Statistica Sinica **8**(1998), 1045-1058

LOGISTIC MODELLING OF LONGITUDINAL SURVEY DATA WITH MEASUREMENT ERROR

C. J. Skinner

University of Southampton

Abstract: A logistic model relating the rates of transition between two states to a vector of covariates is considered. Measurement error on the binary state variable can lead to severely biased parameter estimates. Estimation procedures which adjust for measurement error are proposed for different measurement models. Complex sampling designs are allowed for. The procedures are illustrated using data from the U.S. Panel Study of Income Dynamics, where the response is whether an individual is in a job with a union contract. It is found that adjusting for measurement error can be important.

Key words and phrases: Gross flow, longitudinal, measurement error, misclassification, transition.

1. Introduction

In the analysis of longitudinal survey data it is often of interest to estimate a transition rate, the proportion of units in the population in one state on one occasion which flow into another state on the successive occasion. For example, labour market analysts may be interested in the 3 × 3 matrix of transitions between the states: employed, unemployed, and not in the labour force (Abowd and Zellner (1985)). For analytical purposes, it is often of interest to study how rates vary across different subgroups of the population. For example, in the case of labour force states it may be of interest to study the dependence of transition rates on sex, age and region.

Transition rates are proportions and so may be estimated in the usual way from survey data. The off-diagonal cells of transition matrices are often relatively rare, however, and so, as the number of subgroups increases, the sample sizes upon which some estimates are based can become small, implying large sampling errors. In this case some modelling of the relation between the rates and the covariates defining the subgroups is desirable. In Section 2, we describe a logistic model for representing the dependence of transition probabilities on covariates for the case of two states and two occasions. Similar models have been applied to longitudinal data in both biostatistical applications (e.g. Korn and

Whittemore (1979), Muenz and Rubinstein (1985), Bonney (1987)) and econometric applications (e.g. Hsiao (1986), Sect. 7.4; Maddala (1987)). Our approach to model fitting follows Roberts et al. (1987) and allows for a complex sampling scheme.

Measurement error is a major problem when estimating transition rates from survey data. Random errors in measured states can lead to severe upward bias in standard estimators of the proportions moving between different states. A number of alternative estimators have been proposed which use reinterview data to reduce this bias (Meyer (1988)). The aim of this paper is to extend this work to the estimation of the logistic models referred to above. Specifically, in Section 5 we extend the measurement error model of Chua and Fuller (1987) and propose inference procedures which adjust for measurement error, provided estimates of the parameters of the measurement error process are available from reinterview studies.

2. The Model

Consider a finite population of size N, which is fixed over two occasions $t = 1, 2$ and is partitioned into I cells (also fixed over time) by the levels of one or more factors defined at $t = 1$. Let y_t be a binary indicator variable for the two states and let N_{ijk} be the number of units in cell i for which $y_1 = j$ and $y_2 = k(i = 1, \ldots, I;\ j = 0, 1; k = 0, 1)$.

We assume a superpopulation model under which N is fixed and $E(N_{ijk}) = N\phi_{ijk}(i = 1, \ldots, I; j = 0, 1; k = 0, 1)$ and focus attention on the transition rates in different cells i between states at $t = 1$ and $t = 2$:

$$\pi_{ij} = \phi_{ij1}/\phi_{ij.},$$

where

$$\phi_{ij.} = \phi_{ij0} + \phi_{ij1}, i = 1, \ldots, I; j = 0, 1.$$

The superpopulation model would arise if units fall into cells i and take values $y_1 = j$ and $y_2 = k$ with probabilities $\phi_{ijk} = pr(y_1 = j, y_2 = k,\ \text{cell } i)$. In this case π_{ij} is the conditional probability $\text{pr}(y_2 = 1|y_1 = j,\ \text{cell } i)$ of moving from state j at $t = 1$ to state 1 at $t = 2$ for units in cell i. Under independence between units N_{ijk} would be multinomially distributed with parameters N and ϕ_{ijk}. Dependence between units because of clustering may, however, lead to departures from a multinomial distribution. We take as our objective to study the dependence of the transition rates π_{ij} on i and j and consider the following logistic model:

$$\pi_{ij} = F(x_{ij}\beta), i = 1, \ldots, I; j = 0, 1, \tag{1}$$

where $F(t) = e^t/(1+e^t)$, the x_{ij} are $1 \times s$ vectors of known constants and β is a $s \times 1$ vector of unknown parameters. Note that (1) may be expressed alternatively as

$$\log[\pi_{ij}/(1-\pi_{ij})] = x_{ij}\beta. \tag{2}$$

Some special cases of this model are given below to illustrate its interpretation, but first we record some notation. For a series of $1 \times \rho$ vectors, $a_{ij} (i = 1, \dots, I; j = 0, 1)$, let $[a_{ij}]$ denote the $2I \times \rho$ matrix with rows $a_{10}, a_{11}, a_{20}, a_{21}, \dots, a_{I0}, a_{I1}$. Let $X = [x_{ij}], l = [l_{ij}], \pi = [\pi_{ij}], \phi = [\phi_{ij}]$, $f(\beta) = [f_{ij}(\beta)]$, where $l_{ij} = \log[\pi_{ij}/(1-\pi_{ij})]$, $f_{ij}(\beta) = F(x_{ij}\beta)$. Then (1) may be re-expressed as

$$\pi = f(\beta), \tag{3}$$

and (2) may be re-expressed as

$$l = X\beta. \tag{4}$$

Examples of Models

(i) Constant transition rates

Let $s = 2, x_{ij} = (1, j)$ and $\beta = (\beta_1, \beta_2)'$. Then the transition rates π_{i0} and π_{i1} take the constant values $F(\beta_1)$ and $F(\beta_1 + \beta_2)$ respectively for all cells i.

