SSC 2024

Correlated Shared Frailty Model Incorporating Ascertainment Correction with Missing Covariates in Family-Based Studies

Jiaqi Bi, Osvaldo Espin-Garcia, Yun-Hee Choi

Department of Epidemiology and Biostatistics
Schulich School of Medicine & Dentistry
University of Western Ontario

June 4, 2024

Background

Breast Cancer

- There were estimated 30,500 new cases of breast cancer in Canada in 2024, and approximately 5,500 deaths, making it the second leading cause of cancer-related death among women [1].
- Hereditary breast-ovarian cancer (HBOC) is an autosomal dominant disease characterized by germline pathogenic mutations in the BRCA1/2 genes [2].
- Time-To-Cancer as an outcome, mutation gene status (mgene) & Polygenic Risk Score (PRS) are predictors - Problems: There are missing data!

Background

Frailty Model for Family Data

- Many different frailty models have been proposed for the analysis of BRCA1/2 families by Choi et al. [3], Chen et al. [4]
- Missing data remains a problem

Missing Data

- The issue of the missing data was firstly brought by Rubin [5] in 1976.
- Three missing mechanisms: MCAR, Missing At Random (MAR), Missing Not At Random (MNAR)

Survival Analysis

Survival Analysis

The hazard function is defined as

$$h_{ij}(t_{ij}|\mathbf{x}_{ij}, z_j) = h_0(t_{ij}) \exp(\boldsymbol{\beta}\mathbf{x}_i) z_j \tag{1}$$

In the parametric setting, the baseline hazard $h_0(t_{ij})$ has a closed form.

Frailty Term

The dataset is clustered, so a frailty is required when modelling the time-to-event outcome to introduce random effects, association and unobserved heterogeneity.

$$z_j \sim \text{Gamma}(\kappa, \kappa); \;\; z_j \sim \text{log-Normal}(0, \kappa^2) \tag{2}$$

where κ is the shape and rate parameters for Gamma distribution.

Conditional Likelihood

When there is no missing data

For individual i in family j,

$$L(\boldsymbol{\theta}|z_j) = \prod_{j=1}^{J}\prod_{i=1}^{n_j} h(t_{ij}|\mathbf{x}_{ij}, z_j)^{\delta_{ij}} \exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) \tag{3}$$

Ascertainment Correction

Denote $A(\boldsymbol{\theta})$ be the likelihood for the proband, and p_j be the proband in family j, also denote $I(T_{p_j} < a_{p_j})$ as an indicator of the proband was affected before their entry to the study

$$A(\boldsymbol{\theta}) = \left[1 - S_{p_j}(a_{p_j}|\mathbf{x}_{p_j})\right]^{I(T_{p_j}<a_{p_j})} S_{p_j}\left[(a_{p_j}|\mathbf{x}_{p_j})\right]^{1-I(T_{p_j}<a_{p_j})} \tag{4}$$

Because not all probands are affected.

Conditional Likelihood

Ascertainment Correction

Then the complete conditional likelihood becomes

$$L_C(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{A(\boldsymbol{\theta})} \tag{5}$$

Current multiple imputation methods (Continuous variable)

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 $k = 1, ..., p$
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. Repeat 2. to 7. for M times to obtain M complete datasets

Challenges on current MI methods

It fails to account for the kinship and frailty.

Kinship Matrix

In the genetic epidemiology, covariates are often genetically correlated within one family, current MI for continuous data assumes

$$x_{ij,mis}|\mathbf{x}_{ij,obs} \sim N(\beta\mathbf{x}_{ij,obs}, \sigma^2) \tag{6}$$

Is this an adequate assumption?

Kinship Matrix

An example of a kinship matrix in a family, suppose there are 4 individuals...

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

- In K_{11},the individual is fully related to themselves.
- In $K_{12} = 0.5$, the first and second individuals are half-related (Siblings).
- In $K_{13} = 0.25$, the first and third individuals are a quarter-related (half-siblings or grandparent-grandchild).
- In $K_{14} = 0$, the first and the fourth individuals are not related.

Kinship Matrix

Accounting for the genetic effects (kinship),

$$\mathbf{x}_{j,mis,1}|\mathbf{x}_{j,obs} \sim MVN(\beta\mathbf{x}_{j,obs}, \sigma_g^2 K + \sigma_e^2 I) \tag{7}$$

such that K is the kinship correlation matrix with diagonal of 1.

- σ_g^2 accounts for the genetic variances, and σ_e^2 accounts for the residual variances.
- The multivariate normal distribution is what we are sampling the missing PRS when considering the kinship matrix.
- Denote $\mathbf{\Sigma} = \sigma_g^2 K + \sigma_e^2 I$.

