

ANALYSIS OF LONGITUDINAL SURVEY DATA WITH MISSING OBSERVATIONS: AN APPLICATION OF WEIGHTED GEE TO THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN AND YOUTH (NLSCY)

Ivan Carrillo, Milorad Kovacevic, Changbao Wu¹

ABSTRACT

All surveys, either cross-sectional or longitudinal, contain nonresponse. Under these circumstances one must make assumptions about the response mechanism. There is extensive literature about nonresponse for cross-sectional surveys. For longitudinal surveys, the missing data problem, the properties of different response mechanisms, and their impacts on variances are much less known. Assuming missing at random (MAR), we apply the weighted generalized estimating equations (WGEE) modeling, following the lines of earlier research for longitudinal studies under non-survey settings (Robins et al., 1995), to get estimators of regression coefficients and their joint randomization variance under the situation of either dropouts or intermittent nonresponse. We use the NLSCY for our analyses; our variable of interest is the physical aggression score for kids 2 to 11 years old.

KEY WORDS: Generalized Estimating Equations; longitudinal survey; non-response; regression coefficients.

RÉSUMÉ

Toutes les enquêtes, qu'elles soient transversales ou longitudinales, comportent de la non-réponse. En de telles circonstances, on doit faire des hypothèses à propos du mécanisme de réponse. Il existe une vaste littérature à propos de la non-réponse dans les enquêtes transversales. Pour les enquêtes longitudinales, en ce qui concerne le problème des données manquantes, les propriétés des différents mécanismes de réponse et leur impact sur la variance sont beaucoup moins connus. Si on suppose que le mécanisme de non-réponse est ignorable, on applique la modélisation par les équations d'estimation généralisées pondérées suivant ainsi la voie indiquée par les recherches préalables portant sur les études longitudinales dans un contexte autre que celui des enquêtes (Robins et al., 1995), afin d'obtenir des estimateurs des coefficients de régression et leur variance de randomisation dans la situation soit d'un abandon ou d'une non-réponse intermittente. Nous utilisons l'ELNEJ pour nos analyses. Notre variable d'intérêt est le score d'agression physique pour les enfants de 2 à 11 ans.

KEY WORDS: Coefficient de régression; enquête longitudinale; équations d'estimation généralisées; non-réponse.

1. INTRODUCTION

Statistics Canada conducts many big scale complex surveys. Most of these surveys are longitudinal because of the advantages of this kind of studies. All surveys, either cross-sectional or longitudinal, have some amount of nonresponse. To be able to draw inferences in the presence of nonresponse, the analyst makes assumptions about the response mechanism underlying them. These assumptions can be either explicit or implicit, but are always a must. The three commonly assumed response mechanisms are MCAR, MAR, and NMAR. The MCAR, or missing completely at random, mechanism assumes that the probability of a missing value is independent of the measurement process. MAR, or missing at random, means that the probability of a missing value is conditionally independent of the unobserved measurements given the values observed. And in the NMAR, or non missing at random, one assumes that the probability of a missing value depends on its actual value, even after conditioning on all the observed quantities.

¹ Ivan Carrillo and Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1. Email addresses: iacarril@uwaterloo.ca, cbwu@uwaterloo.ca. Milorad Kovacevic, Statistics Canada, 17 "J" R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Email: milorad.kovacevic@statcan.ca

The additional variability in the estimates, introduced by the response mechanism, should not be ignored. There is extensive literature about the nonresponse issues for cross sectional surveys and their properties. For longitudinal surveys, matched with the GEE methodology, the missing data problem, the properties of the different assumed response mechanisms, and their impacts on variances are much less known. Therefore, it is necessary, to have methods in place to deal with incomplete longitudinal survey data. There is a clear lack of methods for point and variance estimation under these characteristics.

In this paper we study the modeling of longitudinal survey data from a joint randomization perspective. In this case there are three randomization processes. Two are, as in usual joint randomization analyses, model and design randomizations; and also a third one coming from the assumed model of missingness (or response mechanism). The assumed mechanisms of missingness we use in this project are either MCAR or MAR.

We apply either GEE or weighted GEE estimators to the National Longitudinal Survey of Children and Youth (NLSCY), and find their joint randomization variance for different marginal longitudinal models. The variable of interest is the physical aggression score for kids 2 to 11 years old. Two different scenarios are considered: either for drop-outs or intermittent missing data. We follow the lines of earlier research done for longitudinal studies but under the non-survey setting, the work by Robins et al. (JASA, 1995).

2. GENERALIZED ESTIMATING EQUATIONS

The method of Generalized Estimating Equations (GEE) was proposed by Liang and Zeger (1986). This method is applicable to longitudinal studies; it permits estimation of regression coefficients in the presence of the within subject correlation arising in this kind of studies. The GEE method is an attempt to get estimators without the requirement of assuming a distribution for the response, but only a regression model for its mean. Thus this method does not produce MLEs.

We take an independent random sample of n subjects. For each subject i , we take a set of T_i measurements (over time) of a random variable Y . We denote these T_i measurements i by Y_{ij} , $j=1,2,\dots,T_i$, and their expected value $E[Y_{ij}] = \mu_{ij}$. Additionally, associated with each observation Y_{ij} we observe a set of p explanatory variables X_{ijk} , $k=1,2,\dots,p$. We organize the observations for each subject in the following way, $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iT_i})'$, $X_{ij} = (X_{ij1}, X_{ij2}, \dots, X_{ijp})'$, and $X_i = (X_{i1}, X_{i2}, \dots, X_{iT_i})$. Then $E[Y_i] = \mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{iT_i})'$, and the n Y_i vectors are independent. We assume that there exists some variance-covariance matrix of Y_i denoted by Σ_i . We are interested in modeling μ_i but not Σ_i ; we regard this matrix as a nuisance parameter.