(ii) Additive model

Let $s = r + 2, x_{ij} = (1, z_i, j)$ and $\beta = (\beta_1, \beta_2', \beta_3)'$, where z_i is a $1 \times r$ vector of known constants derived from the factor levels defining the I cells and β_2 is an $r \times 1$ vector of unknown parameters. For example, the cells may arise from crossing $I/2$ age groups by 2 sexes and z_i may be (a_i, a_i^2, s_i) where a_i is the midpoint of the age group and s_i is a dummy variable representing sex for cell i. The ratio of the odds that $y_2 = 1$ given $y_1 = 1$ versus $y_1 = 0$ is $\exp(\beta_3)$ which is constant across cells.

(iii) Separate models for different previous states

Let $s = 2r + 2, x_{ij} = (1, z_i, j, jz_i)$ and $\beta = (\beta_1, \beta_2', \beta_2, \beta_4')'$, where z_i is as in (ii). This model allows for interaction between y_1 and cell i. Transition rates are now $\pi_{i0} = F(\beta_1 + z_i\beta_2)$ and $\pi_{i1} = F[(\beta_1 + \beta_3) + z_i(\beta_2 + \beta_2)]$.

(iv) Saturated model

Let $s = 2I$ and suppose X is nonsingular. Then there is a 1-1 mapping between π and β since (4) may be inverted to give $\beta = X^{-1}l$.

In general, we take β to be the vector of parameters of interest. As set out above, β is only well defined if model (1) holds. In practice, however, it may still be of interest to fit a model, such as the main effects model in (ii), even if that model only holds approximately. For a more general definition when model (1) does not necessarily hold, we let β be the solution to

$$\sum_i \sum_j x_{ij}\phi_{ij.}[f_{ij}(\beta) - \pi_{ij}] = 0. \tag{5}$$

In order that a unique solution to (5) exists we assume:
A1: X has full rank s
A2: $0 < \phi_{ij.} < 1, 0 < \pi_{ij} < 1, i = 1, \ldots, I; j = 0, 1$.
Clearly the true β solves (5) when (1) holds. In general, β may be interpreted as follows (cf Scott and Wild (1989), p.194). Let

$$\varepsilon_{ij} = f_{ij}(\beta) - \pi_{ij} \tag{6}$$

be the model approximation error for cell i and $y_i = j$. This will be zero if (1) holds. Suppose x includes an intercept term so $x_{ij} = (1, \tilde{x}'_{ij})'$ and let $(\varepsilon, \tilde{x})$ be random variables taking values $(\varepsilon_{ij}, \tilde{x}_{ij})$ with probabilities $\phi_{ij.}(i = 1, \ldots, I; j = 0, 1)$. Then (5) is equivalent to the constraints that $E(\varepsilon) = 0$, $\text{Cov}(\varepsilon, \tilde{x}) = 0$, i.e. β is defined so that the implied approximation errors have zero means and are uncorrelated with the covariates across cells.

As an alternative interpretation, note that if the N_{ijk} are multinomially distributed with parameters N and ϕ_{ijk} then the likelihood equations for estimating β in model (1), were all the N_{ijk} to be observed and the $\phi_{ij.}$ to be unconstrained, are

$$\sum_i \sum_j x_{ij}[N_{ij.}f_{ij}(\beta) - N_{ij1}] = 0, \tag{7}$$

where $N_{ij.} = N_{ij0} + N_{ij1}$. Dividing by N and taking expectations gives (5), so that in this case β may be interpreted as the limit in probability as $N \to \infty$ of the 'census estimator' of β solving (7).

3. Estimation

Let $\hat{N}_{ijk}$ be an estimator of N_{ijk} based upon the survey data. The estimator may involve weighting adjustments to allow for unequal sampling fractions or for differential nonresponse or to achieve arithmetic consistency under rotating designs as in the maximum likelihood or raking ratio estimators of Chambers et al. (1988) and Holt and Skinner (1989). Consider an asymptotic framework in which N and a suitably defined sample size n increase, but where I and the ϕ_{ijk} are fixed. Let

$$\hat{N}_{ij.} = \sum_k \hat{N}_{ijk}, \hat{\phi}_{ij.} = \hat{N}_{ij.}/\hat{N}, \hat{\pi}_{ij} = \hat{N}_{ij1}/\hat{N}_{ij.}, \hat{N} = \sum_i \sum_j \hat{N}_{ij},$$

$$\hat{\phi} = [\hat{\phi}_{ij.}], \hat{\pi} = [\hat{\pi}_{ij}].$$

We assume:
A3: $(\hat{\pi}', \hat{\phi}')$ is consistent for (π', ϕ') and the asymptotic distribution as $n \to \infty$ of $\sqrt{n}[(\hat{\pi}', \hat{\phi}')' - (\pi', \phi')']$ is normal with mean vector zero and covariance matrix

$$V[(\hat{\pi}', \hat{\phi}')'] = \begin{bmatrix} V(\hat{\pi}), C(\hat{\pi}, \hat{\phi}) \\ C(\hat{\phi}, \hat{\pi}), \quad V(\hat{\phi}) \end{bmatrix}. \tag{8}$$

