Biostatistics (2015), **0**, 0, *pp.* 1–4
doi:10.1093/biostatistics/supp˙materials˙2015˙09˙09

Missing covariates in competing risks analysis Supplementary Materials

JONATHAN W. BARTLETT*

Department of Medical Statistics, London School of Hygiene and Tropical Medicine,

Keppel Street, London, WC1E 7HT, UK

jonathan.bartlett@lshtm.ac.uk

JEREMY M.G. TAYLOR

Department of Biostatistics, School of Public Health, University of Michigan, MI, Ann Arbor,

U.S.A.

APPENDIX

A. Validity of complete case analysis

Without loss of generality, we consider the cause specific hazard function for the first cause in the complete cases, $h_1(t|X,Z,R=1)$. This is equal to

$$
\begin{aligned}
&\lim_{h\to 0}\frac{1}{h}P(t\leqslant Y<t+h, D=1|R=1,X,Z,Y\geqslant t)\\
&=\lim_{h\to 0}\frac{1}{h}\frac{P(t\leqslant T<t+h, D^*=1, C\geqslant t, R=1|X,Z)}{P(R=1,T\geqslant t, C\geqslant t|X,Z)}\\
&=\lim_{h\to 0}\frac{1}{h}\frac{P(t\leqslant T<t+h, D^*=1, C\geqslant t|X,Z)}{P(T\geqslant t, C\geqslant t|X,Z)}\times\frac{P(R=1|t\leqslant T<t+h, D^*=1, C\geqslant t, X,Z)}{P(R=1|T\geqslant t, C\geqslant t,X,Z)}\\
&=\lim_{h\to 0}\frac{1}{h}\frac{P(t\leqslant T<t+h, D^*=1|X,Z)P(C\geqslant t|X,Z)}{P(T\geqslant t|X,Z)P(C\geqslant t|X,Z)}\times\frac{P(R=1|T=t, D^*=1, C\geqslant t, X,Z)}{P(R=1|T\geqslant t, C\geqslant t,X,Z)}\\
&=\lim_{h\to 0}\frac{1}{h}\frac{P(t\leqslant T<t+h, D^*=1|X,Z)}{P(T\geqslant t|X,Z)}\times\frac{P(R=1|T=t, D^*=1, C\geqslant t, X,Z)}{P(R=1|T\geqslant t, C\geqslant t,X,Z)}\\
&=h_1(t|X,Z)\times\frac{P(R=1|T=t, D^*=1, C\geqslant t, X,Z)}{P(R=1|T\geqslant t, C\geqslant t,X,Z)} \qquad \text{(A.1)}
\end{aligned}
$$

A complete case analysis will thus give valid inferences when the second term in the preceding equation is equal to one. An obvious sufficient condition for this to hold is that missingness is covariate dependent, in the sense that $R \perp\!\!\!\perp (T,D^*,C)|(X,Z)$.

We now show that a weaker sufficient condition is that $R \perp\!\!\!\perp (T,D^*)|(C,X,Z)$, which permits missingness to depend on the time to censoring C, in addition to X and Z. To this end, we first show that under this assumption censoring remains independent in the complete cases ($(T,D^*) \perp\!\!\!\perp C|(X,Z,R=1)$), utilizing the assumption that $(T,D^*) \perp\!\!\!\perp C|(X,Z)$:

$$
\begin{aligned}
P(T,D^*,C|X,Z,R=1) &= \frac{P(T,D^*,C,R=1|X,Z)}{P(R=1|X,Z)}\\
&=\frac{P(R=1|T,D^*,C,X,Z)P(T,D^*,C|X,Z)}{P(R=1|X,Z)}\\
&=\frac{P(R=1|C,X,Z)P(T,D^*|X,Z)P(C|X,Z)}{P(R=1|X,Z)}\\
&=P(T,D^*|X,Z)P(C|X,Z,R=1)
\end{aligned}
$$

Using these results we then have that the second term in equation (A.1) is equal to

$$\begin{aligned}\frac{P(R=1|T=t,D^*=1,C\geqslant t,X,Z)}{P(R=1|T\geqslant t,C\geqslant t,X,Z)} &= \frac{\frac{P(T=t,D^*=1,C\geqslant t|X,Z,R=1)P(R=1|X,Z)}{P(T=t,D^*=1,C\geqslant t|X,Z)}}{\frac{P(T\geqslant t,C\geqslant t|X,Z,R=1)P(R=1|X,Z)}{P(T\geqslant t,C\geqslant t|X,Z)}} \\ &= \frac{\frac{P(T=t,D^*=1|X,Z)P(C\geqslant t|X,Z,R=1)}{P(T=t,D^*=1|X,Z)P(C\geqslant t|X,Z)}}{\frac{P(T\geqslant t|X,Z)P(C\geqslant t|X,Z,R=1)}{P(T\geqslant t|X,Z)P(C\geqslant t|X,Z)}} \\ &= 1\end{aligned}$$

such that the complete case analysis is valid.

