Frailty Model Incorporating Ascertainment Correction with Missing Data in Family-Based Study

by

Jiaqi Bi

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Epidemiology and Biostatistics
Schulich School of Medicine & Dentistry
University of Western Ontario

Abstract

Frailty Model Incorporating Ascertainment Correction with Missing Data in Family-Based Study

Jiaqi Bi
Master of Science
Graduate Department of Epidemiology and Biostatistics
Schulich School of Medicine & Dentistry
University of Western Ontario
2024

This is an abstract section...

To my mother, Kai Hua, for all the upbringrings.
To my father, Guangmin Bi, for teaching me not to give up.
To my supervisors, for their unstopping guidance and invaluable lessons.

Acknowledgements

Contents

List of Symbols

i	Individual index
j	Family (Cluster) index
p	Proband index
d_j	Number of events in family j
t	Some time
a	Some Time for the proband
T	Event Time
δ_{ij}	Event indicator for individual i in family j
w	The observed survival data (t, δ)
n	Number of individuals
J	Number of Families (Clusters)
m	Index of the sampled completed dataset in the MCEM
M	Number of the sampled completed dataset in the MCEM
z	Frailty term
q	q-th element of Gauss Hermite Quadrature
ω	q-th weight of Gauss Hermite Quadrature
y_q	q-th node of Gauss Hermite Quadrature
N_q	Total number of quadratures
$h(\cdot)$	Hazard fucntion
$h_0(\cdot)$	Baseline hazard function
$H(\cdot)$	Cumulative hazard fucntion
$S(\cdot)$	Survival fucntion
$A_j(\cdot)$	Ascertainment of family j into the study
$L(\cdot)$	Likelihood function
$\ell(\cdot)$	Log-likelihood function
$\mathscr{L}(\cdot)$	Laplace transform
$\mathbf{x}$	Covariates
$\boldsymbol{\beta}$	Model coefficients vector
$\boldsymbol{\theta}$	Parameter vector
Λ	The combination of $(\boldsymbol{\beta}, \lambda, \alpha)$

λ	Weibull shape parameter
α	Weibull scale parameter
υ	General form of the parameter in an undefined frailty distribution
k	Gamma shape and rate parameters
σ^2	Log-Normal variance parameter
ψ	Missing data distribution parameters

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

The Breast Cancer type 1/2, usually referred as BRCA1/2, are proteins that consists of genes that code for BRCA1 in humans. BRCA1/2 are human tumor suppressor genes, that are responsible for repairing the DNA [1]. When the mutation exists on these genes may cause the impairments of proper functions, which can lead to the possibility of capturing the breast, ovarian, or other specific cancers [2, 3]. Inheriting one of these mutations does not guarantee developing cancer disease, but the mutation can increase the risk of getting those cancers.

In the field of medicine, these cancer types are classified as Hereditary Breast and Ovarian Cancer Syndrome (HBOC). The average life expectancy of individuals with BRCA1, without any interventions, is approximately 4.2 years shorter than that of non-carriers of the BRCA1 gene [4]. Significant advancements have been made in the medical and statistical modeling of breast cancer risk among BRCA1/2 carriers. These include the application of competing risk survival analysis based on breast and ovarian cancer outcomes developed by Choi et al. [5], as well as various clinical trials investigating risk-reducing treatment approaches for breast cancer patients [6]. Despite these efforts, it remains crucial to ensure statistical validity across these studies especially when missing data exists.

From a statistical perspective, the study is centered on a specific disease, which may introduce selection bias due to the sampling process. This bias arises from the selection criteria based on specific probands in each family. To mitigate this sampling bias, an ascertainment correction should be applied to the likelihood calculation, conditioning on the proband information. To accurately capture the heterogeneity between families in the context of time-to-cancer outcomes, the use of a frailty model is recommended. There are various choices for frailty distributions in survival analysis, including the Gamma distribution and the log-Normal distribution.

