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

We discuss alternative approaches for estimating from cross-sectional categorical data in the presence of misclassification. Two parameterisations of the misclassification model are reviewed. The first employs misclassification probabilities and leads to moment-based inference. The second employs calibration probabilities and leads to maximum likelihood inference. We show that maximum likelihood estimation can be alternatively performed by employing misclassification probabilities and a missing data specification. As an alternative to maximum likelihood estimation we propose a quasi-likelihood parameterisation of the misclassification model. In this context an explicit definition of the likelihood function is avoided and a different way of resolving a missing data problem is provided. Variance estimation for the alternative point estimators is considered. The different approaches are illustrated using real data from the UK Labour Force Survey and simulated data.

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 the statistical literature. In the presence of measurement error such methods need to be employed in order to ensure the validity of the inferential process. In a discrete framework, however, conventional errors in variables models (Fuller [7]) cannot 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 framework of double sampling methods (Bross [3];

Tenenbein [17, 18].) In double sampling 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 misclassification mechanism. Inference is then 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 categorical data are subject to misclassification and validation information is available. The structure of the paper is as follows. In Section 2, the framework of double sampling is presented along with moment-based and maximum likelihood inference. In Section 3, we present a quasi-likelihood parameterisation. In Section 4, the alternative approaches are illustrated using data from the UK Labour Force Survey and in Section 5 they are empirically compared using a Monte-Carlo simulation study.

2. Using double sampling to adjust for misclassification

Suppose that the standard measurement device is subject to misclassification. As a result we have biased results. Unbiased estimates can be obtained by using more elaborate measurement tools usually referred to as preferred procedures (Forsman and Schreiner [6]; Kuha and Skinner [10].) An example of a preferred procedure in official statistics is re-interview surveys (Bailar [1].) In bio-statistical applications the term gold standard is more commonly used (Bauman and Koch [2].) Other examples include judgments of experts or checks against administrative records (Greenland [8].) The assumption that the preferred procedure is free of error makes feasible the estimation of the parameters of the misclassification mechanism. On the other hand, preferred procedures are considered to be fairly expensive and thus unsuitable to be used for the entire sample hereafter 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 [10]) depending on how this sample is selected. In the most common case an internal validation sample of size $n^{(v)}$ is obtained by sub-sampling $n^{(v)}$ units from the main sample. Alternatively, an internal validation sample can be selected independently from the main sample and from the same target population. Otherwise, the validation sample is defined as external. In this paper we focus on internal designs.

2.1. Moment-based estimation. Let Y_ξ^* denote a discrete random variable for unit ξ. Denote by $\Pi_i = \mathrm{pr}(Y_\xi^* = i)$ the probability that unit ξ is classified in state i by the standard measurement device, which is subject to measurement error, by $P_k = \mathrm{pr}(Y_\xi = k)$ the probability that unit ξ truly belongs in state k and by $q_{ik} = \mathrm{pr}(Y_\xi^* = i \mid Y_\xi = k)$ 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, one way to describe the misclassification model with r mutually exclusive states is by expressing the marginal distribution of the observed classifications as a product of the

misclassification probabilities and the true classifications as follows

$$\mathrm{pr}\left(Y_\xi^* = i\right) = \sum_{k=1}^{r} \mathrm{pr}\left(Y_\xi^* = i \mid Y_\xi = k\right)\mathrm{pr}\left(Y_\xi = k\right). \tag{2.1}$$

The unknown quantities involved in (2.1) are typically estimated using double sampling. Solving (2.1) with respect to the vector of true classifications P and assuming that Q is non-singular leads to the moment-based estimator $\hat{P} = \hat{Q}^{-1}\hat{\Pi}$, which 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 have been proposed among others by Selén [15] and Greenland [8].

