Journal of Data Science **10**(2012), 157-173

Multivariate Logistic Regression Analysis of Complex Survey Data with Application to BRFSS Data

Minggen Lu* and Wei Yang
University of Nevada, Reno

Abstract: Multiple binary outcomes that measure the presence or absence of medical conditions occur frequently in public health survey research. The multiple possibly correlated binary outcomes may compose of a syndrome or a group of related diseases. It is often of scientific interest to model the interrelationships not only between outcome and risk factors, but also between different outcomes. Applied and practical methods dealing with multiple outcomes from complex designed surveys are lacking. We propose a multivariate approach based on the generalized estimating equation (GEE) methodology to simultaneously conduct survey logistic regressions for each binary outcome in a single analysis. The approach has the following attractive features: 1) It enables modeling the complete information from multiple outcomes in a single analysis; 2) it permits to test the correlations between multiple binary outcomes; 3) it allows of discerning the outcome-specific effect and the overall risk factor effect; and 4) it provides the measurement of difference of the association between risk factors and multiple outcomes. The proposed method is applied to a study on risk factors for heart attack and stroke in 2009 U.S. nationwide Behavioral Risk Factor Surveillance System (BRFSS) data.

Key words: Behavior Risk Factor Surveillance System (BRFSS), cardiovascular disease, generalized estimating equation (GEE), heart attack, odds ratio (OR), stroke.

1. Introduction

Key public health data such as BRFSS are often obtained through complex design, which involves stratification, clustering, multistage sampling, and unequal probability of selection of participants and responding rates. In order to make valid inference for the interested population where samples originated, appropriate statistical methods are required to analyze the complex survey data.

*Corresponding author.

For binary outcomes that measure the presence or absence of certain medical conditions, e.g. with or without a cardiovascular disease (CVD), survey logistic regression model is the standard approach to analyze the relationship between the binary dependent variable and a set of explanatory variables by incorporating the sample design information, including stratification, clustering, and unequal weighting (Greenlund *et al.*, 2004; Kim and Beckles, 2004). If multiple binary outcomes are assessed on the same individual, for example, several CVD related health conditions, including stroke and heart attack, they are subject to share the same characteristics to that individual and are likely to exhibit the correlations within the subject (Fitzmaurice *et al.*, 1995; France *et al.*, 2009; Sergeev and Carpenter, 2010). The standard survey logistic regression method is inapplicable for these situations.

In this manuscript, we propose a marginal multivariate approach using GEE techniques (Liang and Zeger, 1986; Prentice, 1988; Lipsitz *et al.*, 1991) to model the relationship of multiple responses with explanatory variables, and the association between pairs of responses in a single analysis by incorporating the sampling design adjustment for complex survey data. The first-order generalized estimating equations (Liang and Zeger, 1986) are implemented to describe the effects of explanatory variables on each binary response, while the odds ratio estimated by modified second-order estimating equations (Lipsitz *et al.*, 1991) is applied to characterize the degree of association between binary responses. Horton and Fitzmaurice (2004) proposed independence estimating equations for purposes of estimation and made an adjustment to standard errors of estimators to account for the correlation among the outcomes. Our method explicitly models the dependence among the outcomes, while Horton and Fitzmaurice (2004) regarded it as a nuisance feature of data, assuming working independence among outcomes. Modeling the dependence has the potential to yield more efficient estimators. The proposed method has several advantages: First, it captures the complete information about multiple outcomes in a single regression model; second, it allows the assessment of correlation between outcomes by taking into account the covariance structure of responses; third, it provides the test of outcome-specific effects and overall risk factor effect; fourth, it supplies the measurement of difference of associations between risk factors and multiple outcomes; finally, the degree of dependence between the responses from the same individual is measured by the odds ratio rather than the correlation, since the odds ratio is easy to interpret and provides a natural way for modeling the within-subject dependence for binary responses.

This paper is organized as follows. Section 2 describes the GEE method for complex survey sample with multiple correlated binary responses. Also, the comparison between univariate and multivariate estimations is discussed. In Section

3, the proposed method is applied to BFRSS data. Finally, Section 4 concludes with a discussion.

2. Methods

2.1 Complex Survey Data

In many epidemiological studies the source data arise from complex survey sample. There are three main features that need to be accounted in the analysis: (i) stratification, (ii) clustering, and (iii) sampling weight. For example, the BFRSS used a complex survey design with stratification, multi-stage clustering, and unequal sampling weights. It is very common in survey study to divide the population into distinct subpopulations, referred to as strata. Within each stratum, a separate sample is selected from the sampling units independently. The variance of the estimate will decrease if the sampling units within each stratum are homogeneous. Failure to account for the stratification in the analysis will result in overestimation of the standard error, and hence too wide confidence interval. The second common feature in complex survey data is called clustering. In clustering, the total population is divided into some groups (or clusters) and a sample of the groups is selected. In multistage clustering, the clusters selected at first stage are called primary sampling units or PSUs. Further sample selection occurs within PSUs and so on. In general, failure to account for the clustering in the analysis may lead to underestimation of variabilities. Finally, unequal selection in each PSU occurs in many epidemiological surveys. The sampling weight is used as the measure of how many units in the population which the sampled PSU represents. The unequal selection probabilities must be taken into account in analysis to reduce the bias of the estimate and the underestimation of variabilities.