1.2 Motivation

Although numerous studies on risk assessment in susceptible populations and statistical advancements in dynamic prediction have significantly contributed to understanding BRCA1/2 families, the issue of missing data remains a substantial challenge, particularly in the context of survival outcomes. Over the past decade, several methodologies have been proposed to address missing data, including the Expectation-Maximization (EM) algorithm, the Monte-Carlo EM algorithm for cases where the E-step lacks a closed form, and Multiple Imputation (MI). However, when applying frailty models, which incorporate random effects in survival analysis, the literature addressing missing data is relatively sparse.

In genetic epidemiology, research is typically conducted on a family-wise basis. Therefore, considering the family structure when addressing statistical problems is both essential and unavoidable. Moreover, existing techniques for handling missing data must be carefully adapted, as the clustered nature of the dataset introduces additional complexity. Within the genetic framework, many variables, such as genetic information and polygenic risk scores (PRS), are not independent between individuals. Traditional methodologies often fail to account for family correlations and ascertainment bias. Given that families are selected based on a proband, it is crucial to apply ascertainment correction to minimize the selection bias. This situation presents an opportunity to further investigate and develop adequate methods for handling missing data, taking into account family correlations and ascertainment bias.

In this project, we aim to investigate the current implementation of Multiple Imputation (MI) methods for frailty models. Additionally, we propose a novel MI method that explicitly incorporates the kinship matrix during the imputation of genetically related variables. This proposed method will be evaluated by comparing it to existing MI methods and Complete Case Analysis (CCA).

1.3 Objectives

With the proposed MI method and the BRCA1 data, the objectives of this thesis are designed as follows:

1. To adapt the kinship correlations into the imputation step

2. To incorporate the ascertainment correction into the likelihood while considering that not all probands are affected

3. To assess the novel MI method via the calculation of the estimations, biases, and precisions through the simulation study

4. To apply the novel MI method and adjusted likelihood to model the BRCA1 family data

1.4 Organizations of the Thesis

Chapter 2

Literature Reviews

2.1 Survival Analysis

Survival analysis is a robust statistical methodology used to analyze time-to-event data, where the focus is on the time until an event of interest occurs. It has been extensively applied in medical research, particularly in studies involving cancer, where events such as death or relapse are critical endpoints. The literature identifies several key methods and models that form the backbone of survival analysis. Kaplan and Meier [7] introduced the Kaplan-Meier estimator, a nonparametric statistic used to estimate survival functions from incomplete observations, which remains widely used due to its simplicity and effectiveness in handling censored data. Cox [8] proposed the proportional hazards model, which allows for the inclusion of covariates and has become a standard technique for assessing the effect of explanatory variables on survival. Collett [9], Machin et al. [10], and Kleinbaum and Klein [11] provide detailed expositions on survival analysis methods, including parametric and nonparametric approaches. Recent advancements have addressed complex issues such as interval censoring and competing risks, expanding the applicability and precision of survival analysis. These methodological developments have significantly enhanced the ability to make informed inferences from survival data, contributing to more accurate prognostic assessments and treatment evaluations in clinical research. In the survival analysis, there are several key functions that will contribute essentially to nearly all relevant scientific work.

The survival function $S(t)$ measures the probability of study subject's entry age t is less than the event time T,

$$S(t) = P(T > t) \tag{2.1}$$

The cumulative distribution function $F(t)$ represents the probability that the event has occurred by time t. This is the complement of the survival function $S(t)$,

$$F(t) = P(T \leq t) = 1 - S(t) \tag{2.2}$$

The probability density function $f(t)$ is the likelihood of the event occurring at an exact time t. It is the derivative of the cumulative distribution function (CDF) or the negative derivative of the survival function.

