Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data

by James O. Chipperfield, Glenys R. Bishop
and Paul Campbell

June 2011

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 13-24
Statistics Canada, Catalogue No. 12-001-X

Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data

James O. Chipperfield, Glenys R. Bishop and Paul Campbell [1]

Abstract

Data linkage is the act of bringing together records that are believed to belong to the same unit (*e.g.*, person or business) from two or more files. It is a very common way to enhance dimensions such as time and breadth or depth of detail. Data linkage is often not an error-free process and can lead to linking a pair of records that do not belong to the same unit. There is an explosion of record linkage applications, yet there has been little work on assuring the quality of analyses using such linked files. Naively treating such a linked file as if it were linked without errors will, in general, lead to biased estimates. This paper develops a maximum likelihood estimator for contingency tables and logistic regression with incorrectly linked records. The estimation technique is simple and is implemented using the well-known EM algorithm. A well known method of linking records in the present context is probabilistic data linking. The paper demonstrates the effectiveness of the proposed estimators in an empirical study which uses probabilistic data linkage.

Key Words: Data linkage; Probabilistic linkage; Maximum likelihood; Contingency tables; Logistic regression.

1. Introduction

Data linking, also referred to as data linkage or record linkage, is the act of bringing together records that are believed to belong to the same unit (*e.g.*, a person or business), from two or more files. Data linkage is an appropriate technique when data sets must be joined to enhance dimensions such as time and breadth or depth of detail. Ideally, the linkage will be perfect, meaning only records belonging to the same unit are linked and all such links are made. However, in many situations this does not happen, especially when linking records using fields that may have incorrect values, missing values or values that are legitimately different for a given unit.

Probabilistic linking is often used when the files contain a set of common variables or fields that constitute partial identifying information, but which do not constitute a unique unit identifier. In probabilistic linking (Fellegi and Sunter 1969) all possible links are given a score based on the probability that the records belong to the same unit. This score is calculated by comparing the values of linking variables that are common to both files. A link is then declared if the link score is higher than some cut-off. An optimisation algorithm may be used to ensure that each record on one file is linked to no more than one record on the other file. Probabilistic methods for linking files are now well established (see Herzog, Scheuren and Winkler 2007, Winkler 2001 and Winkler 2005) and there is a range of computer packages available to implement them.

This is a consequence of the continued importance of linkage in a variety of fields, particularly relating to health and social policy. Recent examples of probabilistic data linkage from the Australian Bureau of Statistics (ABS) include linking records from the 2006 Australian Census of Population and Housing to a number of data sets including Australian death registrations (Australian Bureau of Statistics 2008), the 2006 Census Dress Rehearsal (Solon and Bishop 2009), and the Australian Migrants Settlements Database (Wright, Bishop and Ayre 2009). In the health arena within Australia, probabilistic linkage methods are used by the Western Australian Data Linkage Unit (Holman, Bass, Rouse and Hobbs 1999) and by the New South Wales Centre for Heath Record Linkage. Internationally, probabilistic methods are used by Statistics Canada (Fair 2004), USBC (see Winkler 2001), the U.S. National Center for Health Statistics (National Center for Health Statistics 2009) and by the Switzerland Statistical agency as part of their Longitudinal Study of People Living in Switzerland.

Data linking offers opportunities for new statistical output and analysis. Naively treating a probabilistically-linked file as if it was perfectly linked will, in general, lead to biased estimates. Lahiri and Larsen (2005) and Scheuren and Winkler (1993) proposed methods to calculate unbiased estimates of coefficients for a linear regression model under probabilistic record linkage. More recently, Chambers, Chipperfield, Davis and Kovačević (2009) and Chambers (2008) extended this work to a wide set of models using generalised estimating equations and, in the case of linking two files, allowing one file to be a subset of the other file.

This paper develops a maximum likelihood (ML) approach for analysis of probabilistically-linked records. The estimation technique is simple and is implemented using the well-known EM algorithm. The approach involves replacing the statistics, which would be observed from perfectly linked

1. James O. Chipperfield, Australian Bureau of Statistics. E-mail: james.chipperfield@abs.gov.au; Glenys R. Bishop, The Australian National University; Paul Campbell, Australian Bureau of Statistics.

data, with their expectation conditional on the linked data. Assuming this expectation is correctly specified, this approach overcomes the following two limitations of the previous work.

First, the previous methods assume only one linkage pass is made, whereas, probabilistic linkage usually involves multiple passes. In the latter case, records not linked in the first pass are eligible to be linked in the second pass, and only records not linked in the first two passes are eligible to be linked in the third pass, and so on. Each pass is designed to link records with a particular common set of characteristics. For example, the first pass may be designed to link records belonging to individuals who have not changed address between the reference dates of the two files. The second pass may be designed to accommodate changes of address. An example of such an approach is given in Table 1 in section 5.

Second, the previous methods assume that either the two files contain records from exactly the same units or the set of units on one file is a subset of those on the other file. The approach proposed can be used when one of the files to be linked is not necessarily a subset of the other file. This situation occurs frequently in practice and occurred in all the ABS examples mentioned above. It is also worth mentioning that the files to be linked do not need to be related via a sampling mechanism, such as the smaller file being a random sub-sample of individuals from the larger file. Removing this restriction means that the two files may be administrative data sets.

Consider linking two files denoted by X and Y. File Y contains the variable y on the population of individuals U_y comprising n_y records. File X contains a vector of variables, $\mathbf{x}$, on the population of individuals U_x comprising n_x records. The target of inference is with respect to the population of n_{xy} individuals, denoted by $U_{xy} = U_x \cap U_y$, who are common to File X and File Y. Files X and Y also contain a vector of fields, denoted by $\mathbf{z}$, which are used to link the files using a probabilistic linkage algorithm. Of course, since we are considering probabilistic linkage here, the variable $\mathbf{z}$ does not constitute a unique unit identifier.

Linking Files X and Y allows the joint distribution of $\mathbf{x}$ and y to be analysed. There are two sources of error that may affect analysis of the joint distribution using the linked file. These errors are referred to as *incorrect links* and *unlinked records*.

A link is correct when the pair of linked records belong to the same individual. A link is incorrect when a pair of linked records do not belong to the same individual. Incorrect links can artificially increase or decrease the correlation between $\mathbf{x}$ and y. An example of the latter is random linkage, where records on File X are randomly linked to records on File Y.

The i^{th} record on File X is defined as an *unlinked record*, if $i \in U_{xy}$ *and* record i was not linked to a record on File Y. Or in other words, an unlinked record is a record on File X that could be correctly linked but was not linked at all (throughout this paper we use the convention of defining unlinked records in terms of File X, though the definition could equally be in terms of records on File Y). It may not always be possible to link a particular record on File X with much confidence that the link is correct. This situation may arise if a record is missing fields that are useful in establishing the correct link. More generally, unlinked records may occur when some sub-populations are relatively difficult to link. For example, fields such as marital status, qualification, field of study, and highest level of schooling would generally not be as powerful when linking children as when linking mature adults. In this situation, the data linker must decide whether or not to link such records. We define the set of linked records by U_l of size n^* so that $n^* \leq n_x$ and $n^* \leq n_y$.

