in the simulation in Section 6 and in the application in Section 7. The importance of modeling assumptions has also been recognized in the literature on multiple imputation for measurement error, where Guo et al. (2012) provide a sensitive analysis to evaluate robustness to violation of a normality assumption and Blackwell et al. (2015) discuss how to incorporate different assumptions about the measurement error.

2. THE FRAMEWORK AND MODEL

We now set out the inferential framework and modelling assumptions. Consider a population of N units, denoted by $U = \{1, ..., N\}$, from which a probability sample A of size n is selected and for which a regression analysis is to be undertaken. Let y_i denote the value of the dependent variable and $(\mathbf{x}_{1i}^\top, u_i)^\top$ the value of the vector of explanatory variables for unit $i \in U$. We suppose that the vector $\mathbf{x}_{1i}$ is observed without error for units in A, but that u_i is measured with error by u_i^*. To permit a departure from model (1), we denote the observed binary accuracy indicator by a_i^* and treat the binary variable a_i which does obey (1) as unobserved. We assume that $a_i = 0$ if $a_i^* = 0$ but that $a_i = 0$ with probability p and $a_i = 1$ with probability $1-p$ if $a_i^* = 1$. When $p = 0$, we have $a_i^* = a_i$ and the observed accuracy indicator obeys (1). In general, we treat p as a specified known value, possibly derived from some external source, which may be varied from 0 in a sensitivity analysis, allowing for the possibility that $a_i^* = 1$ does not guarantee accuracy. We suppose the population values y_i are generated independently by the standard linear regression model

$$y_i = \mathbf{x}_{1i}^\top \boldsymbol{\beta}_x + u_i \beta_u + e_i, \tag{2}$$

where $e_i \sim N(0, \sigma^2)$ is independent of $(\mathbf{x}_{1i}, u_i)$. The objective is to make inference about $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^\top, \beta_u)$ given the observed data, which we assume to be of the form $\{(y_i, \mathbf{x}_i^\top, u_i^*, a_i^*) : i \in A\}$, where $\mathbf{x}_i = (\mathbf{x}_{1i}^\top, \mathbf{x}_{2i}^\top)^\top$ and the vector $\mathbf{x}_{2i}$ contains additional explanatory variables which may affect u_i, as will be discussed later.

We suppose that the model governing the measurement error in u_i is given by (1). We consider two possible forms for the measurement error distribution. The probability density, g, of ϵ_i may be either: (i) standard normal, where

$$g(\epsilon_i \mid u_i, \mathbf{x}_i, a_i = 0) = \phi(\epsilon_i), \tag{3}$$

and $\phi(\cdot)$ denotes the density of the standard normal distribution or (ii) Student's t with pre–specified degrees of freedom ν, where

$$g(\epsilon_i \mid u_i, \mathbf{x}_i, a_i = 0) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\epsilon_i^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{4}$$

The first case represents a classical measurement error model. The latter case represents a measurement error distribution which is more robust than the standard normal against possible outliers in the data (Lange et al. 1989) and seems natural in applications such as ours.

We propose to base inference on a fully parametric model and now set out the features of this model. Our underlying assumptions are that

- $\{(y_i, u_i^*, u_i, a_i, a_i^*, \mathbf{x}_i) : i = 1, ..., N\}$ are independent random vectors;
- y_i and u_i^* are conditionally independent given u_i, a_i and $\mathbf{x}_i$, for all $i \in U$;
- y_i and a_i are conditionally independent given u_i and $\mathbf{x}_i$, for all $i \in U$;
- (y_i, u_i^*, u_i) and a_i^* are conditionally independent given a_i and $\mathbf{x}_i$, for all $i \in U$.

The first assumption represents a standard kind of superpopulation model used in the analysis of complex survey data. The parameters of interest are defined through the marginal unit-level model. Inference about these parameters under complex sampling, such as stratification and clustering, will be handled through survey weighting and variance estimation methods rather than by elaborating the model Skinner et al. (1989). The second assumption is the standard one of nondifferential measurement error, where u_i^* is a surrogate of u_i (Carroll et al. 2006, sec. 2.5). The third assumption is similar to the second in supposing that, not only is the measurement error nondifferential with respect to y_i given u_i and $\mathbf{x}_i$, but this is also true of the accuracy indicator designed to be associated with this measurement error. The fourth assumption enables sensitivity analysis for a simple departure from (1).

Using the above assumptions, we express the basic model as

$$f(y_i, u_i^*, u_i, a_i \mid a_i^*, \mathbf{x}_i) = f(y_i \mid u_i, \mathbf{x}_i; \boldsymbol{\gamma}) f(u_i^* \mid u_i, a_i, \mathbf{x}_i; \tau^2) f(u_i \mid a_i, \mathbf{x}_i; \boldsymbol{\delta}) f(a_i \mid a_i^*). \tag{5}$$

The first three components of (5) are parameterised in terms of $\boldsymbol{\gamma} = (\boldsymbol{\beta}_x^\top, \beta_u, \sigma^2)^\top$, τ^2 and $\boldsymbol{\delta} = (\boldsymbol{\delta}_u^\top, \delta_a, \sigma_u^2)^\top$. The density $f(y_i \mid u_i, \mathbf{x}_i; \boldsymbol{\gamma})$ represents the regression model of interest in (2). The density $f(u_i^* \mid u_i, a_i, \mathbf{x}_i; \tau^2)$ refers to the measurement error model in (1). Our framework allows for dependence of the distribution of measurement error on $\mathbf{x}_i$, but we shall not find this necessary in our application and for simplicity do not include such dependence in the distributions in (3) and (4). The density $f(u_i \mid a_i, \mathbf{x}_i; \boldsymbol{\delta})$ refers to the distribution of the true value of the variable measured with error. The fourth component of (5) depends only on p, treated as specified in a sensitivity analysis and not as an unknown parameter.

Turning to parameter identification, we note first that if $p = 0$ then the parameter vector $\boldsymbol{\gamma}$ could be identified from the $a_i = 1$ observations, but this is not the case for the parameters τ^2 and $\boldsymbol{\delta}$. A basic problem in identifying these parameters from the joint distribution of (u_i^*, a_i) given the $\mathbf{x}_i$ (even if one could assume $p = 0$) is that differences in the observed distribution of u_i^* between cases with $a_i = 1$ and cases with $a_i = 0$ may arise either because of measurement error in the latter cases or because a_i is associated with u_i even in the absence of measurement error. Thus, further assumptions are required for identification. Da Silva and Skinner (2014) deal with this problem by assuming that a_i is conditionally independent of u_i given $\mathbf{x}_i$, analogous to the 'missing at random' assumption in missing data analysis, treating a_i as analogous to the missing data indicator (Little and Rubin 2002). See also Blackwell et al. (2015). In our application we shall assume that $E(\epsilon_i \mid a_i = 0, \mathbf{x}_i) \equiv 0$ in (1) which implies that $E(u_i^* \mid a_i, \mathbf{x}_i) = E(u_i \mid a_i, \mathbf{x}_i)$. Hence, the assumption that a_i is conditionally independent of u_i given $\mathbf{x}_i$ would be testable by regressing u_i^* on a_i and $\mathbf{x}_i$ and testing whether the coefficient of a_i is zero if we could suppose that $p = 0$. In our application we found some evidence that this coefficient is not zero across a number of choices of $\mathbf{x}_i$ and thus allow for a departure from the assumptions in Da Silva and Skinner (2014) by supposing that the conditional distribution of u_i given a_i and $\mathbf{x}_i$ follows a normal regression model, that is

$$f(u_i \mid a_i, \mathbf{x}_i; \boldsymbol{\delta}) = \sigma_u^{-1} \phi((u_i - \mathbf{x}_{2i}^\top \boldsymbol{\delta}_u - a_i \delta_a)/\sigma_u). \tag{6}$$

where δ_a may not be zero.

3. ESTIMATION

Let the parameter vector indexing the overall model described in the previous section be $\boldsymbol{\psi} = (\boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top, \tau^2)$. Suppose that a set of survey weights $\{w_i : i \in A\}$ is available which enables consistent estimation of population totals. We consider estimating $\boldsymbol{\psi}$ via pseudo maximum likelihood (Binder 1983; Godambe and Thompson 1986), as discussed in Da Silva and Skinner (2014). The pseudo maximum likelihood estimator (PML) $\widehat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ is defined as the solution to the pseudo score equations for the observed data $\{(u_i^*, y_i, a_i^*, \mathbf{x}_i) : i \in A\}$, given by

$$\bar{\boldsymbol{S}}_{obs}(\boldsymbol{\psi}) = \sum_{i \in A} w_i \bar{\boldsymbol{S}}_{obs,i}(\boldsymbol{\psi}) = \mathbf{0}, \tag{7}$$

where $\bar{\boldsymbol{S}}_{obs,i}(\boldsymbol{\psi}) \equiv \bar{\boldsymbol{S}}_{obs}(\boldsymbol{\psi} \mid u_i^*, y_i, a_i^*, \mathbf{x}_i) = \partial \ln f(u_i^*, y_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi})/\partial \boldsymbol{\psi}$ is the score function of $\boldsymbol{\psi}$ for the observed vector $(u_i^*, y_i, a_i^*, \mathbf{x}_i)$. Under regularity conditions,

$$\bar{\boldsymbol{S}}_{obs,i}(\boldsymbol{\psi}) = E\big[\boldsymbol{S}_{com,i}(\boldsymbol{\psi}) \mid u_i^*, y_i, a_i^*, \mathbf{x}_i\big], \tag{8}$$

where

$$\boldsymbol{S}_{com,i}(\boldsymbol{\psi}) \equiv \boldsymbol{S}_{com}(\boldsymbol{\psi} \mid u_i, a_i, u_i^*, y_i, a_i^*, \mathbf{x}_i) = \frac{\partial}{\partial \boldsymbol{\psi}} \ln f(u_i, a_i, u_i^*, y_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi}) \tag{9}$$

is the score function of $\boldsymbol{\psi}$ for the complete vector of observations at the i–th unit, (u_i, a_i, y_i, u_i^*, a_i^*, $\mathbf{x}_i$), and the expectation is taken with respect to the joint conditional distribution of u_i and a_i given u_i^*, y_i, a_i^* and $\mathbf{x}_i$.