2.2 GEE Approach for Complex Survey Data

In this section we consider extensions of the population-averaged marginal GEE approach for complex survey data. Assume the population is divided into H distinct strata. In each stratum h, the sample is consisted of n_h clusters and each cluster is comprised of m_{h_i} units, $h = 1, \cdots, H$, $i = 1, \cdots, n_h$. Let $Y_{hij} = (Y_{hij1}, \cdots, Y_{hijT})^T$ and $X_{hij} = (X_{hij1}, \cdots, X_{hijT})^T$ denote the $T \times 1$ vector of binary responses and the $T \times p$ matrix of covariate for hijth unit, respectively, $j = 1, \cdots, m_{h_i}$. The sampling weight for the hijth unit is denoted by ω_{hij}. Suppose that covariates X_{hijt} are associated with each marginal observation Y_{hijt}, $t = 1 \cdots T$. Let K be the sample size. We wish to estimate a logistic regression model in the marginal means

$$\text{logit } \Pr(Y_{hijt} = 1) = \text{logit}\pi_{hijt} = X_{hijt}^T\beta,$$

for $t = 1, \cdots, T$, where β is a p-dimensional vector of regression coefficients. It is assumed that the individuals Y_{hij} are independent, but the marginal observations $Y_{hij1}, \cdots, Y_{hijT}$ may be correlated, $h = 1, \cdots, H$, $i = 1, \cdots, n_h$, and $j = 1, \cdots, m_{h_i}$. We take into account the correlation among the multiple responses by applying the first-order generalized estimating equations proposed by Liang and Zeger (1986) to estimate regression coefficients β and standard errors se(β). The efficiency is expected to increase by utilization of the covariance structure of responses. We refer to $R(\alpha)$ as a working correlation matrix of $Y_{hij1}, \cdots, Y_{hijT}$, where α is a vector which fully characterizes $R(\alpha)$. Then the first estimating equations can be written as

$$U_1(\alpha, \beta) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{h_i}} \omega_{hij} D_{hij}^T V_{hij}^{-1} S_{hij} = 0, \tag{1}$$

where $D_i = \partial\pi_{hij}/\partial\beta = A_{hij}X_{hij}$, $V_{hij} = A_{hij}^{1/2}R(\alpha)A_{hij}^{1/2}$, $A_{hij} = \text{diag}\{\pi_{hijt}(1 - \pi_{hijt})\}$ and $S_{hij} = Y_{hij} - \pi_{hij}$. These generalized estimating equations yield consistent estimators of regression parameters β under the correct specification of the form π_{hij}. The covariance matrix of $U_1(\alpha, \beta)$, which is the negative expected value of $\partial U_1(\beta)/\partial\beta$, can be written as

$$I_1 = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{h_i}} \omega_{hij} D_{hij}^T V_{hij}^{-1} D_{hij}.$$

The correlation α is usually estimated by a $T_{hij} \times (T_{hij} - 1)/2$ vector of empirical correlations with elements

$$r_{hijst} = \frac{(Y_{hijs} - \pi_{hijs})(Y_{hijt} - \pi_{hijt})}{\sqrt{\pi_{hijs}(1 - \pi_{hijs})\pi_{hijt}(1 - \pi_{hijt})}},$$

and then use a second set of equations similar to (1). In this manuscript, we apply the odds ratio method proposed by Lipsitz *et al.* (1991) to measure the association between pairs of binary responses. Let $Z_{hij} = \{Z_{hijst}\}$ be a $T_{hij} \times (T_{hij} - 1)$ random vector, where

$$Z_{hijst} = I[Y_{hijs} = 1, Y_{hijt} = 1] = Y_{hijs}Y_{hijt},$$

and $I[\cdot]$ is an indicator function. The joint probability of success for Y_{hijs} and Y_{hijt} is

$$\pi_{hijst} = E(Z_{hijst}) = \Pr[Y_{hijs} = 1, Y_{hijt} = 1].$$

Now the odds ratio, τ_{hijst}, between the binary responses Y_{hijs} and Y_{hijt} can be written as a function π_{hijs}, π_{hijt}, and π_{hijst}

$$\tau_{hijst} = \frac{\pi_{hijst}(1 - \pi_{hijs} - \pi_{hijt} + \pi_{hijst})}{(\pi_{hijs} - \pi_{hijst})(\pi_{hijt} - \pi_{hijst})}.$$

