BIOS735 Final Project Proposal

Group 5

March 2020

1 Introduction

The dataset we will examine for our project can be found in the JM package in R. This data is from a randomized clinical trial (RCT) examining two treatments in HIV patients. It contains both longitudinal and survival data on 467 patients who failed or were intolerant of zidovudine (AZT) therapy. These treatment groups of the RCT are two anti-retroviral drugs: zalcitabine and didanosine. The dataset contains a total of 1408 observations and 9 variables: patient identifier, time to death or censoring, death indicator, CD4 cell count, time points at which the CD4 cells count were recorded, treatment indicator, gender indicator, previous opportunistic infection (AIDS diagnosis) at study entry indicator, and AZT tolerance vs AZT failure indicator.

Since this dataset contains both longitudinal and survival data, we propose a joint model to simultaneously capture the random effects from the longitudinal data and to estimate the parameters of the survival model. A joint model is preferred to a two-step approach since previous studies have shown that the two-step approach may reduce efficacy which in turn inflates standard errors of estimates. For joint modelling, previous literature suggests that Bayesian methods are more efficient and computationally easier.

2 Aims

Our study aims are as follows:

(1). We will assess the association between CD4 cell count with time, adjusting for other covariates such as obstime, drug, prevOI (indicator of previous opportunistic infection (AIDS diagnosis) at study entry), and AZT tolerance.

(2). We will use the joint model to assess the association between survival outcome (death status and event time) with the drug efficacy and the latent longitudinal mean.

(3). We will assess and compare the model fitting result (mean and standard error) for the MCEM method and Bayesian method, and machine learning method (RSF). We will also do a simulation to compare the model fitting result.

(4). We will also propose the prediction to patients with some longitudinal outcomes available (see Method section for detailed demonstration).

3 Method

Joint modelling framework:

$$\begin{aligned} Y_{ij} =& m_i(t_{ij}) + \epsilon_{ij} && (1) \\ h_i(T_i) =& h_0 \exp(\boldsymbol{w_i'}\boldsymbol{\gamma} + \alpha m_i(T_i)) && (2) \end{aligned}$$

where $m_i(t_{ij}) = \beta_0 + \beta_1 t_{ij} + \boldsymbol{x_{ij}}'\boldsymbol{\beta} + u_i$ and $m_i(T_i) = \beta_0 + \beta_1 T_i + x_i'\beta + u_i$

We have $i = 1, 2, \dots, n$ subjects, $j = 1, 2, \dots, J_i$ visits for each subject. We observe the longitudinal outcome Y_{ij} for subject i at visit j, and we observe the patient's covariate $\boldsymbol{x_{ij}}$ at time t_{ij}. We also observe the event time $T_i = \min(T_i^*, C_i)$, where T_i^* is the actual failure time, and C_i is the censoring time. We also observe the censoring indicator $\delta_i = I(T_i^* \leq C_i)$, and drug type $\boldsymbol{w_i}$.

Here, $m_i(t_{ij})$ is the latent longitudinal mean at time t_{ij}, and u_i is the random effect for each subject i. And $h_i(T_i)$ is the hazard function for subject i, and h_0 is the baseline hazard function.

Assumption: $u_i \sim N(0, \sigma_u^2)$; $\epsilon_{ij} \sim N(0, \sigma_e^2)$; $Y_{ij}|u_i \perp\!\!\!\perp Y_{ij'}|u_i$; $T_i^* \perp\!\!\!\perp C_i$; $T_i \perp\!\!\!\perp Y_{ij}|u_i$.

We write the parameter space $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\beta}, \sigma_u, \sigma_e, \boldsymbol{\gamma}, \alpha)'$. We posit a nonparametric function to the baseline hazard function.

We write the full data likelihood:

$$L(\boldsymbol{\theta}|Y, T) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} [\prod_{j=1}^{J_i} P(Y_{ij}|\beta, u_i)][h_i(T_i)^{\delta_i} S_i(T_i)] P(u_i|\sigma_u^2) du \tag{3}$$

We write the complete data likelihood (assume random effect is known):

$$L_C(\boldsymbol{\theta}|Y, T) = \prod_{i=1}^{n} [\prod_{j=1}^{J_i} P(Y_{ij}|\beta, u_i)][h_i(T_i)^{\delta_i} S_i(T_i)] P(u_i|\sigma_u^2) \tag{4}$$

We can write the complete data likelihood, treating all parameters as random (Bayesian prospective):

$$L_B(\boldsymbol{\theta}|Y,T) = \prod_{i=1}^{n}[\prod_{j=1}^{J_i} P(Y_{ij}|\beta, u_i)][h_i(T_i)^{\delta_i} S_i(T_i)]P(u_i|\sigma_u^2)P(\boldsymbol{\theta}) \tag{5}$$

where we denote $P(\boldsymbol{\theta})$ as the prior distribution of $\boldsymbol{\theta}$.

Suppose for a new subject i', and we observe his longitudinal outcomes and covariates up to time T, where $j_{i'}$ is the number of visits up to time T for subject i'. We want to predict his future longitudinal outcomes, and survival probability at time T'.

