Estimating from Cross-sectional Categorical Data Subject to Misclassification and Double Sampling: Moment-based, Maximum Likelihood and Quasi-Likelihood Approaches

Nikos Tzavidis and Yan-Xia Lin

Abstract

We discuss the analysis of cross-sectional categorical data in the presence of misclassification and double sampling. In a double sampling context we assume that along with the main measurement device, which is subject to misclassification, we have a secondary measurement device, which is free of error but more expensive to apply. Due to its higher cost, the validation survey is employed only for a subset of units. Inference using double sampling is based on combining information from both measurement devices. Previously proposed parameterisations of the misclassification model that utilize either the calibration or the misclassification probabilities are reviewed. We then show that the misclassification model can be alternatively formulated as a missing data problem using the misclassification probabilities. In this context, the model parameters are estimated using maximum likelihood estimation via the EM algorithm. We suggest that the formulation of the misclassification model as a missing data problem using the misclassification probabilities, as opposed to maximum likelihood estimation using the calibration probabilities, offers a robust basis for extending the model to handle more complex situations. We further illustrate that the likelihood-based approaches offer some practical advantages over the moment-based approaches. As an alternative approach, we also present a quasi-likelihood parameterisation of the misclassification model. In this framework, an explicit definition of the likelihood function is avoided and a different way of resolving a missing data problem is provided. The quasi-likelihood method offers further practical advantages to the data analyst over the likelihood-based and the moment-based approaches. Variance estimation under the alternative parameterisations is discussed. The different methods are illustrated using two numerical examples and a Monte-Carlo simulation study.

S^{3}RI Methodology Working Paper M04/03

Estimating from Cross-sectional Categorical Data Subject to Misclassification and Double Sampling: Moment-based, Maximum Likelihood and Quasi-Likelihood Approaches

Nikos Tzavidis [1] and Yan-Xia Lin [2]

1 Southampton Statistical Sciences Research Institute, University of Southampton, Highfield Campus, Southampton, SO17 1BJ, UK

2 School of Mathematics and Applied Statistics, University of Wollongong, Northfields Ave, Wollongong, NSW 2522, Australia

ABSTRACT

We discuss the analysis of cross-sectional categorical data in the presence of misclassification and double sampling. In a double sampling context we assume that along with the main measurement device, which is subject to misclassification, we have a secondary measurement device, which is free of error but more expensive to apply. Due to its higher cost, the validation survey is employed only for a subset of units. Inference using double sampling is based on combining information from both measurement devices. Previously proposed parameterisations of the misclassification model that utilize either the calibration or the misclassification probabilities are reviewed. We then show that the misclassification model can be alternatively formulated as a missing data problem using the misclassification probabilities. In this context, the model parameters are estimated using maximum likelihood estimation via the EM algorithm. We suggest that the formulation of the misclassification model as a missing data problem using the misclassification probabilities, as opposed to maximum likelihood estimation using the calibration probabilities, offers a robust basis for extending the model to handle more complex situations. We further illustrate that the likelihood-based approaches offer some practical advantages over the moment-based approaches. As an alternative approach, we also present a quasi-likelihood parameterisation of the misclassification model. In this framework, an explicit definition of the likelihood function is avoided and a different way of resolving a missing data problem is provided. The quasi-likelihood method offers further practical advantages to the data analyst over the likelihood-based and the moment-based approaches. Variance estimation under the alternative parameterisations is discussed. The different methods are illustrated using two numerical examples and a Monte-Carlo simulation study.

Keywords: Measurement error; Validation surveys; Missing data; EM algorithm; Missing Information Principle; Estimating equations.

Nikos Tzavidis is Research Fellow at the Southampton Statistical Sciences Research Institute, University of Southampton, Highfield Campus, Southampton SO17 1BJ, UK (E-mail: ntzav@socsci.soton.ac.uk). Yan-Xia Lin is Associate Professor at the School of Mathematics and Applied Statistics, University of Wollongong, Northfields Ave, Wollongong, NSW 2522, Australia (E-mail: yanxia_lin@uow.edu.au)

1. INTRODUCTION

The existence of measurement error in data used for statistical analysis can introduce serious bias in the derived results. In a discrete framework, the term measurement error can be replaced by the more natural term misclassification. Methods that account for the existence of measurement error have received great attention in statistical literature. In the presence of measurement error, such methods need to be employed in order to ensure the validity of the inferential process. However, in a discrete framework conventional errors in variables models (Fuller 1987) can not be applied. One of the traditional approaches for adjusting for misclassification in discrete data is by assuming the existence of validation information derived from a validation survey, which is free of error.

The use of validation surveys can be placed into the general framework of double sampling methods (Bross 1954; Tenenbein 1970, 1972). In a double sampling framework, we assume that along with the main measurement device, which is affected by measurement error, we have a secondary measurement device (validation survey), which is free of error but more expensive to apply. Due to its higher cost, the validation survey is employed only for a subset of units. Under the assumption that the validation survey is free of error, one can estimate the parameters of the measurement error mechanism. Inference using double sampling is based on combining information from both measurement devices.

The aim of this paper is to examine and compare alternative parameterisations of the misclassification model when discrete data are subject to misclassification and validation information is available. The organisation of the paper is as follows. In Section 2, the general framework of double sampling is presented and alternative double sampling schemes are examined. A moment-based estimator along with a maximum likelihood estimator is reviewed and the effect of different double sampling designs on the estimation process is studied. In Section 3, the misclassification model is parameterised as a missing data problem and estimation is performed via the EM algorithm. The advantages of this parameterisation are discussed and a procedure for deriving standard errors for the

adjusted estimates is described. In Section 4, the misclassification model is parameterised in a quasi-likelihood framework. The gains from using this parameterisation are described and variance estimation in a quasi-likelihood framework is illustrated. Using a numerical example, in Section 5 we show that under the same conditions utilized also in a maximum likelihood framework the quasi-likelihood approach produces reasonable estimates for the parameters of interest. Using another numerical example, the alternative approaches are contrasted in the presence of intense misclassification. In Section 6, a Monte-Carlo simulation study is designed for empirically comparing the alternative methods.

2. DOUBLE SAMPLING SCHEMES UTILISED TO ADJUST FOR MISCLASSIFICATION

Suppose that the standard measurement device is subject to measurement error. As a result we have biased results. Unbiased estimates can be obtained by utilising more elaborate measurement tools usually referred to as preferred procedures (Deming 1950; Forsman and Schreiner 1991; Kuha and Skinner 1997). Examples of preferred procedures in official statistics are the re-interview surveys (Bailar 1968). In bio-statistical applications the term "gold standard" is more commonly used (Bauman and Koch 1983). Other examples include judgments of experts or checks against administrative records (Greenland 1988). The basic assumption that the preferred procedure is free of error makes possible the estimation of the parameters of the misclassification mechanism. On the other hand, the preferred procedures are considered to be fairly expensive and thus unsuitable to be used for the entire sample (hereinafter main sample). Therefore, these procedures are normally applied to a smaller sample usually referred to as validation sample.

The validation sample can be either internal or external. Kuha and Skinner (1997) make this distinction following literature on misclassification in the context of bio-statistical applications (Greenland 1988). The basic characteristic that distinguishes an internal validation sample from an external validation sample is whether the fallible classifications from the validation sample can be combined with the fallible classifications from the main sample. A validation sample is defined as

internal if it is a sub-sample of $n^{(v)}$ units from the main sample of n units obtained via a randomised double sampling scheme. Alternatively, the validation sample can be regarded as internal if it is selected independently from the main sample and from the same target population. Otherwise, the validation sample is characterised as an external. For example, in the Panel Study of Income Dynamics (Hill 1992) the survey responses were validated by comparing them to company records in a separate sample of employees of one large firm. The parameters of misclassification mechanism estimated from an external validation sample are assumed to be representative of the misclassification process in the target population but the fallible classifications from this validation sample cannot be combined with the fallible classifications from the main sample.

