Imputation model - Competing risks
10 mai 2019

Introduction

— Due to its flexibility, its practicability and its efficiency compared to the complete case analysis, **multiple imputation by chained equations (MICE)** is widely used to impute missing data when covariates have missing values.
— Imputation models should [1] :
  — be **congenial** to the analysis model, i.e, both models should be compatible with some larger model for the data
  — **include the outcome** and the covariates of the analysis model
— In **survival analysis**
  — Outcome is defined by a binary event indicator D and the observed event or censoring time T.
  — Estimates are obtained generally by direct inclusion of D and T (or $\log(T)$) in the imputation model
  — Estimations may still be biaised, even using a MICE procedure with predictive mean matching as recommended by Marshall [2]
  — I. White and P. Royston showed that the imputation model should include the event indicator and the **cumulative baseline hazard** instead of T (or $\log(T)$) , and therefore recommended to include the **Nelson-Aalen estimator** of the cumulative hazard in the imputation model [3]
— In the **competing-risks** setting
  — Subjects may experiment one out of K distinct and exclusive events. Outcome thus is defined by an event indicator ε and the observed event or censoring time T.
  — Two main approaches have been proposed. The most common approach models the cause-specific hazard of the event of interest while the second approach models the subdistribution hazard associated to the cumulative incidence function.
— We propose to extend the work of I. White and P. Royston to the competing-risks setting by including in the imputation model the cumulative hazard associated with the hazard function of the analysis model. Moreover, we will show that **cumulative hazards of all the events should be included** in case of cause-specific analysis.

Notations

Suppose a competing-risks setting, in which subjects may fail from one out of K distinct and exclusive causes of failure.

Let be :
— X a single incomplete variable
— Z a vector of complete variables
— (T^*, ε), where T^* is the minimum of failure time T and the right-censoring time C. $\varepsilon \in \{1, ..., K\}$ denotes the failure cause and $\varepsilon = 0$ denotes a right-censored observation.

$\forall i \in \{1, ..., K\}$, let define the following functions :
— $F_i(t) = P(T \leq t, \varepsilon = i)$, the cumulative incidence of the failure cause i
— $S(t) = 1 - \sum_{i=1}^{K} F_i(t)$, the global survival function
— $h(t) = -\frac{\delta \log(1-F(t))}{\delta t}$
— $H(t)$ cumulative hazard of $h(t)$
— $f_i(t) = \frac{\delta F_i(t)}{\delta t}$
— $h_i(t) = h_i(t, \varepsilon = i) = \lim_{\delta t \to 0} \frac{P(t \leq T < t+\delta t, \varepsilon = i | T \geq t)}{\delta t}$, the cause specific hazard for the failure cause i7
— $h_i(t) = \frac{1}{S(t)} \frac{\delta F_i(t)}{\delta t} = \frac{1}{1-\sum_{j=1}^{K} F_j(t)} \frac{\delta F_i(t)}{\delta t}$
— $H_i(t)$ the cumulative hazard of $h_i(t)$
— $\sum_{j=1}^{K} H_j(t) = -\log(S(t))$
— $\lambda_i(t) = \lambda_i(t, \varepsilon = i) = \lim_{\delta t \to 0} \frac{P(t \leq T < t+\delta t, \varepsilon = i | (T \geq t \cup (T \leq \cap \varepsilon \neq i)))}{\delta t}$, the subdistrubution hazard associated with the cumulative incidence of the failure cause i
— $\lambda_i(t) = \frac{1}{1-F_i(t)} \frac{\delta F_i(t)}{\delta t} = -\frac{\delta \log(1-F_i(t))}{\delta t}$

— $\Lambda_i(t)$ the cumulative hazard of $\lambda_i(t)$

Cause specific approach

We assume that censoring is non-informative. The likelihood for the failures, given complete data, is :

$$P(T^*, \varepsilon|X,Z) = \prod_{i=1}^{K} \left\{f_i(T^*|X,Z)^{1_{[\varepsilon=i]}}\right\} \times S(T^*|X,Z)^{1_{[\varepsilon=0]}} \tag{1}$$

$$= \prod_{i=1}^{K} \left\{h_i(T^*|X,Z)^{1_{[\varepsilon=i]}}\right\} \times S(T^*|X,Z) \tag{2}$$

We obtain :

$$\log(P(T^*, \varepsilon|X,Z)) = \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]} \log(h_i(T^*|X,Z))\right\} + \log(S(T^*|X,Z)) \tag{3}$$

$$= \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]} \log(h_i(T^*|X,Z))\right\} - \sum_{i=1}^{K} H_i(T^*|X,Z) \tag{4}$$

$$= \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]} \log(h_i(T^*|X,Z)) - H_i(T^*|X,Z)\right\} \tag{5}$$

