1352.0.55.098

Australian Bureau of Statistics

Research Paper

Generalised Linear Models with Probabilistically Linked Data

New Issue

Research Paper

Generalised Linear Models with Probabilistically Linked Data

James Chipperfield

Analytical Services Branch

Methodology Advisory Committee

7 November 2008, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 26 FEB 2009

ABS Catalogue no. 1352.0.55.098

© Commonwealth of Australia 2009

This work is copyright. Apart from any use as permitted under the *Copyright Act 1968*, no part may be reproduced by any process without prior written permission from the Commonwealth. Requests and inquiries concerning reproduction and rights in this publication should be addressed to The Manager, Intermediary Management, Australian Bureau of Statistics, Locked Bag 10, Belconnen ACT 2616, by telephone (02) 6252 6998, fax (02) 6252 7102, or email <intermediary.management@abs.gov.au>.

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr James Chipperfield, Analytical Services Branch on Canberra (02) 6252 7301 or email <analytical.services@abs.gov.au>.

GENERALISED LINEAR MODELS WITH PROBABILISTICALLY LINKED DATA: APPLICATION TO THE SIMULATED STATISTICAL LONGITUDINAL CENSUS DATASET

James Chipperfield
Analytical Services

QUESTIONS FOR THE COMMITTEE

1. Were the linkage error models presented in this paper reasonable?

2. Should the methodology that is designed to adjust estimates using probability-linked data be implemented in the Statistical Longitudinal Census Dataset? Does the Committee have any views on this?

3. Is there convincing evidence that sampling bias is the most concerning source of error in estimates calculated from the Bronze-linked data set, formed using a very low cut-off?

CONTENTS

The role of the Methodology Advisory Committee (MAC) is to review and direct research into the collection, estimation, dissemination and analytical methodologies associated with ABS statistics. Papers presented to the MAC are often in the early stages of development, and therefore do not represent the considered views of the Australian Bureau of Statistics or the members of the Committee. Readers interested in the subsequent development of a research topic are encouraged to contact either the author or the Australian Bureau of Statistics.

GENERALISED LINEAR MODELS WITH PROBABILISTICALLY LINKED DATA: APPLICATION TO THE SIMULATED STATISTICAL LONGITUDINAL CENSUS DATASET

James Chipperfield
Analytical Services

ABSTRACT

The Australian Bureau of Statistics has embarked on the Census Data Enhancement project, the key feature of which is to create a Statistical Longitudinal Census Dataset (SLCD) based on a random sample of 5% of person records from the 2006 Census. These will be linked to person records from 2011 and subsequent Censuses without using names and addresses as linking variables. The SLCD will provide a substantial opportunity for longitudinal analysis to see how people and their families change with time, while maintaining the ABS' strong commitment to the confidentiality of its Census respondents. Since a unique person identifier will not be available, some links will be incorrect, so some linked Census records will not correspond to the same individual. The ABS has conducted a quality study to assess the feasibility of forming the SLCD in this way and its likely quality. Part of the assessment has been to fit generalised linear models to longitudinal linked data. This paper describes and implements a method of adjusting regression coefficients in such models to account for incorrect links. Empirical results show that the adjustment method works well, especially as the number of incorrect links increases. Empirical findings also suggest that a possibly more significant source of error arises when certain sub-populations are underrepresented in the linked data set.

1. INTRODUCTION

1.1 Creating a longitudinal dataset

The Australian Census of Population and Housing is conducted every five years and obtains detailed information from all persons in Australia on Census night, which was the 8th August for the 2006 Census.

The Australian Bureau of Statistics has embarked on a project to create a Statistical Longitudinal Census Dataset (SLCD) based on a simple random sample of 5% of person records from the 2006 Census that will be linked to person records from subsequent Censuses. The SLCD will be augmented at each future census with a 5% random sample of people who have been born or have migrated to Australia since the preceding Census. The SLCD provides a substantial opportunity for longitudinal analysis at a relatively small geographical level while maintaining the ABS' strong commitment to maintain the confidentiality of its Census respondents.

Since a unique person identifier will not be available, some links will be incorrect, meaning that some linked Census records will not correspond to the same individual. Linking will be implemented using probabilistic methods (Conn and Bishop, 2005). It is proposed that the 5% sample will be linked to the 2011 Census without using name and address. All names and addresses used by the ABS during the 2006 Census processing period have been destroyed.

The ABS conducted a quality study to assess the feasibility of linking the 5% sample to the 2011 Census without name and address. This study used data from the Census Dress Rehearsal conducted one year before the 2006 Census and comprising approximately 80,000 persons. This was an attempt to simulate the formation of the SLCD. This paper discusses some of the results of the quality study.

1.2 The simulated formation of the SLCD

This quality 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 Census. The 2006 Census collected information from more than 19 million people. The results of this study were used to inform an assessment about the reliability of analysis conducted on the SLCD.

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

- *Gold Standard* uses 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.
- *Silver Standard* used Hash Value (HV), mesh block and selected Census data items. Each name is assigned to 1 of 12,000 HVs, where each HV has a minimum of 1,500 distinct names. All HVs were destroyed at the end of the Census processing period.
- *Bronze Standard* used mesh block and selected Census data items. This is a method proposed to be used for the linking of the sample drawn from the 2006 Census to the 2011 Census.

The role of the Gold Standard in the quality study is critical. It provides a reliable benchmark against which the reliability of the Bronze Standard/Silver Standard can be compared. Its reliability is due to the fact that name and address are powerful variables for the purpose of identifying common individuals on the Census and CDR. If name and address were available, the Gold Standard would be preferred to the Silver and Bronze standards.

As a result, each Gold Standard link is assumed to join records belonging to the same individual and the set of Gold Standard links are assumed to cover all individuals who are common to the CDR and Census. Accordingly, differences between estimates based on the Gold Standard and the Silver Standard/Bronze Standard are interpreted as error. In other words, interest focuses on the reliability of the Silver Standard/Bronze Standard *relative* to the Gold Standard.

