Analysis of Incomplete Binary Longitudinal Data with an Application to the National Population Health Survey

By

Scott McLeish

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Mathematics (Probability and Statistics)

Carleton University
Ottawa, Ontario

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-91544-8

Our file Notre référence
ISBN: 978-0-494-91544-8

Abstract

This thesis explores the impacts of drop-outs on longitudinal data analysis. Weighted generalized estimating equations (Fitzmaurice et al., 1994 and Robins et al., 1995) are employed to mitigate non-response bias associated with ignoring the nature of informative drop-outs. A simulation study, conducted to assess the performance of these methods under various circumstances and for different types of non-response, reveals that when data is not missing completely at random (MCAR), unweighted models are inherently biased. We observe that for drop-outs associated with our covariates, or other information already available on the respondents (MAR), we can use the aforementioned weighted generalized estimating equations to eliminate non-response bias.

We apply this theory to an analysis of incomplete longitudinal data from the National Population Health Survey focussing on the relationship between high blood pressure and some associated covariates such as body mass index. Simply ignoring the drop-outs would bias certain associations in the analysis.

Acknowledgements

I would like to thank Statistics Canada for ongoing support and providing me the opportunity to work with the National Population Health Survey (NPHS) data, especially, Marie Anderson, Dr. Tina Chui, Gina Thompson, Dr. Michael Wendt, and Dr. Georgia Roberts.

Particularly, I would like to thank my supervisor, Dr. Sanjoy Sinha, for his encouragement, guidance, and enduring support throughout my work.

Table of Contents

Chapter 1

Introduction

Whether data are collected through a sample survey, census or through administrative/non-survey data, non-response or missing data is a common issue. Data can be missing for many reasons, some legitimate, and understanding the nature of the missingness is important in order to avoid any possible response bias in statistical analysis. Rubin (1976) examines some of the reasons why non-response can occur and investigates methods for analyzing such incomplete data.

Non-response can be structural, deliberate, or even necessary. For example, surveys about labour force activity do not really apply to individuals who are below the legal working age. So they would not be asked to answer these questions. The same would be true for a health survey with questions on pregnancy (not applicable to male respondents). In these cases, non-response is neither random nor problematic – so long as we account for it properly. If we produce an estimate of the percentage of Canadians who are unemployed, including those who are not yet of legal working age, then this will obviously produce a bias in our estimate. In such cases, a simple solution is to remove the subjects who are out of scope.

In many cases, however, non-response is not structural and is left to the respondent's discretion. In these cases, the questions being answered are applicable to the subject but this does not necessitate a valid response. We can classify 2 types of non-response – complete and item. Complete non-response occurs when a subject chooses not to participate in the study at all. This will imply that no information will be available on the subject and it is difficult to determine if the subject's decision not to respond is associated or even driven by the nature of

the study itself. Item non-response occurs when a subject participates in the study, but does not complete every question. For example, respondents may choose not to answer a question on their smoking habits (i.e. how often they smoke) but they do answer questions about their demographic characteristics (e.g., age, sex, etc.). In this case, while we still do not know if their decision not to respond to the smoking question is associated or driven by what their smoking habits truly are, we do possess some information about the subjects.

In the analysis of non-response data, sometimes we define a missing data mechanism to describe the relationship between non-response and respondent characteristics. When non-response is completely random (not associated whatsoever with the variables of interest), we can simply ignore this mechanism, since there is no relationship between the variables in study and the subject's probability of responding. However, if an association does exist between the variable of interest and the propensity to respond, then ignoring this association will lead to biased results. For example, if people who do smoke are far less likely than non-smokers to respond to the smoking habits question, any analysis ignoring this mechanism will underestimate the prevalence of smoking. In these cases, we refer to the missing data mechanism as informative. Little and Rubin (1987) and Allison (2001) provide a comprehensive overview of missing data analysis.

In this thesis, I investigate the use of weighted generalized estimating equations, as suggested by Fitzmaurice et al. (1994) and Robins et al. (1995), to handle informative non-response in a longitudinal setting. Longitudinal data present a couple of complexities that are not relevant to cross-sectional data. First of all, since observations come from the same subjects over time, these observations cannot (without a certain naivety) be considered independent. For this reason, generalized estimating equations (Liang and Zeger (1986)) are employed, which have the flexibility to include within-subject correlation structures. Secondly, longitudinal studies present a unique type of non-response in that the subject may respond for some cycles, but not for others (which may be treated as a form of item non-

response since some information is available for these respondents). Specifically, this thesis focuses on drop-outs which occur frequently with longitudinal studies, as respondents may tire of the survey, emigrate, become difficult to track or die.

This thesis reviews the theory relating to generalized linear models for longitudinal analysis and generalized estimating equations. Missing data problems and methods for analyzing such missing data in the framework of weighted generalized estimating equations are reviewed. A simulation study was undertaken to investigate the performance of weighted generalized estimating equations for different types of missing data mechanisms and under different assumptions (e.g., correlation structures, response models, etc.). Finally, the theory is applied to the National Population Health Survey, a longitudinal study undertaken by Statistics Canada (1994/95-2008/09). This study focuses on the prevalence of high blood pressure for adult Canadians over time using body mass index (BMI), physical activity and other measures/characteristics as covariates. Different models (logistic regression, and weighted and unweighted generalized estimating equations) are explored for analyzing the data and characteristics of the drop-out or missing data mechanism. Both the simulation study and the analysis of the NPHS data indicate that one can attain considerable gain in efficiency from the weighted generalized estimating equations approach when analyzing incomplete longitudinal data with a drop-out mechanism.

Chapter 2

Theory: Literature review

2.1 Generalized linear models

This section introduces the use of generalized linear models (GLMs) to perform regression analysis. Interest lies in regression analysis to gain a better comprehension of the association between a variable of interest and selected covariates.

Introduced by Nelder and Wedderburn (1972) (although models within the GLM class were developed before), generalized linear models (GLMs) are a generalization of normal linear models in which instead of only measuring the relationship between y and covariates X via a direct linear link such as $E(y) = X'\beta$, we allow for more flexibility and assume that they are related through a link function $g(\cdot)$. This flexibility allows us to perform alternative regression analysis, such as logistic regression for binary data. In addition to the link function, another important characteristic of generalized linear models is that the response, y, has a distribution belonging to the exponential family of distributions including normal, binomial, and Poisson.

The link function $g(\cdot)$ defines the relationship between the expectation of y and the linear predictor, η. Different link functions exist for different circumstances such as:

1. Identity link: $g(\mu) = \mu = X'\beta$ (used for Normal distribution)
2. Log link: $g(\mu) = \log(\mu) = X'\beta$ (used for Poisson distribution)

3. Logit link: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = X'\beta$ (used for Binomial distribution)

In this thesis, our focus is on a binomial response, y, and the logit link will be used for the analysis of the binomial data.

2.1.1 Maximum likelihood estimation for generalized linear models

We define the likelihood function as the probability of the observed data expressed as a function of our parameter(s) of interest, θ.

Let $f(y_i|\theta) = p(y_i = 1|\theta)^{y_i}(1 - p(y_i = 1|\theta))^{1-y_i}$. Essentially, the likelihood of our parameter values, θ, given observed y_i is the same as the joint probability of the observed y_i given θ.

For a set of n independent observations, $y_1, y_2, y_3, \ldots, y_n$, the likelihood can be defined as

$$L(\theta|y) = f(y_1, y_2, y_3, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta). \tag{2.1.1}$$

For computational simplicity, we use the natural logarithm of this function called the log-likelihood function:

$$l(\theta|y) = \sum_{i=1}^{n} \log f(y_i|\theta). \tag{2.1.2}$$

The maximum likelihood (ML) estimator of θ is obtained by maximizing this log-likelihood function. Equivalently, for the ML estimator of θ, we solve the estimating equation (referred to as the Score equation) $S(\theta) = \frac{\partial l(\theta|y)}{\partial \theta} \equiv 0$, with respect to θ.

This score equation can be solved by an iterative method as described in the following section.

2.1.2 Newton – Raphson algorithm and Fisher scoring technique for ML estimation

Often, we cannot find a closed solution to $S(\theta) = 0$. In these cases, we can rely on an optimization algorithm. One such algorithm, which is commonly used for maximum likelihood estimation, is the Newton-Raphson algorithm.

Let us consider a polynomial $S(\theta)$ and assume that θ_0 is the true root of $S(\theta) = 0$. Using a first-order Taylor series approximation, we can write

$$0 \equiv S(\theta_0) \approx S(\theta) + S'(\theta)(\theta_0 - \theta), \tag{2.1.3}$$

which implies that

$$\theta_0 \approx \theta - [S'(\theta)]^{-1}S(\theta), \tag{2.1.4}$$

for a given θ.

To find the best estimate for θ_0, the Newton-Raphson algorithm is applied, which continues until the difference between θ and θ_0 is negligible (our iterative estimates converge). This Newton-Raphson algorithm can be described as follows:

1. Choose starting value θ^1. Set $M = 1$.
2. Calculate $\theta^{M+1} = \theta^M - [S'(\theta^M)]^{-1}S(\theta^M)$
3. Check for convergence. If the difference between θ^M and θ^{M+1} is negligible, then stop. Otherwise, set M = M+1 and return to step 2.

Choosing the starting value of θ is non-trivial. If the likelihood function has many local maximums, then depending on the starting point selected, different answers could be found. It is imperative to try different starting values to determine if the root found is the true maximizing value for the likelihood function.

Let us define the information matrix as $I(\hat{\theta}) = -\frac{\partial s(\hat{\theta})}{\partial \hat{\theta}}$. Maximum likelihood estimation provides a model-based covariance matrix for the fitted model given by $\hat{V}ar(\hat{\theta}) = \{I(\hat{\theta})\}^{-1}$.

Example: Binary data

For binary responses $y_1, y_2, y_3, \ldots, y_n$, we consider a likelihood function given by

$$L(\theta|y) = f(y_1, y_2, y_3, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} p(y_i|\theta)^{y_i}(1 - p(y_i|\theta))^{1-y_i}. \tag{2.1.5}$$

The log-likelihood is obtained as

$$\begin{aligned} l(\theta|y) = \log(L(\theta|y)) &= \sum_{i=1}^{n} \log f(y_i|\theta) = \sum_{i=1}^{n} \log\left(p(y_i|\theta)^{y_i}(1 - p(y_i|\theta))^{1-y_i}\right) \\ &= \sum_{i=1}^{n} \{y_i \log(p(y_i|\theta)) + (1 - y_i)\log(1 - p(y_i|\theta))\}. \end{aligned} \tag{2.1.6}$$

Then the score function is

$$S(\theta) = \frac{\partial l(\theta|y)}{\partial \theta} = \sum_{i=1}^{n} y_i \frac{\partial p(y_i|\theta)}{\partial \theta} \frac{1}{p(y_i|\theta)} - (1 - y_i)\frac{\partial p(y_i|\theta)}{\partial \theta} \frac{1}{(1 - p(y_i|\theta))}. \tag{2.1.7}$$

For simplicity, let $\mu_i = p(y_i|\theta)$. For binomial data where $logit(\mu_i) = X_i'\beta$, we

define our parameter $\theta = \beta$. We can write

$$\mu_i = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)}, \tag{2.1.8}$$

so that

$$\begin{aligned} \frac{\partial \mu_i}{\partial \beta} &= \frac{X_i \exp(X_i'\beta)\,(1 + \exp(X_i'\beta)) - X_i \exp(X_i'\beta)\exp(X_i'\beta)}{(1 + \exp(X_i'\beta))(1 + \exp(X_i'\beta))} \\ &= \frac{X_i \exp(X_i'\beta) + X_i \exp(X_i'\beta)\exp(X_i'\beta) - X_i \exp(X_i'\beta)\exp(X_i'\beta)}{(1 + \exp(X_i'\beta))(1 + \exp(X_i'\beta))} \\ &= \frac{X_i \exp(X_i'\beta)}{(1 + \exp(X_i'\beta))(1 + \exp(X_i'\beta))} = X_i\mu_i(1 - \mu_i). \end{aligned} \tag{2.1.9}$$

Therefore,

$$S(\beta) = \sum_{i=1}^{n} y_i X_i(1 - \mu_i) - (1 - y_i)\, X_i\mu_i = \sum_{i=1}^{n} X_i(y_i - \mu_i). \tag{2.1.10}$$

Finally, the information matrix is given by

$$I(\beta) = -\frac{\partial S(\beta)}{\partial \beta} = \sum_{i=1}^{n} X_i' \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^{n} \mu_i(1 - \mu_i) X_i X_i'. \tag{2.1.11}$$

We can now use the Newton-Raphson algorithm to find $\hat{\beta}$ that maximizes the likelihood function for the binary data.

Fisher scoring is a slight adjustment to the Newton-Raphson in that instead of using the observed derivative of the Score or the information matrix $I_O(\theta) = -\frac{\partial S(\theta)}{\partial \theta}$, we use its expectation or the expected information matrix. This adjusts our algorithm as follows:

1. Choose starting value θ^1. Set $M = 1$.

2. Solve for $\theta^{M+1} = \theta^M + E\left[-\frac{\partial S(\theta^M)}{\partial \theta^M}\right]^{-1} S(\theta^M) = \theta^M + I_E[\theta^M]^{-1} S(\theta^M)$
3. Check for convergence. If the difference between θ^M and θ^{M+1} is negligible, then stop. Otherwise, set M = M+1 and return to step 2.

Fisher scoring is useful when the observed Information matrix is difficult to find.

2.2 Longitudinal data analysis

Many surveys and studies are longitudinal in design to permit the study of effects over time. This is important for health studies, in particular, to understand how different behaviours and other demographic variables interact with the primary variable of interest. For example, the primary variable can be the indicator of heart disease or lung cancer that may be related to eating habits, smoking prevalence, etc.

Clinical trials represent one of the most popular types of longitudinal studies. A good example of a clinical trial is when a pharmaceutical company wishes to test a new drug. They randomly select subjects who are in scope for the study (they could be people who have a particular illness, for example). Some of these subjects will be administered the new drug while the rest will be given some kind of a control (such as the drug currently used or a placebo). They monitor the subjects over time to observe certain things such as:

- Whether the new drug works
- If the new drug works faster than the current drug / placebo
- If the new drug is more effective than the current drug / placebo
- Whether there is evidence of any side effects

In this thesis, we focus on data from a longitudinal health-based survey which is used to monitor individuals' health over time. The respondents are meant to represent the general Canadian public. While some longitudinal studies follow individuals in real time (observations are ongoing – any change is documented immediately), this survey is cycle-based and follows up with respondents every 2 years.

2.2.1 Example: Logistic regression for longitudinal data

Suppose we wish to measure the longitudinal relationship between a binary variable of interest, y, and some covariates X. Let us assume that we have received responses for all times/periods $t = 1, \ldots, T$. Consider some artificial data as shown in Table 2.2.1.

Table 2.2.1 Sample longitudinal data set: Complete data for continuous X and binary Y

Subject	Time	X	Y
1	1	1.216003	0
1	2	1.823071	0
1	3	3.164482	1
1	4	2.11988	1
2	1	1.461545	0
2	2	3.619435	1
2	3	0.096019	0
2	4	2.217957	0
3	1	0.123106	0
3	2	1.181112	0
3	3	0.491442	0
3	4	3.270173	1

Source: None – artificial data.

If we have no interest in the longitudinal nature of the responses, and we assume all observations are independent and identically distributed, then we simply create a master data set that ignores the fact that observations are being received from the same subjects and are being received at different times. In this case, we look to find $\hat{\beta}$ to best represent the logistic relationship between y and X: $p(y) = \exp(X'\beta)/(1 + \exp(X'\beta))$. This can be done using the MLE approach described earlier.

This is, of course, a very hazardous approach to take since the observations are very likely not i.i.d. as assumed and perhaps more pressing, we are completely ignoring the fact that changes over time likely occur. So, now let us consider a slightly less naïve approach while retaining the assumption that the observations are i.i.d., but we will account for the longitudinal nature of the data. To do this, we can simply introduce a new covariate in addition to X to represent the period t. Here let X^* include X as well as t. In this case, we look to find $\hat{\beta}$ to best represent the logistic relationship between y and X^*: $p(y) = \exp(X^{*\prime}\beta)/(1 + \exp(X^{*\prime}\beta))$. This can be done using the MLE approach described earlier.

While this approach does take the longitudinality into account (and rightfully so since time can very well be a determinant), it still ignores any within-subject association.

2.3 Generalized estimating equations

Introduced by Liang and Zeger (1986), unlike the maximum likelihood estimation method shown above, generalized estimating equations allow for within-subject variance when estimating regression parameters.

Specifications that are necessary for generalized estimating equations include:

1. Mean function (Systematic component): Relationship between the dependent variable and covariates is defined by $g(\mu_{it}) = \eta_{it} = X_{it}'\beta$ where $E(Y_{it}|X_{it}) = \mu_{it}$
2. Variance function (Random component): Function representing the variance of Y_{it} given the covariates is defined by $Var(Y_{it}|X_{it}) = v(\mu_{it})$
3. Working correlation matrix: Deviating from the typical GLM framework, we require a correlation structure to be defined to represent the within-subject associations. The working correlation matrix can be defined by $Corr(Y_i) = R(\alpha)$ where α is the correlation parameter.

2.3.1 Derivation of generalized estimating equations

Consider a simple linear case where we have standardized errors defined by

$$\epsilon = \frac{y - \mu(\theta)}{\sqrt{var(y)}}, \tag{2.3.1}$$

with $\mu(\theta)$ being the fitted expectation of y with parameter θ. To have the best possible fit, we would want μ such that we can minimize the sum of the squares of the errors

$$\sum_{i=1}^{N} {\epsilon_i}^2 = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\theta))^2}{var(y_i)}, \tag{2.3.2}$$

where y_i is the response y for individual i with mean function $\mu_i(\theta)$.

To minimize the above, we would simply find its derivative with respect to our parameter θ, and set to 0. This is the basis for solving weighted sum of squares. In matrix notation, we consider the weighted sum of squares formula where we define V_i as the covariance matrix of y_i

$$W(y) = \sum_{i=1}^{K} (y_i - \mu_i)' V_i^{-1} (y_i - \mu_i), \tag{2.3.3}$$

where the weights are given by V_i^{-1}, μ_i is a function of θ, $y_i = (y_{i1}, y_{i2}, \dots, y_{iT})$ and $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{iT})$. If we wish to find $\hat{\theta}$ that minimizes $W(y)$, we take its derivative with respect to θ, set to 0 and solve for $\hat{\theta}$. That is, we solve the equation

$$W'(y) = \sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \theta}' V_i^{-1} (y_i - \mu_i) = 0, \tag{2.3.4}$$

for $\hat{\theta}$.

The generalized estimating equations approach solves equations of the form

$$S(\theta) = \sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \theta}' V_i^{-1} (y_i - \mu_i) = 0, \tag{2.3.5}$$

for an estimator of θ, where V_i is a TxT matrix representing the working covariance matrix of y_i. V_i can be found as

$$V_i = \varphi A_i^{1/2} R(\alpha) A_i^{1/2}, \tag{2.3.6}$$

where $A_i = diag\{v(\mu_{it})\}$, $R(\alpha)$ is the working correlation matrix and φ denotes the dispersion parameter and $D_i = \frac{\partial \mu_i}{\partial \beta}$.

To solve the estimating equation $S(\theta) = 0$ for θ, we can use the Newton-Raphson iterative algorithm as described earlier. We discuss the computations in a later example.

2.3.2 Working correlation structures for generalized estimating equations

To select the working correlation matrix $R(\alpha)$, various options exist including independent, exchangeable and serial correlation.

The assumption of an independent correlation structure is the simplest possible case. Effectively, we assume that no within-subject association exists. Because of this, $R(\alpha)$ does not need to be estimated and is given as:

$$R(\alpha) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Exchangeable correlation structure assumes that for all $j \neq k$, the correlation coefficient between y_{ij} and y_{ik} is common: $\alpha_{jk} = \alpha$. An example of such a structure would be the case where there is no clear relationship (or order) between j and k. For example, if we were considering data associated with units which come from groups (e.g, students in a class or vegetables from a farm), the association between units from the same group could be considered exchangeable.

This type of correlation structure requires us to estimate the working correlation matrix $R(\alpha)$. An estimate of α is

$$\hat{\alpha} = \frac{1}{\varphi(N^* - p)} \sum_{i=1}^{N} \sum_{j<k} e_{ij} e_{ik}, \tag{2.3.7}$$

where

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}, \tag{2.3.8}$$

and

$$N^* = 0.5 \sum_{i=1}^{N} T_i(T_i - 1). \tag{2.3.9}$$

Here e_{ij} represents the standardized Pearson residuals, N represents the total number of subjects and T_i represents the number of observations for subject i. Essentially, this type of estimation can be thought of as the average of the product of residuals for all subjects and all possible combinations of observations where $j \neq k$. Estimation of the dispersion parameter will be discussed later in this section.

Our working correlation matrix is therefore estimated by:

$$R(\hat{\alpha}) = \begin{bmatrix} 1 & \hat{\alpha} & \cdots & \hat{\alpha} \\ \hat{\alpha} & 1 & \cdots & \hat{\alpha} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha} & \hat{\alpha} & \cdots & 1 \end{bmatrix}$$

For serial or autoregressive correlation, we assume that the association between observations decreases as the time between them increases. This form of correlation structure is quite common in longitudinal data. For example, observations that occurred one after the other will have stronger association as compared to the first and last observations. Typically, we assume that the correlation coefficient follows an exponential pattern such that the correlation between cycles j and k is given by $\alpha_{jk} = \alpha^{|j-k|}$.

In this case, an estimate of α can be obtained as

$$\hat{\alpha} = \frac{1}{\varphi(N^* - p)} \sum_{i=1}^{N} \sum_{j<T_i} e_{ij} e_{i(j+1)}, \tag{2.3.10}$$

where

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}, \tag{2.3.11}$$

and

$$N^* = \sum_{i=1}^{N} (T_i - 1). \tag{2.3.12}$$

Essentially, this type of estimator can be thought of as the average of the product of residuals for all subjects and all possible combinations of chronologically adjacent observations.

The estimated working correlation matrix takes the form:

$$R(\hat{\alpha}) = \begin{bmatrix} 1 & \hat{\alpha} & \hat{\alpha}^2 & \cdots & \hat{\alpha}^{|T_i-1|} \\ \hat{\alpha} & 1 & \hat{\alpha} & \cdots & \hat{\alpha}^{|T_i-2|} \\ \hat{\alpha}^2 & \hat{\alpha} & 1 & \cdots & \hat{\alpha}^{|T_i-3|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}^{|T_i-1|} & \hat{\alpha}^{|T_i-2|} & \hat{\alpha}^{|T_i-3|} & \cdots & 1 \end{bmatrix}$$

For the undefined structure, we make no assumption about the correlation matrix. In this case, we need to estimate each possible correlation coefficient independently – as opposed to serial or exchangeable correlation structures, which only require the estimation of a single parameter. Note that this increases the computational requirements significantly.

Here we can use the estimators

$$\hat{\alpha}_{jk} = \frac{1}{\varphi(N - p)} \sum_{i=1}^{N} e_{ij} e_{ik}. \tag{2.3.13}$$

We can think of the estimator of the correlation coefficient between y_{ij} and y_{ik} as the average of the product of residuals for time j and time k for all subjects. The estimated correlation matrix takes the form

$$R(\hat{\alpha}) = \begin{bmatrix} 1 & \hat{\alpha}_{12} & \hat{\alpha}_{13} & \cdots & \hat{\alpha}_{1T_i} \\ \hat{\alpha}_{21} & 1 & \hat{\alpha}_{23} & \cdots & \hat{\alpha}_{2T_i} \\ \hat{\alpha}_{31} & \hat{\alpha}_{32} & 1 & \cdots & \hat{\alpha}_{3T_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}_{T_i1} & \hat{\alpha}_{T_i2} & \hat{\alpha}_{T_i3} & \cdots & 1 \end{bmatrix}$$

For the remainder of the thesis, the working correlation structures consist of independent, exchangeable, and serial. We note that whenever the correlation matrix must be estimated, α is dependent upon μ_{it}, y_{it} and $v(\mu_{it})$, where μ_{it} and $v(\mu_{it})$ rely on a value for β.

2.3.3 Dispersion parameter in generalized estimating equations

Quasi-likelihood estimation (Wedderburn, 1974), from which generalized estimating equations are derived, allows for overdispersion (actual variance is greater than expected variance) or underdispersion (actual variance is less than expected variance). This is due to the fact that instead of using the true variance of y, we use a variance function V. This function includes a scale or dispersion parameter which can be used to account for any over or under dispersion associated with an inaccurate V. For binomial data, since V is a function of the mean, over or under dispersion can arise when $var(y_{it}) \neq \mu_{it}(1 - \mu_{it})$.

In some cases, it is necessary to estimate the dispersion parameter along with the correlation parameters of interest. When sampling weights are not considered, the dispersion parameter can be estimated by

$$\hat{\varphi} = \frac{1}{N-p}\sum_{i=1}^{K}\sum_{t=1}^{T_i} {e_{it}}^2, \tag{2.3.14}$$

where N is total number of observations: $N = \sum_{i=1}^{K} T_i$, p is the number of parameters or degrees of freedom, K is the number of subjects, and T_i the number of observations for subject i. e_{it} represents the Pearson residual for subject i at time t, that is,

$$e_{it} = \frac{y_{it} - \mu_{it}}{\sqrt{v(y_{it})}}. \tag{2.3.15}$$

For binary data, this can be considered as

$$e_{it} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1-p_{it})}}. \tag{2.3.16}$$

In the context of the National Population Health Survey, sampling weights are needed to accurately analyze the data and so we adjust this formula to account for the fact that each respondent, i, is accounting for w_i population units. Therefore, within a survey context, the dispersion parameter is estimated by

$$\hat{\varphi} = \frac{1}{N-p}\sum_{i=1}^{K} w_i \sum_{t=1}^{T_i} {e_{it}}^2, \tag{2.3.17}$$

where $N = \sum_{i=1}^{K} w_i T_i$.

2.3.4 Model-based variance estimation of parameter estimates

As in the case of the maximum likelihood estimation, $\{I(\theta)\}^{-1}$ can be used as an estimate for the model-based covariance matrix of $\hat{\theta}$. That is, for a correctly

specified covariance matrix , V_i, $\widehat{Cov}(\hat{\theta}) = \{I(\theta)\}^{-1} = \left\{\sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \theta}' V_i^{-1} \frac{\partial \mu_i}{\partial \theta}\right\}^{-1}$. This estimator is unfortunately not consistent if the working correlation matrix is misspecified or $Cov(y_i) \neq V_i$. Therefore, introduced by Liang and Zeger (1986), the following empirical estimate (also called sandwich estimate) is used, which remains consistent even if the working correlation matrix is misspecified,

$$\widehat{Cov}(\hat{\theta}) = [I(\theta)]^{-1} I_1 [I(\theta)]^{-1}, \tag{2.3.18}$$

where

$$I_1 = \sum_{i=1}^{K} D_i' V_i^{-1} Cov(y_i) V_i^{-1} D_i. \tag{2.3.19}$$

In this estimation, we replace $Cov(y_i)$ with $(y_i - \mu_i)(y_i - \mu_i)'$. For proof, please refer to Liang and Zeger (1986).

2.3.5 Example: Generalized estimating equations with binary response

Recall from the maximum likelihood estimation, for the case where y_i represents a single observation (instead of T observations), that the score for binomial data can be found as

$$S(\theta) = \frac{\partial l(\theta|x)}{\partial \theta} = \sum_{i=1}^{n} y_i \frac{\partial p(y_i|\theta)}{\partial \theta} \frac{1}{p(y_i|\theta)} - (1 - y_i) \frac{\partial p(y_i|\theta)}{\partial \theta} \frac{1}{\left(1 - p(y_i|\theta)\right)}. \tag{2.3.20}$$

We let $\mu_i = p(y_i|\theta)$. We assume $logit(\mu_i) = X_i'\beta$, so that

$$\mu_i = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)}, \tag{2.3.21}$$

and

$$\frac{\partial \mu_i}{\partial \beta} = X_i \mu_i (1 - \mu_i). \tag{2.3.22}$$

Let us consider the notation $D_i = \frac{\partial \mu_i}{\partial \beta}$ and let $V_i = \mu_i(1 - \mu_i)$. Then,

$$\begin{aligned} S(\theta) &= \sum_{i=1}^{n} \{D_i' y_i \mu_i^{-1} - (1 - y_i) D_i' (1 - \mu_i)^{-1}\} \\ &= \sum_{i=1}^{n} [D_i' y_i (1 - \mu_i) - D_i' (1 - y_i) \mu_i][\mu_i (1 - \mu_i)]^{-1} \\ &= \sum_{i=1}^{n} D_i' V_i^{-1} (y_i - \mu_i). \end{aligned} \tag{2.3.23}$$

Note that this is the standard GEE formulation. Now, let us consider multivariate (longitudinal) y. We still use the *logit* link function, $g(\mu_{it}) = logit(\mu_{it}) = \eta_{it} = X_{it}'\beta$. Therefore, $\mu_{it} = \frac{\exp(X_{it}'\beta)}{1+\exp(X_{it}'\beta)}$ and $v(\mu_{it}) = E(\mu_{it}^2) - E(\mu_{it})^2 = \mu_{it} - \mu_{it}^2 = \mu_{it}(1 - \mu_{it})$.