The problem of analysis with unlinked records has clear parallels with the problem of unit non-response. Both lead to only a subset of legitimate records being available for analysis. The non-response mechanism in survey sampling is, in reality, a function of an unknown set of variables. Here however, we have the slight advantage in knowing that the probability of a record remaining unlinked can only be a function of $\mathbf{z}$. The problem of non-response is often addressed by weighting or by some conditioning argument. This paper considers both approaches to address the issue of unlinked records.

There is a natural trade-off between the number of unlinked records and incorrect links (and consequently the bias that they introduce). Consider the case where File X is a subsample of File Y so that $U_{xy} = U_x$. Linking all records on File X will result, by definition, in no unlinked records but will result in the number of incorrect links being maximised. If instead we decide to only form links which we are very confident are correct, the number of incorrect links will decrease but the number of unlinked records will increase. In practice, finding the optimal balance between the biases due to unlinked records and incorrect links depends upon the analysis to be undertaken, the linkage methodology, and their interaction. For an in-depth practical discussion of this issue see Bishop (2009).

It is worthwhile mentioning that the problem of making inference in the presence of incorrect record linkage is similar to the problem of making inference in the presence of misclassification of the outcome variable, which is a form of measurement error (see Fuller 1987). In the latter case, identifying assumptions separate the misclassification mechanism from the model mechanism and are required since no error-free measurement is typically available. For example,

Hausman, Abrevaya and Scott-Morton (1998) considers misclassification in the outcome variable of a logistic regression model. Their identifying assumption is that the value of the, possibly misclassified, outcome variable is a particular function of the model's explanatory variables. Our proposed method does not require the strong identifying assumptions of measurement error problems essentially because error-free measurement is available from a clerical sample which identifies correct links. The assumptions we make in this paper are outlined in section 3.

Section 2 summarises the ML approach to contingency table and regression analysis under perfect linkage. Section 3 considers the ML approach in the presence of incorrect links. Section 4 considers the ML approach in the presence of both incorrect links and unlinked records. Section 5 demonstrates the effectiveness of many of the proposed estimators in an empirical study. Section 6 summarises the findings.

2. Perfect linkage

By way of introducing notation, this section discusses the case where the linkage is perfect. The estimating approach in this section is standard since, clearly, no special adjustment for incorrect linkage is required. Section 2.1 discusses estimating cell probabilities in a contingency table and section 2.2 discusses estimating regression coefficients in a logistic regression.

2.1 Contingency tables

For notation, it is convenient when considering contingency table analysis to transform $\mathbf{x}_i$ to a single categorical variable x so that $x = 1, 2, ..., g, ..., G$. Define y to be a categorical variable on file Y, where $y = 1, ..., c, ..., C$.

Consider the following factorisation of the distribution of x and y

$$p(y,\ x) = p_1(y \mid x;\ \mathbf{\Pi})\, p_2(x),$$

where $\mathbf{\Pi} = (\boldsymbol{\pi}_1', ..., \boldsymbol{\pi}_g', ..., \boldsymbol{\pi}_G')'$, $\boldsymbol{\pi}_g = (\pi_{1|g}, ..., \pi_{c|g}, ..., \pi_{C|g})'$, $\pi_{c|g}$ is the probability that $y = c$ given $x = g$. We assume that for every value of x there are C possible values of y which implies that the dimension of $\mathbf{\Pi}$ is CG.

We now consider maximum likelihood estimation of the parameter $\mathbf{\Pi}$, characterising p_1, under perfect linkage. Perfect linkage means that all records on file X are correctly linked to their corresponding record on file Y (*i.e.*, there are no incorrect links and no unlinked records). Under perfect linkage, $n_{xy} = n_x$ and the set of linked records is denoted by $\mathbf{d} = \{(y_i,\ x_i)\colon i = 1, ..., n_{xy}\}$. Under perfect linkage, the score function for $\boldsymbol{\pi}_x = (\pi_{1|x}, ..., \pi_{c|x}, ..., \pi_{C|x})'$ characterised by the multinomial distribution, is

$$\text{Score}(\boldsymbol{\pi}_x;\ \mathbf{d}) =$$

$$(\text{Score}(\pi_{1|x};\ \mathbf{d}), ..., \text{Score}(\pi_{c|x};\ \mathbf{d}), ..., \text{Score}(\pi_{C-1|x};\ \mathbf{d}))' \quad (1)$$

where

$$\text{Score}(\pi_{c|x};\ d) = \Sigma_i (w_{ic|x}\pi_{ic|x}^{-1} - w_{iC|x}\pi_{iC|x}^{-1})$$

$$= n_{c|x}\pi_{c|x}^{-1} - n_{C|x}\pi_{C|x}^{-1},$$

for $c = 1, ..., C-1$, where $n_{c|x} = \Sigma_i w_{ic|x}$, $w_{ic|x} = 1$ if $y_i = c$ and $x_i = x$ and $w_{ic|x} = 0$ otherwise, and the category corresponding to $y = C$ is the arbitrarily chosen reference category. Solving $\text{Score}(\boldsymbol{\pi}_x;\ \mathbf{d}) = \mathbf{0}_{C-1}$ for $\boldsymbol{\pi}_x$, where $\mathbf{0}_{C-1}$ is a $C-1$ column vector of zeros, gives the maximum likelihood (ML) estimator

$$\hat{\pi}_{c|x} = n_{c|x}/n_x, \quad (2)$$

where

$$n_x = \Sigma_c \Sigma_i w_{ic|x}$$

and

$$\hat{\pi}_{C|x} = 1 - \sum_{c=1}^{C-1} \hat{\pi}_{c|x}.$$

2.2 Logistic regression

Consider the logistic regression model

$$E(y_i) = \upsilon_i \quad (3)$$

$$\upsilon_i = 1 / [1 + \exp(\boldsymbol{\beta}'\mathbf{x}_i)]. \quad (4)$$

For (4) the K elements of $\mathbf{x}_i$ are dichotomous variables and y_i is now a dichotomous variable available from File Y. If we define $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_{n_{xy}})'$, $\mathbf{y} = (y_1, ...y_i, ..., y_{n_{xy}})'$ and $\boldsymbol{\upsilon} = (\upsilon_1, ...\upsilon_i, ..., \upsilon_{n_{xy}})'$, the score matrix for $\boldsymbol{\beta}$ based on perfectly linked data, $\mathbf{d}$, is

$$\text{Score}(\boldsymbol{\beta};\ \mathbf{d}) = \mathbf{x}'(\mathbf{y} - \boldsymbol{\upsilon}). \quad (5)$$

Solving $\text{Score}(\boldsymbol{\beta};\ \mathbf{d}) = \mathbf{0}_K$ for $\boldsymbol{\beta}$ gives the ML estimate $\hat{\boldsymbol{\beta}}$, which can be found by applying the well-known Newton-Raphson method.

