Simulating from a promotion-time cure rate model

Francisco Javier Rubio

Contents

Promotion Time Cure Models

Promotion-Time Cure Models (PTCM) are a type of survival (regression) models where a portion of the population is assumed to be cured, or to never experience the time-to-event. That is, there is a sub-population with infinite survival times. The survival function of a PTCM is given by

$$S_c(t) = \exp\left\{-\theta \tilde{F}(t)\right\}, \quad t \geq 0,$$

where $\theta > 0$ is a positive parameter, and $\tilde{F}$ is a cumulative distribution function with support on $\mathbb{R}_+$. Note that

$$\lim_{t\to\infty} S_c(t) = \exp\{-\theta\},$$

which represents the cured fraction. Since the survival function associated with the PTCM does not decrease to zero, it is often referred to as an "improper' ' survival function (Peng and Yu 2021).

If a covariates $\mathbf{x} \in \mathbb{R}^p$ are available, these can be incorporated into the two different components of the PTCM.

1. Let $\mathbf{w} \subseteq \mathbf{x}$, and consider the log-link $\theta(\mathbf{w}) = \exp(\mathbf{w}^\top \boldsymbol{\alpha})$.
2. Let $\mathbf{z} \subseteq \mathbf{x}$, and consider the hazard structure $\tilde{H}(t \mid \mathbf{z})$. Then, we can define

$$\tilde{F}(t \mid \mathbf{z}) = 1 - \exp\{-\tilde{H}(t \mid \mathbf{z})\}.$$

The hazard structure can include the proportional hazards, the accelerated failure time, or more general hazard structures Rubio et al. (2019).

The following R code presents an implementation of the simulation of survival times from a PTCM with proportional hazards, accelerated failure time, and general hazard structures coupled with a log-link for θ.

Proportional Hazards

Suppose that the hazard and cumulative hazard follow a proportional hazards (PH) structure:

$$\tilde{h}(t \mid \mathbf{z}) = \tilde{h}_0(t \mid \boldsymbol{\gamma}) \exp\{\mathbf{z}^\top \boldsymbol{\beta}\}, \tag{1}$$
$$\tilde{H}(t \mid \mathbf{z}) = \tilde{H}_0(t \mid \boldsymbol{\gamma}) \exp\{\mathbf{z}^\top \boldsymbol{\beta}\}, \tag{2}$$

where $\tilde{h}_0$ is a baseline hazard associated with a distribution $\tilde{F}_0$, with parameters $\boldsymbol{\gamma}$. Let also $\theta(\mathbf{w}) = \exp(\mathbf{w}^\top \boldsymbol{\alpha})$, and u be a realisation from a $U(0,1)$ distribution. Then, a simulation from the PTCM model with PH structure (PTCM-PH) can be obtained as:

$$t^* = \tilde{F}_0^{-1}\left(1 - \exp\left\{\exp\left\{-\mathbf{z}^\top \boldsymbol{\beta}\right\} \log\left[1 + \log(u) \exp\left\{-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}\right\}\right]\right\}\right).$$

if $1 + \log(u) \exp\left\{-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}\right\} > 0$, and $t^* = \infty$ otherwise.

Accelerated Failure Time

Suppose that the hazard and cumulative hazard follow an accelerated failure time (AFT) hazard structure:

$$\tilde{h}(t \mid \mathbf{z}) = \tilde{h}_0(t \exp\{\mathbf{z}^\top \boldsymbol{\beta}\} \mid \boldsymbol{\gamma}) \exp\{\mathbf{z}^\top \boldsymbol{\beta}\}, \tag{3}$$
$$\tilde{H}(t \mid \mathbf{z}) = \tilde{H}_0(t \exp\{\mathbf{z}^\top \boldsymbol{\beta}\} \mid \boldsymbol{\gamma}), \tag{4}$$

where $\tilde{h}_0$ is a baseline hazard associated with a distribution $\tilde{F}_0$, with parameters $\boldsymbol{\gamma}$. Let also $\theta(\mathbf{w}) = \exp(\mathbf{w}^\top \boldsymbol{\alpha})$, and u be a realisation from a $U(0,1)$ distribution. Then, a simulation from the PTCM model with PH structure (PTCM-AFT) can be obtained as:

$$t^* = \tilde{F}_0^{-1}\left(-\log(u) \exp\left\{-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}\right\}\right) \exp\left\{-\mathbf{z}^\top \boldsymbol{\beta}\right\}.$$

if $\log(u) \exp\left\{-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}\right\} < 1$, and $t^* = \infty$ otherwise.