Let us consider the notation $\mu_{it} = p_{it}$, where p_{it} represents the probability of success for subject i at time t. Then,

$$p_{it} = \frac{\exp(X_{it}'\beta)}{1 + \exp(X_{it}'\beta)}, \tag{2.3.24}$$

and

$$v(\mu_{it}) = p_{it}(1 - p_{it}). \tag{2.3.25}$$

We define

$$A_i = diag\{v(\mu_{it})\} = diag\{p_{it}(1 - p_{it})\}, \tag{2.3.26}$$

and

$$V_i = \varphi A_i^{1/2} R(\alpha) A_i^{1/2}. \tag{2.3.27}$$

In this step, D_i is a $T_i$xp matrix, where p represents the number of parameters in β. We have

$$D_{it} = \frac{\partial \mu_{it}}{\partial \beta} = \left\{\frac{\partial \mu_{it}}{\partial p_{it}} \cdot \frac{\partial p_{it}}{\partial \beta}\right\} = \{p_{it}(1 - p_{it})X_{it}\}. \tag{2.3.28}$$

To define the covariance of the ith response vector, we select a working correlation structure. For those structures where estimation of the correlation matrix is required, the estimators are obtained the as shown above except that the Pearson residuals are defined by

$$e_{ij} = \frac{y_{it} - \mu_{it}}{\sqrt{v(\mu_{it})}} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1 - p_{it})}}. \tag{2.3.29}$$

Finally, the estimator $\hat{\beta}$ of β is obtained by solving the equation

$$S(\beta) = \sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \mu_i) = 0, \tag{2.3.30}$$

with respect to β, where $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ and $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iT})$.

Often, as is the case for MLE, no closed-form solution exists for $S(\beta) = 0$. Therefore, we rely on an iterative method. We
define

$$S'(\beta) = \frac{\partial S(\beta)}{\partial \beta} = -I(\beta) = -\sum_{i=1}^{K} D_i' V_i^{-1} D_i. \tag{2.3.31}$$

Then, $\hat{\beta}$ can be obtained from the iterative equations

$$\beta^{M+1} = \beta^M + [I(\beta^M)]^{-1} S(\beta^M), \tag{2.3.32}$$

for M=0,1,2,...

2.4 Missing data analysis

Statistics is the science of analyzing data to gain a better understanding of a general population of interest. In a perfect world, the data that are being analyzed are perfectly representative of everyone in that population of interest with no error. For this to be the case, the data must be collected:

1. For every member in the population (i.e. census)
2. Free of measurement error (e.g., respondents misinterpreting questions and answering incorrectly)
3. Free of item non-response (response to every variable must be collected from the population)

Unfortunately, censuses are extremely difficult to conduct (they are expensive, require cooperation from everyone in population). Some datasets are collected through administrative records (e.g. registration for a local rugby club), but administrative data are not collected for statistical purposes and are limited in scope. In addition, administrative datasets can still include measurement error (e.g., birth date not being copied into the registration database properly due to unintelligible penmanship on the paper form) and certain fields deemed less important (such as demographic characteristics on a rugby registration form) might be left blank.

Researchers interested in studying the general population of individuals for health, economic, or social topics are often not able to use administrative data. For these purposes, they rely on population surveys. Due to the obstacles surrounding the conduct of a population census (e.g., cost, legally forcing participation, etc.) most countries only run censuses every 5 to 10 years. The content of these surveys is also limited (since it would not be fair to force everyone to answer lengthy surveys). In addition, there is a real danger of measurement error in censuses. With so many people responding, it is not feasible to have interviewers walk each respondent

through the questions and so individual respondents are left to interpret questions for themselves. Finally, with so many people responding, it is impossible to follow-up with everyone who left certain fields blank or provided invalid responses.

All this is to say that while perfect data would be a census, free of measurement error and free of item non-response, this is not realistic. So, instead of relying on a census, we allow for sampling error and only survey a sample of the total population. The major goal here is that the sample be statistically representative of the general population. With a sound sampling design (which is not outlined in this thesis) and no measurement error or non-response error, we accomplish this. However, it is unlikely that everyone in the sample will choose to respond (especially if they are not legally obligated to do so). If the response population (like the sample population) is still representative of the general population of interest, then non-response can be considered a simple shrinking of the sample size. On the other hand, if there is a relationship between the variables being collected and individuals' willingness to respond, then our response population can no longer be seen as representative of the general population.

2.4.1 Complete and item non-response

Non-response can present itself in different varieties. For example, for a simple cross-sectional survey, we incur complete non-response (where the sampled unit fails to respond at all) and item non-response (where the sampled unit responds to the survey but fails to provide a valid response to all questions). Often, we do not have any additional information for complete non-respondents. This can sometimes be rectified by the use of general target population characteristics. Item non-response implies that some information is available for the respondent.

Before classifying types of non-response, it is important to fully understand the impact of item non-response. Suppose respondent A does respond to the survey

and responds to every question except question 4. If we have no interest in the variable derived from question 4, this item non-response is irrelevant. If, however, we do want to study the variable derived from question 4, we can consider this item non-response from 2 different perspectives:

1. Respondent A might as well have not responded at all – item non-response becomes complete non-response.
2. We can rely on the other information provided by respondent A to gain an understanding of what the response to question 4 could have been.

Which approach to use is dictated by the additional or auxiliary information obtained from respondent A. If question 4 is entirely independent of the other variables collected, than approach 1 is appropriate. If, however, there is an association between question 4 and the variables yielded by the other questions on the survey, then we can take approach 2. Understanding the characteristics of non-respondents is paramount to being able to minimize non-response error.

Non-response is classified into 3 different categories:

1. Missing completely at random (MCAR): There is no relationship whatsoever between the variables of interest and propensity to respond.
2. Missing at random (MAR): There is a relationship between the variables of interest and propensity to respond. However, it can be explained by auxiliary information available for all individuals.
3. Non-ignorable (NI): There is a direct relationship between the variables of interest and propensity to respond. It cannot be entirely explained by auxiliary information.

When data are missing completely at random (MCAR), we can treat non-response as the simple shrinking of the sample size as mentioned above. The consequences of this form of non-response are that the sampling variance will increase (due to the smaller sample size). This could be dealt with pre-emptively by the researcher

when developing the sampling strategy by sampling more units. During analysis, we can simply ignore missing responses entirely as this will not bias our results.

When the missingness is informative (MAR or NI), then we cannot simply ignore the records which are missing data. For data with MAR, we can apply different techniques such as weighting or imputation so that the resulting estimates are representative of the population of interest. But without doing something to explicitly tackle the non-response problem, our results will be biased.

2.4.2 Missing data in longitudinal studies

Longitudinal studies offer a complex form of non-response since we try to follow individuals over time. For the first wave of the survey, non-response exists the way it does for a cross-sectional survey; there are sampled units who fail to respond at all and some item non-response for those who did respond to the survey. In fact, for each individual wave of the survey these forms of non-response exist. This creates an issue when looking at the data from a longitudinal perspective. If individuals miss one cycle, they're excluded from the analysis despite all the information they already provided.

For longitudinal studies, complete non-response for a cycle commonly comes through the form of attrition where respondents "disappear" after a certain number of cycles. This is due to many reasons such as:

1. Relocating (making it difficult for the surveyor to follow-up)
2. Getting tired of the survey (refusing to participate anymore)
3. Physical impairment (death, illness, etc.)

The 3rd reason listed above is sometimes considered a valid outcome (and thus a response) depending on the survey topic and of course requiring that this information is relayed to the surveyor.

Consider the following table outlining the response patterns of individuals in a longitudinal survey. "." represents a complete non-response, "P" represents a partial response or some item non-response, and "Y" represents a complete response.

Table 2.4.1 Sample longitudinal data set: Response patterns with partial (P) and complete non-response (.)

Respondent	Time 1	Time 2	Time 3	Time 4	Time 5
1	Y	Y	Y	Y	.
2	Y	Y	Y	.	.
3	Y	Y	.	.	.
4	Y	.	Y	Y	Y
5	P	P	.	.	.
6	Y	P	P	.	.
7	Y	Y	Y	P	Y
8	Y	Y	Y	Y	Y
9	P	Y	Y	Y	.
10	.	.	.	.	.

Source: None – artificial data.

If we are only interested in complete data for all 5 times, only data for respondent 8 could be used. If we need response for all 5 times but partial response is permissible, then respondents 7 and 8 would qualify. By making choices like this, we essentially waste all of the information that was received from those respondents who do not respond at every time.

We consider 2 types of longitudinal non-response not relevant to a cross-sectional survey:

1. Drop-outs (responds for first x waves, then "disappears")
2. Intermittent missingness (periodically misses a wave, but returns to the survey).

Drop-outs described above relate to the attrition of a longitudinal survey's respondents. This form of longitudinal missingness is the focus of this thesis.

In the example shown above, treating partial response as complete response, respondents 1, 2, 3, 5, 6, and 9 are drop-outs. Respondent 4 has intermittent missingness (missed time 2 but responded to all other times). Respondent 10 is a complete non-respondent and respondents 7 and 8 responded to every time.

We can handle longitudinal non-response in a similar way that item non-response is handled since some information is available on the respondents.

2.4.3 Weighting and imputation methods to handle non-response

Different methods are employed to deal with the problems of non-response. Imputation methods involve replacing missing values or "filling in the blanks". These methods vary and can depend on the type of data collected. For example, "carry-forward imputation" is used in longitudinal analysis and involves copying the last known value. Other popular methods include multiple imputation, nearest-neighbour, etc. These methods are examined in the literature, such as Rubin (1987), Chen and Shao (2000), and Little and Rubin (1987). They are not explored in this thesis.

Alternatively, weighting methods apply response weights (similar to sampling weights) to all non-missing values. If the missingness is completely at random (there is no relationship between the data and the missingness) then every non-missing observation will have the same weight. If, however, the missingness is deemed informative during the weighting process, certain observations will be weighted more heavily than others (to best represent those that did not respond).

For example, suppose we were trying to estimate the number of people in a city who play rugby and the true populations are given as follows:

Total: 100,000
Rugby players: 5,000

We take a 1 in 10 simple random sample of 10,000 people but they do not all respond. The counts of respondents are given below:

Total: 8,000
Rugby players: 300

We happen to know that the total population is 100,000 so if we assume that the non-response mechanism is MCAR, we could simply weight all individuals as:

$W = 100,000/8,000 = 12.5$

This would give the following estimates notably underestimating the number of rugby players.

Total: 100,000
Rugby players: 3,750

Unfortunately, without any other information, this bias is unavoidable. But suppose, we knew how many men and women were in the city and, in our survey, in addition to asking about rugby, we asked for the respondent's sex. Our city's population by sex is given as:

Men: 45,000
Women: 55,000

Of our respondents, we can break them down into:

Total: 8,000

Men: 1,620

Women: 6,380

Rugby players: 300

Men: 126

Women: 174

Now, we can weight our estimates based on sex (informative). The weights are given by the ratio between the population by sex and the number of respondents by sex. Our total population by sex with match exactly:

Men: 1,620x(45,000/1,620)=45,000
Women: 6,380x(55,000/6,380)=55,000
Total: 45,000+55,000=100,000

For our rugby players, we will weight the counts by sex first and sum them up to get the total estimate:

Men: 126x(45,000/1,620)=3,500
Women: 174x(55,000/6,380)=1,500
Total: 3,500+1,500=5,000

Therefore, our weighted estimate for the number of rugby players is 5,000 which is unbiased. In this particular case, non-response was entirely driven by sex (and not at all by rugby-playing) which meant that it was Missing at Random (MAR). If non-response was in fact driven by the variable of interest (NI), then we cannot solve the bias through this kind of weighting. The focus of this thesis is on cases where

missingness is at random but in the simulation study, non-ignorable missing data mechanisms are also explored.

2.4.4 Deriving non-response weights for analysis

We've seen that response weights can help correct non-response bias if the data are not missing completely at random (MCAR). In this section, we explore how to derive these weights.

Let R denote a binary variable indicating whether or not y is observed. Completely ignoring any information regarding selection/response, yields the joint distribution $f(y, R)$ for the observed data (y, R). If we only study the distribution of the observed data, then we simply require the conditional distribution $f(y|R)$. However, our true interest lies in the marginal distribution of y in order to draw inferences about the total population not just those who respond.

If we undertake a perfect census, then, $p(R) = p(R|y) = 1$, and $p(y) = p(y, R) = p(y|R)$. In this case, the joint distribution of y and R is sufficient to infer about the true marginal distribution of y. If, however, only a sample responds to the survey but this response is independent of y (MCAR), then, $p(R|y) = p(R)$, and $p(y) = \frac{p(y,R)}{p(R)} = p(y|R)$. In this case, the conditional distribution of y given they responded is sufficient to infer about the true marginal distribution of y. Finally, if the missingness is informative (MAR or NI), then $p(R) \neq p(R|y)$ and we must estimate $p(R|y)$. We define the missing data mechanism by $p(R|y)$.

The Horvitz-Thompson estimator illustrates that we can estimate the true parameter of interest by simply utilizing weights defined as the inverse probability of selection or responding. For example, if we wished to estimate the true mean, μ, of a variable, y, the Horvitz-Thompson estimator gives

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \pi_i^{-1} y_i}{N}, \tag{2.4.1}$$

where n is the number of subjects who responded, N is the true population total and π_i is the probability of subject i responding.

The simplest approach to weighting is to take advantage of characteristics known about the true population of interest to create weights based on ratios. For example, in order to ensure that respondents are adequately represented by all age and sex groupings, we could use the latest census data to ensure that once response weights are applied, the weighted population from the survey will be equal to the true population for all those age and sex groupings.

A more complicated approach to weighting is to use logistic regression. As we've seen earlier, logistic regression can be used to find the probability of success for a binary variable based on given covariates. In this particular case, the binary variable of interest is R. For this method to be used, auxiliary information about the non-respondents is necessary since it would need to be fed into the regression model.

If the missingness is non-informative, we know that $p(R|y) = p(R)$. As discussed above, if the missingness is informative, we need to estimate $p(R|y)$. If we want to determine the missing data mechanism using logistic regression, we look to solve for γ that best fits

$$p(R|y, X) = \frac{\exp(Y^{*\prime}\gamma)}{1 + \exp(Y^{*\prime}\gamma)}. \tag{2.4.2}$$

We have $R = 1$ for response and $R = 0$ for non-response. We define Y^* by the variable which is incurring item non-response, y, plus any auxiliary variables such

that $Y^* = (1, y, X)$. Once we have $\hat{p}(R|y, X)$, we can simply derive the marginal distribution of y, by

$$f(y) = \prod p(y) = \prod \frac{p(y, R)}{\hat{p}(R|y, X)}. \tag{2.4.3}$$

While this is how we would fit the true missing data mechanism, by its very nature, in practice, we cannot approach this problem in such a simple way. For example, in order to fit the above logistic regression, we require valid values for y when $R = 0$ which, of course, do not exist. So, again, we consider the types of missingness when we define Y^*. If the missingness is MCAR, Y^* is only defined by an intercept and $p(R|y) = p(R)$ which is a scalar value. If we have MAR, then Y^* is defined by an intercept and auxiliary information X. Finally, if the missing data NI, Y^* has to be defined as shown above which, in practice, is impossible to fit. So, the strongest assumption we can make to fit this model is MAR and we define $Y^* = (1, X)$ and find γ that best fits

$$p(R|y, X) = \frac{\exp(Y^{*\prime}\gamma)}{1 + \exp(Y^{*\prime}\gamma)}. \tag{2.4.4}$$

The results of the logistic regression are that we have a predicted probability of response for each observation. If there is no model misspecification, then the sum of the inverse predicted probabilities of response for those subjects who did, in fact, respond, should be equal to the total population:

$$\sum_{i=1}^{n} \frac{R_i}{p(R_i|y_i, X_i)} = n. \tag{2.4.5}$$

2.5 Weighted generalized estimating equations

Let $y_i = (y_{i1}, y_{i2}, y_{i3}, \cdots, y_{iT})$. If there were no missing data, then as shown previously, generalized estimating equations would solve

$$S(\theta) = \sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \mu_i) = 0, \tag{2.5.1}$$

where V_i is a function of $var(y_i)$, $\mu_i = E(y_i)$ and $D_i = \frac{\partial \mu_i}{\partial \theta}$.

However, if there are missing components of y_i, then we risk incurring response bias if that missingness is informative. So, we consider a weighted version of the generalized estimating equations to account for the missing data mechanism. For this thesis, we focus on monotone missing patterns or specifically drop-outs.

2.5.1 Weighted generalized estimating equations using inverse probability weights of Robins et al. (1995)

Here we weight the missingness for each time separately. That is, we construct a diagonal matrix of weights whose diagonal elements represent the inverse probability of responding at each time. For example, the kth diagonal element would represent the inverse probability of responding to time k. Obviously, since we are only focussed on weighting the observed data to represent the observed and unobserved data, we define this TxT weight matrix for subject i as:

$$\Lambda_i = \begin{bmatrix} w_{i1}R_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2}R_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{iT}R_{iT} \end{bmatrix}$$

where w_{it} represents the inverse probability of subject i responding at time t and R_{it} represents whether or not subject i responded at time t. This approach is simple to understand since it implies that for all subjects who responded at time t, their weight means they will represent themselves as well as similar subjects who did not respond at time t.

For example, consider a longitudinal survey where missingness is MCAR and we simply look at attrition. For the first wave (and this is the total sample) we have 100 people. For wave 2, we have 80, for wave 3 we have 50, and for wave 4 we have 20. Then weight matrices would look as follows:

For people who only responded to wave 1:

$$\begin{bmatrix} 100/100 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

For people who only responded to wave 1 and 2:

$$\begin{bmatrix} 100/100 & 0 & 0 & 0 \\ 0 & 100/80 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1.25 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

For people who only responded to wave 1, 2, and 3:

$$\begin{bmatrix} 100/100 & 0 & 0 & 0 \\ 0 & 100/80 & 0 & 0 \\ 0 & 0 & 100/50 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1.25 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

For people who only responded to all 4 waves:

$$\begin{bmatrix} 100/100 & 0 & 0 & 0 \\ 0 & 100/80 & 0 & 0 \\ 0 & 0 & 100/50 & 0 \\ 0 & 0 & 0 & 100/20 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1.25 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

In this way, the number of respondents represented at all times is 100. Our sum of respondents by time can be found as:

$$\sum_{R_{i1}=1} w_{i1} = 100 \times 1 = 100$$

$$\sum_{R_{i2}=1} w_{i2} = 80 \times 1.25 = 100$$

$$\sum_{R_{i3}=1} w_{i3} = 50 \times 2 = 100$$

$$\sum_{R_{i4}=1} w_{i4} = 20 \times 5 = 100$$

If we find these weights using information on y, then they can be used to eliminate response bias where missingness is MAR. We define our weighted generalized estimating equations as

$$S(\theta) = \sum_{i=1}^{K} D_i' V_i^{-1} \Lambda_i (y_i - \mu_i) = 0, \tag{2.5.2}$$

where we only consider observed elements of y_i and Λ_i is a diagonal matrix with diagonal elements $w_{it}R_{it}$ as shown above.

How do we find these weights w_{it}? As we reviewed previously, these weights which represent the inverse probability of subject i responding at time t can be derived through logistic regression. First, let us consider a logistic model to define the probability of subject i not responding at time t. It would be safe to say that the probabilities of response for different times are not independent for the same subject. Let us consider the case where they are dependent on the previous outcomes, that is,

$$1 - p_{it}(\gamma) = p\left(R_{it} = 0 | y_{it}, X_{it}, R_{i1}, \ldots R_{i(t-1)}\right) = \frac{\exp\left(Y_{it}^{*\prime}\gamma\right)}{1 + \exp\left(Y_{it}^{*\prime}\gamma\right)}. \tag{2.5.3}$$

Here Y^* represents a vector containing item non-response, y, plus any auxiliary variables such that $Y^* = (1, y, X)$. Since the probabilities of response are not independent, we cannot use the ordinary maximum likelihood estimation. Instead, we use a modified form of the likelihood function, referred to as the pseudolikelihood function (Besag (1975)). The pseudolikelihood function assumes conditional independence of R_{it} given other values for R_i, instead of marginal independence as required for likelihood functions. We undertake maximum likelihood estimation with the pseudolikelihood function in the same was as we would with the true likelihood function.

We define m_i as the last time the ith subject responded or the time of drop-out for the ith subject. Let m_i be a realized value of the random variable M_i with the probability distribution

$$
\begin{aligned}
&P(M_i = m_i | y_{i1}, y_{i2}, \dots, y_{im_i}, X_{i1}, \dots, X_{im_i}, \gamma) \qquad (2.5.4)\\
&\quad = P(R_{im_i} = 0, R_{i1} = 1, \dots, R_{i(m_i-1)} = 1 \,|y_{i1}, y_{i2}, \dots, y_{im_i}, X_{i1}, \dots, X_{im_i}, \gamma)\\
&\quad = p(R_{i1} = 1 | y_{i1}, y_{i2}, \dots, y_{im_i}, X_{i1}, \dots, X_{im_i}, \gamma)\\
&\quad \times p(R_{i2} = 1 | R_{i1} = 1, y_{i1}, y_{i2}, \dots, y_{im_i}, X_{i1}, \dots, X_{im_i}, \gamma) \times \dots\\
&\quad \times p(R_{i(m_i-1)} = 1 | R_{i1} = 1, \dots R_{i(m_i-2)} = 1, y_{i1}, y_{i2}, \dots, y_{im_i}, X_{i1}, \dots, X_{im_i}, \gamma)\\
&\quad \times p(R_{im_i} = 0 | R_{i1} = 1, \dots R_{i(m_i-1)} = 1, y_{i1}, y_{i2}, \dots, y_{im_i}, X_{i1}, \dots, X_{im_i}, \gamma),
\end{aligned}
$$

and $p(R_{it} = 0 | R_{i(t-1)} = 0, y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) = 1$.

If we take $\mathrm{p}(R_{it} = 0 | R_{i1} = 1, \ldots, R_{i(t-1)} = 1, \gamma, y_i, x_i) = 1 - p_{it}(\gamma)$, then we define our pseudolikelihood function as

$$L(\gamma) = \prod_{i=1}^{k} \prod_{t=2}^{T} \left[p_{it}(\gamma)^{R_{it}} \left(1 - p_{it}(\gamma)\right)^{1-R_{it}} \right]^{I(t \le m_i)} \tag{2.5.5}$$

$$= \prod_{i=1}^{k} \left\{ \prod_{t=2}^{m_i} p_{it}(\gamma) \right\} \left(1 - p_{i(m_i+1)}(\gamma)\right)^{I(m_i<T)}$$

$$= \prod_{i=1}^{k} \left\{ \prod_{t=2}^{m_i} \frac{1}{1 + \exp\left(Y_{it}^{*\prime}\gamma\right)} \right\} \left\{ \frac{\exp\left(Y_{i(m_i+1)}^{*\prime}\gamma\right)}{1 + \exp\left(Y_{i(m_i+1)}^{*\prime}\gamma\right)} \right\}^{I(m_i<T)}.$$

We find the log-pseudolikelihood function as $l(\gamma) = \log\left(L(\gamma)\right)$, so that

$$l(\gamma) = \sum_{i=1}^{k} \left\{ \sum_{t=2}^{m_i} -\log\left(1 + \exp\left(Y_{it}^{*\prime}\gamma\right)\right) \right. \tag{2.5.6}$$

$$\left. + I(m_i < T)\left\{\left(Y_{i(m_i+1)}^{*\prime}\gamma\right) - \log\left(1 + \exp\left(Y_{i(m_i+1)}^{*\prime}\gamma\right)\right)\right\} \right).$$

The score function $S(\gamma) = l'(\gamma) = \frac{\partial l(\gamma)}{\partial \gamma}$ is obtained as

$$S(\gamma) = \sum_{i=1}^{k} \left\{ \sum_{t=2}^{m_i} -Y_{it}^{*} \frac{\exp\left(Y_{it}^{*\prime}\gamma\right)}{1 + \exp\left(Y_{it}^{*\prime}\gamma\right)} \right. \tag{2.5.7}$$

$$\left. + I(m_i < T)\left\{ Y_{i(m_i+1)}^{*} - Y_{i(m_i+1)}^{*} \frac{\exp\left(Y_{i(m_i+1)}^{*\prime}\gamma\right)}{1 + \exp\left(Y_{i(m_i+1)}^{*\prime}\gamma\right)} \right\} \right\}$$

$$= \sum_{i=1}^{k} \left\{ \sum_{t=2}^{m_i} -Y_{it}^{*} p_{it}(\gamma) + I(m_i < T)\left\{ Y_{i(m_i+1)}^{*} - Y_{i(m_i+1)}^{*} p_{i(m_i+1)}(\gamma) \right\} \right\}$$

The estimate of γ is obtained by solving $S(\gamma) = 0$ with respect to γ. However, as noted previously, a closed form solution for this often does not exist. So, we require

the information matrix in order to undertake the Newton-Raphson algorithm to find an estimate of γ. We have

$$-I(\gamma) = S'(\gamma) = \sum_{i=1}^{k}\left\{\sum_{t=2}^{m_i} -Y_{it}^{*}\frac{\partial p_{it}(\gamma)}{\partial \gamma} + I(m_i < T)\left\{-Y_{i(m_i+1)}^{*}\frac{\partial p_{i(m_i+1)}(\gamma)}{\partial \gamma}\right\}\right\}. \quad (2.5.8)$$

After some algebra, we have

$$I(\gamma) = \sum_{i=1}^{k}\sum_{t=2}^{\min(m_i+1,T)} p_{it}(\gamma)(1-p_{it}(\gamma))\, Y_{it}^{*}Y_{it}^{*\prime}. \quad (2.5.9)$$

The Newton-Raphson algorithm provides the iterative equation for γ as

$$\gamma^{M+1} = \gamma^{M} + I^{-1}(\gamma^{M})S(\gamma^{M}), \quad (2.5.10)$$

for M=1, 2, ... Once the drop-out model has been determined, we create our weight matrices:

$$\Lambda_i = \begin{bmatrix} w_{i1}R_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2}R_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{iT}R_{iT} \end{bmatrix}$$

In order to define our weights, w_{it}, we calculate the probability of responding to every time prior to t and t as well. To do this, we find

$$\pi_{it} = 1/w_{it} = \prod_{j=1}^{t} p(R_{ij}=1|y_{ij},X_{ij}) = \prod_{j=1}^{t}\{1-p(R_{ij}=0|y_{ij},X_{ij})\} \quad (2.5.11)$$

$$= \prod_{j=1}^{t}\frac{1}{1+\exp(Y_{it}^{*\prime}\gamma)}.$$

So, our weight matrices become:

$$\Lambda_i = \begin{bmatrix} R_{i1}/\pi_{i1} & 0 & \cdots & 0 \\ 0 & R_{i2}/\pi_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{iT}/\pi_{iT} \end{bmatrix}$$

Then, we can solve the weighted generalized estimating equations for θ as

$$S(\theta) = \sum_{i=1}^{K} D_i' V_i^{-1} \Lambda_i (y_i - \mu_i) = 0. \tag{2.5.12}$$

Note that since Λ_i is not a function of our main parameter of interest, θ, finding $I(\theta)$ does not rely on the weight matrix at all. We can find $I(\theta)$ in the same manner as we would in the ordinary GEE method such that $I(\theta) = -S'(\theta)$:

$$I(\theta) = \sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \theta}' V_i^{-1} \Lambda_i \frac{\partial \mu_i}{\partial \theta}. \tag{2.5.13}$$

From here, we can use the Newton-Raphson algorithm and Fisher scoring to find $\hat{\theta}$ that best fits our model of interest.

2.5.2 Weighted generalized estimating equations using inverse probability weights of Fitzmaurice et al. (1994)

While Robins et al.'s method is largely employed and sourced, another approach to this weighting method was introduced by Fitzmaurice et al. In this method, instead of creating a weight matrix with a different set of weights for each time, we create a single weight for each subject. This weight represents the response pattern. That is, if a subject followed a certain response pattern M, their weight would represent themselves as well as similar other subjects who did not follow the same pattern. If the missingness is monotone, then the number of possible response patterns is fairly limited. In fact, if we assume that everyone responds to the first observation,

then there are only T possible patterns where T is the total number of waves in the study.