This covariance matrix may reflect features of a complex sampling design, such as stratification and clustering. The distribution in A3 may in general reflect both a randomised sampling scheme and the superpopulation model.

Given $\hat{\phi}$ and $\hat{\pi}, \beta$ may be estimated by the solution $\hat{\beta}$ of equations (5) with $\phi_{ij.}$ and π_{ij} replaced by $\hat{\phi}_{ij.}$ and $\hat{\pi}_{ij}$ respectively. Under assumptions A1, A2 and A3, $\hat{\beta}$ is consistent for β whether or not the logistic model holds. In order to derive an expression for the asymptotic covariance matrix of $\hat{\beta}$ we now define some further notation. For a series of scalars $a_{ij} (i = 1, \ldots, I; j = 0, 1)$ let $\mathrm{diag}[a_{ij}]$ denote the $2I \times 2I$ diagonal matrix with diagonal elements $a_{10}, a_{11}, \ldots, a_{I0}, a_{I1}$. Let $D(\hat{\phi}) = \mathrm{diag}[\hat{\phi}_{ij.}]$, $D(\phi) = \mathrm{diag}[\phi_{ij.}], D(\varepsilon) = \mathrm{diag}[\varepsilon_{ij}]$, $\Delta = \mathrm{diag}[\phi_{ij.} f_{ij}(\beta)\{1 - f_{ij}(\beta)\}]$. Then $\hat{\beta}$ solves the estimating equations

$$X' D(\hat{\phi}) f(\hat{\beta}) = X' D(\hat{\phi}) \hat{\pi}, \tag{9}$$

(c.f. Roberts et al. (1987), equation 2.3). The asymptotic covariance matrix of $\hat{\beta}$ is given by:

$$V(\hat{\beta}) = n^{-1} (X' \Delta X)^{-1} X' \sum X (X' \Delta X), \tag{10}$$

where

$$\sum = [D(\phi), D(\varepsilon)] V[(\hat{\pi}', \hat{\phi}')'] [D(\phi), D(\varepsilon)]'. \tag{11}$$

The term $D(\varepsilon)$ allows for lack of fit of the logistic model. If the model holds then $D(\varepsilon) = 0, \sum$ reduces to $D(\phi) V(\hat{\pi}) D(\phi)$ and $V(\hat{\beta})$ reduces to an expression analogous to equation (2.4) of Roberts et al. (1987). Given an estimator of $V[(\hat{\pi}', \hat{\phi}')']$, $V(\hat{\beta})$ may be estimated by substituting $\hat{\phi}, \hat{\pi}$ and $\hat{\beta}$ for ϕ, π and β respectively in Δ and $\sum$. The estimation of $V[(\hat{\pi}, \hat{\phi}')']$ will typically employ survey sampling techniques, as described by Wolter (1985).

Example: Simple random sampling

Suppose n units are selected from the population by simple random sampling, and $n_{ij.}$ and n_{ij1}, the sample quantities analogous to $N_{ij.}$ and N_{ij1}, are observed $(i = 1, \ldots, I; j = 0, 1)$. Then the likelihood equations are, by analogy to (7):

$$\sum_i \sum_j x_{ij} [n_{ij.} f_{ij}(\beta) - n_{ij1}] = 0,$$

which corresponds to taking $\hat{\phi}_{ij.} = n_{ij.}/n$, $\hat{\pi}_{ij} = n_{ij1}/n_{ij.}$ in (9). Then $\hat{\pi}$ and $\hat{\phi}$ are asymptotically uncorrelated $(C(\hat{\pi},\hat{\phi}) = 0)$, $V(\hat{\pi}) = \text{diag}[\pi_{ij}(1-\pi_{ij})\phi_{ij.}^{-1}], V(\hat{\phi}) = D(\phi) - \phi\phi'$, and from (11), $\sum = \text{diag}[\pi_{ij}(1-\pi_{ij})\phi_{ij.}] + D(\varepsilon)[D(\phi)\ -\phi\phi']D(\varepsilon)$. If the logistic model holds, then $\sum = \Delta$ and $V(\hat{\beta}) = n^{-1}(X'\Delta X)^{-1}$, the standard formula.