Analysis model

Assuming a cause specific proportional hazard model for each failure cause i :

$$h_i(t|X,Z) = h_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)$$

Then, Equation 3 becomes :

$$\log(P(T^*, \varepsilon|X,Z)) = \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]} \log(h_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)) - H_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)\right\} \tag{6}$$

$$= \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]} \log(h_{i0})(t) + 1_{[\varepsilon=i]}(\beta_{iX}X + \beta_{iZ}Z) - H_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)\right\} \tag{7}$$

Using the Bayes theorem :

$$\begin{aligned}
\log(P(X|T^*, \varepsilon, Z) &= \log(P(X|Z))) + \log(P(T^*, \varepsilon|X,Z)) + const \\
&= \log(P(X|Z)) + \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]} \log(h_{i0}(t)) + 1_{[\varepsilon=i]}(\beta_{iX}X + \beta_{iZ}Z) - H_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)\right\} + const \\
&= \log(P(X|Z)) + \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]} \log(h_{i0}(t))\right\} + \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]}\beta_{iX}X\right\} \\
&\qquad - \sum_{i=1}^{K} \{H_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)\} + const
\end{aligned}$$

where the constant may depend on ε, T and Z but not on X.

Binary X

Writing $\text{logit}(p(X=1|Z)) = \zeta_Z$

$$\begin{aligned}
\text{logit}(p(X=1|T^*, \varepsilon, Z)) &= \log(p(X=1|T^*, \varepsilon, Z)) - \log(p(X=0|T^*, \varepsilon, Z)) \\
&= \zeta_Z + \sum_{i=1}^{K} \left\{1_{[\varepsilon=i]}\beta_{iX}\right\} - \sum_{i=1}^{K} \{H_{i0}(t)\exp(\beta_{iZ}Z)(\exp(\beta_{iX}) - 1)\}
\end{aligned}$$

If there no Z :

$$\begin{aligned}\operatorname{logit}(p(X=1|T^*,\varepsilon)) &= \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\beta_{iX}\right\} - \sum_{i=1}^{K}\left\{H_{i0}(t)(\exp(\beta_{iX})-1)\right\} \\ &= \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\beta_{iX}\right\} + \sum_{i=1}^{K}\left\{\beta'_{iX}H_{i0}(t)\right\}\end{aligned}$$

If we assume $\operatorname{logit}(p(X=1|Z)) = \zeta_0 + \zeta_1 Z$:

$$\begin{aligned}\operatorname{logit}(p(X=1|T^*,\varepsilon,Z)) &\approx \zeta'_0 + \zeta_1 Z + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\beta_{iX}\right\} - \sum_{i=1}^{K}\left\{H_{i0}(t)\exp(\beta_{iZ}\bar{Z})(\exp(\beta_{iX})-1)\right\} \\ &\approx \zeta'_0 + \zeta_1 Z + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\beta_{iX}\right\} + \sum_{i=1}^{K}\left\{\beta'_{iX}H_{i0}(t)\right\}\end{aligned}$$

or more accurately with a interaction term.

$$\operatorname{logit}(p(X=1|T^*,\varepsilon,Z)) \approx \zeta'_0 + \zeta_1 Z + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\beta_{iX}\right\} + \sum_{i=1}^{K}\left\{\beta'_{iX}H_{i0}(t)\right\} + \sum_{i=1}^{K}\left\{\beta''_{iX}H_{i0}(t)Z\right\}$$

General formulas

Following Ian White approximations [**?**], imputation models becomes :

$$X|T^*,\varepsilon,Z \sim \zeta'_0 + \zeta_1 Z + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\beta_{iX}\right\} + \sum_{i=1}^{K}\left\{\beta'_{iX}H_{i0}(t)\right\}$$

or with interaction

$$X|T^*,\varepsilon,Z \sim \zeta'_0 + \zeta_1 Z + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\beta_{iX}\right\} + \sum_{i=1}^{K}\left\{\beta'_{iX}H_{i0}(t)\right\} + \sum_{i=1}^{K}\left\{\beta''_{iX}H_{i0}(t)Z\right\}$$

Subdistribution hazard approach

Using Equation 1, we can write :

$$P(T^*,\varepsilon|X,Z) = \prod_{i=1}^{K}\left\{f_i(T^*|X,Z)^{1_{[\varepsilon=i]}}\right\} \times S(T^*|X,Z)^{1_{[\varepsilon=0]}} \tag{8}$$

$$= \prod_{i=1}^{K}\left\{\lambda_i(T^*|X,Z)^{1_{[\varepsilon=i]}}(1-F_i(T^*|X,Z))^{1_{[\varepsilon=i]}}\right\} \times S(T^*|X,Z)^{1_{[\varepsilon=0]}} \tag{9}$$