B. Derivations for imputation distribution sampling

In this appendix we describe how missing values in X can be sampled, considering separately the case of categorical covariates and non-categorical covariates.

Categorical covariates If X has a finite sample space, we can directly sample from the imputation distribution. Specifically, let k be the constant of proportionality such that

$$P(X|Z,Y,D) = kf(Y,D|X,Z)P(X|Z)$$

Without loss of generality suppose that X has sample space $\{1,..,S\}$, such that

$$1 = \sum_{s=1}^{S} kf(Y,D|X=s,Z)P(X=s|Z)$$

Then we have

$$k = \frac{1}{\sum_{s=1}^{S} f(Y,D|X=s,Z)P(X=s|Z)}$$

and $P(X=s'|Z,Y,D)$ is equal to

$$\begin{aligned}&\frac{f(Y,D|X=s',Z)P(X=s'|Z)}{\sum_{s=1}^{S} f(Y,D|X=s,Z)P(X=s|Z)} \\ &= \frac{\prod_{k=1}^{K} \exp\left[-\exp\left\{g_k(X=s',Z,\beta_k)\right\}H_{0k}(Y)\right]\left[\exp\left\{g_k(X=s',Z,\beta_k)\right\}\right]^{I(D=k)} P(X=s'|Z)}{\sum_{s=1}^{S}\prod_{k=1}^{K} \exp\left[-\exp\left\{g_k(X=s,Z,\beta_k)\right\}H_{0k}(Y)\right]\left[\exp\left\{g_k(X=s,Z,\beta_k)\right\}\right]^{I(D=k)} P(X=s|Z)}\end{aligned}$$

Other covariate types More generally, rejection sampling can be used to draw from the distribution, using $f(X|Z)$ as the proposal distribution. To use rejection sampling the ratio of the target density to the proposal density must be bounded above, up to a constant of proportionality, by a quantity not involving X. From equation 3.1 of the main paper, here this ratio is simply equal to $f(Y, D|X, Z)$.

First suppose that $D = 0$, such that an individual is censored at time Y. Then we have $f(Y, D = 0|X, Z) \leqslant S_0(Y|Z)h_0(Y|Z)$. To sample a missing value of X^*, we sample from $f(X|Z)$, sample $U \sim U(0, 1)$, and accept X^* if

$$\begin{aligned} U &\leqslant \frac{f(Y, D = 0|X^*, Z)}{S_0(Y|Z)h_0(Y|Z)} \\ &= \prod_{k=1}^{K} \exp\left[-\exp\left\{g_k(X^*, Z, \beta_k)\right\} H_{0k}(Y)\right] \end{aligned}$$

Now suppose that $D > 0$. Then we have

$$\begin{aligned} f(Y, D|X, Z) &\leqslant S_0(Y|Z) \exp\left[-\exp\left\{g_D(X, Z, \beta_D)\right\} H_{0D}(Y)\right] h_{0D}(Y) \exp\left\{g_D(X, Z, \beta_D)\right\} \\ &= S_0(Y|Z)h_{0D}(Y) \exp\left[g_D(X, Z, \beta_D) - \exp\left\{g_D(X, Z, \beta_D)\right\} H_{0D}(Y)\right] \end{aligned}$$

Differentiation with respect to $g_D()$ shows that the expression takes its maximum value when $\exp\left\{g_D(X, Z, \beta_D)\right\} H_{0D}(Y) = 1$, so that

$$f(Y, D|X, Z) \leqslant S_0(Y|Z)h_{0D}(Y)\frac{\exp(-1)}{H_{0D}(Y)}$$

To sample a missing value of X^*, we sample from $f(X|Z)$, sample $U \sim U(0, 1)$, and accept X^* if

$$\begin{aligned} U &\leqslant f(Y, D|X^*, Z)\frac{\exp(1)H_{0D}(Y)}{S_0(Y|Z)h_{0D}(Y)} \\ &= H_{0D}(Y) \exp\{1 + g_D(X^*, Z, \beta_D)\} \prod_{k=1}^{K} \exp\left[-\exp\left\{g_k(X^*, Z, \beta_k)\right\} H_{0k}(Y)\right] \end{aligned}$$