3. Analysis with incorrect links

This section considers the situation where the linked file contains incorrect links but does not contain unlinked records. This occurs when all the records on File X are linked to a record on File Y (so $n_x \leq n_y$). Define the linked file of records by $\mathbf{d}^* = \{\mathbf{d}_i^* = (y_i^*, \mathbf{x}_i)\colon i = 1, ..., n_x\}$, where y_i^* is the value of y that is *linked* to record i on file X. To clarify, y_i is the true value of y for record i on file X, so that $y_i^* = y_i$ if record i is correctly linked.

The estimator given by (2), together with the assumption that $y_i^* = y_i$ for $i = 1, ..., n_x$, is naive since it treats the probabilistically linked file as if it were perfectly linked. In general the naive estimator will be biased. This section derives ML estimators which account for the fact that the data have been linked probabilistically or linked imperfectly in some way.

It is common practice to select a subsample of the linked file, denoted by s_c, which is then reviewed clerically. The clerical review classifies a link, $\mathbf{d}_i$, as either correct or incorrect. Let $\delta_i = 1$ if record *i* on File X is correctly linked and $\delta_i = 0$ otherwise.

Designing the clerical subsample is an important problem, especially since clerical review is often a costly exercise. Possible uses of a clerical sample include estimating the proportion of correctly linked and unlinked records, to assist in deciding which records should be linked and which should remain unlinked, to ensure correct inference using $\mathbf{d}^*$ (*i.e.*, the purpose of this paper), and to identify improvements to the way in which records are linked (in the ABS applications mentioned above, clerical samples were designed to ensure that each link had at least a specific probability of being correct). For the purpose of making correct inference using $\mathbf{d}^*$ selecting the clerical sample by simple random sampling is a reasonable approach. A more efficient clerical subsample could possibly be devised but there is no obvious way to do so. This is because the parameters that we need to estimate to implement the ML method described in this paper depend upon the specific analysis (*e.g.*, choice of y and $\mathbf{x}$). Designing a clerical sample for all possible analyses would be difficult.

We factorise the joint distribution $p(y_i, \mathbf{x}_i, \delta_i)$ by

$$p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})\, p(\mathbf{x}_i)\, p(\delta_i \mid \mathbf{x}_i), \tag{6}$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}$ in the regression case, $\boldsymbol{\theta} = \boldsymbol{\Pi}$ in the contingency table case. Factorisation (6) means that the links are incorrect at random (IAR) or, in other words, that the distributions $y_i \mid \mathbf{x}_i$ and $\delta_i \mid \mathbf{x}_i$ are independent. Under this assumption it is only necessary to maximise the likelihood associated with the factor $p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})$. Throughout this section we assume (6). It is important to point out that (6), and the development that follows, makes no assumption requiring File X to be a subset of File Y (*e.g.*, when units on File X are a subsample of the units on File Y) or that the linkage process involves a single pass. We also assume that the correctness of linkage, δ_i, is independent from record to record.

As mentioned in the introduction, each linked record is assigned a score based on the probability that the records belong to the same unit. Denote the score by r_i. A referee suggested using r_i to more accurately parameterise the distribution of δ_i. Technically this suggestion would involve replacing $p(\delta_i \mid \mathbf{x}_i)$ with $p(\delta_i \mid \mathbf{x}_i, r_i)$ in (6) and would likely reduce the variability of the ML estimators discussed in section 3. This would be a useful avenue of further research.

3.1 Contingency tables

Define $w_{ic|x}^* = 1$ if $y_i^* = c$ and $x_i = x$, and $w_{ic|x}^* = 0$ otherwise. The expectation of $w_{ic|x}$ given $\mathbf{d}_i^*$ is

$$\begin{aligned} E_{d|d^*}(w_{ic|x} \mid x_i = x, y_i^* = y^*) = & \\ w_{ic|x}^* p_{xy^*} + (1 - p_{xy^*})\pi_{c|x} \quad & \text{if } i \notin s_c \\ = w_{ic|x}^* \quad & \text{if } i \in s_c \text{ and } \delta_i = 1 \\ = \pi_{c|x} \quad & \text{if } i \in s_c \text{ and } \delta_i = 0 \end{aligned}$$

and p_{xy^*} is the probability that the i^{th} link is correct given $x_i = x$ and $y_i^* = y^*$. The ML estimator of $\pi_{c|x}$ using the probabilistically linked data, $\mathbf{d}_i^*$, is then

$$\tilde{\pi}_{c|x} = \tilde{n}_{c|x} \left(\sum\nolimits_c \tilde{n}_{c|x} \right)^{-1} \tag{7}$$

where

$$\tilde{n}_{c|x} = \Sigma_i \tilde{w}_{ic|x}, \tag{8}$$

$$\begin{aligned} \tilde{w}_{ic|x} &= w_{ic|x}^* \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*})\tilde{\pi}_{c|x} && \text{if } i \notin s_c \\ &= w_{ic|x}^* && \text{if } i \in s_c \\ &= \tilde{\pi}_{c|x} && \text{if } i \in s_c \text{ and } \delta_i = 0 \end{aligned} \tag{9}$$

and

$$\hat{p}_{xy^*} = \left(\sum\nolimits_{i \in s_c} w_{ic|x}^* \delta_i \right) \left(\sum\nolimits_{i \in s_c} w_{ic|x}^* \right)^{-1}. \tag{10}$$

The estimation procedure involves iterating between (7), (8) and (9) until convergence. Specifically the algorithm is:

1. Calculate $\hat{p}_{xy^*}$ from (10).
2. Initialise $\tilde{\pi}_{c|x}^{(0)}$ and then calculate $\tilde{w}_{c|x}^{(0)}$ from (9) and then $\tilde{n}_{c|x}^{(0)}$ from (8).
3. Calculate $\tilde{\pi}_{c|x}^{(t)}$ from (7) using $\tilde{n}_{c|x}^{(t-1)}$.
4. Calculate $\tilde{w}_{c|x}^{(t)}$ from (9) using $\tilde{\pi}_{c|x}^{(t)}$ and then calculate $\tilde{n}_{c|x}^{(t)}$ from (8) using $\tilde{w}_{c|x}^{(t)}$.
5. Iterate between 3 and 4 until convergence.

The initialised value $\tilde{\pi}_{c|x}^{(0)}$ could be set to the naive estimate of $\pi_{c|x}$, which was described in section 3 above. However, our experience was that the choice of initial value was not important.

3.2 Logistic regression

Below we describe two ML methods (Methods 1 and 2) for estimating $\boldsymbol{\beta}$ using the probabilistically linked data, $\mathbf{d}^*$.

Both methods give unbiased estimates under the IAR assumption. The difference between the methods is the level of aggregation at which the probabilities of correct linkage are estimated. Method 1 requires these probabilities at a fine level of aggregation, which may mean its estimates are more variable than those of Method 2.

3.2.1 Method 1

The expectation of y conditional on the linked data is

$$E_{\mathrm{d}|\mathrm{d}^*}(y_i \mid \mathbf{x}_i = \mathbf{x},\, y_i^* = y^*) =$$
$$y_i^* p_{\mathbf{x}y^*} + (1 - p_{\mathbf{x}y^*})\upsilon_i \quad \text{if } i \notin s_c$$
$$= y_i^* \quad \text{if } i \in s_c \text{ and } \delta_i = 1$$
$$= \upsilon_i \quad \text{if } i \in s_c \text{ and } \delta_i = 0$$

and p_{xy^*} is the probability that the i^{th} link is correct given $x = x_i$ and $y_i^* = y^*$.