Multivariate Normal

Multivariate Normal

Assume $\boldsymbol{\mu} = \boldsymbol{\beta}\mathbf{x}_{ij,obs}$. For each individual i, partition the data into

$$\begin{pmatrix} \mu_i \\ \boldsymbol{\mu}_{-i} \end{pmatrix}$$

The covariance can also be partitioned to

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{ii} & \boldsymbol{\Sigma}_{i,-i} \\ \boldsymbol{\Sigma}_{-i,i} & \boldsymbol{\Sigma}_{-i,-i} \end{pmatrix}$$

so the conditional expectation can then be derived to

$$E(x_{mis,i}|\mathbf{x}_{-i}) = \mu_i^* + \hat{\boldsymbol{\Sigma}}_{i,-i}\hat{\boldsymbol{\Sigma}}_{-i,-i}^{-1}(\mathbf{y}_{-i} - \boldsymbol{\mu}_{-i}^*) \tag{8}$$

MI with Kinship Matrix

1. Obtain the kinship matrix K among all individuals
2. Calculate the estimates of $\hat{y} = \mathbf{x}\hat{\boldsymbol{\beta}}$, 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, $\hat{\boldsymbol{\Sigma}}$ is obtained.
3. Draw p-dimensional vector 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

$$E(y_{mis,i}|\mathbf{y}_{-i}) = \mu_i^* + \hat{\boldsymbol{\Sigma}}_{i,-i}\hat{\boldsymbol{\Sigma}}_{-i,-i}^{-1}(\mathbf{y}_{-i} - \boldsymbol{\mu}_{-i}^*) \tag{9}$$

7. Impute $y_{mis,i} = E(y_{mis,i}|\mathbf{y}_{-i})$
8. Repeat 3. to 7. for M times to obtain M complete datasets.

Imputation Model

- Generally, if one has no preliminary knowledge on the dataset, one should use as many variables as the covariate [6].
- In the survival setting, there has been always an argument on whether the $\log(t)$ should be used as the covariate or $H_0(t)$ [7].
- Therefore, we decided to include both scenarios in the simulation, to illustrate the best performed method.

Simulation Study

Table: The Simulation Results for 50% missing on "PRS" covariate using Gamma frailty model, with 200 families with 3228 observations in total. $M = 5$, * denotes a $p < 0.05$.

	Gamma Frailty True value	CCA	MI without Kinship		MI with Kinship	
			$\log(T)$	$\log(H_0(T))$	$\log(T)$	$\log(H_0(T))$
$\log(\alpha)$	-4.135	-3.869^*	-4.269^*	-4.047^*	-4.320^*	-4.126^*
$\log(\lambda)$	1.099	1.125^*	1.031^*	1.131^*	1.035^*	1.120
β_{gender}	1.000	1.112	0.749	0.805	0.824	0.769
β_{mgene}	3.000	2.643	$\mathbf{2.760}^*$	$\mathbf{2.503}^*$	$\mathbf{2.648}^*$	$\mathbf{2.572}^*$
β_{PRS}	3.000	2.703^*	3.313^*	3.273^*	3.364^*	3.336^*
$\log(\kappa)$	0.693	0.651	1.200	1.524	1.019	1.266

BRCA1 Data

Table: The Application Results for BRCA1 family data with 80% missing rate on "PRS" variable using Gamma frailty model, with 498 families and 2650 individuals. $M = 5$, * denotes a $p < 0.05$.

	CCA	MI without Kinship		MI with Kinship	
		$\log(T)$	$\log(H_0(T))$	$\log(T)$	$\log(H_0(T))$
$\log(\alpha)$	-7.344	-4.463^*	-4.541^*	-4.601^*	-4.571^*
$\log(\lambda)$	2.424	0.833^*	0.986^*	0.845^*	0.926^*
β_{mgene}	27.781	2.064^*	2.211^*	2.149^*	2.269^*
β_{PRS}	11.434	0.266	0.665	0.580	1.045
$\log(\kappa)$	-27.300	1.083	0.799	1.046	1.001

References I

[1] Canadian Cancer Statistics Advisory Committee. Canadian cancer statistics 2023, 2023. URL `https://cancer.ca/en/research/cancer-statistics`.

[2] Colin C Pritchard. New name for breast-cancer syndrome could help to save lives. *Nature*, 571(7763):27–29, 2019.

[3] 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.

[4] Lu Chen, Li Hsu, and Kathleen Malone. A frailty-model-based approach to estimating the age-dependent penetrance function of candidate genes using population-based case-control study designs: an application to data on the brca1 gene. *Biometrics*, 65(4):1105–1114, 2009.

[5] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

References II

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

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

Advertisement

More Questions? Interested in being supervised by my fantastic supervisors?
My email: jbi23@uwo.ca