PBC

Stephen Shannon

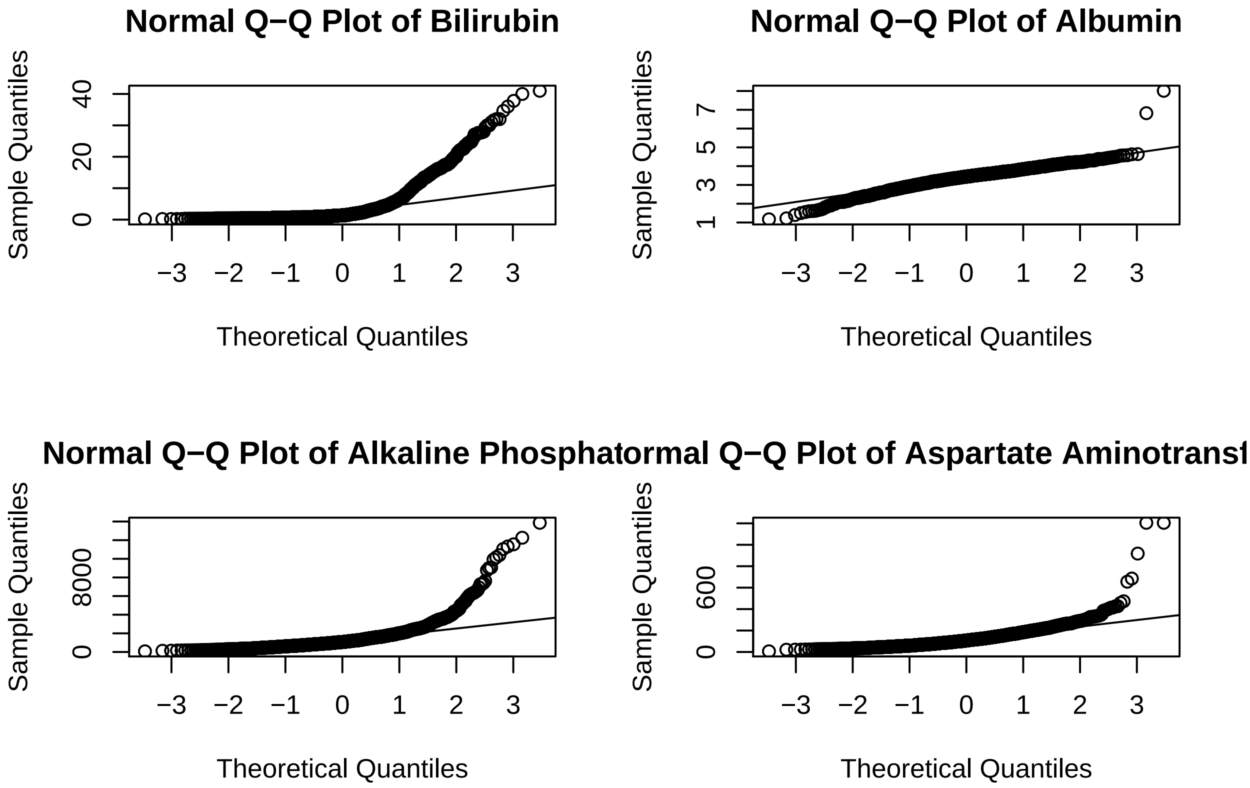
June 28, 2021

```
library(dplyr)
library(tidyverse)
library(survival)
library(survminer)
library(glmnet)
library(vtable)
library(ggplot2)
library(ggfortify)
```

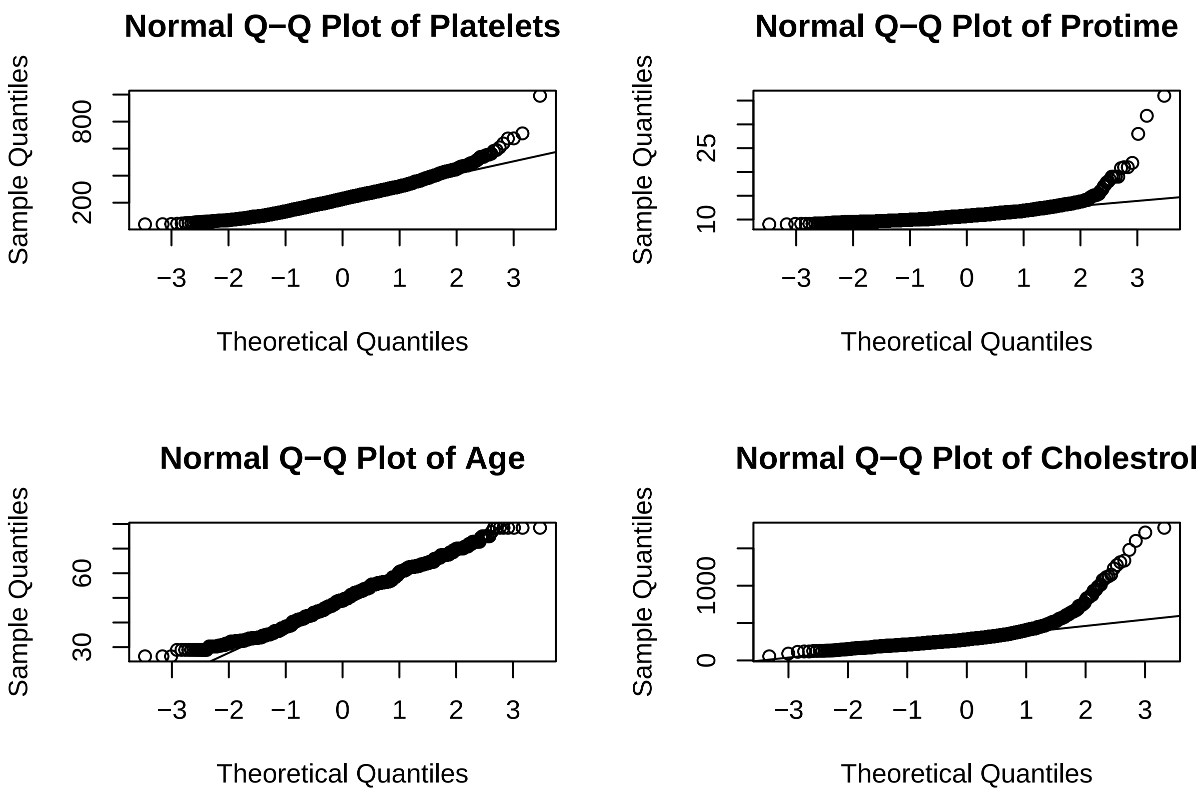
```
  #load data
data(pbc, package="survival")
```

Question 1. Assess the normality of each of the candidate variables and create a table showing the appropriate summary statistics (e.g. mean +- sd or median and interquartile range)

```
  #check normality of the indepedent vars
par(mfrow=c(2,2))
qqnorm(pbcseq$bili, main = "Normal Q-Q Plot of Bilirubin");qqline(pbcseq$bili) #non-normal
qqnorm(pbcseq$albumin,main = "Normal Q-Q Plot of Albumin");qqline(pbcseq$albumin) #normal
qqnorm(pbcseq$alk.phos,main = "Normal Q-Q Plot of Alkaline Phosphate");qqline(pbcseq$alk.phos) #non-nor
qqnorm(pbcseq$ast,main = "Normal Q-Q Plot of Aspartate Aminotransferase");qqline(pbcseq$ast) #non-norma
```



```r
qqnorm(pbcseq$platelet,main = "Normal Q-Q Plot of Platelets");qqline(pbcseq$platelet) #non-normal
qqnorm(pbcseq$protime,main = "Normal Q-Q Plot of Protime");qqline(pbcseq$protime) #non-normal
qqnorm(pbcseq$age,main = "Normal Q-Q Plot of Age");qqline(pbcseq$age) #normal
qqnorm(pbcseq$chol,main = "Normal Q-Q Plot of Cholestrol");qqline(pbcseq$chol) #non-normal
```



Figures 1 - 8: qqplots for each continous variable in the dataset, only albumin, and age are approximately normal.