Solving for π_{hijst} in terms of odds ratios τ_{hijst} and the two marginal probabilities π_{hijs} and π_{hijs} leads to

$$\pi_{hijst} = \begin{cases} \dfrac{f_{hijst} - [f_{hijst}^2 - 4\tau_{hijst}(\tau_{hijst} - 1)\pi_{hijs}\pi_{hijt}]^{1/2}}{2(\tau_{hijst} - 1)}, & \text{if } \tau_{hijst} \neq 1, \\ \pi_{hijs}\pi_{hijt}, & \text{if } \tau_{hijst} = 1, \end{cases}$$

where $f_{hijst} = 1 - (1 - \tau_{hijst})(\pi_{hijs} + \pi_{hijt})$. Now we model the logarithm of odds ratio as the linear combination of α

$$\log \tau_{hijst} = e_{hijst}^T \alpha_{st},$$

$1 \leq s < t \leq T$. Note that $\pi_{hijst} = E(Z_{hijst})$ is a function of β through marginal means π_{hijs} and π_{hijft} and α through odds ratios τ_{hijst}. The first derivative of $\pi_{hij} = E(Z_{hij})$ with respect to parameter α can be expressed as

$$\frac{\partial \pi_{hijst}}{\partial \alpha} = \left(\frac{\pi_{hijs} + \pi_{hijt} - A_{hijst}^{1/2}\eta_{hijst}}{2(\tau_{hijst} - 1)} - \frac{f_{hijst} - A_{hijst}^{1/2}}{2(\tau_{hijst} - 1)^2} \right) \frac{\partial \tau_{hijst}}{\partial \alpha},$$

where

$$\begin{aligned} \eta_{hijst} &= f_{hijst}(\pi_{hijs} + \pi_{hijt}) - (4\tau_{hijst} - 2)\pi_{hijs}\pi_{hijt}, \\ A_{hijst} &= f_{hijst} - 4\tau_{hijst}(\tau_{hijst} - 1)\pi_{hijs}\pi_{hijt}. \end{aligned}$$

Then the second set of estimating equations is

$$U_2(\alpha, \beta) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{h_i}} \omega_{hij} C_{hij}^T W_{hij}^{-1} Q_{hij} = 0, \tag{2}$$

where $C_{hij} = \dfrac{\partial \pi_{hijst}}{\partial \alpha}$, $W_{hij} = \mathrm{diag}\{\pi_{hijst}(1 - \pi_{hijst})\}$, $Q_{hij} = Z_{hij} - \theta_{hij}$, and θ_{hij} is the model for $E(Z_{hij})$. Similarly, the covariance matrix of $U_2(\alpha, \beta)$ can be written as

$$I_2 = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{h_i}} \omega_{hij} C_{hij}^T W_{hij}^{-1} C_{hij}.$$

Let $(\hat{\alpha}, \hat{\beta})$ be solutions of (1) and (2). By Taylor expansion, the valid estimators of variances of $\hat{\alpha}$ and $\hat{\beta}$ that take into accounts the stratification, clustering, and unequal selection probability are provided by

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = \hat{I}_1^{-1} \hat{V}_1 \hat{I}_1^{-1}, \tag{3}$$

$$\widehat{\mathrm{Cov}}(\hat{\alpha}) = \hat{I}_2^{-1} \hat{V}_2 \hat{I}_2^{-1}, \tag{4}$$

where,

$$\begin{aligned}
\hat{V}_1 &= \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (B_{hi} - \bar{B}_{hi})(B_{hi} - \bar{B}_{hi})^T, \\
\hat{V}_2 &= \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (A_{hi} - \bar{A}_{hi})(A_{hi} - \bar{A}_{hi})^T, \\
B_{h_i} &= \sum_{j=1}^{m_{h_i}} \omega_{hij} X_{hij}^T \hat{A}_{hij}^{1/2} \hat{R}^{-1} \hat{A}_{hij}^{1/2} X_{hij}, \\
A_{h_i} &= \sum_{j=1}^{m_{h_i}} \omega_{hij} \hat{C}_{hij}^T \hat{W}_{hij}^{-1} \hat{C}_{hij}, \\
\bar{A}_{h_i} &= 1/n_h \sum_{i=1}^{n_h} A_{h_i}, \quad \bar{B}_{h_i} = 1/n_h \sum_{i=1}^{n_h} B_{h_i}.
\end{aligned}$$

The estimates $\hat{I}_1$, $\hat{I}_2$, $\hat{A}_{hij}$, $\hat{C}_{hij}$, $\hat{W}_{hij}$ and $\hat{R}$ are evaluated at $(\hat{\alpha}, \hat{\beta})$. Note that (3) and (4) are similar to those advocated by Liang and Zeger (1986) and Lipsitz *et al.* (1991), except that (3) and (4) account for unequal selection probability and use V_1 and V_2 , the pooled within-stratum estimators of U_1 and U_2. These variance estimators are robust in the sense of being consistent even if the working covariances are misspecified.

2.3 Computational Implementation

To compute the estimators $(\hat{\alpha}, \hat{\beta})$ of (1) and (2), a Fisher-scoring type iterative algorithm can be applied with starting values (α_0, β_0) for (α, β). Let the mth step estimate be $(\alpha^{(m)}, \beta^{(m)})$. The $(m + 1)$th step estimates are

$$\beta^{(m+1)} = \beta^{(m)} - \left(\sum_{h,i,j} D_{hij}^{(m)T} V_{hij}^{(m)-1} D_{hij}^{(m)} \right)^{-1} \left(\sum_{h,i,j} D_{hij}^{(m)T} V_{hij}^{(m)-1} (Y_{hij} - \pi_{hij}^{(m)}) \right),$$