General Hazards

Suppose that the hazard and cumulative hazard follow a General Hazard (GH) structure (Chen and Jewell 2001) (Rubio et al. 2019):

$$\tilde{h}(t \mid \mathbf{z}) = \tilde{h}_0(t \exp\{\mathbf{z}^\top \boldsymbol{\eta}\} \mid \boldsymbol{\gamma}) \exp\{\mathbf{z}^\top \boldsymbol{\beta}\}, \tag{5}$$
$$\tilde{H}(t \mid \mathbf{z}) = \tilde{H}_0(t \exp\{\mathbf{z}^\top \boldsymbol{\eta}\} \mid \boldsymbol{\gamma}) \exp\{\mathbf{z}^\top \boldsymbol{\beta} - \mathbf{z}^\top \boldsymbol{\eta}\}\}, \tag{6}$$

where $\tilde{h}_0$ is a baseline hazard associated with a distribution $\tilde{F}_0$, with parameters $\boldsymbol{\gamma}$. Let also $\theta(\mathbf{w}) = \exp(\mathbf{w}^\top \boldsymbol{\alpha})$, and u be a realisation from a $U(0,1)$ distribution. Then, a simulation from the PTCM model with PH structure (PTCM-AFT) can be obtained as:

$$t^* = \exp\left\{-\mathbf{z}^\top \boldsymbol{\eta}\right\} \tilde{F}_0^{-1}\left(1 - \exp\left\{\exp\left\{\mathbf{z}^\top \boldsymbol{\eta} - \mathbf{z}^\top \boldsymbol{\beta}\right\} \log\left[1 + \log(u) \exp\left\{-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}\right\}\right]\right\}\right).$$

if $1 - \exp\left\{\exp\left\{\mathbf{z}^\top \boldsymbol{\eta} - \mathbf{z}^\top \boldsymbol{\beta}\right\} \log\left[1 + \log(u) \exp\left\{-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}\right\}\right]\right\} < 1$, and $t^* = \infty$ otherwise.

A general simulation strategy

Provided that one can simulate from the distribution $\tilde{F}(t \mid \mathbf{z})$ using numerical methods, simulating from a PTCM can be done simply by simulating $u \sim U(0,1)$, and solving

$$t^* = \tilde{F}\left(-\log(u) \exp\left\{-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}\right\} \mid \mathbf{z}\right)$$

if $-\log(u) \exp\left\{-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}\right\} < 1$, and $t^* = \infty$ otherwise.

Maximum Likelihood Estimation

No covariates

The log-likelihood function of the parameters $\Psi = (\theta, \boldsymbol{\gamma}^\top)^\top$

$$\begin{aligned}
\ell(\Psi) &= \sum_{i=1}^{n} \delta_i \log h(t_i \mid \Psi) - \sum_{i=1}^{n} H(t_i \mid \Psi) && (7)\\
&= \sum_{i=1}^{n} \delta_i \log(\theta) + \sum_{i=1}^{n} \delta_i \log \tilde{h}(t_i \mid \boldsymbol{\gamma}) - \sum_{i=1}^{n} \delta_i \tilde{H}(t_i \mid \boldsymbol{\gamma}) && (8)\\
&+ \sum_{i=1}^{n} \theta \left[1 - \exp\left\{-\tilde{H}(t_i \mid \boldsymbol{\gamma})\right\}\right] && (9)\\
&= n_O \log(\theta) + \sum_{i=1}^{n} \delta_i \log \tilde{h}(t_i \mid \boldsymbol{\gamma}) - \sum_{i=1}^{n} \delta_i \tilde{H}(t_i \mid \boldsymbol{\gamma}) && (10)\\
&- n\theta - \theta \sum_{i=1}^{n} \exp\left\{-\tilde{H}(t_i \mid \boldsymbol{\gamma})\right\}. && (11)
\end{aligned}$$