Closed-form expressions for $\bar{\boldsymbol{S}}_{obs}(\boldsymbol{\psi})$ and details of how the pseudo maximum likelihood estimator can be computed for the normal measurement error model (3) are set out in the online supplementary materials (section 1). In order to interpret the resulting estimator of $\boldsymbol{\beta}$, we note first that the components of the pseudo score equations in (7) corresponding to $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^\top, \beta_u)$ are given by

$$\begin{aligned}
\bar{\boldsymbol{S}}_{\boldsymbol{\beta}_x}(\boldsymbol{\psi}) &= \frac{1}{\sigma^2} \sum_{i \in A} w_i \left\{ y_i - \mathbf{x}_{1i}^\top \boldsymbol{\beta}_x - \beta_u z_{i,1}(\boldsymbol{\psi}) \right\} \mathbf{x}_{1i} \\
\bar{S}_{\beta_u}(\boldsymbol{\psi}) &= \frac{1}{\sigma^2} \sum_{i \in A} w_i \left\{ (y_i - \mathbf{x}_{1i}^\top \boldsymbol{\beta}_x) z_{i,1}(\boldsymbol{\psi}) - \beta_u z_{i,2}(\boldsymbol{\psi}) \right\},
\end{aligned}$$

where $z_{i,1}(\boldsymbol{\psi}) \equiv \mathrm{E}\big[u_i \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \boldsymbol{\psi}\big]$ and $z_{i,2}(\boldsymbol{\psi}) \equiv \mathrm{E}\big[u_i^2 \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \boldsymbol{\psi}\big]$. It follows that we may express the pseudo maximum likelihood estimator of $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\psi}}) \equiv \begin{bmatrix} \widehat{\boldsymbol{\beta}}_x(\widehat{\boldsymbol{\psi}}) \\ \widehat{\boldsymbol{\beta}}_u(\widehat{\boldsymbol{\psi}}) \end{bmatrix} \equiv \left\{ \sum_{i \in A} w_i \begin{bmatrix} \mathbf{x}_{1i}\mathbf{x}_{1i}^\top & \mathbf{x}_{1i} z_{i,1}(\widehat{\boldsymbol{\psi}}) \\ z_{i,1}(\widehat{\boldsymbol{\psi}})\mathbf{x}_{1i}^\top & z_{i,2}(\widehat{\boldsymbol{\psi}}) \end{bmatrix} \right\}^{-1} \sum_{i \in A} w_i \begin{bmatrix} \mathbf{x}_{1i} y_i \\ z_{i,1}(\widehat{\boldsymbol{\psi}}) y_i \end{bmatrix}. \tag{10}$$

The conditional expectations $z_{i,1}$ and $z_{i,2}$ in this expression may be interpreted as analagous to those arising from a mean score approach to the treatment of missing data, viewing the u_i as missing in our setting (Kim and Shao 2013). In other respects, this expression is familiar as a survey weighted least squares estimator which is design-consistent for a corresponding finite population expression. The key assumptions required for consistency are that the expectation structure of the regression model (1) holds and that the first and second moments of the assumed conditional distribution $f(u_i \mid u_i^*, y_i, a_i^*, \mathbf{x}_i)$ are valid.

Such analytic expressions do not appear to be tractable, however, for the Student's t measurement error model (4). In particular, no closed form expressions appear to be available for the conditional expectations $z_{i,1}$ and $z_{i,2}$ under the Student's t measurement error model. We thus turn to fractional imputation to provide an approach which can be implemented for general parametric models.

4. ESTIMATION USING FRACTIONAL IMPUTATION

In this section, we consider how the parametric fractional imputation (PFI) method of Kim (2011) may be used to overcome the issues of intractability indicated in the previous section. The basic pseudo maximum likelihood method still provides the basis of our approach but now the conditional expectation in (8) is evaluated using imputed data for the unobserved values of u_i and a_i, rather than analytically. The imputed data are generated from the conditional distribution of these unobserved variables given the observed data. We show in the supplementary materials (section 1.2) that this conditional distribution may be expressed as:

$$f(u_i, a_i \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \boldsymbol{\psi}) = \begin{cases} f(u_i \mid u_i^*, y_i, a_i = 0, \mathbf{x}_i; \boldsymbol{\psi}), & a_i = 0, a_i^* = 0, \\ p_i(\boldsymbol{\psi}) f(u_i \mid u_i^*, y_i, a_i = 0, \mathbf{x}_i; \boldsymbol{\psi}), & a_i = 0, a_i^* = 1, \\ (1 - p_i(\boldsymbol{\psi})) f(u_i \mid u_i^*, y_i, a_i = 1, \mathbf{x}_i; \boldsymbol{\psi}), & a_i = 1, a_i^* = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

where

$$f(u_i \mid u_i^*, y_i, a_i, \mathbf{x}_i; \boldsymbol{\psi}) = \begin{cases} \frac{f(u_i|\mathbf{x}_i;\boldsymbol{\delta}) f(y_i|u_i,\mathbf{x}_i;\boldsymbol{\gamma}) g\left((u_i^* - u_i)/\tau\right)}{\int f(u_i|\mathbf{x}_i;\boldsymbol{\delta}) f(y_i|u_i,\mathbf{x}_i;\boldsymbol{\gamma}) g\left((u_i^* - u_i)/\tau\right) du_i}, & a_i = 0, \\ I(u_i = u_i^*) & a_i = 1, \end{cases} \tag{12}$$

and

$$\begin{aligned} p_i(\boldsymbol{\psi}) &\equiv Pr(a_i = 0 \mid u_i^*, y_i, a_i^* = 1, \mathbf{x}_i; \boldsymbol{\psi}) \\ &= \frac{p f(u_i^*, y_i \mid a_i = 0, \mathbf{x}_i; \boldsymbol{\psi})}{p f(u_i^*, y_i \mid a_i = 0, \mathbf{x}_i; \boldsymbol{\psi}) + (1 - p) f(u_i^*, y_i \mid a_i = 1, \mathbf{x}_i; \boldsymbol{\psi})}. \end{aligned} \tag{13}$$

The PFI approach allows the approximation of expectations of quantities such as $h(u_i, a_i)$ by importance sampling using a weighted summation of the form

$$E[h(u_i, a_i) \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \boldsymbol{\psi}] \approx \frac{\sum_{j=1}^M w_{ij}^* h(u_{iI}^{(j)}, a_{iI}^{(j)})}{\sum_{j=1}^M w_{ij}^*},$$

where $h(u,a)$ is an arbitrary integrable function,

$$w_{ij}^* = \frac{f(u_{iI}^{(j)}, a_{iI}^{(j)} \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \boldsymbol{\psi})}{q(u_{iI}^{(j)}, a_{iI}^{(j)} \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \widehat{\boldsymbol{\psi}})},$$

and $(u_{iI}^{(1)}, a_{iI}^{(1)}), ..., (u_{iI}^{(M)}, a_{iI}^{(M)})$ are imputed data generated from a distribution $q(u_i, a_i \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \widehat{\boldsymbol{\psi}})$ having the same support as $f(u_i, a_i \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \boldsymbol{\psi})$. We propose to take:

$$q(u_i, a_i \mid u_i^*, y_i, a_i^*, \mathbf{x}_i; \widehat{\boldsymbol{\psi}}) \equiv q(u_i, a_i \mid u_i^*, a_i^*, \mathbf{x}_i; \widehat{\boldsymbol{\delta}}) = \begin{cases} f(u_i \mid a_i = 0, \mathbf{x}_i; \widehat{\boldsymbol{\delta}}), & a_i = 0, a_i^* = 0, \\ pf(u_i \mid a_i = 0, \mathbf{x}_i; \widehat{\boldsymbol{\delta}}), & a_i = 0, a_i^* = 1, \\ (1-p)I(u_i = u_i^*), & a_i = 1, a_i^* = 1. \end{cases} \quad (14)$$

The PFI algorithm is defined iteratively, with $\widehat{\boldsymbol{\psi}}_{(t)} = (\widehat{\boldsymbol{\delta}}_{(t)}^\top, \widehat{\boldsymbol{\gamma}}_{(t)}^\top, \widehat{\tau}_{(t)}^2)^\top$ denoting the estimate of $\boldsymbol{\psi} = (\boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top, \tau^2)^\top$ at the t–th iteration for $t = 0, 1, \ldots$, where $\widehat{\boldsymbol{\delta}}_{(t)} = (\widehat{\boldsymbol{\delta}}_{u,(t)}^\top, \widehat{\delta}_{a,(t)}, \widehat{\sigma}_{u,(t)}^2)^\top$ and $\widehat{\boldsymbol{\gamma}}_{(t)} = (\widehat{\boldsymbol{\beta}}_{x,(t)}^\top, \widehat{\beta}_{u,(t)}, \widehat{\sigma}_{(t)}^2)^\top$ and the initial estimates for $t = 0$ are described later. The PFI algorithm then consists of the following steps.

Step 1 (Imputation step): For each $i \in A$,

- set $a_{iI}^{(j)} = 0$ for all $j = 1, ..., M$, if $a_i^* = 0$; take $a_{iI}^{(j)} = I(b_{ij} \leq 1 - p)$ for all $j = 1, ..., M$, if $a_i^* = 1$, where $b_{i1}, ..., b_{iM} \sim$ i.i.d $U(0,1)$;
- for all $j = 1, ..., M$, generate $u_{iI}^{(j)} \overset{indep}{\sim} f(u_i \mid a_i = 0, \mathbf{x}_i; \widehat{\boldsymbol{\delta}}_{(0)})$ if $a_{iI}^{(j)} = 0$, where $\widehat{\boldsymbol{\delta}}_{(0)}$ is a preliminary estimate of the vector of parameters $\boldsymbol{\delta}$, and set $u_{iI}^{(j)} = u_i^*$ if $a_{iI}^{(j)} = 1$.