The ML estimator is then obtained by iterating between finding the solution, denoted by $\tilde{\boldsymbol{\beta}}$, for $\boldsymbol{\beta}$ in (5) with y_i replaced by $\tilde{y}_i$, where

$$\begin{aligned} \tilde{y}_i &= y_i^* \hat{p}_{\mathbf{x}y^*} + (1 - \hat{p}_{\mathbf{x}y^*})\tilde{\upsilon}_i \quad \text{if } i \notin s_c \\ &= y_i^* \quad \text{if } i \in s_c \text{ and } \delta_i = 1 \\ &= \tilde{\upsilon}_i \quad \text{if } i \in s_c \text{ and } \delta_i = 0, \end{aligned} \tag{11}$$

$\tilde{\upsilon}_i$ has the same form as υ_i except that $\boldsymbol{\beta}$ is replaced with $\tilde{\boldsymbol{\beta}}$ and $\hat{p}_{\mathbf{x}y^*}$ is the estimated proportion of correct links in the clerical sample for each combination of $\mathbf{x}$ and y^*.

3.2.2 Method 2

Let $\mathbf{x}'\mathbf{y}$ in (5) have k^{th} element

$$r_k = \mathbf{x}_k'\mathbf{y} = \sum_i^n y_i x_{ik} = \sum_i^n r_{ik},$$

where $r_{ik} = y_i x_{ik}$. The expectation of r_{ik} conditional on $\mathbf{d}^*$ is

$$\begin{aligned} &E_{\mathrm{d}|\mathrm{d}^*}(r_{ik} \mid \mathbf{x}_i = \mathbf{x},\, y_i^* = y_i) = \\ &\quad [y_i^* p_{ky^*} + (1 - p_{ky^*})\upsilon_i]x_{ik} \quad \text{if } i \notin s_c \\ &\quad = y_i^* x_{ik} \quad \text{if } i \in s_c \text{ and } \delta_i = 1 \\ &\quad = \upsilon_i x_{ik} \quad \text{if } i \in s_c \text{ and } \delta_i = 0 \end{aligned} \tag{12}$$

and p_{ky^*} is the probability that a link with $x_{ik} = 1$ is correct given $y_i^* = y^*$. The ML estimator is then obtained by iterating between finding the solution, denoted by $\tilde{\tilde{\boldsymbol{\beta}}}$, for $\boldsymbol{\beta}$ in (5) with r_{ik} replaced by $\tilde{r}_{ik}$, where

$$\begin{aligned} \tilde{r}_{ik} &= [y_i^* \hat{p}_{ky^*} + (1 - \hat{p}_{ky^*})\tilde{\tilde{\upsilon}}_i]x_{ik} \quad \text{if } i \notin s_c \\ &= y_i^* x_{ik} \quad \text{if } i \in s_c \text{ and } \delta_i = 1 \\ &= \tilde{\tilde{\upsilon}}_i x_{ik} \quad \text{if } i \in s_c \text{ and } \delta_i = 0, \end{aligned} \tag{13}$$

$\tilde{\tilde{\upsilon}}_i$ has the same form as υ_i except that $\boldsymbol{\beta}$ is replaced with $\tilde{\tilde{\boldsymbol{\beta}}}$ and $\hat{p}_{ky^*}$ is the estimated proportion of correct links in the clerical sample for each combination of $\mathbf{x}$ and y^*. Namely, if $y_i^* = 1$,

$$p_{ky^*} = \left(\sum_{i \in s_c}^n y_i^* x_{ik}\delta_i\right)\left(\sum_{i \in s_c}^n y_i^* x_{ik}\right)^{-1}$$

and if $y^* = 0$,

$$p_{ky^*} = \left(\sum_{i \in s_c}^n (1 - y_i^*) x_{ik}\delta_i\right)\left(\sum_{i \in s_c}^n (1 - y_i^*) x_{ik}\right)^{-1}.$$

This approach requires only $2K$ probabilities to be calculated from the clerical sample and, on this basis, may be preferable to the approach in section 3.2.1 which requires more probabilities to be calculated.

3.3 Estimating the variance using the bootstrap

In this section we describe how to calculate the variance of the ML estimates of section 3. Denote the parameter of interest by $\boldsymbol{\theta}$, introduced earlier, and its ML estimate by $\tilde{\boldsymbol{\theta}}$. The Bootstrap (Rubin and Little 2003) estimate of the variance of $\tilde{\boldsymbol{\theta}}$, denoted by $\hat{v}_{\text{boot}}(\tilde{\boldsymbol{\theta}})$, is obtained by

1. Taking a replicate sample of size n_x from the linked file, $\mathbf{d}^*$, by simple random sampling with replacement. Denote the r^{th} replicate sample by $\mathbf{d}^*(r)$. The r^{th} replicate clerical sample is $s_c(r) = s_c \cap \mathbf{d}^*(r)$.
2. Calculating $\tilde{\boldsymbol{\theta}}(r)$ which has the same form as $\tilde{\boldsymbol{\theta}}$ except that $\mathbf{d}^*(r)$ is used instead of $\mathbf{d}^*$ and $s_c(r)$ is used instead of s_c.
3. Repeating steps 1 and 2 R times, where R is the number of replicates.
4. Calculating

$$\hat{v}_{\text{boot}}(\tilde{\boldsymbol{\theta}}) = \frac{1}{R}\sum\nolimits_{b=1}^{R}(\tilde{\boldsymbol{\theta}}(b) - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}(b) - \tilde{\boldsymbol{\theta}})'.$$

4. Analysis with incorrect links and unlinked records

This section discusses two ways of analysing linked data in the presence of incorrect links and unlinked records. As mentioned in the introduction, the problem of analysis when there are unlinked records has clear parallels with the problem of unit non-response. Unlinked records may result in some characteristics on the linked file being over- or under-represented, thus leading to biased analysis. As discussed in more detail below, we use the fact that the mechanism giving rise to unlinked records can only be a function of $\mathbf{z}$.