Covariates

The log-likelihood function of the parameters $\Psi = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\eta}^\top, \boldsymbol{\gamma}^\top)^\top$

$$\begin{aligned}
\ell(\Psi) &= \sum_{i=1}^{n} \delta_i \log h(t_i \mid \Psi, \boldsymbol{x}) - \sum_{i=1}^{n} H(t_i \mid \Psi, \boldsymbol{x}) && (12)\\
&= \sum_{i=1}^{n} \delta_i \boldsymbol{w}_i^\top \boldsymbol{\alpha} + \sum_{i=1}^{n} \delta_i \log \tilde{h}(t_i \mid \boldsymbol{\gamma}) - \sum_{i=1}^{n} \delta_i \tilde{H}(t_i \mid \boldsymbol{\gamma}) && (13)\\
&- \sum_{i=1}^{n} \exp\{\boldsymbol{w}_i^\top \boldsymbol{\alpha}\} + \sum_{i=1}^{n} \exp\left\{\boldsymbol{w}_i^\top \boldsymbol{\alpha} - \tilde{H}(t_i \mid \boldsymbol{\gamma})\right\}. && (14)
\end{aligned}$$

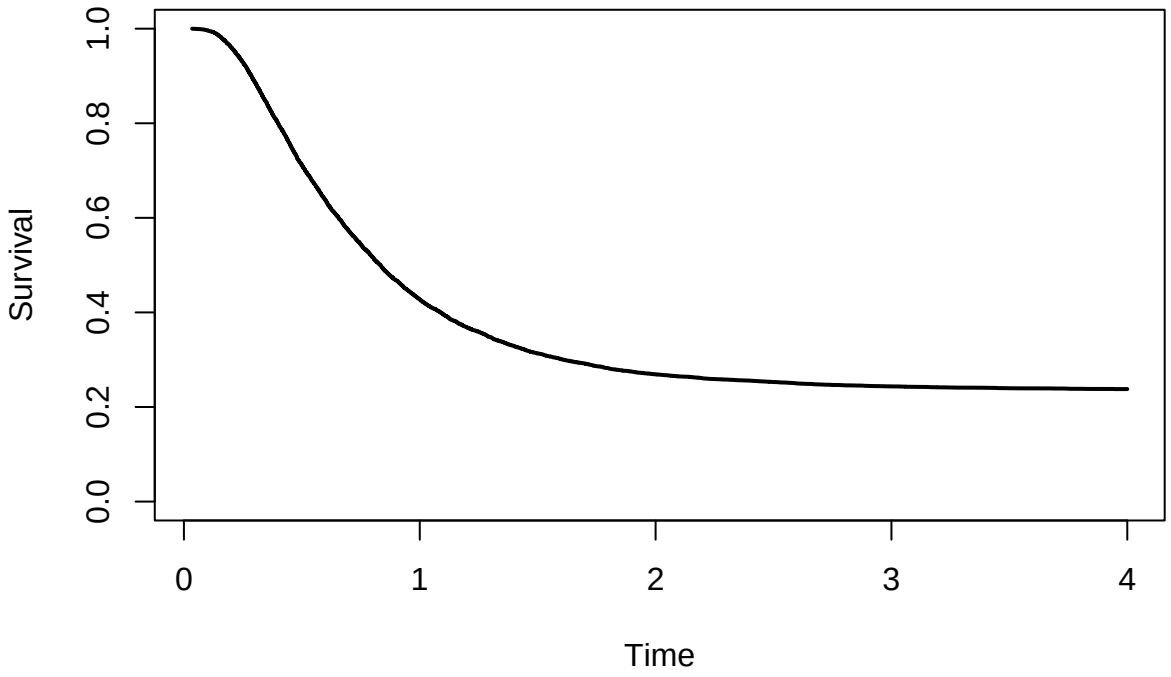
R code

Simulation

```
rm(list=ls())

#library(devtools)
#install_github("FJRubio67/HazReg")
library(HazReg)
#library(devtools)
#install_github("FJRubio67/PTCMGH")
library(PTCMGH)

n = 10000
seed = 123
set.seed(seed)
des0 <- cbind(1, rnorm(n), rnorm(n))


sim = simPTCMGH(n = n,
          seed = seed,
          hstr = "AFT",
          dist = "LogNormal",
          des_theta = des0,
          des_t = NULL,
          des_h = NULL,
          des = des0[, -1],
          par_base = c(-0.25, 0.1),
          alpha = c(0.5, 0.5, 0.5),
          beta_t = NULL,
          beta_h = NULL,
          beta = c(0.5, 0.5))


cens = 4
status <- ifelse(sim < cens, 1, 0)
mean(status)
```

