Al-Kandari, Noriah M.; Lahiri, Partha

**Article**

PREDICTION OF A FUNCTION OF MISCLASSIFIED BINARY DATA

Statistics in Transition New Series

**Provided in Cooperation with:**
Polish Statistical Association


This Version is available at: http://hdl.handle.net/10419/207823

**Terms of use:**

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/
PREDICTION OF A FUNCTION OF MISCLASSIFIED BINARY DATA

Noriah M. Al-Kandari¹, Partha Lahiri²

ABSTRACT

We consider the problem of predicting a function of misclassified binary variables. We make an interesting observation that the naive predictor, which ignores the misclassification errors, is unbiased even if the total misclassification error is high as long as the probabilities of false positives and false negatives are identical. Other than this case, the bias of the naive predictor depends on the misclassification distribution and the magnitude of the bias can be high in certain cases. We correct the bias of the naive predictor using a double sampling idea where both inaccurate and accurate measurements are taken on the binary variable for all the units of a sample drawn from the original data using a probability sampling scheme. Using this additional information and design-based sample survey theory, we derive a bias-corrected predictor. We examine the cases where the new bias-corrected predictors can also improve over the naive predictor in terms of mean square error (MSE).

Key words: binary classification, double sampling, finite population sampling, misclassification, linkage error, sampling design.

1. Introduction

In many disciplines, misclassified binary data are frequently encountered. For example, in device testing, Zhong (2002) studied the specificity and sensitivity of an inaccurate diagnostic test along with a gold standard. Stamey et al. (2007) proposed a Bayesian estimation of an intervention effect with pre and post misclassified binomial data. Lyles et al. (2004) discussed single-armed studies with misclassification of a repeated binary outcome. In epidemiology and medical studies, there are plenty of examples of misclassified binary data. For example, in studying the relationship between low level radiation exposure and cancer death rate using the Cox proportional hazard model, Krewski et al. (2005) noted that misclassified binary data arise in form of imperfect linkages caused by the computerized record linkage method.

Bross (1954) was probably the first to observe that classical estimators of the odds ratio can be heavily biased if the misclassification error in binary data is ignored; see Goldberg (1975) for a follow-up study. Neter et al. (1965) noticed that

¹Department of Statistics and Operations Research, Kuwait University. E-mail: noriah@stat.kuniv.edu.
²Joint Program in Survey Methodology, University of Maryland. E-mail: plahiri@umd.edu.
the matching errors pose an obstacle to the usefulness and correct interpretation of record checks.

There are mainly two different approaches available to correct for the bias in statistical procedures that arise from misclassified binary data. The key ingredient in both the approaches is to use additional data to deal with the identifiability problem. The first approach, pioneered by Tenenbein (1970), employs a double sampling scheme in which a training data set is collected and the binary responses are measured by an accurate instrument (in the case of a random subsample from the original data) or by both an accurate instrument and the same inaccurate instrument used to collect the original data (in the case of an independent new sample). An accurate instrument results in error-free but expensive binary data. On the other hand, an inaccurate instrument results in misclassified but relatively less expensive binary data. Tenenbein’s idea is intuitive and uses both accurate and inaccurate procedures to yield not only model identifiability but also economical viability.

For the single proportion problem, when a training data is obtained using a double sampling scheme, Tenenbein (1970) proposed a maximum likelihood estimator and derived its asymptotic variance. Boese et al. (2006) constructed several likelihood-based confidence intervals for a proportion using data subject to only false positive misclassification. Rahardja and Zhou (2013) proposed a modification of the Wald test in presence of misclassified binary data and applied their test to traffic data. Rahardja and Yang (2015) constructed two likelihood-based confidence intervals for a binomial proportion parameter using a double-sampling scheme with misclassified binary data.

When an accurate instrument is unavailable or prohibitively expensive but certain data related to the cause of misclassification are available, one can develop an identifiable model in an attempt to correct for the misclassification bias in the estimators and predictors. For the single proportion problem using misclassified data with no training data, Gaba and Winkler (1992) and Viana et al. (1993) developed Bayesian approaches with highly informative priors. Bayesian inferences with informative priors were also developed for two-sample problems for two proportions. For example, see Evans et al. (1996) for risk difference (the difference of two proportions) and Gustafson et al. (2001) for odds ratios. Lahiri and Larsen (2005) used a mixture model to correct for the bias of the ordinary least square estimators of regression coefficients due to imperfect linkages.

In this paper, we assume the existence of a training sample such as the one proposed by Tenenbien (1970) and exploit a design-based sample survey approach to predict a function of misclassified binary data. Consider a set $U$ of $N$ units. For unit $i \in U$, we define a binary variable $\delta_i$ taking on values 0 and 1, and a $K \times 1$ vector of measurements $y_i = (y_{i1}, \ldots, y_{iK})'$. We consider a situation when we do not observe $\delta_i$, but instead observe a predictor $\hat{\delta}_i$ subject to a misclassification error $e_i = \hat{\delta}_i - \delta_i$ ($i \in U$). In this paper, we are interested in the prediction of $Y_\delta =$
\[ \sum_{i \in U} \delta y_i \text{ or a non-linear function of the components of } Y_\delta, \text{ say } f(Y_{\delta 1}, \cdots, Y_{\delta K}), \text{ where } Y_{\delta k} = \sum_{i=1}^{N} \delta y_{ik}, \text{ } (k=1, \cdots, K), \text{ based on data } \{(\hat{\delta}_i, y_i), \ i \in S \subseteq U\}.

A natural predictor of \( Y_\delta \) is given by \( \hat{Y}_\delta(S) = \sum_{i \in S} \hat{\delta}_i y_i \). If additional data that explain the mechanism for misclassification errors \( e_i \) are available, it is possible to correct \( \hat{Y}_\delta(S) \) for bias due to the misclassification errors. The misclassification errors could arise due to a variety of reasons. For example, the data set may be obtained by merging two or more data sets using a computerized record linkage method, which may introduce misclassification errors due to incorrect linkages. There are a large number of papers available in the literature that provide valid inferences under linkage errors when data on linkage error mechanism through matching weights are available; for a review of record linkage methodology, see Felligi and Sunter (1969) and Herzog et al. (2007). However, in this paper, we assume that we do not have any data that explain the misclassification error for all records in \( i \in S \). Thus, even for the special case when the misclassification errors is due to incorrect linkages, in this paper we deal with a situation that cannot be handled by a regular record linkage methodology such as the ones given in Scheuren and Winkler (1993) and Lahiri and Larsen (2005).

