Modeling Missing Data In Panel Studies With Multiple Refreshment Samples

by

Yiting Deng

Department of Statistical Science
Duke University

Date: ______________________
Approved:

____
Jerome P. Reiter, Supervisor

____
Fan Li

____
Richard Staelin

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2012

Abstract

Modeling Missing Data In Panel Studies With Multiple Refreshment Samples

by

Yiting Deng

Department of Statistical Science
Duke University

Date: ______________________

Approved:

____
Jerome P. Reiter, Supervisor

____
Fan Li

____
Richard Staelin

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2012

Abstract

Most panel surveys are subject to missing data problems caused by panel attrition. The Additive Non-ignorable (AN) model proposed by Hirano et al. (2001) utilizes refreshment samples in panel surveys to impute missing data, and offers flexibility in modeling the missing data mechanism to incorporate both ignorable and non-ignorable models. We extend the AN model to settings with three waves and two refreshment samples. We address identification and estimation issues related to the proposed model under four different types of survey design, featured by whether the missingness is monotone and whether subjects in the refreshment samples are followed up in subsequent waves of the survey. We apply this approach and multiple imputation techniques to the 2007-2008 Associated Press-Yahoo! News Poll (APYN) panel dataset to analyze factors affecting people's political interest. We find that, when attrition bias is not accounted for, the carry-on effects of past political interest on current political interest are underestimated. This highlights the importance of dealing with attrition bias and the potential of refreshment samples for doing so.

Key words: Refreshment sample, missing data, multiple imputation,panel attrition

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my esteemed committee - Jerry Reiter, Fan Li and Rick Staelin for making my journey in the Masters program fulfilling and enjoyable. Without their help and guidance, I would not have completed this work.

I am especially grateful to my thesis Chair Jerry. He has always been extremely passionate, patient and supportive, and it is his support and great insights that help me overcome various problems I encountered during this process. I am very lucky to have had the chance to work with him.

I would also like to acknowledge the faculty at the Department of Statistical Science and the Fuqua School of Business for the course work and training, for the nurturing and friendly environment, and for encouraging me to be a broader researcher. I am also grateful to Professor Sunshine Hillygus at the Department of Political Science for providing the data and many valuable comments.

Finally, I would like to thank fellow students, especially Yajuan Si and Jouchi Nakajima, for their comments and discussions.

1

Introduction

Panel surveys have become more and more popular over the past decades in economics and other fields of social sciences. However, most panel surveys are subject to missing data problems caused by panel attrition, and missing data problems are more severe in panel studies than in cross-sectional data (e.g., Hausman and Wise (1979); Hirano et al. (2001); Nevo (2003)). If the dropout pattern is correlated with variables of interest, the attrition can lead to biased inferences.

To mitigate panel attrition, panel data sets are sometimes augmented by refreshment samples, i.e., new units randomly sampled from the original population. However, the uses of refreshment samples vary across studies. Some studies use refreshment samples simply with the purpose of increasing sample size or replacing attriters in the original sample with respondents in the refreshment sample (e.g., Dorsett (2004); Fuchs-Schundeln and Schundeln (2005); Huber et al. (2011)). However, this could introduce additional bias, as respondents in refreshment samples might resemble respondents in the original samples more than the non-respondents. Some studies use refreshment samples to disentangle attrition and panel conditioning effects (e.g., Das et al. (2011)). Some studies use refreshment samples to examine panel attrition patterns or to adjust potential estimation bias. For instance, Hirano et al. (2001) show how a single refreshment sample can be used to relax the assumption under which the attrition bias can be identified. They further propose and estimate an Additive Non-ignorable (AN) model which nests both missing at random (MAR) model and the Hausman and Wise (1979) (HW) model as special cases. Nevo (2003) examines a structure similar to the AN model and shows that Generalized Methods of Moments (GMM) estimation can be relied upon to produce the weights needed for reweighting the panel. The refreshment samples in his study are used to attach a weight to each observation in the balanced subpanel such that moments in the weighted sample are set equal to corresponding moments in the refreshment samples. However, while bias from panel attrition can be reduced by such methods, there is wide variability in weights construction. Further, weights are usually created using demographics benchmarks, which alone are not sufficient to correct for attrition (Zheng (2011)).

In this thesis, we extend the model proposed by Hirano et al. (2001) by incorporating multiple refreshment samples. This study is motivated by the increasing availability of multiple refreshment samples in longitudinal survey design and survey data. For instance, the dataset our application is based on, the 2007-2008 Associated Press - Yahoo! News Poll (APYN) panel dataset, is a one year, 11-wave survey with 3 refreshment samples. As the data structure associated with multiple waves of refreshment samples is more complicated than that with a single refreshment sample, it is worth exploring the model selection and implications under a variety of survey designs. Our purpose in this paper is to address model selection and estimation issues related to the proposed method under different designs.

More specifically, we will study issues and implications of four different types of survey design. The potential designs differ in two aspects. The first difference is regarding whether subjects who fail to complete one wave of the survey are sampled again in subsequent waves. If the answer is yes, then the missingness pattern is non-monotone, as missingness in one wave does not necessarily lead to missingness in following waves. If the answer is no, then the missingness pattern is monotone, as missingness in one wave leads to missingness in following waves. The second difference is regarding whether subjects in refreshment samples are followed in subsequent waves. If yes, then the survey is with follow-up to subjects in the refreshment samples; if no, then the survey is without follow-up to those subjects.

The four survey designs (models) are as follows: Model 1 characterizes monotone missingness and no follow-up to subjects in the refreshment samples. Model 2 characterizes non-monotone missingness and no follow-up to subjects in the refreshment samples. Model 3 characterizes monotone missingness and follow-up to subjects in the refreshment samples. Model 4 characterizes non-monotone missingness and follow-up to subjects in the refreshment samples. To keep analysis intuitive and succinct, all the models we consider are 3-wave models with 2 refreshment samples respectively in the second wave and the third wave. These are the simplest models that are able to capture the implications of multiple refreshment samples in longitudinal surveys, and can be generalized to models with more refreshment samples.

The rest of this thesis is organized as follows. In Chapter 2, we briefly review the baseline model with a 2-wave panel data set proposed by Hirano et al. (2001), and discuss how a single refreshment sample can help identify model parameters. In Chapter 3, we extend the 2-wave model to a 3-wave model, and discuss model selection and estimation issues under the four different survey designs described above. In Chapter 4, we apply the method developed in Chapter 3 and multiple imputation to the APYN dataset to demonstrate the applicability of the modeling approach. In Chapter 5, we conclude with limitations, implications of the study and directions for future research.

2

Results For Single Refreshment Sample

In this chapter, we briefly review the 2-wave model with attrition proposed by Hirano et al. (2001). We have adapted some notations for consistency with subsequent chapters. Denote Y_{it} as a vector containing all time-varying variables for subject i at time t; our interest is to make inferences on Y_{it}. Also, denote X_i as a vector containing all time-invariant variables for subject i. For instance, the usual demographic information can be incorporated in X_i. Let N_P be the number of subjects that completed the wave 1 survey. These subjects are referred as the panel members. For each subject $i = 1, \ldots, N_P$, we observe X_i and Y_{1i}. We assume no missing data in X_i and Y_{i1} for the original panel. In wave 2, we observe Y_{2i} for N_{CP} of the N_P panel members, and Y_{2i} is missing for the remaining $N_{IP} = N_P - N_{CP}$ panel members. Meanwhile, a refreshment sample with N_R subjects is obtained in the second wave. For each subject $i = N_P + 1, \ldots, N_P + N_R$, we observe X_i and Y_{2i}, but not Y_{1i}. Further, we denote W_{1i} as an attrition indicator. W_{1i} is defined as 1 if person i responds in wave 2 when interviewed in wave 1 and wave 2. It is defined as 0 if person i does not respond in wave 2 when interviewed in both waves. Hence, W_{1i} is observed if subject i is in the panel and is not observed if subject i is in the refreshment sample. Also denote $N = N_P + N_R$ as the overall sample size. Figure 2.1 displays the pattern of the data.

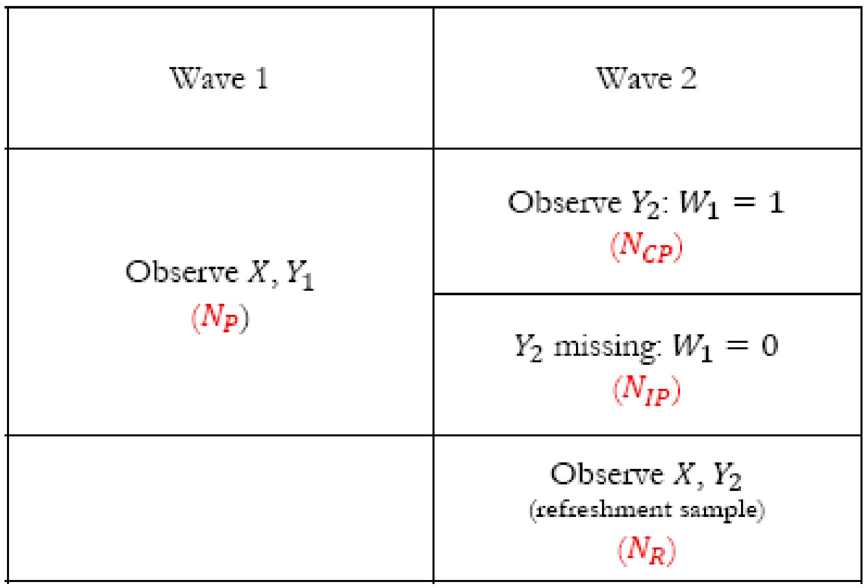

FIGURE 2.1: Graphical representation of the two-wave model

We use the Additive Non-ignorable (AN) model proposed by Hirano et al. (2001) to model the attrition process, which is characterized by the following selection model:

$$W_{1i} \mid X_i, Y_{1i}, Y_{2i}, \alpha \sim Bern\left(\frac{exp\,(\alpha_0 + \alpha_1 X_i + \alpha_2 Y_{1i} + \alpha_3 Y_{2i})}{1 + exp\,(\alpha_0 + \alpha_1 X_i + \alpha_2 Y_{1i} + \alpha_3 Y_{2i})}\right) \quad (2.1)$$

where $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)'$

Both missing at random (MAR) model and the HW model are special cases of this model. The AN model reduces to the HW model when $(\alpha_2 = 0, \alpha_3 \neq 0)$ and reduces to the MAR model when $(\alpha_3 = 0)$.

To provide intuition to model selection, we examine a simple model where X_i, Y_{1i} and Y_{2i} are all Bernoulli distributed binary variables. More specifically, we specify the data generating process in Equation (2.1) and Equation (2.2)-(2.4) that follow:

$$X_i \sim Bern\,(p) \tag{2.2}$$

$$Y_{1i} \mid X_i, \beta \sim Bern\left(\tfrac{exp(\beta_0+\beta_1 X_i)}{1+exp(\beta_0+\beta_1 X_i)}\right) \tag{2.3}$$

$$where\, \beta = (\beta_0, \beta_1)'$$

$$Y_{2i} \mid X_i,\, Y_{1i}, \boldsymbol{\gamma} \sim Bern\left(\tfrac{exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}\right) \tag{2.4}$$

$$where\, \gamma = (\gamma_0, \gamma_1, \gamma_2)'$$

Our interest is to estimate β, γ and α.

For purpose of demonstrating model selection in a concise way, we omit X and focus on Y_1, Y_2 and W_1 in our discussion of identification.

There are three variables Y_1, Y_2 and W_1, each takes value 1 or 0, so there are essentially 8 joint probabilities to be estimated. We know the following 4 probabilities from the complete observations (i.e., the N_{CP} subjects who complete both waves): $P\,(Y_1 = 1, Y_2 = 1, W_1 = 1)$, $P\,(Y_1 = 1, Y_2 = 0, W_1 = 1)$, $P\,(Y_1 = 0, Y_2 = 1, W_1 = 1)$ and $P\,(Y_1 = 0, Y_2 = 0, W_1 = 1)$. The following 4 probabilities are not directly observed: $P\,(Y_1 = 1, Y_2 = 1, W_1 = 0)$, $P\,(Y_1 = 1, Y_2 = 0, W_1 = 0)$, $P\,(Y_1 = 0, Y_2 = 1, W_1 = 0)$ and $P\,(Y_1 = 0, Y_2 = 0, W_1 = 0)$.

So we have 4 probabilities to estimate. However, we only have 3 constraints from the observed data:

$$P\left(Y_1=1, Y_2=0, W_1=0\right)+P\left(Y_1=1, Y_2=1, W_1=0\right)=P\left(Y_1=1, W_1=0\right) \tag{2.5}$$

$$P\left(Y_1=0, Y_2=1, W_1=0\right)+P\left(Y_1=1, Y_2=1, W_1=0\right)$$
$$+P\left(Y_1=0, Y_2=1, W_1=1\right)+P\left(Y_1=1, Y_2=1, W_1=1\right)=P\left(Y_2=1\right) \tag{2.6}$$

$$P\left(Y_1=1, Y_2=0, W_1=0\right)+P\left(Y_1=1, Y_2=1, W_1=0\right)$$
$$+P\left(Y_1=0, Y_2=1, W_1=0\right)+P\left(Y_1=0, Y_2=0, W_1=0\right)=P\left(W_1=0\right) \tag{2.7}$$

Here, $P\left(Y_1=1, W_1=0\right)$ and $P\left(W_1=0\right)$ are observed from the N_P panel members, and $P\left(Y_2=1\right)$ is observed from the refreshment sample.

This motivates the AN model, which does not include a term for Y_1Y_2. If we had included that term, there would be a total of 8 parameters in the model with only 7 total constraints, so that it would not be identified.

We describe our estimation strategy and the estimates for this model in Appendix A. We use 10,000 MCMC iterations and a burn-in of 5,000 in the MCMC algorithm. As an illustration of the validity of the model, we also run a simulation study to compare the performance of the HW model, the MAR model and the AN models. Since the HW model and the MAR model both add restrictions to parameters in the attrition equation, they should only produce unbiased estimates when the actual attrition process is consistent with their respective specifications. The AN model, on the other side, does not add any restrictions to the parameters, beyond excluding Y_1Y_2 interaction in W_1 equation.

Using the simulation design described in Appendix A, we first generate 300 datasets, 100 under the HW assumption, 100 under the MAR assumption, and 100 under the AN assumption. The HW datasets are generated by specifying $\alpha_2 = 0$ and the MAR datasets are generated by specifying $\alpha_3 = 0$. Then we implement the three models for each dataset. Table 2.1, Table 2.2 and Table 2.3 respectively report the estimation results for HW data, MAR data, and AN data. As expected, the AN model is able to produce approximately unbiased estimates regardless of the type of missingness, while the HW model and the MAR model fail to do so for missingness patterns inconsistent with their respective underlying assumptions.