This section considers two methods of making inference in the presence of incorrect links and unlinked records, where linked records are indexed by $i = 1, ..., n^*$. (Remember that the i^{th} record on File X is an *unlinked record* if $i \in U_{xy}$ *and* record i was not linked to any record on File Y.) The methods involve independently modelling the processes that determine which records are incorrectly linked and which are unlinked (see section 5 for an illustration). These models require a subsample, denoted by s_{xc}, of all records on File X to be subjected to clerical review. Records in the subsample will be either linked to records on File Y or not linked. Linked records in the subsample must be identified as either correctly or incorrectly linked by the clerical review process. A subsample record which is not linked must be identified as either *unlinked*, or *otherwise*. *Unlinked* means the corresponding record was found on File Y but not linked to it, whereas *otherwise* indicates the corresponding record was not found on File Y and therefore assumed not to exist. The latter identification is potentially much more difficult and time-consuming than the former because it assumes some other error-free process is available for checking whether links, which were not made, are in fact correct. Unlinked records, by their nature, have limited information that can be used to identify the correct link, even during clerical review. Such a process may not exist, in which case adjusting for unlinked records would seem to be impossible. However, such a process may involve a clerical review of names appearing on the two files to be linked. For example, a clerical reviewer may realise that the names *John O. Smith* and *Joh O. Smith* on two different records may in fact be the same name (with an "n" missing in the latter case, perhaps due to errors in scanning), whereas the automated linking process may treat the two names as completely different. The clerical reviewer may then decide that the above two records correspond to the same individual and so therefore should be linked. (Bishop (2009) and Wright (2009) discuss the benefits of clerical review).

The first method involves conditioning analysis on a variable $\zeta_i = \zeta_i(\mathbf{z}_i)$. The variable ζ is defined so that inference, in the presence of unlinked records, is unbiased conditional on ζ. The term ζ is introduced since, in many cases, it would be impractical or unnecessary to condition on all the information in $\mathbf{z}$. It is possible to give ζ_i a non-missing value even when $\mathbf{z}_i$ contains missing values. The exact form of the function $\zeta(\mathbf{z})$ would need to be justified after analysis of the subsample, s_{xc}. For example, if persons under 20 years of age are under-represented in the linked file, ζ would indicate whether a person is under 20 years of age. One approach to analysis is to include ζ as a covariate in the regression model. The method in section 3 would then apply directly. However, analysts may like to integrate over ζ so that it does not appear in the logistic model or contingency table. Section 4.2 discusses how to do this for contingency tables. Section 4.3 discusses a pseudo-likelihood approach which assigns weights to the linked records that attempt to account for any under- or over-representation of certain subpopulations in the linked data. Again, the choice of weight would need to be justified after analysis of the subsample, s_{xc}, which identifies unlinked records. This is discussed further in the context of the empirical study.

4.1 Can we ignore unlinked records?

Define the variable $\gamma_i = 1$ if record i on File X is unlinked and $\gamma_i = 0$ otherwise. Also let ζ_i be a variable so that $\zeta_i = 1, 2, ..., h, ...H$, where H is the number of categories for ζ. We can ignore the fact that there are unlinked records if we are prepared to assume that, conditional on $\mathbf{x}_i$, the distributions of y_i, γ_i and δ_i are independent. Technically this assumption leads to the factorisation,

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto$$
$$p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})\, p(\delta_i \mid \mathbf{x}_i)\, p(\gamma_i \mid \mathbf{x}_i)\, p(\zeta_i)$$

where again $\boldsymbol{\theta} = \boldsymbol{\beta}$ or $\boldsymbol{\Pi}$. It is worthwhile checking whether this assumption is valid from the clerical subsample. If the assumption is reasonable, then there is no need to apply the methods in section 4.2 and 4.3 and the methods in section 3 will suffice.

We may not be prepared to make the assumption mentioned above. We may however be prepared to assume, conditional on $\mathbf{x}$ and ζ, the distributions of y_i, γ_i and δ_i are independent. In this case, we say unlinked records are not ignorable. Technically this assumption leads to the factorisation,

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto$$
$$p(y_i \mid \mathbf{x}_i, \zeta_i; \boldsymbol{\Lambda})\, p(\delta_i \mid \mathbf{x}_i; \boldsymbol{\tau})\, p(\gamma_i \mid \mathbf{x}_i, \zeta_i)\, p(\zeta_i)$$

where $\boldsymbol{\Lambda}$ is the parameter for the distribution of $y_i \mid \mathbf{x}_i, \zeta_i$. If we are interested in $p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ but not $p(y_i \mid \mathbf{x}_i, \zeta_i; \boldsymbol{\Lambda})$, one approach is to integrate out (*i.e.*, average over) ζ_i from the latter.

4.2 Conditional Maximum Likelihood (CML) for contingency tables

First. parameterise the joint distribution of y_i, x_i and ζ_i by the multinomial distribution with parameter, $\boldsymbol{\Lambda}$. Define $\boldsymbol{\Lambda} = (\boldsymbol{\Pi}_1', ... \boldsymbol{\Pi}_h', ..., \boldsymbol{\Pi}_H')'$, where $\boldsymbol{\Pi}_h = (\boldsymbol{\pi}_{1h}', ..., \boldsymbol{\pi}_{gh}', ..., \boldsymbol{\pi}_{Gh}')'$, $\pi_{gh} = (\pi_{1|gh}, ..., \pi_{c|gh}, ..., \pi_{C|gh})'$ and $\pi_{c|gh}$ is the probability that $y_i = c$, $x_i = g$ and $\zeta_i = h$. The ML estimator of $\boldsymbol{\Pi} = (\pi_{c|x})$ from section 2.1 when linkage errors are not ignorable is $\tilde{\tilde{\boldsymbol{\Pi}}} = (\tilde{\tilde{\pi}}_{c|x})$, where

$$\tilde{\tilde{\pi}}_{c|x} = \sum_{h=1}^{H} \tilde{\pi}_{c|xh} \hat{\pi}_{h|x}, \qquad (14)$$

where

$$\tilde{\pi}_{c|xh} = \tilde{\mathrm{n}}_{c|xh}\left(\sum_c \tilde{\mathrm{n}}_{c|xh}\right)^{-1}, \quad (15)$$

$\tilde{\mathrm{n}}_{c|xh} = \Sigma_{i\in U_l}\tilde{w}_{ic|xh}$, $\Sigma_{i\in U_l}$ is the sum over the n^* linked records and $\hat{\pi}_{h|x}$ for $h = 1, ..., H$ is the standard estimate of the marginal distribution of ζ given x on File X. Further, if $i \notin s_c$

$$\tilde{w}_{ic|xh} = w^*_{ic|xh}\hat{p}_{xy^*h} + (1 - \hat{p}_{xy^*h})\,\tilde{\tilde{\pi}}_{c|xh}, \quad (16)$$

$\hat{p}_{xy^*h}$ is the probability that the i^{th} link is correct given $x_i = x$, $\zeta_{\mathrm{i}} = h$ and $y^*_i = y^*$, $w^*_{ic|xh} = 1$ if $x_i = x$, $\zeta_{\mathrm{i}} = h$ and $y^*_i = y^*$, and $w^*_{ic|xh} = 0$ otherwise. If $i \in s_c$, then $\tilde{w}_{ic|xh} = w^*_{ic|xh} = w_{ic|xh}$ if the link is determined to be correct and $\tilde{w}_{ic|xh} = \tilde{\tilde{\pi}}_{c|xh}$ if it is determined to be incorrect.

The ML estimator $\tilde{\tilde{\pi}}_{c|x}$ is obtained by iterating between (14), (15) and (16) until convergence.

