Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16 No. 1, 2018 (New Series), pp 105-122

Regression Analysis of Binary Data from Complex Surveys with Misclassification in Ordinal Covariates

Zhijian Chen[1], Harold Mantel[2], Changbao Wu[3] and Grace Y. Yi[3]
[1]*Alibaba, Hongzhou, China*
[2] *Statistics Canada, Ottawa, Canada*
[3] *Department of Statistics and Actuarial Science, University of Waterloo, Canada*

Received: December 09, 2017; Revised: December 28, 2017; Accepted: January 15, 2018

Abstract

We develop inferential procedures for logistic regression analysis for a binary response using complex survey data when an ordinal covariate is subject to misclassification. We propose a survey weighted estimating equation procedure based on the expectation correction method. Models for the misclassified ordinal covariate are presented, and issues with estimation of the parameters of the main analysis model and the measurement error model are discussed. Results from a simulation study and an application to the data set from the Canadian Community Health Survey are presented.

Key words: Bootstrap variance estimation; Design-based inference; Logistic regression analysis; Measurement error model.

1. Introduction

Survey sampling has been a widely used method for collecting data. Auxiliary information is often collected and used for improving the estimation of descriptive finite population quantities for particular variables of interest. There exists a rich literature on model-assisted inference for finite populations; see, for instance, the model-calibration approach for efficient use of auxiliary information (Wu and Sitter, 2001; Wu, 2003). Another important development in the past three decades has been the analytic use of survey data to study the relationship between a response variable and auxiliary variables in the target population. Using surveys to answer scientific questions is common in many fields including social science research and medical studies.

Design-based estimating equations approaches have gained increased popularity among users of complex survey data. Estimating functions are motivated by the inferential problems for superpopulation model parameters, and the finite population parameters are defined as the solution to

Corresponding Authors: Changbao Wu and Grace Y. Yi
Email: cbwu@uwaterloo.ca and yyi@uwaterloo.ca

the so-called census estimating equations. The inferential procedures are built based on the survey weighted estimating equations. The general framework was first proposed by Binder (1983), and then formally developed by Godambe and Thompson (1986), with subsequent work by Binder and Patak (1994). Binder and Roberts (2009) contains a detailed discussion on design and model-based inferences for model parameters.

While many methods are available for handling various types of survey data, research gaps still exist. Most commonly used methods are developed under the assumption that survey data are precisely collected. Measurement error, however, arises frequently during the course of the data collection. Chen, Yi and Wu (2011, 2014) developed marginal methods for correlated binary data with misclassified responses and for longitudinal ordinal data with misclassification in both response and covariates. Yi (2017) contains a comprehensive treatment on strategy, method and application on statistical analysis with measurement error or misclassification. Measurement error problems have also been addressed by several authors for survey data. For instance, Ybarra and Lohr (2008) and Gregoire and Salas (2009) considered small area estimation and ratio estimation with measurement error in auxiliary variables.

Many variables collected from surveys are categorical and ordinal. These variables may be subject to misclassifications when the survey is based on self-report. In many health surveys the objective is to investigate the association of binary chronic conditions with categorical exposures that are collected with error. Motivated by this feature, we consider logistic regression analysis of data from complex surveys with misclassification in ordinal covariates. We exploit estimation and inference methods for the regression coefficients associated with the risk factors. An expectation correction method is proposed for simultaneously accounting for misclassification and complex survey features. Results from a simulation study are reported to show the good performance of the proposed method. Finally, we apply the method to a data set from the Canadian Community Health Survey (CCHS).

2. Model Formulation

The choice of a statistical model is often dictated by the types of the variables. Many models are available for continuous response variables. Our discussion is focused on a binary response variable such as the presence of a heart disease and an ordinal covariate which is subject to misclassification.

2.1. Response Model

Suppose that the finite population consists of N individuals. For $i = 1, \ldots, N$, let Y_i denote the binary response variable for individual i such that $Y_i = 1$ if the outcome is present and $Y_i = 0$ otherwise. Let X_i be a $(K + 1)$-level ordinal variable that takes values at $0, 1, \ldots, K$ and is subject to misclassification. Let $X_{i0}, \ldots, X_{iK}$ be the indicators such that $X_{ik} = 1$ if $X_i = k$ and $X_{ik} = 0$ otherwise. Without loss of generality, we treat the lowest category, i.e., 0, as the reference category. Therefore, the vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})^{\mathrm{T}}$ can be used to equivalently represent the original categorical X_i. Let $\mathbf{Z}_i$ be a vector of precisely measured covariates, which may include both continuous and discrete variables. For ease of exposition of the following model, we let 1 be the first component of $\mathbf{Z}_i$.

We assume that the finite population is generated from a superpopulation model ζ. Let $\mu_i = \mathrm{E}_\zeta[Y_i|\mathbf{X}_i, \mathbf{Z}_i]$ be the conditional mean of Y_i under the superpopulation model. A logistic regression model is given by

$$\text{logit } \mu_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_x + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_z,$$

where $\boldsymbol{\beta}_x$ and $\boldsymbol{\beta}_z$ are the vectors of regression coefficients associated with the effects of $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively.

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^{\mathrm{T}}, \boldsymbol{\beta}_z^{\mathrm{T}})^{\mathrm{T}}$. If data on all N individuals were available, the population parameter $\boldsymbol{\beta}_N$ is then defined as the maximizer of the finite population log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)\,,$$

where

$$\ell_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) = Y_i \log \mu_i + (1 - Y_i)\log(1 - \mu_i)\,.$$

The finite population parameter $\boldsymbol{\beta}_N$ can be viewed as an estimator of the model parameter $\boldsymbol{\beta}$ if one has information of the entire finite population. The finite population parameter $\boldsymbol{\beta}_N$ can be viewed as the solution to the so-called census estimating equations

$$\sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{0}\,,$$

where

$$\mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{Y_i - \mu_i}{V_i}\,,$$

and $V_i = \mu_i(1 - \mu_i)$ is the conditional variance of Y_i, given $\mathbf{X}_i$ and $\mathbf{Z}_i$, under the model ζ.