The GEE method can be characterized as being composed of the following four items. 1. A “linear predictor” $\eta_{ij} = X'_{ij}\beta$, where $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ is a vector of unknown regression coefficients. 2. A “link function” $g(\cdot)$ relating the mean of the response to the linear predictor, as $g(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta$. 3. The variance of the response Y_{ij} , given X_{ij} , may change with X_{ij} only as a function of the mean, μ_{ij} , as $Var[Y_{ij}] = (\phi/w_{ij})\nu(\mu_{ij})$; where $\phi > 0$ is the “dispersion parameter,” w_i is a known weight for observation subject i at time j , and $\nu(\mu_{ij})$ is a known “variance function” of the mean. 4. We choose a “working” correlation matrix, $\mathbf{R}_i(\alpha)$, for the elements in Y_i , depending on some parameters α estimated from the data.

Therefore, the “working” variance-covariance matrix of Y_i is composed as $V_i = A_i^{\frac{1}{2}}\mathbf{R}_i(\alpha)A_i^{\frac{1}{2}}$; where A_i is a $T_i \times T_i$ diagonal matrix with $Var[Y_{ij}] = (\phi/w_{ij})\nu(\mu_{ij})$ as j th diagonal element.

The GEE estimator $\hat{\beta}$ is obtained as a solution to the following set of equations:

$$\sum_{i=1}^n \frac{\partial \mu'_i}{\partial \beta} V_i^{-1} (y_i - \mu_i) = 0; \quad (2.1)$$

where $\partial \mu'_i / \partial \beta$ is the $p \times T_i$ matrix of partial derivatives $\partial \mu'_i / \partial \beta = (\partial \mu_{i1} / \partial \beta, \partial \mu_{i2} / \partial \beta, \dots, \partial \mu_{iT_i} / \partial \beta) = [\partial \mu_{ij} / \partial \beta]_{j1}$, and the right-hand side is a $p \times 1$ vector of zeroes. The left-hand side of equation (2.1) is a function of β and α , and in general the equation does not have a closed form solution, but instead has to be solved iteratively.

In this paper we assume that all the subjects share a common unspecified within-subject association. In other words, we assume that the working within-subject correlation matrix is the same for all individuals and we do not constraint this single matrix in any way. We can write these assumptions, mathematically, as: $\mathbf{R}(Y_{ij}, Y_{ik}) = 1$ if $j = k$, and $\mathbf{R}(Y_{ij}, Y_{ik}) = \alpha_{jk}$ if $j \neq k$; and we estimate it with $\hat{\alpha}_{jk} = \sum_{i=1}^n e_{ij} e_{ik} / n - p$, where $e_{ij} = (y_{ij} - \hat{\mu}_{ij}) / \sqrt{\nu(\hat{\mu}_{ij}) / w_{ij}}$ is the residual for subject i at time j . The dispersion parameter ϕ is estimated by

$$\hat{\phi} = \frac{\sum_{i=1}^n \sum_{j=1}^{T_i} e_{ij}^2}{(\sum_{i=1}^n T_i) - p}. \quad (2.2)$$

The procedure is implemented in several statistical packages for different outcome variables and within-subject correlation structures like R (function *gee*) and SAS (procedure *genmod*).

The solution $\hat{\beta}$ of equation (2.1) is consistent for β , as long as the linear predictor and link function are correctly specified. The consistency of $\hat{\beta}$ does not depend on the validity of the assumed correlation matrix \mathbf{R}_i ; however, $\hat{\beta}$ will be more efficient if \mathbf{R}_i resembles Σ_i more closely. We have, asymptotically, $\hat{\beta} \sim \text{MVN}(\beta, \text{Cov}(\hat{\beta}))$, where $\text{Cov}(\hat{\beta}) = B^{-1}$, and $B = \sum_{i=1}^n (\partial \mu_i' / \partial \beta) V_i^{-1} (\partial \mu_i / \partial \beta)$; although B^{-1} is not a robust estimator of the variance of β if \mathbf{R}_i is misspecified. For a more detailed discussion of the GEE methodology see, for example, Fitzmaurice et. al. (2004).

2.1 GEE for Longitudinal Survey Data

In this paper we follow a ‘‘joint randomization’’ approach. We are interested in model parameters; i.e. we are interested in the effect of some covariates on an outcome variable. Beside, the NLSCY is a complex survey, in the sense that the observations are sampled in clusters (more specifically in multiple stages), and are selected with differential probabilities and thus have different weights. A joint randomization approach takes into account both parts of the process, the randomization imposed by the model generating the finite population values and the randomization introduced by the sampling selection. Even under a design-based approach, certain optimality criteria involve relying on a model (as in Wu, 2003). And finally, sometimes it is the only appropriate method of inference because of the way in which the data are collected (as in Chen et. al., 2004).

We assume that we have a GEE model denoted by ξ . We think of an infinite superpopulation satisfying the model

$$\xi : \begin{cases} E[Y_{ij}] = \mu_{ij} = \mathbf{g}^{-1}(\eta_{ij}) = \mathbf{g}^{-1}(X'_{ij}\beta); & i = 1, 2, \dots; j = 1, 2, \dots \\ \text{Var}[Y_{ij}] = (\phi / w_{ij}) \nu(\mu_{ij}); & i = 1, 2, \dots; j = 1, 2, \dots \\ Y_k \text{ and } Y_l \text{ are independent vectors } \forall k \neq l \end{cases}; \quad (2.3)$$

with the requirements specified in the previous section. We assume that the finite population conforms an independent and identically distributed random sample of N elements drawn from model ξ .

