Survival Analysis HW 3

Faizan Khalid Mohsin

March 14, 2019

Contents

1 Question 2

1.1 Question 2 a.

This is done seperately in the attachment provided.

Below we do some data cleaning and present our steps and some thoughts.

```
data(veteran)
# Data Cleaning
veteran$prior = as.factor(veteran$prior/10)
veteran$trt = as.factor(veteran$trt)
str(veteran)
```

```
## 'data.frame':    137 obs. of  8 variables:
##  $ trt     : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ celltype: Factor w/ 4 levels "squamous","smallcell",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ time    : num  72 411 228 126 118 10 82 110 314 100 ...
##  $ status  : num  1 1 1 1 1 1 1 1 1 0 ...
##  $ karno   : num  60 70 60 60 70 20 40 80 50 70 ...
##  $ diagtime: num  7 5 3 9 11 5 10 29 18 6 ...
##  $ age     : num  69 64 38 63 65 49 69 68 43 70 ...
##  $ prior   : Factor w/ 2 levels "0","1": 1 2 1 2 2 1 2 1 1 1 ...
```

```
# Missing data.
all(!is.na(veteran)==TRUE) # No missing data
```

```
## [1] TRUE
```

```
table(veteran$trt, veteran$status)
```

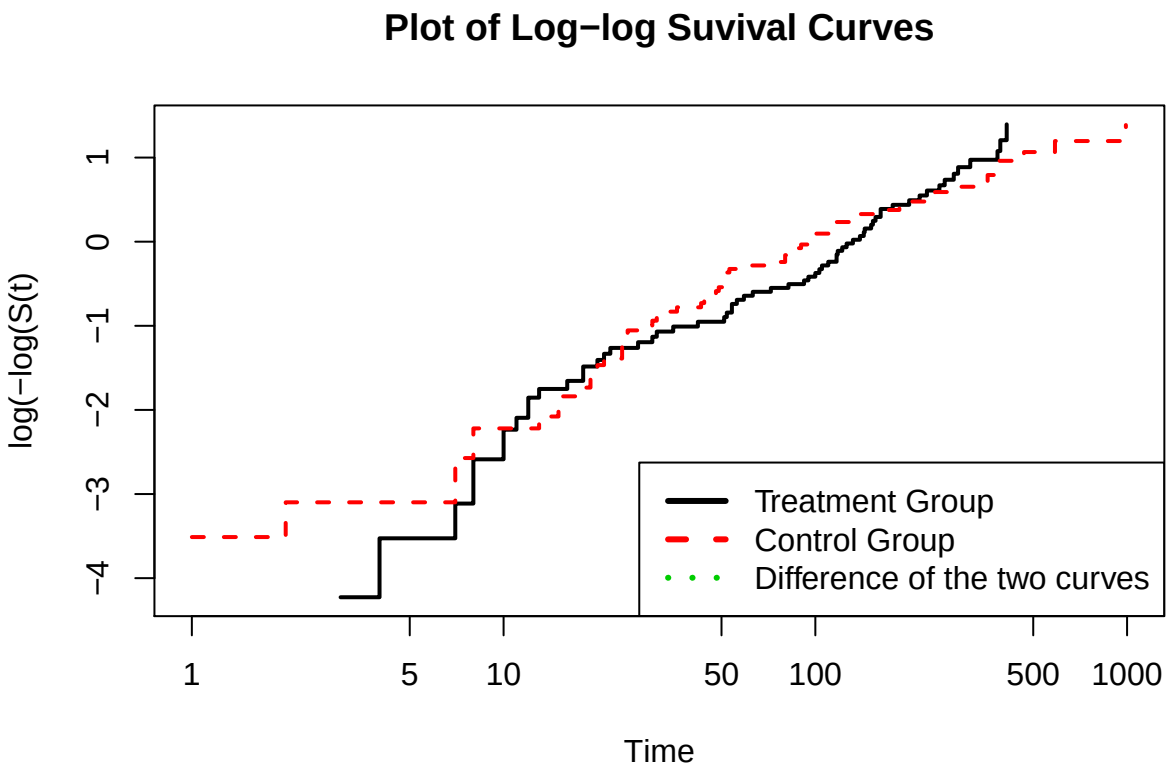
```
##
##      0  1
##   1  5 64
##   2  4 64
```

Note that from the above table we can see that the two groups (treatment and no treatment in the rows 1 and 2) are well balance. Further, note that most of the people died in this study as the majority of the poeple have status 1. Lastly, we did not find any missind data in this data set. The data is now ready for analysis.

1.2 Question 2 b.

We will now fit the model.

```
# Fitting the model.
fit <- survfit(Surv(time, status) ~ trt , data=veteran )

# Plotting the model
plot(fit, fun="cloglog", lty = 1:2, lwd = 2, col = 1:2,
     main = "Plot of Log-log Suvival Curves",
     xlab = "Time", ylab = "log(-log(S(t)")
legend("bottomright", col=1:3,lty = 1:3, lwd=3,
       legend=c("Treatment Group", "Control Group", "Difference of the two curves"))
```




```
#ggsurvplot(fit, data = veteran, risk.table = TRUE, fun="cloglog")
```

From the plot we see that the curves cross/intercept at least three points, they are not parallel. Hence, as the log-log plot of the survival curves are not parallel the assumption of proportional hazards is violated.

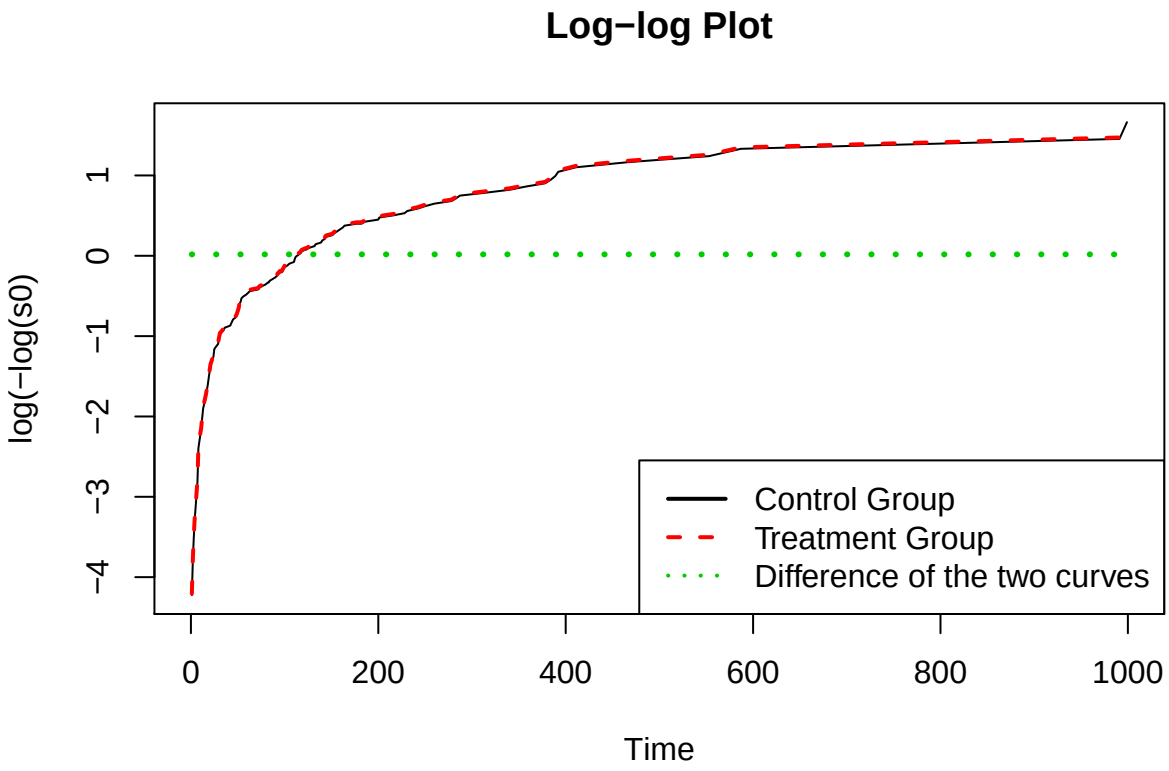
1.3 Question 2 c.

```
model <- coxph(Surv(time, status) ~ as.factor(trt) , data=veteran)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ as.factor(trt), data = veteran)
##
##   n= 137, number of events= 128
##
##                     coef exp(coef) se(coef)     z Pr(>|z|)
## as.factor(trt)2 0.01774   1.01790  0.18066 0.098    0.922
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## as.factor(trt)2     1.018     0.9824    0.7144      1.45
##
## Concordance= 0.525  (se = 0.026 )
## Rsquare= 0   (max possible= 0.999 )
## Likelihood ratio test= 0.01  on 1 df,   p=0.9218
## Wald test            = 0.01  on 1 df,   p=0.9218
```

```
## Score (logrank) test = 0.01  on 1 df,   p=0.9218
```

```
cox<-coxph(Surv(time, status) ~ as.factor(trt) , data=veteran)
b1<-cox$coefficients[1]
bh <- basehaz(cox)
breslow <- bh
s0<-exp(-breslow[,1])
s1<-exp(-breslow[,1]*exp(b1))
difference = log(-log(s1)) - log(-log(s0))

plot(bh[,2],log(-log(s0)), type="l",lty=1, col=1,
     xlab="Time", ylab="log(-log(s0)", main="Log-log Plot")
points(bh[,2], log(-log(s1)),col=2, type="l", lty=2, lwd=2)
points(bh[,2],difference, type="l", col=3, lty=3, lwd=3)
legend("bottomright", col=1:3,lty = 1:3 , lwd=2 ,
       legend=c("Control Group", "Treatment Group", "Difference of the two curves"))
```



