SSC Annual Meeting, June 2009
Proceedings of the Survey Methods Section

ANALYSIS OF BINARY DATA FROM A COMPLEX SURVEY WITH MISCLASSIFICATION IN AN ORDINAL COVARIATE

Zhijian Chen[1] and Harold Mantel[2]

ABSTRACT

Measurement error exists commonly in survey data where variables are collected using non-standard methods. A typical example is the body mass index (BMI) obtained from self-reported weight and height. We consider estimation of logistic regression coefficients using survey data, where measurement error/misclassification of an ordinal covariate is dependent on other variables. We propose the expected score method which employs a parametric assumption for the measurement error process. The method is illustrated using data from the Canadian Community Health Survey, where associations of heart disease with BMI categories are of interest. Results from a simulation study and the data analysis are presented.

KEY WORDS: Bootstrap; Estimating equations; Misclassification; Multinomial logit model.

RÉSUMÉ

Les erreurs de mesure sont fréquentes dans les données d'enquêtes dont les variables sont souvent recueillies au moyen de méthodes inhabituelles. Un exemple type est l'indice de masse corporel (IMC) obtenu à partir des poids et grandeurs déclarés par les répondants. Nous examinons l'estimation de coefficients de régression logistique à l'aide des données d'enquête dans lesquelles l'erreur de mesure ou de classification d'une covariable ordinale dépend d'autres variables. Nous proposons la méthode des scores espérés, qui utilise une hypothèse paramétrique pour le processus des erreurs de mesure. La méthode est mise en évidence à l'aide des données tirées de l'Enquête sur la santé dans les collectivités canadiennes dans laquelle des associations entre les maladies du coeur et les catégories d'IMC sont étudiées. Les résultats d'une étude de simulation et de l'analyse de données sont présentés.

MOTS CLÉS: Bootstrap; Classification erronée; équations d'estimation; Modèle logit multinomial.

1. INTRODUCTION

One of the fundamental tasks in statistical analysis of health data is to study the associations between response or outcome variables and particular exposures (covariates). The associations are usually characterized by regression models involving some parameters of interest. Survey sampling has been a very popular way to collect data. Analytic use of survey data, especially health data, has become more and more common. However, data are often far from perfect. Due to non-standard collection procedures used in surveys, some variables may not be precisely measured, although rough surrogate variables are obtained. For example, a poor diagnostic test for a disease infection may generate a false result. Questionnaires that are not well designed may lead to incorrect interpretations by the respondents, causing error in the collected variables. Examples also include many variables of medical interest, such as exposures to indoor or outdoor pollutants, nutrition or drug intakes. This leads to a measurement error problem in statistical analysis. Many variables collected by surveys are categorical or ordinal. When referring to measurement error in categorical variables, it is termed misclassification. In the presence of covariate error, naive estimators for model parameters are often inconsistent; see Fuller (1987). Measurement error problems have been considered for various types of data within the non-survey context, and many authors have proposed methods for analysis of error-contained data (Suh and Schafer, 2002; Yi, 2008). Different assumptions for the measurement error mechanism correspond to different treatments for the induced bias in estimates. Also, the literature distinguishes between functional modeling, which does not assume a distribution for the mismeasured variable, and structural modeling, which in contrast hypothesizes a distributional structure for the

[1] Zhijian Chen, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada, N2L 3G1, z11chen@math.uwaterloo.ca
[2] Harold Mantel, Research and Innovation Division, 16th floor, R.H. Coats Building, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6, Harold.Mantel@statcan.gc.ca

mismeasured variable (Carroll *et al*. 2006). When measurement error is coupled with complex survey design features, it adds another degree of difficulty and requires development of new statistical tools.

In this article, we consider the error-in-covariate problem in analytic use of survey data, with a focus on logistic regression analysis with the presence of misclassification in an ordinal covariate. This problem arises often in health surveys, for which the objectives include association of a chronic condition with a self-reported ordinal exposure. We propose to use the expected score approach for estimating the regression parameters. This approach employs some parametric assumptions for independent data but can be generalized to the context of survey data by incorporating survey design features. Section 2 presents the formulation and estimation procedure for the proposed method. Section 3 shows a simulation study to investigate the performance of the approaches under correct model assumptions. In section 5, we illustrate the proposed method using a data set from the Canadian Community Health Survey (CCHS) cycle 3.1. Section 6 provides some concluding remarks and discussion.