$$\alpha^{(m+1)} = \alpha^{(m)} - \left(\sum_{h,i,j} C_{hij}^{(m)T} W_{hij}^{(m)-1} C_{hij}^{(m)} \right)^{-1} \left(\sum_{h,i,j} C_{hij}^{(m)T} W_{hij}^{(m)-1} (U_{hij} - \theta_{hij}^{(m)}) \right),$$

where $D_{hij}^{(m)}$, $C_{hij}^{(m)}$, $V_{hij}^{(m)}$, $W_{hij}^{(m)}$, $\pi_{hij}^{(m)}$ and $\theta_{hij}^{(m)}$ are evaluated at $(\alpha^{(m)}, \beta^{(m)})$. The iteration converges at $(m+1)$th step if

$$\frac{|\beta^{(m+1)} - \beta^{(m)}|}{|\beta^{(m)}| + 10^{-6}} < \varepsilon = 10^{-8} \quad \text{and} \quad \frac{|\alpha^{(m+1)} - \alpha^{(m)}|}{|\alpha^{(m)}| + 10^{-6}} < \varepsilon = 10^{-8}.$$

Denote $D = (D_1^T, \cdots, D_K^T)^T$, $C = (C_1^T, \cdots, C_K^T)^T$, $S = (S_1^T, \cdots, S_K^T)^T$ and $Q = (Q_1^T, \cdots, Q_K^T)^T$. Let V and W be block diagonal matrices with V_{hij} and W_{hij} as the diagonal elements, respectively. Then the iterative procedure can be written as

$$\begin{aligned} \beta^{(m+1)} &= (D^T V^{-1} D)^{-1} D^T V^{-1} (D\beta^{(m)} - S), \\ \alpha^{(m+1)} &= (C^T W^{-1} C)^{-1} C^T W^{-1} (C\alpha^{(m)} - Q). \end{aligned}$$

Define the modified outcomes $Z_1 = D\beta - S$ and $Z_2 = C\alpha - Q$. Then the iterative procedure for calculating $\hat{\beta}$ and $\hat{\alpha}$ is equivalent to performing iteratively reweighted regression of Z_1 on D with weight V^{-1} and regression of Z_2 on C with weight W^{-1}. The algorithm was implemented in R 2.13.0.

2.4 Comparison between Univariate and Multivariate Estimates

For correlated binary responses arising from complex survey sample, the standard survey logistic regression is inapplicable. To address this issue, one approach (Greenlund *et al.*, 2005; Hayes *et al.*, 2006) is to pool multiple binary outcomes into a single categorical outcome and then regress the new summary response on predictors. The main limitation in this approach is that the relationship between specific outcome and predicators can be masked by combining the indicators of multiple processes. Another approach (Koziol-McLai *et al.*, 2001; Cumyn *et al.*, 2009; Bitton *et al.*, 2010) is to fit separate logistic models for each outcome. Although the effects of covariates on each outcome can be discerned, this approach is inability to test the difference of risk factor effects on multiple outcomes and the association between multiple responses.

To account for the possible paired correlations, there are three main approaches. The first approach (Horton and Fitzmaurice, 1994) is to assume that the observations are independent for the purposes of estimation, but make a suitable adjustment of standard errors for the possible correlation among the observations. The estimator of β is called independence estimating equation (IEE) estimator. Note that the IEE estimator is equivalent to the univariate estimator with a robust variance correction. Liang and Zeger (1986) showed that the IEE estimator is consistent for β and asymptotically normally distributed. The IEE estimator is preferable, since it is easily implemented in existing statistical packages. The second approach is to devise a generalized estimating equation proposed by Liang and Zeger (1986) and Prentice (1988), which generalize the independent estimating equation to incorporate a working correlation, or some other association such as odds ratio. Under the assumption of correct specification of mean functions, the GEE estimator is consistent and asymptotically normal. The third approach is to completely specify the joint distribution of the observations, by introducing the extra parameters for the association among the observations; see, for example, Rosner (1984), Prentice (1986), Liang and Zeger (1989), and Lipsitz *et al.* (1990) among others.

The asymptotic relative efficiency of the IEE estimator and the GEE estimator has been discussed extensively (Emrich and Piedmonte, 1992; Sharples and Breslow, 1992; Lee *et al.*, 1993; Mancl and Leroux, 1996). McDonald (1993) considered bivariate data with possibly different covariates for each binary observation and recommended the IEE estimator for practical purpose whenever the association between the paired observations is a nuisance. Zhao *et al.* (1992) suggested that incorrectly assuming independence can lead to important losses of efficiency when the correlation between responses is high. Fitzmaurice (1995) demonstrated that the asymptotic relative efficiency depends on not only the strength of correlation between the paired responses, but also the covariate distribution, such as between-cluster and within-cluster covariate designs; that is, a covariate is constant or varies within the cluster. In this manuscript, we adopt the within-cluster covariate design, in which a within-cluster indicator variable is used to fit a survey logistic regression of multiple responses simultaneously. We assume the mean structure is correctly specified, but allow the correlation among the responses to be misspecified, including by incorrectly assuming independence. Following the approach of Fitzmaurice (1995), we can show that when the responses are strongly correlated and the covariate distribution is within cluster, the IEE estimator which assumes the independence can lead to considerable loss of efficiency, compared to the GEE estimator for complex survey sample. Therefore, for correlated binary responses arising from complex survey sample, we adopt the GEE estimator which is relatively straightforward, compared to the full likelihood estimator, and is more efficient than the IEE estimator for within-cluster covariate design.