We draw a sample \mathbf{p} from the finite population, and it will usually be complex. We assume that each subject i has associated with it a ‘‘survey weight’’ w_i . This weight is, basically, be the inverse of the probability of selection of unit i , but also adjusted to account for unit nonresponse and poststratification. The sum of the weights for the elements in the sample will be an estimate of the finite population size, N . In other words, $\hat{N} = \sum_{i=1}^n w_i$.

We solve the following set of equations to get our estimate, $\hat{\beta}$, of β :

$$\sum_{i=1}^n w_i \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} (y_i - \mu_i) = 0. \quad (2.4)$$

In this case we regard w_i as the w_{ij} of equation (2.1) (i.e. $w_{ij} = w_i, j = 1, 2, \dots, T_i$) and A_i is a $T_i \times T_i$ diagonal matrix with $\phi \nu(\mu_{ij})$ as the j th diagonal element; in other words, the weight for a sampled individual i does not vary from measurement to measurement and is equal to the sampling weight.

To get *the point estimate*, $\hat{\beta}$, using survey data, we can use the same procedure as that in the previous section, using equation (2.1), just replacing the w_{ij} , $j = 1, 2, \dots, T_i$ there by w_i . However, some modifications *are* necessary in some of the calculations because the model weights w_{ij} and design weights w_i represent different things. The weights w_{ij} are a measure of model variance. The survey weight w_i , on the other hand, is a measure of the number of subjects that a given individual in the sample represents in the population.

When we replace w_{ij} by the survey weight w_i , we should also modify the quantities n and $\sum_{i=1}^n T_i$ to appropriate values, so that the scale is the same in all the pieces. We earlier used $n = \sum_{i=1}^n 1$ to connote the total number of subjects represented by our sample (which was equal to the number of subjects in the sample). But now the number of subjects represented by our sample is N ; so we should replace n by our best estimate of N , which is $\hat{N} = \sum_{i=1}^n w_i$. In other words, in a place where individual i represented only himself, it now represents w_i subjects. Similarly, we used earlier the term $\sum_{i=1}^n T_i$ to express the total number of measurements represented by our sample (which was equal to the number of measurements in the sample). But now the number of measurements represented by our sample is $\sum_{i=1}^n w_i T_i$; so we replace $\sum_{i=1}^n T_i$ by our best estimate of $\sum_{i=1}^n w_i T_i$, which is $\sum_{i=1}^n w_i T_i$.

With these modifications, the dispersion parameter ϕ is estimated by

$$\hat{\phi} = \frac{\sum_{i=1}^n \sum_{j=1}^{T_i} e_{ij}^2}{(\sum_{i=1}^n w_i T_i) - p} = \frac{\sum_{i=1}^n w_i \sum_{j=1}^{T_i} (y_{ij} - \hat{\mu}_{ij})^2 / \nu(\hat{\mu}_{ij})}{(\sum_{i=1}^n w_i T_i) - p}; \quad (2.5)$$

and if the within-subject association is unspecified (and the same for all subjects), we estimate α by

$$\hat{\alpha}_{jk} = \frac{\sum_{i=1}^n e_{ij} e_{ik}}{(\sum_{i=1}^n w_i) - p} = \frac{\sum_{i=1}^n w_i (y_{ij} - \hat{\mu}_{ij})(y_{ik} - \hat{\mu}_{ik}) / \sqrt{\nu(\hat{\mu}_{ij})\nu(\hat{\mu}_{ik})}}{(\sum_{i=1}^n w_i) - p}. \quad (2.6)$$

Because of these changes, from n to \hat{N} and from $\sum_{i=1}^n T_i$ to $\sum_{i=1}^n w_i T_i$, we do *not* recommend usual GEE software procedures like *gee* in R or *genmod* in SAS for survey data. Even if one inputs the survey weights w_i as the weights, these procedures do not carry out the appropriate modification of $\hat{\phi}$ and $\hat{\alpha}$.

The estimator $\hat{\beta}$ is consistent for β jointly under model and design. We calculate the variance of $\hat{\beta}$ with respect to both randomizations: the one induced by the model *and* the one induced by the sampling scheme. As Kovacevic et. al. (2002) point out, “in general the total variance should be used for inference about the superpopulation parameters because it accounts for both variabilities.” We use subscripts ξ or \mathbf{p} to indicate under the model or under the design respectively.

The joint variance of $\hat{\beta}$ is given by

$$Var_{\xi\mathbf{p}}[\hat{\beta}] = Var_{\xi}[E_{\mathbf{p}}(\hat{\beta})] + E_{\xi}[Var_{\mathbf{p}}(\hat{\beta})] \approx B_N^{-1} + E_{\xi}[Var_{\mathbf{p}}(\hat{\beta})]; \quad (2.7)$$

where $B_N = \sum_{i=1}^N (\partial\mu'_i / \partial\beta) V_i^{-1} (\partial\mu_i / \partial\beta)$. The first component in (2.7), $Var_{\xi}[\hat{\beta}_N] = B_N^{-1}$, is the “model variance component.” It is due to the fact that the N finite population data points scatter according to model ξ . The second component, $E_{\xi}[Var_{\mathbf{p}}(\hat{\beta})]$, is the “design variance component” or “sampling variance component.” It comes from the fact that a sample of n elements is observed rather than the entire finite population of N elements (Särndal et. al., 1992).

