# Survival Analysi Final Project

*Tonghao Zhang, David Bennett, Suman Zhuo*

Data can be found at: http://lib.stat.cmu.edu/datasets/pbc

```
library(survival)
library(KMsurv)
library(MASS)
pbc<-read.table("pbc.txt", na.strings = ".")
colnames(pbc)<-c("id", "fu", "status", "drug", "age", "sex", "ascites", "hepatom", "spiders", "edema",
pbc<-na.omit(pbc)
rownames(pbc)<-c(1:276)
# The original data set was had two different classifications of censored observations (1 and 0). We wa
pbc$status<-ifelse(pbc$status==2,1,0)
sum(pbc$status) #There are 111 deaths and 276-111= 165 censored observations.
```

```
## [1] 111
```

```
#We don't care about the id.
# id: Unique Identification Number
pbc$fu<-pbc$fu/365 #The number in the original data set is the number of days to death or censoring. To
# fu: Follow-Up (Days)
# status: 1 = death, 0 = censored
pbc$drug<-pbc$drug-1
# drug: 0 = Penicillamine, 1 = Placebo
pbc$age<-pbc$age/365 #The number in the original data set is the number of days between birth and study
# age: Years
# sex: 0 = Male; 1 = Female
# ascites: Presence of ascites: 0 = No; 1 = Yes
# hepatom: Presence of Hepatomegaly: 0 = No; 1 = Yes
# spiders: Presence of Spiders: 0 = No; 1 = Yes
# edema: presence of Edema: 0 = No; 0.5 = Edema Present w/out diuretics; 1 = Edema present despite diur
# bili: Amount of serum bilirubin in mg/dl
# chol: Amount of cholesterol in mg/dl
# albumin: Amount of albumin in gm/dl
# copper: Amount of urine copper in ug/day
# alkphos: Amount of alkaline Phosphate in U/liter
# sgot: SGOT in U/ml
# trig: Amount of triglicerides in mg/dl
# platelet: Amount of platelets per cubic ml/1000
# protime: Prothrombin time in seconds
# stage: Histologic stage of disease
```

Here the assumption of interst is whether the drug will improve patients' survival possibility. First we decided to fit a Cox Proportional Hazard model.