3. Data Analysis: Application to BRFSS Data

3.1 Sample Data

The BRFSS is a state-based, random-digit telephone survey of the U.S. noninstitutionalized, civilian population. Self-reported data from 353,280 people aged over 18 years who participated in the 2009 BRFSS were collected. The response rate, based on the Council of American Survey Research Organization (CASRO) guidelines (White, 1984), was 52.48% ranging from 37.90% to 66.85% among states and territories for the 2009 BRFSS. Relative to other surveys, data from BRFSS have acceptable reliability and validity (Bowlin *et al.*, 1993; Nelson *et al.*, 2001). BRFSS data analysis involves the data weighting process which attempts to remove the bias in the sample probability of selection due to nonresponse and noncoverage errors, as well as to adjust variables of age, race, and gender between the sample and the entire population. Weight factors included are the number of residential telephones in a household, the number of adults in a household, and geographic or density stratification. Information on quality assurance and other aspects of this survey is available online at www.cdc.gov/brfss/index.htm. There are two dependent binary variables, heart attack and stroke which were collected as "Have you ever told by a healthcare professional that you had a heart attack (myocardial infarction)?" and "Have you ever told by a healthcare professional that you had a stroke?".

3.2 Multivariate Survey Logistic Regression Analysis

We apply the proposed multivariate survey logistic regression method to U.S. 2009 BRFSS data. In addition to using covariates, such as age, sex, education level, marital status, overweight or obesity, and annual income as independent variables in the model, we also include an indicator variable in the model to allow us to fit a single logistic regression model to the data of multiple responses simultaneously. To simplify the notations, we omit subscriptions for stratum and cluster. Assume I is the indicator variable, where $I = 0$ stands for stroke and $I = 1$ stands for heart attack. Gender is defined as 1 for males and 0 for females. Age is defined as 1 for those aged 55 or older and 0 otherwise. Race is defined as 1 for non-Hispanic white and 0 otherwise. Education is defined as 1 for those attended some college or technical school or graduated from college and 0 otherwise. Overweight or obesity is set to be 1 for body mass index (BMI) greater than or equal to 25 and 0 for BMI less than 25. Marital status is defined as 1 for those are married or a member of unmarried couple, 2 for those who are divorced or widowed or separated, and 3 for those who are never married. Annual income is defined as 1 for those whose annual household income is less than \$35,000, 2 for those whose income is between \$35,000 and \$75,000, 3 for those whose income is above \$75,000. We use two dummy variables for martial status and annual income. Let X be a set of covariates defined above and $\beta = (\beta_0, \cdots, \beta_{10})^T$ and $\gamma = (\gamma_1, \cdots, \gamma_{10})^T$ be the corresponding regression parameters for X and IX, respectively. Then the bivariate survey logistic regression model in the matrix form is

$$\log \frac{\pi}{1-\pi} = \beta_0 + X^T\beta + I\gamma_0 + IX^T\gamma,$$

$$\log OR = \alpha,$$

where π is the probability of a positive response and OR represents the odds ratio between binary outcomes stroke and heart attack, conditional on covariates. The regression parameters for stroke are $\beta_0, \cdots, \beta_{10}$, while $\beta_0 + \gamma_0, \cdots, \beta_{10} + \gamma_{10}$ are corresponding coefficients for heart attack. The parameters for this interaction model are easy to interpret. For the purpose of illustration, we only interpret risk factors age and gender. The other covariates can be interpreted similarly. The response-specific log odds for stroke and heart attack are β_0 and $\beta_0 + \gamma_0$, respectively. The parameters β_1 and $\beta_1 + \gamma_1$ are the average log odds ratios for stroke and heart attack, respectively, for those who are aged fifty five or older, compared to younger individuals. Likewise, β_2 and $\beta_2 + \gamma_2$ are the average log odds ratios for stroke and heart attack, respectively, for males compared to females. The differences of log odds ratios between heart attack and stroke for age and gender are measured by γ_1 and γ_2, respectively. The positivity of γ_1 indicates that the effect of age on heart attack is greater than that on stroke. Likewise, the positivity of γ_2 demonstrates that the odds ratio of heart attack is greater than that of stroke for the risk factor gender. If the effects of risk factors do not vary by multiple responses, the interaction terms may be removed, and the overall effect to responses can be estimated. For example, the model

$$\log \frac{\pi}{1-\pi} = \beta_0 + X^T \beta$$

assumes that the effects of covariates X do not vary by the responses. The heart attack and stroke odds ratio, $\exp(\gamma_0)$, measures the extent to which individuals are more likely to contract heart attack than stroke. A value greater than 1 indicates that heart attack is more likely to occur than stroke, conditional on covariates. On the other hand, the log odds ratio α can be used to compare the strength of association between two binary responses. A value greater than 0 suggests positive association between stroke and heart attack, after adjusting for covariates. Since $\exp(\alpha)$ is estimated with covariates, it can be considered as adjusted for risk factors. The adjusted odds ratio is usually smaller than the corresponding crude odds ratio.

