Damião Nóbrega Da Silva and Chris Skinner

The use of accuracy indicators to correct for survey measurement error

Article (Accepted version) (Refereed)

The Use of Accuracy Indicators to Correct for Survey Measurement Error

Damião N. Da Silva[1] and Chris Skinner[2]

[1]Universidade Federal do Rio Grande do Norte, Natal, Brazil, email: damiao@ccet.ufrn.br

[2]The London School of Economics and Political Science, London, UK, email: C.J.Skinner@lse.ac.uk

July 25, 2014

Abstract

An accuracy indicator is an observed variable which is related to the size of measurement error. Basic and extended models are introduced to represent the properties of a binary accuracy indicator. Under specified assumptions, it is shown that an accuracy indicator can identify a measurement error model. An approach to estimating a distribution function is presented together with methodology for variance estimation. The approach is applied to data on earnings from the British Household Panel Survey, where the accuracy indicator is whether or not a pay slip is observed. A validation study provides justification for the modelling assumptions.

Keywords: accuracy indicator; finite population distribution function; measurement error; pseudo maximum likelihood

1 Introduction

Measurement error is widely recognized as an important potential source of estimation bias in surveys. Correcting for such bias requires information about the measurement error process and this is often difficult to obtain. On some occasions it may be possible to obtain repeated measurements of the variable of interest through test-retest reinterviews or to calibrate the measuring instrument by linking survey data to some accurate external data source (Biemer and Lyberg, 2003, Sect. 8.4), but such exercises are relatively unusual. In other circumstances, it may be possible to record a variable which meets the requirements of instrumental variable estimation (Wansbeek and Meijer, 2000), that is it is related to the true value of the variable of interest but can be assumed to be independent of the measurement error. However, such variables may also often not be available. In this paper we explore the use of an alternative source of information about measurement error.

We consider the use of what we call an accuracy indicator. This is a variable which is related to the measurement error, but not directly to the true value of the variable. It is like an instrumental variable in being an observed auxiliary variable, but the underlying assumptions are quite different. We shall focus on the case when the accuracy indicator is binary, so that it indicates whether a measurement is accurate or inaccurate. Such an indicator may be obtained in various ways, including by self-report, where a respondent expresses uncertainty about the accuracy of their response to a factual question (Mathiowetz, 1998), or by the judgement of an interviewer. This paper is motivated by an application where the variable of interest is earnings and the accuracy indicator is whether or not the respondent refers to a recent pay slip when responding to the question about earnings. Answers which refer to a pay slip are treated as accurate. Otherwise, the answer is assumed to be subject to measurement error. The aim of the paper is to develop an approach to making inference about the population distribution function of the underlying true earnings variable in the presence of such measurement error. See Eltinge (1999) for an approach to the problem of distribution function estimation when estimates of the measurement error variance are available and for some references to related approaches.

The reporting of an accuracy indicator differs from accuracy verification (e.g. Begg and Greenes, 1983) where, in addition to the survey measure, an accurate measurement is taken, usually just for a subsample. In this case, both inaccurate and accurate measurements are recorded and the difference between the two provides an indicator of the accuracy of the first measurement. In our setting, however, we do not have two measurements, even for a subsample. We only have one measurement recorded along with an indicator of its accuracy.

The literature on survey measurement error in earnings and its potential biasing effect on data analysis is large. See e.g. Rodgers et al. (1993), Moore et al. (2000) and Bound et al. (2001). The literature includes a variety of extensions to the classical additive measurement error model. One extension, of relevance to our setting with binary accuracy indicators, is a mixture model where respondents either report accurately with a specified non-zero probability or they report inaccurately with responses subject to a measurement error process (e.g. Kapteyn and Ypma, 2007). Inference for such binary mixture models has been considered by Horowitz and Manski (1995) and subsequent authors, including Kreider and Pepper (2007). They show that, even if some upper bound is known for the population proportion of inaccurate responders, it is not possible to identify parameters in the usual sense. Instead they show that only a set of parameters can be identified and propose associated inference methods. In this literature it is generally assumed that membership of the subpopulation of accurate responders is unobserved. In our setting, our initial assumption is that the accuracy indicator enables us to observe which element of the mixture a respondent belongs to. Dominitz and Sherman (2004, 2006) also consider models for an accuracy indicator, which we refer to in Section 3.

The paper is structured as follows. The data sources used in the application and to help justify the modelling assumptions are described in Section 2. The models and estimation methods are introduced in Sections 3, 4 and 5. The application is described in Section 6 with some brief additional discussion in Section 7. Some further details of the more technical arguments are provided in the Appendix and in Supplementary Materials.

2 Data Sources

We focus on the variable earnings, as a widely used continuous variable in survey research. We consider estimating the distribution function of this variable, since measurement error will typically lead to bias in this estimation problem even if the error has zero mean (Fuller, 1995). We use data from the British Household Panel Survey (BHPS), chosen because its approach to measuring earnings is fairly standard and there is also a related validation study which we use to motivate our measurement error model. We make no use of the panel feature of the BHPS. We now proceed to describe the BHPS and the associated validation study.

2.1 British Household Panel Survey

The BHPS has undertaken annual waves of data collection on around 5,000 households in Great Britain since 1991. We use data from Wave 12, conducted in 2002-3, to match the timing of the validation study. The sample for the survey, drawn at Wave 1, was obtained by multistage sampling with postcode sectors as primary sampling units. A sample of 250 of these sectors was selected with probability proportional to a measure of size using systematic sampling from a list of sectors ordered by region and some socio-economic variables, implying an implicit stratification of the first stage of sampling. From each selected postcode sector, around 20-30 addresses were selected, also by systematic sampling, and households selected from the selected addresses. The survey has been subject to non-response over the successive waves and weights are constructed to compensate for this nonresponse as well as for the relatively minor variation in sample selection probabilities. Further details of the survey and aspects of its data quality are provided in Taylor (2006) and Lynn (2006).

We focus on respondents who were employees and on the variable weekly gross pay, defined as the gross earnings the respondents report for the last time they were paid divided by the number of weeks in their reported pay period. We consider a sample of 3,294 individuals, who were in

Table 1: Measures of dispersion for logarithm of weekly gross pay (£) in BHPS according to whether latest payslip was seen

Accuracy indicator	*Interquartile range*	*Standard deviation*
Latest payslip seen	0.81	0.67
Latest payslip not seen	0.99	0.87
All cases	0.94	0.83

the original wave 1 sample as well as the wave 12 sample, who responded directly rather than via a proxy, who reported at Wave 12 that they had been employed and had been paid a positive amount for a pay period between 1 and 52 weeks. The accuracy indicator records whether the latest payslip was seen, or an earlier payslip was seen or no payslip was seen. Some 78 cases were excluded because of missing or inapplicable data on this accuracy indicator and one case was excluded who reported gross earnings of one pound for a pay period of one year. The principal source of selectivity in the sample is non-response and it is assumed here that this is handled by the use of survey weights provided with BHPS data for cross-sectional estimation.

Among the 3,294 individuals, the latest payslip was reported to have been seen for 33% of cases, an earlier payslip for 2% of cases and no payslip for 65% of cases. The reported weekly gross pay ranged from a minimum of £8 to a maximum of £2,600, with a lower quartile of £175, a median of £296 and an upper quartile of £450. Table 1 provides some summary statistics for the logarithm of this pay variable, according to values of the accuracy indicator. The standard deviation and the interquartile range are 30% and 22% higher, respectively, for cases where the latest payslip was not seen relative to cases where it was seen. Such an increase in dispersion is what might be anticipated as the effect of an increased amount of measurement error. Nevertheless, this is not the only possible explanation for such an increase, as will be discussed in Section 3, and to gain more insight into the measurement error we consider the following validation study.