4.3 Pseudo-Maximum Likelihood (PML)

This section discusses an alternative to the CML, discussed in section 4.2, which is referred to as Pseudo-Maximum Likelihood (see Chambers and Skinner 2003). It is essentially a weighting approach, which may be easier to implement than CML, and relies on the factorisation given in section 4.2. It involves solving weighted versions of the score functions, $\text{Score}(\boldsymbol{\pi}_x; \mathbf{d}) = \mathbf{0}_{C-1}$ and $\text{Score}(\boldsymbol{\beta}; \mathbf{d}) = \mathbf{0}_{\mathrm{K}}$ for $\boldsymbol{\pi}_x$ and $\boldsymbol{\beta}$ respectively, where a record's weight equals the inverse of the probability that the record will remain unlinked. We denote the probability that record i will not remain unlinked by $t_i = E(\gamma_i)$ so that the unit weights are given by $q_i = t_i^{-1}$, where here $i = 1, ..., n^*$. Consequently the PML estimator for $\pi_{c|x}$ is

$$\tilde{\pi}^{\text{PML}}_{c|x} = \breve{n}_{c|x}\left(\sum_c \breve{n}_{c|x}\right)^{-1}, \quad (17)$$

where $\breve{n}_{c|xy} = \Sigma_{i\in U_l} q_i \tilde{w}_{ic|x}$. The estimate of $\tilde{\pi}^{\text{PML}}_{c|x}$ is obtained by iterating between updating $\tilde{w}_{ic|x}$, given by (7), and (17) until convergence. The PML estimator for $\boldsymbol{\beta}$ is the same as the ML estimator but where the estimating equation (5) now has unit weights of q_i. One possible approach to estimating the accuracy of the PML estimates under perfect linkage is to use the Bootstrap method as described earlier, but where now the weight q_i is introduced.

To illustrate when unlinked records are not ignorable, consider linking a data base with personal employment status to another data base with education level. Also assume that age and sex variables, which are correlated with employment and education, are available on one of the data bases. After conducting a clerical review, we may find that records for young males are 50% more likely to remain unlinked than records for females. This could be because males are less likely to provide their personal information, which is useful in linkage. Clearly, records for males on the linked file need to be given a weight double that for females in order for joint analysis of employment status and educational level to be unbiased.

5. Empirical study

A quality study conducted by the Australian Bureau of Statistics involved linking the 2006 Census of Population and Housing to its Dress Rehearsal. The Census Dress Rehearsal collected information from 78,349 persons and was conducted one year before the Census. The 2006 Census collected information from more than 19 million people.

Within a short window, during which the 2006 Census data were being processed, name and address were available for both the Census and the Census Dress Rehearsal. During this time, the two files of person level records were linked using two different standards of information:

- *Gold Standard* (GS) used name, address, mesh block and selected Census data items. Mesh block is a geographic area typically containing 50 dwellings. All names and addresses were destroyed at the end of the Census processing period.
- *Bronze Standard* (BS) used mesh block and selected Census data items (*i.e.*, did not use name and address). This is a method proposed to be used for future linking work by the ABS.

Full details of the quality study and the linkage methodology are given in Solon and Bishop (2009). The role of GS in the quality study is critical. It provides a benchmark against which the reliability of BS can be compared. The usefulness of the GS as a benchmark is due to the fact that name and address are powerful variables for the purpose of identifying common individuals on the Census and CDR and that it was subjected to thorough clerical review. As a result, GS is assumed to correspond to perfect linkage. Accordingly, differences between estimates based on GS and BS are interpreted as error. In other words, interest focuses on the reliability of BS *relative* to GS.

5.1 Linking methodology

5.1.1 Blocking and linking variables and the 1 – 1 assignment algorithm

This subsection provides an overview of the CDR-to-Census linkage methodology for BS. The linking method consisted of a sequence of passes, where each pass is

defined by a set of blocking and linking variables and a 1 - 1 assignment algorithm. In the case of multiple passes, only records not linked in the first pass are eligible to be linked in the second pass, and only records not linked in the second pass are eligible to be linked in the third pass, and so on.

Table 1 gives the blocking variables, denoted by "B" for the BS. For example, during Pass 1, a Census record and a CDR record are only considered as a possible link if they have the same value for mesh block.

Linking variables are used to measure the degree of agreement between a record pair. A high level of agreement suggests that the likelihood of the record pair constituting a correct link is high. Table 1 gives the linking variables, denoted by "L", for BS. For example, during Pass 1 of BS, a range of variables such as day, month and year of birth, country of birth and highest level of qualifications are used as linking variables.

Table 1
An example of blocking (B) and linking (L) variables used when linking 2006 Census data with the Census Dress Rehearsal. Different blocking variables were used on each of the two passes

Variable	Pass 1	Pass 2
Day of birth	L	B
Month of birth	L	B
Year of birth	L	B
Sex	L	B
Indigenous status	L	L
Country of birth	L	L
Language spoken	L	L
Year of arrival	L	L
Marital status	L	L
Religious affiliation	L	L
Field of study of highest qualification	L	L
Level of highest qualification	L	L
Highest level of schooling	L	L
Mesh block	B	L

An output from each pass is a score for all record pairs. The score is a measure of the level of agreement between the pair of records. We defer the formal definition of score (for details see (3.6), Conn and Bishop 2006) but illustrate how it can be interpreted below. Consider BS in Pass 2 where record pairs have the same full date of birth and sex; a record pair would be assigned a score of 23.5 if there is agreement on mesh block (+17) and year of arrival (+8) and disagreement on religion (–1.5) (in this example agreement status for other linking variables would contribute to the score but for illustration purposes we ignore them). The contribution to the score for agreement on mesh block (+17) is greater than that for agreement on year of arrival (+8) because the former is less likely to occur by chance alone.

To formalise the aim of the linkage algorithm, denote the score for record i on the CDR and record j on the Census during pass p of BS by r_{pij}. The set of all record pair scores r_{pij} and the cut-off f_p were used by the linking package *Febrl* (see Christen and Churches 2005) to determine the optimal set of links in pass p. The term f_p is the minimum value for the score in order for a record pair to be assigned as a link during pass p. The *Febrl* algorithm seeks to maximise $\sum_i r_{pij}$, subject to $r_{pij} > f_p$. Clearly, the number of links depends upon f_p.

In what follows, we evaluate BS with two different sets of cut-offs, where a set of cut-offs is defined by the pass 1 and 2 cut-offs. The first is referred to as the Very Low (VL) cut-off and is considered to be optimal cut-off since, for a range of cut-offs, its naive estimates were "closest" to the corresponding GS estimates (see Bishop 2009). The second cut-off is referred to as Ultra-Low (UL) and effectively seeks to maximise the number of linked CDR records. Below we refer to the two BS linked files by their cut-offs, VL and UL.

5.1.2 Linking results

GS linked 70,274 of the 78,349 CDR records. Under the assumption that GS corresponds to perfect linkage, there were 8,075 individuals with CDR records but no Census records. In reality the GS is not perfect. For a discussion on this see Bishop 2009.

VL linked 57,790 CDR records. Of the 70,274 CDR records that were linked by GS, 13,784 remained unlinked by VL, 700 were linked incorrectly by VL and 55,790 were linked correctly by VL. Also, 1,300 CDR records were linked by VL but were not linked by GS- these are also incorrect links. So in total there were 2,000 (= 700 + 1,300) incorrect links.