Table 1. Summary statistics for each continous variable in the dataset

```
  #table of summary statistics using vtable package for sumtable
sumtable(pbcseq, vars = c("bili", "chol", "albumin","alk.phos", "ast", "platelet", "protime", "age"), su
```

```
##   Variable NotNA     Mean       Sd    Min Pctile[25] Median Pctile[75]    Max
## 1     bili  1945    3.672    5.373    0.1        0.8    1.4        3.9     41
## 2     chol  1124  320.472  166.717     55        235    281     349.25   1775
## 3  albumin  1945     3.39    0.503   1.17       3.11   3.44        3.7   8.01
## 4 alk.phos  1885 1381.912 1195.624     73        737   1072       1636  13862
## 5      ast  1945   122.67   78.438    6.2         72    107        155   1205
## 6 platelet  1872  233.681   97.663     40        165    228     290.25    991
## 7  protime  1945   10.998    1.479      9       10.1   10.8       11.5     36
## 8      age  1945    49.26   10.062 26.278     41.793 48.871     56.153 78.439
```

```
  #glmnet requires that there are no NAs, and that the event var is only 0 or 1.
  #Since 1s are given to liver transplant cases, we must filter those who recieved
  #liver transplants and replace it with death, 2.

  #It may be easier to remove cholestrol from the analysis as many entries
  #are missing, and would remove all of these incomplete cases from
  #the analysis

pbcseq <- pbcseq %>% select(-chol)
```

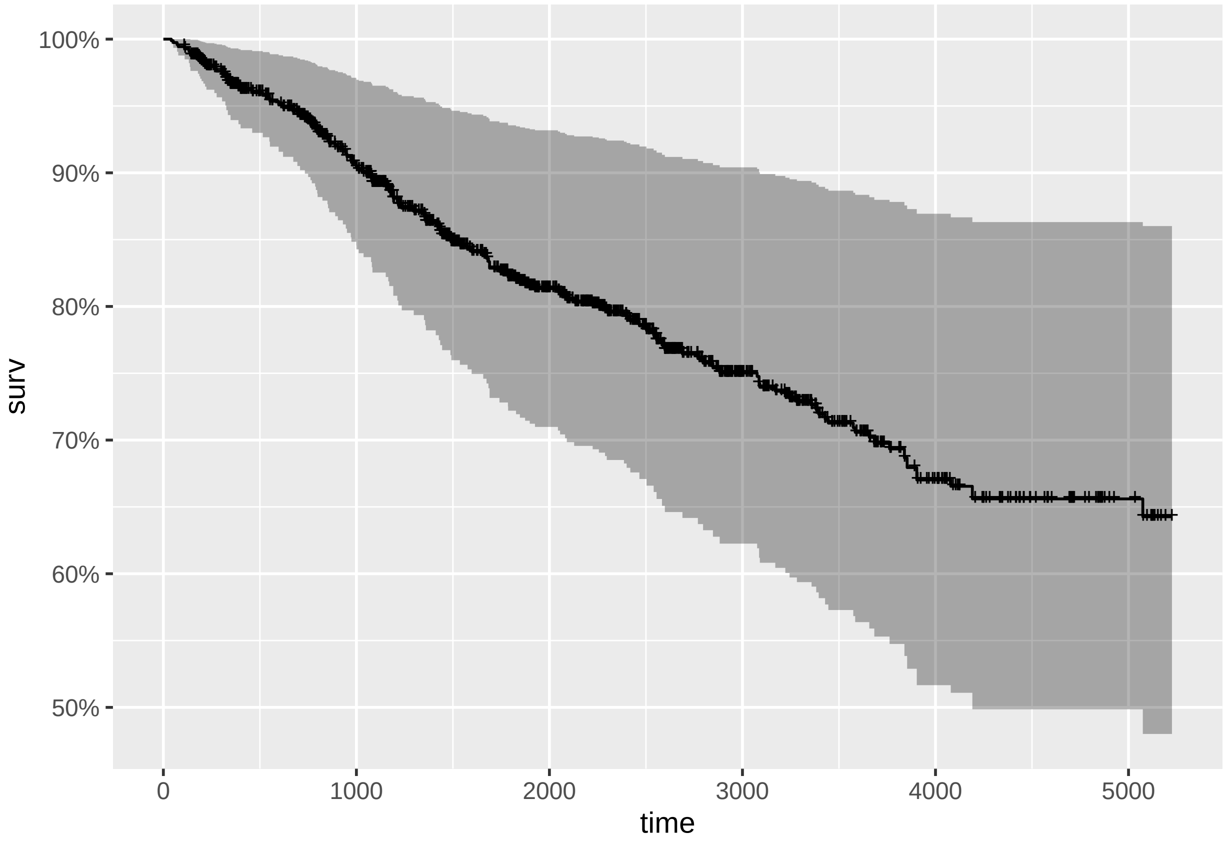
```r
pbcseq <- pbcseq %>% drop_na() %>% filter(status != 1)
pbcseq["status"][pbcseq["status"] == 2] <- 1
```

```r
first <- with(pbcseq, c(TRUE, diff(id) !=0))
last <- c(first[-1], TRUE)

  #setup start, stop times and outcome for coxph
  #if first checkup, choose 0 days, otherwise choose the current day
time1 <- with(pbcseq, ifelse(first, 1, day))

  #if the last checkup, choose the follow up time, otherwise choose the previous check
  #up time (since first checkup is not considered)
time2 <- with(pbcseq, ifelse(last, futime, day[-1]))

  #if last checkup, choose the current status, else choose censored as the outcome
event <- with(pbcseq, ifelse(last, status, 0))

  #basic model from the data source page

#m1 <- coxph(Surv(time1, time2, event) ~ age + sex + log(bili), pbcseq)
#summary(m1)

  #coxph model with every coeffecient
  #transforming some continous variables to the natural log
m2 <- coxph(Surv(time1, time2, event) ~ trt + age + sex + ascites + hepato + spiders + edema + stage +
summary(m2)
```

```
## Call:
## coxph(formula = Surv(time1, time2, event) ~ trt + age + sex +
##     ascites + hepato + spiders + edema + stage + log(bili) +
##     log(albumin) + log(alk.phos) + log(ast) + log(platelet) +
##     log(protime), data = pbcseq)
##
##   n= 1722, number of events= 140
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## trt           -0.20923   0.81121  0.18323 -1.142   0.2535
## age            0.02504   1.02536  0.01012  2.474   0.0134 *
## sexf          -0.32698   0.72110  0.25904 -1.262   0.2068
## ascites        0.40877   1.50496  0.21377  1.912   0.0559 .
## hepato        -0.15878   0.85318  0.23233 -0.683   0.4943
## spiders        0.04984   1.05110  0.19929  0.250   0.8025
## edema          0.86880   2.38406  0.28500  3.048   0.0023 **
## stage          0.33654   1.40009  0.17927  1.877   0.0605 .
## log(bili)      0.89532   2.44811  0.12194  7.343 2.10e-13 ***
## log(albumin)  -2.72070   0.06583  0.57169 -4.759 1.94e-06 ***
## log(alk.phos)  0.05292   1.05434  0.18655  0.284   0.7767
## log(ast)      -0.12281   0.88443  0.18565 -0.661   0.5083
## log(platelet) -0.20906   0.81134  0.21326 -0.980   0.3269
## log(protime)   1.29429   3.64842  0.81966  1.579   0.1143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##                 exp(coef) exp(-coef) lower .95 upper .95
## trt               0.81121     1.2327   0.56646    1.1617
## age               1.02536     0.9753   1.00521    1.0459
## sexf              0.72110     1.3868   0.43401    1.1981
## ascites           1.50496     0.6645   0.98983    2.2882
## hepato            0.85318     1.1721   0.54111    1.3453
## spiders           1.05110     0.9514   0.71123    1.5534
## edema             2.38406     0.4195   1.36373    4.1678
## stage             1.40009     0.7142   0.98528    1.9895
## log(bili)         2.44811     0.4085   1.92770    3.1090
## log(albumin)      0.06583    15.1909   0.02147    0.2019
## log(alk.phos)     1.05434     0.9485   0.73146    1.5198
## log(ast)          0.88443     1.1307   0.61466    1.2726
## log(platelet)     0.81134     1.2325   0.53417    1.2323
## log(protime)      3.64842     0.2741   0.73182   18.1888
##
## Concordance= 0.889  (se = 0.014 )
## Likelihood ratio test= 360.6  on 14 df,   p=<2e-16
## Wald test            = 273.6  on 14 df,   p=<2e-16
## Score (logrank) test = 516.8  on 14 df,   p=<2e-16
```

```
  #general survival curve
autoplot(surv_fit(m2, data=pbcseq))
```



```
  #identifies bili, albumin, edema stage, and age as statistically signifigant variables
```