2.1 A Moment-based Estimator for Adjusting for Misclassification

Let Y_ξ^* denote a discrete random variable for unit ξ. Denote by $\Pi_i = pr\left(Y_\xi^* = i\right)$ the probability that a unit ξ is classified in state i by the standard measurement device, which is subject to measurement error, by $P_k = pr\left(Y_\xi = k\right)$ the probability that a unit ξ truly belongs in state k and by $q_{ik} = pr\left(Y_\xi^* = i \mid Y_\xi = k\right)$ the misclassification probabilities. Define now a vector Π with elements Π_i, a vector P with elements P_k and the misclassification matrix Q with elements q_{ik}. Generally speaking, the misclassification model with r mutually exclusive states can now be described as follows:

$$pr\left(Y_\xi^* = i\right) = \sum_{k=1}^{r} pr(Y_\xi^* = i \mid Y_\xi = k) pr\left(Y_\xi = k\right) \Rightarrow \Pi_i = \sum_{k=1}^{r} q_{ik} P_k. \tag{2.1}$$

Solving (2.1) equation with respect to P, writing it in matrix notation and assuming that Q is non-singular, we obtain the following expression

$$\underset{r\times 1}{P} = \underset{r\times r}{Q^{-1}} \underset{r\times 1}{\Pi}. \tag{2.2}$$

The unknown quantities involved in (2.2) are typically estimated using an appropriate double sampling scheme. An estimator of (2.2) is given by

$$\underset{r\times 1}{\hat{P}} = \underset{r\times r}{\hat{Q}^{-1}}\,\underset{r\times 1}{\hat{\Pi}}. \tag{2.3}$$

The moment-based estimator (2.3) has been used extensively in literature to adjust discrete data for measurement error. A drawback associated with the use of the moment-based estimator is that under certain conditions it can produce estimates that lie outside the parameter space. This can happen due to the inversion of the misclassification matrix involved in the estimation process.

Variance estimation for the moment-based estimator can been performed using linearization techniques and relevant solutions are illustrated among others by Selen (1986) and Greenland (1988). Kuha and Skinner (1997) discuss the use of estimator (2.3) both under an internal and an external validation sample. They conclude that under an internal validation sample the estimator given in (2.3) is more efficient. This is due to the extra information on the observed classifications derived from the internal validation sample. Here, we argue that an external validation sample can be transformed into an internal validation sample. Since the misclassification probabilities estimated from an external validation sample are assumed to be representative of the misclassification process in the target population, we propose to calibrate $pr\left(Y_\xi^* = i, Y_\xi = k\right)$ on the marginal information derived from the main sample. In the simplest case, this calibration procedure can be performed using an Iterative Proportional Fitting (IPF) algorithm (Deming and Stephan 1940). The consequence of this transformation is to make estimator (2.3) under an external validation sample as efficient as the same estimator under an internal validation sample.

2.2 Calibration Probabilities vs. Misclassification Probabilities and Maximum Likelihood Estimation

In order to describe the misclassification mechanism, estimator (2.3) utilises the misclassification probabilities $q_{ik} = pr\left(Y_\xi^* = i \mid Y_\xi = k\right)$. Another way of making inference about the misclassification mechanism is by using what Carroll (1992) refers to as calibration probabilities. The calibration

probabilities are defined as $c_{ki} = pr\left(Y_\xi = k \mid Y_\xi^* = i\right)$. Denote by C the matrix of calibration probabilities. The misclassification model under the calibration probabilities becomes

$$pr\left(Y_\xi = k\right) = \sum_{i=1}^{r} pr\left(Y_\xi = k \mid Y_\xi^* = i\right) pr\left(Y_\xi^* = i\right) \Rightarrow P_k = \sum_{i=1}^{r} c_{ki} \Pi_i. \tag{2.4}$$

In matrix notation,

$$\underset{r\times 1}{P} = \underset{r\times r}{C}\ \underset{r\times 1}{\Pi}. \tag{2.5}$$

Using an appropriate double sampling scheme, an estimator of (2.5) is given by

$$\underset{r\times 1}{\hat{P}} = \underset{r\times r}{\hat{C}}\ \underset{r\times 1}{\hat{\Pi}}. \tag{2.6}$$

Tenenbein (1972) proved that estimator (2.6) is the maximum likelihood estimator of (2.5) and he also provided an expression for its asymptotic variance using the inverse of the information matrix. As noted by Marshall (1990) and Kuha and Skinner (1997), the maximum likelihood estimator (2.6) will be asymptotically more efficient than the moment-based estimator (2.3). However, estimator (2.6) assumes internal validation data. Unlike the misclassification probabilities that condition on the true classifications, the calibration probabilities condition on the observed classifications. The true classifications are perceived as representatives of a universal truth. Therefore, the misclassification probabilities can be regarded as transportable to the population of interest and can be used also in the case of an external validation sample. When only external validation data are available, the poorer performance of the moment-based estimator (2.3) is an important problem. One way to overcome this problem is by transforming the external validation sample into an internal validation sample and then use the maximum likelihood estimator.

3. AN ALTERNATIVE PARAMETERISATION FOR MAXIMUM LIKELIHOOD ESTIMATION

In this section we present an alternative parameterisation for maximum likelihood estimation by utilising the misclassification probabilities instead of the calibration probabilities. We argue that this parameterisation offers a robust basis for extending the model to handle more complex situations.

3.1 The Model

The set up is as follows. For the main sample of n units the classifications are made using only the fallible classifier. For a smaller sample of $n^{(v)}$ units the classifications are made using both the "perfect" (validation survey) and the fallible classifier. Consider the cross-classification of the observed with the true classifications (Tables 1 and 2). Using a subscript $(*)$ to denote unobserved quantities, denote by $n_{ik}^{(*)}, n_{ik}^{(v)}$ the counts referring to this cross-classification in the main sample and in the validation sample respectively. Denote also by $n_{\bullet k}^{(*)}, n_{\bullet k}^{(v)}$ the total number of sample units classified in state k by the "perfect" classifier in the main sample and in the validation sample respectively.

Table 1: Data from Validation sample

		True Classifications			
		(1)	⋯	(r)	Margins
Fallible	(1)	$n_{11}^{(v)}$	⋯	$n_{1r}^{(v)}$	$n_{1\bullet}^{(v)}$
Classifications	⋮	⋮	⋯	⋮	⋮
	(r)	$n_{r1}^{(v)}$	⋯	$n_{rr}^{(v)}$	$n_{r\bullet}^{(v)}$
	Margins	$n_{\bullet 1}^{(v)}$	⋯	$n_{\bullet r}^{(v)}$	$n^{(v)}$

Table 2: Data from Main sample

		True Classifications			
		(1)	⋯	(r)	Margins
Fallible	(1)	$n_{11}^{(*)}$	⋯	$n_{1r}^{(*)}$	$n_{1\bullet}$
Classifications	⋮	⋮	⋯	⋮	⋮
	(r)	$n_{r1}^{(*)}$	⋯	$n_{rr}^{(*)}$	$n_{r\bullet}$
	Margins	$n_{\bullet 1}^{(*)}$	⋯	$n_{\bullet r}^{(*)}$	n

The key concept of the above parameterisation is that both the main sample and the validation sample have a similar structure as this is described in Tables 1 and 2. However, for the validation sample full information exists while for the main sample we have only marginal information about the observed classifications. The idea is to formulate a model by combining information from the main sample and from the validation sample. The basic assumption is that these two samples share common parameters because both are assumed to be representative of the same population. This parameterisation will lead to an optimisation problem that involves missing data. This is due to the fact that the validation procedure is not applied to the units of the main sample. Assuming independence between the main and the validation sample, denoting by $D^{(c)}$ the complete data and by Θ the vector of parameters, the full data likelihood is given by

$$L\left(\Theta; D^{(c)}\right) = \prod_{i=1}^{r}\prod_{k=1}^{r}\left(P_k q_{ik}\right)^{n_{ik}^{(v)}} \prod_{i=1}^{r}\prod_{k=1}^{r}\left(P_k q_{ik}\right)^{n_{ik}^{(*)}}. \tag{3.1}$$

Taking the logarithms in both sides of (3.1) and imposing the additional constraints that

$$\sum_{k=1}^{r} P_k = 1 \tag{3.2}$$

$$\sum_{i=1}^{r} q_{ik} = 1 \text{ for fixed k} \tag{3.3}$$

we obtain the following expression for the full data log-likelihood

$$\begin{aligned} l\left(\Theta; D^{(c)}\right) = &\sum_{k=1}^{r}\sum_{i=1}^{r-1}\left[\left(n_{ik}^{(*)} + n_{ik}^{(v)}\right)\log\left(q_{ik}\right) + \left(n_{rk}^{(*)} + n_{rk}^{(v)}\right)\log\left(1 - \sum_{i=1}^{r-1} q_{ik}\right)\right] + \sum_{k=1}^{r-1}\left(n_{\bullet k}^{(v)} + n_{\bullet k}^{(*)}\right)\log\left(P_k\right) \\ &+ \left(n_{\bullet r}^{(v)} + n_{\bullet r}^{(*)}\right)\log\left(1 - \sum_{k=1}^{r-1} P_k\right). \end{aligned} \tag{3.4}$$

The log-likelihood function (3.4) contains unobserved data. One way of using this likelihood to maximise the likelihood of the observed data is via the EM algorithm (Dempster, Laird and Rubin 1977). The EM algorithm is based on two steps, namely the expectation (E-step) and the maximisation (M-step). Generally speaking, the algorithm is initialised using a set of arbitrarily selected starting values for the parameters involved in the model. Based on these starting values, in the E-step sufficient

statistics defined by the complete data likelihood (e.g. equation (3.4)) are replaced by their conditional expectations given the observed data and the current set of parameter estimates. Having computed these conditional expectations, the full data likelihood is maximised to produce a new set of maximum likelihood estimates. Using this new set of maximum likelihood estimates, new conditional expectations are computed in the E-step and new maximum likelihood estimates are derived in the M-step. The E and M step are iterated until a convergence criterion is satisfied. For the currently described model these steps are described below.