Suppose a sample s consisting of n individuals is drawn from the finite population using a complex survey design p. Let d_i be the survey weights for individual i, $i \in s$. The finite population parameter $\boldsymbol{\beta}_N$ and superpopulation model parameter $\boldsymbol{\beta}$ can be simultaneously estimated by maximizing the pseudo-likelihood, defined as

$$\sum_{i \in s} d_i \ell_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) = \sum_{i \in s} d_i \left\{ Y_i \log \mu_i + (1 - Y_i)\log(1 - \mu_i) \right\}.$$

Since $\sum_{i \in s} d_i \ell_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ is the Horvitz-Thompson estimator of $\sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$, the resulting estimating function, obtained by differentiating $d_i \ell_i(\boldsymbol{\beta};\, Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ with respect to $\boldsymbol{\beta}$, is unbiased for the finite population estimating function under the survey design p, i.e.,

$$\mathrm{E}_p \left[\sum_{i \in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) \right] = \sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i), \tag{2.1}$$

where E_p denotes the expectation with respect to the probability sampling design, p.

2.2. Misclassification Model

Misclassification of a categorical covariate with more than two levels are commonly seen in survey sampling, especially for measurements based on self-reporting. It is reasonable to assume that the misclassification of an ordinal covariate only occurs between adjacent categories (e.g., BMI categories, income levels). Often, the misclassification process may depend on other covariates such as $\mathbf{Z}_i$).

Let W_i be the observed surrogate for X_i. We write $W_{il} = 1$ if $W_i = l$, and $W_{il} = 0$ otherwise, where $l = 0, \ldots, K$. Let $\pi_{ik,l} = \Pr(W_i = l | X_i = k, \mathbf{Z}_i)$ be the probability that the observed category is l given the true category is k for individual i, where $k, l = 0, \ldots, K$. Based on the assumption of adjacent misclassifications, we set $\pi_{ik,l} = 0$ for $|k - l| \geq 2$. The probability of correctly classifying X_i into category k is then given by

$$\pi_{ik,k} = 1 - \pi_{ik,k-1}\mathrm{I}(k > 0) - \pi_{ik,k+1}\mathrm{I}(k < K),$$

where $\mathrm{I}(\cdot)$ is the indicator function.

We assume that the misclassification process is characterized by the generalized logistic models (Pfeffermann et al., 1998)

$$\begin{aligned}
\log\left(\frac{\pi_{ik,k-1}}{\pi_{ik,k}}\right) &= \mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k-1}, \qquad k = 1, \ldots, K, \\
\log\left(\frac{\pi_{ik,k+1}}{\pi_{ik,k}}\right) &= \mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k+1}, \qquad k = 0, \ldots, K-1,
\end{aligned}$$

where $\mathbf{L}_i$ is a set of covariates (usually part of $\mathbf{Z}_i$) associated with the misclassification process, and $\boldsymbol{\varphi}_{k,k-1}$ and $\boldsymbol{\varphi}_{k,k+1}$ are the vectors of regression parameters. Let $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_{01}^{\mathrm{T}}, \ldots, \boldsymbol{\varphi}_{K,K-1}^{\mathrm{T}})^{\mathrm{T}}$. Therefore, the probability of misclassifying an observation into the lower category is given by

$$\pi_{ik,k-1} = \frac{\exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k-1})}{1 + \exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k-1}) + \exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k+1})}, \quad k = 1, \ldots, K,$$

and the probability of misclassifying an observation into the higher category is given by

$$\pi_{ik.k+1} = \frac{\exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k+1})}{1 + \exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k-1}) + \exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k+1})}, \quad k = 0, \ldots, K-1.$$

When both K and the number of covariates in $\mathbf{L}_i$ are large, the dimension of nuisance parameter vector $\boldsymbol{\varphi}$ can be very high. In some cases, the misclassification process may be homogeneous, i.e., the probability of misclassifying the observation into the lower or higher category is consistent for all categories. In such cases, the dimension of $\boldsymbol{\varphi}$ is comparable with $\binom{K}{2}$.

2.3. Model for the Ordinal Covariate

For covariate measurement error problems, the literature distinguishes structural modeling, which hypothesizes a distribution for the error-prone covariate, and functional modeling, which leaves the distribution of the covariates unspecified (Carroll et al., 2006; Yi, 2017).

For ordinal variables, the cumulative probabilities are often used as alternatives to marginals. Let $\lambda_{ik} = \Pr(X_i \geq k|\mathbf{Z}_i)$, for $k = 1, , \ldots, K$. The proportional odds models can be employed to characterize the conditional distribution of X_i given $\mathbf{Z}_i$ (e.g., Agresti, 2002). The kth model is given by

$$\text{logit } \lambda_{ik} = \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\alpha}_k, \quad k = 1, \ldots, K,$$

where $\boldsymbol{\alpha}_k = (\alpha_{0k}, \boldsymbol{\psi}^{\mathrm{T}})^{\mathrm{T}}$, α_{0k} is the intercept and $\boldsymbol{\psi}$ is a vector of regression coefficients associated with the sub-vector of $\mathbf{Z}_i$ with the first element 1 excluded. Here α_{0k} may depend on the index k but $\boldsymbol{\psi}$ is common for all k. Let $\boldsymbol{\alpha} = (\alpha_{01}, \ldots, \alpha_{0K}, \boldsymbol{\psi}^{\mathrm{T}})^{\mathrm{T}}$ be the vector of all regression parameters associated with the distribution of X_i.

In general, the dimension of $\boldsymbol{\alpha}$ mainly depends on K and the dimension of $\mathbf{Z}_i$. However, when X_i and $\mathbf{Z}_i$ are independent, we need only to specify the marginal distribution of X_i, which is given by

$$\Pr(X_i = k) = \alpha_k, \quad k = 0, \ldots, K,$$

where $\sum_{k=0}^{K} \alpha_k = 1$. In this case, we set $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_K)^{\mathrm{T}}$ whose dimension is $K + 1$.