Figure 9. General survival curve from the cox regression model

```
  #can add different strata to view difference between groups
  #making a cut to view range of groups
pbcseq$bili3 <- cut(pbcseq$bili, c(0,1,2.5,40))
strata_m2 <- coxph(Surv(time1, time2, event) ~ trt + age + sex + ascites + hepato + spiders + edema + s
autoplot(survfit(strata_m2))
```

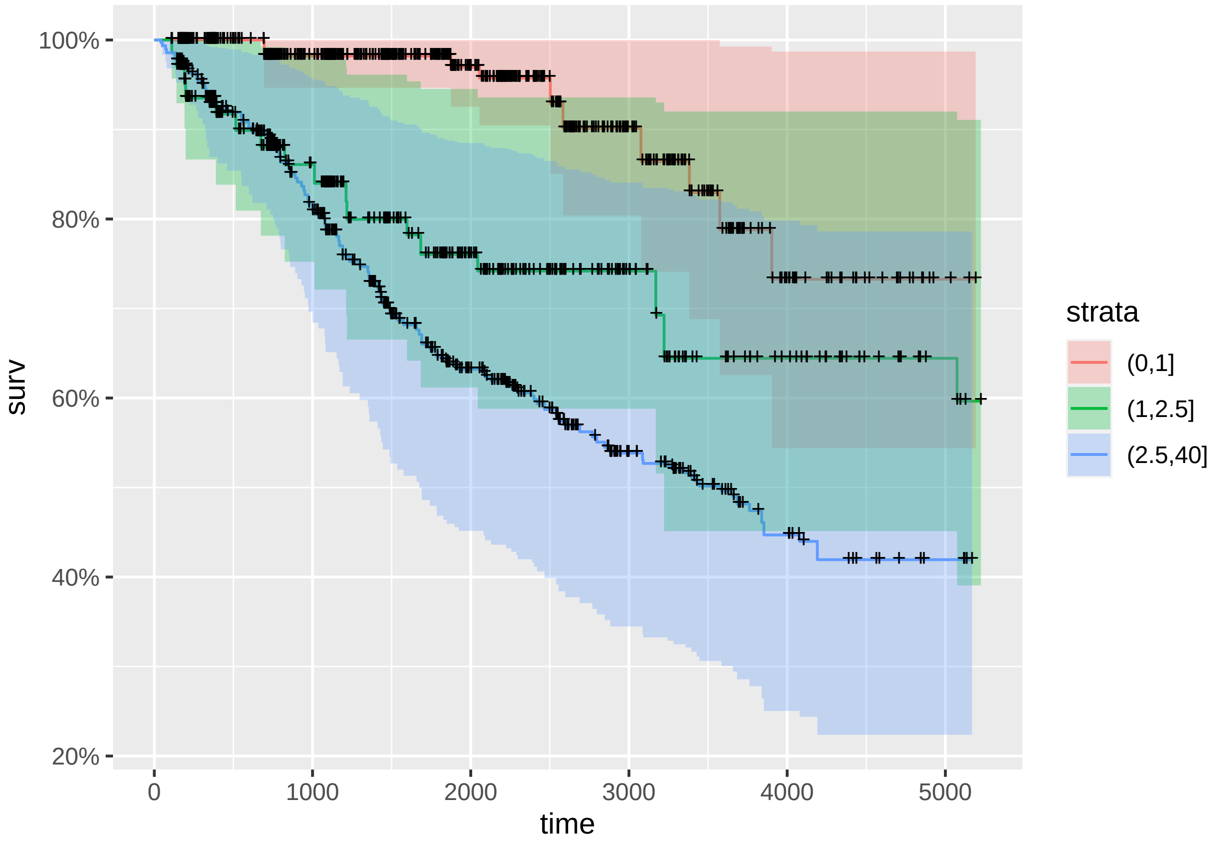


Figure 10. Survival curve with strata added for bilirubin

```
  #stratified survival curve for edema status
strata2_m2 <- coxph(Surv(time1, time2, event) ~ trt + age + sex + ascites + hepato + spiders + strata(e


autoplot(surv_fit(strata2_m2, data=pbcseq))
```