In Section 2.1, we first obtain the bias and mean squared error of the natural predictor for the scaler case \( K = 1 \) and then generalize to \( K \geq 1 \). The extent of the bias clearly depends on the misclassification error distribution. When the positive and negative misclassification errors are equally likely, \( \hat{Y}_\delta(S) \equiv Y_\delta \) turns out to be an unbiased predictor of \( Y_\delta \). This is an interesting observation. We propose a method for correcting the bias in the general case by drawing a probability sample and then obtaining the misclassification errors \( e_i \) for all the units in the sample. Using this extra information, we propose a bias-corrected predictor of \( Y_\delta \). We obtain an exact expression for the MSE of the proposed bias-corrected predictor that incorporates both the sampling and misclassification errors. We also propose an estimator of the MSE of the new predictor. In Section 2.2, we discuss the estimation of relative risk estimation as an illustration of the methodology proposed in Section 2.1. In Section 2.3, we evaluate the method proposed in Section 2.1 using a numerical example. In Section 3, we consider the case \( S \subseteq U \). Finally, some conclusions are presented in Section 4.

2. Prediction of \( Y_\delta \) when \( S = U \)

2.1. The Methodology

For the simplicity of exposition, we first consider the scaler case, i.e., \( K = 1 \). We define the bias and mean squared error (MSE) of \( \hat{Y}_\delta \) as follows:

\[
\text{Bias}_M(\hat{Y}_\delta) = \mathbb{E}_M(\hat{Y}_\delta - Y_\delta),
\]

\[
\text{MSE}_M(\hat{Y}_\delta) = \text{Var}_M(\hat{Y}_\delta - Y_\delta),
\]
Table 1. The probability distribution of the misclassification error \( e_i \) (probabilities are given within parenthesis)

<table>
<thead>
<tr>
<th>( \hat{\delta}_i )</th>
<th>( \delta_i )</th>
<th>( \hat{\delta}_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(p_{i00})</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>(p_{i10})</td>
<td>0</td>
</tr>
</tbody>
</table>

where \( E_M \) and \( \text{Var}_M \) denote the expectation and variance with respect to a misclassification model described by the two-way table given in Table 1.

For unit \( i \in U \), the table displays the misclassification error distribution, where \( p_{i11} + p_{i10} + p_{i01} + p_{i00} = 1 \). We call \( p_{i10} \) and \( p_{i01} \) false positive and false negative probabilities, respectively. We say that we have high, moderate and low linkage errors if \( p_{i10} + p_{i01} = p_i;T \) (say) is close to 1, 0.5 and 0, respectively.

**Theorem 1.** Under the misclassification model given in Table 1, we have

(i) \( \text{Bias}_M(\hat{Y}_\delta) = \sum_{i \in U} y_i p_{i:D} = Y_{pD} \) (say),

(ii) \( \text{MSE}_M(\hat{Y}_\delta) = \sum_{i \in U} [p_i;T - p_{i:D}^2] y_i^2 \),

where \( p_i;D = p_{i10} - p_{i01} \) and \( p_i;T = p_{i10} + p_{i01} \) (\( i \in U \)).

**Proof:** First note that \( Y_\delta - Y_\delta = \sum_{i \in U} y_i e_i = Y_e \) (say). Under the misclassification model, we have \( E_M(e_i) = p_{i:D} \) and \( E_M(e_i^2) = p_i;T \). Thus, part (i) follows immediately. To prove part (ii), using \( \text{Cov}_M(e_i, e_j) = 0 \) \((i \neq j \in U)\), we have

\[
\text{MSE}_M(Y_\delta) = E_M \left( \sum_{i \in U} y_i e_i \right)^2 - \left[ E_M \left( \sum_{i \in U} y_i e_i \right) \right]^2
\]

\[
= E_M \left( \sum_{i \in U} y_i^2 e_i^2 + \sum_{i \neq j} y_i y_j e_i e_j \right) - \left( \sum_{i \in U} y_i p_{i:D} \right)^2
\]

\[
= \sum_{i=1}^N y_i^2 p_i;T + \sum_{i \neq j} y_i y_j p_i;D p_j;D - \left( \sum_{i=1}^N y_i p_{i:D} \right)^2 .
\]

Part (ii) now follows using algebra.

Throughout the paper, we assume that we do not have any additional data that explain the misclassification errors \( e_i \) for all units in \( U \). Thus, we propose to draw a sample \( s_1 \) of size \( n \) from \( U \), using a probability sampling scheme. For each unit in the sample, we assume that we can obtain \( e_i \) with some extra effort. Let \( \pi_i = \Pr(s_1 \ni i) \) denote the first-order inclusion probability of unit \( i \in U \). We propose to
Theorem 2. Under the sampling design and misclassification model, we have

(i) Bias(\(\hat{Y}\)) = 0,
(ii) MSE(\(\hat{Y}\)) = \(\sum_{i \in U} \sum_{j > i \in U} (\pi_i \pi_j - \pi_{ij}) \psi_{ij}\),
(iii) \(E[\text{mse}(\hat{Y})] = \text{MSE}(\hat{Y})\),

where \(\pi_{ij} = \Pr(s_1 \ni \{ij\})\), the second-order inclusion probability, \(\psi_{ij} = \pi_i^{-2} y_i^2 p_{iT} + \pi_j^{-2} y_j^2 p_{jT} - 2(\pi_i \pi_j)^{-1} y_i y_j p_{iD} p_{jD}\), \(\text{mse}(\hat{Y}) = \sum_{i \in s_1} \sum_{j > i \in s_1} \pi_{ij}^{-1} (\pi_i \pi_j - \pi_{ij}) \left(\pi_i^{-1} y_i e_i - \pi_j^{-1} y_j e_j\right)^2\).