4. The Effect of Measurement Error

Let us now consider the effect of measuring y_1 and y_2 with error. Let y_1^* and y_2^* denote the observed variables measuring y_1 and y_2 respectively. Measurement error arises if misclassification of the states at $t = 1$ or $t = 2$ occurs, that is if $(y_1^*, y_2^*) \neq (y_1, y_2)$. Let $\hat{N}_{ijk}^*$ be the estimator of N_{ijk} which takes the same form as $\hat{N}_{ijk}$ in section 3 but with (y_1, y_2) replaced by (y_1^*, y_2^*). Let $\hat{\phi}_{ijk}^* = \hat{N}_{ijk}^*/\hat{N}^*$, where $\hat{N}^* = \sum\sum\sum \hat{N}_{ijk}^*$, and similarly define $\hat{N}_{ij.}^*, \hat{\phi}_{ij.}^*, \hat{\pi}_{ij}^*, \hat{\phi}^*, \hat{\pi}^*$, and $D(\hat{\phi}^*)$ analogously to their non-asterisked versions. We assume that an asterisked version A3* of A3 holds with $(\hat{\pi}^{*'}, \hat{\phi}^{*'})$ being consistent for $(\pi^{*'}, \phi^{*'})$, where $\pi^* = [\pi_{ij}^*]$ and $\phi^* = [\phi_{ij.}^*]$. The distribution involved in the statement of A3* now involves not only the randomised sampling scheme and the model generating y_1 and y_2 but also a measurement error model (misclassification mechanism) generating y_1^* and y_2^* from y_1 and y_2. We may interpret the elements of π^* and ϕ^* as $\pi_{ij}^* = \text{pr}(y_2^* = 1|y_1^* = j,\ \text{cell } i)$ and $\phi_{ij.}^* = \text{pr}(y_1^* = j,\ \text{cell } i)$. Under general misclassification mechanisms there is no reason why $(\pi_{ij}^*, \phi_{ij.}^*)$ should be the same as $(\pi_{ij}, \phi_{ij.})$.

If $\hat{\beta}*$ is the solution of estimating equations (9) with $\hat{\phi}$ and $\hat{\pi}$ replaced by $\hat{\phi}^*$ and $\hat{\pi}^*$ repectively, then $\hat{\beta}*$ will be consistent for the solution $\beta*$ of the equations:

$$X'D(\phi^*)f(\beta*) = X'D(\phi^*)\pi^*, \tag{12}$$

provided π^* and ϕ^* obey the asterisked version of A2. In general, $\beta*$ will not equal β unless $\phi^* = \phi$ and $\pi^* = \pi$. Hence, measurement error induces bias even in large samples.

5. Adjustment for Measurement Error

The nature of the measurement error adjustment will depend on the specification of the measurement error model which will in turn depend, in practice, on the nature and extent of the validation data available. In order to specify our measurement error model we first define ϕ_{ijk}^* by extending the assumption that $\hat{\phi}_{ij.}^*$ is consistent for $\phi_{ij.}^*$ to assume that $\hat{\phi}_{ijk}^*$ is consistent for ϕ_{ijk}^*, where $\phi_{ij0}^* + \phi_{ij1}^* = \phi_{ij.}^*$. We suppose that only cross-sectional validation data is available in which case it is natural, following e.g. Abowd and Zellner (1985), to make the following assumption:

A4 (*independent measurement errors within cells*):

$$\phi^*_{ijk} = \sum_{l=0}^{1}\sum_{m=0}^{1} \theta^1_{ij1}\theta^2_{ikm}\phi_{ilm}, \tag{13}$$

where θ^t_{ijk} =pr$(y^*_t = j|y_t = k,\ \text{cell } i)$ is the probability of misclassifying state k as state j in cell i at time t. Without longitudinal validation data it is difficult to know how to specify a model for dependent errors, although some sensitivity analysis to departures from independence is possible (Singh and Rao (1995)).

Letting $\theta^t(i)$, $\phi(i)$ and $\phi^*(i)$ denote the 2×2 matrices with jkth elements $\theta^t_{ijk}, \phi_{ijk}$ and ϕ^*_{ijk} respectively, (13) may be reexpressed as $\phi^*(i) = \theta^1(i)\phi(i)\theta^2(i)'$. If estimators $\hat{\theta}^t(i)$ of the $\theta^t(i)$ are available from validation studies then an adjusted estimator of $\phi(i)$ is

$$\check{\phi}(i) = [\hat{\theta}^t(i)]^{-1}\hat{\phi}^*(i)[\hat{\theta}^2(i)']^{-1}, \tag{14}$$

where the ikth element of $\hat{\phi}^*(i)$ is $\hat{\phi}^*_{ijk}$. An adjusted estimator of β is then obtained by solving (5) with $\phi_{ij.}$ and π_{ij} replaced by $\check{\phi}_{ij.}$ and $\check{\pi}_{ij} = \check{\phi}_{ij1}/\check{\phi}_{ij.}$ respectively, where $\check{\phi}_{ijk}$, is the jkth element of $\check{\phi}(i)$ and $\check{\phi}_{ij.} = \check{\phi}_{ij1} + \check{\phi}_{ij2}$. Assuming consistency of $\hat{\theta}^t_{ijk}$ for θ^t_{ijk}, the adjusted estimator will be consistent, and its asymptotic covariance matrix is as defined by (10) and (11), where $V[(\hat{\pi}', \hat{\phi}')']$ is replaced by the asymptotic covariance matrix of the vector of $\check{\pi}_{ij}$ and $\check{\phi}_{ij.}$. This matrix can be estimated by the δ-method provided an estimate of the covariance matrix of the $\hat{\theta}^*_{ijk}$ is available.

