Carleton University
School of Mathematics and Statistics

# Technical Report of a Survival Analysis Using Proportional Hazards Models

Hua Ye

December 2009

In this paper: 1) run a complete data analysis using an assumed Proportional Hazards model; 2) select the appropriate parameters for the PH model; 3) assess the fit and perform inference. Software for analyzing is R.

# Contents

**Purpose**

1) Run a complete data analysis using an assumed Proportional Hazards (PH) model;
2) Select the appropriate parameters for the PH model;
3) Assess the fit and perform inference.

**Data description**

The data consists of 863 kidney transplant patients.

We consider the data set from a study designed to assess the effect of some factors on the survival time of patients who took kidney transplant. The *TIME* variable contains survival time or on-study time in days after a kidney transplant. The variable *STATUS* has a value of 1 (dead) for those events at time, and has a value of 0 (alive) for those right censored.

The covariates included in the analyses are:

(i) *gender*: 1 = male, 2 = female;

(ii) *race*: 1 = white, 2 = black;

(iii) *age*: age in years.

We assume a PH model beforehand and let *gamma = 1269* for the baseline hazard rate $\lambda_0(t)$.

## 1 Import data into R

Before starting the data analysis, we need to load the *survival* library in R. We can do this by running *library(survival)*.

The data is stored in the text file *stat_5603-a3d1.txt*. To import the data set, use:

```
> a<-read.table("m:/stat_5603-a3d1.txt", header=T)
```

The *header=T* option tells R that the variable names are stored in the first row of the data set.

## 2 Evaluate data using Kaplan–Meier survival curves

Data can be evaluated based upon empirical Kaplan–Meier (K–M) survival curves.

## 2.1 General K–M survival curve

First, the general K–M survival curve can be obtained in R:

```
> fit.km=survfit(Surv(time,status)~1,data=a,conf)
> plot(fit.km, ylab='K-M estimate', xlab='Time (days)',
cex.main=0.8,cex.lab=0.7, cex.axis=0.7)
```

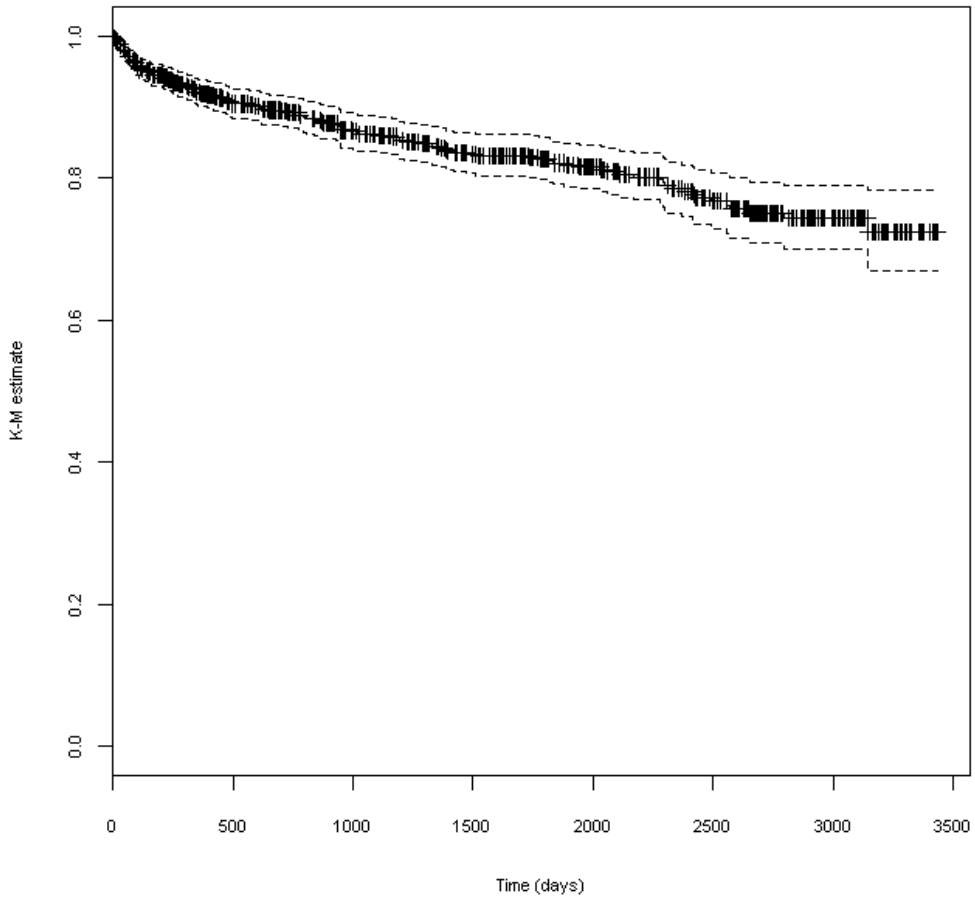


**Figure 1 General K–M survival curve with a 95% confidence interval**

## 2.2 Categorize the data in terms of gender

We categorize gender (the variable is *gender*: *1 = male, 2 = female*) to depict K–M survival curves:

```
> fit.km.byg=survfit(Surv(time,status)~gender,data=a,conf.type="none")
> plot(fit.km.byg, ylab='K-M estimate',xlab='Time
(days)',cex.main=0.8,cex.lab=0.7,cex.axis=0.7,col=c("black","red"),lty=1:2)
> legend(230, 0.1,lty=1:2, cex=.7,col=c("black","red"),c("male","female"))
```

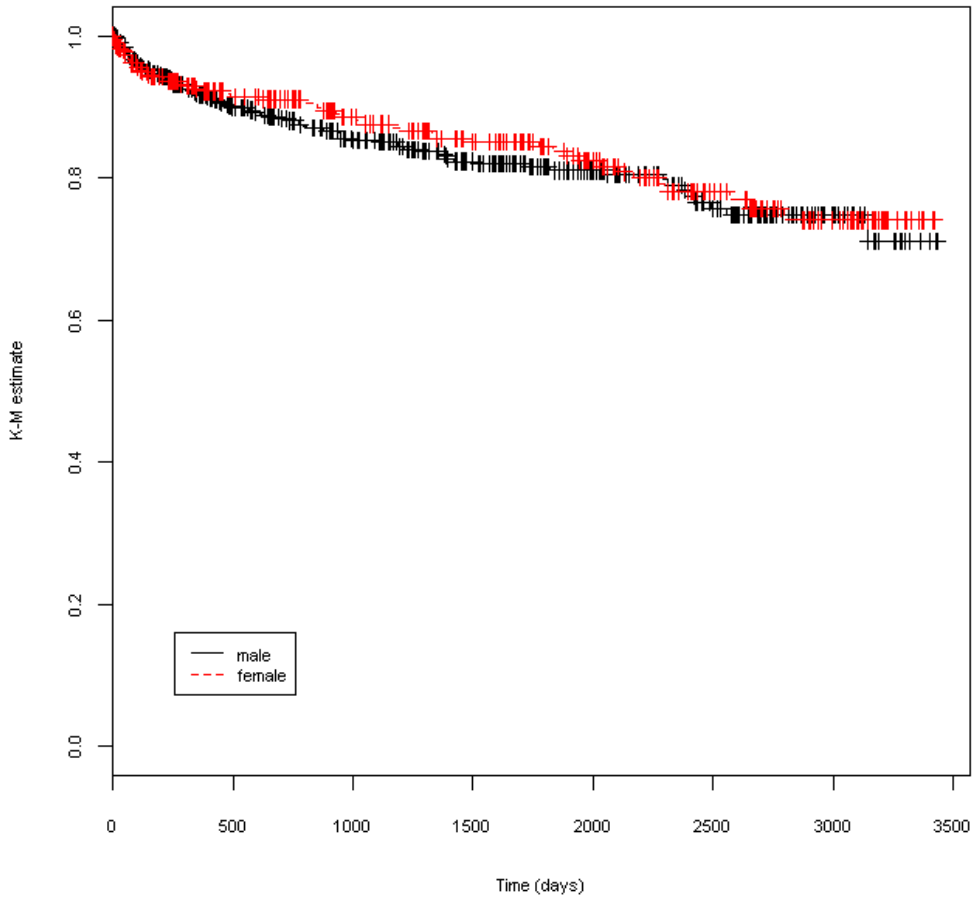


**Figure 2 K-M survival curves by gender**

## 2.3 Categorize the data in terms of race

We categorize race (the variable is *race*: *1 = white*, *2 = black*) to depict K–M survival curves:

```
> fit.km.byr=survfit(Surv(time,status)~race,data=a,conf.type="none")
> plot(fit.km.byr,ylab='K-M estimate',xlab='Time
(days)',cex.main=0.8,cex.lab=0.7,cex.axis=0.7,col=c(3:4),lty=3:4)
> legend(260, .16,lty=3:4, cex=.7,col=c(3:4),c("white","black"))
```

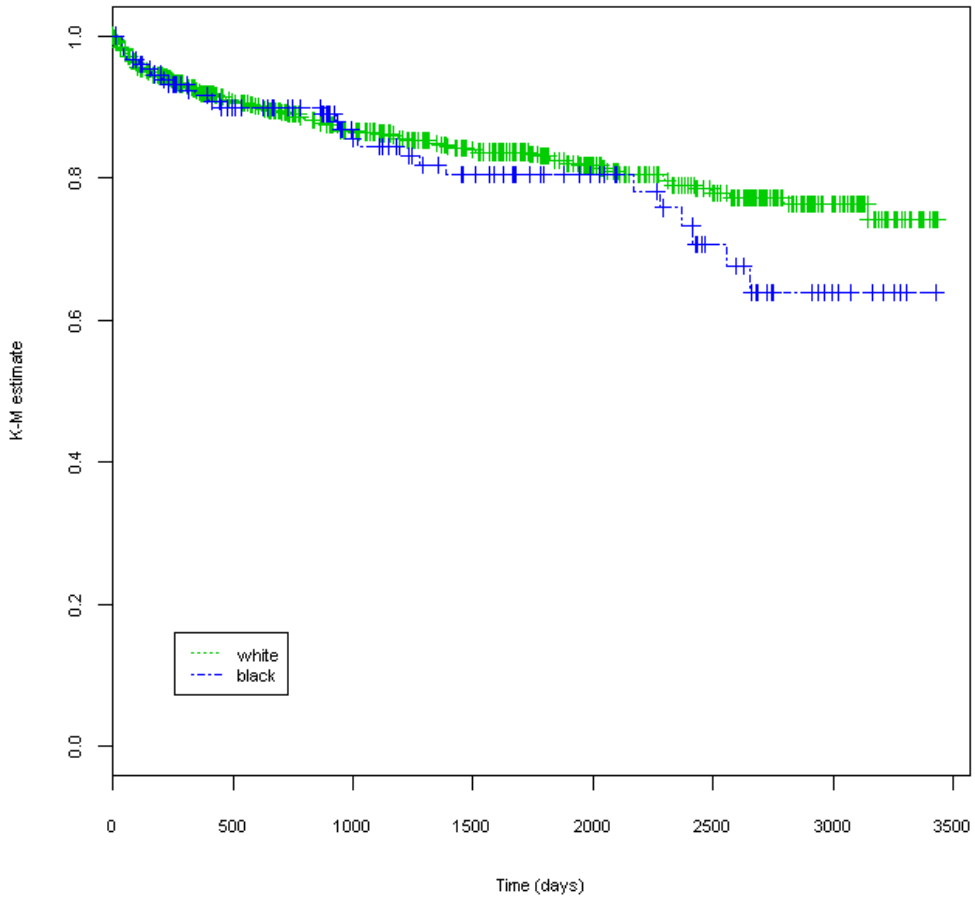


**Figure 3 K-M survival curves by race**

# 3 Approach to apply a parametric PH model

One of the interests of survival analysis is to understand the relationship between time to failure and other covariates measured at the studied subjects. In this case, we assume a PH model beforehand.

## 3.1 PH model

Let $t_i$ be a random variable denoting the failure time for the *ith* subject, and let $x_{i1}$, $x_{i2}$, …, $x_{ip}$ be the values of $p$ covariates for that same subject. In this case, we assume that $\underline{X_i}$ is independent with time. A PH model is given as

$$\lambda_i(t;\alpha,\beta,\underline{x}) = \lambda_0(t,\alpha)g(\underline{x},\beta)$$

Since the hazard function of $\lambda_i(t;\alpha,\beta,\underline{x})$ is strictly positive, consider the *log* link,

$$\log(g(\underline{x},\beta)) = \underline{x}^T\beta,$$

this means that: $g(\underline{x},\beta) = \exp\{\underline{x}^T\beta\}$.

Then, the PH model can be usually expressed as Cox Proportional Hazards Model:

$$\lambda_i(t;\alpha,\beta,\underline{x}) = \lambda_0(t,\alpha)\exp\{\underline{x}^T\beta\}.$$

Assume beforehand that this PH model has a baseline hazard rate as

$$\lambda_0(t,\alpha) = \exp\{\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t-\gamma)_+^3\}$$

Where $(t-\gamma)_+ = \max\{0, t-\gamma\}$ and $\gamma$ = 1269.

Thus, the hazard ratio for subjects with covariate vectors $\underline{x}_a$ and $\underline{x}_b$ is

$$\frac{\lambda_a(t;\alpha,\beta,\underline{x}_a)}{\lambda_b(t;\alpha,\beta,\underline{x}_b)} = \exp\{(\underline{x}_a^T - \underline{x}_b^T)\beta\}$$

### 3.2 Use R to fit the PH model

In this case, since the baseline hazard rate $\lambda_0(t,\alpha)$ is parametric with coefficient vector $\alpha$. Thus, we need two steps to obtain all estimates.

#### 3.2.1 Approach to obtain $\hat{\beta}$

The *coxph()* function in R produces estimates $\hat{\beta}$ of Cox PH Models with censored survival data using the method of maximum likelihood. The *coxph()* function uses the *Efron*[1] method for handling ties as a default. For example, in this case:

```
> phmodel<-coxph(Surv(time,status)~gender+race+age, data=a)
> phmodel
Call:
coxph(formula = Surv(time, status) ~ gender + race + age, data = a)

          coef exp(coef) se(coef)     z       p
gender 0.0265      1.03  0.17490 0.152 8.8e-01
race   0.1164      1.12  0.21151 0.550 5.8e-01
age    0.0510      1.05  0.00718 7.107 1.2e-12

Likelihood ratio test=57.1  on 3 df, p=2.50e-12  n= 863
```

[1] Efron method: this is the default method and is more intensive computationally but also more precise than the Breslow method; the Efron method is more precise because it tries to account for how the risk set changes depending on the sequence of tied events. Cox Proportional Hazards Regression for Duration Dependent Variables, http://gking.harvard.edu/zelig/docs/coxph.pdf.

The output of *coxph()* function gives the estimated coefficients and their standard errors.

- The values of *gender*, *race*, and *age* refer to the estimates of $\beta_1, \beta_2, \beta_3$.
- It also gives the results of likelihood ratio, Wald and Score tests (with *summary()*).

### 3.2.2 Approach to obtain $\hat{\alpha}$

There are four main steps to obtain $\hat{\alpha}$ **:**

(i) Generate the values of the baseline hazard rate $\lambda_0(t,\alpha)$ in terms of $\hat{\beta}$;

(ii) Get the estimate of $\log \lambda_0(t,\alpha \mid \hat{\beta})$;

(iii) Add new columns to the original event time data (*status* = 1) for expressing $t^2$, $t^3$, and $(t-\gamma)_+ = \max\{0, t-\gamma\}$ where $\gamma$ = *1269*;

(iv) Use the *lm()* function in R to fit $\log \lambda_0(t,\alpha \mid \hat{\beta})$ that can be seen as a linear model:

$$\log \lambda_0(t,\alpha \mid \hat{\beta}) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t-\gamma)_+^3 .$$

*The details are presented in Section 4.*

## 4 Fit the PH model

This section consists of two parts: (i) find the full model; (ii) obtain the reduced model.

### 4.1 Find the full model

1) Obtain $\hat{\beta}$ for the full model

```
> phmodel<-coxph(Surv(time,status)~gender+race+age, data=a)
> summary(phmodel)
Call:
coxph(formula = Surv(time, status) ~ gender + race + age, data = a)
  n= 863

           coef exp(coef) se(coef)     z Pr(>|z|)
gender 0.026540  1.026895 0.174900 0.152    0.879
race   0.116365  1.123406 0.211512 0.550    0.582
age    0.051035  1.052360 0.007181 7.107 1.19e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
gender     1.027     0.9738    0.7289     1.447
race       1.123     0.8902    0.7422     1.700
age        1.052     0.9502    1.0377     1.067
```

```
Rsquare= 0.064   (max possible= 0.87 )
Likelihood ratio test= 57.06  on 3 df,   p=2.495e-12
Wald test            = 51.27  on 3 df,   p=4.284e-11
Score (logrank)  test= 53.7   on 3 df,   p=1.303e-11
```

| Parameter | Estimate | Std.Error | z | p | |
|---|---|---|---|---|---|
| $\beta_1$ gender | 0.026540 | 0.174900 | 0.152 | 0.879 | |
| $\beta_2$ race | 0.116365 | 0.211512 | 0.550 | 0.582 | |
| $\beta_3$ age | 0.051035 | 0.007181 | 7.107 | 1.19e-12 | *** |

**Table 1 Estimates of parameter $\beta$ for the full model**

2) Obtain $\hat{\alpha}$ for the full model

(i) Get the values of the baseline hazard rate $\lambda_0(t,\alpha)$ in terms of $\hat{\beta}$

- The *coxph.detail()* function in R returns the estimate of the hazard for an average individual $\lambda_m(t,\alpha)$, i.e. with all covariates evaluated at **their average over the sample**.

```
> d.phmodel =coxph.detail(phmodel)
> hazm=c(d.phmodel$hazard)
> hazm
  [1] 0.0009291547 0.0009312235 0.0018681775 0.0018838977 0.0009468440
  [6] 0.0009527388 0.0019123916 0.0009586419 0.0009633320 0.0009644059
 [11] 0.0019334706 0.0009744288 0.0009757577 0.0009797398 0.0009855676
 [16] 0.0009880177 0.0019811358 0.0009923592 0.0019880142 0.0020037476
 [21] 0.0010038892 0.0010069269 0.0010093569 0.0010219802 0.0010233692
 [26] 0.0010286548 0.0020649718 0.0010355589 0.0020808591 0.0010611124
 [31] 0.0010627076 0.0010680423 0.0010743612 0.0010784738 0.0010849420
 [36] 0.0010867228 0.0010891214 0.0011106332 0.0011215194 0.0011260060
 [41] 0.0011339626 0.0011359565 0.0011453105 0.0011490903 0.0011505970
 [46] 0.0011593386 0.0012001569 0.0012119932 0.0012166565 0.0024534657
 [51] 0.0012466032 0.0012526122 0.0012595749 0.0012627201 0.0012678615
 [56] 0.0012712654 0.0012953902 0.0013079141 0.0013257902 0.0013450786
 [61] 0.0013664240 0.0013753896 0.0013772917 0.0013826479 0.0013910202
 [66] 0.0013995276 0.0014372829 0.0014538750 0.0014797726 0.0014827651
 [71] 0.0014855450 0.0015166750 0.0015850861 0.0016011898 0.0016520832
 [76] 0.0016537708 0.0016734089 0.0016806609 0.0016826474 0.0016993526
 [81] 0.0017019749 0.0017140706 0.0017375464 0.0018125406 0.0018213015
 [86] 0.0018270055 0.0036649955 0.0018370716 0.0018869343 0.0018937253
 [91] 0.0019067116 0.0019640231 0.0020582821 0.0020800664 0.0020831858
 [96] 0.0020876609 0.0021021937 0.0022200311 0.0022915447 0.0022983704
[101] 0.0023284515 0.0023905097 0.0024304732 0.0024387062 0.0025624816
[106] 0.0025910309 0.0029872519 0.0031886748 0.0034135914 0.0034256242
[111] 0.0034754521 0.0035766432 0.0040413054 0.0042223758 0.0045103137
[116] 0.0045890098 0.0050633396 0.0051145509 0.0052055311 0.0052912776
[121] 0.0056035296 0.0057784709 0.0058403724 0.0065449394 0.0070315084
[126] 0.0071185459 0.0080298535 0.0108984614 0.0310141762
```

- In order to evaluate the estimate of the baseline hazard rate $\lambda_0(t,\alpha)$ at $\underline{x} = \underline{0}$, we must **subtract this value from the mean** $\bar{\underline{x}}$.