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) = h(t)S(t) \tag{2.3}$$

The hazard function $h(t)$ measures the instantaneous rate at which the event occurs, given that the individual has survived up to the time t. It is the probability that an event occurs in a very small time interval, given survival until the beginning of the interval.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{2.4}$$

The accumulated risk of experiencing the event up to time t can be expressed using the cumulative hazard function. It is the integral of the hazard function over time, providing a cumulative measure of risk.

$$H(t) = \int_0^t h(u)du = -\log S(t) \tag{2.5}$$

Whenever we can define one of above functions, it is straighforward to derive the rest.

2.1.1 Frailty Model for Family Based Study

Parametric survival analysis methods assume that the time-to-event data follow a specific probability distribution. This approach provides a more detailed and flexible framework for modeling survival data, allowing for more precise estimates and interpretations of survival functions and hazard rates. Hosmer Jr et al. [12] has comprehensively summarized most of parametric baseline hazard

Chapter 3

Methods

3.1 Introduction

When making the imputation on the continuous variable, one common way is to assume a conditionally normal distribution. This conditional distribution can be estimated from a linear regression. The original multiple imputation structure was brought by Rubin [13], that this method has been widely used by different scientific researchers with different models. There are many softwares that based on the multiple imputations as well, such as Blimp software [14], `MICE` package in R [15], and `jomo` package in R [16]. Apparently, this thesis cannot provide enough rooms for those well-designed softwares that I have not mentioned. The development of Multiple Imputation within certain specific models remains incomplete, and many details have not yet been comprehensively addressed in the current published literature. In genetic epidemiology, studies are mostly conducted with family clusters. Current implementation of the multiple imputation does not account for the genetic correlations. Furthermore, there remains room for exploration in the application of the frailty model with ascertainment correction. Therefore, this research is designed to develop a computationally efficient multiple imputation method to account for the kinship correlations, and apply it to frailty models with ascertainment correction.

In this chapter, a comprehensive guideline and adjusted multiple imputation formulas are provided. We explicitly show how the kinship matrix works in the imputation step, and how the ascertainment correction handles the analysis step in this research. The special imputation model is introduced as well, while considering the genetic variance and residual variance (sometimes refer to the environmental variance). The variance estimation using Rubin's Rule is provided, as well as the confidence interval based on completed data following the proposed multiple imputation.

3.2 Kinship Matrix

The kinship matrix, also known as the relatedness matrix, is a fundamental concept in statistical genetics, particularly in the context of quantitative genetics and genetic epidemiology. It quantifies the genetic relatedness between individuals based on pedigree information [17]. This matrix is essential for estimating heritability and for controlling for familial relatedness in genetic association studies. The elements of the kinship matrix represent the probability that a randomly chosen allele from one individual is identical by descent (IBD) to a randomly chosen allele from another individual. Once the relationship between each individual in a study is known, the kinship matrix can be generated either for each family or as a whole [18]. For example, the kinship matrix K for a family may look like the following:

$$K = \begin{bmatrix} 1 & 0.5 & 0.25 & 0 \\ 0.5 & 1 & 0.5 & 0 \\ 0.25 & 0.5 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

such that assume there are 4 individuals in a family, the row and the column represent the exactly the same individuals. Therefore, the diagonal of the kinship matrix has a value of 1, which measures the correlation of one person and himself. In K_{11}, the individual index 1 and himself is fully related to himself. In $K_{12} = 0.5$, the first and second individuals are half-related (e.g., Siblings). In $K_{13} = 0.25$, the first and third individuals are a quarter-related (e.g., half-siblings or grandparent-grandchild). In $K_{14} = 0$, the first and the fourth individuals are not related (e.g., spousal relationship, or lawful adoption). Note that in nature, the kinship matrix is fully symmetric. Being a symmetric matrix yields real eigenvalues and orthogonal eigenvectors, which is beneficial in principal component analysis (PCA) that is often conducted within a GWAS study to reduce dimensionality [19]. Also, this symmetric property ensures the genetic relatedness between any two individuals is consistent regardless of their order in a dataset, which further ensures an accurate interpretation in the the statistical inference and valid and unbiased statistical tests. In statistical analyses, the kinship matrix is not typically estimated; rather, it is derived from a precise understanding of familial relationships. This direct acquisition ensures an accurate representation of genetic relatedness, which is essential for the integrity of subsequent genetic analyses.