3.3 Results

The results of estimated regression coefficients and standard errors from univariate survey logistic regression model, pooled survey logistic regression model, and bivariate survey logistic regression model are presented in Tables 1 and 2. The estimated odds ratios are based on the estimated coefficients reported in Tables 1 and 2. Define a summary heart disease response variable HD = 1 if the subject was diagnosed with stroke or heart attack or both and 0 otherwise for

Table 1: Results of univariate survey logistic regressions for BFRSS heart disease study. Stroke and heart attack are binary outcomes in models 1 and 2, respectively. In pooled model, the binary outcome is stroke or heart disease or both. * represents p-value < 0.05

	Model 1	Model 2	Pooled Model
Parameter	Estimate ± SE	Estimate ± SE	Estimate ± SE
Intercept	-4.262±0.071*	-4.448±0.066 *	-3.831 ± 0.054*
Age	1.622±0.046*	1.784±0.042 *	1.769 ± 0.033*
Gender	0.117±0.038*	0.804±0.033 *	0.553 ± 0.027*
Race	-0.014±0.046	0.171±0.043 *	0.122 ± 0.035*
Education	-0.086±0.039*	-0.166±0.033 *	-1.134 ± 0.028*
Widowed/divorced/separated	0.386±0.041*	0.239±0.034 *	0.305 ± 0.029*
Never married	-0.382±0.096*	-0.711±0.075 *	-0.577 ± 0.065*
Overweight/obesity	0.073±0.039	0.281±0.034 *	0.211 ± 0.029*
Smoking	0.362±0.048*	0.336±0.043 *	0.361 ± 0.036*
Middle income	-0.665±0.046*	-0.559±0.037 *	-0.586 ± 0.032*
High income	-1.149±0.063*	-1.085±0.050 *	-1.101 ± 0.042*

Table 2: Results of bivariate survey logistic regression for BFRSS heart disease study. * represents p-value < 0.05. ** represents the estimates of interactions of covariates and I, where I is defined as 0 for stroke and 1 for heart attack

	Stroke	Heart Attack	Interaction**
Parameter	Estimate ± SE	Estimate ± SE	Estimate ± SE
Intercept	-4.268±0.072*	-4.445±0.067*	-0.177±0.089*
Age	1.628±0.046*	1.782±0.042*	0.153±0.056*
Gender	0.120±0.038*	0.803±0.032*	0.683±0.046*
Race	-0.009±0.046	0.168±0.043*	0.178±0.057*
Education	-0.088±0.039*	-0.168±0.033*	-0.080±0.047
Widowed/divorced/separated	0.384±0.041*	0.240±0.034*	-0.144±0.048*
Never married	-0.381±0.097*	-0.710±0.077*	-0.329±0.114*
Overweight/obesity	0.070±0.040	0.282±0.034*	0.212±0.048*
Smoking	0.359±0.048*	0.339±0.043*	-0.021±0.058
Middle income	-0.661±0.046*	-0.559±0.037*	0.103±0.055
High income	-1.143±0.064*	-1.082±0.051*	0.061±0.075

pooled survey logistic regression. The univariate survey logistic regression model and the pooled survey logistic regression model were fitted using SAS SURVEYLOGISTC Procedure in SAS 9.2 software, while the bivariate survey logistic regression model (GEE model) was fitted in R 2.13.0 using a Fisher-scoring type algorithm to estimate the regression coefficients and the corresponding odds ratios. The pooled survey logistic regression yielded that age (OR = 5.86, $p < .0001$), male (OR = 1.74, $p < .0001$), non-Hispanic white (OR = 1.13, $p = 0.0005$), widowed or divorced or separate (OR = 1.36, $p < .0001$), overweight or obesity (OR = 1.24, $p < .0001$), and smoking (OR = 1.44, $p < .0001$) are significant risk factors for heart disease, while high education level (OR = 0.87, $p < .0001$), never married (OR = 0.56, $p < .0001$), middle annual income (OR = 0.56, $p < .0001$), and high annual income (OR = 0.33, $p < .0001$) appeared to be significant beneficial factors for heart disease. The univariate survey logistic regression models generated similar results to those of the pooled model, except that neither gender (OR = 0.98, $p = 0.753$) nor overweight or obesity (OR = 1.08, $p = 0.065$) has significant effect on stroke. The bivariate model demonstrated similar estimates of the relationships between the covariates and binary outcomes (Table 2). Moreover, the bivariate model showed that the effects of age (OR = 5.94 vs. OR = 5.09), gender (OR = 2.23 vs. OR = 0.99), race (OR = 1.18 vs. OR = 1.13), and overweight or obesity (OR = 1.33 vs. OR = 1.07) on heart attack are significantly stronger than those on stroke, with p-values 0.006, < .0001, 0.002, and < .0001, respectively, while the models demonstrated that effects of widowed or divorced or separate (OR = 1.27 vs. OR = 1.47) and never married (OR = 0.49 vs. OR = 0.68) have significantly smaller effects on heart attack than those on stroke, with p-values 0.003 and 0.004, respectively. The bivariate model did not reject the possibility that there are no differences of odds ratios between heart attack and stroke in education level (OR = 0.86 vs. OR = 0.91), smoking (OR = 1.40 vs. OR = 1.43), middle annual income (OR = 0.57 vs. OR = 0.52), and high annual income (OR = 0.33 vs. OR = 0.32), with p-values 0.086, 0.721, 0.062, and 0.414, respectively. The GEE model estimated the odds ratio for within cluster (within-subject) dependence, i.e., $\exp(\alpha)$, to be 7.53, $p < .0001$, after adjusting for covariates.