We define a variable denoting the last observed time as m_i. We know that $1 \leq m_i \leq T$. Our weight, w_i, is defined by the inverse of the probability of dropping out after time m_i. So we define our weight as the inverse of

$$v_{im} = p(M_i = m_i|y). \tag{2.5.14}$$

Similar to the approach undertaken by Robins et al., we define the probability of not responding at time t using a logistic model

$$1 - p_{it}(\gamma) = p\big(R_{it} = 0|y_{it}, X_{it}, R_{i1}, \dots R_{i(t-1)}\big) = \frac{\exp\big(Y_{it}^{*\prime}\gamma\big)}{1 + \exp(Y_{it}^{*\prime}\gamma)}. \tag{2.5.15}$$

where Y^* is a vector containing our variable of interest (which is incurring item non-response), y, plus any auxiliary variables such that $Y^* = (1, y, X)$. As is the case in the Robins et al. approach, the pseudolikelihood function is again defined as

$$L(\gamma) = \prod_{i=1}^{k}\left\{\prod_{t=2}^{m_i}\frac{1}{1 + \exp\,(Y_{it}^{*\prime}\gamma)}\right\} \times \left\{\frac{\exp\,(Y_{i(m_i+1)}^{*\prime}\gamma)}{1 + \exp\,(Y_{i(m_i+1)}^{*\prime}\gamma)}\right\}^{I(m_i<T)}. \tag{2.5.16}$$

We find the log-pseudolikelihood function as $l(\gamma) = \log\big(L(\gamma)\big)$, so that

$$l(\gamma) = \sum_{i=1}^{k}\left\{\sum_{t=2}^{m_i} - \log\big(1 + \exp\big(Y_{it}^{*\prime}\gamma\big)\big)\right. \tag{2.5.17}$$

$$\left. + I(m_i < T)\big\{\big(Y_{i(m_i+1)}^{*\prime}\gamma\big) - \log\,\big(1 + \exp\big(Y_{i(m_i+1)}^{*\prime}\gamma\big)\big)\big\}\right\}.$$

The score function $S(\gamma) = l'(\gamma) = \frac{l(\gamma)}{\partial\gamma}$ is obtained as

$$S(\gamma) = \sum_{i=1}^{k}\left\{\sum_{t=2}^{m_i} -Y_{it}^{*}p(\gamma)_{it} + I(m_i < T)\{Y_{i(m_i+1)}^{*} - Y_{i(m_i+1)}^{*}p(\gamma)_{i(m_i+1)}\}\right\}. \quad (2.5.18)$$

The estimate of γ is obtained by solving $S(\gamma) = 0$ with respect to γ. However, as noted previously, a closed form solution for this often does not exist. So, we require the information matrix in order to undertake the Newton-Raphson algorithm to find an estimate of γ. We have

$$I(\gamma) = \sum_{i=1}^{k}\sum_{t=2}^{\min(m_i+1,T)} p_{it}(\gamma)(1-p_{it}(\gamma))\,Y_{it}^{*}Y_{it}^{*\prime}. \quad (2.5.19)$$

The Newton-Raphson iterative equation for γ takes the form

$$\gamma^{M+1} = \gamma^{M} + I^{-1}(\gamma^{M})S(\gamma^{M}), \quad (2.5.20)$$

for M=1, 2, ...

From this point, the weighting method of Fitzmaurice et al. Differs from that of Robins et al. Instead of being interested in the inverse probability of responding at each time where a response existed, we simply want the probability of the response pattern denoted by v_{im}, which is modeled as

$$\begin{aligned} v_{im} = p(M_i = m_i|Y_i^{*}) &= p\big(R_{i(m_i+1)} = 0, R_{i1} = 1, \ldots, R_{im_i} = 1\,|Y_i^{*}\big) \quad (2.5.21)\\ &= p(R_{i1} = 1|Y_i^{*}) \times p(R_{i2} = 1|R_{i1} = 1, Y_i^{*}) \times \ldots \\ &\times p\big(R_{im_i} = 1\big|R_{i1} = 1, \ldots R_{i(m_i-1)} = 1, Y_i^{*}\big) \\ &\times p\big(R_{i(m_i+1)} = 0\big|R_{i1} = 1, \ldots R_{im_i} = 1, Y_i^{*}\big). \end{aligned}$$

Note that since we are considering drop-outs, $p(R_{it} = 0 | R_{i(t-1)} = 0, Y_i^*) = 1$ and we assume that we have a valid observation for the first wave. Therefore,

$$\begin{aligned} v_{im} &= p(M_i = m_i | Y_i^*) \\ &= \prod_{t=2}^{m_i} p(R_{it} = 1 | R_{i1} = 1, \dots, R_{i(t-1)} = 1, Y_i^*) \\ &\quad \times p(R_{i(m_i+1)} = 0 | R_{i1} = 1, \dots, R_{im_i} = 1, Y_i^*)^{I(m_i<T)}. \end{aligned} \tag{2.5.22}$$

If we take our probabilities of being missing at time t as $p(R_{it} = 0 | R_{i1} = 1, \dots, R_{i(t-1)} = 1, Y_i^*) = \pi_{it}$, then our fitted probabilities are given by

$$\hat{\pi}_{it} = \frac{\exp(Y_{it}^{*\prime} \hat{\gamma})}{1 + \exp(Y_{it}^{*\prime} \hat{\gamma})}. \tag{2.5.23}$$

Then using our fitted logistic regression model to find $\hat{\gamma}$, we estimate v_{im} by

$$\hat{v}_{im} = \hat{p}(M_i = m_i | Y_i^*) = \left\{ \prod_{t=2}^{m_i} (1 - \hat{\pi}_{it}) \right\} \times \{\hat{\pi}_{i(m_i+1)}\}^{I(m_i<T)}. \tag{2.5.24}$$

As $w_i = v_{im}$, our weighted generalized estimating equations can take the form

$$S(\theta) = \sum_{i=1}^{K} \frac{1}{v_{im}} D_i' V_i^{-1} (y_i - \mu_i) = 0, \tag{2.5.25}$$

where we only consider the observed values for y_i. As is often the case, the information matrix may need to be estimated in order to find parameter estimates using Fisher scoring. We can find $I(\theta)$ in the same manner as we would in the ordinary GEE method such that $-I(\theta) = S'(\theta)$:

$$I(\theta) = \sum_{i=1}^{K} \frac{1}{v_{im}} \frac{\partial \mu_i}{\partial \theta}' V_i^{-1} \frac{\partial \mu_i}{\partial \theta}. \tag{2.5.26}$$

From here, we can use Newton-Raphson to find θ that best fits our model of interest.

This approach by Fitzmaurice et al. is employed for the duration of this thesis in both the simulation study as well as the data analysis. Either the Fitzmaurice et al. method or the Robins et al. method would yield unbiased results where the missingness is MCAR or MAR.

Chapter 3

Simulation study

In this chapter, we explore the empirical properties of the weighted generalized estimating equations through simulations. We ran a series of simulations under different correlation structures and missing data mechanisms, and study empirical biases, mean squared errors and coverage probabilities of the estimators.

3.1 Model for simulation study

In our simulations, we consider that the response, y, is a binary longitudinal variable. While generating random binary values is simple enough, it becomes more complicated as we introduce covariates and a correlation structure between binary responses for the same subject (due to the longitudinal nature of the data).

The covariate, x, could be discrete or continuous or multivariate. For the purposes of this simulation study, x is considered to be a binary variable representing two treatment groups. We define a covariate vector X as (1,x) where 1 represents the intercept. The mean response of y is assumed to be related to X by the logistic regression $p(y) = \frac{\exp(X'\beta)}{1+\exp(X'\beta)}$ where β represents the treatment effect.

For a given value of x, we generate y from a Bernoulli distribution with probability of success $p(y)$. For each combination of simulation runs, we can generate binary observations from the given model. But, in a longitudinal setting, this becomes more complicated as we expand y to become a longitudinal variable with T measurements. If we assume that no relationship exists between the responses

from a given subject, then we would simply use the method above to generate y_t for t=1,...T. We also include a variable to indicate the effect of time or cycle, t. Here we redefine our covariate X as (1, x, t) where 1 represents the intercept, and x is the original binary covariate. In this thesis, the method developed by Bahadur (1961) is employed to generate longitudinal data, as described in the next section.

3.1.1 Bahadur model for generation of correlated data

In order to undertake the simulation study, data has to be generated with the possibility of within-subject observations having association. A method that is commonly used to undertake this comes from Bahadur (1961). He demonstrated that multivariate binary distributions could be expressed jointly. Specifically, if all correlation coefficients with order higher than 2 are ignored, Bahadur's model reduces to

$$f(Y_1, \dots, Y_T) = \left\{\prod_{t=1}^{T} \mu_t^{Y_t}(1-\mu_t)^{1-Y_t}\right\}\left\{1 + \sum_{1\le t<j\le T} \rho_{tj}\tilde{Y}_t\tilde{Y}_j\right\}, \tag{3.1.1}$$

where $\tilde{Y}_t = (Y_t - \mu_t)/\sqrt{\mu_t(1-\mu_t)}$ is standardized Y_t, μ_t is the expected value of Y_t (or probability of success at time t), and ρ_{tj} is the correlation coefficient of Y_t and Y_j. This notation is quite useful for generating correlated binary data since we can define the correlation coefficients between any observations from the same subject.

In order to actually generate the data, we need to transform this joint distribution into a conditional one where we need to determine $p(Y_t = 1|Y_1, \dots, Y_{t-1})$. To do this, we simply think of the joint distribution as a product of conditional probabilities

$$\begin{aligned} p(Y_1, \dots, Y_T) &= p(Y_T = 1|Y_1, \dots, Y_{T-1})p(Y_1, \dots, Y_{T-1}) \\ &= p(Y_T = 1|Y_1, \dots, Y_{T-1})p(Y_{T-1} = 1|Y_1, \dots, Y_{T-2})p(Y_1, \dots, Y_{T-2}) \\ &= p(Y_T = 1|Y_1, \dots, Y_{T-1})p(Y_{T-1} = 1|Y_1, \dots, Y_{T-2}) \dots p(Y_2 = 1|Y_1)p(Y_1). \end{aligned} \tag{3.1.2}$$

This implies that we can find $p(Y_1)$ independent of any future observations. To do this, we simply find the binomial probability of success for Y_1 which is equal to μ_1. For future observations t=2,...,T, we determine the conditional probability based on

$$p(Y_1, \dots, Y_t) = p(Y_t = 1|Y_1, \dots, Y_{t-1})p(Y_1, \dots, Y_{t-1}), \tag{3.1.3}$$

where

$$p(Y_t = 1|Y_1, \dots, Y_{t-1}) = \frac{p(Y_1, \dots, Y_t)}{p(Y_1, \dots, Y_{t-1})}. \tag{3.1.4}$$

To generate multivariate binary data, first we calculate

$$p(Y_1) = \mu_1 = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}, \tag{3.1.5}$$

for given values of β and X. We then generate Y_1 as a random Bernoulli variable with probability μ_1. Subsequent generations of Y_t rely on previously generated values for $Y_1, \dots, Y_{t-1}$. So for $t > 1$, we generate the data by assuming the marginal distribution for Y_t

$$p(Y_t) = \mu_t = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}. \tag{3.1.6}$$

Now, for t=2, we have

$$p(Y_2 = 1|Y_1) = \frac{p(Y_1, Y_2)}{p(Y_1)} = \frac{\{\mu_1^{Y_1}(1-\mu_1)^{1-Y_1}\mu_2^{Y_2}(1-\mu_2)^{1-Y_2}\}\{1 + \rho_{12}\tilde{Y}_1\tilde{Y}_2\}}{\mu_1^{Y_1}(1-\mu_1)^{1-Y_1}} \tag{3.1.7}$$

$$= \{\mu_2^{Y_2}(1-\mu_2)^{1-Y_2}\}\{1 + \rho_{12}\tilde{Y}_1\tilde{Y}_2\} = \mu_2\{1 + \rho_{12}\tilde{Y}_1\tilde{Y}_2\}.$$

We can then generate Y_2 as a random Bernoulli variable with probability $p(Y_2 = 1|Y_1)$.

In general, for a given value of t, we have

$$p(Y_t = 1|Y_1, \dots, Y_{t-1}) = \frac{p(Y_1, \dots, Y_t)}{p(Y_1, \dots, Y_{t-1})} \tag{3.1.8}$$

$$= \frac{\{\prod_{k=1}^{t} \mu_k^{Y_k}(1-\mu_k)^{1-Y_k}\}\{1 + \sum_{1\le k<j\le t} \rho_{tj}\tilde{Y}_k\tilde{Y}_j\}}{\{\prod_{k=1}^{t-1} \mu_k^{Y_k}(1-\mu_k)^{1-Y_k}\}\{1 + \sum_{1\le k<j\le t-1} \rho_{tj}\tilde{Y}_k\tilde{Y}_j\}}$$

$$= \frac{\mu_t^{Y_t}(1-\mu_t)^{1-Y_t}\{1 + \sum_{1\le k<j\le t} \rho_{tj}\tilde{Y}_k\tilde{Y}_j\}}{\{1 + \sum_{1\le k<j\le t-1} \rho_{tj}\tilde{Y}_k\tilde{Y}_j\}} = \frac{\mu_t\{1 + \sum_{1\le k<j\le t} \rho_{tj}\tilde{Y}_k\tilde{Y}_j\}}{\{1 + \sum_{1\le k<j\le t-1} \rho_{tj}\tilde{Y}_k\tilde{Y}_j\}}.$$

So we can generate Y_t as a random Bernoulli variable with probability $p(Y_t = 1|Y_1, \dots, Y_{t-1})$. Note that the correlation coefficients ρ_{tj} are given for the generation as well. In our simulation study, we use different correlation structures as shown in the next section. We use a binary value of x (not time-varying) with a probability of success of 0.2. Values for β vary throughout the study.

3.1.2 Correlation structures used to generate data

Creating correlation structures for the purposes of simulation is fairly straight-forward. In this study we consider 3 types of correlation matrices: Independent, exchangeable, and serial. For independent correlation, we simply create a TxT identity matrix where T represents the total number of waves in a longitudinal study. For exchangeable correlation, this requires a correlation parameter α. We create another TxT identity matrix but replace all non-diagonal elements with α. For serial correlation, we also require a correlation parameter α and begin with a TxT matrix. The jkth element of the matrix is denoted by: $R_{jk} = \alpha^{|j-k|}$ for all j,k.

Generating these correlation structures is imperative for the Bahadur correlated data generation. Fitting these matrices afterwards is described in the theory portion of this thesis.

3.1.3 Generation of missingness or drop-out indicators

Since the principal purpose of this simulation study is to examine missing data, we must generate missingness once our data sets have been created. This study focuses on drop-outs where we have complete covariate information. So, for each subject, we must generate a drop-out time. To do this, we recall from the theory section (2.5.2) that we have a drop-out mechanism or model where the probability of not responding at time t is given by

$$1 - p_{it}(\gamma) = p(R_{it} = 0|y_{it}, X_{it}) = \frac{\exp(Y_{it}^{*\prime}\gamma)}{1 + \exp(Y_{it}^{*\prime}\gamma)}, \tag{3.1.9}$$

where we define Y^* by the variable which is incurring non-response, y, plus any auxiliary variables. In this particular study, we consider $Y_{it}{}^* = (1, y_{i(t-1)}, y_{it})$ and we define our probability of missing as

$$p(R_{it} = 0|y_{it}, X_{it}) = \frac{\exp(\gamma_0 + \gamma_1 y_{i(t-1)} + \gamma_2 y_{it})}{1 + \exp(\gamma_0 + \gamma_1 y_{i(t-1)} + \gamma_2 y_{it})}. \tag{3.1.10}$$

This is practical since we can easily define missingness as MCAR, MAR or NI. If missingness is MCAR, then $\gamma_1 = \gamma_2 = 0$. If missingness is MAR, then $\gamma_2 = 0$ but $\gamma_1 \neq 0$. Finally, if missingness is NI, then it depends on the current value of y, which implies that $\gamma_2 \neq 0$.

To determine the drop-out time, we refer to the probability of dropping out at each time given by

$$v_{im} = p(M_i = m_i | Y_i^*)$$
$$= \prod_{t=2}^{m_i} p\left(R_{it} = 1 | R_{i1} = 1, \dots, R_{i(t-1)} = 1, Y_i^*\right) \tag{3.1.11}$$
$$\times p\left(R_{i(m_i+1)} = 0 | R_{i1} = 1, \dots, R_{im_i} = 1, Y_i^*\right)^{I(m_i<T)}.$$

So, for each $1 \leq m \leq T$, we calculate

$$P(M_i = m_i | y_{i1}, y_{i2}, \dots, y_{im}, \gamma) \tag{3.1.12}$$
$$= \left\{\prod_{t=2}^{m} \frac{1}{1 + \exp\left(\gamma_0 + \gamma_1 y_{i(t-1)} + \gamma_2 y_{it}\right)}\right\} \times \left\{\frac{\exp\left(\gamma_0 + \gamma_1 y_{im} + \gamma_2 y_{i(m+1)}\right)}{1 + \exp\left(\gamma_0 + \gamma_1 y_{im} + \gamma_2 y_{i(m+1)}\right)}\right\}^{I(m<T)}.$$

We can then generate the drop out times for each subject i, as a multinomial distribution (for a count of 1) with probabilities given above. We then remove all observations after the point of drop-out.

3.1.4 Fitting generalized estimating equations

As illustrated in the theory chapter of this thesis (2.3), we estimate generalized estimating equations using the Fisher scoring technique. For the purposes of this thesis, we define $x_{it} = (1, x_i, t)$ as the covariate matrix where we find the probability of our variable of interest using the logit link function

$$p(y_{it}) = \mu_{it} = \frac{\exp\left(x_{it}'\beta\right)}{1 + \exp\left(x_{it}'\beta\right)}. \tag{3.1.13}$$

To find $\hat{\beta}$ that best fits the above, we employ generalized estimating equations and look to solve

$$S(\beta) = \sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \mu_i) = 0, \tag{3.1.14}$$

where V_i is a $T_i x T_i$ matrix representing the working covariance matrix of y_i. V_i can be found as

$$V_i = \varphi A_i^{1/2} R(\alpha) A_i^{1/2}, \tag{3.1.15}$$

where φ is the dispersion parameter, $R(\alpha)$ is the working correlation matrix, and A_i is a $T_i x T_i$ matrix with diagonal elements $v_{it} = \widehat{var}(y_{it}) = p_{it}(1 - p_{it})$.
In addition, D_i is a $T_i$xp matrix (p represents the number of parameters or the length of β) and is found as

$$D_i = \frac{\partial \mu_i}{\partial \beta} = \left\{\frac{\partial \mu_{it}}{\partial p_{it}} \cdot \frac{\partial p_{it}}{\partial x_{it}}\right\} = p_{it}(1 - p_{it}) x_{it}. \tag{3.1.16}$$

Finally, $\mu_{it} = E(y_{it}) = p_{it}$ where $logit(p_{it}) = \beta_0 + \beta_1 x_i + \beta_2(t-1) = x_{it}'\beta$. We define the expected Information matrix as

$$I(\beta) = \sum_{i=1}^{K} D_i' V_i^{-1} D_i. \tag{3.1.17}$$

We use the Newton-Raphson iterative algorithm as described earlier to solve the equation $S(\beta) = 0$ with respect to β.

It should be noted that all unweighted generalized estimating equations use all observed data (as opposed to complete-case analysis which would discard subjects without an observation for each time).

For weighted generalized estimating equations, prior to solving $S(\beta) = 0$, we must derive the response weights given by the probability of the subject dropping out when they did: $v_{im} = p(M_i = m_i | Y_i^*)$ where $Y_i^* = (1, y_{i(t-1)}, y_{it})$. This is solved as shown in the theory section (2.5.2), where we find γ that maximizes the pseudolikelihood function

$$L(\gamma) = \left\{\prod_{t=2}^{m} \frac{1}{1+\exp(\gamma_0 + \gamma_1 y_{i(t-1)} + \gamma_2 y_{it})}\right\} \times \left\{\frac{\exp(\gamma_0 + \gamma_1 y_{im} + \gamma_2 y_{i(m+1)})}{1+\exp(\gamma_0 + \gamma_1 y_{im} + \gamma_2 y_{i(m+1)})}\right\}^{I(m<T)}. \quad (3.1.18)$$

We let $p(\gamma)_{it} = \frac{\exp(\gamma_0+\gamma_1 y_{im}+\gamma_2 y_{i(m+1)})}{1+\exp(\gamma_0+\gamma_1 y_{im}+\gamma_2 y_{i(m+1)})}$ and find the estimate for γ by taking the Score and Information matrix as illustrated in the theory section (2.5.2). For the ML estimator of γ, we have

$$S(\gamma) = \sum_{i=1}^{k}\left\{\sum_{t=2}^{m_i} -Y_{it}^{*} p_{it}(\gamma) + I(m_i < T)\{Y_{i(m_i+1)}^{*} - Y_{i(m_i+1)}^{*} p_{i(m_i+1)}(\gamma)\}\right\}, \quad (3.1.19)$$

and

$$I(\gamma) = \sum_{i=1}^{k} \sum_{t=2}^{\min(m_i+1,T)} p_{it}(\gamma)(1-p_{it}(\gamma)) Y_{it}^{*} Y_{it}^{*\prime}. \quad (3.1.20)$$

Using the ML estimate $\hat{\gamma}$, we find the predicted probabilities of responding for each time as

$$\hat{\pi}_{it} = \frac{\exp(Y_{it}^{*\prime}\hat{\gamma})}{1+\exp(Y_{it}^{*\prime}\hat{\gamma})}. \quad (3.1.21)$$

Then using the ML estimate $\hat{\gamma}$, we estimate our probability of dropping out at time m, v_{im}, by

$$\hat{v}_{im} = \hat{p}(M_i = m_i | Y_i^{*}) = \left\{\prod_{t=2}^{m_i}(1-\hat{\pi}_{it})\right\} \times \{\hat{\pi}_{i(m_i+1)}\}^{I(m_i<T)}. \quad (3.1.22)$$

Now, we can solve the weighted generalized estimating equations

$$S(\beta) = \sum_{i=1}^{K} \frac{1}{\hat{v}_{im}} D_i' V_i^{-1}(y_i - \mu_i) = 0. \quad (3.1.23)$$

with respect to β. We undertake this using the same methods used for the regular generalized estimating equations where the expected Information matrix is given by

$$I(\beta) = \sum_{i=1}^{K} \frac{1}{\hat{v}_{im}} D_i' V_i^{-1} D_i. \tag{3.1.24}$$

3.1.5 Diagnostic methods used in simulation study

Several diagnostics are utilized in this thesis to measure the data quality of different analytical methods. For the simulation study, we focus on 3 principal quality measures:

Bias

Bias is perhaps the most important diagnostic we consider in the simulation study. It can indicate the presence of non-response bias. The weighted generalized estimating equations should minimize / eliminate non-response bias if the missing data mechanism is MAR.

We find bias for a given estimate $\hat{\theta}$ by

$$bias(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta, \tag{3.1.25}$$

which is approximated by

$$bias(\hat{\theta}) \approx \frac{\sum_{s=1}^{S} \hat{\theta}_s}{S} - \theta, \tag{3.1.26}$$

where $\hat{\theta}_s$ is the estimate for θ for the sth simulated dataset. We find percentage relative bias by $\frac{bias(\hat{\theta})}{\theta} \times 100$.

Mean Squared Error

Mean Squared Error (MSE) is an overall error measure indicating the combined effects of bias and variance. This value is found by

$$MSE(\hat{\theta}) = bias(\hat{\theta})^2 + var(\hat{\theta}). \tag{3.1.27}$$

Coverage Probability

We find coverage probabilities for 95% confidence intervals, $\hat{\theta} \pm 1.96\, S.E.(\hat{\theta})$, where $S.E.(\hat{\theta})$ is obtained by taking the square root of the variance of the estimates $\hat{\theta}$. The coverage probability is obtained from

$$\text{CoverageProbability}(\hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S} I\left[\left|\hat{\theta}_s - \theta\right| < 1.96\, S.E.(\hat{\theta})\right], \tag{3.1.28}$$

where I is an indicator function.

3.2 Results for complete data

Before generating missingness, we will consider some simulations on complete data sets. For these simulations, we generate the data using $p(x) = 0.2$ and $\beta = (-1,1,0.2)'$ where the covariate X consists of the intercept, x, and t indicating the interval or wave in question. For each simulation, we generate 5000 data sets with 500 subjects and 5 observations per subject.

We consider the simplest case where there is no relationship between the observations from the same subject – that is, we generate the data using an independence correlation structure between the responses. This data is fit using 3 different approaches: Generalized estimating equations (GEE) using independent correlation structure, using exchangeable correlation structure, and using serial correlation structure.

Table 3.2.1 Simulation results: Complete data

	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
GEE-Independent	0.192	0.403	0.150	0.011	0.011	0.001	0.949	0.950	0.950
GEE-Exchangeable	0.192	0.403	0.150	0.011	0.011	0.001	0.949	0.950	0.950
GEE-Serial	0.191	0.402	0.148	0.011	0.011	0.001	0.949	0.950	0.949

Source: Simulation study – McLeish, S.

We see, not surprisingly that the level of bias here is trivial since generalized estimating equations are unbiased estimators of β for complete data. In conclusion, we see that generalized estimating equations (with different working correlation structures) provide unbiased estimates for our true parameters with minimal variance when there is no missingness and the true correlation structure is independent.

3.3 Results for data with drop-outs completely at random (MCAR)

We will now run some simulations on incomplete data sets. Here the missing data mechanism is MCAR which implies that while there are missing values of y, this phenomenon is completely at random and entirely independent of the covariates and the true value of y itself. Here we let $\gamma' = (-2,0,0)$.

For these simulations, we generate the data using $p(x) = 0.2$ and $\beta = (-1,1,0.2)'$ where the covariate X consists of the intercept, x, and t indicating the interval or wave in question. For each simulation, we generate 5000 data sets with 500 subjects and a maximum of 5 observations per subject.

For this simulation study, we assume a serial correlation between observations from the same subject. The correlation parameter is set as $\alpha = 0.5$. This data is fit using 6 different approaches: Weighted generalized estimating equations (WGEE) using independent correlation structure, using exchangeable correlation structure, and using serial correlation structure, and unweighted generalized estimating equations (GEE) using independent correlation structure, using exchangeable correlation structure, and using serial correlation structure.

Table 3.3.1 Simulation results: Data with MCAR

	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	-0.242	2.606	-1.884	0.026	0.055	0.002	0.949	0.949	0.952
WGEE-Exchangeable	-0.273	2.473	-1.955	0.025	0.060	0.002	0.946	0.948	0.950
WGEE-Serial	-0.076	2.756	-1.850	0.027	0.056	0.003	0.950	0.949	0.949
GEE-Independent	-0.574	2.449	-2.872	0.016	0.031	0.001	0.948	0.947	0.950
GEE-Exchangeable	-0.578	2.203	-2.922	0.016	0.030	0.001	0.946	0.946	0.953
GEE-Serial	-0.341	2.921	-2.955	0.016	0.029	0.001	0.949	0.948	0.951

Source: Simulation study – McLeish, S.

We can see that since the missingness occurred completely at random, the weighted GEEs and unweighted GEEs performed similarly with negligible bias. On the other hand, the unweighted models have slightly less variance than the weighted models leading to lesser values for mean squared error. This is due to the fact that additional estimation is required for the weighted models since the drop-out model must also be found. So we see that when data is missing completely at random, unweighted generalized estimating equations are sufficient to perform unbiased analysis with minimal variance.

3.4 Results for data with drop-outs at random (MAR)

Here the missing data mechanism is MAR which implies that the probability of y being missing is related to the covariate values or other auxiliary information such as previous values of y but remains entirely independent of the current value of y itself. Here we take on different values for the drop-out mechanism. In order for missingness to be MAR, the 2nd element of γ (denoting the effect of the previous value of y on the drop-out probability) must be non-zero while the 3rd element of γ denoting the effect of the current value of y on the drop-out probability) must still be zero.

For these simulations, we generate the data using $p(x) = 0.2$ and use varying values of β where the covariate X consists of the intercept, x, and t indicating the interval or wave in question. For each simulation, we generate 5000 data sets with 500 subjects and a maximum of 5 observations per subject.

For this simulation study, we assume a serial correlation between observations from the same subject. The correlation parameter is set as $\alpha = 0.5$. This data is fit using 6 different approaches: Weighted generalized estimating equations (WGEE) using independent correlation structure, using exchangeable correlation structure, and using serial correlation structure, and unweighted generalized estimating equations (GEE) using independent correlation structure, using exchangeable correlation structure, and using serial correlation structure.

3.4.1 Results for principal model of study with drop-outs at random (MAR)

To begin with, we consider $\beta' = (-1,1,.2)$ and $\gamma' = (-2,2,0)$. This model will be our principal model of study within these simulations. The following sections will explore deviations from this model.

Table 3.4.1 Simulation results: Data with MAR

	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	-0.030	0.622	-1.313	0.041	0.071	0.004	0.947	0.948	0.947
WGEE-Exchangeable	0.096	0.619	-0.761	0.040	0.064	0.004	0.948	0.949	0.947
WGEE-Serial	0.235	0.965	-0.845	0.039	0.064	0.004	0.950	0.948	0.949
GEE-Independent	-5.664	-0.655	-59.568	0.021	0.033	0.016	0.934	0.949	0.232
GEE-Exchangeable	9.115	2.775	14.765	0.025	0.035	0.003	0.887	0.946	0.896
GEE-Serial	-1.167	2.667	-12.832	0.016	0.033	0.003	0.948	0.947	0.912

Source: Simulation study – McLeish, S.