Table 2.1: Estimation results for HW datasets based on 100 simulations

	True value	Mean (HW)	95% Cov. rate (HW)	Mean (MAR)	95% Cov. rate (MAR)	Mean (AN)	95% Cov. rate (AN)
β_0	0.3	0.30	0.98	0.28	0.94	0.30	1.00
β_1	-0.4	-0.40	0.98	-0.39	0.98	-0.40	0.97
γ_0	0.3	0.31	1.00	0.57	0.00	0.31	0.99
γ_1	-0.3	-0.31	0.99	-0.40	0.50	-0.31	1.00
γ_2	0.7	0.71	0.98	0.72	0.90	0.71	0.98
α_0	-0.4	-0.37	1.00	0.36	0.00	-0.36	1.00
α_1	1.0	0.99	1.00	0.82	0.12	0.99	1.00
α_2	0	-	-	0.00	1.00	0.00	1.00
α_3	1.3	1.25	1.00	-	-	1.24	1.00

Table 2.2: Estimation results for MAR datasets based on 100 simulations

	True value	Mean (HW)	95% Cov. rate (HW)	Mean (MAR)	95% Cov. rate (MAR)	Mean (AN)	95% Cov. rate (AN)
β_0	0.3	0.29	0.94	0.30	1.00	0.30	0.99
β_1	-0.4	-0.40	0.95	-0.40	1.00	-0.40	0.94
γ_0	0.3	0.36	0.93	0.29	0.99	0.29	0.96
γ_1	-0.3	-0.30	1.00	-0.30	1.00	-0.30	0.97
γ_2	0.7	0.80	0.71	0.68	0.99	0.71	0.98
α_0	-0.4	-0.49	0.83	-0.45	0.98	-0.40	1.00
α_1	1.0	1.01	0.99	1.05	1.00	1.00	1.00
α_2	-0.7	-	-	-0.69	1.00	-0.70	1.00
α_3	0	-0.42	0.01	-	-	0.00	0.98

Table 2.3: Estimation results for AN datasets based on 100 simulations

	True value	Mean (HW)	95% Cov. rate (HW)	Mean (MAR)	95% Cov. rate (MAR)	Mean (AN)	95% Cov. rate (AN)
β_0	0.3	0.29	0.94	0.28	0.99	0.29	0.95
β_1	-0.4	-0.39	0.96	-0.40	0.95	-0.40	0.96
γ_0	0.3	0.45	0.22	0.54	0.00	0.32	0.97
γ_1	-0.3	-0.35	0.98	-0.39	0.71	-0.30	0.99
γ_2	0.7	0.69	0.93	0.83	0.45	0.69	0.96
α_0	-0.4	-0.44	0.98	-0.01	0.00	-0.37	1.00
α_1	1.0	0.95	0.93	0.87	0.47	0.98	1.00
α_2	-0.7	-	-	-0.00	0.00	-0.69	0.99
α_3	1.3	0.72	0.04	-	-	1.25	1.00

Some agencies or analysts may want to use only the original panel for analyses. One potential way to do so is via multiple imputation, in which the agency completes the missing values in Y_2 in the original panel. We now provide a brief review of multiple imputation (Rubin (1987)), followed by an investigation of its usefulness for this context. Multiple imputation offers a way to deal with item nonresponse by drawing missing data to create full datasets, which then can be analyzed with complete data techniques. The inference method proposed by Rubin (1987) is summarized as follows. With m imputed datasets, we can compute m different sets of the point and variance estimates for the quantity of interest, Q. Denote $q^{(l)}$ and $u^{(l)}$ respectively as the point and variance estimates for Q from the lth imputed dataset, where $l = 1, \ldots, m$. The point estimate for Q from the multiple imputation is the average of the m complete-data estimates,

$$\bar{q}_m = \frac{\sum_{l=1}^{m} q^{(l)}}{m}. \tag{2.8}$$

The variance for $\bar{q}_m$ is

$$T_m = \left(1 + \frac{1}{m}\right) b_m + \bar{u}_m, \tag{2.9}$$

where $b_m = \frac{\sum_{j=1}^{m}\left(q^{(l)} - \bar{q}_m\right)^2}{m-1}$, and $\bar{u}_m = \frac{\sum_{l=1}^{m} u^{(l)}}{m}$.

For large samples, inferences for Q are obtained from the t-distribution,

$$\frac{\bar{q}_m - Q}{\sqrt{T_m}} \to t_{v_m}(0, 1), \tag{2.10}$$

where $v_m = (m-1)\left[1 + \frac{m\bar{u}_m}{(m+1)b_m}\right]^2$.

However, according to Reiter (2008), when some of the records used to estimate the imputation models in multiple imputation are not used or available for analysis, the usual multiple imputation variance estimator has positive bias. Therefore, if we make inference based only on the original panel, using a refreshment sample in the process of imputation would create a positive bias in the variance estimator for the one-step multiple imputation. To correct for that bias, Reiter (2008) proposes a two-step multiple imputation routine. Denote observed data as D, the model parameters as θ and the missing values in the original panel as Y_{mis}. First, we sample m values of θ from $f(\theta|D)$. For each $\theta^{(l)}$ ($l = 1, \ldots, m$), we draw n sets of Y_{mis} from $f\left(Y_{mis}|D, \theta^{(l)}\right)$. We can use the combining rules below to draw unbiased inferences for means and variances in the following ways. [1]

$$\bar{q}_m = \frac{\sum_{l=1}^{m} \sum_{i=1}^{n} q^{(l,i)}}{mn} = \frac{\sum_{l=1}^{m} \bar{q}_n^{(l)}}{m} \tag{2.11}$$

$$\bar{w}_m = \frac{\sum_{l=1}^{m} \sum_{i=1}^{n} \left(q^{(l,i)} - \bar{q}_n^{(l)}\right)^2}{m(n-1)} = \frac{\sum_{l=1}^{m} w_n^{(l)}}{m} \tag{2.12}$$

$$b_m = \frac{\sum_{l=1}^{m} \left(\bar{q}_n^{(l)} - \bar{q}_m\right)^2}{m-1} \tag{2.13}$$

$$\bar{u}_m = \frac{\sum_{l=1}^{m} \sum_{i=1}^{n} u^{(l,i)}}{mn} \tag{2.14}$$

The variance of $\bar{q}_m$ is estimated by $T_m = \bar{u}_m - \bar{w}_m + \left(1 + \frac{1}{m}\right) b_m - \frac{\bar{w}_m}{n}$. When $T_m < 0$, we use $T_m = \left(1 + \frac{1}{m}\right) b_m$.

When the sample size is large, inferences are based on a t-distribution:

$$\frac{\bar{q}_m - Q}{\sqrt{T_m}} \rightarrow t_{v_M}(0, 1) \tag{2.15}$$

Where $v_m = \left\{ \frac{\left[\left(1+\frac{1}{m}\right)b_m\right]^2}{(m-1)T_m^2} + \frac{\left[\left(1+\frac{1}{n}\right)\bar{w}_m\right]^2}{m(n-1)T_m^2} \right\}^{-1}$ when $T_m > 0$, and $v_m = m - 1$ when $T_m < 0$.

[1] Notations are similar to the previous ones.

We implement both one-step and two-step multiple imputation and compare the results. In our one-step multiple imputation study, we first generate 500 datasets using parameters in Table 2.3 . Then for each dataset, we implement the estimation routine described in Appendex A to obtain a set of parameter estimates. We then implement the multiple imputation routine discussed in Equation (2.8)-(2.10) with $m = 100$. In the two-step multiple imputation study, we apply Equation (2.11)-(2.15) with $m = 100$ and $n = 10$ for the same 500 datasets. Table 2.4 and Table 2.5 report the results obtained under the one-step and two-step multiple imputation respectively.

Table 2.4: One-step multiple imputation results based on 500 simulations

Parameter	True value	95% Coverage rate	Mean of MI point estimates	Variance of MI point estimates	Mean of estimated MI variances
β_0	0.3	0.958	0.30	0.0008	0.0008
β_1	-0.4	0.952	-0.40	0.0016	0.0016
γ_0	0.3	0.992	0.30	0.0018	0.0034
γ_1	-0.3	0.984	-0.30	0.0022	0.0031
γ_2	0.7	0.964	0.70	0.0031	0.0032

Table 2.5: Two-step multiple imputation results based on 500 simulations

Parameter	True value	95% Coverage rate	Mean of MI point estimates	Variance of MI point estimates	Mean of estimated MI variances
β_0	0.3	0.958	0.30	0.0008	0.0008
β_1	-0.4	0.952	-0.40	0.0016	0.0016
γ_0	0.3	0.972	0.30	0.0018	0.0023
γ_1	-0.3	0.910	-0.30	0.0021	0.0017
γ_2	0.7	0.872	0.70	0.0031	0.0017

As we can see, these two methods result in point estimates that are very close. This is because we use the same draws of parameters when imputing the missing values. And since we generate a very large sample (10,000 panel members), the resulting estimates are almost identical. However, the variance estimators differ between the two methods. The bias in the variance estimators can be evaluated by comparing the variance estimator T_m (the last column in the above two tables) with the variance of the mean estimator q_m (the second-to-last column in the above two tables). We can see that there exists a positive bias in the variance estimators under the one-step multiple imputation. The variance estimators under the two-step multiple imputation are smaller. Comparing the variance estimators under the two approaches, we see that the one-step multiple imputation produces better variance estimators for γ_0 and γ_1, while the two-step multiple imputation produces a better variance estimator for γ_2. Therefore, under this specific case, the superiority of the two-step multiple imputation over the standard (one-step) multiple imputation is not clear.

3

Multiple Refreshment Samples

In this chapter, we extend the 2-wave model described in Chapter 2 to 3-wave models, i.e., models with refreshment samples in the second wave and the third wave. The models are specified similarly to the previous chapter, except that we add one wave (wave 3), in which Y_{3i} is collected. Meanwhile, a second refreshment sample with N_{R2} subjects is obtained in wave 3. We add W_{2i} as an attrition indicator for wave 3. W_{2i} is defined as 1 if person i responds in wave 3 when interviewed in wave 2 and wave 3. It is defined as 0 if person i does not respond in wave 3 when interviewed in wave 2 and wave 3. We specify the (full) data generating process similarly to Chapter 2, as shown in Equation (3.1)- (3.5) below:

$$Y_{1i} \mid X_i, \beta \sim Bern\left(\frac{exp\left(v_\beta\right)}{1+exp\left(v_\beta\right)}\right) \tag{3.1}$$

$$Y_{2i} \mid X_i,\, Y_{1i}, \gamma \sim Bern\left(\frac{exp(v_\gamma)}{1+exp(v_\gamma)}\right) \tag{3.2}$$

$$W_i \mid X_i, Y_{1i}, Y_{2i}, \alpha \sim Bern\left(\frac{exp(v_\alpha)}{1+exp(v_\alpha)}\right) \tag{3.3}$$

$$Y_{3i} \mid X_i, Y_{1i}, Y_{2i}, W_i, \theta \sim Bern\left(\frac{exp(v_\theta)}{1+exp(v_\theta)}\right) \tag{3.4}$$

$$W_{2i} \mid X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_i, \lambda \sim Bern\left(\frac{exp(v_\lambda)}{1+exp(v_\lambda)}\right), \tag{3.5}$$

here we have:

$v_\beta = \beta_0 + \beta_1 X_i$

$v_\gamma = \gamma_0 + \gamma_1 X_i + \gamma_2 Y_{1i}$

$v_\alpha = \alpha_0 + \alpha_1 X_i + \alpha_2 Y_{1i} + \alpha_3 Y_{2i} + \alpha_4 Y_{1i} Y_{2i}$

$v_\theta = \theta_0 + \theta_1 X_i + \theta_2 Y_{1i} + \theta_3 Y_{2i} + \theta_4 W_{1i} + \theta_5 Y_{1i} Y_{2i} + \theta_6 Y_{1i} W_{1i} + \theta_7 Y_{2i} W_{1i} + \theta_8 Y_{1i} Y_{2i} W_{1i}$

$$\begin{aligned} v_\lambda = {} & \lambda_0 + \lambda_1 X_i + \lambda_2 Y_{1i} + \lambda_3 Y_{2i} + \lambda_4 Y_{3i} + \lambda_5 W_{1i} \\ & + \lambda_6 Y_{1i} Y_{2i} + \lambda_7 Y_{1i} W_{1i} + \lambda_8 Y_{2i} W_{1i} + \lambda_9 Y_{1i} Y_{3i} + \lambda_{10} Y_{2i} Y_{3i} + \lambda_{11} Y_{3i} W_{1i} \\ & + \lambda_{12} Y_{1i} Y_{2i} Y_{3i} + \lambda_{13} Y_{1i} Y_{2i} W_{1i} + \lambda_{14} Y_{1i} Y_{3i} W_{1i} + \lambda_{15} Y_{2i} Y_{3i} W_{1i} + \lambda_{16} Y_{1i} Y_{2i} Y_{3i} W_{1i} \end{aligned}$$

$\beta = (\beta_0, \beta_1)'$

$\gamma = (\gamma_0, \gamma_1, \gamma_2)'$

$\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)'$

$\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8)'$

$\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8, \lambda_9, \lambda_{10}, \lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{14}, \lambda_{15}, \lambda_{16})'$

Note that it is possible to include interaction terms with X in the above equations, but we omit them here for brevity. Our interest is to estimate β, γ, α, θ and λ.

However, not all parameters are identifiable. Under some models, certain elements in θ and λ take values of zero. As mentioned earlier, we will consider four different models corresponding to different survey designs and discuss model selection under each model.