In some variables within a familial study, the structure could be highly dependent on the kinship matrix. For example, the PRS score may be used as a covariate when researchers try to model certain disease risks. Especially, in a genetic study where the missing data occurs, without the extensive considerations on the kinship correlation structure, the inference of parameters may be biased. In current imputation models, there is a challenge that those methods do not handle the kinship correlations. Thus, the following section shows how to conduct a proper multiple imputation without ignoring the pedigree information.

3.3 Multiple Imputation without Kinship Matrix

In the existing multiple imputation methods for continuous variable, one easy way is to assume a conditionally normal distribution. Suppose y is the variable contains missing values, $\mathbf{x}$ are other variables are fully observed. We assume there are total p parameters being estimated in this linear regression. The conditionally normal model (linear regression) can be written as

$$y|\mathbf{x}, \boldsymbol{\beta} \sim N(\boldsymbol{\beta}\mathbf{x}, \sigma^2) \tag{3.1}$$

In the linear regression setting,

$$y = \mathbf{x}\boldsymbol{\beta} + \epsilon,\ \epsilon \sim N(0, \sigma^2) \tag{3.2}$$

This simply corresponds to the likelihood

$$f(y|\mathbf{x}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}(y - \mathbf{x}\boldsymbol{\beta})^\top (y - \mathbf{x}\boldsymbol{\beta})\Big) \tag{3.3}$$

because of the conditional normality of y. We can solve that

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top y \tag{3.4}$$

In the Bayesian framework, we need to find the prior, which is the joint distribution $f(\sigma^2, \boldsymbol{\beta})$. Note that $(y - \mathbf{x}\boldsymbol{\beta})^\top (y - \mathbf{x}\boldsymbol{\beta})$ can be written as

$$\begin{aligned}
(y - \mathbf{x}\boldsymbol{\beta})^\top (y - \mathbf{x}\boldsymbol{\beta}) &= \big((y - \mathbf{x}\boldsymbol{\beta}) + (\mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{x}\boldsymbol{\beta})\big)^\top \big((y - \mathbf{x}\boldsymbol{\beta}) + (\mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{x}\boldsymbol{\beta})\big) && (3.5)\\
&= (y - \mathbf{x}\boldsymbol{\beta})^\top (y - \mathbf{x}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top (\mathbf{x}^\top \mathbf{x})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + 2(\mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{x}\boldsymbol{\beta}) && (3.6)\\
&= (y - \mathbf{x}\boldsymbol{\beta})^\top (y - \mathbf{x}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top (\mathbf{x}^\top \mathbf{x})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) && (3.7)
\end{aligned}$$

So Equation 3.3 will become

$$f(y|\mathbf{x}, \boldsymbol{\beta}, \sigma^2) \propto \underbrace{(\sigma^2)^{-v/2} \exp(-\frac{vs^2}{2\sigma^2})}_{f(\sigma)} \underbrace{(\sigma^2)^{-\frac{n-v}{2}} \exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top (\mathbf{x}^\top \mathbf{x})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\Big)}_{f(\boldsymbol{\beta}|\sigma)} \tag{3.8}$$

where $vs^2 = (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}) = SSE$, such that $v = n_{obs} - p$. Then $f(\sigma^2)$ can be written as a proportional density to the inverse gamma distribution,

$$f(\sigma^2) \propto (\sigma^2)^{-\frac{v}{2}-1} \exp(-\frac{vs^2}{2\sigma^2}) = (\sigma^2)^{-\frac{v}{2}-1} \exp(-\frac{SSE}{2\sigma^2}) \tag{3.9}$$