  The baseline hazard rate $\lambda_0(t,\alpha)$ can be obtained as:

$$\lambda_0(t,\alpha) = \frac{\lambda_m(t,\alpha)}{\exp\{\underline{\bar{x}}^T \beta\}}$$

```
> attach(a)
> meanx=c(mean(gender),mean(race),mean(age))
> meanx
[1]  1.392816  1.174971 42.836616
> beta=phmodel$coef
> beta
    gender       race        age
0.02653956 0.11636477 0.05103510
> t(meanx)%*%beta
         [,1]
[1,] 2.359861
> ex=exp(t(meanx)%*%beta)
> ex
         [,1]
[1,] 10.58948
> haz0=hazm/ex
> haz0            # ---------- the values of baseline hazard rate λ0(t,α)
  [1] 8.774318e-05 8.793854e-05 1.764182e-04 1.779027e-04 8.941364e-05
  [6] 8.997031e-05 1.805935e-04 9.052775e-05 9.097065e-05 9.107207e-05
 [11] 1.825841e-04 9.201856e-05 9.214405e-05 9.252010e-05 9.307044e-05
 [16] 9.330181e-05 1.870853e-04 9.371178e-05 1.877348e-04 1.892206e-04
 [21] 9.480060e-05 9.508746e-05 9.531694e-05 9.650899e-05 9.664017e-05
 [26] 9.713930e-05 1.950022e-04 9.779128e-05 1.965025e-04 1.002044e-04
 [31] 1.003550e-04 1.008588e-04 1.014555e-04 1.018439e-04 1.024547e-04
 [36] 1.026229e-04 1.028494e-04 1.048808e-04 1.059088e-04 1.063325e-04
 [41] 1.070839e-04 1.072722e-04 1.081555e-04 1.085124e-04 1.086547e-04
 [46] 1.094802e-04 1.133348e-04 1.144526e-04 1.148929e-04 2.316889e-04
 [51] 1.177209e-04 1.182883e-04 1.189459e-04 1.192429e-04 1.197284e-04
 [56] 1.200498e-04 1.223280e-04 1.235107e-04 1.251988e-04 1.270203e-04
 [61] 1.290360e-04 1.298826e-04 1.300622e-04 1.305680e-04 1.313587e-04
 [66] 1.321621e-04 1.357274e-04 1.372943e-04 1.397399e-04 1.400225e-04
 [71] 1.402850e-04 1.432247e-04 1.496850e-04 1.512057e-04 1.560117e-04
 [76] 1.561711e-04 1.580256e-04 1.587104e-04 1.588980e-04 1.604755e-04
 [81] 1.607232e-04 1.618654e-04 1.640823e-04 1.711643e-04 1.719916e-04
 [86] 1.725302e-04 3.460977e-04 1.734808e-04 1.781895e-04 1.788308e-04
 [91] 1.800571e-04 1.854693e-04 1.943704e-04 1.964276e-04 1.967222e-04
 [96] 1.971448e-04 1.985172e-04 2.096449e-04 2.163982e-04 2.170428e-04
[101] 2.198834e-04 2.257438e-04 2.295177e-04 2.302952e-04 2.419837e-04
[106] 2.446797e-04 2.820962e-04 3.011172e-04 3.223568e-04 3.234931e-04
[111] 3.281985e-04 3.377543e-04 3.816339e-04 3.987330e-04 4.259240e-04
[116] 4.333555e-04 4.781480e-04 4.829841e-04 4.915757e-04 4.996730e-04
[121] 5.291600e-04 5.456803e-04 5.515258e-04 6.180605e-04 6.640088e-04
[126] 6.722280e-04 7.582858e-04 1.029178e-03 2.928772e-03
```

Thus, the values of the baseline hazard rate $\lambda_0(t,\alpha)$ in terms of $\hat{\beta}$ are obtained as the above.

(ii) Get the estimate of $\log \lambda_0(t,\alpha \mid \hat{\beta})$

```
> loghaz0=log(haz0)
> loghaz0
  [1] -9.341096 -9.338872 -8.642653 -8.634274 -9.322237 -9.316031 -8.619262
  [8] -9.309854 -9.304974 -9.303859 -8.608300 -9.293520 -9.292157 -9.288085
 [15] -9.282154 -9.279671 -8.583946 -9.275287 -8.580480 -8.572597 -9.263735
 [22] -9.260713 -9.258303 -9.245874 -9.244516 -9.239365 -8.542500 -9.232675
 [29] -8.534836 -9.208299 -9.206796 -9.201789 -9.195890 -9.192070 -9.186090
 [36] -9.184450 -9.182245 -9.162686 -9.152932 -9.148940 -9.141898 -9.140141
 [43] -9.131941 -9.128646 -9.127335 -9.119767 -9.085164 -9.075350 -9.071510
```

```
 [50] -8.370115 -9.047194 -9.042385 -9.036842 -9.034348 -9.030285 -9.027604
 [57] -9.008804 -8.999183 -8.985608 -8.971164 -8.955419 -8.948879 -8.947497
 [64] -8.943616 -8.937579 -8.931482 -8.904862 -8.893384 -8.875728 -8.873708
 [71] -8.871835 -8.851096 -8.806978 -8.796869 -8.765579 -8.764558 -8.752754
 [78] -8.748429 -8.747248 -8.737369 -8.735827 -8.728745 -8.715142 -8.672887
 [85] -8.668065 -8.664938 -7.968789 -8.659444 -8.632663 -8.629070 -8.622236
 [92] -8.592621 -8.545745 -8.535217 -8.533718 -8.531572 -8.524635 -8.470095
 [99] -8.438390 -8.435416 -8.422413 -8.396110 -8.379530 -8.376149 -8.326640
[106] -8.315561 -8.173263 -8.108011 -8.039851 -8.036333 -8.021892 -7.993192
[113] -7.871049 -7.827218 -7.761250 -7.743952 -7.645590 -7.635527 -7.617895
[120] -7.601557 -7.544220 -7.513477 -7.502822 -7.388924 -7.317215 -7.304913
[127] -7.184450 -6.878995 -5.833172
```

Then, the values of $\log \lambda_0(t, \alpha \mid \hat{\beta}) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t-\gamma)_+^3$ are obtained as above.

(iii) Add new columns to the original event time data (*status* = 1)

Add new columns to the original event time data (*status* = 1) for expressing $t^2$, $t^3$, and $(t-\gamma)_+^3$ where $(t-\gamma)_+ = \max\{0, t-\gamma\}$ and $\gamma$ = 1269.

**# Retrieve the original event time data (status = 1 and no repeat) t**

```
> haz.time=c(d.phmodel$time)
> haz.time
  [1]    2    3    7   10   17   21   26   28   37   40   43   44   45   50   52
 [16]   56   57   59   62   68   69   78   79   88   91   97   98  104  106  119
 [31]  121  135  143  150  154  158  162  190  206  209  228  229  242  248  249
 [46]  252  273  291  297  311  334  340  344  346  354  366  391  402  421  439
 [61]  450  470  478  481  490  495  570  583  614  615  621  652  697  730  773
 [76]  776  790  793  806  840  852  864  875  929  939  943  945  946 1001 1013
 [91] 1016 1105 1164 1186 1191 1196 1210 1275 1326 1331 1357 1384 1388 1418 1473
[106] 1509 1734 1777 1820 1835 1877 1940 2034 2056 2108 2171 2276 2291 2301 2313
[121] 2369 2414 2421 2489 2557 2567 2650 2795 3146
```

**# Generate $t^2$**

```
> haz.time2<-haz.time^2
> haz.time2
  [1]       4       9      49     100     289     441     676     784    1369
 [10]    1600    1849    1936    2025    2500    2704    3136    3249    3481
 [19]    3844    4624    4761    6084    6241    7744    8281    9409    9604
 [28]   10816   11236   14161   14641   18225   20449   22500   23716   24964
 [37]   26244   36100   42436   43681   51984   52441   58564   61504   62001
 [46]   63504   74529   84681   88209   96721  111556  115600  118336  119716
 [55]  125316  133956  152881  161604  177241  192721  202500  220900  228484
 [64]  231361  240100  245025  324900  339889  376996  378225  385641  425104
 [73]  485809  532900  597529  602176  624100  628849  649636  705600  725904
 [82]  746496  765625  863041  881721  889249  893025  894916 1002001 1026169
 [91] 1032256 1221025 1354896 1406596 1418481 1430416 1464100 1625625 1758276
[100] 1771561 1841449 1915456 1926544 2010724 2169729 2277081 3006756 3157729
[109] 3312400 3367225 3523129 3763600 4137156 4227136 4443664 4713241 5180176
[118] 5248681 5294601 5349969 5612161 5827396 5861241 6195121 6538249 6589489
[127] 7022500 7812025 9897316
```

**# Generate $t^3$**

```
> haz.time3<-haz.time^3
> haz.time3
  [1]           8          27         343        1000        4913        9261
  [7]       17576       21952       50653       64000       79507       85184
 [13]       91125      125000      140608      175616      185193      205379
 [19]      238328      314432      328509      474552      493039      681472
 [25]      753571      912673      941192     1124864     1191016     1685159
 [31]     1771561     2460375     2924207     3375000     3652264     3944312
```

```
 [37]     4251528     6859000     8741816     9129329    11852352    12008989
 [43]    14172488    15252992    15438249    16003008    20346417    24642171
 [49]    26198073    30080231    37259704    39304000    40707584    41421736
 [55]    44361864    49027896    59776471    64964808    74618461    84604519
 [61]    91125000   103823000   109215352   111284641   117649000   121287375
 [67]   185193000   198155287   231475544   232608375   239483061   277167808
 [73]   338608873   389017000   461889917   467288576   493039000   498677257
 [79]   523606616   592704000   618470208   644972544   669921875   801765089
 [85]   827936019   838561807   843908625   846590536  1003003001  1039509197
 [91]  1048772096  1349232625  1577098944  1668222856  1689410871  1710777536
 [97]  1771561000  2072671875  2331473976  2357947691  2498846293  2650991104
[103]  2674043072  2851206632  3196010817  3436115229  5213714904  5611284433
[109]  6028568000  6178857875  6612913133  7301384000  8414975304  8690991616
[115]  9367243712 10232446211 11790080576 12024728171 12182876901 12374478297
[121] 13295209409 14067333944 14190064461 15419656169 16718302693 16915218263
[127] 18609625000 21834609875 31136956136
```

**# Generate $(t-\gamma)_+ = \max\{0, t-\gamma\}$ where $\gamma$ = 1269**

```
> new.time<-haz.time
> for(i in 1:length(new.time)) {if (new.time[i]-1269<0) new.time[i]=0
else new.time[i]= new.time[i]- 1269}
> new.time
  [1]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [16]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [31]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [46]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [61]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [76]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [91]    0    0    0    0    0    0    0    6   57   62   88  115  119  149  204
[106]  240  465  508  551  566  608  671  765  787  839  902 1007 1022 1032 1044
[121] 1100 1145 1152 1220 1288 1298 1381 1526 1877
```

**# Generate $(t-\gamma)_+^3$**

```
> new.time3<- new.time^3
> new.time3
  [1]          0          0          0          0          0          0
  [7]          0          0          0          0          0          0
 [13]          0          0          0          0          0          0
 [19]          0          0          0          0          0          0
 [25]          0          0          0          0          0          0
 [31]          0          0          0          0          0          0
 [37]          0          0          0          0          0          0
 [43]          0          0          0          0          0          0
 [49]          0          0          0          0          0          0
 [55]          0          0          0          0          0          0
 [61]          0          0          0          0          0          0
 [67]          0          0          0          0          0          0
 [73]          0          0          0          0          0          0
 [79]          0          0          0          0          0          0
 [85]          0          0          0          0          0          0
 [91]          0          0          0          0          0          0
 [97]          0        216     185193     238328     681472    1520875
[103]    1685159    3307949    8489664   13824000  100544625  131096512
[109]  167284151  181321496  224755712  302111711  447697125  487443403
[115]  590589719  733870808 1021147343 1067462648 1099104768 1137893184
[121] 1331000000 1501123625 1528823808 1815848000 2136719872 2186875592
[127] 2633789341 3553559576 6612913133
```

**# Generate a new dataset for** $\log \lambda_0(t, \alpha \mid \hat{\beta}) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t-\gamma)_+^3$

```
> dfull<-cbind(loghaz0, haz.time, haz.time2, haz.time3, new.time3)
> dfull.a<-data.frame(dfull)
> dfull.a
      loghaz0 haz.time haz.time2   haz.time3  new.time3
1   -9.341096        2         4           8          0
2   -9.338872        3         9          27          0
3   -8.642653        7        49         343          0
4   -8.634274       10       100        1000          0
```

| | | | | | |
|---|---|---|---|---|---|
| 5 | -9.322237 | 17 | 289 | 4913 | 0 |
| 6 | -9.316031 | 21 | 441 | 9261 | 0 |
| 7 | -8.619262 | 26 | 676 | 17576 | 0 |
| 8 | -9.309854 | 28 | 784 | 21952 | 0 |
| 9 | -9.304974 | 37 | 1369 | 50653 | 0 |
| 10 | -9.303859 | 40 | 1600 | 64000 | 0 |
| 11 | -8.608300 | 43 | 1849 | 79507 | 0 |
| 12 | -9.293520 | 44 | 1936 | 85184 | 0 |
| 13 | -9.292157 | 45 | 2025 | 91125 | 0 |
| 14 | -9.288085 | 50 | 2500 | 125000 | 0 |
| 15 | -9.282154 | 52 | 2704 | 140608 | 0 |
| 16 | -9.279671 | 56 | 3136 | 175616 | 0 |
| 17 | -8.583946 | 57 | 3249 | 185193 | 0 |
| 18 | -9.275287 | 59 | 3481 | 205379 | 0 |
| 19 | -8.580480 | 62 | 3844 | 238328 | 0 |
| 20 | -8.572597 | 68 | 4624 | 314432 | 0 |
| 21 | -9.263735 | 69 | 4761 | 328509 | 0 |
| 22 | -9.260713 | 78 | 6084 | 474552 | 0 |
| 23 | -9.258303 | 79 | 6241 | 493039 | 0 |
| 24 | -9.245874 | 88 | 7744 | 681472 | 0 |
| 25 | -9.244516 | 91 | 8281 | 753571 | 0 |
| 26 | -9.239365 | 97 | 9409 | 912673 | 0 |
| 27 | -8.542500 | 98 | 9604 | 941192 | 0 |
| 28 | -9.232675 | 104 | 10816 | 1124864 | 0 |
| 29 | -8.534836 | 106 | 11236 | 1191016 | 0 |
| 30 | -9.208299 | 119 | 14161 | 1685159 | 0 |
| 31 | -9.206796 | 121 | 14641 | 1771561 | 0 |
| 32 | -9.201789 | 135 | 18225 | 2460375 | 0 |
| 33 | -9.195890 | 143 | 20449 | 2924207 | 0 |
| 34 | -9.192070 | 150 | 22500 | 3375000 | 0 |
| 35 | -9.186090 | 154 | 23716 | 3652264 | 0 |
| 36 | -9.184450 | 158 | 24964 | 3944312 | 0 |
| 37 | -9.182245 | 162 | 26244 | 4251528 | 0 |
| 38 | -9.162686 | 190 | 36100 | 6859000 | 0 |
| 39 | -9.152932 | 206 | 42436 | 8741816 | 0 |
| 40 | -9.148940 | 209 | 43681 | 9129329 | 0 |
| 41 | -9.141898 | 228 | 51984 | 11852352 | 0 |
| 42 | -9.140141 | 229 | 52441 | 12008989 | 0 |
| 43 | -9.131941 | 242 | 58564 | 14172488 | 0 |
| 44 | -9.128646 | 248 | 61504 | 15252992 | 0 |
| 45 | -9.127335 | 249 | 62001 | 15438249 | 0 |
| 46 | -9.119767 | 252 | 63504 | 16003008 | 0 |
| 47 | -9.085164 | 273 | 74529 | 20346417 | 0 |
| 48 | -9.075350 | 291 | 84681 | 24642171 | 0 |
| 49 | -9.071510 | 297 | 88209 | 26198073 | 0 |
| 50 | -8.370115 | 311 | 96721 | 30080231 | 0 |
| 51 | -9.047194 | 334 | 111556 | 37259704 | 0 |
| 52 | -9.042385 | 340 | 115600 | 39304000 | 0 |
| 53 | -9.036842 | 344 | 118336 | 40707584 | 0 |
| 54 | -9.034348 | 346 | 119716 | 41421736 | 0 |
| 55 | -9.030285 | 354 | 125316 | 44361864 | 0 |
| 56 | -9.027604 | 366 | 133956 | 49027896 | 0 |
| 57 | -9.008804 | 391 | 152881 | 59776471 | 0 |
| 58 | -8.999183 | 402 | 161604 | 64964808 | 0 |
| 59 | -8.985608 | 421 | 177241 | 74618461 | 0 |
| 60 | -8.971164 | 439 | 192721 | 84604519 | 0 |
| 61 | -8.955419 | 450 | 202500 | 91125000 | 0 |
| 62 | -8.948879 | 470 | 220900 | 103823000 | 0 |
| 63 | -8.947497 | 478 | 228484 | 109215352 | 0 |
| 64 | -8.943616 | 481 | 231361 | 111284641 | 0 |
| 65 | -8.937579 | 490 | 240100 | 117649000 | 0 |
| 66 | -8.931482 | 495 | 245025 | 121287375 | 0 |
| 67 | -8.904862 | 570 | 324900 | 185193000 | 0 |
| 68 | -8.893384 | 583 | 339889 | 198155287 | 0 |
| 69 | -8.875728 | 614 | 376996 | 231475544 | 0 |
| 70 | -8.873708 | 615 | 378225 | 232608375 | 0 |
| 71 | -8.871835 | 621 | 385641 | 239483061 | 0 |
| 72 | -8.851096 | 652 | 425104 | 277167808 | 0 |
| 73 | -8.806978 | 697 | 485809 | 338608873 | 0 |
| 74 | -8.796869 | 730 | 532900 | 389017000 | 0 |
| 75 | -8.765579 | 773 | 597529 | 461889917 | 0 |

```
76  -8.764558      776    602176   467288576          0
77  -8.752754      790    624100   493039000          0
78  -8.748429      793    628849   498677257          0
79  -8.747248      806    649636   523606616          0
80  -8.737369      840    705600   592704000          0
81  -8.735827      852    725904   618470208          0
82  -8.728745      864    746496   644972544          0
83  -8.715142      875    765625   669921875          0
84  -8.672887      929    863041   801765089          0
85  -8.668065      939    881721   827936019          0
86  -8.664938      943    889249   838561807          0
87  -7.968789      945    893025   843908625          0
88  -8.659444      946    894916   846590536          0
89  -8.632663     1001   1002001  1003003001          0
90  -8.629070     1013   1026169  1039509197          0
91  -8.622236     1016   1032256  1048772096          0
92  -8.592621     1105   1221025  1349232625          0
93  -8.545745     1164   1354896  1577098944          0
94  -8.535217     1186   1406596  1668222856          0
95  -8.533718     1191   1418481  1689410871          0
96  -8.531572     1196   1430416  1710777536          0
97  -8.524635     1210   1464100  1771561000          0
98  -8.470095     1275   1625625  2072671875        216
99  -8.438390     1326   1758276  2331473976     185193
100 -8.435416     1331   1771561  2357947691     238328
101 -8.422413     1357   1841449  2498846293     681472
102 -8.396110     1384   1915456  2650991104    1520875
103 -8.379530     1388   1926544  2674043072    1685159
104 -8.376149     1418   2010724  2851206632    3307949
105 -8.326640     1473   2169729  3196010817    8489664
106 -8.315561     1509   2277081  3436115229   13824000
107 -8.173263     1734   3006756  5213714904  100544625
108 -8.108011     1777   3157729  5611284433  131096512
109 -8.039851     1820   3312400  6028568000  167284151
110 -8.036333     1835   3367225  6178857875  181321496
111 -8.021892     1877   3523129  6612913133  224755712
112 -7.993192     1940   3763600  7301384000  302111711
113 -7.871049     2034   4137156  8414975304  447697125
114 -7.827218     2056   4227136  8690991616  487443403
115 -7.761250     2108   4443664  9367243712  590589719
116 -7.743952     2171   4713241 10232446211  733870808
117 -7.645590     2276   5180176 11790080576 1021147343
118 -7.635527     2291   5248681 12024728171 1067462648
119 -7.617895     2301   5294601 12182876901 1099104768
120 -7.601557     2313   5349969 12374478297 1137893184
121 -7.544220     2369   5612161 13295209409 1331000000
122 -7.513477     2414   5827396 14067333944 1501123625
123 -7.502822     2421   5861241 14190064461 1528823808
124 -7.388924     2489   6195121 15419656169 1815848000
125 -7.317215     2557   6538249 16718302693 2136719872
126 -7.304913     2567   6589489 16915218263 2186875592
127 -7.184450     2650   7022500 18609625000 2633789341
128 -6.878995     2795   7812025 21834609875 3553559576
129 -5.833172     3146   9897316 31136956136 6612913133
```

```
# Export and save the new dataset
> write.table(dfull.a, "M:/dfull.a.txt", sep=" ", col.names=TRUE,
row.names=FALSE, quote=FALSE, na="NA")
```

(iv) Use the *lm()* function in R to fit $\log \lambda_0(t, \alpha \mid \hat{\beta})$ that can be seen as a linear model

$$\log \lambda_0(t, \alpha \mid \hat{\beta}) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t - \gamma)_+^3 .$$

```
> linear1 <- lm(loghaz0 ~ haz.time + haz.time2 + haz.time3 + new.time3,
data=dfull.a)
> summary(linear1)
```

```
Call:
lm(formula = loghaz0 ~ haz.time + haz.time2 + haz.time3 + new.time3,
    data = dfull.a)

Residuals:
      Min        1Q    Median        3Q       Max
-0.234444 -0.077203 -0.020356  0.006947  0.683959

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -9.107e+00  3.970e-02 -229.379   <2e-16 ***
haz.time     4.905e-05  2.494e-04    0.197   0.8444
haz.time2    6.740e-07  3.446e-07    1.956   0.0527 .
haz.time3   -2.302e-10  1.251e-10   -1.840   0.0681 .
new.time3    5.374e-10  1.987e-10    2.705   0.0078 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1835 on 124 degrees of freedom
Multiple R-squared: 0.9133,     Adjusted R-squared: 0.9105
F-statistic: 326.5 on 4 and 124 DF,  p-value: < 2.2e-16
```

| Parameter | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| $\alpha_0$ (Intercept) | -9.107e+00 | 3.970e-02 | -229.379 | <2e-16 | *** |
| $\alpha_1$ haz.time | 4.905e-05 | 2.494e-04 | 0.197 | 0.8444 | |
| $\alpha_2$ haz.time2 | 6.740e-07 | 3.446e-07 | 1.956 | 0.0527 | . |
| $\alpha_3$ haz.time3 | -2.302e-10 | 1.251e-10 | -1.840 | 0.0681 | . |
| $\alpha_4$ new.time3 | 5.374e-10 | 1.987e-10 | 2.705 | 0.0078 | ** |

**Table 2 Estimates of parameter $\alpha$ for the full model**

## 4.2 Fit the PH model and find the reduced model

Suppose we want a 95% confidence level for all parameters.