2.2 Validation Study

This study, entitled 'Improving Survey Measurement of Income and Employment' (ISMIE), was based on the UK part of the European Community Household Panel Survey (ECHP), a study which includes data collected in a very similar way to the BHPS (Jäckle et al., 2004). The two surveys were administered jointly at the University of Essex until 2001, when the funding for ECHP expired. The ISMIE study was undertaken through interviews in early 2003 with respondents drawn from the final ECHP wave. Validation data were obtained through both linkage to welfare records and a survey of employers of ISMIE respondents who were in employment. We make use of just the latter data where the employer data on pay is considered as an alternative measure for comparing against the responses provided by the individual employees. We treat the employer measure as the true variable *prima facie*, although recognize that it may itself be subject to occasional errors. There were 156 individuals in the ISMIE study for whom matching data on pay are available from both the individuals and their employers and for whom the accuracy variable is recorded. This sample is subject to a series of forms of selection. The final ECHP wave had been subject to attrition from earlier waves and there was nonresponse to both the ISMIE survey of individuals and to the survey of their employers. The ISMIE study was based on a low income subsample from the final wave. Moreover, respondents had to give permission for their data to be linked to welfare records and to their employers being contacted. Finally, a small number of cases were removed because their pay periods were unclear. Given all this selection, we look to the validation data to inform qualitative assumptions about the measurement error process rather than to provide estimates of its parameters.

Fig. 1 displays a plot of log weekly pay recorded in the employer survey vs. log weekly pay recorded in the ECHP survey. The two are in reasonable agreement. We use the logarithm of the ratio of the pay reported by the employer and the employee as a measure of error.

Box plots of this measure are presented in Fig. 2 separately for cases where the latest pay slip was

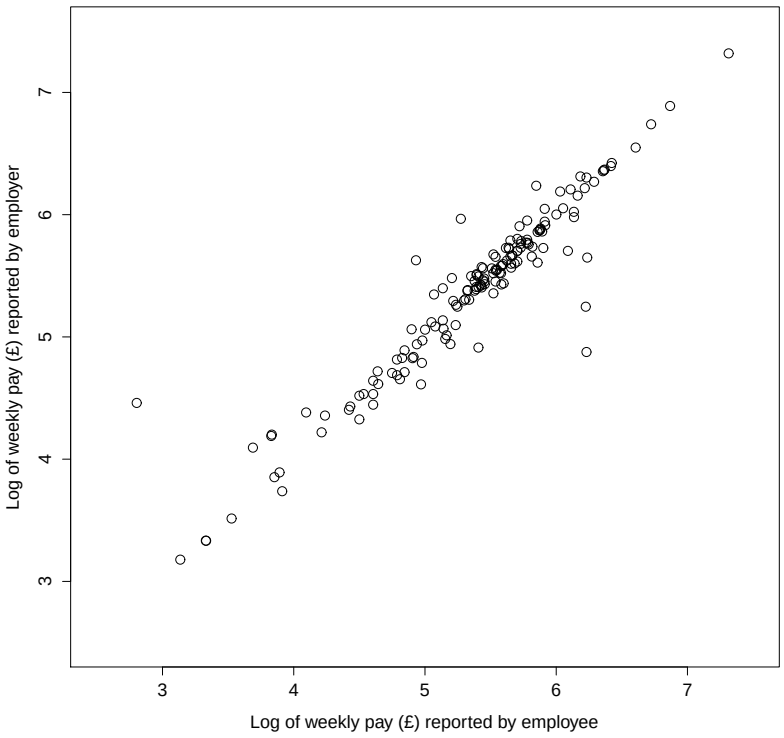

Figure 1: Logarithm of gross weekly pay reported by employer vs. logarithm of gross weekly pay reported by employee from ISMIE validation study

seen (50 cases) and when it was not (106 cases). We observe that, for the former group, the upper and lower quartiles are visually indistinguishable. This provides evidence that the ECHP measure can be viewed as accurate for over 50% of the cases when the latest pay slip is seen. Nevertheless, there are some cases where measurement error appears to arise, although it is possible that these discrepancies between the two sources arise from error in the employer survey. On the other hand, the distribution of the errors seems reasonably well approximated by as classical normal error model with zero mean when the latest pay slip is not seen.

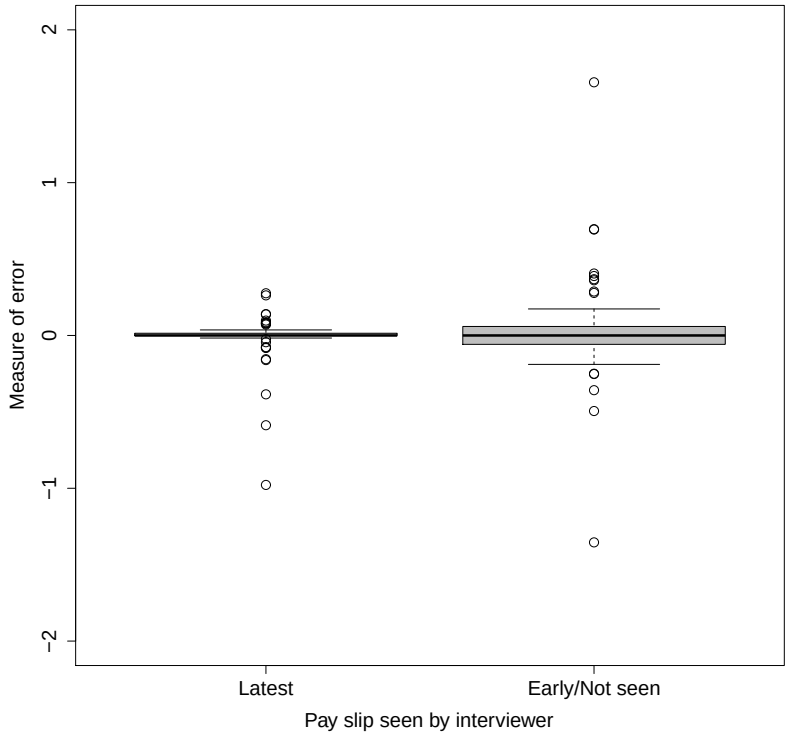

Figure 2: Boxplots of the logarithm of the ratio of the gross weekly pay (£) reported by the employer and the employee according to whether the payslip is seen by the interviewer in the ECHP survey

3 Models and Estimation Problem

Let y_i be the true value of a survey variable of interest for unit i in a finite population $U = \{1, \ldots, N\}$. Let a_i be a binary indicator of whether y_i is measured accurately and let y_i^* denote the measured value of y_i. As our basic model, we assume that

$$y_i^* = \begin{cases} y_i + \epsilon_i & \text{if } a_i = 0 \\ y_i & \text{if } a_i = 1, \end{cases}$$

where ϵ_i denotes measurement error. It is assumed that a sample is drawn from the population using a probability design and that y_i^* and a_i are observed for sample units.