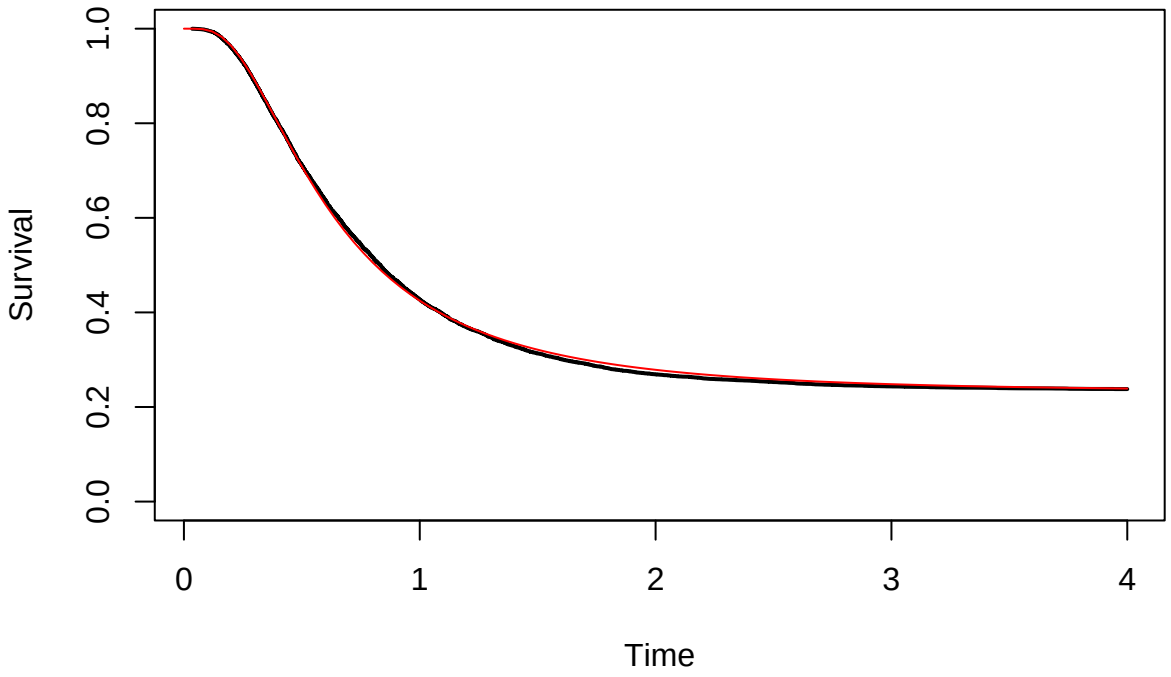
```
## [1] 0.7621
```

```
times <- ifelse(sim< cens, sim, cens)

library(survival)
# Kaplan-Meier estimator for the survival times
km <- survfit(Surv(times, status) ~ 1)

plot(km$time, km$surv, type = "l", col = "black", lwd = 2, lty = 1,
     ylim = c(0,1), xlab = "Time", ylab = "Survival")
```



