May 17, 2007

Supplement to "Baseline and Treatment Effect Heterogeneity for Survival Times between Centers in a Random Effects Accelerated Failure Time Model with Flexible Error Distribution"

Arnošt Komárek[1], Emmanuel Lesaffre[1] and Catherine Legrand[2]

[1] *Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B–3000, Leuven, Belgium*

[2] *European Organisation for Research and Treatment of Cancer, E. Mounierlaan 83/11, B–1200, Brussel, Belgium*

E-mail: `Arnost.Komarek@med.kuleuven.be`

This supplement is devoted to a brief description of the R (R Development Core Team, 2007) package bayesSurv to perform the analysis presented in Section 4 of the main paper (Komárek, Lesaffre, and Legrand, 2007). We assume that the data are stored in a `data.frame` called `eortc` which has a structure as indicated in Table 1. The column `id` identifies patients, column `center` different centers. The DFS time is found in the column `DFStime` and a censoring indicator (0 for right-censored and 1 for observed event times) is given in the column `DFSevent`. The values of covariates are given in columns labeled `trtmt`, `ageGroup`, `typeSur`, `tumSize`, `nodStat`, `otDis`, `region`. The columns corresponding to non-dichotomous covariates (`ageGroup` and `region`) are assumed to be created by the R function `factor` with appropriately chosen reference category.

Firstly, we specify the number of knots (`K`), the distance between 2 consecutive knots expressed as a multiple of the basis standard deviation σ (`c4delta`), order of the penalty s (`order`) and prior choices for the intercept α, scale τ and the smoothing hyperparameter λ. Specified choices are stored in a list `prior.error`:

```
> prior.error <- list(K=15, c4delta=1.5, order=3,
+   prior.intercept="normal", mean.intercept=0, var.intercept=100,
+   prior.scale="gamma", shape.scale=1, rate.scale=0.005,
+   prior.lambda="gamma", shape.lambda=1, rate.lambda=0.005)
```

Secondly, the prior choices for fixed effects $\boldsymbol{\beta}$, the mean of the random effects b_2 (parameter γ_2) and the random effects $\boldsymbol{b}$ are specified and stored as lists `prior.betaGamma` and `prior.b`.

```
> prior.betaGamma <- list(mean.prior=rep(0, 11), var.prior=rep(100, 11))
> prior.b <- list(prior.D = "inv.wishart", df.D = 2, scale.D = 0.002*c(1,0,1))
```

Table 1: Structure of the R `data.frame` `eortc` holding the data.

id	center	DFStime	DFSevent	trtmt	ageGroup
1	11	5 139	0	1	40--50
2	31	4 163	0	0	<40
3	41	733	1	1	>50
⋮	⋮	⋮	⋮	⋮	⋮

typeSur	tumSize	nodStat	otDis	region
0	1	0	0	NL
1	0	0	0	F
0	1	1	0	SE
⋮	⋮	⋮	⋮	⋮

Note that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})'$ in the model with region so that there are 11 'β' and 'γ' parameters to be estimated.

To start the MCMC sampling, it is useful to give some reasonable initial values for the regression parameters. These can be found, e.g., by fitting an AFT model without random effects using the standard R function survreg with, e.g., normal error distribution:

```
> library(survival)
> fit0 <- survreg(Surv(DFStime, DFSevent)~trtmt+ageGroup+typeSur+
+   tumSize+nodStat+otDis+region, dist="lognormal", data=eortc)
```

Initial $\boldsymbol{\beta}$ vector will be the vector of the estimates from survreg after removing intercept and treatment effect, initial γ parameter will be the estimate of the treatment effect from survreg:

```
> beta.init <- fit0$coeff[-(1:2)]
> gamma.init <- fit0$coeff["trtmt"]
```

Initial values of some model parameters, specification of the middle knot μ_0 and the basis standard deviation σ will be stored in a list called `init`.

```
> init <- list(beta = c(gamma.init, beta.init), D = c(1, 0, 1), lambda = 100,
+   intercept = fit0$coeff["(Intercept)"], scale = fit0$scale,
+   gamma = 0, sigma = 0.2)
```

In the list `init`, component `beta` determines initials for the regression parameters $(\boldsymbol{\beta}', \gamma_2)'$, component `D` gives a lower triangle of the initial covariance matrix $\mathbb{D}$ of the random effects, component λ stores the initial value of the smoothing hyperparameter λ. Further, components `intercept` and `scale` give the initial values of the error intercept α and scale τ, respectively. Finally, components `gamma` and `sigma` determine the value of the middle knot μ_0 and the basis standard deviation σ, respectively.

The core part of the analysis, MCMC sampling, is then performed using the function `bayessurvreg2` in the following way:

```
> library(bayesSurv)
> sample <- bayessurvreg2(Surv(DFStime, DFSevent)~trtmt+ageGroup+typeSur+
+   tumSize+nodStat+otDis+region+cluster(center), random=~trtmt,
```

```
+   prior=prior.error, init=init,
+   prior.beta=prior.betaGamma, prior.b=prior.b,
+   nsimul=list(niter=125000, nthin=5, nburn=100000), store=list(b=TRUE),
+   dir="/home/userAK/", data=eortc)
```

Sampled chains are then found in the form of ASCII files having an extension `.sim` in the directory called "`/home/userAK/`" and can be further worked out, e.g., using the R package `coda` (Plummer et al., 2006). For example, data for Table I in the paper were obtained using the following commands:

```
> library(coda)
> betaGamma <- read.table("/home/userAK/beta.sim", header=TRUE)
> exp.betaGamma <- mcmc(exp(betaGamma))
> summary(exp.betaGamma)
> HPDinterval(exp.betaGamma)
```

To compute the predictive hazard and survival functions as shown in Figure 2, we have to specify the combinations of covariates for which the hazard and survival functions would be computed:

```
> eortc.pred <- data.frame(DFStime=c(1, 1), DFSevent=c(0, 0), trtmt=c(1, 0),
+   ageGroup=factor(c(0, 0), levels=0:2, labels=c("<40", "40--50", ">50")),
+   typeSur=c(0, 0), tumSize=c(0, 0), nodStat=c(0, 0), otDis=c(0, 0),
+   region=factor(c(0, 0), levels=0:4, labels=c("F", "NL", "P", "SE", "SA")),
+   center=c(1, 2))
```

Computation of the values of predictive survival and hazard functions on the equidistant grid of 100 time values from 1 to 5 002 days is then performed using the following code:

```
> pred <- predictive2(Surv(DFStime, DFSevent)~trtmt+ageGroup+typeSur+
+   tumSize+nodStat+otDis+region+cluster(center), random=~trtmt,
+   grid=seq(1, 5002, length=100), Gspline=list(dim=1, K=15),
+   quantile=c(0.025, 0.975), only.aver=FALSE, dir="/home/userAK/",
+   predict=list(Surv=TRUE, density=FALSE, hazard=TRUE, cum.hazard=FALSE),
+   data=eortc.pred)
```

By the argument `quantile`, the user can obtain also pointwise posterior predictive quantiles for the hazard and survival function.

References

Komárek, A., Lesaffre, E., and Legrand, C. (2007). Baseline and treatment effect heterogeneity in disease free survival between centers in a random effects accelerated failure time model with flexible error distribution. *Submitted*.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). *coda: Output analysis and diagnostics for MCMC*. R package version 0.10-7.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.