For purpose of demonstrating model selection in a concise way, we omit X and focus on Y_1, Y_2, Y_3, W_1 and W_2 in our discussion of model selection in this chapter. Hence, β_1, γ_1, α_1, θ_1, and λ_1 are dropped. This leaves us with 31 parameters to estimate. As each of the five variables can take value of 0 or 1, there are essentially $2^5 = 32$ probabilities to be estimated. In addition, these 32 probabilities sum up to one, so that estimating these 32 probabilities is equivalent to estimating the 31 parameters. All model information is included in these 32 probabilities summarized in vector P. If we were able to estimate all the elements in P, then we will be able to estimate the full model specified in Equation (3.1)- (3.5).

$$P = \begin{bmatrix} P\left(Y_1=1,Y_2=1,Y_3=1,W_1=1,W_2=1\right) \\ P\left(Y_1=1,Y_2=1,Y_3=0,W_1=1,W_2=1\right) \\ P\left(Y_1=1,Y_2=0,Y_3=1,W_1=1,W_2=1\right) \\ P\left(Y_1=1,Y_2=0,Y_3=0,W_1=1,W_2=1\right) \\ P\left(Y_1=0,Y_2=1,Y_3=1,W_1=1,W_2=1\right) \\ P\left(Y_1=0,Y_2=1,Y_3=0,W_1=1,W_2=1\right) \\ P\left(Y_1=0,Y_2=0,Y_3=1,W_1=1,W_2=1\right) \\ P\left(Y_1=0,Y_2=0,Y_3=0,W_1=1,W_2=1\right) \\ P\left(Y_1=1,Y_2=1,Y_3=1,W_1=1,W_2=0\right) \\ P\left(Y_1=1,Y_2=1,Y_3=0,W_1=1,W_2=0\right) \\ P\left(Y_1=1,Y_2=0,Y_3=1,W_1=1,W_2=0\right) \\ P\left(Y_1=1,Y_2=0,Y_3=0,W_1=1,W_2=0\right) \\ P\left(Y_1=0,Y_2=1,Y_3=1,W_1=1,W_2=0\right) \\ P\left(Y_1=0,Y_2=1,Y_3=0,W_1=1,W_2=0\right) \\ P\left(Y_1=0,Y_2=0,Y_3=1,W_1=1,W_2=0\right) \\ P\left(Y_1=0,Y_2=0,Y_3=0,W_1=1,W_2=0\right) \\ P\left(Y_1=1,Y_2=1,Y_3=1,W_1=0,W_2=1\right) \\ P\left(Y_1=1,Y_2=1,Y_3=0,W_1=0,W_2=1\right) \\ P\left(Y_1=1,Y_2=0,Y_3=1,W_1=0,W_2=1\right) \\ P\left(Y_1=1,Y_2=0,Y_3=0,W_1=0,W_2=1\right) \\ P\left(Y_1=0,Y_2=1,Y_3=1,W_1=0,W_2=1\right) \\ P\left(Y_1=0,Y_2=1,Y_3=0,W_1=0,W_2=1\right) \\ P\left(Y_1=0,Y_2=0,Y_3=1,W_1=0,W_2=1\right) \\ P\left(Y_1=0,Y_2=0,Y_3=0,W_1=0,W_2=1\right) \\ P\left(Y_1=1,Y_2=1,Y_3=1,W_1=0,W_2=0\right) \\ P\left(Y_1=1,Y_2=1,Y_3=0,W_1=0,W_2=0\right) \\ P\left(Y_1=1,Y_2=0,Y_3=1,W_1=0,W_2=0\right) \\ P\left(Y_1=1,Y_2=0,Y_3=0,W_1=0,W_2=0\right) \\ P\left(Y_1=0,Y_2=1,Y_3=1,W_1=0,W_2=0\right) \\ P\left(Y_1=0,Y_2=1,Y_3=0,W_1=0,W_2=0\right) \\ P\left(Y_1=0,Y_2=0,Y_3=1,W_1=0,W_2=0\right) \\ P\left(Y_1=0,Y_2=0,Y_3=0,W_1=0,W_2=0\right) \end{bmatrix}$$

3.1 Monotone Missingness Pattern, No Follow-up For Refreshment Samples

In this section, we discuss the model selection issues under the 3-wave model with monotone missingness pattern and no follow-up for subjects in refreshment samples. Monotone missingness means that subjects who fail to complete one wave of the survey are not surveyed in subsequent waves. No follow-up means that subjects in both refreshment samples are only observed in the corresponding wave and not sampled in later waves.

More specifically, the first two waves are sampled as described in Chapter 2. In the third wave, the second refreshment sample including N_{R2} subjects is sampled. For each subject $i = N_P + N_R + 1, \ldots, N_P + N_R + N_{R2}$, we observe X_i and Y_{3i}, but neither Y_{1i} nor Y_{2i} is observed. This means that subjects in the first refreshment sample are not being followed in subsequent surveys. Because of the monotone missingness pattern, i.e., the missingness is caused by "dropout", when $W_{1i} = 0$, we do not observe subsequent choices for subject i. Among the N_{CP} subjects in the panel with $W_1 = 1$, we observe Y_{3i} for N_{CP2} subjects, and Y_{3i} is missing for the rest $N_{IP2} = N_{CP} - N_{CP2}$ subjects. For the N_{IP} subjects in the panel with $W_1 = 0$, we observe neither Y_3 nor W_2. For the N_R subjects in the first refreshment sample, we observe X_i and Y_2, but not Y_1. W_1 and W_2 are not observed for all subjects in the refreshment samples. Figure 3.1 shows the pattern of the data.

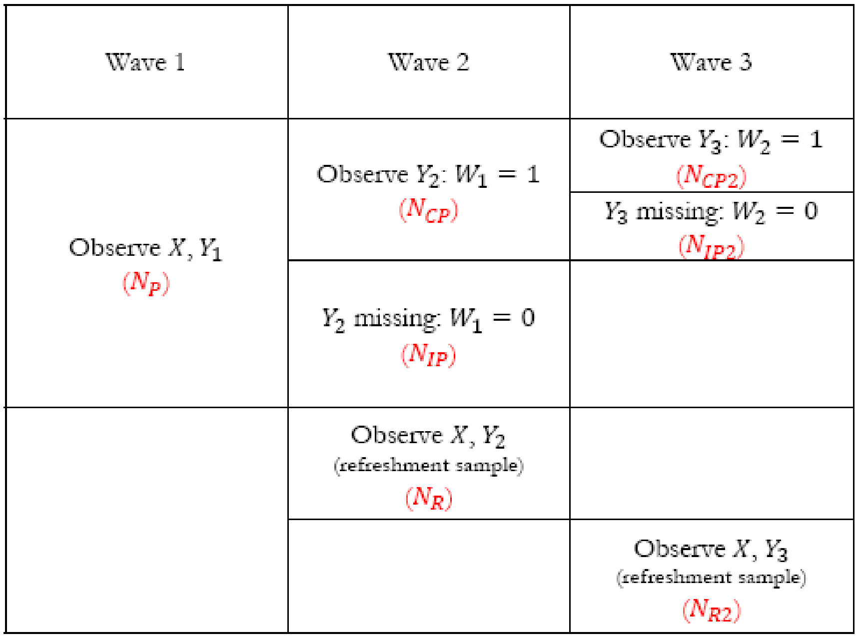

FIGURE 3.1: Graphical representation of the 3-wave model with monotone missingness pattern and no follow-up for subjects in refreshment samples

There are 21 probabilities that are directly observable from the observed data. From the N_{CP2} subjects who complete all 3 waves, we observe the first 8 elements in P. From the N_{CP} subjects who complete the first 2 waves, we observe $P(W_1 = 1, W_2 = 1)$ and $P(W_1 = 1, W_2 = 0)$. From the N_P panel members, we observe $P(W_1 = 0)$. From the N_{IP2} subjects who complete the first 2 waves but not the third wave, we observe $P(Y_1 = 1, Y_2 = 1, W_1 = 1, W_2 = 0)$, $P(Y_1 = 1, Y_2 = 0, W_1 = 1, W_2 = 0)$, $P(Y_1 = 0, Y_2 = 1, W_1 = 1, W_2 = 0)$, and $P(Y_1 = 0, Y_2 = 0, W_1 = 1, W_2 = 0)$. From the N_{IP} subjects who only complete the first wave, we observe $P(Y_1 = 1, W_1 = 0)$ and $P(Y_1 = 0, W_1 = 0)$. From the N_R subjects in the first refreshment sample, we observe $P(Y_2 = 1)$ and $P(Y_2 = 0)$. From the N_{R2} subjects in the second refreshment sample, we observe $P(Y_3 = 1)$ and

$P(Y_3 = 0)$. Obviously, some of the probabilities are redundant. Hence, to see the number of independent constraints we have on P, we summarize the 21 observable probabilities in vector R_1 and write it as a function of P:

$$R_1 \equiv \begin{bmatrix} P(Y_1 = 1, Y_2 = 1, Y_3 = 1, W_1 = 1, W_2 = 1) \\ P(Y_1 = 1, Y_2 = 1, Y_3 = 0, W_1 = 1, W_2 = 1) \\ P(Y_1 = 1, Y_2 = 0, Y_3 = 1, W_1 = 1, W_2 = 1) \\ P(Y_1 = 1, Y_2 = 0, Y_3 = 0, W_1 = 1, W_2 = 1) \\ P(Y_1 = 0, Y_2 = 1, Y_3 = 1, W_1 = 1, W_2 = 1) \\ P(Y_1 = 0, Y_2 = 1, Y_3 = 0, W_1 = 1, W_2 = 1) \\ P(Y_1 = 0, Y_2 = 0, Y_3 = 1, W_1 = 1, W_2 = 1) \\ P(Y_1 = 0, Y_2 = 0, Y_3 = 0, W_1 = 1, W_2 = 1) \\ P(W_1 = 1, W_2 = 1) \\ P(W_1 = 1, W_2 = 0) \\ P(W_1 = 0) \\ P(Y_1 = 1, Y_2 = 1, W_1 = 1, W_2 = 0) \\ P(Y_1 = 1, Y_2 = 0, W_1 = 1, W_2 = 0) \\ P(Y_1 = 0, Y_2 = 1, W_1 = 1, W_2 = 0) \\ P(Y_1 = 0, Y_2 = 0, W_1 = 1, W_2 = 0) \\ P(Y_1 = 1, W_1 = 0) \\ P(Y_1 = 0, W_1 = 0) \\ P(Y_2 = 1) \\ P(Y_2 = 0) \\ P(Y_3 = 1) \\ P(Y_3 = 0) \end{bmatrix} = C_1 P$$

where:

$$C_1 = \begin{bmatrix}
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1\\
1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0\\
0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1\\
1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0\\
0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1
\end{bmatrix}$$

The number of independent constraints equals to the rank of C_1. As $rank\,(C_1) = 16$, there are 16 independent constraints from the observed data. The model specified in Equation (3.1)- (3.5) has 31 parameters excluding $\beta_1, \gamma_1, \alpha_1, \theta_1$ and λ_1, so we need to drop 15 parameters to ensure identification. Because we have the same amount of information for subjects in the first two waves as in Chapter 2, we continue to set $\alpha_4 = 0$ as before. And since we cannot observe Y_3 and W_2 when $W_1 = 0$, we are not able to distinguish between the N_{IP2} subjects and the N_{IP} subjects in wave 3, θ_4 and λ_5 are not identifiable. So in Equation (3.4), we specify $\theta_4 = \theta_6 = \theta_7 = \theta_8 = 0$, and in Equation (3.5), we specify $\lambda_5 = \lambda_7 = \lambda_8 = \ldots = \lambda_{16} = 0$. We require Y_3 and W_2 to be conditionally independent of W_1. Then the N_{CP2} complete cases and the N_{CP} subjects with fully observed (Y_1, Y_2) and the second refreshment sample can identify the interaction of Y_1Y_2 respectively in Equation (3.4) and Equation (3.5). Essentially, the N_{CP} subjects with fully observed (Y_1, Y_2) and the second refreshment sample considered in isolation are akin to a two-wave panel sample with Y_1 and Y_2 as integrally observed overrates in the first wave and Y_3 in the second wave. As a general AN model, we can identify the main effect term Y_3 in Equation (3.5). The requirement that Y_3 and W_2 be conditionally independent of W_1 means that conditional on a subject's attitudes in the first two waves (either observable or non-observable), whether she has responded in the second wave does not affect her attitudes in the third wave, nor does it affect her willingness to respond in the third wave. We note that this is a rather strong requirement and may not be satisfied in case of panel conditioning where previous participation in an interview alters respondents' true values and/or their reports of the true values.

We run a MCMC simulation to test the model and most parameters are well recovered. We use $200,000$ MCMC iterations and a burn-in sample of $100,000$. The details on the MCMC algorithm and the results are provided in Appendix B.

3.2 Non-Monotone Missingness Pattern, No Follow-up For Refreshment Samples

In this section, we discuss the model selection issues under the 3-wave model with non-monotone missingness pattern and no follow-up for subjects in refreshment samples. This model is specified similarly to the model discussed in Section 3.1, except that people who fail to complete one wave are surveyed again in subsequent waves. Hence, W_{2i} is observed for all subjects in the panel. Among the N_{CP} subjects in the panel with $W_1 = 1$, we observe Y_{3i} for N_{CP2} subjects, and Y_{3i} is missing for the rest $N_{IP2} = N_{CP} - N_{CP2}$ subjects. Among the N_{IP} subjects in the panel with $W_1 = 0$, we observe Y_{3i} for N_{CP3} subjects, and Y_{3i} is missing for the rest $N_{IP3} = N_{IP} - N_{CP3}$ subjects. Again, W_{1i} and W_{2i} are missing for all subjects in the refreshment samples. Figure 3.2 shows the pattern of the data.