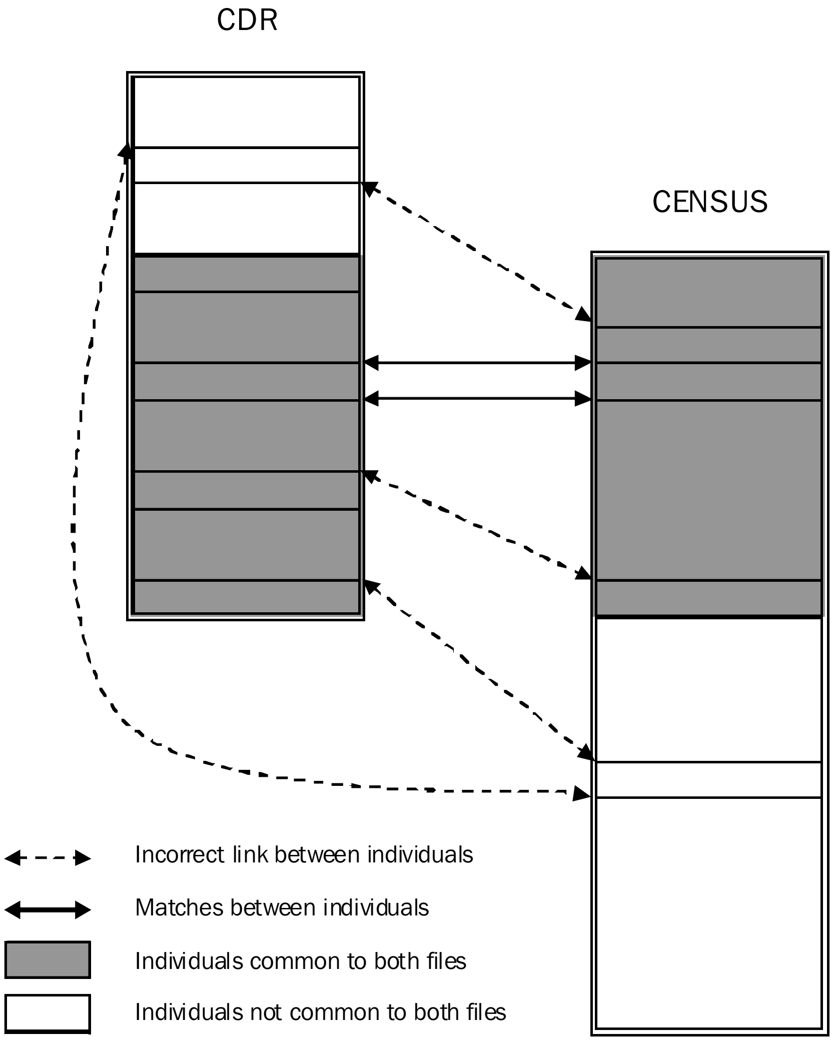
It was thought that a major error affecting the Bronze Standard/Silver Standard estimates would be due to linkage error. Linkage error arises when a pair of linked records do not correspond to the same individual. An important effect of linkage error is to bias estimates, such as those estimates used in longitudinal analysis. The main aim of this paper is to assess whether the effects of linkage error could be reduced by using a methodology recently developed by Chambers (2008).

Now we discuss linkage error in more detail. Through the linking process, the aim was to link the records on the CDR file and the Census file corresponding to an individual for those individuals common to both the CDR and Census. A pair of records belonging to the same individual is called a match and we have defined the set of matches as the linked pairs of records in the Gold Standard. These are shown with a grey background in figure 1.1. Records with a white background correspond to individuals who are not common to the Census and CDR, i.e. are present in one or the other file only, there being millions of these in the Census file but only a few thousand in the CDR file.

When performing a Silver or Bronze Standard linkage, a pair of records may be linked correctly, i.e. correspond to a match, and these are shown by unbroken lines in figure 1.1. Alternatively a pair of records may be linked incorrectly. The broken lines in figure 1.1 illustrate the different types of incorrect linkages that were made for the Bronze and Silver standards.

Figure 1.1 illustrates another type of error referred to as a missed link. This error arises for the Bronze Standard or the Silver Standard when a CDR record corresponding to an individual who is common to the CDR and Census (i.e. has a grey background) is not linked (i.e. has no corresponding arrow, whether broken or unbroken).

1.1 Matches and linkage errors


CDR
CENSUS
Incorrect link between individuals
Matches between individuals
Individuals common to both files
Individuals not common to both files

The main aim of this paper is to assess whether the effects of linkage error could be reduced when fitting a logistic model. Section 2 broadly describes the process of linking the CDR to the 2006 Census under the Gold, Silver and Bronze standards. Section 3 describes the methodology for adjusting regression coefficients to account for incorrect links. Section 4 describes how the methodology was implemented. Section 5 evaluates the performance of the adjustment method. Section 6 makes some concluding remarks.

2. THE SIMULATED SLCD LINKING METHODOLOGY

This section provides an overview of the CDR-to-Census linkage methodology for the Bronze, Silver and Gold standards. The linking methodology consists 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 for all the passes.

When linking two files of size N and M, there are a total of NM possible record pairs. A record pair is one record from each file that is considered as a possible link. Blocking is a method by which the number of record pairs is reduced to computationally feasible levels while attempting not to discard any record pair that corresponds to a match. Blocking means any two records can only be considered as a possible link if they take the same value for the blocking variable(s). Table 2.1 gives the blocking variables, denoted by "B" for the Bronze Standard. For example, Table 2.1 shows that during Pass 1 of the Bronze Standard, two records can only be considered as a possible link if they have the same value for mesh block.

Linking variables are used to measure the degree of agreement (discussed more below) between a pair of records. A high level of agreement suggests that the likelihood of the record pair being a match is high. Table 2.1 gives the linking variables, denoted by "L", for the Bronze Standard. For example, table 2.1 shows that during Pass 1 of the Bronze Standard, a range of person-level information including day, month and year of birth are linking variables.

2.1 Bronze Standard: Blocking (B) and linking (L) variables

Variable	*Pass 1*	*Pass 2*
Day of birth	L	
Month of birth	L	
Year of birth	L	
Full date of birth (Valid values only)		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
Non school qualification field of study	L	L
Non school qualification level of education	L	L
Highest level of schooling	L	L
Mesh block	B	L

An output from each pass is a weight for each pair of records that were compared. The weight is a measure of the level of agreement between the pair of records. We defer the formal definition of the record pair or comparison weight to Conn and Bishop (2006) (see equation (3.6)). For example, consider the Bronze Standard in Pass 2 where record pairs have the same full date of birth and sex; a record pair would be assigned a weight 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 comparison weight but for illustration purposes we ignore them.) The weight for agreement on mesh block (+17) is greater than the weight for agreement on year of arrival – the former is less likely to occur by chance alone. To formalise, the record pair comparison weight for record i on the CDR and record j on the Census during pass p of linking standard s is denoted by c_{spij}, where i indexes CDR records and j indexes Census records available for linking in pass p.