Proof: Let \(E_d\) and \(\text{Var}_d\) denote the expectation and variance with respect to the sample design. First note that \(E_d(\hat{Y}_{\pi^{-1}e}) = Y_e = Y_{\delta} - Y_{\delta}\), since \(\hat{Y}_{\pi^{-1}e}\) is the Horvitz-Thompson estimator of \(Y_e\). To prove part (i) of Theorem 2, note that

\[
\text{Bias}(\hat{Y}) = E(\hat{Y} - Y_{\delta}) = E(Y_{\delta} - \hat{Y}_{\pi^{-1}e} - Y_{\delta})
= E_M E_d(Y_{\delta} - \hat{Y}_{\pi^{-1}e} - Y_{\delta})
= E_M (Y_{\delta} - Y_e - Y_{\delta})
= E_M(0)
= 0,
\]

since \(Y_{\delta} - Y_{\delta} = 0\). To prove part (ii), we first apply the iterated formula for variance to obtain

\[
\text{MSE}(\hat{Y}) = \text{Var}(\hat{Y} - Y_{\delta})
= E_M \text{Var}_d(\hat{Y} - Y_{\delta}) + \text{Var}_M E_d(\hat{Y} - Y_{\delta})
= E_M \text{Var}_d(Y_{\delta} - \hat{Y}_{\pi^{-1}e} - Y_{\delta}) + \text{Var}_M E_d(Y_{\delta} - \hat{Y}_{\pi^{-1}e} - Y_{\delta})
= E_M \text{Var}_d(\hat{Y}_{\pi^{-1}e}) + \text{Var}_M (Y_{\delta} - Y_e - Y_{\delta})
= E_M \left[\sum_{i \in U} \sum_{j > i \in U} (\pi_i \pi_j - \pi_{ij}) y_i e_i y_j e_j / \pi_i \pi_j\right],
\]

since \(Y_{\delta} - Y_e - Y_{\delta} = 0\). Now, part (ii) follows using \(E_M(e_i) = p_{iD}\), \(E_M(e_i^2) = p_{iT}\), and \(\text{Cov}_M(e_i, e_j) = 0\, (i \neq j)\), and algebra. Part (iii) follows using the iterated formula for expectation and the design-unbiasedness of the well-known Yates-Grundy estimator, Yates and Grundy (1953).
2.2. An illustrative example: bias adjusted SMR and relative regression coefficients in the presence of linkage errors

We now turn our attention to the estimation of \( f(\mathbf{Y_\delta}) \), where \( \mathbf{Y_\delta} = (Y_{\delta 1}, \cdots, Y_{\delta K})' \). A natural estimator is given by \( f(\mathbf{Y_\delta}) \). Using the Taylor’s series argument, it can be shown that

\[
E_M \left[ f(\mathbf{Y_\delta}) - f(\mathbf{Y_\delta}) \right] = \left[ \nabla f(\mathbf{Y_\delta}) \right]' \cdot E_M (\mathbf{Y_\delta} - \mathbf{Y_\delta}) = \left[ \nabla f(\mathbf{Y_\delta}) \right]' \cdot \mathbf{Y_{PD}},
\]

where \( \nabla f(\mathbf{Y_\delta}) \) is the gradient of \( f(\mathbf{Y_\delta}) \), \( \mathbf{Y_{PD}} = (Y_{pd1}, \cdots, Y_{pdK})' \) and \( Y_{pdk} = \sum_{i \in U} P_i \cdot D_i Y_{ik} \) \((k = 1, \cdots, K)\). Thus, \( f(\mathbf{Y_\delta}) \) is biased for \( f(\mathbf{Y_\delta}) \). A bias-adjusted estimator is given by \( f(\hat{\mathbf{Y}}) \), where \( \hat{\mathbf{Y}} = (Y_{\delta 1} - \hat{Y}_{\pi^{-1}e1}, \cdots, Y_{\delta 1} - \hat{Y}_{\pi^{-1}eK})' \).

Theorem 3. Under the misclassification model, we have

\[
\text{MSE} \left[ f(\hat{\mathbf{Y}}) \right] \approx \left[ \nabla f(\mathbf{Y_\delta}) \right]' \cdot \Sigma \cdot \left[ \nabla f(\mathbf{Y_\delta}) \right],
\]

where

\[
\sigma_{kk} = E_M \left\{ \sum_{i \in U} \sum_{j > i \in U} (\pi_i \cdot \pi_j - \pi_{ij}) \left( \pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j \right)^2 \right\}
\]

\[
= \sum_{i \in U} \sum_{j > i \in U} (\pi_i \cdot \pi_j - \pi_{ij}) \psi_{ij;kk},
\]

with

\[
\psi_{ij;kk} = \pi_i^{-2} p_i;T Y_{ik}^2 + \pi_j^{-2} p_j;T Y_{jk}^2 - 2(\pi_i \cdot \pi_j)^{-1} p_i;DP_j;D Y_{ik} Y_{jk}.
\]

We propose to estimate \( \sigma_{kk} \) by \( \hat{\sigma}_{kk} = \sum_{i \in s_1} \sum_{j > i \in s_1} \pi_{ij}^{-1} (\pi_i \cdot \pi_j - \pi_{ij}) (\pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j)^2 \), \((k = 1, \cdots, K)\). Using the property of the Yates-Grundy estimator, we have \( E_d(\hat{\sigma}_{kk}) = \sum_{i \in U} \sum_{j > i \in U} (\pi_i \cdot \pi_j - \pi_{ij}) (\pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j)^2 \) and hence \( E(d(\hat{\sigma}_{kk}) = E_M E_d(\hat{\sigma}_{kk}) = \sigma_{kk} \). Thus, \( \hat{\sigma}_{kk} \) is an unbiased estimator of \( \sigma_{kk} \) \((k = 1, \cdots, K)\). Thus, an unbiased estimator of \( \sigma_{kl} \) is given by \( \hat{\sigma}_{kl} = (\hat{\sigma}_{k+l} - \hat{\sigma}_{kk} - \hat{\sigma}_{ll})/2 \), \((k \neq l, k, l = 1, \cdots, K)\). Thus, an approximately unbiased estimator of \( \text{MSE} \left[ f(\hat{\mathbf{Y}}) \right] \) is given by

\[
\text{mse} \left[ f(\hat{\mathbf{Y}}) \right] = \left[ \nabla f(\hat{\mathbf{Y}}) \right]' \cdot \hat{\Sigma} \cdot \left[ \nabla f(\hat{\mathbf{Y}}) \right].
\]

2.2. An illustrative example: bias adjusted SMR and relative regression coefficients in the presence of linkage errors

There has been an increasing use of computerized record linkage (CRL) method in various studies such as historical cohort mortality studies, cancer studies, political
studies and crime studies, in several countries, Howe (1985, 1998), Bennell et al. (2012), Giraud-Carrier et al. (2015). With very little effort, the method enables us to collect a large amount of data by linking records of human exposure to environmental hazards with records on health status. Since CRL utilizes already existing databases, it saves a substantial amount of money to collect new data. Various government agencies have developed sophisticated software to implement CRL, usually attaching weights reflecting the likelihood of a match to pairs of records.