Step 2 (Weighting step): Compute

$$\bar{\boldsymbol{S}}^*(\boldsymbol{\psi} \mid \widehat{\boldsymbol{\psi}}_{(t)}) = \sum_{i \in A} w_i \sum_{j=1}^{M} w_{ij,t}^* \boldsymbol{S}_{com}(\boldsymbol{\psi} \mid u_{iI}^{(j)}, a_{iI}^{(j)}, u_i^*, y_i, a_i^*, \mathbf{x}_i),$$

where $\boldsymbol{S}_{com}(\boldsymbol{\psi} \mid u_i, a_i, u_i^*, y_i, a_i^*, \mathbf{x}_i)$ is given in (9) and

$$w_{ij,t}^* = \begin{cases} \frac{f(y_i \mid u_{iI}^{(j)}, \mathbf{x}_{1i}; \widehat{\boldsymbol{\gamma}}_{(t)}) g((u_i^* - u_{iI}^{(j)})/\tau_{(t)}) f(u_{iI}^{(j)} \mid a_{iI}^{(j)} = 0, \mathbf{x}_{2i}; \widehat{\boldsymbol{\delta}}_{(t)}) / f(u_{iI}^{(j)} \mid a_{iI}^{(j)} = 0, \mathbf{x}_{2i}; \widehat{\boldsymbol{\delta}}_{(0)})}{\sum_{j=1}^{M} f(y_i \mid u_{iI}^{(j)}, \mathbf{x}_{1i}; \widehat{\boldsymbol{\gamma}}_{(t)}) g((u_i^* - u_{iI}^{(j)})/\tau_{(t)}) f(u_{iI}^{(j)} \mid a_{iI}^{(j)} = 0, \mathbf{x}_{2i}; \widehat{\boldsymbol{\delta}}_{(t)}) / f(u_{iI}^{(j)} \mid a_{iI}^{(j)} = 0, \mathbf{x}_{2i}; \widehat{\boldsymbol{\delta}}_{(0)})}, & a_{iI}^{(j)} = 0, \\ \frac{f(y_i \mid u_i^*, \mathbf{x}_{1i}; \widehat{\boldsymbol{\gamma}}_{(t)}) f(u_i^* \mid a_{iI}^{(j)} = 1, \mathbf{x}_{2i}; \widehat{\boldsymbol{\delta}}_{(t)})}{\sum_{j=1}^{M} f(y_i \mid u_i^*, \mathbf{x}_{1i}; \widehat{\boldsymbol{\gamma}}_{(t)}) f(u_i^* \mid a_{iI}^{(j)} = 1, \mathbf{x}_{2i}; \widehat{\boldsymbol{\delta}}_{(t)})}, & a_{iI}^{(j)} = 1. \end{cases}$$

Step 3 (Maximisation step): Update the current estimate of $\widehat{\boldsymbol{\psi}}$ as

$$\widehat{\boldsymbol{\psi}}_{(t+1)} \longleftarrow \text{ solution to } \bar{S}^*(\boldsymbol{\psi} \mid \widehat{\boldsymbol{\psi}}_{(t)}) = \mathbf{0}.$$

The procedure continues by iterating Steps 2 and 3 until a specified convergence criterion for the estimates is met. When this happens, w^*_{ij} and $\widehat{\boldsymbol{\psi}}$ are taken as the corresponding values of $w^*_{ij,t}$ and $\widehat{\boldsymbol{\psi}}_{(t)}$ obtained at the last iteration.

The components of the imputed pseudo score function $\bar{\boldsymbol{S}}^*(\boldsymbol{\psi} \mid \widehat{\boldsymbol{\psi}}_{(t)})$ in Step 2 corresponding to $\boldsymbol{\beta}_x$ and β_u are given by

$$\bar{\boldsymbol{S}}^*_{\boldsymbol{\beta}_x}(\boldsymbol{\psi}) = \sum_{i\in A} w_i \sum_{j=1}^{M} w^*_{ij,t} \frac{\mathbf{x}_{1i}}{\sigma^2}\Big\{(y_i - \mathbf{x}_{1i}^\top\boldsymbol{\beta}_x - u_{iI}^{(j)}\beta_u)(1-a_{iI}^{(j)}) + (y_i - \mathbf{x}_{1i}^\top\boldsymbol{\beta}_x - u_i^*\beta_u)a_{iI}^{(j)}a_i^*\Big\},$$

$$\bar{S}^*_{\beta_u}(\boldsymbol{\psi}) = \sum_{i\in A} w_i \sum_{j=1}^{M} w^*_{ij,t} \frac{1}{\sigma^2}\Big\{(y_i - \mathbf{x}_{1i}^\top\boldsymbol{\beta}_x - u_{iI}^{(j)}\beta_u)u_{iI}^{(j)}(1-a_{iI}^{(j)}) + (y_i - \mathbf{x}_{1i}^\top\boldsymbol{\beta}_x - u_i^*\beta_u)u_i^*a_{iI}^{(j)}a_i^*\Big\}.$$

Because $\sum_{j=1}^{M} w^*_{ij,t}\{(1-a_{iI}^{(j)}) + a_{iI}^{(j)}a_i^*\} = 1$, setting these expressions equal to zero implies that

$$\begin{aligned}
\sum_{i\in A} w_i\mathbf{x}_{1i}\mathbf{x}_{1i}^\top\widehat{\boldsymbol{\beta}}_x + \sum_{i\in A} w_i\mathbf{x}_{1i}\widehat{z}^*_{i,1t}\widehat{\beta}_u &= \sum_{i\in A} w_i\mathbf{x}_{1i}y_i \\
\sum_{i\in A} w_i\widehat{z}^*_{i,1t}\mathbf{x}_{1i}^\top\widehat{\boldsymbol{\beta}}_x + \quad \sum_{i\in A} w_i\widehat{z}^*_{i,2t}\widehat{\beta}_u &= \sum_{i\in A} w_i\widehat{z}^*_{i,1t}y_i,
\end{aligned}$$

where $\widehat{z}^*_{i,1}$ and $\widehat{z}^*_{i,2}$ are the values of $\widehat{z}^*_{i,1t} \equiv \sum_{j=1}^{M} w^*_{ij,t}\{u_{iI}^{(j)}(1-a_{iI}^{(j)}) + u_i^*a_{iI}^{(j)}a_i^*\}$ and $\widehat{z}^*_{i,2t} \equiv \sum_{j=1}^{M} w^*_{ij,t}\{u_{iI}^{(j)2}(1-a_{iI}^{(j)}) + u_i^{*2}a_{iI}^{(j)}a_i^*\}$ at convergence. Thus, the PFI estimates of $\boldsymbol{\beta_x}$ and β_u are equivalent to using (10) with the $z_{i,\ell}(\widehat{\boldsymbol{\psi}})$ terms replaced by $\widehat{z}^*_{i,\ell}$, $\ell = 1, 2$. This is true regardless of whether the measurement errors are normal or t_3 because the equations to be solved are based only on the score functions of the model for y given $\mathbf{x}, u$. The measurement error model enters into the estimation method only via the $w^*_{ij,t}$.

Initial estimates $\widehat{\boldsymbol{\psi}}_{(0)} = (\widehat{\boldsymbol{\delta}}_{(0)}^\top, \widehat{\boldsymbol{\gamma}}_{(0)}^\top, \widehat{\tau}^2_{(0)})^\top$ for the PFI algorithm are obtained as follows: $\widehat{\boldsymbol{\gamma}}_{(0)} = (\widehat{\boldsymbol{\beta}}_{x,(0)}^\top, \widehat{\beta}_{u,(0)}, \widehat{\sigma}^2_{(0)})^\top$ is computed by fitting model (2) by (survey) weighted least squares with u^* replacing u using just the $a^* = 1$ cases, so that $\widehat{\boldsymbol{\beta}}_{x,(0)}^\top$ and $\widehat{\beta}_{u,(0)}$ are the corresponding estimated regression coefficients and

$$\widehat{\sigma}^2_{(0)} = \Big\{\sum_{i\in A} w_i a_i^*\Big\}^{-1} \sum_{i\in A} w_i a_i^*(y_i - \mathbf{x}_{1i}^\top\widehat{\boldsymbol{\beta}}_{x,(0)} - u_i^*\widehat{\beta}_{u,(0)})^2.$$

The sub–vector $\widehat{\boldsymbol{\delta}}_{(0)} = (\widehat{\boldsymbol{\delta}}_{u,(0)}^\top, \widehat{\delta}_{a,(0)}, \widehat{\sigma}^2_{u,(0)})^\top$ is obtained in two steps. First, model (6) is fitted by weighted least squares using the $a^* = 0$ cases and again replacing u by u^*, giving

$$\widehat{\boldsymbol{\delta}}_{u,(0)}^\top = \Big\{\sum_{i\in A} w_i(1-a_i^*)\mathbf{x}_{2i}\mathbf{x}_{2i}^\top\Big\}^{-1} \sum_{i\in A} w_i(1-a_i^*)\mathbf{x}_{2i}u_i^*.$$

Second, $u^* - \mathbf{x}_2^\top \widehat{\boldsymbol{\delta}}_{u,(0)}$ is regressed on 1 using the $a^* = 1$ cases. The resulting estimates $\widehat{\delta}_{a,(0)}$ and $\widehat{\sigma}^2_{u,(0)}$ are computed by

$$\widehat{\delta}_{a,(0)} = \widehat{n}_1^{-1} \sum_{i \in A} w_i a_i^* (u_i^* - \mathbf{x}_{2i}^\top \widehat{\boldsymbol{\delta}}_{u,(0)}), \quad \widehat{\sigma}^2_{u,(0)} = \widehat{n}_1^{-1} \sum_{i \in A} w_i a_i^* (u_i^* - \mathbf{x}_{2i}^\top \widehat{\boldsymbol{\delta}}_{u,(0)} - \widehat{\delta}_{a,(0)})^2, \quad \text{and}$$

$$\widehat{\tau}^2_{(0)} = \widehat{n}_0^{-1} \sum_{i \in A} w_i (1 - a_i^*)(u_i^* - \mathbf{x}_{2i}^\top \widehat{\boldsymbol{\delta}}_{u,(0)})^2 - \widehat{n}_1^{-1} \sum_{i \in A} w_i a_i^* (u_i^* - \mathbf{x}_{2i}^\top \widehat{\boldsymbol{\delta}}_{u1,(0)})^2,$$

where $\widehat{n}_0 = \sum_{i \in A} w_i (1 - a_i^*)$, $\widehat{n}_1 = \sum_{i \in A} w_i a_i^*$ and $\widehat{\boldsymbol{\delta}}_{u1,(0)}$ is the estimated vector of coefficients in the regression with the $a^* = 1$ cases. Further details are given in the supplementary materials (section 2).

