Survival Analysis on the Primary Biliary Cirrhosis

Lucas Lassus, Mélanie Maharjan and Prashanth Pushparasah
Data ScienceTech Institute Spring-18

November 2018

Contents

1 Introduction

Multiple tools and data for survival analysis are available in R packages such as "survival" from where we picked the PBC dataset. It comes from a clinical trial in the field of primary biliary cirrhosis conducted at the Mayo Clinic between 1974 and 1984. Primary biliary cirrhosis is a fatal chronic liver disease.

A total of 418 PBC patients were randomized to either a placebo or a drug called D-penicillamine. Each of them was followed until death or censoring (the duration is measured in days). The status at endpoint is coded as follows: 0/1/2 for censored, transplant and dead respectively. In addition, 17 covariates are recorded for this study. These include a treatment variable, patient age, gender and clinical, biochemical and histologic measurements made at the time of randomization.

After preparing the data for in depth survival analysis (in part 2), the present work aimed to answer various questions related to the survival of the studied patients with biliary cirrhosis. From the very first one being "What does the survival of the patients look like overall?" we explicitely confronted their survival to selected covariates (part 3.3). We then studied the potentially existing differences between groups through unique covariates (part 3.4) and assessed both isolated and joint impact of progressively selected covariates on the survival through statistical significance (part 3.5). We finally produced the tests diagnostics (part 3.6) and concluded.

2 Data preparation

2.1 Libraries importation

We also ground the present document on various other packages for model elaboration (ie. glmnet, survival) and data presentation (ie. gglopt2, readr, glmnet . . . etc.).

2.2 Specify death of patients as the survival event

Declare data importation and event association to the death of the patient corresponding to the status parameter returning “2”. Transplantation cases (status parameter returning “1”) were not considered in the context of a survival analysis, stricto sensu.

```
# assign data set to a labelled object
data <- pbc
# create event parameter corresponding to death of the patient
data$event <- 0 + (data$status == 2)
```

2.3 Convert the necessary covariates into factor

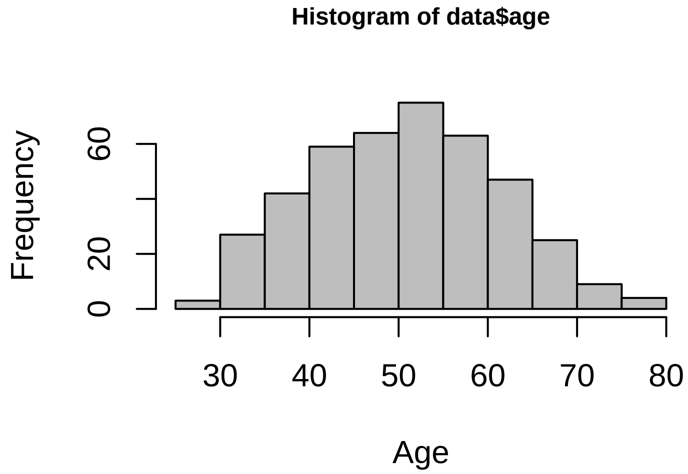
The reader might be interested in an example line of code as displayed below:

```
data$trt <- factor(data$trt)
```

2.4 Create age intervals

Printing the histogram of the age variable helped naively identifying gaps or cuts in its distribution. This would highlight the number of modes encompassed in the variable. Obvious gaps would suggest cutoffs on which basing the creation of age groups variable.

```
hist(data$age, xlab="Age", col="gray", cex.main=0.75)
```



As the output shown, we observed a “properly” distributed age variable with no cuts or gaps thus implying an ageGroup variable “evenly”" distributed too.

```
data$ageGroup <- cut(data$age, breaks = c(0,10,20,30,40,50,60,70,80,90,Inf))
```

2.5 Input some missing values

Few observable missing values required us to apply linear model predictions in order to be refined. The reader might be interested in an example line of code as displayed below:

```
# for chol (cholesterol) parameter
fit.chol <- (lm(chol ~ age, data = data))
data$chol[is.na(data$chol)] <-
  predict(fit.chol, newdata = subset(data, is.na(chol)))
```

Raw data having been treated at this point, we could enter the exploratory analysis phase of the data set.

3 Exploratory analysis

It started with the creation of the data subset to be studied according the following rule.

3.1 Locate patients for survival analysis

Our patient’s type-profile encompassed:

- patients that followed a treatment, thus excluding the 106 patients that did not;
- patients concerned by the event of their death or consored, thus excluding the patients that were transplanted.

```
specimen <- subset(data, data$trt != "NA" & data$status != 1)
```

Getting a 293 observations data set on, called “specimen”, which we now would be able to run the survival analysis related tests as below.

3.2 Create survival objects

Survival objects are created through the Surv(time, status) function from the “survival” package. To create right-censored data, this function needs two arguments:

- time: returns the observed duration in days;
- status: returns a boolean regarding whereas the observation corresponds to a censored one or not.

In the situation where status returns more than two modalities or if the modalities are not returning a boolean, conditioned by the fact that the observations are censored or not, the formula creating the survival object must precise the proper modalities corresponding to censored observations.

```
survival <- Surv(specimen$time / 365.25, specimen$event)
```

Hereby computed with time parameter alteration to show yearly-basis scale for lisibility purpose of the reader.

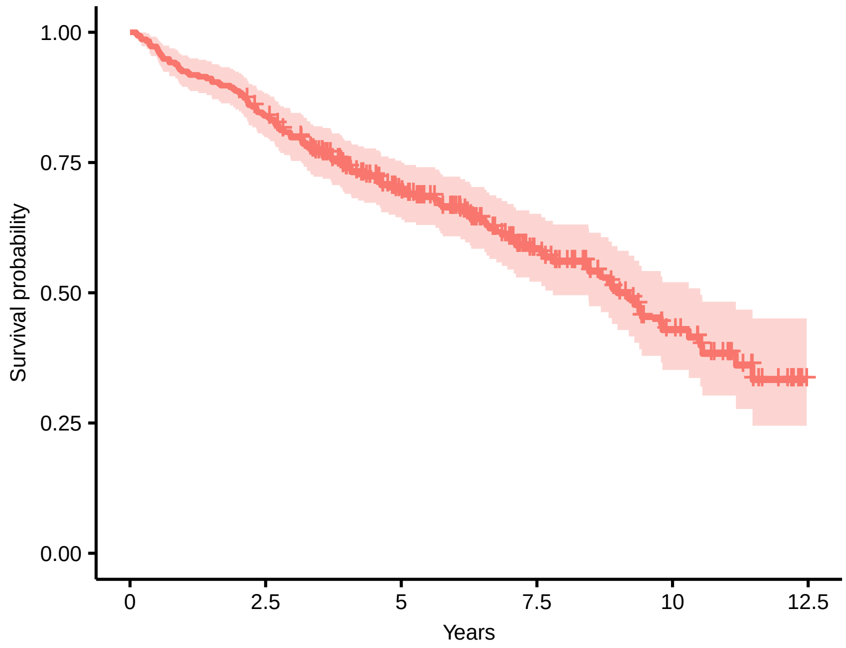
3.3 Kaplan-Meyer estimator - estimation of the survival function