<table>
<tr><th>Wave 1</th><th>Wave 2</th><th>Wave 3</th></tr>
<tr><td rowspan="4">Observe X, Y_1
(N_P)</td><td rowspan="2">Observe Y_2: $W_1 = 1$
(N_{CP})</td><td>Observe Y_3: $W_2 = 1$
(N_{CP2})</td></tr>
<tr><td>Y_3 missing: $W_2 = 0$
(N_{IP2})</td></tr>
<tr><td rowspan="2">Y_2 missing: $W_1 = 0$
(N_{IP})</td><td>Observe Y_3: $W_2 = 1$
(N_{CP3})</td></tr>
<tr><td>Y_3 missing: $W_2 = 0$
(N_{IP3})</td></tr>
<tr><td rowspan="2"></td><td>Observe X, Y_2
(refreshment sample)
(N_R)</td><td></td></tr>
<tr><td></td><td>Observe X, Y_3
(refreshment sample)
(N_{R2})</td></tr>
</table>

FIGURE 3.2: Graphical representation of the 3-wave model with non-monotone missingness pattern and no follow-up for subjects in refreshment samples

Model selection of this model is similar to the case in Section 3.1.However, as subjects who fail to complete wave 2 are sampled again in wave 3, we can obtain more constraints from the presence of the N_{CP3} and N_{IP3} subjects. We observe the same 8 probabilities from the N_{CP2} subjects who complete all 3 waves. From the N_P panel members, we observe $P\left(W_1 = 1, W_2 = 1\right)$, $P\left(W_1 = 1, W_2 = 0\right)$, $P\left(W_1 = 0, W_2 = 1\right)$ and $P\left(W_1 = 0, W_2 = 0\right)$. We observe the same 4 probabilities from the N_{IP2} subjects who complete the first two waves but not the third wave. From the N_{CP3} subjects who fail to complete wave 2 but complete wave 3, we observe $P\left(Y_1 = 1, Y_3 = 1, W_1 = 0, W_2 = 1\right)$, $P\left(Y_1 = 1, Y_3 = 0, W_1 = 0, W_2 = 1\right)$, $P\left(Y_1 = 0, Y_3 = 1, W_1 = 0, W_2 = 1\right)$, and $P\left(Y_1 = 0, Y_3 = 0, W_1 = 0, W_2 = 1\right)$. From the N_{IP3} subjects who only complete wave 1, we observe $P\left(Y_1 = 1, W_1 = 0, W_2 = 0\right)$ and

$P(Y_1 = 0, W_1 = 0, W_2 = 0)$. From the N_R subjects in the first refreshment sample and the N_{R2} subjects in the second refreshment sample, we observe the same probabilities as before. To see the number of independent constraints we have on P, we summarize these observable probabilities in vector R_2 and write it as a function of P:

$$R_2 \equiv \begin{bmatrix} P(Y_1 = 1, Y_2 = 1, Y_3 = 1, W_1 = 1, W_2 = 1) \\ P(Y_1 = 1, Y_2 = 1, Y_3 = 0, W_1 = 1, W_2 = 1) \\ P(Y_1 = 1, Y_2 = 0, Y_3 = 1, W_1 = 1, W_2 = 1) \\ P(Y_1 = 1, Y_2 = 0, Y_3 = 0, W_1 = 1, W_2 = 1) \\ P(Y_1 = 0, Y_2 = 1, Y_3 = 1, W_1 = 1, W_2 = 1) \\ P(Y_1 = 0, Y_2 = 1, Y_3 = 0, W_1 = 1, W_2 = 1) \\ P(Y_1 = 0, Y_2 = 0, Y_3 = 1, W_1 = 1, W_2 = 1) \\ P(Y_1 = 0, Y_2 = 0, Y_3 = 0, W_1 = 1, W_2 = 1) \\ P(W_1 = 1, W_2 = 1) \\ P(W_1 = 1, W_2 = 0) \\ P(W_1 = 0, W_2 = 1) \\ P(W_1 = 0, W_2 = 0) \\ P(Y_1 = 1, Y_2 = 1, W_1 = 1, W_2 = 0) \\ P(Y_1 = 1, Y_2 = 0, W_1 = 1, W_2 = 0) \\ P(Y_1 = 0, Y_2 = 1, W_1 = 1, W_2 = 0) \\ P(Y_1 = 0, Y_2 = 0, W_1 = 1, W_2 = 0) \\ P(Y_1 = 1, Y_3 = 1, W_1 = 0, W_2 = 1) \\ P(Y_1 = 1, Y_3 = 0, W_1 = 0, W_2 = 1) \\ P(Y_1 = 0, Y_3 = 1, W_1 = 0, W_2 = 1) \\ P(Y_1 = 0, Y_3 = 0, W_1 = 0, W_2 = 1) \\ P(Y_1 = 1, W_1 = 0, W_2 = 0) \\ P(Y_1 = 0, W_1 = 0, W_2 = 0) \\ P(Y_2 = 1) \\ P(Y_2 = 0) \\ P(Y_3 = 1) \\ P(Y_3 = 0) \end{bmatrix} = C_2 P$$

where:

$$
C_2 = \begin{bmatrix}
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1\\
0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1\\
1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0\\
0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1\\
1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0\\
0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1
\end{bmatrix}
$$

As $rank\,(C_2) = 20$, there are 20 independent constraints. If there are no interaction terms in the model, then we obtain another 16 constraints. Therefore, we have 36 constraints on the 32 probabilities in P. We can keep a total number of 4 interaction terms in Equation (3.4) and Equation (3.5). However, in Equation (3.5), we cannot keep interaction terms that involve Y_3, as it is only observed when $W_2 = 1$ for the panel members, and is not observed along with Y_1 and Y_2 for subjects in the second refreshment sample. In addition, we cannot keep the three-way interactions in Equation (3.4) and Equation (3.5) if not all two-way interactions are kept. Therefore, we keep Y_1Y_2 and Y_1W_1 in Equation (3.4) and Equation (3.5) by specifying $\theta_7 = \theta_8 = 0$ in Equation (3.4) and $\lambda_8 = \ldots = \lambda_{16} = 0$ in Equation (3.5). We run a MCMC simulation to test the model and most parameters are well recovered. We use $200,000$ MCMC iterations and a burn-in sample of $100,000$. The details on the MCMC algorithm and the results are provided in Appendix B.

3.3 Monotone Missingness Pattern, With Follow-up For Refreshment Samples

In this section, we discuss the model selection issues under the 3-wave model with monotone missingness pattern and follow-up for subjects in refreshment samples. This model is specified similarly to the model discussed in Section 3.1, except that we also observe W_2 for subjects in the first refreshment sample, and further observe Y_3 for those with $W_2 = 1$. Figure 3.3 shows the pattern of the data.

<table>
<tr><th>Wave 1</th><th>Wave 2</th><th>Wave 3</th></tr>
<tr><td rowspan="3">Observe X, Y_1
(N_P)</td><td rowspan="2">Observe Y_2: $W_1 = 1$
(N_{CP})</td><td>Observe Y_3: $W_2 = 1$
(N_{CP2})</td></tr>
<tr><td>Y_3 missing: $W_2 = 0$
(N_{IP2})</td></tr>
<tr><td>Y_2 missing: $W_1 = 0$
(N_{IP})</td><td></td></tr>
<tr><td rowspan="3"></td><td rowspan="2">Observe X, Y_2
(refreshment sample)
(N_R)</td><td>Observe Y_3: $W_2 = 1$
(N_{CR2})</td></tr>
<tr><td>Y_3 missing: $W_2 = 0$
(N_{IR2})</td></tr>
<tr><td></td><td>Observe X, Y_3
(refreshment sample)
(N_{R2})</td></tr>
</table>

FIGURE 3.3: Graphical representation of the 3-wave model with monotone missingness pattern and follow-up for subjects in refreshment samples

Model selection of this model is similar to the case in Section 3.1. However, as subjects in the first refreshment sample are sampled again in wave 3, we can obtain more constraints from the presence of the N_{CR2} and N_{IR2} subjects. In addition to the probabilities in R_1, we also observe $P(Y_2 = 1, Y_3 = 1, W_2 = 1)$, $P(Y_2 = 1, Y_3 = 0, W_2 = 1)$, $P(Y_2 = 0, Y_3 = 1, W_2 = 1)$, and $(Y_2 = 0, Y_3 = 0, W_2 = 1)$ from the N_{CR2} subjects who are in the first refreshment sample and complete both wave 2 and wave 3. To see the number of independent constraints we have on P, we summarize these observable probabilities in vector R_3 and write it as a function of P:

$$R_3 \equiv \begin{bmatrix} P\left(Y_1=1,Y_2=1,Y_3=1,W_1=1,W_2=1\right) \\ P\left(Y_1=1,Y_2=1,Y_3=0,W_1=1,W_2=1\right) \\ P\left(Y_1=1,Y_2=0,Y_3=1,W_1=1,W_2=1\right) \\ P\left(Y_1=1,Y_2=0,Y_3=0,W_1=1,W_2=1\right) \\ P\left(Y_1=0,Y_2=1,Y_3=1,W_1=1,W_2=1\right) \\ P\left(Y_1=0,Y_2=1,Y_3=0,W_1=1,W_2=1\right) \\ P\left(Y_1=0,Y_2=0,Y_3=1,W_1=1,W_2=1\right) \\ P\left(Y_1=0,Y_2=0,Y_3=0,W_1=1,W_2=1\right) \\ P\left(W_1=1,W_2=1\right) \\ P\left(W_1=1,W_2=0\right) \\ P\left(W_1=0\right) \\ P\left(Y_1=1,Y_2=1,W_1=1,W_2=0\right) \\ P\left(Y_1=1,Y_2=0,W_1=1,W_2=0\right) \\ P\left(Y_1=0,Y_2=1,W_1=1,W_2=0\right) \\ P\left(Y_1=0,Y_2=0,W_1=1,W_2=0\right) \\ P\left(Y_1=1,W_1=0\right) \\ P\left(Y_1=0,W_1=0\right) \\ P\left(Y_2=1\right) \\ P\left(Y_2=0\right) \\ P\left(Y_3=1\right) \\ P\left(Y_3=0\right) \\ P\left(Y_2=1,Y_3=1,W_2=1\right) \\ P\left(Y_2=1,Y_3=0,W_2=1\right) \\ P\left(Y_2=0,Y_3=1,W_2=1\right) \\ P\left(Y_2=0,Y_3=0,W_2=1\right) \end{bmatrix} = C_3 P$$

where:

$$C_3 = \begin{bmatrix}
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1\\
1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0\\
0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1\\
1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0\\
0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1\\
1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0
\end{bmatrix}$$

As $rank\,(C_3) = 20$, there are 20 independent constraints. If there are no interaction terms in the model, then we obtain another 16 constraints. Therefore, we have 36 constraints on the 32 probabilities in P. We can keep a total number of 4 interaction terms in Equation (3.4) and Equation (3.5). Again, in Equation (3.5), we cannot keep interaction terms that involve Y_3, as it is only observed for the panel members when $W_2 = 1$, and is not observed along with Y_1 and Y_2 for subjects in the second refreshment sample. We cannot keep Y_1W_1 in Equation (3.4) and Equation (3.5), because Y_3 is observed only when $W_1 = 1$. In addition, we cannot keep the three-way interactions in both equations if not all two-way interactions are kept. Therefore, we keep Y_1Y_2 and Y_2W_1 in Equation (3.4) and Equation (3.5), specifying $\theta_6 = \theta_8 = 0$ in Equation (3.4) and $\lambda_7 = \lambda_9 = \ldots = \lambda_{16} = 0$ in Equation (3.5). We run a MCMC simulation to test the model and most parameters are well recovered. We use $200,000$ MCMC iterations and a burn-in sample of $100,000$. The details on the MCMC algorithm and the results are provided in Appendix B.

3.4 Non-Monotone Missingness Pattern, With Follow-up For Refreshment Samples

In this section, we discuss the model selection issues under the 3-wave model with non-monotone missingness pattern and follow-up for subjects in refreshment samples. This model has more information than the three models described above, as it combines the information in the model presented in Section 3.2 and that in Section 3.3. Figure 3.4 shows the pattern of the data.

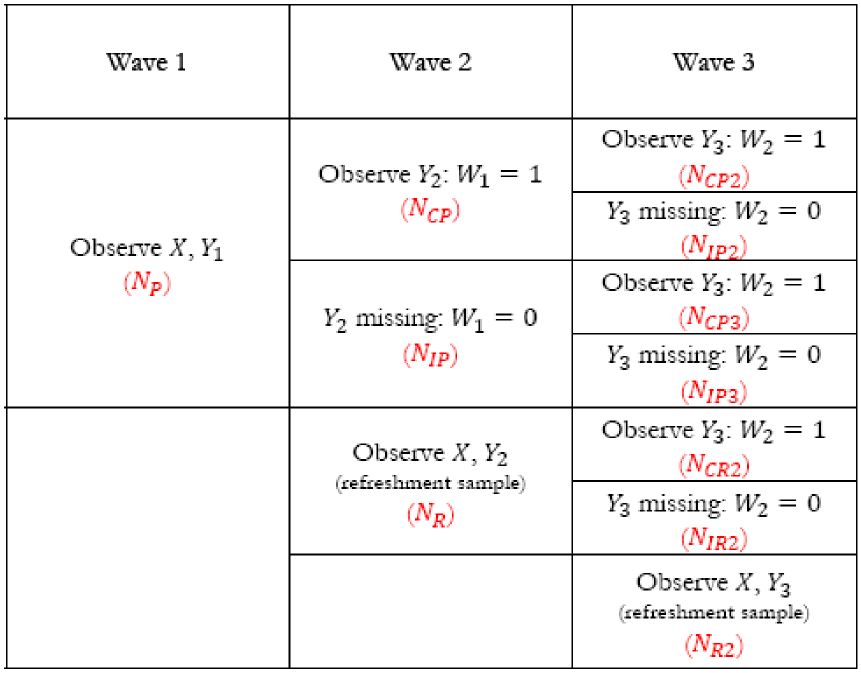

FIGURE 3.4: Graphical representation of the 3-wave model with non-monotone missingness pattern and follow-up for subjects in refreshment samples

Under this model, we combine information observed in the models in Section 3.2 and Section 3.3. To see the number of independent constraints we have on P, we summarize the observable probabilities in vector R_4 and write it as a function of P:

$$
R_4 \equiv \begin{bmatrix}
P\left(Y_1=1,Y_2=1,Y_3=1,W_1=1,W_2=1\right)\\
P\left(Y_1=1,Y_2=1,Y_3=0,W_1=1,W_2=1\right)\\
P\left(Y_1=1,Y_2=0,Y_3=1,W_1=1,W_2=1\right)\\
P\left(Y_1=1,Y_2=0,Y_3=0,W_1=1,W_2=1\right)\\
P\left(Y_1=0,Y_2=1,Y_3=1,W_1=1,W_2=1\right)\\
P\left(Y_1=0,Y_2=1,Y_3=0,W_1=1,W_2=1\right)\\
P\left(Y_1=0,Y_2=0,Y_3=1,W_1=1,W_2=1\right)\\
P\left(Y_1=0,Y_2=0,Y_3=0,W_1=1,W_2=1\right)\\
P\left(W_1=1,W_2=1\right)\\
P\left(W_1=1,W_2=0\right)\\
P\left(W_1=0,W_2=1\right)\\
P\left(W_1=0,W_2=0\right)\\
P\left(Y_1=1,Y_2=1,W_1=1,W_2=0\right)\\
P\left(Y_1=1,Y_2=0,W_1=1,W_2=0\right)\\
P\left(Y_1=0,Y_2=1,W_1=1,W_2=0\right)\\
P\left(Y_1=0,Y_2=0,W_1=1,W_2=0\right)\\
P\left(Y_1=1,Y_3=1,W_1=0,W_2=1\right)\\
P\left(Y_1=1,Y_3=0,W_1=0,W_2=1\right)\\
P\left(Y_1=0,Y_3=1,W_1=0,W_2=1\right)\\
P\left(Y_1=0,Y_3=0,W_1=0,W_2=1\right)\\
P\left(Y_1=1,W_1=0,W_2=0\right)\\
P\left(Y_1=0,W_1=0,W_2=0\right)\\
P\left(Y_2=1\right)\\
P\left(Y_2=0\right)\\
P\left(Y_3=1\right)\\
P\left(Y_3=0\right)\\
P\left(Y_2=1,Y_3=1,W_2=1\right)\\
P\left(Y_2=1,Y_3=0,W_2=1\right)\\
P\left(Y_2=0,Y_3=1,W_2=1\right)\\
P\left(Y_2=0,Y_3=0,W_2=1\right)
\end{bmatrix} = C_4 P
$$

where:

$$
C_4 = \begin{bmatrix}
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1\\
0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1\\
0&0&0&0&0&0&0&0&1&0&1&0&1&0&1&0&0&0&0&0&0&0&0&0&1&0&1&0&1&0&1&0\\
0&0&0&0&0&0&0&0&0&1&0&1&0&1&0&1&0&0&0&0&0&0&0&0&0&1&0&1&0&1&0&1\\
1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0
\end{bmatrix}
$$

As $rank\,(C_4) = 22$, there are 22 independent constraints. If there are no interaction terms in the model, then we obtain another 16 constraints. Therefore, we have 38 constraints on the 32 probabilities in P. We can keep a total number of 6 interaction terms in Equation (3.4) and Equation (3.5). Again, in Equation (3.5), we cannot keep interaction terms that involve Y_3, as it is only observed when $W_2 = 1$ for the panel members, and is not observed along with Y_1 and Y_2 for subjects in the second refreshment sample. In addition, we cannot keep the three-way interactions in Equation (3.4) and Equation (3.5) if not all two-way interactions are kept. Therefore, we keep the 3 two-way interaction terms Y_1Y_2, Y_1W_1 and Y_2W_1 in Equation (3.4) and Equation (3.5). We run a MCMC simulation to test the model and all parameters are well recovered. We use $200,000$ MCMC iterations and a burn-in sample of $100,000$. The details on the MCMC algorithm and the results are provided in Appendix B.