We start by taking the conditional expectation of the full data log-likelihood given the observed data and the current estimates. We denote by $D^{(v)}$ the data derived from the validation sample, by $D^{(m)}$ the data derived from the main sample and by (h) the current EM iteration. The form of the full data log-likelihood after taking the conditional expectations becomes

$$\begin{aligned} E\left[l\left(\Theta;D^{(c)}\right) \mid D^{(v)},D^{(m)},\Theta^{(h)}\right] &= \sum_{k=1}^{r}\sum_{i=1}^{r-1}\left[E\left[\left(n_{ik}^{(*)}+n_{ik}^{(v)}\right) \mid D^{(m)},D^{(v)},\Theta^{(h)}\right]\log\left(q_{ik}\right)\right. \\ &+E\left[\left(n_{rk}^{(*)}+n_{rk}^{(v)}\right) \mid D^{(m)},D^{(v)},\Theta^{(h)}\right]\log\left(1-\sum_{i=1}^{r-1}q_{ik}\right)\Bigg] + \sum_{k=1}^{r-1}E\left[\left(n_{\bullet k}^{(v)}+n_{\bullet k}^{(*)}\right) \mid D^{(m)},D^{(v)},\Theta^{(h)}\right]\log\left(P_k\right) \\ &+E\left[\left(n_{\bullet r}^{(v)}+n_{\bullet r}^{(*)}\right) \mid D^{(m)},D^{(v)},\Theta^{(h)}\right]\log\left(1-\sum_{k=1}^{r-1}P_k\right). \end{aligned} \tag{3.5}$$

Under this parameterisation, unobserved quantities exist only in the main sample. The expectation step (E-step) and the maximisation step (M-step) can be performed using the following two results.

Result 3.1

For the E-step, the conditional expectations of the missing data in the main sample are estimated using the following expressions

$$\hat{E}\left(n_{ik}^{(*)} \mid D^{(m)},\Theta^{(h)}\right) = n_{i\bullet}\left[\frac{\hat{P}_k^{(h)}\,\hat{q}_{ik}^{(h)}}{\sum_{k=1}^{r}\hat{P}_k^{(h)}\,\hat{q}_{ik}^{(h)}}\right] \text{ and } \hat{E}\left(n_{\bullet k}^{(*)} \mid D^{(m)},\Theta^{(h)}\right) = \sum_{i=1}^{r}\hat{E}\left(n_{ik}^{(*)} \mid D^{m},\Theta^{(h)}\right). \tag{3.6}$$

Proof

Proof of Result 3.1 is given in Appendix A.

Result 3.2

For the M-step, the maximum likelihood estimators are given by the following expressions

$$\hat{q}_{ik} = \frac{\hat{E}\left(n_{ik}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{ik}^{(v)}}{\hat{E}\left(n_{\bullet k}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{\bullet k}^{(v)}} \text{ and } \hat{P}_k = \frac{\hat{E}\left(n_{\bullet k}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{\bullet k}^{(v)}}{\sum_{k=1}^{r} \hat{E}\left(n_{\bullet k}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{\bullet k}^{(v)}}. \quad (3.7)$$

Proof

Proof of Result 3.2 is given in Appendix A.

Identification of the model parameters can be checked by initialising the EM algorithm from different starting values and seeing whether the algorithm converges to the same solution. We assume that convergence is achieved when the difference between the maximum likelihood estimates obtained from two successive iterations of the EM algorithm, as this is measured for example by the $L^2 -$ norm, is less than a small value δ.

Independence between the main and the validation sample is not guaranteed only under a double sampling scheme where the validation sample is selected independently from the main sample. Alternatively, one can select a validation sample by sub-sampling $n^{(v)}$ units from the main sample and then form two samples i.e. one in which the sampled units participate only in the main survey and another in which the sampled units participate both in the main and in the validation survey.

As we will later illustrate, in a cross-sectional framework the parameterisation of the misclassification model using the calibration probabilities (Section 2.2) or the misclassification probabilities (current section) will lead to identical results. However, in some cases the use of the misclassification probabilities is more reasonable than the use of the calibration probabilities. Such a case, where the standard measurement device is a panel survey and a cross-sectional validation survey is used, is presented in Singh and Rao (1995). Due to the cross-sectional nature of the validation data that are used in this paper, a conditional independence assumption is employed in order the parameters of the longitudinal misclassification mechanism to be identified. More specifically, the authors assume that the misclassification at time t depends only on the current true state and not on previous or future

true states. It has been proven by Meyer (1988) that this assumption should be used only in conjunction with the misclassification and not with the calibration probabilities. Therefore, the formulation of the misclassification model using the misclassification probabilities offers a robust basis for extending the model to handle more complex situations.

3.2 Variance Estimation for the Maximum Likelihood Adjusted Estimates

Variance estimation for the maximum likelihood adjusted estimates can be placed into the general framework of maximum likelihood estimation. This implies the use of the inverse of the information matrix. However, due to the formulation of the misclassification model as a missing data problem, the variance estimates should account for the additional variability introduced by the existence of missing data in the main sample. One way to perform variance estimation in an EM framework is by applying the Missing Information Principle (Louis 1982).

Denote by $\hat{\Theta}$ the vector of maximum likelihood estimates and by $Z^{(m)}, Z^{(v)}$ the missing data in the main and in the validation sample respectively. The missing data and the observed data $D^{(m)}, D^{(v)}$ define the complete data $D^{(c)}$. The Missing Information Principle is defined as

$$Observed\ Information = Complete\ Information - Missing\ Information \tag{3.8}$$

Lemma 3.1 (Louis 1982)

The complete information is evaluated at $\hat{\Theta}$ using the following expression

$$Complete\ Information = E\left[-\frac{\partial^2 l\left(\Theta; D^{(c)}\right)}{\partial\Theta\partial\Theta^T} \mid D^{(m)}, D^{(v)}\right]. \tag{3.9}$$

Lemma 3.2 (Louis 1982)

The missing information is evaluated at $\hat{\Theta}$ using the following expression

$$Missing\ Information = Var\left[\frac{\partial l\left(\Theta; D^{(c)}\right)}{\partial\Theta} \mid D^{(m)}, D^{(v)}\right]. \tag{3.10}$$

Lemma 3.3

Conditionally on the information in the main sample, there are r multinomial distributions defined by the r rows of the cross-classification of the observed with the true classifications (see for example Table 2).

Proof

Proof of Lemma 3.3 is given in Appendix A.