5. VARIANCE ESTIMATION

We now consider the estimation of the variance of the PML and PFI estimators of $\boldsymbol{\psi}$. Two variance estimation approaches that can be used are the linearization and replication methods. See, e.g., Wolter (2007) and Shao and Tu (1995). Both of these approaches can be formulated in terms of the observed information matrix of $\boldsymbol{\psi}$, namely

$$\bar{\mathbf{I}}_{obs}(\boldsymbol{\psi}) = \sum_{i \in A} w_i \bar{\mathbf{I}}_{obs,i}(\boldsymbol{\psi}) = \sum_{i \in A} w_i [\mathbf{I}_{1i}(\boldsymbol{\psi}) + \mathbf{I}_{2i}(\boldsymbol{\psi})], \tag{15}$$

where $\mathbf{I}_{1i}(\boldsymbol{\psi}) = -E[\dot{\mathbf{S}}_{com,i}(\boldsymbol{\psi}) \mid u_i^*, y_i, a_i^*, \mathbf{x}_i]$, $\mathbf{I}_{2i}(\boldsymbol{\psi}) = -E[(\boldsymbol{S}_{com,i}(\boldsymbol{\psi}) - \bar{\boldsymbol{S}}_{obs,i}(\boldsymbol{\psi}))^{\otimes 2} \mid u_i^*, y_i, a_i^*, \mathbf{x}_i]$ $\dot{\mathbf{S}}_{com,i}(\boldsymbol{\psi}) \equiv \dot{\mathbf{S}}_{com,i}(\boldsymbol{\psi} \mid u_i, a_i, u_i^*, y_i, a_i^*, \mathbf{x}_i) = \partial \boldsymbol{S}_{com,i}^\top(\boldsymbol{\psi})/\partial \boldsymbol{\psi}$, $\bar{\boldsymbol{S}}_{obs,i}(\boldsymbol{\psi})$ and $\boldsymbol{S}_{com,i}(\boldsymbol{\psi})$ are defined in (8) and (9) respectively, and $B^{\otimes 2} = BB^\top$. Decomposition (15) corresponds to the Louis (1982) formula.

The linearization variance estimator of $\widehat{\boldsymbol{\psi}}_{PML}$ can be computed, under simple random sampling, by the inverse of the observed information matrix in (15) evaluated at $\boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}_{PML}$. Expressions for this matrix are provided in Section 1.3 of the supplementary materials for the case of Gaussian measurement errors. For complex designs with first and second–order inclusion probabilities π_i and π_{ij}, respectively, the variance estimator can be obtained by the sandwich formula

$$\widehat{V}\left(\widehat{\boldsymbol{\psi}}\right) = \left\{\bar{\mathbf{I}}_{obs}(\widehat{\boldsymbol{\psi}})\right\}^{-1} \widehat{V}\{\bar{\boldsymbol{S}}_{obs}(\widehat{\boldsymbol{\psi}})\} \left\{\bar{\mathbf{I}}_{obs}(\widehat{\boldsymbol{\psi}})\right\}^{-1}, \tag{16}$$

where $\widehat{\boldsymbol{\psi}} = \widehat{\boldsymbol{\psi}}_{PML}$ and

$$\widehat{V}\{\bar{\boldsymbol{S}}_{obs}(\widehat{\boldsymbol{\psi}})\} = \sum_{i \in A} \sum_{j \in A} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} w_i w_j \bar{\boldsymbol{S}}_{obs,i}(\widehat{\boldsymbol{\psi}}) \bar{\boldsymbol{S}}_{obs,j}^\top(\widehat{\boldsymbol{\psi}}).$$

For PFI estimation, the linearization variance estimator can also be obtained from (16) by taking $\widehat{\boldsymbol{\psi}} = \widehat{\boldsymbol{\psi}}_{PFI}$ and replacing the matrices $\widehat{V}\{\bar{\boldsymbol{S}}_{obs}(\widehat{\boldsymbol{\psi}})\}$ and $\bar{\mathbf{I}}_{obs}(\widehat{\boldsymbol{\psi}})$ by their corresponding weighted–imputed versions

$$\widehat{V}^*\{\bar{\boldsymbol{S}}(\widehat{\boldsymbol{\psi}})\} = \sum_{i\in A}\sum_{k\in A}\frac{(\pi_{ik}-\pi_i\pi_k)}{\pi_{ik}} w_i w_k \bar{\boldsymbol{S}}_i^*(\widehat{\boldsymbol{\psi}})\bar{\boldsymbol{S}}_k^*(\widehat{\boldsymbol{\psi}})$$

and

$$\bar{\mathbf{I}}_{obs}^*(\widehat{\boldsymbol{\psi}}) = -\sum_{i\in A} w_i \sum_{j=1}^{M} w_{ij}^* \dot{\mathbf{S}}_{ij}^*(\widehat{\boldsymbol{\psi}}) - \sum_{i\in A} w_i \sum_{j=1}^{M} w_{ij}^* \left\{\boldsymbol{S}_{ij}^*(\widehat{\boldsymbol{\psi}}) - \bar{\boldsymbol{S}}_i^*(\widehat{\boldsymbol{\psi}})\right\}^{\otimes 2}, \tag{17}$$

with $\bar{\boldsymbol{S}}_i^*(\widehat{\boldsymbol{\psi}}) = \sum_{j=1}^{M} w_{ij}^* \boldsymbol{S}_{ij}^*(\widehat{\boldsymbol{\psi}})$ and $\boldsymbol{S}_{ij}^*(\widehat{\boldsymbol{\psi}}) = \boldsymbol{S}_{com}(\widehat{\boldsymbol{\psi}} \mid u_{iI}^{(j)}, a_{iI}^{(j)}, u_i^*, y_i, a_i^*, \mathbf{x}_i)$ and $\dot{\mathbf{S}}_{ij}^*(\widehat{\boldsymbol{\psi}}) = \dot{\mathbf{S}}_{com,i}(\widehat{\boldsymbol{\psi}} \mid u_{iI}^{(j)}, a_{iI}^{(j)}, u_i^*, y_i, a_i^*, \mathbf{x}_i)$. Section 2.2 of the supplementary materials details the computation of the matrix $\bar{\mathbf{I}}_{obs}^*(\widehat{\boldsymbol{\psi}})$.

In the replication method, replication weights $w_i^{(k)}$ are used in the estimation procedures. If we are only interested in estimating the variance of $\widehat{\boldsymbol{\psi}}$, then we can use $\widehat{\boldsymbol{\psi}}$ to obtain the replicated variance estimator given by

$$\widehat{V}(\widehat{\boldsymbol{\psi}}) = \sum_{k=1}^{L} c_k \left(\widehat{\boldsymbol{\psi}}^{(k)} - \widehat{\boldsymbol{\psi}}\right)^2,$$

where $\widehat{\boldsymbol{\psi}}^{(k)}$ is the estimate of $\boldsymbol{\psi}$ for the kth replicate which is obtained in the same as $\widehat{\boldsymbol{\psi}}$ by replacing $w_i^{(k)}$ for the w_i. For PFI estimation, the imputation step does not change. However, the weighting step and maximization step use the replication weights and the replicated version of $\widehat{\boldsymbol{\psi}}$. Because the EM algorithm is used for each replication, the replication method can be computationally unattractive in practice. To avoid this computation difficulty, one may devise a Newton–Raphson method instead of the EM algorithm and then apply the replication method to the Newton–Raphson method. The Newton–Raphson method can be implemented by

$$\widehat{\boldsymbol{\psi}}_{(t+1)} = \widehat{\boldsymbol{\psi}}_{(t)} + \left\{\bar{\mathbf{I}}_{obs}(\widehat{\boldsymbol{\psi}}_{(t)})\right\}^{-1} \bar{\boldsymbol{S}}(\widehat{\boldsymbol{\psi}}_{(t)}).$$

In each replicate k, we can use one-step approximation

$$\widehat{\boldsymbol{\psi}}^{(k)} = \widehat{\boldsymbol{\psi}} + \left\{\bar{\mathbf{I}}_{obs}(\widehat{\boldsymbol{\psi}})\right\}^{-1} \bar{\boldsymbol{S}}^{(k)}(\widehat{\boldsymbol{\psi}}),$$

where $\bar{\boldsymbol{S}}^{(k)}(\boldsymbol{\psi}) = \sum_{i\in A} w_i^{(k)} \bar{\boldsymbol{S}}_i^*(\boldsymbol{\psi})$.