Also known as "product-limit estimator", the Kaplan-Meyer estimator (KM) is a non-parametric statistic (ie. not based on the assumption of an underlying probability distribution) that allows one to estimate the survival function. It gives the probability that an individual patient will survive past a particular time "t". It is based on the assumption that the probability of surviving past this point is equal to the product of the observed survival rates until time point "t". It is similar to the censoring version of empirical survival function, generating a stair-step curve but not accounting for effect of other covariates.

Applying the Kaplan-Meyer estimator helped answer the question "How is the survival of the overall studied sample shaped like?"

```
KM <- survfit(survival ~ 1, data = specimen)
```

```
## Call: survfit(formula = survival ~ 1, data = specimen)
##
##       n  events  median 0.95LCL 0.95UCL
##  293.00  125.00    8.99    7.79   10.51
```

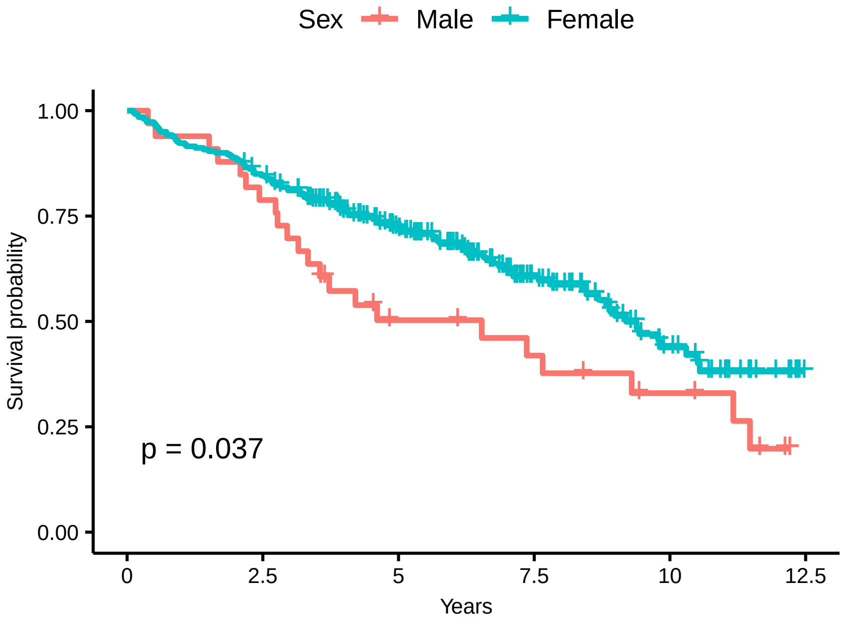


The KM test returned a median survival of 8.99 years, the moment at which 50% of the patients were alive and 50% were reaching the event point ie. here, death. On a broader note, the reader may be interested in visualizing the survival regarding other parameters. This have been realised by crossing the survival object with the specific parameters through additional KM tests as shown below.

One might state that interesting parameters to be confronted to survival are sex, trt (for treatment parameter) and age, the later requiring a preparation to its study (ie. "binarizing" the sample into "younger" vs "older" patients for example).

First trying to answer the question: "How is the overall survival shaped like regarding patients gender?"

```
# fitting the survival to sex parameter
fit.sex <- survfit(survival ~ sex, data = specimen)
```

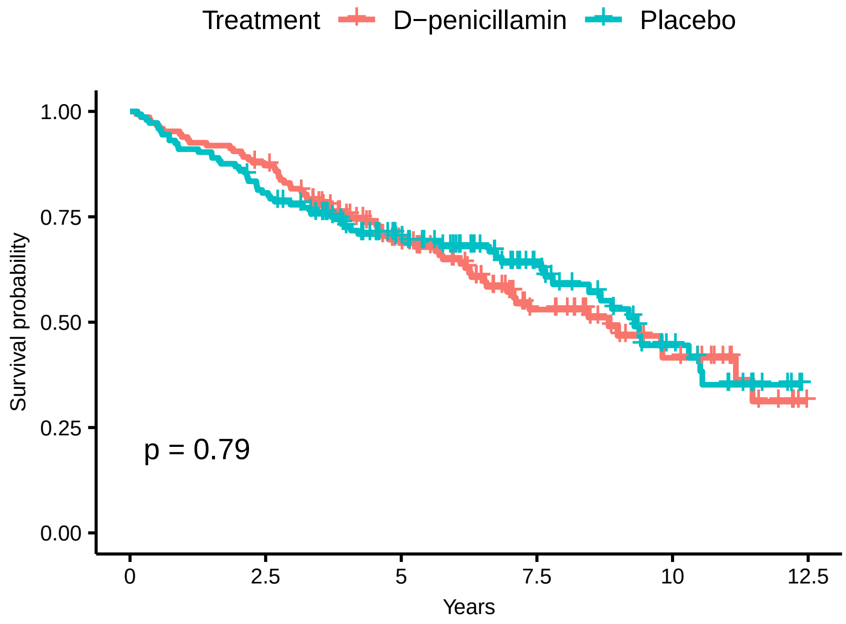


Additionally to the general shape of the curves, the reader might be interested in the p-value shown at the bottom-left of the figure which is the corresponding log-rank test p-value result. Here statistically significant, as under an arbitrary threshold of 5% (corresponding to a 95% confidence interval) it conveys enough significance to reject the log-rank null hypothesis and affirm that the two groups, here male & female, survive differently to the biliary cirrhosis.