### 4.2.1 Variable selection for $\beta$

It is necessary to determine which variables should be included in the fitted PH model. Variable selection is performed using **a forward and backward stepwise procedure**[2] that searches all possible models to determine which model minimized the Akaike Information Criterion (AIC).

$$AIC = -2logL + 2p$$

[2] Swindell, W. 2009. Accelerated failure time models provide a useful statistical framework for aging research, Experimental Gerontology 44 (2009): 190–200.

The selecting criterion can be also based on the ***z* statistics** associated with estimated regression parameters, which are equivalent to the common **chi-square Wald test statistics**[3].

We can use R function *stepAIC()* in the *MASS* library for automatic model selection:

```
> library(MASS)
> phmodel<-coxph(Surv(time,status)~gender+race+age, data=a)
> stepAIC(phmodel)
Start:  AIC=1707.28
Surv(time, status) ~ gender + race + age

         Df    AIC
- gender  1 1705.3
- race    1 1705.6
<none>      1707.3
- age     1 1761.0

Step:  AIC=1705.3
Surv(time, status) ~ race + age

       Df    AIC
- race  1 1703.6
<none>    1705.3
- age   1 1759.3

Step:  AIC=1703.6
Surv(time, status) ~ age

       Df    AIC
<none>    1703.6
- age   1 1758.3
Call:
coxph(formula = Surv(time, status) ~ age, data = a)


      coef exp(coef) se(coef)    z       p
age 0.0511      1.05  0.00714 7.16 8.3e-13

Likelihood ratio test=56.7  on 1 df, p=4.97e-14  n= 863
```

From the output of *stepAIC(phmodel)*, it appears that:

- ***Age*** is perhaps the only important covariate for this data set.

Using the selecting criterion based on the ***z* statistics** associated with estimated regression parameters, which are equivalent to the common **chi-square Wald test statistics**, we have the same results: ***age*** is perhaps the only important covariate for this data set.

Thus, we get:

```
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> summary(phmodel.a)
Call:
coxph(formula = Surv(time, status) ~ age, data = a)

  n= 863
```

[3] Parametric regression in survival analysis, http://www.stat.ncu.edu.tw/teacher/Tsengyk/Handout2a.htm.

```
        coef exp(coef) se(coef)     z Pr(>|z|)
age 0.051068  1.052394 0.007136 7.156  8.3e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
age     1.052     0.9502     1.038     1.067

Rsquare= 0.064   (max possible= 0.87 )
Likelihood ratio test= 56.74  on 1 df,   p=4.974e-14
Wald test            = 51.21  on 1 df,   p=8.301e-13
Score (logrank) test = 53.44  on 1 df,   p=2.666e-13
```

| Parameter | Estimate | Std.Error | z | p |
|---|---|---|---|---|
| $\beta_3$ age | 0.051068 | 0.007136 | 7.156 | 8.3e-13 *** |

**Table 3 Estimate of parameter $\beta$ for the reduced model**

Note:
We can also check the cross products such as *age*gender*, *age*race*, and *race*gender*. The output leads to the same results as table 3.

```
> phmodel.ra<-coxph(Surv(time,status)~race*age, data=a)
> phmodel.ra
Call:
coxph(formula = Surv(time, status) ~ race * age, data = a)

             coef exp(coef) se(coef)     z     p
race       1.0653      2.90   0.9493  1.12 0.2600
age        0.0732      1.08   0.0234  3.13 0.0017
race:age -0.0193      0.98   0.0191 -1.01 0.3100

Likelihood ratio test=58  on 3 df, p=1.56e-12  n= 863

> phmodel.rg<-coxph(Surv(time,status)~race*gender, data=a)
> phmodel.rg
Call:
coxph(formula = Surv(time, status) ~ race * gender, data = a)

               coef exp(coef) se(coef)     z     p
race         -0.834     0.434    0.662 -1.26 0.210
gender       -0.994     0.370    0.548 -1.81 0.070
race:gender  0.745     2.107    0.427  1.75 0.081

Likelihood ratio test=4.37  on 3 df, p=0.224  n= 863

> phmodel.ag<-coxph(Surv(time,status)~age*gender, data=a)
> phmodel.ag
Call:
coxph(formula = Surv(time, status) ~ age * gender, data = a)

               coef exp(coef) se(coef)     z       p
age          0.0725     1.075   0.0217  3.34 0.00085
gender       0.7557     2.129   0.7140  1.06 0.29000
age:gender -0.0150     0.985   0.0143 -1.05 0.29000

Likelihood ratio test=57.9  on 3 df, p=1.68e-12  n= 863
```

All the cross products *age*gender*, *age*race*, and *race*gender* have big *p-values*, which means that they are insignificant.

### **4.2.2** Variable selection for $\alpha$

(i) Obtain $\hat{\alpha}$ for the reduced model

**# Get the estimate of the hazard for an average individual $\lambda_m(t,\alpha)$, i.e. with all covariates evaluated at their average over the sample**

```
> d.phmodel.a =coxph.detail(phmodel.a)
> hazm.a=c(d.phmodel.a$hazard)
> hazm.a
  [1] 0.0009305548 0.0009326398 0.0018711436 0.0018871527 0.0009484691
  [6] 0.0009545158 0.0019160407 0.0009605395 0.0009653551 0.0009663451
 [11] 0.0019370146 0.0009762250 0.0009776009 0.0009812297 0.0009871629
 [16] 0.0009896329 0.0019844702 0.0009939758 0.0019913303 0.0020074466
 [21] 0.0010057774 0.0010089244 0.0010113759 0.0010240866 0.0010255253
 [26] 0.0010305193 0.0020688216 0.0010375936 0.0020850592 0.0010626065
 [31] 0.0010641937 0.0010693751 0.0010757584 0.0010799963 0.0010862599
 [36] 0.0010880531 0.0010902608 0.0011121985 0.0011232298 0.0011273627
 [41] 0.0011352496 0.0011373112 0.0011465126 0.0011503524 0.0011519099
 [46] 0.0011603195 0.0012016754 0.0012133942 0.0012182187 0.0024558939
 [51] 0.0012478176 0.0012539999 0.0012611194 0.0012639382 0.0012692100
 [56] 0.0012727283 0.0012972700 0.0013099960 0.0013275927 0.0013474209
 [61] 0.0013682120 0.0013769377 0.0013788522 0.0013841519 0.0013928097
 [66] 0.0014015073 0.0014398204 0.0014570067 0.0014836055 0.0014866269
 [71] 0.0014895094 0.0015199346 0.0015878509 0.0016043579 0.0016553046
 [76] 0.0016570514 0.0016758830 0.0016833895 0.0016853906 0.0017026904
 [81] 0.0017053341 0.0017165749 0.0017388600 0.0018117617 0.0018199219
 [86] 0.0018257981 0.0036618171 0.0018356174 0.0018844917 0.0018909336
 [91] 0.0019039478 0.0019616549 0.0020561897 0.0020773747 0.0020805829
 [96] 0.0020851868 0.0020997532 0.0022176704 0.0022878723 0.0022947063
[101] 0.0023237748 0.0023872241 0.0024261219 0.0024345830 0.0025600742
[106] 0.0025887394 0.0029880272 0.0031915114 0.0034160959 0.0034285249
[111] 0.0034791716 0.0035794334 0.0040484492 0.0042311875 0.0045253518
[116] 0.0046060111 0.0050776212 0.0051236282 0.0052141110 0.0053005443
[121] 0.0056225644 0.0057873060 0.0058466425 0.0065499214 0.0070489242
[126] 0.0071291642 0.0080546959 0.0108888918 0.0313329011
```

**# Get the estimate of the baseline hazard rate $\lambda_0(t,\alpha)$**

```
> attach(a)
> meanx.a=c(mean(age))
> meanx.a
[1] 42.83662
> beta.a=phmodel.a$coef
> beta.a
      age
0.05106769
> t(meanx.a)%*%beta.a
         [,1]
[1,] 2.187567
> ex.a=exp(t(meanx.a)%*%beta.a)
> ex.a
         [,1]
[1,] 8.9135
> haz0.a=hazm.a/ex.a
> haz0.a
  [1] 0.0001043984 0.0001046323 0.0002099224 0.0002117185 0.0001064082
  [6] 0.0001070865 0.0002149594 0.0001077623 0.0001083026 0.0001084136
 [11] 0.0002173124 0.0001095221 0.0001096764 0.0001100835 0.0001107492
 [16] 0.0001110263 0.0002226365 0.0001115135 0.0002234061 0.0002252142
 [21] 0.0001128375 0.0001131906 0.0001134656 0.0001148916 0.0001150530
 [26] 0.0001156133 0.0002320998 0.0001164070 0.0002339215 0.0001192132
 [31] 0.0001193912 0.0001199725 0.0001206887 0.0001211641 0.0001218668
 [36] 0.0001220680 0.0001223157 0.0001247769 0.0001260144 0.0001264781
 [41] 0.0001273629 0.0001275942 0.0001286265 0.0001290573 0.0001292320
 [46] 0.0001301755 0.0001348152 0.0001361299 0.0001366712 0.0002755252
 [51] 0.0001399919 0.0001406855 0.0001414842 0.0001418004 0.0001423919
```

```
 [56] 0.0001427866 0.0001455399 0.0001469676 0.0001489418 0.0001511663
 [61] 0.0001534988 0.0001544778 0.0001546926 0.0001552871 0.0001562584
 [66] 0.0001572342 0.0001615325 0.0001634607 0.0001664448 0.0001667837
 [71] 0.0001671071 0.0001705205 0.0001781400 0.0001799919 0.0001857076
 [76] 0.0001859035 0.0001880162 0.0001888584 0.0001890829 0.0001910238
 [81] 0.0001913204 0.0001925815 0.0001950816 0.0002032604 0.0002041759
 [86] 0.0002048351 0.0004108169 0.0002059368 0.0002114199 0.0002121426
 [91] 0.0002136027 0.0002200768 0.0002306826 0.0002330594 0.0002334193
 [96] 0.0002339358 0.0002355700 0.0002487990 0.0002566749 0.0002574416
[101] 0.0002607028 0.0002678212 0.0002721851 0.0002731343 0.0002872131
[106] 0.0002904290 0.0003352249 0.0003580536 0.0003832496 0.0003846440
[111] 0.0003903261 0.0004015744 0.0004541929 0.0004746942 0.0005076964
[116] 0.0005167455 0.0005696551 0.0005748166 0.0005849678 0.0005946647
[121] 0.0006307919 0.0006492742 0.0006559311 0.0007348315 0.0007908143
[126] 0.0007998164 0.0009036512 0.0012216179 0.0035152183
```

**# Get the estimate of** $\log\lambda_0(t,\alpha\,|\,\hat{\beta})$

```
> loghaz0.a=log(haz0.a)
> loghaz0.a
  [1] -9.167297 -9.165059 -8.468773 -8.460253 -9.148228 -9.141873 -8.445061
  [8] -9.135583 -9.130582 -9.129557 -8.434174 -9.119385 -9.117976 -9.114271
 [15] -9.108243 -9.105744 -8.409970 -9.101365 -8.406519 -8.398459 -9.089562
 [22] -9.086437 -9.084011 -9.071521 -9.070117 -9.065260 -8.368343 -9.058418
 [29] -8.360525 -9.034597 -9.033105 -9.028248 -9.022296 -9.018365 -9.012582
 [36] -9.010932 -9.008905 -8.988984 -8.979114 -8.975441 -8.968470 -8.966655
 [43] -8.958598 -8.955254 -8.953901 -8.946627 -8.911606 -8.901901 -8.897933
 [50] -8.196832 -8.873926 -8.868984 -8.863323 -8.861090 -8.856928 -8.854159
 [57] -8.835060 -8.825298 -8.811955 -8.797130 -8.781818 -8.775460 -8.774071
 [64] -8.770235 -8.763999 -8.757774 -8.730804 -8.718938 -8.700847 -8.698813
 [71] -8.696876 -8.676655 -8.632941 -8.622599 -8.591337 -8.590283 -8.578982
 [78] -8.574513 -8.573325 -8.563113 -8.561561 -8.554991 -8.542093 -8.501023
 [85] -8.496529 -8.493305 -7.797363 -8.487941 -8.461664 -8.458252 -8.451393
 [92] -8.421534 -8.374468 -8.364217 -8.362674 -8.360464 -8.353503 -8.298865
 [99] -8.267700 -8.264717 -8.252129 -8.225191 -8.209028 -8.205547 -8.155286
[106] -8.144151 -8.000709 -7.934828 -7.866824 -7.863192 -7.848528 -7.820118
[113] -7.696988 -7.652840 -7.585627 -7.567960 -7.470479 -7.461460 -7.443954
[120] -7.427513 -7.368534 -7.339655 -7.329455 -7.215869 -7.142447 -7.131128
[127] -7.009067 -6.707579 -5.650654
```

**# Generate a new dataset for** $\log\lambda_0(t,\alpha\,|\,\hat{\beta}) = \alpha_0+\alpha_1 t+\alpha_2 t^2+\alpha_3 t^3+\alpha_4(t-\gamma)_+^3$ **where *haz.time2*, *haz.time3*, and *new.time3* are already known from the full model**

```
> dreduced<-cbind(loghaz0.a, haz.time, haz.time2, haz.time3, new.time3)
> dreduced.a<-data.frame(dreduced)
> dreduced.a
    loghaz0.a haz.time haz.time2   haz.time3  new.time3
1   -9.167297        2          4           8          0
2   -9.165059        3          9          27          0
3   -8.468773        7         49         343          0
4   -8.460253       10        100        1000          0
5   -9.148228       17        289        4913          0
6   -9.141873       21        441        9261          0
7   -8.445061       26        676       17576          0
8   -9.135583       28        784       21952          0
9   -9.130582       37       1369       50653          0
10  -9.129557       40       1600       64000          0
11  -8.434174       43       1849       79507          0
12  -9.119385       44       1936       85184          0
13  -9.117976       45       2025       91125          0
14  -9.114271       50       2500      125000          0
15  -9.108243       52       2704      140608          0
16  -9.105744       56       3136      175616          0
17  -8.409970       57       3249      185193          0
18  -9.101365       59       3481      205379          0
19  -8.406519       62       3844      238328          0
20  -8.398459       68       4624      314432          0
21  -9.089562       69       4761      328509          0
22  -9.086437       78       6084      474552          0
```

