Biometrics 00, 1–3 DOI: 000
000 000

Supporting Information for "Biased Estimation With Shared Parameter Models in the Presence of Competing Dropout Mechanisms" by Edward F. Vonesh and Tom Greene

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A: Log-likelihood functions for competing risk SP models

The conditional log-likelihood function for a competing risk shared parameter (SP) model can be derived directly from Kalbfleisch and Prentice (2002, Section 8.2.3). Specifically, to accommodate an observed time-dependent covariate y_{ij} as a MAR predictor of dropout along with subject-specific random intercept and slope effects, b_{0i} and b_{1i} as MNAR predictors of dropout, one can formulate the event-time likelihood function for two competing causes of dropout using the Nelson-Aalen estimator of the cumulative hazard function.

Following Kalbfleisch and Prentice (2002), let (T_i, δ_i) be the observed values of $(T_i^{\mathrm{MIN}}, \delta_i^{\mathrm{MIN}})$ under the three competing dropout mechanisms and let k_i be the i^{th} subject's cause of dropout which is unobserved when $\delta_i = 0$ (i.e., when subjects complete the study). We let $k = 1$ represent non-terminal dropout events while $k = 2$ represents terminal dropout events. Let $\boldsymbol{Z}_i(T_i)$ be the covariate path for the i^{th} subject through time T_i which, for the simulation study, is defined by the vector of covariates $\boldsymbol{Z}_i(t) = (y_{ij},\, b_{0i},\, b_{1i})$ associated with each of the competing dropout mechanisms at observation times $t = t_{ij}$. Since censoring is independent, we may express the overall likelihood function for $(T_i^{\mathrm{MIN}}, \delta_i^{\mathrm{MIN}})$, $i = 1, \ldots n$, as

$$
\begin{aligned}
L(\boldsymbol{\eta}) &= \prod_{i=1}^{n}\left(\Big\{h_{k_i}[T_i;\ \boldsymbol{Z}_i(T_i)]\Big\}^{\delta_i}\prod_{k=1}^{2}\exp\Big\{-\int_0^{T_i} h_k[u; \boldsymbol{Z}_i(u)]du\Big\}\right) \qquad (1)\\
&= \prod_{k=1}^{2}\left(\prod_{i=1}^{n}\Big\{h_k[T_i; \boldsymbol{Z}_i(T_i), \boldsymbol{\eta}_k]\Big\}^{\delta_{ki}}\exp\Big\{-\int_0^{\infty}\sum_{i=1}^{n} Y_i(t)h_k[t; \boldsymbol{Z}_i(t), \boldsymbol{\eta}_k]dt\Big\}\right)
\end{aligned}
$$

where $Y_i(t)$ is the at risk indicator at time t with $Y_i(t) = 1$ if $T_i > t$ and 0 otherwise, δ_{ki} is the indicator of a type k dropout event for the i^{th} subject where, in the simulation but unbeknowst to the investigator, $\delta_{1i} = \delta_i^{\mathrm{MAR}}$ and $\delta_{2i} = \delta_i^{\mathrm{MNAR}}$. Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ where $\boldsymbol{\eta}_k$ is the vector of parameters associated with the risk-specific hazard function, $h_k[t; \boldsymbol{Z}_i(t), \boldsymbol{\eta}_k]$. Under the piecewise exponential survival model, the i^{th} subject's contribution to the overall log-likelihood function for $(T_i^{\text{MIN}}, \delta_i^{\text{MIN}})$ can be expressed as

$$l_i(\boldsymbol{\eta}; T_i, \delta_i) = \log\Bigg[\prod_{k=1}^{2}\Bigg(\Big\{h_k[T_i; \boldsymbol{Z}_i(T_i), \boldsymbol{\eta}_k]\Big\}^{\delta_{ki}} \exp\Big\{-\int_0^\infty Y_i(t)h_k[t; \boldsymbol{Z}_i(t), \boldsymbol{\eta}_k]dt\Big\}\Bigg)\Bigg] \quad (2)$$
$$= \sum_{k=1}^{2}\Bigg(\delta_{ki}\log\Big\{h_k[T_i; \boldsymbol{Z}_i(T_i), \boldsymbol{\eta}_k]\Big\} + \Big\{-\int_0^\infty Y_i(t)h_k[t; \boldsymbol{Z}_i(t), \boldsymbol{\eta}_k]dt\Big\}\Bigg)$$
$$= \sum_{k=1}^{2}\Bigg(\delta_{ki}\log\Big\{h_k[T_i; \boldsymbol{Z}_i(T_i), \boldsymbol{\eta}_k]\Big\} - \Lambda_k[T_i;\ \boldsymbol{Z}_i(T_i), \boldsymbol{\eta}_k]\Bigg)$$

where $\Lambda_k[T_i;\ \boldsymbol{Z}_i(T_i), \boldsymbol{\eta}_k]$ are the cumulative hazard functions under non-terminal dropout events ($k = 1$) and terminal dropout events ($k = 2$), respectively.