A problem with this approach is that the values of $\check{\phi}^*_{ijk}$ implied by (14) may fall outside the interval [0,1], a situation which may often arise since the $\hat{\phi}^*(i)$ are likely to display appreciable sampling variability, given that this is the reason that a logistic model is being used in the first place. This suggests either imposing a constrained inference procedure or considering a narrower measurement error model specification. One such more restricted assumption, following Chua and Fuller (1987), is

A5 (*unbiased measurement error*): $\phi^*_{ij.} = \phi_{ij.}, \phi^*_{i.k} = \phi_{i.k}\ i = 1,\ldots,I,\ j = 0,1,\ k = 0,1$. In order to study the impact of this assumption, it is conventient to define

$$\alpha^t_i = \text{pr}(y^*_t = 1|y_t = 0,\ \text{cell } i)/\text{pr}(y^*_t = 1|\ \text{cell } i), \tag{15}$$

which measures the ‘amount’ of measurement error at time t in cell i. A consequence of the unbiasedness condition is that the right-hand side of (15) is unchanged if 1 is replaced by 0 and vice-versa. It follows from the assumption that the measurement errors are both independent and unbiased that

$$\pi^*_{ij} = (1-\alpha^1_i)(1-\alpha^2_i)\pi_{ij} + [1-(1-\alpha^1_i)(1-\alpha^2_i)]\phi_{i.1}/\phi_{1..}. \tag{16}$$

Given an estimator $\tilde{\gamma}_i$ of $\gamma_i = [(1-\alpha_i^1)(1-\alpha_i^2)]^{-1}$ (see Fuller (1990), for the case when $\alpha_i^1 = \alpha_i^2$) we may estimate π_{ij} by

$$\tilde{\pi}_{ij} = \tilde{\gamma}_i \hat{\pi}_{ij}^* - (\tilde{\gamma}_i - 1)\hat{\pi}_i^*, \tag{17}$$

$$\text{where } \hat{\pi}_i^* = (\hat{\phi}_{i0.}^* \hat{\pi}_{i0}^* + \hat{\phi}_{i1.}^* \hat{\pi}_{i1}^*)/(\hat{\phi}_{i0}^* + \hat{\phi}_{i1.}^*). \tag{18}$$

An adjusted estimator of β is then obtained by solving (5) with $\phi_{ij.}$ and π_{ij} replaced by $\hat{\phi}_{ij.}^*$ and $\tilde{\pi}_{ij}$ respectively, that is $X'D(\hat{\phi}^*)f(\tilde{\beta}) = X'D(\hat{\phi}^*)\tilde{\pi}$, where $\tilde{\pi} = [\tilde{\pi}_{ij}]$. Letting $\gamma = (\gamma_1, \ldots, \gamma_I)'$ and $\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_I)'$, we make the further assumption:
A6: $n^{1/2}(\tilde{\gamma} - \gamma) \rightarrow N[0, V(\tilde{\gamma})]$ in law as $n \rightarrow \infty$ and $\tilde{\gamma}$ is asymptotically independent of $(\hat{\pi}^{*'}, \hat{\phi}^{*'})$.

The assumption of independence seems plausible if the $\tilde{\gamma}_i$ are derived from a separate study. The asymptotic distribution of $\tilde{\gamma}$ is indexed by n in order to simplify results, even though the separate study may be based on a very different sample size. Since $\hat{\phi}^*$ and $\tilde{\pi}$ are consistent for $\phi^* = \phi$ and π respectively, $\tilde{\beta}$ is consistent for β.

Note that the computation of $\tilde{\beta}$ only requires a simple adjustment to standard packages, such as PROC CATMOD in SAS, which obtain maximum likelihood estimates of logit models. Such packages require as standard input the design matrix X, the sample size n, the vector of cell proportions $\hat{\phi}^*$ and the vector of response proportions $\hat{\pi}^*$. Only the last of these needs adjusting to $\tilde{\pi}$, which from (17) and (18) is simply obtained by multiplying $\hat{\pi}^*$ by a block diagonal matrix, the elements of which depend on $\tilde{\gamma}$ and $\hat{\phi}^*$. Then the standard computational procedure will generate $\tilde{\beta}$, although, of course, the associated standard errors and test procedures will be incorrect.