The evaluation of the expectation of the complete information matrix involves the second order derivatives of the log-likelihood function with respect to the vector of parameters. The evaluation of the covariance matrix of the score functions involves the first order derivatives of the log-likelihood function with respect to the vector of parameters. Under simple random sampling, the variance of the score functions can be computed using Lemma 3.3 and standard results for the variance of a sum of binomial random variables. However, even for the 2-state misclassification model, the evaluation of the elements of this covariance matrix is tedious. Instead, we can approximate the components of the Missing Information Principle using a simulation-based procedure. Having arrived at the maximum likelihood estimator, we generate H complete data sets (main samples) by drawing

$$Z_1^{(m)}, Z_2^{(m)}, \ldots, Z_H^{(m)} \overset{iid}{\sim} p\left(Z^{(m)} \mid D^{(m)}, \hat{\Theta}\right) \tag{3.11}$$

where $p\left(Z^{(m)} \mid D^{(m)}, \hat{\Theta}\right)$ denotes the conditional distribution of the missing data in the main sample given the observed data and the maximum likelihood estimates and H denotes the total number of simulations. This conditional distribution is defined by Lemma 3.3. This first step of the simulation can be viewed as the imputation step. Having replaced the missing data with imputed values in simulation (h), we derive complete data $D^{(c)(h)}$ that are employed for evaluating the complete information matrix and the missing information matrix. This is done by using the simulation-based (empirical) estimators for the complete information matrix and for the variance of the score functions over simulations defined as

$$E\left[-\frac{\partial^2 l\left(\Theta; D^{(c)}\right)}{\partial\Theta\ \partial\Theta^T} \mid D^{(m)}, D^{(v)}\right] = \frac{1}{H}\sum_{h=1}^{H} -\frac{\partial^2 l\left(\Theta; D^{(c)(h)}\right)}{\partial\Theta\ \partial\Theta^T}, \tag{3.12}$$

$$Var\left[\frac{\partial l\left(\Theta; D^{(c)}\right)}{\partial\Theta} \mid D^{(m)}, D^{(v)}\right] = \frac{1}{H}\sum_{h=1}^{H}\left\{\frac{\partial l\left(\Theta; D^{(c)(h)}\right)}{\partial\Theta} - E\left[\frac{\partial l\left(\Theta; D^{(c)(h)}\right)}{\partial\Theta}\right]\right\}^2 \tag{3.13}$$

Having evaluated the complete information matrix and the missing information matrix, the covariance matrix of $\hat{\Theta}$ is then determined by the inverse of the matrix defined by the difference of these two matrices

$$Var\left(\hat{\Theta}\right) = \left\{E\left[-\frac{\partial^2 l\left(\Theta; D^{(c)}\right)}{\partial\Theta\ \partial\Theta^T} \mid D^{(m)}, D^{(v)}\right] - Var\left[\frac{\partial l\left(\Theta; D^{(c)}\right)}{\partial\Theta} \mid D^{(m)}, D^{(v)}\right]\right\}^{-1}. \tag{3.14}$$

4. A QUASI-LIKELIHOOD PARAMETERISATION OF THE MISCLASSIFICATION MODEL

In this section we present a quasi-likelihood parameterisation of the misclassification model. This parameterisation offers an alternative to the EM algorithm way of resolving a missing data problem. The advantage of this approach is that it does not require any explicit definition of the likelihood function.

The approach we follow was introduced by Wedderburn (1974) as a basis for fitting generalised linear regression models. As described in Heyde (1997), Wedderburn observed that from a computational point of view the only assumptions for fitting such a model are the specification of the mean and of the relationship between the mean and the variance and not necessarily a fully specified likelihood. Under this approach, Wedderburn replaced the assumptions about the underlying probability distribution by assumptions based solely on the mean variance relationship leading to an estimating function with properties similar to those of the derivative of a log-likelihood. This estimating function is usually referred to as the quasi-score estimating function. The quasi-likelihood estimator is then defined as the solution of the system of equations defined by the quasi-score estimating function. To illustrate, consider the following model

$$Y = \mu(\Theta) + \varepsilon \tag{4.1}$$

where Y is a $n \times 1$ data vector and $E(\varepsilon) = 0$. The quasi-score estimating function $G(\Theta)$ is then defined (Heyde 1997 Theorem 2.3) as

$$G(\Theta) = \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\right)^T [Var(\varepsilon)]^{-1} \varepsilon. \tag{4.2}$$

The quasi-score estimating function defined by (4.2) is also referred in the literature as Wedderburn's quasi-score estimating function.

4.1 The Model

Denote by $P_k^{(v)}$ the probability of correct classification in category k for units in the validation sample, by $q_{ik}^{(v)}$ the probability of misclassification for units in the validation sample, by $n_{i\cdot}$ the number of units in the main survey classified in category i by the standard measurement device and by n the sample size of the main survey. Without loss of generality, we describe the model for the case of two mutually exclusive states to which a sample unit can be classified. Instead of specifying the form of the likelihood function, the model can now be described by a system of equations. The number of equations we need is defined by the smallest possible set of independent estimating equations that can be established for the underlying problem. For the two-state cross-sectional misclassification model one possible system of equations is

$$\left.\begin{aligned} \hat{P}_1^{(v)} &= P_1 + \varepsilon_1 \\ \hat{q}_{11}^{(v)} &= q_{11} + \varepsilon_2 \\ \hat{q}_{12}^{(v)} &= q_{12} + \varepsilon_3 \\ n_{1\cdot} &= n\left[P_1 q_{11} + (1 - P_1) q_{12}\right] + \varepsilon_4. \end{aligned}\right\} \tag{4.3}$$

Note that $n_{1\cdot} = n \overset{\wedge}{pr}\left(Y_\xi^* = 1\right)$. The left hand side of the equations given in (4.3) describes estimates obtained from the main and the validation sample whereas the right hand side describes the unknown parameters of interest plus an error term. Equations described by (4.3) incorporate the extra constraints that are also utilised by the maximum likelihood approach. For the currently described model,

$P_2 = 1 - P_1, q_{21} = 1 - q_{11}$ and $q_{22} = 1 - q_{12}$. Likewise in a maximum likelihood framework, in a quasi likelihood framework we assume that the main and the validation sample share common parameters due to the fact that both are representative of the same population.

Assuming the general form of the model defined by (4.1), denote by ε the vector of errors, by $\mu(\Theta)$ the vector of means and by $\Theta = (P_1, q_{11}, q_{12})$ the vector of parameters. Following Heyde (1997), Wedderburn's quasi-score estimating function is then defined as

$$G(\Theta) = \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\right)^T [Var(\varepsilon)]^{-1} \varepsilon. \tag{4.4}$$

Equation (4.4) for the two-state model can be expressed as follows

$$G(\Theta) = \begin{pmatrix} 1 & 0 & 0 & n(q_{11} - q_{12}) \\ 0 & 1 & 0 & nP_1 \\ 0 & 0 & 1 & n(1 - P_1) \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}^{-1} \begin{pmatrix} \hat{P}_1^{(v)} - P_1 \\ \hat{q}_{11}^{(v)} - q_{11} \\ \hat{q}_{12}^{(v)} - q_{12} \\ \hat{n}_{1\cdot} - n\left[P_1 q_{11} + (1 - P_1) q_{12}\right] \end{pmatrix}. \tag{4.5}$$

Setting (4.5) equal to zero then leads to three quasi-score normal equations. These equations need to be solved using numerical techniques.

In the system of quasi-score normal equations defined by (4.5), the elements of the covariance matrix of the error terms are unknown and need to be estimated using the sample data.

Under simple random sampling (i.e. assuming a multinomial distribution), σ_1^2, σ_4^2 can be estimated by

$$\left.\begin{aligned} \hat{\sigma}_1^2 &= \frac{\hat{P}_1^{(v)}\left(1 - \hat{P}_1^{(v)}\right)}{n^{(v)}} \\ \hat{\sigma}_4^2 &= n\,\hat{pr}\left(Y_\xi^* = 1\right)\left[1 - \hat{pr}\left(Y_\xi^* = 1\right)\right]. \end{aligned}\right\} \tag{4.6}$$

In order to estimate the covariance matrix of the estimates of the misclassification probabilities, we denote by $n_{ik}^{(v)}$ the number of sample units in the validation sample classified by the standard measurement device in state i when they truly belong in state k. The estimated misclassification

probabilities are denoted by $\hat{q}_{ik} = \dfrac{n_{ik}^{(v)}}{\sum_{i=1}^{r} n_{ik}^{(v)}}$ and the estimated matrix of misclassification probabilities by $\hat{Q}$. While $n^{(v)} = \sum_{i=1}^{r}\sum_{k=1}^{r} n_{ik}^{(v)}$ can be considered as fixed, $\sum_{i=1}^{r} n_{ik}^{(v)}$ must be considered as random. Consequently, in the computation of this covariance matrix we must take into account the non-linearity introduced by the fact that both the numerator and the de-numerator of $\hat{q}_{ik}$ are random quantities. Therefore, we apply the Delta method (Bishop, Fienberg and Holland 1975). Let $\hat{\Theta}^{*} = \left(n_{11}^{(v)}, n_{21}^{(v)}, n_{12}^{(v)}, n_{22}^{(v)}\right)$ and $vec\left[Q\left(\hat{\Theta}^{*}\right)\right] = \left[f_1\left(\hat{\Theta}^{*}\right),\ldots,f_{r^2}\left(\hat{\Theta}^{*}\right)\right]^{T}$ be an $r^2 \times 1$ vector of functions of $\hat{\Theta}^{*}$. Applying the delta method to $vec\left[Q\left(\hat{\Theta}^{*}\right)\right]$ we obtain the following approximation

$$vec\left[Q\left(\hat{\Theta}^{*}\right)\right] - vec\left[Q\left(\Theta^{*}\right)\right] \approx \nabla_{\Theta^{*}}\left(\hat{\Theta}^{*} - \Theta^{*}\right), \quad \nabla_{\Theta^{*}} = \frac{\partial vec\left[Q\left(\Theta^{*}\right)\right]}{\partial \Theta^{*}}\Big|_{\Theta^{*}=\hat{\Theta}^{*}} . \tag{4.7}$$