We observe, as expected based on the theory outlined in section 2.5, that the weighted generalized estimating equations outperform the unweighted GEEs with respect to bias and coverage probability where drop-outs are at random (MAR). On the other hand, when the missingness is not properly dealt with (i.e. unweighted GEE), the correlation structure plays a role in the accuracy of results. The data being generated with serial correlation has less biased estimates with unweighted GEE with serial working correlation than exchangeable and independent. The latter being the weakest model by far.

The elimination of bias comes at the price of efficiency with higher variance for the weighted models than the unweighted ones. However, this efficiency is not enough to bring the mean squared error for the independent unweighted GEE models lower than that for the weighted GEEs. We see from our coverage probability estimates that the weighted GEE models provide accurate confidence intervals for their estimates while the coverage is significantly weaker for the unweighted models.

3.4.2 Results under different models with drop-outs at random (MAR)

Here we compare the original results above against a stronger model association between the variable of interest and the covariates, x, and t. Recall for the original model we had $\beta' = (-1,1,.2)$ and $\gamma' = (-2,2,0)$. We will compare this against model 2 which has $\beta' = (-1,2,.3)$ and $\gamma' = (-2,2,0)$.

We are interested in whether or not the change in the model association will impact our results.

Table 3.4.2 Simulation results: Data with MAR for different models of study

β'=(-1,1,.2)	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	-0.030	0.622	-1.313	0.041	0.071	0.004	0.947	0.948	0.947
WGEE-Exchangeable	0.096	0.619	-0.761	0.040	0.064	0.004	0.948	0.949	0.947
WGEE-Serial	0.235	0.965	-0.845	0.039	0.064	0.004	0.950	0.948	0.949
GEE-Independent	-5.664	-0.655	-59.568	0.021	0.033	0.016	0.934	0.949	0.232
GEE-Exchangeable	9.115	2.775	14.765	0.025	0.035	0.003	0.887	0.946	0.896
GEE-Serial	-1.167	2.667	-12.832	0.016	0.033	0.003	0.948	0.947	0.912
β'=(-1,2,.3)	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	1.077	2.909	0.626	0.040	0.116	0.005	0.951	0.942	0.953
WGEE-Exchangeable	1.084	1.773	0.985	0.041	0.101	0.005	0.951	0.945	0.949
WGEE-Serial	1.209	2.825	0.898	0.040	0.107	0.005	0.950	0.944	0.948
GEE-Independent	-6.129	-0.327	-40.550	0.023	0.060	0.017	0.928	0.953	0.259
GEE-Exchangeable	10.793	2.319	15.382	0.030	0.061	0.005	0.870	0.942	0.845
GEE-Serial	1.198	2.828	-0.859	0.018	0.060	0.003	0.947	0.942	0.953

Source: Simulation study – McLeish, S.

We can observe that there is minimal impact associated with the change in parameter values. The weighted generalized estimating equations continue to offer more accurate parameter estimates than the unweighted GEEs. This accuracy comes at the price of efficiency with higher variance for the weighted models. While there are some slight differences between the results from the 2 models for bias and

variance (typically lower for the 2nd model), overall these discrepancies are negligible.

3.4.3 Results under different drop-out mechanisms with drop-outs at random (MAR)

Now we are interested in whether the original results will be affected if we apply a stronger or weaker MAR drop-out model. Recall for the original model we had $\beta' = (-1,1,.2)$ and $\gamma' = (-2,2,0)$. We will continue to use $\beta' = (-1,1,.2)$ but compare our original against models with $\gamma' = (-2,3,0)$ (Stronger association between drop-out probability and previous value of y) and $\gamma' = (-2,1,0)$ (Weaker association between drop-out probability and previous value of y) respectively.

We are interested in whether or not the change in the drop-out model will impact our results.

Table 3.4.3 Simulation results: Data with MAR for different drop-out mechanisms

γ'=(-2,2,0)	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	-0.030	0.622	-1.313	0.041	0.071	0.004	0.947	0.948	0.947
WGEE-Exchangeable	0.096	0.619	-0.761	0.040	0.064	0.004	0.948	0.949	0.947
WGEE-Serial	0.235	0.965	-0.845	0.039	0.064	0.004	0.950	0.948	0.949
GEE-Independent	-5.664	-0.655	-59.568	0.021	0.033	0.016	0.934	0.949	0.232
GEE-Exchangeable	9.115	2.775	14.765	0.025	0.035	0.003	0.887	0.946	0.896
GEE-Serial	-1.167	2.667	-12.832	0.016	0.033	0.003	0.948	0.947	0.912
γ'=(-2,3,0)	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	1.378	-4.058	-4.222	0.149	0.281	0.016	0.947	0.950	0.948
WGEE-Exchangeable	3.205	-1.938	1.937	0.164	0.321	0.019	0.949	0.986	0.950
WGEE-Serial	2.135	-2.331	-1.694	0.145	0.224	0.017	0.948	0.951	0.947
GEE-Independent	-8.092	-2.351	-96.967	0.027	0.034	0.040	0.915	0.951	0.023
GEE-Exchangeable	12.713	4.060	15.781	0.032	0.039	0.004	0.833	0.947	0.915
GEE-Serial	-4.359	2.588	-42.090	0.020	0.035	0.010	0.939	0.947	0.620
γ'=(-2,1,0)	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	0.209	1.634	-1.075	0.024	0.046	0.002	0.948	0.949	0.948
WGEE-Exchangeable	0.180	1.474	-1.053	0.023	0.048	0.002	0.951	0.949	0.950
WGEE-Serial	0.327	1.741	-0.967	0.024	0.045	0.002	0.948	0.948	0.950
GEE-Independent	-1.886	1.129	-22.520	0.017	0.032	0.004	0.948	0.950	0.793
GEE-Exchangeable	4.006	1.907	5.992	0.018	0.031	0.002	0.941	0.946	0.937
GEE-Serial	0.006	2.232	-3.336	0.016	0.030	0.002	0.952	0.947	0.945

Source: Simulation study – McLeish, S.

We do observe some changes when using different drop-out models – particularly for the bias associated with unweighted models. Not surprisingly, the stronger the association between the probability of dropping out and our auxiliary information (previous value of y), the more non-response bias is present in the unweighted generalized estimating equations. Thankfully, no matter the level of missingness, our weighted GEEs continue to provide unbiased estimates as is further evidence by the coverage probability estimates.

We also note that the more missingness present (due to stronger MAR with model staying constant), the variance and thus the mean squared error associated with our weighted models increases. The added missingness has ignorable effect on the variance of the unweighted model estimates.

3.4.4 Results for different correlation structures with drop-outs at random (MAR)

Considering the fact that our drop-out model is based on the previous values of y, the between-subject association is essentially the link between the drop-out mechanism and the current value of y for MAR missingness. So we now consider a weaker correlation parameter (Still using serial correlation to generate the data) to analyze if this has any effect on our results. For both models, we employ $\beta' = (-1,1,.2)$ and $\gamma' = (-2,2,0)$. However, our original model has serial correlation structure with α=0.5 while the 2nd model in this analysis has serial correlation structure with α=0.3.

Table 3.4.4 Simulation results: Data with MAR for different correlation structures

T=5, α=0.5	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	-0.030	0.622	-1.313	0.041	0.071	0.004	0.947	0.948	0.947
WGEE-Exchangeable	0.096	0.619	-0.761	0.040	0.064	0.004	0.948	0.949	0.947
WGEE-Serial	0.235	0.965	-0.845	0.039	0.064	0.004	0.950	0.948	0.949
GEE-Independent	-5.664	-0.655	-59.568	0.021	0.033	0.016	0.934	0.949	0.232
GEE-Exchangeable	9.115	2.775	14.765	0.025	0.035	0.003	0.887	0.946	0.896
GEE-Serial	-1.167	2.667	-12.832	0.016	0.033	0.003	0.948	0.947	0.912
T=5, α=0.3	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	0.733	0.665	1.005	0.038	0.056	0.004	0.952	0.951	0.952
WGEE-Exchangeable	0.798	0.817	1.238	0.038	0.054	0.004	0.952	0.947	0.952
WGEE-Serial	0.756	0.785	1.172	0.038	0.054	0.004	0.953	0.946	0.953
GEE-Independent	-1.714	-0.728	-28.594	0.017	0.029	0.005	0.947	0.949	0.739
GEE-Exchangeable	6.880	1.666	13.207	0.021	0.030	0.003	0.913	0.947	0.913
GEE-Serial	0.050	1.170	-3.504	0.016	0.029	0.002	0.947	0.948	0.946

Source: Simulation study – McLeish, S.

For the weighted generalized estimating equations, we witness negligible difference between the results of these 2 approaches. However, the weaker between-subject association implies that the MAR association between the previous value of y and the drop-out mechanism is less important to the determination of the current value of y. This means that the unweighted GEE models are less biased when this association is less (since we essentially have a weaker MAR assumption).

3.4.5 Results for different sample sizes with drop-outs at random (MAR)

Finally, while almost all of these simulations were conducted with 5000 samples of size 500. We also compare our original results (with sample size 500) against the same approach but with a sample size of 100. For both models, we employ

$\beta' = (-1,1,.2)$ and $\gamma' = (-2,2,0)$. We also return to a serial correlation structure with α=0.5.

Table 3.4.5 Simulation results: Data with MAR for different sample sizes

k=500	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	-0.030	0.622	-1.313	0.041	0.071	0.004	0.947	0.948	0.947
WGEE-Exchangeable	0.096	0.619	-0.761	0.040	0.064	0.004	0.948	0.949	0.947
WGEE-Serial	0.235	0.965	-0.845	0.039	0.064	0.004	0.950	0.948	0.949
GEE-Independent	-5.664	-0.655	-59.568	0.021	0.033	0.016	0.934	0.949	0.232
GEE-Exchangeable	9.115	2.775	14.765	0.025	0.035	0.003	0.887	0.946	0.896
GEE-Serial	-1.167	2.667	-12.832	0.016	0.033	0.003	0.948	0.947	0.912
k=100	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	2.613	2.381	-0.874	0.205	0.412	0.022	0.948	0.949	0.950
WGEE-Exchangeable	3.627	3.478	3.027	0.203	0.372	0.024	0.947	0.947	0.948
WGEE-Serial	3.559	4.088	2.231	0.199	0.375	0.023	0.946	0.946	0.951
GEE-Independent	-4.535	2.144	-58.964	0.101	0.194	0.025	0.948	0.952	0.789
GEE-Exchangeable	10.318	5.385	15.768	0.101	0.201	0.011	0.936	0.951	0.937
GEE-Serial	0.580	6.081	-10.183	0.092	0.254	0.012	0.952	0.974	0.956

Source: Simulation study – McLeish, S.

Not surprisingly, the smaller data sets offer less accurate results for all models. The smaller number of subjects leads to higher variance and mean squared error for both weighted and unweighted models, while bias estimates remain fairly consistent.

However, we continue to see the same outcomes. Weighted generalized estimating equations provide unbiased estimates while unweighted generalized estimating equations provide biased estimates. The weighted models have less bias at the cost of higher variance.

3.4.6 Comparison of results for weighted generalized estimating equation methods by Robins et al. (1995) and Fitzmaurice et al. (1994)

As described in the theory section of this thesis (2.5.1 and 2.5.2), while the Fitzmaurice et al. approach to weighted generalized estimating equations is used throughout this thesis, it is a modification of the Robins et al. approach. Robins et al. used a weighting matrix with diagonal elements representing the inverse probability of responding for each time while Fitzmaurice used a scalar weight representing the inverse probability of following the response pattern that was followed. We compare these 2 methods against an unweighted generalized estimating equation for a specific case. We again generate the data using $\beta' = (-1,1,.2)$ and $\gamma' = (-2,2,0)$ with a serial correlation structure with α=0.5. However, we only focus on fitting the generalized estimating equations using the independent working correlation structure.

Table 3.4.6 Simulation results: Comparison of methods by Robins et al. (1995) and Fitzmaurice et al. (1994) for data with MAR

	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
GEE-Independent	-5.600	-0.189	-59.638	0.013	0.015	0.015	0.910	0.962	0.048
WGEE-Fitzmaurice	-0.221	2.127	-2.156	0.022	0.038	0.002	0.940	0.956	0.944
WGEE-Robins	-0.219	3.317	-2.344	0.014	0.029	0.002	0.954	0.944	0.936

Source: Simulation study – McLeish, S.

We can observe that both approaches provide unbiased estimates of our parameter estimates at a similar expense of variance while the unweighted model is inherently biased. So either approach can obviously be used for weighted generalized estimating equations.

3.5 Results for non-ignorable drop-outs (NI)

Moving away from data missing at random (MAR), we now consider the worst possible case of missingness – non-ignorable. In this case, the missingness depends on the current value of our variable of interest y, and not necessarily on any auxiliary variables. Using weighted generalized estimating equations, we can only assume MAR and so we would fit the models assuming that the informative missingness is still at random. Ultimately, this implies that our weighted generalized estimating equations will be inherently biased as we cannot properly specify the drop-out model.

We again generate the data using $\beta' = (-1,1,.2)$ but this time we employ $\gamma' = (-2,0,1)$ (Missingness is weak NI with no association with previous value of y) and $\gamma' = (-2,2,2)$ (Missingness is stronger NI with association with previous value of y) respectively.

Table 3.5.1 Simulation results: Data with NI for different drop-out mechanisms

γ'=(-2,0,1)	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	-6.369	-2.739	-41.672	0.027	0.044	0.009	0.933	0.947	0.569
WGEE-Exchangeable	-6.821	-3.243	-42.272	0.027	0.045	0.009	0.928	0.950	0.549
WGEE-Serial	-8.013	-2.829	-44.566	0.029	0.042	0.010	0.922	0.950	0.532
GEE-Independent	-4.275	0.967	-46.313	0.019	0.032	0.010	0.940	0.954	0.375
GEE-Exchangeable	-1.535	1.045	-31.959	0.018	0.031	0.006	0.946	0.952	0.651
GEE-Serial	-4.947	1.367	-38.439	0.019	0.030	0.007	0.934	0.952	0.503
γ'=(-2,2,2)	% Relative Bias			Mean Squared Error			Coverage Probability		
	Int	x	t	Int	x	t	Int	x	t
WGEE-Independent	-33.603	-43.880	-215.730	0.281	0.447	0.209	0.883	0.879	0.191
WGEE-Exchangeable	-33.619	-42.865	-215.483	0.284	0.414	0.209	0.884	0.875	0.199
WGEE-Serial	-36.373	-42.074	-219.617	0.292	0.402	0.215	0.866	0.874	0.160
GEE-Independent	-27.814	-7.058	-233.104	0.102	0.038	0.221	0.572	0.932	0.000
GEE-Exchangeable	-21.357	-5.458	-199.398	0.069	0.037	0.164	0.721	0.940	0.000
GEE-Serial	-29.579	-5.546	-213.690	0.110	0.037	0.187	0.508	0.938	0.000

Source: Simulation study – McLeish, S.

We observe that in both models, with non-ignorable missingness, our weighted generalized estimating equations incur non-response bias just as the unweighted models. So, we see that weighted generalized estimating equations can be quite useful when missingness can be explained by auxiliary variables or information (such as the previous value of y). However, when the probability of dropping out depends on the current value of our variable of interest, we cannot correct this bias using weighted GEEs.

Chapter 4

Application to the National Population Health Survey

4.1 Introduction to the National Population Health Survey

The National Population Health Survey (NPHS) is a voluntary longitudinal survey conducted by Statistics Canada. The program follows a sample of 17,276 respondents from 1994/1995. Information is collected every 2 years with the most recent data available for 2008/09. The survey captures information on a wide range of health-related topics such as:

- Chronic illness and disease
- Physical activity and lifestyle
- Nutrition
- Mental health
- Medication and health care
- Socio-economic characteristics

The ultimate goal of the survey is to gain insight into the factors influencing health outcomes over time. This comes in the way of understanding health trends in Canada, the association between health care and status, and the relationships between an individual's health and their socio-economic or lifestyle characteristics.

4.1.1 Sources of error in the National Population Health Survey

As a voluntary sample survey, the NPHS is subject to coverage error, sampling error, measurement error and non-response error. In this thesis, we assume that any sampling/coverage bias or initial non-response bias (response for first cycle) is corrected through the sampling weights included on the dataset.

As a longitudinal survey, one of the principal methodological issues facing this survey is attrition. As time goes on, respondents may feel less and less inclined to continue to participate in the survey. In addition, death and emigration (or migration within Canada) can play a major role in the disappearance of respondents over time. The overall attrition rate of the NPHS considers those individuals who are deceased to continue to exist on the full dataset (since death may be a useful indicator/result for some health analysis). The overall attrition of the NPHS is demonstrated in the table below.

Table 4.1.1 Number and percentage of NPHS respondents by cycle

Cycle	Number of individuals in full subset	% of original sample
1994/1995	17,276	100.0%
1996/1997	15,672	90.7%
1998/1999	14,630	84.7%
2000/2001	13,596	78.7%
2002/2003	12,559	72.7%
2004/2005	11,619	67.3%
2006/2007	10,992	63.6%
2008/2009	9,982	57.8%

Source: National Population Health Survey (1994/95-2008/09)

There is also a fair amount of item non-response present within the longitudinal NPHS dataset. This occurs when the respondent has responded for a given cycle but did not provide complete responses for all questions. This could be due to refusal (the respondent does not wish to disclose the response to certain questions), not

knowing the answer (could be answering by proxy or does not recall), or not being in scope for the question. The latter reflects valid non-response and is not an issue if the universe is simply restricted to those individuals in scope (for example, information on employment status is not applicable to individuals younger than 15 years of age – so we would limit the scope of our analysis of this question to individuals aged 15 or older). Item non-response replaces overall non-response (both initial non-response and attrition) when those affected variables are required for analysis.

4.1.2 Sample design for the National Population Health Survey

The NPHS sample was not a simple random sample of the Canadian population in 1994/1995. For all provinces except Québec, it relied on the same methodology used for the Labour Force Survey (a prominent survey undertaken by Statistics Canada, which reveals important labour statistics such as the unemployment rate). For Québec, a provincial survey "Enquête sociale et de santé" (1992/1993) was used.

The LFS survey design consisted of a stratified 2-stage sample. Strata were defined by geographic and socio-demographic information for different areas within each province (urban cities, urban towns, and rural areas). First, clusters were selected – these tended to be small geographic areas – then within each cluster, dwellings were selected. The "Enquête sociale et de santé" also used a 2-stage design where areas were created by a cross between provincially determined health regions and urban vs. rural areas. As with the LFS design, within each area, clusters were selected based on socio-demographic characteristics from which dwellings were then randomly selected.

For the NPHS, the sample was more complex as it first selected households and then selected individuals within the household. For the longitudinal component, the

sample was restricted to individuals aged 12 and over and was then supplemented by 1994/1995 respondents to the National Longitudinal Survey of Children and Youth (NLSCY) to capture data on those under the age of 12. This analysis focuses on individuals aged 15+ in 1994/1995 and therefore does not consider this portion of the NPHS sample.

Because of the complexity associated with the NPHS sample, traditional formulae for variance cannot be used even for simple point estimates (Means, counts, etc.). Therefore, bootstrap methods must be employed for these purposes (Appendix 4).

4.1.3 Variables of interest in this study

This thesis concentrates on the role of physical activity and body mass in the incidence of high blood pressure. Therefore, a variable to measure incidence of high blood pressure, a variable to measure physical activity, a variable to measure body mass, and other control variables such as socio-economic characteristics are required. The variables included in this analysis are:

- High blood pressure
- Physical activity category: Active, Moderate, Inactive
- Body mass Index (BMI)
- Marital status
- Employment status
- Age in 1994/95
- Sex
- Urban/Rural: Geographic location
- Cycle

High blood pressure

This is provided by variable CCCX_1F where X indicates the cycle number. Respondents were asked if they had high blood pressure diagnosed by a physician / health professional as a long-term condition. Responses are binary – Yes or No. The question was limited to respondents aged 12 and older.

Table 4.1.2 Weighted NPHS counts for High blood pressure for Cycle 1

Response	Population
Yes	2,076,756
No	21,555,499
Not applicable	4,655,277
Don't know	17,400
Refusal	13,376

Source: National Population Health Survey (1994/95-2008/09)

We can observe that in 1994/95, most Canadians in-scope (Aged 12 and older) did not have high blood pressure. However those that did still represented about 10% of the adult population. We note that item non-response for this question comes in the form of refusals and "Don't know" responses and makes up a negligible proportion of the population.

Physical activity category: Active, Moderate, Inactive

This is provided by variable PACXDPAI where X indicates the cycle number. This is a derived variable based on respondents' answers to several questions. The variable relies on the respondents' answers to questions about their daily physical activities and an energy cost per activity. Their average daily energy expenditure is determined by taking the sum of total energy spent on each activity per day. The respondents are then classified by:

- Active: 3.0 kcal/kg/day or more
- Moderate: 1.5 to 2.9 kcal/kg/day
- Inactive: Less than 1.5 kcal/kg/day

For this analysis, this variable was converted into a binary field indicating whether or not the respondent is either active or moderately active. The question was limited to respondents aged 12 and older.

Table 4.1.3 Weighted NPHS counts for Physical activity for Cycle 1

Response	Population
Active	4,206,304
Moderate	4,711,344
Inactive	12,507,509
Not applicable	4,604,917
Not stated	2,288,235

Source: National Population Health Survey (1994/95-2008/09)

As is the case for the high blood pressure variable, physical activity was limited to Canadians aged 12 and over in 1994/95. These individuals were more likely to be inactive than moderately or physically active. Non-response in the form of "Not stated" does not represent a trivial subsample of our population. Ignoring these non-responses could carry considerable risk as a result.

Body Mass Index (BMI)

This is provided by variable HWCXDBMI where X indicates the cycle number. This is a derived variable based on respondents' answers to weight and height. BMI is measured by weight (in kilograms) divided by height (in metres) squared. The question was limited to respondents living in a household excluding pregnant women. Here responses are continuous and summarized in the table below.

Table 4.1.4 Weighted NPHS counts for Body Mass Index (BMI) for Cycle 1

Response	Population
Valid response	25,597,364
Not applicable	896,746
Not stated	1,824,199

Source: National Population Health Survey (1994/95-2008/09)

Body mass index (BMI) provides a measurement of an individual's physical "shape" and is commonly used to determine if an individual is obese, overweight, or underweight. Like the physical activity variable, non-response in the form of "Not stated" is non-trivial and will need to be considered further.

Marital status

This is provided by variable DHCX_MAR where X indicates the cycle number. This comes from a question asking for the respondent's current marital status. For this analysis, this variable was converted into a binary field indicating whether or not the respondent is in a couple (includes married individuals and those living in a common-law relationship or with a partner).

Table 4.1.5 Weighted NPHS counts for Marital status for Cycle 1

Response	Population
Now married	12,394,605
Common-Law	1,488,462
Living with a partner	66,350
Single (Never married)	11,498,474
Widowed	1,264,451
Separated	556,449
Divorced	1,044,962
Refusal	4,555

Source: National Population Health Survey (1994/95-2008/09)

Marital status in Canada in 1994/95 shows that the vast majority of the population at that time as either single (never married) or legally married. It should be noted that many of the single individuals are youth and will not be in scope for this thesis. Item non-response in the form of refusals is negligible.

Employment status

This is provided by LSCXDLFS, where X indicates the cycle number. This is a derived variable based on several variables that indicate the respondent's labour force status. For this analysis, this variable was converted into a binary field indicating whether or not the respondent is unemployed (if they were not in the labour force they were not considered to be unemployed). The question was limited to respondents aged 15 and older.

Table 4.1.6 Weighted NPHS counts for Employment status for Cycle 1

Response	Population
Employed	13,244,753
Unemployed	1,053,374
Not in the labour force	7,588,528
Not applicable	5,944,849
Not stated	486,803

Source: National Population Health Survey (1994/95-2008/09)

We observe that for Canadians aged 15 and older in 1994/95, the majority were employed. The second largest group consists of those who are not in the labour force which would indicate that they are retired, stay-at-home parents, or simply not looking for work. Item non-response for this question in the form of "Not stated" is fairly low.

Age in 1994/95

This is provided by DHCX_AGE, where X indicates the cycle number. This is provided by the respondent who indicates their age in years. For this analysis, only the value for 1994/95 is used. There is no item non-response for this question.

Sex

This is provided by SEX. This is provided by the respondent who indicates their sex. For this analysis, this variable is a binary indicator of whether or not the respondent is female. There is no item non-response for this question.

Table 4.1.7 Weighted NPHS counts for Sex for Cycle 1

Response	Population
Male	14,020,059
Female	14,298,249

Source: National Population Health Survey (1994/95-2008/09)

Urban/Rural: Geographic location

This is provided by GE34DPOP for the first cycle. This is a derived variable based on the respondent's postal code. For this analysis, this variable was converted into a binary field indicating whether or not the respondent lives in an urban area.

Table 4.1.8 Weighted NPHS counts for Urban/Rural for Cycle 1

Response	Population
Rural area	7,140,472
Urban area: Less than 30,000 people	977,817
Urban area: 30,000 - 99,999 people	2,877,105
Urban area: 100,000 - 499,999 people	3,975,261
Urban area: 500,000 or more people	12,467,291
Not stated	880,362

Source: National Population Health Survey (1994/95-2008/09)

We observe that the majority of Canadians in 1994/95 lived in urban centres (particularly larger metropolitan areas such as Toronto, Vancouver and Montréal). Item non-response of this question comes in the form of "Not stated" and is non-trivial.

4.1.4 Attrition rate for this study

If we consider only our primary variable of interest: Incidence of high blood pressure, our attrition rate would be defined by the number of individuals who respond to that specific question by cycle.

Table 4.1.9 Percentage of NPHS respondents who responded to high blood pressure question by cycle and age group

Row Labels	Aged 15+ in 1994/95 (%)	Total Population (%)
1994/1995	99.8%	88.2%
1996/1997	91.2%	82.7%
1998/1999	84.2%	78.2%
2000/2001	77.3%	74.1%
2002/2003	70.9%	69.7%
2004/2005	65.4%	66.1%
2006/2007	62.7%	64.7%
2008/2009	55.3%	56.8%

Source: National Population Health Survey (1994/95-2008/09)

We see that by the 8th wave (2008/2009), we have lost almost 45% of our original sample (Slightly more than the overall attrition rate). Restricting the attrition rate to the response for the high blood pressure variable is generous. The principal element of this analysis is to study the drop-out mechanism which relies on the previous value to be modelled. Therefore, when we determine our drop-out time, M, we must have complete responses for all times $t < M$. So we must define our drop-out time as the first time that a response is missing.

Note that although we have both drop-outs and intermittent missing data, we only consider drop-outs when analyzing the data. The 20 most prevalent response patterns are provided in the table below. Each digit represents a response where the first digit (from left) reflects the response to the first wave and so on. 1 indicates a valid response to the high blood pressure question and 0 indicates non-response to that particular question.

Table 4.1.10 Most prevalent response patterns (relating to high blood pressure question) for NPHS respondents aged 15+ in 1994/95

Response Pattern	%	Drop-out time (M)
11111111	46.1%	8
11111110	6.5%	7
10000000	6.5%	1
11100000	5.8%	3
11000000	5.3%	2
11110000	5.1%	4
11111000	4.4%	5
11111100	4.1%	6
11111011	1.6%	5
11110111	1.1%	4
11111101	1.1%	6
11111010	1.0%	5
11101111	0.9%	3
11011111	0.7%	2
11010000	0.6%	2
11110100	0.5%	4
11110110	0.5%	4
11101000	0.4%	3
11111001	0.4%	5
11110011	0.4%	4

Source: National Population Health Survey (1994/95-2008/09)

We see that fortunately, the most prevalent pattern is a complete data set from the first observation to the 8th. The next 7 most prevalent patterns are pure drop-outs in the sense that once they disappear from the survey, they do not return. These first 8 patterns represent about 83.7% of all response patterns for the first 8 cycles of the NPHS. The remaining response patterns include a degree of intermittent missingness where a subject may not respond one cycle but will respond in a future cycle. So, as described above, any observations that are made after a subject has registered a non-response will be ignored and the time of the first missing value for high blood pressure will be considered the time of drop-out.

Table 4.1.11 Percentage of NPHS respondents who responded to the high blood pressure question every cycle prior to and including the current cycle by cycle and age group

Row Labels	Aged 15+ in 1994/95 (%)	Total Population (%)
1994/1995	99.8%	88.2%
1996/1997	91.1%	80.6%
1998/1999	82.7%	73.3%
2000/2001	73.5%	65.2%
2002/2003	65.1%	57.8%
2004/2005	57.8%	51.2%
2006/2007	52.6%	46.5%
2008/2009	46.1%	40.7%

Source: National Population Health Survey (1994/95-2008/09)

We obviously see that this reduces our response rates for later cycles significantly (by about 10 percentage points for the population aged 15 and over in 1994/95 for cycles 7 and 8). The true response rates for our analysis are, in fact, even smaller since we also need to consider non-response for our covariates.