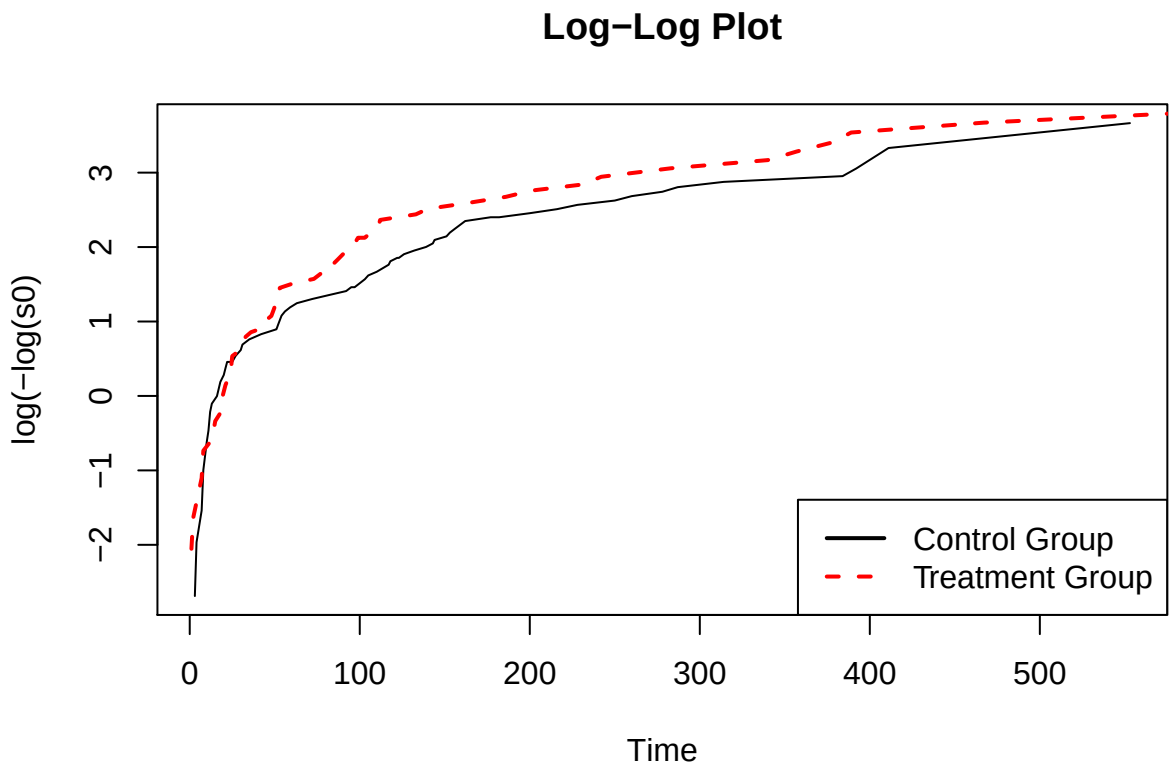
This is not useful for testing proportionallity of hazard functions because that is an assumption of the model. So once we fit the model which assumes proportionality, the estimates of the model will already have proportionality assumption built in them. Hence, there is no point in testing it this way.

2 Question 3

Separate Breslow estimates for the two baseline hazards can now be obtained to compare the baseline survival functions. Present and interpret a log-log plot of the baseline survival functions to check the proportionality

assumption of the treatment effect in the adjusted model.

```
model <- coxph(Surv(time, status) ~ strata(trt) + karno + age
               + as.factor(prior)
               + as.factor(celltype) + diagtime
               , data=veteran)


breslow <- basehaz(model, centered = FALSE)
breslow0 <- breslow[1:61,]
breslow1 <- breslow[62:114,]
s0<-exp(-breslow0[,1])
s1<-exp(-breslow1[,1])

plot(breslow0[,2],log(-log(s0)),
     type="l",lty=1, col=1,
     xlab="Time",
     ylab="log(-log(s0)",
     main="Log-Log Plot")
points(breslow1[,2], log(-log(s1)),col=2, type="l", lty=2, lwd=2)
legend("bottomright", col=1:2,lty = 1:2 ,
       legend=c("Control Group", "Treatment Group")
       , lwd=2)
```



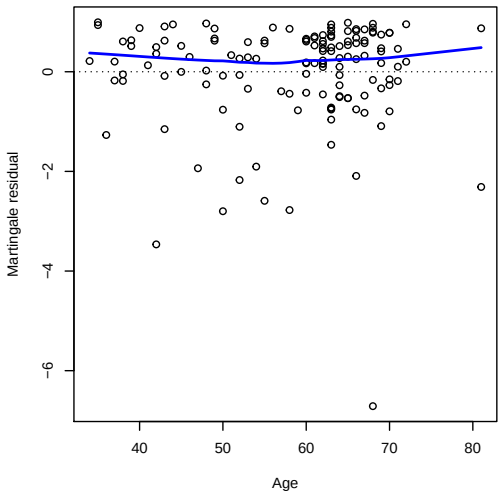
From the above plot we see that the two curves are roughly parallel. Further, the diffence plot shows that the diffence over time practically constant, hence, β is constant overtime. Therefore, the proportionality assumption of the treatment effect in the adjusted model holds.

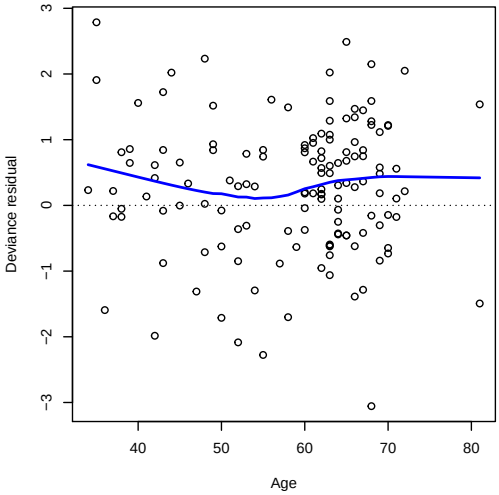
3 Question 4

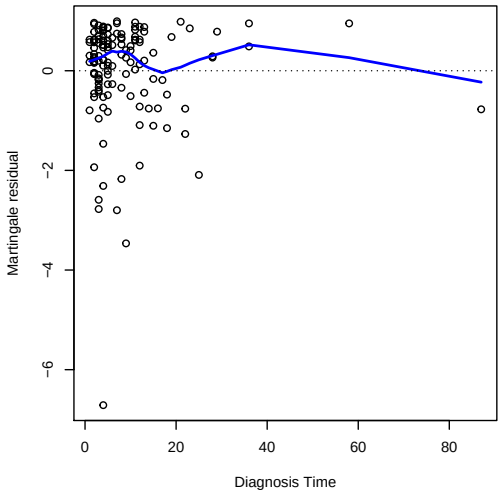
3.1 Question 4 a.