Explicitly, the output above shown that men have a worse survival expectancy than women to biliary cirrhosis. The reader might be interested in noting that the number of censored data for female patients appear to be greater than the ones of male patients, one might naively state that it would be concuring the above results.

Now considering the treatment parameter (D-penicillamin vs. placebo), thus answering the question: "How is the overall survival shaped like regarding patients administred treatment?"

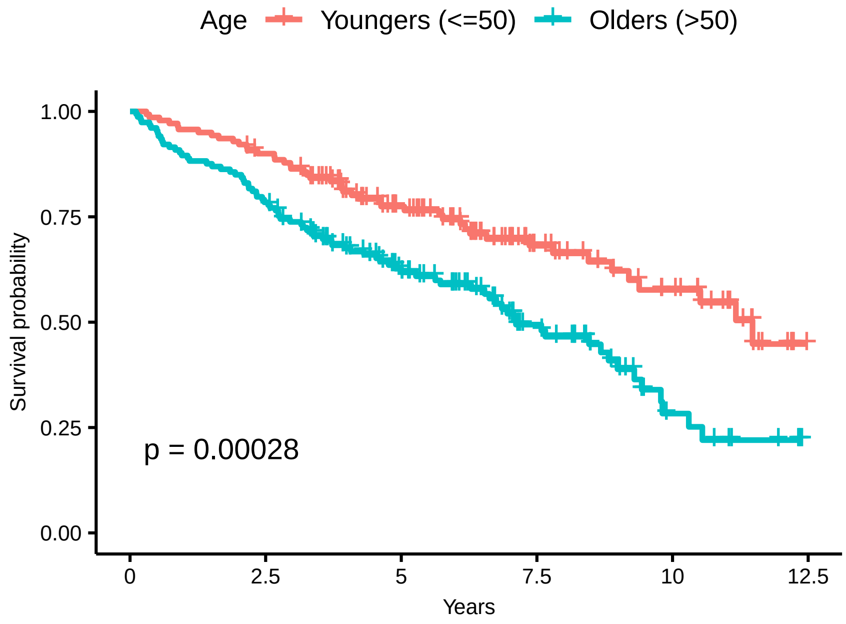
```
# fitting the survival to treatment parameter
fit.trt <- survfit(survival ~ trt, data = specimen)
```



As the reader might visually build an intuition of the difference in treatment parameter, the resulting non-significant p-value (79%) indicates that there is not enough statistical material in order to reject the null hypothesis and thus leads one to conclude that there is no significant difference between both treatment protocols. Explicitly, whereas taking a D-penicillamin treatment or a placebo treatment has no impact on patients survival expectancy.

Finally one may think an additional useful visualization to the reader would be the study of the survival object regarding the age parameter. As stated earlier it appeared necessary to retreat the age parameter in order to make it senseful to the KM and log-rank tests by "binarizing" it as shown in the following, aiming to answer the question: "How is the overall survival shaped like regarding patients relative age?"

```r
# age parameter retreatment named "ageBin" parameter
specimen$ageBin <- ifelse(specimen$age > 50, ">50", "<=50")
# converting the ageBin parameter into factor
specimen$ageBin <- as.factor(specimen$ageBin)
# fitting the survival to the new age parameter
fit.age <- survfit(survival ~ ageBin, data = specimen)
```



Here again as the reader may have an intuition of the potential difference in survival regarding the age parameter as shown by the shapes of the curves, the resulting p-value (0.028%) indicates that there exists a statistically significant difference in the survival of the two groups segmented through the age parameter. Explicitly the "olders", the patients who's age is greater than fifty years old, have a worse survival expectancy over time than the "youngers", the patients who's age is lower or equal to fifty years old.

The provided p-value to the KM visualization of the survival object introduced the reader to the observation of differences in some parameters variable to be explicited in the Mantel-Haenzel test, also called the log-rank test.

3.4 Mantel-Haenzel test - comparing two groups' own survival

Also known as log-rank test, it is a statistical hypothesis test that tests the null hypothesis that survival curves of two populations do not differ. It will output an indicator of the two groups being significantly different in terms of survival when its p-value will be inferior to risk threshold.

It is efficient in comparing groups differed by categorical variables, but not continuous ones. Its validity conditions might appear quite delicate to the reader as the log-rank test, to be considered as applicable, requires or an important number of death times which mathcs the situation of our sample study, or an important number of deads at each death time.

Nonetheless it appeared necessary to answer the question: "Is survival different for patients who were administred one treatment rather than the other?"

```
MH <- survdiff(survival ~ specimen$trt)
```

```
## Call:
## survdiff(formula = survival ~ specimen$trt)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## specimen$trt=1 148       65     63.5    0.0354    0.0722
## specimen$trt=2 145       60     61.5    0.0366    0.0722
```

```
##
##  Chisq= 0.1  on 1 degrees of freedom, p= 0.8
```

The log-rank test returned a non-significant p-value (80%) indicating that one does not have enough elements to reject the null hypothesis allowing to interprete that there is no statistically significant difference between the two treatments. Concurring the earlier interpretation, whereas a patient was administered D-penicillamin or placebo had no impact on the patient's survival expectancy.

An alternative test, the Wilcoxon test may be applied in order to compare the significance of the result with the one from the log-rank test. However the reader will be advised that:

- the log-rank test is more effective when the survival curves do not cross each other;
- when instantaneous hazard rates are proportional, the log-rank test is the "best" to be run.

```
W <- survdiff(survival ~ specimen$trt, rho=1)
```

```
## Call:
## survdiff(formula = survival ~ specimen$trt, rho = 1)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## specimen$trt=1 148     49.1     48.7   0.00370   0.00949
## specimen$trt=2 145     46.3     46.7   0.00385   0.00949
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

Having returned a greater p-value for the log-rank test, it prevented from rejecting the null hypothesis, concurring the earlier interpretation, leading to conclude that there is no statistically significant difference between the two studied groups ie. whereas patients were administred D-penicillamin or placebo.

Various groups having been studied and compared regarding different perspectives, another complementary approach would be association of survival to a quantitative variable, allowed by the Cox model as presented in the following part.

3.5 Cox Model Regressions

Also known as proportional hazard model, it conveniently accesses the effect of continuous and categorical variable using partial likelihood to get inference even without knowledge of baseline hazard.

While the log-rank test compares two Kaplan-Meier survival curves, which might be derived from splitting a patient population into treatment subgroups, Cox proportional hazards model regressions are derived from the underlying baseline hazard functions of the specific patient populations studied and an arbitrary number of dichotomized covariates. Again, it does not assume an underlying probability distribution but it does assume that the hazards of the compared patient groupsare constant over time.

The reader be advised that the presented approach was first to process univariate Cox regressions fitting the three covariates : sex, treatment & age. This would supposedly help answering the question "Do specifically selected covariates (sex, treatment & age) independently and significantly impact survival and how?"

The second step of the approach has been to process a multivariate Cox regression on all the covariates of the sample in order to identify the most significant ones on patients survival expectancy and then return a multivariate Cox regression on the selected covariates. This would help us answer the question "Which of covariates from the data set jointly and significantly impact survival and how?"

First then, one may want to describe if and how the sex, treatment & age parameters independently impact on survival:

```r
# univariate cox regression on sex parameter
cox.sex <- coxph(survival ~ sex, data = specimen)
```

```
## Call:
## coxph(formula = survival ~ sex, data = specimen)
##
##         coef exp(coef) se(coef)     z      p
## sexf -0.4872    0.6143   0.2365 -2.06 0.0394
##
## Likelihood ratio test=3.82  on 1 df, p=0.05064
## n= 293, number of events= 125
```

```r
# univariate cox regression on treatment parameter
cox.trt <- coxph(survival ~ trt, data = specimen)
```

```
## Call:
## coxph(formula = survival ~ trt, data = specimen)
##
##          coef exp(coef) se(coef)      z     p
## trt2 -0.04823   0.95292  0.17917 -0.269 0.788
##
## Likelihood ratio test=0.07  on 1 df, p=0.7877
## n= 293, number of events= 125
```

```r
# univariate cox regression on age parameter
cox.age <- coxph(survival ~ age, data = specimen)
```

```
## Call:
## coxph(formula = survival ~ age, data = specimen)
##
##         coef exp(coef) se(coef)     z        p
## age 0.036526  1.037201 0.008903 4.103 4.08e-05
##
## Likelihood ratio test=16.81  on 1 df, p=4.13e-05
## n= 293, number of events= 125
```

The reader may be interested in the description of the numerous interpretable outputs provided by the Cox Model regressions for the sake of clarity:
The “*input p-value*” indicates whereas there is a statistically significant association between a given variable and the hazard (risk of the event, here death).
The *statistical significance* marked by the “z” column assessesg whether the beta coefficient of a given variable is statistically significantly different from 0 by measuring each regression coefficient to its standard error.
The sign of the regression coefficient with a positive (negative) sign implies a higher (lower) hazard, with the specificity for variables encoded as numeric vectors, here as for sex parameter (1=male, 2=female) and treatment parameter (1=D-penicillamin, 2=placebo), that the coefficient assesses the second group relative to the first one.
The *hazard ratio* gives the effect size of covariates.
The *global statistical significance* of the model is brought by the “output p-value” given for overall significance of the model, the likelihood-ratio test.

Now from the output above, one would carrefully interprete the results by stating that:

- both sex & age parameters have a statistically significant association with the hazard of the patients produced by biliary cirrhosis (ie. p-value of 3.9% & c.0% respectively);
- both sex & age parameters have highly statistically significant coefficients;
- on one hand a beta of -0.49 indicates that females have lower risk of death than males and that on another hand older patients have a higher risk of death regarding biliary cirrhosis;
- being a female patient reduces the hazard by a factor of 0.61 or 39% thus associated with a relatively better prognostic. Besides, one can estimate that a 3.7% greater risk of death is associated with a 1-year increase in age at diagnosis;
- the p-value being associated to c.0%, the model is indeed significant.

With these elements in mind, one may now want to describe how the factors jointly impact on survival. Answering this question required performing a multivariate Cox regression analysis. As the treatment parameter was not significant in the univariate Cox analysis, it was omitted from the multivariate analysis.

```
# multivariate cox regression
coxph <- coxph(survival  ~ age + edema + hepato + platelet + sex + spiders + ascites
               + log(albumin) + log(alk.phos) + log(ast) + log(bili) + log(chol)
               + log(copper) + log(trig) + log(protime), data = specimen)
```