| | | | | | |
|---|---|---|---|---|---|
| 23 | -9.084011 | 79 | 6241 | 493039 | 0 |
| 24 | -9.071521 | 88 | 7744 | 681472 | 0 |
| 25 | -9.070117 | 91 | 8281 | 753571 | 0 |
| 26 | -9.065260 | 97 | 9409 | 912673 | 0 |
| 27 | -8.368343 | 98 | 9604 | 941192 | 0 |
| 28 | -9.058418 | 104 | 10816 | 1124864 | 0 |
| 29 | -8.360525 | 106 | 11236 | 1191016 | 0 |
| 30 | -9.034597 | 119 | 14161 | 1685159 | 0 |
| 31 | -9.033105 | 121 | 14641 | 1771561 | 0 |
| 32 | -9.028248 | 135 | 18225 | 2460375 | 0 |
| 33 | -9.022296 | 143 | 20449 | 2924207 | 0 |
| 34 | -9.018365 | 150 | 22500 | 3375000 | 0 |
| 35 | -9.012582 | 154 | 23716 | 3652264 | 0 |
| 36 | -9.010932 | 158 | 24964 | 3944312 | 0 |
| 37 | -9.008905 | 162 | 26244 | 4251528 | 0 |
| 38 | -8.988984 | 190 | 36100 | 6859000 | 0 |
| 39 | -8.979114 | 206 | 42436 | 8741816 | 0 |
| 40 | -8.975441 | 209 | 43681 | 9129329 | 0 |
| 41 | -8.968470 | 228 | 51984 | 11852352 | 0 |
| 42 | -8.966655 | 229 | 52441 | 12008989 | 0 |
| 43 | -8.958598 | 242 | 58564 | 14172488 | 0 |
| 44 | -8.955254 | 248 | 61504 | 15252992 | 0 |
| 45 | -8.953901 | 249 | 62001 | 15438249 | 0 |
| 46 | -8.946627 | 252 | 63504 | 16003008 | 0 |
| 47 | -8.911606 | 273 | 74529 | 20346417 | 0 |
| 48 | -8.901901 | 291 | 84681 | 24642171 | 0 |
| 49 | -8.897933 | 297 | 88209 | 26198073 | 0 |
| 50 | -8.196832 | 311 | 96721 | 30080231 | 0 |
| 51 | -8.873926 | 334 | 111556 | 37259704 | 0 |
| 52 | -8.868984 | 340 | 115600 | 39304000 | 0 |
| 53 | -8.863323 | 344 | 118336 | 40707584 | 0 |
| 54 | -8.861090 | 346 | 119716 | 41421736 | 0 |
| 55 | -8.856928 | 354 | 125316 | 44361864 | 0 |
| 56 | -8.854159 | 366 | 133956 | 49027896 | 0 |
| 57 | -8.835060 | 391 | 152881 | 59776471 | 0 |
| 58 | -8.825298 | 402 | 161604 | 64964808 | 0 |
| 59 | -8.811955 | 421 | 177241 | 74618461 | 0 |
| 60 | -8.797130 | 439 | 192721 | 84604519 | 0 |
| 61 | -8.781818 | 450 | 202500 | 91125000 | 0 |
| 62 | -8.775460 | 470 | 220900 | 103823000 | 0 |
| 63 | -8.774071 | 478 | 228484 | 109215352 | 0 |
| 64 | -8.770235 | 481 | 231361 | 111284641 | 0 |
| 65 | -8.763999 | 490 | 240100 | 117649000 | 0 |
| 66 | -8.757774 | 495 | 245025 | 121287375 | 0 |
| 67 | -8.730804 | 570 | 324900 | 185193000 | 0 |
| 68 | -8.718938 | 583 | 339889 | 198155287 | 0 |
| 69 | -8.700847 | 614 | 376996 | 231475544 | 0 |
| 70 | -8.698813 | 615 | 378225 | 232608375 | 0 |
| 71 | -8.696876 | 621 | 385641 | 239483061 | 0 |
| 72 | -8.676655 | 652 | 425104 | 277167808 | 0 |
| 73 | -8.632941 | 697 | 485809 | 338608873 | 0 |
| 74 | -8.622599 | 730 | 532900 | 389017000 | 0 |
| 75 | -8.591337 | 773 | 597529 | 461889917 | 0 |
| 76 | -8.590283 | 776 | 602176 | 467288576 | 0 |
| 77 | -8.578982 | 790 | 624100 | 493039000 | 0 |
| 78 | -8.574513 | 793 | 628849 | 498677257 | 0 |
| 79 | -8.573325 | 806 | 649636 | 523606616 | 0 |
| 80 | -8.563113 | 840 | 705600 | 592704000 | 0 |
| 81 | -8.561561 | 852 | 725904 | 618470208 | 0 |
| 82 | -8.554991 | 864 | 746496 | 644972544 | 0 |
| 83 | -8.542093 | 875 | 765625 | 669921875 | 0 |
| 84 | -8.501023 | 929 | 863041 | 801765089 | 0 |
| 85 | -8.496529 | 939 | 881721 | 827936019 | 0 |
| 86 | -8.493305 | 943 | 889249 | 838561807 | 0 |
| 87 | -7.797363 | 945 | 893025 | 843908625 | 0 |
| 88 | -8.487941 | 946 | 894916 | 846590536 | 0 |
| 89 | -8.461664 | 1001 | 1002001 | 1003003001 | 0 |
| 90 | -8.458252 | 1013 | 1026169 | 1039509197 | 0 |
| 91 | -8.451393 | 1016 | 1032256 | 1048772096 | 0 |
| 92 | -8.421534 | 1105 | 1221025 | 1349232625 | 0 |
| 93 | -8.374468 | 1164 | 1354896 | 1577098944 | 0 |

```
94  -8.364217     1186   1406596  1668222856          0
95  -8.362674     1191   1418481  1689410871          0
96  -8.360464     1196   1430416  1710777536          0
97  -8.353503     1210   1464100  1771561000          0
98  -8.298865     1275   1625625  2072671875        216
99  -8.267700     1326   1758276  2331473976     185193
100 -8.264717     1331   1771561  2357947691     238328
101 -8.252129     1357   1841449  2498846293     681472
102 -8.225191     1384   1915456  2650991104    1520875
103 -8.209028     1388   1926544  2674043072    1685159
104 -8.205547     1418   2010724  2851206632    3307949
105 -8.155286     1473   2169729  3196010817    8489664
106 -8.144151     1509   2277081  3436115229   13824000
107 -8.000709     1734   3006756  5213714904  100544625
108 -7.934828     1777   3157729  5611284433  131096512
109 -7.866824     1820   3312400  6028568000  167284151
110 -7.863192     1835   3367225  6178857875  181321496
111 -7.848528     1877   3523129  6612913133  224755712
112 -7.820118     1940   3763600  7301384000  302111711
113 -7.696988     2034   4137156  8414975304  447697125
114 -7.652840     2056   4227136  8690991616  487443403
115 -7.585627     2108   4443664  9367243712  590589719
116 -7.567960     2171   4713241 10232446211  733870808
117 -7.470479     2276   5180176 11790080576 1021147343
118 -7.461460     2291   5248681 12024728171 1067462648
119 -7.443954     2301   5294601 12182876901 1099104768
120 -7.427513     2313   5349969 12374478297 1137893184
121 -7.368534     2369   5612161 13295209409 1331000000
122 -7.339655     2414   5827396 14067333944 1501123625
123 -7.329455     2421   5861241 14190064461 1528823808
124 -7.215869     2489   6195121 15419656169 1815848000
125 -7.142447     2557   6538249 16718302693 2136719872
126 -7.131128     2567   6589489 16915218263 2186875592
127 -7.009067     2650   7022500 18609625000 2633789341
128 -6.707579     2795   7812025 21834609875 3553559576
129 -5.650654     3146   9897316 31136956136 6612913133
```

**# Export and save the new dataset for the reduced model**

```
> write.table(dreduced.a, "M:/dreduced.a.txt", sep=" ", col.names=TRUE,
row.names=FALSE, quote=FALSE, na="NA")
```

**# Use the *lm()* function in R to fit** $\log\hat{\lambda}_0(t,\alpha\,|\,\hat{\beta})$

```
> linear2 <- lm(loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3,
data=dreduced.a)
> summary(linear2)

Call:
lm(formula = loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3, data
= dreduced.a)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23444 -0.07772 -0.02158  0.00818  0.68299

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -8.933e+00  3.972e-02 -224.906   <2e-16 ***
haz.time     5.135e-05  2.495e-04    0.206   0.8373
haz.time2    6.674e-07  3.448e-07    1.936   0.0552 .
haz.time3   -2.275e-10  1.252e-10   -1.817   0.0716 .
new.time3    5.340e-10  1.988e-10    2.687   0.0082 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1836 on 124 degrees of freedom
Multiple R-squared: 0.9133,     Adjusted R-squared: 0.9105
```

```
F-statistic: 326.4 on 4 and 124 DF,  p-value: < 2.2e-16
```

(ii) Variable selection for $\alpha$ using the **F-test**

**# Use the *lm()* function in R to fit** $\log \lambda_0(t, \alpha \mid \hat{\beta})$

```
> linear2 <- lm(loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3,
data=dreduced.a)
> summary(linear2)

Call:
lm(formula = loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3, data
= dreduced.a)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23444 -0.07772 -0.02158  0.00818  0.68299

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -8.933e+00  3.972e-02 -224.906   <2e-16 ***
haz.time     5.135e-05  2.495e-04    0.206   0.8373
haz.time2    6.674e-07  3.448e-07    1.936   0.0552 .
haz.time3   -2.275e-10  1.252e-10   -1.817   0.0716 .
new.time3    5.340e-10  1.988e-10    2.687   0.0082 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1836 on 124 degrees of freedom
Multiple R-squared: 0.9133,     Adjusted R-squared: 0.9105
F-statistic: 326.4 on 4 and 124 DF,  p-value: < 2.2e-16
```

- Step 1

Since the variable *haz.time* is associated with the least *t-value* and hence the largest *p-value*, it is excluded from the linear regression model first.

**# Exclude *haz.time***

```
> linear2.1 <- lm(loghaz0.a ~ haz.time2 + haz.time3 + new.time3,
data=dreduced.a)
> summary(linear2.1)

Call:
lm(formula = loghaz0.a ~ haz.time2 + haz.time3 + new.time3, data =
dreduced.a)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24084 -0.07578 -0.01985  0.01164  0.68380

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -8.926e+00  2.397e-02 -372.458  < 2e-16 ***
haz.time2    7.362e-07  8.529e-08    8.631 2.36e-14 ***
haz.time3   -2.514e-10  4.658e-11   -5.397 3.28e-07 ***
new.time3    5.679e-10  1.105e-10    5.140 1.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1828 on 125 degrees of freedom
```

```
Multiple R-squared: 0.9132,     Adjusted R-squared: 0.9112
F-statistic: 438.5 on 3 and 125 DF,  p-value: < 2.2e-16

# Compare linear2 and linear2.1 using the anova() function
> anova(linear2,linear2.1)
Analysis of Variance Table

Model 1: loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3
Model 2: loghaz0.a ~ haz.time2 + haz.time3 + new.time3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    124 4.1777
2    125 4.1792 -1 -0.001427 0.0424 0.8373
```

The **F-test** returns a *p-value=0.8373 >0.05*. This means that *linear2* and *linear2.1* are not significant different after excluding the variable *haz.time*.

- Step 2

In the linear model *linear2.1*, since the variable *new.time3* is associated with the least **t-value** and hence the largest **p-value**, we try to exclude it and check the results.

```
> linear2.2 <- lm(loghaz0.a ~ haz.time2 + haz.time3, data=dreduced.a)
> summary(linear2.2)

Call:
lm(formula = loghaz0.a ~ haz.time2 + haz.time3, data = dreduced.a)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29231 -0.12209 -0.01858  0.06185  0.78748

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -8.875e+00  2.387e-02 -371.806  < 2e-16 ***
haz.time2    3.491e-07  4.391e-08    7.951 9.08e-13 ***
haz.time3   -2.559e-11  1.699e-11   -1.506    0.135
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2004 on 126 degrees of freedom
Multiple R-squared: 0.8949,     Adjusted R-squared: 0.8932
F-statistic: 536.4 on 2 and 126 DF,  p-value: < 2.2e-16

# Compare linear2 and linear2.2 using the anova() function
> anova(linear2,linear2.2)
Analysis of Variance Table

Model 1: loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3
Model 2: loghaz0.a ~ haz.time2 + haz.time3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    124 4.1777
2    126 5.0625 -2  -0.88477 13.130 6.725e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **F-test** returns a *p-value=6.725e-06 <0.05*. This means that *linear2* and *linear2.2* are significant different. So, we should keep the variable *new.time3*.

Since at this step, the variable *haz.time3* has extraordinary values of *t-value* and *p-value*, we repeat the step 2 on the variable *haz.time3*.

```
> linear2.3 <- lm(loghaz0.a ~ haz.time2 + new.time3, data=dreduced.a)
> summary(linear2.3)

Call:
lm(formula = loghaz0.a ~ haz.time2 + new.time3, data = dreduced.a)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31068 -0.13736  0.00506  0.09216  0.80704

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -8.857e+00  2.231e-02 -396.980   <2e-16 ***
haz.time2    2.824e-07  1.587e-08   17.795   <2e-16 ***
new.time3    5.647e-12  4.067e-11    0.139     0.89
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2022 on 126 degrees of freedom
Multiple R-squared: 0.893,      Adjusted R-squared: 0.8913
F-statistic: 525.9 on 2 and 126 DF,  p-value: < 2.2e-16

> anova(linear2,linear2.3)
Analysis of Variance Table

Model 1: loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3
Model 2: loghaz0.a ~ haz.time2 + new.time3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    124 4.1777
2    126 5.1528 -2  -0.97511 14.471 2.246e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **F-test** returns a *p-value=2.246e-06 <0.05*. This means that *linear2* and *linear2.3* are significant different. So, we should keep the variable *haz.time3*.

- Step 3

After several **forward and backward** rounds check, such as:

```
> linear2.4 <- lm(loghaz0.a ~ haz.time3 + new.time3, data=dreduced.a)
> anova(linear2,linear2.4)
> linear2.5 <- lm(loghaz0.a ~ haz.time + haz.time3 + new.time3,
data=dreduced.a)
> anova(linear2,linear2.5)
> linear2.6 <- lm(loghaz0.a ~ haz.time + haz.time2 + new.time3,
data=dreduced.a)
> anova(linear2,linear2.6)
…….
…….
…….
```

Finally, it appears that:

- ***haz.time2*** ($t^2$)**, *haz.time3*** ($t^3$)**,** and ***new.time3*** ($(t-\gamma)_+^3$) <u>are perhaps the only important variables</u> in the linear model of $\log \lambda_0(t, \alpha \mid \hat{\beta}) =$ $\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t-\gamma)_+^3$.

**Note:**
We can also use R function *stepwise()* to obtain the same result more readily:

```
> stepwise(lm, direction='backward/forward', criterion='AIC')

Direction:  backward/forward
Criterion:  AIC

Start:  AIC=-432.48
loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3

            Df Sum of Sq    RSS     AIC
- haz.time   1  0.001427 4.1792 -434.43
<none>                   4.1777 -432.48
- haz.time3  1  0.111263 4.2890 -431.09
- haz.time2  1  0.126270 4.3040 -430.63
- new.time3  1  0.243181 4.4209 -427.18

Step:  AIC=-434.43
loghaz0.a ~ haz.time2 + haz.time3 + new.time3

            Df Sum of Sq    RSS     AIC
<none>                   4.1792 -434.43
+ haz.time   1   0.00143 4.1777 -432.48
- new.time3  1   0.88334 5.0625 -411.70
- haz.time3  1   0.97368 5.1528 -409.41
- haz.time2  1   2.49054 6.6697 -376.13

Call:
lm(formula = loghaz0.a ~ haz.time2 + haz.time3 + new.time3, data =
dreduced.a)

Coefficients:
(Intercept)    haz.time2    haz.time3    new.time3
 -8.926e+00    7.362e-07   -2.514e-10    5.680e-10
```

The selected model has the following performance:

```
> linear2.1 <- lm(loghaz0.a ~ haz.time2 + haz.time3 + new.time3,
data=dreduced.a)
> summary(linear2.1)

Call:
lm(formula = loghaz0.a ~ haz.time2 + haz.time3 + new.time3, data =
dreduced.a)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24084 -0.07578 -0.01985  0.01164  0.68380

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -8.926e+00  2.397e-02 -372.458  < 2e-16 ***
haz.time2    7.362e-07  8.529e-08    8.631 2.36e-14 ***
haz.time3   -2.514e-10  4.658e-11   -5.397 3.28e-07 ***
```

```
new.time3     5.679e-10  1.105e-10    5.140 1.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1828 on 125 degrees of freedom
Multiple R-squared: 0.9132,     Adjusted R-squared: 0.9112
F-statistic: 438.5 on 3 and 125 DF,  p-value: < 2.2e-16
```

Thus we determine the final $\hat{\alpha}$:

| Parameter | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| $\alpha_0$ (Intercept) | -8.926e+00 | 2.397e-02 | -372.458 | < 2e-16 | *** |
| $\alpha_2$ haz.time2 | 7.362e-07 | 8.529e-08 | 8.631 | 2.36e-14 | *** |
| $\alpha_3$ haz.time3 | -2.514e-10 | 4.658e-11 | -5.397 | 3.28e-07 | *** |
| $\alpha_4$ new.time3 | 5.679e-10 | 1.105e-10 | 5.140 | 1.03e-06 | *** |

**Table 4 Estimates of parameter $\alpha$ for the reduced model**

### 4.3 Influence analysis for the PH model fit

From the information obtained in 4.1 and 4.2, we know that:

| | Parameter | Estimate | Std.Error | z or t | P | |
|---|---|---|---|---|---|---|
| Full model | $\beta_1$ gender | 0.026540 | 0.174900 | 0.152 | 0.879 | |
| | $\beta_2$ race | 0.116365 | 0.211512 | 0.550 | 0.582 | |
| | $\beta_3$ age | 0.051035 | 0.007181 | 7.107 | 1.19e-12 | *** |
| | $\alpha_0$ (Intercept) | -9.107e+00 | 3.970e-02 | -229.379 | <2e-16 | *** |
| | $\alpha_1$ haz.time (t) | 4.905e-05 | 2.494e-04 | 0.197 | 0.8444 | |
| | $\alpha_2$ haz.time2 ($t^2$) | 6.740e-07 | 3.446e-07 | 1.956 | 0.0527 | . |
| | $\alpha_3$ haz.time3 ($t^3$) | -2.302e-10 | 1.251e-10 | -1.840 | 0.0681 | . |
| | $\alpha_4$ new.time3 $(t-\gamma)_+^3$ | 5.374e-10 | 1.987e-10 | 2.705 | 0.0078 | ** |
| Reduced model | $\beta_3$ age | 0.051068 | 0.007136 | 7.156 | 8.3e-13 | *** |
| | $\alpha_0$ (Intercept) | -8.926e+00 | 2.397e-02 | -372.458 | < 2e-16 | *** |
| | $\alpha_2$ haz.time2 ($t^2$) | 7.362e-07 | 8.529e-08 | 8.631 | 2.36e-14 | *** |
| | $\alpha_3$ haz.time3 ($t^3$) | -2.514e-10 | 4.658e-11 | -5.397 | 3.28e-07 | *** |
| | $\alpha_4$ new.time3 $(t-\gamma)_+^3$ | 5.679e-10 | 1.105e-10 | 5.140 | 1.03e-06 | *** |

**Table 5 Parameter estimates of the full model and the reduced model**

The reduced PH model is expressed as:

$$\lambda_i(t;\hat{\alpha},\hat{\beta},\underline{x}) = \lambda_0(t,\hat{\alpha})\exp\{\underline{x}^T\hat{\beta}\}$$

$$= \exp\{\hat{\alpha}_0+\hat{\alpha}_2 t^2+\hat{\alpha}_3 t^3+\hat{\alpha}_4(t-\gamma)_+^3\}\exp\{x_{age}\hat{\beta}_3\},$$

For $\beta$, from the output of R, we know that:

| | Loglik without covariates | Loglik with covariates |
|---|---|---|
| Full model | -879.1705 | -850.6405 |
| Reduced model | -879.1705 | -850.8007 |

**Table 6 Log-likelihood for the full model and the reduced model**

Assuming our diagnostics give some credibility to our chosen model, we can do some **inference analysis**. Let $\underline{\theta}=[\underline{\theta_1},\underline{\theta_2}]=[\beta,\alpha]$.

1) Test: $H_0: \beta_1=\beta_2=\beta_3=0$

$$\Lambda_0 = 2\,(l(\hat{\theta})-l(\tilde{\theta}))\; = 2\,(\text{loglik}_1-\text{loglik}_0)$$

$$= 2\,(-850.6405+879.1705)\; = 57.06007$$

```
# The full model
> phmodel<-coxph(Surv(time,status)~gender+race+age, data=a)
> phmodel$loglik
```

```
[1] -879.1705 -850.6405
# The first element is the log likelihood for a model

# Thus the likelihood ratio test is given by
> 2*(phmodel$loglik[2]- phmodel$loglik[1])
[1] 57.06007

# The p-value is given by
> pchisq(2*(phmodel$loglik[2]-phmodel$loglik[1]),df=3,lower.tail=F)
[1] 2.495107e-12

# Use another way to work out p-value with d.f.= 3
> 1-pchisq(57.06007,3)
[1] 2.495115e-12
```

The *p-value* is much smaller than 0.05. This means that: at the 95% confidence level, we may reject $H_0$, and find strong evidence that $\beta_1, \beta_2, \beta_3$ may not be equal to 0 at the same time.

2) Test: $H_0$: $\beta_1 = \beta_2 = 0$

Let $\hat{\theta}$ refer to the parameters of the full model, and $\tilde{\theta}$ be for the reduced model. This is a PH model with the partial likelihood leading to $\beta$ directly (we can neglect $\alpha$ when checking $\beta$), thus,

$$\Lambda_1 = 2\,(l(\hat{\theta}) - l(\tilde{\theta}))$$

$$= 2\,(l(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \underline{\hat{\alpha}}) - l(0,0,\tilde{\beta}_3, \underline{\tilde{\alpha}}))$$

$$= 2\,(\text{loglik}_1 - \text{loglik}_2) = 2\,(-850.6405 + 850.8007)$$

$$= 0.3204133$$

```
# The full model
> phmodel<-coxph(Surv(time,status)~gender+race+age, data=a)
> phmodel$loglik
[1] -879.1705 -850.6405

# The reduced model
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> phmodel.a$loglik
[1] -879.1705 -850.8007


# Thus the likelihood ratio test is given by
> 2*(phmodel$loglik[2]- phmodel.a$loglik[2])
[1] 0.3204133


# Work out p-value with d.f.=2
> 1-pchisq(0.3204133,2)
[1] 0.8519677
```

The *p-value* is bigger than 0.05. This means that: at the 95% confidence level, there is **hardly** any log-likelihood difference between the full model and the reduced model, and we find no evidence against the model only the variable ***age*** included.

This conclusion corresponds to the previous results using the AIC approach.