2. METHODOLOGY

2.1 Model Formulation

We consider a finite population of N individuals. Let Y_i denote the binary response for individual i such that $Y_i = 1$ if the outcome is present and $Y_i = 0$ otherwise, $i = 1, \ldots, N$. We assume Y_i is generated from a superpopulation model ξ (Godambe and Thompson, 1986). Let X_i be an ordinal variable with K levels subject to misclassification, and let $\mathbf{Z}_i$ be a vector of correctly measured covariates. Misclassification of a categorical variable is common in surveys based on self-reporting. Without loss of generality, we treat the last category of X_i as the reference group. Let $\mu_i = \mathrm{E}_\xi[Y_i|X_i, \mathbf{Z}_i]$ be the conditional mean of Y_i given true covariates, where E_ξ denotes expectation under the superpopulation model. We assume a logistic regression model for Y_i given by

$$\log\{\mu_i/(1-\mu_i)\} = \beta_0 + \sum_{k=1}^{K-1} \beta_k \mathrm{I}(X_i = k) + \mathbf{Z}_i^T \boldsymbol{\beta}_Z, \tag{1}$$

where $\mathrm{I}(\cdot)$ is the indicator function, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{K-1}, \boldsymbol{\beta}_Z^T)^T$ is a vector of regression parameters. If error-free data were available on all individuals in the finite population, superpopulation parameter $\boldsymbol{\beta}$ can be estimated by the finite population parameter $\boldsymbol{\beta}_N$ which is the solution to the "census" estimating equations

$$\sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i) = 0, \tag{2}$$

where $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i) = \partial\mu_i/\partial\boldsymbol{\beta}\,(Y_i - \mu_i)\,/V_i$ is the score function of $\boldsymbol{\beta}$ contributed from individual i, and $V_i = \mu_i(1-\mu_i)$ is the variance of Y_i under the superpopulation model. Suppose s is a sample of n individuals obtained by a complex survey p. Let d_i for $i \in s$ be the design weights that satisfy

$$\mathrm{E}_p\left[\sum_{i\in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, Z_i)\right] = \sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, Z_i), \tag{3}$$

where E_p denotes expectation under the design p. We estimate $\boldsymbol{\beta}_N$ by solving estimating equations

$$\sum_{i\in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i) = \mathbf{0}. \tag{4}$$

In analytic uses of survey data, solving equation (4) simultaneously estimates the superpopulation parameter $\boldsymbol{\beta}$ and the finite population parameter $\boldsymbol{\beta}_N$ (Godambe and Thompson, 1986).

When X_i is subject to misclassification, we observe an error-contained version W_i. Different from additive measurement error in a continuous variable where the observed surrogate can be expressed as a summation of the true measurement plus a noise term, the misclassification process is expressed in terms of (mis)classification probabilities. Let

$\tau_{k,l}(\mathbf{Z}_i) = \Pr(W_i = l | X_i = k, \mathbf{Z}_i)$ be the probability of classifying individual i into the lth category when the true one is k $(k, l = 1, \ldots, K)$. It is reasonable to assume that misclassification only occurs between adjacent categories of an ordinal variable (e.g., BMI category). Under this assumption, $\tau_{k,l}(\mathbf{Z}_i) = 0$ for $|k - l| > 1$. Therefore, the probability of correctly observing X_i in the k-th category is given by $\tau_{k,k}(\mathbf{Z}_i) = 1 - \tau_{k,k-1}(\mathbf{Z}_i) - \tau_{k,k+1}(\mathbf{Z}_i)$. A useful model to describe the process is the multinomial logit model (Pfeffermann *et a*l., 1998) given by

$$\log\{\tau_{k,k-1}(\mathbf{Z}_i)/\tau_{k,k}(\mathbf{Z}_i)\} = \mathbf{Z}_i^T \boldsymbol{\gamma}_{k,k-1}, \qquad k = 2, \ldots, K, \tag{5}$$