```
## Call:
## coxph(formula = survival ~ age + edema + hepato + platelet +
##     sex + spiders + ascites + log(albumin) + log(alk.phos) +
##     log(ast) + log(bili) + log(chol) + log(copper) + log(trig) +
##     log(protime), data = specimen)
##
##   n= 293, number of events= 125
##
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## age            0.0297466  1.0301935  0.0102532  2.901 0.003717 **
## edema0.5       0.2095631  1.2331392  0.2955846  0.709 0.478338
## edema1         0.8182917  2.2666245  0.3516666  2.327 0.019971 *
## hepato1        0.2644691  1.3027392  0.2307070  1.146 0.251654
## platelet       0.0001284  1.0001284  0.0011422  0.112 0.910467
## sexf          -0.0178119  0.9823458  0.2927984 -0.061 0.951492
## spiders1      -0.0264097  0.9739360  0.2262825 -0.117 0.907089
## ascites1       0.3116815  1.3657197  0.3259339  0.956 0.338935
## log(albumin)  -2.4454742  0.0866850  0.8238775 -2.968 0.002995 **
## log(alk.phos) -0.1066629  0.8988286  0.1375068 -0.776 0.437931
## log(ast)       0.4021523  1.4950390  0.2797417  1.438 0.150552
## log(bili)      0.6291907  1.8760916  0.1670954  3.765 0.000166 ***
## log(chol)      0.0923531  1.0967520  0.2694179  0.343 0.731758
## log(copper)    0.3452784  1.4123830  0.1620346  2.131 0.033098 *
## log(trig)     -0.0870289  0.9166506  0.2509698 -0.347 0.728764
## log(protime)   3.2968547 27.0274966  1.1738557  2.809 0.004976 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## age             1.03019     0.9707   1.00970    1.0511
## edema0.5        1.23314     0.8109   0.69089    2.2010
## edema1          2.26662     0.4412   1.13773    4.5156
## hepato1         1.30274     0.7676   0.82886    2.0476
## platelet        1.00013     0.9999   0.99789    1.0024
## sexf             0.98235     1.0180   0.55339    1.7438
## spiders1         0.97394     1.0268   0.62506    1.5175
## ascites1         1.36572     0.7322   0.72098    2.5870
## log(albumin)     0.08669    11.5360   0.01724    0.4357
## log(alk.phos)    0.89883     1.1126   0.68648    1.1769
## log(ast)         1.49504     0.6689   0.86404    2.5868
## log(bili)        1.87609     0.5330   1.35214    2.6031
## log(chol)        1.09675     0.9118   0.64681    1.8597
## log(copper)      1.41238     0.7080   1.02808    1.9403
## log(trig)        0.91665     1.0909   0.56050    1.4991
## log(protime)    27.02750     0.0370   2.70781  269.7700
##
## Concordance= 0.858  (se = 0.017 )
## Rsquare= 0.51   (max possible= 0.987 )
## Likelihood ratio test= 209.2  on 16 df,   p=<2e-16
## Wald test            = 206.1  on 16 df,   p=<2e-16
## Score (logrank) test = 308.3  on 16 df,   p=<2e-16
```

First, this time the output gave p-values for three alternative tests for overall significance of the model: the likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent. For N large enough, they will give similar results. For small N, they may differ, literature indicating the likelihood-ratio test would be preferred in such case. From the output the reader may be interested in the observation that for all three overall tests, the p-value is significant thus indicating the model is indeed significant. These tests evaluate the null hypothesis that all of the beta coefficients are 0. Here the test statistics were in close agreement, consequently, the null hypothesis was soundly rejected.

Six covariates appeared to be significant with some notable results:

- age parameter remained significant;
- sex parameter failed to be significant (p-value = 0.95);
- the p-value for bili parameter (serum bilirunbin) returned 0.000166 with hazard ratio of 1.88, allowing to estimate that, holding all other covariates equal, a 88% greater risk of death is associated with a 1mg increase by dl of blood at diagnosis.

By contrast all covariates with confidence interval including 1 were not significant and thus rejected from the selection towards refined analysis.

```
# multivariate Cox regression with significant covariates only
fit.coxph <- coxph(survival  ~ age + as.factor(edema) + log(albumin) + log(bili)
                   + log(protime) + log(copper), data = specimen)
```

```
## Call:
## coxph(formula = survival ~ age + as.factor(edema) + log(albumin) +
##     log(bili) + log(protime) + log(copper), data = specimen)
##
##   n= 293, number of events= 125
##
##                          coef exp(coef)  se(coef)      z Pr(>|z|)
## age                  0.029276  1.029709  0.008604  3.402 0.000668 ***
## as.factor(edema)0.5  0.136814  1.146615  0.277344  0.493 0.621800
## as.factor(edema)1    0.861945  2.367762  0.305261  2.824 0.004748 **
## log(albumin)        -2.825567  0.059275  0.735043 -3.844 0.000121 ***
## log(bili)            0.745536  2.107571  0.113283  6.581 4.67e-11 ***
## log(protime)         3.083341 21.831229  1.098555  2.807 0.005005 **
## log(copper)          0.368473  1.445525  0.136325  2.703 0.006874 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## age                   1.02971    0.97115   1.01249    1.0472
## as.factor(edema)0.5   1.14662    0.87213   0.66580    1.9747
## as.factor(edema)1     2.36776    0.42234   1.30167    4.3070
## log(albumin)          0.05928   16.87051   0.01403    0.2503
## log(bili)             2.10757    0.47448   1.68794    2.6315
## log(protime)         21.83123    0.04581   2.53505  188.0051
## log(copper)           1.44552    0.69179   1.10659    1.8883
##
## Concordance= 0.852  (se = 0.018 )
## Rsquare= 0.502   (max possible= 0.987 )
## Likelihood ratio test= 204.1  on 7 df,   p=<2e-16
## Wald test            = 200.9  on 7 df,   p=<2e-16
## Score (logrank) test = 287.4  on 7 df,   p=<2e-16
```

The model held its overall significance according to the "output p-values" returned by the three tests (likelihood, Wald, and score) and additional notable results were to be reported to the reader:

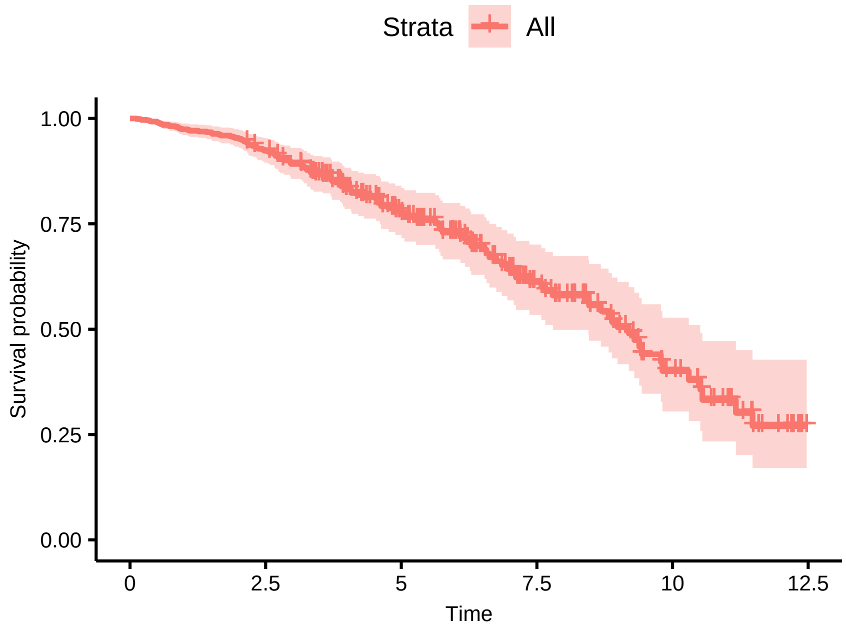
- all covariates remain significant;
- with an ever closer to 0% p-value and a still extremely high hazard ratio bili parameter kept its position of most significant and affective covariate on survival. The reader be invited to precaution regarding such results especially its reported hazard ratio being probably the consequence of the unit scale in which bili parameter is measured (mg/dl) as an increase of 1mg/dl might be a very unlikely phenomenon;
- also significant (p-value = 0.5%) protime parameter reported a suspiciously high hazard ratio of 21.83 which may be explained by the fact that time parameter has been converted from days to years for reader lisibility in the earlier steps of the present study. As a consequence, protime was not considered for further analysis.

The present approach helped identifying the most significant covariates to survival of the present data set. A naive interpretation of the final selection of the most significant continuous covariates may look like the following:

- the older the patient the lower the survival expectancy;
- the higher the serum albumin of the patient the higher the survival expectancy;
- the higher the serum bilirunbin of the patient the lower the survival expectancy;
- the higher the urine copper of the patient the lower the survival expectancy.

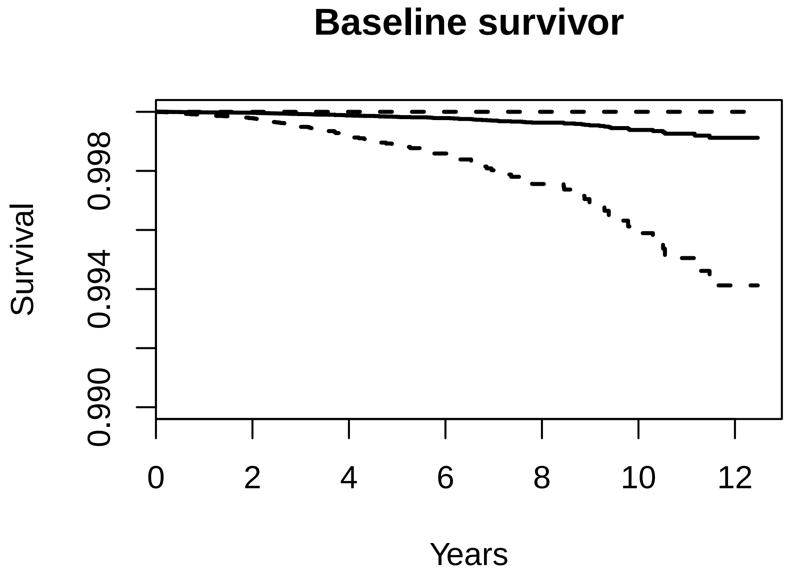
The Cox model having been fit to the data, the reader may now be interested in visualizing the predicted survival proportion at any given point in time for a particular risk group. The function survfit() estimates the survival proportion, by default at the mean values of covariates.