Taking the variance operator on both sides of (4.7) leads to

$$Var\left[vec\left(Q\right)\right] \approx \nabla_{\Theta^{*}} Var\left(\hat{\Theta}^{*}\right)\left(\nabla_{\Theta^{*}}\right)^{T} . \tag{4.8}$$

Under simple random sampling, $Var\left(\hat{\Theta}^{*}\right)$ can be estimated using the following results

$$\begin{cases} \hat{Var}\left(n_{ik}^{(v)}\right) = n^{(v)}\ \hat{pr}\left(Y_{\xi}^{*} = i, Y_{\xi} = k\right)\left[1 - \hat{pr}\left(Y_{\xi}^{*} = i, Y_{\xi} = k\right)\right] \\ \hat{Cov}\left(n_{ik}^{(v)}, n_{i^{*}k^{*}}^{(v)}\right) = -n^{v}\ \hat{pr}\left(Y_{\xi}^{*} = i, Y_{\xi} = k\right)\hat{pr}\left(Y_{\xi}^{*} = i^{*}, Y_{\xi} = k^{*}\right) \quad (i,k) \neq \left(i^{*}, k^{*}\right) \end{cases} \tag{4.9}$$

Substituting (4.9) and the Jacobian matrix into (4.8) we obtain estimates for σ_2^2, σ_3^2 and $\sigma_{23} = \sigma_{32}$.

For estimating the covariance terms $\sigma_{14}, \sigma_{24}, \sigma_{34}$ we need to consider the double sampling scheme that we employ. In Section 3.1, we mentioned that independence between the main and the validation sample can be assumed both when the validation sample is selected independently from the main sample and when the validation sample is selected by sub-sampling units from the main sample. While

for the former case this argument is clear, for the latter case more explanation is required. More specifically, for the latter case independence is guaranteed by splitting the sample into sample units that participate only in the main survey and sample units that participate both in the main and in the validation survey. Under the assumption of independence between the main and the validation sample, the following holds

$$\sigma_{14} = \sigma_{41} = \sigma_{24} = \sigma_{42} = \sigma_{34} = \sigma_{43} = 0. \tag{4.10}$$

It only remains to estimate the following covariance terms: $\sigma_{12} = \sigma_{21}$ and $\sigma_{13} = \sigma_{31}$. These covariance terms can be more generally defined as follows

$$Cov\left(\hat{q}_{ik}^{(v)}, \hat{P}_{k}^{(v)}\right) = Cov\left(\frac{n_{ik}^{(v)}}{\sum_{i=1}^{r} n_{ik}^{(v)}}, \frac{\sum_{i=1}^{r} n_{ik}^{(v)}}{n^{(v)}}\right). \tag{4.11}$$

Estimation of these covariance terms is performed using the results below.

Lemma 4.1 (Mood et.al. 1963)

An approximate expression for the expectation of a function $g(X,Y)$ of two random variables X,Y using a Taylor's series expansion around (μ_X, μ_Y) is given by

$$\begin{aligned} E[g(X,Y)] \approx\ & g(\mu_X,\mu_Y) + \frac{1}{2}\frac{\partial^2}{\partial y^2} g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(Y) + \frac{1}{2}\frac{\partial^2}{\partial x^2} g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(X) \\ & + \frac{\partial^2}{\partial x \partial y} g(X,Y)\,|_{\mu_X,\mu_Y}\ Cov(X,Y). \end{aligned} \tag{4.12}$$

Result 4.1

Assume that X,Y,A are three random variables and n is fixed. A first order approximation for $Cov\left(\frac{X}{Y}, \frac{A}{n}\right)$ is given by

$$Cov\left(\frac{X}{Y}, \frac{A}{n}\right) \approx \frac{1}{nE(Y)}\left[Cov(A,X) - \frac{E(X)}{E(Y)} Cov(A,Y)\right]. \tag{4.13}$$

Proof

Proof of this result is given in Appendix B.

Setting $X = n_{ik}^{(v)},\ Y = \sum_{i=1}^{r} n_{ik}^{(v)},\ A = \sum_{i=1}^{r} n_{ik}^{(v)}$ and $n = n^{v}$ in Result 4.1, we can then estimate the remaining covariance terms of interest.

Having obtained estimates for the variance terms, the final step in deriving the quasi- likelihood estimators is to solve the system of equations defined by (4.5). This can be done using a Newton-Raphson algorithm. Define by Θ the vector of parameters of dimension $\omega \times 1$, and by A a $\omega \times \omega$ matrix with elements $A_{ij} = \dfrac{\partial G_i(\Theta)}{\partial \vartheta_j}$ $i, j = 1, \cdots, \omega$. The system of quasi-score normal equations defined by (4.5) can be now solved numerically as follows. Assume a vector of initial solutions $\hat{\Theta}^{(0)}$. The vector of initial solutions can be updated using

$$\hat{\Theta}^{(1)} = \hat{\Theta}^{(0)} - A^{-1}\left[\hat{\Theta}^{(0)}\right] G\left[\hat{\Theta}^{(0)}\right]. \tag{4.14}$$

The iterations continue until a pre-specified convergence criterion is satisfied.

4.2 Variance Estimation for the Quasi-likelihood Adjusted Estimates

Variance estimation for the quasi-likelihood adjusted estimates is performed using the following result.

Result 4.2

The variance of the quasi-likelihood adjusted estimates is estimated using the expression below

$$\hat{Var}\left(\hat{\Theta}\right) \approx \left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^{T}\left[\hat{Var}(\varepsilon)\right]^{-1}\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)\right\}^{-1}. \tag{4.15}$$

Proof

Proof of this result is given in Appendix B.

Although some work is required in order to derive $\hat{Var}(\varepsilon)$, the evaluation of the covariance matrix of the quasi-likelihood estimates is straightforward since it requires the utilization of matrix quantities that have been already used during the estimation process. Therefore, unlike variance estimation in an

EM context that requires the use of computer intensive techniques, variance estimation under the quasi-likelihood parameterisation may be more appealing to the data analyst.

5. APPLICATIONS

5.1 Application 1

The alternative approaches are illustrated using the following numerical example. A firm wishes to assess the quality of the units that it produces. The units can be classified into two categories: Defective (1) or Satisfactory (2). There are two classification methods. One, which is currently used, is inexepnsive and is subject to measurement error. Altenatively, the firm can use an accurate but more expensive classification method. The firm suspects that a number of truly satisfactory units are classified as defective. The management team is interested in investigating the trade-off between the loss of satisfactory units and the extra cost of improving the currently used classifier. A sample of $n = 60000$ production units is selected and the units are classified using the inexpensive classification method. In order to validate the inexpensive classifier, another sample of $n^{(v)} = 10000$ production units is selected and these units are classified using both the expensive and the inexpensive classifier. The data for this numerical example are given in Tables 3 and 4. The estimators we consider are the undajusted for misclassification estimator (Naïve), the maximum likelihood estimator (MLE) with callibration probabilities (Section 2.2), the maximum likelihood estimator (MLE) with misclassification probabilities (Section 3.1) and the quasi-likelihood estimator (Section 4.1).