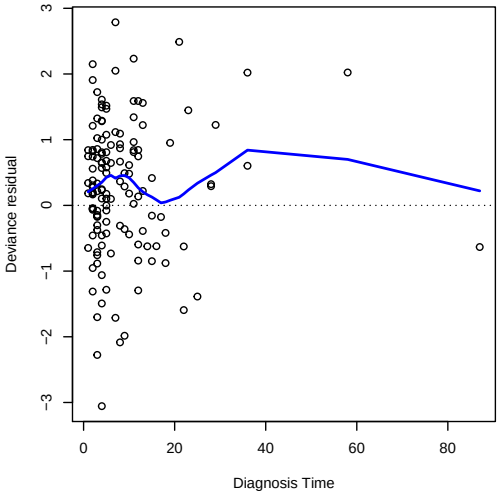
Investigate the appropriateness of the linearity assumptions made on the effects of the continuous covariates using appropriate residual plots.

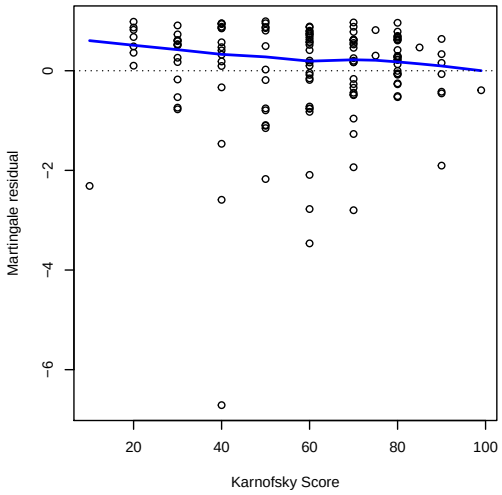
```
model <- coxph(Surv(time, status) ~ as.factor(trt) + karno + age
                                  + as.factor(prior)
                                  + as.factor(celltype) + diagtime
                                  , data=veteran)

martingaleres <- residuals(model, type=c('martingale'))
devianceres <- residuals(model, type=c('deviance'))
dfbeta <- residuals(model, type=c('dfbeta'))
dfbetas <- residuals(model, type=c('dfbetas'))

# Check martingale and deviance residuals for continuous covariates:

par(mfrow=c(3,1))

plot(veteran$age, martingaleres, xlab='Age', ylab='Martingale residual')
lines(lowess(veteran$age, martingaleres), lwd=2, col='blue')
abline(h=0, lty='dotted')

plot(veteran$diagtime, martingaleres, xlab='Diagnosis Time', ylab='Martingale residual')
lines(lowess(veteran$diagtime, martingaleres), lwd=2, col='blue')
abline(h=0, lty='dotted')

plot(veteran$karno, martingaleres, xlab='Karnofsky Score', ylab='Martingale residual')
lines(lowess(veteran$karno, martingaleres), lwd=2, col='blue')
abline(h=0, lty='dotted')

# Deviance residuals

plot(veteran$age, devianceres, xlab='Age', ylab='Deviance residual')
lines(lowess(veteran$age, devianceres), lwd=2, col='blue')
abline(h=0, lty='dotted')

plot(veteran$diagtime, devianceres, xlab='Diagnosis Time', ylab='Deviance residual')
lines(lowess(veteran$diagtime, devianceres), lwd=2, col='blue')
abline(h=0, lty='dotted')

plot(veteran$karno, devianceres, xlab='Karnofsky Score', ylab='Deviance residual')
lines(lowess(veteran$karno, devianceres), lwd=2, col='blue')
abline(h=0, lty='dotted')
```

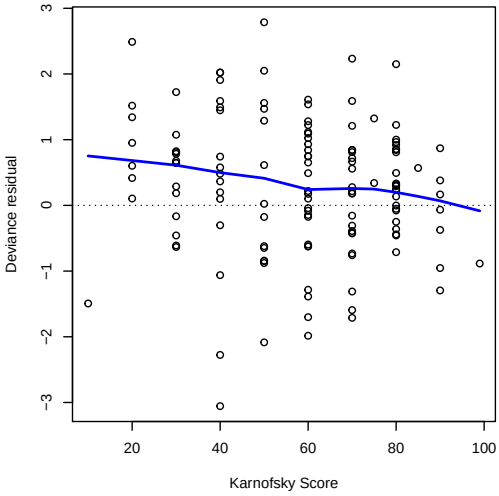

Martingale residual
Age


Deviance residual
Age


Martingale residual
Diagnosis Time


Deviance residual
Diagnosis Time


Martingale residual
Karnofsky Score


Deviance residual
Karnofsky Score

First, note that we will mainly be using the deviance residual plots and not the martingale plots because they are skewed, whereas the deviance residual plots are rescaled versions of the martingale residuals, to make them more symmetric around zero.

From the "age" residual deviance plot we see that there is no apparent patern. Hence, linearity assumption of for "age" is not violated.

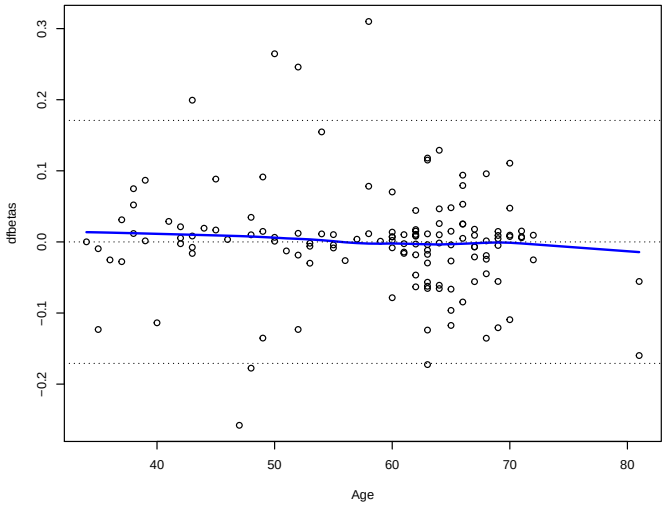
From the "diagnosis time" deviance residual plot, we see that the obsevations are mostly concentrated to the left handside (right skewed), however, again no outright clear pattern is observed apart from the skewness of the data. Therefore, the linearity assumption for "daignosis time" is not violated. However, perhaps a log transformation of the diagnosis time should be done to decrease the skewness.

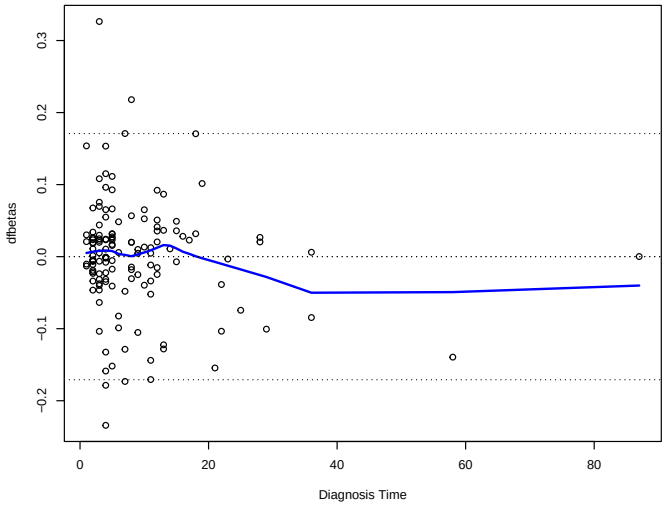
From the "Karnofsky score" deviance residual plot, again there is no distinct pattern. The data points appear to be completely random. Hence, the linearity assumption of for "Karnofsky score" is not violated.

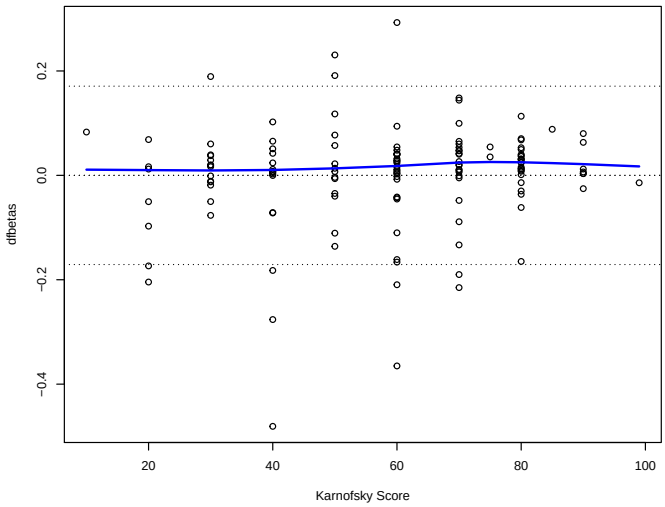
3.2 Question 4 b.