In this $\sigma^2 \sim$ Inverse-Gamma(α, ϕ), we have $\alpha = \frac{v}{2} = \frac{n_{obs}-p}{2}$ and $\phi = \frac{1}{2}vs^2 = \frac{SSE}{2}$. From the inverse-gamma property, when

$$\sigma^2 \sim \text{Inverse-Gamma}(\frac{n_{obs}-p}{2}, \frac{SSE}{2}) \tag{3.10}$$

and $\exists\lambda$ such that

$$\sigma^2 = \frac{SSE}{2}/\lambda \tag{3.11}$$

then λ can be transformed

$$\lambda = \frac{\chi^2_{n_{obs}-p}}{2} \tag{3.12}$$

since $\chi^2_{df} = \text{Gamma}(\frac{df}{2}, 2)$, so

$$\sigma^2 = \frac{\frac{SSE}{2}}{\frac{\chi^2_{n_{obs}-p}}{2}} = \frac{SSE}{\chi^2_{n_{obs}-p}} \tag{3.13}$$

Thus, we can sample the standard deviation of the missing data distribution from

$$\sigma^* = \hat{\sigma}\sqrt{\frac{SSE}{\chi^2_{n_{obs}-p}}} = \hat{\sigma}\sqrt{\frac{SSE}{g}} \tag{3.14}$$

from sampling $g \sim \chi^2_{n_{obs}-p}$. Moreover, we know

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{x}^\top\mathbf{x}^{-1}) = \mathbf{V} \tag{3.15}$$

so the marginal distribution for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{x}^\top\mathbf{x})^{-1}) \tag{3.16}$$

Note that when a random variable $T \sim N(\mathbf{m}, \mathbf{c})$, then T can be generated from a standard normal variable $\mathbf{u}$ with

$$T = \mathbf{m} + \mathbf{L}\mathbf{u} \tag{3.17}$$

where $\mathbf{L}$ is the cholesky decomposition of $\mathbf{c}$ such that $\mathbf{c} = \mathbf{L}\mathbf{L}^\top$. For $\boldsymbol{\beta}$, from 3.16, it can be derived as

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} + \sigma(\mathbf{x}^\top\mathbf{x})^{-1/2}\mathbf{u}_1 \tag{3.18}$$

$$= \hat{\boldsymbol{\beta}} + \mathbf{u}_1\mathbf{V}^{-1/2} \tag{3.19}$$

Adjusting for σ^* to make sure $\boldsymbol{\beta}$ matches the variability implied by the random draw of σ^*,

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\mathbf{u}_1\mathbf{V}^{-1/2} \tag{3.20}$$

such that $\mathbf{u}_1$ is a row vector of p independent draws from a standard normal distribution, $u_{1k} \overset{iid}{\sim} N(0,1)$, and $k = \ldots, p$. The imputation for y_i^* is computed as

$$y_i^* = \boldsymbol{\beta}^* \mathbf{x}_i + u_{2i}\sigma^*, \ s.t.\ u_{2i} \sim N(0,1) \tag{3.21}$$

where u_{2i} adds the uncertainty to the imputation as well to ensure the imputation is not solely based on the predicted value of y_i^*. This prevents the underestimation of the variability. Therefore, the comprehensive steps of the multiple imputation on the continuous missing data can be summarized to the following steps:

1. Calculate $\hat{y} = \hat{\boldsymbol{\beta}}\mathbf{x}$ using y_{obs}, and $\hat{\boldsymbol{\beta}}$ can be obtained easily, as well as $\hat{\sigma}$, and $\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{V}$
2. Draw $g \sim \chi^2_{n_{obs}-p}$ for one random draw
3. Calculate $\sigma^* = \hat{\sigma}/\sqrt{SSE/g}$
4. Draw a p dimensional vector $\mathbf{u}_1$ such that $u_{1k} \overset{iid}{\sim} N(0,1)$ and $k = 1, \ldots, p$
5. Calculate $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\mathbf{u}_1\mathbf{V}^{1/2}$ such that $\mathbf{V}^{1/2}$ is the cholesky decomposition of $\mathbf{V}$
6. Draw $u_{2i} \overset{iid}{\sim} N(0,1)$
7. Impute $y_{mis,i} = \boldsymbol{\beta}^* \mathbf{x}_i + u_{2i}\sigma^*$
8. (Option: PMM) Match the imputed value to the nearest observed value.
9. Repeat 2. to 7. (2. to 8. for PMM) for M times to obtain M complete datasets