We estimate the joint variance of $\hat{\beta}$ in (2.7) by estimating the two components separately, as $\hat{V}ar_{\xi\mathbf{p}}[\hat{\beta}] = \hat{B}_N^{-1} + \hat{V}ar_{\mathbf{p}}(\hat{\beta})$; where $\hat{V}ar_{\mathbf{p}}(\hat{\beta})$ is an estimator of the design variance of $\hat{\beta}$, which in our case, we estimate using the bootstrap technique; and $\hat{B}_N = \sum_{i=1}^n w_i (\partial\mu'_i / \partial\beta) V_i^{-1} (\partial\mu_i / \partial\beta)$, in which we replace the values of α , ϕ , and β by their estimated values $\hat{\alpha}$, $\hat{\phi}$, and $\hat{\beta}$.

3. MISSING VALUES IN LONGITUDINAL STUDIES

In longitudinal studies the nonresponse patterns and mechanisms are more complicated than in cross-sectional ones. We intend to measure a response variable and a set of covariates on each of n subjects at each of say T times (cycles). However, generally a variety of different missing patterns occur and we cannot observe all the intended measures.

Some subjects respond at all cycles and to all variables of interest. These units are called complete cases. Some units fail to respond altogether, at all cycles and all variables of interest. This situation is referred to as unit nonresponse. It is usually dealt with by reweighting the respondent units to account for these unit-nonrespondents. In our analysis of the NLSCY we will not deal with this kind of nonresponse. The longitudinal weight for cycle 1 (and the funnel weight) is already adjusted for these units.

Another situation arises when a subject is observed for some cycles but not for others. Those cycles in which a subject is not observed at all are sometimes referred to as wave nonresponse. In the NLSCY this kind of nonresponse is handled cycle by cycle. Each cycle's longitudinal weight is adjusted so that the wave respondents for that cycle account for the wave nonrespondents and the weight for the latter are set to zero. Additionally, for the funnel weight, all individuals with at least one wave nonresponse get a zero weight and they are accounted for by those subjects who were respondents at all cycles. For some units, it may happen that, at some cycles, some of the covariates are not observed, whereas the response variable is observed. This situation is sometimes referred to as missing covariates. At some cycles, some units may have the response variable Y not observed, whereas all the covariates are observed. This situation is sometimes referred to as missing outcome or missing response. There are also some units which, at some cycles, have missing outcomes and missing covariates at the same time.

Additionally, in longitudinal studies it often occurs that when a subject misses one wave, that subject never returns to the study. In other words, once there is a wave nonresponse for some unit, that unit is likely to have wave nonresponse from then on. These units are usually called dropouts. On the other hand, in some cases, some units who miss a wave do come back to the study at a later wave. These units can be called intermittent observations or units.

We restrict our attention to complete-wave observations. For each subject, we will ignore any wave in which either the response and/or any covariate is missing. So, our analyses concentrate on datasets with wave-nonresponse; which can be complete, monotone, or intermittent. We consider only those kids who have at least one completely observed wave. There are 5,570 kids like that in the NLSCY. We call this dataset, the "available case" dataset. The GEE methodology has been developed also for the case of missing covariates, but we do not elaborate on it here. Readers interested in this subject can consult for example Robins et. al. (1994).

A dropout mechanism is missing completely at random (MCAR) if it is independent of the measurement process. In other words, dropout is MCAR if the probability of dropout at any given wave is independent of all observed (past) and unobserved (present and future) outcomes. The dropout mechanism is missing at random (MAR) if it depends on the past (observed) outcomes but is independent of the current (missing) and future (missing) outcomes.

Intermittent patterns are a bit more complicated. We say the intermittent missing mechanism is missing completely at random (MCAR) if it is independent of the measurement process. This is, if the probability of missingness at a given wave is independent of all observed and unobserved outcomes either in the past, present, or future waves. The intermittent missing mechanism is missing at random (MAR) if it is independent of the unobserved portion of the measurement process. This is, if at a given wave, the probability of nonresponse depends on the observed outcomes (either past or future) but is independent of the unobserved ones.

The GEE analysis using only the complete cases, i.e. those subjects with all waves completely observed, is valid (consistent estimators) only under the strongest assumption of MCAR. However, even if the nonresponse mechanism is MCAR, complete case analyses are usually inefficient (high variances) because of the reduction in number of observations compared to an analysis which uses all the observed waves for each subject. We call this approach "available case" (AC) analysis. This method is more efficient than a complete case analysis because it uses more data; but nonetheless, it yields consistent estimators of the regression coefficients only under a MCAR mechanism (Albert, 1999), just like complete case analysis. For an illustration of the sort of biases in point estimates using GEE when the missing data are not MCAR see for example Rotnitzky and Wypij (1994).

4. WEIGHTED GEE UNDER MAR FOR MONOTONE LONGITUDINAL SURVEY DATA

Complete case or available case GEE analysis with either monotone or intermittent missing data produces consistent estimates of the regression coefficients only in the case of MCAR data. Robins et. al. (1995) propose an extension of the GEE method, applicable to longitudinal studies with missing observations, when the missing mechanism is MAR.

Diggle et. al. (2002) summarize the idea of this method, for monotone datasets, along the following lines. If p_{ij} is the probability that subject i has not dropped out by time j , given his/her observed history, then (under MAR) the observation y_{ij} is representative of all subjects who do drop out and have the same history; therefore, in an available case GEE analysis, the contribution of y_{ij} needs to be weighted by the inverse of p_{ij} , to account for those who dropped out and have the same history. This methodology is sometimes called weighted generalized estimating equations (WGEE).

The WGEE is well suited for our NLSCY dataset because it “requires, inevitably, that we can consistently estimate the dropout probabilities for each subject given their observed measurement history and any relevant covariates. This makes the method best suited to large-scale studies” (Diggle et. al., 2002). We concentrate on the how the method extends to survey data; readers interested in non-survey situations can consult the original paper by Robins et. al. (1995).