Fair (1989) listed a number of health studies where environmental exposure data were linked to the Canadian Mortality Data Base (CMDB). Krewski et al. (2005) provided an example where National Dose Registry (NDR) of Canada has been linked to CMDB in order to study the associations between excess mortality due to cancer and occupational exposure to low levels of ionizing radiation. In Beauchamp et al. (2011), a sample of 2000 participants from a cohort study was linked to a statewide hospitalisations dataset in Victoria, Australia using the national health insurance (Medicare) number and demographic data as identifying variables. Kabudula et al. (2014) applied deterministic and probabilistic record linkage approaches to mortality records from 2006 to 2009 from the Agincourt Health and Demographic Surveillance Systems (HDSS) to those in the national civil registration (CR) in South Africa.

In a cohort mortality study, CRL method introduces two types of linkage errors. The Type I linkage error (usually called a false positive) occurs when a member of the cohort who is alive is incorrectly identified as dead. The Type II (or a false negative) error occurs when a member of the cohort who is dead is incorrectly identified as alive. Krewski et al. (2005) investigated the impact of linkage errors on estimates of epidemiological indicators of risk such as standardized mortality ratio (SMR) and the parameters of relative risk regression model. Their analytical and simulation results indicate that these indicators are, in general, subject to biases and additional variabilities in the presence of linkage errors.

In this subsection, we use the notation used in Krewski et al. (2005). In the analysis of cohort studies, mortality is usually characterized by the hazard function which relates death rate as a function of time. Denoting $T$ the time of death, the hazard function at time $u$ is defined as

$$\lambda(u) = \lim_{\Delta u \downarrow 0} \frac{\Pr\{u \leq T < u + \Delta u \mid T \geq u\}}{\Delta u}.$$ 

The corresponding survival function and the probability density function are given by $S(u) = \exp\left(-\int_0^u \lambda(t)\,dt\right)$ and $f(u) = \lambda(u)S(u)$, respectively. Let $\lambda_i(u)$ and $Z_i(u)$ be the hazard function for a specific cause of death and the value of the vector of covariates at time $u$ for the $i$th member of the cohort, $i = 1, \cdots, N$. The relative risk regression is then described as

$$\lambda_i(u) = \lambda^*(u) \gamma(Z_i(u)')\beta,$$
where \( \lambda^*(u) \), value of \( \lambda_i(u) \) when \( \beta = 0 \), is known as the baseline hazard and \( \gamma \) is a positive function of the covariates and \( \beta \).

Let \( t_i^0 \) and \( t_i^1 \) be the age at the time of entry into the study and time of loss to follow up for the \( i \)th member of the cohort \( i = 1, \ldots, N \). Let \( \delta_i = 1 \) if the \( i \)th individual has died at the time of loss to follow-up and \( \delta_i = 0 \) otherwise. The likelihood based on the relative risk regression model is given by \( L = \prod_{i=1}^{N} f(t_i) \delta_i S(t_i)^{1-\delta_i} \). The corresponding log-likelihood is given by

\[
\log L = \sum_{i=1}^{N} \left\{ \delta_i \log \left( \gamma(z_i(t_i)^\gamma) \right) - \int_{t_i^0}^{t_i^1} \gamma(z_i(u)^\gamma) \lambda^*(u) \, du \right\}.
\]

The maximum likelihood estimate \( \hat{\beta} \) of \( \beta \) is obtained as a solution of \( \frac{\partial \log L}{\partial \beta} = 0 \).

Note that when \( z_i(u) = 1 \), \( \gamma(\hat{\beta}) \) reduces to the standardized mortality rate given by \( SMR = OBS/EXP \), where \( OBS = \sum_{i=1}^{N} \delta_i = \) observed number of death before time to follow-up and \( EXP = \sum_{i=1}^{N} e_i = \) expected number of deaths, with \( e_i = \int_{t_i^0}^{t_i^1} \lambda^*(u) \, du \).

Due to time-dependent covariates \( z_i(u) \), the integral must be re-evaluated at each iteration of the maximization process. Thus, it is very computer-intensive, specially when the cohort size is large. Breslow et al. (1983) simplified the likelihood by assuming \( z_i(u) = z_j \) whenever the \( i \)th cohort member passes through the state \( S_j \) \((j = 1, \ldots, J)\). The states can be defined by cross-classification of the covariates of interest. In this case, the log-likelihood can be written as

\[
\log L = \sum_{i=1}^{N} \left\{ d_j \log(\gamma(z_j^\gamma)) - \gamma(z_j^\gamma) e_j \right\},
\]

where \( e_j = \sum_{i=1}^{N} \int_{z_i(u) \in S_j} \lambda^*(u) \, du \) is the contribution to the expected number of deaths from all person-years of observation in the state \( S_j \) and \( d_j \) is the total number of death in that state. The maximum likelihood estimate of \( \beta \) is then obtained as a solution to the following equation:

\[
\sum_{i=1}^{J} \frac{\partial \Lambda_j(\beta)}{\partial \beta} \left\{ d_j - \exp(\Lambda_j(\beta)) e_j \right\} = 0,
\]

where \( \Lambda_j(\beta) = \log(\gamma(z_j^\gamma)) \).

Note that the results of Section 2.1 are valid for each state \( S_j, j = 1, \ldots, J \). We introduce an additional suffix \( j \) to indicate that the parameter or estimator refers to the state \( S_j, j = 1, \ldots, J \). For example, we shall have \( Y_j \) in place of \( Y \). Note that with \( Y_{1ij} = 1 \) and \( Y_{0ij} = 0 \), \( Y_j \) reduces to \( d_j \) of Krewski et al. (2005). We have \( Y_j = e_j \)
if we choose $Y_{1ij} = \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du$ and $Y_{0ij} = \int_{t_{ij}^2}^{t_{ij}^3} \lambda^*(u)du$. Consider $f(X, Y) = X/Y$ with $X = \sum_{j=1}^{f} d_j = d$ and $Y = \sum_{j=1}^{e} e_j = e$. With these choices, $f(X, Y) = SMR$.

Suppose a sample of size $n_j$ is selected using a probability sampling scheme from each state $S_j, j = 1, \cdots, J$. Let $\pi_{i(j)}$ and $\pi_{ik(j)}$ denote the first order and second order inclusion probabilities in state $S_j, j = 1, \cdots, J$. Using the results of Section 2.1, SMR can be adjusted for its bias due to linkage errors. The bias-corrected SMR is given by $\text{SMR} = (\hat{d}/\hat{e})$, where $\hat{d} = \sum_{j=1}^{J} \hat{d}_j, \hat{e} = \sum_{j=1}^{J} \hat{e}_j, \Delta \hat{d}_j = \sum_{i\in s} \pi_{i(j)}^{-1} \Delta \delta_{i(j)}, \Delta \hat{e}_j = \sum_{i\in s} \pi_{i(j)}^{-1} \Delta \delta_{i(j)}, \hat{d}_j = d_j - \Delta \hat{d}_j \text{ and } \hat{e}_j = e_j - \Delta \hat{e}_j$.