If record i and j are compared in passes 1 and 2, then $c_{s1ij} \neq c_{s2ij}$. This is because the linking variables, which in part determine the record pair comparison weight, used in pass 1 and 2 are not the same.

The record pair weights c_{spij} and the cut-off f_{sp} are used by the linking package *Febrl* to determine the optimal set of links in pass p. The term f_{sp} is the minimum weight that is required for a record pair to form a link during pass p. Clearly, the number of links depends upon f_{sp}. The cut-off is usually set after some clerical review so that the linked data set contains as many matches as possible while at the same time not including too many non-matches.

A series of cut-offs were used for the Bronze and Silver standards. The higher the cut-off, the less likely is a link to be a non-match and the more likely that a match is not linked. In what follows, we focus only on Silver Standard-VL, Silver Standard-UL, Bronze Standard-VL and Bronze Standard-UL, where VL and UL denote a very low and ultra low cut-off respectively. This is because extensive univariate and multivariate analysis showed that the very low cut-off resulted in parameter estimates that were closest to the corresponding Gold Standard estimates. The ultra low cut-off was an attempt to link as many CDR records as possible and is considered here as a point of contrast.

3. STATISTICAL METHODOLOGY

3.1 Literature review of probabilistic record linkage

The goal of probabilistic record linkage is to join together two files that contain information on an overlapping set of individuals but which lack sufficient unique identifier information. A natural consequence of this type of record linkage is that the two files will be joined imperfectly. When joining two files, such as the CDR and Census, for analysis, interest would naturally focus on making inference about individuals that are *common* to the two files.

Naively treating the joined files as if they were without error will lead to biased estimates. Lahiri and Larsen (2005) and Scheuren and Winkler (1993) propose methods to calculate unbiased estimates of coefficients in a linear regression model under probabilistic record linkage. These methods are not directly useful in the CDR-to-Census situation for two reasons. First, linear regression is not well suited to analysis of Census data items which are dichotomous, nominal or ordinal. Second, these methods assume that the files to be joined are of the same size and include the same individuals; in the CDR-to-Census linkage there are many individuals in the Census file that are not in the CDR file.

More recently, Chambers (2008) extended this work to a much wider set of models using estimating equations (see Chambers and Skinner, 2003). The set of models includes those that are suitable for analysing Census data items (e.g. logistic regression). Chambers (2008) also allows for individuals in one of the files, say file X, to be a subset of the individuals in the other file, say file Y. Chambers (2008) does not allow for the more general case, as illustrated in figure 1.1, where there are individuals on the CDR (file X) that are not on the Census file (file Y). This would occur, for example, if a CDR respondent was not in Australia on Census night.

All the above methods rely on knowing the probability that record i on file X is a correctly linked. These probabilities define the model for the linkage error. In general, these probabilities are unknown and must be estimated. Estimating these probabilities for the CDR-to-Census linkage is discussed in Section 4.

3.2 Estimation

Consider a situation when file Y contains a scalar variable y and file X contains a K row vector of variables x, and where individuals on file X are a subsample of the individuals on file Y. In this section $i = 1, \dots, n$ indexes the records on file X, $j = 1, \dots, N$ indexes the records on file Y, and we assume that all records on file X are linked. The rest of this section closely follows the development in Chambers (2008).

A general approach to parameter estimation is estimating equations (see Chambers and Skinner, 2003). An estimate of the K vector of parameters θ in the regression of y_i on x_i, characterised by the function $f(\theta)$, is obtained by solving the estimating equation $H(\theta) = 0$, where

$$H(\theta) = \sum_i^n G_i \left[y_i - f_i(\theta) \right] \tag{1}$$

where G_i is some function of x_i, $E[y_i] = f_i(\theta)$, and $E[]$ is the expectation with respect to the model. The form of f depends upon the assumed model for y – e.g. for the logistic regression model,

$$f_i(\theta) = \frac{\exp(X_i\theta)}{1+\exp(X_i\theta)}$$

Under probabilistic record linkage, the estimating equation which naively treats all links as correct, is

$$H_N(\theta) = \sum_i^n G_i \left[y_i^* - f_i(\theta) \right] \tag{2}$$

where y_i^* is the value of y that is linked to record i on file X rather than the true value, y_i. The estimate of θ obtained by solving $H_N(\theta) = 0$ will be biased.

Using a general result given by Breckling *et al.* (1995), solving the adjusted estimating equation $H_{Adj}(\theta) = E_M E_X\{H_N(\theta)\} = 0$ will give an unbiased estimate of θ, where E_M and E_X are the expectations with respect to the linkage error model (discussed more below) and the model respectively.

We define A to be the $n \times N$ permutation matrix of 0s except for 1s in the $(i,,j)$th element corresponding to where record i on file X is linked to record j on file Y. The i and j labels are consistent such that records i and j corresponds to a match when $i = j$. This means that if record i is correctly linked, a 1 will appear in the (i,i)th element of A. The matrix A is a realisation of an underlying process referred to as the linkage error model. We define $E = (E_{ij})$ to be an $n \times N$ matrix: with (i,i) th element equal to the probability that record i is correctly linked; and with (i,j) th element equal to the probability that record i is incorrectly linked to record j, where $i \neq j$.

Accordingly, we define an unbiased estimating equation

$$H_{Adj}(\theta) = \sum_i^n G_i \left[y_i^* - E_M E_X\{y_i^*\} \right] = \sum_i^N G_i \left[y_i^* - E_i^T F_i(\theta) \right] \tag{3}$$

Equation (3) follows by noting that

$$E_M E_X\{y_i^*\} = E_M E_X\{A_i y_i\} = E_M\{A_i\} E_X\{y_i\} = E_i^T F_i(\theta)$$

where F_i and E_i are row vectors of dimension N with j th elements $f_j(\theta)$ and E_{ij} respectively and $E = (E_1, \ldots, E_i, \ldots, E_n)'$. We now make three important points about $H_{Adj}(\theta)$.

First, to simplify the form of E, we assume that each record j on file Y has a non-zero probability of being linked (either correctly or incorrectly) with one and only one record on file X. The validity of this assumption for the CDR-to-Census linkage is tested in Section 4.