$$\log\{\tau_{k,k+1}(\mathbf{Z}_i)/\tau_{k,k}(\mathbf{Z}_i)\} = \mathbf{Z}_i^T \boldsymbol{\gamma}_{k,k+1}, \qquad k = 1, \ldots, K-1, \tag{6}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{12}^T, \boldsymbol{\gamma}_{21}^T, \ldots, \boldsymbol{\gamma}_{K,K-1}^T)^T$ is a vector of parameters in the misclassification process. It is convenient to take a structural modeling approach, which hypothesizes a distribution for X_i conditional on $\mathbf{Z}_i$. Let $\lambda_k(\mathbf{Z}_i) = \Pr(X_i = k | \mathbf{Z}_i)$, $k = 1, \ldots, K$. The probability distribution of X_i follows a multinomial logit model given by

$$\log\{\lambda_k(\mathbf{Z}_i)/\lambda_K(\mathbf{Z}_i)\} = \mathbf{Z}_i^T \boldsymbol{\alpha}_k, \qquad k = 1, \ldots, K-1, \tag{7}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T)^T$ is a vector of regression parameters. When X_i has many levels and/or $\mathbf{Z}_i$ contains many discrete covariates, the dimension of $\boldsymbol{\alpha}$ can become very large. In that case we may assume independence between X_i and $\mathbf{Z}_i$ and only specify the marginal distribution of X_i, which is given by $\Pr(X_i = k) = \alpha_k$, where $\sum_{k=1}^K \alpha_k = 1$.

2.2 Estimation and Inference

When W_i rather than X_i is available in the data, $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)$ is no longer unbiased under the model ξ. Therefore, naïve inference based on equation (4) with X_i replaced by W_i is invalid. However, if we can obtain an estimating function, say, $\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)$, that approximates $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i)$ reasonably well around the true value of $\boldsymbol{\beta}$, then solving

$$\sum_{i \in s} d_i \mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i) = 0 \tag{8}$$

still yields a consistent estimator under the design and the model. One approximation can be given by the conditional mean of $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i)$ given observed data $(Y_i, W_i, \mathbf{Z}_i)$, i.e., $\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)$=$\mathrm{E}\left[\mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i) | Y_i, W_i, \mathbf{Z}_i\right]$, where E denotes expectation under all probability models but not the survey design. Let $\pi_{ik} = \Pr(X_i = k | Y_i, W_i, \mathbf{Z}_i)$ be the posterior probability of $\{X_i = k\}$. We have

$$\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i) = \sum_{k \in \Omega_i} \pi_{ik} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i = k, \mathbf{Z}_i), \tag{9}$$

where $\Omega_i = \{l : \max(W_i - 1, 1) \leq l \leq \min(W_i + 1, K)\}$ is a set of possible values for X_i conditional on W_i. Furthermore, π_{ik} can be expressed in terms of the response model, misclassification model, and the distribution of X_i, i.e.,

$$\pi_{ik} = \frac{\Pr(Y_i | X_i = k, \mathbf{Z}_i)\Pr(W_i | X_i = k, \mathbf{Z}_i)\Pr(X_i = k)}{\sum_l \Pr(Y_i | X_i = l, \mathbf{Z}_i)\Pr(W_i | X_i = l, \mathbf{Z}_i)\Pr(X_i = l)}. \tag{10}$$

Therefore, the expected score function $\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)$ involves nuisance parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, which are suppressed in the notation. We note here that the misclassification model is unidentifiable without additional information on X_i besides the surrogate W_i. This additional information can be a replicate measures study or a validation subsample in which bothX_i and W_i are available (Spiegelman *et al*., 2000). In the present paper we consider the last case. The sample s can be partitioned into three subsets: $s_1 = \{i : (Y_i, X_i, \mathbf{Z}_i)\}$, $s_2 = \{i : (Y_i, W_i, X_i, \mathbf{Z}_i)\}$, and $s_3 = \{i : (Y_i, W_i, \mathbf{Z}_i)\}$. An improved alternative to equation (9) can be given by

$$\sum_{i\in s_1\cup s_2} d_i\mathbf{U}_i(\boldsymbol{\beta};Y_i,X_i,\mathbf{Z}_i)+\sum_{i\in s_3}\sum_{k\in\Omega_i} d_i\pi_{ik}\mathbf{U}_i(\boldsymbol{\beta};Y_i,X_i=k,\mathbf{Z}_i)=0. \quad (11)$$