An application of Theorem 3 provides us with an estimator of the variance of $\text{SMR} - \text{SMR}$. It is given by

$$\text{var}(\hat{d} - d) + \text{var}(\hat{e} - e) + 2\text{cov}(\hat{d} - d, \hat{e} - e),$$

where $\text{var}(\hat{d} - d) = \sum_{j=1}^{J} v_j$ with $v_j = \sum_{\pi_{ik(j)}^{-1} (\pi_{i(j)} \pi_{ik(j)} - \pi_{ik(j)})(\pi_{i(j)} \pi_{ik(j)} - \pi_{ik(j)})^2}. We can define $\text{var}(\hat{e} - e)$ and $\text{cov}(\hat{d} - d, \hat{e} - e)$ similarly.

Let us now consider estimation of the regression coefficient $\beta$ in the relative risk regression model. First, we propose to adjust the log-likelihood given in Krewski et al. (2005). We shall find the estimator of $\beta$ as a solution, say, $\hat{\beta}$, of the following score function

$$Q = Q(\beta; (\hat{d}_j, \hat{e}_j), j = 1, \cdots, J) = \sum_{j=1}^{J} \frac{\partial \Lambda_j(\beta)}{\partial \beta} \{ \hat{d}_j - \exp(\Lambda_j(\beta)) \hat{e}_j \} = 0.$$

Since $E \{ Q(\beta; (\hat{d}_j, \hat{e}_j), j = 1, \cdots, J) \} - Q(\beta; (d_j, e_j), j = 1, \cdots, J) = 0$, we can expect $\hat{\beta}$ to perform well. The covariance matrix of the proposed estimator, $\hat{\beta}$, can be estimated by $\left[ \frac{\partial Q}{\partial \beta} \right]^2 \beta = \hat{\beta}$.

### 2.3. Evaluation

In this subsection, we consider the prediction of $Y = \sum_{j=1}^{N} \delta_i$. A naive predictor of $Y$ that ignores the misclassification errors is given by $\hat{Y} = \sum_{i=1}^{N} \hat{\delta}_i$. Under the misclassification error model of Section 2.1 with $p_{ikl} = p_{kl}, (i = 1, \cdots, N; k, l = 0, 1)$, the misclassification error bias of $\hat{Y}$ can be obtained as

$$\text{Bias}_M(\hat{Y}) = N(p_{10} - p_{01}) = Nb,$$

where $b = p_{10} - p_{01}$. Thus, $\hat{Y}$ is positively (negatively) biased if the false positive probability is more (less) than the false negative probability. It is interesting to
note that there is no bias in $\hat{Y}$ due to misclassification even if there is a large misclassification error as long as the false positive and false negative probabilities are identical.

Using the theory developed in Section 2.1, we can correct the misclassification bias of $\hat{Y}$ by drawing a simple random sample (SRS) of size $n$ from $U$ and determining the status of each sampled unit for misclassification error. We propose the misclassification bias-corrected predictor of $Y$ as

$$\tilde{Y} = \hat{Y} - \hat{\text{Bias}},$$

where

$$\hat{\text{Bias}} = \frac{N}{n} \sum_{i \in s} e_i.$$ 

Evidently, the proposed predictor is unbiased with respect to the combined distribution of misclassification and sampling errors. But, the proposed method introduces some costs. Also, the bias correction is expected to increase the variability. Thus, we study how the above bias-correction affects the mean squared error that incorporates both the misclassification and sampling errors. We define the total mean square error of a predictor $\hat{Y}$ of $Y$ as

$$\text{MSE}(\hat{Y}) = \mathbb{E}(\hat{Y} - Y)^2,$$

where the expectation $\mathbb{E}$ is with respect to the SRS and the multinomial misclassification error model. After considerable algebra, we obtain

$$\text{MSE}(\hat{Y}) = N(a + Nb^2),$$

$$\text{MSE}(\tilde{Y}) = \frac{N^2}{n} (1 - f)a,$$

where $a = p_{10}(1 - p_{10}) + p_{01}(1 - p_{01}) + 2p_{10}p_{01}$ and $f = \frac{n}{N}$.

We define the relative improvement in MSE (MSERI) as follows:

$$\text{MSERI} = \frac{\text{MSE}(\hat{Y}) - \text{MSE}(\tilde{Y})}{\text{MSE}(\hat{Y})}.$$ 

It can be shown that

$$\text{MSERI} = \left[\frac{n}{(1 - f)} \times \frac{b^2}{a}\right] - \frac{1 - 2f}{1 - f},$$

where $a > 0$, $0 < f < 1$ and $0 < b^2 < 1$. In order for the bias-corrected predictor $\tilde{Y}$ to improve on the naive predictor $\hat{Y}$ in terms of MSE, $b$ must satisfy one of the following two conditions:

$$b > \sqrt{\frac{1}{n} - \frac{2}{N}}a = b_2$$

(2)
or

\[ b < - \sqrt{\frac{1}{n} - \frac{2}{N}} a = b_1. \] (3)

In many situations, the sampling fraction \( f \) is negligible in which case \( \text{MSREI} \approx n \frac{b^2}{2} a \).

Define the relative bias (RB) as

\[ \text{RB}(\hat{Y}) = \frac{Nb}{\text{E}(Y)} = \frac{Nb}{N(p_{11} + p_{01})} = \frac{p_{10} - p_{01}}{p_{11} + p_{01}}. \]

Since both MSRI and RB depend on \( N \) only through \( f \), we arbitrarily fix \( N \) and vary \( f \). For a numerical comparison, we fix \( N = 100 \). Table 2 displays \( \text{RB}(\hat{Y}) \), \( \text{MSE}(\hat{Y}) \), \( \text{MSE}(\hat{Y}) \) and MSERI for \( f = 0.05, 0.10 \) and two levels of misclassification errors (LE): High (H) and Moderate (M). Table 3 reports the results for low misclassification errors (L). First of all, we notice that the relative bias of \( \hat{Y} \) depends on the configurations of false positive \( (p_{10}) \) and false negative \( (p_{01}) \) probabilities. Clearly, a high relative bias in \( \hat{Y} \) is possible. In this case, the bias-corrected estimator \( \tilde{Y} \) can have substantially smaller MSE than \( \hat{Y} \) even when the sampling fraction \( f \) is small. When the relative bias in \( \hat{Y} \) is small, one needs much higher sampling fraction for \( \tilde{Y} \) to improve on \( \hat{Y} \) in terms of mean squared error.

