Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

18-09

Inference Based on Estimating Equations and Probability-Linked Data

Ray Chambers, James Chipperfield, Walter Davis, Milorad Kovacevic

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Inference Based on Estimating Equations and Probability-Linked Data

Ray Chambers, University of Wollongong

James Chipperfield, Australian Bureau of Statistics

Walter Davis, Statistics New Zealand

Milorad Kovacevic, Statistics Canada

Abstract

Data obtained after probability linkage of administrative registers will include errors due to the fact that some linked records contain data items sourced from different individuals. Such errors can induce bias in standard statistical analyses if ignored. In this paper we describe some approaches to eliminating this bias when parametric inference is based on solution of an estimating equation, with an emphasis on linear and logistic regression analysis. Simulation results that illustrate the gains from allowing for linkage error when using probabilistically linked data to carry out these analyses are presented, as are extensions of the approach to more complex linkage situations. In particular, we explore issues that arise when sample records are linked to administrative records and also where the target of inference is the solution to the estimating equation defined by the perfectly linked data. A substantial application that illustrates the use of these ideas in identifying the major sources of error when modeling data obtained by probabilistically linking two successive Australian censuses is described.

1. Introduction

Record linkage (Fellegi and Sunter, 1969) allows data for a single individual to be compiled from different data sources, enabling more powerful and effective analyses to be carried out than would otherwise be the case. In particular, datasets created by linking individual records constitute a critical resource for research in health, epidemiology, economics, demography, sociology and many other scientific areas. For example, Iron *et al*. (2003) describe a longitudinal dataset created by linking hospital admission and general practitioner records to private health insurance expenditure records with the aim of building models for how changes in health expenditure influence subsequent uptake of medical services. National statistical agencies increasingly rely on linking surveys to administrative registers to provide more accurate measurement and to reduce respondent burden. Frequently, one or more datasets (whether all administrative data or a mix of administrative and survey data) are linked to answer a broader array of research questions than can be addressed through any of the datasets individually. Thus, the linked longitudinal employer-employee dataset based on linking administrative data held on the New Zealand (NZ) Inland Revenue Department's tax system and Statistics New Zealand's list of NZ businesses allows the analysis of job and worker flows, employment tenure, multiple job holding and business demography. Similarly, the Census Data Enhancement Initiative of the Australian Bureau of Statistics (ABS) aims to create a Statistical Longitudinal Census Dataset that integrates census data from the same individuals over a number of censuses, with the objective of building a research resource for longitudinal analysis of the Australian population.

Linking datasets often involves a probabilistic matching of records from one dataset to another. In particular, matching variables present in both datasets are used to maximise the probability that the values of the variables making up the linked record are the correct

measurements for the population unit corresponding to that record. Statistical methods for linking datasets are now well established (Herzog *et al*., 2007), with recent statistical research in this area mainly focused on the confidentiality issues that arise as a consequence of linkage. In contrast, aside from the notable contributions of Neter *et al*. (1965), Scheuren and Winkler (1993) and Lahiri and Larsen (2005), there has been comparatively little methodological research carried out on the impact of linkage errors on subsequent statistical analysis. Linkage errors are the errors caused by incorrectly linking different population units as well as the errors caused by not linking the same population units in the datasets that are linked. These errors are a particular type of measurement error, and can lead to biased inference.

In this paper we describe a methodological framework that can be used to provide appropriate modifications to standard statistical analysis methods to ensure that they remain unbiased when used with probabilistically linked data. The framework is based on modelling the relationship between the probabilistically linked data and the 'true' data that would be obtained if error free linkage were possible. Our assumptions about the data linkage situation and a description of a simple model for linkage errors are set out in the following Section. In Section 3 we apply these ideas to where the statistical model of interest is fitted via the solution of an estimating equation, with applications to linear and logistic regression. Simulation results described in Section 4 illustrate the potential gains from the modified analytic methods that we propose. In Section 5 we show how the model-based inferential methods described in Section 3 can be extended to fixed finite population inference based on probabilistically linked data, and in Section 6 we describe a further extension that can accommodate sample to population linkage, together with an application to assessment of potential sources of bias in a linked census data set under

development by the ABS. Finally, in Section 7 we conclude the paper with a short discussion of avenues for further research.

2. A Model for Linkage Error

We assume the existence of a population of N units, indexed by $i = 1,\ldots,N$, such that, for each unit, it is possible to measure the values of a scalar random variable Y and a vector random variable X. We are interested in the relationship between Y and X in this population, and in particular our aim is to fit a model of the form $E(Y|X) = g(X;\theta)$ for the regression of Y on X. Here g corresponds to a known functional form while the parameter θ is unknown and needs to be estimated. This is usually straightforward if we have the values of Y and X for a random sample of units from this population. Unfortunately, we do not have such a sample. Instead, we have access to two registers that separately contain the population values of Y and X. We refer to these as the Y-register and X-register respectively from now on. We also assume that both registers are for the same population and have no duplicates, so each is made up of N records.

If each unit in the population has a unique identifier, and this identifier is stored on both registers, we can use it to link records from the two registers, and then estimate θ using the Y and X values associated with the N linked records. Unfortunately, such a unique identifier does not exist. Instead, a probability-linking algorithm is used to associate (i.e. link) records on the X-register with records on the Y-register. This algorithm makes it is possible (at least conceptually) to link every record on the X-register with a record on the Y-register. That is, linkage is *complete* and *one to one* between the Y and X-registers. Clearly, the data set constructed by this process (the linked data) can contain linkage errors, i.e. records where the values of Y and X actually come from different population units.

Although it may be theoretically possible for any two records on the *Y* and *X*-registers to be linked, most reasonable probability linking algorithms will only attempt to link records that are similar in some sense. Consequently, we shall assume that the linked records can be partitioned into *Q* distinct 'strata' such that there is no possibility that linked records in different strata contain data for the same population unit. We model this situation by assuming that the different strata correspond to different values of a categorical population variable *Z* that can be derived from the information on either register, and which is defined in such a way that if a record on one register does not have the same value of *Z* as the record on the other register, then it is reasonable to assume that these two records cannot correspond to the same unit in the original population. Conversely, the fact that a *Y*-register record and an *X*-register record have the same value for *Z* does not guarantee that they correspond to the same unit, and so linkage errors can still occur within a stratum. We refer to *Z* as a stratifying variable, and those population units with the same value of *Z* as being in the same stratum. Note that errors in measurement of *Z* can lead to the same population unit having one value of *Z* on the *Y*-register and another on the *X*-register, which invalidates the assumption of no linkage errors when *Y* and *X*-register records have different values of *Z*. Consequently, we shall assume that *Z* is measured without error on both the *Y*-register and the *X*-register. With this set up, data linkage errors only occur among records in both registers in the same stratum.

Without loss of generality, we denote the *Q* distinct values taken by *Z* by $1,2,\ldots,Q$. Let stratum q correspond to the M_q population units with $Z = q$, so $N = \sum_q M_q$. Since *Z* is measured without error in both registers, and linkage is complete, the number of records in stratum q in each register is the same, i.e. M_q. Let *i* index the records in the linked data set. Again, without loss of generality we assume that this index is the same as the one used to index

the *X*-register, i.e. the linkage process associates a record from the *Y*-register, with its associated *Y*-value, with each record on the *X*-register. In stratum q we then have M_q linked data pairs (y_i^*, X_i), where y_i^* denotes the *Y*-value from stratum q on the *Y*-register that is matched to X_i. More accurately, the record with $Y = y_i^*$ in stratum q on the *Y*-register is matched to the record with $X = X_i$ in stratum q on the *X*-register. We use $\mathbf{y}_q^*$ to denote the vector of order M_q defined by the linked values y_i^* in stratum q and $\mathbf{X}_q$ as the matrix with rows defined by the values X_i in the same stratum. Also, let $\mathbf{y}_q$ denote the unknown vector of order M_q with entries indexed as in the *X*-register that corresponds to the 'true' *Y* values associated with $\mathbf{X}_q$.

Since one and only one of the M_q records in stratum q on the *Y*-register can be matched to each distinct record in the corresponding stratum on the *X*-register, we model randomness in the outcome of the linkage process via the identity

$$\mathbf{y}_q^* = \mathbf{A}_q \mathbf{y}_q \qquad (1)$$

where $\mathbf{A}_q = [a_{ij}^q]$ is an unknown random permutation matrix of order M_q. Note that the entries a_{ij}^q of $\mathbf{A}_q$ are either zero or one, with a value of one occurring just once in each row and column. That is, $\mathbf{A}_q$ is doubly stochastic. Also, since we are assuming that linkage errors are confined to strata, it is natural to impose the condition that for $q_1 \neq q_2$, $\mathbf{A}_{q_1}$ and $\mathbf{A}_{q_2}$ are independent.