2.2. Maximum likelihood estimation. In order to describe the misclassification mechanism estimator $\hat{P} = \hat{Q}^{-1}\hat{\Pi}$ employs misclassification probabilities defined as $q_{ik} = \mathrm{pr}(Y_\xi^* = i \mid Y_\xi = k)$. An alternative way of quantifying the misclassification mechanism is by using what Carroll [4] refers to as calibration probabilities. The calibration probabilities are defined as $c_{ki} = \mathrm{pr}(Y_\xi = k \mid Y_\xi^* = i)$. Denote by C the matrix of calibration probabilities. The misclassification model can be alternatively described with calibration probabilities as follows

$$\mathrm{pr}\left(Y_\xi = k\right) = \sum_{i=1}^{r} \mathrm{pr}\left(Y_\xi = k \mid Y_\xi^* = i\right)\mathrm{pr}\left(Y_\xi^* = i\right). \tag{2.2}$$

Using double sampling, an estimator of (2.2) is given by $\hat{P} = \hat{C}\hat{\Pi}$. Tenenbein [18] showed that estimator $\hat{P} = \hat{C}\hat{\Pi}$ is the maximum likelihood estimator of (2.2) and he also provided an expression for its asymptotic variance using the inverse of the information matrix. As noted by Marshall [12] and Kuha and Skinner [10] the maximum likelihood estimator $\hat{P} = \hat{C}\hat{\Pi}$ will be asymptotically more efficient than the moment-based estimator $\hat{P} = \hat{Q}^{-1}\hat{\Pi}$.

Maximum likelihood estimation can be also performed by employing misclassification probabilities 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 error free classifier and the fallible classifiers. Consider the cross-classification of the observed with the true classifications. Denote by $n_{ik}^{(*)}$, $n_{ik}^{(v)}$ the counts referring to this cross-classification in the main and in the validation samples respectively. Denote also by $n_{\cdot k}^{(*)}$, $n_{\cdot k}^{(v)}$, $n_{k\cdot}$, $n_{k\cdot}^{(v)}$ the total number of sample units classified in state k by the the error free and the fallible classifiers in the main and in the validation samples respectively. Note that a superscript $(*)$ is used to denote missing data. While for the main sample we have only marginal information about the observed classifications, for the validation sample full information exists. The idea is to estimate the model parameters by combining information from both samples. This parameterisation will eventually lead to

an optimisation problem that involves missing data. This is because the validation procedure is not applied to the units of the main sample. Assuming independence between the main and the validation samples and denoting by D^{Complete} the complete data and by Θ the vector of parameters, the full data likelihood is given by

$$L(\Theta; D^{\text{Complete}}) = \prod_{i=1}^{r} \prod_{k=1}^{r} (P_k q_{ik})^{n_{ik}^{(v)}} \prod_{i=1}^{r} \prod_{k=1}^{r} (P_k q_{ik})^{n_{ik}^{(*)}} \tag{2.3}$$

subject to the following constraints $\sum_{k=1}^{r} P_k = 1$ and $\sum_{i=1}^{r} q_{ik} = 1$ for fixed k. The likelihood function (2.3) contains unobserved data. One way of using this likelihood to maximise the likelihood of the observed data is via the EM algorithm (Dempster et al. [5].) The EM algorithm is based on two steps, namely the expectation step (E-step) and the maximisation step (M-step.) For the currently described model these steps are described below. Denote by $D^{(v)}$ the observed data from the validation sample, by $D^{(m)}$ the observed data from the main sample and by (h) the current EM iteration.

RESULT 2.1. *For the E-step, the conditional expectations of the missing data in the main sample are estimated using the following expression*

$$\widehat{E}\left(n_{ik}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) = n_{i.}\left(\frac{P_k^{(h)} q_{ik}^{(h)}}{\sum_{k=1}^{r} P_k^{(h)} q_{ik}^{(h)}}\right). \tag{2.4}$$

RESULT 2.2. *For the M-step, the maximum likelihood estimators are given below*

$$\widehat{q}_{ik} = \frac{\widehat{E}\left(n_{ik}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{ik}^{(v)}}{\widehat{E}\left(n_{.k}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{.k}^{(v)}}, \qquad \widehat{P}_k = \frac{\widehat{E}\left(n_{.k}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{.k}^{(v)}}{\sum_{k=1}^{r} \widehat{E}\left(n_{.k}^{(*)} \mid D^{(m)}, \Theta^{(h)}\right) + n_{.k}^{(v)}}. \tag{2.5}$$

Results 2.1 and 2.2 are obtained by implementing the EM algorithm with multinomial data (see also Tanner [16].)

Variance estimation for the maximum likelihood estimates requires 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. One way to perform variance estimation in an EM framework is by applying the Missing Information Principle (Louis [11].) Denote by $\widehat{\Theta}$ the vector of maximum likelihood estimates. The Missing Information Principle is defined as

Observed Information = Complete Information − Missing Information.   (2.6)

Following (Louis [11]), the complete information and the missing information are evaluated at $\widehat{\Theta}$ using respectively the expectation of the complete information matrix and the variance of the score functions.