If, on the other hand, the interest is on estimating the variance of other parameters, such as η obtained by solving the imputed estimating equation

$$\sum_{i\in A} w_i \sum_{j=1}^{M} w_{ij}^* U(\eta \mid u_{iI}^{(j)}, a_{iI}^{(j)}, u_i^*, y_i, a_i^*, \mathbf{x}_i) = 0,$$

then the replication method can be applied by considering as the k-th replicate of $\widehat{\eta}$ the solution to

$$\sum_{i\in A} w_i^{(k)} \sum_{j=1}^{M} w_{ij}^{*(k)} U(\eta \mid u_{iI}^{(j)}, a_{iI}^{(j)}, u_i^*, y_i, a_i^*, \mathbf{x}_i) = 0,$$

where

$$w_{ij}^{*(k)} = \begin{cases} \frac{f(y_i|u_{iI}^{(j)},\mathbf{x}_{1i};\widehat{\boldsymbol{\gamma}}^{(k)})g((u_i^*-u_{iI}^{(j)})/\widehat{\tau}^{(k)})f(u_{iI}^{(j)}|a_{iI}^{(j)}=0,\mathbf{x}_{2i};\widehat{\boldsymbol{\delta}}^{(k)})/f(u_{iI}^{(j)}|a_{iI}^{(j)}=0,\mathbf{x}_{2i};\widehat{\boldsymbol{\delta}}_{(0)})}{\sum_{j=1}^{M} f(y_i|u_{iI}^{(j)},\mathbf{x}_{1i};\widehat{\boldsymbol{\gamma}}^{(k)})g((u_i^*-u_{iI}^{(j)})/\widehat{\tau}^{(k)})f(u_{iI}^{(j)}|a_{iI}^{(j)}=0,\mathbf{x}_{2i};\widehat{\boldsymbol{\delta}}^{(k)})/f(u_{iI}^{(j)}|a_{iI}^{(j)}=0,\mathbf{x}_{2i};\widehat{\boldsymbol{\delta}}_{(0)})}, & a_{iI}^{(j)}=0, \\ \frac{f(y_i|u_i^*,\mathbf{x}_{1i};\widehat{\boldsymbol{\gamma}}^{(k)})f(u_i^*|a_{iI}^{(j)}=1,\mathbf{x}_{2i};\widehat{\boldsymbol{\delta}}^{(k)})}{\sum_{j=1}^{M} f(y_i|u_i^*,\mathbf{x}_{1i};\widehat{\boldsymbol{\gamma}}^{(k)})f(u_i^*|a_{iI}^{(j)}=1,\mathbf{x}_{2i};\widehat{\boldsymbol{\delta}}^{(k)})}, & a_{iI}^{(j)}=1. \end{cases}$$

and $\widehat{\boldsymbol{\delta}}^{(k)}$, $\widehat{\boldsymbol{\gamma}}^{(k)}$ and $\widehat{\tau}^{(k)}$ give the corresponding estimates of $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$ and τ^2 for the kth replicate.

6. SIMULATION EXPERIMENT

We now conduct a small simulation study to compare the properties of alternative methods for point estimation of $\boldsymbol{\beta}$, when the models considered earlier hold and where the measurement error follows either a normal or t_3 distribution. We also compare the properties of the linearized variance estimator of Section 5 for the proposed methods.

We created a finite population of N = 20,000 units with values of the variables generated as follows: $\mathbf{x}_{1i} = \mathbf{x}_{2i} = (x_{1i1}, x_{1i2})^\top$, where $x_{1i1} \sim$ Poisson($\mu = 3$), and $x_{1i2} \sim 1 + \mathrm{B}(1, 0.6)$; $a_i^* \sim B(1, p_i)$, where $p_i = 1/[1 + \exp\{-(0.3 + 0.4x_{1i1} - 1.3x_{1i2})\}]$; The resulting frequencies of $a_i^* = 0$ and $a_i^* = 1$ cases were 12,412 and 7,588, respectively. One set of a values was generated from the a^* according to the extended measurement error model with true value of p equal to zero. Based on these set, we generated $u_i = \delta_0 + \delta_1 x_{1i1} + \delta_2 x_{1i2} + \delta_a a_i + \sigma_u \epsilon_{iu}$, $\epsilon_{iu} \overset{indep}{\sim} N(0,1)$ and $y_i \sim \beta_0 + \beta_1 x_{1i1} + \beta_2 x_{1i2} + \beta_u u_i + e_i$, $e_i \overset{indep}{\sim} N(0, \sigma^2)$ for $i = 1, 2, ..., N$. Finally, the following two sets of u^* values was generated according to the Normal and t_3 measurement error models: $u_{i1}^* = (u_i + \tau\epsilon_{i1})(1 - a_i) + u_i a_i$, where $\epsilon_{i1} \overset{indep}{\sim} N(0,1)$, and $u_{i2}^* = (u_i + \tau\epsilon_{i2})(1 - a_i) + u_i a_i$, where $\epsilon_{i2} \overset{indep}{\sim} t_3$. The parameter values used to generated the values of these variables were $\delta_0 = 300$, $\delta_1 = 1$, $\delta_2 = -2$, $\delta_a = 2$, $\sigma_u = 3$, $\beta_0 = 50$, $\beta_1 = 2$, $\beta_2 = 3$, $\beta_u = 0.5$, $\sigma = 2$ and $\tau = 2$. We

selected $2,000$ independent simple random samples (without replacement) of sizes $n = 500$ from the population and computed various estimators from the data for each sample under both the normal and t_3 measurement error model cases. The PML and PFI methods were applied with $p = 0.0$ and $p = 0.2$. The PFI algorithm was implemented with $M = 200$ imputations at Step 1.

Table 1 presents the simulation mean, relative bias, standard deviation and square root of the mean square error of the unadjusted, PML and PFI (specified both with $p = 0.0$) estimators of the elements of the vector $\boldsymbol{\beta}$ for the normal measurement error case. The results of the PML and PFI estimators for mispecifying p by the value 0.2 are given in the supplementary materials (section 3). The unadjusted estimator displays non-negligible bias for each element of $\boldsymbol{\beta}$, as expected, and this bias is severe for β_u. The PML and PFI estimators are effective in correcting for this bias, with their simulation RMSEs being dominated by their simulation standard deviations for each parameter. Little difference is observed between the three adjusted estimators, other than very slight gains for the two methods based on the correct normality assumption versus the method based on the incorrect t_3 assumption. The table also contains results of an unadjusted analysis applied just to the accurate cases with $a_i^* = 1$. This approach also removes the bias but has variances substantially larger than the PML and PFI approaches. The corresponding results for the case where the measurements errors have a t_3 distribution are given in Table 2. Again, the unadjusted estimator shows serious bias. The PFI estimators are effective in correcting for this bias and generally perform similarly. There is little evidence of gain from the method based on the correct t_3 assumption.

The misspecification of p as 0.2 when it is actually 0 has little effect on the estimation of the main parameters of interest, β_1, β_2 and β_u, although it does lead to non-negligible bias in the estimation of β_0. Similarly, the misspecification of p as 0 when it is 0.2 has little effect on the estimation of β_1, β_2 and β_u, but does lead to non-negligible bias in the estimation of β_0.

Table 1: Monte Carlo properties of alternative estimators of the parameters of interest under normal measurement errors and $p = 0.0$

Method	Parameter	True	Mean	RB	SD	RMSE
All cases	β_0	50.00	79.46	58.9	7.89	30.50
	β_1	2.00	2.12	6.0	0.06	0.14
	β_2	3.00	2.76	-8.0	0.20	0.31
	β_u	0.50	0.40	-19.6	0.03	0.10
a^*=1 cases	β_0	50.00	49.05	-1.9	14.07	14.10
	β_1	2.00	1.99	-0.3	0.10	0.10
	β_2	3.00	3.00	0.0	0.32	0.32
	β_u	0.50	0.50	0.6	0.05	0.05
PML Normal ($p = 0.0$)	β_0	50.00	49.93	-0.1	10.62	10.62
	β_1	2.00	2.00	0.2	0.07	0.07
	β_2	3.00	3.01	0.2	0.21	0.21
	β_u	0.50	0.50	0.0	0.04	0.04
PFI Normal ($p = 0.0$)	β_0	50.00	47.44	-5.1	10.46	10.77
	β_1	2.00	1.99	-0.3	0.07	0.07
	β_2	3.00	3.03	0.9	0.21	0.21
	β_u	0.50	0.51	1.7	0.03	0.04
PFI t_3 ($p = 0.0$)	β_0	50.00	51.42	2.8	10.37	10.47
	β_1	2.00	2.01	0.5	0.07	0.07
	β_2	3.00	3.00	-0.1	0.21	0.21
	β_u	0.50	0.50	-1.0	0.03	0.03

Table 2: Monte Carlo properties of alternative estimators of the parameters of interest under t_3 measurement errors and $p = 0.0$

Method	Parameter	True	Mean	RB	SD	RMSE
All cases	β_0	50.00	112.28	124.6	10.49	63.15
	β_1	2.00	2.24	12.0	0.07	0.25
	β_2	3.00	2.48	-17.2	0.22	0.56
	β_u	0.50	0.29	-41.4	0.03	0.21
a^*=1 cases	β_0	50.00	49.05	-1.9	14.07	14.10
	β_1	2.00	1.99	-0.3	0.10	0.10
	β_2	3.00	3.00	0.0	0.32	0.32
	β_u	0.50	0.50	0.6	0.05	0.05
PML Normal ($p = 0.0$)	β_0	50.00	49.90	-0.2	10.77	10.77
	β_1	2.00	2.00	-0.2	0.07	0.07
	β_2	3.00	3.01	0.2	0.22	0.22
	β_u	0.50	0.50	0.1	0.04	0.04
PFI Normal ($p = 0.0$)	β_0	50.00	47.99	-4.0	10.60	10.79
	β_1	2.00	1.99	-0.6	0.07	0.07
	β_2	3.00	3.02	0.7	0.22	0.22
	β_u	0.50	0.51	1.3	0.04	0.04
PFI t_3 ($p = 0.0$)	β_0	50.00	48.89	-2.2	10.49	10.55
	β_1	2.00	1.99	-0.3	0.07	0.07
	β_2	3.00	3.02	0.5	0.21	0.22
	β_u	0.50	0.50	0.7	0.03	0.03

Table 3 displays the properties of the linearized variance estimator for the PML and PFI estimators for the Normal and t_3 measurement error cases with true value of p taken equal to zero. These tables give for each adjusted estimator its corresponding Monte Carlo variance, the mean and relative bias of the variance estimators, the z statistics for the test that the variance estimator is unbiased, the coverage of 95% confidence interval and the average width of these intervals. The results regarding the remaining model parameters are given in Section 3.1 of the supplementary materials. The z statistics were computed by a formula given in Kim (2004). Linearization variance estimation of the PML-N estimator of the parameters in the main regression yields negligible relatives biases and coverages near the 95% nominal levels under both Normal and t_3 measurements when p is correctly specified at the value 0. For the the case where the PML–N method is implemented with the value $p = 0.2$, the relative biases of the linearization variance estimator are still negligible, but the resulting confidence intervals show undercoverage (86–93%), possibly as a result of the bias of the point estimator due the misspecification of p. Results for the case when the true model generating the data has t_3 measurement errors are very similar to those in Table 3 although the degree of undercoverage of the confidence intervals was worse (75–91%) when p is misspecified as 0.2.