3. Estimation Procedures

We present the procedures for estimating the parameter $\boldsymbol{\beta}$ in the main model as well as the parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ of the measurement error models. Variance estimation is also discussed.

3.1. Expected Score for Estimation of $\boldsymbol{\beta}$

If data were free of measurement error, the estimating function $\sum_{i \in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ is unbiased under the sampling scheme and the superpopulation model. In the presence of misclassification, however, $\mathbf{X}_i$ is not available. Let $\mathbf{W}_i = (W_{i1}, \ldots, W_{iK})^{\mathrm{T}}$. Ignoring misclassification and naively solving the set of equations

$$\sum_{i \in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathbf{0}$$

for $\boldsymbol{\beta}$ no longer yields a valid estimate of $\boldsymbol{\beta}$. If there exists a set of estimating functions, say, $\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, \mathbf{W}_i, \mathbf{Z}_i)$, that is "close" to $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$, then solving

$$\sum_{i \in s} d_i \mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathbf{0}$$

may lead to a consistent estimator for $\boldsymbol{\beta}$ under certain regularity conditions.

We here construct an approximate version of $\sum_{i\in s} d_i\mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ by taking conditional expectation with respect to the underlying unobserved variables $\mathbf{X}_i$ given the observed data $(Y_i, W_i, \mathbf{Z}_i)$. The evaluation of the conditional expectation pertains to the response model, the measurement error model, as well as the covariate distributions. Without additional information, $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ cannot be estimated from the available data $\{(Y_i, W_i, \mathbf{Z}_i), i \in s\}$. We require additional data sources to feature the misspecification probabilities (Carroll et al. 2006; Yi 2017). Here we consider cases where validation data are available.

Suppose internal validation data is available where the true error-prone covariate is partially observed. The original sample s can be divided into three subsets as follows:

$$\begin{aligned} s_1 &= \{i : (Y_i, X_i, \mathbf{Z}_i)\}\,; \\ s_2 &= \{i : (Y_i, W_i, X_i, \mathbf{Z}_i)\}\,; \\ s_3 &= \{i : (Y_i, W_i, \mathbf{Z}_i)\}. \end{aligned}$$

For $i \in s_3$, let $\mathbf{U}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathrm{E}_\zeta[\mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)|Y_i, \mathbf{W}_i, \mathbf{Z}_i]$. Given $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, the response parameter $\boldsymbol{\beta}$ can be estimated by solving

$$\sum_{i\in s_1\cup s_2} d_i\mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) + \sum_{i\in s_3} d_i\mathbf{U}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathbf{0}. \tag{3.1}$$

For $k = 1, \ldots, K$, let $\mathbf{e}_k$ denote a K-dimensional vector whose kth element is 1 and 0 otherwise; let $\mathbf{e}_0 = \mathbf{0}$. Let $\Omega_i(W_i) = \{k : \max(0, W_i - 1) \le k \le \min(W_i + 1, K)\}$ be a set of possible values for the underlying true covariate, given W_i. The function $\mathbf{U}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}; Y_i, \mathbf{W}_i, \mathbf{Z}_i)$ can be shown as a weighted sum of the $\mathbf{U}_i$

$$\mathbf{U}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \sum_{k\in\Omega_i(W_i)} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{e}_k, \mathbf{Z}_i)\mathrm{Pr}(X_i = k|Y_i, W_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}),$$

where $\mathrm{Pr}(X_i = k|Y_i, W_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})$ is the posterior weight of $(X_i = k)$, given the observed data $(Y_i, W_i, \mathbf{Z}_i)$. This posterior weight can be expressed as

$$\begin{aligned} &\mathrm{Pr}(X_i = k|Y_i, W_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}) \\ = &\frac{\mathrm{Pr}(Y_i, W_i, X_i = k|\mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})}{\sum_{k'\in\Omega_i(W_i)} \mathrm{Pr}(Y_i, W_i, X_i = k'|\mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})} \\ = &\frac{\mathrm{Pr}(Y_i|W_i, X_i = k, \mathbf{Z}_i; \boldsymbol{\beta})\mathrm{Pr}(W_i, X_i = k|\mathbf{Z}_i; \boldsymbol{\varphi}, \boldsymbol{\alpha})}{\sum_{k'\in\Omega_i(W_i)} \mathrm{Pr}(Y_i|W_i, X_i = k', \mathbf{Z}_i; \boldsymbol{\beta})\mathrm{Pr}(W_i, X_i = k'|\mathbf{Z}_i; \boldsymbol{\varphi}, \boldsymbol{\alpha})} \\ = &\frac{\mathrm{Pr}(Y_i|X_i = k, \mathbf{Z}_i; \boldsymbol{\beta})\mathrm{Pr}(W_i|X_i = k, \mathbf{Z}_i; \boldsymbol{\varphi})\mathrm{Pr}(X_i = k|\mathbf{Z}_i; \boldsymbol{\alpha})}{\sum_{k'\in\Omega_i(W_i)} \mathrm{Pr}(Y_i|X_i = k', \mathbf{Z}_i; \boldsymbol{\beta})\mathrm{Pr}(W_i|X_i = k', \mathbf{Z}_i; \boldsymbol{\varphi})\mathrm{Pr}(X_i = k'|\mathbf{Z}_i; \boldsymbol{\alpha})}, \end{aligned}$$

which involves the response model, the misclassification model and the covariate distribution. If X_i and $\mathbf{Z}_i$ are independent, then

$$\begin{aligned} &\mathrm{Pr}(X_i = k|Y_i, W_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}) \\ = &\frac{\mathrm{Pr}(Y_i|X_i = k, \mathbf{Z}_i; \boldsymbol{\beta})\mathrm{Pr}(W_i|X_i = k, \mathbf{Z}_i; \boldsymbol{\varphi})\mathrm{Pr}(X_i = k; \boldsymbol{\alpha})}{\sum_{k'\in\Omega_i(W_i)} \mathrm{Pr}(Y_i|X_i = k', \mathbf{Z}_i; \boldsymbol{\beta})\mathrm{Pr}(W_i|X_i = k', \mathbf{Z}_i; \boldsymbol{\varphi})\mathrm{Pr}(X_i = k'; \boldsymbol{\alpha})}. \end{aligned}$$