As a heuristic way of thinking about this adjustment procedure, note that in (16) measurement error shrinks π_{ij}^* towards $\phi_{i.1}/\phi_{i..}$, a weighted average of π_{i0} and π_{i1}, so that the effect of y_1 on y_2 in cell i, as measured by $\pi_{i1} - \pi_{i0,}$ is attenuated by a factor $\gamma_i^{-1}(0 < \gamma_i^{-1} \leq 1)$:

$$\pi_{i1}^* - \pi_{i0}^* = \gamma_i^{-1}(\pi_{i1} - \pi_{i0}^*).$$

The aim of the adjustment is to disattenuate this effect by setting

$$\tilde{\pi}_{i1} - \tilde{\pi}_{i0} = \tilde{\gamma}_i(\hat{\pi}_{i1}^* - \hat{\pi}_{i0}^*). \tag{19}$$

The asymptotic covariance matrix of $\tilde{\beta}$, normalised by $n^{1/2}$, is given by expression (10) with $V[(\hat{\pi}', \hat{\phi}')']$ in (11) replaced by the (normalised) asymptotic covariance matrix of $(\tilde{\pi}', \hat{\phi}^{*\prime})'$. Writing $\tilde{\pi} = g(\hat{\pi}^*, \hat{\phi}^*, \tilde{\gamma})$ and assuming A6 and that $\hat{\pi}^*$ and $\hat{\phi}^*$ obey the asterisked version of A3, we have

$$\begin{aligned} V(\tilde{\pi}) &= \nabla_\pi V(\hat{\pi}^*)\nabla'_\pi + 2\nabla_\pi C(\hat{\pi}^*, \hat{\phi}^*)\nabla'_\phi + \nabla_\phi V(\hat{\phi}^*)\nabla'_\phi + \nabla_\gamma V(\tilde{\gamma})\nabla'_\gamma, \quad (20) \\ C(\tilde{\pi}, \hat{\phi}^*) &= \nabla_\phi V(\hat{\phi}^*) + \nabla_\pi C(\hat{\pi}^*, \hat{\phi}^*), \end{aligned}$$

where ∇_π, ∇_ϕ and ∇_γ are the first partial derivative matrices of $g(\pi^*, \phi, \gamma)$ with respect to π^*, ϕ and γ respectively. Each of these matrices is block diagonal with diagonal element matrices $\nabla_{\pi i}(2 \times 2), \nabla_{\phi i}(2 \times 2)$, and $\nabla_{\gamma i}(2 \times 1)$ respectively $(i = 1, \ldots, I)$, where the elements of $\nabla_{\pi i,} \nabla_{\phi i}$ and $\nabla_{\gamma i}$ are

$$\begin{aligned} \nabla_{\pi ijl} &= \gamma_i \delta_{jl} - (\gamma_i - 1)\phi_{il.}/\phi_{i..} & j, l &= 0, 1, \\ \nabla_{\phi ijl} &= -(\gamma_i - 1)\pi^*_{il}(1 - \phi_{il.}/\phi_{i..})/\phi_{i..}, & j, l &= 0, 1, \\ \nabla_{\gamma ij} &= \pi^*_{ij} - (\pi^*_{i0}\phi_{i0.} + \pi^*_{i1}\phi_{i1.})/\phi_{i..}, & j &= 0, 1, \end{aligned}$$

and where $\phi_{i..} = \phi_{i0.} + \phi_{i1.}$ and δ_{jl} is the Kronecker delta. Given consistent estimators of $V[(\hat{\pi}^{*\prime}, \hat{\phi}^{*\prime})']$ and $V(\tilde{\gamma})$, a consistent estimator of $V(\tilde{\beta})$ may be obtained by substituting $\hat{\pi}^*, \hat{\phi}^*, \tilde{\gamma}$ and $\tilde{\beta}$ for π^*, ϕ, γ and β respectively in $\nabla_\pi, \nabla_{\phi,} \nabla_{\gamma,} \Delta$ and $D(\phi)$, and substituting $f(\tilde{\beta}) - \tilde{\pi}$ for ε in $D(\varepsilon)$.

Whilst the adjustment in $\tilde{\beta}$ removes the inconsistency in $\hat{\beta}^*$, there may be an associated cost in terms of increased variance. There are two ways in which $V(\tilde{\beta})$ can be increased by measurement error. First, there is the uncertainty arising out of estimation of γ, which affects $V(\tilde{\beta})$ through the term $V(\tilde{\gamma})$ in (20). This arises even if all the $\gamma_i = 1$ and there is no measurement error. For in this case, $\nabla_\phi = 0, \nabla_\pi = I$ and ∇_γ has elements $\nabla_{\gamma ij} = \pi_{ij} - \phi_{i.1}/\phi_{i..}$. Hence, $V(\tilde{\beta})$ has the same form as $V(\hat{\beta}^*)$ except that $V(\hat{\pi}^*)$ is replaced by $V(\hat{\pi}^*) + \nabla_\gamma V(\tilde{\gamma})\nabla'_\gamma$ and so $V(\tilde{\beta}) - V(\hat{\beta}^*)$ is a non-negative definite matrix. Often $\tilde{\gamma}$ will be based on a reinterview study with much smaller sample size than the study on which $\hat{\pi}^*$ is based and so the error in estimating γ could, in principle, increases the variance of $\tilde{\beta}$ compared to $\hat{\beta}^*$ substantially.