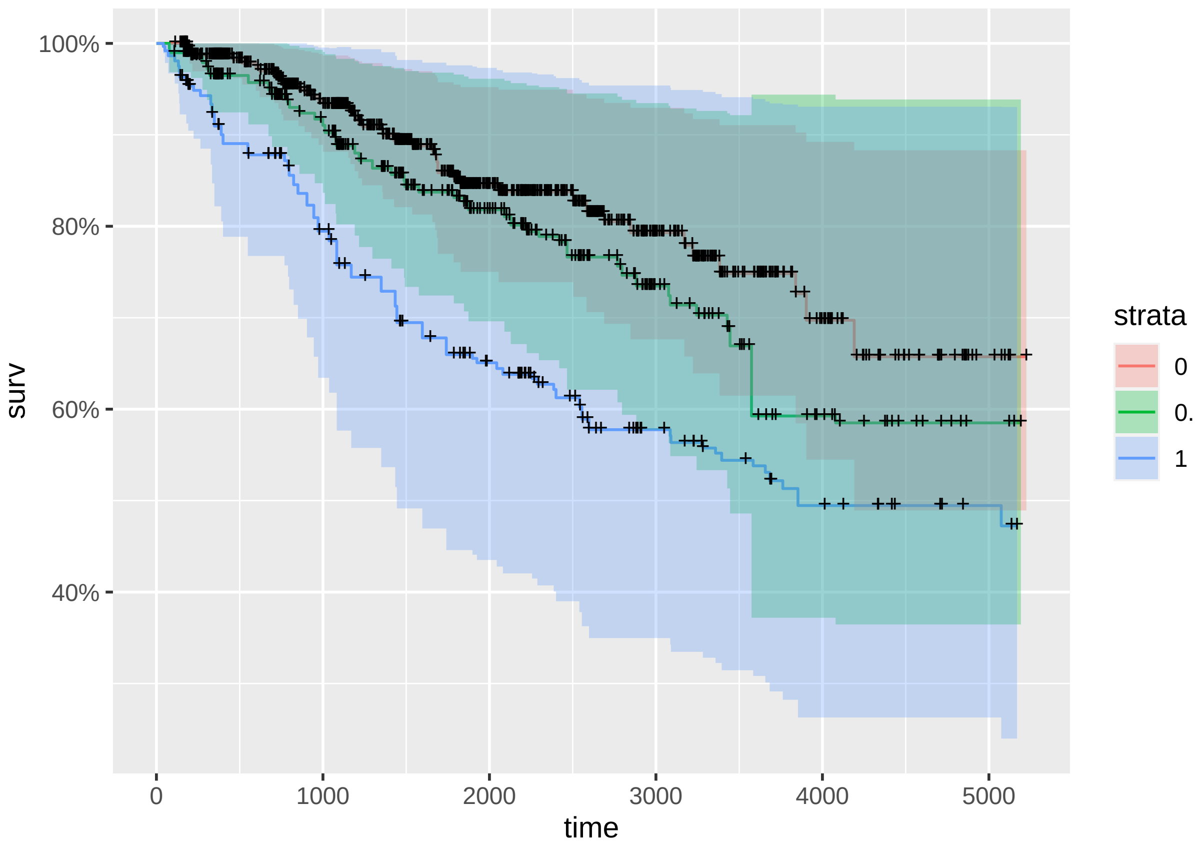


Figure 11. Survival curve with strata added for edema status

Question 2. Create a standard Cox regression model for survival using each of the variables as a predictor. Then use LASSO regression to identify a parsimonious set of variables predictive of survival. Propose what you believe to be the best model for the prediction of survival.

```
  #lasso with cox regression, start, stop, status triplet used
  #cross validation method for lambda selection

y <- Surv(time1, time2, event)
x <- model.matrix(y ~ trt + sex + ascites + hepato + spiders + edema + stage + age + log(bili) + log(al

  #glmnet model will not work unless low values of lambda are specified, otherwise
  #cv.glmnet will choose lambda > 10000 which imposes an absurdly strong penalty
  #function, leaving no params left in the model. Thus it is necessary to choose
  #lambda values manually
  #inspecting the algorithm with trace.it = 2, the models are indeed converging to a single
  #value but they are slightly off the target warm up number, triggering the
  #cox.fit algorithm did not converge warning

m4 <- cv.glmnet(x,y, family="cox", standardize = TRUE, lambda = c(0.5, 0.1, 0.05, 0.01, 0.005, 0.004, 0
coef(m4)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                          1
## (Intercept)   -3.916279372
## trt             .
## sexf          -0.049764971
## ascites        0.545429893
## hepato         .
## spiders        .
## edema          0.748103991
## stage          0.032920358
## age            0.024889741
## log(bili)      0.725316021
## log(albumin)  -2.092580395
## log(alk.phos)  .
## log(ast)       .
## log(platelet) -0.008910614
## log(protime)   1.517152966
```

```
  #from this model, the best predictors appear to be sex, ascites, edema, stage,
  #age, log(bili), log(albumin), log(platelet), log(protime). Of these factors,
  #only ascites, edema, log(bili), log(albumin), and log(protime) have
  #covariate scores greater than abs(0.1), while the other variables are still
  #included at the optimal value of lambda, their score has little effect
  #on the hazard ratio for a patient. Age should still be included as it
  #despite being 0.02 as it is not log transformed and ranges from 26 to 78.
  #comparing the two models, the penalized model with LASSO regression has
  #4 variables with coeffecients larger than 0.1, which greatly reduces
  #the complexity of the model. The lower model complexity can reduce potential
  #overfitting present in the unpenalized cox regression model,
```

```
plot(m4)
```

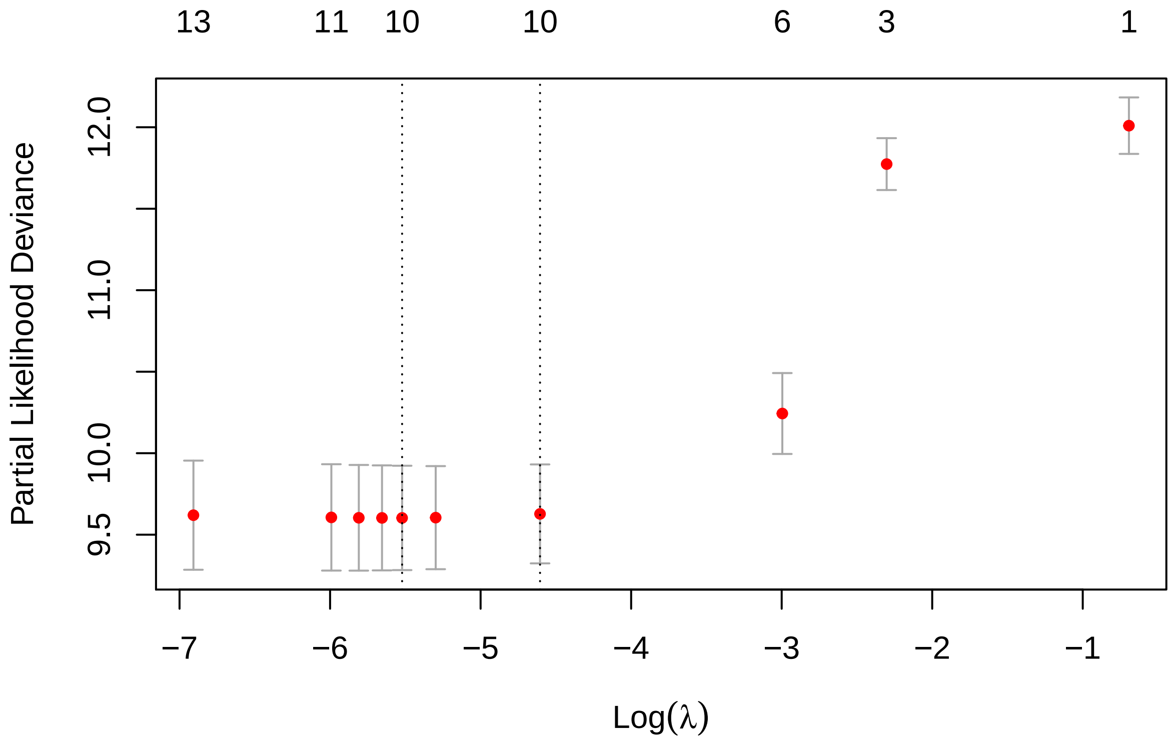


Figure 12. Partial likelihood Deviance for different lambda selections

```
  #survival curve for a subject with covariates equal to the means of each variable.
  #glmnet is not able to produce confidence intervals
plot(survival::survfit(m4, s = "lambda.min", x = x, y = y))
```