## Question 1

### Which of the predictors should be included in the model selection?

We want to use AIC method to do a model selection to see which covariates are significant in the Cox Proportional Hazard Model. The assumption of interest is about drug, so it must be included. Using function `stepAIC`, the model selection went backwards from all main effects and and two-way interactions with `drug` to lower model containing only `drug` indicator. One thing worth noting is that we have to make sure certain covariates to be in or out of the model as a whole (for example, if backgroup knowledge indicates potential intercation effect between drug and gender, we have to make sure these three covariates to be included in or excluded from the model at the same time).

```
ovfit<-coxph(Surv(fu,status)~drug+sex+ascites+hepatom+spiders+as.factor(edema)+bili+chol+albumin+copper+
selection<-stepAIC(ovfit, scope = list(upper=.~.+drug*sex+drug*ascites+drug*hepatom+drug*spiders+drug*as
```

```
## Start:  AIC=975
## Surv(fu, status) ~ drug + sex + ascites + hepatom + spiders + 
##     as.factor(edema) + bili + chol + albumin + copper + alkphos + 
##     sgot + trig + platelet + protime + as.factor(stage)
## 
##                        Df    AIC
## + drug:as.factor(edema)  2 971.42
## - hepatom               1 973.00
## - alkphos               1 973.02
## - ascites               1 973.04
## - spiders               1 973.08
## - platelet              1 973.38
## - chol                  1 973.39
## + drug:hepatom          1 973.85
## - trig                  1 973.85
## - sgot                  1 974.91
## <none>                    975.00
## - as.factor(edema)      2 976.09
## + drug:ascites          1 976.17
## + drug:sex              1 976.54
## + drug:spiders          1 976.61
## - sex                   1 977.07
## - copper                1 977.66
## - protime               1 979.15
## - as.factor(stage)      3 980.21
## - albumin               1 981.52
## - bili                  1 984.12
## 
## Step:  AIC=971.42
## Surv(fu, status) ~ drug + sex + ascites + hepatom + spiders + 
##     as.factor(edema) + bili + chol + albumin + copper + alkphos + 
##     sgot + trig + platelet + protime + as.factor(stage) + drug:as.factor(edema)
## 
##                        Df    AIC
## + drug:ascites          1 967.11
## + drug:hepatom          1 969.37
## - alkphos               1 969.42
## - platelet              1 969.46
## - hepatom               1 969.48
## - trig                  1 969.53
## - ascites               1 969.61
## - spiders               1 969.67
## - chol                  1 970.27
## - sgot                  1 970.86
## <none>                    971.42
## + drug:sex              1 972.50
## - sex                   1 972.74
## + drug:spiders           1 973.38
## - copper                 1 974.97
## - drug:as.factor(edema)  2 975.00
## - as.factor(stage)       3 976.39
## - protime                1 977.98
## - albumin                1 978.37
## - bili                   1 981.61
##
## Step:  AIC=967.11
## Surv(fu, status) ~ drug + sex + ascites + hepatom + spiders +
##     as.factor(edema) + bili + chol + albumin + copper + alkphos +
##     sgot + trig + platelet + protime + as.factor(stage) + drug:as.factor(edema) +
##     drug:ascites
##
##                         Df    AIC
## - alkphos                1 965.11
## - hepatom                1 965.11
## - trig                   1 965.11
## + drug:hepatom           1 965.22
## - platelet               1 965.34
## - chol                   1 965.74
## - spiders                1 966.00
## - sgot                   1 966.96
## <none>                     967.11
## + drug:sex               1 968.19
## - protime                1 968.44
## + drug:spiders           1 969.10
## - copper                 1 969.52
## - as.factor(stage)       3 969.88
## - sex                    1 970.09
## - drug:ascites           1 971.42
## - albumin                1 974.85
## - drug:as.factor(edema)  2 976.17
## - bili                   1 978.17
##
## Step:  AIC=965.11
## Surv(fu, status) ~ drug + sex + ascites + hepatom + spiders +
##     as.factor(edema) + bili + chol + albumin + copper + sgot +
##     trig + platelet + protime + as.factor(stage) + drug:as.factor(edema) +
##     drug:ascites
##
##                         Df    AIC
## - hepatom                1 963.11
## - trig                   1 963.11
## + drug:hepatom           1 963.28
## - platelet               1 963.35
## - chol                   1 963.74
## - spiders                1 964.03
## - sgot                   1 964.96
## <none>                     965.11
## + drug:sex               1 966.21
## - protime                1 966.44
## + drug:spiders           1 967.10
## + alkphos                1 967.11
## - copper                 1 967.78
## - as.factor(stage)       3 967.99
## - sex                    1 968.09
## - drug:ascites           1 969.42
## - albumin                1 972.97
## - drug:as.factor(edema)  2 974.21
## - bili                   1 976.20
##
## Step:  AIC=963.11
## Surv(fu, status) ~ drug + sex + ascites + spiders + as.factor(edema) +
##     bili + chol + albumin + copper + sgot + trig + platelet +
##     protime + as.factor(stage) + drug:as.factor(edema) + drug:ascites
##
##                         Df    AIC
## - trig                   1 961.11
## - platelet               1 961.36
## - chol                   1 961.77
## - spiders                1 962.04
## - sgot                   1 962.98
## <none>                     963.11
## + drug:sex               1 964.22
## - protime                1 964.45
## + drug:spiders           1 965.10
## + hepatom                1 965.11
## + alkphos                1 965.11
## - copper                 1 965.82
## - sex                    1 966.09
## - drug:ascites           1 967.48
## - as.factor(stage)       3 968.13
## - albumin                1 970.97
## - drug:as.factor(edema)  2 972.21
## - bili                   1 974.38
##
## Step:  AIC=961.11
## Surv(fu, status) ~ drug + sex + ascites + spiders + as.factor(edema) +
##     bili + chol + albumin + copper + sgot + platelet + protime +
##     as.factor(stage) + drug:as.factor(edema) + drug:ascites
##
##                         Df    AIC
## - platelet               1 959.40
## - chol                   1 959.78
## - spiders                1 960.04
## - sgot                   