4

Application to APYN Poll

In this chapter, we apply the proposed model developed in Section 3.2 to a panel data set, the 2007-2008 Associated Press - Yahoo! News Poll (APYN). We also implement multiple imputation to create a complete dataset for the original panel only. We compare the results obtained under our proposed model with the estimation results based only on complete observations (i.e., without imputation for missing data) to illustrate the advantage of using refreshment samples.

4.1 Data Description

The APYN is a one year, 11-wave survey with 3 refreshment samples aimed to measure attitudes about the 2008 U.S. Presidential election and politics. The baseline data collection occurred prior to the start of the political primaries (November 2007), and the final data collection took place after the November 2008 general election. Wave 1 fielded 3,548 participants and obtained 2,735 respondents, for a 77.1% survey completion rate. All baseline respondents who were still available for survey sampling were fielded in follow-up waves, even if they had failed to respond to a previous wave (Kruse et al. (2009)). Participation rates for follow-up surveys varied from 69% to 87%, with rates decreasing towards the end of the panel. Refreshment samples were collected during wave 3 (697 completed), wave 6 (567 completed), and wave 9 (464 completed). Subjects in refreshment samples were not followed up in subsequent waves. Therefore, the model pattern is non-monotone missingness without follow-up to subjects in refreshment samples, which is in line with the model developed in Section 3.2.

We use only data collected in wave 1, wave 3 and wave 9 in our analysis. Wave 3 and wave 9 include respondents in both the original panel and the corresponding refreshment samples. The total numbers of panel members in wave 1, wave 3 and wave 9 are 2,735, 2,279 and 1,674 respectively. Our main interest is to investigate factors that affect political interest, and compare the results obtained under the proposed data imputation method with the results obtained using only complete observations (i.e., subjects that complete all three waves). We will use *CND1* as the measure of political interest. It measures how much thought a subject has given to candidates. [1] Table 4.1 summarizes the distribution of the answers in the three waves.

[1] The question is framed as "How much thought, if any, have you given to candidates who may be running for president in 2008?"

Table 4.1: Political interest in Wave 1, 3 and 9

Wave	P1	P3	P9	R3	R9
% Choose "A lot"	29.8	40.3	65.0	42.0	72.2
% Choose "Some"	48.6	44.3	25.9	43.3	20.3
% Choose "Not much"	15.3	10.8	5.8	10.2	5.0
% Choose "None at all"	6.1	4.4	2.9	3.6	1.9
% Not answered	0.2	0.2	0.5	0.9	0.7

"P" denotes the panel, and "R" denotes the refreshment sample in the corresponding wave.

4.2 Estimation Using Refreshment Samples

In this section, we address the question of interest using the model developed in Section 3.2. The notations are the same as in that section.

We first derive the variable of interest, Y_{ti}, from subject i's response to survey question *CND1*. $Y_{ti} = 1$ if subject i chooses "A lot" at time t, [2] and $Y_{ti} = 0$ otherwise. X_i include potential predictors for Y_{ti}. For this illustration, we use demographic variables because other variables tend to have many missing values themselves. Table 4.2 summarizes the independent variables.

[2] To keep notation simple and consistent, we denote $t = 1$ for wave 1, $t = 2$ for wave 3, and $t = 3$ for wave 9.

Table 4.2: Independent variables used in the model

Variable	Definition
AGE1	=1: 30-44; =0: otherwese;
AGE2	=1: 45-59; =0: otherwese;
AGE3	=1: above 60; =0: otherwese;
GENDER	=1: male; =0: female
COLLEGE	=1: has college degree; =0: otherwise
BLACK	=1: African American; =0: otherwise

For this illustration, we keep the data pattern consistent with the missingness pattern described in Section 3.2. We first remove all subjects with missingness in X_i's. We then remove subjects in the panel with Y_{1i} missing, subjects in the first refreshment sample with Y_{2i} missing, and subjects in the second refreshment sample with Y_{3i} missing. Table 4.3 displays the number of observations in the original panel and the two refreshment samples before and after this manipulation. After this manipulation, the numbers of observations for different groups corresponding to Figure 3.2 are: $N_P = 2730$, $N_{CP} = 2316$, $N_{IP} = 414$, $N_{CP2} = 1632$, $N_{IP2} = 684$, $N_{CP3} = 83$, $N_{IP3} = 331$, $N_R = 691$, $N_{R2} = 461$.

Table 4.3: Number of observations before and after the manipulation

	Original	After the manipulation
Wave 1, panel	2735	2730
Wave 3, refreshment sample	697	691
Wave 9, refreshment sample	464	461

Our model is specified in Equation (4.1) - (4.5):

$$Y_{1i} \mid \boldsymbol{X_i}, \boldsymbol{\beta} \sim Bern\left(\frac{exp(\boldsymbol{\beta X_i})}{1+exp(\boldsymbol{\beta X_i})}\right) \tag{4.1}$$

$$Y_{2i} \mid \boldsymbol{X_i}, Y_{1i}, \boldsymbol{\gamma} \sim Bern\left(\frac{exp(\boldsymbol{\gamma_1 X_i}+\gamma_2 Y_{1i})}{1+exp(\boldsymbol{\gamma_1 X_i}+\gamma_2 Y_{1i})}\right) \tag{4.2}$$

$$where\, \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \gamma_2)'$$

$$W_i \mid \boldsymbol{X_i}, Y_{1i}, Y_{2i}, \boldsymbol{\alpha} \sim Bern\left(\frac{exp(\boldsymbol{\alpha_1 X_i}+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}{1+exp(\boldsymbol{\alpha_1 X_i}+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}\right) \tag{4.3}$$

$$where\, \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \alpha_2, \alpha_3)'$$

$$Y_{3i} \mid \boldsymbol{X_i}, Y_{1i}, Y_{2i}, W_i, \boldsymbol{\theta}, \sim Bern\left(\frac{exp(\boldsymbol{\theta_1 X_i}+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_i+\theta_5 Y_1 Y_2+\theta_6 Y_1 W_1)}{1+exp(\boldsymbol{\theta_1 X_i}+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_i+\theta_5 Y_1 Y_2+\theta_6 Y_1 W_1)}\right), \tag{4.4}$$

$$where\, \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)'$$

$$W_{2i} \mid \boldsymbol{X_i}, Y_{1i}, Y_{2i}, Y_{3i}, W_i, \boldsymbol{\lambda} \sim Bern\left(\frac{exp(\boldsymbol{\lambda_1 X_i}+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_i+\lambda_6 Y_1 Y_2+\lambda_7 Y_1 W_1)}{1+exp(\boldsymbol{\lambda_1 X_i}+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_i+\lambda_6 Y_1 Y_2+\lambda_7 Y_1 W_1)}\right) \tag{4.5}$$

$$where\, \boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7)'$$

We implement the model developed in Section 3.2 and report the estimation results in Table 4.4-Table 4.8. We use $200,000$ MCMC iterations and a burn-in sample of $100,000$. The mean acceptance rates in the Metropolis-Hastings algorithms for β, γ, α, θ and λ are respectively 22.5%, 22.8%, 28.3%, 21.3% and 35.4%. Predictors that are significant are marked with "*".

Table 4.4: Estimates for determinants of Y_1

Variable	Mean	s.d.	2.5% quantile	97.5% quantile
CONSTANT	-1.60 *	0.17	-1.93	-1.28
AGE1	0.25	0.19	-0.11	0.62
AGE2	0.75 *	0.18	0.41	1.10
AGE3	1.26 *	0.18	0.91	1.61
COLLEGE	0.11	0.10	-0.09	0.30
GENDER	-0.05	0.10	-0.23	0.13
BLACK	0.75 *	0.13	0.49	1.00

Table 4.5: Estimates for determinants of Y_2

Variable	Mean	s.d.	2.5% quantile	97.5% quantile
CONSTANT	-1.72 *	0.24	-2.16	-1.25
AGE1	0.27	0.20	-0.12	0.66
AGE2	0.61 *	0.19	0.23	0.99
AGE3	0.94 *	0.20	0.55	1.33
COLLEGE	0.51 *	0.12	0.29	0.74
GENDER	-0.02	0.11	-0.23	0.18
BLACK	0.01	0.19	-0.36	0.37
Y_1	2.47 *	0.13	2.21	2.72

Table 4.6: Estimates for determinants of W_1

Variable	Mean	s.d.	2.5% quantile	97.5% quantile
CONSTANT	1.72 *	0.30	1.20	2.34
AGE1	-0.08	0.22	-0.52	0.35
AGE2	0.26	0.24	-0.20	0.73
AGE3	0.41	0.26	-0.10	0.92
COLLEGE	0.38 *	0.17	0.07	0.73
GENDER	0.13	0.14	-0.14	0.40
BLACK	-0.53 *	0.21	-0.92	-0.12
Y_1	0.64	0.35	-0.06	1.25
Y_2	-0.85	0.70	-2.14	0.43

Table 4.7: Estimates for determinants of Y_3

Variable	Mean	s.d.	2.5% quantile	97.5% quantile
CONSTANT	-0.44	0.65	-1.75	0.77
AGE1	0.04	0.22	-0.39	0.46
AGE2	0.17	0.22	-0.26	0.59
AGE3	0.90 *	0.23	0.44	1.36
COLLEGE	0.59 *	0.15	0.30	0.90
GENDER	-0.02	0.14	-0.30	0.24
BLACK	0.10	0.27	-0.42	0.64
Y_1	3.88 *	1.64	0.44	6.18
Y_2	2.08 *	0.23	1.63	2.55
W_1	-0.01	0.58	-1.05	1.18
Y_1Y_2	-0.43	0.45	-1.31	0.44
Y_1W_1	-2.43	1.57	-4.56	0.86

Table 4.8: Estimates for determinants of W_2

Variable	Mean	s.d.	2.5% quantile	97.5% quantile
CONSTANT	-1.62 *	0.29	-2.18	-1.02
AGE1	0.28	0.18	-0.07	0.63
AGE2	0.27	0.18	-0.08	0.62
AGE3	0.39 *	0.19	0.03	0.76
COLLEGE	0.56 *	0.13	0.31	0.81
GENDER	0.08	0.11	-0.13	0.29
BLACK	-0.11	0.19	-0.49	0.26
Y_1	0.86 *	0.37	0.18	1.63
Y_2	0.23	0.20	-0.19	0.60
Y_3	-1.01 *	0.53	-2.19	-0.08
W_1	2.50 *	0.17	2.17	2.85
Y_1Y_2	-0.04	0.28	-0.59	0.48
Y_1W_1	-0.62 *	0.30	-1.20	-0.07

There are two general conclusions that can be drawn from the estimation results: First, MAR does not hold for wave 9 responses, as the estimated coefficient of Y_3 in W_2 equation is significantly negative. This indicates that people with higher political interest are more likely to drop out in wave 9 of the survey. Second, past political interest is a good predictor for future political interest, as Y_1 is significantly positive in both Y_2 equation and Y_3 equation, and Y_2 is significantly positive in Y_3 equation.

4.3 Multiple Imputation Approach

We next investigate multiple imputation based approaches to releasing a complete original panel. Before doing so, we perform a brief simulation study to compare the one-step and two-step variance estimators for the 3-wave case.

4.3.1 Simulation Study Based on One-Step and Two-Step Multiple Imputation

We generate 100 datasets and run multiple imputation for each dataset. In each simulation, we first implement the estimation routine developed in Section 3.2 to obtain a set of parameter estimates, then implement the multiple imputation routine described in Equation (2.8) - (2.10). In our simulation, we take $m = 100$. Table 4.9 reports the results. In Table 4.9, column 3 reports the coverage rates of individual parameters with a Bayesian credible interval under multiple imputation. For purpose of comparison, we also fit the model using the two alternative methods discussed above for each of the 100 datasets. We first fit the model with MLE using complete observations, i.e., observations with $W_1 = W_2 = 1$. Column 4 reports the coverage rate of individual parameter estimates under this method. We can see this method performs significantly worse than our proposed model in recovering true parameter values. We then fit the model using MLE with original data before missingness was introduced for all units in the original panel. Column 5 reports the coverage rate of individual parameter estimates under this method. Comparing Column 3 and Column 5, we can see that our proposed model performs nearly as well as the best model where all information is observed. This confirms the effectiveness of our proposed model in recovering missing data.

Table 4.9: One-step multiple imputation results based on 100 simulations

	True value	Cov. rate of MI	Cov. rate of MLE with complete observations	Cov. rate of MLE assuming no missing	Mean of MI point estimates	Variance of MI point estimates	Mean of MI variances
β_0	.10	.98	.00	.98	.10	.0008	.0008
β_1	.50	.99	.94	.99	.51	.0017	.0017
γ_0	.10	1.00	.00	.92	.11	.0032	.0071
γ_1	.20	.94	.86	.91	.20	.0027	.0030
γ_2	.70	1.00	.78	.99	.70	.0026	.0036
θ_0	-.15	.99	.00	.99	-.10	.0121	.0315
θ_1	.30	1.00	.95	.98	.29	.0015	.0032
θ_2	-.55	.99	.36	.94	-.55	.0168	.0179
θ_3	.87	.98	.39	.97	.85	.0100	.0228
θ_4	.37	.93	.67	.98	.36	.0252	.0278

We then implement the two-step multiple imputation discussed in Equation (2.11) - (2.15) for the same 100 datasets. In our simulation, we take $m = 100$ and $n = 10$. Table 4.10 reports the results. Again, we see our proposed method significantly outperforms MLE implemented to complete observations. However, with this limited simulation, we do not find sufficient evidence that supports the superiority of the two-step multiple imputation routine over the standard (one-step) multiple imputation routine. So we will apply both multiple imputation methods in the APYN analysis, and compare the results with results obtained using only complete observations.

