PSTAT175 Project

Joseph Chang, Jonathan Rosal, Rohan Majumdar

December 11th, 2022

LIBRARIES

```
library(knitr)
library(dplyr)
library(tidyverse)
library(survival)
library(survminer)
```

Introduction

Cardiovascular diseases, also known as CVDs, are the number 1 cause of deaths around the world. Every year, CVDs account for an estimated 17.9 million deaths, which is 31% of all deaths worldwide. Most cardiovascular diseases can be caused by daily behaviors such as tobacco use, unhealthy diet and obesity, physical inactivity, and harmful use of alcohol. The most common event caused by CVDs is heart failure. People with CVDs, hypertension, diabetes, hyperlipidaemia or another established disease, are at high risk. They need early detection and management. This project uses a survival analysis model and contains 12 features that can predict mortality by heart failure (Larxel 2020).

The features include:

- Age
- Anaemia - decrease of red blood cells, or hemoglobin (Yes = 1 / No =0)
- Creatinine Phosphokinase - Level of the CPK enzyme in the blood (mcg/L)
- Diabetes - If the patient has Diabetes (Yes = 1 / No = 0)
- Ejection Fraction - Percentage of blood leaving the heart at each contraction (percentage)
- High Blood Pressure - If patient has hypertension (Yes = 1 / No = 0)
- Platelets - Platelets in the blood (kilo-platelets/mL)
- Serum Creatinine - Level of serum creatinine in the blood (mg/dl)
- Serum Sodium - Level of serum sodium in the blood (mEq/L)
- Sex - Female or Male (Male = 1 / Female = 0)
- Smoking - If the patient smokes or not (Yes = 1 / No = 0)
- Time - The follow-up period
- Death Event - If the patient deceased during the follow-up period (Yes = 1 / No = 0)

With the exception of the feature Age (we think age is an important indicator of heart failure), we want to reduce the number of features and observe only the features that give blood measurements and are associated with blood sampling. In particular, the features we want to examine are age, anaemia, creatinine_phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, and serum_sodium, all of which will be measured against time. Thus, the objective of this project is to determine which features related to blood sampling are most significant with time. Measuring the features that are related to blood measurements play an important role in researching into heart failures, which could crucially help doctors, researchers, and those in the medical field focus on the cause of heart failures, help advance medical practices, and potentially save lives in the future.

Some limitations of our project include the inability to change the known factors of age or sex. The Age factor is only between 40 to 95, and thus does not consider younger or older people than that age range. This could affect predictions and alter our data. The Sex factor shows that there is nearly double the amount of males than females. Since there is more data on males (194) than females (105), this could result in a prediction that applies more closely to males than females. Other than these factors, all the other factors can be moderated to some extent.