**Table 2.** RB(\( \hat{Y}^* \)), MSE(\( \hat{Y}^* \)), MSE(\( \hat{Y} \)) and MSERI for high and medium misclassification errors

<table>
<thead>
<tr>
<th>( n )</th>
<th>( f )</th>
<th>LE</th>
<th>( p_{10} )</th>
<th>( p_{01} )</th>
<th>( b_2 )</th>
<th>( b )</th>
<th>( b_1 )</th>
<th>RB(( \hat{Y}^* ))</th>
<th>MSE(( \hat{Y}^* ))</th>
<th>MSE(( \hat{Y} ))</th>
<th>MSERI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>H</td>
<td>0.80</td>
<td>0.15</td>
<td>0.31</td>
<td>0.65</td>
<td>-0.31</td>
<td>3.71</td>
<td>4277.75</td>
<td>1002.25</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.75</td>
<td>0.15</td>
<td>0.31</td>
<td>0.60</td>
<td>-0.31</td>
<td>3.00</td>
<td>3654.00</td>
<td>1026.00</td>
<td>2.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.60</td>
<td>0.10</td>
<td>0.28</td>
<td>0.50</td>
<td>-0.28</td>
<td>3.33</td>
<td>2545.00</td>
<td>855.00</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.60</td>
<td>0.28</td>
<td>-0.50</td>
<td>-0.28</td>
<td>-0.77</td>
<td>2545.00</td>
<td>855.00</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.20</td>
<td>0.65</td>
<td>0.34</td>
<td>-0.45</td>
<td>-0.34</td>
<td>-0.60</td>
<td>2089.75</td>
<td>1230.25</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.50</td>
<td>0.35</td>
<td>0.39</td>
<td>0.15</td>
<td>-0.39</td>
<td>0.38</td>
<td>307.75</td>
<td>1572.25</td>
<td>-0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.30</td>
<td>0.60</td>
<td>0.38</td>
<td>-0.30</td>
<td>-0.38</td>
<td>-0.49</td>
<td>981.00</td>
<td>1539.00</td>
<td>-0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
<td>0.25</td>
<td>0.34</td>
<td>0.20</td>
<td>-0.34</td>
<td>0.44</td>
<td>466.00</td>
<td>1254.00</td>
<td>-0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.40</td>
<td>0.27</td>
<td>-0.30</td>
<td>-0.27</td>
<td>-0.43</td>
<td>941.00</td>
<td>779.00</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.35</td>
<td>0.26</td>
<td>-0.25</td>
<td>-0.26</td>
<td>-0.45</td>
<td>663.75</td>
<td>736.25</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.35</td>
<td>0.20</td>
<td>0.31</td>
<td>0.15</td>
<td>-0.31</td>
<td>0.27</td>
<td>277.75</td>
<td>1002.25</td>
<td>-0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
<td>0.35</td>
<td>0.29</td>
<td>-0.20</td>
<td>-0.29</td>
<td>-0.33</td>
<td>446.00</td>
<td>874.00</td>
<td>-0.49</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>H</td>
<td>0.80</td>
<td>0.15</td>
<td>0.21</td>
<td>0.65</td>
<td>-0.21</td>
<td>3.71</td>
<td>4277.75</td>
<td>474.75</td>
<td>8.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.75</td>
<td>0.15</td>
<td>0.21</td>
<td>0.60</td>
<td>-0.21</td>
<td>3.00</td>
<td>3654.00</td>
<td>486.00</td>
<td>6.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.60</td>
<td>0.10</td>
<td>0.19</td>
<td>0.50</td>
<td>-0.19</td>
<td>3.33</td>
<td>2545.00</td>
<td>405.00</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.60</td>
<td>0.19</td>
<td>-0.50</td>
<td>-0.19</td>
<td>-0.77</td>
<td>2545.00</td>
<td>405.00</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.20</td>
<td>0.65</td>
<td>0.23</td>
<td>-0.45</td>
<td>-0.23</td>
<td>-0.60</td>
<td>2089.75</td>
<td>582.75</td>
<td>2.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.50</td>
<td>0.35</td>
<td>0.26</td>
<td>0.15</td>
<td>-0.26</td>
<td>0.38</td>
<td>307.75</td>
<td>744.75</td>
<td>-0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.30</td>
<td>0.60</td>
<td>0.25</td>
<td>-0.30</td>
<td>-0.25</td>
<td>-0.49</td>
<td>981.00</td>
<td>729.00</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
<td>0.25</td>
<td>0.23</td>
<td>0.20</td>
<td>-0.23</td>
<td>0.44</td>
<td>466.00</td>
<td>594.00</td>
<td>-0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.40</td>
<td>0.18</td>
<td>-0.30</td>
<td>-0.18</td>
<td>-0.43</td>
<td>941.00</td>
<td>369.00</td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.35</td>
<td>0.18</td>
<td>-0.25</td>
<td>-0.18</td>
<td>-0.45</td>
<td>663.75</td>
<td>348.75</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.35</td>
<td>0.20</td>
<td>0.21</td>
<td>0.15</td>
<td>-0.21</td>
<td>0.27</td>
<td>277.75</td>
<td>474.75</td>
<td>-0.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
<td>0.35</td>
<td>0.19</td>
<td>-0.20</td>
<td>-0.19</td>
<td>-0.33</td>
<td>446.00</td>
<td>414.00</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Table 3. \( \text{RB}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}) \) and MSERI for low misclassification errors