Clearly inference based on linked data will involve assumptions about the distribution of the $\mathbf{A}_q$. In this paper we assume that linkage is non-informative at each level of *Z*, i.e. the distribution of $\mathbf{A}_q$ is independent of $\mathbf{y}_q$ given $\mathbf{X}_q$, and we put

$$E\left(\mathbf{A}_q \middle| \mathbf{X}_q\right) = \mathbf{E}_q . \qquad (2)$$

Given the care that typically goes into the construction of a linked data set, it seems reasonable that a 'declared' link is more likely to be correct than incorrect. Although the probability that such a link is correct will typically vary between the records that make up the linked dataset, as a first approximation we assume that the probability of correct linkage is the same for all records in a stratum. We also assume that it is equally likely that any two *Y*-records in the same stratum that are not linked to a particular *X*-record in that stratum could in fact be the 'correct' link for this record. A simple way of characterising both of these assumptions is via an exchangeable linkage errors model, where for each value of q

$$\Pr\left(\text{correct linkage}\right) = \Pr\left(a_{ii}^{q} = 1\right) = \lambda_q \tag{3}$$

and, for $i \neq j$,

$$\Pr\left(\text{incorrect linkage}\right) = \Pr\left(a_{ij}^{q} = 1\right) = \gamma_q . \tag{4}$$

Given (3) and (4) hold, it follows that (2) is then of the form

$$\mathbf{E}_q = \left(\lambda_q - \gamma_q\right)\mathbf{I}_q + \gamma_q \mathbf{J}_q \tag{5}$$

where $\mathbf{I}_q$ is the identity matrix of order M_q and $\mathbf{J}_q$ denotes a square matrix of ones of order M_q. Since $\mathbf{A}_q$ is doubly stochastic, $\mathbf{E}_q$ must be as well, which means that (5) implies

$$\lambda_q + (M_q - 1)\gamma_q = 1 . \tag{6}$$

In other words, we just need to specify λ_q in order to completely specify the first order properties of the linkage mechanism under the model (5). This will be particularly useful later since estimation of λ_q requires only that we know whether a defined link is correct or incorrect, and not the identity of the correct link.

The model specified by (3) and (4) represents what is probably the simplest way of characterising the behaviour of a probability-based linkage process, and will form the basis for

the theory developed in this report. It was originally suggested by Neter *et al*. (1965) in a groundbreaking paper that investigated its use in assessing the impact of linkage error on response error analysis, where alternative data sources were linked to respondent records in order to assess the extent of response error in these records. As these authors note, and as we shall see in next Section, the impact of linkage error defined by (3) and (4) is to attenuate the relationship between the study variable (in their case the difference between the survey value and the linked alternative value) and explanatory covariates.

Although we do not explore them in this paper, it is clear that more sophisticated models for linkage error can be formulated. For example, it may be the case that the *Y* and *X*-registers are ordered so that only 'nearby' records in the linked data can possibly correspond to the correct link. Another extension is where there exists another variable on the *X*-register, say *W*, with values w_i that vary within a stratum, such that the probability of a correct link depends on these values. Note however that if *W* is categorical and available on both registers then by including it in the definition of the stratifying variable *Z* we recover the situation implicit in the exchangeable linkage errors model, where all linked records in the same stratum have the same probability of being incorrectly linked.

3. Estimating Equations and Probability-Linked Data

In what follows the aim of inference is to fit a regression model with parameter θ using an unbiased estimating function $H(\theta)$ and data that have been probability-linked. In particular we consider the case where, given the true values of y_i, $H(\theta)$ is of the form

$$H(\theta) = \sum_{i=1}^{N} G_i(\theta)\{y_i - f_i(\theta)\}. \qquad (7)$$

Here $f_i(\theta_0) = E\left(y_i \middle| X_i\right) = E_X\left(y_i\right)$ and $G_i(\theta)$ is a vector of order *p* which is a function of θ and X_i but not of y_i. Clearly (7) defines an unbiased estimating function for θ_0, which we can write in 'stratified' form as

$$H(\theta) = \sum_q \mathbf{G}_q(\theta)\left\{\mathbf{y}_q - \mathbf{f}_q(\theta)\right\} \tag{8}$$

where $\mathbf{G}_q(\theta)$ is the $p \times M_q$ matrix with columns defined by the vectors $G_i(\theta)$ associated with the population units making up stratum *q*, and $\mathbf{f}_q(\theta)$ is the vector of order M_q defined by their corresponding values of $f_i(\theta)$.

Now consider the situation where instead of $\mathbf{y}_q$, we have access to a probability-linked version of this vector, $\mathbf{y}_q^* = \mathbf{A}_q\mathbf{y}_q$. Here $\mathbf{A}_q$ is a random permutation matrix of order M_q distributed independently of $\mathbf{y}_q$ given the values in $\mathbf{X}_q$ (i.e. linkage is non-informative given the values of the explanatory variables), with values of $\mathbf{A}_q$ distributed independently between strata and where $E_X\left(\mathbf{A}_q\right) = \mathbf{E}_q$ is known. Let $H^*(\theta)$ denote the value of (8) when we use $\mathbf{y}_q^*$ instead of $\mathbf{y}_q$, with $\hat{\theta}^*$ defined by $H^*(\hat{\theta}^*) = 0$. Since $E_X\left\{H^*(\theta_0)\right\} = \sum_q \mathbf{G}_q(\theta_0)\left\{\left(\mathbf{E}_q - \mathbf{I}_q\right)\mathbf{f}_q(\theta_0)\right\} \neq 0$, we see that $H^*(\theta)$ is biased if linkage is not perfect, and so the naive estimator $\hat{\theta}^*$ is also biased in this case. Since we know the value of $\mathbf{E}_q$, we can correct for this bias, replacing the estimating function $H^*(\theta)$ by its bias-corrected version:

$$H^*_{adj}(\theta) = H^*(\theta) - \sum_q \mathbf{G}_q(\theta)\left\{\left(\mathbf{E}_q - \mathbf{I}_q\right)\mathbf{f}_q(\theta)\right\} = \sum_q \mathbf{G}_q(\theta)\left\{\mathbf{y}_q^* - \mathbf{E}_q\mathbf{f}_q(\theta)\right\}. \tag{9}$$

A bias-adjusted estimator of θ based on the linked data is then $\hat{\theta}^*_{adj}$ where

$$H^*_{adj}(\hat{\theta}^*_{adj}) = 0. \tag{10}$$

In what follows we use ∂_x to denote partial differentiation with respect to the components of *x* and apply standard results for inference based on unbiased estimating functions to (9). In particular, the large sample variance of $\hat{\theta}^*_{adj}$ is then

$$Var_X(\hat{\theta}^*_{adj}) \approx \left(\partial_\theta H^*_{adj}\Big|_{\theta=\theta_0}\right)^{-1} Var_X\left\{H^*_{adj}(\theta_0)\right\}\left\{\left(\partial_\theta H^*_{adj}\Big|_{\theta=\theta_0}\right)^{-1}\right\}^T$$

with plug-in sandwich-type estimator

$$\hat{V}_X(\hat{\theta}^*_{adj}) = \left\{\partial_\theta H^*_{adj}(\hat{\theta}^*_{adj})\right\}^{-1} \hat{V}_X\left\{H^*_{adj}(\theta_0)\right\}\left[\left\{\partial_\theta H^*_{adj}(\hat{\theta}^*_{adj})\right\}^{-1}\right]^T. \qquad (11)$$

where $\partial_\theta H^*_{adj}(\hat{\theta}^*_{adj}) = \left(\partial_\theta H^*_{adj}\right)\Big|_{\theta=\hat{\theta}^*_{adj}}$ and $\hat{V}_X\left\{H^*_{adj}(\theta_0)\right\}$ is an estimator of $Var_X\left\{H^*_{adj}(\theta_0)\right\}$.