A prerequisite for formulating a measurement error model is the specification of the measurement error process. We have already described two ways of doing this, that is, via calibration or misclassification probabilities. Although it is more natural to parametrise

the measurement error process in terms of misclassification probabilities, in a cross-sectional framework the parameterisation of the misclassification model using calibration probabilities or misclassification probabilities leads to identical results. In general, for an unconstrained (saturated) model, like the one described in this paper, there will be a one to one correspondence between (P, Q) and (Π, C). This is not the case, however, with a constrained model. For example, assume that the standard measurement device is a panel survey but only cross-sectional validation data are available (Tzavidis [19].) The cross-sectional nature of the validation data dictates the use of a conditional independence assumption for estimating the parameters of the longitudinal misclassification mechanism. More specifically, the author assumes that misclassification at time t depends only on the current true state and not on previous or future true states. This assumption implies that the misclassification probabilities are the same at each wave of the panel survey. As shown by Meyer [13] this assumption should be used only with misclassification probabilities and not with calibration probabilities. Therefore, parameterising the model using misclassification probabilities is a more general method.

3. 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 [20] as a basis for fitting generalised linear regression models. As described in Heyde [9], 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

$$Y = \mu(\Theta) + \varepsilon, \tag{3.1}$$

where Y is a $n \times 1$ data vector and $E(\varepsilon) = 0$. The quasi-score estimating function $G(\Theta)$ is then defined (Heyde, [9, Theorem 2.3]) as

$$G(\Theta) = \left\{ \frac{\partial \mu(\Theta)}{\partial \Theta} \right\}^{T} \left\{ \mathrm{var}(\varepsilon) \right\}^{-1} \left\{ Y - \mu(\Theta) \right\}. \tag{3.2}$$

The quasi-score estimating function defined by (3.2) is also referred to as Wedderburn's quasi-score estimating function.

3.1. 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_k the number of units in the main sample classified in category k by the standard measurement device and by n the sample size of the main survey. Without loss of generality, we describe the model in 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 the following

$$\widehat{P}_1^{(v)} = P_1 + \varepsilon_1, \qquad \widehat{q}_{11}^{(v)} = q_{11} + \varepsilon_2$$
$$\widehat{q}_{12}^{(v)} = q_{12} + \varepsilon_3, \qquad n_1 = n[P_1 q_{11} + (1 - P_1) q_{12}] + \varepsilon_4. \tag{3.3}$$

The left-hand side of the equations given in (3.3) describes the estimates obtained from the main and the validation samples whereas the right hand side describes the unknown parameters of interest plus an error term. Equations described by (3.3) incorporate the extra constraints also employed under maximum likelihood estimation, that is, $P_2 = 1 - P_1$, $q_{21} = 1 - q_{11}$ and $q_{22} = 1 - q_{12}$. As with maximum likelihood estimation, we assume that the main and the validation samples share common parameters since both are representative of the same population. Assuming the general form of the model defined by (3.1) 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 [9], Wedderburn's quasi-score estimating function for the two-state model is defined as follows

$$G(\Theta) = \begin{pmatrix} 1 & 0 & 0 & n(q_{11} - q_{12}) \\ 0 & 1 & 0 & nP_1 \\ 0 & 0 & 1 & n(1 - P_1) \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}. \tag{3.4}$$

Setting (3.4) equal to zero leads to three quasi-score normal equations. These equations need to be solved using numerical techniques. In addition, 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 are estimated respectively by

$$\widehat{\sigma}_1^2 = \frac{\widehat{P}_1^{(v)} \left(1 - \widehat{P}_1^{(v)}\right)}{n^{(v)}}, \qquad \widehat{\sigma}_4^2 = n\widehat{\text{pr}}(Y_\xi^* = 1)\{1 - \widehat{\text{pr}}(Y_\xi^* = 1)\}. \tag{3.5}$$

Regarding the covariance matrix of the estimated misclassification probabilities, let us 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 then defined as $\widehat{q}_{ik} = n_{ik}^{(v)} / \sum_{i=1}^{r} n_{ik}^{(v)}$ and the estimated matrix of misclassification probabilities by $\widehat{Q}$. While $n^{(v)} = \sum_{i=1}^{r} \sum_{k=1}^{r} n_{ik}^{(v)}$ can be treated as fixed,