4.2 Sample design and selection weights for analysis

Once the NPHS data is collected, sampling weights are assigned to account for the sampling design and general non-response. These weights are included on the dataset and are based on the 1994/1995 Canadian population. We assume these weights are reflective of the Canadian population at that time. Let w_i be the weight associated with respondent i, n be the sample size, and N be the population of Canadia in 1994/95. The Horvitz-Thompson estimator of the true population mean can be found from

$$\hat{\mu}_y = \frac{\sum_{i=1}^{n} w_i y_i}{N}, \tag{4.2.1}$$

where weight w_i is the inverse probability of selection for individual i and $\sum_{i=1}^{n} w_i = N$. We limit the data set to those variables which we are interested in studying (the primary variable of interest, incidence of high blood pressure, and our covariates) and to our population of interest (all Canadian residents aged 15+ in 1994/1995). It is important to note that while an individual may have responded to some questions on the survey (and were therefore classified as a respondent during the overall survey weighting process), they did not necessarily respond to the questions within the scope of our study. For example, person A might have responded to some questions about cancer but did not answer the question on high blood pressure.

Therefore, we must create new weights to account for any informative selection amongst those who chose to answer the questions for our variables of study. To do this, we define a logistic regression where we model

$$logit\big(p(selection|sample)\big) = X'\beta, \tag{4.2.2}$$

where X represents variables associated with the probability of being in the selection.

Selection is based on the original NPHS respondent sample (restricted to our population of interest), where individuals who responded to the variables under study are selected. The definition of "complete response for our variables of study" is not fixed. This depends on the type of analysis we wish to perform.

The covariates used for this model should have a non-missing value for all respondents in the sample (including those not selected). Otherwise, the analysis will need to simply ignore individuals with missing values. This could result in a meaningless selection model if selection is based on whether these values are missing or not and ultimately will lead to some observations missing selection weights. To ensure complete covariate information, we turn all covariates into categorical variables and let non-response be represented by a category.

It should be noted that this selection model must be weighted using the original sample design weights (since we are considering everyone in the sample for the selection model).

Once we determine which individuals are selected, or not selected, based on item non-response, we calculate the above logistic regression to find $\hat{\beta}$. This provides us with an estimate of the relationship between the probability of being selected and some associated covariates. Based on this estimated relationship we can derive predicted probabilities $logit(\hat{p}_i) = X_i{}'\hat{\beta}$ for each respondent i. As demonstrated in section 2.4, we can then restrict the sample to those selected and find their selection weight as the inverse of the probability of being selected: $\widehat{w}_i = 1/\hat{p}_i$.

4.2.1 Calibration of weights to ensure consistent population totals

Once we have determined the selection weights: $\widehat{w}_i(selection|sample)$ for each respondent, we can determine new individual weights by finding the inverse

probability of Canadians being in the analytical subsample. This is found by taking the product of the original sampling weight (which represents the inverse probability of being in the NPHS sample) and the selection weight (which represents the inverse probability of being in the analytical subsample conditional upon being in the original NPHS sample):

$$p_i(selection) = p_i(sample) \cdot p_i(selection|sample) \tag{4.2.3}$$

$$\begin{aligned} &\Rightarrow 1/p_i(selection) = 1/(p_i(sample) \cdot p_i(selection|sample)) \\ &\Rightarrow w_i(selection) = w_i(sample) \cdot w_i(selection|sample). \end{aligned} \tag{4.2.4}$$

We use our estimates $\widehat{w}_i$ and find the selection weight for each of the respondents in scope for the analysis (note that $\widehat{w}_i(selection|sample) = 0$ for all those records not selected):

$$\widehat{w}_i(selection) = w_i(sample) \cdot \widehat{w}_i(selection|sample). \tag{4.2.5}$$

Now, this new weight should still satisfy the condition

$$\sum_{i=1}^{n} \widehat{w}_i(selection) = \sum_{i=1}^{n} w_i(sample) = N. \tag{4.2.6}$$

However, due to model misspecification in the logistic regression (since $\widehat{w}_i(selection|sample) \neq w_i(selection|sample)$), this condition may not hold. Therefore, we must calibrate all of our weights to ensure that this condition is met. We find our final weight by

$$\begin{aligned} \widehat{w}_i(final) &= w_i(sample) \cdot \widehat{w}_i(selection|sample) \cdot \frac{\sum_{i=1}^{n} w_i(sample)}{\sum_{i=1}^{n} \widehat{w}_i(selection)} \\ &= \widehat{w}_i(selection) \cdot \frac{\sum_{i=1}^{n} w_i(sample)}{\sum_{i=1}^{n} \widehat{w}_i(selection)}. \end{aligned} \tag{4.2.7}$$

Therefore:

$$\sum_{i=1}^{n} \widehat{w}_i(final) = \sum_{i=1}^{n} \widehat{w}_i(selection) \cdot \frac{\sum_{i=1}^{n} w_i(sample)}{\sum_{i=1}^{n} \widehat{w}_i(selection)} \tag{4.2.8}$$

$$= \frac{\sum_{i=1}^{n} w_i(sample)}{\sum_{i=1}^{n} \widehat{w}_i(selection)} \cdot \sum_{i=1}^{n} \widehat{w}_i(selection) = \sum_{i=1}^{n} w_i(sample) = N.$$

In theory if our selection mechanism based on item non-response was completely random, we could have simply calibrated by multiplying the original sample weights by the sum of all sample weights divided by the sum of the sample weights in the selection subset. However, by modelling the selection model, we assume that selection is informative based on those characteristics used in the logistic regression (MAR).

4.3 Initial findings: Blood pressure analysis

4.3.1 Cursory analysis of high blood pressure and associated variables

High blood pressure or Hypertension is a chronic health condition of growing concern in Canada. The condition exists when the pressure of blood against arterial walls is higher than "normal" which means that the heart needs to work harder than it should to circulate blood throughout the body.

Generally, the prevalence of high blood pressure in Canada has been increasing steadily. According to the Canadian Community Health Survey (CCHS), more than 17% of Canadians aged 12 and over in 2010 have been diagnosed with high blood pressure.

Table 4.3.1 Percentage of population with high blood pressure by age group, sex, and year according to the CCHS

	2005	2007	2008	2009	2010
	percent				
Total, 12 years and over	15.0	16.0	16.4	16.9	17.1
Males	14.2	15.1	15.9	16.4	17.0
Females	15.7	16.8	16.9	17.3	17.2
12 to 19 years	0.5	0.6[E]	0.5[E]	0.6[E]	0.6[E]
Males	0.4[E]	0.7[E]	0.6[E]	0.8[E]	0.5[E]
Females	0.6[E]	0.5[E]	0.4[E]	0.5[E]	0.8[E]
20 to 34 years	2.5	2.6	2.4	2.1	2.3
Males	3.1	3.0	2.9	2.7	2.8
Females	2.0	2.1	1.8	1.4	1.9
35 to 44 years	6.6	7.0	7.1	7.6	8.0
Males	7.4	8.2	7.6	8.4	10.0
Females	5.8	5.9	6.7	6.9	6.1
45 to 64 years	21.7	22.6	23.2	23.2	23.3
Males	22.1	22.3	24.2	24.3	25.1
Females	21.4	22.8	22.2	22.1	21.5
65 years and over	44.2	46.5	47.1	48.9	47.7
Males	39.5	42.7	43.6	44.9	43.7
Females	48.0	49.5	50.0	52.2	51.1

[E] : use with caution.

Source: Canadian Community Health Survey, Statistics Canada, Summary tables, CANSIM, table 105-0501 and Catalogue no. 82-221-X.

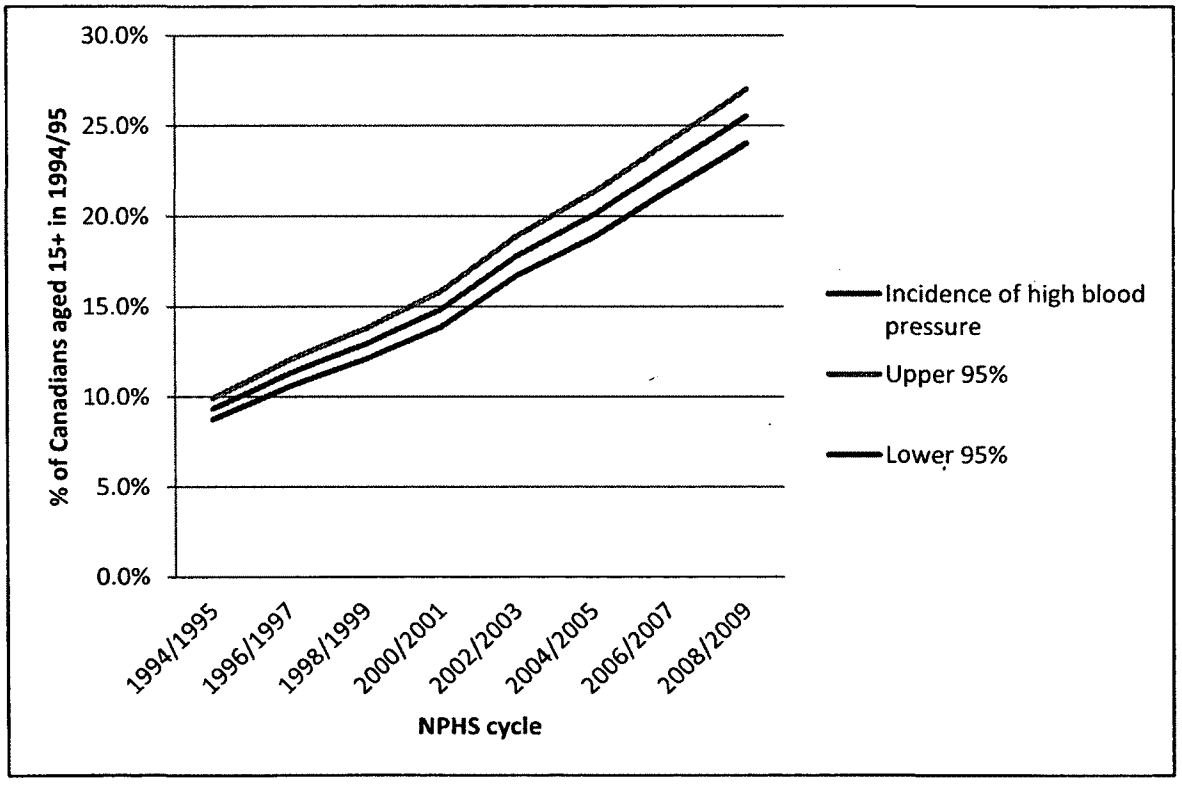
According to the Public Health Agency of Canada, high blood pressure increases the risk of contracting various serious health conditions such as dementia, heart disease, kidney failure, and strokes (heart disease is the most common cause of death in Canada). They note that high blood pressure is related to 13% of deaths and it is the most common reason to visit a doctor or take medication.

High blood pressure is caused by many different factors such as diet and physical activity. According to the Public Health Agency of Canada, high sodium diets are a

major cause of hypertension in Canadians. In this analysis, we explore the relationship between incidence of high blood pressure and covariates body mass index (BMI) and level of physical activity.

According to the NPHS, in 1994/1995, 9.3% of Canadians aged 15 and above reported that they had been diagnosed with high blood pressure by a health care physician. For that same cohort (Canadians who were 15 years or older in 1994/1995), that percentage had risen to over 25% by 2008/09. Variance estimates are derived using the bootstrap method as described in Appendix 4.

Figure 4.3.1 NPHS % of Canadians aged 15+ in 1994/95 with high blood pressure by cycle with upper and lower 95% confidence intervals



Source: National Population Health Survey (1994/95-2008/09)

We can see that there is a large increase in the incidence of high blood pressure. This may be the result of the population aging roughly 14 years (between 1994/1995 and 2008/2009) if there is a strong correlation between age and high

blood pressure. Alternatively, the course of time may be influencing other factors such as level of physical activity and body mass index (BMI) which could be contributing to the rising levels of high blood pressure among this cohort.

We begin our analysis focussing on the first cycle of the NPHS in 1994/1995. We wish to identify any effects which may be correlated with incidence of high blood pressure for this cycle. The following tables outline the mean values for different characteristics for the population with high blood pressure and the population without high blood pressure:

Table 4.3.2 NPHS Cycle 1 characteristics of Canadians aged 15+ in 1994/95 with high blood pressure by presence of high blood pressure

Does not have high blood pressure	Average	Lower 95%	Upper 95%
Female	52.2%	51.5%	52.8%
Lives in urban area	75.0%	73.6%	76.3%
Married or Common-Law	62.1%	61.1%	63.1%
Body Mass Index (BMI)	24.8	24.6	24.9
Weight (lbs)	156.5	155.7	157.4
Height (inches)	56.5	56.5	56.6
Unemployed	5.0%	4.5%	5.5%
Physically active or moderately active	40.6%	39.4%	41.8%
Age (in 1994/1995)	40.7	40.5	41.0
Has high blood pressure	**Average**	**Lower 95%**	**Upper 95%**
Female	60.1%	56.6%	63.5%
Lives in urban area	69.9%	66.2%	73.6%
Married or Common-Law	63.3%	59.7%	66.9%
Body Mass Index (BMI)	27.3	26.9	27.6
Weight (lbs)	167.1	164.3	170.0
Height (inches)	55.5	55.2	55.9
Unemployed	3.0%	1.8%	4.2%
Physically active or moderately active	36.1%	32.2%	40.0%
Age (in 1994/1995)	60.2	59.1	61.3

Source: National Population Health Survey (1994/95-2008/09)

We can see that some of these variables are related to incidence of high blood pressure. For example, women appear to be far more likely to have high blood pressure than men, those individuals with high body mass index values are more likely to have high blood pressure (interestingly, those with high blood pressure are both shorter and heavier), and there is a large difference between the average age of those with high blood pressure and those without. We see that within the sample, physical activity is not a statistically significant indicator of hypertension.

Now, we consider how these relationships evolve. The results for cycles 2-7 can be found in Appendix 3. For now, we will focus on how these associations have changed for the 8th and most recent cycle of the National Population Health Survey (2008/09).

Table 4.3.3 NPHS Cycle 8 characteristics of Canadians aged 15+ in 1994/95 with high blood pressure by presence of high blood pressure

Does not have high blood pressure	Average	Lower 95%	Upper 95%
Female	50.3%	48.7%	51.9%
Lives in urban area	74.2%	72.3%	76.2%
Married or Common-Law	70.4%	68.7%	72.1%
Body Mass Index (BMI)	26.3	26.1	26.4
Weight (lbs)	167.3	165.8	168.7
Height (inches)	56.8	56.6	56.9
Unemployed	2.7%	2.1%	3.4%
Physically active or moderately active	57.3%	55.3%	59.3%
Age (in 1994/1995)	37.7	37.2	38.1
Has high blood pressure	**Average**	**Lower 95%**	**Upper 95%**
Female	56.5%	53.4%	59.7%
Lives in urban area	72.2%	69.2%	75.1%
Married or Common-Law	65.3%	62.1%	68.6%
Body Mass Index (BMI)	28.9	28.4	29.3
Weight (lbs)	178.0	175.0	180.9
Height (inches)	55.7	55.4	55.9
Unemployed	1.3%	0.6%	2.0%
Physically active or moderately active	42.5%	39.1%	46.0%
Age (in 1994/1995)	51.2	50.3	52.0

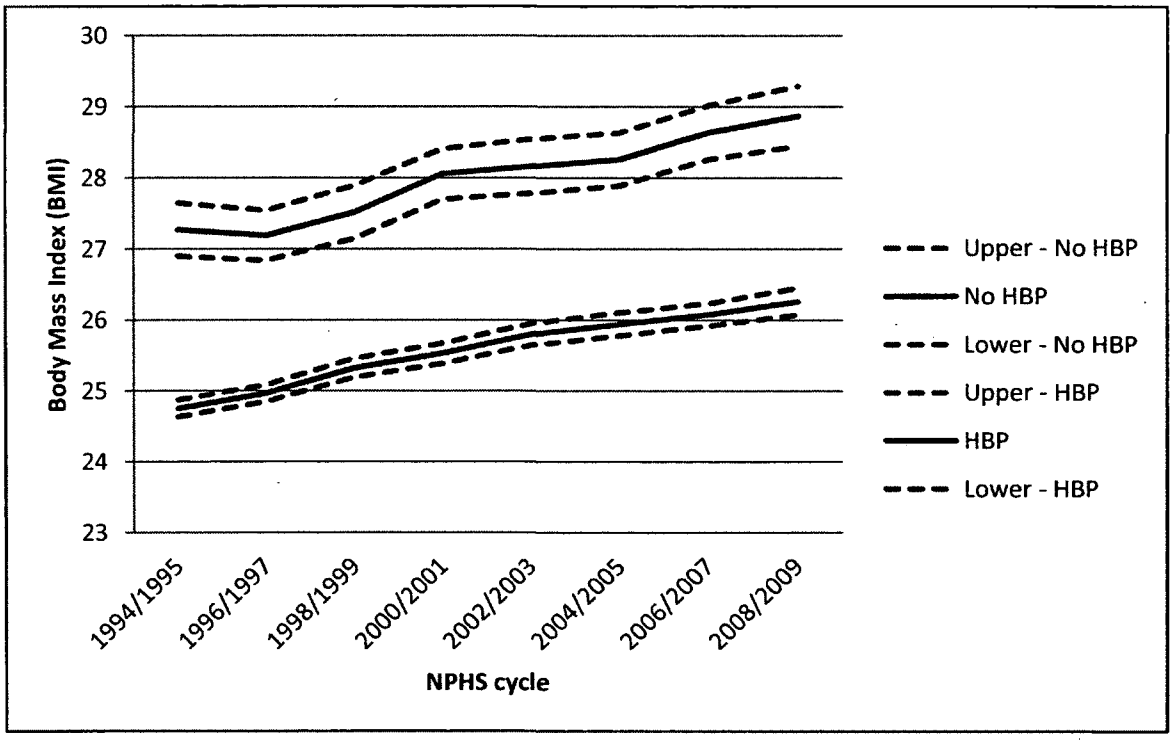
Source: National Population Health Survey (1994/95-2008/09)

Several of the relationships evolve over time as our population in scope ages and changes due to death and emigration. By the most recent cycle, 2008/2009, women are still more likely to have high blood pressure than men but the difference is reduced, BMI continues to be significantly different for those with and without high

blood pressure (those with HBP are still shorter and heavier), and age continues to be a major effect. The most notable change is that the level of physical activity is now significantly different for those with and without high blood pressure. This could be indicative of physical activity being more of an important effect for older individuals.

The following graph illustrates the average Body Mass Index (BMI) for those Canadians aged 15 and above in 1994/95 with and without High Blood Pressure by cycle.

Figure 4.3.2 Average Body Mass Index (BMI) for Canadians aged 15+ in 1994/95 with and without high blood pressure (HBP) by cycle with upper and lower 95% confidence intervals



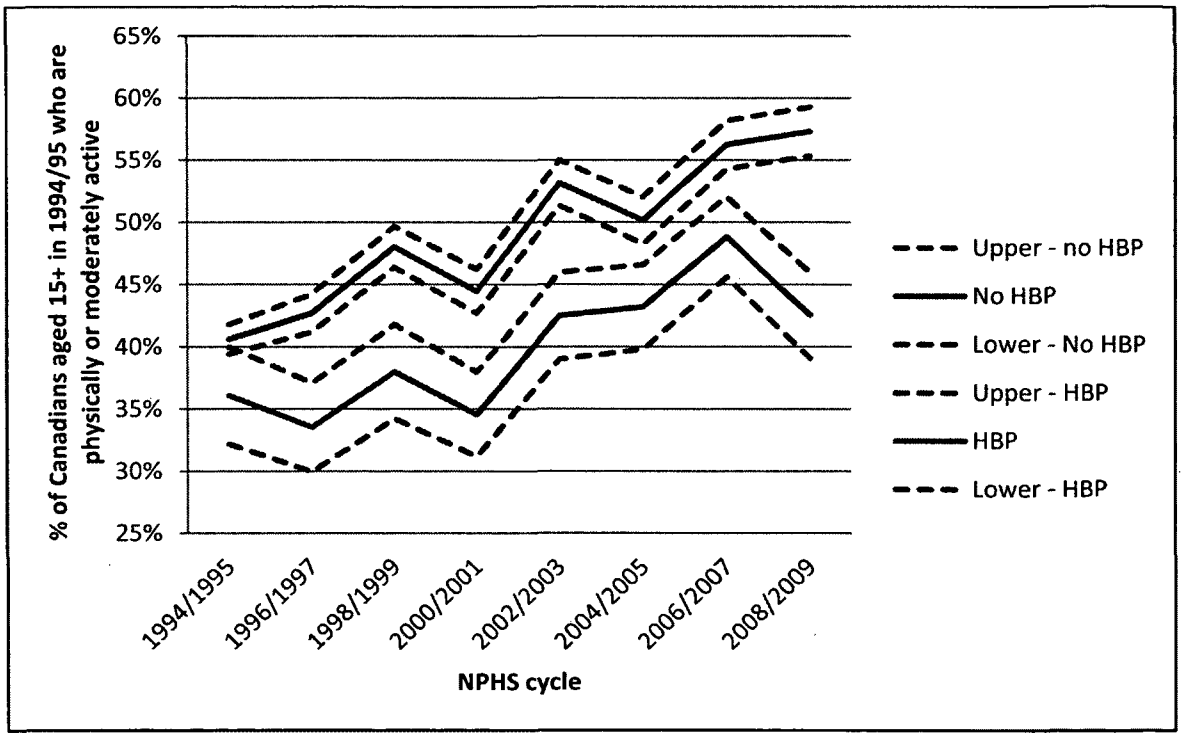
Source: National Population Health Survey (1994/95-2008/09)

We see that as the population ages, the average body mass index increases steadily over time. We also note that the average BMI is significantly higher for those with

high blood pressure than those without. This difference is consistent for all cycles of the NPHS and demonstrates that BMI is related to incidence of high blood pressure.

The following graph illustrates the percentage of Canadians aged 15 and above in 1994/95 who are physically or moderately physically active for those with and without High Blood Pressure by cycle.

Figure 4.3.3 Percentage of Canadians aged 15+ in 1994/95 who are physically or moderately active with and without high blood pressure (HBP) by cycle with upper and lower 95% confidence intervals



Source: National Population Health Survey (1994/95-2008/09)

We can see that physical activity is related to incidence of high blood pressure for all cycles of the NPHS. With the exception of the first cycle, the percentage of Canadians aged 15+ in 1994/95 with high blood pressure who are physically active is significantly less than those without high blood pressure. This indicates that individuals who are physically active are less likely to have high blood pressure.

Physical activity, while a variable that can incur a fair amount of flux, can be a consistent characteristic of some individuals. Those who are active may remain active and those who are inactive may remain so as well.

The following table outlines the percentage of Canadians aged 15+ in 1994/95 who are physically active by cycle and by physical activity in 1994/95.

Table 4.3.4 Percentage of Canadians aged 15+ in 1994/95 who are physically active by cycle and by physical activity in 1994/95

Cycle	Not active in 1994/95	Active in 1994/95	Grand total
1996/97	26.5%	64.2%	41.6%
1998/99	33.7%	65.7%	46.7%
2000/01	31.1%	60.1%	43.0%
2002/03	39.8%	67.5%	51.2%
2004/05	36.8%	65.4%	48.7%
2006/07	45.0%	68.0%	54.5%
2008/09	42.7%	68.6%	53.4%

Source: National Population Health Survey (1994/95-2008/09)

We can see that the overall level of physical activity has increased between the 1st and 8th cycles but there is considerable fluctuation between. Meanwhile, more than 60% of those individuals who were physically active in 1994/95 have remained active. However, those who were not active in 1994/95 have seen a large increase in physical activity from 0% in 1994/95 to 26.5% in 1996/97 up to 42.7% in 2008/09. Still we do see that those who were active in the first cycle are more likely to be active in subsequent cycles than those who were inactive in the first cycle.

4.4 Analysis of high blood pressure using generalized estimating equations

As illustrated in the theory section (2.2 and 2.3), logistic regression using maximum likelihood estimation can be used but it makes the major assumption that no correlation structure exists between observations. For longitudinal data, this assumption is fairly naïve since there are several observations coming from the same subject (and it would follow that observations from the same subject would be related). For this particular study, we examine the correlation structure of high blood pressure responses for the 8 cycles. Here we are simply estimating the correlation structure by finding the Pearson product-moment correlation coefficients $\rho_{xy} = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}}$.

Table 4.4.1 Estimated correlation matrix for high blood pressure responses by cycle

	HBP1	HBP2	HBP3	HBP4	HBP5	HBP6	HBP7	HBP8
HBP1	1.000	0.753	0.647	0.583	0.529	0.479	0.442	0.407
HBP2	0.753	1.000	0.788	0.681	0.606	0.553	0.512	0.467
HBP3	0.647	0.788	1.000	0.764	0.674	0.610	0.550	0.516
HBP4	0.583	0.681	0.764	1.000	0.789	0.701	0.617	0.572
HBP5	0.529	0.606	0.674	0.789	1.000	0.808	0.716	0.661
HBP6	0.479	0.553	0.610	0.701	0.808	1.000	0.808	0.747
HBP7	0.442	0.512	0.550	0.617	0.716	0.808	1.000	0.838
HBP8	0.407	0.467	0.516	0.572	0.661	0.747	0.838	1.000

Source: National Population Health Survey (1994/95-2008/09)

We observe that not only is there a relationship between observations from the same subjects, it is fairly strong – especially between consecutive waves. In fact, similar to serial correlation, we see that the correlation coefficients decrease as the

time between cycles increases. For a simple comparison, let us quickly consider an 8x8 working serial correlation matrix with alpha = 0.85.

Table 4.4.2 Working correlation matrix with serial correlation (α=0.85) with 8 observations per subject

	HBP1	HBP2	HBP3	HBP4	HBP5	HBP6	HBP7	HBP8
HBP1	1.000	0.850	0.723	0.614	0.522	0.444	0.377	0.321
HBP2	0.850	1.000	0.850	0.723	0.614	0.522	0.444	0.377
HBP3	0.723	0.850	1.000	0.850	0.723	0.614	0.522	0.444
HBP4	0.614	0.723	0.850	1.000	0.850	0.723	0.614	0.522
HBP5	0.522	0.614	0.723	0.850	1.000	0.850	0.723	0.614
HBP6	0.444	0.522	0.614	0.723	0.850	1.000	0.850	0.723
HBP7	0.377	0.444	0.522	0.614	0.723	0.850	1.000	0.850
HBP8	0.321	0.377	0.444	0.522	0.614	0.723	0.850	1.000

Source: None

We see that this is fairly similar to the estimated correlation structure for high blood pressure. This would indicate that a serial correlation structure might be appropriate when fitting generalized estimating equations for this data.

4.4.1 Formulation of survey generalized estimating equations

Generalized estimating equations look to find β to solve

$$\sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \mu_i) = 0, \tag{4.4.1}$$

where V_i is a $T_i x T_i$ matrix representing the working covariance matrix of y_i. V_i can be found as

$$V_i = \varphi A_i^{1/2} R(\alpha) A_i^{1/2}, \tag{4.4.2}$$

where φ is the dispersion parameter, $R(\alpha)$ is the working correlation matrix, and A_i is a T_ixT_i matrix with diagonal elements $v_{it} = \widehat{var}(y_{it})$. For binomial data, $v_{it} = p_{it}(1 - p_{it})$.

In addition, D_i is a $T_i$xp matrix (p represents the number of parameters or the length of β) and is found as

$$D_i = \frac{\partial \mu_i}{\partial \beta} = \left\{\frac{\partial \mu_{it}}{\partial p_{it}} \cdot \frac{\partial p_{it}}{\partial x_{it}}\right\} = p_{it}(1 - p_{it})x_{it}. \tag{4.4.3}$$

Finally, $\mu_{it} = E(y_i|x_{it}, \beta) = p_{it}$ where $logit(p_{it}) = \beta_0 + \beta_1 x_i + \beta_2(t-1) = x_{it}'\beta$. For survey data with design weights, the generalized estimating equations must be modified slightly. We include w_i as the design weight of subject i in our formula and we now find β to solve

$$\sum_{i=1}^{K} w_i D_i' V_i^{-1} (y_i - \mu_i) = 0. \tag{4.4.4}$$

Here we are simply applying the weight of w_i to the effect of subject i on the estimating equations. As, by the nature of survey design weights, subject i is representing w_i subjects instead of just 1. The estimation of the dispersion parameter φ and the working correlation matrix $R(\alpha)$ may be altered as well. These modifications will be illustrated later.

4.4.2 Dispersion parameter estimates for generalized estimating equations analysis

While in the simulation study, dispersion was controlled and thus no dispersion parameter was necessary, it may have an influence on this survey data. It is important to understand the full impact of dispersion on the working correlation matrix and ultimately on the final results.

When sampling weights are not considered, the dispersion parameter is estimated

by

$$\hat{\varphi} = \frac{1}{N-p}\sum_{i=1}^{K}\sum_{t=1}^{T_i} {e_{it}}^2, \tag{4.4.5}$$

where N is total number of observations: $N = \sum_{i=1}^{K} T_i$, p is the number of parameters or degrees of freedom, K is the number of subjects, and T_i the number of observations for subject i. e_{it} represents the Pearson residual for subject i at time t:

$$e_{it} = \frac{y_{it} - \mu_{it}}{\sqrt{v(y_{it})}}. \tag{4.4.6}$$

For binary data, this can be considered as

$$e_{it} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1-p_{it})}}. \tag{4.4.7}$$

Under the context of the National Population Health Survey, sampling weights are needed to accurately analyze the data and so we adjust this formula to account for the fact that each respondent, i, is accounting for w_i population units. Therefore, within a survey context, the dispersion parameter is estimated by

$$\hat{\varphi} = \frac{1}{N-p}\sum_{i=1}^{K} w_i \sum_{t=1}^{T_i} {e_{it}}^2, \tag{4.4.8}$$

where $N = \sum_{i=1}^{K} w_i T_i$.