In order to define $\hat{V}_X\left\{H^*_{adj}(\theta_0)\right\}$ we put $Var_X(\mathbf{y}_q) = \boldsymbol{\Omega}_q(\theta_0)$ and $\mathbf{V}_q(\theta_0) = Var_X\left\{\mathbf{A}_q\mathbf{f}_q(\theta_0)\right\}$, and observe that

$$Var_X(\mathbf{y}^*_q) = E_X\left\{\mathbf{A}_q\boldsymbol{\Omega}_q(\theta_0)\mathbf{A}_q^T\right\} + \mathbf{V}_q(\theta_0) = \Sigma_q(\theta_0). \qquad (12)$$

Hence $Var_X\left\{H^*_{adj}(\theta_0)\right\} = \sum_q \mathbf{G}_q(\theta_0) Var_X\left(\mathbf{y}^*_q\right)\mathbf{G}_q^T(\theta_0) = \sum_q \mathbf{G}_q(\theta_0)\Sigma_q(\theta_0)\mathbf{G}_q^T(\theta_0)$, which suggests the plug-in estimator

$$\hat{V}_X\left\{H^*_{adj}(\theta_0)\right\} = \sum_q \mathbf{G}_q(\hat{\theta}^*_{adj})\Sigma_q(\hat{\theta}^*_{adj})\mathbf{G}_q^T(\hat{\theta}^*_{adj}). \qquad (13)$$

Before we can compute (13) we need to estimate the covariance matrix $\Sigma_q(\theta)$ specified by (12). Clearly, estimation of its first component $E_X\left\{\mathbf{A}_q\boldsymbol{\Omega}_q(\theta_0)\mathbf{A}_q^T\right\}$ will depend on the second order properties of our model for $\mathbf{y}_q$. On the other hand, in Appendix 1 of Chambers (2008) it is shown that a first order approximation to $Var_X\left\{\mathbf{A}_q\mathbf{f}_q\right\}$ under the exchangeable linkage errors model defined by (3) and (4) is

$$Var_X\left\{\mathbf{A}_q\mathbf{f}_q(\theta)\right\} \approx diag\left[(1-\lambda_q)\left\{\lambda_q\left(f_i(\theta)-\bar{f}_q\right)^2+\bar{f}_q^{(2)}-\bar{f}_q^2\right\}\right]$$

where $\bar{f}_q$ is the stratum q mean of $f_i(\theta)$ and $\bar{f}_q^{(2)}$ is the corresponding mean of $f_i^2(\theta)$. This expression can be used to estimate $\mathbf{V}_q(\theta_0)$ after replacing $f_i(\theta)$ by $f_i(\hat{\theta}^*_{adj})$.

In order to define the matrix of partial derivatives $\partial_\theta H^*_{adj}(\hat{\theta}^*_{adj})$ in (11) we note that although in theory $\partial_\theta H^*_{adj} = \sum_q \partial_\theta\left[\mathbf{G}_q(\theta)\left\{\mathbf{y}^*_q - \mathbf{E}_q\mathbf{f}_q(\theta)\right\}\right]$, it is often the case that $\mathbf{G}_q(\theta)$ varies little as θ changes. Consequently, we put

$$\partial_\theta H^*_{adj}(\hat{\theta}^*_{adj}) = -\sum_q \mathbf{G}_q(\hat{\theta}^*_{adj})\mathbf{E}_q\left\{\partial_\theta \mathbf{f}_q(\theta)\big|_{\theta=\hat{\theta}^*_{adj}}\right\}. \qquad (14)$$

The final variance estimator for $\hat{\theta}^*_{adj}$ is then obtained by substituting (13) and (14) into (11).

3.1 Application to Linear Regression

Here our focus is on the linear regression model defined by

$$E_X\left(\mathbf{y}_q\right) = \mathbf{X}_q\beta = \mathbf{f}_q \qquad (15)$$

$$Var_X\left(\mathbf{y}_q\right) = \sigma^2\mathbf{I}_q. \qquad (16)$$

Note that in addition to the regression parameter β in (15), which corresponds to θ above and is the target of inference, (16) now includes an unknown nuisance parameter σ^2. Given the $\mathbf{y}_q$ and $\mathbf{X}_q$, the optimal estimator of β is

$$\hat{\beta} = \left[\sum_q \mathbf{X}_q^T\mathbf{X}_q\right]^{-1}\left[\sum_q \mathbf{X}_q^T\mathbf{y}_q\right]. \qquad (17)$$

Unfortunately, unless the linkage is perfect, (17) cannot be calculated. Instead, what is usually done is to substitute the linked data values $\mathbf{y}^*_q$ for $\mathbf{y}_q$ in this expression, which leads to the naïve linked data OLS estimator

$$\hat{\beta}^{*} = \left[\sum_q \mathbf{X}_q^T \mathbf{X}_q\right]^{-1}\left[\sum_q \mathbf{X}_q^T \mathbf{y}_q^{*}\right]. \qquad (18)$$

However, under non-informative linkage, $E_X\left(\mathbf{y}_q^{*}\right) = E_X\left(\mathbf{A}_q\right)E_X\left(\mathbf{y}_q\right) = \mathbf{E}_q\mathbf{X}_q\beta$, and so the $\mathbf{y}_q^{*}$ also follow a linear model with regression coefficient β but with a modified set of explanatory variables $\mathbf{H}_q = \mathbf{E}_q\mathbf{X}_q$ in stratum *q*. Lahiri and Larsen (2005) note this relationship and suggest estimation of β using the OLS estimator for this situation,

$$\hat{\beta}_A = \left(\sum_q \mathbf{H}_q^T\mathbf{H}_q\right)^{-1}\left(\sum_q \mathbf{H}_q^T\mathbf{y}_q^{*}\right) = \left(\sum_q \mathbf{X}_q^T\mathbf{E}_q^T\mathbf{E}_q\mathbf{X}_q\right)^{-1}\left(\sum_q \mathbf{X}_q^T\mathbf{E}_q^T\mathbf{y}_q^{*}\right). \qquad (19)$$

This estimator can be improved upon. Under (15) and (16), $Var_X\left(\mathbf{y}_q^{*}\right) = \sigma^2\mathbf{I}_q + \mathbf{V}_q(\beta) = \Sigma_q$, where $\mathbf{V}_q(\beta) = Var_X\left\{\mathbf{A}_q\mathbf{X}_q\beta\right\}$. Given Σ_q, the Best Linear Unbiased (BLU) estimator for β is therefore

$$\hat{\beta}_C = \left(\sum_q \mathbf{H}_q^T\Sigma_q^{-1}\mathbf{H}_q\right)^{-1}\left(\sum_q \mathbf{H}_q^T\Sigma_q^{-1}\mathbf{y}_q^{*}\right) = \left(\sum_q \mathbf{X}_q^T\mathbf{E}_q^T\Sigma_q^{-1}\mathbf{E}_q\mathbf{X}_q\right)^{-1}\left(\sum_q \mathbf{X}_q^T\mathbf{E}_q^T\Sigma_q^{-1}\mathbf{y}_q^{*}\right). \qquad (20)$$

An 'empirical' version of (20), or EBLUE, is defined by iterating between (20) and the following unbiased estimator of σ^2:

$$\hat{\sigma}^2 = N^{-1}\left\{\sum_q\left(\mathbf{y}_q^{*} - \mathbf{f}_q\right)^T\left(\mathbf{y}_q^{*} - \mathbf{f}_q\right) - 2\sum_q \mathbf{f}_q^T\left(\mathbf{I}_q - \mathbf{E}_q\right)\mathbf{f}_q\right\}. \qquad (21)$$

It is not difficult to see that the Lahiri-Larsen estimator (19) and the BLUE (20) are special cases of solution of an estimating equation of the form $\sum_q \mathbf{G}_q\left\{\mathbf{y}_q^{*} - \mathbf{E}_q\mathbf{X}_q\beta\right\} = 0$, and so fit into the general estimating equation theory set out at the start of the Section. In particular, they can be obtained from (10) by setting $\theta \equiv \beta$ and $\mathbf{f}_q(\beta) = \mathbf{X}_q\beta$ with $\mathbf{G}_q = \mathbf{X}_q^T\mathbf{E}_q^T$ in the case of (19) and $\mathbf{G}_q = \mathbf{X}_q^T\mathbf{E}_q^T\Sigma_q^{-1}$ in the case of (20). In either case, a sandwich-type estimator of variance can be obtained by appropriate substitution in (11).