As an extended model, motivated by the validation study in Section 2.2, we suppose that some measurement error may also arise for a proportion of the cases observed to be accurate. Thus, rather than observe a_i, we suppose that a_i^* is observed, where

$$a_i = 0 \qquad\qquad a_i^* = 0$$

$$a_i = \begin{cases} 0 \text{ with probability } p \\ \\ 1 \text{ with probability } 1 - p \end{cases} \qquad a_i^* = 1.$$

We suppose that p is a specified known value derived from some external source, which we may wish to vary in a sensitivity analysis. When $p = 0$, the extended model reduces to the basic model. Dominitz and Sherman (2004, 2006) also present models for an observed accuracy indicator a_i^*, although Dominitz and Sherman (2004) assumes that $y_i^* = y_i$ necessarily when $a_i^* = 1$, which we assume only in our basic model. Even this model differs from that of Dominitz and Sherman (2004, 2006) since they assume that $y_i^* = y_i$ with non-zero probability if $a_i = 0$ unlike in our more classical measurement error model, where we shall suppose that ϵ_i is continuously distributed. As noted by a referee, the former model may be plausible if the reason why some respondents do not consult their payslip is because they have accurate knowledge of their pay. However, we shall not consider this model further here.

A fundamental identification problem affecting the estimation of the distribution of y_i arises from the potential confounding between measurement error and the selection of cases with $a_i = 0$ vs. 1. When a_i is an indicator of whether the respondent refers to documentary records, there are at least two possible explanations of a difference in the distribution of y_i^* between cases with $a_i = 0$ vs. $a_i = 1$. The difference could be due to the presence of measurement error when $a_i = 0$ and its absence when $a_i = 1$ or it could be due to the fact that the use of documentary records may depend on characteristics which are related to true earnings. For example, imagine that some people have a simple source of earnings and are always paid the same amount. The task of remembering their pay may be so simple that there is no need to refer to documentary records. On the other hand, other people may have more complex sources of earnings, increasing the difficulty of answering accurately a question about pay and thus, potentially, increasing the chance that documents are referred to. The distribution of earnings may differ between these two types of employee and thus could explain differences in the distribution of y_i^* between cases with $a_i = 0$ vs. $a_i = 1$ even if no measurement error arises in either case. To address this identification challenge of disentangling measurement error from selection, we shall suppose that an additional vector $\mathbf{x}_i$ of variables is observed in the survey and that the following assumption holds:

- a_i and y_i are conditionally independent given $\mathbf{x}_i$, $i \in U$.

We also assume that the vectors

- $\left(a_i, a_i^*, y_i, y_i^*, \mathbf{x}_i\right)$ are independent, $i \in U$,
- $\left(y_i, y_i^*\right)$ is conditionally independent of a_i^* given $(a_i, \mathbf{x}_i)$, $i \in U$.

We shall make parametric assumptions regarding the distribution of $\left(y_i, y_i^*\right)$ given $(a_i, \mathbf{x}_i)$. Specifically, we suppose that this distribution depends on $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\eta})$, where the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ index the following conditional distributions:

- $y_i \mid \mathbf{x}_i \overset{indep}{\sim} f(y_i \mid \mathbf{x}_i; \boldsymbol{\gamma})$,

- $y_i^* \mid \mathbf{x}_i, y_i, a_i = 0 \overset{indep}{\sim} g(y_i^* \mid \mathbf{x}_i, y_i, a_i = 0; \boldsymbol{\eta})$.

The first of these conditional distributions represents the superpopulation model which generates the population values y_i. The second conditional distribution represents the measurement error model. The target of inference is taken to be the finite population distribution function

$$F(c) = N^{-1} \sum_{i \in U} I(y_i \leq c)$$

for an arbitrary fixed value c, where $I(y \leq c)$ is the indicator function that takes the value one if $y \leq c$ and zero otherwise. The methodology developed in this paper is also applicable to other characteristics of the distribution of the y_i, but we focus on this parameter for concreteness.

The observed data are given by $\{(y_i^*, a_i^*, \mathbf{x}_i) \ : \ i \in A\}$, where A is the index set of the observed units. We refer to A as the sample, but note that in practice it refers to the selected sample after removing nonresponding units. The basic problem of interest then is how to use these data to estimate $F(c)$.

4 Estimation

Our aim is to construct an estimator of $F(c)$ which is consistent with respect to the joint distribution induced by the parametric models, specified in the last section, and the mechanism by which the sample A is selected. We refer to the latter mechanism as the design mechanism and refer to moments with respect to this mechanism as design moments. In practice, we note that this mechanism represents a combination of sampling and nonresponse. We suppose that, without measurement error, sample weights w_i can be determined so that $(\sum_{i \in A} w_i)^{-1} \sum_{i \in A} w_i I(y_i \leq c)$ is a consistent estimator of $F(c)$ with respect to the design and hence with respect to the joint distribution. In the presence of measurement error with y_i replaced by y_i^*, the direct estimator

$$\hat{F}_d(c) = \Big(\sum_{i \in A} w_i \Big)^{-1} \sum_{i \in A} w_i I(y_i^* \leq c) \tag{1}$$

will, in general, be inconsistent for $F(c)$ even if ϵ_i has zero mean (Fuller, 1995). A typical pattern would be for $\hat{F}_d$ to be upwardly biased for values of c in the lower tail of the distribution and downwardly biased for values of c in the upper tail.

To correct for this inconsistency, we consider instead the corresponding weighted estimator of the expectation of $F(c)$ conditional on the population values of y_i^*, a_i^*, and $\mathbf{x}_i$, given by $\hat{F}(c; \hat{\boldsymbol{\psi}})$, where

$$\hat{F}(c, \boldsymbol{\psi}) = \left(\sum_{i \in A} w_i \right)^{-1} \sum_{i \in A} w_i E \left\{ I(y_i \leq c) \mid y_i^*, a_i^*, \mathbf{x}_i; \boldsymbol{\psi} \right\}$$

depends on the parameter vector $\boldsymbol{\psi}$ through the conditional expectation

$$E \left\{ I(y_i \leq c) \mid y_i^*, a_i^*, \mathbf{x}_i; \boldsymbol{\psi} \right\}$$

and $\hat{\boldsymbol{\psi}}$ is the pseudo-maximum likelihood estimator of $\boldsymbol{\psi}$ to be defined below. We assume here that sampling is noninformative given the observed variables, that is that

$$E \left\{ I(y_i \leq c) \mid y_i^*, a_i^*, \mathbf{x}_i, i \in A; \boldsymbol{\psi} \right\} = E \left\{ I(y_i \leq c) \mid y_i^*, a_i^*, \mathbf{x}_i; \boldsymbol{\psi} \right\}.$$

We wish to consider estimation for both the basic and extended models, but since the former is a special case of the latter model (by taking $p = 0$), we will assume the observations follow the extended model. In this case, we may write

$$E \left\{ I(y_i \leq c) \mid y_i^*, a_i^*, \mathbf{x}_i; \boldsymbol{\psi} \right\} = \begin{cases} (1 - p_i(\boldsymbol{\psi})) I(y_i^* \leq c) + p_i(\boldsymbol{\psi}) P_i(c; \boldsymbol{\psi}) & \text{if } a_i^* = 1 \\ \\ P_i(c; \boldsymbol{\psi}) & \text{if } a_i^* = 0, \end{cases}$$

where $p_i(\boldsymbol{\psi}) \equiv \Pr[a_i = 0 \mid y_i^*, a_i^* = 1, \mathbf{x}_i; \boldsymbol{\psi}]$ and $P_i(c; \boldsymbol{\psi}) \equiv E \left\{ I(y_i \leq c) \mid y_i^*, a_i = 0, \mathbf{x}_i; \boldsymbol{\psi} \right\}$ and we have

$$\hat{F}(c; \hat{\boldsymbol{\psi}}) = \left[\sum_{i \in A} w_i \right]^{-1} \left[\sum_{i \in A} w_i a_i^* \{ (1 - \hat{p}_i) I(y_i^* \leq c) + \hat{p}_i \hat{P}_i(c) \} + \sum_{i \in A} w_i (1 - a_i^*) \hat{P}_i(c) \right], \tag{2}$$

where $\hat{p}_i = p_i(\hat{\boldsymbol{\psi}})$ and $\hat{P}_i(c) = P_i(c; \hat{\boldsymbol{\psi}})$.