Check also the presence of potential influential observations.

```
# Scaled dfbeta influence measures (compare to the threshold of 2/sqrt(n)):
par(mfrow=c(3,1))

plot(veteran$age, dfbetas[,7], xlab='Age', ylab='dfbetas')
lines(lowess(veteran$age, dfbetas[,7]), lwd=2, col='blue')
abline(h=c(-2/sqrt(nrow(dfbetas)), 0, 2/sqrt(nrow(dfbetas))), lty='dotted')

plot(veteran$diagtime, dfbetas[,6], xlab='Diagnosis Time', ylab='dfbetas')
lines(lowess(veteran$diagtime, dfbetas[,6]), lwd=2, col='blue')
abline(h=0, lty='dotted')
abline(h=c(-2/sqrt(nrow(dfbetas)), 0, 2/sqrt(nrow(dfbetas))), lty='dotted')

plot(veteran$karno, dfbetas[,5], xlab='Karnofsky Score', ylab='dfbetas')
lines(lowess(veteran$karno, dfbetas[,5]), lwd=2, col='blue')
abline(h=0, lty='dotted')
abline(h=c(-2/sqrt(nrow(dfbetas)), 0, 2/sqrt(nrow(dfbetas))), lty='dotted')
```


dfbetas
0.3
0.2
0.1
0.0
−0.1
−0.2
40
50
60
70
80
Age


dfbetas
0.3
0.2
0.1
0.0
−0.1
−0.2
0
20
40
60
80
Diagnosis Time


dfbetas
0.2
0.0
−0.2
−0.4
20
40
60
80
100
Karnofsky Score

All the observations that are outside of the bands are potential outliers or influencer observations.

```
# EXTRACT THE INLUENTIAL POINTS.
upper = 2/sqrt(nrow(dfbetas))

aged_out<-ifelse( abs(dfbetas[,7]) >= upper, 1, 0)
influentialaged_out<-subset(aged_out, aged_out==1)

diagtime_out<-ifelse( abs(dfbetas[,6]) >= upper, 1, 0)
influentialdiagtimed_out<-subset(diagtime_out, diagtime_out==1)

karno_out<-ifelse( abs(dfbetas[,5]) >= upper, 1,0)
influentialkarno_out<-subset(karno_out, karno_out==1)

age_outliers = names(influentialaged_out)
diagtime_outliers = names(influentialdiagtimed_out)
karno_outliers = names(influentialkarno_out)

age_outliers
```

```
## [1] "9"  "12" "15" "58" "73" "75" "78"
```

```
diagtime_outliers
```

```
## [1] "11"  "73"  "75"  "78"  "85"  "118"
```

```
karno_outliers
```

```
##  [1] "6"  "9"  "11" "12" "13" "15" "17" "21" "44" "73" "77" "78" "91"
```

The outliers that we found for the variable age were 9, 12, 15, 58, 73, 75, 78.

The outliers that we found for the variable diagnosis time were 11, 73, 75, 78, 85, 118.

The outliers that we found for the variable Karnofsky scores were 6, 9, 11, 12, 13, 15, 17, 21, 44, 73, 77, 78, 91.

3.3 Question 4 c.

Investigate the appropriateness of the proportinal hazards assumptions through appropriate residual plots, and statistical tests. Report the tests for both the residual-time correlations, and covariate-time interactions added to the Cox model (the latter can be implemented using the tt argument of the coxph function).

```
# Checks for proportionality:
output.cox.zph = cox.zph(model, global=FALSE)
output.cox.zph
```

```
##                                rho   chisq        p
## as.factor(trt)2             -0.0273  0.1227 0.726104
## karno                        0.3073 13.0449 0.000304
## age                          0.1890  5.3476 0.020750
## as.factor(prior)1           -0.1767  4.4714 0.034467
## as.factor(celltype)smallcell  0.0128  0.0261 0.871621
## as.factor(celltype)adeno      0.1424  2.9794 0.084329
## as.factor(celltype)large      0.1712  4.1093 0.042649
## diagtime                      0.1491  2.9436 0.086217
```

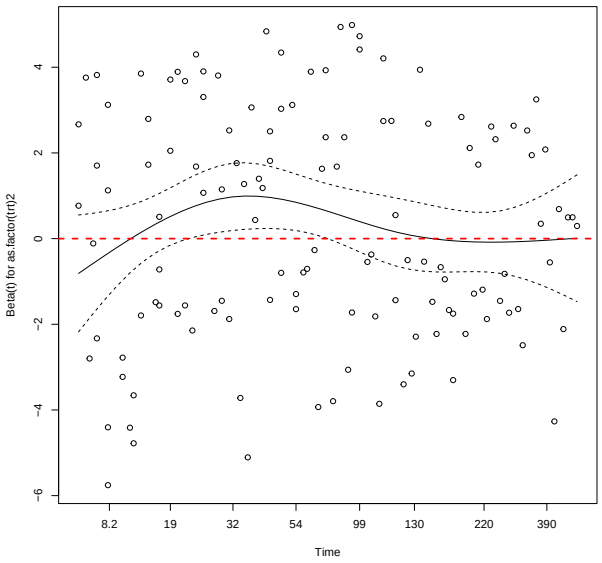
From the test for proportional hazards, it can be seen that we conducted eight tests. Hence, we need to correct for multiple testing, hence, our significance level will be $5\%/8 = 0.00625$. Hence, only the variable Karnofsky scores has a p-values which is statistically significant at the bonferroni level. Hence, we have strong evidence only for the variable Karnofsky scores that the hazard ratios are not proportional overtime.

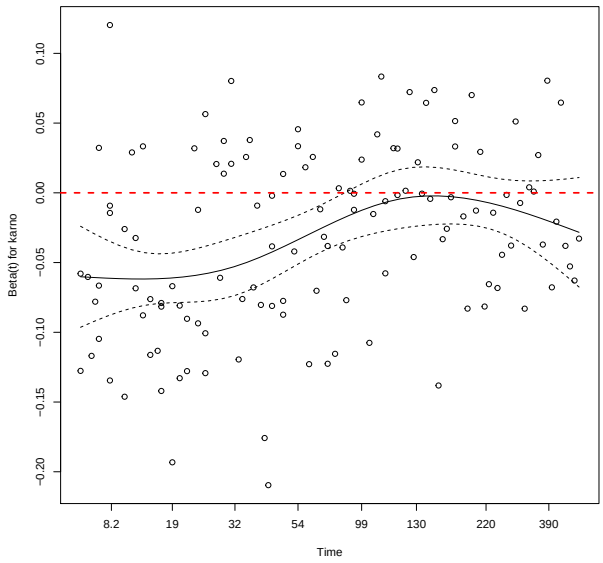
The variables age, prior and the celltype with factor large, have statistically significant p-values at the significance level of 5%. But do not have "strong" evidence to reject the Null (Null:hazard ratios are proportional).

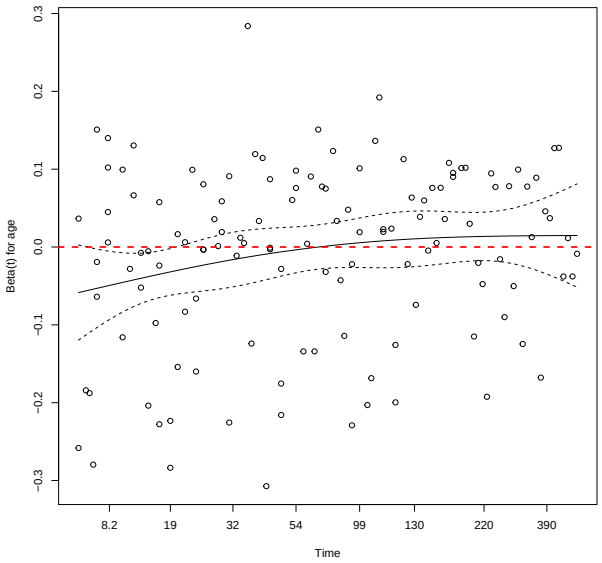
The statistical tests used is the was a Time-weighted score tests of the proportional hazards. The statistical test we performed is a Schoenefeld residual test. It is a chi-square test.