3.4 Multiple Imputation with Kinship Matrix

In order to account for the kinship correlations, the conditional normal distribution needs to be adjusted where the variable contains the missing components are said to be multivariate. Denote $\mathbf{y}$ as the continuous variable that is subject to missing, $\mathbf{x}$ is other covariates that formed a design matrix that determined to be eligible to be part of the imputation model, and $\boldsymbol{\beta}$ is simply the parameter of the imputation model.

$$\mathbf{y}|\mathbf{x}, \boldsymbol{\beta} \sim MVN(\boldsymbol{\beta}\mathbf{x}, \sigma_g^2 K + \sigma_e^2 I) \tag{3.22}$$

where in the linear mixed effect regression form with flexible covariance matrix, the model can be written as

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e} \tag{3.23}$$

such that two random parts of the model have certain multivariate normal distribution,

$$\mathbf{u} \sim MVN(0, \sigma_g^2 K), \text{ and } \mathbf{e} \sim MVN(0, \sigma_e^2 I)$$

K is the kinship matrix with the diagonal of 1, and I is the identity matrix. Note that the flexible covariance matrix can be written as

$$\mathbf{\Sigma} = \sigma_g^2 K + \sigma_e^2 I \tag{3.24}$$

So in this multivariate version of linear mixed effects model, the kinship matrix is being adapted. With unknown mean and fully unknown covariance matrix, the prior of the covariance matrix can be selected as an Inverse Wishart distribution with some degrees of freedom and the scale parameter in Bayesian sense. However, in this case, the covariance matrix is not fully unknown. The only unknown parts of this covariance matrix is introduced by simply σ_g^2 and σ_e^2. Deriving the Bayesian prior of these two parameters is challenging, especially it needs to incorporate with partially known multivariate parameter $\mathbf{\Sigma}$. Therefore, prior definition steps are adjusted to empirical estimates of these two variances. Although, in a normal model, the inverse-gamma distribution can be an option as a conjugate prior for variance component. The conjugacy simplifies the mathematical derivation, but there are two variance components when the flexible covariance matrix for this designed imputation model, which defers from the regular choice. With the above preliminary settings before the imputation step, the adjusted multiple imputation can be concluded into following steps:

1. Obtain the kinship matrix K among all individuals
2. Calculate the estimates of $\hat{y} = \mathbf{x}\boldsymbol{\beta}$, obtain estimates of $\hat{\boldsymbol{\beta}}$, $\hat{\sigma_g}^2$, $\hat{\sigma_e}^2$, $\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{V}$. In this step, naturally, $\hat{\mathbf{\Sigma}}$ is obtained.
3. Suppose there are p predictors in the imputation model, draw p-dimensional vector $\mathbf{w}_1$ such that $w_{1k} \overset{iid}{\sim} N(0,1)$ where $k = 1, ..., p$
4. Calculate $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + w_1 \mathbf{V}^{1/2}$ such that $\mathbf{V}^{1/2}$ is the cholesky decomposition of $\mathbf{V}$
5. Obtain $\mu_i^* = \boldsymbol{\beta}^* \mathbf{x}_i$
6. Obtain the conditional expectations

$$E(y_{mis,i}|\mathbf{y}_{-i}) = \mu_i^* + \hat{\mathbf{\Sigma}}_{i,-i}\hat{\mathbf{\Sigma}}_{-i,-i}^{-1}(\mathbf{y}_{-i} - \boldsymbol{\mu}_{-i}^*) \tag{3.25}$$

7. Impute $y_{mis,i} = E(y_{mis,i}|\mathbf{y}_{-i})$
8. (Option: PMM) Match the imputed value to the nearest observed value
9. Repeat 3. to 7. (3. to 8. for PMM) for M times to obtain M completed datasets.