In our application, we consider a particular Gaussian parametric model where

$$y_i \mid \mathbf{x}_i \sim \mathrm{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2), \quad y_i^* \mid \mathbf{x}_i, y_i, a_i = 0 \sim \mathrm{N}(y_i, \tau^2),$$

so that $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \sigma^2)$ and $\boldsymbol{\eta} = \tau^2$. This model provides a reasonable fit to the kind of log weekly pay data discussed earlier in Section 2.2. Under this model, it may be shown that

$$y_i \mid y_i^*, a_i = 0, \mathbf{x}_i \;\sim\; \mathrm{N}(\mu_i^*,\, \sigma^{*2}),$$

where $\mu_i^* = (1-\rho)\mathbf{x}_i^\top \boldsymbol{\beta} + \rho y_i^*$, $\sigma^{*2} = \sigma^2(1-\rho)$ and $\rho = \sigma^2/(\sigma^2+\tau^2)$, implying that

$$P_i(c;\boldsymbol{\psi}) = \Phi\left(\frac{c-\mu_i^*}{\sigma^*}\right) \tag{3}$$

where $\Phi(.)$ denotes the standard normal distribution function. It can also be shown that

$$p_i(\boldsymbol{\psi}) = \frac{p}{\sqrt{\sigma^2+\tau^2}}\phi\left(\frac{y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}}{\sqrt{\sigma^2+\tau^2}}\right)\Big/ f(y_i^* \mid a_i^* = 1, \mathbf{x}_i; \boldsymbol{\psi}), \tag{4}$$

where

$$f(y_i^* \mid a_i^* = 1, \mathbf{x}_i; \boldsymbol{\psi}) = \frac{p}{\sqrt{\sigma^2+\tau^2}}\phi\left(\frac{y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}}{\sqrt{\sigma^2+\tau^2}}\right) + \frac{1-p}{\sigma}\phi\left(\frac{y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right)$$

and $\phi(\cdot)$ denotes the standard normal density function.

In order to determine $\hat{F}(c;\hat{\boldsymbol{\psi}})$, it only remains to define the pseudo maximum likelihood estimator of the parameter vector $\boldsymbol{\psi}$, which we take to be the solution of the pseudo score equation $\mathbf{S}_{obs}(\boldsymbol{\psi}) = \mathbf{0}$ (Skinner, 1989), where the pseudo–score function for $\boldsymbol{\psi}$ based on the observed data is defined as

$$\mathbf{S}_{obs}(\boldsymbol{\psi}) = \sum_{i\in A} w_i \frac{\partial}{\partial\boldsymbol{\psi}} \ln f(y_i^* \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi}) \equiv \sum_{i\in A} w_i \mathbf{S}_{obs,i}(\boldsymbol{\psi}), \tag{5}$$

and $f(y_i^* \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi})$ is the conditional density of y_i^* given a_i^* and $\mathbf{x}_i$. Under regularity conditions, we may write

$$\mathbf{S}_{obs,i}(\boldsymbol{\psi}) \;\;=\;\; E[\mathbf{S}_{com,i}(\boldsymbol{\psi}) \mid y_i^*, a_i^*, \mathbf{x}_i], \tag{6}$$

where

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

is the score function for $\boldsymbol{\psi}$ based on the complete vector of observations, i.e., $(y_i^*, y_i, a_i^*, a_i, \mathbf{x}_i)$, for the i–th unit.

An expression for $\mathbf{S}_{com,i}(\boldsymbol{\psi})$ is given in the Appendix both for the general case and for the the particular Gaussian parametric model introduced earlier. In the latter case, it is also shown that the three components of $\mathbf{S}_{obs,i}(\boldsymbol{\psi})$ are

$$\mathbf{S}_{obs,i}(\boldsymbol{\beta}) = \frac{\rho}{\sigma^2} w_{i,1}^* \mathbf{x}_i (y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}), \tag{7}$$

$$\mathbf{S}_{obs,i}(\sigma^2) = \frac{\rho}{2\sigma^2} w_{i,1}^* + \frac{\rho^2}{2\sigma^4} w_{i,2}^* (y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \tag{8}$$

and

$$\mathbf{S}_{obs,i}(\tau^2) = -\frac{\rho}{2\sigma^2} w_{i,3}^* + \frac{\rho^2}{2\sigma^4} w_{i,3}^* (y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})^2. \tag{9}$$

where $w_{i,1}^* = (1 - a_i^*) + a_i^* \{p_i + (1 - p_i)/\rho\}$, $w_{i,2}^* = (1 - a_i^*) + a_i^* \{p_i + (1 - p_i)/\rho^2\}$, $w_{i,3}^* = (1 - a_i^*) + a_i^* p_i$ and $p_i \equiv p_i(\boldsymbol{\psi})$.

It follows that the pseudo–maximum likelihood estimator of $\boldsymbol{\psi} = (\boldsymbol{\beta}, \sigma^2, \tau^2)$ can be expressed as

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left\{ \sum_{i \in A} \mathbf{x}_i w_i \hat{w}_{i,1}^* \mathbf{x}_i^\top \right\}^{-1} \sum_{i \in A} \mathbf{x}_i w_i \hat{w}_{i,1}^* y_i^*, \\
\hat{\sigma}^2 &= \frac{\sum_{i \in A} w_i a_i^* (1 - \hat{p}_i)(y_i^* - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2}{\sum_{i \in A} w_i a_i^* (1 - \hat{p}_i)} \\
\hat{\tau}^2 &= \frac{\sum_{i \in A} w_i \hat{w}_{i,3}^* (y_i^* - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2}{\sum_{i \in A} w_i \hat{w}_{i,3}^*} - \hat{\sigma}^2,
\end{aligned} \tag{10}$$

where $\hat{w}_{i,1}^*$, $\hat{w}_{i,3}^*$ and $\hat{p}_i$ take the same form as $w_{i,1}^*$, $w_{i,3}^*$ and p_i with $\boldsymbol{\psi}$ replaced by $\hat{\boldsymbol{\psi}}$. Expressing (10) as $\hat{\boldsymbol{\beta}} = g_1(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\tau}^2)$, $\hat{\sigma}^2 = g_2(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\tau}^2)$ and $\hat{\tau}^2 = g_3(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\tau}^2)$, we compute $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$ and $\hat{\tau}^2$ by iterating between $\hat{\sigma}^2_{(t+1)} = g_2(\hat{\boldsymbol{\beta}}_{(t)}, \hat{\sigma}^2_{(t)}, \hat{\tau}^2_{(t)})$, $\hat{\tau}^2_{(t+1)} = g_3(\hat{\boldsymbol{\beta}}_{(t)}, \hat{\sigma}^2_{(t)}, \hat{\tau}^2_{(t)})$ and $\hat{\boldsymbol{\beta}}_{(t+1)} = g_1(\hat{\boldsymbol{\beta}}_{(t)}, \hat{\sigma}^2_{(t+1)}, \hat{\tau}^2_{(t+1)})$ until convergence. A formal demonstration of the consistency of the pseudo–maximum likelihood estimator $\hat{\boldsymbol{\psi}}$ for $\boldsymbol{\psi}$ and of the pointwise consistency of $\hat{F}(c; \hat{\boldsymbol{\psi}})$ for $F(c)$, assuming a fixed known value of p, is provided in the Supplementary Materials, adapting Theorem 1.3.9 of Fuller (2009).