Second, measurement error could increase $V(\tilde{\beta})$, even if $V(\tilde{\gamma}) = 0$, via the terms γ_i appearing in ∇_π and ∇_ϕ. For example, if $V(\tilde{\gamma}) = 0$, then from (19) the adjusted estimator $\tilde{\pi}_{i1} - \tilde{\pi}_{i0}$ of the contrast $\pi_{i1} - \pi_{10}$ has variance γ_i^2 times the variance of the unadjusted estimator. Whilst one might conjecture that, in practice, the adjusted estimators will tend to have higher variance than the unadjusted estimators even when $V(\tilde{\gamma}) = 0$, this is not necessary theoretically.

For example, supposed that the parameter of interest is $\Psi = \pi_{i1} + 3\pi_{i0}$ and that $\phi_{i0.}/\phi_{i..} = \phi_{i1.}/\phi_{i..} = 0.5$, $\tilde{\gamma}_i = 2$, $V(\hat{\phi}^*) = 0$, $V(\tilde{\gamma}) = 0$; then the unadjusted estimator is $\hat{\Psi}^* = \hat{\pi}^*_{i1} + 3\hat{\pi}^*_{i0}$ and the adjusted estimator is $\tilde{\Psi} = \tilde{\pi}_{i1} + 3\tilde{\pi}_{i0} = 4\hat{\pi}^*_{i0}$. If $\hat{\pi}^*_{i1}$ and $\hat{\pi}^*_{i0}$ are uncorrelated and $\hat{\pi}^*_{i1}$ has much greater variance than $\hat{\pi}^*_{i0}$, then

$\hat{\Psi}^*$ will have greater variance than $\tilde{\Psi}$. Such an example could easily be extended to a comparison of $\tilde{\beta}$ and $\hat{\beta}^*$.

6. An Example

We now illustrate the application of the proposed methods on data from the U.S. Panel Study of Income Dynamics (PSID). Table 1 cross-classifies the variable:

$$\begin{aligned} y_t^* &= 1 \text{ if individual is recorded to be in job covered by union contract} \\ &= 0 \text{ otherwise} \end{aligned}$$

for the two years $t = 1$ (1983) and $t = 2$ (1987), based on men in the self-weighting 'Survey Research Centre sample' (Hill (1992), p.9) who are currently working in both years but are not self-employed nor working for government.

Table 1. Sample counts for observed variables

		In union job in 1987 (y_2^*)	
		No(0)	Yes(1)
In union job in 1983 (y_1^*)	No(0)	684	33
	Yes(1)	43	191

Two factors which might be expected to affect transitions between the states are considered. The first factor is age, which is divided into four categories, 18-29, 30-34, 35-44 45+, which are of roughly equal size for the sample considered. The second factor partitions employment sectors into two categories, roughly according to tendency to be unionized. The first less-unionized category includes professional, managerial, sales and farming employment. The second more-unionized category includes manual and clerical employment. These two factors together define $I = 8$ cells.

Because of difficulties in identifying strata and clusters in the available data file, we ignore here the complexity of the sampling design and assume simple random sampling. Fitting alternative logistic models with y_2^* as the response and examining the likelihood-ratio chi-squared statistic suggests a model with $x_{ij} = (1, j,$ age(2), age(3), age(4), work, j.age(2), j.age(3) j.age(4), where j is the value of y_1^*, age(2)-age(4) are binary indicators representing the age factor and 'work' is a binary indicator of the second factor. Thus the model includes an interaction between age and y_1^*, which reflects the fact that as age increases there is declining mobility either from $y_1^* = 0$ to $y_2^* = 1$ or from $y_1^* = 1$ to $y_2^* = 0$. On the other hand, there seems little evidence of an interaction between

y_1^* and the second factor or between the two factors. Parameter estimates and standard errors are given in the first three columns of Table 2. The standard errors labelled 'model-based' set $D(\varepsilon) = 0$ in (11), whereas those labelled 'robust' allow for non-zero ε. The fact that these two sets of standard errors are similar provides further evidence that the model represents a reasonable approximation.

Table 2. Parameter estimates for logistic model

	Measurement error ignored			Adjusted for measurement error	
Covariate	Estimated Coefficient	Standard Error		Estimated Coefficient	Standard Error
		Model-based	robust		
Constant	−2.81	0.34	0.33	−2.75	0.33
y_1	3.13	0.44	0.44	3.61	0.48
age(2)	−0.69	0.47	0.47	−1.11	0.49
age(3)	−1.02	0.53	0.53	−1.74	0.54
age(4)	−0.93	0.53	0.53	−2.26	0.55
$y_1$age(2)	0.80	0.65	0.65	1.19	0.71
$y_1$age(3)	1.92	0.75	0.74	2.74	0.80
$y_1$age(4)	2.54	0.76	0.76	4.76	0.84
work	0.63	0.30	0.29	0.32	0.29

One source of information on measurement error is the PSID Validation Study (Bound et al. (1990), Hill (1992), p.29) This study involved comparing responses to the PSID instrument with company records for a sample of workers from one large firm. A cross-classification of validated and survey responses on the reponse variable in 1987 is given in Table 3.

Table 3. Survey responses by responses from validation study in 1987

		In union job in survey (y_2^*)	
		No(0)	Yes(1)
In union job in Validation Study (y_2)	No(0)	140	8
	Yes(1)	2	302

Assuming that the misclassification matrices for each cell i in the validation study and the general population are the same and that errors are independent (assumption A4) and identically distributed over time, observed counts in Table 1 are adjusted according to the approach in (14) to the first table in Table 4.