3.2 Application to Logistic Regression

Linear logistic regression is widely used for regression modelling of categorical data. Here $\mathbf{f}_q(\theta) = \{f_i(\theta); i \in q\}$ where

$$f_i(\theta) = \left(1 + \exp(X_i^T\theta)\right)^{-1} \exp(X_i^T\theta). \qquad (22)$$

It follows that $\partial_\theta \mathbf{f}_q(\theta) = \mathbf{D}_q(\theta)\mathbf{X}_q$ where $\mathbf{D}_q(\theta) = diag\left[f_i(\theta)\{1 - f_i(\theta)\}\right]$. The standard maximum likelihood estimating function (i.e. the score function) for the logistic regression model puts $\mathbf{G}_q(\theta) = \mathbf{X}_q^T$ in (8). However, this is not efficient when we estimate θ via the adjusted estimating equation (9), and it is preferable to use the expressions for $\mathbf{G}_q(\theta)$ that lead to the Lahiri-Larsen and BLU estimators in the linear regression case. In the former case this implies use of $\mathbf{G}_q(\theta) = \mathbf{X}_q^T\mathbf{E}_q^T$, while in the latter case the same efficiency argument that motivated the BLUE (20) in this case leads us to use the second order efficient version of (8), which in the logistic case is given by

$$\mathbf{G}_q(\theta) = \partial_\theta \left\{E_X\left(\mathbf{y}_q^*\right)\right\} Var_X^{-1}\left(\mathbf{y}_q^*\right) = \left\{\partial_\theta \mathbf{f}_q(\theta)\right\}\mathbf{E}_q^T\Sigma_q^{-1}(\theta) = \mathbf{X}_q^T\mathbf{D}_q(\theta)\mathbf{E}_q^T\Sigma_q^{-1}(\theta). \qquad (23)$$

It is easy to see that the corresponding optimal version of $\mathbf{G}_q(\theta)$ in the linear case is $\mathbf{G}_q = \mathbf{X}_q^T\mathbf{E}_q^T\Sigma_q^{-1}$ and leads to (20). In either case, variance estimation for the solution to the estimating equation (10) uses the plug-in sandwich estimator (11), with $\partial_\theta H_{adj}^*(\hat{\theta}_{adj}^*)$ defined by (14) and with $\hat{V}_X\left\{H_{adj}^*(\theta_0)\right\}$ given by (13). In order to compute the latter expression, we observe that under (22), $E_X\left\{\mathbf{A}_q\Omega_q(\theta_0)\mathbf{A}_q^T\right\} = E_X\left\{\mathbf{A}_q\mathbf{D}_q(\theta_0)\mathbf{A}_q^T\right\} = diag\left\{\sum_{j=1}^{M_q} f_j(\theta)\{1 - f_j(\theta)\}e_{ij}^q\right\}$, where $\mathbf{A}_q = \left[a_{ij}^q\right]$ and $\mathbf{E}_q = \left[e_{ij}^q\right]$.

3.3 Using Estimated Linkage Error Probabilities

The development so far has assumed that the matrix of expected values $\mathbf{E}_q$ for the stochastic linkage matrix $\mathbf{A}_q$ is known. If this matrix is specified using the exchangeable errors model (5) then this is equivalent to assuming that the probabilities λ_q of correct linkage are known. This is unlikely, and these probabilities will usually be estimated in some way. The extra uncertainty arising from this estimation needs to be accounted for when carrying out variance estimation for the estimators of θ that use $\mathbf{E}_q$ to correct for bias induced by linkage errors.

Let λ denote the vector defined by the stratum-specific values of λ_q. The estimating function (9) then needs to be replaced by $H^*_{adj}(\theta,\lambda) = \sum_q \mathbf{G}_q(\theta)\left\{\mathbf{y}^*_q - \mathbf{E}_q(\lambda)\mathbf{f}_q(\theta)\right\} = \sum_q \mathbf{U}_q(\theta,\lambda_q)$, which is now considered to be a function of both θ and λ, allowing us to develop a first order Taylor series approximation of the form $\hat{\theta} \approx \theta_0 + \left(\partial_\theta H^*_0\right)^{-1}\left\{H^*_0 + \partial_\lambda H^*_0\left(\hat{\lambda} - \lambda_0\right)\right\}$, where $H^*_0 = H^*_{adj}(\theta_0,\lambda_0)$. Here θ_0 and λ_0 denote the 'true' values of these parameters with $\hat{\theta}$ and $\hat{\lambda}$ their corresponding estimators. It follows that we can approximate the variance of $\hat{\theta}$ by

$$Var_X\left(\hat{\theta}\right) \approx \left\{\sum_q \partial_\theta \mathbf{U}_q(\theta_0,\lambda_{0q})\right\}^{-1}\left[\sum_q\left(\mathbf{G}_q(\theta_0)\Sigma_{0q}\mathbf{G}^T_q(\theta_0) + \Psi_q\right)\right]\left[\left\{\sum_q \partial_\theta \mathbf{U}_q(\theta_0,\lambda_{0q})\right\}^{-1}\right]^T$$

where $\Psi_{2q} = \left(\partial_{\lambda_q}\mathbf{U}_q(\theta_0,\lambda_{0q})\right)Var_X\left(\hat{\lambda}_q\right)\left(\partial_{\lambda_q}\mathbf{U}_q(\theta_0,\lambda_{0q})\right)^T$. Note that we have also assumed that the distribution of $\hat{\lambda}$ is (at least approximately) independent of the distribution of H^*_0. Under (5), $\partial_{\lambda_q}\mathbf{U}_q(\theta_0,\lambda_{0q}) = -(M_q-1)^{-1}\mathbf{G}_{0q}\left(M_q\mathbf{I}_q - \mathbf{J}_q\right)\mathbf{f}_{0q}$. Consequently

$$Var_X\left(\hat{\theta}\right) \approx \left(\sum_q \partial_\theta \mathbf{U}_{0q}\right)^{-1}\left\{\sum_q \mathbf{G}_{0q}\left(\Sigma_{0q} + \Delta_{0q}\right)\mathbf{G}^T_{0q}\right\}\left\{\left(\sum_q \partial_\theta \mathbf{U}_{0q}\right)^{-1}\right\}^T \qquad (24)$$

where $\Delta_{0q} = (M_q-1)^{-2}Var_X\left(\hat{\lambda}_q\right)\left(M_q\mathbf{I}_q - \mathbf{J}_q\right)\mathbf{f}_{0q}\mathbf{f}^T_{0q}\left(M_q\mathbf{I}_q - \mathbf{J}_q\right)$.

A simple way of estimating the linkage probabilities is to check a random 'audit' sample of linked records in each stratum, in which case $Var_X\left(\hat{\lambda}_q\right) = m_q^{-1}\lambda_{0q}\left(1-\lambda_{0q}\right)$. Variance estimation based on (24) then follows in the usual way, by plugging in estimates for unknown quantities. That is, our estimator of $Var_X\left(\hat{\theta}\right)$ is

$$\hat{V}_X\left(\hat{\theta}\right) = \left(\sum\nolimits_q \partial_\theta \hat{\mathbf{U}}_q\right)^{-1}\left[\sum\nolimits_q \hat{\mathbf{G}}_q\left(\hat{\Sigma}_q + \hat{\Delta}_q\right)\hat{\mathbf{G}}_q^T\right]\left\{\left(\sum\nolimits_q \partial_\theta \hat{\mathbf{U}}_q\right)^{-1}\right\}^T \qquad (25)$$

where a 'hat' denotes a plug-in estimate. Note that for linear regression $\partial_\theta \hat{\mathbf{U}}_q = -\hat{\mathbf{G}}_q\hat{\mathbf{E}}_q\mathbf{X}_q$, while for logistic regression $\partial_\theta \hat{\mathbf{U}}_q = -\hat{\mathbf{G}}_q\hat{\mathbf{E}}_q\mathbf{D}_q(\hat{\theta})\mathbf{X}_q$.

4. Simulation Results

In this Section we illustrate the comparative performance of the different linear and logistic regression estimators described so far by simulating their performance under repeated application of probabilistic linkage based on the exchangeable linkage error model defined by (1) – (5).

4.1 Scenario 1: A Small Number of Large Blocks

Data were generated for a population of size $N = 2000$ made up of three strata of size $M_1 = 1500$, $M_2 = 300$ and $M_3 = 200$ respectively. In each simulation, values of the scalar explanatory variable *X* were independently drawn from the uniform distribution over [0,1]. Two situations were simulated. In the first a linear regression model was simulated with

$$y = 1 + 5x + e$$

where the regression errors e were independently drawn from the N(0,1) distribution. True data pairs (y_i, x_i) were then randomly allocated to strata. In the second, population values of *Y* were generated as independently distributed Bernoulli variables with

$$logit\left\{\Pr\left(y_i = 1 \middle| x_i\right)\right\} = 1 - x_i$$

and the distribution of the explanatory variable *X* was allowed to vary between strata. In particular, *X* values in stratum 1 were drawn from the uniform distribution on [5,20], while those in stratum 2 were drawn from the uniform distribution on [-5,5] and those in stratum 3 were drawn from the uniform distribution on [-20,-5]. In both situations, linked data pairs (y_i^*, x_i) were generated according to the exchangeable linkage errors model defined by (1) – (5) with correct linkage probabilities $\lambda_1 = 1$, $\lambda_2 = 0.95$ and $\lambda_3 = 0.75$. It was assumed that all links for stratum 1 were known to be correct, while links for strata 2 and 3 were known to possibly have errors. The (unknown) probabilities of correct linkage in these strata were then estimated by taking random audit samples of $m_q = 20$ linked pairs from each stratum and counting the number of correct links in each sample, with the estimate of λ_q calculated as

$$\hat{\lambda}_q = \min\left\{m_q^{-1}\left(m_q - 0.5\right), \max\left(M_q^{-1}, l_q\right)\right\}$$

where l_q denotes the proportion of correctly linked pairs identified in the audit sample in stratum *q*. Note that the upper limit on $\hat{\lambda}_q$ above is to allow for the fact that when the true value of λ_q is near one, many audit samples will produce 'false negatives', i.e. indicate there are no errors when in fact there are errors in the linked records.