Table 4.10: Two-step multiple imputation results based on 100 simulations

	True value	Cov. rate of MI	Cov. rate of MLE with complete observations	Cov. rate of MLE assuming no missing	Mean of MI point estimates	Variance of MI point estimates	Mean of MI variances
β_0	.10	.98	.00	.98	.10	.0008	.0008
β_1	.50	.99	.94	.99	.51	.0017	.0017
γ_0	.10	1.00	.00	.92	.11	.0031	.0063
γ_1	.20	.87	.86	.91	.20	.0029	.0019
γ_2	.70	.94	.78	.99	.70	.0027	.0026
θ_0	-.15	.98	.00	.99	-.10	.0118	.0290
θ_1	.30	1.00	.95	.98	.29	.0015	.0022
θ_2	-.55	.92	.36	.94	-.56	.0164	.0128
θ_3	.87	.97	.39	.97	.85	.0100	.0184
θ_4	.37	.88	.67	.98	.36	.0254	.0196

4.3.2 Estimation Using Refreshment Samples And Multiple Imputation

We first apply the one-step multiple imputation routine described in Section 4.3.1 using $m = 100$ in our application to APYN data. We use $200,000$ MCMC iterations and a burn-in sample of $100,000$. The estimation results for Y_1, Y_2 and Y_3 are reported in Table 4.11-Table 4.13.

Table 4.11: Estimates for determinants of Y_1

Variable	Mean	s.d.	95% CI	P value
CONSTANT	-1.56	0.14	(-1.82, -1.29)	0.00
AGE1	0.19	0.16	(-0.11, 0.50)	0.22
AGE2	0.74	0.15	(0.45, 1.03)	0.00
AGE3	1.18	0.15	(0.88, 1.47)	0.00
COLLEGE	0.12	0.10	(-0.07, 0.30)	0.21
GENDER	-0.05	0.09	(-0.22, 0.12)	0.53
BLACK	0.72	0.15	(0.43, 1.00)	0.00

Table 4.12: Estimates for determinants of Y_2 using one-step multiple imputation

Variable	Mean	s.d.	95% CI	P value
CONSTANT	-1.66	0.22	(-2.11, -1.21)	0.00
AGE1	0.18	0.17	(-0.16, 0.52)	0.31
AGE2	0.62	0.16	(0.31, 0.92)	0.00
AGE3	0.80	0.16	(0.49, 1.11)	0.00
COLLEGE	0.50	0.11	(0.28, 0.72)	0.00
GENDER	-0.02	0.11	(-0.24, 0.20)	0.87
BLACK	-0.01	0.19	(-0.38, 0.36)	0.95
Y_1	2.48	0.14	(2.19, 2.77)	0.00

Table 4.13: Estimates for determinants of Y_3 using one-step multiple imputation

Variable	Estimate	s.d.	95% CI	P value
CONSTANT	-0.24	0.34	(-0.98, 0.50)	0.50
AGE1	-0.06	0.20	(-0.48, 0.35)	0.77
AGE2	0.04	0.21	(-0.38, 0.46)	0.85
AGE3	0.72	0.25	(0.21, 1.23)	0.01
COLLEGE	0.59	0.17	(0.23, 0.94)	0.00
GENDER	-0.10	0.11	(-0.31, 0.11)	0.36
BLACK	0.02	0.25	(-0.49, 0.53)	0.94
Y_1	1.45	0.28	(0.89, 2.02)	0.00
Y_2	2.03	0.19	(1.64, 2.41)	0.00
Y_1Y_2	-0.36	0.36	(-1.07, 0.36)	0.33

We next apply the two-step multiple imputation with $m = 100$ and $n = 10$. The estimation results for Y_2 and Y_3 are reported in Table 4.14 and Table 4.15 below. The estimation results for Y_1 are the same as in Table 4.11.

Table 4.14: Estimates for determinants of Y_2 using two-step multiple imputation

Variable	Mean	s.d.	95% CI	P value
CONSTANT	-1.62	0.22	(-2.06, -1.18)	0.00
AGE1	0.18	0.18	(-0.17, 0.52)	0.32
AGE2	0.60	0.17	(0.26, 0.93)	0.00
AGE3	0.78	0.18	(0.43, 1.14)	0.00
COLLEGE	0.48	0.11	(0.26, 0.71)	0.00
GENDER	-0.01	0.10	(-0.20, 0.18)	0.92
BLACK	-0.05	0.20	(-0.44, 0.34)	0.80
Y_1	2.43	0.13	(2.18, 2.69)	0.00

Table 4.15: Estimates for determinants of Y_3 using two-step multiple imputation

Variable	Estimate	s.d.	95% CI	P value
CONSTANT	-0.23	0.23	(-0.68, 0.23)	0.33
AGE1	-0.02	0.18	(-0.37, 0.34)	0.93
AGE2	0.12	0.19	(-0.25, 0.50)	0.53
AGE3	0.78	0.21	(0.37, 1.18)	0.00
COLLEGE	0.57	0.13	(0.31, 0.84)	0.00
GENDER	-0.11	0.12	(-0.34, 0.13)	0.37
BLACK	0.06	0.23	(-0.40, 0.52)	0.78
Y_1	1.38	0.26	(0.87, 1.88)	0.00
Y_2	1.57	0.24	(1.10, 2.04)	0.00
Y_1Y_2	0.07	0.42	(-0.77, 0.90)	0.88

Based on the estimates, we can draw the following general insights regarding political interest: First, elder people tend to have higher political interest, as shown by the positively significant coefficients of AGE2 and AGE3 in Y_1 equation and Y_2 equation as well as the positively significant coefficient of AGE3 in Y_3 equation. Second, people with higher education background tend to have higher political interest, as shown by the positively significant coefficient of COLLEGE in Y_2 equation and Y_3 equation. Third, past political interest is a good predictor on future political interest, as Y_1 is significantly positive in both Y_2 equation and Y_3 equation, and Y_2 is significantly positive in Y_3 equation.

Table 4.16 compares the means of Y_1, Y_2 and Y_3 in the original panel based on the complete case and the imputed data (using both one-step and two-step multiple imputation). As we can see, the means of Y_3 are higher in the imputed datasets than in the complete case. Actually, the mean of Y_3 in the refreshment sample collected in wave 9 is 0.722, much higher than that in the complete case (0.651). This explains why the imputed datasets have higher means on Y_3.

Table 4.16: Confidence intervals for the mean of Y_1, Y_2 and Y_3 in the original panel

	Complete case		Imputed data (one-step)		Imputed data (two-step)	
	Point est.	Interval	Point est.	Interval	Point est.	Interval
Y_1	0.298	(0.281, 0.315)	0.298	(0.281, 0.315)	0.298	(0.281, 0.315)
Y_2	0.407	(0.389, 0.425)	0.420	(0.375, 0.466)	0.423	(0.382, 0.464)
Y_3	0.651	(0.633, 0.669)	0.698	(0.626, 0.771)	0.697	(0.654, 0.739)

4.3.3 Estimation With Complete Observations

As we mentioned in Section 4.2, MAR does not hold for wave 9 responses. Therefore, treating the data as MAR and ignoring the attrition bias would lead to inference problems, as shown in simulation studies in Section 2. To see how our results differ if we treat the missingness as being completely random, we drop all subjects with missingness in Y_{1i}, Y_{2i} or Y_{3i}, and re-estimate the model using the $N_{CP2} = 1633$ subjects with complete observations (balanced panel). We estimate the model using MLE and report the estimation results for Y_2 and Y_3 in Table 4.17 and Table 4.18. The estimation results for Y_1 are the same as in Table 4.11.

Table 4.17: MLE estimates for determinants of Y_2 with complete observations

Variable	Estimate	s.d.	95% CI	P value
CONSTANT	-2.05	0.15	(-2.34, -1.76)	0.00
AGE1	0.16	0.16	(-0.16, 0.48)	0.32
AGE2	0.63	0.16	(0.33, 0.94)	0.00
AGE3	0.79	0.16	(0.48, 1.11)	0.00
COLLEGE	0.57	0.10	(0.38, 0.77)	0.00
GENDER	0.03	0.09	(-0.16, 0.21)	0.79
BLACK	-0.30	0.17	(-0.63, 0.04)	0.08
Y_1	2.21	0.10	(2.02, 2.41)	0.00

Table 4.18: MLE estimates for determinants of Y_3 with complete observations

Variable	Estimate	s.d.	95% CI	P value
CONSTANT	-1.44	0.13	(-1.69, -1.19)	0.00
AGE1	0.11	0.14	(-0.17, 0.38)	0.45
AGE2	0.20	0.14	(-0.07, 0.47)	0.15
AGE3	0.56	0.14	(0.28, 0.84)	0.00
COLLEGE	0.67	0.09	(0.49, 0.85)	0.00
GENDER	0.01	0.09	(-0.16, 0.17)	0.95
BLACK	-0.18	0.16	(-0.48, 0.13)	0.25
Y_1	0.69	0.15	(0.41, 0.98)	0.00
Y_2	1.37	0.12	(1.13, 1.61)	0.00
Y_1Y_2	-0.30	0.20	(-0.69, 0.01)	0.14

Comparing Table 4.17-4.18 with Table 4.12-4.15, we can see the main differences lie in estimates for the Y_3 equation. Clearly, the coefficients of Y_1 and Y_2 are smaller in Table 4.18 than in Table 4.13 and Table 4.15. This means that the carry-on effects of past political interest on current political interest are underestimated when the attrition bias is not accounted for.

5

Conclusion

The Additive Non-ignorable model proposed by Hirano et al. (2001) utilizes refreshment samples in panel surveys, and offers flexibility in modeling the missing data mechanism. In this thesis, we extend this model by incorporating multiple refreshment samples. We address model selection and estimation issues related to our proposed model under four different survey designs.

By applying the one-step (Rubin (1987)) and two-step multiple imputation (Reiter (2008)) routines to the 2-wave model, we find that one-step multiple imputation tends to lead to positive biased variance estimators in the presence of single refreshment sample. However, under this specific case, the superiority of the two-step multiple imputation over the one-step multiple imputation is not clear.

To demonstrate the applicability of our proposed approach, we apply our approach along with multiple imputation to the APYN 2008 Election panel. The simulation studies applying the one-step and two-step multiple imputation to the

3-wave model show that our proposed model performs nearly as well as the first-best model where there is no missingness. This confirms the effectiveness of our proposed model in recovering missing data. The advantage of two-step multiple imputation over the standard one-step multiple imputation, however, is again not obvious. This is possibly because of the binary variables we use. Hence we apply both multiple imputation methods in our real data analysis that explores factors affecting people's political interest. We find that when attrition bias is not accounted for, the carry-on effects of past political interest on current political interest are underestimated. This again highlights the importance of dealing with attrition bias through appropriate ways, and our proposed approach provides one possible solution.

One of the limitations of our current model is that it can only deal with binary missing variables. Our next step of work is to extend this model to allow for multivariate missing variables and to deal with issues related to missingness in the original panel and refreshment samples.

Appendix A

MCMC Algorithm In Chapter 2

We use flat priors on β, α and γ.

Step 1 For units with $W_{1i} = 0$, impute Y_{2i}:

$P\left(Y_{2i} \mid X_i,\, Y_{1i}, W_{1i} = 0, \gamma, \alpha, \beta\right)$

$\propto P\left(Y_{2i} \mid X_i,\, Y_{1i}, \gamma\right) P\left(W_{1i} = 0 \mid X_i, Y_{1i}, Y_{2i}, \alpha\right)$

$\propto \left[\frac{exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}\right]^{Y_{2i}} \left[\frac{1}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}\right]^{1-Y_{2i}} \frac{1}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}$

$\Rightarrow$

$\frac{P(Y_{2i}=1|X_i,Y_{1i},W_{1i}=0,\gamma,\alpha,\beta)}{P(Y_{2i}=0|X_i,Y_{1i},W_{1i}=0,\gamma,\alpha,\beta)} = exp\left(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_{1i}\right) \frac{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i})}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3)}$

$\equiv q_1$

Hence $P\left(Y_{2i} = 1 \mid X_i,\, Y_{2i}, W_{1i}, \gamma, \alpha, \beta\right) = \frac{q_1}{1+q_1}$.

Step 2 For units in the refreshment sample, impute Y_{1i}:

$P\left(Y_{1i} \mid X_i,\, Y_{2i}, W_{1i}, \gamma, \alpha\right)$

$\propto P\left(Y_{1i} \mid X_i, \beta\right) P\left(Y_{2i} \mid X_i,\, Y_{1i}, \gamma\right) P\left(W_i \mid X_i, Y_{1i}, Y_{2i}, \alpha\right)$

$\propto \left[\frac{exp(\beta_0+\beta_1 X_i)}{1+exp(\beta_0+\beta_1 X_i)}\right]^{Y_{1i}} \left[\frac{1}{1+exp(\beta_0+\beta_1 X_i)}\right]^{1-Y_{1i}}$

$\times \left[\frac{exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}\right]^{Y_{2i}} \left[\frac{1}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}\right]^{1-Y_{2i}}$

$\times \left[\frac{exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}\right]^{W_{1i}} \left[\frac{1}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}\right]^{1-W_{1i}}$

$\Rightarrow$

$\frac{P(Y_{1i}=1|X_i, Y_{2i}, W_{1i}, \gamma, \alpha, \beta)}{P(Y_{1i}=0|X_i, Y_{2i}, W_{1i}, \gamma, \alpha, \beta)}$

$= exp\left(\beta_0 + \beta_1 X_i + \gamma_2 Y_{2i} + \alpha_2 W_{1i}\right) \times \frac{1+exp(\gamma_0+\gamma_1 X_i)}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2)} \times \frac{1+exp(\alpha_0+\alpha_1 X_i+\alpha_3 Y_{2i})}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2+\alpha_3 Y_{2i})}$

$\equiv q_2$

Hence $P\left(Y_{1i} = 1 \mid X_i,\, Y_{2i}, W_{1i}, \gamma, \alpha, \beta\right) = \frac{q_2}{1+q_2}$.

Step 3 For units in the refreshment sample, impute W_{1i}:

$P\left(W_{1i} = 1 \mid X_i,\, Y_{1i}, Y_{2i}, \gamma, \alpha, \beta\right) = \frac{exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}$.