1 960.98
## <none>                     961.11
## + drug:sex               1 962.23
## - protime                1 962.52
## + drug:spiders           1 963.10
## + trig                   1 963.11
## + hepatom                1 963.11
## + alkphos                1 963.11
## - copper                 1 963.83
## - sex                    1 964.13
## - drug:ascites           1 965.58
## - as.factor(stage)       3 966.14
## - albumin                 1 969.07
## - drug:as.factor(edema)  2 970.85
## - bili                    1 974.44
##
## Step:  AIC=959.4
## Surv(fu, status) ~ drug + sex + ascites + spiders + as.factor(edema) +
##     bili + chol + albumin + copper + sgot + protime + as.factor(stage) +
##     drug:as.factor(edema) + drug:ascites
##
##                          Df    AIC
## - chol                    1 957.91
## - spiders                 1 958.41
## <none>                      959.40
## - sgot                    1 959.79
## + drug:sex                1 960.38
## - protime                 1 960.80
## + platelet                1 961.11
## + trig                    1 961.36
## + drug:spiders            1 961.38
## + hepatom                 1 961.39
## + alkphos                 1 961.39
## - copper                  1 961.84
## - sex                     1 963.51
## - drug:ascites            1 963.65
## - as.factor(stage)        3 965.03
## - albumin                 1 967.43
## - drug:as.factor(edema)  2 968.98
## - bili                    1 972.46
##
## Step:  AIC=957.91
## Surv(fu, status) ~ drug + sex + ascites + spiders + as.factor(edema) +
##     bili + albumin + copper + sgot + protime + as.factor(stage) +
##     drug:as.factor(edema) + drug:ascites
##
##                          Df    AIC
## - spiders                 1 957.14
## <none>                      957.91
## + drug:sex                1 958.73
## - sgot                    1 958.89
## - protime                 1 959.24
## + chol                    1 959.40
## + platelet                1 959.78
## + drug:spiders            1 959.85
## + trig                    1 959.86
## + hepatom                 1 959.87
## + alkphos                 1 959.90
## - copper                  1 960.00
## - sex                     1 961.97
## - drug:ascites            1 962.44
## - as.factor(stage)        3 963.21
## - albumin                 1 966.23
## - drug:as.factor(edema)  2 967.48
## - bili                    1 976.53
##
```

```
## Step:  AIC=957.14
## Surv(fu, status) ~ drug + sex + ascites + as.factor(edema) + 
##     bili + albumin + copper + sgot + protime + as.factor(stage) + 
##     drug:as.factor(edema) + drug:ascites
## 
##                         Df    AIC
## <none>                     957.14
## + spiders                1 957.91
## + drug:sex               1 958.05
## - sgot                   1 958.05
## + chol                   1 958.41
## + platelet               1 958.98
## - protime                1 959.05
## + alkphos                1 959.07
## + hepatom                1 959.07
## + trig                   1 959.13
## - copper                 1 960.08
## - sex                    1 960.43
## - drug:ascites           1 960.81
## - albumin                1 965.03
## - as.factor(stage)       3 965.25
## - drug:as.factor(edema)  2 965.61
## - bili                   1 976.62
```

```
summary(selection)
```

```
## Call:
## coxph(formula = Surv(fu, status) ~ drug + sex + ascites + as.factor(edema) + 
##     bili + albumin + copper + sgot + protime + as.factor(stage) + 
##     drug:as.factor(edema) + drug:ascites, data = pbc, method = "breslow")
## 
##   n= 276, number of events= 111 
## 
##                              coef exp(coef)  se(coef)      z Pr(>|z|)    
## drug                    -0.271495  0.762239  0.235906 -1.151  0.24979    
## sex                     -0.674390  0.509467  0.279493 -2.413  0.01583 *  
## ascites                 -0.932187  0.393692  0.514417 -1.812  0.06997 .  
## as.factor(edema)0.5      0.084399  1.088063  0.394628  0.214  0.83065    
## as.factor(edema)1        2.515699 12.375254  0.624830  4.026 5.67e-05 ***
## bili                     0.102598  1.108045  0.020572  4.987 6.13e-07 ***
## albumin                 -0.922720  0.397437  0.289200 -3.191  0.00142 ** 
## copper                   0.002435  1.002438  0.001060  2.298  0.02155 *  
## sgot                     0.003289  1.003295  0.001849  1.779  0.07519 .  
## protime                  0.251222  1.285596  0.125019  2.009  0.04449 *  
## as.factor(stage)2        1.422204  4.146247  1.057958  1.344  0.17885    
## as.factor(stage)3        1.759089  5.807145  1.027237  1.712  0.08681 .  
## as.factor(stage)4        2.271380  9.692764  1.027595  2.210  0.02708 *  
## drug:as.factor(edema)0.5  0.634976  1.886977  0.649903  0.977  0.32855    
## drug:as.factor(edema)1   -2.247824  0.105629  0.746751 -3.010  0.00261 ** 
## drug:ascites             1.719046  5.579203  0.730032  2.355  0.01854 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                         exp(coef) exp(-coef) lower .95 upper .95
## drug                       0.7622    1.31192   0.48005    1.2103
## sex                         0.5095    1.96284   0.29459    0.8811
## ascites                     0.3937    2.54006   0.14364    1.0790
## as.factor(edema)0.5         1.0881    0.91906   0.50205    2.3581
## as.factor(edema)1          12.3753    0.08081   3.63662   42.1124
## bili                        1.1080    0.90249   1.06426    1.1536
## albumin                     0.3974    2.51612   0.22548    0.7005
## copper                      1.0024    0.99757   1.00036    1.0045
## sgot                        1.0033    0.99672   0.99967    1.0069
## protime                     1.2856    0.77785   1.00621    1.6426
## as.factor(stage)2           4.1462    0.24118   0.52134   32.9754
## as.factor(stage)3           5.8071    0.17220   0.77549   43.4859
## as.factor(stage)4           9.6928    0.10317   1.29347   72.6336
## drug:as.factor(edema)0.5    1.8870    0.52995   0.52792    6.7447
## drug:as.factor(edema)1      0.1056    9.46711   0.02444    0.4565
## drug:ascites                5.5792    0.17924   1.33404   23.3333
##
## Concordance= 0.839  (se = 0.031 )
## Rsquare= 0.47   (max possible= 0.981 )
## Likelihood ratio test= 175.3  on 16 df,   p=0
## Wald test            = 173.9  on 16 df,   p=0
## Score (logrank) test = 341.9  on 16 df,   p=0
```