<table>
<thead>
<tr>
<th>( n )</th>
<th>( f )</th>
<th>( p_{10} )</th>
<th>( p_{01} )</th>
<th>( b_2 )</th>
<th>( b )</th>
<th>( b_1 )</th>
<th>( \text{RB}(\hat{Y}^*) )</th>
<th>( \text{MSE}(\hat{Y}^*) )</th>
<th>( \text{MSE}(\hat{Y}) )</th>
<th>( \text{MSERI} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.07</td>
<td>0.03</td>
<td>0.13</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.05</td>
<td>25.84</td>
<td>186.96</td>
<td>-0.86</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.13</td>
<td>0.19</td>
<td>-0.06</td>
<td>-0.19</td>
<td>-0.11</td>
<td>55.64</td>
<td>373.16</td>
<td>-0.86</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>0.07</td>
<td>0.03</td>
<td>0.09</td>
<td>0.04</td>
<td>-0.09</td>
<td>0.05</td>
<td>25.84</td>
<td>88.56</td>
<td>-0.71</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.13</td>
<td>0.13</td>
<td>-0.06</td>
<td>-0.13</td>
<td>-0.11</td>
<td>55.64</td>
<td>176.76</td>
<td>-0.69</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.15</td>
<td>0.07</td>
<td>0.03</td>
<td>0.07</td>
<td>0.04</td>
<td>-0.07</td>
<td>0.05</td>
<td>25.84</td>
<td>55.76</td>
<td>-0.54</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.13</td>
<td>0.10</td>
<td>-0.06</td>
<td>-0.10</td>
<td>-0.11</td>
<td>55.64</td>
<td>111.29</td>
<td>-0.50</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.30</td>
<td>0.07</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.04</td>
<td>0.05</td>
<td>25.84</td>
<td>22.96</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.13</td>
<td>0.05</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.11</td>
<td>55.64</td>
<td>48.83</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>

Suppose \( p_{10} = p_{01} \), then \( b = 0 \). Hence, \( \text{RB}(\hat{Y}) \) will always be zero. Also, the MSERI will be a function of the sampling fraction \( f \) as shown below

\[
\text{MSERI} = -\frac{1 - 2f}{1 - f}.
\]  

Figure 1 displays the MSERI for different choices of \( f \).

![Figure 1](image_url)

**Figure 1.** MSE Relative Improvement when \( p_{10} = p_{01} \)

3. **Prediction of \( Y_\delta \) when \( S \subset U \)**

In this subsection, we consider prediction of a finite population total \( Y = \sum_{i=1}^{N} \delta_i \) using a sample \((s_1)\) of size \( n_1 \) drawn by the *probability proportional to size with*
replacement (PPSWR) sampling design, where size is defined as:

\[ \phi_i = \frac{x_i}{\sum_{i=1}^{N} x_i}, \]

where \( x_i > 0 \) known, \( i \in U \). A natural estimator of \( Y \) is given by \( \hat{Y}^* = \sum_{i \in s_1} \omega_i \delta_i^* \), where \( \delta_i^* \) is observed value of \( \delta_i \) with misclassification error and \( \omega_i = \frac{1}{n_1 \phi_i}, i \in s_1 \).

Under the PPSWR sample design and the multinomial classification model of Section 2.3, a heavy algebra yields

\[ \text{Bias}(\hat{Y}^*) = Nb \]

and

\[ \text{MSE}(\hat{Y}^*) = \frac{1}{n_1} \sum_{i=1}^{N} \frac{1}{\phi_i} p_{1+} - \frac{N}{n_1} p_{1+} \{1 + (N-1)p_{1+}\} + N[a + b^2(N-1)], \]

where \( p_{1+} = p_{11} + p_{10} \) and \( p_{1+} = p_{11} + p_{01} \).

In order to correct the bias of the predictor \( \hat{Y}^* \), a second sample, say \( s_2 \), of size \( n_2 \) is drawn from the first sample \( s_1 \) using a SRS design and true \( \delta_i \) is measured without misclassification error. We define a bias-corrected predictor \( \hat{Y} \) as follows:

\[ \hat{Y} = \hat{Y}^* - \hat{\text{Bias}}, \]

where

\[ \hat{\text{Bias}} = \frac{n_1}{n_2} \sum_{i \in s_2} \omega_i (\delta_i^* - \delta_i). \]

It is easy to show that \( \hat{Y} \) is an unbiased predictor of \( Y \). Using heavy algebra, the exact MSE of \( \hat{Y} \) under different sources of uncertainty is obtained as follows:

\[ \text{MSE}(\hat{Y}) = \frac{1}{n_1} \left\{ -2Np_{11} + \sum_{i=1}^{N} \frac{1}{\phi_i} [p_{1+} + a \frac{1}{f_2} (1 - f_2)] \right. \]

\[ \left. - Np_{1+} \frac{1}{f_2} (1 - f_2) [1 + p_{1+}(N-1)] \right\}, \]

where \( f_2 = \frac{n_2}{n_1} \).

It can be shown that

\[ \text{MSE}(\hat{Y}^*) - \text{MSE}(\hat{Y}) = \]

\[ \frac{1}{n_1} \left\{ \sum_{i=1}^{N} \frac{1}{\phi_i} \left[ b - a \frac{(1 - f_2)}{f_2} \right] - Np_{1+} [1 + p_{1+}(N-1)] - \frac{1}{f_2} (1 - f_2) [1 + p_{1+}(N-1)] \right\} \]

\[ + 2Np_{11} + N[a + b^2(N-1)]. \] (5)
The expression for MSERI is obtained by dividing Eq.(5) by \( \text{MSE}(\hat{Y}) \). Also, we can show that

\[
\text{RB}(\hat{Y}^*) = \frac{Nb}{E(Y)} = \frac{Nb}{N(p_{11} + p_{01})} = \frac{b}{p_{11} + p_{01}}.
\]

Tables 4, 5 and 6 display \( \text{RB}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}) \) and MSERI for high, medium and low misclassification probabilities given by

\[
P_H = \begin{pmatrix}
0.10 \\
0.80 - p_{10} \\
0.10
\end{pmatrix},
P_M = \begin{pmatrix}
0.25 \\
0.50 - p_{10} \\
0.25
\end{pmatrix},
P_L = \begin{pmatrix}
0.40 \\
0.20 - p_{10} \\
0.40
\end{pmatrix},
\]

respectively.

Different configurations of the high, medium and low misclassification errors are considered by varying the false positive probability \( p_{10} \). For each case, we consider \( f_2 = 0.2, N = 1000 \) when \( n_1 = 300 \), and \( N = 100,000 \) when \( n_1 = 10,000 \).