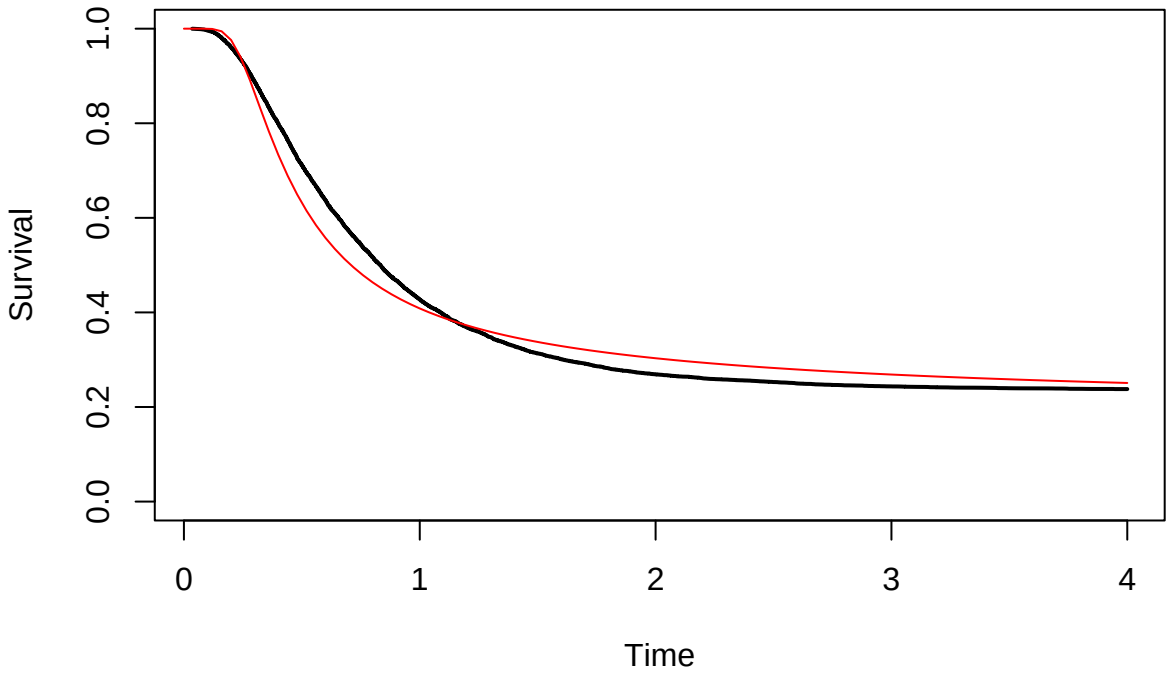
Model fit: baseline

```
OPT <- PTCMMLE(init = c(0,0,0),
                    times = times,
                    status = status,
                    hstr = "baseline",
                    dist = "LogNormal",
                    des_theta = NULL,
                    des_t = NULL,
                    des_h = NULL,
                    des = NULL,
                    method = "nlminb",
                    maxit = 10000)


MLE <- c(OPT$OPT$par[1],exp(OPT$OPT$par[2]),exp(OPT$OPT$par[3]))

spt <- Vectorize(function(t) exp(- MLE[3]*(1-exp(-chlnorm(t,MLE[1],MLE[2])))) )

plot(km$time, km$surv, type = "l", col = "black", lwd = 2, lty = 1,
     ylim = c(0,1), xlab = "Time", ylab = "Survival")
curve(spt,0,4, col = "red", add= T)
```



Model fit: PH

```
OPT_PH <- PTCMMLE(init = c(0,0,0,0,0,0,0),
                  times = times,
                  status = status,
                  hstr = "PH",
                  dist = "LogNormal",
                  des_theta = des0,
                  des_t = NULL,
                  des_h = NULL,
                  des = des0[,-1],
                  method = "nlminb",
                  maxit = 10000)


MLE_PH <- c(OPT_PH$OPT$par[1],exp(OPT_PH$OPT$par[2]),OPT_PH$OPT$par[-c(1:2)])

spt_ph <- Vectorize(function(t){

  theta_i <- as.vector(exp(des0 %*% MLE_PH[3:5]))
  F_i <- 1 - exp(-chlnorm(t,MLE_PH[1],MLE_PH[2])*exp(des0[,-1]%*%MLE_PH[6:7]))
  survs <- exp(-theta_i*F_i)
  return(mean(survs))
  })

plot(km$time, km$surv, type = "l", col = "black", lwd = 2, lty = 1,
     ylim = c(0,1), xlab = "Time", ylab = "Survival")
curve(spt_ph,0,4, col = "red", add= T)
```