$\sum_{i=1}^{r} n_{ik}^{(v)}$ is random. Consequently, in the computation of this covariance matrix we need to account for the non-linearity introduced by the fact that both the numerator and the de-numerator of $\hat{q}_{ik}$ are random. To estimate the covariance matrix of interest we apply the δ-method. Let $\hat{\Theta}^* = (n_{11}^{(v)}, n_{21}^{(v)}, n_{12}^{(v)}, n_{22}^{(v)})$ and $\text{vec}\{Q(\hat{\Theta}^*)\} = \{f_1(\hat{\Theta}^*), \ldots, f_{r^2}(\hat{\Theta}^*)\}^T$ be an $r^2 \times 1$ vector of nonlinear functions of $\hat{\Theta}^*$. Applying the δ-method to $\text{vec}\{Q(\hat{\Theta}^*)\}$ we obtain the following approximation

$$\text{vec}\{Q(\hat{\Theta}^*)\} - \text{vec}\{Q(\Theta^*)\} \approx \nabla_{\Theta^*}(\hat{\Theta}^* - \Theta^*), \tag{3.6}$$

where $\nabla_{\Theta^*} = \partial \text{vec}\{Q(\Theta^*)\}/\partial \Theta^* \mid_{\Theta^* = \hat{\Theta}^*}$. Taking the variance operator on both sides of (3.6) leads to

$$\text{var}\{\text{vec}(Q)\} \approx \nabla_{\Theta^*} \text{var}(\hat{\Theta}^*)(\nabla_{\Theta^*})^T. \tag{3.7}$$

In (3.7), under simple random sampling, $\text{var}(\hat{\Theta}^*)$ is estimated by

$$\hat{\text{var}}\left(n_{ik}^{(v)}\right) = n^{(v)}\hat{\text{pr}}(Y_\xi^* = i, Y_\xi = k)\{1 - \hat{\text{pr}}(Y_\xi^* = i, Y_\xi = k)\},$$

$$\hat{\text{cov}}\left(n_{ik}^{(v)}, n_{i^*k^*}^{(v)}\right) = -n^{(v)}\hat{\text{pr}}(Y_\xi^* = i, Y_\xi = k)\hat{\text{pr}}(Y_\xi^* = i^*, Y_\xi = k^*), \tag{3.8}$$

for $(i,k) \neq (i^*, k^*)$. Thus, we are able to obtain estimates for σ_2^2, σ_3^2, σ_{23} and σ_{32}, where $\sigma_{23} = \sigma_{32}$.

For estimating the covariance terms σ_{14}, σ_{24} and σ_{34} we need to consider the way we select the validation sample. Independence is assumed when the validation sample is selected independently from the main sample. Independence is also assumed when the validation sample is selected by sub-sampling units from the main sample. This is achieved by dividing the sample into units that belong only in the main sample and units that belong both in the main and in the validation samples. Under the assumption of independence it follows that $\sigma_{14} = \sigma_{41} = \sigma_{24} = \sigma_{42} = \sigma_{34} = \sigma_{43} = 0$.

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

$$\text{cov}\left(\hat{q}_{ik}^{(v)}, \hat{P}_k^{(v)}\right) = \text{cov}\left(\frac{n_{ik}^{(v)}}{\sum_{i=1}^{r} n_{ik}^{(v)}}, \frac{\sum_{i=1}^{r} n_{ik}^{(v)}}{n^{(v)}}\right). \tag{3.9}$$

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

LEMMA 3.1 (Mood et al. [14]). *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*

$$E\{g(X,Y)\} \approx g(\mu_X, \mu_Y) + \frac{1}{2}\frac{\partial^2}{\partial y^2}g(X,Y) \mid_{\mu_X, \mu_Y} \text{var}(Y) + \frac{1}{2}\frac{\partial^2}{\partial x^2}g(X,Y) \mid_{\mu_X, \mu_Y} \text{var}(X)$$

$$+ \frac{\partial^2}{\partial x \partial y}g(X,Y) \mid_{\mu_X, \mu_Y} \text{cov}(X,Y). \tag{3.10}$$