For each fitted model, the dispersion parameter will be different since it is based on

$$p_{it} = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}. \tag{4.4.9}$$

The following table outlines the estimated dispersion parameters for GEE models based on different working correlation structures using the same covariates as was used during logistic regression.

Table 4.4.3 Dispersion parameter estimates for generalized estimating equations solving longitudinal marginal probability of having high blood pressure by correlation structure

Model	Dispersion parameter estimate
GEE-Independent	0.8976311
GEE-Exchangeable	0.8484257
GEE-Serial	0.8423306

Source: National Population Health Survey (1994/95-2008/09)

Since the assumption with generalized linear models is that the dispersion parameter for binomial data is 1, we witness underdispersion which can signal that variance is actually less than estimated for these models. Fortunately, they are somewhat close to 1 which provides less reason to be alarmed. This analysis is conducted, nonetheless, with a set ($\varphi = 1$) and estimated ($\varphi = \hat{\varphi}$) dispersion parameter to verify that this has minimal effect on our analysis.

4.4.3 Variance estimation for generalized estimating equations

It should be noted that variance estimates for generalized estimating equations (as is the case for all estimates from the National Population Health Survey) are derived using the bootstrap method as outlined in Appendix 4.

4.4.4 Results for independent working correlation structure

Now, we begin our analysis of generalized estimating equations by considering the independent working correlation matrix. Note that this is essentially equivalent to the MLE estimation for logistic regression since it ignores any interaction between observations from the same subject. Because of this, $R(\alpha)$ does not need to be estimated and is given as:

$$R(\alpha) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

The following table outlines the parameter estimates for GEE for independent working correlation using an estimated dispersion parameter as shown above:

Table 4.4.4 Generalized Estimating Equations: Parameter estimates for Independent correlation structure and estimated dispersion parameter

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-9.214	0.263	1224.520	0.000
Sex	0.235	0.072	10.729	0.001
Couple	-0.046	0.071	0.422	0.516
Unemployed	-0.027	0.121	0.050	0.824
Age in 1994/95	0.074	0.002	1031.412	0.000
Physically Active	-0.154	0.053	8.342	0.004
BMI	0.113	0.007	297.003	0.000
Urban	-0.046	0.071	0.418	0.518
Cycle	0.224	0.008	716.961	0.000

Source: National Population Health Survey (1994/95-2008/09)

The variables which are statistically significant for modelling high blood pressure include sex, age in 1994/95, physical activity, body mass index (BMI) and cycle. If we undertake the same analysis but with a fixed dispersion parameter (φ=1), we

note that for the independent working correlation structure, no difference exists between our estimates. This is due to the fact that we assumed independent correlation which means that dispersion plays no role in the working correlation matrix. The scalar effect on the variance function has no effect on our estimation since it can be factored out.

4.4.5 Results for exchangeable working correlation structure

Now, we shift our analysis to the real purpose of generalized estimating equations which is to incorporate the interactions between observations from the same subject into our analysis. We begin this analysis using the exchangeable correlation structure. Recall that this assumes that for all cycles $j \neq k$, the correlation coefficient is common: $\alpha_{jk} = \alpha$.

This requires us to estimate the working correlation matrix $R(\alpha)$. Without survey weights, we find $\hat{\alpha}$ such that

$$\hat{\alpha} = \frac{1}{\varphi(N^* - p)} \sum_{i=1}^{N} \sum_{j<k} e_{ij} e_{ik}, \tag{4.4.10}$$

where

$$e_{ij} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1 - p_{it})}}, \tag{4.4.11}$$

and

$$N^* = 0.5 \sum_{i=1}^{N} T_i(T_i - 1), \tag{4.4.12}$$

where p is the number of parameters or length of β and T_i represents the number of observations for subject i. For the dropout model, $T_i = m_i$. To account for survey weights, we modify these formulae slightly to incorporate our design weight w_i:

$$\hat{\alpha} = \frac{1}{\varphi(N^* - p)} \sum_{i=1}^{N} w_i \sum_{j<k} e_{ij} e_{ik}, \tag{4.4.13}$$

where

$$e_{ij} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1 - p_{it})}}, \tag{4.4.14}$$

and

$$N^* = 0.5 \sum_{i=1}^{N} w_i T_i (T_i - 1). \tag{4.4.15}$$

Our working correlation matrix is therefore estimated by:

$$R(\alpha) = \begin{bmatrix} 1 & \hat{\alpha} & \cdots & \hat{\alpha} \\ \hat{\alpha} & 1 & \cdots & \hat{\alpha} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha} & \hat{\alpha} & \cdots & 1 \end{bmatrix}$$

Table 4.4.5 Generalized Estimating Equations: Parameter estimates for Exchangeable correlation structure and estimated dispersion parameter

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.374	0.249	1126.900	0.000
Sex	0.208	0.072	8.344	0.004
Couple	0.022	0.057	0.148	0.700
Unemployed	-0.179	0.085	4.443	0.035
Age in 1994/95	0.073	0.002	1226.341	0.000
Physically Active	-0.021	0.031	0.485	0.486
BMI	0.083	0.006	172.803	0.000
Urban	-0.117	0.069	2.897	0.089
Cycle	0.225	0.008	809.778	0.000

Source: National Population Health Survey (1994/95-2008/09)

Here, we see some changes from the GEE using independence correlation structure. Notably, physical activity is no longer a significant factor in the model while

unemployment status (for the first time in our analysis – including the logistic regression by cycle) takes its place as a statistically significant effect. We observe that the more likely people are to be unemployed, the less likely they are to have high blood pressure. Sex, age in 1994/95, BMI, and cycle continue to be significant effects.

We now conduct the same analysis but using a fixed dispersion parameter set at 1 which would be the case if our variance function was perfectly specified and no under or overdispersion existed.

Table 4.5.6 Generalized Estimating Equations: Parameter estimates for Exchangeable correlation structure and fixed dispersion parameter (φ=1)

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.485	0.243	1222.148	0.000
Sex	0.218	0.069	9.864	0.002
Couple	0.012	0.056	0.043	0.836
Unemployed	-0.171	0.086	3.961	0.047
Age in 1994/95	0.073	0.002	1269.251	0.000
Physically Active	-0.030	0.031	0.904	0.342
BMI	0.087	0.006	203.722	0.000
Urban	-0.105	0.067	2.411	0.120
Cycle	0.225	0.008	812.428	0.000

Source: National Population Health Survey (1994/95-2008/09)

Unlike the independent case, for exchangeable correlation structure, we observe some minimal differences in our parameter estimates. Our overall results do not change as the same variables which are significant for the estimated dispersion parameter remain so for the fixed parameter. These changes are only due to the effect of the dispersion parameter in the working correlation matrix estimation. For the duration of our analysis, we will focus on the case where we take dispersion parameter fixed at 1 since its estimation causes minimal effect on our results.

4.4.6 Results for serial or autoregressive working correlation structure

Finally, we will run this analysis for GEE using serial correlation structure where the correlation between cycles j and k is given by $\alpha_{jk} = \alpha^{|j-k|}$.

This requires us to estimate the working correlation matrix $R(\alpha)$. Without survey weights, we find $\hat{\alpha}$ such that

$$\hat{\alpha} = \frac{1}{\varphi(N^* - p)} \sum_{i=1}^{N} \sum_{j<T_i} e_{ij} e_{i(j+1)}, \tag{4.4.16}$$

where

$$e_{ij} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1 - p_{it})}}, \tag{4.4.17}$$

and

$$N^* = \sum_{i=1}^{N} (T_i - 1). \tag{4.4.18}$$

where p is the number of parameters or length of β and T_i represents the number of observations for subject i. For the dropout model, $T_i = m_i$. To account for survey weights, we modify these formulae slightly to incorporate our design weight w_i:

$$\hat{\alpha} = \frac{1}{\varphi(N^* - p)} \sum_{i=1}^{N} w_i \sum_{j<T_i} e_{ij} e_{i(j+1)}, \tag{4.4.19}$$

where

$$e_{ij} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1 - p_{it})}}, \tag{4.4.20}$$

and

$$N^* = \sum_{i=1}^{N} w_i (T_i - 1). \tag{4.4.21}$$

Our working correlation matrix is therefore estimated by:

$$R(\alpha) = \begin{bmatrix} 1 & \hat{\alpha} & \hat{\alpha}^2 & \cdots & \hat{\alpha}^{|T_i-1|} \\ \hat{\alpha} & 1 & \hat{\alpha} & \cdots & \hat{\alpha}^{|T_i-2|} \\ \hat{\alpha}^2 & \hat{\alpha} & 1 & \cdots & \hat{\alpha}^{|T_i-3|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}^{|T_i-1|} & \hat{\alpha}^{|T_i-2|} & \hat{\alpha}^{|T_i-3|} & \cdots & 1 \end{bmatrix}$$

Table 4.4.7 Generalized Estimating Equations: Parameter estimates for Serial correlation structure and fixed dispersion parameter (φ=1)

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.322	0.245	1158.243	0.000
Sex	0.207	0.067	9.432	0.002
Couple	-0.025	0.054	0.220	0.639
Unemployed	-0.082	0.076	1.161	0.281
Age in 1994/95	0.072	0.002	1149.552	0.000
Physically Active	-0.040	0.028	2.079	0.149
BMI	0.083	0.006	183.628	0.000
Urban	-0.078	0.067	1.344	0.246
Cycle	0.222	0.008	837.495	0.000

Source: National Population Health Survey (1994/95-2008/09)

For this model, we observe that unemployment is once again a negligible effect in our model. Physical activity is again not significant while sex, age in 1994/95, BMI, and cycle continue to be significant factors. As is the case for the exchangeable working correlation, our overall results encounter ignorable change when we estimate the dispersion parameter. These changes are only due to the effect of the dispersion parameter in the working correlation matrix estimation.

4.5 Analysis of high blood pressure using weighted generalized estimating equations

As illustrated in the theory section (2.5) and reinforced in the simulation study, weighted generalized estimating equations can be used to properly account for any drop-out mechanism and eliminate the bias associated with ignoring missingness at random (MAR). Before we begin to perform our analysis of weighted generalized estimating equations, we attempt to get a better understanding of why respondents are dropping out of the study.

4.5.1 Formulation of survey weighted generalized estimating equations

Similar to regular generalized estimating equations, weighted generalized estimating equations look to find β to solve

$$\sum_{i=1}^{K} \frac{1}{v_{im}} D_i' V_i^{-1} (y_i - \mu_i) = 0. \tag{4.5.1}$$

V_i is an $m_i x m_i$ matrix representing the working covariance matrix of y_i. V_i can be found as

$$V_i = \varphi A_i^{1/2} R(\alpha) A_i^{1/2}, \tag{4.5.2}$$

where φ is the dispersion parameter, $R(\alpha)$ is the working correlation matrix, and A_i is an $m_i x m_i$ matrix with diagonal elements $v_{it} = \widehat{var}(y_{it})$. For binomial data, $v_{it} = p_{it}(1 - p_{it})$. In addition, D_i is an $m_i$xp matrix (p represents the number of parameters or the length of β) and is found as

$$D_i = \frac{\partial \mu_i}{\partial \beta} = \left\{ \frac{\partial \mu_{it}}{\partial p_{it}} \cdot \frac{\partial p_{it}}{\partial x_{it}} \right\} = p_{it}(1 - p_{it}) x_{it}, \tag{4.5.3}$$

and, $\mu_{it} = E(y_i|x_{it}, \beta) = p_{it}$, where $logit(p_{it}) = \beta_0 + \beta_1 x_i + \beta_2(t-1) = x_{it}'\beta$, and $v_{im} = p_i(drop - out\ at\ time\ m_i)$ is estimated prior to solving the weighted generalized estimating equations.

For survey data with design weights, the generalized estimating equations must be modified slightly. We include w_i as the design weight of subject i in our formula and we now find β to solve

$$\sum_{i=1}^{K} \frac{w_i}{v_{im}} D_i' V_i^{-1} (y_i - \mu_i) = 0. \tag{4.5.4}$$

Here we are simply applying the weight of w_i to the effect of subject i on the estimating equations. As, by the nature of survey design weights, subject i is representing w_i subjects instead of just 1. This formulation is of note since the scalar $\frac{w_i}{v_{im}}$ is the combined weight of subject i representing the design weight as well as the inverse probability of dropping out at m_i. The estimation of the dispersion parameter φ and the working correlation matrix $R(\alpha)$ may be altered as well. These modifications will be illustrated later.

4.5.2 Drop-out models for weighted generalized estimating equations analysis

To begin with, we need to find v_{im} for each subject. This is found through a simple logistic regression. We wish to find the probability of dropping out at time m_i. If the response at time t for respondent i is denoted by R_{it} (where 1 = response and 0 = non-response), then as illustrated in the theory section (2.5.2), let us define the conditional probability of missing time t given that all previous cycles were responded to as

$$logit(R_{it} = 0|R_{i1} = 1, \dots, R_{i(t-1)} = 1, \gamma, y_i, x_i) = \gamma' Y_{it}, \tag{4.5.5}$$

where we define Y_{it} based on selected variables which may influence the subject's propensity to drop-out of the study. If we consider the case used in the simulation study such that drop-out is based on the previous (MAR) value of y_i, then we find the joint probability of the response pattern (the unconditional probability of dropping out at time m_i) given by

(4.5.6)

$$
\begin{aligned}
&P(M_i = m_i | y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) \\
&\quad = P(R_{im_i} = 0, R_{i1} = 1, \dots, R_{i(m_i-1)} = 1 \,|y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) \\
&\quad = p(R_{i1} = 1 | y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) \\
&\quad \times p(R_{i2} = 1 | R_{i1} = 1, y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) \times \dots \\
&\quad \times p(R_{i(m_i-1)} = 1 | R_{i1} = 1, \dots R_{i(m_i-2)} = 1, y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) \\
&\quad \times p(R_{im_i} = 0 | R_{i1} = 1, \dots R_{i(m_i-1)} = 1, y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma).
\end{aligned}
$$

Note that since we are considering drop-outs,
$p(R_{it} = 0 | R_{i(t-1)} = 0, y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) = 1$. Therefore,

(4.5.7)

$$
\begin{aligned}
P(M_i = m_i | y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) &= \prod_{t=1}^{m_i-1} p(R_{it} = 1 | R_{i1} = 1, \dots, R_{i(t-1)} = 1, \gamma, y_i, x_i) \\
&\quad \times p(R_{it} = 0 | R_{i1} = 1, \dots, R_{i(t-1)} = 1, \gamma, y_i, x_i)^{I(m_i \le T)}.
\end{aligned}
$$

If we take $p(R_{it} = 0 | R_{i1} = 1, \dots, R_{i(t-1)} = 1, \gamma, y_i, x_i) = \pi_{it}$ then

$$
\pi_{it} = \frac{\exp(\gamma_0 + \gamma_1 y_{i(t-1)} + \gamma_2 y_{it})}{1 + \exp(\gamma_0 + \gamma_1 y_{i(t-1)} + \gamma_2 y_{it})}, \tag{4.5.8}
$$

and

(4.5.9)

$$
P(M_i = m_i | y_{i1}, y_{i2}, \dots, y_{im_i}, \gamma) = \left\{ \prod_{t=2}^{m_i-1} \frac{1}{1 + \exp(\gamma_0 + \gamma_1 y_{i(t-1)})} \right\} \times \left\{ \frac{\exp(\gamma_0 + \gamma_1 y_{i(m_i-1)})}{1 + \exp(\gamma_0 + \gamma_1 y_{i(m_i-1)})} \right\}^{I(m_i \le T+1)}.
$$

As shown in the theory section (2.5.2), to find v_{im}, we first find $\hat{\gamma}$ that maximizes the pseudolikelihood function:

$$L(\gamma) = \prod_{i=1}^{k}\left\{\prod_{t=2}^{m_i-1}\frac{1}{1+\exp\left(\gamma_0+\gamma_1 y_{i(t-1)}\right)}\right\} \times\left\{\frac{\exp\left(\gamma_0+\gamma_1 y_{i(m_i-1)}\right)}{1+\exp\left(\gamma_0+\gamma_1 y_{i(m_i-1)}\right)}\right\}^{I(m_i\leq T+1)}. \tag{4.5.10}$$

We then find

$$v_{im} = P\left(M_i = m_i | y_{i1}, y_{i2}, \ldots, y_{im_i}, \hat{\gamma}\right) \tag{4.5.11}$$
$$= \left\{\prod_{t=2}^{m_i-1}\frac{1}{1+\exp\left(\hat{\gamma}_0+\hat{\gamma}_1 y_{i(t-1)}\right)}\right\} \times\left\{\frac{\exp\left(\hat{\gamma}_0+\hat{\gamma}_1 y_{i(m_i-1)}\right)}{1+\exp\left(\hat{\gamma}_0+\hat{\gamma}_1 y_{i(m_i-1)}\right)}\right\}^{I(m_i\leq T+1)}.$$

Now, let us consider different drop-out models.

We start with a simple regression model with only 1 covariate: the previous response to the high blood pressure question. This model was used for the simulation study and is very important since it can indicate a direct relationship between the variable of interest (albeit from an earlier observation) and probability of dropping out.

Table 4.5.1 Drop-out model 1: Logistic regression measuring probability of dropping out depending only on previous value of high blood pressure

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-2.151	0.023	9005.591	0.000
HBP (previous value)	0.100	0.052	3.681	0.055

Source: National Population Health Survey (1994/95-2008/09)

We witness a positive relationship between the previous value of high blood pressure and probability of dropping out indicating that individuals who have high blood pressure may be more likely to drop out of the study than those without. This is an important observation since it implies that ignoring the drop-out mechanism may result in underestimating the incidence of high blood pressure. We note, however, that at 95% confidence, we cannot consider the previous value of high blood pressure to be a significant effect (while at 90% confidence it would be statistically significant).

A slightly different approach to this problem is to model the relationship between the covariates and the probability of dropping out.

Table 4.5.2 Drop-out model 2: Logistic regression measuring probability of dropping out depending on previous values of covariates

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-1.291	0.123	109.786	0.000
BMI (previous value)	-0.033	0.004	60.783	0.000
Physically Active (previous value)	-0.192	0.041	22.382	0.000
Unemployed (previous value)	0.463	0.093	24.885	0.000
Couple (previous value)	-0.527	0.043	147.662	0.000
Age in 1994/95	0.008	0.001	57.905	0.000
Sex	-0.099	0.039	6.487	0.011
Urban	0.085	0.041	4.317	0.038
Cycle (previous value)	0.009	0.010	0.773	0.379

Source: National Population Health Survey (1994/95-2008/09)

We observe that some covariates seem to have very strong associations with the drop-out mechanism. These include BMI, physical activity, unemployed status, and marital status. In fact, every covariate except for cycle is statistically significant in relation to the probability of dropping out.

We know from the Canadian Census of Population that single / unattached individuals are significantly more mobile than those in marriages. This is, in part, related to age as younger individuals (who are less likely to be married) are very mobile as they move to attend school or begin to establish careers. Unemployed individuals may also be more mobile since they are not committed to a job in a specific place. Mobility increases the risks associated with respondent "drop out" since it becomes difficult for Statistics Canada to locate the respondent for future waves. Fortunately, neither unemployment status nor marital status are significant effects in our model of interest.

BMI and Physical activity are interesting characteristics affecting the drop-out mechanism. It's not immediately clear as to why they play a role from a survey methods perspective. They could be subject-matter specific effects and since they do tend to influence the incidence of high blood pressure, they have an impact on our model. It is also of note that women are less likely than men to drop-out and since women are more likely to have high blood pressure, ignoring this effect could impact our variable of interest or its association with sex.

Finally, we combine the above 2 models to find perhaps the most accurate but also most complex drop-out model.

Table 4.5.3 Drop-out model 3 parameter estimates: Logistic regression measuring probability of dropping out depending on previous values of covariates and high blood pressure

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-1.265	0.126	100.082	0.000
HBP (previous value)	0.051	0.056	0.817	0.366
BMI (previous value)	-0.034	0.004	60.829	0.000
Physically Active (previous value)	-0.191	0.041	22.176	0.000
Unemployed (previous value)	0.464	0.093	24.890	0.000
Couple (previous value)	-0.525	0.044	144.149	0.000
Age in 1994/95	0.008	0.001	46.348	0.000
Sex	-0.101	0.039	6.684	0.010
Urban	0.086	0.041	4.357	0.037
Cycle (previous value)	0.008	0.010	0.577	0.447

Source: National Population Health Survey (1994/95-2008/09)

We see that the effect of the previous value for high blood pressure is lessened by the inclusion of the covariates which are accounting for certain effects related to the probability of dropping out. For the covariates, their coefficients in this model are very similar to those in the previous model.

For weighted generalized estimating equations, we weight our original GEE model based on the inverse probability of the individual respondent's response pattern (which is essentially the probability of responding for all times prior to their dropout time and dropping out when they did). Therefore, these probabilities are found using the fitted drop-out model (selected from above) and are then inversed to be used as weights in the GEE model.

Before proceeding, we examine if there are large differences in the results depending on the type of drop-out model selected. For the purposes of the following set of tables, Model 1 refers to the model including only the previous value

of high blood pressure, Model 2 refers to the model including the covariates, and Model 3 refers to the model including the previous value of high blood pressure in addition to the covariates. These tables outline the parameter coefficient estimates for the weighted generalized estimating equations using these different drop-out mechanisms.

Table 4.5.4 Comparison of weighted generalized estimating equations parameter estimates for Independent correlation structure and estimated dispersion parameter by drop-out model

Parameter	Model 1	Model 2	Model 3
Intercept	-8.92885102	-8.99865975	-9.00363387
BMI	0.10680164	0.10793756	0.10750336
Physically Active	-0.05047309	-0.03695937	-0.03949957
Unemployed	-0.16446695	-0.13349809	-0.12872097
Couple	-0.01336394	-0.04464297	-0.05324653
Age in 1994/95	0.07065820	0.07260673	0.07269442
Sex	0.36304676	0.31484902	0.31741220
Urban	-0.10196044	-0.09514666	-0.09408188
Cycle	0.21534951	0.21404401	0.21630712

Source: National Population Health Survey (1994/95-2008/09)

Table 4.5.5 Comparison of weighted generalized estimating equations parameter estimates for Exchangeable correlation structure and estimated dispersion parameter by drop-out model

Parameter	Model 1	Model 2	Model 3
Intercept	-8.011317088	-8.06498830	-8.05550194
BMI	0.076550578	0.07670640	0.07615259
Physically Active	0.032166153	0.03620356	0.03435484
Unemployed	-0.178606131	-0.18922217	-0.18327732
Couple	0.000637679	-0.02748793	-0.03062682
Age in 1994/95	0.069060192	0.07152316	0.07142406
Sex	0.346735346	0.29983348	0.30380833
Urban	-0.156931916	-0.16845456	-0.16736080
Cycle	0.211858716	0.21603335	0.21460388

Source: National Population Health Survey (1994/95-2008/09)

Table 4.5.6 Comparison of weighted generalized estimating equations parameter estimates for Serial correlation structure and estimated dispersion parameter by drop-out model

Parameter	Model 1	Model 2	Model 3
Intercept	-7.64452276	-7.72503568	-7.72136156
BMI	0.06743783	0.06774706	0.06721850
Physically Active	0.01228412	0.01927207	0.01733783
Unemployed	-0.12936633	-0.12979174	-0.12252159
Couple	-0.04361290	-0.05635803	-0.06032018
Age in 1994/95	0.06725460	0.06981269	0.06977220
Sex	0.30933074	0.26292442	0.26781167
Urban	-0.16046970	-0.16797883	-0.16612178
Cycle	0.21193003	0.21660781	0.21577059

Source: National Population Health Survey (1994/95-2008/09)

We observe minimal difference between the parameter estimates for the different models. In particular, the estimates for those variables which have tended to be significant effects in the unweighted GEE models and logistic regression (BMI, physical activity, age in 1994/95, sex, and cycle) are similar for the different drop-

out mechanisms. This illustrates that any of the above drop-out models could be used and would give similar results.

Therefore, for simplicity and since it was widely used in the simulation study, we will use drop-out model 1 which is based on the previous value of our variable of interest, high blood pressure. In addition, we will use drop-out model 3 which is more complex but better account for covariate associations with non-response. We will compare the results from using these different drop-out mechanisms.

4.5.3 Dispersion parameter estimates for weighted generalized estimating equations analysis

The following table outlines the estimated dispersion parameters for GEE models based on different working correlation structures using the same covariates as was used during logistic regression. These estimates are found using drop-out model 1.

Table 4.5.7 Dispersion parameter estimates for weighted generalized estimating equations solving longitudinal marginal probability of having high blood pressure by correlation structure

Model	Dispersion parameter estimate
WGEE-Independent	0.8739663
WGEE-Exchangeable	0.8401887
WGEE-Serial	0.8420009

Source: National Population Health Survey (1994/95-2008/09)

As we witnessed with the unweighted generalized estimating equations, since the estimated dispersion parameters are sufficiently close to 1, applying our analysis with an estimated or fixed (at 1) dispersion parameter has minimal impact on our coefficient estimates.

4.5.4 Variance estimation for weighted generalized estimating equations

It should be noted that variance for the weighted generalized estimating equations (as is the case for all estimates from the National Population Health Survey) are derived using the bootstrap method as described in Appendix 4. The fact that estimation of the drop-out mechanism is required before the weighted GEE itself, introduces additional variability. To account for this, each bootstrap sample incorporates both estimation processes so that variance of final estimates for β reflect the total variance associated with this method.

4.5.5 Results for independent working correlation structure

Now, we begin our analysis of weighted generalized estimating equations by considering the independent working correlation matrix. As shown in the previous section, this ignores any interaction between observations from the same subject. Because of this, $R(\alpha)$ does not need to be estimated and is given as:

$$R(\alpha) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

We first consider the simplest drop-out model, Model 1, which depends only on the previous values for High Blood Pressure.

Table 4.5.8 Weighted Generalized Estimating Equations: Parameter estimates for Independent correlation structure, fixed dispersion parameter (φ=1), and drop-out model 1

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.929	0.305	856.558	0.000
Sex	0.363	0.089	16.532	0.000
Couple	-0.013	0.088	0.023	0.879
Unemployed	-0.164	0.167	0.969	0.325
Age in 1994/95	0.071	0.003	696.218	0.000
Physically Active	-0.050	0.064	0.624	0.430
BMI	0.107	0.008	194.275	0.000
Urban	-0.102	0.093	1.210	0.271
Cycle	0.215	0.012	325.851	0.000

Source: National Population Health Survey (1994/95-2008/09)

While the same covariates are significant effects (sex, age in 1994/95, BMI, and cycle) as in the unweighted case, we do witness some interesting differences. For instance, in the weighted model, sex is a much stronger variable (coefficient increases from 0.235 to 0.363). This implies that women who have high blood pressure are not well represented in the unweighted model since the weighting added by the response mechanism has increased this relationship. While physical activity was a significant effect in the unweighted model, this is no longer the case once the drop-out mechanism is accounted for. As was the case for the unweighted generalized estimating equations, we observe that there is no difference between our estimates using an estimated or a fixed dispersion parameter.

Considering the more complex drop-out model, Model 3, where we account for the previous value of high blood pressure as well as previous values for the covariates, we run our analysis again.

Table 4.5.9 Weighted Generalized Estimating Equations: Parameter estimates for Independent correlation structure, fixed dispersion parameter (φ=1), and drop-out model 3

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-9.004	0.306	867.799	0.000
Sex	0.317	0.091	12.288	0.000
Couple	-0.053	0.090	0.348	0.555
Unemployed	-0.129	0.164	0.617	0.432
Age in 1994/95	0.073	0.003	731.858	0.000
Physically Active	-0.039	0.064	0.380	0.538
BMI	0.108	0.008	191.681	0.000
Urban	-0.094	0.095	0.990	0.320
Cycle	0.216	0.012	328.631	0.000

Source: National Population Health Survey (1994/95-2008/09)

While we do witness some parameter estimate changes with the more complex drop-out model. The differences are fairly negligible and do not result in any change in which variables are considered significant effects. We do see the effect for sex decrease (albeit minor) implying that we may be overestimating this effect when using the simpler drop-out model. Once again, for independent working correlation, no difference exists between parameter estimates for an estimated or fixed dispersion parameter.