From computation, our Cox Proportional Hazard model will include **drug type, patient gender, presence of ascites, indicator of edema, bili, albumin, copper, sgot, protime, stage and the interaction between drug & edema, drug & ascites.**

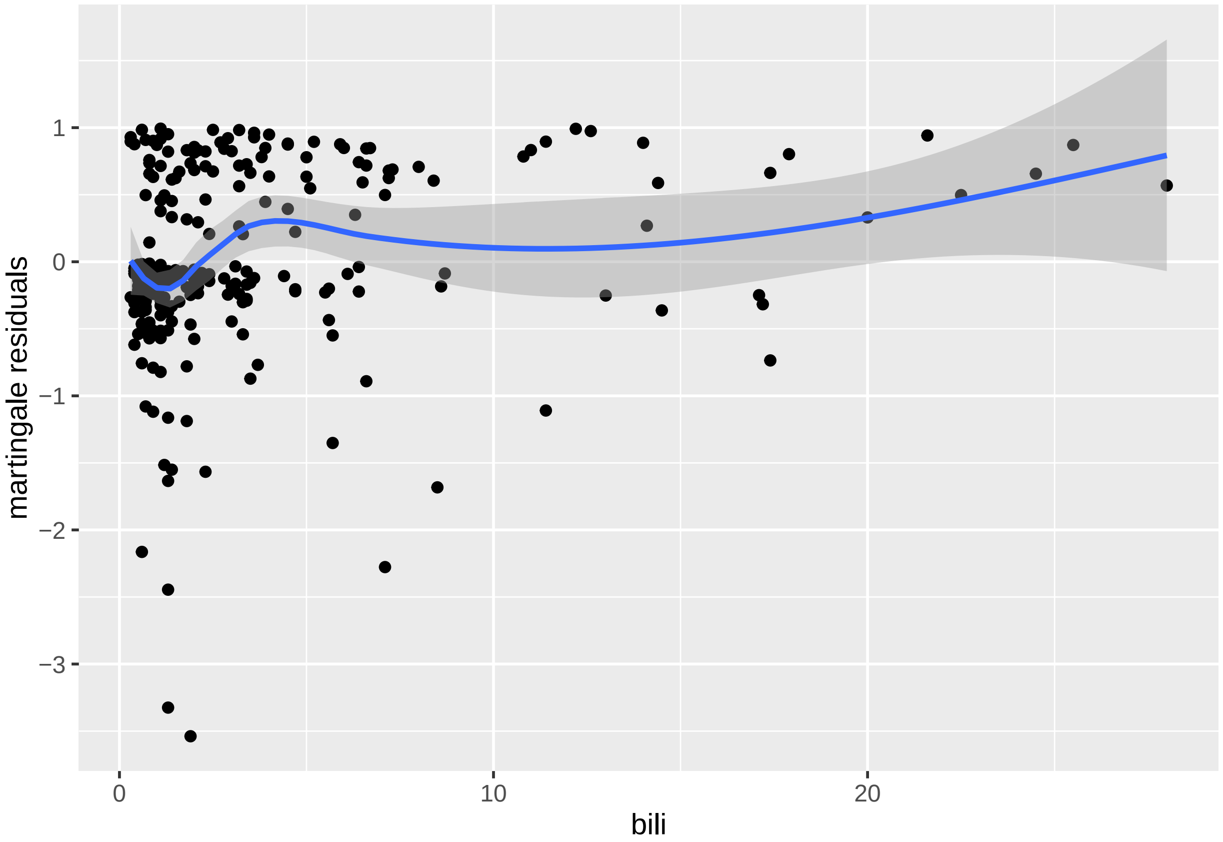
## Question 2

**Should any of the variables be stratified?**

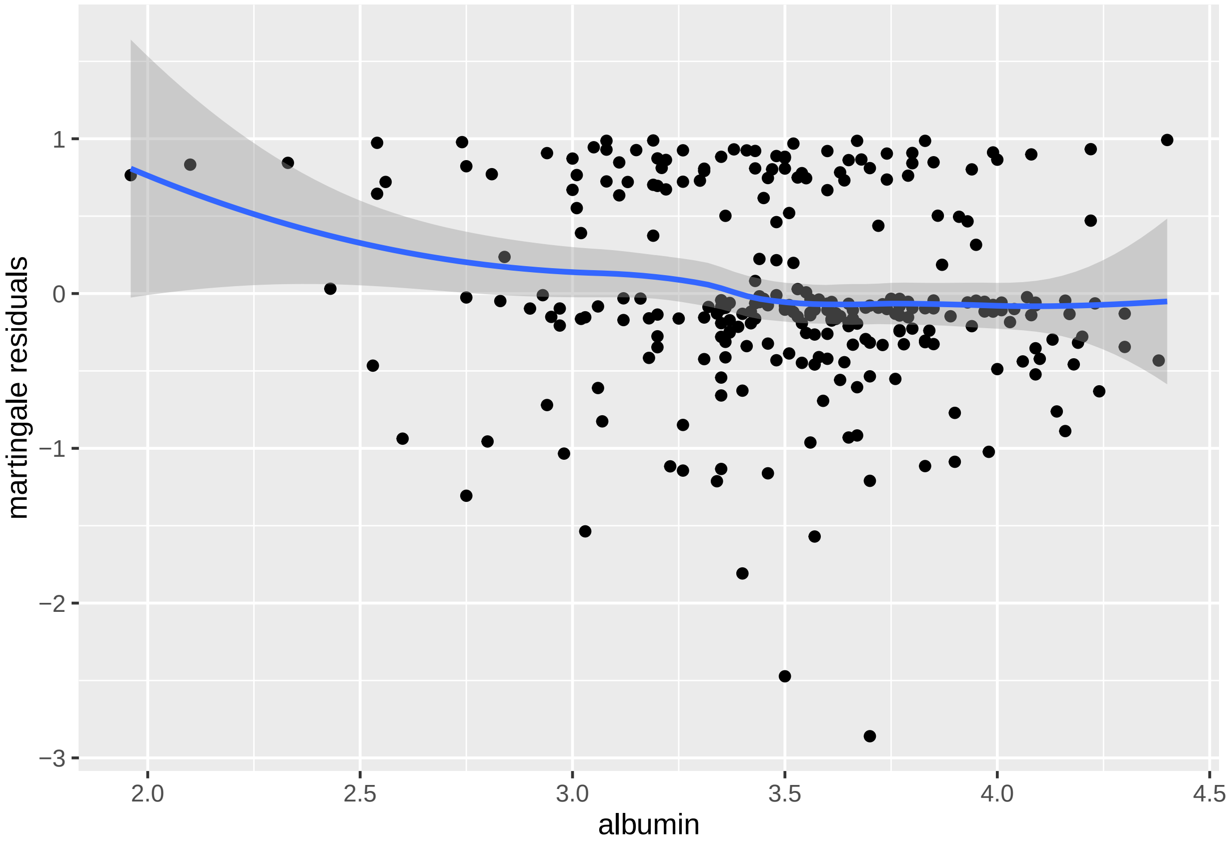