RESULT 3.2. *Assume that X, Y, A are three random variables and n is fixed. A first order approximation for* cov(X/Y, A/n) *is given by*

$$\operatorname{cov}\left(\frac{X}{Y}, \frac{A}{n}\right) \approx \frac{1}{nE(Y)}\left\{\operatorname{cov}(A, X) - \frac{E(X)}{E(Y)}\operatorname{cov}(A, Y)\right\}. \tag{3.11}$$

Proof of this result is given in the appendix. 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 3.2 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 estimates requires solving the system of equations defined by (3.4). This is achieved 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 $A_{ij} = \partial G_i(\Theta)/\partial \vartheta_j$, $i, j = 1, \dots, \omega$. The system of quasi-score normal equations is then solved numerically as follows. Assume a vector of initial solutions $\hat{\Theta}^{(0)}$. The vector of initial solutions is updated using

$$\hat{\Theta}^{(1)} = \hat{\Theta}^{(0)} - A^{-1}\{\hat{\Theta}^{(0)}\}G\{\hat{\Theta}^{(0)}\}, \tag{3.12}$$

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

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

RESULT 3.3. *The variance of the quasi-likelihood estimates is estimated using the expression below*

$$\hat{\operatorname{var}}(\hat{\Theta}) \approx \left[\left\{\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right\}^{T}\{\hat{\operatorname{var}}(\varepsilon)\}^{-1}\left\{\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right\}\right]^{-1}. \tag{3.13}$$

Proof of this result is given in the appendix.

The system of (3.3) can be modified for tackling more complex situations. Let us consider the case described at the end of Section 2 where the standard measurement device is a panel survey but only cross-sectional validation data are available. In this case we need to incorporate a conditional independence assumption to enable estimation of the longitudinal misclassification mechanism. A quasi-likelihood solution can be offered using a system of equations similar to (3.3). This system will consist of equations for P_i and for q_{ij}. However, the difference now is that the conditional independence assumption needs to be incorporated appropriately into the system of equations.

4. A numerical example

The alternative approaches are illustrated using data form the UK labour force survey (LFS). The UK LFS is a panel survey of households living at private addresses. One of its main purposes is to provide cross-sectional estimates of the proportion of individuals in each of the main labour force states, that is, employed, unemployed and inactive. However, as with every sample survey, the UK LFS is subject to response error. Validation data (Table 4.1) are obtained from a validation survey, which is not explicitly defined due to confidentiality restrictions. In addition, we use unweighted UK LFS data (Table 4.2) between summer-autumn 1997.

Table 4.1. Data from the validation sample.

| | | Correct classifications | | |
		Employed	Unempl./Inact.	Margins
Fallible	Employed	2178	56	2234
classifications	Unempl./Inact.	22	744	766
	Margins	2200	800	3000

Table 4.2. Data from the main sample.

| | | Correct classifications | | |
		Employed	Unempl./Inact.	Margins
Fallible	Employed	$n_{11}^{(*)}$	$n_{12}^{(*)}$	44460
Classifications	Unemp./Inact.	$n_{21}^{(*)}$	$n_{22}^{(*)}$	15540
	Margins	$n_{\cdot 1}^{(*)}$	$n_{\cdot 2}^{(*)}$	60000

Table 4.3. Proportion of units classified as employed, estimated standard errors in parentheses.

Point estimator	Proportion of employed
Naive (unadjusted estimator)	0.741 ($1.79*10-3$)
Moment-based	0.730 ($3.61*10-3$)
Maximum likelihood	
(Calibration probabilities)	0.730 ($3.20*10-3$)
Maximum likelihood	
(Misclassification probabilities)	0.730 ($3.00*10-3$)
Quasi-likelihood	0.729 ($3.21*10-3$)

The target is to adjust cross-sectional (summer 1997) UK LFS estimates for response error. For simplicity we consider a two-state model where individuals can be classified in two states, that is, employed and unemployed or inactive. The estimators we consider are the unadjusted estimator (Naive), the moment-based estimator (Section 2), the maximum likelihood estimator with callibration probabilities (Section 2), the maximum likelihood estimator with misclassification probabilities (Section 2) and the quasi-likelihood estimator (Section 3.) The convergence criterion for the EM and the Newton-Raphson algorithms is $\delta = 10^{-6}$. Variance estmation for the naive (unadjusted) estimator is performed assuming a multinomial distribution. The variance of the moment-based estimator is estimated using results from Selén [15]. The variance of the maximum likelihood estimator that employs calibration probabilities is estimated using the results of Tenenbein [18] while variance estimation for the maximum likelihood estimator, using the EM algorithm, and for the quasi-likelihood estimator is performed using the Missing Information Principle and Result 3.3 respectively. The results are summarised in Table 4.3.