5 Variance estimation

For simplification, we shall assume that the sampling fraction is negligible, as seems reasonable in our application, so that the variance of $\hat{F}(c;\hat{\boldsymbol{\psi}}) - F(c)$ with respect to the joint distribution induced by the design and the model can be adequately captured by the model expectation of its design variance and so we seek an approximately unbiased estimator of this design variance. Following standard arguments (e.g. Fuller, 2009, p. 379), a linearization estimator of the design covariance matrix of the pseudo maximum likelihood estimator $\hat{\boldsymbol{\psi}}$ is given by

$$\hat{V}(\hat{\boldsymbol{\psi}}) = \{I_{obs}(\hat{\boldsymbol{\psi}})\}^{-1}\hat{V}\{\mathbf{S}_{obs}(\hat{\boldsymbol{\psi}})\}\{I_{obs}(\hat{\boldsymbol{\psi}})\}^{-1}, \tag{11}$$

where

$$I_{obs}(\boldsymbol{\psi}) = \sum_{i\in A} w_i\Big\{ -\frac{\partial}{\partial\boldsymbol{\psi}}\mathbf{S}_{obs,i}^{\top}(\boldsymbol{\psi})\Big\} \tag{12}$$

is the observed information matrix and $\hat{V}\{\mathbf{S}_{obs}(\hat{\boldsymbol{\psi}})\}$ is an estimator of the design covariance matrix of $\mathbf{S}_{obs}(\boldsymbol{\psi})$, evaluated at $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$. An expression for $I_{obs}(\boldsymbol{\psi})$, in the case of the Gaussian parametric model, is given in the Appendix.

The estimator of interest $\hat{F}(c;\hat{\boldsymbol{\psi}})$ in (2) may be expressed as

$$\hat{F}(c;\hat{\boldsymbol{\psi}}) = \frac{\sum_{i\in A} w_i z_i(\hat{\boldsymbol{\psi}})}{\sum_{i\in A} w_i},$$

where

$$z_i(\boldsymbol{\psi}) = a_i^*\{(1-p_i(\boldsymbol{\psi}))I(y_i^* \le c) + p_i(\boldsymbol{\psi})P_i(c;\boldsymbol{\psi})\} + (1-a_i^*)P_i(c;\boldsymbol{\psi}).$$

A linearization estimator of the design variance of $\hat{F}(c;\hat{\boldsymbol{\psi}})$, which takes account of the estimation error in $\hat{\boldsymbol{\psi}}$, is given by

$$\hat{V}[\hat{F}(c;\hat{\boldsymbol{\psi}})] = \frac{1}{\big[\sum_{i\in A} w_i\big]^2}\hat{V}\Big\{\sum_{i\in A} w_i\eta_i(c;\hat{\boldsymbol{\psi}})\Big\} \tag{13}$$

where $\hat{V}\big[\sum_{i\in A} w_i\eta_i(c;\hat{\boldsymbol{\psi}})\big]$ is an estimator of the design variance of $\sum_{i\in A} w_i\eta_i(c;\boldsymbol{\psi})$, evaluated at $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$, and

$$\eta_i(c;\boldsymbol{\psi}) = z_i(\boldsymbol{\psi}) + \Big\{\sum_{i\in A} w_i\frac{\partial z_i(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big\}^{\top} I_{obs}^{-1}(\boldsymbol{\psi})\mathbf{S}_{obs,i}(\boldsymbol{\psi}) - \hat{F}(c;\boldsymbol{\psi}). \tag{14}$$

An expression for $\partial z_i(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ is given in the Appendix. Fuller justifications and further details of the expressions in this section are given in the Supplementary Materials. The nature of the variance estimators $\hat{V}$ on the right hand sides of (11) and (13) will depend on the sampling scheme and we shall leave specification of these until the next section.

A simulation study was conducted with values of n increasing from 200 to 1,000. Confidence intervals based upon the variance estimation approach above were found to have appropriate coverage even when $n = 200$ across a range of values of c with $F(c)$ varying between 0.03 and 0.97. The study is described in the Supplementary Materials.

6 Application to BHPS data

In this section we apply the estimation methods set out in Sections 4 and 5 to the BHPS data described in Section 2.1. For the purpose of variance estimation, we treat the sampling design of the BHPS as a stratified multistage design where the primary sampling units (PSUs) can be treated as if they were selected from the strata with replacement (Skinner et al., 1989, Sect. 2.13). We approximate the stratification scheme by one with $H = 11$ design strata consisting of Government Office regions. Based on these assumptions, we use the following expression for the $\hat{V}$ term on the right–hand side of (11)

$$\hat{V}\{\mathbf{S}_{obs}(\hat{\boldsymbol{\psi}})\} = \sum_{h=1}^{H} \frac{n_h}{n_h - 1}\Big\{\sum_{j=1}^{n_h}(\mathbf{u}_{hj} - \bar{\mathbf{u}}_h)(\mathbf{u}_{hj} - \bar{\mathbf{u}}_h)^\top\Big\},$$

where h denotes stratum, j denotes PSU, n_h is the number of selected PSUs in stratum h, $\mathbf{u}_{hj} = \sum_{i\in A_{hj}} w_i \mathbf{S}_{obs,i}(\hat{\boldsymbol{\psi}})$ and $\bar{\mathbf{u}}_h = n_h^{-1}\sum_{j=1}^{n_h}\mathbf{u}_{hj}$. We use the following similar expression for the $\hat{V}$ term on the right–hand side of (13):

$$\hat{V}\Big\{\sum_{i\in A} w_i\eta_i(c;\hat{\boldsymbol{\psi}})\Big\} = \sum_{h=1}^{H}\frac{n_h}{n_h-1}\Big\{\sum_{j=1}^{n_h}\hat{\eta}_{hj}(c)^2 - \frac{\hat{\eta}_{h\cdot}(c)^2}{n_h}\Big\},$$

where $\hat{\eta}_{hj}(c) = \sum_{i\in A_{hj}} w_i\eta_i(c;\hat{\boldsymbol{\psi}})$ and A_{hj} denotes the set of sample units in PSU j in stratum h.

Before constructing such estimators, however, our first step was to investigate the choice of $\mathbf{x}$ for the regression model $f(y_i \mid \mathbf{x}_i;\boldsymbol{\gamma})$, where we took y to be the logarithm of weekly gross pay. We

wish to choose $\mathbf{x}$ so that the assumption of conditional independence between a_i and y_i given $\mathbf{x}_i$ in Section 3 is reasonable. For this purpose, we consider the kinds of predictors of pay in a wage equation, assuming that any systematic reason for a respondent to refer to a pay slip will relate to such kinds of observable characteristics of the respondent or their employer and not to any additional factors which are directly related to pay. We included the logarithm of hours worked as well as an indicator of part-time status (above or below 30 hours per week) in $\mathbf{x}$ in order to control for the dependence of pay on time worked and then considered standard variables used in wage equations including qualifications, region, industry, occupation and employer size as well as socio-demographic variables including gender and marital status.

The nature of the period according to which someone was paid was found not only to have a significant effect on the reported pay variable, after controlling for the above $\mathbf{x}$ variables, but it also appeared to lead to different amounts of measurement error. It seems plausible that the task of recalling pay and hence measurement error will differ between such periods. After some exploratory analysis, we decided to divide the sample into two groups, defined according to whether the pay period was less than a month or was greater than or equal to a month (the percentages of the sample in these two groups were 37% and 63% respectively).