## Question 3

**Should any of the variables be transformed?**

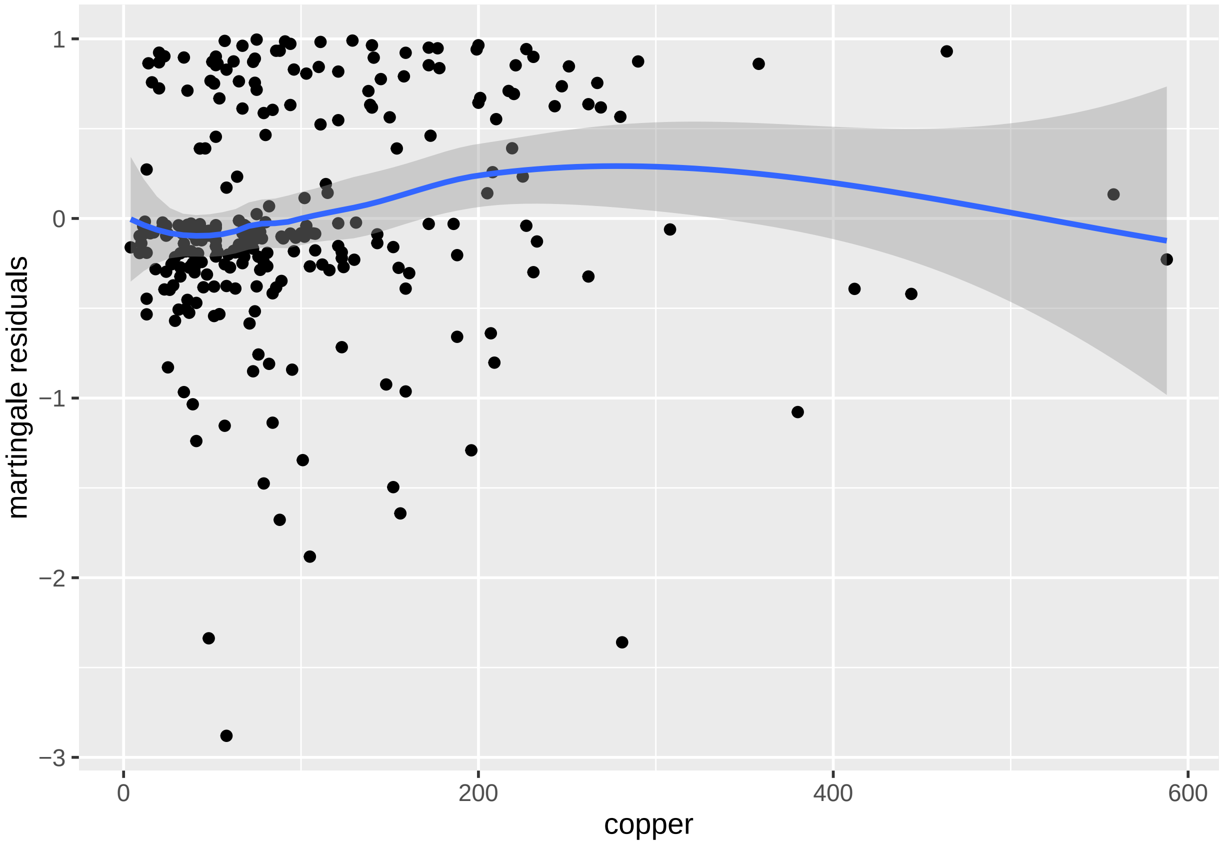
```
library(ggplot2)
model.bili<-coxph(Surv(fu, status)~drug+sex+ascites+as.factor(edema)+albumin+copper+sgot+protime+as.fac
mres.bili<-resid(model.bili)
data.bili<-data.frame(x=pbc$bili,y=mres.bili)
ggplot(data.bili,aes(x=x,y=y))+geom_point()+labs(x="bili",y="martingale residuals")+stat_smooth(method =
```