We obtain :

$$\log(P(T^*,\varepsilon|X,Z)) = \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\log(\lambda_i(T^*|X,Z))\right\} + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\log(1-F_i(T^*|X,Z))\right\} + 1_{[\varepsilon=0]}\log(S(T^*|X,Z)) \tag{10}$$

$$= \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\log(\lambda_i(T^*|X,Z))\right\} - \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\Lambda_i(T^*|X,Z)\right\} + 1_{[\varepsilon=0]}\log(S(T^*|X,Z)) \tag{11}$$

Analysis model

Assuming a proportional hazard model for the subdistribution hazard of failure cause i :

$$\lambda_i(t|X,Z) = \lambda_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)$$

Then, Equation 10 becomes :

$$\log(P(T^*,\varepsilon|X,Z)) = \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\log(\lambda_{i0})(t)\right\} + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX}X + \beta_{iZ}Z)\right\} - \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\Lambda_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)\right\} \tag{12}$$

$$+ 1_{[\varepsilon=0]}\log(S(T^*|X,Z)) \tag{13}$$

Using the Bayes theorem :

$$\begin{aligned}
\log(P(X|T^*,\varepsilon,Z) &= \log(P(X|Z))) + \log(P(T^*,\varepsilon|X,Z)) + const \\
&= \log(P(X|Z)) + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\log(\lambda_{i0})(t)\right\} + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX}X + \beta_{iZ}Z)\right\} \\
&\quad - \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\Lambda_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)\right\} + 1_{[\varepsilon=0]}\log(S(T^*|X,Z)) + const \\
&= \log(P(X|Z)) + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\log(\lambda_{i0})(t)\right\} + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX}X)\right\} \\
&\quad - \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\Lambda_{i0}(t)\exp(\beta_{iX}X + \beta_{iZ}Z)\right\} + 1_{[\varepsilon=0]}\log(S(T^*|X,Z)) + const
\end{aligned}$$

where the constant may depend on ε, T and Z but not on X.

Binary X

Writing $\text{logit}(p(X=1|Z)) = \zeta_Z$

$$\begin{aligned}
\text{logit}(p(X=1|T^*,\varepsilon,Z)) &= \log(p(X=1|T^*,\varepsilon,Z)) - \log(p(X=0|T^*,\varepsilon,Z)) \\
&= \zeta_Z + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX})\right\} - \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\Lambda_{i0}(t)\exp(\beta_{iZ}Z)(\exp(\beta_{iX})-1)\right\} \\
&\quad + 1_{[\varepsilon=0]}\left\{\log(S(T^*|1,Z)) - \log(S(T^*|0,Z))\right\}
\end{aligned}$$

If there no Z :

$$\begin{aligned}
\text{logit}(p(X=1|T^*,\varepsilon)) &= \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX})\right\} - \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}\Lambda_{i0}(t)(\exp(\beta_{iX})-1)\right\} + 1_{[\varepsilon=0]}\left\{\log(S(T^*|1)) - \log(S(T^*|0))\right\} \\
&= \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX} + \beta'_{iX}\Lambda_{i0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\log(S(T^*|1)) - \log(S(T^*|0))\right\} \\
&= \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX} + \beta'_{iX}\Lambda_{i0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\log(1-\sum_{i=1}^{K}F_i(T^*|1)) - \log(1-\sum_{i=1}^{K}F_i(T^*|0))\right\} \\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX} + \beta'_{iX}\Lambda_{i0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}F_i(T^*|0) - \sum_{i=1}^{K}F_i(T^*|1)\right\} \\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX} + \beta'_{iX}\Lambda_{i0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}F_i(T^*|0) - \sum_{i=1}^{K}F_i(T^*|1)\right\} \\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX} + \beta'_{iX}\Lambda_{i0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}\exp(\Lambda_i(T^*|0)) - \sum_{i=1}^{K}\exp(\Lambda_i(T^*|1))\right\} \\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX} + \beta'_{iX}\Lambda_{i0}(t))\right\} - 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}\left\{\Lambda_{i0}(t)(\exp(\beta_{iX})-1)\right\}\right\} \\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=i]}(\beta_{iX} + \beta'_{iX}\Lambda_{i0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}\left\{\beta''_{iX}\Lambda_{i0}(t)\right\}\right\}
\end{aligned}$$