A total of 1000 independent simulations were carried out for each situation above. For each simulation, four estimators were calculated. These were TR, the 'true' data estimator (the OLS estimator (17) in the linear regression case and the MLE under an independent Bernoulli variables assumption in the logistic regression case), the naive estimator ST defined by substituting the linked data values y_i^* for their true values y_i in TR, the Lahiri-Larsen type bias-

corrected estimator defined by setting $\mathbf{G}_q = \mathbf{X}_q^T \mathbf{E}_q^T$ in (9), denoted LL, and the second order efficient estimator defined by setting $\mathbf{G}_q = \mathbf{X}_q^T \mathbf{E}_q^T \Sigma_q^{-1}$ in (9), denoted BL. Note that both LL and BL use the estimated value of λ_q generated by the audit sample.

Table 1 sets out the relative biases and relative root mean squared errors of the estimators of the slope parameter β in each situation (equal to 5 in the linear model simulations and -1 in the logistic model simulations) as well as the actual coverages of the nominal 95% confidence intervals for this parameter generated by their associated variance estimators (these were the sandwich estimators of variance defined by (25) in the case of LL and BL), while Figure 1 shows boxplots of the distribution of relative errors for the different estimators of β.

Examination of Table 1 shows clearly the bias of the naive estimator (ST) that ignores linkage errors. As expected, since this type of error is a particular case of measurement error, this bias leads to an attenuated estimate of β. Both LL and BL correct this bias, with the more efficient being BL. In the linear model case we see that the variances of ST, LL and BL are all similar, so this bias correction substantially reduces MSE. In contrast, in the logistic simulations the bias in the naive estimator ST is somewhat overshadowed by the increased variability of the bias-corrected estimators LL and BL, so that in pure MSE terms ST is superior because of its very low variance. However, the bias of ST means that the coverage of its associated nominal 95 percent confidence intervals is very low. On the other hand, the coverages recorded by both LL and BL are very close to the required nominal level. One reason for the better relative performance of ST in the logistic case relates to the fact that linkage error and measurement error are not quite the same thing in the case of logistic regression. This is because the zero-one nature of the response variable means that values of y_i and y_i^* can be the same even when there is linkage error.

Finally, it is interesting to see that even small audit samples are sufficient for specifying the parameters of the linkage error process, and that variance estimators like (25) that allow for the extra variability induced by estimation of these parameters lead to confidence intervals with good coverage properties. Although not reported here, simulation experiments where the estimated values of λ_q were treated as fixed led to variance estimators that were biased downwards with much poorer coverage for associated confidence intervals.

4.2 Scenario 2: A Large Number of Small Blocks

The assumption that linkage errors are homogeneous within strata could require that these strata be quite small. In this sub-Section we therefore present simulation results that illustrate the performance of the bias-corrected methods described in Section 3 when this is the case. In particular, we focus on estimation of the average values of the study variable *Y* in each of 4 cells defined by a two-way cross-classification of the study population. Given zero-one classifiers x_1 and x_2, individual population values y were generated using the model

$$y = 1 + x_1 + x_2 + x_1 x_2 + e \qquad (26)$$

with values for x_1 and x_2 stored on the *X*-register and values for y stored on the *Y*-register. As usual, we assume that the *Y*-register and the *X*-register are linked probabilistically, subject to linkage errors generated under the model specified by (1) – (5). Two scenarios were investigated. In the first, values of the error term e in (26) were independently drawn from the standard normal distribution, while in the second they were independently drawn from a mean corrected standard lognormal distribution. In both cases, the aim was to use the information in the linked *Y* and *X*-registers to estimate the cross-classification cell averages

$$\mu_{ij} = \left\{ \sum I(x_{1k} = i) I(x_{2k} = j) \right\}^{-1} \sum y_k I(x_{1k} = i) I(x_{2k} = j) \qquad (27)$$

where the summations in (27) are over the entire population and $I(u)$ is the indicator function for whether u holds or not. Using (26), data were generated for a population of size $N = 3000$, with individual records allocated sequentially to the 4 cells of the cross-classification, i.e. records 1, 5, 9, ... were allocated to cell 1, records 2, 6, 10, ... were allocated to cell 2 and so on. A data set containing linked values y^* of y, together with values of x_1 and x_2 was created by simulating independent linkage errors within 600 strata, each of size 5, where these strata were grouped so that the first 1500 records (group 1, strata 1 – 300) all had $\lambda = 1$, the next 800 records (group 2, strata 301 – 460) had $\lambda = 0.95$, the next 500 records (group 3, strata 461 – 560) had $\lambda = 0.85$ and the final 200 (group 4, strata 561 – 600) had $\lambda = 0.7$. It was assumed that this grouping of records into strata with the same linkage error probabilities is known. Where these linkage error probabilities are less than one (groups 2 – 4) the group-specific values of λ were estimated by collapsing strata into groups, and taking a random audit sample of 30 records from each group. The same methodology as described in Section 4.1 was then used to estimate the common value of λ for the strata making up each of these groups. A total of 1000 independent simulations were carried out for each of the two scenarios (Y normal, Y lognormal) above. Note that the cell means were estimated by first fitting the linear regression model (26), with appropriate linkage error bias corrections, with the estimated regression coefficients and their estimated covariances then converted to estimates of cell means with associated estimated standard errors.

Table 2 shows the relative bias, relative RMSE and actual coverage of nominal 95% confidence intervals for each cell mean based on the 'no linkage error' estimator TR, the estimator ST that does not adjust for linkage errors, and the linkage error adjusted estimators LL and BL. We see that estimation methods based on LL and BL successfully correct for biases induced by the linkage errors, and that BL is (marginally) more efficient than LL, with

confidence interval coverages very close to their nominal values. However, it is also interesting to note that in this case the bias in ST is not really an issue for the cell means μ_{01} and μ_{10}, a consequence of mutual cancellation in the biases associated with the estimated regression parameters from (26) that define these two estimated means.

4.3 Scenario 3: A Large Population Simulation

Linked databases created by national statistics agencies tend to hold millions of records. If strata within these databases are large, this poses some challenges for computational efficiency if the full-sized $\mathbf{E}_q$ matrix is included in the calculations. For the linear model presented in Section 3.1, given the assumption of exchangeable linkage errors, there are several methods for computationally efficient estimation. An implementation in SAS was deemed most appropriate for use in a national statistics agency, and some simple re-expressions of (13), (14), (19), (20), (21) and (25) was then used to reduce the dimensionality of the matrices in these expressions. For example, working from (5), and assuming that the dimension of $\mathbf{X}_q$ is $M_q \times p$, the components of (19) can be written as $\mathbf{X}_q^T\mathbf{E}_q^T\mathbf{E}_q\mathbf{X}_q = (\lambda_q - \gamma_q)^2\mathbf{X}_q^T\mathbf{X}_q + 2(\lambda_q - \gamma_q)\gamma_q\mathbf{t}_{Xq}^T\mathbf{t}_{Xq} + \gamma_q^2 M_q\mathbf{t}_{Xq}^T\mathbf{t}_{Xq}$ and $\mathbf{X}_q^T\mathbf{E}_q^T\mathbf{y}_q^* = (\lambda_q - \gamma_q)\mathbf{X}_q^T\mathbf{y}_q^* + \gamma_q\mathbf{t}_{Xq}^T\mathbf{t}_{yq}^*$, where $\mathbf{t}_{Xq}$ is a *p*-dimension row vector of the column sums of $\mathbf{X}_q$ and $\mathbf{t}_{yq}^*$ is the sum of y_q^* over the M_q records in stratum *q*. That is, (19) can be expressed as a function of the 'raw' cross-product matrix, 'raw' column sums and some scalars, all of which are of dimension no larger than $p \times p$ and can be efficiently calculated via SAS procedures. When λ_q is unknown, it is replaced by an estimate. Estimation of (20) requires iteration and the creation of weighted *x* vectors and their cross-products and sums but follows the same logic. Similar manipulations can be used to simplify variance estimation under (13), (14) and (21) and incorporating the uncertainty of an estimated λ_q. These simplifications have been implemented

in a SAS macro (*LinkReg*) that allows a 10-variable linear regression model to be fitted to a million cases in just under a minute of real time using minimal computer memory. A copy of this macro, and a User's Guide, can be obtained from the authors on request.