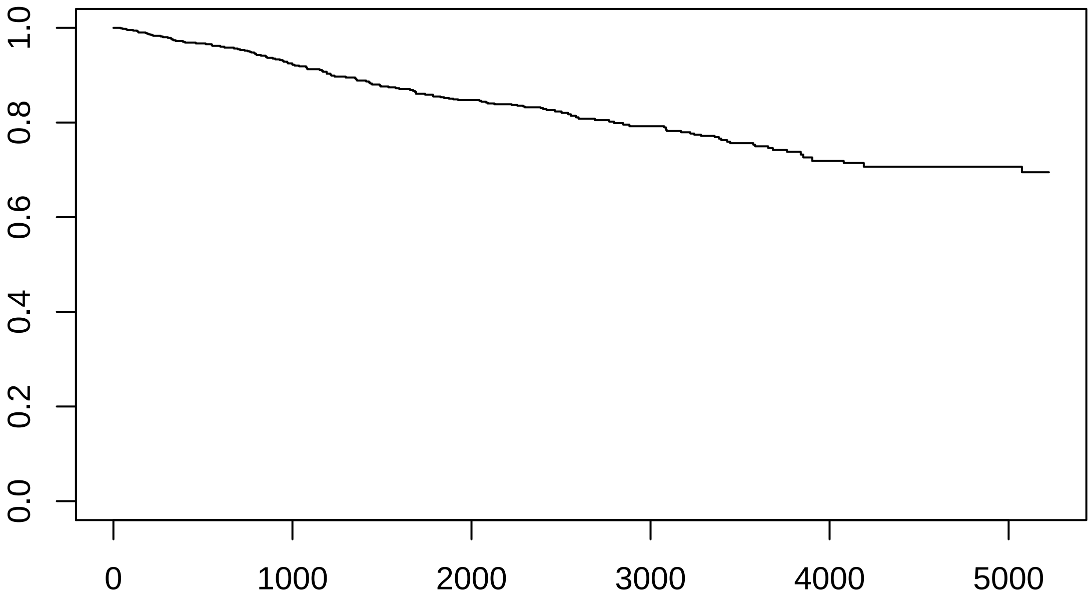
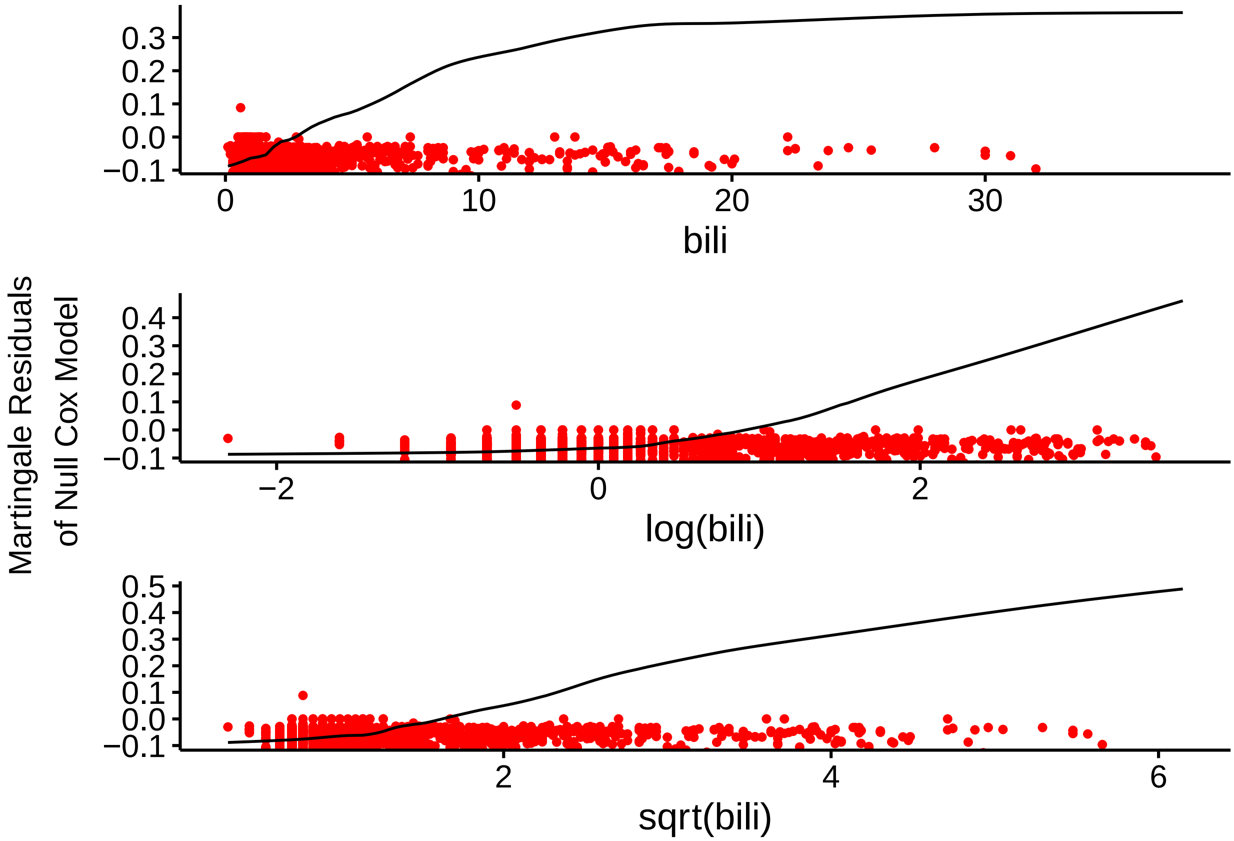


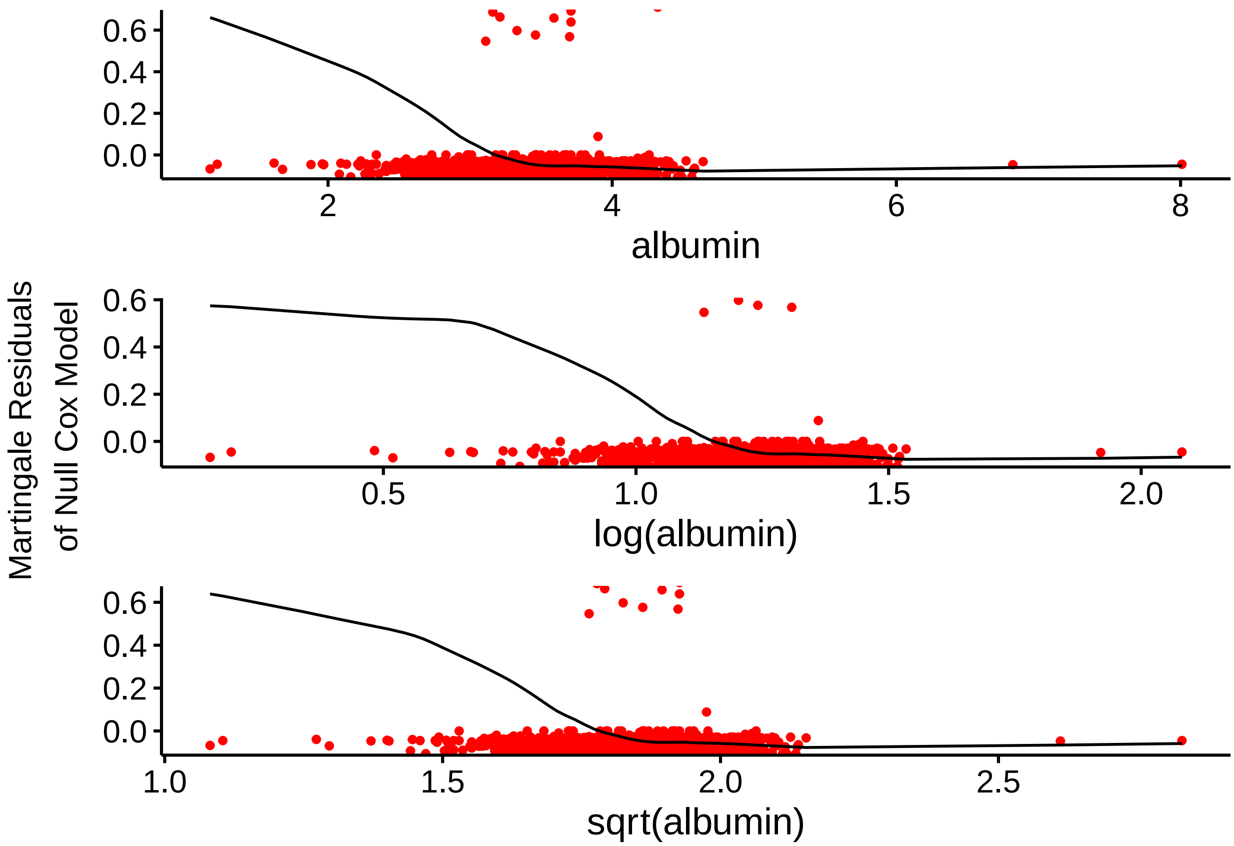
Figure 13. Survival curve generated from the L1 regularized glmnet model.

Question 3. Employ a valid way to assess whether any variables may be non-linearly associated with the time to event outcome.