Step 4 Drawβ using a symmetric random-walk Metropolis-Hastings algorithm:

$P\left(\beta \mid X_i,\, Y_{1i}, Y_{2i}, W_{1i}, \alpha, \gamma\right)$

$\propto \prod_{i=1}^{N} \frac{[exp(\beta_0+\beta_1 X_i)]^{Y_{1i}}}{1+exp(\beta_0+\beta_1 X_i)}$

The standard deviation of the proposal distribution is σ_β.

We accept the proposal value β^{new} with the probability:

$pr = min\left\{1,\, \prod_{i=1}^{N} \left(\frac{P(\beta^{new}|X_i, Y_{1i}, Y_{2i}, W_{1i}, \alpha, \gamma)}{P\left(\beta^{old}|X_i, Y_{1i}, Y_{2i}, W_{1i}, \alpha, \gamma\right)}\right)\right\}$

Step 5 Draw γ using a symmetric random-walk Metropolis-Hastings algorithm:

$P\left(\gamma \mid X_i,\, Y_{1i}, Y_{2i}, W_{1i}, \alpha, \beta\right)$

$\propto \prod_{i=1}^{N} \frac{[exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})]^{Y_{2i}}}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}$

The standard deviation of the proposal distribution is σ_γ.

We accept the proposal value γ^{new} with the probability:

$$pr = min\left\{1,\ \prod_{i=1}^{N} \left(\frac{P(\gamma^{new}|X_i, Y_{1i}, Y_{2i}, W_{1i}, \alpha, \beta)}{P\left(\gamma^{old}|X_i, Y_{1i}, Y_{2i}, W_{1i}, \alpha, \beta\right)}\right)\right\}$$

Step 6 Draw α using a symmetric random-walk Metropolis-Hastings algorithm:

$P\left(\alpha \mid X_i,\ Y_{1i}, Y_{2i}, W_{1i}, \gamma, \beta\right)$

$\propto \prod_{i=1}^{N} \frac{[exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})]^{W_{1i}}}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}$

The standard deviation of the proposal distribution is σ_α.

We accept the proposal value α^{new} with the probability:

$$pr = min\left\{1,\ \prod_{i=1}^{N} \left(\frac{P(\alpha^{new}|X_i, Y_{1i}, Y_{2i}, W_{1i}, \gamma, \beta)}{P\left(\alpha^{old}|X_i, Y_{1i}, Y_{2i}, W_{1i}, \gamma, \beta\right)}\right)\right\}$$

In the simulation, $N_p = 10,000$, $N_R = 5,000$. We use $10,000$ MCMC iterations and a burn-in sample of $5,000$. We take $\sigma_\beta = \sigma_\gamma = \sigma_\alpha = 0.05$. The mean acceptance rates in the Metropolis-Hastings algorithms for β, γ and α are respectively 31.0%, 28.0% and 25.7%. The results are presented in Table A.1.

Table A.1: Simulation of the 2-wave model

Parameter	True value	Estimates	s.d.	2.5% quantile	97.5% quantile
β_0	0.3	0.32	0.03	0.26	0.38
β_1	-0.4	-0.42	0.04	-0.50	-0.34
γ_0	0.3	0.28	0.05	0.17	0.38
γ_1	-0.3	-0.33	0.05	-0.41	-0.24
γ_2	0.7	0.71	0.06	0.60	0.82
α_0	-0.4	-0.39	0.07	-0.52	-0.23
α_1	1.0	1.00	0.06	0.87	1.08
α_2	-0.7	-0.76	0.06	-0.87	-0.64
α_3	1.3	1.31	0.13	1.05	1.55

Appendix B

MCMC Algorithm In Chapter 3

Step 1 For the N_{IP} units with $W_{1i} = 0$ as well as the N_{R2} units in the second refreshment sample, impute Y_{2i}:

$$P\left(Y_{2i} \mid X_i,\, Y_{1i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda\right)$$

$$\propto P\left(Y_{2i} \mid X_i,\, Y_{1i}, \gamma\right) P\left(W_{1i} \mid X_i, Y_{1i}, Y_{2i}, \alpha\right) P\left(Y_{3i} \mid X_i, Y_{1i}, Y_{2i}, W_{1i}, \theta\right)$$

$$\times P\left(W_{2i} \mid X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, \lambda\right)$$

$$\propto \frac{[exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})]^{Y_{2i}}}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})} \times \frac{[exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})]^{W_{1i}}}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}$$

$$\times \frac{[exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_{1i} Y_{2i}+\theta_6 Y_{1i} W_{1i}+\theta_7 Y_{2i} W_{1i})]^{Y3i}}{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_{1i} Y_{2i}+\theta_6 Y_{1i} W_{1i}+\theta_7 Y_{2i} W_{1i})}$$

$$\times \frac{[exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})]^{W_{2i}}}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})}$$

$$\Rightarrow$$

$$\frac{P(Y_{2i}=1|X_i, Y_{1i}, Y_{3i}, W_i, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda)}{P(Y_{2i}=0|X_i, Y_{1i}, Y_{3i}, W_i, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda)}$$

$$= exp\left(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_{1i} + \alpha_3 W_{1i} + \theta_3 Y_{3i} + \theta_5 Y_{1i} Y_{3i} + \theta_7 W_{1i} Y_{3i} + \lambda_3 W_{2i} + \lambda_6 Y_{1i} W_{2i} + \lambda_8 W_{1i} W_{2i}\right)$$

$$\times \frac{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i})}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3)}$$

$$\times \frac{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_4 W_{1i}+\theta_6 Y_{1i}W_{1i})}{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3+\theta_4 W_{1i}+\theta_5 Y_{1i}+\theta_6 Y_{1i}W_{1i}+\theta_7 W_{1i})}$$

$$\times \frac{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_7 Y_{1i}W_{1i})}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i}+\lambda_7 Y_{1i}W_{1i}+\lambda_8 W_{1i})}$$

$$\equiv q_1$$

Hence $P\left(Y_{2i}=1 \mid X_i,\, Y_{1i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda\right) = \frac{q_1}{1+q_1}$.

Step 2 For the N_R units in the first refreshment sample as well as the N_{R2} units in the second refreshment sample, impute Y_{1i}:

$$P\left(Y_{1i} \mid X_i,\, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda\right)$$

$$\propto P\left(Y_{1i} \mid X_i, \beta\right) P\left(Y_{2i} \mid X_i,\, Y_{1i}, \gamma\right) P\left(W_{1i} \mid X_i, Y_{1i}, Y_{2i}, \alpha\right) P\left(Y_{3i} \mid X_i, Y_{1i}, Y_{2i}, W_{1i}, \theta\right)$$

$$\times P\left(W_{2i} \mid X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, \lambda\right)$$

$$\propto \frac{[exp(\beta_0+\beta_1 X_i)]^{Y_{1i}}}{1+exp(\beta_0+\beta_1 X_i)} \times \frac{[exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})]^{Y_{2i}}}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}$$

$$\times \frac{[exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})]^{W_{1i}}}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}$$

$$\times \frac{[exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_{1i}Y_{2i}+\theta_6 Y_{1i}W_{1i}+\theta_7 Y_{2i}W_{1i})]^{Y_{3i}}}{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_{1i}Y_{2i}+\theta_6 Y_{1i}W_{1i}+\theta_7 Y_{2i}W_{1i})}$$

$$\times \frac{[exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i}Y_{2i}+\lambda_7 Y_{1i}W_{1i}+\lambda_8 Y_{2i}W_{1i})]^{W_{2i}}}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i}Y_{2i}+\lambda_7 Y_{1i}W_{1i}+\lambda_8 Y_{2i}W_{1i})}$$

$$\Rightarrow$$

$$\frac{P(Y_{1i}=1|X_i, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda)}{P(Y_{1i}=0|X_i, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda)}$$

$$= exp\left(\beta_0 + \beta_1 X_i + \gamma_2 Y_{2i} + \alpha_2 W_{1i} + \theta_2 Y_{3i} + \theta_5 Y_{2i} Y_{3i} + \theta_6 W_{1i} Y_{3i} + \lambda_2 W_{2i} + \lambda_6 Y_{2i} W_{2i} + \lambda_7 W_{1i} W_{2i}\right)$$

$$\times \frac{1+exp(\gamma_0+\gamma_1 X_i)}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2)} \times \frac{1+exp(\alpha_0+\alpha_1 X_i+\alpha_3 Y_{2i})}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2+\alpha_3 Y_{2i})}$$

$$\times \frac{1+exp(\theta_0+\theta_1 X_i+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_7 Y_{2i}W_{1i})}{1+exp(\theta_0+\theta_1 X_i+\theta_2+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_{5i} Y_{2i}+\theta_{6i} W_{1i}+\theta_7 Y_{2i}W_{1i})}$$

$$\times \frac{1+exp(\lambda_0+\lambda_1 X_i+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_8 Y_{2i}W_{1i})}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{2i}+\lambda_7 W_{1i}+\lambda_8 Y_{2i}W_{1i})}$$

$$\equiv q_2$$

Hence $P\left(Y_{1i}=1 \mid X_i,\, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda\right) = \frac{q_2}{1+q_2}$.

Step 3 For the N_R units in the first refreshment sample as well as the N_{R2} units in the second refreshment sample, impute W_{1i}:

$$P\left(W_{1i} \mid X_i, Y_{1i} Y_{2i}, Y_{3i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda\right)$$

$$\propto P\left(W_{1i} \mid X_i, Y_{1i}, Y_{2i}, \alpha\right) P\left(Y_{3i} \mid X_i, Y_{1i}, Y_{2i}, W_{1i}, \theta\right) P\left(W_{2i} \mid X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, \lambda\right)$$

$$\propto \frac{[exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})]^{W_{1i}}}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}$$

$$\times \frac{[exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_{1i} Y_{2i}+\theta_6 Y_{1i} W_{1i}+\theta_7 Y_{2i} W_{1i})]^{Y_{3i}}}{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_{1i} Y_{2i}+\theta_6 Y_{1i} W_{1i}+\theta_7 Y_{2i} W_{1i})}$$

$$\times \frac{[exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})]^{W_{2i}}}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})}$$

$$\Rightarrow$$

$$\frac{P(W_{1i}=1|X_i,Y_{1i}\,Y_{2i},Y_{3i},W_{2i},\gamma,\alpha,\beta,\theta,\lambda)}{P(W_{1i}=0|X_i,Y_{1i}\,Y_{2i},Y_{3i},W_{2i},\gamma,\alpha,\beta,\theta,\lambda)}$$

$$= exp\left(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i}+\theta_4 Y_{3i}+\theta_6 Y_{1i} Y_{3i}+\theta_7 Y_{2i} Y_{3i}+\lambda_5 W_{2i}+\lambda_7 Y_{1i} W_{2i}+\lambda_8 Y_{2i} W_{2i}\right)$$

$$\times \frac{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_5 Y_{1i} Y_{2i})}{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4+\theta_5 Y_{1i} Y_{2i}+\theta_6 Y_{1i}+\theta_7 Y_{2i})}$$

$$\times \frac{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_6 Y_{1i} Y_{2i})}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i}+\lambda_8 Y_{2i})}$$

$$\equiv q_3$$

Hence $P\left(W_{1i}=1 \mid X_i, Y_{1i} Y_{2i}, Y_{3i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda\right) = \frac{q_3}{1+q_3}$.

Step 4 Impute Y_{3i} for:

1) the N_{IP2} units with $W_{1i}=1$ and $W_{2i}=0$, the N_{IP} units with $W_{1i}=0$, as well as the N_R units in the first refreshment sample for Chapter 3.1.

2) the N_{IP2} units with $W_{1i}=1$ and $W_{2i}=0$, the N_{IP3} units with $W_{1i}=0$ and $W_{2i}=0$, as well as the N_R units in the first refreshment sample for Chapter 3.2.

3) the N_{IP2} units with $W_{1i}=1$ and $W_{2i}=0$, the N_{IP} units with $W_{1i}=0$, as well as the N_{IR2} units in the first refreshment sample with $W_{2i}=0$ for Chapter 3.3.

4) the N_{IP2} units with $W_{1i}=1$ and $W_{2i}=0$, the N_{IP3} units with $W_{1i}=0$ and

$W_{2i} = 0$, as well as the N_{IR2} units in the first refreshment sample with $W_{2i} = 0$ for Chapter 3.4.

$$P\left(Y_{3i} \mid X_i,\, Y_{1i}, Y_{2i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda\right)$$

$$\propto P\left(Y_{3i} \mid X_i, Y_{1i}, Y_{2i}, W_{1i}, \theta\right) P\left(W_{2i} \mid X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_i, \lambda\right)$$

$$\propto \frac{\left[exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_1 Y_2+\theta_6 Y_1 W_1+\theta_7 Y_2 W_1)\right]^{Y_{3i}}}{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_1 Y_2+\theta_6 Y_1 W_1+\theta_7 Y_2 W_1)}$$

$$\times \frac{\left[exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})\right]^{W_{2i}}}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})}$$

$$\Rightarrow$$

$$\frac{P(Y_{3i}=1 | X_i, Y_{1i}, Y_{2i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda)}{P(Y_{3i}=0 | X_i, Y_{1i}, Y_{2i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda)}$$

$$= exp\left(\theta_0 + \theta_1 X_i + \theta_2 Y_{1i} + \theta_3 Y_{2i} + \theta_4 W_{1i} + \theta_5 Y_1 Y_2 + \theta_6 Y_1 W_1 + \theta_7 Y_2 W_1 + \lambda_4 W_{2i}\right)$$

$$\times \frac{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})}$$

$$\equiv q_4$$

Hence $P\left(Y_{3i} = 1 \mid X_i,\, Y_{1i}, Y_{2i}, W_{1i}, W_{2i}, \gamma, \alpha, \beta, \theta, \lambda\right) = \frac{q_4}{1+q_4}$.

Step 5 Impute W_{2i} for:

1) the N_{IP} units with $W_{1i} = 1$, the N_{IP} units with $W_{1i} = 0$, the N_R units in the first refreshment sample and the N_{R2} units in the second refreshment sample for Chapter 3.1.

2) the N_R units in the first refreshment sample and the N_{R2} units in the second refreshment sample for Chapter 3.2.

3) the N_{IP} units with $W_{1i} = 1$, and the N_{R2} units in the second refreshment sample for Chapter 3.3.

4) the N_{R2} units in the second refreshment sample for Chapter 3.4.

$$P\left(W_{2i} = 1 \mid X_i,\, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, \gamma, \alpha, \beta, \theta, \lambda\right)$$

$$= \frac{exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i}Y_{2i}+\lambda_7 Y_{1i}W_{1i}+\lambda_8 Y_{2i}W_{1i})}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i}Y_{2i}+\lambda_7 Y_{1i}W_{1i}+\lambda_8 Y_{2i}W_{1i})}$$

Step 6 Drawβ using a symmetric random-walk Metropolis-Hastings algorithm:

$$P\left(\beta \mid X_i,\, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \theta, \lambda\right)$$

$$\propto \prod_{i=1}^{N} \frac{[exp(\beta_0+\beta_1 X_i)]^{Y_{1i}}}{1+exp(\beta_0+\beta_1 X_i)}$$

The standard deviation of the proposal distribution is σ_β.