Table 4.4. Simulation results (averages over 1000 simulations).

Estimators	$\hat{P}$	Bias($\hat{P}$)	var($\hat{P}$)	MSE($\hat{P}$)
Moment-based	0.6061	$1*10-4$	$1.40*10-5$	$1.40*10-5$
Maximum likelihood				
(Misclassification)	0.6059	$-1*10-4$	$1.28*10-5$	$1.28*10-5$
Quasi-likelihood	0.6059	$-1*10-4$	$1.28*10-5$	$1.28*10-5$

The estimators that adjust for response error produce reasonable estimates, which are close to the proportion of truly employed people estimated from the validation sample. Although the alternative estimators produce identical point estimates, differences exist in the estimated standard errors.

5. A simulation study

In this section the alternative methods are empirically compared using a Monte-Carlo simulation study. The simulation algorithm consists of the following steps. Step 1: 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 ($P_1 = 0.606$, $P_2 = 0.394$). Using this distribution, we draw a with replacement sample of size $n = 60000$. Step 2: At the second step we assume the existence of misclassification described by the misclassification probabilities q_{ik} ($q_{11} = 0.98$, $q_{22} = 0.96$). Using the misclassification probabilities, we generate the observed status, given the true status (Step 1), for each sample unit ξ. Step 3: At the third step we generate validation data ($n^{(v)} = 3000$). After all three steps have been completed, the generated data are employed for computing the alternative estimators. The properties of the alternative point estimators $\hat{P}$ are evaluated using (a) the bias of a point estimator, (b) the variance of a point estimator and (c) the mean squared error (MSE) of a point estimator. The results from the Monte-Carlo simulation are summarised in Table 4.4. Note that since for the simple case considered by this paper there is only one set of maximum likelihood estimates, obtained using either calibration or misclassification probabilities, we only report the maximum likelihood estimates derived with the use of misclassification probabilities.

The simulation results verify that using the maximum likelihood or the quasi-likelihood estimators, instead of the moment-based estimator, leads to gains in efficiency.

6. Discussion

Two alternative parameterisations for maximum likelihood estimation using either calibration or misclassification probabilities are presented. In a cross-sectional framework both parameterisations lead to identical results. However, using misclassification probabilities instead of calibration probabilities is more reasonable with complex data such as longitudinal misclassified data. We therefore suggest that the formulation of the misclassification model as a missing data problem is a more general method. As an alternative approach we further presented a quasi-likelihood formulation 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. Results from a simulation study indicate that the quasi-likelihood estimator is almost as efficient as the maximum likelihood estimator. Regarding the moment-based estimator we suggest that this should be avoided since it can produce estimates that lie outside the parameter space and is less efficient compared to either the maximum likelihood estimators or the quasi-likelihood estimator.

Appendix

A. Proofs of the results in Section 3

Proof of Result 3.2. Apply Lemma 3.1 to $E(AX/Y)$ and $E(X/Y)$, we obtain

$$E\left(\frac{AX}{Y}\right) \approx \frac{\mu_X \mu_A}{\mu_Y} + \frac{1}{2}\frac{2\mu_X \mu_A}{\mu_Y^3}\mathrm{var}(Y) - \frac{\mu_A}{\mu_Y^2}\mathrm{cov}(X,Y)$$

$$+ \frac{1}{\mu_Y}\mathrm{cov}(X,A) - \frac{\mu_X}{\mu_Y^2}\mathrm{cov}(A,Y), \tag{A.1}$$

$$E\left(\frac{X}{Y}\right) \approx \frac{\mu_X}{\mu_Y} + \frac{1}{2}\frac{2\mu_X}{\mu_Y^3}\mathrm{var}(Y) - \frac{1}{\mu_Y^2}\mathrm{cov}(X,Y).$$