UL linked 74,350 CDR records. Of the 70,274 CDR records that were linked by GS, 2,811 remained unlinked by UL, 9,793 were linked incorrectly by UL and 57,670 were linked correctly by UL. Also, 6,887 CDR records were linked by UL but were not linked by GS.

In summary, 97% of the VL links are correct and 20% (= 13,784/70,274) of the GS' CDR records remain unlinked. The corresponding figures for UL are 78% and 4% (= 2,811/70,274).

5.1.3 Modelling the probability of a link being correct

All UL and VL links were known to be correct or incorrect (*e.g.*, if a UL link is also made by GS then the UL link is correct. Otherwise the UL link is incorrect). As a result, p_{xy^*} in section 3.1 was known from GS. However, to simulate reality, p_{xy^*} was estimated from a clerical sample of size 1,000 that was selected from the linked files by simple random sampling.

5.1.4 Modelling the probability of a record remaining unlinked

Each CDR record linked by the GS was assigned a variable which indicated whether the record was unlinked by BS. Namely, if the record remained unlinked by BS then the indicator variable was assigned a '1' otherwise a '0'. A logistic model was fitted using GS, where the response variable was the above indicator variable and the explanatory variables were obtained from the CDR. The more than 20 explanatory variables that are in the model were selected by standard forward-backward model selection. The explanatory variables included educational level, language, born overseas, Indigenous status, and indicators of missing key variables such as *meshblock*. The resulting prediction resulted in t_i and was used below to implement the Pseudo-ML method for both contingency tables and logistic regression.

5.2 Results of tabular analysis

Table 2 gives the results of cross-tabulating employment status of indigenous people as reported on the CDR and Census. Table 2a shows that the GS estimate of the proportion of indigenous people employed in the Census, given they were employed in CDR, is 78.3%. The corresponding naive estimate for VL, which assumes the data are perfectly linked, is 86.7%. Even after replacing each of the 700 incorrect VL links by their corresponding correct link and discarding the 1,300 linked records for which no correct link exists, the naive estimate is largely unchanged at 86.0% (referred to as *Gold Links* in Table 2a). This shows that the difference between the VL and GS estimates is not so much due to incorrect links but is mainly due to unlinked records. This explains in part why the ML estimate (86.4%) for VL (see section 3.1), which only corrects for incorrect links, did not lead to much improvement. Conditional ML (CML) (see section 4) was considered in an attempt to reduce the error due to unlinked records that may have led to a misrepresentation, with respect to age and sex characteristics, in the linked file. The CML employment estimate was 86.6%. Unfortunately, CML did not make much of an improvement, indicating that the underlying mechanism generating unlinked records did not depend upon age and sex. PML estimates (see section 4) also did not make much of an improvement, indicating that the logistic model described in section 5.1.4 did not explain the mechanism generating unlinked records. Interestingly, the ML estimate using UL was 81.8%- by far the closest estimate to the GS estimate of 78.3%. The UL's main source of error is due to incorrect links, the type of linkage error which the ML estimator addresses. This indicates that correcting for errors due to incorrect links was much more successful than correcting for errors due to unlinked records.

Standard errors of the GS, naive and ML estimates are shown in parentheses in Table 2a. For VL and UL, ML standard errors are respectively about 25% and 75% larger than the corresponding naive standard errors. Also, the ML standard errors for UL are slightly smaller than for VL indicating that the extra links made by UL were worthwhile. Clearly, naive inference with UL over-states the level of confidence in estimates. For VL, naive and ML standard errors and estimates are very close.

Irrespective of the cut-off, the ML estimates in Table 2 a, b and c are always closer to the GS estimates than the corresponding naive estimate. For example in Table 2b the ML estimates for VL is 36.9%, noticeably closer to the GS estimate of 37.9% than the naive estimate of 33.3%. Based on the estimates in Table 2 it could be argued that the choice of whether to use VL or UL is not so important, as along as the ML estimator is used.

Table 2
Percentages of Indigenous persons in various employment categories in 2006 given their employment category in 2005. For each linked data set, Very Low and Ultra Low, the estimation methods can be compared with the Gold

	Estimates for different methods and linked data set							
a: Indigenous persons employed in 2005								
Status in 2006	**Gold**			**Very Low Cut-off**			**Ultra Low Cut-off**	
		Naive	**Gold links**	**ML**	**PML**	**CML**	**Naive**	**ML**
Employed	78.3	86.7	86.0	86.4	86.6	86.1	71.9	81.8
	(1.7)	(2.4)		(3.0)			(1.7)	(2.9)
Unemployed	3.7	4.2	4.3	4.1	4.1	4.2	6.3	3.3
	(0.84)	(1.2)		(2.5)			(0.82)	(2.1)
Not in the labour force	17.8	9.0	9.6	9.3	9.1	9.6	21.6	14.7
	(1.6)	(2.4)		(3.1)			(1.6)	(2.8)
b: Indigenous persons unemployed in 2005								
Status in 2006	**Gold**		**Very Low**				**Ultra Low**	
			Naive	**ML**			**Naive**	**ML**
Employed	27.5		27.7	27.2			35.2	23.8
Unemployed	34.4		38.9	36.4			32.3	38.0
Not in the labour force	37.9		33.3	36.3			32.3	38.0
c: Indigenous persons not-in-the-labour force in 2005								
Employed	13.7		10.8	10.7			24.3	10.5
Unemployed	5.8		7.6	7.4			6.3	5.8
Not in the labour force	80.4		81.5	81.8			69.2	83.5

Table 3 is the same as Table 2 except that it describes analyses of linked records from all persons 15 and over rather than only Indigenous persons. Again the ML always makes an improvement for the UL, though this is not the case for VL. Table 4 gives the student status in 2006 for persons who were students in 2005. Again the ML generally makes the estimates closer to the corresponding Gold estimate, especially for UL.

Table 3
Percentages of all persons aged over 15 in various employment categories in 2006 given their employment category in 2005. For each linked data set, Very Low and Ultra Low, the estimation methods can be compared with the Gold

		Estimates for different methods and linked data set			
Status in 2006	**Gold**	**Very Low**		**Ultra Low**	
		Naive	**ML**	**Naive**	**ML**
a: Persons employed in 2005					
Employed	91.8	92.2	92.6	89.7	92.4
Unemployed	1.8	1.7	1.6	1.9	1.6
Not in the labour force	6.2	6.1	5.6	8.3	5.8
b: Persons unemployed in 2005					
Employed	44.5	44.3	44.0	49.4	43.8
Unemployed	26.8	26.6	27.5	22.8	27.6
Not in the labour force	28.6	28.7	28.4	27.6	28.5
c: Persons not-in-the-labour force in 2005					
Employed	12.1	12.3	11.1	16.8	11.0
Unemployed	3.1	3.1	3.0	3.0	3.0
Not in the labour force	84.7	84.5	85.7	80.1	85.9