For fixed $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, estimation of $\boldsymbol{\beta}$ can be performed through an iterative procedure for solving (3.1). We now describe the estimation algorithm as follows:

1. For $i \in s_3$, obtain the set of all possible values of X_i, given W_i.
2. Given a current estimate $\hat{\boldsymbol{\beta}}^{(t)}$ and fixed $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, calculate the pseudo-survey weight for each enumerated possibility in the set $\Omega_i(W_i)$
$$d_{ik}^{(t)} = d_i \, \Pr(X_i = k | Y_i, W_i, \mathbf{Z}_i; \hat{\boldsymbol{\beta}}^{(t)}, \boldsymbol{\varphi}, \boldsymbol{\alpha})$$
3. Obtain a new estimate $\hat{\beta}^{(t+1)}$ by solving
$$\sum_{i \in s_1 \cup s_2} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) + \sum_{i \in s_3} \sum_{k \in \Omega_i(W_i)} d_{ik}^{(t)} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{e}_k, \mathbf{Z}_i) = \mathbf{0}$$
for $\boldsymbol{\beta}$.
4. The algorithm iterates between steps 2 and 3 until the resulting estimates of $\boldsymbol{\beta}$ converge.

Let $\hat{\boldsymbol{\beta}}$ denote the final estimate at convergence.

3.2. Estimation of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$

The estimation procedure for $\boldsymbol{\beta}$ requires knowledge of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, which can be estimated from the validation data. An estimate of $\boldsymbol{\varphi}$ can be obtained by fitting the misclassification model to subsample s_2, while an estimate of $\boldsymbol{\alpha}$ can be obtained from the combined information from s_1 and s_2.

When the dimensions of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are very high, the validation data may not be able to provide sufficient information for the estimation. In this situation, we may impose certain assumptions to constrain the dimension of the parameters. For example, assuming the independence between X_i and $\mathbf{Z}_i$ gives that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$, whose entries can be estimated by
$$\hat{\alpha}_k = \frac{\sum_{i \in s_1 \cup s_2} d_i \mathrm{I}(X_i = k)}{\sum_{i \in s_1 \cup s_2} d_i}, \quad k = 1, \ldots, K.$$

3.3. Variance Estimation

Variance estimation for model parameters using complex survey data is a challenge task. It is known that model-based variance estimators for the survey weighted estimator $\hat{\boldsymbol{\beta}}$ do not work (Binder and Roberts, 2009). When the sampling fraction n/N is small, the design-based variance estimator provides valid inference on the model parameter under mild conditions on the model and the finite population (Binder and Roberts, 2009). With misclassification in the ordinal covariate, additional modeling of the misclassification process makes the variance estimation even harder.

We suggest to use a resampling method such as the bootstrap approach for variance estimation. Bootstrap variance estimators such as those of Rao and Wu (1988) and of Sitter (1992) are popular in survey practice due to their straightforward implementation.

Suppose that $\hat{\boldsymbol{\beta}}_{(b)}$ is the estimate of $\boldsymbol{\beta}$ from an estimation procedure using the bth bootstrap sample, where $b = 1, 2, \ldots, B$, and B is a user-specified positive integer. Given fixed $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, the bootstrap variance estimate of $\hat{\boldsymbol{\beta}}$ is given by

$$BV(\hat{\boldsymbol{\beta}}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\boldsymbol{\beta}}_{(b)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{(b)} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}. \tag{3.2}$$

When $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are estimated from internal validation data, the uncertainty in $(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\alpha}})$ needs to be accounted for when calculating the variance of $\hat{\boldsymbol{\beta}}$. This can be done by re-estimating $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ within each bootstrap sample.

4. Simulation Study

We conducted a simulation study to investigate the performance of the proposed method and compared our method to the naive approach which ignores measurement error and the complete-case approach using the validation subsample. The configuration of the simulation, especially the choice of a very large sample size n, is based on the data set from the Canadian Community Health Survey (CCHS) Cycle 3.1. Details of the survey are described in the next section.

4.1. Design of the Simulation Study

We only considered simple random sampling from a superpopulation. When the finite population is extremely large and follows the superpopulation model, the differences between the finite population parameters and the superpopulation parameters can be ignored. We set the sample size to be $n = 100,000$. The covariates included a three-level ordinal X_i (valued at 1, 2 and 3) that was subject to misclassification, and a continuous Z_i free of measurement error. We first generated X_i with probabilities 0.2, 0.5, and 0.3 for levels 1, 2, and 3, respectively. We then generated Z_i, independent of X_i, using the standard normal distribution $\mathrm{Normal}(0, 1)$ for all subjects. The binary response variable Y_i was generated under the logistic regression model

$$\text{logit }(\mu_i) = \beta_0 + \beta_1 \mathrm{I}(X_i = 1) + \beta_2 \mathrm{I}(X_i = 3) + \beta_z Z_i\,, \tag{4.1}$$

where $\mathrm{logit}(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$ and $\mu_i = E(Y_i \mid X_i, Z_i) = \Pr(Y_i = 1 \mid X_i, Z_i)$. The parameters were specified as $\beta_0 = -3$, $\beta_1 = 0.3$, $\beta_2 = 0.5$, and $\beta_z = 0.5$. The X_i can be viewed as the ordinal variable for the Body Mass Index (BMI), for which the three levels represent underweight, normal weight, and overweight or obese categories. The coefficients β_1 and β_2 were specified in such a way that both level 1 and level 3 have positive effect on the risk of developing the outcome event (i.e., $Y_i = 1$) compared to the normal level 2.