To assess the performance of the estimators for large samples, data were generated for populations of size $N = 1{,}000{,}000$, made up of 100 strata containing 10,000 observations each, with $\lambda = 1$ for the first forty strata, $\lambda = 0.95$ for the next forty and $\lambda = 0.75$ for the remaining twenty. The values of λ are estimated from random audit samples of 50 taken from each of these three groups. The model used to generate the population data defined the response variable (i.e. the values on the *Y*-register) as a linear function of 10 independent variables (stored on the *X*-register) with correlations between the different independent variables set at 20 per cent. Due to time and computing constraints, only 500 simulations of this model were run. The top half of Table 3 shows the resulting relative bias, relative RMSE and actual coverage of nominal 95% confidence intervals for the regression coefficient of the first independent variable in the regression model. Similar results are obtained for the other coefficients. As can be seen from these results, the bias in OLS resulting from ignoring linkage error (estimator ST) persists in large samples. Both the LL and BL estimators again work well, with LL in this case marginally outperforming BL.

5. Linkage Errors and Fixed Population Inference

The emphasis so far has been on the impact of linkage errors on inference related to parameters of statistical models for populations. Thus, even in Section 4.2 where we considered estimation of the cell means in a cross-classified population, our approach was based on fitting a model to the population using the probabilistically linked data defined by the *Y* and *X*-registers. In this

Section we consider a different target of inference that is often of interest to statistical agencies, where the data values making up the finite population are treated as fixed (rather than generated via a model), and the only source of variability is the outcome of the probabilistic linking process.

Let *U* define a finite population of *N* units, where *N* is known. As usual, we assume that *U* can be divided into *Q* exhaustive and mutually exclusive strata of size M_q, $q = 1,\ldots,Q$, based on a stratifying variable *Z*, as explained in Section 2. Let *Y* and *X* be characteristics associated with each unit, with values related through the linear regression model (15) and (16). If these values are known for each unit in *U*, then $\hat{\beta}$ defined by (17) can be regarded as a corresponding finite population regression coefficient, which we denote by *B*. Within the finite population inference paradigm, *B* is often considered as a target of inference (see Binder and Roberts, 2003). In the rest of this Section we therefore refer to *B* as the target finite population (TFP) parameter.

Clearly, the naïve linked data OLS estimator (18), which we denote by B^*, defines a 'linked finite population' (LFP) parameter that is supposed to correspond to *B*. However, linkage error means that *B* and B^* are not the same. The expected value of this difference under the linkage error process will then drive the 'fixed population' bias of any inference that uses the LFP parameter B^* as its target instead of the TFP parameter *B*. Denoting moments of the linkage error process by a 'star' superscript, $E^*\left(B^* - B\right) = 0$ only under perfect linkage. One might therefore think of using $\hat{\beta}_A$ defined by (19) as the LFP parameter corresponding to *B*, or $\hat{\beta}_C$ defined by (20). However, fixed population inference about *B* that uses either $\hat{\beta}_A$ or $\hat{\beta}_C$ as its target will still be biased, since both $E^*\left(\hat{\beta}_A - B\right)$ and $E^*\left(\hat{\beta}_C - B\right)$ are not equal to zero. That is, although the estimators (19) and (20) are model-unbiased for the regression parameter β of a linear model for

the population data, neither are appropriate as a LFP parameter that can be used as the target of inference. Such a 'fixed population unbiased' parameter corresponding to *B* under (5) is

$$B_K = \left[\sum_q \mathbf{X}_q^T \mathbf{X}_q\right]^{-1} \sum_q \mathbf{X}_q^T \mathbf{E}_q^{-1} \mathbf{y}_q^* . \qquad (28)$$

Note that B_K is both finite population unbiased for *B* under the non-informative linkage model defined by (1) – (5) and also a model-unbiased estimator of the linear model parameter β. From the latter perspective, we see that (28) can also be written as the solution of the estimating equation $\sum_q \mathbf{G}_q \left\{\mathbf{y}_q^* - \mathbf{E}_q \mathbf{X}_q \beta\right\} = 0$, with $\mathbf{G}_q = \mathbf{X}_q^T \mathbf{E}_q^{-1}$, and so the model-based theory developed in Section 3 can be directly applied, including specification for a model-based sandwich-type variance estimator. However, use of (28) as an estimator of β is not recommended from a model-based viewpoint, since it does not have the same model-based efficiency as (19) or (20). Our proposition, rather, is that (28) be used to define the finite population target of inference when working with the linked data, since any statistical inference about B_K is equivalent to the inference about the actual target finite population parameter *B*.

To illustrate the fixed population performance of (28), we return to Table 3. In particular, the lower half of this Table shows results from a fixed population simulation that compares (28), which is denoted by BK in Table 3, with the model-based estimators LL and BL. The simulation setup here is identical to the one described in Section 4.3, with the exception that the population values were generated once and then kept fixed, with the only source of variation due to the 500 random linkages that were performed. Similarly to the top half of Table 3, results are only presented for the regression coefficient of the first independent variable in the linear model, since similar results were obtained for the other coefficients.

The LFP parameter (BK in the table) noticeably outperforms the LL and BL estimators in terms of relative bias but not relative RMSE. Moreover, the sandwich type variance estimator (25) turns out to be extremely conservative for all three estimators, substantially overestimating their repeated linkage error variance, and leading to confidence intervals with 100% coverage.

6. Regression Analysis under Sample to Register Linkage

A key assumption in the development so far has been that the *Y*-register and the *X*-register refer to the same population (and so are the same size) and that all records in them are linked on a one to one basis. In this Section we modify this theory to allow for linkage error in the case where only a sample of records on the *X*-register (rather than the complete *X*-register) is linked to the *Y*-register. These linked sample data are subsequently used to fit regression models of *Y* on *X*, using a weighted estimating function. The weights are usually functions of sample weights, and the weighted estimating function is an estimate of the 'census' estimating function under perfect linkage. We then describe an application of this theory to identification of the main sources of error when using probabilistically linked Australian census data to fit logistic models.

6.1 Methodology

As in Section 2, we assume that linkage errors can be 'stratified' following linkage. We also assume that if all records in stratum *q* had been linked, then the observed linked values $\mathbf{y}_q^*$ of *Y* would be related to the true values $\mathbf{y}_q$ of this variable via the identity (1). The impact of sampling is that we only observe a subset s_q of m_q records from $\mathbf{y}_q^*$, which we denote by $\mathbf{y}_{sq}^*$. We use a subscript of *sq* (*rq*) to denote quantities that depend on the m_q ($M_q - m_q$) linked sample

(unlinked non-sample) records associated with the population units in sample s_q (non-sample r_q) in stratum *q*.

Suppose now that under perfect linkage of the sample data we estimate the true value θ_0 that characterises the vector $\mathbf{f}(\theta)$ of the regression of *Y* on *X* by solving the set of estimating equations $H_s(\theta) = 0$, where

$$H_s(\theta) = \sum_q \mathbf{G}_{sq} \left\{ \mathbf{y}_{sq} - \mathbf{f}_{sq}(\theta) \right\} \tag{29}$$

is the sample-level estimating function. Here $\mathbf{G}_{sq}$ is an appropriately defined weighting matrix that will usually depend on the sample weights $\mathbf{w}_{sq}$, and $\mathbf{f}_{sq}(\theta)$ denotes the sample component of $\mathbf{f}_q(\theta)$. When we ignore linkage errors, we effectively replace (29) by

$$H_s^*(\theta) = \sum_q \mathbf{G}_{sq} \left\{ \mathbf{y}_{sq}^* - \mathbf{f}_{sq}(\theta) \right\} \tag{30}$$

where $\mathbf{y}_{sq}^* = \mathbf{A}_{sq}\mathbf{y}_q$ and

$$\mathbf{A}_q = \begin{bmatrix} \mathbf{A}_{sq} \\ \mathbf{A}_{rq} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{ssq} & \mathbf{A}_{srq} \\ \mathbf{A}_{rsq} & \mathbf{A}_{rrq} \end{bmatrix}$$

denotes the sample/non-sample decomposition of the outcome of the 'complete' linkage process in stratum *q*. As in Section 3, we can show that the solution to (30) is biased for θ_0 under linkage error. Correcting this bias leads to the adjusted estimating function

$$H_s^{adj}(\theta) = \sum_q \mathbf{G}_{sq} \left\{ \mathbf{y}_{sq}^* - \mathbf{E}_{sq}\mathbf{f}_q(\theta) \right\} = \sum_q \mathbf{G}_{sq} \left\{ \mathbf{y}_{sq}^* - \mathbf{E}_{ssq}\mathbf{f}_{sq}(\theta) \right\} - \sum_q \mathbf{G}_{sq}\mathbf{E}_{srq}\mathbf{f}_{rq}(\theta) \tag{31}$$