4. Discussion

A health survey is often conducted to obtain information about the prevalence of diseases and unhealthy behaviors, exposures to potential risk factors, and cost and utilization of health-care services of a population. Binary outcomes that measure the presence or absence of certain medical conditions are common in survey research. Many survey studies measure a vector of health conditions to make an overall assessment about an individual health status. It is often an important task for health researchers to exam the relationship between multiple health conditions and predicators, including behaviors and socioeconomic measures in complex surveys. Since multiple binary outcomes are obtained from the same individual, they are likely to be correlated within the subject. Ignoring the interrelations among the multiple outcomes essentially will lead to inefficient estimates in statistical analysis. It is neither scientifically appropriate nor statistically efficient to fit separate logistic models for each binary outcome. Therefore, the standard univariate survey logistic regression method is inapplicable for complex survey data with multiple binary outcomes.

Most of the statistical approaches that simultaneously model the correlated binary response variables can be grouped into two groups: population-averaged marginal modeling using generalized estimating equations and cluster-specific hierarchical modeling using generalized linear mixed models (Stiratelli *et al.*, 1984; Breslow and Calyton, 1993; Fahrmeir and Tutz, 1994; Das *et al.*, 2004; Molenberghs and Verbeke, 2005). There are several advantages to utilize the population averaged marginal modeling approach. From a public health perspective, the GEE model provides population averaged estimates of the relationships between risk factors and clinic outcomes. Furthermore, the GEE approach is robust to model misspecification in terms of the covariance structure among the multiple responses. Moreover, compared with the cluster-specific modeling approach involving intensive multidimensional integral computation, the GEE estimation algorithm is relatively computationally efficient for large data sets (in our data, $n = 353,280$) and widely implemented in standard statistical software, such as SAS, SPSS, Splus/R, STATA, and SUDAAN.

Different from previous approaches when dealing with correlated binary responses arising from complex survey sample, e.g. combining multiple binary outcomes into a single categorical or using separated logistic models, our approach utilized GEE techniques to model the relationships between multiple clinical outcomes and risk factors, and the degree of dependence between the outcomes simultaneously. The results from BRFSS sample data can not only provide detailed odds ratios for the predicting variables to heart attack and stroke, either from univariate or bivariate survey logistic models, but also estimate the odds ratio for within cluster (within-subject) dependence (e.g. $\exp(\alpha)$, to be 7.53, $p < .0001$).

Our multivariate approach has added following advantages to analyze complex survey data. First, we may gain more precision by utilizing all information about multiple outcomes in a single analysis, instead of modeling separate logistic regression for each binary outcome; second, simultaneous modeling of all the outcomes allows not only to provide the test of outcome-specific effects and the overall risk factor effect, but also to characterize the association between the pairs of outcomes by taking into account the covariance structure of responses; finally, it permits the assessment of differences of associations between risk factors and multiple outcomes. This paper has illustrated the use of multivariate logistic interaction model as a flexible method for analysis of complex survey data involving multiple binary outcomes. The parameters in this interaction model have simple interpretations. The effects of covariates on specific outcome can be expressed as odds ratios, which are preferable for binary outcomes. The multivariate logistic regression interaction model has the advantage of allowing of testing whether the effects of risk factors vary by responses. When the outcome-risk factor interaction term is present, separate outcome specific risk factor estimates can be computed. The difference of log odds ratio between two outcomes for specific risk factor can also be estimated and may have important interpretations. For example, the stronger association between sex and heart attack was observed than that between sex and stroke in BFRSS study.

The multivariate regression approach proposed in this paper can be generalized to more general settings where multiple outcomes are measured from the same subject and the distributions of outcomes are exponential family. In fact, it is applicable to continuous, categorical, and count response variables by changing the link function of expected value of the outcome. For continuous and count variables, the link function would be identity function and logarithm function, respectively. It would also be interesting to adopt cluster specific hierarchical modeling using generalized linear mixed models (GLMM) to multistage stratified cluster sampling, especially in the situation that the estimation of effects of within-cluster covariates is of interest, and study the differences in conceptualization and interpretation between the population-averaged model and the cluster specific model for complex survey data.

Acknowledgements

This study was supported by University of Nevada, Reno, Junior Faculty Research Grant for first author.

References

Bitton, A., Zaslavsky, A. M. and Ayanian, J. Z. (2010). Health risks, chronic diseases, and access to care among US pacific islanders. *Journal of General Internal Medicine* **25**, 435-440.