Generally, when there is no preliminary knowledge on the dataset, one should use as many variables as possible for the imputation model covariate [20]. Also, in the survival analysis, there has been always an argument on whether the $\log(T)$ or $\log(H_0(T))$ should be used within

the imputation model [21]. In this thesis, we have used both $\log(T)$ and $\log(H_0(T))$ in the simulation study to determine the differences to avoid potential biases caused by misspecification of the model.

3.5 Variance Estimation

Based on the Rubin's rule, the pooling step of the analysis is defined as

$$\bar{\theta} = \frac{1}{M}\sum_{i=1}^{M}\theta_i \tag{3.26}$$

where θ_i is the parameter estimate that we are making the inference in a study for i-th imputation after M imputations. So $\bar{\theta}$ is the pooled parameter estimates. Based on Barnard and Rubin [22], the variance estimation is also defined in a two-level structure, which are within imputation variance and between imputation variance. The within imputation variance is similar to the pooled parameter estimation, where

$$V_W = \frac{1}{M}\sum_{i=1}^{m} SE_i^2 \tag{3.27}$$

such that V_W determines the within-imputation variance, and it's simply "pooled" variance among all analyses from the imputed dataset i. The between-imputation variance, on the other hand, accounts for the extra variances caused by the missing data. The between-imputation variance V_B is defined as

$$V_B = \frac{\sum_{i=1}^{M}(\theta_i - \bar{\theta})^2}{M-1} \tag{3.28}$$

where V_B is the unbiased estimation. Note that θ is estimated using the only finite M imputed datasets, according to Van Buuren [23], V_B is the approximation when $M \to \infty$. Therefore, the total variance can be written as

$$V_{\text{Total}} = V_W + V_B + \frac{V_B}{M} \tag{3.29}$$

when M is large enough, V_{Total} tends to have the only two components of within and between variances. The test statistic (Wald statistic) is

$$\frac{(\bar{\theta} - \theta_0)^2}{V_{\text{Total}}} \sim F_{1,df_{adj}} \tag{3.30}$$

which

$$\frac{\bar{\theta} - \theta_0}{\sqrt{V_{\text{Total}}}} \sim t_{df_{adj},\frac{1-\alpha}{2}} \tag{3.31}$$

such that α is the significance level. Define

$$\rho = \frac{V_B + \frac{V_B}{M}}{V_{\text{Total}}} \tag{3.32}$$

and

$$r = \frac{V_B + \frac{V_B}{M}}{V_W} \tag{3.33}$$

where ρ is the fraction of missing information, which quantifies the proportion of the total variance that is attributable to the fact that the data have been imputed M times. Also, r is the relative increase in variance due to non-responses (missing data), which measures the increase in variance due to the missing data compared to if there were no missing data. In the interpretation, ρ gives the proportion of the total variance that is due to the uncertainty introduced by the missing data. A higher ρ indicates a larger fractino of the total uncertainty comes from the fact that the data were imputed. A higher r indicates that the missing data have a larger impact on the overall variance. In the first MI literature by Rubin [13], the degrees of freedom was defined as

$$df_{old} = (M-1) \times (1 + \frac{1}{r})^2 \tag{3.34}$$

The estimated degrees of freedom for the observed data, adjusted for the missing information, are

$$df_{obs} = \frac{(n-k)+1}{(n-k)+3}(n-k)(1-\rho) \tag{3.35}$$

where n is the number of observations in each imputed data, and k is the number of parameters to be estimated. Barnard and Rubin [22] further alternated the calculation of the degrees of freedom by combining Equations 3.34 and 3.35