```r
# automatically visualizing the estimated distribution of survival times
ggsurvplot(survfit(fit.coxph), data = specimen,
           font.x =  8, font.y = 8, font.tickslab = 8)
```

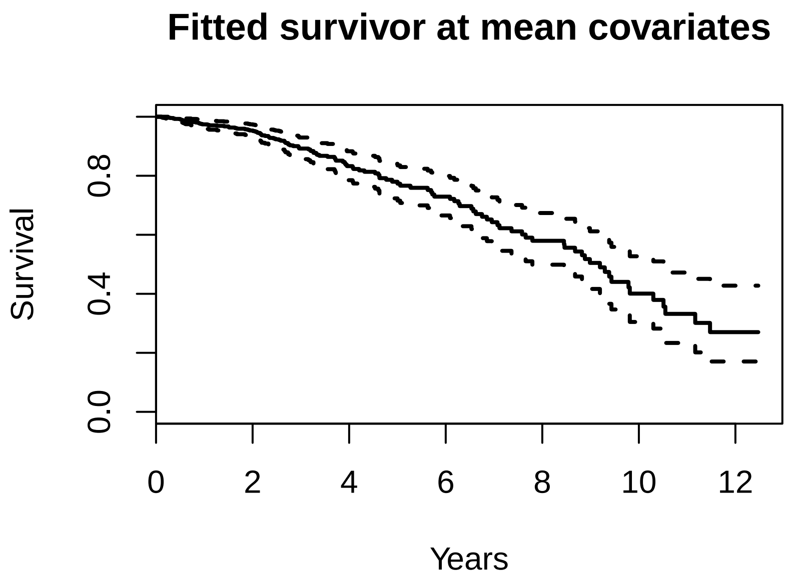


A more "manual" approach allowed to separate a baseline survivor model from the mean values of covariates. The reader be advised that mentionning the "as.factor()" function on the edema parameter, although already ran earlier, helped fix the plot of the baseline.

```
# Manually visualizing the estimated distribution of survival times
specimen.null<-data.frame(age=rep(0,1), edema=rep(0,1), bili=rep(1,1), albumin=rep(1,1),
                          protime=rep(1,1), copper=rep(1,1))
# for baseline
plot(survfit(fit.coxph, newdata=specimen.null), lwd=2,ylim=c(.99,1),
     main='Baseline survivor', xlab='Years', ylab='Survival', conf.int=T)
```