3) Test: $H_0$: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$

Since this is for a linear model

$$\log \lambda_0(t, \alpha \mid \hat{\beta}) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t-\gamma)_+^3,$$

we can use the *F-test* to check.

```
# The full linear model
> linear1 <- lm(loghaz0 ~ haz.time + haz.time2 + haz.time3 + new.time3,
data=dfull.a)
> lmm<-lm(formula = loghaz0 ~ 1, data = dfull.a)
> anova(linear1,lmm)
Analysis of Variance Table

Model 1: loghaz0 ~ haz.time + haz.time2 + haz.time3 + new.time3
Model 2: loghaz0 ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    124  4.174
2    128 48.132 -4   -43.958 326.46 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The value of *F-test 326.46* is big enough and the *p-value=2.2e-16* is small enough. This means that: at the 95% confidence level, we may reject $H_0$, and find strong evidence that $\alpha_1, \alpha_2, \alpha_3$, and $\alpha_4$ may not be equal to 0 at the same time.

4) Test: $H_0$: $\alpha_1 = 0$

Use the *F-test* to check.

```
> linear2 <- lm(loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3,
data=dreduced.a)
> linear2.1 <- lm(loghaz0.a ~ haz.time2 + haz.time3 + new.time3,
data=dreduced.a)
> anova(linear2,linear2.1)
Analysis of Variance Table

Model 1: loghaz0.a ~ haz.time + haz.time2 + haz.time3 + new.time3
Model 2: loghaz0.a ~ haz.time2 + haz.time3 + new.time3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    124 4.1777
2    125 4.1792 -1 -0.001427 0.0424 0.8373
```

The **F-test** returns a *p-value=0.8373 >0.05*. This means that *linear2* and *linear2.1* are not significant different after excluding the variable *haz.time (t)*. There is **hardly** any difference between the full **linear** model and the reduced **linear** model, and we find no evidence against the model only the variable *haz.time2* ($t^2$), *haz.time3* ($t^3$), and *new.time3* ($(t-\gamma)_+^3$) included.

This conclusion corresponds to the previous results using the AIC approach through R function *stepwise()*.

# 5 Get inferences for the survivor functions at a point from the reduced model

## 5.1 Obtain inferences for the hazard ratio

The reduced PH model in this case can be expressed as:

$$\lambda_i(t;\hat{\alpha},\hat{\beta},\underline{x}) = \lambda_0(t,\hat{\alpha})\exp\{\underline{x}^T\hat{\beta}\}$$

$$= \exp\{\hat{\alpha}_0+\hat{\alpha}_2 t^2+\hat{\alpha}_3 t^3+\hat{\alpha}_4(t-\gamma)_+^3\}\exp\{x_{age}\hat{\beta}_3\},$$

Where $(t-\gamma)_+ = \max\{0, t-\gamma\}$ and $\gamma = 1269$.

| | Parameter | Estimate | Std.Error | z or t | p |
|---|---|---|---|---|---|
| Reduced model | $\beta_3$ age | 0.051068 | 0.007136 | 7.156 | 8.3e-13 *** |
| | $\alpha_0$ (Intercept) | -8.926e+00 | 2.397e-02 | -372.458 | < 2e-16 *** |
| | $\alpha_2$ haz.time2 ($t^2$) | 7.362e-07 | 8.529e-08 | 8.631 | 2.36e-14 *** |
| | $\alpha_3$ haz.time3 ($t^3$) | -2.514e-10 | 4.658e-11 | -5.397 | 3.28e-07 *** |
| | $\alpha_4$ new.time3 $(t-\gamma)_+^3$ | 5.679e-10 | 1.105e-10 | 5.140 | 1.03e-06 *** |

**Table 7 Parameter estimates of the reduced model**

From table 7, $\hat{\beta}_3$ = *0.051068*, the standard error for the relative risk is *0.007136*. The results indicate that the relative risk of dying increases about 5% for each one year increase in age, and the *age* effect is nearly consistent for *gender* (males and females) and for *race* (blacks and whites).

The *coxph()* function gives the hazard ratio for a *one unit change* in the predictor as well as the 95% confidence interval. Also given is the Wald statistic for each parameter as well as overall likelihood ratio, Wald and Score tests.

We can use the output of the *coxph()* function to estimate the point for the hazard ratio:

$$\hat{hr}(\underline{x}^{*T} : \underline{x}^{T}) = \frac{\exp\{\underline{x}^{*T}\hat{\beta}\}}{\exp\{\underline{x}^{T}\hat{\beta}\}} = \exp\{(\underline{x}^{*T} - \underline{x}^{T})\hat{\beta}\}.$$

We can construct 95% confidence intervals for the hazard ratio as:

$$\exp\{(\underline{x}^{*T} - \underline{x}^{T})\hat{\beta} \pm 1.96\,\hat{se}((\underline{x}^{*T} - \underline{x}^{T})\hat{\beta})\}$$

For example, in this case, we want to estimate *hr(age=60 : age=50)*, the point estimate is obtained as:

$$\hat{hr}\,(age=60 : age=50)$$

$$= \exp\{(\underline{x}^{*T} - \underline{x}^{T})\hat{\beta}\} = \exp\{(x^{*}_{age} - x_{age})\hat{\beta}_3\}$$

$$= \exp\{(60-50)\times 0.051068\}$$

$$\approx 1.666424$$

```
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> phmodel.a$var
             [,1]
[1,] 5.092617e-05
> sqrt(diag(phmodel.a$var))
[1] 0.007136258
```

Then the estimate of the variance of $(x^{*}_{age} - x_{age})\hat{\beta}_3 = (60-50)\hat{\beta}_3$:

$$\hat{\text{var}}[(60-50)\hat{\beta}_3] = 100\,\hat{\text{var}}[\hat{\beta}_3] = 100*(5.092617\text{e-}05) = 0.005092617$$

Thus, the 95% confidence interval for *hr(age=60 : age=50)* is:

$$\exp\{(\underline{x}^{*T} - \underline{x}^{T})\hat{\beta} \pm 1.96\,\hat{se}((\underline{x}^{*T} - \underline{x}^{T})\hat{\beta})\}$$

$$= \exp\{(60-50)\hat{\beta}_3 \pm 1.96\,\hat{se}((60-50)\hat{\beta}_3)\}$$

$$= 1.666424 * \exp(\pm 1.96 \times \sqrt{0.005092617})$$

$$= 1.666424 * \exp(\pm 0.1398706)$$

That is, a 95% confidence interval for the hazard ratio comparing *age=60* and *age=50* is obtained as: (1.448907, 1.916596).

Since possible relationships in the hazard ratio is:

- If the hazard ratio = 1, the changing in *age* does not effect survival.
- If the hazard ratio < 1, the changing in *age* is associated with increased survival.

- If the hazard ratio > 1, the changing in *age* is associated with decreased survival.

In this case, $\hat{hr}$ *(age=60 : age=50)>1* and *the confidence interval* also *>1*, it means the increase in *age* is associated with decreased survival.

We can also use graph to depict the above scenario:

```
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> phmodel.a
Call:
coxph(formula = Surv(time, status) ~ age, data = a)

      coef exp(coef) se(coef)    z       p
age 0.0511      1.05  0.00714 7.16 8.3e-13

Likelihood ratio test=56.7  on 1 df, p=4.97e-14  n= 863
> phmodel.a1<-survfit(phmodel.a, list(age=60))
> phmodel.a2<-survfit(phmodel.a, list(age=50))
> plot(phmodel.a1$time, phmodel.a1$surv, type="l", ylim=c(0,1), col= c(1:2),
xlab="Time (days)", ylab="Survival probability", cex.main=0.8, cex.lab=0.7,
cex.axis=0.7)
> lines(phmodel.a2$time, phmodel.a2$surv, lty=2, col=2)
> legend(2, 0.3, c("age=60", "age=50"), lty=c(1,2), col= c(1:2), cex=.7)
```

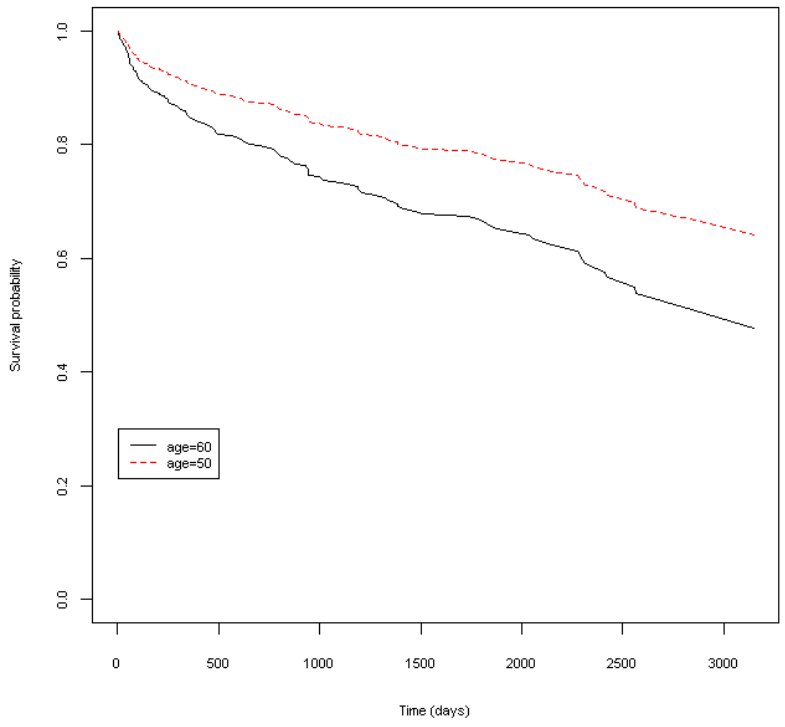


**Figure 4 Compare the survival curves between *age=60* and *age=50***

From figure 4, we also see that the increase in *age* is associated with decreased survival.

## 5.2 Plot survival curves using the reduced PH model

We know that the cumulative hazard function is: $H(t) = \int_0^t \lambda(u)du$. In this case, under a discrete time framework $H(t_k) = \sum_{j=1}^{k} \lambda(t_k)$, then the survival function is given as:

$S(t_k) = \exp\{-H(t_k)\}$.

Consider 10 patients with the following covariate combinations:

- $X_1$ = *(age = 0)*
- $X_2$ = *(age = 5)*
- $X_3$ = *(age = 10)*
- $X_4$ = *(age = 20)*
- $X_5$ = *(age = 30)*
- $X_6$ = *(age = 40)*
- $X_7$ = *(age = 50)*
- $X_8$ = *(age = 60)*
- $X_9$ = *(age = 70)*
- $X_{10}$ = *(age = 80)*

We can use *survfit()* to extract *S(t)* in this case.

```
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> phmodel.a0<-survfit(phmodel.a, list(age=0))
> phmodel.a1<-survfit(phmodel.a, list(age=5))
> phmodel.a2<-survfit(phmodel.a, list(age=10))
> phmodel.a3<-survfit(phmodel.a, list(age=20))
> phmodel.a4<-survfit(phmodel.a, list(age=30))
> phmodel.a5<-survfit(phmodel.a, list(age=40))
> phmodel.a6<-survfit(phmodel.a, list(age=50))
> phmodel.a7<-survfit(phmodel.a, list(age=60))
> phmodel.a8<-survfit(phmodel.a, list(age=70))
> phmodel.a9<-survfit(phmodel.a, list(age=80))
> plot(phmodel.a0$time, phmodel.a0$surv, type="l", xlab="Time (days)",
ylim=c(0,1), ylab="Survival probability", main="Survival curves of the
reduced PH model with the unspecified baseline hazard function",lty= c(1:10),
cex.main=0.8, cex.lab=0.7, cex.axis=0.7, col=c(1:10))
> lines(phmodel.a1$time, phmodel.a1$surv, lty=2, col=2)
> lines(phmodel.a2$time, phmodel.a2$surv, lty=3, col=3)
> lines(phmodel.a3$time, phmodel.a3$surv, lty=4, col=4)
> lines(phmodel.a4$time, phmodel.a4$surv, lty=5, col=5)
> lines(phmodel.a5$time, phmodel.a5$surv, lty=6, col=6)
> lines(phmodel.a6$time, phmodel.a6$surv, lty=7, col=7)
> lines(phmodel.a7$time, phmodel.a7$surv, lty=8, col=8)
> lines(phmodel.a8$time, phmodel.a8$surv, lty=9, col=9)
> lines(phmodel.a9$time, phmodel.a9$surv, lty=10, col=10)
> legend(2, 0.38, c("Subject X1: age=0", "Subject X2: age=5", "Subject X3:
age=10", "Subject X4: age=20", "Subject X5: age=30", "Subject X6: age=40",
"Subject X7: age=50", "Subject X8: age=60", "Subject X9: age=70", "Subject
X10: age=80"), lty=c(1:10), col= c(1:10), cex=.7)
```

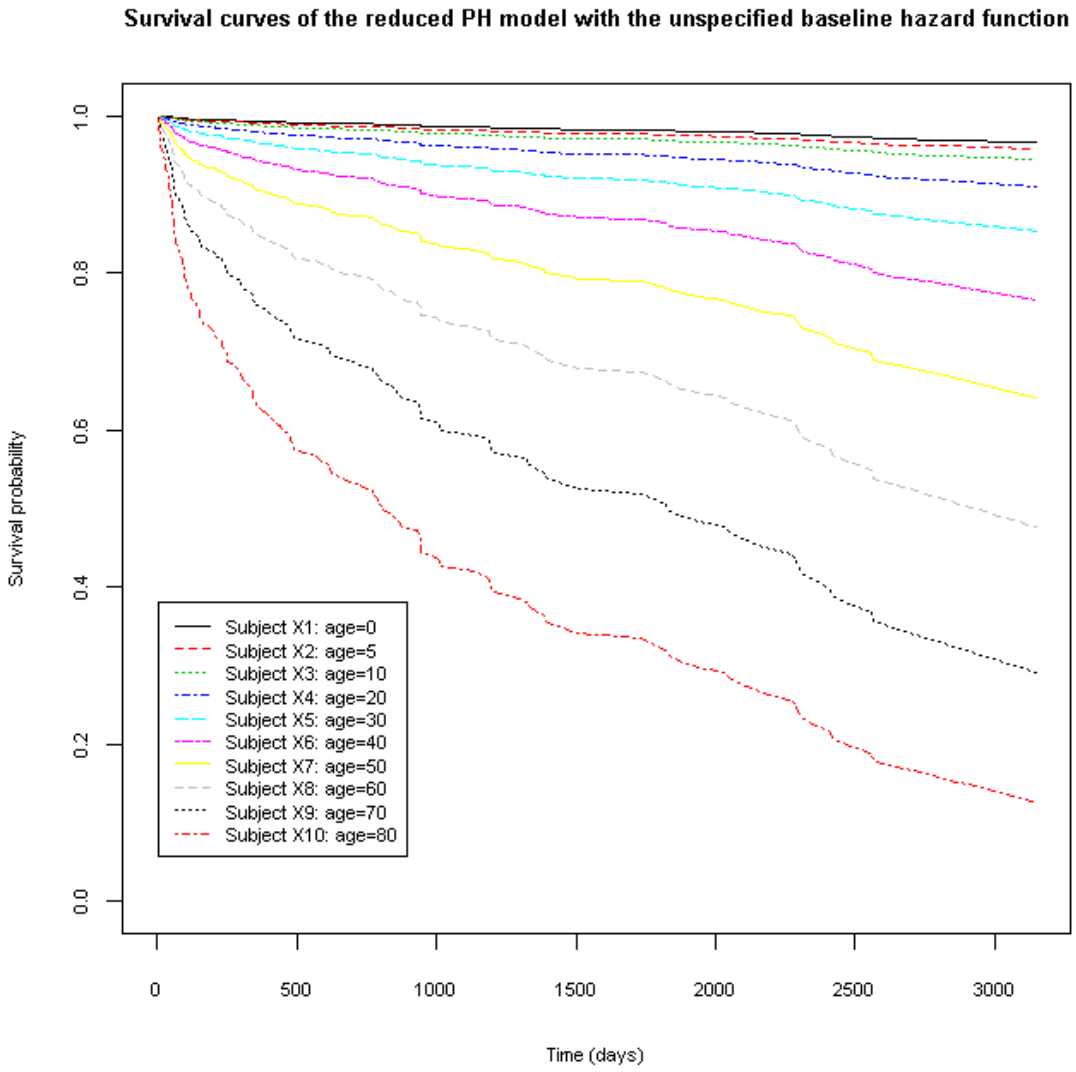


**Figure 5 Survival curves of the reduced PH model with the *unspecified* baseline hazard function**

We see that the increase in age is associated with decreased survival.

## 5.3 Plot the baseline hazard curve using the reduced PH model

The *basehaz()* function in R is to compute the baseline hazard function

$$H_0(t) = \Lambda_0(t) = \int_0^t \lambda_0(u)du$$

where $\underline{x} = \underline{0}$.

In this case, $H_0(t_k, \alpha) = \Lambda_0(t_k, \alpha) = \sum_{j=1}^{k} \lambda_0(t_k, \alpha)$ since it is under a discrete time framework.

```
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> phmodel.a
> hazardr<-basehaz(phmodel.a, centered=F)  # baseline hazard at x=0
> hazardr$hazard
  [1] 0.0001043984 0.0002090306 0.0004189530 0.0006306715 0.0007370797 0.0008441662
  [7] 0.0010591256 0.0011668879 0.0012751905 0.0013836041 0.0016009166 0.0017104386
 [13] 0.0018201150 0.0019301986 0.0020409478 0.0021519741 0.0023746105 0.0024861240
 [19] 0.0027095301 0.0029347443 0.0030475818 0.0031607724 0.0032742380 0.0033891296
 [25] 0.0035041827 0.0036197960 0.0038518958 0.0039683027 0.0042022242 0.0043214374
 [31] 0.0044408286 0.0045608011 0.0046814898 0.0048026539 0.0049245207 0.0050465887
 [37] 0.0051689043 0.0052936812 0.0054196956 0.0055461737 0.0056735367 0.0058011309
 [43] 0.0059297574 0.0060588147 0.0061880468 0.0063182223 0.0064530375 0.0065891674
 [49] 0.0067258386 0.0070013638 0.0071413556 0.0072820411 0.0074235253 0.0075653257
 [55] 0.0077077176 0.0078505041 0.0079960440 0.0081430117 0.0082919535 0.0084431198
 [61] 0.0085966186 0.0087510964 0.0089057889 0.0090610760 0.0092173345 0.0093745687
 [67] 0.0095361012 0.0096995619 0.0098660067 0.0100327904 0.0101998975 0.0103704180
 [73] 0.0105485580 0.0107285499 0.0109142575 0.0111001610 0.0112881773 0.0114770357
 [79] 0.0116661186 0.0118571423 0.0120484627 0.0122410442 0.0124361258 0.0126393862
 [85] 0.0128435621 0.0130483972 0.0134592141 0.0136651509 0.0138765708 0.0140887135
 [91] 0.0143023162 0.0145223930 0.0147530756 0.0149861350 0.0152195542 0.0154534900
 [97] 0.0156890600 0.0159378591 0.0161945340 0.0164519757 0.0167126785 0.0169804996
[103] 0.0172526847 0.0175258190 0.0178130322 0.0181034612 0.0184386861 0.0187967397
[109] 0.0191799894 0.0195646334 0.0199549595 0.0203565338 0.0208107268 0.0212854210
[115] 0.0217931174 0.0223098628 0.0228795179 0.0234543346 0.0240393024 0.0246339671
[121] 0.0252647590 0.0259140332 0.0265699643 0.0273047959 0.0280956102 0.0288954266
[127] 0.0297990779 0.0310206957 0.0345359141
> hazardr$time
  [1]    2    3    7   10   17   21   26   28   37   40   43   44   45   50   52   56   57
 [18]   59   62   68   69   78   79   88   91   97   98  104  106  119  121  135  143  150
 [35]  154  158  162  190  206  209  228  229  242  248  249  252  273  291  297  311  334
 [52]  340  344  346  354  366  391  402  421  439  450  470  478  481  490  495  570  583
 [69]  614  615  621  652  697  730  773  776  790  793  806  840  852  864  875  929  939
 [86]  943  945  946 1001 1013 1016 1105 1164 1186 1191 1196 1210 1275 1326 1331 1357 1384
[103] 1388 1418 1473 1509 1734 1777 1820 1835 1877 1940 2034 2056 2108 2171 2276 2291 2301
[120] 2313 2369 2414 2421 2489 2557 2567 2650 2795 3146
> plot(hazardr$time, hazardr$hazard, type="l",xlab="time",ylab="H(t)",
main="Baseline hazard (age=0)", cex.main=0.8, cex.lab=0.7, cex.axis=0.7)
```

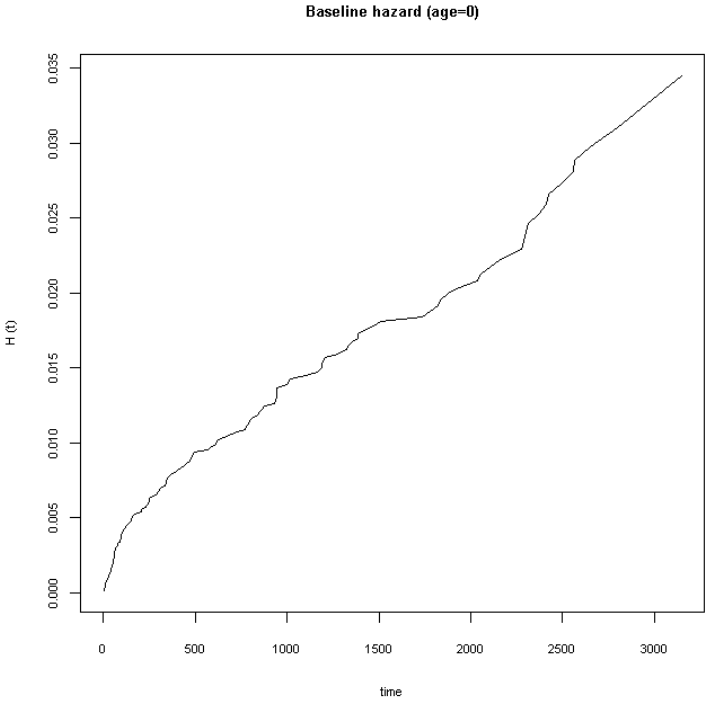


**Figure 6 Cumulative baseline hazard when *age=0***