Table 3: Data from the Validation Sample Used in Application 5.1

		True Classifications		
		Defective (1)	Satisfactory (2)	Margins
Fallible Classifications	Defective (1)	672	918	1590
	Satisfactory (2)	28	8382	8410
	Margins	700	9300	$n^{(v)} = 10000$

Table 4: Data from the Main Sample Used in Application 5.1

		True Classifications		
		Defective (1)	Satisfactory (2)	Margins
Fallible Classifications	Defective (1)	$n_{11}^{(*)}$	$n_{12}^{(*)}$	9000
	Satisfactory(2)	$n_{21}^{(*)}$	$n_{22}^{(*)}$	51000
	Margins	$n_{\bullet 1}^{(*)}$	$n_{\bullet 2}^{(*)}$	$n = 60000$

Table 5: Proportion of Units Classified as Defective Under the Alternative Parameterisations of the Misclassification Model, Standard Deviations in Parentheses

Point Estimator	Proportion of Units Classified as Defective
Naïve (Unadjusted Estimator)	0.1512 ($1.35*10^{-3}$)
MLE (Calibration Probabilities)	0.0667 ($2.08*10^{-3}$)
MLE (Misclassification Probabilities)	0.0667 ($2.07*10^{-3}$)
Quasi-likelihood	0.0669 ($2.11*10^{-3}$)

The convergence criterion for the EM algorithm and for the Newton-Raphson is $\delta = 10^{-6}$. An empirical investigation of the identifiability of the model parameters is provided by initialising the EM and the Newton-Raphson algorithms using different sets of starting values and checking whether these algorithms converge to the same value. Using this diagnostic, we conclude that the model parameters are identified. All three estimators that correct for measurement error produce similar results. The management team of the firm can now decide whether it worths investing in improving the inexpensive classification procedure. The variance of the naïve (unadjusted) estimator is computed under simple random sampling assuming a multinomial distribution. The variance of the maximum likelihood estimator that employs the calibration probabilities is computed using the result of Tenenbein (1972, p.191). The variance of the maximum likelihood estimator, under the EM algorithm,

and of the quasi-likelihood estimator is computed using the results from Sections 3.2 and 4.2 respectively.

Assuming that the validation survey provides an unbiased estimate of the proportion of units that are trully classified as Defective, we can then examine the efficiency of the alternative estimators. Although the estimators that account for measurement error have higher variances than the estimator that ignores measurement error, they are unbiased. On the other hand, the estimator that ignores measurement error is seriously biased. Therefore, in mean squared error terms, we conclude that accounting for measurement error will produce more efficient estimates than ignoring measurement error.

5.2 Application 2: Contratsing the Alternative Parameterisations in the Presence of Intense Misclassification

In Section 2.1 we mentioned that in the presence of intense misclassification the moment-based estimator can produce estimates that lie outside the parameter space. We now utilise a numerical example for comparing the alternative parameterisations when intense misclassification exists. A sample of $n = 60000$ units is selected and the units are classified into two mutually exclusive categories, denoted by (1) and (2), using an inexpensive classification method. In order to validate the inexpensive classifier, another sample of $n^{(v)} = 20000$ units is selected and these units are classified using both the expensive and the inexpensive classifier. The data for this numerical example are given in Tables 6 and 7. The estimators we consider are the moment-based estimator (Section 2.1), the maximum likelihood estimator (MLE) with callibration probabilities (Section 2.2), the maximum likelihood estimator (MLE) with misclassification probabilities (Section 3.1) and the quasi-likelihood estimator (Section 4.1).

Table 6: Data from the Validation Sample Used in Application 5.2

		True Classifications		
		(1)	(2)	Margins
Fallible	(1)	500	4400	4900
Classifications	(2)	500	14600	15100
	Margins	1000	19000	$n^{(v)} = 20000$

Table 7: Data from the Main Sample Used in Application 5.2

		True Classifications		
		(1)	(2)	Margins
Fallible	(1)	$n_{11}^{(*)}$	$n_{12}^{(*)}$	13600
Classifications	(2)	$n_{21}^{(*)}$	$n_{22}^{(*)}$	46400
	Margins	$n_{\bullet 1}^{(*)}$	$n_{\bullet 2}^{(*)}$	$n = 60000$

Table 8: Adjusted Proportion of Units Classified in Each Category Under the Alternative Parameterisations of the Misclassification Model

Point Estimators	Category (1)	Category (2)
Moment-based	-0.0012	1.0012
MLE (Calibration Probabilities)	0.0491	0.9509
MLE (Misclassification Probabilities)	0.0491	0.9509
Quasi-likelihood	0.0488	0.9512

The intensity of the misclassification problem can be seen by noticing that 50% of the units that truly belong in category (1) are misclassified as being in category (2) (see Table 6). The results (Table 8) indicate that when high misclassification exists, the moment-based estimator can lead to ackward estimates (in this case negative proportions). On the other hand, both the maximum likelihood estimators and the quasi-likelihood estimator produce estimates that lie within the boundaries of the parameter space.

6. SIMULATION STUDY

6.1 Design and Implementation of the Simulation Algorithm

The alternative parameterisations are empirically compared using a Monte-Carlo simulation study. The simulation algorithm simulates the measurement error process in a double sampling framework. The algorithm involves four steps. At the first step we generate true classifications for each sample unit ξ. This is done by assuming the probability distribution function of the true classifications. Using this distribution, we draw a with replacement sample of size $n = 60000$. The probabilities of correct classification that we use are $pr\left(Y_\xi = 1\right) = 0.606$ and $pr\left(Y_\xi = 2\right) = 0.394$. At the second step we assume the existence of measurement error that is described by the misclassification probabilities q_{ik}. Using these misclassification probabilities, we generate the observed status, given the true status (Step 1), for each sample unit ξ. The diagonal elements of the misclassification matrix that we employ are $q_{11} = 0.98$ and $q_{22} = 0.96$. According to the methodology that assumes the availability of validation information the estimation of the matrix of misclassification probabilities is based on the validation sample. We simulate a validation sample by selecting a with replacement sample of $n^{(v)} = 3000$ independently from the main sample and from the same target population. The probability structure of $pr\left(Y_\xi^* = i, Y_\xi = k\right)$ is defined as follows: $pr\left(Y_\xi^* = 1, Y_\xi = 1\right) = 0.593$, $pr\left(Y_\xi^* = 1, Y_\xi = 2\right) = 0.016$, $pr\left(Y_\xi^* = 2, Y_\xi = 1\right) = 0.012$, $pr\left(Y_\xi^* = 2, Y_\xi = 2\right) = 0.379$. The first three steps summarise the generation process. At the final step we employ the generated data for computing the alternative estimators. We conduct a total of $H = 1000$ simulations and we empirically evaluate the properties of the alternative point estimators (averages over simulations) using (a) the bias of a point estimator, (b) the variance of a point estimator and (c) the mean squared error of a point estimator.

6.2 Results

The results from the Monte-Carlo simulation are summarised in Table 9.

Table 9: Monte-Carlo Simulation Results for Comparing the Alternative Parameterisations of the Misclassification Model

Point Estimators	Estimates	Bias	Variance	Mean Squared Error
Moment-based	0.6061	$1*10^{-4}$	$1.40*10^{-5}$	$1.40*10^{-5}$
MLE (Calibration Probabilities)	0.6059	$-1*10^{-4}$	$1.28*10^{-5}$	$1.28*10^{-5}$
MLE (Misclassification Probabilities)	0.6059	$-1*10^{-4}$	$1.28*10^{-5}$	$1.28*10^{-5}$
Quasi-likelihood	0.6059	$-1*10^{-4}$	$1.28*10^{-5}$	$1.28*10^{-5}$

7. DISCUSSION

Two alternative parameterisations for maximum likelihood estimation using either the calibration or the misclassification probabilities are presented. In a cross-sectional framework, both parameterisations for maximum likelihood estimation lead to identical results. However, it has been shown (Mayer 1988) that the use of misclassification probabilities instead of the calibration probabilities is more reasonable when handling more complex situations for example, when analysing longitudinal misclassified data. Thus, we suggest that the parameterisation of the misclassification model as a missing data problem using the misclassification probabilities provides a more general method and it should be preferred.

As an alternative approach, we further presented a quasi-likelihood parameterisation of the misclassification model. This approach offers an alternative to the EM algorithm way of resolving a missing data problem, which at the same time does not require full specification of the likelihood function. The results from the simulation study show that the quasi-likelihood estimator is as efficient as the maximum likelihood estimator. A further advantage that the quasi-likelihood parameterisation

offers is an easier way of performing variance estimation for the adjusted estimates. More specifically, variance estimation for the maximum likelihood estimator when using the EM framework involves the application of the Missing Information Principle and may require the use of computer intensive methods. On the other hand, variance estimation in a quasi-likelihood framework is much simpler and requires the utilization of matrix quantities that have been already used during the estimation process.