```r
# for mean covariates
plot(survfit(fit.coxph),lwd=2,main= 'Fitted survivor at mean covariates',
     xlab='Years', ylab='Survival')
```



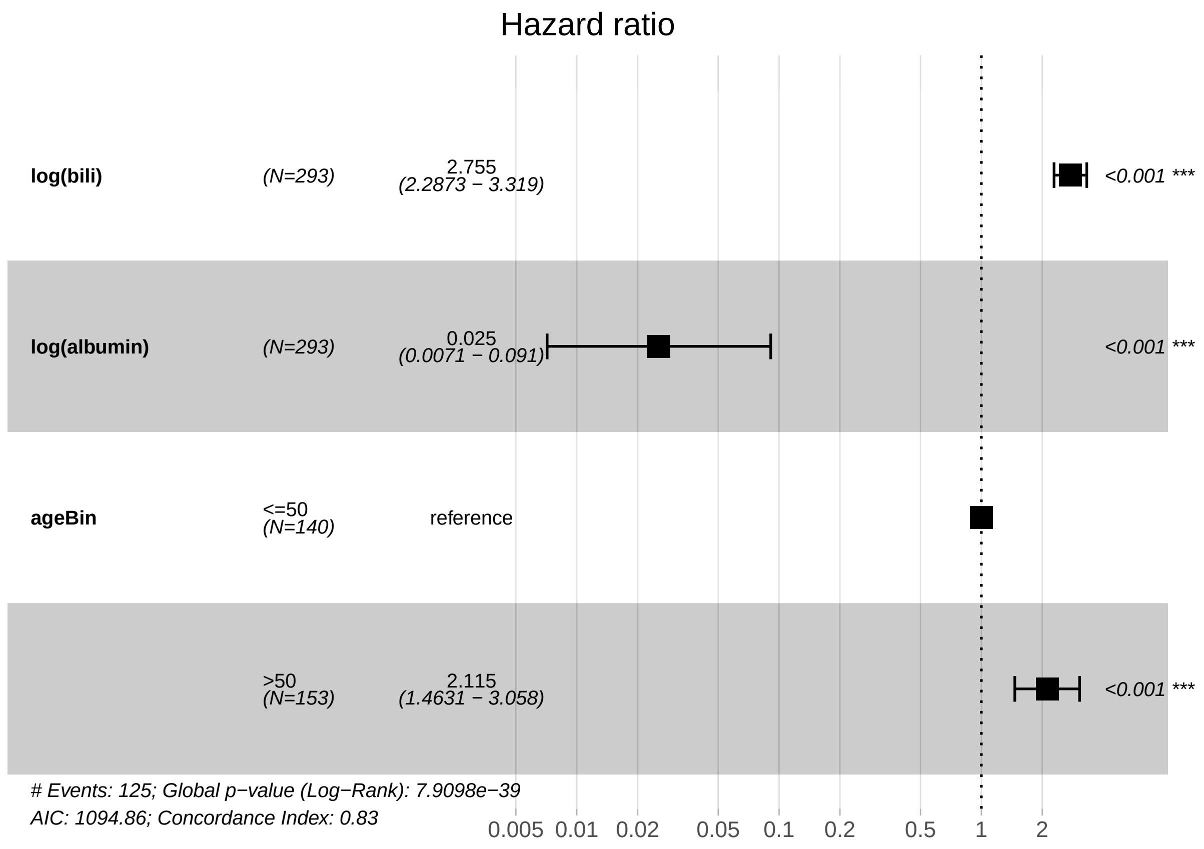
Returning a unique curve for all the patients in the data set with a confidence interval.

Various other visualizations exist such as the function ggforest() from the survminer package which creates a forest plot for a Cox regression model fit. Hazard ratio estimates along with confidence intervals and p-values are plotter for each variable. One may be interested in the forest plot for the three most significant covariates : age, bili and albumin parameters.

```r
ggforest(coxph(survival ~ log(bili) + log(albumin) + ageBin, data = specimen))
```

```
## Warning in .get_data(model, data = data): The `data` argument is not
## provided. Data will be extracted from model fit.
```

```
## Warning: Removed 1 rows containing missing values (geom_errorbar).
```



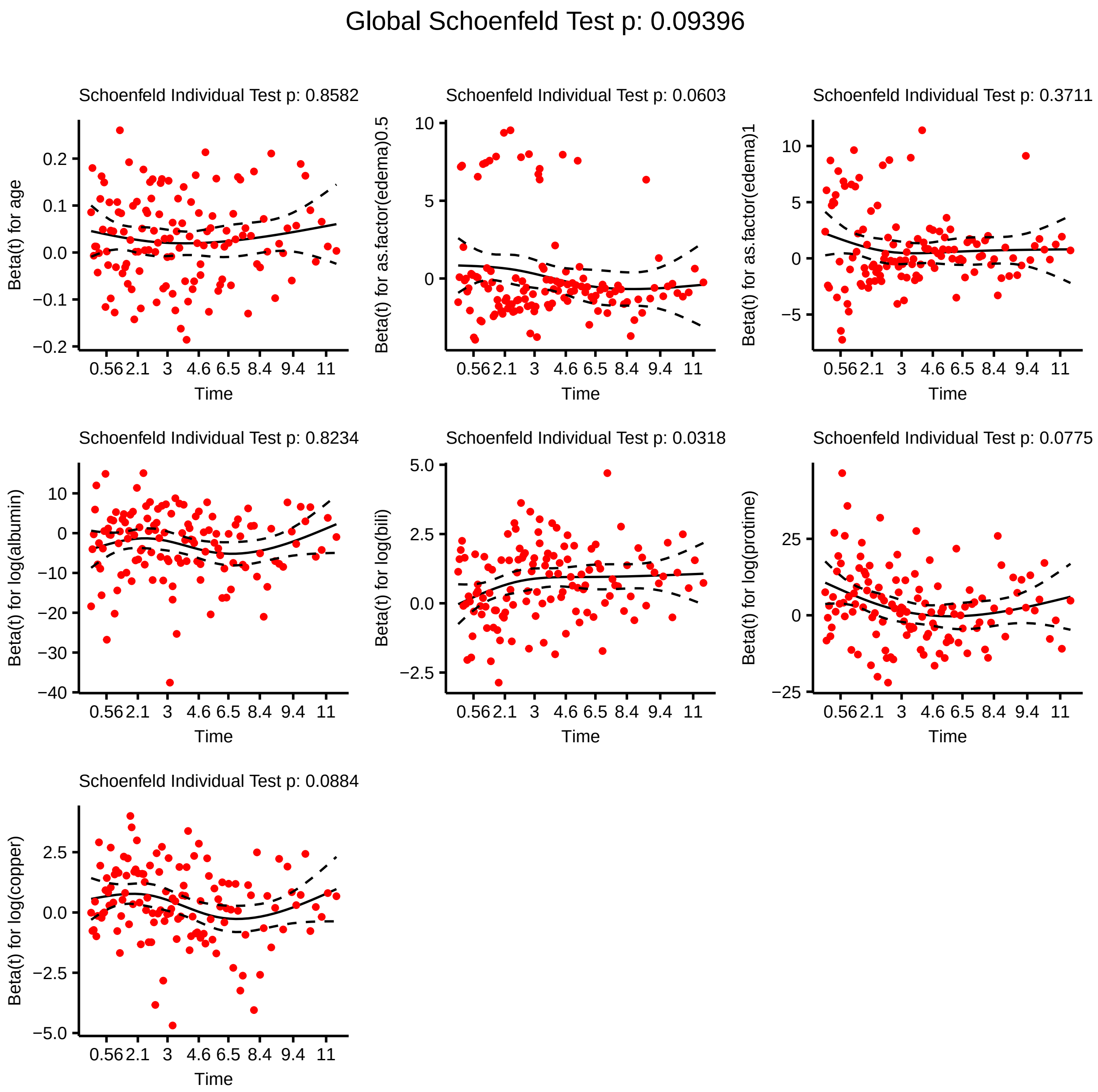
From the output above one can state that the forest plot provided an alternative view concurring the earlier stated conclusions about the most significant covariates. The value 1 being the point at which the covariate has no impact on the survival, the reader clearly sees that the selected coviates do have a statistically significant impact ie. a positive one on the left side of the 1 value and conversely a negative one on the right side. The reader be advised of the scale especially when interpreting the confidence intervals.

Both the isolated and joint impacts of the covariates having been assessed through Cox Model regressions, it is of scientific requirement and one should be now dedicated to the assurance that the Cox Model can be applicated to the studied data set.

3.6 Diagnostic of Cox Model

The function cox.zph() from survival package may be used to test the proportional hazards assumption for a Cox regression model fit. The graphical verification of this assumption may be performed with the function ggcoxzph() from the survminer package. For each covariate it produces plots with scaled Schoenfeld residuals against the time.