We define the set P_i of size N_i to be the set of records on file Y that have a non-zero chance of being linked (either correctly or incorrectly) with record i on file X. Clearly then $P = \bigcup_i P_i$ is the set of all N records on file Y. By definition, the set P_i must include the records on file Y that were linked and are a match with record i on file X. In the context of the CDR-to-Census under Bronze Standard linkage, the set P_i must include the Census records that were linked to record i by both the Bronze Standard and the Gold Standard, where Gold Standard links are by definition matches.

Second, $H_{Adj}(\theta)$ requires knowledge of E_i for $i = 1, \ldots, n$. Further, we assume that the probability of record i on file X being correctly linked is λ_i and the probability of record i being incorrectly linked to one of the remaining $N_i - 1$ records in P_i is $(1 - \lambda_i)/(N_i - 1)$. The problem of estimating E_i then becomes one of estimating λ_i. Estimating λ_i for the CDR-to-Census situation is described in Section 4.

Third, since x_i and $F_i(\theta)$ are only available on the sample, the last term on the right hand side of (3) must be estimated. After some algebra (Chambers, 2008), (3) simplifies to:

$$H_{Adj}(\theta) = \sum_i^n G_i \left[y_i^* - L_i f_i(\theta) \right] \tag{4}$$

where $L_i = [\lambda_i N_i - 1 + (b_i - 1)(1 - \lambda_i)]/(N_i - 1)$, $b_i = n/N$ is the probability of selecting record i on file X, and $G_i = x_i$. If $L_i = 1$, which occurs if $\lambda_i = 1$, for all i then $H_{Adj}(\theta) = H_N(\theta)$.

Another unbiased estimating equation is given by (5) and is obtained by replacing $G_i = x_i$ in (3) with its expectation with respect to the linkage error given by $E_M[x_i] = Q_i E_i^T$, where $Q_i = (x_1, x_2, \ldots, x_j, \ldots, x_N)$.

$$H_{Adj2}(\theta) = \sum_i Q_i E_i^T \left[y_i^* - E_M E_X\{y_i^*\} \right] = \sum_i Q_i E_i^T \left[y_i^* - E_i^T F_i(\theta) \right] \tag{5}$$

Following the same steps as above, this estimating equation simplifies to

$$H_{Adj2}(\theta) = \sum_i x_i L_i \left[y_i^* - L_i f_i(\theta) \right] \tag{6}$$

The empirical evaluation of Section 5 estimated θ by solving (6), rather than solving (4). Though not reported in this paper, (6) performed considerably better than (4). The corresponding variance estimator for estimates of θ obtained from (6) is provided in Chambers (2008) but is not provided in this paper.

We now specifically consider estimating θ for the logistic model given by

$$\ln\left[\pi_i \,(1 - \pi_i)^{-1} \right] = x_i^T \theta + \varepsilon_i \tag{7}$$

where y_i is the response (containing 0s and 1s) for unit i, π_i is the probability that $y_i = 0$, and ε_i are residuals that are independent and normally distributed. An estimate of θ, obtained by solving $H_{Adj2}(\theta) = 0$ is obtained using the Newton–Raphson algorithm (see Chong and Zak, 1996). This algorithm is described below:

1. Choose initial estimates of the regression coefficients $\theta^{(0)} = 0$.

2. At each iteration, update the coefficients:

   $\theta^{(t+1)} = \theta^{(t)} + (\Sigma_i L_i^2 v_i^{(t)} x_i^T x_i)^{-1} \Sigma_i L_i x_i^T (y_i - p_i^{(t)})$

   where $p_i^{(t)} = [1 + exp(-x_i^T \theta^{(t)})]^{-1}$ and $v_i^{(t)} = p_i^{(t)}(1 - p_i^{(t)})$.

3. Repeat Step 2 until $\theta^{(t+1)} - \theta^{(t)}$ is small (i.e. <0.001).

4. MODELLING THE LINKAGE ERROR

A major step in implementing the method described in Section 3 was to model the linkage error defined by E. Section 4.1 describes the broad steps involved in modelling the linkage error for the CDR (file X) to the Census (file Y) linkage. Sections 4.2 and 4.3 describe the proposed approach for estimating the linkage error for the Bronze Standard and the Silver Standard while Section 4.4 mentions some of the other sub-optimal approaches that were evaluated.

4.1 Broad steps in modelling the linkage error

A major step in modelling the linkage error was to define the sets P_i for $i = 1, \ldots n$, where P_i is the set of Census records on file Y that have a non-zero chance of being linked (either correctly or incorrectly) with record i on the CDR and n is the number of linked CDR records. The basic idea was to allocate Census records to P_i if they had a high record pair comparison weight with the ith CDR record. Many Census records were assigned small, negative or no record pair weight with all of the CDR records; these Census records were not allocated to any of the sets P_i and accordingly they are assumed to have a zero probability of being linked with any of the CDR records.

For example, if a Census record did not report a mesh block or date of birth during the Bronze Standard linkage, it would not be assigned to a record pair, or assigned a record pair comparison weight, in either pass 1 or 2 and so would not be allocated to any of the sets P_i. For the Bronze Standard with a very low cut-off, 57,790 CDR records were linked ($n = 57{,}790$) and these CDR records had a non-zero probability of being linked with only 82,000 Census records ($N = 82{,}000$). This meant that while there were over 19 million Census records in scope of being linked to the 57,790 CDR records, we assumed that only 82,000 Census records had a non-zero probability of being linked with one of the 57,790 CDR records.

On many occasions, a Census record had a high weight with more than one CDR record. However, as mentioned above, the structure imposed on E requires that only one of these record pairs with a high weight has a non-zero chance of being linked. In terms of notation, this means a Census record is allocated to no more than one of the sets P_i. Section 4.2 discusses this allocation process for the Bronze Standard.

The assumptions mentioned above are strong assumptions. The validity of these assumptions were tested, found to be reasonable, and greatly simplified the estimation of E. This is discussed further in Section 4.2.

Another major step in estimating E was estimating λ_i, the probability that record i on the CDR is correctly linked. The idea was to allocate linked CDR records to one of a number of strata, where strata were formed to be homogeneous in λ. The probability that a CDR record in stratum h is correctly linked using linking standard s is

$\lambda_{hs} = m_{hs}/n_{hs}$, where m_{hs} is the number of matches in stratum h and n_{hs} is the number of CDR records in stratum h that were linked using standard s.