Therefore,

$$\mathrm{cov}\left(\frac{X}{Y},\frac{A}{n}\right) = \frac{1}{n}\left\{E\left(\frac{AX}{Y}\right) - E\left(\frac{X}{Y}\right)E(A)\right\}$$

$$\approx \frac{1}{nE(Y)}\left\{\mathrm{cov}(X,A) - \frac{E(X)}{E(Y)}\mathrm{cov}(A,Y)\right\} \tag{A.2}$$

as required. □

Proof of Result 3.3. Let $\hat{\Theta}$ denote the vector of quasi-likelihood estimates and ε the vector of errors. The quasi-score estimating function is defined by $G(\Theta) = \{\partial\mu(\Theta)/\partial\Theta\}^T \times\{\mathrm{var}(\varepsilon)\}^{-1}\varepsilon$. By Taylor expansion, $G(\hat{\Theta})$ can be approximated by

$$G(\Theta) + \left\{\frac{\partial\mu(\Theta)}{\partial\Theta}\mid_{\Theta=\hat{\Theta}}\right\}^T \{\mathrm{var}(\varepsilon)\}^{-1}\left[\left\{\frac{\partial\mu(\Theta)}{\partial\Theta}\mid_{\Theta=\hat{\Theta}}\right\}^T\right]^T (\hat{\Theta}-\Theta). \tag{A.3}$$

Thus, $\mathrm{var}\{G(\hat{\Theta})\} = \{\partial\mu(\Theta)/\partial\Theta\mid_{\Theta=\hat{\Theta}}\}^T\{\mathrm{var}(\varepsilon)\}^{-1}[\{\partial\mu(\Theta)/\partial\Theta\mid_{\Theta=\hat{\Theta}}\}^T]^T$ can be approximated by

$$\mathrm{var}\left[\left\{\frac{\partial\mu(\Theta)}{\partial\Theta}\mid_{\Theta=\hat{\Theta}}\right\}^T \{\mathrm{var}(\varepsilon)\}^{-1}\left[\left\{\frac{\partial\mu(\Theta)}{\partial\Theta}\mid_{\Theta=\hat{\Theta}}\right\}^T\right]^T (\hat{\Theta}-\Theta)\right]. \tag{A.4}$$

This leads to

$$\hat{\text{var}}(\hat{\Theta}) \approx \left[\left\{ \frac{\partial \mu(\Theta)}{\partial \Theta} \mid_{\Theta=\hat{\Theta}} \right\}^{T} \left\{ \hat{\text{var}}(\varepsilon) \right\}^{-1} \left\{ \frac{\partial \mu(\Theta)}{\partial \Theta} \mid_{\Theta=\hat{\Theta}} \right\} \right]^{-1}. \qquad (A.5)$$

$\square$

References

[1] B. A. Bailar, *Recent research in reinterview procedures*, Journal of the American Statistical Association **63** (1968), no. 321, 41–63.

[2] K. E. Bauman and G. G. Koch, *Validity of self-reports and descriptive and analytical conclusions: the case of cigarette smoking by adolescents and their mothers*, American Journal of Epidemiology **118** (1983), no. 1, 90–98.

[3] I. Bross, *Misclassification in 2×2 tables*, Biometrics **10** (1954), 478–486.

[4] R. J. Carroll, *Approaches to estimation with errors in predictors*, Advances in GLIM and Statistical Modelling (L. Fahrmeir, B. Francis, R. Gilchrist, and G. Tutz, eds.), Springer, New York, 1992, pp. 40–47.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society. Series B (Methodological) **39** (1977), no. 1, 1–38.

[6] G. Forsman and I. Schreiner, *The design and analysis of reinterview: an overview*, Measurement Error in Surveys (P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiewetz, and S. Sudman, eds.), John Wiley & Sons, New York, 1991, pp. 279–302.

[7] W. A. Fuller, *Measurement Error Models*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, New York, 1987.

[8] S. Greenland, *Variance estimation for epidemiologic effect estimates under misclassification*, Statistics in Medicine **7** (1988), no. 7, 745–757.

[9] C. C. Heyde, *Quasi-Likelihood and Its Application. A General Approach to Optimal Parameter Estimation*, Springer Series in Statistics, Springer, New York, 1997.

[10] J. Kuha and C. J. Skinner, *Categorical data analysis and misclassification*, Survey Measurement and Process Quality (L. E. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, eds.), John Wiley & Sons, New York, 1997, pp. 633–670.

[11] T. A. Louis, *Finding the observed information matrix when using the EM algorithm*, Journal of the Royal Statistical Society. Series B (Methodological) **44** (1982), no. 2, 226–233.