Model fit: AFT

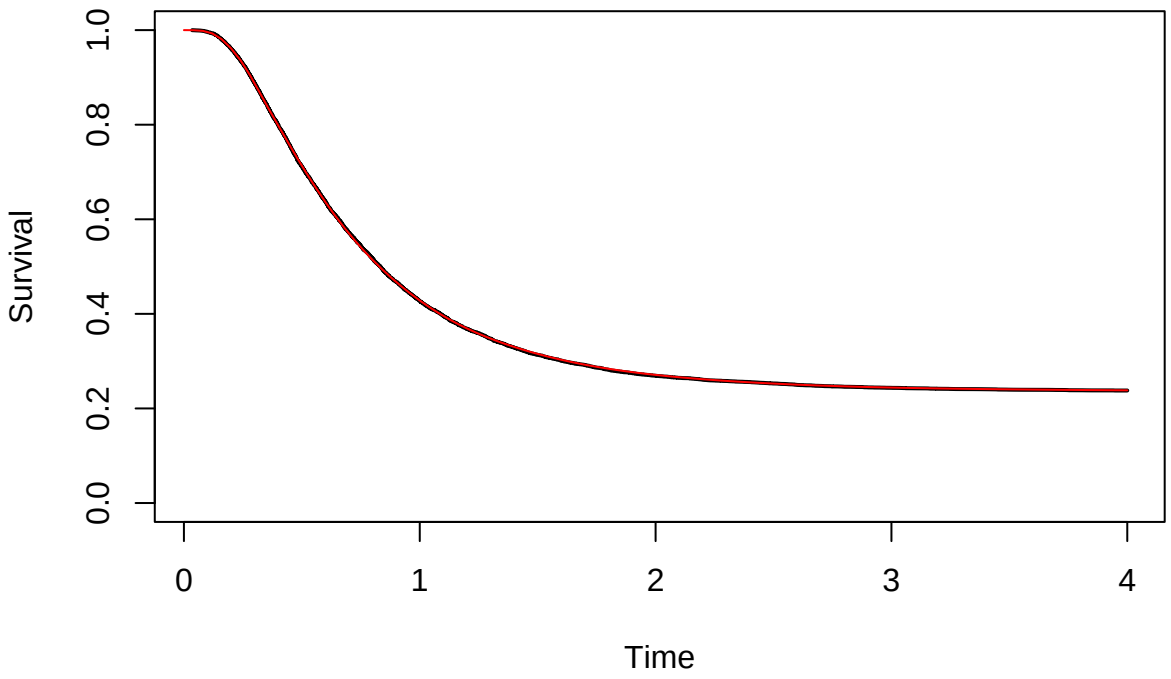
```
OPT_AFT <- PTCMMLE(init = c(0,0,0,0,0,0,0),
                   times = times,
                   status = status,
                   hstr = "AFT",
                   dist = "LogNormal",
                   des_theta = des0,
                   des_t = NULL,
                   des_h = NULL,
                   des = des0[,-1],
                   method = "nlminb",
                   maxit = 10000)


MLE_AFT <- c(OPT_AFT$OPT$par[1],exp(OPT_AFT$OPT$par[2]),OPT_AFT$OPT$par[-c(1:2)])

cbind(MLE_AFT, c(-0.25, 0.1,0.5, 0.5, 0.5,0.5, 0.5))
```

```
##           MLE_AFT
## [1,] -0.24945988 -0.25
## [2,]  0.09896654  0.10
## [3,]  0.49402475  0.50
## [4,]  0.50035275  0.50
## [5,]  0.49520539  0.50
## [6,]  0.50125675  0.50
## [7,]  0.50087607  0.50
```

```
spt_aft <- Vectorize(function(t){

  theta_i <- as.vector(exp(des0 %*% MLE_AFT[3:5]))
  F_i <- 1 - exp(-chlnorm(as.vector(t*exp(des0[,-1]%*%MLE_AFT[6:7])),MLE_AFT[1],MLE_AFT[2]))
```

```
  survs <- exp(-theta_i*F_i)
  return(mean(survs))
})
```

```
plot(km$time, km$surv, type = "l", col = "black", lwd = 2, lty = 1,
     ylim = c(0,1), xlab = "Time", ylab = "Survival")
curve(spt_aft,0,4, col = "red", add= T)
```



Model fit: Accelerated Hazards

```
OPT_AH <- PTCMMLE(init = c(0,0,0,0,0,0,0),
                  times = times,
                  status = status,
                  hstr = "AH",
                  dist = "LogNormal",
                  des_theta = des0,
                  des_t = des0[,-1],
                  des_h = NULL,
                  des = NULL,
                  method = "nlminb",
                  maxit = 10000)
```