Under a piecewise exponential survival model with constant baseline hazard rates λ_{0kh} reparameterized as $\eta_{kh} = \log(\lambda_{0kh})$, the parameter vectors for the two competing risks for dropout can be written as

$$\boldsymbol{\eta}_k' = (\eta_{k1}, \eta_{k2}, \eta_{k3}, \eta_{k4}, \eta_{ky}, \eta_{kb_0}, \eta_{kb_1}),\ k = 1, 2$$

corresponding to the log-hazard functions

$$\log\Big\{h_1[t; \boldsymbol{Z}_i(t), \boldsymbol{\eta}_1]\Big\} = \log\Big\{h_{01}(t)\exp(\eta_{1y}y_{ij} + \eta_{1b_0}b_{0i} + \eta_{1b_1}b_{1i})\Big\} \text{ for } k = 1 \quad (3)$$
$$\log\Big\{h_2[t; \boldsymbol{Z}_i(t), \boldsymbol{\eta}_2]\Big\} = \log\Big\{h_{02}(t)\exp(\eta_{2y}y_{ij} + \eta_{2b_0}b_{0i} + \eta_{2b_1}b_{1i})\Big\} \text{ for } k = 2.$$

Here $h_{0k}(t) = \exp\Big\{\sum_{h=1}^4 \eta_{kh} I(t_{h-1} < t \leq t_h)\Big\} = \exp(\eta_{kh}) = \lambda_{0kh}$ serves as the stepwise cause-specific baseline hazard function for $t \in (t_{h-1},\ t_h]$.

Given this set-up one can fit a competing risk SP model with the following marginal joint log-likelihood function for the i^{th} subject

$$l_i(\boldsymbol{\theta}; \boldsymbol{y}_i, T_i,\ \delta_i) = \log\int_{\boldsymbol{b}} \exp\Big\{l_i(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}_i|\boldsymbol{b}_i) + l_i(\boldsymbol{\eta}; T_i, \delta_i|y_{ij},\ \boldsymbol{b}_i) + l_i(\boldsymbol{\tau}; \boldsymbol{b}_i)\Big\}d\boldsymbol{b}_i \quad (4)$$
$$= \log\int_{\boldsymbol{b}} \exp\Big\{l_i(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}_i|\boldsymbol{b}_i) + \sum_{k=1}^{2} l_{ki}(\boldsymbol{\eta}_k; T_i, \delta_{ki}|y_{ij},\ \boldsymbol{b}_i) + l_i(\boldsymbol{\tau}; \boldsymbol{b}_i)\Big\}d\boldsymbol{b}_i$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}, \boldsymbol{\eta})$ is the overall vector of parameters to be estimated with $\boldsymbol{\beta}$ and σ^2 being the regression parameters and variance of $Y_{ij}|\boldsymbol{b}_i$ assuming conditional independence;

$\boldsymbol{\tau}$ is the vector of variance-covariance parameters of $\boldsymbol{b}_i$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ are the parameters for $(T_i^{\mathrm{MIN}}, \delta_i^{\mathrm{MIN}})$ as defined above. The term $l_i(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}_i|\boldsymbol{b}_i)$ is the conditional log-likelihood function for $\boldsymbol{Y}_i|\boldsymbol{b}_i$ and $l_{ki}(\boldsymbol{\eta}_k; T_i, \delta_{ki}|y_{ij},\, \boldsymbol{b}_i) = \delta_{ki} \log\Big\{h_k[T_i; \boldsymbol{Z}_i(T_i), \boldsymbol{\eta}_k]\Big\} - \Lambda_k[T_i;\, \boldsymbol{Z}_i(T_i), \boldsymbol{\eta}_k]$ are the competing risk conditional log-likelihood functions of $(T_i, \delta_i|y_{ij},\, \boldsymbol{b}_i)$ for non-terminal $(k = 1)$ and terminal $(k = 2)$ dropout events. Finally, $l_i(\boldsymbol{\tau}; \boldsymbol{b}_i)$ is the log-likelihood function for the random effects $\boldsymbol{b}_i$. Maximum likelihood estimation is achieved by maximizing the overall integrated log-likelihood function

$$l(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}, \boldsymbol{\eta} \,|\, \boldsymbol{y}, \boldsymbol{T}, \boldsymbol{\delta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}; \boldsymbol{y}_i, T_i,\, \delta_i) \tag{5}$$

were $(\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{\delta})$ are the observable quantities for the marker and event times.

The competing risk SP model represented by (1)-(5) corresponds to just two unique observable causes of dropout which, for the simulation study, happen to coincide with a single MAR dropout mechanism and a single MNAR dropout mechanism. However, one can easily generalize this approach by including several observable causes of dropout as well as by expanding the number of covariates included in the hazard functions whether those covariates are functions of the observed data (i.e., predictors of a MAR type dropout) or functions of unobserved random-effects (i.e., predictors of a MNAR type dropout). Of course such generalizations must take into account the frequency of the cause-specific dropout events so as to be adequately powered to detect a possible association between the covariates and dropout.

References

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, 2nd Edition.* Hoboken, NJ: John Wiley & Sons, Inc.