Table 3: Monte Carlo properties of the linearization variance estimator for the estimation of the main regression coefficients by the Normal pseudo maximum likelihood estimator and the Normal and t_3 parametric fractional imputation estimators (with $M = 200$), all evaluated with $p = 0.0$ and $p = 0.2$. True model generating the data has Normal measurement errors with $p = 0$

Estimation Method	θ	$Var(\widehat{\theta})$	$E[\widehat{V}(\widehat{\theta})]$	RB	z	Coverage	Width
PML Normal ($p = 0.0$)	β_0	112.71	117.02	3.8	1.20	94.8	42.3
	β_1	0.01	0.01	-0.5	-0.14	95.0	0.3
	β_2	0.04	0.04	-0.0	0.01	94.8	0.8
	β_u	0.00	0.00	3.7	1.17	94.9	0.1
PML Normal ($p = 0.2$)	β_0	160.12	155.99	-2.6	-0.76	85.7	48.6
	β_1	0.01	0.01	-3.2	-0.98	91.8	0.3
	β_2	0.05	0.05	-1.2	-0.34	93.4	0.9
	β_u	0.00	0.00	-2.7	-0.80	85.8	0.2
PFI Normal ($p = 0.0$)	β_0	109.38	112.52	2.9	0.89	94.2	41.4
	β_1	0.01	0.01	-1.2	-0.36	94.7	0.3
	β_2	0.04	0.04	-1.2	-0.35	94.9	0.8
	β_u	0.00	0.00	2.8	0.87	94.2	0.1
PFI Normal ($p = 0.2$)	β_0	115.34	121.96	5.7	1.77	93.6	43.1
	β_1	0.01	0.01	0.5	0.19	94.8	0.3
	β_2	0.04	0.05	0.7	0.24	94.7	0.8
	β_u	0.00	0.00	5.6	1.74	93.7	0.1
PFI t_3 ($p = 0.0$)	β_0	107.56	107.73	0.2	0.06	93.8	40.5
	β_1	0.01	0.01	-1.7	-0.54	94.5	0.3
	β_2	0.04	0.04	-1.3	-0.39	94.8	0.8
	β_u	0.00	0.00	0.1	0.04	93.8	0.1
PFI t_3 ($p = 0.2$)	β_0	108.97	111.82	2.6	0.82	94.8	41.3
	β_1	0.01	0.01	-1.3	-0.40	94.8	0.3
	β_2	0.04	0.04	0.6	0.20	94.8	0.8
	β_u	0.00	0.00	2.4	0.77	94.9	0.1

7. APPLICATION

In this section, we illustrate the performance of the proposed methods in a regression analysis motivated by a study described by Berthoud et al. (2009), using data from the British Household Panel Survey (BHPS), a stratified clustered sample of the UK resident population (Taylor 2006).

The study is concerned with how a given level of income may lead to different standards of living according to personal circumstances. For example, it may be expected that people living together will be able to live more efficiently from their joint income than if they live separately. Less obvious is the effect of age which is the focus of the study. There is evidence in some settings that older people may experience less hardship than might be expected given their income (Berthoud et al. 2009). This may be studied by fitting a regression model with a measure of living standards as the dependent variable and with a measure of income, age and other variables as covariates.

Our analysis is at the individual level, taking earnings as a proxy for income and restricting attention to employed survey respondents aged 30-65 years old. We base the dependent variable y on the overall deprivation index considered by Berthoud et al. (2009). Their index, denoted DI, was obtained by computing an average of four sub–indexes, two representing daily living deprivation, one reflecting lack of possession of consumer durables and the last one measuring deprivation due to financial strain. The index DI was standardized to have mean zero and variance one with higher values indicating more deprivation. To improve normality of the overall deprivation index, we apply the transformation $-(DI+5)^{-1.75}$ and y is taken as this transformed variable standardized to have mean zero and variance one.

The covariate u^* measured with error is the natural logarithm of the gross pay (in pounds) reported at last payment. The accuracy variable a is taken as the indicator that the respondent's last payslip was seen by the interviewer. The existence of measurement error in pay has already been referred to in the Introduction. Evidence that measurement error in the logarithm of pay can be non-normal has been given by (Rodgers et al. 1993) and they suggest that 'the departure from normality may be a reflection primarily of a small number of outliers' (pp. 1215–6). We propose to employ our t_3 model for the measurement error as well as the normal model in order to allow for the possible effect of such outliers, in line with the discussion in (Lange et al. 1989).

The first additional covariate in the vector $\mathbf{x}_1$ in model (2) is age at the date of interview (years),

rescaled by the transformation (age - 30)/10. There is slight evidence of nonlinear dependence on age in this model but we choose to specify the dependence on age here as linear to simplify the interpretation. Further covariates, reflecting living arrangements are added as controls for studying the joint effect of age and earnings on y. These further covariates consist of number of children in household, whether living with a spouse, household size, whether 'head of household', whether owning or renting accommodation. The vector $\mathbf{x}_2$ in model (6) is taken to include $\mathbf{x}_1$ as well as the following potential predictors of earnings: a quadratic term in age, indicators of occupation (professional, managerial & technical, skilled, unskilled & armed forces), an indicator of academic qualifications, indicators of size of workplace ($<$ 25, 25–99, 100–499, 500+, don't know).

For simplicity, we undertake a cross-sectional analysis of a single wave of BHPS data (Wave 6) collected between 1996 and 1997. We consider respondents who are employed, aged 30-65 and without missing values on the variables considered, giving a dataset of 2,262 observations, of which 946 are from individuals who had consulted their latest payslip and the remaining 1,316 cases were from those who had not. We approximate the BHPS sampling design by a stratified design with independent sampling of the primary sampling units (PSUs) within each stratum, taken to be region, defining $H = 11$ strata. The weight variable w is the BHPS cross–sectional respondent weight.

We estimate the parameters in the regression model (2) using five different approaches. Firstly, as a reference for comparison, we obtain least squares estimates, weighted by the w, both using all 2,262 cases and using just the 946 "accurate" ($a^* = 1$) cases. Secondly, we apply the proposed pseudo–maximum likelihood estimation approach assuming Normal measurement errors (PML-N) and the parametric fractional imputation method with Normal (PFI-N) and t_3 errors (PFI–t_3). These three methods are applied to produce estimates of all parameters in the vector $\boldsymbol{\psi}$. The PFI estimators are implemented with $M = 100$ imputations in Step 1 of the algorithm. Initial estimates for the model parameters are computed using the svyglm() function of the Survey package in R (R Core Team 2015). The estimated regression coefficients correspond to the those presented in Section 4. However, the estimated variances $\widehat{\sigma}^2_{u,(0)}$, $\widehat{\sigma}^2_{(0)}$ and $\widehat{\tau}^2_{(0)}$ obtained from the package, namely 0.9945, 0.8118 and 0.6188 respectively, are slightly different from the ones computed by the expressions given in Section 4 0.9935, 0.8109 and 0.6186. Since the difference seems quite small,

the PFI algorithm is for simplicity started with the former set.

The standard errors of the PML and PFI estimates are obtained from the estimated variance-covariance matrix for $\widehat{\boldsymbol{\psi}}$ of each method. These matrices are computed following the sandwich formula theory described in Section 5. Because of the ultimate cluster approximations for the BHPS sample design, we use in the PML case

$$\widehat{V}\{\bar{\boldsymbol{S}}_{obs}(\widehat{\boldsymbol{\psi}})\} = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left\{ \sum_{k=1}^{n_h} (\boldsymbol{z}_{hk} - \bar{\boldsymbol{z}}_h)(\boldsymbol{z}_{hk} - \bar{\boldsymbol{z}}_h)^\top \right\},$$

where n_h is the number of PSUs in the h–th stratum, A_{hk} is the index set of individuals in the k–th PSU of the h–th stratum, $\boldsymbol{z}_{hk} = \sum_{i \in A_{hk}} w_i \bar{\boldsymbol{S}}_{obs,i}(\widehat{\boldsymbol{\psi}})$ and $\bar{\boldsymbol{z}}_h = n_h^{-1} \sum_{k=1}^{n_h} \boldsymbol{z}_{hk}$. The corresponding variance for the PFI method is estimated similarly using $\boldsymbol{z}_{hk} = \sum_{i \in A_{hk}} w_i \bar{\boldsymbol{S}}_i^*(\widehat{\boldsymbol{\psi}})$. The variances for the 'All cases' and '$a^* = 1$ cases' estimates follow also sandwich–based formulae and these are computed directly using the svyglm() function in the R Survey package.

Table 4 presents the estimated coefficients of age and ln(gross pay) in the regression model (2) of interest, together with their corresponding standard errors for the five estimation methods. Estimates of the remaining coefficients in $\boldsymbol{\beta}_x$ are given in the supplementary materials (Section 4). We observe negative coefficients of pay (with high levels of statistical significance) in Table 4, as expected, i.e. deprivation decreases on average as pay increases. The size of the point estimate of the coefficient is smaller for the first method, which ignores measurement error, than for all the other methods which seek to control for measurement error, in line with what might be expected from attenuation bias (Carroll et al. 2006, chap. 3).

The point estimates of the coefficient of age are also negative suggesting that older respondents report on average less deprivation than younger respondents for given levels of earnings and the other control variables. A conventional interpretation of the 't value' $0.083/0.031 = 2.7$ of age for the first approach, which ignores measurement error, would be that this coefficient differs significantly from zero. However, the second approach using just the $a^* = 1$ cases to control for measurement error leads to a t-value of $0.072/0.046 = 1.6$, which would conventionally be interpreted as age no longer having a significant effect. This appears, however, to be primarily the effect of the loss of efficiency of this point estimate, arising from the reduction in the number of observations from 2,262 to 946. The estimated coefficients from the three proposed methods all have t-values similar

to that for the 'all cases' approach, illustrating how the proposed methods may make efficient use of all the observations, while controlling for measurement error. We obtain similar findings on repeating the analysis with different choices of variables in the vector $\mathbf{x}_1$.