We use an iterative estimation procedure for solving (11). In the first step, we evaluate π_{ik} for all $k\in\Omega_i$ and $i\in s_3$ given a current estimate $\boldsymbol{\beta}^{(t)}$ and fixed values of $\boldsymbol{\gamma}$ an $\boldsymbol{\alpha}$. In the second step, we replace π_{ik} in equation (11) with the realized posterior probabilities, say, $\pi_{ik}^{(t)}$, and solve to obtain a new estimate $\beta^{(t+1)}$. The algorithm iterates between the two steps until convergence. Let $\hat{\boldsymbol{\beta}}$ denote the final estimate. This procedure is conditional on the nuisance $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, which can be estimated from s_2 and $s_1\cup s_2$, respectively. As an example, the marginal distribution of X_i can be estimated by $\hat{\alpha}_k=\left\{\sum_{i\in s_1\cup s_2} d_i\mathrm{I}(X_i=k)\right\}/\sum_{i\in s_1\cup s_2} d_i, k=1,\ldots,K.$

For variance estimation, we suggest using a replication approach such as bootstrap variance estimation (Efron, 1981). It is well known that the model-based covariance matrix for estimator $\hat{\boldsymbol{\beta}}$ is often invalid, as it does not take into account the sampling design. Because of the features of the complex survey and the non-response issue, design weights in each bootstrap sample need to be re-calculated to account for these issues. Let $\hat{\boldsymbol{\beta}}_{(b)}$ be the realization of $\hat{\boldsymbol{\beta}}$ using the b-th bootstrap sample, $b=1,\ldots,B$. Given $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, an approximate covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by

$$BSV\left(\hat{\boldsymbol{\beta}}\right)=\frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat{\boldsymbol{\beta}}_{(b)}-\hat{\boldsymbol{\beta}}\right\}\left\{\hat{\boldsymbol{\beta}}_{(b)}-\hat{\boldsymbol{\beta}}\right\}^T.$$

Since $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are estimated from internal validation data, extra variability should be accounted for when calculating the variance of $\hat{\boldsymbol{\beta}}$. In the present paper we treat the estimated nuisance parameters as known.

3. SIMULATION STUDY

In this section we conduct a limited simulation study to investigate the performance of the proposed method. We generate independent data from probability models without considering survey features. Each data set contains $n=100000$ subjects. A three-level ordinal covariate X_i, which is subject to misclassification, takes value at 1, 2, and 3 with probability .2, .5, and .3, respectively. An error-free continuous Z_i, which is independent of X_i, follows a normal distribution with mean 0 and variance 1. The binary response Y_i is generated under a logistic model given by

$$\mathrm{logit}(\mu_i)=\beta_0+\beta_1\mathrm{I}(X_i=1)+\beta_2\mathrm{I}(X_i=3)+\beta_3 Z_i. \quad (12)$$

We obtain W_i, namely the surrogate version of X_i, under a multinomial logit model

$$\log\{\tau_{k,l}(Z_i)/\tau_{k,k}(Z_i)\}=\gamma_{k,l(0)}+\gamma_{k,l(1)}Z_i, \qquad k=1,2,3,\ \ |k-l|=1. \quad (13)$$

The response parameters are given by $\beta_0=-3,\beta_1=.3,\beta_2=.5$ and $\beta_3=.5$. Thus, the probability of developing the outcome is higher for subjects in categories 1 and 3 than those in category 2. Misclassification parameters are given by $\gamma_{1,2(0)}=-1.5,\ \gamma_{1,2(1)}=-.05,\ \gamma_{2,1(0)}=-3,\ \gamma_{2,1(1)}=.05,\ \gamma_{2,3(0)}=-3,\ \gamma_{2,3(1)}=\ -.05,\ \gamma_{3,2(0)}=-1.5$, and $\gamma_{3,2(1)}=.5$. Thus, Z_i has a positive effect on increasing the probability of classifying subjects from higher categories into lower categories but has a positive effect on decreasing the probability of classifying subjects from lower categories into higher categories. A validation subsample is obtained by randomly selecting subjects with probability .04. We replicate the simulation 500 times and compare results from the proposed method to those from the naive analysis, which ignores misclassification, and the complete-cases analysis, which uses only the validation subsample.