$$df_{adj} = \frac{df_{old} df_{obs}}{df_{old} + df_{obs}} \tag{3.36}$$

When conducting the statistical tests on the pooled estimates, df_{adj} is used for the t-distribution degrees of freedom. The confidence interval is straightforward, where we obtain

$$SE_{pooled} = \sqrt{V_{\text{Total}}} \tag{3.37}$$

then we simply calculate

$$CI = \bar{\theta} \pm t_{df_{adj}, \frac{1-\alpha}{2}} \times SE_{pooled} \tag{3.38}$$

Bibliography

[1] JA Duncan, JR Reeves, and TG Cooke. BRCA1 and BRCA2 proteins: roles in health and disease. *Molecular pathology*, 51(5):237, 1998.

[2] Bruce G Haffty, Elizabeth Harrold, Atif J Khan, Pradip Pathare, Tanya E Smith, Bruce C Turner, Peter M Glazer, Barbara Ward, Daryl Carter, Ellen Matloff, et al. Outcome of conservatively managed early-onset breast cancer by BRCA1/2 status. *The Lancet*, 359 (9316):1471–1477, 2002.

[3] Yong-Wen Huang. Association of BRCA1/2 mutations with ovarian cancer prognosis: an updated meta-analysis. *Medicine*, 97(2), 2018.

[4] Michael P Lux, Peter A Fasching, and Matthias W Beckmann. Hereditary breast and ovarian cancer: review and future perspectives. *Journal of molecular medicine*, 84:16–28, 2006.

[5] Yun-Hee Choi, Hae Jung, Saundra Buys, Mary Daly, Esther M John, John Hopper, Irene Andrulis, Mary Beth Terry, and Laurent Briollais. A competing risks model with binary time varying covariates for estimation of breast cancer risks in brca1 families. *Statistical Methods in Medical Research*, 30(9):2165–2183, 2021.

[6] Yun-Hee Choi, Mary Beth Terry, Mary B Daly, Robert J MacInnis, John L Hopper, Sarah Colonna, Saundra S Buys, Irene L Andrulis, Esther M John, Allison W Kurian, et al. Association of risk-reducing salpingo-oophorectomy with breast cancer risk in women with brca1 and brca2 pathogenic variants. *JAMA oncology*, 7(4):585–592, 2021.

[7] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[8] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[9] David Collett. *Modelling survival data in medical research*. Chapman and Hall/CRC, 2023.

[10] David Machin, Yin Bun Cheung, and Mahesh Parmar. *Survival analysis: a practical approach*. John Wiley & Sons, 2006.

[11] David G Kleinbaum and Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.

[12] David W Hosmer Jr, Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time-to-event data*, volume 618. John Wiley & Sons, 2008.

[13] Donald B Rubin. Multiple imputation for survey nonresponse, 1987.

[14] BT Keller and CK Enders. Blimp user's guide (version 3). *Blimp Software: Los Angeles, CA, USA*, 2021.

[15] Patrick Royston and Ian R White. Multiple imputation by chained equations (mice): implementation in stata. *Journal of statistical software*, 45:1–20, 2011.

[16] Matteo Quartagno, Simon Grund, and James Carpenter. Jomo: a flexible package for two-level joint modelling multiple imputation. *R Journal*, 9(1), 2019.

[17] Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

[18] Jason P Sinnwell, Terry M Therneau, and Daniel J Schaid. The kinship2 r package for pedigree data. *Human heredity*, 78(2):91–93, 2014.

[19] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[20] Donald B Rubin. Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pages 29–62. Chapman and Hall/CRC, 2018.

[21] Ian R White and Patrick Royston. Imputing missing covariate values for the cox model. *Statistics in medicine*, 28(15):1982–1998, 2009.

[22] John Barnard and Donald B Rubin. Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999.

[23] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.