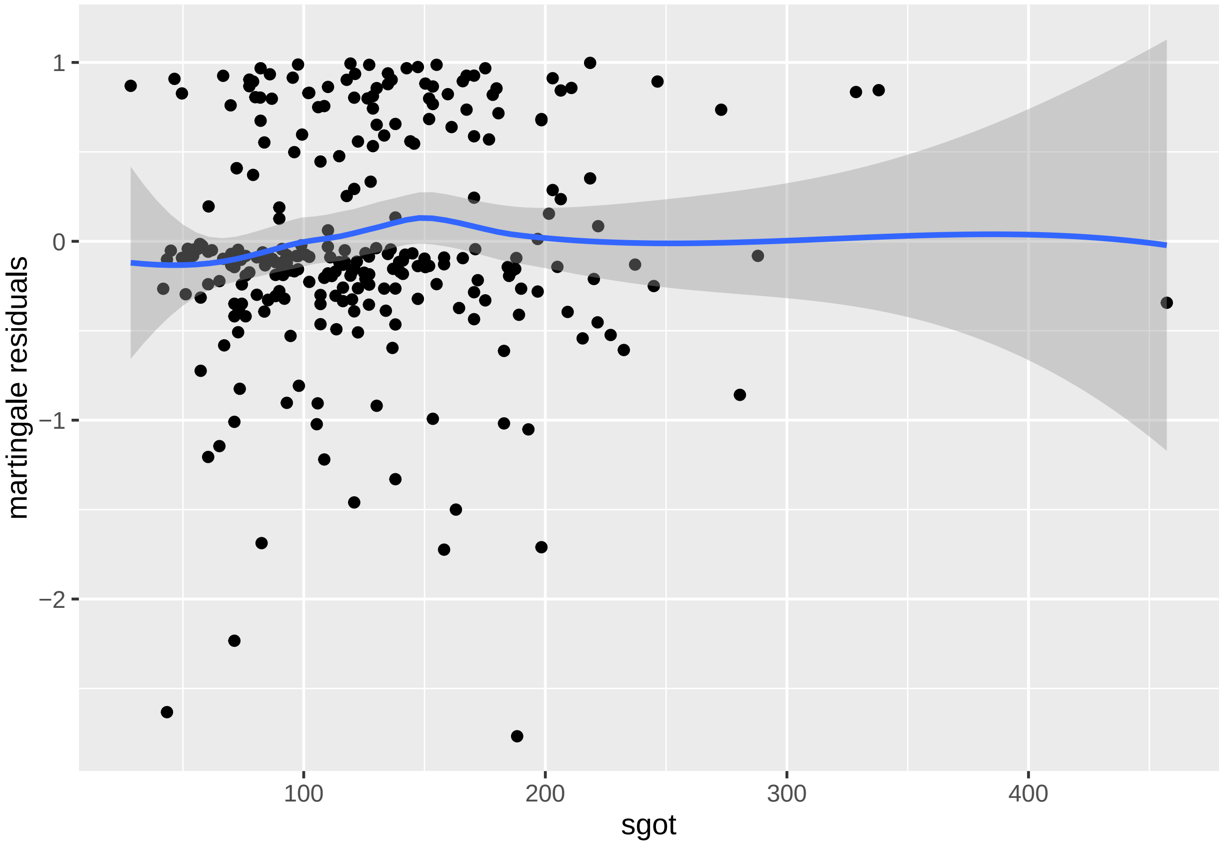
```
model.albumin<-coxph(Surv(fu, status)~drug+sex+ascites+as.factor(edema)+bili+copper+sgot+protime+as.fac
mres.albumin<-resid(model.albumin)
data.albumin<-data.frame(x=pbc$albumin,y=mres.albumin)
ggplot(data.albumin,aes(x=x,y=y))+geom_point()+labs(x="albumin",y="martingale residuals")+stat_smooth(m
```