In section 2.1, we reviewed a moment-based estimator and mentioned that one of the main disadvantages associated with the use of this estimator is that it can produce estimates that lie outside the parameter space for example, negative adjusted estimates. Unlike the moment-based estimator, both the maximum likelihood estimators and the quasi-likelihood estimator constrain the derived estimates to lie within the boundaries of the parameter space. Moreover, the simulation study verifies the superiority of the maximum likelihood and the quasi-likelihood estimators when compared to the moment-based estimator.

Currently, we investigate the extension of the quasi-likelihood approach for analysing longitudinal misclassified data. We also examine ideas for applying the misclassification model in other areas of statistical inference. For example, in demographic applications one of the most commonly encountered problems is heaping. A traditional way of resolving this problem is via the use smoothing techniques. An alternative solution can be offered by viewing heaping as a misclassification problem. A further possible application is in statistical disclosure control. More specifically, one way of protecting the data is by misclassifying them and providing to the data analyst the misclassified data along with the misclassification probabilities (Van den Hout and Van der Heijden 2002). The basic difference between the approach utilised in statistical disclosure control and our approach is that in the former case the misclassification probabilities are treated as fixed whereas in the latter case the misclassification probabilities are random since they are estimated from the validation survey. Finally, another application regards adjustments in the Census. For example, the existence of inaccurate addresses can result in the erroneous estimation of the population size in an area. This problem can be

described in a misclassification context and a model that combines information from the Census (main survey) and from a post-enumeration survey (validation survey) can provide adjustments to the Census-based estimates.

APPENDIX A: PROOFS OF THE RESULTS IN SECTION 3

Proof of Result 3.1

The number of sample units that belong in the ik cell of the cross-classification of the observed by the true classification is denoted by $n_{ik}^{(*)}$. Note that while a superscript $(*)$ refers to the unobserved quantities, a superscript $*$ refers to the observed classifications. The expectation of an unobserved quantity is given by

$$E\left(n_{ik}^{(*)}\right) = nE\left(Y_{\xi}^{*} = i, Y_{\xi} = k\right). \tag{A.1}$$

Equation (A.1) can be re-expressed as follows

$$E\left(n_{ik}^{(*)}\right) = nE\left(Y_{\xi}^{*} = i \mid Y_{\xi} = k\right) E\left(Y_{\xi} = k\right). \tag{A.2}$$

From the main sample we have information about the observed classifications $n_{i\bullet}$.

$$n_{i\bullet} = n\sum_{k=1}^{r} E\left(Y_{\xi}^{*} = i \mid Y_{\xi} = k\right) E\left(Y_{\xi} = k\right). \tag{A.3}$$

Given the data constraints $n_{i\bullet}$, the conditional expectations of the unobserved quantities are given below

$$E\left(n_{ik}^{(*)} \mid D^{(m)}\right) = n_{i\bullet}\left[\frac{E\left(Y_{\xi}^{*} = i \mid Y_{\xi t} = k\right) E\left(Y_{\xi} = k\right)}{\sum_{k=1}^{r} E\left(Y_{\xi}^{*} = i \mid Y_{\xi} = k\right) E\left(Y_{\xi} = k\right)}\right]. \tag{A.4}$$

The expectations of the random variables involved in the expression above can be computed using well known results for binomial random variables. More specifically,

$$pr\left(Y_{\xi}^{*} = i \mid Y_{\xi} = k\right) = q_{ik}\,,\ pr\left(Y_{\xi} = k\right) = P_{k}. \tag{A.5}$$

Substituting (A.5) into (A.4) we obtain the required result

$$\hat{E}\left(n_{ik}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) = n_{i\bullet}\left[\frac{\hat{P}_k^{(h)}\,\hat{q}_{ik}^{(h)}}{\sum_{k=1}^{r}\hat{P}_k^{(h)}\,\hat{q}_{ik}^{(h)}}\right].$$

It follows that

$$\hat{E}\left(n_{\bullet k}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) = \sum_{i=1}^{r}\hat{E}\left(n_{ik}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right).$$

Proof of Result 3.2

The system of normal equations that we need to solve in order to obtain the maximum likelihood estimators is defined by setting the score functions equal to zero i.e.

$$\frac{\partial E\left(l \mid D^{(m)}, D^{(v)}, \Theta\right)}{\partial q_{ik}} = 0. \tag{A.6}$$

The $\left(r^2 - r\right) \times \left(r^2 - r\right)$ system of normal equations and the corresponding maximum likelihood estimator for q_{ik} is given below.

$$\left.\begin{aligned} &\frac{E\left(n_{11}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{11}^{(v)}}{q_{11}} - \frac{E\left(n_{r1}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{r1}^{(v)}}{\left(1 - q_{11} - \cdots - q_{r-11}\right)} = 0 \\ &\vdots \\ &\frac{E\left(n_{r-1r}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{r-1r}^{(v)}}{q_{r-1r}} - \frac{E\left(n_{rr}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{rr}^{(v)}}{\left(1 - q_{1r} - \cdots - q_{r-1r}\right)} = 0 \end{aligned}\right\} \tag{A.7}$$

$$\hat{q}_{ik} = \frac{\hat{E}\left(n_{ik}^{(*)} \mid D^{(m)}, \Theta\right) + n_{ik}^{(v)}}{\hat{E}\left(n_{\bullet k}^{(*)} \mid D^{(m)}, \Theta\right) + n_{\bullet k}^{(v)}}. \tag{A.8}$$

Similarly, for P_k, the $(r-1)\times(r-1)$ system of normal equations and the corresponding maximum likelihood estimator is given by the following expressions

$$\left.\begin{aligned} &\frac{E\left(n_{\bullet 1}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{\bullet 1}^{(v)}}{P_1} - \frac{E\left(n_{\bullet r}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{\bullet r}^{(v)}}{\left(1 - P_1 - \ldots P_{r-1}\right)} = 0 \\ &\vdots \\ &\frac{E\left(n_{\bullet r-1}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{\bullet r-1}^{(v)}}{P_{r-1}} - \frac{E\left(n_{\bullet r}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{\bullet r}^{(v)}}{\left(1 - P_1 - \ldots P_{r-1}\right)} = 0 \end{aligned}\right\} \tag{A.9}$$

$$\hat{P}_k = \frac{\hat{E}\left(n_{\bullet k}^{(*)} \mid D^{(m)}, \Theta\right) + n_{\bullet k}^{(v)}}{\sum_{k=1}^{r} \hat{E}\left(n_{\bullet k}^{(*)} \mid D^{(m)}, \Theta\right) + n_{\bullet k}^{(v)}}. \tag{A.10}$$

Proof of Lemma 3.3

Before selecting the main sample, $n_{1\bullet}, n_{2\bullet}, \cdots, n_{r\bullet}$ are assumed to be random and the only fixed quantity is the size of the main sample n. However, the EM algorithm conditions on the information available from the main sample. Thus, conditionally on the main sample $n_{1\bullet}, n_{2\bullet}, \cdots, n_{r\bullet}$ are assume to be fixed. This implies the existence of r multinomial distributions defined by the r rows of the cross-classification of the observed with the true classifications.