We accept the proposal value β^{new} with the probability:

$$pr = min\left\{1,\, \prod_{i=1}^{N}\left(\frac{P(\beta^{new}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \theta, \lambda)}{P\left(\beta^{old}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \gamma, \alpha, \theta, \lambda\right)}\right)\right\}$$

Step 7 Draw γ using a symmetric random-walk Metropolis-Hastings algorithm:

$$P\left(\gamma \mid X_i,\, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \alpha, \beta, \theta, \lambda\right)$$

$$\propto \prod_{i=1}^{N} \frac{[exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})]^{Y_{2i}}}{1+exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_{1i})}$$

The standard deviation of the proposal distribution is σ_γ.

We accept the proposal value β^{new} with the probability:

$$pr = min\left\{1,\, \prod_{i=1}^{N}\left(\frac{P(\gamma^{new}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \alpha, \beta, \theta, \lambda)}{P\left(\gamma^{old}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \alpha, \beta, \theta, \lambda\right)}\right)\right\}$$

Step 8 Draw α using a symmetric random-walk Metropolis-Hastings algorithm:

$$P\left(\alpha \mid X_i,\, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \theta, \lambda\right)$$

$$\propto \prod_{i=1}^{N} \frac{[exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})]^{W_{1i}}}{1+exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_{1i}+\alpha_3 Y_{2i})}$$

The standard deviation of the proposal distribution is σ_α.

We accept the proposal value α^{new} with the probability:

$$pr = min\left\{1,\, \prod_{i=1}^{N}\left(\frac{P(\alpha^{new}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \theta, \lambda)}{P\left(\alpha^{old}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \theta, \lambda\right)}\right)\right\}$$

Step 9 Draw θ using a symmetric random-walk Metropolis-Hastings algorithm:

$$P\left(\theta \mid X_i,\, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \alpha, \lambda\right)$$

$$\propto \prod_{i=1}^{N} \frac{[exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_1 Y_2+\theta_6 Y_1 W_1+\theta_7 Y_2 W_1)]^{Y_{3i}}}{1+exp(\theta_0+\theta_1 X_i+\theta_2 Y_{1i}+\theta_3 Y_{2i}+\theta_4 W_{1i}+\theta_5 Y_1 Y_2+\theta_6 Y_1 W_1+\theta_7 Y_2 W_1)}$$

The standard deviation of the proposal distribution is σ_θ.

We accept the proposal value θ^{new} with the probability:

$$pr = min\left\{1,\, \prod_{i=1}^{N}\left(\frac{P(\theta^{new}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \alpha, \lambda)}{P\left(\theta^{old}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \alpha, \lambda\right)}\right)\right\}$$

Step 10 Draw λ using a symmetric random-walk Metropolis-Hastings algorithm:

$$P\left(\lambda \mid X_i,\, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \alpha, \theta\right)$$

$$\propto \prod_{i=1}^{N} \frac{[exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})]^{W_{2i}}}{1+exp(\lambda_0+\lambda_1 X_i+\lambda_2 Y_{1i}+\lambda_3 Y_{2i}+\lambda_4 Y_{3i}+\lambda_5 W_{1i}+\lambda_6 Y_{1i} Y_{2i}+\lambda_7 Y_{1i} W_{1i}+\lambda_8 Y_{2i} W_{1i})}$$

The standard deviation of the proposal distribution is σ_λ.

We accept the proposal value λ^{new} with the probability:

$$pr = min\left\{1,\, \prod_{i=1}^{N}\left(\frac{P(\lambda^{new}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \alpha, \theta)}{P\left(\lambda^{old}|X_i, Y_{1i}, Y_{2i}, Y_{3i}, W_{1i}, W_{2i}, \beta, \gamma, \alpha, \theta\right)}\right)\right\}$$

In the simulations below, $N_p = 10,000$, $N_R = N_{R2} = 2,500$.

Estimation Results for Chapter 3.1

We take $\sigma_\beta = \sigma_\gamma = \sigma_\alpha = 0.05$ and $\sigma_\beta = \sigma_\gamma = 0.1$. The mean acceptance rates in the Metropolis-Hastings algorithms for β, γ, α, θ and λ are respectively 37.1%, 28.2%, 28.0%, 23.6% and 25.4%. We use $200,000$ MCMC iterations and a burn-in sample of $100,000$. The results are presented in Table B.1.

Table B.1: Simulation of the 3-wave model with monotone missingness pattern and no follow-up for subjects in refreshment samples

Parameter	True value	Estimates	s.d.	2.5% quantile	97.5% quantile
β_0	0.1	0.18	0.03	0.12	0.24
β_1	0.5	0.43	0.04	0.35	0.52
γ_0	0.1	0.03	0.06	-0.10	0.14
γ_1	0.2	0.14	0.05	0.05	0.24
γ_2	0.7	0.79	0.06	0.68	0.90
α_0	0.1	-0.04	0.08	-0.19	0.11
α_1	-0.3	-0.29	0.06	-0.40	-0.18
α_2	0.7	0.65	0.06	0.54	0.77
α_3	1.0	1.30	0.17	0.98	1.65
θ_0	0.1	0.10	0.14	-0.18	0.36
θ_1	0.3	0.30	0.06	0.19	0.41
θ_2	-1.0	-0.97	0.14	-1.25	-0.71
θ_3	1.0	0.96	0.13	0.70	1.21
θ_4	0	-	-	-	-
θ_5	0.2	0.18	0.15	-0.11	0.50
θ_6	0	-	-	-	-
θ_7	0	-	-	-	-
λ_0	0.1	0.34	0.17	0.03	0.69
λ_1	0.2	0.19	0.07	0.05	0.33
λ_2	0.8	0.52	0.11	0.29	0.73
λ_3	1.0	0.89	0.11	0.67	1.11
λ_4	-0.5	-0.63	0.26	-1.10	-0.15
λ_5	0	-	-	-	-
λ_6	0.1	0.13	0.19	-0.23	0.47
λ_7	0	-	-	-	-
λ_8	0	-	-	-	-

Estimation Results for Chapter 3.2

We take $\sigma_\beta = \sigma_\gamma = \sigma_\alpha = \sigma_\theta = \sigma_\lambda = 0.2$. The mean acceptance rates in the Metropolis-Hastings algorithms for β, γ, α, θ and λ are respectively 29.7%, 27.2%, 25.9%, 26.8% and 31.6%. We use 200,000 MCMC iterations and a burn-in sample of 100,000. The results are presented in Table B.2.

Table B.2: Simulation of the 3-wave model with non-monotone missingness pattern and no follow-up for subjects in refreshment samples

Parameter	True value	Estimates	s.d.	2.5% quantile	97.5% quantile
β_0	0.1	0.12	0.03	0.06	0.17
β_1	0.5	0.45	0.04	0.37	0.53
γ_0	0.1	0.16	0.07	0.03	0.29
γ_1	0.2	0.26	0.05	0.17	0.36
γ_2	0.7	0.63	0.06	0.52	0.75
α_0	0.1	0.20	0.09	0.03	0.39
α_1	-0.3	-0.26	0.05	-0.37	-0.16
α_2	0.7	0.78	0.05	0.67	0.88
α_3	1.0	0.70	0.18	0.32	1.03
θ_0	0.1	0.23	0.13	-0.03	0.47
θ_1	0.3	0.34	0.05	0.24	0.44
θ_2	-1.0	-1.11	0.15	-1.41	-0.82
θ_3	1.0	0.70	0.12	0.47	0.94
θ_4	-0.5	-0.35	0.10	-0.54	-0.14
θ_5	0.2	0.60	0.16	0.31	0.91
θ_6	0.8	0.60	0.13	0.32	0.84
θ_7	0	-	-	-	-
λ_0	0.1	0.12	0.17	-0.18	0.53
λ_1	0.2	0.19	0.06	0.07	0.31
λ_2	0.8	0.78	0.13	0.52	1.02
λ_3	1.0	1.13	0.11	0.90	1.35
λ_4	-0.5	-0.74	0.24	-1.29	-0.29
λ_5	0.5	0.60	0.09	0.41	0.78
λ_6	0.1	0.13	0.13	-0.11	0.38
λ_7	-1.0	-1.07	0.11	-1.28	-0.85
λ_8	0	-	-	-	-

Estimation Results for Chapter 3.3

We take $\sigma_\beta = \sigma_\gamma = \sigma_\alpha = 0.05$ and $\sigma_\theta = \sigma_\lambda = 0.03$. The mean acceptance rates in the Metropolis-Hastings algorithms for β, γ, α, θ and λ are respectively 31.7%, 28.5%, 27.1%, 31.9% and 35.2%. We use $200,000$ MCMC iterations and a burn-in sample of $100,000$. The results are presented in Table B.3.

Table B.3: Simulation of the 3-wave model with monotone missingness pattern and follow-up for subjects in refreshment samples

Parameter	True value	Estimates	s.d.	2.5% quantile	97.5% quantile
β_0	0.1	0.07	0.03	0.01	0.13
β_1	0.5	0.57	0.04	0.49	0.65
γ_0	0.1	0.16	0.07	0.03	0.29
γ_1	0.2	0.17	0.05	0.08	0.27
γ_2	0.7	0.71	0.06	0.60	0.81
α_0	0.1	0.13	0.09	-0.04	0.30
α_1	-0.3	-0.29	0.05	-0.40	-0.19
α_2	0.7	0.67	0.06	0.55	0.78
α_3	1.0	0.94	0.18	0.61	1.32
θ_0	0.1	-0.47	0.36	-1.28	0.19
θ_1	0.3	0.25	0.06	0.13	0.38
θ_2	-1.0	-1.01	0.15	-1.30	-0.72
θ_3	1.0	2.01	0.73	0.52	3.36
θ_4	-0.5	0.19	0.39	-0.59	1.01
θ_5	0.2	0.23	0.16	-0.08	0.56
θ_6	0	-	-	-	-
θ_7	-1.0	-2.10	0.75	-3.47	-0.53
λ_0	0.1	0.34	0.29	-0.21	0.93
λ_1	0.2	0.13	0.08	-0.03	0.28
λ_2	0.8	0.63	0.15	0.33	0.91
λ_3	1.0	0.67	0.36	0.04	1.43
λ_4	-0.5	-0.66	0.31	-1.28	-0.06
λ_5	0.5	0.41	0.24	-0.05	0.85
λ_6	0.1	0.17	0.15	-0.11	0.48
λ_7	0	-	-	-	-
λ_8	-0.5	-0.10	0.37	-0.86	0.54

Estimation Results for Chapter 3.4

We take $\sigma_\beta = \sigma_\gamma = \sigma_\alpha = 0.05$ and $\sigma_\theta = \sigma_\lambda = 0.03$.. The mean acceptance rates in the Metropolis-Hastings algorithms for β, γ, α, θ and λ are respectively 31.5%, 28.2%, 27.5%, 26.1% and 27.2%. We use 200,000 MCMC iterations and a burn-in sample of 100,000. The results are presented in Table B.4.

Table B.4: Simulation of the 3-wave model with non-monotone missingness pattern and follow-up for subjects in refreshment samples

Parameter	True value	Estimates	s.d.	2.5% quantile	97.5% quantile
β_0	0.1	0.10	0.03	0.04	0.16
β_1	0.5	0.49	0.04	0.41	0.57
γ_0	0.1	0.03	0.06	-0.08	0.16
γ_1	0.2	0.29	0.05	0.20	0.39
γ_2	0.7	0.73	0.06	0.62	0.84
α_0	0.1	0.05	0.07	-0.08	0.21
α_1	-0.3	-0.29	0.06	-0.40	-0.18
α_2	0.7	0.70	0.06	0.59	0.81
α_3	1.0	1.06	0.16	0.75	1.38
θ_0	0.1	0.05	0.19	-0.34	0.42
θ_1	0.3	0.24	0.05	0.14	0.34
θ_2	-1.0	-1.14	0.22	-1.64	-0.75
θ_3	1.0	1.26	0.52	0.35	2.23
θ_4	-0.5	-0.43	0.20	-0.85	-0.01
θ_5	0.5	0.34	0.16	0.03	0.65
θ_6	0.2	0.82	0.22	0.45	1.29
θ_7	-1.0	-1.30	0.53	-2.31	-0.34
λ_0	0.1	0.05	0.18	-0.33	0.39
λ_1	0.2	0.19	0.05	0.09	0.29
λ_2	0.8	0.96	0.13	0.71	1.21
λ_3	1.0	1.04	0.44	0.33	2.08
λ_4	-0.5	-0.41	0.18	-0.77	-0.07
λ_5	0.5	0.49	0.18	0.18	0.86
λ_6	0.1	-0.02	0.12	-0.27	0.24
λ_7	-1.0	-1.16	0.13	-1.41	-0.92
λ_8	-0.5	-0.43	0.46	-1.54	0.30

Bibliography

Das, M., V. Toepoel, and A. v. Soest (2011). Nonparametric tests for panel conditioning and attrition bias in panel surveys. *Sociological Methods & Research 40*, 32–54.

Dorsett, R. (2004). Using matched substitutes to adjust for nonignorable nonresponse: an empirical investigation using labour market data. *Policy Studies Centre Research Discussion Paper*.

Fuchs-Schundeln, N. and M. Schundeln (2005, August). Precautionary savings and self-selection: Evidence from the german reunification "experiment". *The Quarterly Journal of Economics 120*(3), 1085–1120.

Hausman, J. A. and D. A. Wise (1979, March). Attrition bias in experimental and panel data: The gary income maintenance experiment. *Econometrica 47*(2), 455–473.

Hirano, K., G. W. Imbens, G. Ridder, and D. B. Rubin (2001, November). Combining panel data sets with attrition and refreshment samples. *Econometrica 69*(6), 1645–1659.

Huber, M., M. Lechner, C. Wunsch, and T. Walter (2011). Do german welfare-to-work programmes reduce welfare dependency and increase employment? *German Economic Review 12*(2), 182–204.

Kruse, Y., M. Callegaro, J. M. Dennis, S. Subias, and M. Lawrence (2009). Panel conditioning and attrition in the ap-yahoo news election panel study.

Nevo, A. (2003, January). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics 21*(1), 43–52.

Reiter, J. P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika 95*, 933–946.

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

Zheng, S. (2011). Modeling non-ignorable missing data mechanism in panel studies using refreshment samples.