Given our choice of $\mathbf{x}$, we then applied the pseudo maximum likelihood method to estimate the model parameters and the population distribution function for each of the two pay period subpopulations separately. In each case, we considered the following values of p: $0, 0.2$ and 0.4, viewed as plausible in the light of the validation study. Table 2 shows the pseudo maximum likelihood estimates of τ^2 together with standard errors in parentheses, calculated as described earlier. As in the exploratory data analysis, we found virtually no evidence of measurement error when the pay period was one month or more. There was, however, evidence of a non-zero value of τ^2 when the pay period was less than a month, especially as p increased.

We compared the distribution function of y estimated using our proposed method with that estimated using the direct method in (1) based on the reported values. For the pay period of one

Table 2: Pseudo maximum likelihood estimates of τ^2 according to pay period and choice of p. Figures in parentheses are the standard errors of the estimates

	p		
Pay period	*0.0*	*0.2*	*0.4*
1–4 weeks	0.0406	0.0705	0.0919
	(0.0209)	(0.0286)	(0.0276)
1 month or more	0.0004	0.0006	0.0015
	(0.0145)	(0.0236)	(0.0651)

month or more there was negligible difference between these estimated distribution functions for each of the three values of p. For the other pay period there were noticeable differences. The estimated lower part of the distribution function for the pay period of under a month is shown in Fig. 3 for alternative choices of p. There is a tendency for the proposed method to lead to lower estimates than the direct method for low values of pay and for this adjustment to be larger the larger the value of p. For example, when weekly pay is £20, the estimated distribution function is adjusted down from 1.61% to 1.52% when $p = 0$ and to 1.26% when $p = 0.2$. When weekly pay is £30, the estimated distribution function is adjusted down from 4.27% to 3.95% when $p = 0$ and to 3.56 when $p = 0.2$. The general pattern of downward adjustment for measurement error at the low end of a pay distribution seems plausible (c.f e.g. Skinner et al., 2002). The dependence on p might be anticipated by the fact that the larger the value of p the larger the proportion of cases for which measurement error applies and, hence, the greater the potential for adjustment. However, as p changes the parameter estimates change, as we saw in Table 2, so the sensitivity of results to the value of p is rather more complex.

Pointwise confidence intervals for $F(c)$ were calculated by obtaining standard errors first for $\text{logit}\{\hat{F}(c)\}$ using the methods described earlier and in Section 5 together with a linearization adjustment for the logit transformation, calculating standard normal theory confidence intervals for

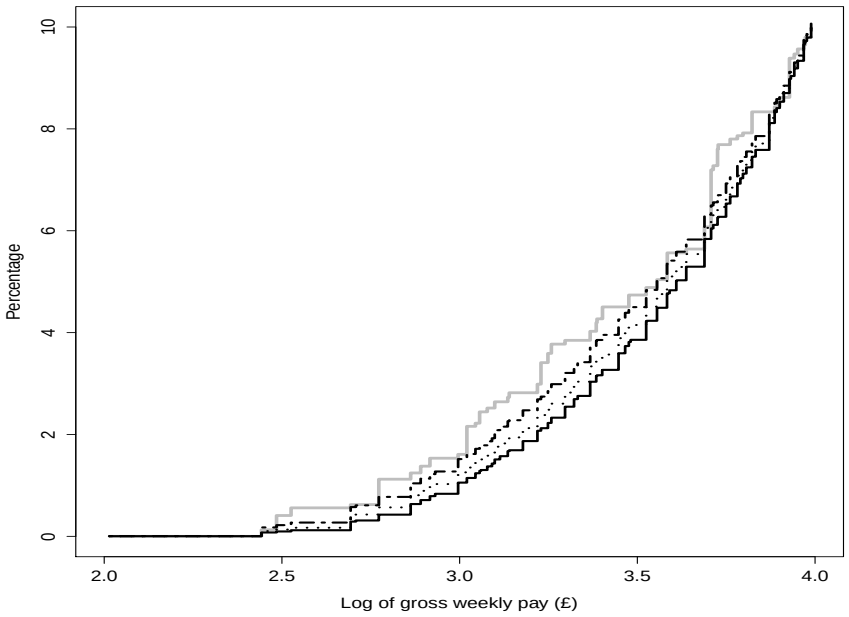

Figure 3: Estimated distribution functions of the logarithm of gross weekly pay for the direct method (——) and for the pseudo-maximum likelihood method with three values of p (0.0 – – –, 0.2 ····· and 0.4 ——) using BHPS data

logit$\{F(c)\}$ and then transforming back to obtain intervals for $F(c)$ which would fall within the bounds $[0, 1]$. The intervals are displayed in Fig. 4 for $p = 0.4$ together with point estimates. The plot relates, as in Fig. 3, to those paid at periods of less than a month. The adjustment might be viewed as reasonably modest in the sense that the direct estimator mainly lies within the proposed confidence intervals. Simultaneous intervals would, of course, be even wider.

7 Discussion

We have shown that it is feasible to use an accuracy indicator to adjust the estimation of a distribution function for measurement error. Although we have focussed on the estimation of a distribution function for concreteness, our approach could in principle be applied to other estimation problems,

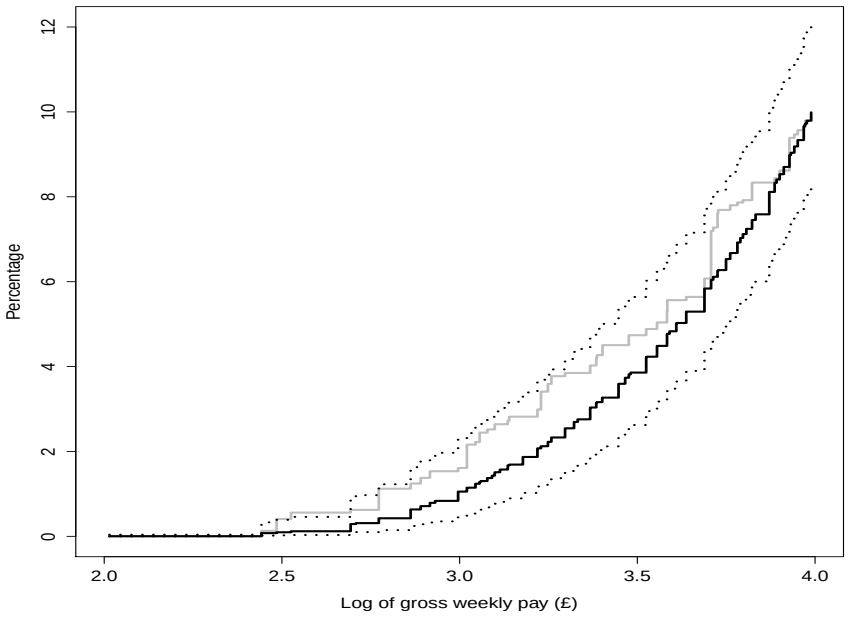

Figure 4: Estimated distribution functions of the logarithm of gross weekly pay for the direct method (——) and for the pseudo-maximum likelihood method with $p = 0.4$ (——) using BHPS data. Dotted lines give the 95% pointwise confidence limits using the proposed method with $p = 0.4$

for example in regression with covariates measured with error. The tractability of our approach has benefited in this paper from certain Gaussian assumptions. Of course, departures from these assumptions may arise. Rodgers et al. (1993) illustrate such departures for similar earnings data and we observe some suggestion of non-normality in the tails of the distributions for our data. To develop methodology to handle other less tractable distributions we are exploring the potential use of parametric fractional imputation (Kim, 2011). Dominitz and Sherman (2006) show how to identify and estimate bounds for a distribution function with an underlying binary mixture model for measurement error. As suggested by a referee, it would be interesting to compare empirically the results of their approach with ours.