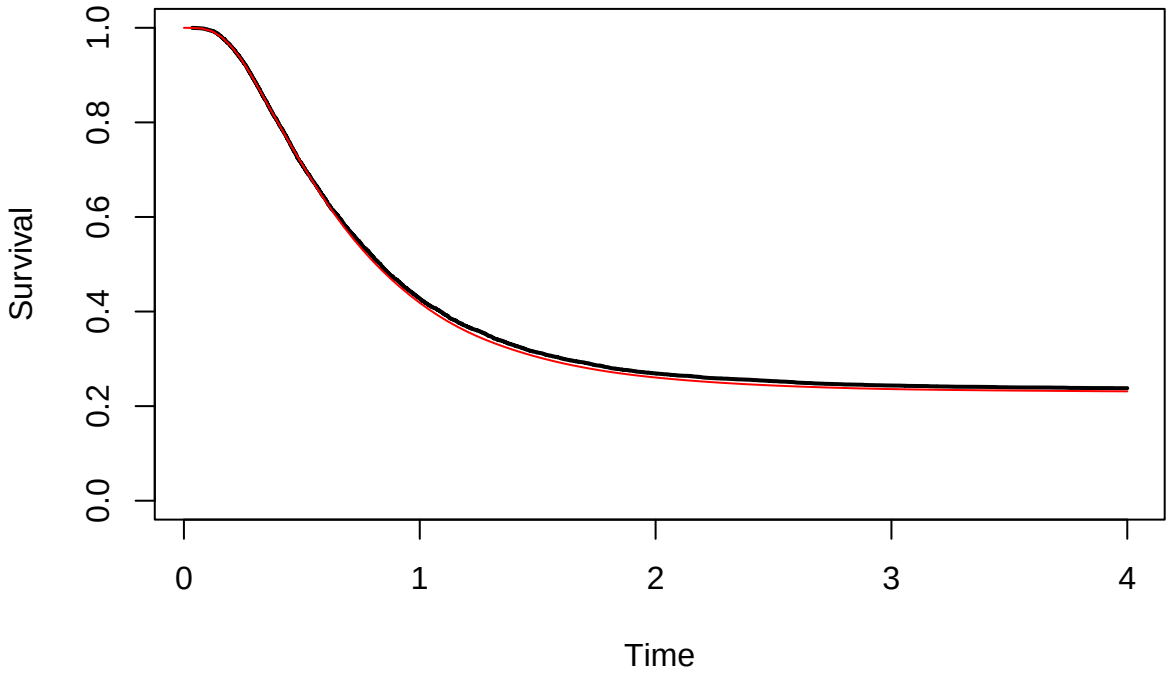
```
MLE_AH <- c(OPT_AH$OPT$par[1],exp(OPT_AH$OPT$par[2]),OPT_AH$OPT$par[-c(1:2)])

cbind(MLE_AH, c(-0.25, 0.1,0.5, 0.5, 0.5,0.5, 0.5))
```

```
##          MLE_AH
```

```
## [1,] -0.2413794 -0.25
## [2,]  0.1003200  0.10
## [3,]  0.5598464  0.50
## [4,]  0.5713208  0.50
## [5,]  0.5721592  0.50
## [6,]  0.5336055  0.50
## [7,]  0.5317930  0.50
```

```r
spt_ah <- Vectorize(function(t){

  theta_i <- as.vector(exp(des0 %*% MLE_AH[3:5]))
  F_i <- 1 - exp(-chlnorm(as.vector(t*exp(des0[,-1]%*%MLE_AH[6:7])),MLE_AH[1],MLE_AH[2])*as.vector(exp(
  survs <- exp(-theta_i*F_i)
  return(mean(survs))
})
```

```r
plot(km$time, km$surv, type = "l", col = "black", lwd = 2, lty = 1,
     ylim = c(0,1), xlab = "Time", ylab = "Survival")
curve(spt_ah,0,4, col = "red", add= T)
```