Simulation results are shown in Table 1. The relative percentage bias (RB%) is small in the estimates of β_3 from all three approaches, although the naïve estimate is slightly larger than the other two. However, the naïve estimates of β_1 and β_2 are attenuated by 42.3% and 25.2%, respectively. The complete-cases approach and the proposed method perform reasonably well in estimating β_2. The bias in the estimate of β_1 from the complete-cases analysis is significantly large (attenuated by 4.42%) compared to that from the expected score approach, which is attenuated by 1.42%. Because both the naïve approach and the expected score approach use the whole sample, the empirical variances (var.) of the estimates are very small for all parameters. The empirical variances from the complete-cases analysis, however, are much larger due to the use of only the validation subsample.

Table 1: Simulation results for three approaches.

	Naïve		Complete-cases		Proposed	
Parameter	RB%	var.	RB%	var.	RB%	var.
β_0	-2.90	.0004	-.12	.0115	-.02	.0007
β_1	-42.31	.0011	-4.42	.0293	-1.42	.0020
β_2	-25.24	.0009	-.39	.0224	-.24	.0016
β_3	1.71	.0002	-.03	.0044	-.01	.0002

4. DATA ANALYSIS

We illustrate the proposed method using a data set from the Canadian Community Health Survey (CCHS) cycle 3.1. CCHS is an ongoing large scale survey conducted by Statistics Canada. Cycle 3.1 in 2005 targets persons aged 12 years or older who are living in private dwellings in the ten provinces and the three territories (Statistics Canada, 2005). Our interest is in the association of heart disease with risk factors such as age, sex, physical activity index (PAI), and body mass index (BMI) among adults in the Canadian population. We do not include subjects who were less than 18 years old, because children are in a stage of development where BMI may change over a short period of time (Connor Gorber *et al*., 2008). Based on Canadian guidelines, which are in line with those of the World Health Organization, BMI for adults is classified into six categories: underweight, normal weight, overweight, obese class I, obese class II, and obese class III, with corresponding BMI (kg/m^2) ranges <18.5, 18.5-24.9, 25.0-29.9, 30.0-34.9, 35.0-39.9 and 40.0+, respectively. As BMI was obtained by self-reporting for most of the respondents, the derived category may be different from the true category. A validation subsample containing both self-reported and measured BMI is available. We define four age categories: 18-34 (reference), 35-49, 50-64, and 65+. For this study, women who were pregnant or breastfeeding are excluded. Observations in the validation subsample with self-reported and measured BMI differ by more than one category are considered outliers. Subjects with missing correctly measured covariates or missing both the self-reported and the measured BMI values are also excluded from the analysis. This left a sample of 114547 respondents with 4125 in the validation subsample. The heart disease status follows the logistic regression model given by (1), and the misclassification model for BMI category is given by (5) and (6). The variance estimations are based on 500 bootstrap samples.

The results from the proposed method are reported in Table 2 with a comparison to those from the complete-cases analysis using only the validation subsample. We focus on the estimates of regression parameters for BMI categories. One can see that except for the overweight BMI category, all four other BMI categories are significantly associated with increased probability of developing heart disease compared to the normal weight BMI category. The risk increases as obese class increases. For instance, the odds of developing heart disease for subjects in obese class III is approximately $\exp(.96) = 2.61$ times that for normal weight subjects when controlling other factors. Age, sex, and physical activity index are also significantly associated with the heart disease. Due to a relatively smaller sample used in the complete-cases analysis, the standard error associated with estimates for BMI categories are very large. Therefore, a conclusion of insignificant association of heart disease with BMI from complete-cases analysis may be invalid.