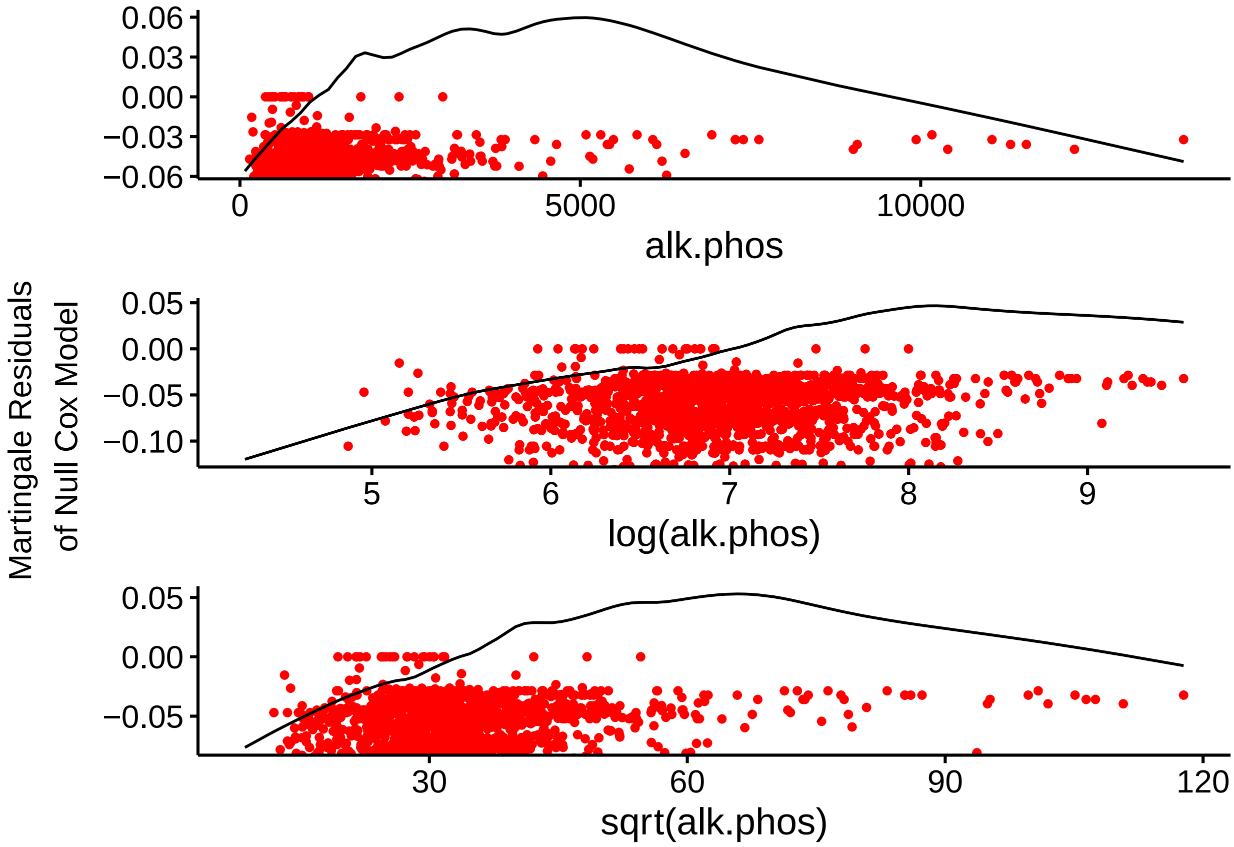
```
ggcoxfunctional(Surv(time1, time2, event) ~ bili + log(bili) + sqrt(bili), data = pbcseq) #close to lin
```



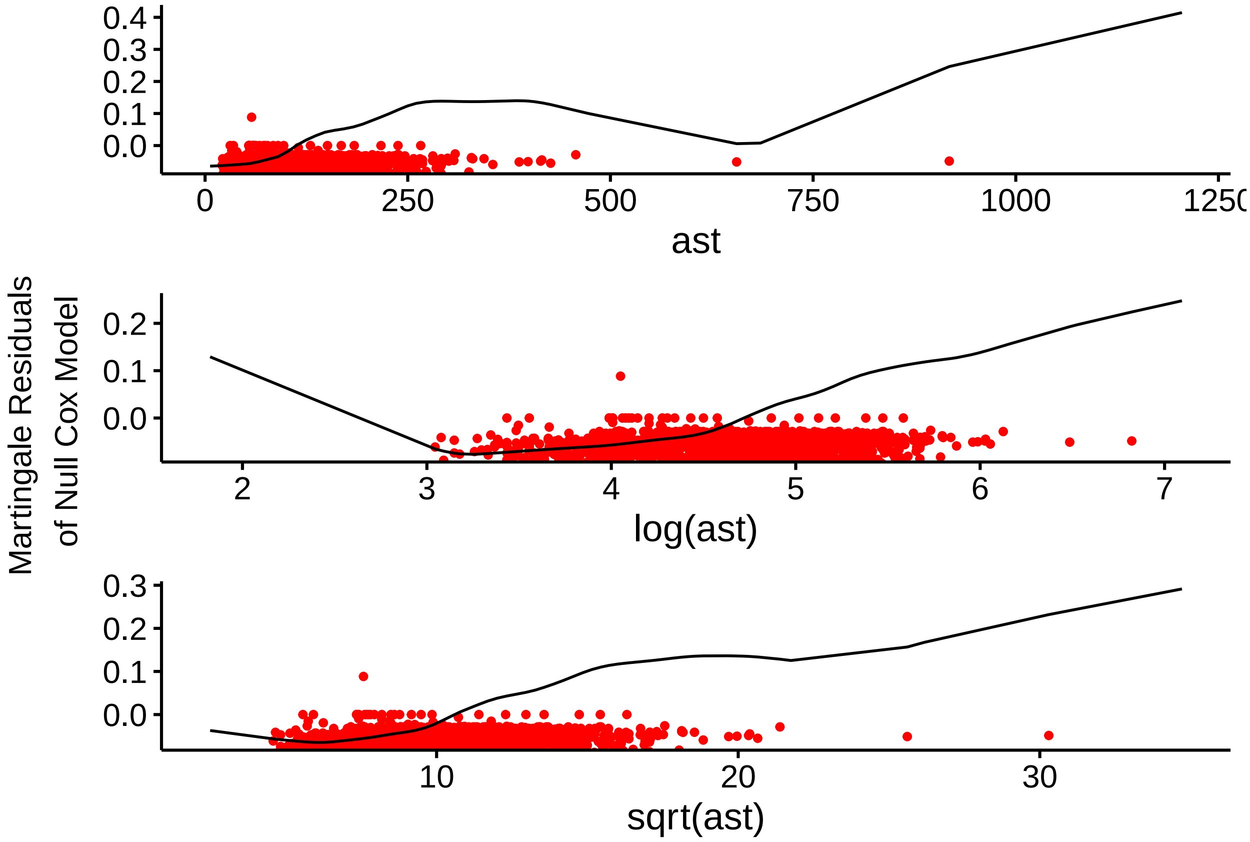
```
ggcoxfunctional(Surv(time1, time2, event) ~ albumin + log(albumin) + sqrt(albumin), data = pbcseq) #lin
```



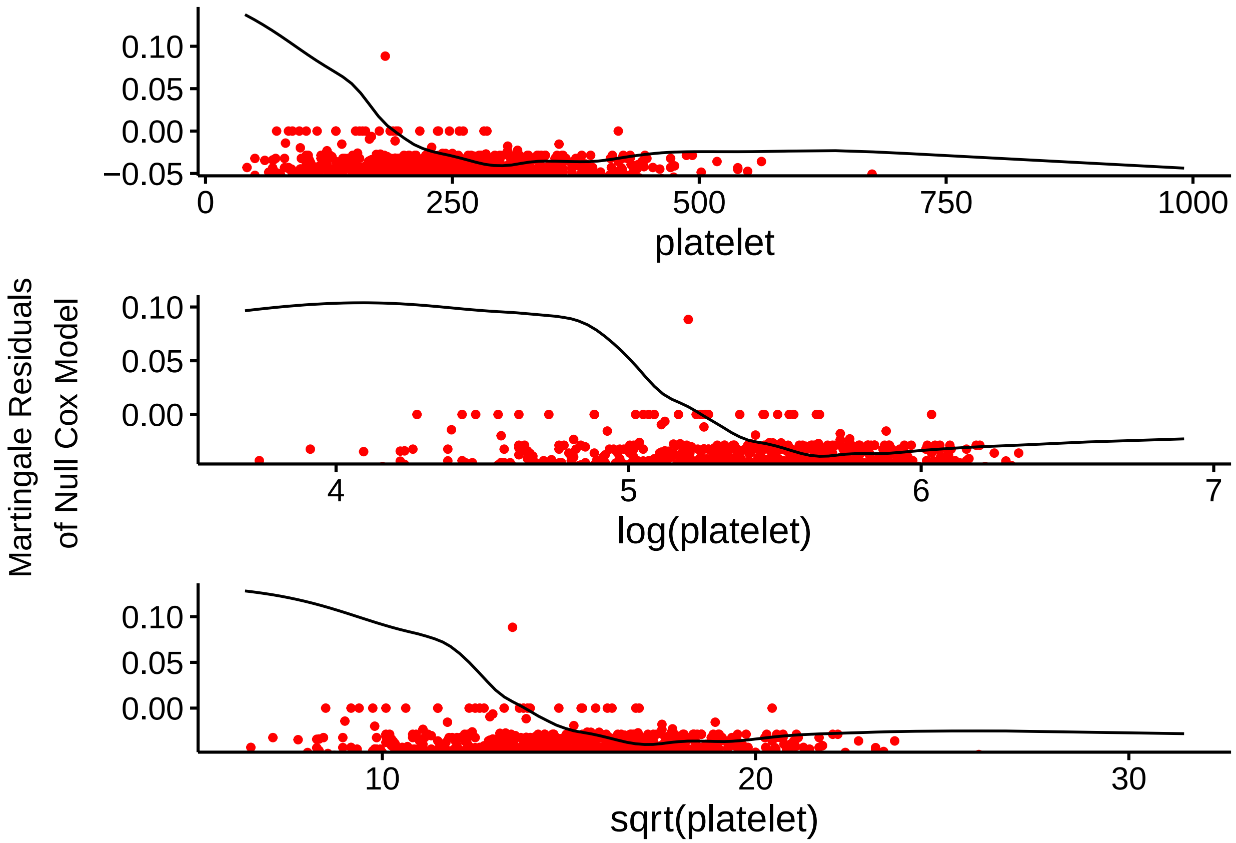
```
ggcoxfunctional(Surv(time1, time2, event) ~ alk.phos + log(alk.phos) + sqrt(alk.phos), data = pbcseq) #
```