```
> plot(hazardr$time, exp(-hazardr$hazard), type="l", xlab="time",ylab="S(t)",
main="Baseline survival (age=0)", cex.main=0.8, cex.lab=0.7, cex.axis=0.7)
```

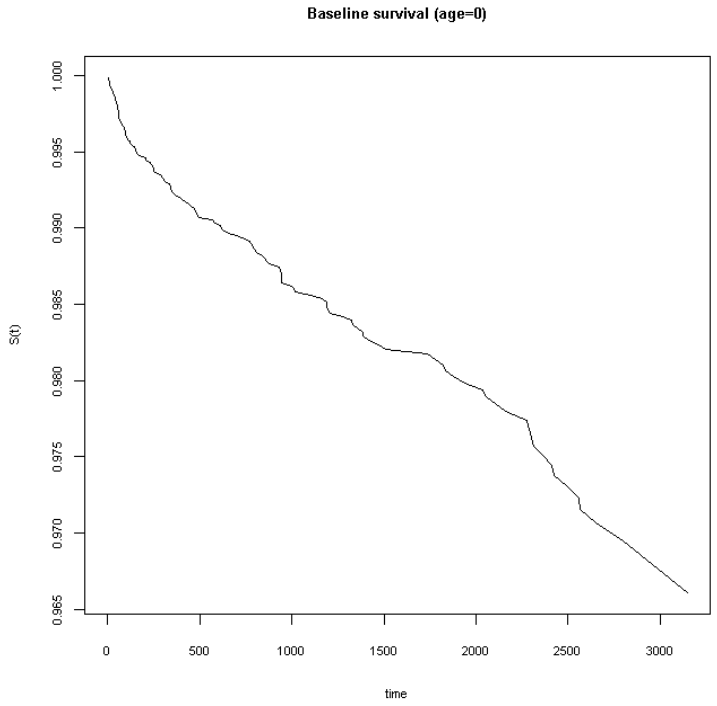


**Figure 7 Baseline survival curve when *age=0***

Note: the same baseline survival can be also obtained by

```
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> phmodel.a0<-survfit(phmodel.a, list(age=0))
> plot(phmodel.a0$time, phmodel.a0$surv, type="l", xlab="Time (days)",
ylab="Survival Probability", lty= c(1:10), cex.main=0.8, cex.lab=0.7,
cex.axis=0.7, col=c(1:10))
```

# 6 Perform diagnostic analyses to evaluate the adequacy of model fit

## 6.1 Residual analysis

Residual analysis is used to evaluate whether or not observations are characterized by the model. Tests and graphical diagnostics[4] for proportional hazards may be based on the scaled *Schoenfeld* residuals; these can be obtained directly as *residuals(model, "scaledsch")*, where model is a *coxph* model object. The matrix returned by residuals has one column for each covariate in the model.

[4] Fox, J. 2002. Cox Proportional-Hazards Regression for Survival Data. http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf.

More conveniently, the *cox.zph* function calculates tests of the PH assumption for each covariate, by correlating the corresponding set of scaled *Schoenfeld* residuals with a suitable transformation of time.

```
# Test the Proportional Hazards Assumption of a Cox Regression
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> coxtest<- cox.zph(phmodel.a)
> coxtest
       rho chisq     p
age 0.0941  1.14 0.285
```

There is, therefore, strong evidence of proportional hazards for *age*.

```
# Plotting the object returned by cox.zph produces graphs of the scaled
Schoenfeld residuals against transformed time
> plot(coxtest, cex.lab=0.7)
```

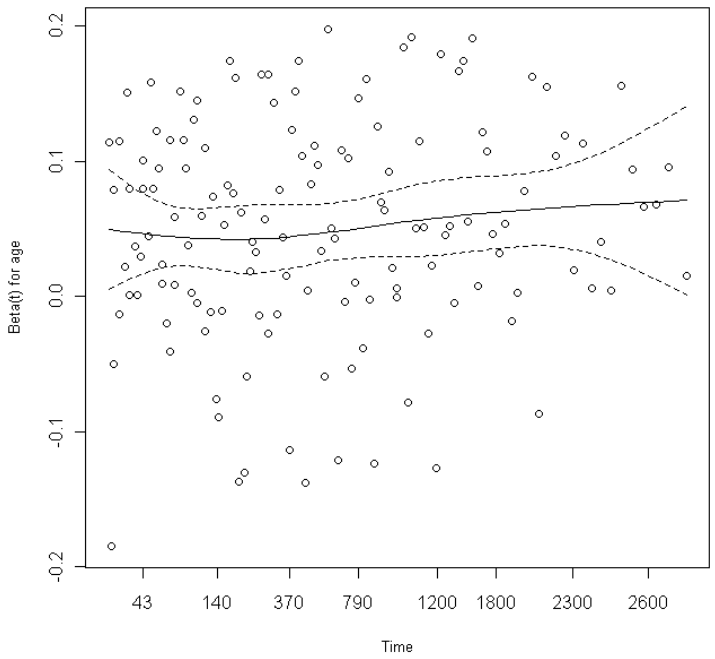


**Figure 8 Scaled *Schoenfeld* residuals**

Plots of scaled *Schoenfeld* residuals against transformed time for each covariate in a model fit to the recidivism data. The solid line is a smoothing-spline fit to the plot, with the broken lines representing a ±2-standard-error band around the fit. Systematic departures from a horizontal line are indicative of non-proportional hazards. Thus, **the assumption of proportional hazards appears to be supported for the covariates *age*** since its departures from a horizontal line are not systematic.

## 6.2 Check influential observations

Specifying the argument *type=dfbeta* to residuals produces a matrix of estimated changes in the regression coefficients upon deleting each observation in turn[5].

```
> phmodel.a<-coxph(Surv(time,status)~age, data=a)

# Cox regression index plots of dfbetas to identify influential observations
# The following R code can be generated from 'Cox Influence Plot' in the
  Package RcmdrPlugin.Survival and RcmdrPlugin.SurvivalT for R Commander
> ResidualsDFBeta <- residuals(phmodel.a, type='dfbeta')
> NumberCoeficients <- length(names(coef(phmodel.a)))
> NumberRows <- round(NumberCoeficients+0.1/2)
> if (NumberRows >= 2) {
+              par(mfrow=c(NumberRows,2))
+                        for(i in
+   1:NumberCoeficients) {
+             plot(ResidualsDFBeta[,i],
+   ylab=names(coef(phmodel.a))[i])
+             abline(h=0, lty=3)
+          }
+                } else {
+                        plot(ResidualsDFBeta, ylab=names(coef(phmodel.a))[1])
+     abline(h=0, lty=3)
+                }
```

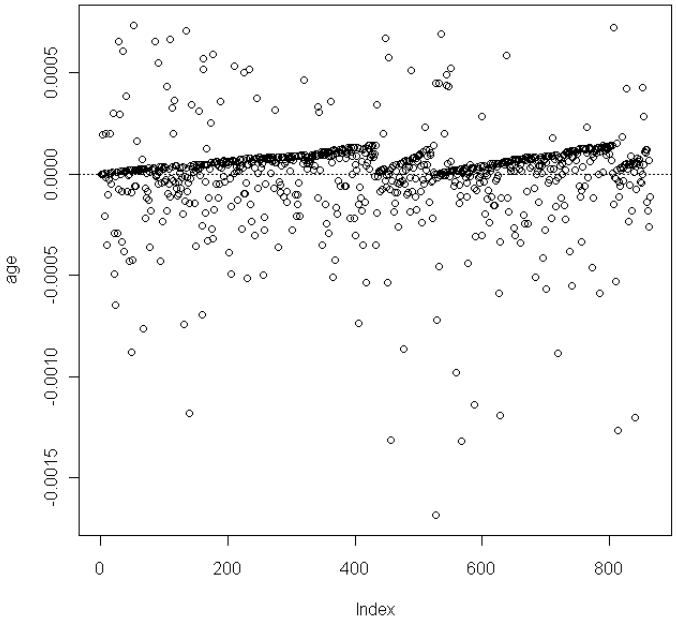


**Figure 9 Plots of *dfbeta* for influential observations**

Note: Index in figure 9 means the number of observations.

Comparing the magnitudes of the largest *dfbeta* values to the coefficient of *age* ( $\hat{\beta}_3$ *=0.051068*) suggests that none of the observations is terribly influential individually.

[5] Fox, J. 2002. Cox Proportional-Hazards Regression for Survival Data. http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf.

### 6.3 Model fit based upon graphical comparisons

Model fit can be evaluated based upon a graphical comparison between empirical K–M survival curves and fitted or "predicted" survival curves generated from the final PH model.

The reduced PH model in this case is:

$$\lambda_i(t;\hat{\alpha},\hat{\beta},\underline{x}) = \lambda_0(t,\hat{\alpha})\exp\{\underline{x}^T\hat{\beta}\}$$

$$= \exp\{\hat{\alpha}_0+\hat{\alpha}_2 t^2+\hat{\alpha}_3 t^3+\hat{\alpha}_4(t-\gamma)_+^3\}\exp\{x_{age}\hat{\beta}_3\},$$

where $\hat{\beta}_3$*=0.051068,* $\hat{\alpha}_0$*=-8.926e+00,* $\hat{\alpha}_2$*=7.362e-07,* $\hat{\alpha}_3$*=-2.514e-10,* $\hat{\alpha}_4$*=5.679e-10.*

We know that the cumulative hazard function is: $H(t)=\int_0^t \lambda(u)du$. In this case, under a discrete time framework $H(t_k)=\sum_{j=1}^{k}\lambda(t_k)$, then the survival function is given be:

$S(t_k)=\exp\{-H(t_k)\}$.

#### 6.3.1 Examine the estimated distribution of survival times

Having fit a Cox PH model to the data, it is often of interest to examine the estimated distribution of survival times. The *survfit()* function estimates *S(t)*, by default <u>at the mean values of the covariates</u>.

```
> fit.km=survfit(Surv(time,status)~1,data=a)
> phmodel.a<-coxph(Surv(time,status)~age, data=a)
> plot(fit.km, main="Compare the estimated distribution of survival times
with the general K-M estimate", ylab='K-M estimate & survival probability',
xlab='Time (days)', cex.main=0.8, cex.lab=0.7, cex.axis=0.7, lty=2, col=
c("red","blue"))
> lines(survfit(phmodel.a, conf.type="none"), lty=1, col= c("blue"))
> legend(200, 0.1, c("the general K-M estimate", "the estimated
distribution"), lty=c(2,1), col= c("red","blue"), cex=.7)
```

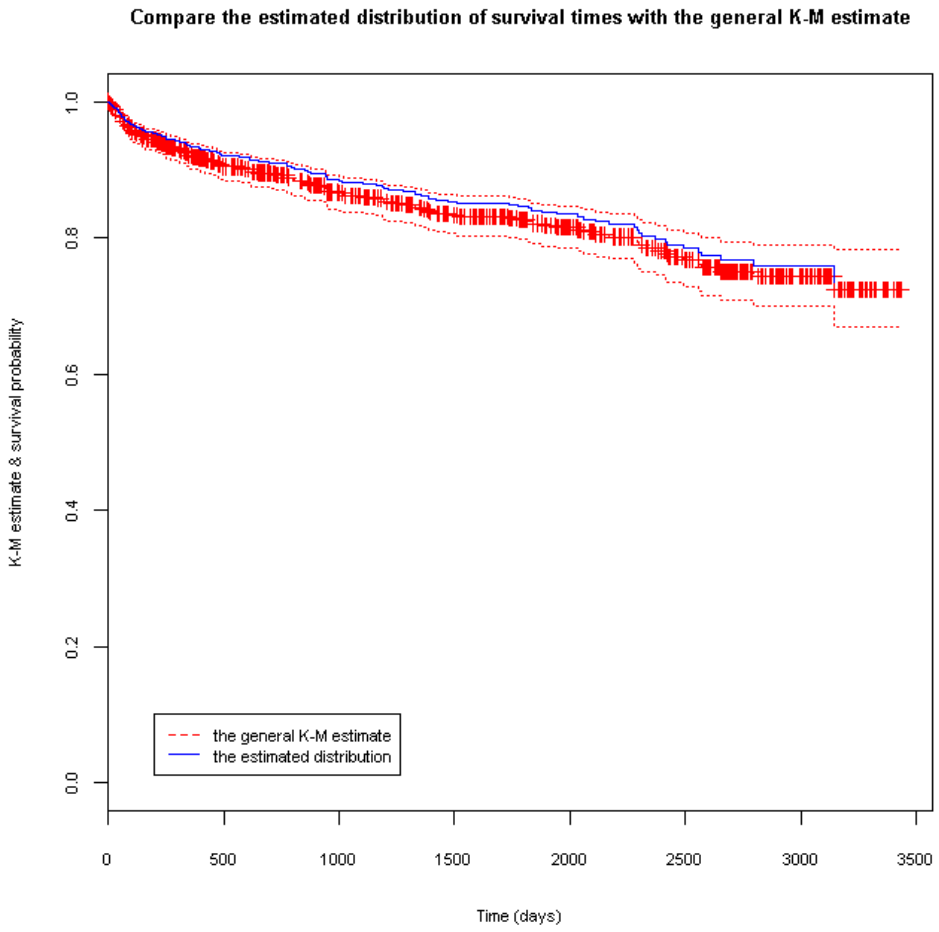


**Figure 10 Compare the estimated distribution of survival times with the general K-M estimate**

From figure 10, the estimated distribution of survival times generated by the reduced PH model **match**es the general K-M estimate and falls into its 95% confidence interval.

### 6.3.2 Check $\hat{\alpha}$ given $\hat{\beta}$

First, we can take a look at the trend of the baseline hazard rate $\lambda_0(t,\hat{\alpha})$, and compare it with $\lambda_0(t)$.

```
# Plot λ0(t,α^)
> dreduced.a <- read.table("M:/dreduced.a.txt", header=TRUE)
> attach(dreduced.a)
> alpha0 <--8.926e+00
> alpha2 <-7.362e-07
> alpha3 <--2.514e-10
> alpha4 <-5.679e-10
> hrate0 <- exp(alpha0+alpha2*haz.time2+alpha3*haz.time3+alpha4*new.time3)
```

```
> plot(hrate0, xlab="Time (days) ", ylab="Baseline hazard rate", main="
Baseline hazard rate of specified alpha's", cex.main=1, cex.lab=1,
cex.axis=1)

# Plot λ0(t) --- this process is almost the same as that in 4.2.2
> d.phmodel.a =coxph.detail(phmodel.a)
> hazm.a=c(d.phmodel.a$hazard)
> detach(dreduced.a)
> attach(a)
> meanx.a=c(mean(age))
> beta.a=phmodel.a$coef
> ex.a=exp(t(meanx.a)%*%beta.a)
> d.phmodel.a =coxph.detail(phmodel.a)
> hazm.a=c(d.phmodel.a$hazard)
> haz0.a=hazm.a/ex.a
> plot(haz0.a, xlab="Time (days) ", ylab="Baseline hazard rate", main="
Baseline hazard rate of unspecified alpha's", cex.main=1, cex.lab=1,
cex.axis=1)
```

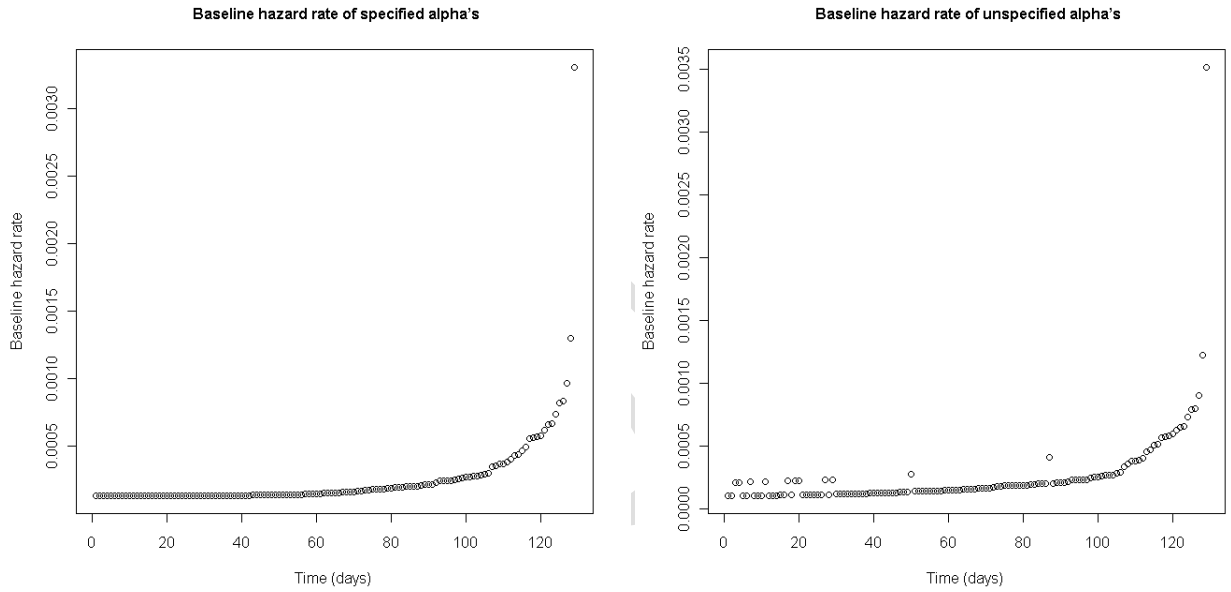


**Figure 11 Compare $\lambda_0(t,\hat{\alpha})$ with $\lambda_0(t)$**

The plots of $\lambda_0(t,\hat{\alpha})$ are very close to those of $\lambda_0(t)$.

Then, let us consider 10 patients with the following covariate combinations:

- $X_1$ = *(age = 0)*
- $X_2$ = *(age = 5)*
- $X_3$ = *(age = 10)*
- $X_4$ = *(age = 20)*
- $X_5$ = *(age = 30)*
- $X_6$ = *(age = 40)*
- $X_7$ = *(age = 50)*
- $X_8$ = *(age = 60)*

- $X_9$ = *(age = 70)*
- $X_{10}$ = *(age = 80)*

```
# Import the time data for α̂ that was created previously
> dreduced.a <- read.table("M:/dreduced.a.txt", header=TRUE)
# dreduced.a includes t, t^2, t^3, and (t-γ)^3_+ for the hazard function
> dreduced.a
    loghaz0.a haz.time haz.time2   haz.time3  new.time3
1   -9.167297        2         4           8          0
2   -9.165059        3         9          27          0
3   -8.468773        7        49         343          0
.      .             .         .           .          .
.      .             .         .           .          .
.      .             .         .           .          .
.      .             .         .           .          .
.      .             .         .           .          .
.      .             .         .           .          .
127 -7.009067     2650   7022500 18609625000 2633789341
128 -6.707579     2795   7812025 21834609875 3553559576
129 -5.650654     3146   9897316 31136956136 6612913133
> attach(dreduced.a)