We have considered only the case of a binary accuracy indicator. The use of ordinal and continuous accuracy indicators is also left for exploration in further work.

We have presented an approach which allows for a sensitivity analysis to departures from a basic binary accuracy indicator model according to values of p. In simulation work not reported here, we have studied the effect of misspecification of p. On the whole, the results of the study were encouraging, for example showing that if the true value of p was 0.3 then it made little difference in bias terms whether a value of $p = 0.1, 0.3$ or 0.5 was specified. The effect of specifying $p = 0$ when the true value of p was 0.3 was a little more damaging with the bias adjustment still being in the right direction but with some underadjustment. The effect of specifying $p = 0.3$ when the true value of p was 0 seemed less damaging. In general, it does seem sensible to consider the possibility that $p \neq 0$ unless there are strong *a priori* grounds to suppose that when the accuracy indicator indicates accurate measurement then no measurement error can have occurred.

8 Acknowledgements

Research was supported by the Economic and Social Research Council. Damião Da Silva is on a postdoctoral research fellowship at the Southampton Statistical Sciences Research Institute, University of Southampton, U.K. We are grateful to Jae-Kwang Kim for discussion.

Appendix

Under the extended model we may write

$$
\begin{aligned}
f(y_i^*, y_i, a_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi}) &= f(a_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi}) f(y_i \mid a_i^*, a_i, \mathbf{x}_i; \boldsymbol{\psi}) f(y_i^* \mid y_i, a_i^*, a_i, \mathbf{x}_i; \boldsymbol{\psi}) \\
&= f(a_i \mid a_i^*) f(y_i \mid \mathbf{x}_i; \boldsymbol{\gamma}) f(y_i^* \mid y_i, a_i, \mathbf{x}_i; \boldsymbol{\psi}) \\
&= \begin{cases}
f(y_i \mid \mathbf{x}_i, \boldsymbol{\gamma}) g(y_i^* \mid y_i, a_i = 0, \mathbf{x}_i; \boldsymbol{\eta}), & a_i^* = 0, a_i = 0 \\
p f(y_i \mid \mathbf{x}_i, \boldsymbol{\gamma}) g(y_i^* \mid y_i, a_i = 0, \mathbf{x}_i; \boldsymbol{\eta}), & a_i^* = 1, a_i = 0 \\
(1-p) f(y_i^* \mid \mathbf{x}_i; \boldsymbol{\gamma}), & a_i^* = 1, a_i = 1, y_i = y_i^* \\
0, & \text{otherwise.}
\end{cases}
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\mathbf{S}_{com,i}(\boldsymbol{\psi}) &= (1 - a_i)\left\{ \frac{\partial}{\partial \boldsymbol{\psi}} \ln f(y_i \mid \mathbf{x}_i; \boldsymbol{\gamma}) + \frac{\partial}{\partial \boldsymbol{\psi}} \ln g(y_i^* \mid y_i, a_i = 0, \mathbf{x}_i; \boldsymbol{\eta}) \right\} + \\
& \quad a_i^* a_i \frac{\partial}{\partial \boldsymbol{\psi}} \ln f(y_i^* \mid \mathbf{x}_i; \boldsymbol{\gamma}).
\end{aligned}
$$

Under the further assumptions that $y_i \mid \mathbf{x}_i \sim \mathrm{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ and $y_i^* \mid \mathbf{x}_i, y_i, a_i = 0 \sim \mathrm{N}(y_i, \tau^2)$, we have

$$
\ln f(y_i \mid \mathbf{x}_i; \boldsymbol{\gamma}) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2
$$

and

$$
\ln g(y_i^* \mid \mathbf{x}_i, y_i, a_i = 0; \boldsymbol{\eta}) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \tau^2 - \frac{1}{2\tau^2} (y_i^* - y_i)^2
$$

and so we obtain

$$
\mathbf{S}_{com,i}(\boldsymbol{\beta}) = (1 - a_i) \frac{1}{\sigma^2} \mathbf{x}_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) + a_i^* a_i \frac{1}{\sigma^2} \mathbf{x}_i (y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}), \tag{15}
$$

$$
\mathbf{S}_{com,i}(\sigma^2) = (1 - a_i)\left\{ -\frac{1}{2\sigma^2} + \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^4} \right\} + a_i^* a_i \left\{ -\frac{1}{2\sigma^2} + \frac{(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^4} \right\} \tag{16}
$$

and

$$\mathbf{S}_{com,i}(\tau^2) = (1-a_i)\Big\{ -\frac{1}{2\tau^2} + \frac{(y_i^* - y_i)^2}{2\tau^4}\Big\}. \tag{17}$$

Since $y_i \mid y_i^*, a_i = 0, \mathbf{x}_i \sim \mathrm{N}\{(1-\rho)\mathbf{x}_i^\top\boldsymbol{\beta} + \rho y_i^*,\ \sigma^2(1-\rho)\}$ we have

$$E[y_i - \mathbf{x}_i^\top\boldsymbol{\beta} \mid y_i^*, a_i = 0, \mathbf{x}_i] = \rho(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}),$$

$$E[(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2 \mid y_i^*, a_i = 0, \mathbf{x}_i] = \sigma^2(1-\rho) + \rho^2(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta})^2,$$

$$E[y_i^* - y_i \mid y_i^*, a_i = 0, \mathbf{x}_i] = (1-\rho)(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta})$$

and

$$E[(y_i^* - y_i)^2 \mid y_i^*, a_i = 0, \mathbf{x}_i] = \sigma^2(1-\rho) + (1-\rho)^2(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta})^2.$$

We may use these results to evaluate the conditional expectations of (15)–(17) given $(y_i^*, a_i^*, \mathbf{x}_i)$, as in (6), and hence obtain expressions (7)–(9).

Turning to variance estimation, the information matrix in (12) may be expressed, in the case of the Gaussian parametric model, as:

$$I_{obs}(\boldsymbol{\psi}) = -\sum_{i\in A} w_i \begin{Bmatrix} \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{obs,i}^\top(\boldsymbol{\beta}) & \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{obs,i}(\sigma^2) & \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{obs,i}(\tau^2) \\ \frac{\partial}{\partial\boldsymbol{\beta}^\top}\mathbf{S}_{obs,i}(\sigma^2) & \frac{\partial}{\partial\sigma^2}\mathbf{S}_{obs,i}(\sigma^2) & \frac{\partial}{\partial\sigma^2}\mathbf{S}_{obs,i}(\tau^2) \\ \frac{\partial}{\partial\boldsymbol{\beta}^\top}\mathbf{S}_{obs,i}(\tau^2) & \frac{\partial}{\partial\sigma^2}\mathbf{S}_{obs,i}(\tau^2) & \frac{\partial}{\partial\tau^2}\mathbf{S}_{obs,i}(\tau^2) \end{Bmatrix},$$

where the derivatives of $\mathbf{S}_{obs,i}(\boldsymbol{\psi})$ with respect to the components of $\boldsymbol{\psi}$ are as follows