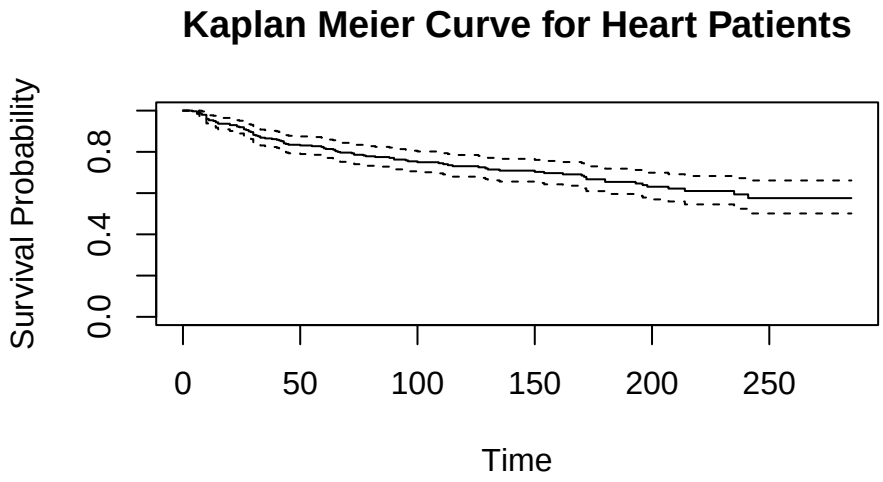
Data Importation

```
heart <- read.csv("heart_failure_clinical_records_dataset.csv")
summary(heart)
```

```
##       age           anaemia       creatinine_phosphokinase    diabetes
##  Min.   :40.00   Min.   :0.0000   Min.   :  23.0           Min.   :0.0000
##  1st Qu.:51.00   1st Qu.:0.0000   1st Qu.: 116.5           1st Qu.:0.0000
##  Median :60.00   Median :0.0000   Median : 250.0           Median :0.0000
##  Mean   :60.83   Mean   :0.4314   Mean   : 581.8           Mean   :0.4181
##  3rd Qu.:70.00   3rd Qu.:1.0000   3rd Qu.: 582.0           3rd Qu.:1.0000
##  Max.   :95.00   Max.   :1.0000   Max.   :7861.0           Max.   :1.0000
##  ejection_fraction high_blood_pressure   platelets      serum_creatinine
##  Min.   :14.00     Min.   :0.0000      Min.   : 25100   Min.   :0.500
##  1st Qu.:30.00     1st Qu.:0.0000      1st Qu.:212500   1st Qu.:0.900
##  Median :38.00     Median :0.0000      Median :262000   Median :1.100
##  Mean   :38.08     Mean   :0.3512      Mean   :263358   Mean   :1.394
##  3rd Qu.:45.00     3rd Qu.:1.0000      3rd Qu.:303500   3rd Qu.:1.400
##  Max.   :80.00     Max.   :1.0000      Max.   :850000   Max.   :9.400
##   serum_sodium        sex            smoking           time
##  Min.   :113.0   Min.   :0.0000   Min.   :0.0000   Min.   :  4.0
##  1st Qu.:134.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 73.0
##  Median :137.0   Median :1.0000   Median :0.0000   Median :115.0
##  Mean   :136.6   Mean   :0.6488   Mean   :0.3211   Mean   :130.3
##  3rd Qu.:140.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:203.0
##  Max.   :148.0   Max.   :1.0000   Max.   :1.0000   Max.   :285.0
##   DEATH_EVENT
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3211
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

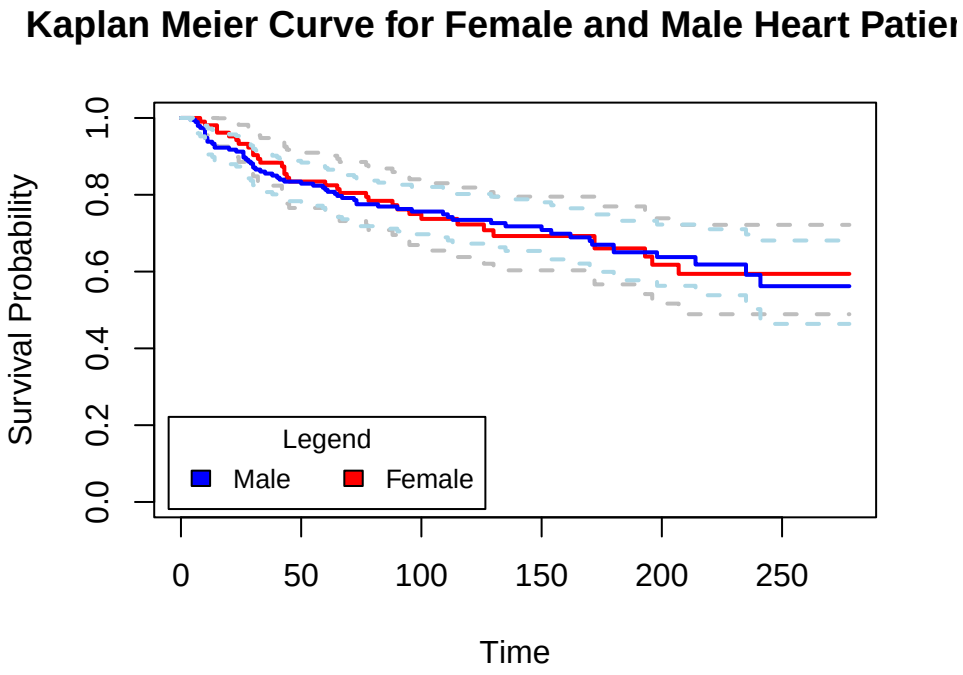
Kaplan Meier Plot

```
# fit kaplan meier curve
heart_surv <- Surv(heart$time, heart$DEATH_EVENT)
heart_fit <- surv_fit(heart_surv ~ 1, data = heart)