The linked CDR records were potentially allocated to strata on the basis of the pass in which they were linked, their link comparison weight, and N_i. For example, *pass* was used in the stratification of CDR records linked by the Bronze Standard since the link-accuracy for pass 1 and for pass 2 were quite different.

4.2 Modelling the linkage error for the Bronze Standard

This section describes the proposed method for modelling the linkage error for the Bronze Standard-VL.

The information used in the modelling process is a file of all record pairs and corresponding record pair comparison weights. Record pairs were only available on the file if their record pair comparison weight was greater than 0. In the case of Bronze Standard-VL, this file includes over four million record pairs that were available after the 57,790 linked CDR records were compared with the over 19 million Census records; from this file of comparison weights, the sets P_i for $i = 1, \ldots, 57\,790$ were formed. These sets were formed by steps 1 and 2 below.

Step 1: Define the initial sets P_i for $i = 1, \ldots, n$.

We define P_i to be the set of Census records that have the top 10 record pair weights with the i th CDR record, where the record pair weight must be greater than 9. For example, if record i has comparison weights greater than 9 with five Census records then P_i is defined by that set of five Census records.

Through empirical evaluations, we found that the values '10' and '9' just mentioned resulted in the set of adjusted regression coefficients that were closest to the corresponding Gold coefficients, where the distance measure is given by (8).

To illustrate what this means in practice, consider a 30 year old, Indigenous, female, post graduate student who responded to the Census, and a 17 year old non-Indigenous, male, high school student who responded to the CDR. Even if these individuals lived in the same mesh block, their Bronze Standard record pair weight would be less than 9 because their values for key linking variables do not agree. Accordingly the 30 year old would not appear in the set for the 17 year old, and these two records would consequently be assigned a zero chance of being linked. Further, if the record for the 30 year old Census respondent was not included in *any* record pair that had a weight greater than 9 then the record would not be assigned to any set and consequently not a member of P.

Step 2: Finalise the sets P_i for $i = 1, \ldots, n$.

It is possible that a particular Census record is in more than one of the initial sets P_i for $i = 1, \ldots, n$. However, the structure imposed on E requires that a Census record is allocated to no more than one of the final sets P_i for $i = 1, \ldots, n$. To form final versions of the sets P_i , Census records *j* are removed from all initial sets except the *i*th set, where the *i*th CDR record:

(a) is linked to Census record *j*; and

(b) has the highest record pair comparison weight with Census record *j*, if record *j* is not linked.

The final 57,790 sets were made up of N = 82,000 Census records. This meant that while there were over 19 million Census records in scope to be linked to the 57,790 CDR records, we assumed that there were only 82,000 Census records that had a non-zero chance of being linked with the 57,790 CDR records.

The theoretical development in Section 3 requires that the set P_i for $i = 1, \ldots, n$ contain the Census records that were linked to the *i* th CDR record by both the Bronze Standard and the Gold Standard (i.e. the match). It was easy to ensure that this was the case for the Bronze Standard link. However, this was not always the case for the Gold Standard link. In particular, 2,152 out of the 57,790 Bronze Standard-VL sets did not contain the Gold Standard linked Census record because of one of the following:

1. The Gold Standard link did not exist – 1,300 CDR records were linked by Bronze Standard-VL but were not linked by the Gold Standard. Because these 1,300 Bronze-Standard-VL links did not appear in the set of Gold Standard links they are, by definition, incorrect.

2. The Gold Standard link was not present on the file of four million record pairs because it was not assigned a positive record pair weight by Bronze Standard-VL. This occurred in 610 sets and could arise in situations where Gold Standard links relied on a high level of agreement on name and address but a low level of agreement on the Bronze Standard linking variables.

3. The Gold Standard link was discarded as a result of steps 1 and 2 above. This occurred in 242 sets.

In terms of implementing the method described in Section 3, we assumed that the Gold Standard link *was* in the set, even though we knew this *was not* true for 2,152 sets. It was appropriate to do this because for the SLCD linkage, where Gold Standard links will not be available, we will be making such an assumption.

We know that at least 400 of the 852 (= 610 + 242) errors described by points 2 and 3 arose for individuals living in Indigenous communities. Investigations showed that 400 individuals living in Indigenous communities were linked by the Gold Standard, and none of these individuals were linked by the Bronze Standard. The latter was due in part to poor reporting of the Bronze Standard linking and blocking variables and difficulties of assigning a mesh block to remote communities. Obtaining high quality information would reduce the number of times this situation arises.

Regarding point 3, discarding 242 Gold Standard links during steps 1 and 2 means that there is a small error in the value of N_i for 242 sets (i.e. these sets should have included at least an additional Census record corresponding to the Gold Standard link). When compared to the 57,790 Bronze Standard-VL linked records, 242 is an acceptably small error. Section 4.3 briefly mentions sub-optimal options that do not involve discarding any of the 242 links.

Step 3: Calculate λ_i

As mentioned above, the idea was to allocate each linked Bronze Standard/Silver Standard CDR record to a stratum so that CDR records within the same stratum had a similar probability of being correctly linked. Strata were formed pragmatically. The strata for Bronze Standard-VL are given in table 4.1. The table shows, for example, that links with a weight greater than 26 in pass 2 were all matches.

4.1 Strata for Bronze Standard-VL

Stratum	*Pass*	*Weight range*	*Set size N_i*	*Probability that link is a match λ*
1	1	<18	1	0.73
2	1	<18	>1	0.40
3	2	<18	1	0.91
4	2	<18	2	0.85
5	2	<18	>2	0.47
6	2	18–26	n/a	0.93
7	1	18–26	1	0.98
8	1	18–26	>1	0.95
9	1	>26	n/a	0.99
10	2	>26	n/a	1.00

Steps 1,2 and 3 were repeated for Bronze Standard-UL and Silver Standard-UL though the results are not reported here.

4.3 Modelling the linkage error for the Silver Standard

The steps described in Section 4.1 were taken to model the linkage error for Silver Standard-VL. The results are given in table 4.2. The slight difference was that all links with weights greater than 20 were assumed to be matches (i.e. $\lambda=1$). This was a reasonable assumption since 99.2% of the links were matches.

