Using Design-based and Model-based method to study the relationship of asthma and smoking using longitudinal survey data

S. Ghosh[1], P. Pahwa[1,2]
I.ARE.H. University of Saskatchewan, Saskatoon, Canada[1]
CH&EP, University of Saskatchewan, Saskatoon, Canada[2]

Abstract

Survey data analysis using complex sampling designs ought to account for clustering, stratification and unequal probability of selection. Design-based and model-based methods are two commonly used routes taken to account for such survey designs. Several studies of cross-sectional survey designs have shown that these two approaches provide similar results when the model fits the data well. The present paper aims at comparing these two approaches using the longitudinal National Population Health Survey (NPHS) dataset. The NPHS is an ongoing longitudinal study, for which data collection was based on a stratified multi-stage sampling design. A marginal modeling approach proposed by Rao (1998) was used by way of a design-based method. The Generalized Estimating Equation (GEE) method, proposed by Liang and Zeger (1986), was used as a typical model-based approach. Results obtained from these methods were compared to study the relationship of asthma and smoking.

Keywords: Multistage sampling, GEE, survey GEE, Taylor linearization, bootstrap, NPHS.

1. Introduction

Data collection for large surveys uses a complex sampling design. The growing interest in this field is mainly due to the fact that the results obtained from these surveys can be easily generalized to the target population under study, if analyzed correctly. The primary objective of analyzing survey data is to make inferences about the population of interest (LaVange, Koch et al. 2001). However, to obtain correct estimates and inferences the sampling design should be taken into account. In this paper, we primarily focus on comparing the GEE approach (model based method), proposed by Liang and Zeger (1986) and the design-based marginal modeling approach proposed by Rao (1998).

2 Models

2.1 Generalized Estimating Equation (GEE):

Consider Y_{it} a dichotomous outcome variable which assumes the logit model for the first order marginal probabilities: logit $[Pr(Y_{ij}=1)]$=logit (μ_{ij})=

$$\log\left(\frac{\mu_{it}}{1-\mu_{it}}\right) = \beta_s^T x_{is} + \beta_t^T x_{it}, \quad \text{eq (2.1)}$$; where t = *1,……..,T*(occasion) and *i= 1,……..,M* (subjects), β_s^T is a vector of stationary covariates, and β_t^T is a vector of time varying covariates

$\mu_{it} = \frac{\exp\{\beta_s' x_{is} + \beta_t' x_{it}\}}{1+\exp\{\beta_s' x_{is} + \beta_t' x_{it}\}}$, where x_{is} = design-matrix of time stationary covariates and x_{it} = design-matrix of time varying covariates

Liang and Zeger (1986) and Zeger and Liang (1986) proposed generalized estimating equations. A set of score equations for a marginal normal model is given by

$$U(\beta) = \sum_{i=1}^{N} D_i^T (\Delta_i^{1\backslash 2} \Re_i(\alpha) \Delta_i^{1\backslash 2})^{-1} (y_i - \mu_i) = 0, \quad \text{eq (2.2)}$$

where $D_i = \partial\mu_i / \partial\beta^T$ and μ_i is the mean function, and V_i is a working covariance matrix of outcome variable $Y_i = (Y_{i1,\ldots,} Y_{ij})^T$ a t x 1 vector of i=1,……,M individuals observed at t occasions, $X_i = (X_{i1},\ldots,X_{ij})^T$ is t X P matrix of covariates for individual i. The working covariance structure can be decomposed into $V_i = \Delta_i^{1\backslash 2} \Re_i(\alpha) \Delta_i^{1\backslash 2}$, where Δ_i =diag [var(Y_{i1}),……..,var(Y_{ij})], and $\Re_i(\alpha)$=corr (Y_i) is a $T X T$ "working" correlation matrix and α is a vector of parameters which are usually associated with a specified model for corr (Y_i) (Fitzmaurice and Laird 1993). The above equations reduce to independence equations if $\Re_i(\alpha)$ is the identity matrix.

The main inference is on the model-based coefficients, while the intra-cluster dependence is merely a nuisance characteristic, merely accounted for, but not subject to modeling in the classical sense. It can be used for Gaussian and non-Gaussian outcomes alike (Zeger and Liang 1986).

2.2 Marginal Modeling approach accounting for survey design [1] Consider a longitudinal study with T occasion of measurement and the finite longitudinal population of size *M* is clustered into *N* primary sampling units also known as primary sampling units (psu). The subscript i in equation 2.1 is changed to *hik* in survey data, where *h* represents strata, *i* represents clusters and *k* represents subject. For each stratum *h*, N_h and M_{hi} are respectively the number of clusters in

1 This section was modified from the notes of Suzanne Rubin-Bluer from an Internal report, January 2006, Statistics Canada

stratum h and the number of secondary units in the cluster *hi*, *i = 1,……., N_h* and *h = 1,……..,L*
Assume the same logit model for the first order marginal probabilities as in equation 2.1
The Survey independent estimating equation (IEE) estimator are given by

$$\hat{U}_{IEE}(\beta) = \sum_{hik \in S_l} \omega_{hik} D_{hik}^{'} A_{hik}^{-1} (y_{hik} - \mu_{hik}) = 0 \text{ eq (2.3)}$$

S_l is the longitudinal sample and ω_{hik} is the longitudinal weights.