The true values of X_i were not observed in the sample data. We instead observed a surrogate W_i for X_i, which was generated using the multinomial logit models given by

$$\log\left(\pi_{ik,l}/\pi_{ik,k}\right) = \varphi_{kl(0)} + \varphi_{kl(z)} Z_i \quad \text{for } |k-l| = 1\,,$$

where $\pi_{ik,l} = \Pr(W_i = l|X_i = k, Z_i)$. The parameters $\boldsymbol{\varphi}$ associated with the misclassification process were specified by Table 1. The misclassification of X_i depends on Z_i in such a way that Z_i has a positive effect on increasing the probability of misclassifying a higher level into a lower one. The dependence is stronger for misclassification of $X_i = 3$ into $W_i = 2$ than for other cases.

Table 1: Values of $\boldsymbol{\varphi}$

X	W	$\varphi_{kl(0)}$	$\varphi_{kl(z)}$
1	2	-1.5	-0.05
2	1	-3.0	0.05
	3	-3.0	-0.05
3	2	-1.5	0.50

We obtained the final observed sample $s = \{(Y_i, W_i, Z_i), i = 1, \ldots, n\}$. Also, we obtained a validation subsample $s_2 = \{(Y_i, W_i, X_i, Z_i)\}$ by randomly selecting subjects from s with probability 0.04 to be included in s_2. Therefore, the size of the validation subsample s_2 was around 4,000. The data were then analyzed using following approaches: (i) the naive approach which ignors measurement error; (ii) the complete-cases analysis which uses only the validation subsample; and (iii) the expectation correction method that accounts for misclassification and uses the full sample. The simulation was repeated 500 times. It turned out that obtaining bootstrap variance estimates for $\hat{\boldsymbol{\beta}}$ under the current setting with the chosen sample size $n = 100,000$ was extremely time consuming for repeated simulations. We only included the empirical variance of the estimators based on the 500 repeated simulation samples. Our simulations are conducted using the R software and the codes are available from the authors upon requests.

4.2. Simulation Results

Table 2: Simulation Results for the Estimation of $\boldsymbol{\beta}$

	Naive Method		Complete-case		Proposed Method	
Parameter	%RB [†]	EV[‡]	%RB	EV	%RB	EV
β_0	-2.90	0.00036	-0.12	0.01148	-0.02	0.00068
β_1	-42.31	0.00110	-4.42	0.02926	-1.42	0.00200
β_2	-25.24	0.00085	-0.39	0.02244	-0.24	0.00157
β_z	1.71	0.00017	-0.03	0.00444	-0.01	0.00017

[†] %RB = $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\boldsymbol{\beta} \times 100$

[‡] Empirical variance based on 500 simulation samples

The simulated relative biases (%RB, in percentage) and the empirical variances of the estimator $\hat{\boldsymbol{\beta}}$ for the parameters in the main logistic regression model (4.1) are reported in Table 2. Major observations form the simulation results can be summarized as follows: (i) For the estimation of β_z and β_0, all three estimators have small biases, although the naive estimator is slightly larger than the other two. (ii) The naive estimators of β_1 and β_2, which ignore the measurement error, are seriously biased with the values of RB at 42.3% and 25.2%, respectively. (iii) The complete-cases analysis based on the validation subsample produces acceptable results but is less efficient than the proposed method. (iv) The magnitude of the empirical variance of the estimators reinforces the statement from (iii) and the proposed expectation correction method should be preferred in dealing with the measurement error in the ordinal covariate.

5. Application to the CCHS Data Set

In this section, we presents results from the application of the proposed inferential procedures to the data set from the Cycle 3.1 of the Canadian Community Health Survey (CCHS), an ongoing large scale survey conducted by Statistics Canada.

5.1. The Canadian Community Health Survey

The Canadian Community Health Survey (CCHS) initiative began in 2000 with its main goals being the provision of population-level information on health determinants, health status and health system utilization across Canada and gathering data at the sub-provincial levels of geography (Statistics Canada, 2005). The Cycle 3.1 of CCHS was conducted in 2005 and targeted persons aged 12 years or older who live in private dwellings in the ten provinces and the three territories. Persons living on Indian Reserves or Crown lands, clientele of institutions, full-time members of the Canadian Armed Forces and residents of certain remote regions are excluded from the survey.

For administrative purposes, each province is divided into health regions (HR) according to the types of regions: major urban centres, cities, and rural regions. Each territory is designated as a single HR. During Cycle 3.1 of the CCHS, data were collected in 122 HRs in the ten provinces, in addition to one HR per territory, totalling 125 HRs. Three sampling frames are used to select the sample of households: 49% of the sample of households came from an area frame, 50% came from a list frame of telephone numbers and the remaining 1% came from a Random Digit Dialling (RDD) sampling frame. The CCHS uses the area frame designed for the Canadian Labour Force Survey (LFS). The sampling plan of the LFS is a multistage stratified cluster design in which the dwelling is the final sampling unit. Geographic or socio- economic strata are created within each HR. Within the strata, between 150 and 250 dwellings are regrouped to create clusters. Some urban centres have separate strata for apartments or for census Enumeration Areas (EA) to pinpoint households with high income, immigrants and the native people.

In each stratum, six clusters or residential buildings (sometimes 12 or 18 apartments) are chosen with probability proportional to size (PPS), with the number of households as the size variable. The list frame of telephone numbers was used in all but five HRs (the two RDD only HRs and the three territories) to complement the area frame. One list frame stratum was then created for each HR based on postal codes that were obtained from names, addresses and telephone numbers. Within each stratum the required number of telephone numbers was selected using simple random sampling from the list. As for the RDD frame, additional telephone numbers were selected to account for the numbers not in service or out-of-scope. The hit rate observed under the list frame approach varied from 75% to 88% depending on the province, which was much higher than that for the RDD frame. In four HRs, a Random Digit Dialling (RDD) sampling frame of telephone numbers was used to select the sample of households.