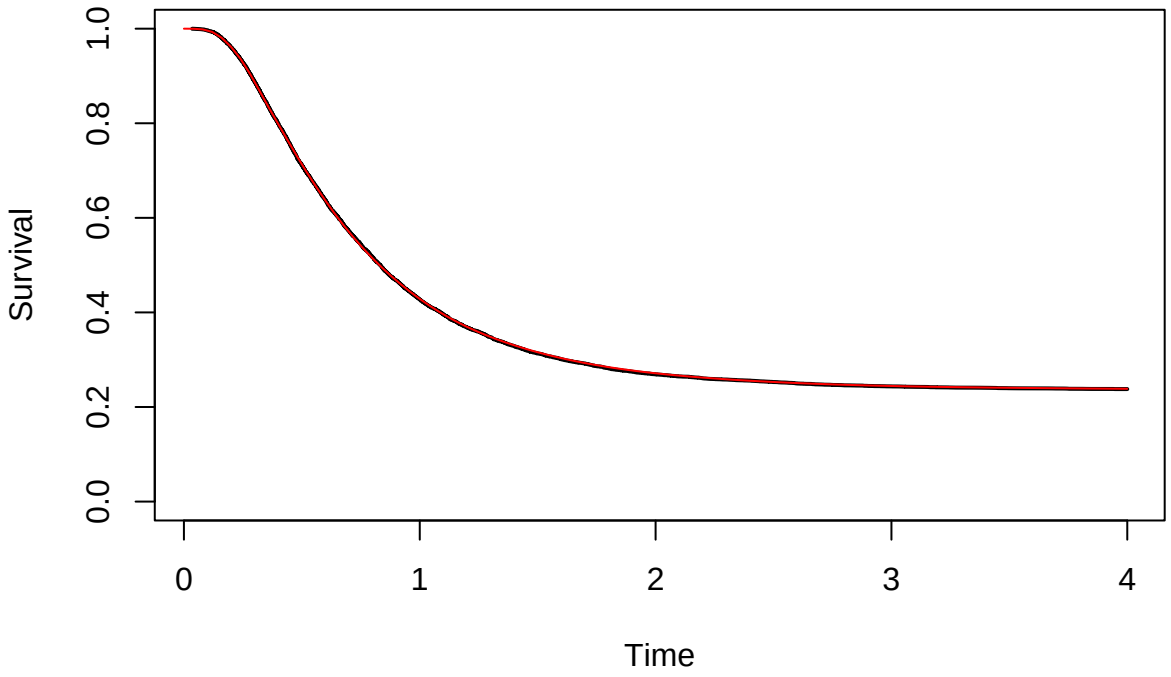
Model fit: General Hazards

```r
OPT_GH <- PTCMMLE(init = c(0,0,0,0,0,0,0,0,0),
                  times = times,
                  status = status,
                  hstr = "GH",
                  dist = "LogNormal",
                  des_theta = des0,
                  des_t = des0[,-1],
```

```
                  des_h = des0[,-1],
                  des = NULL,
                  method = "nlminb",
                  maxit = 10000)


MLE_GH <- c(OPT_GH$OPT$par[1],exp(OPT_GH$OPT$par[2]),OPT_GH$OPT$par[-c(1:2)])

cbind(MLE_GH, c(-0.25, 0.1,0.5, 0.5, 0.5,0.5, 0.5, 0.5, 0.5))
```

```
##            MLE_GH
##  [1,] -0.24974331 -0.25
##  [2,]  0.09896675  0.10
##  [3,]  0.49207223  0.50
##  [4,]  0.49927476  0.50
##  [5,]  0.49122992  0.50
##  [6,]  0.50182697  0.50
##  [7,]  0.49799340  0.50
##  [8,]  0.49609252  0.50
##  [9,]  0.54552263  0.50
```

```
spt_gh <- Vectorize(function(t){

  theta_i <- as.vector(exp(des0 %*% MLE_GH[3:5]))
  F_i <- 1 - exp(-chlnorm(as.vector(t*as.vector(exp(des0[,-1]%*%MLE_GH[6:7]))),MLE_GH[1],MLE_GH[2])*as.
  survs <- exp(-theta_i*F_i)
  return(mean(survs))
})




plot(km$time, km$surv, type = "l", col = "black", lwd = 2, lty = 1,
     ylim = c(0,1), xlab = "Time", ylab = "Survival")
curve(spt_gh,0,4, col = "red", add= T)
```



Model comparison

```
AIC <- 2*OPT$OPT$objective + 2*length(OPT$OPT$par)
AIC_PH <- 2*OPT_PH$OPT$objective + 2*length(OPT_PH$OPT$par)
AIC_AFT <- 2*OPT_AFT$OPT$objective + 2*length(OPT_AFT$OPT$par)
AIC_AH <- 2*OPT_AH$OPT$objective + 2*length(OPT_AH$OPT$par)
AIC_GH <- 2*OPT_GH$OPT$objective + 2*length(OPT_GH$OPT$par)

c(AIC, AIC_PH, AIC_AFT, AIC_AH, AIC_GH)
```

```
## [1]  19573.199   2581.728 -13723.042 -13300.643 -13721.128
```

References

Chen, Y. Q., and N. P. Jewell. 2001. "On a General Class of Semiparametric Hazards Regression Models." *Biometrika* 88 (3): 687–702.

Peng, Y., and B. Yu. 2021. *Cure Models: Methods, Applications, and Implementation*. CRC Press.

Rubio, F. J., L. Remontet, N. P. Jewell, and A. Belot. 2019. "On a General Structure for Hazard-Based Regression Models: An Application to Population-Based Cancer Research." *Statistical Methods in Medical Research* 28 (8): 2404–17.