To calculate the survey IEE estimator $\hat{\beta}_{IEE}$ we do the iteration

$$\hat{\beta}_{IEE}(K) = \hat{\beta}_{IEE}(K-1) - \left(\frac{\partial \hat{U}_{IEE}}{\partial \beta}\right)^{-1} \left(\beta_{IEE}(\hat{K-1})\right).\hat{U}_{IEE}\left(\hat{\beta}_{IEE}(K-1)\right) \quad \text{eq (2.4)}$$

Where $\hat{U}_{IEE}$ is the survey estimate defined above and $\left(\frac{\partial \hat{U}_{IEE}}{\partial \beta}\right)$ is replaced by its expectation: $\left(\frac{\partial \hat{U}_{IEE}}{\partial \beta}\right) \approx \sum_{hik \in S_l} \omega_{hik} D_{hik}^{'} A_{hik}^{-1} D_{hik}$

Where $A_{hik}^{-1} = Diag\left(\frac{1}{\mu_{hik1} - \mu_{hik1}^2}, ..., \frac{1}{\mu_{hik5} - \mu_{hik5}^2}\right)$

The Survey Generalized Estimating Equation (GEE) estimator proposed by Rao (1998) is of the form:

$$\hat{U}_{GEE}(\beta) = \sum_{hik \in S_l} \omega_{hik} D_{hik}^{`}(\beta) \Delta_{hik}^{-1/2}(\beta) \hat{R}^{-1} \Delta_{hik}^{-1/2}(\beta)(y_{hik} - \mu_{hik}(\beta)) = 0$$

eq (2.5)

Where the matrix of "correlation" $\hat{R}$ now has the form $\hat{R} = (r_{tu})_{tu}$

The estimator $\hat{\beta}_{GEE}$ is defined as the solution of the survey GEE.

$\hat{\beta}_{GEE}$ is calculated through iteration, where the $\hat{\beta}_{GEE}$ (K-1) change at each iterations, but $\hat{R} = (r_{tu})_{tu}$ is fixed throughout the iterations to calculate $\hat{\beta}_{GEE}$

The variance matrix of $\hat{\beta}_{GEE}$ can be consistently estimated by

$$v\left(\hat{\beta}_{GEE}\right) = \hat{J}_G^{-1}\left(\hat{\beta}_{GEE}\right).v\left(\hat{U}_{GEE}\right).\hat{J}_G^{-1}\left(\hat{\beta}_{GEE}\right) \text{ eq (2.6)}$$

evaluated at $\beta = \hat{\beta}_{GEE}$ with

$$\hat{J}_G(\beta) = -\sum_{hik \in S_l} \omega_{hik} D_{hik}^{'}(\beta) A_{hik}^{-1}(\beta) D_{hik}(\beta) \quad \text{and} \quad v\left(\hat{U}_{GEE}\right),$$

eq (2.7) evaluated at $\beta = \hat{\beta}_{GEE}$, is the survey design variances of a survey total and can be estimated by bootstrap, calculating for each one of the 500 sets of bootstrap weights estimated.

$$\hat{U}_{GEE}(b)\left(\hat{\beta}_{GEE}\right) = \sum_{hik \in S_l} \omega_{hik}(b) D_{hik}^{'} \Delta_{hik}^{-1/2} \hat{R}^{-1} \Delta_{hik}^{-1/2}\left(y_{hik} - \mu_{hik}\right)$$

eq (2.8) b= 1,.., 500
And then calculate:

$$v\left(\sqrt{n}\,\hat{\mathbf{U}}_{GEE}\right) = n\frac{1}{500}\sum_{b=1}^{500}\left(\hat{U}_{GEE}^{(b)} - \overline{\hat{U}}_{GEE}^{(b)}\right)\left(\hat{U}_{GEE}^{(b)} - \overline{\hat{U}}_{GEE}^{(b)}\right)'$$

3.1 Conclusion

In this paper we compared the model based and design based methods. The results shows that estimates obtained from both methods and using exchangeable correlation matrix are very close to each other. The weight variables used for purpose of analysis were specifically created for longitudinal survey design. In order to calculate correct standard error we used linearized estimating function bootstrap technique (Binder, Kovacevic et al. 2004). The method proposed by Rao (1998) does account for the complex sampling design as well as the longitudinal nature. When there is larger variation of weights, it can result in larger standard errors for the weighted estimators (Reiter, Zanutto et al. 2005). In our previous paper (Ghosh and Pahwa 2006) where we used a smaller subset of the NPHS data set as we were studying the asthma prevalence of female Canadian population. The results of that study indicated that there were differences in the standard error when using these two approaches. One of the conclusions made for the variability in the standard errors was attributed to the smaller sub-sample population used. The same analysis when repeated using a larger sub-population produced minimal differences in the standard errors.