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{obs,i}^\top(\boldsymbol{\beta}) = -\frac{\rho}{\sigma^2}\Big\{w_{i,1}^*\mathbf{x}_i\mathbf{x}_i^\top + \frac{(1-\rho)}{\rho}a_i^*\frac{\partial p_i}{\partial\boldsymbol{\beta}}\mathbf{x}_i^\top(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta})\Big\},$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{obs,i}(\sigma^2) = -\frac{\rho^2}{\sigma^4}w_{i,2}^*\mathbf{x}_i(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}) + a_i^*\frac{\partial p_i}{\partial\boldsymbol{\beta}}\Big\{-\frac{\rho-1}{2\sigma^2} + \frac{\rho^2-1}{2\sigma^4}(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta})^2\Big\},$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{obs,i}(\tau^2) = -\frac{\rho^2}{\sigma^4}w_{i,3}^*\mathbf{x}_i(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}) + a_i^*\frac{\partial p_i}{\partial\boldsymbol{\beta}}\Big\{-\frac{\rho}{2\sigma^2} + \frac{\rho^2}{2\sigma^4}(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta})^2\Big\},$$

$$\frac{\partial}{\partial\sigma^2}\mathbf{S}_{obs,i}(\sigma^2) = -\frac{\rho^2}{2\sigma^4}w^*_{i,1} + \frac{(1-\rho)}{2\sigma^2}a^*_i\Big\{\frac{\partial p_i}{\partial\sigma^2} + \frac{1-p_i}{\rho^2}\Big\} - \Big[\frac{\rho^3}{\sigma^6}w^*_{i,2} + \frac{(1-\rho)}{2\sigma^4}a^*_i\Big\{(1+\rho)\frac{\partial p_i}{\partial\sigma^2} + \frac{2(1-p_i)}{\sigma^2}\Big\}\Big](y^*_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2,$$

$$\frac{\partial}{\partial\sigma^2}\mathbf{S}_{obs,i}(\tau^2) = -\frac{\rho^2}{2\sigma^4}w^*_{i,3} - \frac{\rho}{2\sigma^2}a^*_i\frac{\partial p_i}{\partial\sigma^2} + \Big\{-\frac{\rho^3}{\sigma^6}w^*_{i,3} + \frac{\rho^2}{2\sigma^4}a^*_i\frac{\partial p_i}{\partial\sigma^2}\Big\}(y^*_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2$$

and

$$\frac{\partial}{\partial\tau^2}\mathbf{S}_{obs,i}(\tau^2) = \frac{\rho^2}{2\sigma^4}w^*_{i,3} - \frac{\rho}{2\sigma^2}a^*_i\frac{\partial p_i}{\partial\tau^2} + \Big\{-\frac{\rho^3}{\sigma^6}w^*_{i,3} + \frac{\rho^2}{2\sigma^4}a^*_i\frac{\partial p_i}{\partial\tau^2}\Big\}(y^*_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2,$$

where

$$\frac{\partial p_i}{\partial\boldsymbol{\beta}} = -p_i(1-p_i)\frac{(1-\rho)}{\sigma^2}\mathbf{x}_i(y^*_i - \mathbf{x}_i^\top\boldsymbol{\beta}), \tag{18}$$

$$\frac{\partial p_i}{\partial\sigma^2} = -p_i(1-p_i)\Big\{-\frac{(1-\rho)}{2\sigma^2} + \frac{(1-\rho^2)}{2\sigma^4}(y^*_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2\Big\} \tag{19}$$

and

$$\frac{\partial p_i}{\partial\tau^2} = p_i(1-p_i)\Big\{-\frac{\rho}{2\sigma^2} + \frac{\rho^2}{2\sigma^4}(y^*_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2\Big\}. \tag{20}$$

The term $\partial z_i(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ in (14) may be expressed as

$$\frac{\partial z_i(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}} = a^*_i\Big\{-\frac{\partial p_i(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big(I(y^*_i < c) - P_i(c;\boldsymbol{\psi})\Big) + p_i(\boldsymbol{\psi})\frac{\partial P_i(c;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big\} + (1-a^*_i)\frac{\partial P_i(c;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}},$$

where $\partial p_i(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ is obtained from (18)–(20) and

$$\frac{\partial P_i(c;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}} = -\phi\Big(\frac{c-\mu^*_i}{\sigma^*}\Big)\begin{bmatrix} \frac{(1-\rho)}{\sigma\sqrt{1-\rho}}\mathbf{x}_i \\ \\ \frac{(1-\rho)}{2\sigma^2}\frac{\{(c-\mathbf{x}_i^\top\boldsymbol{\beta})+\rho(y^*_i-\mathbf{x}_i^\top\boldsymbol{\beta})\}}{\sigma\sqrt{1-\rho}} \\ \frac{\rho^2}{2\sigma^2(1-\rho)}\frac{\{(c-y^*_i)-(1-\rho)(y^*_i-\mathbf{x}_i^\top\boldsymbol{\beta})\}}{\sigma\sqrt{1-\rho}} \end{bmatrix}.$$

References

Begg, C. B. and R. A. Greenes (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics 39*, 207–215.

Biemer, P. P. and L. Lyberg (2003). *Introduction to Survey Quality*. Hoboken, New Jersey: Wiley-Interscience.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement error in survey data. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5, pp. 3705 – 3843. Elsevier.

Dominitz, J. and R. P. Sherman (2004). Sharp bounds under contaminated or corrupted sampling with verification, with an application to environmental pollutant data. *Journal of Agricultural, Biological, and Environmental Statistics 9*(3), 319–338.

Dominitz, J. and R. P. Sherman (2006). Identification and estimation of bounds on school performance measures: a nonparametric analysis of a mixture model with verification. *Journal of Applied Econometrics 21*(8), 1295–1326.

Eltinge, J. L. (1999). Accounting for non-gaussian measurement error in complex survey estimators of distribution functions and quantiles. *Statistica Sinica 9*, 425–449.

Fuller, W. A. (1995). Estimation in the presence of measurement error (with discussion). *International Statistical Review 63*, 121–147.

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley & Sons, Inc.

Horowitz, J. and C. F. Manski (1995). Identification and robustness with contaminated and corrupted data (STMA V36 4949). *Econometrica 63*, 281–302.

Jäckle, A., E. Sala, S. P. Jenkins, and P. Lynn (2004). Validation of survey data on income and employment: The ismie experience. Working paper 2004–14 488, Institute for Social and Economics Research, University of Essex.

Kapteyn, A. and J. Y. Ypma (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics 25*, 513–551.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika 98*(1), 119–132.

Kreider, B. and J. V. Pepper (2007). Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors. *Journal of the American Statistical Association 102*(478), 432–441.

Lynn, P. (2006). *Quality Profile: British Household Panel Survey, Waves 1 to 13: 1991-2003*. Colchester: University of Essex.

Mathiowetz, N. A. (1998). Respondent expressions of uncertainty: Data source for imputation. *Public Opinion Quarterly 62*(1), 47–56.

Moore, J. C., L. L. Stinson, and J. Welniak, Edward J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics 16*(4), 331–361.

Rodgers, W. L., C. Brown, and G. J. Duncan (1993). Errors in survey reports of earnings, hours worked, and hourly wages. *Journal of the American Statistical Association 88*(424), 1208–1218.

Skinner, C., N. Stuttard, G. Beissel-Durrant, and J. Jenkins (2002). The measurement of low pay in the UK labour force Survey. *Oxford Bulletin of Economics and Statistics 64*(s1), 653–676.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, and T. M. F. Smith (Eds.), *Analysis of Complex Surveys*, pp. 59–87. John Wiley & Sons.

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

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

Wansbeek, T. J. and E. Meijer (2000). *Measurement Error and Latent Variables in Econometrics*. Amsterdam: North-Holland Publishing Co.