Bowlin, S. J., Morrill, B. D., Nafziger, A. N., Jenkins, P. L., Lewis, C. and Pearson, T. A. (1993). Validity of cardiovascular disease risk factors assessed by telephone survey: the behavioral risk factor survey. *Journal of Clinical Epidemiology* **46**, 561-571.

Breslow, N. E. and Calyton, D. G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association* **88**, 9-25.

Cumyn, L., French, L. and Hechtman, L. (2009). Comorbidity in adults with attention-deficit hyperactivity disorder. *Canadian Journal of Psychiatry* **54**, 673-683.

Das, A., Poole, W. K. and Bada, H. S. (2004). A repeated measure approach for simultaneous modeling of multiple neurobehavioral outcomes in newborn exposed to cocaine in utero. *American Journal of Epidemiology* **159**, 891-899.

Emrich, L. J. and Piedmonte, M. R. (1992). On some small sample properties of generalized estimating equations for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation* **41**, 19-29.

Fahrmeir, L and Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models*, 2nd edition. Springer, New York.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**, 309-317.

Fitzmaurice, G. M., Laird, N. M., Zahner, G. E. P. and Daskalakis, C. (1995). Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. *American Journal of Epidemiology* **142**, 1194-1203.

France, E. and Velanovich, V. (2009). The relative influence of surgical disease and co-morbidities on patient responses to a generic health-related quality-of-life instrument. *American Surgeon* **75**, 1084-1090.

Greenlund, K. J., Zheng, Z. J., Keenan, N. L., Giles, W. H., Casper, M. L., Mensah, G. A. and Croft, J. B. (2004). Trends in self-reported multiple cardiovascular disease risk factors among adults in the United States, 1991-1999. *Archives of Internal Medicine* **164**, 181-188.

Greenlund, K. J., Denny, C. H., Mokdad, A. H., Watkins, N., Croft, J. B. and Mensah, G. A. (2005). Using behavioral risk factor surveillance data for heart disease and stroke prevention programs. *American Journal of Preventive Medicine* **29**, 81-87.

Hayes, D. K., Denny, C. H., Keenan, N. L., Croft, J. B., Sundaram, A. A. and Greenlund, K. J. (2006). Racial/ethnic and socioeconomic differences in multiple risk factors for heart disease and stroke in women: behavioral risk factor surveillance system. *Journal of Women's Health* **15**, 1000-1008.

Horton, J. P. and Fitzmaurice, G. M. (2004). Regression analysis of multiple source and multiple informant data from complex survey samples. *Statistics in Medicine* **23**, 2911-2933.

Kim, C. and Beckles, G. L. (2004). Cardiovascular disease risk reduction in the behavioral risk factor surveillance system. *American Journal of Preventive Medicine* **27**, 1-7.

Koziol-McLai, J., Coates, C. J. and Lowenstein, S. R. (2001). Predictive validity of a screen for partner violence against women. *American Journal of Preventive Medicine* **21**, 93-100.

Lee, A. J., Scott, A. J. and Soo, S. C. (1993). Comparing Liang-Zeger estimates with maximum likelihood in bivariate logistic regression. *Journal of Statistical Computation and Simulation* **44**, 133-148.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Liang, K. Y. and Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series. *Journal of the American Statistical Association* **84**, 447-451.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1990). Maximum likelihood regression methods for paired binary data. *Statistics in Medicine* **9**, 1517-1525.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* **78**, 153-160.

Mancl, L. A. and Leroux, B. G. (1996). Efficiency of regression estimates for clustered data. *Biometrics* **52**, 500-511.

McDonald, B. W. (1993). Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society, Series B* **55**, 391-397.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* Springer, Verknpfüngen.

Nelson, D. E., Holtzman, D., Bolen, J., Stanwyck, C. A. and Mack, K. A. (2001). Reliability and validity of measures from the Behavioral Risk Factor Surveillance System (BRFSS). *International Journal of Public Health* **46**, S3-S42.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of American Statistical Association* **81**, 321-327.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

Rosner, B. (1984). Multivariate methods in opthalmology with application to other paired-data situations. *Biometrics* **40**, 1025-1035.

Sergeev, A. V. and Carpenter, D. O. (2010). Residential proximity to environmental sources of persistent organic pollutants and first-time hospitalizations for myocardial infarction with comorbid diabetes mellitus: a 12-year population-based study. *International Journal of Occupational and Environmental Health* **23**, 5-13.

Sharples, K. and Breslow, N. E. (1992). Regression analysis of correlated binary data: some small sample results for estimating equations. *Journal of the Statistical Computation and Simulation* **42**, 1-20.

Stiratelli, R., Laird, N. M. and Ware, J. H. (1984). Random effects models for serial observations with binary response. *Biometrics* **40**, 961-971.

White, A. A. (1984). Response rate calculation in RDD telephone health surveys: current practices. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, District of Columbia.

Zhao, L. P., Prentice, R. L. and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B* **54**, 805-811.

Received July 27, 2011; accepted October 4, 2011.

Minggen Lu
School of Community Health Sciences
University of Nevada, Reno
Reno, Nevada, USA
minggenl@unr.edu

Wei Yang
School of Community Health Sciences
University of Nevada, Reno
Reno, Nevada, USA
weiyang@unr.edu