### Table 4. \( \text{RB}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}) \) and MSERI for high misclassification errors

<table>
<thead>
<tr>
<th>( n_1 )</th>
<th>( p_{10} )</th>
<th>Bias(( \hat{Y}^* ))</th>
<th>( \text{RB}(\hat{Y}^*)(%) )</th>
<th>MSE(( \hat{Y}^* ))</th>
<th>MSE(( \hat{Y} ))</th>
<th>MSERI(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>0.10</td>
<td>-600</td>
<td>-75.00</td>
<td>387092.90</td>
<td>535198.40</td>
<td>-27.67</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>-400</td>
<td>-57.14</td>
<td>200519.40</td>
<td>521137.70</td>
<td>-61.52</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>-200</td>
<td>-33.33</td>
<td>93799.32</td>
<td>506810.60</td>
<td>-81.49</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0</td>
<td>0.00</td>
<td>66932.65</td>
<td>492217.10</td>
<td>-86.40</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>200</td>
<td>50.00</td>
<td>119919.40</td>
<td>477357.20</td>
<td>-74.88</td>
</tr>
<tr>
<td></td>
<td>0.60</td>
<td>400</td>
<td>133.33</td>
<td>252759.50</td>
<td>462230.90</td>
<td>-45.32</td>
</tr>
<tr>
<td></td>
<td>0.70</td>
<td>600</td>
<td>300.00</td>
<td>465453.00</td>
<td>446838.20</td>
<td>-4.17</td>
</tr>
</tbody>
</table>

### Table 5. \( \text{RB}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}) \) and MSERI for medium misclassification errors

<table>
<thead>
<tr>
<th>( n_1 )</th>
<th>( p_{10} )</th>
<th>Bias(( \hat{Y}^* ))</th>
<th>( \text{RB}(\hat{Y}^*)(%) )</th>
<th>MSE(( \hat{Y}^* ))</th>
<th>MSE(( \hat{Y} ))</th>
<th>MSERI(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>0.10</td>
<td>-60000</td>
<td>-75.00</td>
<td>3615422483</td>
<td>308209682</td>
<td>1073.04</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>-40000</td>
<td>-57.14</td>
<td>1623101725</td>
<td>300300437</td>
<td>440.50</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>-20000</td>
<td>-33.33</td>
<td>430752967</td>
<td>292311194</td>
<td>47.36</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0</td>
<td>0.00</td>
<td>38376209</td>
<td>284241951</td>
<td>-86.50</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>20000</td>
<td>50.00</td>
<td>445971451</td>
<td>276092709</td>
<td>61.53</td>
</tr>
<tr>
<td></td>
<td>0.60</td>
<td>40000</td>
<td>133.33</td>
<td>1653538694</td>
<td>267863468</td>
<td>517.31</td>
</tr>
<tr>
<td></td>
<td>0.70</td>
<td>60000</td>
<td>300.00</td>
<td>3661077936</td>
<td>259554228</td>
<td>1310.53</td>
</tr>
</tbody>
</table>

### Table 6. \( \text{RB}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}^*) \), \( \text{MSE}(\hat{Y}) \) and MSERI for low misclassification errors

<table>
<thead>
<tr>
<th>( n_1 )</th>
<th>( p_{10} )</th>
<th>Bias(( \hat{Y}^* ))</th>
<th>( \text{RB}(\hat{Y}^*)(%) )</th>
<th>MSE(( \hat{Y}^* ))</th>
<th>MSE(( \hat{Y} ))</th>
<th>MSERI(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>0.05</td>
<td>-400</td>
<td>-57.14</td>
<td>1623071725</td>
<td>207789527</td>
<td>681.11</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>-20000</td>
<td>-33.33</td>
<td>430722967</td>
<td>316635.8</td>
<td>-62.22</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0</td>
<td>0.00</td>
<td>38346209</td>
<td>191731041</td>
<td>-80</td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>20000</td>
<td>50.00</td>
<td>445941451</td>
<td>183581799</td>
<td>842.96</td>
</tr>
</tbody>
</table>
Table 6. RB($\hat{Y}^*$), MSE($\hat{Y}^*$), MSE($\hat{Y}$) and MSERI for low misclassification errors

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$p_{10}$</th>
<th>Bias($\hat{Y}^*$)</th>
<th>RB($\hat{Y}^*$)(%)</th>
<th>MSE($\hat{Y}^*$)</th>
<th>MSE($\hat{Y}$)</th>
<th>MSERI(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>-180</td>
<td>-30.51</td>
<td>86919.25</td>
<td>183920.5</td>
<td>-52.74</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>-100</td>
<td>-18.18</td>
<td>69784.31</td>
<td>178104.4</td>
<td>-60.82</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>0.10</td>
<td>0</td>
<td>66332.65</td>
<td>170774.4</td>
<td>-61.16</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>100</td>
<td>22.22</td>
<td>82844.34</td>
<td>163377.7</td>
<td>-49.29</td>
<td></td>
</tr>
<tr>
<td>0.199</td>
<td>198</td>
<td>49.38</td>
<td>118394.2</td>
<td>156064.4</td>
<td>-24.14</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>-10000</td>
<td>-30.51</td>
<td>355456551</td>
<td>106486050</td>
<td>233.81</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>-10000</td>
<td>-18.18</td>
<td>134508088</td>
<td>103264753</td>
<td>30.26</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>0.10</td>
<td>0</td>
<td>38316209</td>
<td>99220131</td>
<td>-61.38</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>10000</td>
<td>22.22</td>
<td>142117330</td>
<td>95155510</td>
<td>49.35</td>
<td></td>
</tr>
<tr>
<td>0.199</td>
<td>19800</td>
<td>49.38</td>
<td>437875637</td>
<td>91152778</td>
<td>380.38</td>
<td></td>
</tr>
</tbody>
</table>

The results for MSE improvement achieved by the bias-corrected estimator over the naive estimator for different situations are plotted in Figures 2 and 3.

![Figure 2](image1.png)

(a) High Misclassification Error

![Figure 2](image2.png)

(b) Medium Misclassification Error

![Figure 2](image3.png)

(c) Low Misclassification Error

**Figure 2.** MSE Relative Improvement for $f_2 = 0.2$, $N = 1000$ and $n_1 = 100, 200, 300$
Figure 3. MSE Relative Improvement for $f_2 = 0.2$, $N = 10,000$ and $n_1 = 10,000$, 20,000, 30,000

4. Conclusions

Our research shows that it is possible to correct bias due to misclassification in predictors by drawing a probability sample from the original data, determining the status of misclassification error for the sample and then applying the standard sample survey method. The bias-correction increases variance in the predictor, which impacts the mean squared error. The improvement depends on the distribution of the misclassification error and the sampling fraction in the drawn sample. If additional data that generates the misclassification error are available as in the record linkage literature, it may be possible to improve on the proposed method, we plan to investigate this direction in the future.
5. Acknowledgements

We are also grateful to an anonymous referee for making a number of constructive suggestions, which led to a significant improvement of our paper.

REFERENCES