# Give values to the variable age
> x1<-0
> x2<-5
> x3<-10
> x4<-20
> x5<-30
> x6<-40
> x7<-50
> x8<-60
> x9<-70
> x10<-80

# Give values to α and β in terms of α̂ and β̂ for the reduced model
> beta3 <-0.051068
> alpha0 <--8.926e+00
> alpha2 <-7.362e-07
> alpha3 <--2.514e-10
> alpha4 <-5.679e-10

# Work out survival probabilities for different values of age
> hx1 <- exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x1*beta3)
> chx1 <- cumsum(hx1)
> sur.px1<- exp(-chx1)
> sur.px2<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3*  haz.time3+alpha4*
new.time3)*exp(x2*beta3)))
> sur.px3<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x3*beta3)))
> sur.px4<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x4*beta3)))
> sur.px5<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x5*beta3)))
> sur.px6<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x6*beta3)))
> sur.px7<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x7*beta3)))
> sur.px8<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x8*beta3)))
> sur.px9<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x9*beta3)))
```

```
> sur.px10<- exp(-cumsum(exp(alpha0+alpha2* haz.time2+alpha3* haz.time3+alpha4*
new.time3)*exp(x10*beta3)))
> plot(haz.time,sur.px1,type="l",main="Check the reduced PH model fit with
specified alpha's for the baseline hazard",xlab="Time (days)",
ylab="Survival probability", ylim=c(0,1),lty= c(1:10), cex.main=0.8,
cex.lab=0.7, cex.axis=0.7, col=c(1:10))
> lines(haz.time,sur.px2,lty=2,col=2)
> lines(haz.time,sur.px3,lty=3,col=3)
> lines(haz.time,sur.px4,lty=4,col=4)
> lines(haz.time,sur.px5,lty=5,col=5)
> lines(haz.time,sur.px6,lty=6,col=6)
> lines(haz.time,sur.px7,lty=7,col=7)
> lines(haz.time,sur.px8,lty=8,col=8)
> lines(haz.time,sur.px9,lty=9,col=9)
> lines(haz.time,sur.px10,lty=10,col=10)
> legend(2, 0.38, c("subject X1: age=0", "subject X2: age=5", "subject X3:
age=10", "subject X4: age=20", "subject X5: age=30", "subject X6: age=40",
"subject X7: age=50", "subject X8: age=60", "subject X9: age=70", "subject
X10: age=80"), lty=c(1:10), col= c(1:10), cex=.7)
```

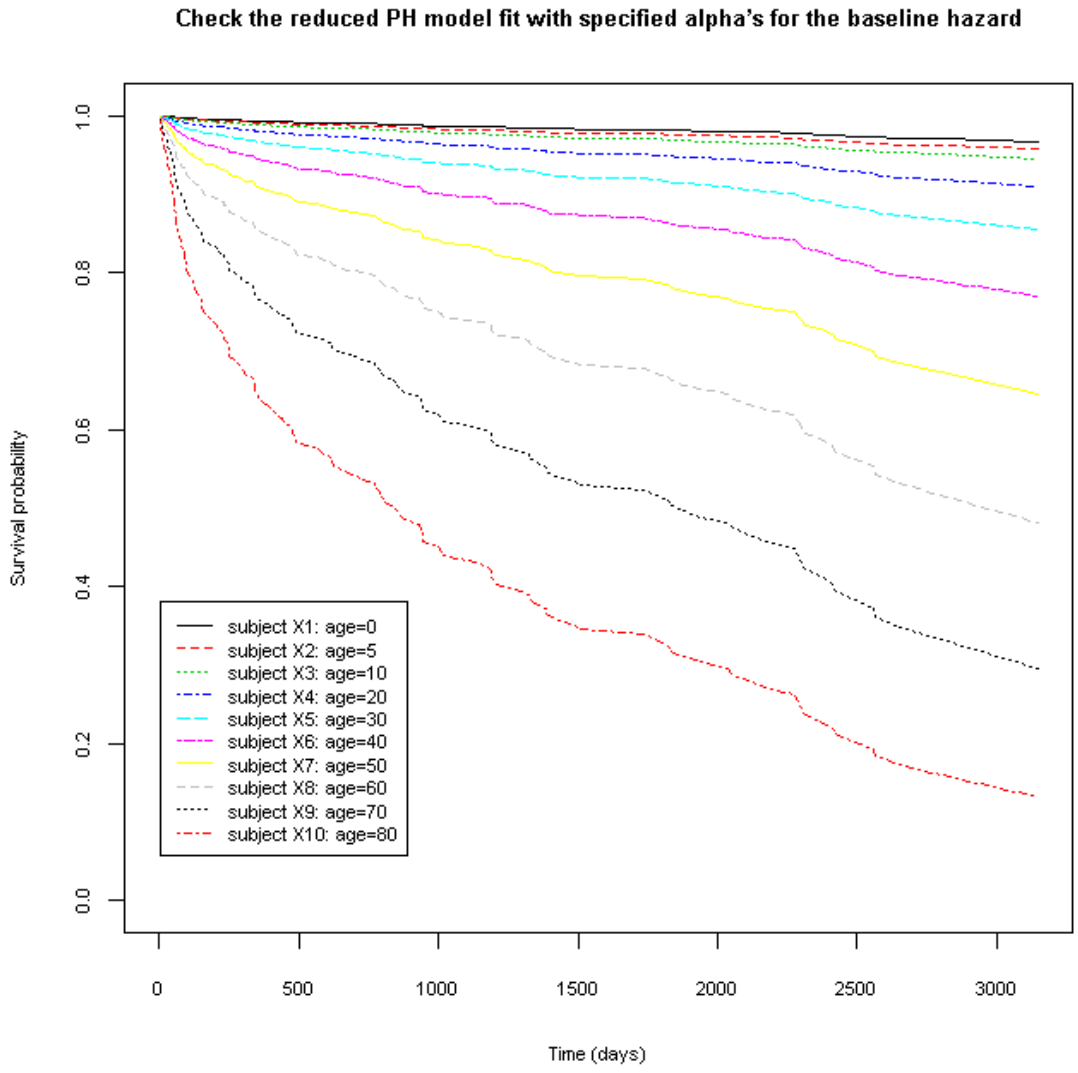


**Figure 12 Check the reduced PH model fit with *specified* alpha's for the baseline hazard**

A graphical comparison between figure 5 and figure 12 shows that:

- Figure 12 is very close to figure 5.
- Given $\hat{\beta}_3 = 0.051068$, $\hat{\alpha}_0$=-8.926e+00, $\hat{\alpha}_2$=7.362e-07, $\hat{\alpha}_3$=-2.514e-10, $\hat{\alpha}_4$ =5.679e-10 are good to estimate the cumulative hazard function and also the survival probabilities.
- We see that the increase in age is associated with decreased survival.
- The survival probability for an individual who took kidney transplant decreases more quickly as age increases.

### 6.3.3 Compare the final PH model with the K–M survival curves

We compare the reduced PH model with the general K–M survival estimate:

```
> fit.km=survfit(Surv(time,status)~1,data=a,conf.type="none")
> plot(fit.km, main="Compare the reduced PH model with the general K-M
estimate", ylab='K-M estimate & survival probability', xlab='Time (days)',
cex.main=0.8,cex.lab=0.7, cex.axis=0.7, lty=c(1:11), col= c(1:11))
> lines(haz.time,sur.px1,lty=2,col=2)
> lines(haz.time,sur.px2,lty=3,col=3)
> lines(haz.time,sur.px3,lty=4,col=4)
> lines(haz.time,sur.px4,lty=5,col=5)
> lines(haz.time,sur.px5,lty=6,col=6)
> lines(haz.time,sur.px6,lty=7,col=7)
> lines(haz.time,sur.px7,lty=8,col=8)
> lines(haz.time,sur.px8,lty=9,col=9)
> lines(haz.time,sur.px9,lty=10,col=10)
> lines(haz.time,sur.px10,lty=11,col=11)
> legend(100, 0.33, c("K-M estimate", "Subject X1: age=0", "Subject X2:
age=5", "Subject X3: age=10", "Subject X4: age=20", "Subject X5: age=30",
"Subject X6: age=40", "Subject X7: age=50", "Subject X8: age=60", "Subject
X9: age=70", "Subject X10: age=80"), lty=c(1:11), col= c(1:11), cex=.7)
```

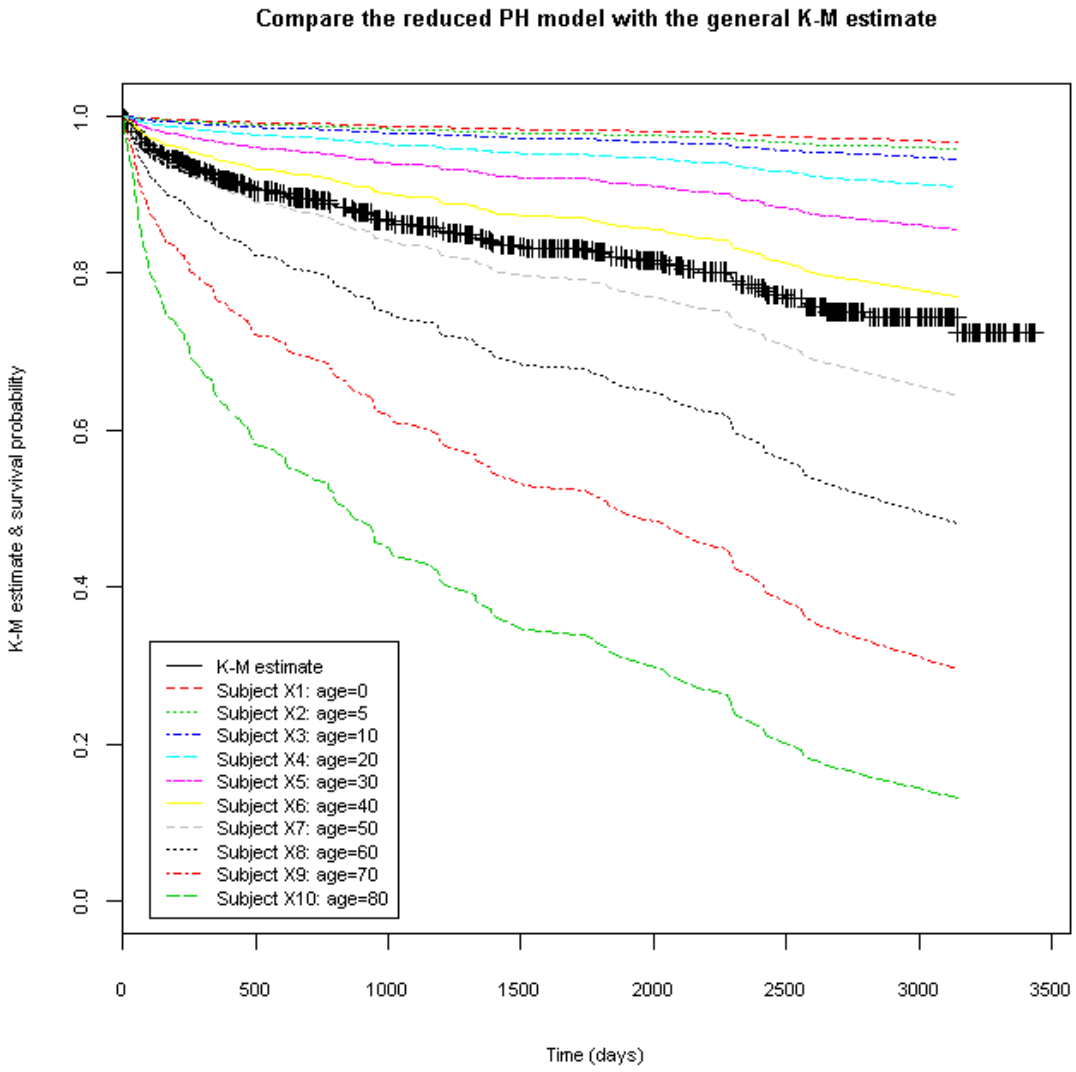


**Figure 13 Compare the reduced PH model with the general K-M estimate**

From figure 13, we know that the general K-M estimate falls into the interval of the survival curves of *age=40* and *age=50* of the final reduced PH model.

In terms of the histogram of *age* in this case, we see the frequency about *age*:

```
> Hist(a$age, scale="frequency", breaks="Sturges", col="darkgray", xlab='Age
(years)', ylab='Frequency ', cex.main=0.8, cex.lab=0.9, cex.axis=0.9)
```

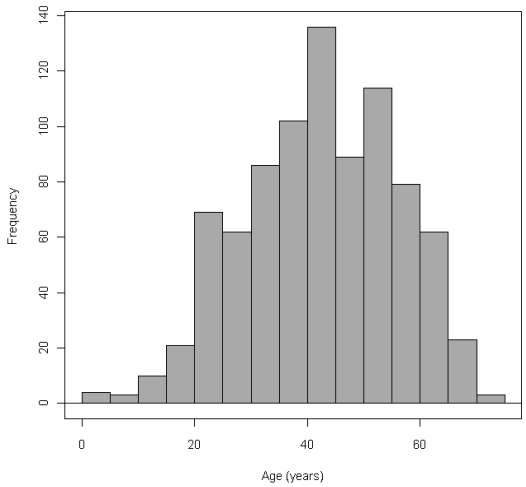


**Figure 14 Histogram of *age***

We can also compare it with the **categorized** K–M survival curves.

For example, let us categorize the variable *age*, and set a new variable *age1* where the patients with *age1 (age >= 43) = 1* and *age1 (age < 43) = 0* (since the median is *43*).

```
> attach(a)
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   33.00   43.00   42.84   54.00   75.00
> age1<-age>=43
> length(age1)
[1] 863
> for(i in 1:length(age1)) {if (age1[i]!=FALSE) age1[i]=1 else age1[i]=0}
> age1
> age1
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 0 1 1
 [38] 0 1 1 0 1 0 1 0 1 0 0 1 0 0 1 1 0 1 0 1 1 0 0 1 0 1 0 1 0 0 0 0 1 1 1 1 1
 [75] 1 0 1 1 1 0 0 0 1 0 1 1 1 1 1 1 1 0 1 0 0 1 0 1 1 0 1 1 0 1 1 1 1 1 1 1 1
[112] 0 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 0 1 1 1 1 0 1 0 1 1 1 1 1 0 1
[149] 1 0 0 0 1 0 1 1 1 0 0 1 1 1 0 1 1 0 1 1 0 1 1 1 0 1 0 0 1 1 0 0 0 1 0 0 1
[186] 1 1 0 1 1 0 0 1 0 1 0 0 0 0 1 0 1 0 1 0 1 0 0 0 1 1 0 1 1 1 0 1 0 1 1 0 1
[223] 0 1 1 1 1 1 0 1 1 0 1 0 0 1 1 0 0 0 0 1 1 0 0 1 1 1 0 0 0 0 1 1 0 1 1 1 1
[260] 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 1 1 1 1 1 0 0 1 0 0 1 0 1 1 0 0 0
[297] 0 1 0 0 1 0 0 0 0 0 1 0 1 1 1 1 0 1 0 0 0 0 1 0 1 0 0 0 1 1 0 0 1 0 0 0 0 0
[334] 0 1 0 1 0 0 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 1 0 0 1 0 1 0 0
[371] 1 1 0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 1 1 0 1 0 1 1 1
[408] 0 1 1 1 1 0 1 0 0 0 1 1 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 1 1 1 0 1 1 1 1 0 0 1 1
[445] 0 0 1 1 1 1 1 0 1 1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 0 1 1 0 0 1 1 0 0
[482] 1 1 1 0 1 1 1 0 1 0 1 0 1 1 0 1 0 0 0 0 0 1 1 0 1 0 1 1 1 1 0 1 1 0 1 1 0 0
[519] 0 1 1 1 1 1 0 0 1 0 0 0 1 1 0 0 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 1 1 0 1 1 0 0 0 0
[556] 1 1 0 0 1 1 0 1 1 1 1 0 0 1 1 0 1 0 0 1 1 1 1 1 0 1 1 1 1 0 0 0 0 0 0 1 1 1 1 0
[593] 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 0 1 1 1 0 1 0 0 0 0 0
[630] 1 0 0 0 0 0 0 0 0 1 1 0 1 0 1 1 1 1 0 1 0 1 1 1 0 0 0 0 0 0 0 1 0 1 0 1 0 0 1 0
[667] 0 0 0 1 0 1 0 0 1 0 0 0 0 1 0 1 0 0 1 1 0 1 1 1 1 0 0 1 1 0 0 0 0 0 0 1 0 0 0
[704] 1 0 1 1 0 1 1 0 1 0 0 1 1 0 0 1 0 0 1 1 0 1 1 0 0 0 1 1 0 0 0 0 0 1 0 0 0 1
[741] 0 0 0 0 0 0 1 1 1 0 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
[778] 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 0 1 0 1 0 1
[815] 1 0 1 1 1 0 1 1 0 0 0 0 1 1 0 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 0 1 1 1 1
[852] 1 1 1 0 0 1 1 0 1 1 1 1
> a.a1<-cbind(time,status,gender,race,age,age1)
> a.age1<-data.frame(a.a1)
> detach(a)
> attach(dreduced.a)
> fit.km.bya1=survfit(Surv(time,status)~age1,data=a.age1,conf.type="none")
```

```
> plot(fit.km.bya1, main="Compare the reduced PH model compare with the
categorized K-M survival curves", ylab='K-M estimate & survival probability',
xlab='Time (days)', cex.main=0.8,cex.lab=0.7, cex.axis=0.7, lty=c(1:12), col=
c(1:12))
> lines(haz.time,sur.px1,lty=3,col=3)
> lines(haz.time,sur.px2,lty=4,col=4)
> lines(haz.time,sur.px3,lty=5,col=5)
> lines(haz.time,sur.px4,lty=6,col=6)
> lines(haz.time,sur.px5,lty=7,col=7)
> lines(haz.time,sur.px6,lty=8,col=8)
> lines(haz.time,sur.px7,lty=9,col=9)
> lines(haz.time,sur.px8,lty=10,col=10)
> lines(haz.time,sur.px9,lty=11,col=11)
> lines(haz.time,sur.px10,lty=12,col=12)
> legend(100, 0.4, c("K-M for age<43", "K-M for age>=43", "Subject X1: age=0",
"Subject X2: age=5", "Subject X3: age=10", "Subject X4: age=20", "Subject X5:
age=30", "Subject X6: age=40", "Subject X7: age=50", "Subject X8: age=60",
"Subject X9: age=70", "Subject X10: age=80"), lty=c(1:12), col= c(1:12), cex=.7)
```

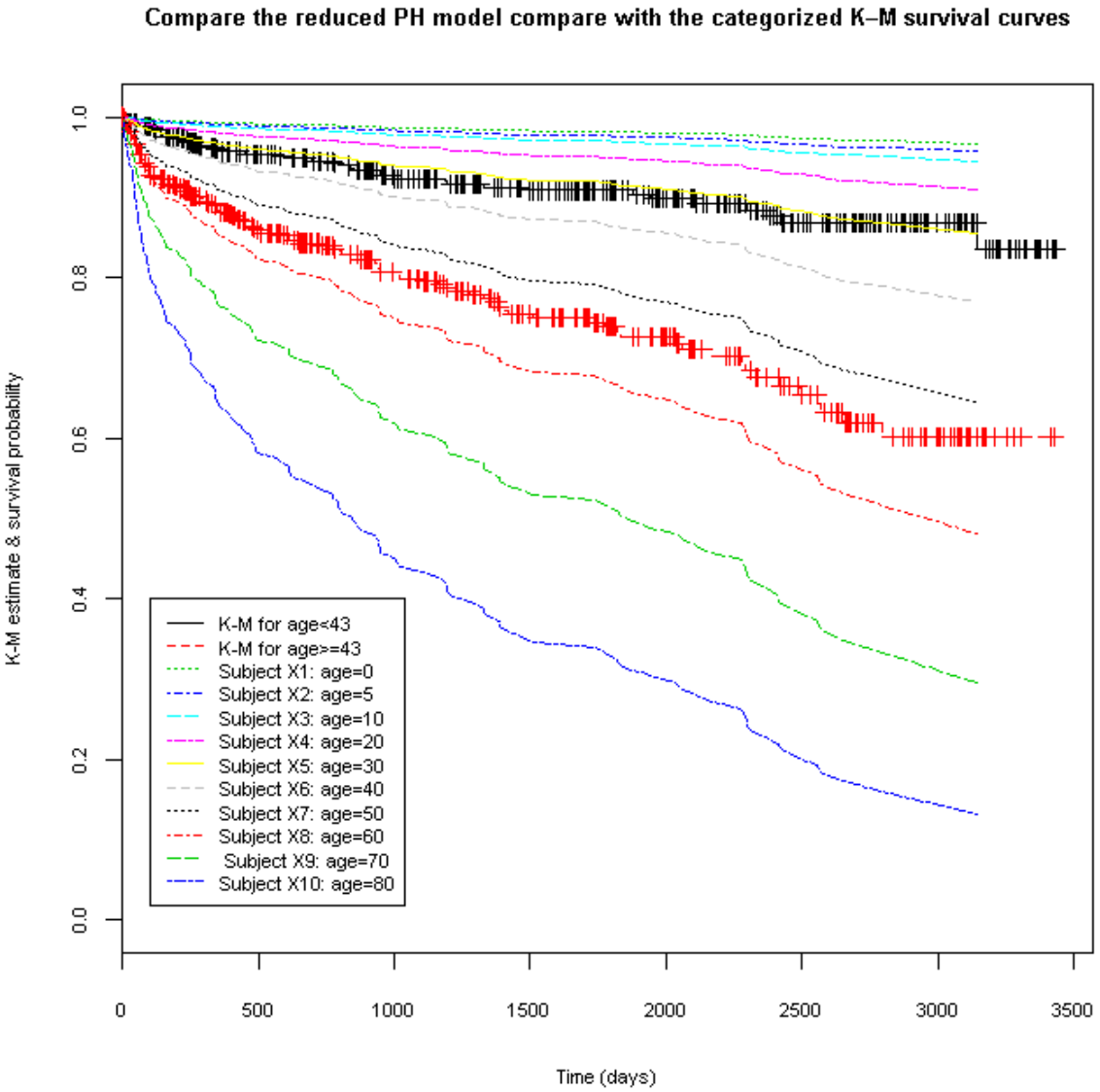


**Figure 15 Compare the reduced PH model compare with the categorized K–M survival curves**

From figure 15, we see that the K-M estimate for (*age* < *43*) **match**es the survival curves of *age=30* of the reduced PH model, and the K-M estimate for (*age* >= *43*) falls into the interval of the survival curves of *age=50* and *age=60* of the reduced PH model.

Moreover, we can use the distinguishably categorized K-M curves to verify the reduced PH model. We create a new variable *age2* to categorize patients' *age* with eight groups as: *age2=1* if *age* between 0 to 10; *age2=2* if *age* between 10 to 20; *age2=3* if *age* between 20 to 30; *age2=4* if *age* between 30 to 40; *age2=5* if *age* between 40 to 50; *age2=6* if *age* between 50 to 60; *age2=7* if *age* between 60 to 70; *age2=8* if *age* between 70 to 80.