Table 4. Adjusted counts under alternative measurement models

		Common Misclassification Matrices		Unbiased Errors, common α	
		y_2		y_2	
		No(0)	Yes(1)	No(0)	Yes(1)
y_1	No(0)	764	-8	695	22
	Yes(1)	3	192	32	202

Under this measurement model, it appears that essentially all the observed transitions can be explained by measurement error and hence there is no purpose in continuing to fit a logistic model. As an alternative measurement model suppose that measurement errors are now not only independent and identically distributed over time but they are also unbiased (assumption A5) in both the general population within cells and in the Validation Study. We estimate the parameter α in (15) from the 2×2 table in Table 3 by numerically maximising the likelihood under a multinomial model with equal marginal distributions. The estimate is $\hat{\alpha} = 0.051$ (with a standard error calculated from the observed information matrix of 0.016) and the Pearson chi-squared test statistic for the hypothesis of unbiased measurement error is 3.6 on 1 d.f. (not significant at a 95% level). Assuming that α (rather than the entire misclassification matrix) is the same in the Validation Study and within each cell in the general population and is the same over time, the adjusted count matrix, following the approach in (17) and summing across cells, is the second table in Table 4. This adjustment is quite different from the first and implies a more moderate adjustment of Table 1. The reason for the difference is that the marginal distribution of y_2^* is very different in the Validation Study and the general population. Hence assuming unbiased measurement errors with common α in both populations implies very different misclassification matrices.

Extending the adjustment under the unbiased error model to the logistic model following the approach in Section 5 and assuming a common α over all cells gives the adjusted estimates in Table 2. Note that even though the adjustment may appear small with $\hat{\alpha}$ only equal to 0.05, the effect on the coefficients of age and of the interaction between age and y_1 are very marked. For example, the estimated ratio of the odds of staying in a job covered by a union contract for men aged 45+ compared to men aged 18-29 rises from about 5 to about 12 as we adjust for measurement error. The adjusted standard errors allow for the error in estimating α and it is reassuring that these are not much larger than the original standard errors.

Acknowledgements

Research was supported by grant numbers R000 23 2522 and H519255005 from the Economic and Social Research Council. The author is grateful to Keith Humphreys for carrying out the computational work in Section 6. The data were made available through the ESRC Data Archive by the Inter-University Consortium for Political and Social Research, Ann Arbor, Michigan. The data were originally collected by J. N. Morgan. Neither the original investigator, nor the Consortium, nor the Archive bear any responsibility for the analysis presented here.

References

Abowd, H. M. and Zellner, A. (1985). Estimating gross flows. *J. Business and Economic Statist.* **3**, 254-283.

Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics* **43**, 951-973.

Bound, J., Brown, C., Duncan, G. J. and Rodgers, W. L. (1990). Measurement error in cross-sectional and longitudinal labor market surveys: validation survey evidence. In *Panel Data and Labor Market Studies* (Edited by J. Hartog, G. Ridder, J. Theeuwes). Elsevier Science Publishers BV.

Chambers, R. L., Woyzbun, L. and Pillig, R. (1988). Maximum likelihood estimation of gross flows. *Austral. J. Statist.* **30**, 149-162.

Chua, T. C. and Fuller, W. A. (1987). A model for multinomial response error applied to labor flows. *J. Amer. Statist. Assoc.* **82**, 46-51.

Fuller, W. A. (1990). Analysis of repeated surveys. *Survey Methodology* **16**, 167-180.

Hill, M. S. (1992). *The Panel Study of Income Dynamics: A User's Guide.* Newbury Park, Sage.

Hogue, C. R. and Flaim, P. O. (1986). Measuring gross flows in the labor force: an overview of a spatial conference. *J. Business and Economic Statist.* **4**, 111-121.

Holt, D. and Skinner, C. J. (1989). Components of change in repeated surveys. *Int. Statist. Rev.* **57**, 1-18.

Hsiao, C. (1986). *Analysis of Panel Data.* Cambridge: Cambridge University Press.

Korn, E. L. and Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* **35**, 795-802.

Maddala, G. S. (1987). Limited dependent variable models using panel data. *J. Human Resources* **22**, 307-338.

Meyer, B. D. (1988). Classification-error models and labor-market dynamics. *J. Business and Economic Statist.* **6**, 385-390.

Muenz, L. R. and Rubinstein, L. V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics* **41**, 91-101.

Roberts, G., Rao, J. N. K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**, 1-12.

Scott, A. J. and Wild, C. J. (1989). Selection based on the response variable in logistic regression. In *Analysis of Complex Surveys* (Edited by C. J., Skinner, D. Holt, and T. M. F. Smith), 191-205. Chichester: Wiley, Singh, A. C. and Rao, J. N. K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *J. Amer. Statist. Assoc.* **90**, 478-488.

Wolter, K. M. (1985). *Introduction to Variance Estimation.* Springer Verlag, New York.

Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom.

E-mail: cjs@soton.ac.uk

(Received May 1996; accepted August 1997)