Each subject i has associated with it the (cycle 1’s longitudinal) survey weight w_i . We intend to measure each subject i T times (4 in our NLSCY), but some subjects fail to respond after a given cycle and so, drop out of the study. Due to dropouts the full vector Y_i and matrix X_i' are not always observed. We assume that in addition to Y_{it} and X_{it} , at time t we also intend to measure a vector of covariates V_{it} , $t = 1, 2, \dots, T$. These covariates V_{it} help us in the estimation of the nonresponse probabilities. We define the response indicator variable $R_{it} = 1$ if subject i is observed at time t , and $R_{it} = 0$ otherwise. We assume that at any give time t Y_{it} , X_{it} , and V_{it} are either all observed or all missing. In this section we only deal with dropouts; i.e. if $R_{it} = 0$ then $R_{i(t+1)} = 0$. We also assume $R_{i1} = 1$ for all subjects.

The MAR mechanism implies the following, weaker, assumption; which is sufficient for the WGEE method to be valid;

$$P(R_{it} = 1 | R_{i(t-1)} = 1, X_{i1}, \dots, X_{i(t-1)}, V_{i1}, \dots, V_{i(t-1)}, Y_{i1}, \dots, Y_{iT}) = P(R_{it} = 1 | R_{i(t-1)} = 1, X_{i1}, \dots, X_{i(t-1)}, V_{i1}, \dots, V_{i(t-1)}, Y_{i1}, \dots, Y_{i(t-1)}) = p_{it}. \quad (4.1)$$

In words of Robins et. a. (1995), this assumption means that, among subjects observed at time $t-1$, nonresponse at time t is unrelated to the current and future outcomes Y_{it}, \dots, Y_{iT} , conditional on the observed past $X_{i1}, \dots, X_{i(t-1)}$, $V_{i1}, \dots, V_{i(t-1)}$, $Y_{i1}, \dots, Y_{i(t-1)}$. We assume that the probability p_{it} of being observed at time t , having been observed at time $t-1$ (conditional on $X_{i1}, \dots, X_{i(t-1)}$, $V_{i1}, \dots, V_{i(t-1)}$, $Y_{i1}, \dots, Y_{i(t-1)}$), is bigger than zero for all times $t = 2, \dots, T$.

We assume that the response probabilities p_{it} are a known function (taking values on $[0, 1]$) of an unknown parameter λ , and the observed past $X_{i1}, \dots, X_{i(t-1)}$, $V_{i1}, \dots, V_{i(t-1)}$, $Y_{i1}, \dots, Y_{i(t-1)}$; i.e. $p_{it} = p_{it}(X_{i1}, \dots, X_{i(t-1)}, V_{i1}, \dots, V_{i(t-1)}, Y_{i1}, \dots, Y_{i(t-1)}; \lambda)$. Robins et. al. (1995) also argue that “standard procedures can be used to investigate the functional form of $p_{it}(\cdot)$, [and...] augmenting the model for p_{it} will usually lead to an improvement of the efficiency with which we estimate β .”

We let $\hat{\lambda}$ be the partial pseudo maximum likelihood estimator of λ . And we define $\pi_{it}(\lambda) = p_{i1}(\lambda) \times \dots \times p_{it}(\lambda)$; which, under MAR, is the probability that subject i is observed at time t given X_{i1}, \dots, X_{iT} , V_{i1}, \dots, V_{iT} , Y_{i1}, \dots, Y_{iT} ; and $\pi_{it}(\hat{\lambda}) = p_{i1}(\hat{\lambda}) \times \dots \times p_{it}(\hat{\lambda})$. We also define the $T \times T$ diagonal matrix $\Delta_i(\lambda) = \text{diag}[R_{it}/\pi_{it}(\lambda)]_t$; and similarly the matrix $\Delta_i(\hat{\lambda})$.

Instead of equations (2.4), we solve iteratively the following set of equations to get our estimate, $\hat{\beta}$, of β :

$$\sum_{i=1}^n w_i \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \Delta_i(\hat{\lambda}) (y_i - \mu_i) = 0. \quad (4.2)$$

Note that equations (4.2) differ from equations (2.4) only in the inclusion of the “weighting” matrix $\Delta_i(\hat{\lambda})$. This matrix has the effect of setting to zero any unobserved residual in the vector (of residuals) $(y_i - \mu_i)$, and weighting by the inverse of $p_{it}(\hat{\lambda})$ the corresponding observed residual in this vector.

Under (4.1), and provided the model for p_{it} is correctly specified, equation (4.2) has a root $\hat{\beta}$ that is consistent for β jointly under model and design. Additionally, $\hat{\beta}$ is unique with probability approaching to one and asymptotically normal, under mild regularity conditions. The joint variance of $\hat{\beta}$ is given by

$$Var_{\mathfrak{p}}[\hat{\beta}] = Var_{\xi}[E_{\mathfrak{p}}(\hat{\beta})] + E_{\xi}[Var_{\mathfrak{p}}(\hat{\beta})]; \quad (4.3)$$

Where the first component, $Var_{\xi}[E_{\mathfrak{p}}(\hat{\beta})]$, is the model variance component and the second one, $E_{\xi}[Var_{\mathfrak{p}}(\hat{\beta})]$, is the design component. We estimate the two components separately, using bootstrap to estimate the design component. For the specific form of the components of the variance and the corresponding estimators, the reader is referred to Carrillo (2006).