Table 2: Analysis of CCHS cycle 3.1 data

		Complete-cases (n=4123)			Proposed (n=114370)		
Parameter		Estimate	SE	p-value	Estimate	SE	p-value
Intercept		-5.05	.65	$< .001$	-5.66	.12	$< .001$
BMI	Underweight	.49	.80	.539	.85	.27	.002
	Overweight	-.17	.25	.494	-.01	.09	.957
	Obese class I	.26	.33	.434	.48	.08	$< .001$
	Obese class II	.25	.42	.555	.51	.16	.001
	Obese class III	.88	.58	.126	.96	.15	$< .001$
Age	35-49	.16	.74	.832	.93	.13	$< .001$
	50-64	1.86	.69	.007	2.40	.11	$< .001$
	65+	3.18	.68	$< .001$	3.70	.11	$< .001$
Sex	Male	.69	.20	$< .001$	.47	.04	$< .001$
PAI	Active	.21	.33	.536	.16	.06	.070
	Inactive	.44	.29	.130	.23	.05	$< .001$

5. DISCUSSION

In this article we consider logistic regression analysis using survey data when an ordinal categorical covariate is subject to misclassification. We propose to use the expected score estimation method, which employs parametric models for the misclassification process as well as the error-prone ordinal covariate. A limited numerical study shows that the proposed method performs reasonably well. The implementation in SAS or R/S-plus is straightforward.

During the bootstrap variance estimation, nuisance parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are estimated from a validation subsample and are treated as known. Therefore, the standard error of $\hat{\boldsymbol{\beta}}$ is generally underestimated. One possible approach to accounting for the extra uncertainty is to obtain new estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ in each bootstrap sample. The main problem with this approach is that some bootstrap samples may not contain enough data from s_2 to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, especially when dimensions of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are large.

In some studies, the marginal distribution of X_i may be of interest. For example, frequencies of BMI categories in the target population can be one of the objectives in health surveys. We can simultaneously update the estimate of $\boldsymbol{\alpha}$ using survey weights and the posterior probabilities of imputed BMI categories, which is given by $\alpha_k^{(t+1)} = \left\{\sum_{i \in s_1 \cup s_2} d_i \mathrm{I}(X_i = k) + \sum_{i \in s_3} \sum_{k \in \Omega_i} d_i \pi_{ik}^{(t)}\right\} / \sum_{i \in s} d_i, k = 1, \ldots, K.$

Misclassification can also exist in the response variable. In our current analysis, the BMI category variable is used as a risk factor. However, BMI category can be viewed as a response variable, and association of obesity with risk factors such as food intake may be of interest. Our future research includes a generalization of the approach of Neuhaus (1999) for misclassified binary responses to misclassification in ordinal responses as well as to the context of surveys.

ACKNOWLEDGEMENTS

This project was supported by Statistics Canada and MITACS/ NPCDS. The authors would like to thank Milorad Kovacevic, Georgia Roberts and Ivan Carrillo-Garcia in Statistics Canada and Changbao Wu and Grace Y. Yi in the University of Waterloo for discussion.

REFERENCES

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

Connor Gorber, S., Shields, M., Tremblay, M.S. and McDowell, I. (2008). "The feasibility of establishing correction factors to adjust self-reported estimates of obesity". *Health Reports*, **19**(3), 71-82.

Efron, B. (1981). "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods". *Biometrika*, **68**, 589-599.

Fuller, W.A. (1995). "Estimation in the presence of measurement error". *International Statistical Review*, **63**, 121-141.

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

Neuhaus, J.M. (1999). "Bias and efficiency loss due to misclassified responses in binary regression". *Biometrika*, **86**, 843-855.

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.

Spiegelman, D., Rosner, B. and Logan, R. (2000). "Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs". *Journal of the American Statistical Association*, **95**, 51-61.

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

Suh, E.Y., and Schafer, D.W. (2002). "Semiparametric maximum likelihood for nonlinear regression with measurement errors". *Biometrics*, **58**, 448-453.

Yi, G.Y. (2008). "A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates". *Biostatistics*, **9**, 501-512.