However, based on previous literature when analyzing data set obtained from multi-stage sampling, the design based approach is the most appropriate. As suggested by Pfefferman (Pfeffermann 1996), the design based methods should be preferred as bias is the main issue in large sample surveys, and use of these methods removes bias so they should be preferred even if at the cost of large variances. The design-based estimators using the weighted estimates allows to obtain unbiased coefficient estimates of the independent variables in the regression model (Reiter, Zanutto et al. 2005). Binder (1983) and Kott (1991) suggest the use of design-based approaches as the weighted estimates provide unbiased estimates of the coefficients of the independent variables in the model, even when the other relevant independent variables are excluded from the model (Binder 1983; Kott 1991).

One should also keep in mind that the uses of design-based approach are limited. Large samples are required for hypothesis tests to be valid (Pfeffermann 1993). The generalization of the result obtained from

design-based approach are not readily made to different populations as the inferences are specific to a particular finite population (Kalton 1983). The model-based methods have gained popularity over the design-based methods as these methods can be readily implemented using standard commercial software. Model-based approaches are more valid and powerful than design-based approaches when the model assumption is reasonable. The standard errors obtained using model based methods are smaller when compared to design based methods. This is because unweighted estimates have smaller variances, as the variation in magnitude of weights are not included (Reiter, Zanutto, 2005). However, if the model does not fit the data well, biased estimation can result. Model-based standards errors have smaller standard errors than design-based estimators (Pfeffermann 1993).

However, caution should be used while applying any of these methods as both have their own advantages and disadvantages as discussed above. Hence, it is important to consider the relative advantages of these two approaches carefully.

Table 1: Estimates (Standard Errors) and Odds Ratio (95% confidence intervals) using Generalized Estimating equation proposed by Liang and Zeger (1986) for model based approach and Rao (1998) Survey GEE for design based approach

Covariates	Liang and Zeger GEE estimates and odds ratio		Survey GEE estimates and odds ratio	
	Estimate (S.E.)	Odds Ratio (95% CI)	Estimate (S.E.)	Odds Ratio (95% CI)
Immigration (Citizen)				
Immigrants	-0.65***	0.52	-0.62***	0.54
	(0.14)	(0.39-0.69)	(0.14)	(0.41-0.71)
Smoking Status (Non-Smokers)				
Current Smokers	0.03	1.03	0.03	1.03
	(0.09)	(0.87-1.21)	(0.09)	(0.87-1.22)
Ex-Smokers	0.03	1.03	0.03	1.04
	(0.08)	(0.89-1.20)	(0.08)	(0.89-1.20)
Location * Smoking (Urban Non-smoker)				
Rural Smokers	0.28*	1.33	0.29*	1.34
	(0.13)	(1.03-1.72)	(0.13)	(1.03-1.74)
Rural Ex-Smokers	0.25	1.28	0.26	1.30
	(0.14)	(0.97-1.68)	(0.14)	(0.99-1.71)

Reference category are specified in the parentheses
*** p<0.0001;** p<0.01; * p<0.05

References

Binder, D. (1983). "On the variances of Asymptotically Normal estimators from Complex Surveys." International Statistical Review **51**(2): 279-292.

Binder, D., M. Kovacevic, et al. (2004). Design Based Methods for Survey Data: Alternative uses of estimating Functions. Proceeding of the Section on Survey Resrearch Methods of the American Statistical Association.

Fitzmaurice, G. M. and N. M. Laird (1993). "A likelihood based method for analysing longitudinal binary data." Biometrika **80**: 141-151.

Ghosh, S. and P. Pahwa (2006). Design-based Versus Model-based Methods: A Comparative Study Using Longitudinal Survey Data. Proceedings of Statistical Society of Canada Survey Methods Section, London, Ontario.

Kalton, G. (1983). "Model in the Practice of Survey Sampling." International Statistical Review **51**(1): 175-188.

Kott, P. S. (1991). "A model based look at the Linear Regression with Survey Data." The American Statistician. **45**(2): 107-112.

LaVange, L. M., G. G. Koch, et al. (2001). "Applying sample survey methods to clinical trials data." Statistics in Medicine **20**: 2609-2623.

Pfeffermann, D. (1993). "The role of sampling weights when Modelling Survey Data." International Statistical Review **61**(2): 317-337.

Pfeffermann, D. (1996). "The use of sampling weights for Survey Data Analysis." Statistical Methods in Medical Research **5**(1): 239-261.

Reiter, J. P., E. L. Zanutto, et al. (2005). "Analytical Modeling in Complex Surveys of Work Practices." Industrial and Labor Relations Review **59**(1): 82-100.

Zeger, S. L. and K. Y. Liang (1986). "Longitudinal data analysis for discrete and continuous outcomes." Biometrics **42**(1): 121-30.