Table 4
Student outcomes in 2006 for high school students in 2005

Student Status in 2006	**Gold**	**Very Low**		**Ultra Low**	
		Naive	**ML**	**Naive**	**ML**
High School Student	79.3	79.3	79.6	77.4	79.6
Completed High School	14.0	14.3	13.7	14.7	14.1
Did not Complete High School	6.6	6.3	6.6	7.8	6.2

5.3 Simulation

The following simulation study illustrates the problems with naive analysis and the benefit of using the method outlined in this paper. Files X and Y in the simulation, each containing 2,000 records, are independently generated 400 times, where each generated file is denoted by X(r) and Y(r), and $r = 1, ..., 400$. Specifically, on X(r) x_i is randomly generated from the Bernoulli distribution with parameter 0.5. On Y(r), y_i is randomly generated from the Bernoulli distribution with parameter υ_i, where $\upsilon_i = 1 / [1 + \exp(\beta_0 + \beta_1 x_i)]$, $\boldsymbol{\beta} = (\beta_0, \beta_1)', \beta_0 = -0.5, \beta_1 = 1.5$. The r^{th} set of imperfectly linked data, $\mathbf{d}^*(r)$, is generated by correctly linking each record on File Y(r) to one record on File X(r) with probability $p = 0.8, 0.90, 0.95$ and 1. For each r^{th} set of linked data a clerical sample of 300 links is selected. Each link in the clerical sample is assigned as being correct or incorrect. We summarise the performance of the ML estimator from section 3.2.2 and the naive method, which assumes there is no linkage error, by their 95% coverage rates and their Mean Squared Error (MSE). The coverage rates are based on the standard errors calculated from the Bootstrap described in section 3.3 with $R = 40$ replicates. The MSE of $\tilde{\boldsymbol{\beta}}$ is calculated by

$$\text{MSE}(\tilde{\boldsymbol{\beta}}) = \frac{1}{400}\sum_{r=1}^{400}(\tilde{\boldsymbol{\beta}}_r - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}_r - \boldsymbol{\beta})'$$

where $\tilde{\boldsymbol{\beta}}_r$ is the ML estimate of $\boldsymbol{\beta}$ from $\mathbf{d}^*(r)$.

Table 5 shows that the naive approach has poor coverage rates, due to its significant bias in the presence of linkage error, and consequently a relatively high MSE. The coverage rates for ML-Method 1 are very close to their nominal levels. The results show that, as the percentage of correct links reduces from 100% to 80%, the MSE of ML increases by a factor of about 3 for β_0 and β_1. (The coverage rates and MSE of ML Method 1and 2 were very similar so only the former are reported).

Table 5
Mean squared error and coverage rates for linked simulated data, where correct linkage occurs with probability, p

		Mean Squared Error				**95% Coverage Rates**		
		0.8	0.9	0.95	1	0.8	0.9	0.95
Naive	β_0	0.024	0.010	0.0056	0.0043*	0.35	0.80	0.93
	β_1	0.11	0.038	0.016	0.011*	0.05	0.62	0.88
ML-Method 1	β_0	0.013	0.0078	0.0055	0.0043*	93.0	94.25	93.5
	β_1	0.031	0.018	0.013	0.011*	96.0	94.5	96.25

* when $p = 1$ the naive and ML estimators are the same by definition.

6. Discussion

Data linkage is an appropriate technique when data sets must be joined to enhance dimensions such as time and breadth or depth of detail. Data linkage is increasingly being used by statistical organisations around the world. It is well-known that errors can arise when linking files, for example when applying probabilistic linking methods. However, there has been little work reported in the literature about how to make valid inferences in the presence of such errors.

This paper provides methodological and practical advice to support analysts in this area.

In general, naively treating a linked file as if it were perfectly linked will lead to biased estimates. The analyst should only use the naive approach when both the number of unlinked records, defined as records that could be correctly linked but were not linked at all, and the number of incorrect links are negligible. This paper has presented a maximum likelihood approach to making valid inferences in the presence of both sources of error. The approach uses the well-known EM algorithm and is easy to apply in practice. The method can be applied when one of the files is not necessarily a subset of the other and when the linkage involves multiple passes. These situations often arise in practice, including many recent examples in the Australian Bureau of Statistics. The empirical study shows that the ML approach makes significant and meaningful improvements to the estimates from the linked data.

In the special case where File X is obtained by taking a random sample from File Y, the estimation procedure described is not 'full' maximum likelihood. This is because it does not use the fact that population totals for File Y are known. While inference using the method described here are still valid in this case, it could perhaps be made more efficient (see Scott and Wild 1997).

Acknowledgements

The authors would like to thank Raymond Chambers and two reviewers from Survey Methodology for their contributions to this paper.

References

Australian Bureau of Statistics (2008). Census Data Enhancement - Indigenous Mortality Quality Study, 2006-07. Information Paper catalogue no. 4723.0.

Bishop, G. (2009). Assessing the Likely Quality of the Statistical Longitudinal Census Dataset. Methodology Research Papers, catalogue no. 1351.0.55.026, Australian Bureau of Statistics, Canberra.

Chambers, R., Chipperfield, J.O., Davis, W. and Kovačević, M. (2009). Regression Inference Based on Estimating Equations and Probability-Linked Data. Submitted for publication.

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

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

Christen, P., and Churches, T. (2005). Febrl – Freely extensible biomedical record linkage. Release 0.3.1, viewed 17 November 2008, http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/contents.html.

Conn, L., and Bishop, G. (2006). Exploring Methods for Creating a Longitudinal Census Dataset. Methodology Advisory Committee Papers, catalogue no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Fair, M. (2004). Generalized record linkage system-Statistics Canada's record linkage software. *Austrian Journal of Statistics*, 33(1 and 2), 37-53.

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

Fuller, W.A. (1987). *Measurement Error Models*. New York: John Wiley & Sons, Inc.

Hausman, J.A., Abrevaya, J. and Scott-Morton, F.M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87, 239-269.

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

Holman, C.D.J., Bass, A.J., Rouse, I.L. and Hobbs, M.S.T. (1999). Population-based linkage of health records in Western Australia: Development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, 23(5), 453-459.

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

National Center for Health Statistics (2009). Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Social Security Administration. Methodology Report, http://www.cdc.gov/nchs/data/datalinkage/ssa_methods_report_2009.pdf.

Rubin, D.B., and Little, R.J.A. (2003). *Statistical analysis of missing data*, 2nd Edition. New York: John Wiley & Sons, Inc.

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

Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57-71.

Solon, R., and Bishop, G. (2009). A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset. Methodology Research Papers, catalogue no. 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Winkler, W.E. (2001). Record Linkage Software and Methods for Merging Administrative Lists. Statistical Research Report Series, No. RR2001/03, Bureau of the Census.

Winkler, W.E. (2005). Approximate String Comparator Search Strategies for Very Large Administrative Lists. Statistical Research Report Series, no. RRS2005/02, Bureau of the Census.

Wright, J., Bishop, G. and Ayre, T. (2009). Assessing the Quality of Linking Migrant Settlement Records to Census Data. Methodology Research Papers, catalogue no. 1351.0.55.027, Australian Bureau of Statistics, Canberra.