We end up this section with an important note. Equation (4.2) can also be written as

$$\sum_{i=1}^n \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} diag[w_i R_{it} / \pi_{it}(\hat{\lambda})] (y_i - \mu_i) = 0. \quad (4.4)$$

This form is “appealing” because it indicates that, for monotone longitudinal survey data, WGEE adjusts the original survey weight (for example cycle 1’s longitudinal weight) for the inverse of the estimated probability that subject i responds at the given wave; and then weights the corresponding residual in $(y_i - \mu_i)$ by this wave-specific “adjustment factor.” Since Statistics Canada provides longitudinal weights for each cycle, which are the cycle 1’s longitudinal weights adjusted for nonresponse at the given cycle, it is likely that using Statistics Canada’s cycle-specific weights in equation (4.4) produces similar results to the ones obtained by WGEE along with the estimation of the π_{it} ’s. Nonetheless, this approach is not directly applicable to the NLSCY because this survey has nonmonotone (i.e. intermittent) missing patterns, and equation (4.4) is only applicable to monotone patterns.

5. WGEE UNDER MAR FOR INTERMITTENT LONGITUDINAL SURVEY DATA

The method in the previous chapter only deals with (artificially) monotone datasets. When the dataset contains intermittent patterns, that method is obviously inefficient, and maybe even inconsistent if the reasons for dropping out of the study differ from the reasons for missing a wave (and coming back to the study). Robins et. al. (1995) also propose a WGEE method applicable to longitudinal studies with intermittent missing observations, when this missing mechanism can be considered MAR. Here we show how the method extends to survey data; again, for non-survey situations, we refer the reader to the original paper.

We make the same assumptions about the model and the measurements as in the previous section, but here we additionally permit that some subjects who fail to respond at a given cycle may come back to the study at a later cycle. We define the response indicator variable as $R_{it}^* = 1$ if subject i is observed at time t , and $R_{it}^* = 0$ otherwise. We assume that at any give time t Y_{it} , X_{it} , and V_{it} are either all observed or all missing. We allow dropouts and missing waves coming back; i.e. we allow the vector $R_i^* = (R_{i1}^*, R_{i2}^*, \dots, R_{iT}^*)'$ to take on any of 2^{T-1} possible realizations (i.e. any vector $r = (r_1, r_2, \dots, r_T)'$ of zeros and ones of length T with first component equal to one). This method assumes that all subjects are observed at wave 1; i.e. $R_{i1}^* = 1$ for all subjects. And we redefine R_{it} in the following way: $R_{it} = 1$ if $R_{i1}^* = R_{i2}^* = \dots = R_{it}^* = 1$, and $R_{it} = 0$ otherwise. So, R_{it} is zero once a subject misses one wave.

This method assumes the following, which is stronger than equation (4.1) (and stronger than MAR):

$$\begin{aligned} P(R_{it}^* = 1 \mid R_{i1}^*, \dots, R_{i(t-1)}^*, R_{i1}^* X_{i1}, \dots, R_{i(t-1)}^* X_{i(t-1)}, R_{i1}^* V_{i1}, \dots, R_{i(t-1)}^* V_{i(t-1)}, R_{i1}^* Y_{i1}, \dots, R_{i(t-1)}^* Y_{i(t-1)}, X_{i1}, \dots, X_{iT}, V_{i1}, \dots, V_{iT}, Y_{i1}, \dots, Y_{iT}) \\ = P(R_{it}^* = 1 \mid R_{i1}^*, \dots, R_{i(t-1)}^*, R_{i1}^* X_{i1}, \dots, R_{i(t-1)}^* X_{i(t-1)}, R_{i1}^* V_{i1}, \dots, R_{i(t-1)}^* V_{i(t-1)}, R_{i1}^* Y_{i1}, \dots, R_{i(t-1)}^* Y_{i(t-1)}) = p_{it}. \end{aligned}$$

In words of Robins et. a. (1995), this equation means that, the probability of being observed at time t , given the observed past $R_{i1}^*, \dots, R_{i(t-1)}^*$, $R_{i1}^* X_{i1}, \dots, R_{i(t-1)}^* X_{i(t-1)}$, $R_{i1}^* V_{i1}, \dots, R_{i(t-1)}^* V_{i(t-1)}$, $R_{i1}^* Y_{i1}, \dots, R_{i(t-1)}^* Y_{i(t-1)}$ through time $t-1$, does not depend on the unobserved past or present or on the future. We still assume that the probability p_{it} of being observed at time t ,

conditional on the observed past $R_{i1}^*, \dots, R_{i(t-1)}^*, R_{i1}^* X_{i1}, \dots, R_{i(t-1)}^* X_{i(t-1)}, R_{i1}^* V_{i1}, \dots, R_{i(t-1)}^* V_{i(t-1)}, R_{i1}^* Y_{i1}, \dots, R_{i(t-1)}^* Y_{i(t-1)}$, is bigger than zero for all times $t = 2, \dots, T$.

We assume that the response probabilities p_{it} are a known function (taking values on $[0, 1]$) of an unknown parameter λ , and the observed past $R_{i1}^*, \dots, R_{i(t-1)}^*, R_{i1}^* X_{i1}, \dots, R_{i(t-1)}^* X_{i(t-1)}, R_{i1}^* V_{i1}, \dots, R_{i(t-1)}^* V_{i(t-1)}, R_{i1}^* Y_{i1}, \dots, R_{i(t-1)}^* Y_{i(t-1)}$; i.e.

$$p_{it} = p_t(R_{i1}^*, \dots, R_{i(t-1)}^*, R_{i1}^* X_{i1}, \dots, R_{i(t-1)}^* X_{i(t-1)}, R_{i1}^* V_{i1}, \dots, R_{i(t-1)}^* V_{i(t-1)}, R_{i1}^* Y_{i1}, \dots, R_{i(t-1)}^* Y_{i(t-1)}; \lambda). \quad (5.2)$$