4.5.6 Results for exchangeable working correlation structure

Now, we conduct our analysis using the exchangeable correlation structure. Recall that this assumes that for all cycles $j \neq k$, the correlation coefficient is common: $\alpha_{jk} = a$ where we estimate the correlation parameter a as

$$\hat{\alpha} = \frac{1}{\varphi(N^* - p)} \sum_{i=1}^{N} w_i \sum_{j<k} e_{ij} e_{ik}, \tag{4.5.12}$$

where

$$e_{ij} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1 - p_{it})}}, \tag{4.5.13}$$

and

$$N^* = 0.5 \sum_{i=1}^{N} w_i m_i (m_i - 1). \tag{4.5.14}$$

Our working correlation matrix is therefore estimated by:

$$R(\alpha) = \begin{bmatrix} 1 & \hat{\alpha} & \cdots & \hat{\alpha} \\ \hat{\alpha} & 1 & \cdots & \hat{\alpha} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha} & \hat{\alpha} & \cdots & 1 \end{bmatrix}$$

We first consider the simplest drop-out model, Model 1 which depends only on the previous values for High Blood Pressure.

Table 4.5.10 Weighted Generalized Estimating Equations: Parameter estimates for Exchangeable correlation structure, fixed dispersion parameter (φ=1), and drop-out model 1

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.147	0.295	765.281	0.000
Sex	0.348	0.088	15.549	0.000
Couple	-0.004	0.069	0.003	0.956
Unemployed	-0.176	0.123	2.063	0.151
Age in 1994/95	0.069	0.002	828.998	0.000
Physically Active	0.026	0.043	0.353	0.552
BMI	0.081	0.008	107.347	0.000
Urban	-0.147	0.090	2.656	0.103
Cycle	0.212	0.010	455.435	0.000

Source: National Population Health Survey (1994/95-2008/09)

There are some notable differences between the results for weighted and unweighted GEE for the exchangeable working correlation structure. Sex, age in 1994/95, BMI, and cycle continue to be statistically significant effects. However, while unemployed status was a significant effect in the unweighted model, that is no longer the case (owing to an increase in variability – the coefficient is almost the same). As is the case for the independent correlation case, sex is a much stronger variable in the weighted model (coefficient increases from 0.218 to 0.348). Physical activity is still not a significant effect. As was the case for unweighted generalized estimating equations, we observe minimal differences in the parameter estimates where the dispersion parameter is either estimated or fixed. Changes are due to the impact of the dispersion parameter on working correlation matrix estimation.

We conduct our analysis using the more complex drop-out model which depends on the previous value of high blood pressure as well as the previous values for our covariates.

Table 4.5.11 Weighted Generalized Estimating Equations: Parameter estimates for Exchangeable correlation structure, fixed dispersion parameter (φ=1), and drop-out model 3

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.214	0.320	659.299	0.000
Sex	0.304	0.090	11.328	0.001
Couple	-0.037	0.076	0.245	0.621
Unemployed	-0.178	0.129	1.914	0.166
Age in 1994/95	0.072	0.003	767.160	0.000
Physically Active	0.028	0.045	0.390	0.532
BMI	0.081	0.008	93.467	0.000
Urban	-0.155	0.094	2.724	0.099
Cycle	0.215	0.010	456.301	0.000

Source: National Population Health Survey (1994/95-2008/09)

Similar to the independent correlation case, we observe minimal differences in our estimates. Once again the parameter for sex has declined indicating that the simple model may overestimate the effect of sex on high blood pressure. Once again, we observe minimal differences in the parameter estimates where the dispersion parameter is either estimated or fixed. Changes are due to the impact of the dispersion parameter on working correlation matrix estimation.

4.5.7 Results for serial or autoregressive working correlation structure

Finally, we will run this analysis for weighted GEE using serial correlation structure where the correlation between cycles j and k is given by $\alpha_{jk} = \alpha^{|j-k|}$.

This requires us to estimate the working correlation matrix $R(\alpha)$. Recall, we find $\hat{\alpha}$ such that

$$\hat{\alpha} = \frac{1}{\varphi(N^* - p)} \sum_{i=1}^{N} w_i \sum_{j<T_i} e_{ij} e_{i(j+1)}, \tag{4.5.15}$$

where

$$e_{ij} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1 - p_{it})}}, \tag{4.5.16}$$

and

$$N^* = \sum_{i=1}^{N} w_i (T_i - 1). \tag{4.5.17}$$

Our working correlation matrix is therefore estimated by:

$$R(\alpha) = \begin{bmatrix} 1 & \hat{\alpha} & \hat{\alpha}^2 & \cdots & \hat{\alpha}^{|T_i-1|} \\ \hat{\alpha} & 1 & \hat{\alpha} & \cdots & \hat{\alpha}^{|T_i-2|} \\ \hat{\alpha}^2 & \hat{\alpha} & 1 & \cdots & \hat{\alpha}^{|T_i-3|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}^{|T_i-1|} & \hat{\alpha}^{|T_i-2|} & \hat{\alpha}^{|T_i-3|} & \cdots & 1 \end{bmatrix}$$

We first consider the simplest drop-out model, Model 1 which depends only on the previous values for High Blood Pressure.

Table 4.5.12 Weighted Generalized Estimating Equations: Parameter estimates for Serial correlation structure, fixed dispersion parameter (φ=1), and drop-out model 1

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.041	0.296	737.290	0.000
Sex	0.323	0.086	14.080	0.000
Couple	-0.037	0.070	0.274	0.601
Unemployed	-0.136	0.118	1.339	0.247
Age in 1994/95	0.068	0.002	774.633	0.000
Physically Active	0.005	0.046	0.011	0.916
BMI	0.080	0.008	98.764	0.000
Urban	-0.144	0.090	2.546	0.111
Cycle	0.213	0.011	364.128	0.000

Source: National Population Health Survey (1994/95-2008/09)

Similar to the independent correlation case, while the same covariates are significant effects (sex, age in 1994/95, BMI, and cycle) as in the unweighted case, we do witness some interesting differences. As is the case for the other 2 correlation structures, sex is a much stronger variable (coefficient increases from 0.207 to 0.309). Physical activity is still not a significant effect. The effects of using a fixed or estimated dispersion parameter are again minimal on our model estimates.

We conduct our analysis using the more complex drop-out model which depends on the previous value of high blood pressure as well as the previous values for our covariates.

Table 4.5.13 Weighted Generalized Estimating Equations: Parameter estimates for Serial correlation structure, fixed dispersion parameter (φ=1), and drop-out model 3

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.136	0.299	741.898	0.000
Sex	0.282	0.088	10.173	0.001
Couple	-0.062	0.074	0.683	0.409
Unemployed	-0.124	0.128	0.935	0.334
Age in 1994/95	0.071	0.003	783.377	0.000
Physically Active	0.011	0.048	0.050	0.823
BMI	0.080	0.008	100.127	0.000
Urban	-0.147	0.091	2.579	0.108
Cycle	0.216	0.011	370.350	0.000

Source: National Population Health Survey (1994/95-2008/09)

As we've already witnessed with the other 2 correlation structures, we observe minimal differences in our estimates using the more sophisticated drop-out model. Once again the parameter for sex has declined indicating that the simple model may overestimate the effect of sex on high blood pressure. The urban indicator becomes a more important effect in the model (although we would still reject it if we required 95% confidence). Similar to the case with the simple drop-out model, the use of the dispersion parameter has little effect on the results of our analysis.

4.6 Summary of analysis of high blood pressure

Throughout this analysis, we are interested in determining the factors which influence hypertension propensity in adults (Canadians aged 15+) in Canada in 1994/95. From a longitudinal perspective, we are capable of understanding how effects can change over time.

Not surprisingly, Body Mass Index (BMI) was a consistent positive effect influencing high blood pressure. This was true for each individual cycle as well as in the longitudinal analysis. This effect was maintained through different approaches to correlation structures and non-response weighting methods. The same was true for age and sex. Older individuals and women were more likely to develop high blood pressure than their demographic counterparts and this was consistent in our analysis. Physical activity was not a consistent factor. While for some individual cycles, this was a significant effect in our logistic regression model, high blood pressure variance was better explained by the other covariates in the longitudinal analysis. One of these covariates which played a strong role was the cycle. Treated as a continuous variable, this effect demonstrated that over time, people in the study became more and more likely to have hypertension. The other variables, which were included only as controls, were not significant effects in our model.

From a data quality perspective, we observed that missingness in the study was not completely at random as we were able to construct drop-out models with significant effects. Ignoring the drop-out mechanism would result in underestimating the effect of sex on our variable of interest. The models fit well and bootstrap variance was sufficient to provide Wald tests for significance for parameter estimates.

Of the models considered, we examined the use of an estimated or fixed dispersion parameter, different working correlation structures, and different drop-out models. We determined that the estimation of the dispersion parameter was unnecessary

since it had minimal impact on our results and since it was near enough to 1. We know from the simulation study that if the data followed a serial correlation structure (which was similar to the true correlation structure), then our results may not be as accurate if we assumed anything different such as an independent working correlation structure. Deciding which drop-out mechanism to use is a more difficult choice. On one hand, our "saturated" drop-out model (which included all covariates as well as the previous value of high blood pressure – also referred to as "drop-out model 3") will inevitably provide more accurate weights than the simpler model using only the previous value of high blood pressure. However, this comes at the price of the additional estimation required for each of the supplementary covariates in the drop-out model hence increasing the variability of the overall weighted generalized estimating equations model. Considering that the individual parameter estimates did not suffer great losses in terms of efficiency from the use of the "saturated" drop-out model, it is recommended to use the more detailed or "saturated" drop-out model since it will best reflect the nature of the drop-outs.

The value of undertaking longitudinal analysis cannot be overstated. It is important, however, to consider attrition as potentially informative to one's analysis. Otherwise, there is substantial risk (as illustrated in the simulation study) of producing biased results. Finally, it should be noted that we cannot explicitly determine if the drop-out mechanism was non-ignorable or NI (as opposed to missing at random or MAR). This analysis makes the assumption of MAR drop-out mechanism and if the missingness was in fact non-ignorable, we could not approach this problem in the way we did.

Chapter 5

Conclusions

Ideally, data sets would be complete and perfectly representative of the populations about which they infer. Practically, however, this is seldom the case. Surveys incur complete non-response as well as item non-response while even censuses and administrative data sources incur item non-response. If this missingness is driven by our variable of interest either directly (NI) or indirectly through other variables (MAR), then ignoring the missingness will lead to response or selection bias.

A simple framework for missing data problems where we are interested in the marginal distribution of y irrespective of the effect of the probability of response can be thought of as: $p(y) = \frac{p(y,R)}{p(R|y)} = p(y|R)\frac{p(R)}{p(R|y)}$. Essentially, it shows, as is also implied by the Horvitz-Thompson estimator, that we can find the marginal distribution of y by simply weighting the joint distribution of y and the response mechanism, R, (which can be thought of as simply the incomplete data) by the inverse probability of selection or response which can depend on y. Robins et al. and Fitzmaurice et al. applied response weights to generalized estimating equations to account for longitudinal non-response – specifically drop-outs. This theory was explored in depth throughout this thesis.

We observed during the simulation study that when data is complete or missingness is completely at random, unweighted generalized estimating equations provided unbiased parameter estimates regardless of the working covariance matrix (Benefit of the robustness of generalized estimating equations). When drop-outs were missing at random or had association with covariates / previous value of y, then unweighted generalized estimating equations (as well as maximum likelihood

estimation) were inherently biased. On the other hand, bias associated with this type of informative missingness was corrected using weighted generalized estimating equations (though at the cost of efficiency). This remained consistent for various simulation modifications. Finally, non-ignorable drop-outs (which cannot be specified properly using weighted generalized estimating equations) caused response bias for all methods explored in this thesis. It's clear that when data is missing, it is imperative to understand the extent to which that missingness is informative. If it is, any misspecification of the missing data model (such as assuming MCAR and using unweighted GEEs where data is missing MAR or NI) will lead to response bias.

The National Population Health Survey provided a concrete example of a longitudinal survey incurring attrition over time. Simply ignoring all records that do not have a response for all available cycles implies removing or wasting valuable data that was obtained on those respondents with incomplete records (not to mention increasing the risk associated with selection bias as the % of records removed from analysis increases). On the contrary, if we perform our analysis on a "drop-out" data set but ignore the nature of the missingness, we risk response bias as shown in the simulation study. While we cannot, in isolation, examine the possibility of drop-outs being non-ignorable for our variable of interest, high blood pressure, we did see associations between the drop-out mechanism and some covariates such as sex. This had an impact on our generalized estimating equations parameter estimates since sex is a significant factor in prevalence of high blood pressure.

The longitudinal analysis on the NPHS yielded interesting results. We observed that women are more likely than men to develop hypertension. As one ages or as one's body mass index (BMI – a measure of weight over height) increases, then their likelihood of having high blood pressure also increases. While we were interested in studying the effect of physical activity on our variable of interest, in marginal longitudinal analysis, it was not a significant effect.

Other potential areas of study would be interesting with this data set. This would include examining weighted generalized estimating equations for intermittent missingness (not just limited to drop-outs) as shown by Robins et al. (1995). This extension could increase the accuracy of our analysis as we would include more data points in our analysis (with drop-outs, we are forced to remove records from the study once they miss once even if they return again afterwards). The analysis could be modified to base associations on odds ratios rather than correlation coefficients as shown in Fitzmaurice et al. (1994). This change would be sensible considering our variable of interest is binary. Other simulations could also be conducted comparing weighted generalized estimating equations against alternative methods such as multiple imputation. Finally, exploring methods to identify or handle non-ignorable missingness would also be of interest.

It is important to continue to look for methods to improve the analysis of incomplete data. In practice, perfect complete data sets do not exist and assuming that selection or response is completely random can have dangerous repercussions analytically.

References

Agresti, A. (2002) *Categorical Data Analysis*, Second Edition (2002), John Wiley & Sons, Hoboken.

Allison, P. D. (2001) *Missing Data*, Sage Publications, Thousand Oaks.

Bahadur, R.T. (1961) A representation of the joint distribution of responses to n dichotomous items, *Studies in Item Analysis and Prediction*, pp. 158-168, Stanford University Press.

Besag, J. (1975) Statistical Analysis of Non-Lattice Data, *The Statistician*, Vol. 24, No. 3, pp. 179-195.

Besaj, J. (1977) Efficiency of pseudo-likelihood estimation for simple Gaussian fields, *Biometrika*, Vol. 64, pp. 616-618.

Carrillo, I., Kovacevic, M. and Wu, C. (2006) Analysis of longitudinal survey data with missing observations: An application of weighted GEE to the national longitudinal survey of children and youth (NLSCY), *Proceedings of the Survey Methods Section*, Statistical Society of Canada, London.

Carrillo, I.A., Chen, J. and Wu, C. (2010) The pseudo-GEE approach to the analysis of longitudinal surveys, *The Canadian Journal of Statistics*, Vol. 38, No. 4, pp. 540–554.

Carrillo-García, I.A. (2006) Analysis of longitudinal survey data with missing observations: An application of weighted GEE to the national longitudinal survey of children and youth (NLSCY), *Technical Report: MITACS/NPCDS Internship Program*, Statistics Canada.

Chen, J., and Shao, J. (2000) Nearest neighbor imputation for survey data, *Journal of Official Statistics*, Vol. 16, No. 2.

Diggle P.J., Heagerty P., Liang K.-Y., Zeger S.L. (2002) *Analysis of Longitudinal Data*, 2nd edition, Oxford University Press, New York.

Farrell, P.J. and Rogers-Stewart, K. (2008) Methods for Generating Longitudinally Correlated Binary Data, *International Statistical Review*, Vol. 76, No. 1, pp. 28–38.

Fitzmaurice, G.M. , Laird, N.M. and Ware, J.H. (2004) *Applied Longitudinal Analysis*, John Wiley & Sons, New York.

Fitzmaurice, G.M., Molenberghs, G. and Lipsitz, S. (1995) Regression Models for Longitudinal Binary Responses with Informative Drop-outs, *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 57, No. 4, pp. 691-704.

Hedeker, D. and Gibbons, R.D. (2006) *Longitudinal Data Analysis*, John Wiley & Sons, New York.

Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, Vol. 47, pp. 663-685.

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

Little, R.J. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models, Second Edition*, Chapman and Hall/CRC, Boca Raton.

Nelder, J.A. and Wedderburn, R.W. (1972) Generalized linear models, *Journal of the Royal Statistical Society Series A*, Vol. 135, No. 3, pp. 370–384.

Public Health Agency of Canada (2011) Hypertension, www.phac-aspc.gc.ca/cd-mc/cvd-mcv/hypertension-eng.php, Public Health Agency of Canada.

Rao, J.N.K. (2006) Bootstrap Methods for Analyzing Complex Sample Survey Data, *Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health*, Cat. 11-522-X, Statistics Canada

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*, Vol. 90, pp. 106–121.

Rubin, D.B. (1976) Inference and missing data, *Biometrika*, Vol. 63, pp. 581-592.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Sinha, S. K., Troxel, A. B., Lipsitz, S. R., Sinha, D., Fitzmaurice, G. M., Molenberghs, G. and Ibrahim, J. G. (2011) A bivariate pseudo-likelihood for incomplete longitudinal binary data with nonignorable non-monotone missingness. *Biometrics*, Vol. 67, pp. 1119-1126.

Statistics Canada (2010) *Data Dictionary - Master File: Longitudinal Square, National Population Health Survey (NPHS) - Cycle 1 to 8 (1994/1995 to 2008/2009)*, Statistics Canada.

Statistics Canada (2010) *Longitudinal Documentation, National Population Health Survey (NPHS) Household Component - Cycle 1 to 8 (1994/1995 to 2008/2009)*, Statistics Canada.

Statistics Canada (2011) *Health Indicators*, Cat. 82-221-X, Statistics Canada.

Statistics Canada (2011) High blood pressure, by age group and sex, *http://www40.statcan.gc.ca/l01/cst01/health03a-eng.htm*, Statistics Canada.

Wedderburn R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika*, Vol. 61, pp.439-47.

Wedderburn, R.W.M. (1976) On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models, *Biometrika*, Vol. 63, pp. 27-32.

Wilkins, K., Campbell, N.R.C., Joffres, M.R., McAlister, F.A., Nichol, M., Quach, S., Johansen, H.L., and Tremblay, M.S. (2010) Blood Pressure in Canadian Adults, *Health Reports*, Cat.. 82-003-X, Vol. 21, No. 1, Statistics Canada.

Appendix 1

List of tables and figures

Tables

Figures

Appendix 2

Analytical selection sample and corresponding weights

First, let us consider our primary variable of interest, y (incidence of high blood pressure). Since we will be handling non-response from the perspective of this variable later, we do not have to be as strict in its completeness.

For the drop-out model, we model our missing data mechanism based on the previous response or y_t depends on y_{t-1}. Therefore, for all t<=m (for all waves prior to drop-out), we must have valid responses for y_t. Since we define m based on the first time y is missing, we have valid responses for y_t for all t<=m. However, we must assume that a response was received for t=1 in order to model y_t on y_{t-1}. Therefore, under the drop-out model, the sole condition for selection based on response to the variable of primary interest is that a valid response must have been received for t=1.

The covariates encounter different requirements for completeness. First it is important to determine which covariates will be used in the analysis (in order minimize the exclusion of records on the basis of item non-response). Once the covariates have been selected, their requirement to have a response or not depends on how they are defined (for the first wave only or time-varying) and whether they are used to model the missing data mechanism.

For covariates that are only defined for the 1st wave, in order for the record to be useful in analysis, a valid value must be present for each covariate.

In this thesis, we focus on drop-outs with some time-varying covariates and other covariates only defined for the first cycle. This means that our initial selection requires:

1. A response for our variable of interest for the first cycle
2. A response for the covariates we only define at the first cycle
3. A response for each time-varying covariate for all cycles prior to the drop-out time (defined by the variable of interest)

Let us examine the number of records that will need to be removed from our analysis as a result of incomplete data. First, we restrict our analysis to the population aged 15 and above in 1994/95 which requires no reweighting since the original weights took age into account. This decision is based on the fact that some of the covariates were not asked of individuals aged under 15 (e.g. employment status). Next we consider incomplete covariate information which will lead to the exclusion of some records. The following tables outline the weighted non-response rates for each cycle using the initial design/sampling weights. After cycle 1, only those records that provide a valid response for high blood pressure are considered for these rates:

Table A.2.1 Weighted count and % of NPHS respondents aged 15+ in 1994/95 by validity of response and variable for Cycle 1

Cycle 1	Valid Response	Missing	Grand Total	% Valid
Unemployed	21,886,656	486,803	22,373,459	97.8%
Urban / Rural	21,713,307	660,152	22,373,459	97.0%
Sex	22,373,459	0	22,373,459	100.0%
Age in 1994/95	22,373,459	0	22,373,459	100.0%
Weight	21,244,394	1,129,065	22,373,459	95.0%
Height	21,463,509	909,950	22,373,459	95.9%
Physical activity	20,335,294	2,038,166	22,373,459	90.9%
Body Mass Index (BMI)	20,994,601	1,378,858	22,373,459	93.8%
Marital status	22,373,459	0	22,373,459	100.0%

Source: National Population Health Survey (1994/95-2008/09)

Table A.2.2 Weighted count and % of NPHS respondents aged 15+ in 1994/95, who provided a valid response to the high blood pressure question, by validity of response and variable for Cycle 2

Cycle 2	Valid Response	Missing	Grand Total	% Valid
Unemployed	19,992,359	420,508	20,412,867	97.9%
Urban / Rural	19,844,243	568,624	20,412,867	97.2%
Sex	20,412,867	0	20,412,867	100.0%
Age in 1994/95	20,412,867	0	20,412,867	100.0%
Weight	19,866,791	546,076	20,412,867	97.3%
Height	20,037,660	375,207	20,412,867	98.2%
Physical activity	19,748,318	664,549	20,412,867	96.7%
Body Mass Index (BMI)	19,626,936	785,931	20,412,867	96.1%
Marital status	20,412,867	0	20,412,867	100.0%

Source: National Population Health Survey (1994/95-2008/09)

Table A.2.3 Weighted count and % of NPHS respondents aged 15+ in 1994/95, who provided a valid response to the high blood pressure question, by validity of response and variable for Cycle 3

Cycle 3	Valid Response	Missing	Grand Total	% Valid
Unemployed	18,192,177	304,079	18,496,256	98.4%
Urban / Rural	18,026,070	470,186	18,496,256	97.5%
Sex	18,496,256	0	18,496,256	100.0%
Age in 1994/95	18,496,256	0	18,496,256	100.0%
Weight	18,037,378	458,879	18,496,256	97.5%
Height	18,175,171	321,085	18,496,256	98.3%
Physical activity	17,832,215	664,041	18,496,256	96.4%
Body Mass Index (BMI)	17,861,718	634,539	18,496,256	96.6%
Marital status	18,496,256	0	18,496,256	100.0%

Source: National Population Health Survey (1994/95-2008/09)

Table A.2.4 Weighted count and % of NPHS respondents aged 15+ in 1994/95, who provided a valid response to the high blood pressure question, by validity of response and variable for Cycle 4

Cycle 4	Valid Response	Missing	Grand Total	% Valid
Unemployed	16,117,339	301,890	16,419,228	98.2%
Urban / Rural	15,993,966	425,263	16,419,228	97.4%
Sex	16,419,228	0	16,419,228	100.0%
Age in 1994/95	16,419,228	0	16,419,228	100.0%
Weight	16,151,525	267,704	16,419,228	98.4%
Height	16,239,744	179,484	16,419,228	98.9%
Physical activity	15,693,821	725,407	16,419,228	95.6%
Body Mass Index (BMI)	15,939,727	479,501	16,419,228	97.1%
Marital status	16,419,228	0	16,419,228	100.0%

Source: National Population Health Survey (1994/95-2008/09)

Table A.2.5 Weighted count and % of NPHS respondents aged 15+ in 1994/95, who provided a valid response to the high blood pressure question, by validity of response and variable for Cycle 5

Cycle 5	Valid Response	Missing	Grand Total	% Valid
Unemployed	14,156,522	389,346	14,545,868	97.3%
Urban / Rural	14,215,112	330,756	14,545,868	97.7%
Sex	14,545,868	0	14,545,868	100.0%
Age in 1994/95	14,545,868	0	14,545,868	100.0%
Weight	14,282,224	263,644	14,545,868	98.2%
Height	14,374,223	171,645	14,545,868	98.8%
Physical activity	13,878,215	667,653	14,545,868	95.4%
Body Mass Index (BMI)	13,864,613	681,255	14,545,868	95.3%
Marital status	14,545,868	0	14,545,868	100.0%

Source: National Population Health Survey (1994/95-2008/09)

Table A.2.6 Weighted count and % of NPHS respondents aged 15+ in 1994/95, who provided a valid response to the high blood pressure question, by validity of response and variable for Cycle 6

Cycle 6	Valid Response	Missing	Grand Total	% Valid
Unemployed	13,453,771	271,542	12,912,528	104.2%
Urban / Rural	12,628,310	284,218	12,912,528	97.8%
Sex	12,912,528	0	12,912,528	100.0%
Age in 1994/95	12,912,528	0	12,912,528	100.0%
Weight	12,708,211	204,318	12,912,528	98.4%
Height	12,778,252	134,276	12,912,528	99.0%
Physical activity	12,458,184	454,344	12,912,528	96.5%
Body Mass Index (BMI)	12,342,652	569,876	12,912,528	95.6%
Marital status	12,912,528	0	12,912,528	100.0%

Source: National Population Health Survey (1994/95-2008/09)

Table A.2.7 Weighted count and % of NPHS respondents aged 15+ in 1994/95, who provided a valid response to the high blood pressure question, by validity of response and variable for Cycle 7

Cycle 7	Valid Response	Missing	Grand Total	% Valid
Unemployed	11,406,042	351,781	11,757,823	97.0%
Urban / Rural	11,519,054	238,769	11,757,823	98.0%
Sex	11,757,823	0	11,757,823	100.0%
Age in 1994/95	11,757,823	0	11,757,823	100.0%
Weight	11,527,120	230,703	11,757,823	98.0%
Height	11,595,884	161,939	11,757,823	98.6%
Physical activity	10,961,805	796,018	11,757,823	93.2%
Body Mass Index (BMI)	11,170,149	587,674	11,757,823	95.0%
Marital status	11,757,823	0	11,757,823	100.0%

Source: National Population Health Survey (1994/95-2008/09)

Table A.2.8 Weighted count and % of NPHS respondents aged 15+ in 1994/95, who provided a valid response to the high blood pressure question, by validity of response and variable for Cycle 8

Cycle 8	Valid Response	Missing	Grand Total	% Valid
Unemployed	10,044,141	339,461	10,383,602	96.7%
Urban / Rural	10,169,663	213,939	10,383,602	97.9%
Sex	10,383,602	0	10,383,602	100.0%
Age in 1994/95	10,383,602	0	10,383,602	100.0%
Weight	10,152,953	230,649	10,383,602	97.8%
Height	10,236,483	147,119	10,383,602	98.6%
Physical activity	9,833,597	550,005	10,383,602	94.7%
Body Mass Index (BMI)	9,934,599	449,003	10,383,602	95.7%
Marital status	10,383,602	0	10,383,602	100.0%

Source: National Population Health Survey (1994/95-2008/09)

Note: Urban/rural status, age in 1994/95, and sex are not time-varying covariates – only the first observation for these variables is used in analysis.

We see that the item response rates are fairly high for the records in scope by cycle. We will still need to remove those records where a missing value is ever present for any cycle in which a high blood pressure response was received (prior to the time of drop-out – defined by the first missing observation of high blood pressure).

Next, let us restrict our dataset to include only those records where a response was received for high blood pressure prior to the time of drop-out. This implies that we do not consider any auxiliary information received after the subject has "dropped out". It also implies, as noted above, that we must remove those subjects who did not have a valid response to the high blood pressure question in the first cycle. The following table shows the weighted count of all individuals aged 15+ in 1994/95 by time or cycle of drop-out:

Table A.2.9 Weighted count of NPHS respondents aged 15+ in 1994/95 by time of drop-out

Time of drop-out	Count	%
0	30,776	0.1%
1	1,929,816	8.6%
2	1,916,611	8.6%
3	2,077,028	9.3%
4	1,873,361	8.4%
5	1,633,340	7.3%
6	1,154,706	5.2%
7	1,374,220	6.2%
8	10,383,602	46.5%
Grand Total	22,342,683	100.0%

Source: National Population Health Survey (1994/95-2008/09)

If the time of drop-out is 0, this implies that they didn't even respond to the high blood pressure question on the first wave of the survey. So, we remove these individuals from our analysis as described above.

Finally, we conduct our selection of valid records for analysis. For the drop-out model, this implies removing those records where the first value of high blood pressure is missing and removing those records where covariate information is missing for any wave in which a valid high blood pressure response was recorded.

Table A.2.10 Weighted count of NPHS respondents aged 15+ in 1994/95 by completeness of covariates and initial value of high blood pressure

	Complete covariates	Missing covariates	Grand Total
Complete HBP	14,716,672	7,626,011	22,342,683
Missing first HBP	11,975	18,801	30,776
Grand Total	14,728,647	7,644,812	22,373,459

Source: National Population Health Survey (1994/95-2008/09)

Table A.2.11 Weighted % of NPHS respondents aged 15+ in 1994/95 by completeness of covariates and initial value of high blood pressure

	Complete covariates	Missing covariates	Grand Total
Complete HBP	65.8%	34.1%	99.9%
Missing first HBP	0.1%	0.1%	0.1%
Grand Total	65.8%	34.2%	100.0%

Source: National Population Health Survey (1994/95-2008/09)

So, we reduce our data set by about 34.2%. We could simply assume that this selection is completely at random (or the removal of these records would be Missing completely at random (MCAR). However, if there is a relationship between any of these variables and their likelihood to go missing, we would bias our analysis. Considering that 34.2% is a fairly large proportion of the records in scope, the risk for this bias is non-trivial. Therefore, we run a logistic regression to calculate the probability of each record being selected using variables from our eventual analytical model as covariates.