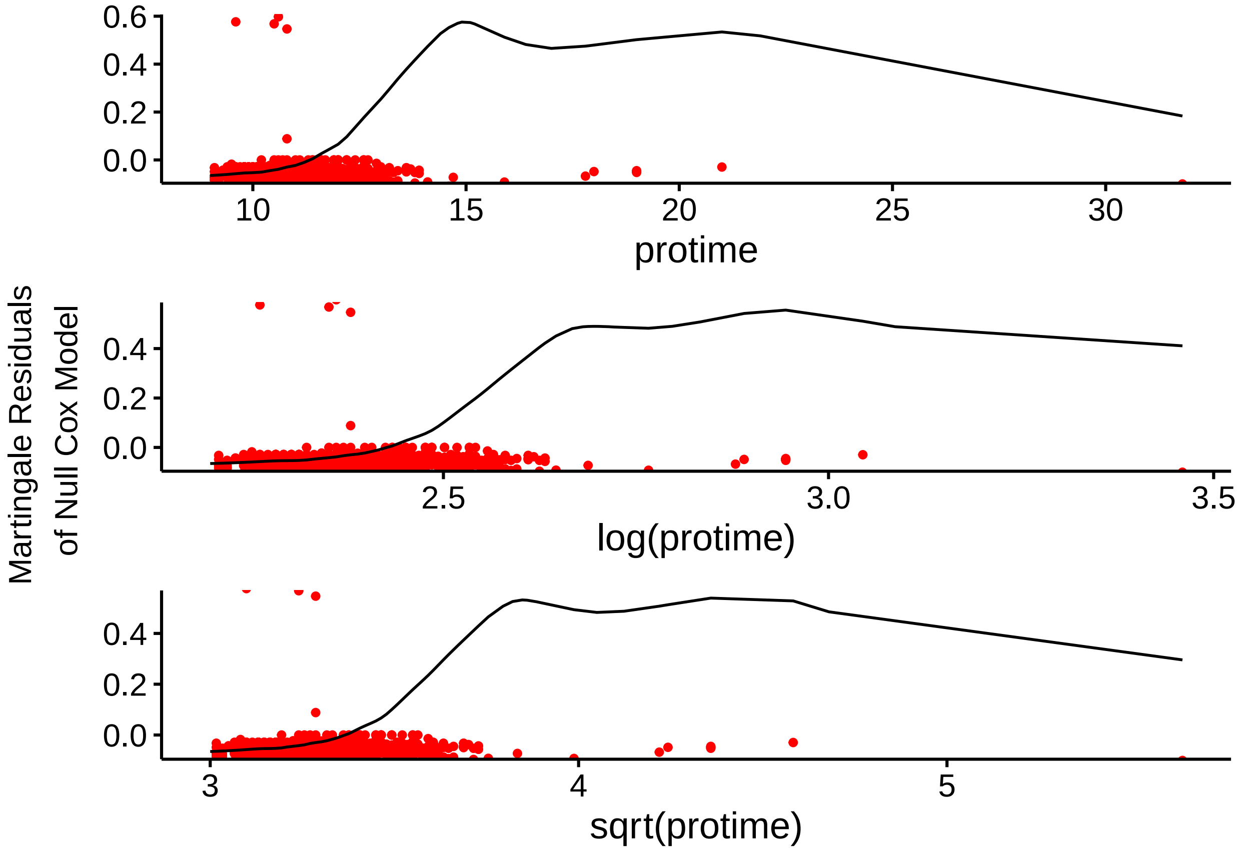
```
ggcoxfunctional(Surv(time1, time2, event) ~ ast + log(ast) + sqrt(ast), data = pbcseq) #not linear, log
```



```
ggcoxfunctional(Surv(time1, time2, event) ~ platelet + log(platelet) + sqrt(platelet), data = pbcseq) #
```



```r
ggcoxfunctional(Surv(time1, time2, event) ~ protime + log(protime) + sqrt(protime), data = pbcseq) #not
```



```
ggcoxfunctional(Surv(time1, time2, event) ~ age + log(age) + sqrt(age), data = pbcseq) #linear or close
```

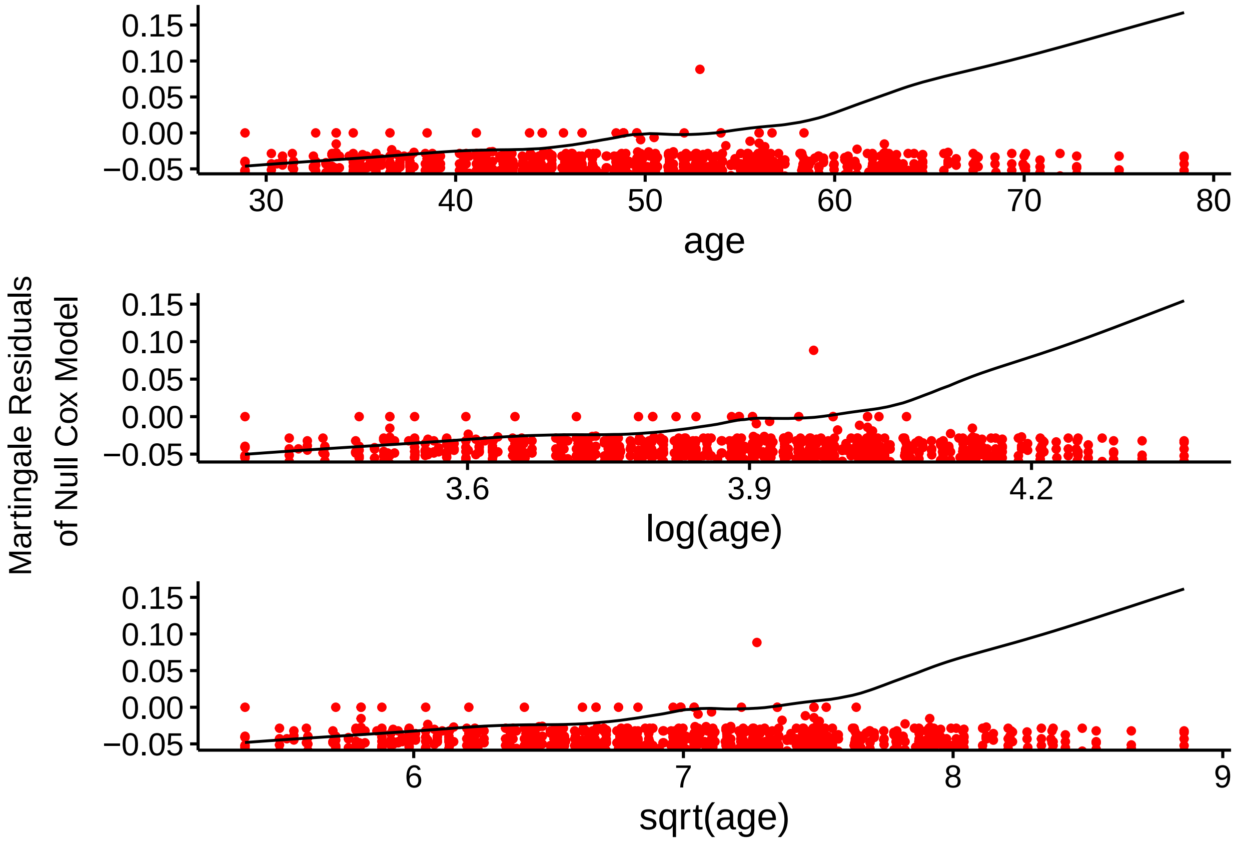


Figure 14 - 20. Plots of Martingale residuals against the continous covariates. The closer the curve is to a linear form (consistent slope, generally monotonic increase or decrease), the closer the functional form is to showing a linear association between the covariate and the hazards ratio.