4.2 Strata for Silver Standard-VL

Stratum	*Pass*	*Set size N_i*	*Weight range*	*Probability that link is a match λ*
1	1	1	<20	0.97
2	2	1	<20	0.95
3	3	1	<20	0.86
4	1	2	<20	0.81
5	2	2	<20	0.76
6	3	2	<20	0.60
7	1	3	<20	0.53
8	2	3	<20	0.49
9	3	3	<20	0.46
10	1	4	<20	0.33
11	2	4	<20	0.38
12	3	4	<20	0.28
13	2	5	<20	0.42
14	3	5	<20	0.49
15	2	6	<20	0.36
16	3	6	<20	0.23
17	2	>6	<20	0.30
18	3	>6	<20	0.14
19	123	n/a	>20	1.00

It is worthwhile noting that the distributions of the Bronze Standard and Silver Standard record pair comparison weights are not comparable. This is due to the fact that the Bronze and Silver standards use different linking variables.

4.4 Alternative linkage error models

A total of eight variations to step 1, described in Section 4.1, were considered for modelling the linkage error. These included defining P_i to be the set of Census records that had the top 5, 10, 25 and 200 (four options) record pair comparison weights with the ith CDR record, where the record pair comparison weight must be greater than 0 or 9 (two options). None of these alternatives performed as well as the proposed approach described in Sections 4.2 and 4.3 and so are not reported in this paper. Performance was measured by (8) (see below), which measures how close a set of regression parameters is to the corresponding set of Gold Standard parameters. The performance of the proposed approach is evaluated in Section 5.

5. EVALUATION

This section evaluates the method of Section 3 that is designed to give unbiased estimates of regression coefficients under probabilistic linkage. We refer to these estimates as 'adjusted'. We refer to estimates that assume no linkage error as 'naive'. This evaluation considers adjusted and naive estimates for Bronze Standard-VL, Bronze Standard-UL, Silver Standard-VL and Silver Standard-UL.

This evaluation includes fitting logistic regression models that predict the odds that:

1. a person moves between 2005 and 2006;
2. a person 15 years and older is employed in 2006; or
3. a person 15 years and older is a student in 2006.

All models' explanatory variables were from the CDR and all models' dependent variables were from the Census. Each model is computed on records where there are no missing model variables (i.e. only complete cases were used).

Section 5.1 gives a simple presentation of the differences between the regression coefficients based on the Bronze Standard/Silver Standard and the Gold Standard. Section 5.2 quantifies the contributing factors behind these differences.

5.1 Simplistic evaluation

The model parameters estimated from the Bronze Standard and the Silver Standard, whether adjusted or naive, were compared with those estimated from the Gold Standard using (8).

$$\text{Deviance of coefficients for standard } S = \frac{1}{K}\sum_k \frac{|G_k - S_k|}{se\{G_k\}} \qquad (8)$$

where

S denotes the set of regression coefficients estimated using Bronze Standard-VL, Bronze Standard-UL, Silver Standard-VL or Silver Standard-UL, and may be adjusted or naive,

S_k is the kth model parameter for standard S,

G_k is the kth model parameter using the Gold Standard,

$se\{G_k\}$ is the standard error of the kth model parameter using the Gold Standard,

$k = 1, \ldots, K$,

K is the number of parameters in the model.

While (8) does not any obvious statistical properties, it has intuitive appeal. It can be interpreted as the average difference between a Bronze and its corresponding Gold regression coefficient, where the difference is measured in terms of the number of standard deviations of the coefficient. A small deviance for standard *S*'s coefficients means they are close to the corresponding Gold Standard coefficients.

Table 5.1 gives the deviance measures for the three logistic models. The table shows that the deviance for the naive and adjusted Bronze Standard-VL coefficients for the model, *Probability of employment in 2006*, are both 0.66. This means that the adjusted coefficients are not closer to the Gold Standard coefficients than the naive coefficients.

5.1(a) Deviance of adjusted and unadjusted (naive) coefficients – Bronze Standard

	Bronze Standard-VL		*Bronze Standard-UL*	
Model	*Unadjusted*	*Adjusted*	*Unadjusted*	*Adjusted*
Person moves between 2005 and 2006	0.78	0.77	1.06	0.61
Employment in 2006	0.66	0.66	2.32	1.79
Student in 2006	0.59	0.57	1.36	1.00
Average	0.68	0.67	1.58	1.13

5.1(b) Deviance of adjusted and unadjusted (naive) coefficients – Silver Standard

	Silver Standard-VL		*Silver Standard-UL*	
Model	*Unadjusted*	*Adjusted*	*Unadjusted*	*Adjusted*
Person moves between 2005 and 2006	0.43	0.44	1.09	0.89
Employment in 2006	0.42	0.40	2.49	2.07
Student in 2006	0.39	0.31	3.05	2.51
Average	0.41	0.38	2.21	1.82

On average, across the three models, the deviance for the naive and adjusted coefficients are:

- 0.68 and 0.67 for Bronze Standard-VL and 0.41 and 0.38 for Silver Standard-VL – this means the adjustment reduces the deviance by 1.5% and 7.5% respectively.
- 1.58 and 1.13 for Bronze Standard-UL and 2.21 and 1.82 for Silver Standard-UL – this means the adjustment reduces the deviance by 28% and 18% respectively.

With the possible exception of the ultra low cut-off, the adjustment method only marginally reduces the deviance. The deviances for the ultra-low cut-off is generally larger than the deviances for the very-low cut-off because of the increased number of incorrect links. This is discussed further in Section 5.2.

The Silver Standard coefficients are always closer (indicated by a smaller deviance) to the corresponding Gold Standard coefficients than the Bronze Standard coefficients, and so we would conclude that in general the Silver Standard coefficients are of higher quality than the Bronze Standard coefficients. Also, on average a very low cut-off results in a much smaller deviance than an ultra low cut-off.