We let $\hat{\lambda}$ solve the estimating equation $\sum_{i=1}^n w_i S_{\lambda,i}(\lambda) = 0$, where

$$S_{\lambda,i}(\lambda) = \frac{\partial}{\partial \lambda} \log \left\{ \prod_{t=2}^T [p_{it}(\lambda)]^{R_{it}^*} [1 - p_{it}(\lambda)]^{1 - R_{it}^*} \right\} = \sum_{t=2}^T [R_{it}^* - p_{it}(\lambda)] \frac{\partial \log p_{it}(\lambda)}{\partial \lambda}. \quad (5.3)$$

In this case, instead of equations (4.2), we solve iteratively the following set of equations to get our estimate, $\tilde{\beta}$, of β :

$$\sum_{i=1}^n w_i \left(\frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \Delta_i(\hat{\lambda})(y_i - \mu_i) - \hat{\theta} A_i \right) = 0; \quad (5.4)$$

where $\Delta_i(\hat{\lambda})$ is defined as in the previous chapter but replacing R_{it} by the one defined earlier in this chapter. A_i is a function of the observed data and the estimated probabilities of nonresponse for the different cycles and all the different patterns; and has expected value equal to zero for all subjects. And θ is chosen so that the solution to (5.4) is at least as efficient as the solution to equations (4.2) applied to the artificial monotone dataset. For the exact form of θ , A_i , and $\hat{\theta}$ see Carrillo (2006).

Under (5.1), and provided the model for p_{it} is correctly specified, equation (5.4) has a root $\tilde{\beta}$ that is consistent for β jointly under model and design. Additionally, $\tilde{\beta}$ is unique with probability approaching to one and asymptotically normal, under mild regularity conditions. The joint variance of $\tilde{\beta}$ is given by $Var_{\mathbb{P}}[\tilde{\beta}] = Var_{\varepsilon}[E_{\mathbb{P}}(\tilde{\beta})] + E_{\varepsilon}[Var_{\mathbb{P}}(\tilde{\beta})]$. Model variance component plus design variance component. Which we estimate separately, using bootstrap for the design component. Again, we refer the reader to Carrillo (2006) for the detailed form of the components of the variance and the estimators. Robins et. al. (1995) claim that the estimator obtained by applying the method in the previous section to the artificially monotone dataset is never more efficient than $\tilde{\beta}$.

6. APPLICATION TO THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN AND YOUTH (NLSCY)

6.1 Brief Description of the NLSCY and Variables of Interest

The NLSCY is a longitudinal survey by Human Resources Development Canada designed to measure child development and well-being. The main objective of the survey is to study the development of children's behaviour problems as they grow as well as examining the factors that contribute to change. It consists of (so far) five biennial cycles conducted from 1994 to 2003, and looked at households with children from 0 to 11 years old at the first cycle.

One very important measurement of the NLSCY is the aggressive behaviour of young children. "Aggression in childhood has been linked with later aggression, delinquency, and crime in adolescence and adulthood; with poor school outcomes; with unemployment in adulthood; and with other negative circumstances" (Thomas, 2004). This is our outcome of interest. The response variable is the "Physical Aggression Score," (PAS). This variable is a scale from 0-12 based on eight or six questions (depending on the age); a high score indicates behaviours associated with conduct disorders, physical aggression, and opposition. PAS is scaled from 0 to 16 based on eight questions for children who are 2 to 3 years old, and is scaled from 0 to 12 based on six questions for children who are 4 to 11 years old. For the results to be comparable across different age groups, PAS's are unified to a scale of 0 to 12. To do this we simply multiply the score for 2-3 year-old children by $12/16=0.75$ and leave the score for 4-11 year-old children unchanged. Although PASs are ordinal data, we treated them as continuous variables in our analyses since it has more than seven categories (Carrillo et. al., 2005). In the dataset there are 5,570 kids who were 2-5 years old at cycle 1 and 9-11 years old at cycle 4. Our population of interest are those children who were 2-5 years old from 1994 to 1995 and 9-11 years old from 2000 to 2001.

Earlier studies (Thomas 2004, Carrillo et. al. 2005) found some significant factors contributing to change in aggressive behaviours as children grow. On these grounds we included the following 12 covariates in our models, as potentially being significant in explaining the PAS. For a detailed description of these variables see Carrillo (2006). Age, Age² (the square of Age), Depression of the PMK (person most knowledgeable about the kid), Punitive Parenting Status, Region, Gender, Family Status, Household Income Status, Hours in Daycare, the Age by Punitive Parenting Interaction, the Age by Household Income Status interaction, and the Age by Region interaction. We abbreviate these variables as Age, Age², DeprePMK, Punitive, Region, Gender, FamStat, Income, Hours, Age*Puni, Age*Inco, and Age*Regi, respectively.

6.2 Results of the WGEE Analysis for the Intermittent Dataset

Table 1 summarizes the results we obtained when we applied the WGEE method to the largest dataset, the one that included all the completed waves for all the kids with at least one completed wave. We only include the variables that turned out to be significant.

Table 1 - Estimates for intermittent dataset with cycle 1's longitudinal weights.

Parameter	(Level)	Estimate	Variance Model Compo.	Variance Design Compo.	Joint Variance	Joint Standard Error	Joint p- value
Intercept		3.0782	0.00017	0.10985	0.110	0.332	<.0001
Age		-0.9279	0.00001	0.00422	0.004	0.065	<.0001
Age ²		0.0577	0.00000	0.00001	0.000	0.003	<.0001
DeprePMK		0.0390	0.00000	0.00003	0.000	0.005	<.0001
Punitive		0.2333	0.00000	0.00082	0.001	0.029	<.0001
Gender	Female	-0.3473	0.00001	0.00449	0.004	0.067	<.0001

Among all our covariates of interest we found that Age, Age², DeprePMK, Punitive, and Gender influence the PAS. This means that, as age increases from 2 to about 6.5 years old, kids become less and less aggressive, then aggression stabilizes until around 9 years old, when aggression tends to begin increasing somewhat again, until 11 years old. All other things being equal, if a PMK is one point more depressive than other, then the kid will be about 0.04 points more aggressive than the other PMK's. All other things being equal, if a PMK scores one point higher on the punitive scale, then the kid will be about 0.23 points more aggressive. And finally, a girl will have on average a PA score about 0.35 lower than a boy.