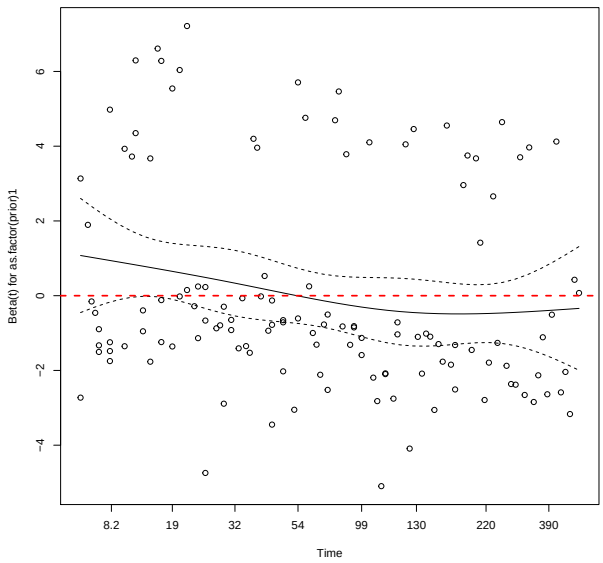
We will now plot the Schoenefeld residual plots.

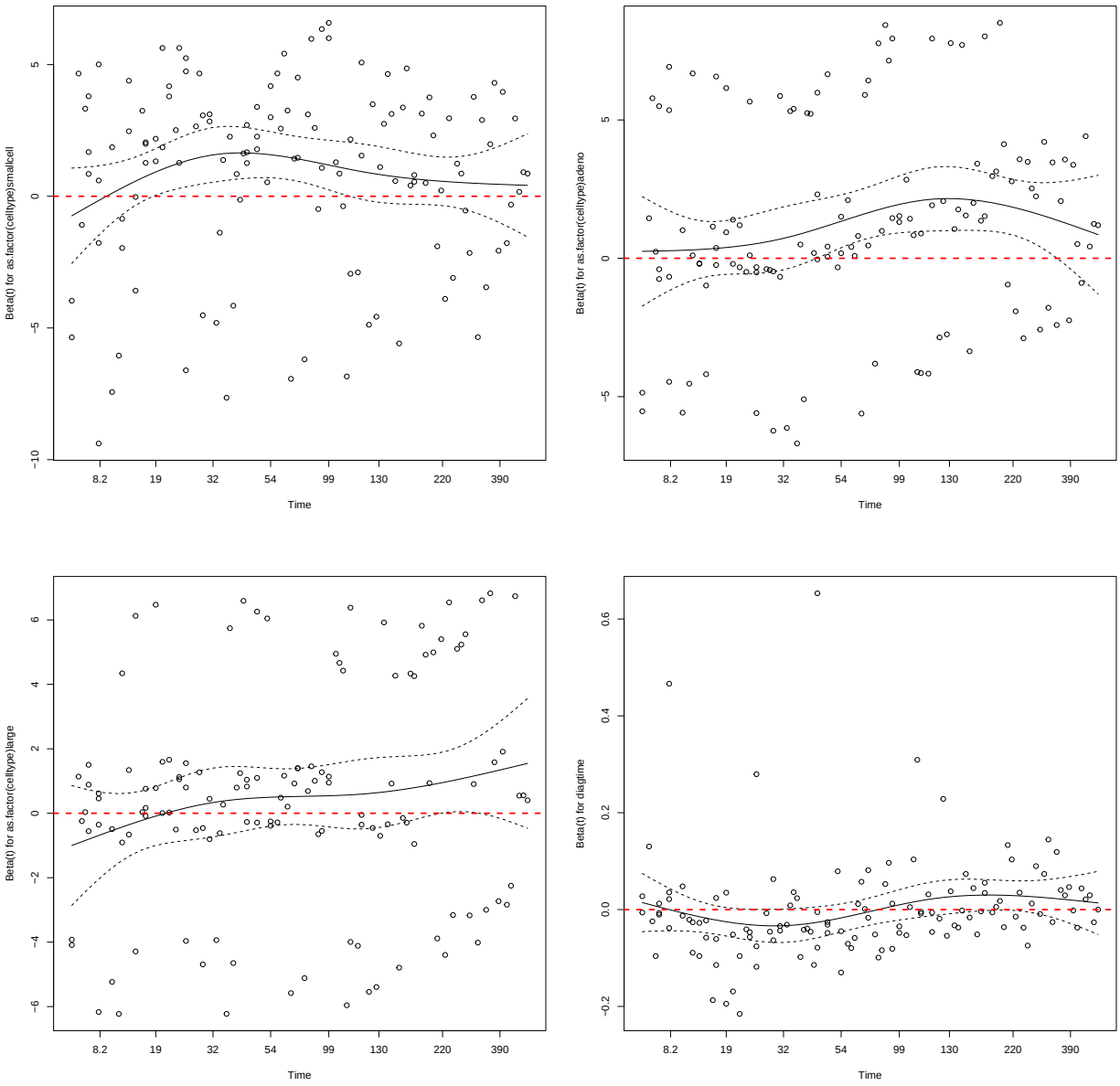
```
# Plotting
par(mfrow=c(2,2))

for (i in 1:8){
  plot(cox.zph(model, global=FALSE), var=i)
  abline(h=0, lty=2, col = "red", lwd = 2)
}
```


Beta(t) for as.factor(trt)2
Time
8.2
19
32
54
99
130
220
390


Beta(t) for karno
Time
8.2
19
32
54
99
130
220
390


Beta(t) for age
Time
8.2
19
32
54
99
130
220
390


Beta(t) for as.factor(prior)1
Time
8.2
19
32
54
99
130
220
390



From the above Schoenefeld residual plots we see that for the variable Karnofsky scores, majority of the time the confidence interval bands do not include zero in them. Only around time 99 do they begin to include zero. Hence, this residual plot also confirms that the hazard ratios are not proportional for this variable. This is what we found in the statistical test performed above.

For the variable age we see that there is a distinct increasing pattern, however, most of the time zero is in the confidence interval bands. Hence, there may be some mild departure from the assumption of proportional hazards but nothing as bad as the Karnofsky scores variable.

Similarly, there is a distinct decreasing pattern for the prior variable in the residual plots and and increasing pattern for the variable celltype with factor large. And again, similarly, zero is in the confidence interval band with for the majority of the time.