This logistic regression which is measuring $p(selection|sample) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$ will employ as covariates: Age in 1994/95, sex, urban/rural indicator, marital status at wave 1, physical activity at wave 1, employment status at wave 1, and high blood pressure at wave 1. Since some of these variables are incomplete (thus contributing to their selection or non-selection), we treat item non-response as a valid categorical value for each variable in question (e.g., the binary high blood pressure variable becomes a 3-category variable: 1. Has high blood pressure; 2. Does not have high blood pressure; 3. Non-response).

Table A.2.12 Selection model parameter estimates: Logistic regression where variable of interest is selected into analytical sample from survey respondent sample

Parameter	Estimate	Standard Error	Wald Chi-Square	P(>Chi-Square)
Intercept	-23.978	14.809	2.622	0.105
Age in 1994/95	0.002	0.000	3707.825	<.0001
Sex	-0.527	0.001	29738.815	<.0001
Couple1	-0.010	0.001	74.497	<.0001
Unemployed1=0	6.444	3.954	2.656	0.103
Unemployed1=1	6.120	3.954	2.396	0.122
Urban=0	6.296	3.455	3.321	0.068
Urban=1	6.391	3.455	3.422	0.064
HBP1=0	6.089	13.677	0.198	0.656
HBP1=1	5.867	13.677	0.184	0.668
Physical activity1=0	6.404	2.162	8.774	0.003
Physical activity1=1	6.593	2.162	9.299	0.002

Source: National Population Health Survey (1994/95-2008/09)

We see that, for the most part, the covariates in this model are not significant factors in determining selection. In fact, only age in 1994/95, sex, marital status for the first wave and physical activity for the first wave are significant covariates for $\alpha = 0.05$. While age in 1994/95, sex, and marital status have no missing values and are

therefore only influencing this model in a topic-based fashion, physical activity is affecting this model because of its missing values. Physical activity as a covariate is missing more values than any other variable in this study and as such is the most influential variable in determining which records are selected.

We consider the odds ratios estimated through this logistic regression for the categorical covariates. These indicate the ratio of the odds of this parameter being true for those in our selection against the odds of this parameter being true for those not in our selection.

Table A.2.13 Selection model odds ratio estimates: Logistic regression where variable of interest is selected into analytical sample from survey respondent sample

Parameter	Estimate	Lower 95%	Upper 95%
Sex (female)	0.591	0.589	0.592
Couple1	0.990	0.988	0.992
Unemployed1=0	>999.999	0.014	>999.999
Unemployed1=1	>999.999	0.010	>999.999
Urban=0	>999.999	0.264	>999.999
Urban=1	>999.999	0.290	>999.999
HBP1=0	>999.999	<0.001	>999.999
HBP1=1	>999.999	<0.001	>999.999
Physical activity1=0	>999.999	802.774	>999.999
Physical activity1=1	>999.999	969.579	>999.999

Source: National Population Health Survey (1994/95-2008/09)

We see that men and singles are more likely to be selected than women and individuals in couples (married or common-law). As expected we see a very large and significant (The upper and lower 95% estimates are on the same side of 1) odds ratio for the physical activity valid variables. This again, is indicating that if physical activity was missing, it was very indicative that the record would not be selected.

As described above, using the results from this model, we determine the predicted probability of selection for each record. We then remove all those records that are not selected and find our selection weight for those that have been selected within our sample as

$$\widehat{w}_i(selection|sample) = \frac{1}{\hat{p}_i(selection|sample)}. \qquad \text{(A.2.1)}$$

Appendix 3

Cross-sectional analysis of high blood pressure

A.3.1 Characteristics of Canadians with high blood pressure for NPHS cycles 2-7

Table A.3.1 NPHS Cycle 2 characteristics of Canadians aged 15+ in 1994/95 with high blood pressure by presence of high blood pressure

Does not have high blood pressure	Average	Lower 95%	Upper 95%
Female	51.4%	50.6%	52.2%
Lives in urban area	74.4%	73.0%	75.8%
Married or Common-Law	62.1%	60.9%	63.3%
Body Mass Index (BMI)	25.0	24.8	25.1
Weight (lbs)	158.7	157.8	159.6
Height (inches)	56.7	56.6	56.8
Unemployed	5.2%	4.5%	5.8%
Physically active or moderately active	42.7%	41.2%	44.2%
Age (in 1994/1995)	40.2	39.9	40.5
Has high blood pressure	**Average**	**Lower 95%**	**Upper 95%**
Female	62.1%	58.6%	65.6%
Lives in urban area	70.7%	67.3%	74.0%
Married or Common-Law	64.0%	60.7%	67.2%
Body Mass Index (BMI)	27.2	26.8	27.5
Weight (lbs)	166.4	163.7	169.2
Height (inches)	55.4	55.1	55.7
Unemployed	3.0%	1.5%	4.5%
Physically active or moderately active	33.5%	30.0%	37.1%
Age (in 1994/1995)	58.8	57.9	59.8

Source: National Population Health Survey (1994/95-2008/09)

Table A.3.2 NPHS Cycle 3 characteristics of Canadians aged 15+ in 1994/95 with high blood pressure by presence of high blood pressure

Does not have high blood pressure	Average	Lower 95%	Upper 95%
Female	50.7%	49.8%	51.6%
Lives in urban area	74.1%	72.6%	75.7%
Married or Common-Law	63.1%	61.8%	64.5%
Body Mass Index (BMI)	25.3	25.2	25.5
Weight (lbs)	161.1	160.1	162.0
Height (inches)	56.7	56.6	56.8
Unemployed	3.4%	2.8%	4.0%
Physically active or moderately active	48.0%	46.4%	49.7%
Age (in 1994/1995)	39.7	39.4	40.1
Has high blood pressure	**Average**	**Lower 95%**	**Upper 95%**
Female	62.2%	58.8%	65.5%
Lives in urban area	70.1%	66.4%	73.7%
Married or Common-Law	63.0%	59.5%	66.4%
Body Mass Index (BMI)	27.5	27.1	27.9
Weight (lbs)	168.4	165.7	171.1
Height (inches)	55.4	55.1	55.7
Unemployed	1.9%	0.9%	2.9%
Physically active or moderately active	38.0%	34.2%	41.8%
Age (in 1994/1995)	57.8	56.9	58.8

Source: National Population Health Survey (1994/95-2008/09)

Table A.3.3 NPHS Cycle 4 characteristics of Canadians aged 15+ in 1994/95 with high blood pressure by presence of high blood pressure

Does not have high blood pressure	Average	Lower 95%	Upper 95%
Female	51.2%	50.2%	52.2%
Lives in urban area	74.5%	72.9%	76.1%
Married or Common-Law	65.2%	63.7%	66.7%
Body Mass Index (BMI)	25.5	25.4	25.7
Weight (lbs)	162.1	161.1	163.1
Height (inches)	56.7	56.6	56.8
Unemployed	3.7%	3.0%	4.3%
Physically active or moderately active	44.4%	42.7%	46.2%
Age (in 1994/1995)	39.2	38.9	39.5
Has high blood pressure	**Average**	**Lower 95%**	**Upper 95%**
Female	62.9%	59.6%	66.2%
Lives in urban area	69.2%	65.8%	72.6%
Married or Common-Law	65.0%	61.6%	68.5%
Body Mass Index (BMI)	28.1	27.7	28.4
Weight (lbs)	171.1	168.5	173.7
Height (inches)	55.4	55.1	55.7
Unemployed	1.8%	0.9%	2.7%
Physically active or moderately active	34.6%	31.2%	38.0%
Age (in 1994/1995)	56.2	55.3	57.1

Source: National Population Health Survey (1994/95-2008/09)

Table A.3.4 NPHS Cycle 5 characteristics of Canadians aged 15+ in 1994/95 with high blood pressure by presence of high blood pressure

Does not have high blood pressure	Average	Lower 95%	Upper 95%
Female	51.0%	49.9%	52.2%
Lives in urban area	74.2%	72.6%	75.9%
Married or Common-Law	66.8%	65.2%	68.4%
Body Mass Index (BMI)	25.8	25.6	25.9
Weight (lbs)	164.1	163.0	165.3
Height (inches)	56.7	56.6	56.8
Unemployed	3.2%	2.6%	3.7%
Physically active or moderately active	53.1%	51.3%	55.0%
Age (in 1994/1995)	38.7	38.3	39.1
Has high blood pressure	**Average**	**Lower 95%**	**Upper 95%**
Female	62.0%	58.9%	65.0%
Lives in urban area	72.5%	69.4%	75.5%
Married or Common-Law	62.3%	58.9%	65.8%
Body Mass Index (BMI)	28.2	27.8	28.5
Weight (lbs)	171.7	168.9	174.4
Height (inches)	55.3	55.1	55.6
Unemployed	1.3%	0.7%	2.0%
Physically active or moderately active	42.5%	39.0%	46.0%
Age (in 1994/1995)	54.7	53.9	55.6

Source: National Population Health Survey (1994/95-2008/09)

Table A.3.5 NPHS Cycle 6 characteristics of Canadians aged 15+ in 1994/95 with high blood pressure by presence of high blood pressure

Does not have high blood pressure	Average	Lower 95%	Upper 95%
Female	51.1%	49.8%	52.4%
Lives in urban area	74.5%	72.8%	76.3%
Married or Common-Law	68.4%	66.7%	70.2%
Body Mass Index (BMI)	25.9	25.8	26.1
Weight (lbs)	165.2	164.0	166.4
Height (inches)	56.7	56.6	56.9
Unemployed	2.4%	2.0%	2.9%
Physically active or moderately active	50.1%	48.2%	52.0%
Age (in 1994/1995)	38.2	37.8	38.6
Has high blood pressure	**Average**	**Lower 95%**	**Upper 95%**
Female	59.7%	56.4%	63.0%
Lives in urban area	70.3%	67.2%	73.4%
Married or Common-Law	63.0%	59.7%	66.4%
Body Mass Index (BMI)	28.3	27.9	28.6
Weight (lbs)	173.7	170.8	176.6
Height (inches)	55.6	55.3	55.9
Unemployed	1.1%	0.5%	1.7%
Physically active or moderately active	43.2%	39.8%	46.6%
Age (in 1994/1995)	53.8	52.9	54.7

Source: National Population Health Survey (1994/95-2008/09)

Table A.3.6 NPHS Cycle 7 characteristics of Canadians aged 15+ in 1994/95 with high blood pressure by presence of high blood pressure

Does not have high blood pressure	Average	Lower 95%	Upper 95%
Female	51.0%	49.6%	52.4%
Lives in urban area	74.3%	72.5%	76.1%
Married or Common-Law	69.7%	68.0%	71.4%
Body Mass Index (BMI)	26.1	25.9	26.2
Weight (lbs)	165.7	164.5	167.0
Height (inches)	56.7	56.6	56.8
Unemployed	3.0%	2.3%	3.7%
Physically active or moderately active	56.2%	54.3%	58.2%
Age (in 1994/1995)	37.6	37.2	38.0
Has high blood pressure	**Average**	**Lower 95%**	**Upper 95%**
Female	57.3%	54.2%	60.4%
Lives in urban area	71.2%	68.2%	74.1%
Married or Common-Law	63.7%	60.4%	67.0%
Body Mass Index (BMI)	28.6	28.3	29.0
Weight (lbs)	176.9	174.1	179.7
Height (inches)	55.7	55.5	56.0
Unemployed	1.1%	0.6%	1.7%
Physically active or moderately active	48.8%	45.6%	52.0%
Age (in 1994/1995)	51.9	51.1	52.7

Source: National Population Health Survey (1994/95-2008/09)

A.3.2 Logistic regression analysis of high blood pressure by cycle

We run some simple logistic regression models to better understand the interactions between the covariates and our variable of interest. Here, we will ignore the longitudinality altogether and run separate logistic models for each cycle.

We model $p(y) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$ where the covariates, x, used include sex, marital status, unemployment status, age in 1994/95, physical activity, body mass index (BMI), and an urban/rural indicator. So we model for cycle t

$$logit(p(y_{it})) = \beta_0 + sex_{it}\beta_{sex} + age_{it}\beta_{age} + \cdots. \quad \text{(A.3.1)}$$

For all i with an observation at cycle t. For this analysis, marital status, employment status, physical activity and body mass index are time-varying (values exist for each cycle). Age, sex, and urban/rural indicator are always referring to the first cycle (1994/95). To determine which effects can be considered statistically significant, p-values are derived using Wald tests. These represent the probability of the coefficient being equal to 0 with Wald scores having a chi-squared distribution and are given by

$$Wald = \frac{(\hat{\theta} - \theta)^2}{var(\hat{\theta})} \sim \chi^2. \quad \text{(A.3.2)}$$

When a p-value is less than 0.05, then we can say that we have at least 95% confidence that the coefficient is not 0 and is therefore a statistically significant effect.

The following tables outline the results of these logistic regression models for each cycle.

Table A.3.7 Logistic regression parameter estimates: Probability of having high blood pressure in Cycle 1, 1994/95

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.418	0.403	437.056	0.000
Sex	0.191	0.088	4.690	0.030
Couple	-0.124	0.094	1.720	0.190
Unemployed	0.096	0.239	0.163	0.687
Age in 1994/95	0.069	0.003	559.484	0.000
Physically active	0.065	0.096	0.467	0.494
Body Mass Index (BMI)	0.101	0.012	75.283	0.000
Urban	-0.073	0.106	0.476	0.490

Source: National Population Health Survey (1994/95-2008/09)

For the first cycle of the NPHS, variables with a significant effect on the incidence of high blood pressure include sex, age in 1994/95, and body mass index (BMI). Level of physical activity was not a statistically significant effect. Body mass index proves to have a significantly positive relationship with incidence of high blood pressure where the higher the BMI, the more likely someone is to have HBP. We also note that women are more likely than men to have high blood pressure. The probability of having HBP increases as individuals get older.

Table A.3.8 Logistic regression parameter estimates: Probability of having high blood pressure in Cycle 2, 1996/97

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.401	0.333	638.092	0.000
Sex	0.308	0.099	9.746	0.002
Couple	-0.027	0.091	0.085	0.771
Unemployed	0.187	0.305	0.378	0.539
Age in 1994/95	0.072	0.003	664.458	0.000
Physically active	-0.104	0.091	1.320	0.251
Body Mass Index (BMI)	0.100	0.009	125.619	0.000
Urban	0.019	0.098	0.037	0.847

Source: National Population Health Survey (1994/95-2008/09)

Similar to the scenario for the first cycle of the NPHS, the second cycle variables with a significant effect on the incidence of high blood pressure include sex, age in 1994/95, and body mass index (BMI). Once again, level of physical activity was not a statistically significant effect. Similar relationships exist for the statistically significant variables as was the case for the first cycle. The primary difference is that sex has become a more considerable factor in determining incidence of high blood pressure (women more likely than men).

Table A.3.9 Logistic regression parameter estimates: Probability of having high blood pressure in Cycle 3, 1998/99

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.184	0.356	527.247	0.000
Sex	0.313	0.097	10.535	0.001
Couple	-0.046	0.101	0.209	0.647
Unemployed	0.136	0.292	0.219	0.640
Age in 1994/95	0.074	0.003	598.420	0.000
Physically active	-0.190	0.096	3.917	0.048
Body Mass Index (BMI)	0.099	0.009	119.756	0.000
Urban	-0.031	0.109	0.081	0.776

Source: National Population Health Survey (1994/95-2008/09)

If we define statistical significance at α =0.05 = 5% then, for the first time using this model, physical activity would be a statistically significant effect in determining probability to have high blood pressure. The less likely individuals are to be physically active, the more likely they are to have HBP. The other statistically significant effects are consistent with the first 2 cycles: sex, age in 1994/95, and body mass index (BMI). Similar relationships exist for these variables as was the case for the last cycle.

Table A.3.10 Logistic regression parameter estimates: Probability of having high blood pressure in Cycle 4, 2000/01

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.592	0.373	531.583	0.000
Sex	0.365	0.091	16.183	0.000
Couple	0.031	0.102	0.091	0.763
Unemployed	0.073	0.336	0.048	0.827
Age in 1994/95	0.076	0.003	599.905	0.000
Physically active	-0.189	0.092	4.241	0.039
Body Mass Index (BMI)	0.117	0.010	131.892	0.000
Urban	-0.145	0.093	2.438	0.118

Source: National Population Health Survey (1994/95-2008/09)

For the fourth cycle, the statistically significant effects are consistent with the third cycle: sex, age in 1994/95, body mass index (BMI), and physical activity. Similar relationships exist for these variables as was the case for the last cycle.

Table A.3.11 Logistic regression parameter estimates: Probability of having high blood pressure in Cycle 5, 2002/03

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.269	0.329	631.829	0.000
Sex	0.315	0.087	12.987	0.000
Couple	-0.075	0.103	0.530	0.467
Unemployed	-0.210	0.291	0.522	0.470
Age in 1994/95	0.077	0.003	609.412	0.000
Physically active	-0.165	0.094	3.048	0.081
Body Mass Index (BMI)	0.114	0.009	164.133	0.000
Urban	0.020	0.092	0.047	0.829

Source: National Population Health Survey (1994/95-2008/09)

If we define statistical significance at α =0.05 = 5% then, once again, physical activity would no longer be a statistically significant effect in determining

probability to have high blood pressure. The other statistically significant effects are consistent with the other cycles: sex, age in 1994/95, and body mass index (BMI).

Table A.3.12 Logistic regression parameter estimates: Probability of having high blood pressure in Cycle 6, 2004/05

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-8.105	0.365	494.457	0.000
Sex	0.199	0.099	4.027	0.045
Couple	-0.096	0.112	0.737	0.391
Unemployed	-0.215	0.335	0.411	0.522
Age in 1994/95	0.079	0.004	465.533	0.000
Physically active	0.002	0.090	0.001	0.981
Body Mass Index (BMI)	0.116	0.009	167.849	0.000
Urban	-0.109	0.089	1.503	0.220

Source: National Population Health Survey (1994/95-2008/09)

The statistically significant effects are consistent with the other cycles: sex, age in 1994/95, and body mass index (BMI). Physical activity is again not a statistically significant factor.

Table A.3.13 Logistic regression parameter estimates: Probability of having high blood pressure in Cycle 7, 2006/07

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-7.937	0.368	465.042	0.000
Sex	0.080	0.092	0.756	0.385
Couple	-0.083	0.113	0.544	0.461
Unemployed	-0.469	0.326	2.065	0.151
Age in 1994/95	0.077	0.003	545.866	0.000
Physically active	0.006	0.093	0.004	0.950
Body Mass Index (BMI)	0.123	0.009	181.916	0.000
Urban	-0.082	0.086	0.913	0.339

Source: National Population Health Survey (1994/95-2008/09)

For the first time, sex is not a statistically significant effect – meaning that we can no longer say that women are more likely than men to have high blood pressure. The other statistically significant effects are consistent with the other cycles: age in 1994/95 and body mass index (BMI). Physical activity is again not a statistically significant factor.

Table A.3.14 Logistic regression parameter estimates: Probability of having high blood pressure in Cycle 8, 2008/09

Parameter	Estimate	Standard error	Wald	P-value
Intercept	-7.787	0.376	430.058	0.000
Sex	0.051	0.097	0.278	0.598
Couple	0.072	0.110	0.429	0.513
Unemployed	-0.145	0.314	0.211	0.646
Age in 1994/95	0.079	0.004	455.038	0.000
Physically active	-0.425	0.104	16.813	0.000
Body Mass Index (BMI)	0.122	0.010	162.275	0.000
Urban	-0.010	0.094	0.010	0.919

Source: National Population Health Survey (1994/95-2008/09)

For the final and most recent cycle, physical activity returns as a statistically significant effect while sex is still not a significant effect. The other statistically significant effects are consistent with the other cycles: age in 1994/95 and body mass index (BMI).

We see that the effects influencing incidence of high blood pressure evolve over time. Age in 1994/95 and body mass index (BMI) are consistently statistically significant positive factors – an increase in either leads to a higher probability of HBP. While sex was a major effect for the first 6 cycles, it ceased to be a factor in the last 2 cycles. This could be due to the aging population where sex plays less of a role in determining probability of high blood pressure for older individuals than younger individuals. Physical activity is a significant effect for some cycles but not others and no linear pattern can be established reflecting its role in determining high blood pressure over time.

We consider the odds ratios estimated through this logistic regression for the categorical covariates. These indicate the ratio of the odds of this parameter being

true for with high blood pressure against the odds of this parameter being true for those without high blood pressure.

Table A.3.15 Logistic regression odds ratio estimates: Probability of having high blood pressure by cycle and covariate

Parameter	1994/ 95	1996/ 97	1998/ 99	2000/ 01	2002/ 03	2004/ 05	2006/ 07	2008/ 09
Sex	1.211	1.360	1.368	1.441	1.371	1.220	1.084	1.052
Couple	0.884	0.974	0.955	1.031	0.927	0.909	0.920	1.074
Unemployed	1.101	1.206	1.146	1.076	0.810	0.807	0.626	0.865
Physically active	1.067	0.901	0.827	0.827	0.848	1.002	1.006	0.654
Urban	0.929	1.019	0.970	0.865	1.020	0.897	0.921	0.990

Source: National Population Health Survey (1994/95-2008/09)

Table A.3.16 Upper 95% confidence interval for logistic regression odds ratio estimates: Probability of having high blood pressure by cycle and covariate

Parameter	1994/ 95	1996/ 97	1998/ 99	2000/ 01	2002/ 03	2004/ 05	2006/ 07	2008/ 09
Sex	1.439	1.650	1.653	1.722	1.627	1.482	1.299	1.273
Couple	1.063	1.165	1.163	1.259	1.136	1.131	1.148	1.332
Unemployed	1.759	2.193	2.030	2.079	1.433	1.555	1.186	1.603
Physically active	1.287	1.077	0.998	0.991	1.020	1.196	1.207	0.801
Urban	1.144	1.235	1.199	1.038	1.221	1.068	1.090	1.192

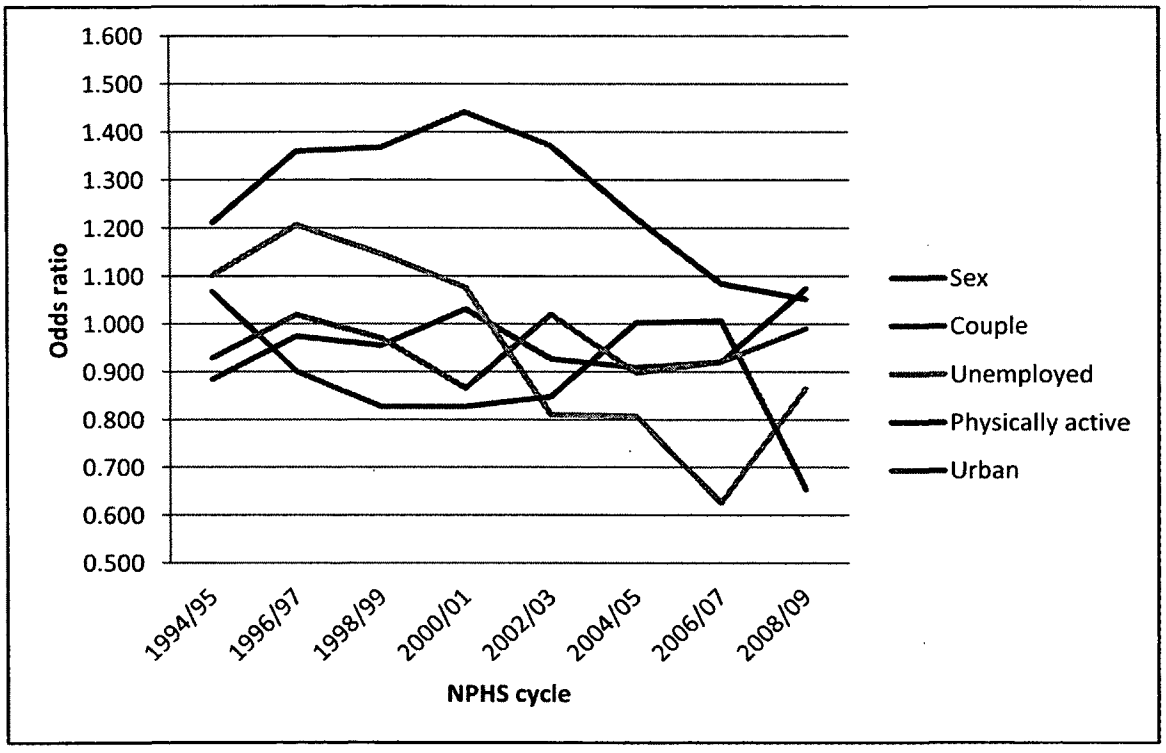
Source: National Population Health Survey (1994/95-2008/09)

Table A.3.17 Lower 95% confidence interval for logistic regression odds ratio estimates: Probability of having high blood pressure by cycle and covariate

Parameter	1994/95	1996/97	1998/99	2000/01	2002/03	2004/05	2006/07	2008/09
Sex	1.018	1.121	1.132	1.206	1.155	1.005	0.904	0.870
Couple	0.735	0.814	0.784	0.845	0.757	0.730	0.737	0.867
Unemployed	0.689	0.663	0.647	0.557	0.458	0.419	0.330	0.467
Physically active	0.885	0.754	0.685	0.691	0.705	0.840	0.838	0.534
Urban	0.755	0.841	0.784	0.721	0.852	0.753	0.778	0.823

Source: National Population Health Survey (1994/95-2008/09)

Figure A.3.1 Logistic regression odds ratio estimates: Probability of having high blood pressure by cycle and covariate



Source: National Population Health Survey (1994/95-2008/09)

It is observed that the relationships between the incidence of high blood pressure and the covariates are fluid and change over time. For instance, in the first waves,

women were much more likely than men to have high blood pressure but this difference is greatly reduced for the last cycle. This is reflected in that variable no longer being a significant determinant of high blood pressure as per our logistic regression models for the last 2 cycles. We also observe that the relationship between high blood pressure and physical activity is somewhat unpredictable over time.

Appendix 4

Bootstrapping for variance estimation

Due to the complex design of the NPHS sample, bootstrap methods are required to undertake any analysis of variance. The theory of using bootstrapping to calculate variance for complex survey designs is relatively straight forward and outlined in Rao (2006).

We accept the fact that any measures derived from the NPHS reflect a subsample of the Canadian population and are thus simply estimates. The true values of these measures remain unknown (unless a perfect census is taken and no error exists). However, if the NPHS is a representative sample (after weighting) of the Canadian population then the estimate should be an unbiased estimator of the true value. We encounter sampling error in the form of variance, however, as if we had selected a different sample, we may have derived a slightly different result and more importantly these results may differ slightly from the true value.

If we were to take a large number of samples from the population of interest, we would end up with a large number of estimates. We assume that these estimates should be focused around the true value with a certain level of variability. The variability of these estimates is thus an excellent measure of the variance of a single estimate. So, instead of relying on trying to derive a nontrivial formula to calculate variance of an estimate, we can simply calculate the estimates for a large number of samples and then find the variance of these estimates

$$var(\hat{y}) = \frac{\sum_{s=1}^{S}(\hat{y}_s - \hat{y})^2}{S-1}, \qquad \text{(A.4.1)}$$

where $\hat{y}$ is the original estimate (alternatively the mean value from the sample estimates can be used) and we have found $\hat{y}_i$ for S samples.

Unfortunately, it is incredibly expensive and generally not feasible to run the survey for many different samples. So, we employ a method of bootstrapping. Essentially, we begin by making the assumption that our original sample is truly representative of the general population. Therefore, any random subsamples derived from the original sample would be equivalent of taking random samples from the general population. Then we select S subsamples of the original sample for the purposes of this analysis. The subsamples are taken with replacement implying that the same individuals can be selected by more than 1 subsample.

Now, each of these subsamples undergoes the same weighting process that the overall sample went through to make sure that each one is representative of the general Canadian population in 1994/1995. So each subsample S, representing a fraction of the original sample, has its own set of weights $w_{s,i}$ where, for all $s \in S$,

$$\sum_{i=1}^{n} w_{s,i} = \sum_{i=1}^{n} w_i = N. \tag{A.4.2}$$

Now, if we run analysis on the overall sample and arrive at an estimate $\hat{y}$, we can run the same analysis for each subsample s and calculate the variance of $\hat{y}$ as

$$var(\hat{y}) = \frac{\sum_{s=1}^{S}(\hat{y}_s - \hat{y})^2}{S-1}. \tag{A.4.3}$$

The NPHS uses a SAS program called Bootvar to calculate bootstrap variances. This program cannot be used for some of the methods employed in this thesis but it was utilized for some of the simpler confidence intervals (e.g., totals). Essentially, it calculates the bootstrap variance as shown above for pre-determined functions.