where

$$\mathbf{E}_q = \begin{bmatrix} \mathbf{E}_{sq} \\ \mathbf{E}_{rq} \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{ssq} & \mathbf{E}_{srq} \\ \mathbf{E}_{rsq} & \mathbf{E}_{rrq} \end{bmatrix}$$

denotes the corresponding sample/non-sample decomposition of the expected value $\mathbf{E}_q$ of $\mathbf{A}_q$ and we have used the fact that $E_X\left(\mathbf{y}^*_{sq}\right) = E_X\left(\mathbf{A}_{sq}\mathbf{y}_q\right) = \mathbf{E}_{sq}\mathbf{f}_q(\theta)$. If we assume the exchangeable linkage error model (1) - (5) and replace $\sum_q \mathbf{G}_{sq}\mathbf{E}_{srq}\mathbf{f}_{rq}(\theta)$ in (31) by a weighted sample estimate, it can be shown that an unbiased estimating equation for θ under this linkage error model is

$$H^{adj}_{ws}(\theta) = \sum_q \mathbf{G}_{sq}\left\{\mathbf{y}^*_{sq} - \tilde{\mathbf{E}}_{sq}\mathbf{f}_{sq}(\theta)\right\} \qquad (32)$$

where $\tilde{\mathbf{E}}_{sq} = \left(M_q - 1\right)^{-1}\left(\lambda_q M_q - 1\right)\mathbf{I}_{sq} + \left(1-\lambda_q\right)\mathbf{1}_{sq}\mathbf{w}^T_{sq}$ and $\mathbf{1}_{sq}$ is the vector of ones of size m_q. Let $\hat{\theta}$ denote the solution to $H^{adj}_{ws}(\theta) = 0$. Assuming that the λ_q are known, and that the regression errors are uncorrelated, we can adapt the development leading to the sandwich variance estimator (11) to define a corresponding variance estimator for the solution to (32):

$$\hat{V}_X(\hat{\theta}) = \left\{\sum_q \mathbf{G}_{sq}\tilde{\mathbf{E}}_{sq}\partial_\theta \mathbf{f}_{sq}(\hat{\theta})\right\}^{-1}\left\{\sum_q \mathbf{G}_{sq}\hat{\mathbf{\Sigma}}_{sq}\mathbf{G}^T_{sq}\right\}\left[\left\{\sum_q \mathbf{G}_{sq}\tilde{\mathbf{E}}_{sq}\partial_\theta \mathbf{f}_{sq}(\hat{\theta})\right\}^{-1}\right]^T \qquad (33)$$

where

$$\hat{\Sigma}_{sq} = \underset{i\in s_q}{diag}\left[\left(M_q - 1\right)^{-1}\left\{(\lambda_q M_q - 1)d_i + M_q(1-\lambda_q)\bar{d}_{wq}\right\} + (1-\lambda_q)\left\{\lambda_q\left(f_i - \bar{f}_{wq}\right)^2 + \bar{f}^{(2)}_{wq} - \bar{f}_{wq}{}^2\right\}\right].$$

Note that a subscript of *w* above indicates a weighted sample mean, so $\bar{d}_{wq}$ is the weighted stratum mean of $d_i = Var_X(y_i)$, and $\bar{f}_{wq}$ and $\bar{f}^{(2)}_{wq}$ are weighted versions of $\bar{f}_q$ and $\bar{f}^{(2)}_q$ respectively. Note also that since f_i and d_i depend on unknown parameters, their fitted values $\hat{f}_i$ and $\hat{d}_i$ are substituted in $\hat{\Sigma}_{sq}$.

6.2 Application to Logistic Modelling of Linked Census Data

The Australian Bureau of Statistics plans to create a Statistical Longitudinal Census Dataset (SLCD) based on linking a random sample of 5% of the person records from the 2006 Census

with the records for the same individuals in the 2011 and subsequent Censuses (Conn and Bishop, 2005). The SLCD will be created using probability linking, but will not use names and addresses as linking variables. In order to evaluate the impact of this linking process on the quality of analyses carried out using the linked data, the ABS conducted a study to assess the main sources of error when analysing data from a SLCD that had been created without using a unique person identifier to define the linked records. This study involved linking the 2006 Census Dress Rehearsal (CDR) to the 2006 Census. The CDR collected information from about 78,000 people and was conducted one year before the 2006 Census. The 2006 Census collected information from more than 19 million people. The CDR and Census person level records were linked using two levels of linking information:

- *Gold Standard*. This used name, address, mesh block and selected Census data items to define links. Mesh block is a geographic area typically containing 50 dwellings. Since all name and address information is destroyed at the end of the Census processing period, this type of linking will not be available in 2011. However it was available at the time of this study and was therefore used to define an error-free linking standard. In what follows we refer to this method of linking as *G* linking.
- *Bronze Standard*. This is a probability-based linkage method that uses mesh block and selected Census data items. It represents the type of linkage method that will be used for linking the sample drawn from the 2006 Census to the 2011 Census. Details of this linking method are given in Chipperfield (2009). This is referred to as *B* linking below.

Since G linking is (hypothetically) error-free, it can be used to estimate λ_q for *B* linking and also to provide a benchmark against which estimates based on the *B* linked data can be compared. There were 70,274 *G* links, or approximately 90 percent of the 78,000 CDR records. In contrast,

there were only 57,790 *B* links, or approximately 74 percent of the CDR records. Further, 54,860 or approximately 95 percent of the *B* links were correct (i.e. were also *G* Links). That is, there were 2,930 incorrect *B* links. One obvious concern with *B* linking was the impact of incorrect links on analysis of the linked data. Another concern was the fact that under *B* linking, 12,484 fewer CDR records were linked compared with *G* linking. If these 12,484 records are unusual in some way it could lead to bias in estimates obtained from the *B* linked data. This second concern is analogous to concerns about non-ignorable non-response in sample surveys, and is substantively based. For example children were under-represented in the *B* linked data because there were relatively fewer useful linking variables for them; also, data collected from people living in indigenous communities had high levels of missingness, which resulted in there being no *B* linked individuals from these communities.

In order to apply the linkage error model (1) - (5) under *B* linking, a separate stratum was defined for each *B* linked CDR record. In addition to this record, this stratum contained further unlinked Census records, chosen so that there was negligible probability that the *B* linked CDR record in the stratum could be linked to any census record not in the stratum. The linkage error probability λ_q associated with each stratum was then estimated by grouping strata and comparing the *B* links in these grouped strata with those defined by *G* linking.

An evaluation of the impact of *B* linking on analysis of the linked data was then carried out. This consisted of fitting logistic models for the probability that (a) a person moves between 2005 and 2006; (b) a person 15 years and older is employed in 2006; and (c) a person 15 years and older is a student in 2006. Zero-one values indicating the non-occurrence/occurrence of these events were taken from the Census, while all explanatory variables were taken from the CDR. Naive and bias-corrected estimates of the logistic regression model coefficients for these

variables were then calculated, by applying (32) and (33) to the *B* linked data, using $\mathbf{G}_q(\theta)$ defined by (23). Corresponding model estimates based on the *G* linked data were also calculated.

Examination of the differences between these *B* linked and *G* linked model estimates allows one to assess the comparative importance of different sources of linkage error when analysing *B* linked data. In particular, we note the following important sources of linkage error:

- *Missing links.* This occurs where a CDR record that could have been linked was not linked at all. That is, where a CDR record was *G* linked but not *B* linked. There were 13,784 CDR records that were *G* linked but not *B* linked. The mechanism generating these linkage failures could be a purely random, in which case the only impact on analysis is a loss of precision due to a smaller *B* linked sample, or could be informative (see Chapter 2 of Chambers and Skinner, 2003), in which case analysis based on the *B* linked sample will be biased compared with that based on the *G* linked sample.
- *Incorrect links.* This occurs when a CDR record is both *B* linked as well as *G* linked, but the two links are not the same. That is, the *B* link is incorrect. There were 1,630 such incorrect links.
- *Impossible links*. This occurs when a correct link with a sample record does not exist but, nevertheless, the sample record is linked. In particular, this error occurred when a *B* link was declared for a record that could not be *G* linked. There were 1,300 such impossible links. These records could correspond to individuals who responded to the CDR but were out of Australia during the 2006 Census. Such links fall outside the theoretical framework developed in this paper, which assumes a correct link exists for all sample records.