For all selected households, a single person aged 12 and older was randomly chosen from members of the household. After removing the out-of-scope units, 168,464 households were selected to participate in the CCHS Cycle 3.1. Data were obtained from 132,947 respondents, yielding a response rate of 79%. Data were collected on general health, chronic health conditions, drinking or smoking status, including self-reported weight and height. A subsample of 7,376 respondents aged 12 or older were also selected, who were asked later in the interview to directly measure weight and height. Among the 7,376 individuals selected in the subsample, 4,735 individuals responded. The main reason for non-response was refusal (Statistics Canada, 2005). Such validation subsample is useful in studies of risk factors for obesity as well as the effect of obesity on health conditions. It provides information on the relationship between a precise measurement and an error-contaminated measurement of weight or height that makes it possible to correct estimation bias induced by the self-reported survey data.

5.2. Application of the Proposed Method

We applied the method developed in this paper to the survey data set from the Cycle 3.1 of CCHS in 2005. Our interest was in studying the association of health conditions with risk factors including age, sex, physical activity, and body mass index (BMI). Based on the Canadian guidelines, which are in line with those of the World Health Organization, the BMI for adults was divided into six categories: underweight, normal weight, overweight, and three obese classes; see Table 3 for the range of each category.

As BMI was derived from self-reported weight and height, the recorded category might be different from the true category for some subjects. The subsample contains both self-reported and the measured weight and height and hence can be used as validation data. Five age groups were formed with 18-24 being the reference group. Physical activity index is an ordinal variable with three levels: active, moderate, and inactive. Here the error-contaminated variable is the self-reported BMI category, and the true underlying variable is the measured BMI category. For this study, we excluded subjects who were less than 18 years old, as children are in a stage of development where weight and height may change over a short period of time. Women who were pregnant or breastfeeding were also excluded. Observations in the subsample with self-reported and measured BMI two categories apart were considered as outliers, and the frequency for such instances is less than 0.1%. Subjects with missing any of the error-free covariates or missing both the self-reported and the measured BMI were also excluded from the analysis. This led to a sample of 114,547 respondents with 4,125 in the validation subsample.

We first present results from an exploratory analysis using the validation subsample. Figures 1 and 2 show the weighted estimates of population proportions for high blood pressure and heart disease in each of BMI categories. There is a clear trend of increasing proportion of subjects with high blood pressure as the BMI category level increases, indicating that obesity is a strong risk factor in developing high blood pressure. We observe a similar pattern on heart disease, except

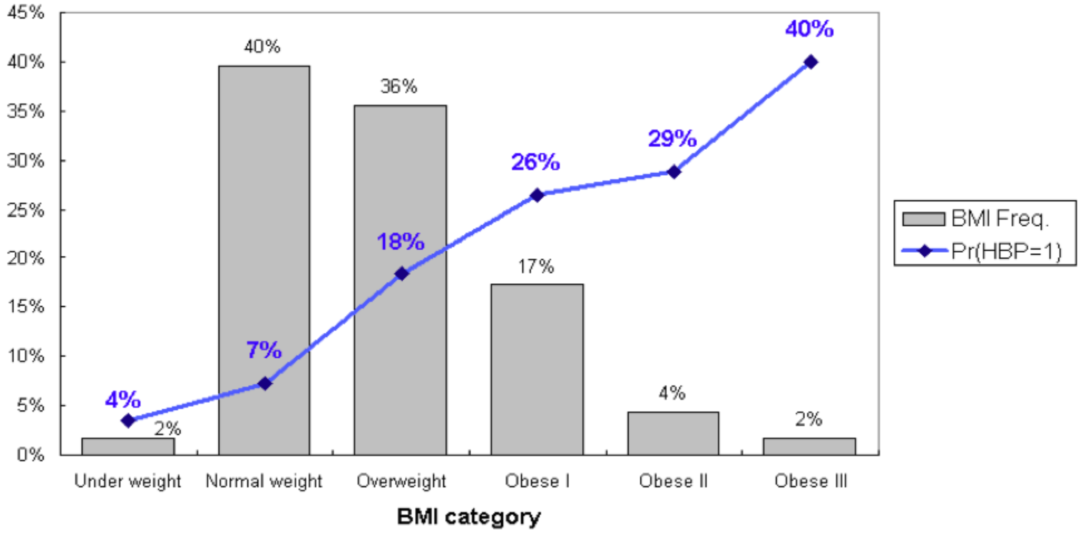
Table 3: Body Mass Index Categories

Category	BMI (kg/m^2)
Underweight (UW)	Less than 18.5
Normal weight (NW)	18.5 to 24.9
Overweight (OW)	25.0 to 29.9
Obese class I (OB I)	30.0 to 34.9
Obese class II (OB II)	35.0 to 39.9
Obese class III (OB III)	40.0 or more

that the proportion is higher in the underweight category than in the normal-weight category.

Table 4 summarizes the sample percentages of the classification rates of the self-reported BMI in the validation subsample which contains accurately measured BMI. It can be seen that subjects with normal weight tend to report the BMI more accurately while subjects in the overweight or obese group are more likely to self-report the value to a lower category. In general, the misclassification rates increase as the level of the BMI category moves up and subjects with high level of BMI tend to under-report the value of the BMI.

Figure 1: Population Proportions for High Blood Pressure in BMI Categories



Our formal analysis applied the proposed expectation correction method to the Cycle 3.1 of the CCHS data set. It also included the naive approach, which treats the self-reported BMIs as if they are the true values, and the complete-case approach, which uses only the validation subsample, for the purpose of comparisons. The normal-weight category of the BMI was treated as the reference group, and the relative risk of the five other BMI categories on the probabilities of having chronic conditions are of scientific interest. The variance estimates were based on the 500 bootstrap samples with adjusted survey weights which were included as part of the data set.