We will now do tests for covariate-time interactions by adding those to our model.

```r
# Examples of tests for covariate-time interactions:

final_data = as.data.frame(model.matrix(~., data=veteran))

model <- coxph(Surv(time, status) ~ trt2 + tt(trt2) +
                                  + karno
                                  + age
                                  + prior1
                                  + celltypeadeno
                                  + celltypesmallcell
                                  + celltypelarge
                                  + diagtime
                                  , tt=function(x,t, ...) x * t
                                  , data=final_data)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt2 + tt(trt2) + +karno +
##     age + prior1 + celltypeadeno + celltypesmallcell + celltypelarge +
##     diagtime, data = final_data, tt = function(x, t, ...) x *
##     t)
##
##   n= 137, number of events= 128
##
##                          coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt2                2.992e-01  1.349e+00  2.663e-01  1.124 0.261177
## tt(trt2)           -5.011e-05  9.999e-01  1.814e-03 -0.028 0.977968
## karno              -3.278e-02  9.677e-01  5.627e-03 -5.826 5.69e-09 ***
## age                -8.690e-03  9.913e-01  9.319e-03 -0.933 0.351056
## prior1              7.212e-02  1.075e+00  2.331e-01  0.309 0.757001
## celltypeadeno       1.194e+00  3.299e+00  3.128e-01  3.816 0.000136 ***
## celltypesmallcell   8.597e-01  2.362e+00  2.837e-01  3.030 0.002446 **
## celltypelarge       3.991e-01  1.490e+00  2.938e-01  1.358 0.174396
## diagtime            6.600e-05  1.000e+00  9.153e-03  0.007 0.994246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## trt2                  1.3488     0.7414    0.8003    2.2731
## tt(trt2)              0.9999     1.0001    0.9964    1.0035
## karno                 0.9677     1.0333    0.9571    0.9785
## age                   0.9913     1.0087    0.9734    1.0096
## prior1                1.0748     0.9304    0.6807    1.6971
## celltypeadeno         3.2993     0.3031    1.7871    6.0909
## celltypesmallcell     2.3623     0.4233    1.3547    4.1195
## celltypelarge         1.4904     0.6709    0.8379    2.6510
## diagtime              1.0001     0.9999    0.9823    1.0182
##
## Concordance= 0.736  (se = 0.234 )
## Rsquare= 0.364   (max possible= 0.999 )
## Likelihood ratio test= 62.1  on 9 df,   p=5.259e-10
## Wald test            = 62.38  on 9 df,   p=4.646e-10
## Score (logrank) test = 66.87  on 9 df,   p=6.222e-11
```

```
model <- coxph(Surv(time, status) ~ trt2 +
                                  + karno + tt(karno)
                                  + age
                                  + prior1
                                  + celltypeadeno
                                  + celltypesmallcell
                                  + celltypelarge
                                  + diagtime
                                  , tt=function(x,t, ...) x * t
                                  , data=final_data)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt2 + +karno + tt(karno) +
##     age + prior1 + celltypeadeno + celltypesmallcell + celltypelarge +
##     diagtime, data = final_data, tt = function(x, t, ...) x *
##     t)
##
##   n= 137, number of events= 128
##
##                         coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt2                0.2123886  1.2366283  0.2110982  1.006 0.314361
## karno              -0.0420272  0.9588437  0.0067431 -6.233 4.59e-10 ***
## tt(karno)           0.0001137  1.0001137  0.0000477  2.383 0.017189 *
## age                -0.0096917  0.9903551  0.0092968 -1.042 0.297192
## prior1              0.0417820  1.0426672  0.2336226  0.179 0.858060
## celltypeadeno       1.2276209  3.4130998  0.3047047  4.029 5.60e-05 ***
## celltypesmallcell   0.9357855  2.5492151  0.2806760  3.334 0.000856 ***
## celltypelarge       0.4078135  1.5035267  0.2835417  1.438 0.150353
## diagtime           -0.0009430  0.9990575  0.0090982 -0.104 0.917454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## trt2                   1.2366     0.8087    0.8176    1.8704
## karno                  0.9588     1.0429    0.9463    0.9716
## tt(karno)              1.0001     0.9999    1.0000    1.0002
## age                    0.9904     1.0097    0.9725    1.0086
## prior1                 1.0427     0.9591    0.6596    1.6482
## celltypeadeno          3.4131     0.2930    1.8784    6.2017
## celltypesmallcell      2.5492     0.3923    1.4706    4.4189
## celltypelarge          1.5035     0.6651    0.8625    2.6210
## diagtime               0.9991     1.0009    0.9814    1.0170
##
## Concordance= 0.745  (se = 0.234 )
## Rsquare= 0.391   (max possible= 0.999 )
## Likelihood ratio test= 68.02  on 9 df,   p=3.714e-11
## Wald test            = 67.37  on 9 df,   p=4.985e-11
## Score (logrank) test = 73.33  on 9 df,   p=3.368e-12
```

```
model <- coxph(Surv(time, status) ~ trt2
                                  + karno
                                  + age  + tt(age)
                                  + prior1
```

```
                                  + celltypeadeno
                                  + celltypesmallcell
                                  + celltypelarge
                                  + diagtime
                                  , tt=function(x,t, ...) x * t
                                  , data=final_data)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt2 + karno + age + tt(age) +
##     prior1 + celltypeadeno + celltypesmallcell + celltypelarge +
##     diagtime, data = final_data, tt = function(x, t, ...) x *
##     t)
##
##   n= 137, number of events= 128
##
##                         coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt2                2.925e-01  1.340e+00  2.070e-01  1.413  0.15755
## karno              -3.300e-02  9.675e-01  5.501e-03 -5.999 1.98e-09 ***
## age                -1.548e-02  9.846e-01  1.195e-02 -1.295  0.19535
## tt(age)             7.837e-05  1.000e+00  8.945e-05  0.876  0.38098
## prior1              6.243e-02  1.064e+00  2.325e-01  0.269  0.78831
## celltypeadeno       1.195e+00  3.302e+00  3.005e-01  3.976 7.01e-05 ***
## celltypesmallcell   8.430e-01  2.323e+00  2.768e-01  3.045  0.00232 **
## celltypelarge       3.923e-01  1.480e+00  2.831e-01  1.386  0.16582
## diagtime            2.395e-04  1.000e+00  9.137e-03  0.026  0.97909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## trt2                  1.3398     0.7464    0.8930     2.010
## karno                 0.9675     1.0336    0.9572     0.978
## age                   0.9846     1.0156    0.9618     1.008
## tt(age)               1.0001     0.9999    0.9999     1.000
## prior1                1.0644     0.9395    0.6748     1.679
## celltypeadeno         3.3024     0.3028    1.8326     5.951
## celltypesmallcell     2.3233     0.4304    1.3505     3.997
## celltypelarge         1.4803     0.6755    0.8500     2.578
## diagtime              1.0002     0.9998    0.9825     1.018
##
## Concordance= 0.737  (se = 0.234 )
## Rsquare= 0.368   (max possible= 0.999 )
## Likelihood ratio test= 62.89  on 9 df,   p=3.711e-10
## Wald test            = 63.15  on 9 df,   p=3.301e-10
## Score (logrank) test = 67.9  on 9 df,   p=3.928e-11
```

```
model <- coxph(Surv(time, status) ~ trt2 +
                                  + karno
                                  + age
                                  + prior1 + tt(prior1)
                                  + celltypeadeno
                                  + celltypesmallcell
                                  + celltypelarge
                                  + diagtime
```

```
                                  , tt=function(x,t, ...) x * t
                                  , data=final_data)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt2 + +karno + age + prior1 +
##     tt(prior1) + celltypeadeno + celltypesmallcell + celltypelarge +
##     diagtime, data = final_data, tt = function(x, t, ...) x *
##     t)
##
##   n= 137, number of events= 128
##
##                         coef exp(coef)  se(coef)      z Pr(>|z|)
## trt2                0.305504  1.357309  0.207816  1.470  0.14154
## karno              -0.032309  0.968207  0.005534 -5.839 5.27e-09 ***
## age                -0.007965  0.992066  0.009330 -0.854  0.39323
## prior1              0.324992  1.384020  0.302440  1.075  0.28257
## tt(prior1)         -0.002235  0.997768  0.001833 -1.219  0.22271
## celltypeadeno       1.189874  3.286667  0.298845  3.982 6.85e-05 ***
## celltypesmallcell   0.849403  2.338251  0.274110  3.099  0.00194 **
## celltypelarge       0.369889  1.447573  0.282628  1.309  0.19062
## diagtime           -0.001280  0.998721  0.009093 -0.141  0.88804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## trt2                  1.3573     0.7368    0.9032    2.0397
## karno                 0.9682     1.0328    0.9578    0.9788
## age                   0.9921     1.0080    0.9741    1.0104
## prior1                1.3840     0.7225    0.7651    2.5037
## tt(prior1)            0.9978     1.0022    0.9942    1.0014
## celltypeadeno         3.2867     0.3043    1.8297    5.9038
## celltypesmallcell     2.3383     0.4277    1.3664    4.0014
## celltypelarge         1.4476     0.6908    0.8319    2.5189
## diagtime              0.9987     1.0013    0.9811    1.0167
##
## Concordance= 0.739  (se = 0.234 )
## Rsquare= 0.372   (max possible= 0.999 )
## Likelihood ratio test= 63.7  on 9 df,   p=2.578e-10
## Wald test            = 63.69  on 9 df,   p=2.59e-10
## Score (logrank) test = 68.31  on 9 df,   p=3.257e-11
```