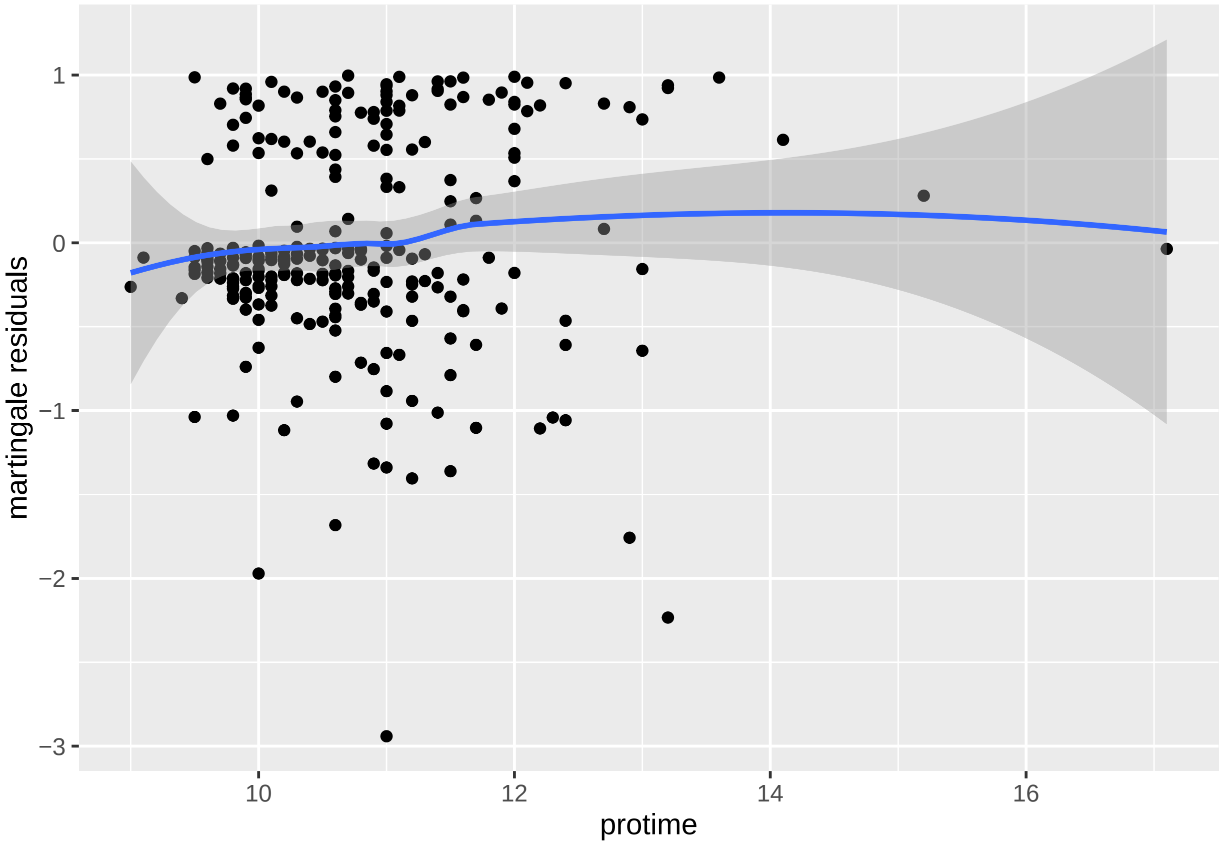
```
model.copper<-coxph(Surv(fu, status)~drug+sex+ascites+as.factor(edema)+bili+albumin+sgot+protime+as.fac
mres.copper<-resid(model.copper)
data.copper<-data.frame(x=pbc$copper,y=mres.copper)
ggplot(data.copper,aes(x=x,y=y))+geom_point()+labs(x="copper",y="martingale residuals")+stat_smooth(met
```



```
model.sgot<-coxph(Surv(fu, status)~drug+sex+ascites+as.factor(edema)+bili+copper+albumin+protime+as.fac
mres.sgot<-resid(model.sgot)
data.sgot<-data.frame(x=pbc$sgot,y=mres.sgot)
ggplot(data.sgot,aes(x=x,y=y))+geom_point()+labs(x="sgot",y="martingale residuals")+stat_smooth(method =
```



```
model.protime<-coxph(Surv(fu, status)~drug+sex+ascites+as.factor(edema)+bili+copper+sgot+albumin+as.fac
mres.protime<-resid(model.protime)
data.protime<-data.frame(x=pbc$protime,y=mres.protime)
ggplot(data.protime,aes(x=x,y=y))+geom_point()+labs(x="protime",y="martingale residuals")+stat_smooth(me
```



## Question 4

**Are the hazard rates proportional in the drug group vs. the placebo group?**

```
time.dependence<-log(pbc$fu)*pbc$drug
cox.data<-cbind(pbc,time.dependence)
model.full<-coxph(Surv(fu, status)~drug+sex+ascites+as.factor(edema)+bili+albumin+copper+sgot+protime+a

model.reduce<-coxph(Surv(fu, status)~drug+sex+ascites+as.factor(edema)+bili+albumin+copper+sgot+protime
LR<--2*(model.reduce$loglik[2]-model.full$loglik[2])
```