Figure 2: Population Proportions for Heart Disease in BMI Categories

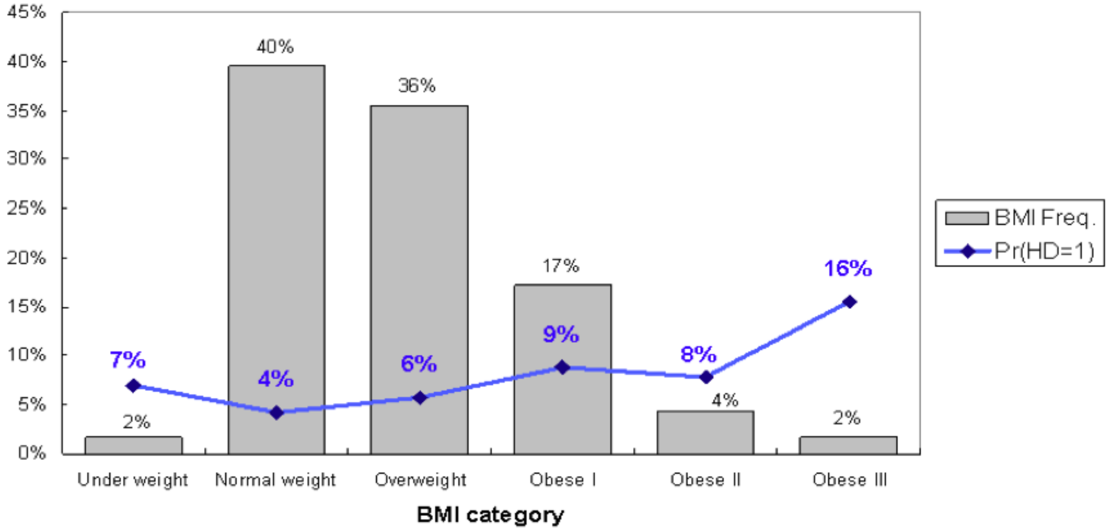


Table 4: BMI Classification Rates in the CCHS Subsample

Measured	Self-reported BMI						
BMI	UW	NW	OW	OB I	OB II	OB III	Missing
UW	70.00%	27.50%					2.5%
NW	4.21%	90.06%	4.34%				1.38%
OW		29.33%	66.93%	2.00%			1.74%
OB I			46.42%	51.24%	0.58%		1.46%
OB II				57.79%	37.19%	3.02%	2.01%
OB III					32.43%	60.81%	6.76%

The results on the estimation of model parameters for the response variable "high blood pressure" are shown in Table 5. It can be seen that the expectation correction approach does not differ much from the naive approach in terms of estimating $\boldsymbol{\beta}_z$, the regression coefficients associated with the error-free covariates: age, sex and physical activity index. The estimates of $\boldsymbol{\beta}_z$ from the complete-cases analysis, however, are noticeably different in terms of the magnitude or the direction. The three approaches do not quite agree in the risk estimates of BMI categories, although the trend of increasing risk across BMI categories is consistent. The direction of the risk estimates of the underweight category is positive for the expectation correction approach but is negative for the naive approach and the complete-case approach. Variance estimates, or equivalently the standard error (SE), using the three methods are all very large for the underweight category compared to those of other BMI categories. We conclude that the risk of having high blood pressure is not significantly higher among underweight people than people with normal-weight.

The analysis results for the response variable "heart disease" are presented in Table 6. We observe similar patterns in the estimates of model parameters. The results from the expectation correction approach indicate that the risk of having heart disease increases as the level of the BMI category increases. However, subjects in the underweight BMI category have relatively higher risk than those in normal-weight category. In contrast, the risk for subjects in overweight category is not significantly different from those in normal-weight category. The variance estimates from the complete-case approach are significantly larger that those of the other two methods, which leads to the statistical conclusion of non-significant BMI effect on heart disease. This is partially due to the much smaller size of the validation subsample and the result might not be very reliable.

6. Concluding Remarks

We consider logistic regression analysis of survey data with a binary response and an ordinal covariate which is subject to misclassification. We propose to use the expectation correction estimation method (Yi, 2017, Section 2.5) for analysis of this type of error-contaminated data with survey weights incorporated. The implementation of the algorithm is relatively easy.

The proposed method requires calculations of the posterior weights for all possible values of the unobserved true covariate, hence it relies on full parametric assumptions for the misclassification mechanism as well as the covariate distribution. Robustness to model misspecification needs to be investigated. Also, the parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are estimated from the validation data and are treated as fixed "plus-in" estimators in the estimation of $\boldsymbol{\beta}$. When calculating the bootstrap variance of $\hat{\boldsymbol{\beta}}$, one can account for the extra uncertainty by obtaining estimates of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ in each bootstrap sample. Otherwise, the standard error of $\hat{\boldsymbol{\beta}}$ would be generally underestimated. The main problem is that some bootstrap samples do not contain enough validation data to obtain stable estimators for $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, especially for the cases where the ordinal covariate has many levels, and the misclassification process involves large number of precisely measured covariates.

To validate the our proposed correction model, we first need to validate the appropriateness of the misclassification model. This can be done by using cross-validation method by training the model and parameters in 70% of the subsample and testing in the remaining 30%, or by collecting more measured samples. Since the number of parameters was large in the application presented in

Section 5, we trained the misclassification model using the entire subsample and did not perform cross-validations.

In some practical situations the marginal distribution of X_i may be of interest, e.g., estimation of population frequency of each BMI category can be an objective of health surveys. When the dimensions of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are small, we can simultaneously estimate $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\alpha}$. Specifically, one can use the extended data with pseudo-survey weights $d_{ik}^{(t)}$ to update the estimates of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$. When X_i is independent of $\mathbf{Z}_i$, for instance, the estimate of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$ can be updated during each iteration by

$$\alpha_k^{(t+1)} = \frac{\sum_{i \in s_1 \cup s_2} d_i \mathrm{I}(X_i = k) + \sum_{i \in s_3} d_{ik}^{(t)}}{\sum_{i \in s} d_i}, \quad k = 1, \ldots, K.$$

In the data analysis example on the CCHS survey data set, the BMI variable is used as a risk factor for health conditions. However, BMI itself can be viewed as a response variable, and studying the association of obesity with covariates such as age, sex and physical activity index may be of interest. Misclassifications in both categorical response and categorical covariate are commonly seen in large scale surveys. It would be interesting to extend our current work to account for this type of survey data.