[12] R. J. Marshall, *Validation study methods for estimating exposure proportions and odds ratios with misclassified data*, Journal of Clinical Epidemiology **43** (1990), no. 9, 941–947.

[13] B. D Meyer, *Classification-error models and labor-market dynamics*, Journal of Business and Economic Statistics **6** (1988), no. 3, 385–390.

[14] A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics*, McGraw-Hill, Singapore, 1963.

[15] J. Selén, *Adjusting for errors of classification and measurement in the analysis of partly and purely categorical data*, Journal of the American Statistical Association **81** (1986), 75–81.

[16] M. A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer, New York, 1997.

[17] A. Tenenbein, *A double sampling scheme for estimating from binomial data with misclassifications*, Journal of the American Statistical Association **65** (1970), 1350–1361.

[18] ———, *A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection*, Technometrics **14** (1972), 187–202.

[19] N. Tzavidis, *Correcting for misclassification error in gross flows using double sampling: Moment-based inference vs. likelihood-based inference*, S3RI Methodology Series Working Papers M04/11, 1–33, University of Southampton, Southampton, 2004, http://www.s3ri.soton.ac.uk/publications/methodology.php.

[20] R. W. M. Wedderburn, *Quasi-likelihood functions, generalized linear models and the Gauss-Newton method*, Biometrika **61** (1974), no. 3, 439–447.

Nikos Tzavidis: Centre for Longitudinal Studies (CLS), Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK; Southampton Statistical Sciences Research Institute, University of Southampton, UK
E-mail address: n.tzavidis@ioe.ac.uk

Yan-Xia Lin: School of Mathematics and Applied Statistics, University of Wollongong, Northfields Ave, Wollongong, NSW 2500, Australia
E-mail address: yanxia@uow.edu.au

Special Issue on
Intelligent Computational Methods for Financial Engineering

Call for Papers

As a multidisciplinary field, financial engineering is becoming increasingly important in today's economic and financial world, especially in areas such as portfolio management, asset valuation and prediction, fraud detection, and credit risk management. For example, in a credit risk context, the recently approved Basel II guidelines advise financial institutions to build comprehensible credit risk models in order to optimize their capital allocation policy. Computational methods are being intensively studied and applied to improve the quality of the financial decisions that need to be made. Until now, computational methods and models are central to the analysis of economic and financial decisions.

However, more and more researchers have found that the financial environment is not ruled by mathematical distributions or statistical models. In such situations, some attempts have also been made to develop financial engineering models using intelligent computing approaches. For example, an artificial neural network (ANN) is a nonparametric estimation technique which does not make any distributional assumptions regarding the underlying asset. Instead, ANN approach develops a model using sets of unknown parameters and lets the optimization routine seek the best fitting parameters to obtain the desired results. The main aim of this special issue is not to merely illustrate the superior performance of a new intelligent computational method, but also to demonstrate how it can be used effectively in a financial engineering environment to improve and facilitate financial decision making. In this sense, the submissions should especially address how the results of estimated computational models (e.g., ANN, support vector machines, evolutionary algorithm, and fuzzy models) can be used to develop intelligent, easy-to-use, and/or comprehensible computational systems (e.g., decision support systems, agent-based system, and web-based systems)

This special issue will include (but not be limited to) the following topics:

- **Computational methods**: artificial intelligence, neural networks, evolutionary algorithms, fuzzy inference, hybrid learning, ensemble learning, cooperative learning, multiagent learning

- **Application fields**: asset valuation and prediction, asset allocation and portfolio selection, bankruptcy prediction, fraud detection, credit risk management

- **Implementation aspects**: decision support systems, expert systems, information systems, intelligent agents, web service, monitoring, deployment, implementation

Authors should follow the Journal of Applied Mathematics and Decision Sciences manuscript format described at the journal site http://www.hindawi.com/journals/jamds/. Prospective authors should submit an electronic copy of their complete manuscript through the journal Manuscript Tracking System at http://mts.hindawi.com/, according to the following timetable:

Manuscript Due	December 1, 2008
First Round of Reviews	March 1, 2009
Publication Date	June 1, 2009

Guest Editors

Lean Yu, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong; yulean@amss.ac.cn

Shouyang Wang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; sywang@amss.ac.cn

K. K. Lai, Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong; mskklai@cityu.edu.hk