```
model <- coxph(Surv(time, status) ~ trt2
                                  + karno
                                  + age
                                  + prior1
                                  + celltypeadeno
                                  + tt(celltypeadeno)
                                  + celltypesmallcell
                                  + celltypelarge
                                  + diagtime
                                  , tt=function(x,t, ...) x * t
                                  , data=final_data)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt2 + karno + age + prior1 +
##     celltypeadeno + tt(celltypeadeno) + celltypesmallcell + celltypelarge +
##     diagtime, data = final_data, tt = function(x, t, ...) x *
##     t)
##
##   n= 137, number of events= 128
##
##                         coef exp(coef)  se(coef)      z Pr(>|z|)
## trt2                0.284411  1.328979  0.203765  1.396  0.16278
## karno              -0.036448  0.964208  0.005682 -6.414 1.41e-10 ***
## age                -0.008085  0.991948  0.009333 -0.866  0.38636
## prior1              0.133832  1.143200  0.233502  0.573  0.56654
## celltypeadeno       0.351523  1.421231  0.427664  0.822  0.41110
## tt(celltypeadeno)   0.015238  1.015355  0.004896  3.113  0.00185 **
## celltypesmallcell   0.778606  2.178432  0.274054  2.841  0.00450 **
## celltypelarge       0.415980  1.515855  0.281570  1.477  0.13958
## diagtime           -0.001282  0.998719  0.008843 -0.145  0.88473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## trt2                  1.3290     0.7525    0.8914     1.981
## karno                 0.9642     1.0371    0.9535     0.975
## age                   0.9919     1.0081    0.9740     1.010
## prior1                1.1432     0.8747    0.7234     1.807
## celltypeadeno         1.4212     0.7036    0.6147     3.286
## tt(celltypeadeno)     1.0154     0.9849    1.0057     1.025
## celltypesmallcell     2.1784     0.4590    1.2731     3.728
## celltypelarge         1.5159     0.6597    0.8729     2.632
## diagtime              0.9987     1.0013    0.9816     1.016
##
## Concordance= 0.743  (se = 0.234 )
## Rsquare= 0.404   (max possible= 0.999 )
## Likelihood ratio test= 70.92  on 9 df,   p=1.004e-11
## Wald test            = 71.87  on 9 df,   p=6.529e-12
## Score (logrank) test = 78.74  on 9 df,   p=2.878e-13
```

```
model <- coxph(Surv(time, status) ~ trt2
                                  + karno
                                  + age
                                  + prior1
                                  + celltypeadeno
                                  + celltypesmallcell + tt(celltypesmallcell)
                                  + celltypelarge
                                  + diagtime
                                  , tt=function(x,t, ...) x * t
                                  , data=final_data)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt2 + karno + age + prior1 +
##     celltypeadeno + celltypesmallcell + tt(celltypesmallcell) +
##     celltypelarge + diagtime, data = final_data, tt = function(x,
##     t, ...) x * t)
```

```
##
##   n= 137, number of events= 128
##
##                              coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt2                    0.2250844  1.2524284  0.2081220  1.082 0.279474
## karno                  -0.0349964  0.9656089  0.0056027 -6.246 4.20e-10 ***
## age                    -0.0084494  0.9915862  0.0095013 -0.889 0.373847
## prior1                  0.0815627  1.0849812  0.2312174  0.353 0.724273
## celltypeadeno           1.4091778  4.0925892  0.3230468  4.362 1.29e-05 ***
## celltypesmallcell       1.3714677  3.9411309  0.3613934  3.795 0.000148 ***
## tt(celltypesmallcell) -0.0051179  0.9948951  0.0023474 -2.180 0.029237 *
## celltypelarge           0.4522454  1.5718376  0.2855955  1.584 0.113304
## diagtime               -0.0005637  0.9994365  0.0088417 -0.064 0.949169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## trt2                     1.2524     0.7984    0.8329    1.8832
## karno                    0.9656     1.0356    0.9551    0.9763
## age                      0.9916     1.0085    0.9733    1.0102
## prior1                   1.0850     0.9217    0.6896    1.7070
## celltypeadeno            4.0926     0.2443    2.1728    7.7086
## celltypesmallcell        3.9411     0.2537    1.9409    8.0028
## tt(celltypesmallcell)    0.9949     1.0051    0.9903    0.9995
## celltypelarge            1.5718     0.6362    0.8981    2.7511
## diagtime                 0.9994     1.0006    0.9823    1.0169
##
## Concordance= 0.735  (se = 0.234 )
## Rsquare= 0.389   (max possible= 0.999 )
## Likelihood ratio test= 67.47  on 9 df,   p=4.763e-11
## Wald test            = 64  on 9 df,   p=2.262e-10
## Score (logrank) test = 69.94  on 9 df,   p=1.564e-11
```

```
model <- coxph(Surv(time, status) ~ trt2
                                  + karno
                                  + age
                                  + prior1
                                  + celltypeadeno
                                  + celltypesmallcell
                                  + celltypelarge + tt(celltypelarge)
                                  + diagtime
                                  , tt=function(x,t, ...) x * t
                                  , data=final_data)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt2 + karno + age + prior1 +
##     celltypeadeno + celltypesmallcell + celltypelarge + tt(celltypelarge) +
##     diagtime, data = final_data, tt = function(x, t, ...) x *
##     t)
##
##   n= 137, number of events= 128
##
##                        coef exp(coef)  se(coef)      z Pr(>|z|)
## trt2                0.330164  1.391196  0.208241  1.585 0.112855
```

```
## karno             -0.033429  0.967124  0.005583 -5.987 2.13e-09 ***
## age               -0.009397  0.990647  0.009496 -0.990 0.322352
## prior1             0.118485  1.125790  0.233057  0.508 0.611176
## celltypeadeno      1.067139  2.907052  0.300942  3.546 0.000391 ***
## celltypesmallcell  0.796408  2.217561  0.271517  2.933 0.003355 **
## celltypelarge     -0.271250  0.762426  0.412248 -0.658 0.510552
## tt(celltypelarge)  0.004758  1.004770  0.001982  2.400 0.016374 *
## diagtime          -0.001010  0.998990  0.008990 -0.112 0.910542
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## trt2                 1.3912     0.7188    0.9250    2.0924
## karno                0.9671     1.0340    0.9566    0.9778
## age                  0.9906     1.0094    0.9724    1.0093
## prior1               1.1258     0.8883    0.7130    1.7776
## celltypeadeno        2.9071     0.3440    1.6117    5.2434
## celltypesmallcell    2.2176     0.4509    1.3024    3.7756
## celltypelarge        0.7624     1.3116    0.3399    1.7104
## tt(celltypelarge)    1.0048     0.9953    1.0009    1.0087
## diagtime             0.9990     1.0010    0.9815    1.0167
##
## Concordance= 0.743  (se = 0.234 )
## Rsquare= 0.391   (max possible= 0.999 )
## Likelihood ratio test= 67.88  on 9 df,   p=3.963e-11
## Wald test            = 65.46  on 9 df,   p=1.176e-10
## Score (logrank) test = 71.15  on 9 df,   p=9.067e-12
```

```
model <- coxph(Surv(time, status) ~ trt2
                                  + karno
                                  + age
                                  + prior1
                                  + celltypeadeno
                                  + celltypesmallcell
                                  + celltypelarge
                                  + diagtime + tt(diagtime)
                                  , tt=function(x,t, ...) x * t
                                  , data=final_data)
summary(model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt2 + karno + age + prior1 +
##     celltypeadeno + celltypesmallcell + celltypelarge + diagtime +
##     tt(diagtime), data = final_data, tt = function(x, t, ...) x *
##     t)
##
##   n= 137, number of events= 128
##
##                         coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt2                2.982e-01  1.347e+00  2.082e-01  1.432  0.15203
## karno              -3.288e-02  9.677e-01  5.512e-03 -5.966 2.44e-09 ***
## age                -8.720e-03  9.913e-01  9.292e-03 -0.938  0.34799
## prior1              5.728e-02  1.059e+00  2.401e-01  0.239  0.81143
## celltypeadeno       1.192e+00  3.294e+00  3.015e-01  3.954 7.70e-05 ***
## celltypesmallcell   8.582e-01  2.359e+00  2.754e-01  3.116  0.00183 **
```

```
## celltypelarge      4.027e-01  1.496e+00  2.828e-01  1.424  0.15445
## diagtime          -1.149e-03  9.989e-01  1.062e-02 -0.108  0.91382
## tt(diagtime)       3.103e-05  1.000e+00  1.295e-04  0.240  0.81062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## trt2                 1.3474     0.7422    0.8960    2.0263
## karno                0.9677     1.0334    0.9573    0.9782
## age                  0.9913     1.0088    0.9734    1.0095
## prior1               1.0590     0.9443    0.6615    1.6953
## celltypeadeno        3.2938     0.3036    1.8241    5.9475
## celltypesmallcell    2.3590     0.4239    1.3749    4.0474
## celltypelarge        1.4959     0.6685    0.8593    2.6040
## diagtime             0.9989     1.0012    0.9783    1.0199
## tt(diagtime)         1.0000     1.0000    0.9998    1.0003
##
## Concordance= 0.736  (se = 0.234 )
## Rsquare= 0.365   (max possible= 0.999 )
## Likelihood ratio test= 62.16  on 9 df,   p=5.129e-10
## Wald test            = 62.35  on 9 df,   p=4.714e-10
## Score (logrank) test = 66.75  on 9 df,   p=6.589e-11
```