```
> detach(dreduced.a)
> attach(a)
> age2<-age
> for(i in 1:length(age2)) if (age2[i]<= 10) {age2[i] =1} else {if (age2[i]<= 20)
{age2[i] =2} else {if (age2[i]<= 30) {age2[i] =3} else {if (age2[i]<= 40) {age2[i]
=4} else {if (age2[i]<= 50) {age2[i] =5} else {if (age2[i]<= 60) {age2[i] =6}
else {if (age2[i]<=70) {age2[i] =7}else age2[i]=8}}}}}}
> age2
  [1] 5 6 6 6 5 5 5 7 6 5 5 6 4 7 6 5 5 4 3 6 5 5 4 6 5 5 5 7 7 6 6 4 5 5 3 7 5
 [38] 2 5 6 4 7 4 6 5 6 3 4 7 5 5 7 7 2 7 5 6 5 5 4 5 5 5 3 6 4 4 4 3 5 5 5 5 7
 [75] 7 4 8 5 7 4 5 4 5 4 7 7 6 6 6 6 7 3 5 5 3 5 5 5 5 4 6 6 5 6 7 7 6 5 7 7 6
[112] 3 6 6 6 6 6 3 3 6 5 6 6 5 5 6 6 5 4 7 4 6 7 2 7 6 6 5 3 7 2 6 6 5 5 5 3 7
[149] 5 5 5 4 5 3 6 6 5 2 4 5 7 7 2 7 6 4 6 7 4 6 5 5 4 6 5 5 7 7 4 2 3 6 3 5 6
[186] 6 5 4 6 5 4 5 6 4 6 4 4 3 3 5 4 7 4 6 4 5 4 4 3 7 5 4 5 5 6 4 5 4 5 5 4 7
[223] 4 5 7 6 6 6 3 7 5 1 7 3 4 6 7 5 5 4 4 7 5 4 3 6 5 6 3 4 3 4 5 5 4 7 6 7 6
[260] 3 4 4 4 4 4 4 5 4 5 4 4 6 4 6 5 5 4 3 3 7 6 6 6 5 4 3 6 4 4 5 5 6 5 4 4 2
[297] 2 5 4 4 7 4 5 3 4 6 4 6 6 6 6 3 6 3 3 3 3 5 4 7 4 5 4 5 5 5 4 5 2 5 4 2 4
[334] 4 5 3 5 2 3 4 6 6 2 6 5 3 4 3 7 4 4 4 4 6 4 4 4 6 4 4 2 7 5 4 2 7 4 7 4 4
[371] 6 6 4 5 1 3 6 4 4 4 3 2 4 6 6 5 4 4 3 4 4 6 3 4 5 5 3 3 6 6 5 5 5 3 7 6 6
[408] 4 6 5 6 4 6 3 4 4 5 7 5 4 3 5 3 5 3 5 1 3 5 4 5 4 5 6 5 6 3 5 7 5 4 3 7 6
[445] 3 4 7 6 6 6 4 6 7 5 6 3 6 5 4 3 7 6 7 5 3 6 6 7 6 6 6 4 3 5 6 3 4 6 5 4 4
[482] 6 6 5 4 5 6 7 3 6 4 5 3 6 6 3 6 3 5 3 3 4 5 6 4 5 3 6 5 6 4 5 5 5 5 6 3 4
[519] 1 5 6 5 5 3 5 6 2 5 4 8 6 5 3 6 7 4 6 6 3 3 6 6 7 6 5 6 6 6 3 7 6 4 5 4 4
[556] 7 6 4 4 6 6 2 6 7 7 4 3 5 5 5 5 4 4 6 6 7 6 5 5 6 6 3 4 5 5 5 3 6 8 6 5 3
[593] 5 6 6 6 6 4 5 6 5 6 3 7 5 5 7 6 6 5 7 5 4 5 6 3 7 4 4 7 6 6 5 6 3 4 5 3 2
[630] 6 3 3 4 3 5 3 4 7 6 3 7 4 5 5 5 4 5 3 7 5 7 4 3 4 2 4 3 5 4 5 3 5 4 4 6 3
[667] 5 4 3 6 3 7 4 4 5 4 4 4 5 3 6 4 3 7 5 4 5 6 5 5 4 4 5 5 4 4 5 3 4 7 1 4 3
[704] 6 3 6 5 4 6 5 4 5 2 3 6 6 3 3 5 2 5 5 6 5 5 5 3 3 5 6 5 3 3 3 2 1 7 3 4 7
[741] 1 4 5 3 3 5 5 5 4 4 3 3 6 6 6 6 4 5 3 3 6 3 3 6 2 5 3 3 5 3 4 7 4 5 6 2 3
[778] 3 4 4 3 4 5 7 4 3 4 3 5 3 2 4 3 5 4 4 2 2 4 3 3 3 5 3 4 5 7 6 4 6 4 6 3 7
[815] 6 3 6 5 6 4 6 5 4 4 4 4 6 5 5 7 5 6 4 5 5 5 5 5 2 3 6 5 5 3 5 5 4 6 5 6 7
[852] 5 6 5 3 2 5 6 2 6 6 5 6
> a.a2<-cbind(time,status,gender,race,age,age2)
> a.age2<-data.frame(a.a2)
> fit.km.bya2=survfit(Surv(time,status)~age2,data=a.age2,conf.type="none")
> plot(fit.km.bya2, main="The detailed categorized K-M survival curves", ylab='K-
M estimate', xlab='Time (days)', cex.main=0.8,cex.lab=0.7, cex.axis=0.7,
lty=c(1:8), col= c(1:8))
> legend(230, 0.38,lty=1:8, cex=.7,col=c(1:8),c("age: 0-10","age: 10-20", "age:
20-30", "age: 30-40", "age: 40-50", "age: 50-60", "age: 60-70", "age: 70-80"))
```

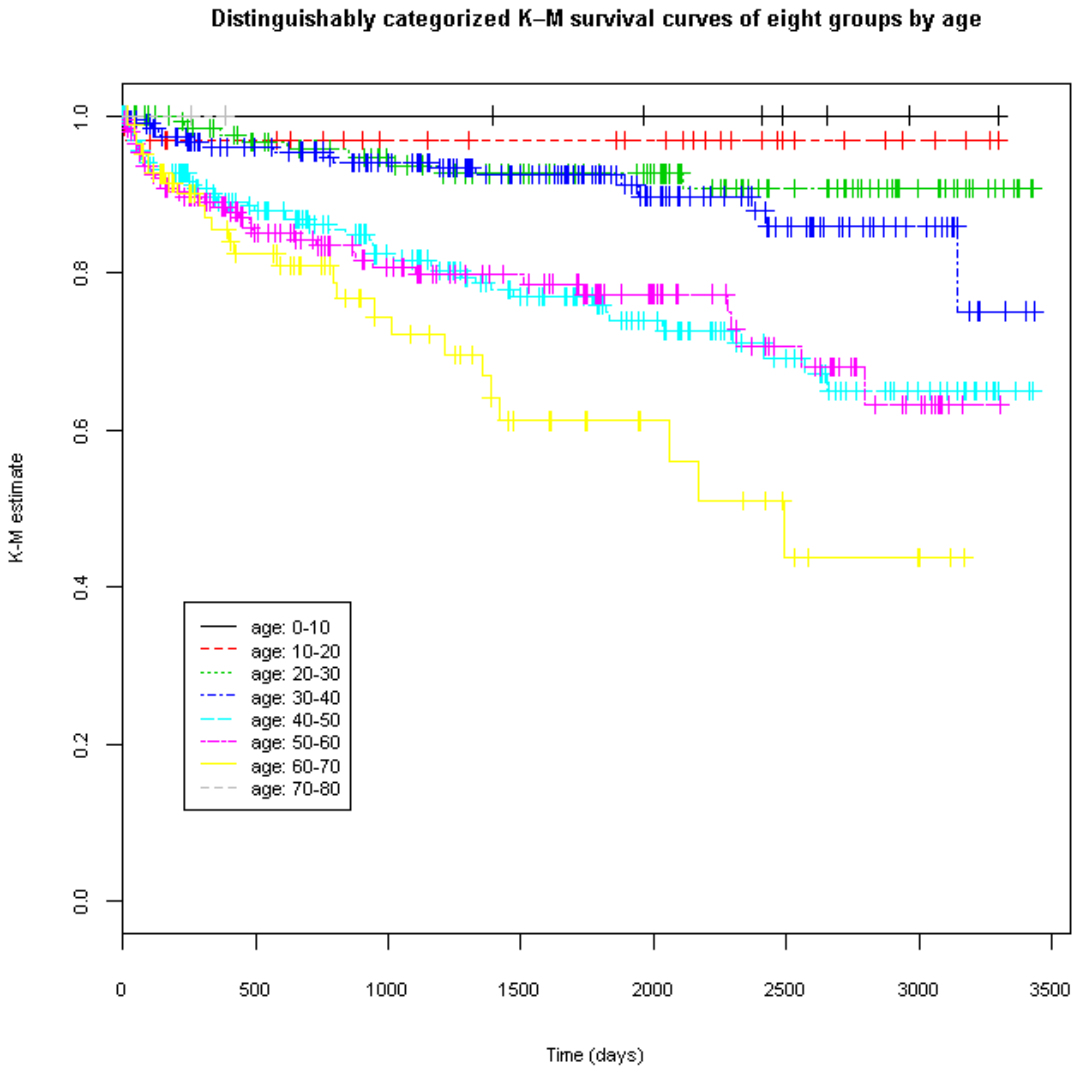


**Figure 16 Distinguishably categorized K–M survival curves of eight groups by age**

Comparing the distinguishably categorized K-M curves in figure 16 with the curves generated by the final reduced PH model in figure 12, we observe that: the final reduced PH model can not well estimate the groups with age 40 to 50, 50 to 60, and 70 to 80 years old.

### 6.3.4 Model fit under a simulated time framework

The reduced PH model in this case is:

$$\lambda_i(t;\hat{\alpha},\hat{\beta},\underline{x}) = \lambda_0(t,\hat{\alpha})\exp\{\underline{x}^T\hat{\beta}\}$$

$$= \exp\{\hat{\alpha}_0+\hat{\alpha}_2 t^2+\hat{\alpha}_3 t^3+\hat{\alpha}_4(t-\gamma)_+^3\}\exp\{x_{age}\hat{\beta}_3\},$$

where $\hat{\beta}_3$ *=0.051068,* $\hat{\alpha}_0$ *=-8.926e+00,* $\hat{\alpha}_2$ *=7.362e-07,* $\hat{\alpha}_3$ *=-2.514e-10,* $\hat{\alpha}_4$ *=5.679e-10.*

Thus, we may use this function to depict some trends under a simulated time framework. For example, suppose the *TIME* variable contains each day, consider 10 patients with the following covariate combinations:

- $Y_1$ *= (age = 0)*
- $Y_2$ *= (age = 5)*
- $Y_3$ *= (age = 10)*
- $Y_4$ *= (age = 20)*
- $Y_5$ *= (age = 30)*
- $Y_6$ *= (age = 40)*
- $Y_7$ *= (age = 50)*
- $Y_8$ *= (age = 60)*
- $Y_9$ *= (age = 70)*
- $Y_{10}$ *= (age = 80)*

```
> y1<-0
> y2<-5
> y3<-10
> y4<-20
> y5<-30
> y6<-40
> y7<-50
> y8<-60
> y9<-70
> y10<-80
> t <- seq(0,3434,1)
> t2 <- t^2
> t3 <- t^3
> new.t <- t
> for(i in 1:length(new.t)) {if (new.t[i]-1269<0) new.t[i]=0 else new.t[i]=
new.t[i]- 1269}
> new.t3<- new.t^3
> beta3 <-0.051068
> alpha0 <--8.926e+00
> alpha2 <-7.362e-07
> alpha3 <--2.514e-10
> alpha4 <-5.679e-10
> hy1 <- exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y1*beta3)
> chy1 <- cumsum(hy1)
> sur.py1<- exp(-chy1)
> sur.py2<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y2*beta3)))
> sur.py3<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y3*beta3)))
> sur.py4<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y4*beta3)))
> sur.py5<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y5*beta3)))
> sur.py6<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y6*beta3)))
> sur.py7<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y7*beta3)))
> sur.py8<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y8*beta3)))
> sur.py9<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y9*beta3)))
> sur.py10<- exp(-cumsum(exp(alpha0+alpha2*t2+alpha3*t3+alpha4*new.t3)*exp(y10*beta3)))
> plot(t,sur.py1,type="l",main="Check the reduced PH model with the
specified baseline hazard under a simulated time framework",xlab="Time
```

(i) We can change this value to simulate different time frames with equal time intervals, e.g. 1 day for this example.

(ii) Or we can write R code to create a new life-like time data set to simulate different time frames.

```
(days)", ylab="Survival probability", ylim=c(0,1),lty= c(1:10), cex.main=0.8,
cex.lab=0.7, cex.axis=0.7, col=c(1:10))
> lines(t,sur.py2,lty=2,col=2)
> lines(t,sur.py3,lty=3,col=3)
> lines(t,sur.py4,lty=4,col=4)
> lines(t,sur.py5,lty=5,col=5)
> lines(t,sur.py6,lty=6,col=6)
> lines(t,sur.py7,lty=7,col=7)
> lines(t,sur.py8,lty=8,col=8)
> lines(t,sur.py9,lty=9,col=9)
> lines(t,sur.py10,lty=10,col=10)
> legend(2500, 0.9, c("Subject Y1: age=0", "Subject Y2: age=5", "Subject Y3:
age=10", "Subject Y4: age=20", "Subject Y5: age=30", "Subject Y6: age=40",
"Subject Y7: age=50", "Subject Y8: age=60", "Subject Y9: age=70", "Subject
Y10: age=80"), lty=c(1:10), col= c(1:10), cex=.7)
```

**Check the reduced PH model with the specified baseline hazard under a simulated time framework**

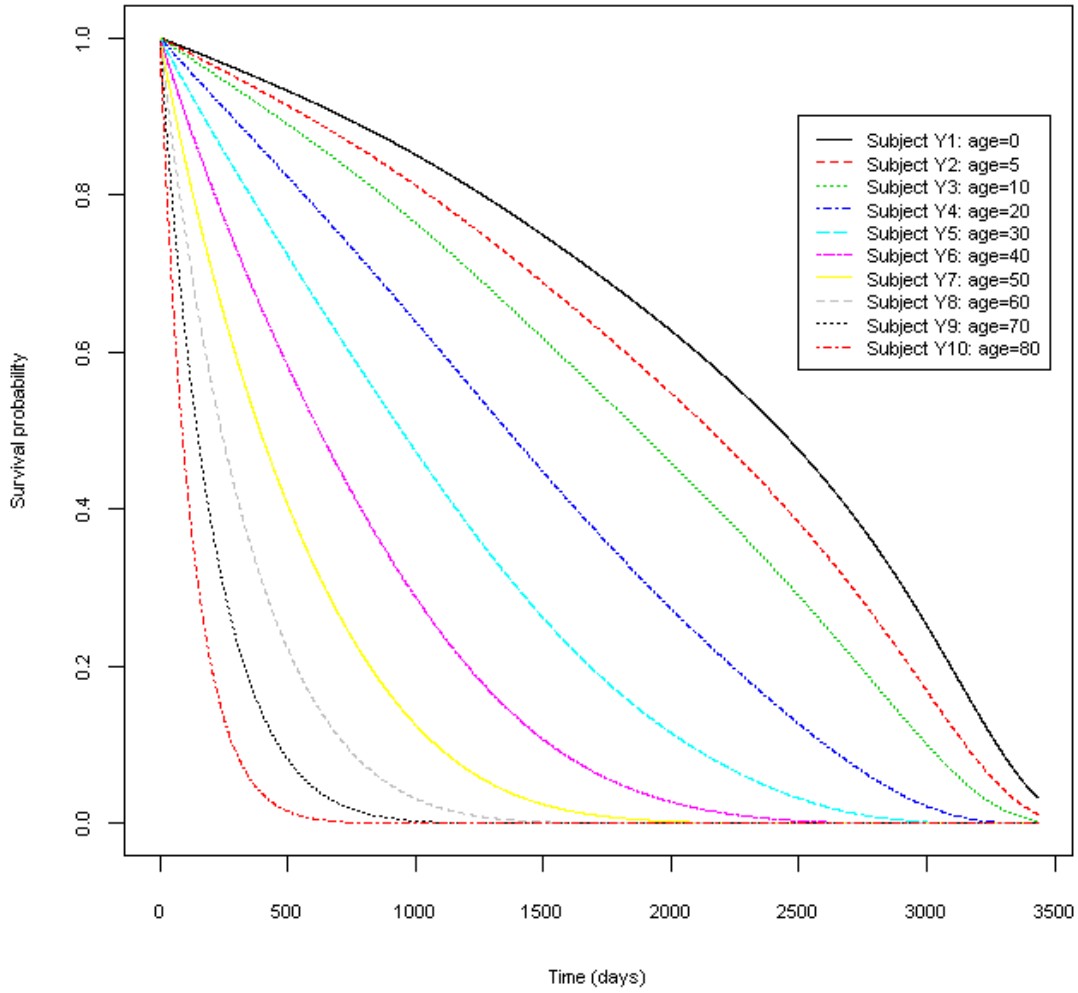


**Figure 17 Check the reduced PH model with the specified baseline hazard under a simulated time framework**

We see that the increase in age is associated with decreased survival, and the older the worse.