To help assess the comparative importance of these three sources of error, let B_k and G_k denote the estimates of parameter *k* in a model with *K* parameters based on the *B* linked and *G*

linked data sets respectively. Recollect that these data sets contain 57,790 and 70,274 records respectively. Similarly, let B_k^c and G_k^c be the same parameter estimates, but this time calculated using the 56,490 (= 57,790 – 1,300) records for which the *B* link is possible. Note that for this group of records, these two links correspond to the same Census record, or, if they do not, then the Census record corresponding to the *G* link is in the same stratum as the Census record corresponding to the *B* link. A scale-free and intuitive measure of the total linkage error under *B* linking is then $W = K^{-1}\sum_k (G_k - B_k)^2 / se(B_k)^2$, which is crudely approximated by

$$\tilde{W} = K^{-1}\sum_k \left[(G_k - G_k^c)^2 + (B_k^c - G_k^c)^2 + (B_k^c - B_k)^2 \right] se(G_k)^{-2} \qquad (34)$$

where $se(G_k)$ is the estimated standard error of parameter *k* in the model, calculated using the *G* linked data.

The approximation (34) is useful because its first, second and third terms correspond respectively to the three sources of linkage error discussed in the preceding paragraph. For example, the second term in (34) is zero if all possible *B* links are correct, and is smaller for a linkage error adjusted estimator compared with an estimator that ignores linkage errors. Table 4 sets out the values of these three components of linkage error for the three logistic models that were investigated in this study. This clearly shows that missing links are generally the biggest source of error; while the contribution from incorrect links is always smaller for the linkage error adjusted estimator compared with the estimator that ignores these errors.

Given its magnitude, testing whether the linkage error component due to missing links is random or not becomes important. Consequently, a simulation test was carried out of the null hypothesis that the CDR records missed under *B* linking constitute a random sample of the *G* linked records. For all three models, this null hypothesis was rejected at the one percent level of significance, i.e. there is strong evidence that there is an informative mechanism underpinning

whether a *B* link is declared or not. Adjusting for this informativeness is an issue that is currently being investigated.

7. Comments and Extensions

As we noted at the start of this paper, there is growing use of some form of probability-based linking in many areas of applied statistics. The advantages of being able to create 'new' databases by linking existing databases are obvious, especially when weighed up against the costs involved in attempting to capture the linked data via a special purpose collection. However, even though methods of record linking are improving, there will always be situations where the linked database contains errors, i.e. incorrectly linked records. Our concern in this paper has been to address the impact of these linkage errors on subsequent statistical analysis of the linked data. In this context, we noted that linkage error can be viewed as a form of measurement error. However, it differs from the standard notion of additive measurement error, where the bias arises because of inflation in variation of the explanatory variables. In the linkage error case this bias arises because the covariance between the response and explanatory variables is attenuated. As a consequence, different methods need to be applied to the linked data in order to correct this bias. This paper uses a simple, but widely applicable, model for linkage errors to develop the appropriate adjustments for a wide class of statistical estimators that can be represented as solutions to estimating equations. In simulations, as well as in a real data application, we then show that these linkage error adjustments work and are practically useful.

Significant challenges remain in applying these estimators to real-world linked data. Although in Section 6 we describe how our adjustment methods can be adapted to sample to population linkage, our basic assumption is that the linkage error process and the sampling process operate

independently of each other. However, it is more likely that this process is sequential, with only sampled records linked. The issue of how to deal with sample records that cannot be linked is a related and highly important problem, as the analysis of Australian census-linked data discussed in Section 6.2 showed. Secondly, our development has assumed that only two databases are linked, with all explanatory variables held on one of them. This is rather unlikely in practice, where multiple databases are often linked, with explanatory variables drawn from a number of them. A simple example is longitudinal modelling using time varying covariates where covariates for different time periods are stored on different databases, which first need to be linked. A third area that needs development is more complex and more realistic modelling of the linkage error process itself. The model (1) - (5) that we focus on in this paper is rather simple and should be capable of being improved using information from the actual statistical matching process. Some extensions to this model were discussed at the send of Section 2, and these need following up. Finally, more efficient model-based methods (e.g. maximum likelihood) for dealing with probabilistically linked data need to developed, as does the extension of the fixed finite population approach (see Section 5) to analysis using linked data. We are currently researching all of these areas.

References

Binder, D.A and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. Chapter 3 in *Analysis of Survey Data* (eds R.L. Chambers and C.J. Skinner). John Wiley and Sons: Chichester.

Chambers, R.L. and Skinner, C.J. (2003). *Analysis of Survey Data*. John Wiley and Sons: Chichester.

Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, Volume 4. http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

Chipperfield, J.O. (2009). Analysis of probability-linked data. *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.098, Australian Bureau of Statistics, Canberra.

Conn, L. and Bishop, G. (2005). Exploring methods for creating a longitudinal census dataset. *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Felligi, I.P. and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1183-1210.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. Springer: New York

Iron, K. S., Manuel, D. G. and Williams, J. (2003). Using a linked data set to determine the factors associated with utilization and costs of family physician services in Ontario: effects of self-reported chronic conditions. *Chronic Diseases in Canada*, **24:4**, Public Health Agency of Canada. See http://www.phac-aspc.gc.ca/publicat/cdic-mcc/24-4/g_e.html

Lahiri, P. and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, **100**, 222-230.

Neter, J., Maynes, E.S. and Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, **60**, 1005-1027.

Scheuren, F. and Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, **19**, 39-58.

Table 1 Relative bias (RB, in percent), relative root mean squared error (RRMSE, in percent) and actual coverage (COVER, in percent) of nominal 95 percent confidence intervals for β in Scenario 1 simulations.

Estimator	RB	RRMSE	COVER
	Linear regression simulation		
TR	-0.12	3.35	95.4
ST	-3.38	8.42	46.3
LL	-0.10	3.99	96.0
BL	-0.21	3.71	95.7
	Logistic regression simulation		
TR	2.37	9.87	96.8
ST	-8.53	11.83	76.0
LL	3.31	15.77	96.8
BL	2.57	12.53	95.6

Table 2 Relative bias (RB, in percent), relative root mean squared error (RRMSE, in percent) and actual coverage (COVER, in percent) of nominal 95 percent confidence intervals for cell means in Scenario 2 simulations. For each mean, column (N) corresponds to normal within cell distributions for *Y* and column (Ln) corresponds to lognormal within cell distributions.

Estimator	μ_{00}		μ_{10}		μ_{01}		μ_{11}	
	N	Ln	N	Ln	N	Ln	N	Ln
				RB				
TR	0.09	0.00	-0.08	0.14	0.06	-0.03	-0.02	0.03
ST	7.51	7.25	0.69	0.95	0.92	0.77	-2.69	-2.58
LL	0.10	-0.01	-0.10	0.17	0.10	-0.04	-0.04	0.03
BL	0.23	0.16	-0.08	0.17	0.12	-0.03	-0.10	-0.03
				RRMSE				
TR	3.76	3.58	1.80	1.75	1.78	1.78	0.92	0.90
ST	8.53	8.18	1.97	2.06	2.04	2.00	2.87	2.77
LL	4.44	4.34	1.92	1.90	1.90	1.93	1.19	1.16
BL	4.34	4.20	1.91	1.89	1.89	1.92	1.13	1.10
				COVER				
TR	94.3	94.9	95.6	96.1	95.3	96.0	94.7	95.3
ST	49.7	55.6	94.7	93.7	92.8	94.3	21.7	24.6
LL	93.7	94.9	95.3	96.3	95.4	95.3	93.6	94.3
BL	93.9	94.4	95.2	95.7	95.1	95.6	93.7	94.6

Table 3 Relative bias (RB, in percent), relative root mean squared error (RRMSE, in percent) and actual coverage of nominal 95 percent confidence intervals (COVER, in percent) for β_1 in Large Population (Scenario 3) simulations. Note that in the Fixed Population Simulations, estimation errors are defined relative to the value of TR for this population.

Estimator	RB	RRMSE	COVER
	Model-Based Simulations		
TR	-0.02	0.64	93.3
ST	-7.04	7.07	0.0
LL	-0.15	0.78	92.9
BL	-0.24	0.79	91.2
	Fixed Population Simulations		
ST	-7.18	7.16	0.0
LL	-0.17	0.31	100
BL	-0.20	0.31	100
BK	-0.01	0.35	100

Table 4 Components of (34) for three logistic models fitted to *B* and *G* linked data. Note that first component is contribution from missing links; second is contribution from incorrect links and third is contribution from impossible links.

Modeled Variable	No adjustment for linkage error	Linkage error adjusted
Mover 2005 - 2006	1.36, 0.17, 0.08	1.36, 0.12, 0.07
Employed in 2006	0.98, 0.19, 0.20	0.98, 0.13, 0.20
Student in 2006	0.41, 0.18, 0.10	0.41, 0.13, 0.08

Figure 1 Distributions of estimation error, expressed as a percentage of the absolute value of the parameter of interest, for slope parameter β in Scenario 1 simulations. Left plot is linear regression, right plot is logistic regression.