# plot kaplan meier curve
plot(heart_fit, xlab = "Time", ylab = "Survival Probability",
     main = "Kaplan Meier Curve for Heart Patients")
```



Kaplan Meier Plot for each Gender

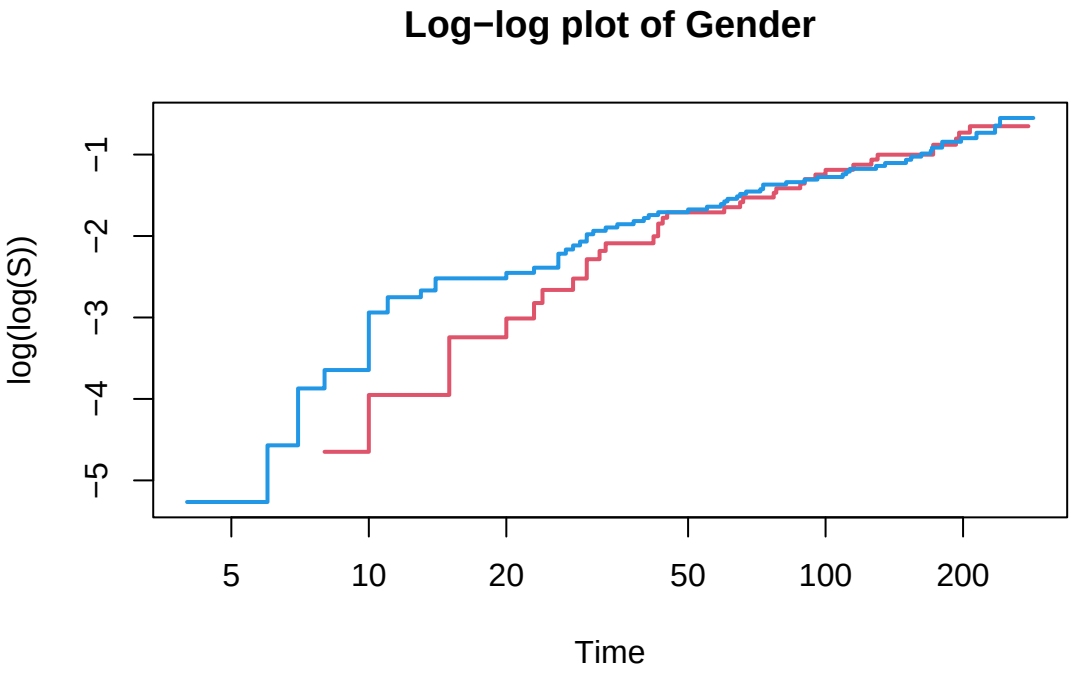
```
# survival vector for female
heart_surv_female <- Surv(heart$time[heart$sex == 0], heart$DEATH_EVENT[heart$sex == 0])

# survival vector for male
heart_surv_male <- Surv(heart$time[heart$sex == 1], heart$DEATH_EVENT[heart$sex == 1])

# fit kaplan meier curve
heart_fit_female <- survfit(heart_surv_female ~ 1)
heart_fit_male <- survfit(heart_surv_male ~ 1)

# plot the male and female survival probability
plot(heart_fit_female, xlab= "Time", ylab = "Survival Probability",
     main = "Kaplan Meier Curve for Female and Male Heart Patients",
     col = c("red","grey","grey"), lwd = 2)
lines(heart_fit_male, xlab= "Time", ylab = "Survival Probability",
      col = c("blue","light blue","light blue"), lwd = 2)
legend("bottomleft", inset=.02, title="Legend",
   c("Male","Female"), fill=c("Blue", "Red"), horiz=TRUE, cex=0.8)
```



From the plot, we see that survival function for women and men are around the same.

Log-Log Plot for each Gender on Time

```
plot(survfit(Surv(time, DEATH_EVENT) ~ sex, data = heart), fun = "cloglog",
     lwd = 2, col =c(2, 4), xlab="Time", ylab = "log(log(S))",
     main = "Log-log plot of Gender")
```



We see that the log-log survival plot is parallel at first, but starts to intersect as at around time 45. There is evidence that the Cox Proportional Hazards assumption is not reasonable for this model.

Model Fitting

Stepwise Function

```
library(MASS)
library(tidyr)
heart1 <- heart[,-c(12, 13)]
fit1 <- coxph(heart_surv ~ ., data = heart1)
fit2 <- coxph(heart_surv ~ 1, heart1) ## base model
stepAIC(fit2, direction="forward", scope=list(upper=fit1,lower=fit2))
```

```
## Start:  AIC=1018.41
## heart_surv ~ 1
##
##                             Df     AIC
## + age                        1  996.90
## + ejection_fraction          1  999.86
## + serum_creatinine           1 1002.47
## + serum_sodium               1 1010.18
## + high_blood_pressure        1 1016.22
## + anaemia                    1 1017.73
## <none>                         1018.41
## + creatinine_phosphokinase   1 1019.31
## + platelets                  1 1019.86
## + diabetes                   1 1020.37
## + sex                        1 1020.41
```