Table 4: Point estimates with standard errors (in parentheses) for the coefficients of age and pay in the regression model of interest

Coefficient	All cases	$a = 1$ cases	PML-N	PFI-N	PFI-t_3
Age	-0.083	-0.072	-0.088	-0.088	-0.089
	(0.031)	(0.046)	(0.031)	(0.031)	(0.031)
Pay (logged)	-0.109	-0.150	-0.142	-0.143	-0.142
	(0.017)	(0.034)	(0.024)	(0.023)	(0.023)

We next consider the sensitivity of these results to alternative departures from assumptions. We propose two approaches to assessing such sensitivity. First, our model enables us to assess the possibility that the observation $a_i^* = 1$ does not guarantee accuracy by specifying values of p greater than 0. For our application, there is evidence in the validation study referred to in Da Silva and Skinner (2014) that this may be the case, although it suggests that it is implausible that p exceeds 0.2. Table 5 presents comparable results for values of p of 0.1 and 0.2 as well as 0. The principal conclusion from Table 4 that the coefficient of age differs significantly from zero remains unaffected by the variation in p. The estimated coefficients using PFI are less sensitive to variation in p than the PML estimates and the estimated coefficients of age less sensitive than those for pay.

Our second proposed approach to assessing the sensitivity of estimates to departures from assumptions is to simulate K repeated sets of observations (y_i, u_i^*) from our fitted model with a given value of p with the $(a_i^*, \mathbf{x}_i)$ fixed at their observed values. This exercise is then repeated with various modifications to the model. The PML and PFI estimates are obtained for each set of observations and the difference in the estimates assessed. Unless they are modified, the random terms in the model (e_i, u_i, ϵ_i) are kept the same to improve the efficiency of comparison between the estimates. As in the case of Table 5 across different values of p, we have found the sensitivity of the standard errors of less concern than the possibility of biased estimation of coefficients arising from the departure from assumptions. As evidence of such possible bias we have found it sufficient

Table 5: Point estimates with standard errors (in parentheses) for the coefficients of age and pay in the regression model of interest by alternative estimation methods for various values of p

	Adjusted	p		
Coefficient	estimator	0.00	0.10	0.20
Age	PML-N	-0.088	-0.094	-0.097
		(0.031)	(0.031)	(0.031)
	PFI–N	-0.088	-0.088	-0.090
		(0.031)	(0.031)	(0.032)
	PFI–t_3	-0.089	-0.090	-0.090
		(0.031)	(0.031)	(0.032)
Pay (logged)	PML-N	-0.142	-0.172	-0.185
		(0.023)	(0.028)	(0.028)
	PFI–N	-0.145	-0.145	-0.154
		(0.023)	(0.024)	(0.027)
	PFI–t_3	-0.145	-0.147	-0.150
		(0.023)	(0.023)	(0.025)

to take $K = 100$ and to compute the mean value of the estimators for the data simulated from the assumed and modified models.

We apply this approach first to assessing sensitivity to distributional assumptions. We fixed $p = 0$ and simulated sets of observations from the fitted model with $e_i \sim N(0, \widehat{\sigma}^2)$ in (2) and then, as a modified model, with e_i generated as $\widehat{\sigma}$ times a t_3 random variable truncated to the interval $(-10, 10)$ and standardised to have variance 1 (detailed further in the supplementary materials, Section 5.2). The average values of the estimated coefficients of age and pay were found to differ little between the normal and the t_3 simulated values. For example, the average PML estimated coefficients of (age, pay) were $(-0.086, -0.143)$ for normal and $(-0.091, -0.145)$ for t_3. The true values were $(-0.088, -0.142)$. We noted after expression (10) how the PML estimation approach is primarily based upon first and second moment assumptions and so this lack of sensitivity might

have been anticipated. The findings for the PFI methods were similar.

As a further departure from distributional assumptions, we consider the effect of replacing u_i in the data generation process by $u_i^{1.1}$ (supplementary materials, Section 5.3). This implies a departure from the symmetric normal distributional assumption in model (6). Moreover, it also introduces a dependence in the conditional variance of u_i given $(a_i, \mathbf{x}_i)$ on a_i. We simulated 100 sets of observations as in the previous sensitivity analysis. We found there was little impact of this change on the estimated coefficient of age, with the average PML estimate of this coefficient remaining at -0.086 (to this number of decimal places). There was an effect on the estimated coefficient of pay, however, with the average PML estimate of this coefficient changing from -0.143 to -0.190. Results for the PFI estimates were similar. We consider the likely source of effect on the pay coefficient to be the dependence of the conditional variance of u_i given $(a_i, \mathbf{x}_i)$ on a_i in the auxiliary model (6). The distributional assumption seems a less likely source given our moment-based approach.

Given this potential importance of the auxiliary model (6) to the estimates, we now explore it further. We note first that estimates of the coefficients $\boldsymbol{\delta}_u$ of the covariate vector $\mathbf{x}_2$ in this model correspond to what we may expect in a model predicting earnings and they appear in Table 7 in the supplementary materials. Estimates of the parameter δ_a are presented in Table 6. This reflects the dependence of the conditional expectation of u_i given $(a_i, \mathbf{x}_i)$ on a_i in the model. We see here evidence of a departure from the assumption in Da Silva and Skinner (2014) that u_i and a_i are conditionally independent given $\mathbf{x}_i$ (taken to be $\mathbf{x}_{2i}$ here). The positive estimated value of δ_a in Table 6 suggests that respondents referring to their payslips have higher levels of earnings on average, even controlling for $\mathbf{x}_{2i}$. It is possible that the paradata variable a_i is acting as a proxy for some unobserved factors which are influencing the earnings variable u_i. Alternatively, it is even possible that the positive estimated value of δ_a is arising from a direct dependence of a_i on u_i, in the sense that whether an individual chooses to refer to their payslip to confirm their earnings may depend on the nature of the earnings. If we allow p to increase from 0 to 0.2, we obtain similar findings for the PFI-N and PFI-t_3 methods, but, curiously, the estimate of δ_a is attenuated and falls below two standard errors for $p \geq 0.5$ for the PML-N method (supplementary materials, section 4).

Features of the measurement error process and the parameter δ_a can be observed graphically by

Table 6: Estimates of δ_a with standard errors (in parentheses)

All cases	PML-N	PFI-N	PFI-t_3
0.127	0.133	0.131	0.194
(0.055)	(0.054)	(0.055)	(0.054)

considering the relation between the residuals $u_i^* - \mathbf{x}_{2i}^\top \widehat{\boldsymbol{\delta}}_u$ versus a_i. Box plots of these residuals for $a_i^* = 0$ and $a_i^* = 1$, where $\boldsymbol{\delta}_u$ is estimated by PML, are presented in Figure 1. Under our assumption of zero-mean measurement error, the difference in means between the two box plots ($a_i^* = 1$ versus $a_i^* = 0$) represents an estimate of the parameter δ_a. The approximately symmetric distributions in Figure 1 are consistent with our assumption of symmetric measurement error distributions and normal error in model (6). The difference in variances of the boxplot for $a_i^* = 0$ to the one for $a_i^* = 1$ represents an estimate of the parameter τ^2 times the variance of the measurement error term. Estimates of the variances σ_u^2 and τ^2 (and also of σ^2) are presented in Table 7. If we allow p to increase from 0 to 0.2, we obtain similar findings for σ_u^2 for the PML and PFI methods (supplementary materials, section 4). The estimates of τ^2 and σ^2 show some sensitivity to the choice of p with the estimate of τ^2 tending to increase and the estimate of σ^2 tending to decrease as p increases, for each of the PML-N, PFI-N methods.

Table 7: Estimates of model variances with standard errors (in parentheses)

parameter	PML-N	PFI-N	PFI-t_3
σ^2	0.888	0.887	0.887
	(0.033)	(0.032)	(0.033)
σ_u^2	0.872	0.850	0.828
	(0.179)	(0.054)	(0.055)
τ^2	0.640	0.700	0.320
	(0.197)	(0.096)	(0.045)

We next consider departures from the assumption that y_i and a_i are conditionally independent given $(u_i, \mathbf{x}_i)$, which might arise if each are dependent upon the same unobserved variable. To modify the model (2) used to generate y_i, we include an additional term $a_i\beta_a$, with β_a chosen so

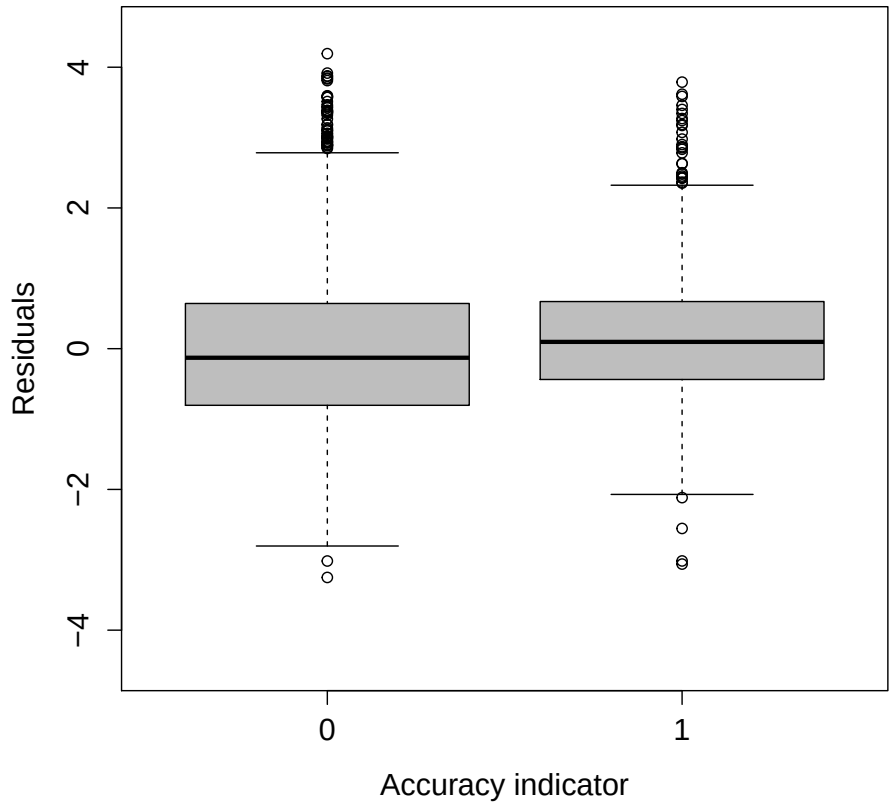

Figure 1: Boxplots of the residuals $u_i^* - \mathbf{x}_{2i}^\top \widehat{\boldsymbol{\delta}}_u$ for both accuracy groups. The estimated regression coefficient $\widehat{\boldsymbol{\delta}}_u$ used was taken from PML estimates of $\boldsymbol{\delta} = (\boldsymbol{\delta}_u^\top, \delta_a, \sigma_u^2)^\top$ given in the supplementary materials.