Another future research problem is to deal with scenarios where both measurement error and missing values are present in survey data, the features which are common for large scale surveys such as CCHS. Developing valid inferential procedures for survey data with measurement error and missing values is an interesting yet very challenging research topic.

Acknowledgements

This research was supported by grants to Wu and Yi from the Natural Sciences and Engineering Research Council (NSERC) of Canada and Mathematics of Information Technology and Complex Systems (MITACS).

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience, 2nd edition.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

Binder, D. A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, **89**, 1035–1043.

Binder, D. A. and Roberts, G. (2009). Design- and model-based inference for model parameters. In *Handbook of Statistics, Volume 29B, Sample Surveys: Inference and Analysis*, editors: D. Pfeffermann and C.R. Rao, 33–54.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, London, 2nd edition.

Chen, Z., Yi, G. Y. and Wu, C. (2011). Marginal methods for correlated binary data with misclassified responses. *Biometrika*, **98**, 647–662.

Chen, Z., Yi, G. Y. and Wu, C. (2014). Marginal analysis of longitudinal ordinal data with misclassification in both response and covariates. *Biometrical Journal*, **56(1)**, 69–85.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, **54**, 127–138.

Gregoire, T. G. and Salas, C. (2009). Ratio estimation with measurement error in the auxiliary variate. *Biometrics*, **65**, 590–598.

Pfeffermann, D., Skinner, C., and Humphreys, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society, Series A*, **161**, 13–32.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, **83**, 231–241.

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, **87**, 755–765.

Statistics Canada (2005). *Canadian Community Health Survey (CCHS) Cycle 3.1 Public Use Microdata File User Guide*. Statistics Canada.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, **90**, 937–951.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185–193.

Ybarra, L. M. R. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**, 919–931.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer, New York.

Table 5: Analysis Results for Modeling “High Blood Pressure”

	Naive Method [†] ($n = 114325$)			Complete-case [‡] ($n = 4120$)			Expectation Correction ($n = 114325$)		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	-4.137	0.045	< 0.001	-4.670	0.375	< 0.001	-4.345	0.075	< 0.001
BMI									
Underweight	-0.099	0.106	0.349	-1.082	1.280	0.398	0.298	0.487	0.540
Normal weight	0.000	.	.	0.000	.	.	0.000	.	.
Overweight	0.660	0.036	< 0.001	0.645	0.173	< 0.001	0.787	0.052	< 0.001
Obese I	1.178	0.021	< 0.001	1.043	0.201	< 0.001	1.345	0.040	< 0.001
Obese II	1.638	0.042	< 0.001	1.488	0.316	< 0.001	1.849	0.100	< 0.001
Obese III	1.806	0.099	< 0.001	2.548	0.574	< 0.001	2.084	0.106	< 0.001
Age									
18-34	0.000	.	.	0.000	.	.	0.000	.	.
35-49	1.152	0.047	< 0.001	1.307	0.384	< 0.001	1.133	0.021	< 0.001
50-64	2.468	0.052	< 0.001	2.655	0.356	< 0.001	2.444	0.080	< 0.001
65+	3.431	0.063	< 0.001	3.812	0.355	< 0.001	3.369	0.058	< 0.001
Sex									
Male	-0.105	0.016	< 0.001	0.036	0.130	0.784	-0.115	0.067	0.084
Female	0.000	.	.	0.000	.	.	0.000	.	.
PAI									
Active	-0.119	0.043	0.006	0.149	0.217	0.494	-0.130	0.038	< 0.001
Moderate	0.000	.	.	0.000	.	.	0.000	.	.
Inactive	0.109	0.042	0.009	0.206	0.186	0.267	0.111	0.010	< 0.001

[†] Self-reported BMIs are used except for subjects in the validation subsample

[‡] Only the validation subsample is used

Table 6: Analysis Results for Modeling “Heart Disease”

	Naive Method † ($n = 114370$)			Complete-case ‡ ($n = 4123$)			Expectation Correction ($n = 114370$)		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	-5.638	0.112	< 0.001	-5.052	0.646	< 0.001	-5.663	0.117	< 0.001
BMI									
Underweight	0.381	0.142	0.007	0.494	0.803	0.539	0.846	0.268	0.002
Normal weight	0.000	.	.	0.000	.	.	0.000	.	.
Overweight	0.118	0.044	0.007	-0.170	0.248	0.494	-0.005	0.092	0.957
Obese I	0.458	0.057	< 0.001	0.255	0.326	0.4339	0.480	0.079	< 0.001
Obese II	0.648	0.098	< 0.001	0.248	0.420	0.555	0.510	0.159	0.001
Obese III	0.850	0.130	< 0.001	0.883	0.578	0.126	0.955	0.152	< 0.001
Age									
18-34	0.000	.	.	0.000	.	.	0.000	.	.
35-49	0.918	0.130	< 0.001	0.156	0.735	0.832	0.930	0.131	< 0.001
50-64	2.382	0.107	< 0.001	1.859	0.686	0.007	2.403	0.108	< 0.001
65+	3.692	0.106	< 0.001	3.179	0.675	< 0.001	3.695	0.107	< 0.001
Sex									
Male	0.460	0.041	< 0.001	0.689	0.202	< 0.001	0.470	0.042	< 0.001
Female	0.000	.	.	0.000	.	.	0.000	.	.
PAI									
Active	-0.122	0.063	0.052	-0.206	0.333	0.536	-0.115	0.063	0.070
Moderate	0.000	.	.	0.000	.	.	0.000	.	.
Inactive	0.223	0.047	< 0.001	0.442	0.292	0.130	0.225	0.047	< 0.001

† Self-reported BMIs are used except for subjects in the validation subsample

‡ Only the validation subsample is used