Question 4. Create a figure that you feel adequately displays the results of your analysis that you have found

```
  #survival curve for two subjects close to the same age, one with edema and ascites and high
  #bilirubin, the other with low bilirubin, no edema and no ascites, the patient with
  #presence of edema and ascites, high bilirubin creates very strong hazard ratio

x[2:3,]
```

```
##   (Intercept) trt sexf ascites hepato spiders edema stage      age  log(bili)
## 2           1   1    1       1      1       1     1     4 58.76523 3.05870707
## 3           1   1    1       0      1       1     0     3 56.44627 0.09531018
##   log(albumin) log(alk.phos) log(ast) log(platelet) log(protime)
## 2     1.078410      7.385231 1.824549      5.209486     2.415914
## 3     1.420696      8.908559 4.731803      5.398163     2.360854
```

```
plot(survival::survfit(m4, s = "lambda.min", x = x, y = y, newx = x[2:3,]))
```

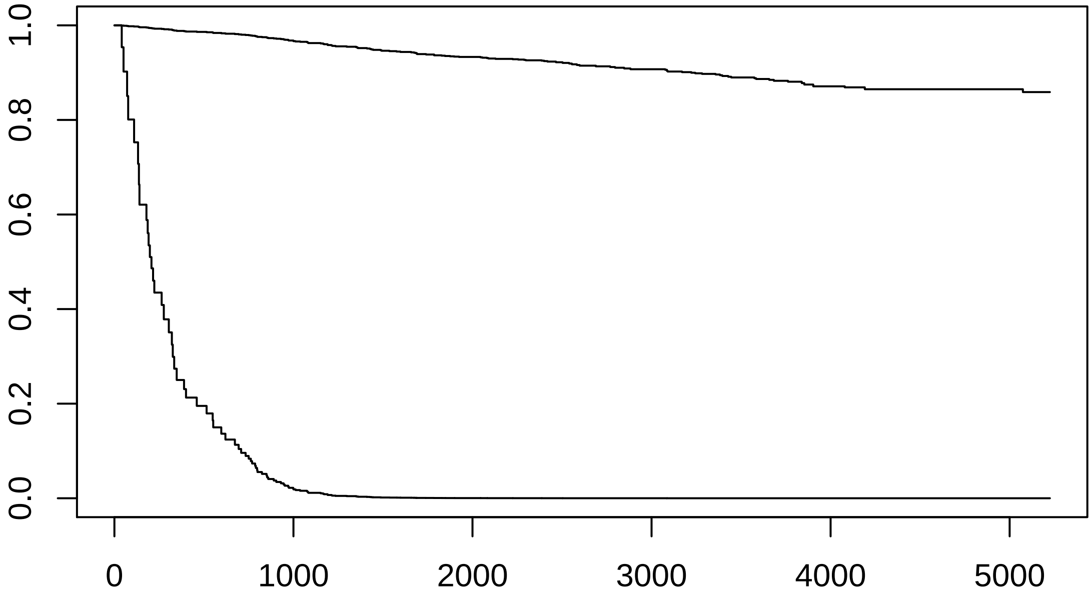


Figure 21. Survival curve of two patients described above, the patient with edema, ascites, and high bilirubin has a much higher hazard ratio than the patient with no edema, ascites and low bilirubin.

```
strata_m2 <- coxph(Surv(time1, time2, event) ~ strata(trt) + age + sex + ascites + hepato + spiders + e

  #looks as if the treatment makes little difference compared
  #to no treatment

autoplot(survfit(strata_m2))
```

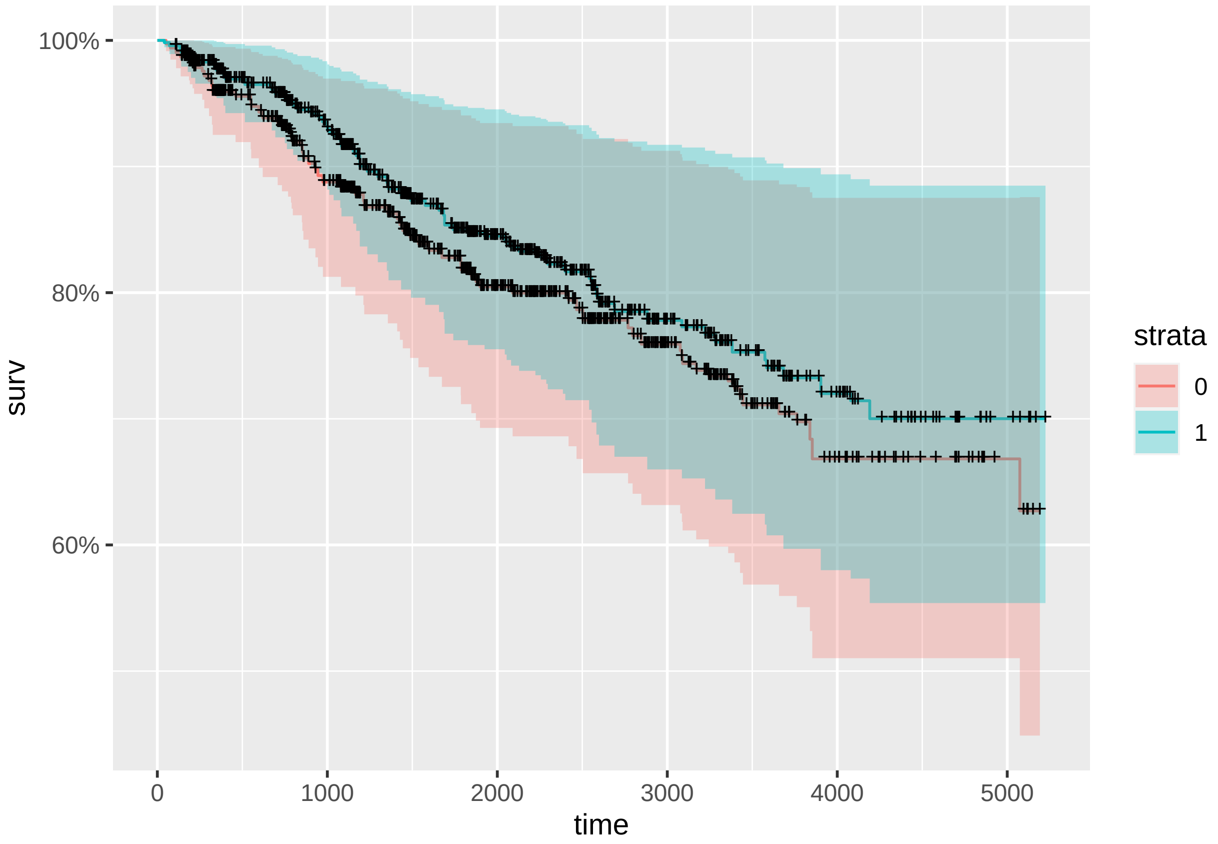


Figure 22. Two survival curves based on the strata of the treatment levels. Treatment makes little to no difference compared to no treatment.