5.2 Explanation of results

The deviance presented in table 5.1 is a simple measure of difference between the Bronze Standard/Silver Standard and the Gold Standard model coefficients. Equivalently the deviance is a measure of error in the Bronze Standard/Silver Standard coefficients. This section goes some way to quantifying the different sources of error. The sources of error in the estimates of the Bronze Standard/Silver Standard coefficients are:

1. Sampling error. This error is caused by matches that were not linked. Two sources of sampling error are due to:

   a. variability. Sampling variability arises because CDR records linked by the Silver Standard/Bronze Standard are a sub-sample of the CDR records that are linked by the Gold Standard.

   b. informativeness due to sampling bias. Sampling bias arises when the characteristics of the CDR records that are linked by the Bronze Standard/Silver Standard are different to the characteristics of the CDR records that are linked by the Gold Standard. Informativeness arises when there is sampling bias and it is not accounted for by the regression model. Sampling informativeness could arise if Indigenous people were under- represented in the Silver Standard/Bronze Standard and Indigenous status is not included as an explanatory variable in the logistic model. Clearly, if the characteristics of the CDR records linked by the Bronze Standard/Silver Standard were a random sample of the CDR records links by the Gold Standard, then this component of error would be zero.

2. Linkage error. Linkage error occurs when a linked pair of records does not correspond to the same individual. The effects of linkage error are due to:

   a. variability induced by the adjustment methodology described in Section 3. This source of error can be measured (see Chambers , 2008). Clearly this source of error does not apply to naive estimates, which by definition are not adjusted in any way.

b. Bias due to systematic differences between the estimated Gold Standard and Bronze Standard/Silver Standard coefficients. The methodology in Chambers (2008) attempts to reduce or eliminate the bias observed in naive estimates due to incorrect linkage at the cost of inducing some variability (see 2a.) However, adjusted estimates may be biased if the linkage error model is mis-specified.

3. Out-of-scope links. This error occurs when a CDR record is linked to a Census record, but the CDR record does not have a matching record on the Census. In other words, these are CDR records that were linked by the Bronze Standard/Silver Standard but not by the Gold Standard and are illustrated in figure 1.1 by the CDR records with a white background. These CDR records should not have been linked by the Bronze Standard/Silver Standard and so cause a bias. For example, there were 1300 such CDR records for Bronze Standard-VL.

To help measure the three sources of error just mentioned, we define some notation. We define $S^{(c)}$ to be the subset of the Bronze Standard/Silver Standard links that include CDR records that have a matching record on the Census. (These CDR records have a grey background in figure 1.1.) For example, for Bronze Standard-VL, $S^{(c)}$ is made up of 56,490 (= 57,790 – 1,300) links. Also, we define $G^{(S_c)}$ to be the set of Gold Standard links that only include the CDR records that are present in $S^{(c)}$. It follows that for Bronze Standard-VL, $G^{(S_c)}$ and $S^{(c)}$ are made up of the same 56,490 CDR records but the former has Gold Standard links and the latter has Bronze Standard links. The difference between estimates based on $G^{(S_c)}$ and $S^{(c)}$ is therefore only due to linkage error (point 2). We identify the kth coefficient from the logistic regression by subscript k, so that for example, $S_k^{(c)}$ is the kth regression coefficient that was estimated from data $S^{(c)}$.

A measure of the total error in the estimates for standard S is

$$Error = \frac{1}{K}\sum_k \frac{(G_k - S_k)^2}{se\{G_k\}^2} \tag{9}$$

which is crudely approximated by

$$Crude\ Error = \frac{1}{K}\sum_k \frac{\left[(G_k - G_k^{(S_c)})^2 + (S_k^{(c)} - G_k^{(S_c)})^2 + (S_k^{(c)} - S_k)^2\right]}{se\{G_k\}^2} \tag{10}$$

The approximation is crude but it is useful to make because the three components of error mentioned above are neatly separated. Namely, the first, second and third terms in (10) correspond to the errors 1, 2 and 3 described above. For example, in the case of Bronze Standard-VL, the second term in (10) would be zero if all the Bronze Standard-VL links were matches. Table 5.2 gives the Crude Error and table 5.3 gives its corresponding three components.

5.2(a) Crude Error – Bronze Standard

	Bronze Standard-VL		*Bronze Standard-UL*	
Model	*Unadjusted*	*Adjusted*	*Unadjusted*	*Adjusted*
Person moves between 2005 and 2006	1.63	1.55	1.48	1.02
Employment in 2006	1.36	1.25	6.55	3.75
Student in 2006	0.69	0.61	8.50	5.91

5.2(b) Crude Error– Silver Standard

	Silver Standard-VL		*Silver Standard-UL*	
Model	*Unadjusted*	*Adjusted*	*Unadjusted*	*Adjusted*
Person moves between 2005 and 2006	0.30	0.27	1.20	1.12
Employment in 2006	0.49	0.42	7.24	5.38
Student in 2006	0.19	0.12	8.50	5.90

5.3(a) Sampling error, linkage error and out-of-scope links – Bronze Standard

	Bronze Standard-VL		*Bronze Standard-UL*	
Model	*Unadjusted*	*Adjusted*	*Unadjusted*	*Adjusted*
Person moves between 2005 and 2006	(1.36, 0.17, 0.08)	(1.36, 0.12, 0.07)	(0.04, 1.09, 0.36)	(0.04, 0.48, 0.53)
Employment in 2006	(0.98, 0.19, 0.20)	(0.98, 0.13, 0.20)	(0.07, 5.41, 1.07)	(0.07, 1.08, 2.50)
Student in 2006	(0.41, 0.18, 0.10)	(0.41, 0.13, 0.08)	(0.02, 1.54, 0.65)	(0.02, 0.40, 0.98)

5.3(b) Sampling error, linkage error and out-of-scope links – Silver Standard

	Silver Standard-VL		*Silver Standard-UL*	
Model	*Unadjusted*	*Adjusted*	*Unadjusted*	*Adjusted*
Person moves between 2005 and 2006	(0.18, 0.07, 0.06)	(0.18, 0.04, 0.05)	(0.01, 0.43, 0.75)	(0.01, 0.20, 0.90)
Employment in 2006	(0.20, 0.14, 0.16)	(0.20, 0.11, 0.12)	(0.00, 4.70, 2.50)	(0.00, 1.19, 4.18)
Student in 2006	(0.04, 0.06, 0.08)	(0.04, 0.03, 0.05)	(0.01, 5.88, 2.62)	(0.01, 1.71, 1.42)

Table 5.3 (a) shows, for the *Person moves between 2005 and 2006* model that the errors due to sampling, linking and out-of-scope links are 1.36, 0.17 and 0.08 for the unadjusted estimates and are 1.36, 0.12, and 0.07 for the adjusted estimates. More generally, table 5.3 (a) and (b) shows that for:

- Bronze Standard-VL and Silver Standard-VL, the sampling error is generally the biggest source of error, though this is more marked for Bronze Standard-VL
- Bronze Standard-VL and Silver Standard-VL, the error due to linkage and out-of-scope links are roughly equal.
- Bronze Standard-UL and Silver Standard-UL, the sampling error is the smallest source of error and in most cases is negligible. This is because many more records are linked with an ultra low cut-off than a very low cut-off.
- Bronze Standard-UL and Silver Standard-UL, the errors due to linkage and out-of-scope links are much larger than Bronze Standard-VL and Silver Standard-VL respectively.