7. CONCLUSIONS AND RECOMMENDATIONS

We showed how the WGEE method, originally proposed by Robins et. al. (1995), can be extended to accommodate survey data, under the joint randomization perspective.

We applied a variety of methods to either the complete cases, the available cases, the monotone cases, and the intermittent cases (all the results can be consulted in Carrillo, 2006), we found that the missing mechanism in the NLSCY dataset that we analyzed is not MCAR. Neither the dropouts nor the intermittent missingness behave as having been generated by a MCAR mechanism. On the contrary, the probability of dropping out and the probability of missing one visit (and coming back to the study) are shown to depend, sometimes strongly, on variables such as the age, gender, depression, and level of school completed by the person most knowledgeable about the kid; the region of residence; the urban-rural status; the child's parent status (family status); the household income status; and the number of hours in daycare. Additionally, sometimes the probability of responding at a given cycle, or the probability of dropping out, even depends on the outcome variable of interest, the physical aggression score.

We found differences in point estimates when using the WGEE for the (artificially) monotone dataset and for the intermittent dataset. This is due to the fact that the factors influencing the dropouts are different to those influencing the likelihood of missing a visit and then coming back; as is evidenced by the models for dropouts in the (artificially) monotone dataset and the response models for the intermittent dataset. It is also evident because for the intermittent dataset, the models for the response probabilities for a given cycle *are* different depending on the observed history. This indicates that an analysis of the monotone dataset or of the artificially monotone dataset (which ignores all the data after a missed cycle) is not the wisest method to choose; even though it is easier than the most appropriate one. This is why, for

the dataset at hand, we recommend using all observations (before and after a missed cycle) in an intermittent WGEE analysis. (For more information see Carrillo, 2006)

So that, methods of analysis that naively assume a MCAR mechanism can be very misleading for this dataset. In fact we found that when we analyzed it using techniques such as complete case or available case analysis the results were very different to the ones obtained with more appropriate methods. Since a less stringent assumption about the missingness mechanism, such as MAR, is more suitable for this dataset, we claim that the WGEE method should be preferred instead of the GEE, although it requires more computational effort.

The standard errors obtained when using the WGEE approach are generally bigger than those for the GEE because in the former one needs to estimate the response probabilities for each cycle. And then, the uncertainty about the parameters in the response probability models gets added to the uncertainty in estimating the parameter of interest β . Nonetheless, if the missingness is MAR, as opposed to MCAR, a WGEE methodology should be adopted because it will eliminate, or at least reduce, the bias in the GEE estimates. The sort of biases that WGEE will correct for are those that may occur due to the differential missingness patterns accounted for by the factors included in the response probability models.

Finally, we want to note that due to missing items, throughout we ignored about 300 kids. For the same reason we also ignored, for some kids (among the 5,216 actually used), some cycles in which they had actually responded. This is a waste of information, especially in those cases in which there is only a couple of items missing. Therefore, a generalization of the present WGEE for survey data, or some other approach, is needed, to accommodate, in addition to missing waves, the possibility of missing items (or covariates).

REFERENCES

- Albert, P. S. (1999). "Longitudinal data analysis (repeated measures) in clinical trials". *Statistics in Medicine*, **18**, 1707-32
- Carrillo, I. (2006). Technical Report, MITACS/NPCDS Internship Program, Methodology Branch, Statistics Canada
- Carrillo, I., Chu, C., Su, W., and Xie, X. (2005). "A Longitudinal Study of Factors Affecting Children's Behaviour". *Proceedings of the Survey Methods Section*, Saskatoon 12-15 June 2005, Statistical Society of Canada
- Chen, J., Thompson, M. E., and Wu, C. (2004). "Estimation of Fish Abundance Indices Based on Scientific Research Trawl Surveys". *Biometrics*, **60**, 116-123
- Diggle, P. J., Heagerty, P., Liang, K-Y, and Zeger, S. (2002). *Analysis of Longitudinal Data* (2nd ed.). New York: Oxford University Press
- Fitzmaurice, G., Laird, N., and Ware, J. (2004). *Applied Longitudinal Analysis*, Hoboken: Wiley
- Kovacevic, M. S., and Shesh, N. R. (2002). "Log-Linear Modelling of Change Using Longitudinal Survey Data". *Communications in Statistics A*, **31**, 1815-35
- Liang, K-Y., and Zeger, S. L. (1986). "Longitudinal Data Analysis Using Generalized Linear Models". *Biometrika*, **73**, 13-22
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed". *Journal of the American Statistical Association*, **89**, 846-866
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data". *Journal of the American Statistical Association*, **90**, 106-121
- Rotnitzky, A., and Wypij, D. (1994). "A Note on the Bias of Estimators with Missing Data". *Biometrics*, **50**, 1163-70
- Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag

- Thomas, E. M. (2004). "Aggressive Behaviour Outcomes for Young Children: Change in Parenting Environment Predicts Change in Behaviour, Children and Youth Research Paper Series". Catalogue number 89-599-MIE, Statistics Canada
- Wu, C. (2003). "Optimal Calibration Estimators in Survey Sampling". *Biometrika*, **90**, 937-951