that the partial correlation between y_i and a_i conditional on $(u_i, \mathbf{x}_i)$ is 0.1 and where the methods and models used for estimation remained the same (supplementary materials, section 5.1). We simulated 100 sets of observations as in the previous sensitivity analysis. We found there was little impact of this change on the estimated coefficient of age, with the average PML estimate of this coefficient changing from -0.086 to -0.084. There was a little attenuation of the estimated coefficient of pay, with the average PML estimate of this coefficient changing from -0.136 to -0.123. The results for the PFI estimates were very similar.

In summary, the results of our sensitivity analyses suggest some robustness in the estimation of

the coefficient of age and in main conclusions drawn from Table 5. We noticed some sensitivity in assumptions to the estimation of the coefficient of pay, the variable with measurement error, with possibly the most important assumption relating to our assumed constant conditional variance of u_i given $(a_i, \mathbf{x}_i)$ in model (6).

In Tables 4 and 6 we found no strong differences between the estimates obtained from the three methods PML-N, PFI-N and PFI-t_3. We did, however, observe a somewhat larger estimated value of δ_a in Table 6 for PFI-t_3 relative to the PML-N and PFI-N methods. It is possible in this case, that the PFI-t_3 method protects better against measurement error outliers in the distribution of u_i^* given u_i when estimating δ_a, supporting the value of having the PFI-t_3 method to provide an alternative robust estimation approach. We did find some cases when the PFI-t_3 estimate performed better than the other approaches, in terms of the estimate of β_u being closer to the $a^* = 1$ cases estimate than the PML-N and PFI-N methods but this effect was not consistent across a range of models with different choices of $\mathbf{x}_1$. It is not easy in plots of residuals, such as that for cases with $a_i^* = 0$ in Figure 1 which represents a convolution of distributions, to detect whether there are more measurement error outliers than expected under normality. It is possible in our case that the similarity of findings for PML-N, PFI-N and PFI-t_3 may be a reflection of the absence of such outliers.

8. DISCUSSION

In this paper we have considered the classical problem of correcting for measurement error in one of the covariates of a regression model. We have shown how binary paradata, indicating the presence of the measurement error, can be used to correct for bias. In this final section, we comment on the potential broader uses of our approach.

First, the approach needs extending to richer paradata variables than the binary case discussed in detail here, including ordinal, continuous and multivariate cases. This raises questions about how to extend models (1) and (6). Extension to multivariate measurement error is needed also. The extension of univariate imputation for measurement error to the multivariate case has already been considered in the multiple imputation literature (He and Zaslavsky 2009), but the extension of models in (1) and (6) to represent the relationship between multivariate measurement error and

paradata (possibly also multivariate) needs further research.

We have assumed so far, at least implicitly, that the measurement error correction takes place after the survey has been completed. In fact, much of the interest in paradata among survey researchers relates to stages of the survey process prior to this point.

Much of the use of paradata on measurement error so far has been to assist the improvement of the survey measurement process, such as computer-assisted interviewing systems, with the general objective of reducing measurement error. Yan and Olson (2013) review a series of studies of this kind. They note that, although paradata may be used as an indicator of measurement error, it often requires auxiliary variables to be interpreted suitably. For example, the time taken to respond to a question may reflect measurement error, but it may be desirable to take account of the length of the question if this variable is to provide a useful indicator of measurement error. Our approach might be seen as an extension of such an attempt to control for auxiliary variables. Paradata is essentially observational and so selection effects will invariably be an issue. Attempts to design measuring instruments so that the paradata ostensibly suggest less measurement error may be misleading in the presence of the kinds of selection considered in this paper. The kinds of adjustment method we have considered may then be useful.

Another main area of use of paradata in surveys is in adaptive design during data collection. As discussed by Schouten and Calinescu (2013), the aim may be to define a measure of survey quality, embracing both nonresponse and measurement error, and to identify ways of maximising quality under budget constraints at stages during the data collection process, making use of paradata collected up to these stages. Again, selection effects could lead to misleading estimates of the contribution of measurement error to the quality measure and adjustment methods of the kind considered in this paper might lead to a more appropriate measure of quality.

SUPPLEMENTARY MATERIALS

The supplementary materials provide further information for the pseudo–maximum likelihood estimator, expressions for pseudo–score functions used by the parametric fractional imputed estimator, variance estimation and additional tables for the application.

REFERENCES

Barrett, K., Sloan, M., and Wright, D. (2006), "Interviewer Perception and Interview Quality," in *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp. 4026–4033.

Battistin, E., Miniaci, R., and Weber, G. (2003), "What Do We Learn from Recall Consumption Data?" *Journal of Human Resources*, 38, 354–386.

Berthoud, R., Blekesaune, M., and Hancock, R. (2009), "Ageing, income and living standards: evidence from the British Household Panel Survey," *Ageing and Society*, 29, 1105–1122.

Binder, D. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279–292.

Blackwell, M., Honaker, J., and King, G. (2015), "A Unified Approach to Measurement Error and Missing Data: Overview and Applications," *Sociological Methods & Research*, (To appear).

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement error in nonlinear models: a modern perspective (2nd ed.)*, Boca Raton, FL: Chapman and Hall.

Cole, S. R., Chu, H., and Greenland, S. (2006), "Multiple-imputation for measurement-error correction," *International Journal of Epidemiology*, 35, 1074–1081.

Da Silva, D. N. and Skinner, C. (2014), "The use of accuracy indicators to correct for survey measurement error," *Journal of the Royal Statistical Society*, Series C, 63, 303–319.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Godambe, V. P. and Thompson, M. E. (1986), "Parameters of Superpopulation and Survey Population: Their Relationships and Estimation," *International Statistical Review*, 54, 127–138.

Guo, Y., Little, R. J., and McConnell, D. S. (2012), "On using summary statistics from an external calibration sample to correct for covariate measurement error." *Epidemiology (Cambridge, Mass.)*, 23, 165–174.

He, Y. and Zaslavsky, A. M. (2009), "Combining Information from Cancer Registry and Medical Records Data to Improve Analyses of Adjuvant Cancer Therapies," *Biometrics*, 65, 946–952.

Kim, J. K. (2004), "Finite sample properties of multiple imputation estimators," *Ann. Statist.*, 32, 766–783.

— (2011), "Parametric Fractional Imputation for Missing Data Analysis," *Biometrika*, 98, 119–132.

Kim, J. K. and Shao, J. (2013), *Statistical methods for handling incomplete data*, Boca Raton, FL: CRC Press.

Kreuter, F. (2013), *Improving Surveys with Paradata: Analytic Uses of Process Information*, Hoboken, NJ: Wiley.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Modeling Using the t-distribution," *Journal of the American Statistical Association*, 84, 881–896.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data (2nd ed.)*, Chichester: Wiley.

Louis, T. A. (1982), "Finding the Observed Information Matrix when Using the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 44, 226–233.

Mathiowetz, N. A. (1998), "Respondent Expressions of Uncertainty: Data Source for Imputation," *Public Opinion Quarterly*, 62, 47–56.

Moore, J. C., Stinson, L. L., and Welniak, Edward J., J. (2000), "Income Measurement Error in Surveys: A Review," *Journal of Official Statistics*, 16, 331–361.

Olson, K. and Parkhurst, B. (2013), "Collecting Paradata for Measurement Error Evaluations," in *Improving Surveys with Paradata: Analytic Uses of Process Information*, ed. Kreuter, F., Hoboken, NJ: Wiley, pp. 43–72.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rodgers, W. L., Brown, C., and Duncan, G. J. (1993), "Errors in Survey Reports of Earnings, Hours Worked, and Hourly Wages," *Journal of the American Statistical Association*, 88, 1208–1218.

Schouten, B. and Calinescu, M. (2013), "Paradata as Input to Monitoring Representativeness and Measurement Profiles: A Case Study of the Dutch Labour Force Survey," in *Improving Surveys with Paradata: Analytic Uses of Process Information*, ed. Kreuter, F., Hoboken, NJ: Wiley, pp. 231–258.

Shao, J. and Tu, D. (1995), *The jackknife and bootstrap*, Berlin; New York: Springer-Verlag Inc.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989), *Analysis of Complex Surveys*, Chichester: Wiley.

Taylor, M. F. (2006), *British Household Panel Survey User Manual, Volume A: Introduction, Technical Report and Appendices*, Colchester: University of Essex.

Wolter, K. M. (2007), *Introduction to variance estimation (2nd ed.)*, Berlin; New York: Springer-Verlag Inc.

Yan, T. and Olson, K. (2013), "Analyzing Paradata to Investigate Measurement Error," in *Improving Surveys with Paradata: Analytic Uses of Process Information*, ed. Kreuter, F., Hoboken, NJ: Wiley, pp. 73–95.