```
ftest <- cox.zph(fit.coxph)
ggcoxzph(ftest, font.main=8, font.x =  8, font.y = 8, font.tickslab = 8)
```



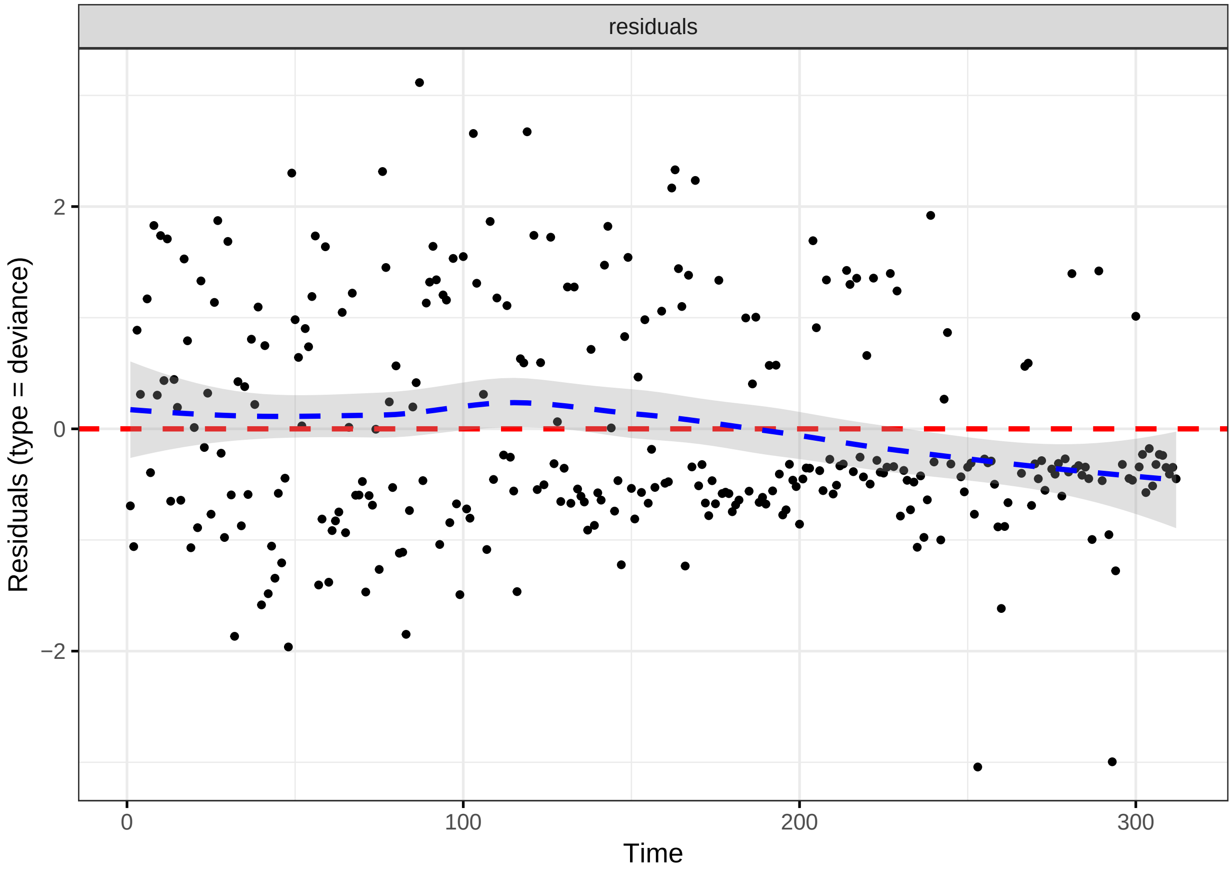
The Schoenfeld Residuals Test is used to test the independence between residuals and time and hence is used to test the proportional Hazard assumption in Cox Model. It is analogous to testing whether the slope of scaled residuals on time is zero or not. If the slope is not zero then the proportional hazard assumption has been violated.

Consequently a first observation we are expecting to a flat resulting curve in order to consider that the hazard ratio hypothesis (or instantaneous proportional risks hypothesis) is verified. Which happens to be the case for our selected covariates overall. And thus the Cox regression model was validated.

Additionally the function ggcoxdiagnostics() plots different types of residuals. The reader may be especially interested in the diagnostics of “deviance” for its clarity as providing an overall diagniostic by returning a unique plot for all selected covariates.

```
# deviance vs. time
ggcoxdiagnostics(fit.coxph, type = "deviance", ox.scale = "time")
```

```
## Warning in ggcoxdiagnostics(fit.coxph, type = "deviance", ox.scale =
## "time"): ox.scale='time' works only with type=schoenfeld/scaledsch
```



Similarly the returning curve (in blue) was expected to be as flat as possible, distributed around 0 and the data to be homogeneously distributed on the graph in order to consider the hazard ratio hypothesis to validate the applied Cox model on the studied data set. Which once again appeared to be the case overall.

4 Conclusion

The present document aimed to provide the reader with rigorous statistical analysis material for answering various questions related to the survival of the studied patients with biliary cirrhosis. In essence, it conclusively allowed one to establish that the median survival of the studied sample was 9 years with significant differences between patients gender on one hand and patients age on the other. Also, five covariates were identified as statistically signficant impactors of the patients survival.