From the model output above we see that the covariate-time interaction is only statistically significant for the variables Karnofsky scores and celltype for all the three levels at the 5% significance level. This is similar to what we found previously in the test for proportionality (Schoenfeld residual test).

4 Question 5

4.1 Question 5 a.

Below is the code, as well as the step what we are doing as comments in the code to explain the algorithm for computing the martingales manually. We also print dim() and head() so it is easy to understand what is happening.

We basically use the $M_i(t)$ formula from the lecture notes.

```
# model_initial <- coxph(Surv(time, status) ~ as.factor(trt) + as.factor(celltype)
#                                  + karno  + diagtime + age
#                                  + as.factor(prior)
#                                  , data=veteran)
#
# summary(model_initial)

# Below we create the data that we will use.
final_data = as.data.frame(model.matrix(~., data=veteran))

# We fit the model.
model <- coxph(Surv(time, status) ~ trt2 + celltypesmallcell + celltypeadeno
                                  + celltypelarge + karno + diagtime + age + prior1
                                  , data=final_data)
summary(model)
```

```
## Call:
```

```
## coxph(formula = Surv(time, status) ~ trt2 + celltypesmallcell +
##     celltypeadeno + celltypelarge + karno + diagtime + age +
##     prior1, data = final_data)
##
##   n= 137, number of events= 128
##
##                         coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt2                2.946e-01  1.343e+00  2.075e-01  1.419  0.15577
## celltypesmallcell   8.616e-01  2.367e+00  2.753e-01  3.130  0.00175 **
## celltypeadeno       1.196e+00  3.307e+00  3.009e-01  3.975 7.05e-05 ***
## celltypelarge       4.013e-01  1.494e+00  2.827e-01  1.420  0.15574
## karno              -3.282e-02  9.677e-01  5.508e-03 -5.958 2.55e-09 ***
## diagtime            8.132e-05  1.000e+00  9.136e-03  0.009  0.99290
## age                -8.706e-03  9.913e-01  9.300e-03 -0.936  0.34920
## prior1              7.159e-02  1.074e+00  2.323e-01  0.308  0.75794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## trt2                  1.3426     0.7448    0.8939    2.0166
## celltypesmallcell     2.3669     0.4225    1.3799    4.0597
## celltypeadeno         3.3071     0.3024    1.8336    5.9647
## celltypelarge         1.4938     0.6695    0.8583    2.5996
## karno                 0.9677     1.0334    0.9573    0.9782
## diagtime              1.0001     0.9999    0.9823    1.0182
## age                   0.9913     1.0087    0.9734    1.0096
## prior1                1.0742     0.9309    0.6813    1.6937
##
## Concordance= 0.736  (se = 0.03 )
## Rsquare= 0.364   (max possible= 0.999 )
## Likelihood ratio test= 62.1  on 8 df,   p=1.799e-10
## Wald test            = 62.37  on 8 df,   p=1.596e-10
## Score (logrank) test = 66.74  on 8 df,   p=2.186e-11
```

```
# We extract the correct martigales from the model.
martingaleres_correct <- residuals(model, type=c('martingale'))
head(martingaleres_correct)
```

```
##          1          2          3          4          5          6
##  0.7452837 -0.4900714 -0.1846769  0.4136253  0.6080666  0.8671830
```

```
sum(martingaleres_correct) # They sum upto zero as expected.
```

```
## [1] 1.057487e-14
```

```
# We extract the beta's from our model.
coeffi = (as.matrix(model$coefficients, nrow = 8, ncol = 1))
dim(coeffi)
```

```
## [1] 8 1
```

```
coeffi
```

```
##                           [,1]
## trt2                2.946028e-01
## celltypesmallcell   8.615605e-01
## celltypeadeno       1.196066e+00
```

```
## celltypelarge      4.012917e-01
## karno             -3.281533e-02
## diagtime           8.132051e-05
## age               -8.706475e-03
## prior1             7.159360e-02
```

```
# We only choose the variables we need for the data matrix.
data_set = as.matrix(select(final_data, - time, -status)[,-1])
dim(data_set)
```

```
## [1] 137   8
```

```
# Now we matrix mulitply the data matrix with the coefficients to get the
# X*betas which we will then exponentiate exp(X*betas)
coef_matrix= t(t(coeffi) %*% t(data_set))
dim(coef_matrix)
```

```
## [1] 137   1
```

```
#  exponentiate exp(X*betas)
exp_coef_matrix= exp(coef_matrix)
dim(exp_coef_matrix)
```

```
## [1] 137   1
```

```
head(exp_coef_matrix)
```

```
##         [,1]
## 1 0.07660468
## 2 0.06189679
## 3 0.10030681
## 4 0.08671729
## 5 0.06139017
## 6 0.33874177
```

```
# Now we will extract the cumulative base hazard which we need to multiply with
# exp(X*betas). This will give us exp(X*betas)*hz.
bh = basehaz(model, centered = FALSE)
hz  = left_join(final_data, bh) %>% select(hazard)
```

```
## Joining, by = "time"
```

```
dim(hz)
```

```
## [1] 137   1
```

```
head(hz)
```

```
##       hazard
## 1  3.3250753
## 2 24.0734863
## 3 11.8105327
## 4  6.7619120
## 5  6.3843022
## 6  0.4081395
```

```
# Now we get the exp_coef_matrix*hz = exp(X*betas)*hz which is the expected
# number of deaths at time interval t.
expected = hz*exp_coef_matrix
colnames(expected) = "expected"
```

```r
dim(expected)
```

```
## [1] 137   1
```

```r
head(expected)
```

```
##    expected
## 1 0.2547163
## 2 1.4900714
## 3 1.1846769
## 4 0.5863747
## 5 0.3919334
## 6 0.1382539
```

```r
# Now we subtract the observed number of deaths by our expected number of deaths
# at each time t and this gives us the manually calculated martingale residuals.

observed = final_data$status
martingale_res = observed - expected
```

We will now compare our manually obtained martingales with the correct martingales obtained from the model directly.

```r
head(martingale_res)
```

```
##     expected
## 1  0.7452837
## 2 -0.4900714
## 3 -0.1846769
## 4  0.4136253
## 5  0.6080666
## 6  0.8617461
```

```r
head(martingaleres_correct)
```

```
##          1          2          3          4          5          6
##  0.7452837 -0.4900714 -0.1846769  0.4136253  0.6080666  0.8671830
```

```r
diff = martingale_res - martingaleres_correct

head(diff)
```

```
##        expected
## 1  0.000000e+00
## 2 -2.220446e-16
## 3  0.000000e+00
## 4  1.110223e-16
## 5  1.110223e-16
## 6 -5.436953e-03
```

```r
sum(diff)
```

```
## [1] -0.79699
```

From the above we can see that the actual martingales "martingaleres_correct" and our martingales are equal, upto approximation error. The approximation error is -0.79699.

```r
summary(martingale_res)
```

```
##     expected
```

```
##  Min.   :-6.712211
##  1st Qu.:-0.333488
##  Median : 0.274991
##  Mean   :-0.005817
##  3rd Qu.: 0.626345
##  Max.   : 0.989810
```

```
summary(martingaleres_correct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -6.7122 -0.3335  0.2750  0.0000  0.6380  0.9924
```

```
sum(martingale_res)
```

```
## [1] -0.79699
```

```
sum(martingaleres_correct)
```

```
## [1] 1.057487e-14
```

From the above we can see that martingales that we obtained equal to zero, upto the approximation error.

4.2 QUestion 5 b.

We do question 5 b in a separate attachement.