We can sample the random effect for the subject i' using the following equation:

$$P(u_{i'}|Y_{i'}^{(T)}, T_{i'} > T, \hat{\boldsymbol{\theta}}) \propto \prod_{j=1}^{j_{i'}} P(Y_{i'j}|u_i', \hat{\boldsymbol{\theta}})P(T_{i'} > T|u_{i'}, \hat{\boldsymbol{\theta}})P(u_{i'}|\hat{\boldsymbol{\theta}}) \tag{6}$$

We sample from the above distribution, using sampling methods from module 2.

Suppose we have D samples of $u_{i'}$, we can then predict his longitudinal outcomes at time T', and his conditional survival probability at time T' as:

$$m_{i'}(T') = \hat{\beta}_0 + \hat{\beta}_1 T' + \boldsymbol{x'_{i'j}}\hat{\boldsymbol{\beta}} + u_{i'}^{(d)} \tag{7}$$

$$P(T_{i'} > T'|T_{i'} > T) = S_{i'}(T'|\hat{\boldsymbol{\theta}}, u_{i'}^{(d)})/S_{i'}(T|\hat{\boldsymbol{\theta}}, u_{i'}^{(d)}) \tag{8}$$

We can calculate the area under curve (AUC) and Brier Score (BS) to compare the prediction accuracy.

4 Analysis Plan

We propose to use EM algorithm, Bayesian methods, and machine learning methods to estimate parameters.

For EM algorithm: Suppose we have the parameter estimate at iteration t. We can sample the random effect u_i from the posterior distribution of u_i, and do the maximization step.

We propose the Q-function as:

$$Q(Y|\theta^{(t)}) = E(l_c(\theta^{(t)}|Y,T)) \tag{9}$$

$$= \sum_{i=1}^{n} \int l_i(\theta|Y_i, T_i) f(u_i|\theta^{(t)}, Y_i, T_i) du \tag{10}$$

where $l_i(\theta|Y_i, T_i) = \sum_{j=1}^{J_i} logP(Y_{ij}|\theta, u_i) + log[h_i(T_i)^{\delta_i} S_i(T_i)] + logP(u_i|\theta)$,

$f(u_i|\theta^{(t)}, Y_i, T_i) \propto \prod_{j=1}^{j_i} P(Y_{ij}|u_i, \boldsymbol{\theta^{(t)}}) h_i(T_i|\theta^{(t)})^{\delta_i} S_i(T_i|\theta^{(t)}) P(u_i|\boldsymbol{\theta^{(t)}})$.

We will use the Metropolis Hastings Algorithm with a random walk to draw the samples of random effect and calculate the integral by Monte Carlo method. To tune the variance of the proposal distributions, we will first test the algorithm for a small number of iterations to achieve around 25% acceptance rate for optimal efficiency. We will use non-informative priors for all the parameters.

For Bayesian methods: We will use Stan to perform sampling, where we will use No-U-Turn Sampler (NUTS), based on Hamiltonian Monte Carlo (HMC).

We will use normal prior distributions for β..., and inverse Gamma prior distributions for the variance parameters. We will use the following prior distribution for the parameters for Bayesian methods:

(1) $\beta \sim N(0, 10^2)$.
(2) $\gamma \sim N(0, 10^2)$.
(3) $\alpha \sim N(0, 10^2)$.
(4) $\sigma_u \sim$ Inverse Gamma$(0.1, 0.1)$.
(5) $\sigma_e \sim$ Inverse Gamma$(0.1, 0.1)$.

We will use a constant baseline hazard function for simplicity.

For machine learning methods: We plan to use random survival forest to check prediction accuracy, where we will use the CD4 count nearest to survival time as the substitute of latent longitudinal mean.

We will compare the model estimates, standard error, computation speed, and prediction accuracy. We will use cross-validation to check prediction accuracy.

5 Simulation

We compare the model fitting results using the simulation set-up as follows:

First, we simulate the longitudinal responses using the following formula: $Y_{ij} = 20 + (-1) * t_{ij} + (-4) * x_i + u_i + \epsilon_{ij}$, where $u_i \sim N(0, 2), \epsilon_{ij} \sim N(0, 1)$, $x_i \sim binom(1, 0.4)$.

Second, we simulate N random uniform distribution, which corresponds to the survival probability. And we calculate the failure time as: $h_i(T_i) = exp(-2)\exp(-1 * w_i + 0.2 * m_i(T_i))$, where $m_i(T_i) = 20 + (-1) * T_i + (-4) * x_i + u_i$. And we can calculate the survival time using an explicit formula.

We simulate N=500 subjects, and time points as (0, 2, 6, 12, 18) as from the 'aids' data in the longitudinal model, and another 200 subjects as the testing dataset.

6 References

Li, Kan, and Sheng Luo. "Dynamic Predictions in Bayesian Functional Joint Models for Longitudinal and Time-to-Event Data: An Application to Alzheimer's Disease." Statistical Methods in Medical Research 28, no. 2 (2017): 327–42. https://doi.org/10.1177/0962280217722177.

http://www.drizopoulos.com/vignettes/multivariate%20joint%20models

http://www.drizopoulos.com/vignettes/dynamic_predictions