Importantly, the amount of linkage error for adjusted estimates is always smaller than the linkage error for the corresponding naive estimates, particularly for the ultra low cut-off. For example, the linkage error on the naive and adjusted estimates are 5.88 and 1.71 respectively for the Student in 2006 model. This means that the method described in Section 3 is reducing the linkage error.

It is important to distinguish between errors that can lead to inappropriate inference and those than do not. Errors 1a. and 2a., defined above do not lead to inappropriate inference since they can easily be measured and incorporated into inference. However, the errors due to 1b., 2b. and 3. lead to bias. In practice this bias cannot be measured and can lead to inappropriate inference.

In this paper, we do not measure the relative size of errors 2a and 2b, only their combined total. This means that we cannot measure the component of the linkage error that may lead to inappropriate inference (2a.) and that which will not (2b.) However, next we do obtain some indication of the relative size of errors 1a. and 1b.

5.4 Test for sampling informativeness in the linked data

	Sampling error [50% and 1% upper tail of distribution under the Null]	
Model	*Bronze Standard-VL*	*Silver Standard-VL*
Person moves between 2005 and 2006	1.36 [0.17, 0.64]	0.18 [0.06, 0.14]
Employment in 2006	0.97 [0.17, 0.64]	0.20 [0.06, 0.11]
Student in 2006	0.41 [0.17, 0.37]	0.04 [0.05, 0.12]

We now test the null hypothesis that the sampling error (i.e. error 1) in table 5.3 can be fully explained by sampling variability (i.e. error 1a).

Ho :Sampling error in table 5.3 is due to only sampling variability

Ha :Sampling error in table 5.3 is not due to only sampling variability.

The distribution of the sampling errors in table 5.3 under the null hypothesis can readily be generated by simulation. The 50% and upper 0.01% tail of the distribution are presented in table 5.4. For example, the sampling error for the Bronze Standard-VL is 1.36 for the *Person Moves between 2005 and 2006 model*; under the null hypothesis we would expect the sampling error to be greater than 0.17 about 50% of the time and greater than 0.64 about 1% of the time. In all three models under Bronze Standard-VL, the null hypothesis is rejected at the 1% significance level. This is strong evidence of informative sampling. Further, we could conclude that informativeness accounts for a significant majority of the sampling error and that informativeness may contribute much more to the overall error than linkage error.

For Silver Standard-VL there is no evidence of informativeness in the *Student in 2006* model but there is strong evidence of informativeness for the other two models. Compared with Bronze Standard-VL, the impact of sampling error is a lot smaller in magnitude and closer in magnitude to the other sources of error. This suggests that the error due to informativeness, which is a component of sampling error, is much less of an issue for Silver Standard-VL than Bronze Standard-VL.

6. CONCLUSIONS

The main aim of this paper was to assess whether the effects of linkage error on estimates of regression parameters could be reduced when fitting a logistic model. The main conclusions about the CDR-to-Census linkage are:

1. The adjustment method significantly reduces the effects of linkage error for estimates of regression coefficients, particularly when the link accuracy is below 90%. The adjustment method could provide some protection against relatively low link accuracy when creating the SLCD.
2. When the link accuracy is above 95%, the error due to linkage error appears small and so the adjustment method is less important.
3. There is strong evidence of sample informativeness in the linked data that caused bias in regression estimates. There is evidence to suggest the bias due to sample informativeness is greater than the bias due to linkage error for the Bronze Standard, which does not use Hash Value as one of the linking variables. The impact of informativeness is significantly less for the Silver Standard, which uses Hash Value.
4. The quality of the Silver Standard is significantly higher than the Bronze Standard, largely because it is less exposed to bias due to informativeness in the linked sample.

The recommendation of this paper is that the adjustment method should be made available to analysts of the SLCD.

The model for the Bronze Standard/Silver Standard CDR-to-Census linkage error predicts the probability that any given record pair is a match. Matches were identified by Gold Standard links, which used name and address and consequently were of high quality. However, there will be no equivalent Gold Standard for the creation of the SLCD following the 2011 Census, since name and address will not be available. This means it will be very difficult to model the linkage error underlying the creation of the SLCD. A reasonable option may be to assume that the model for the CDR-to-Census linkage error is the same as the model for the SLCD linkage error.

ACKNOWLEDGMENTS

I would very much like to thank Raymond Chambers for his technical assistance. I would like to thank Glenys Bishop for her attention to detail that led to correcting technical errors. I would also like to acknowledge the contributions of Tim Ayre, Paul Campbell and Shaun McNaughton for their programming work and for our thoughtful discussions.

REFERENCES

Breckling, J.U.; Chambers, R.L.; Dorfman, A.H.; Tam, S.M. and Welsh, A.H. (1994) "Maximum Likelihood Inference from Sample Survey Data", *International Statistical Review*, 62(3), pp. 349–363.

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

Chambers, R.L. (2008) "Regression Analysis of Probability-Linked Data", unpublished report.

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>

Chong, E.K.P. and Zak, S.H. (1996) *An Introduction to Optimisation*, Wiley.

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.

Lahiri, P. and Larsen, M.D. (2005) "Regression Analysis with Linked Data", *Journal of the American Statistical Association*, 100(469), pp. 222–230.

Scheuren, F. and Winkler, W. (1993) "Regression Analysis of Data Files that are Computer Matched", *Survey Methodology*, 19, pp. 39–58.

Scheuren, F. and Winkler, W. (1997) "Regression Analysis of Data Files that are Computer Matched – Part II", *Survey Methodology*, 23, pp. 157–165.

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS	www.abs.gov.au

1352.0.55.098 • RESEARCH PAPER: GENERALISED LINEAR MODELS WITH PROBABILISTICALLY LINKED DATA (METHODOLOGY ADVISORY COMMITTEE) • November 2008

© Commonwealth of Australia 2009
Produced by the Australian Bureau of Statistics