APPENDIX B: PROOFS OF THE RESULTS IN SECTION 4

Proof of Result 4.1

We start the proof by expanding the covariance term of interest using the standard definition for the covariance between random variables.

$$Cov\left(\frac{X}{Y}, \frac{A}{n}\right) = E\left(\frac{X}{Y}\frac{A}{n}\right) - E\left(\frac{X}{Y}\right)E\left(\frac{A}{n}\right) = \frac{1}{n}\left[E\left(\frac{AX}{Y}\right) - E\left(\frac{X}{Y}\right)E(A)\right]. \tag{B.1}$$

We evaluate the different components of the expression above using Lemma 4.1. More specifically, we approximate $E\left(\frac{AX}{Y}\right)$ by utilising the Taylor series expansion of $\frac{AX}{Y}$ around (μ_X, μ_Y, μ_A). This Taylor series expansion is given by

$$\begin{aligned} E\left[g(X,Y,A)\right] \approx\ & g(\mu_X, \mu_Y, \mu_A) + \frac{1}{2}\frac{\partial^2}{\partial y^2} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\, Var(Y) + \frac{1}{2}\frac{\partial^2}{\partial x^2} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\, Var(X) \\ & + \frac{1}{2}\frac{\partial^2}{\partial a^2} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\, Var(A) + \frac{\partial^2}{\partial x \partial y} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\, Cov(X,Y) \\ & + \frac{\partial^2}{\partial x \partial a} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\, Cov(X,A) + \frac{\partial^2}{\partial a \partial y} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\, Cov(Y,A). \end{aligned}$$

It follows that

$$E\big[g(X,Y,A)\big] \approx \frac{\mu_X \mu_A}{\mu_Y} + \frac{1}{2}\frac{2\mu_X \mu_A}{{\mu_Y}^3} Var(Y) - \frac{\mu_A}{{\mu_Y}^2} Cov(X,Y) + \frac{1}{\mu_Y} Cov(X,A) - \frac{\mu_X}{{\mu_Y}^2} Cov(A,Y). \tag{B.2}$$

Next, we approximate $E\left(\frac{X}{Y}\right)$ using a Taylor series expansion of $\frac{X}{Y}$ around (μ_X, μ_Y).

$$E\big[g(X,Y)\big] \approx g(\mu_X,\mu_Y) + \frac{1}{2}\frac{\partial^2}{\partial y^2} g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(Y) + \frac{1}{2}\frac{\partial^2}{\partial x^2} g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(X) + \frac{\partial^2}{\partial x \partial y} g(X,Y)\,|_{\mu_X,\mu_Y}\ Cov(X,Y)$$

It follows that

$$E\big[g(X,Y)\big] \approx \frac{\mu_X}{\mu_Y} + \frac{1}{2}\frac{2\mu_X}{{\mu_Y}^3} Var(Y) - \frac{1}{{\mu_Y}^2} Cov(X,Y). \tag{B.3}$$

Substituting results (B.2) and (B.3) into (B.1) we derive the following

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) \approx \frac{1}{n}\left[\frac{\mu_X \mu_A}{\mu_Y} + \frac{1}{2}\frac{2\mu_X \mu_A}{{\mu_Y}^3} Var(Y) - \frac{\mu_A}{{\mu_Y}^2} Cov(X,Y) + \frac{1}{\mu_Y} Cov(X,A) - \frac{\mu_X}{{\mu_Y}^2} Cov(A,Y) - \frac{\mu_X \mu_A}{\mu_Y} - \frac{1}{2}\frac{2\mu_X \mu_A}{{\mu_Y}^3} Var(Y) + \frac{\mu_A}{{\mu_Y}^2} Cov(X,Y)\right].$$

It follows that

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) \approx \frac{1}{n}\left[\frac{1}{\mu_Y} Cov(X,A) - \frac{\mu_X}{{\mu_Y}^2} Cov(A,Y)\right] = \frac{1}{n\mu_Y}\left[Cov(X,A) - \frac{\mu_X}{\mu_Y} Cov(A,Y)\right].$$

Finally,

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) \approx \frac{1}{nE(Y)}\left[Cov(X,A) - \frac{E(X)}{E(Y)} Cov(A,Y)\right].$$

Proof of Result 4.2

Let $\hat{\Theta}$ denote the vector of quasi-likelihood estimates and ε the vector of errors. The quasi-score estimating function is defined by

$$G(\Theta) = \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\right)^T [Var(\varepsilon)]^{-1} \varepsilon. \tag{B.4}$$

It follows that

$$\begin{aligned} Var\left[G\left(\hat{\Theta}\right)\right] = Var\left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1} \varepsilon\right\} = \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1} Var(\varepsilon) \\ \left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1}\right\}^T. \end{aligned} \tag{B.5}$$

Taken into account that $[Var(\varepsilon)]^{-1}$ is symmetric, it follows that

$$Var\left[G\left(\hat{\Theta}\right)\right] = \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1} \left[\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T. \tag{B.6}$$

Now, $G\left(\hat{\Theta}\right)$ can be expanded as follows

$$G\left(\hat{\Theta}\right) \approx G(\Theta) + \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1} \left[\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T \left(\hat{\Theta} - \Theta\right). \tag{B.7}$$

Thus,

$$Var\left[G\left(\hat{\Theta}\right)\right] \approx Var\left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1} \left[\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T \left(\hat{\Theta} - \Theta\right)\right\}. \tag{B.8}$$

It follows that

$$\begin{aligned} Var\left[G\left(\hat{\Theta}\right)\right] \approx \left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1} \left[\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T\right\} \left[Var\left(\hat{\Theta}\right)\right] \\ \left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1} \left[\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T\right\}^T = \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1} \left[\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T. \end{aligned} \tag{B.9}$$

Solving equation (B.9) with respect to $Var\left(\hat{\Theta}\right)$ we obtain the required result.

$$\hat{Var}\left(\hat{\Theta}\right) \approx \left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T \left[\hat{Var}(\varepsilon)\right]^{-1} \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)\right\}^{-1}. \tag{B.10}$$

REFERENCES

Bailar, A.B. (1968), "Recent Research in Re-interview Procedures", *Journal of the American Statistical Association*, 63, 41-63.

Bauman, K.E., and Koch, G.C. (1983), "Validity of Self-reports and Descriptive and Analytical Conclusions: The Case of Cigarette Smoking by Adolescents and Their Mothers", *American Journal of Epidemiology*, 118, 90-98.

Bishop, Y., Fienberg, S., and Holland, P. (1975), "Discrete Multivariate Analysis", MIT Press.

Bross, I. (1954), "Misclassification in 2×2 Tables", *Biometrics*, 10, 478-486.

Caroll, R.J. (1992), "Approaches to Estimation with Error in Predictors", in *Advances in GLIM and Statistical Modelling,* eds. Fahrmeir, L., Francis, B., Gilchrist, R. and Tutz, G., 40-47, Springer.

Deming, W.E., and Stephan, F.F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known", *Annals of Mathematical Statistics*, 11, 427-444.

Deming, W.E. (1950). "Some Theory of Sampling", Wiley.

Dempster, P.A., Laird, M.N., and Rubin, B.D. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society Ser. B*, 39, 1-38.

Forsman G., and Schreiner I. (1991), "The Design and Analysis of Re-interview: An Overview", in *Measurement Error in Surveys*, eds. P.P. Biemer, R.M. Grooves, L.E. Lyberg, N.A. Mathiowetz, S. Sudman, 279-301, Wiley.

Fuller, W. A. (1987), "Measurement Error Models", Wiley.

Greenland, S. (1988), "Variance Estimation for Epidemiologic Effect Estimates Under Misclassification", *Statistics in Medicine*, 7, 745-757.

Heyde, C.C. (1997), "Quasi-Likelihood and Its Application, A General Approach to Optimal Parameter Estimation", Springer.

Hill, M.S. (1992), "The Panel Study of Income Dynamics, A User's Guide to Major Science Data Bases", 2, Sage.

Kuha, J., and Skinner, C. J. (1997), "Categorical Data Analysis and Misclassification", in *Survey Measurement and Process Quality,* eds. Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin, 633-670, Wiley.

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

Marshall, R.J. (1990), "Validation Study Methods for Estimating Exposure Proportions and Odds Ratios with Misclassified Data", *Journal of Clinical Epidemiology*, 43, 95-109.

Meyer, D, B. (1988), "Classification Error Models and Labour-Market Dynamics", *Journal of Business and Economic Statistics*, 6, 385-390.

Mood, A.M., Graybill, A.F., and Boes, C.D. (1963), "Introduction to the Theory of Statistics", McGraw-Hill.

Selen, J. (1986), "Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical", *Journal of the American Statistical Association*, 81*,* 75-81.

Singh, A.C., and Rao, J.N.K. (1995), "On the Adjustment of Gross Flows Estimates for Classification Error with Application to Data from the Canadian Labour Force Survey", *Journal of the American Statistical Association*, 90, 478-488.

Tenenbein, A. (1970), "A Double Sampling Scheme for Estimating from Misclassified Binomial Data", *Journal of the American Statistical Association*, 65*,* 1350-1361.

_____ (1972), "A Double Sampling Scheme for Estimating from Misclassified Multinomial Data", Technometrics, 14, 187-202.

Van den Hout, A., and Van der Heijden, M.G. (2002), "Randomised Response, Statistical Disclosure Control and Misclassification: A Review", *International Statistical Review*, 70, 269-288.

Wedderburn, R.W.M. (1974), "Quasi-likelihood Functions, Generalised Linear Models and Gauss-Newton Method", *Biometrika*, 61, 439-447.