## Question 5

**Is there a significant difference in the drug and the placebo group?**

```
# km.fit <-survfit(Surv(fu, status)~drug, data=data, type="kaplan-meier")
# summary(km.fit)
# plot(km.fit, xlab="Time (Years)", ylab="S(t)", main = 'Penicillamine vs. Placebo Patients', col = c('
# text(8, .75, "Placebo", col = 'blue')
# text(8, .4, "Penicillamine", col = 'red')
#
#
# #Estimate the hazard rate with a uniform kernel and a bandwidth of 5 years, at 1,3,5,7,9,11, and 13 y
# #Penicillamine Group
# pen <- data[data$drug == 0,]
# pen.fit <-survfit(Surv(fu, status)~1, data=pen, type="kaplan-meier")
# times = summary(pen.fit)$time
# n = length(times)
# hazards = rep(NA,n)
# for (i in 1:n){hazards[i]=sum(summary(pen.fit)$n.event[1:i]/summary(pen.fit)$n.risk[1:i])}
# hazards = append(hazards,0,0) #Append basically concatenates the vector.
# deltah = rep(NA,n)
# for (i in 1:n){deltah[i] = hazards[i+1]-hazards[i]}
#
#
# # t=1
# q = 1/5
# aq = (4*(1 + q^3))/((1 + q)^4)
# bq = (6*(1 - q))/((1 + q)^3)
# kernel1 = (1-times)/5
# h1 = 0.2*sum((aq+bq*kernel1[1:15])*deltah[1:15])
# # t=3
# q = 3/5
# aq = (4*(1 + q^3))/((1 + q)^4)
# bq = (6*(1 - q))/((1 + q)^3)
# kernel3 = (3-times)/5
# h3 = 0.2*sum((aq+bq*kernel3[1:16])*deltah[1:16])
# # t=5
# kernel5 = (5-times)/5
# h5 = 0.2*sum(0.5*deltah[1:17])
# # t=7
# kernel7 = (7-times)/5
# h7 = 0.2*sum(0.5*deltah[8:19])
# # t=9
# kernel9 = (9-times)/5
# h9 = 0.2*sum(0.5*deltah[14:20])
# # t=11
# kernel11 = (11-times)/5
# h11 = 0.2*sum(0.5*deltah[16:20])
# # t=13
# kernel13 = (13-times)/5
# h13 = 0.2*sum(0.5*deltah[17:20])
#
#
# x=c(1,3,5,7,9,11,13)
# h=c(h1,h3,h5,h7,h9,h11,h13)
# plot(x,h,type="b",xlab="Time (Years)",ylab="Hazard Rate",main="Estimated Hazard Rates", col ='red')
#
#
# #Placebo Group
# pla <- data[data$drug == 1,]
# pla.fit <-survfit(Surv(fu, status)~1, data=pla, type="kaplan-meier")
# times = summary(pla.fit)$time
# n = length(times)
# hazards = rep(NA,n)
# for (i in 1:n){hazards[i]=sum(summary(pla.fit)$n.event[1:i]/summary(pla.fit)$n.risk[1:i])}
# hazards = append(hazards,0,0) #Append basically concatenates the vector.
# deltah = rep(NA,n)
# for (i in 1:n){deltah[i] = hazards[i+1]-hazards[i]}
#
#
# # t=1
# q = 1/5
# aq = (4*(1 + q^3))/((1 + q)^4)
# bq = (6*(1 - q))/((1 + q)^3)
# kernel1 = (1-times)/5
# h1 = 0.2*sum((aq+bq*kernel1[1:15])*deltah[1:15])
# # t=3
# q = 3/5
# aq = (4*(1 + q^3))/((1 + q)^4)
# bq = (6*(1 - q))/((1 + q)^3)
# kernel3 = (3-times)/5
# h3 = 0.2*sum((aq+bq*kernel3[1:16])*deltah[1:16])
# # t=5
# kernel5 = (5-times)/5
# h5 = 0.2*sum(0.5*deltah[1:17])
# # t=7
# kernel7 = (7-times)/5
# h7 = 0.2*sum(0.5*deltah[8:19])
# # t=9
# kernel9 = (9-times)/5
# h9 = 0.2*sum(0.5*deltah[14:20])
# # t=11
# kernel11 = (11-times)/5
# h11 = 0.2*sum(0.5*deltah[16:20])
# # t=13
# kernel13 = (13-times)/5
# h13 = 0.2*sum(0.5*deltah[17:20])
#
#
# x=c(1,3,5,7,9,11,13)
# h=c(h1,h3,h5,h7,h9,h11,h13)
# lines(x,h,type="b",xlab="Time (Years)",ylab="Hazard Rate", main="Estimated Hazard Rate \n Placebo Gro
# text(5,0.02, 'Penicillamine', col = 'red')
# text(5, 0.009, 'Placebo', col = 'blue')
#
#




# Fit a parametric and a non-parametric model to see which is better.
# Use diagnostic plots (residuals) to see if the model is a good fit.
```