```
## + smoking                   1 1020.41
##
## Step:  AIC=996.9
## heart_surv ~ age
##
##                            Df    AIC
## + ejection_fraction         1 974.11
## + serum_creatinine          1 981.95
## + serum_sodium              1 988.88
## + high_blood_pressure       1 995.05
## <none>                        996.90
## + anaemia                   1 996.99
## + creatinine_phosphokinase  1 997.57
## + platelets                 1 998.40
## + diabetes                  1 998.60
## + smoking                   1 998.86
## + sex                       1 998.89
##
## Step:  AIC=974.11
## heart_surv ~ age + ejection_fraction
##
##                            Df    AIC
## + serum_creatinine          1 957.90
## + serum_sodium              1 970.24
## + high_blood_pressure       1 971.06
## + anaemia                   1 973.20
## <none>                        974.11
## + creatinine_phosphokinase  1 975.12
## + sex                       1 975.46
## + platelets                 1 975.93
## + diabetes                  1 975.94
## + smoking                   1 976.03
##
## Step:  AIC=957.9
## heart_surv ~ age + ejection_fraction + serum_creatinine
##
##                            Df    AIC
## + high_blood_pressure       1 955.10
## + anaemia                   1 956.91
## <none>                        957.90
## + serum_sodium              1 957.91
## + creatinine_phosphokinase  1 958.04
## + diabetes                  1 959.08
## + sex                       1 959.33
## + platelets                 1 959.89
## + smoking                   1 959.90
##
## Step:  AIC=955.1
## heart_surv ~ age + ejection_fraction + serum_creatinine + high_blood_pressure
##
##                            Df    AIC
## + anaemia                   1 954.58
## + serum_sodium              1 954.66
## + creatinine_phosphokinase  1 954.78
## <none>                        955.10
## + diabetes                  1 956.43
```

```
## + sex                        1 956.80
## + platelets                  1 957.04
## + smoking                    1 957.10
##
## Step:  AIC=954.58
## heart_surv ~ age + ejection_fraction + serum_creatinine + high_blood_pressure +
##     anaemia
##
##                             Df    AIC
## + creatinine_phosphokinase  1 953.39
## + serum_sodium              1 953.44
## <none>                        954.58
## + diabetes                  1 955.95
## + sex                       1 956.25
## + platelets                 1 956.50
## + smoking                   1 956.58
##
## Step:  AIC=953.39
## heart_surv ~ age + ejection_fraction + serum_creatinine + high_blood_pressure +
##     anaemia + creatinine_phosphokinase
##
##                Df    AIC
## + serum_sodium  1 951.83
## <none>            953.39
## + diabetes      1 954.68
## + sex           1 954.89
## + platelets     1 955.31
## + smoking       1 955.39
##
## Step:  AIC=951.83
## heart_surv ~ age + ejection_fraction + serum_creatinine + high_blood_pressure +
##     anaemia + creatinine_phosphokinase + serum_sodium
##
##             Df    AIC
## <none>         951.83
## + sex        1 953.16
## + diabetes   1 953.42
## + platelets  1 953.80
## + smoking    1 953.82

## Call:
## coxph(formula = heart_surv ~ age + ejection_fraction + serum_creatinine +
##     high_blood_pressure + anaemia + creatinine_phosphokinase +
##     serum_sodium, data = heart1)
##
##                                coef  exp(coef)   se(coef)      z        p
## age                         4.357e-02  1.045e+00  8.831e-03  4.934 8.05e-07
## ejection_fraction          -4.747e-02  9.536e-01  1.027e-02 -4.621 3.82e-06
## serum_creatinine            3.139e-01  1.369e+00  6.895e-02  4.552 5.31e-06
## high_blood_pressure         4.965e-01  1.643e+00  2.137e-01  2.324   0.0201
## anaemia                     4.460e-01  1.562e+00  2.150e-01  2.074   0.0380
## creatinine_phosphokinase    2.101e-04  1.000e+00  9.825e-05  2.138   0.0325
## serum_sodium               -4.569e-02  9.553e-01  2.336e-02 -1.956   0.0505
##
## Likelihood ratio test=80.58  on 7 df, p=1.048e-14
## n= 299, number of events= 96
```

```
best_model <- coxph(heart_surv ~ age + ejection_fraction + serum_creatinine +
                      high_blood_pressure + anaemia + creatinine_phosphokinase +
                      serum_sodium, heart)
summary(best_model)
```

```
## Call:
## coxph(formula = heart_surv ~ age + ejection_fraction + serum_creatinine +
##     high_blood_pressure + anaemia + creatinine_phosphokinase +
##     serum_sodium, data = heart)
##
##   n= 299, number of events= 96
##
##                               coef  exp(coef)   se(coef)      z Pr(>|z|)
## age                       4.357e-02  1.045e+00  8.831e-03  4.934 8.05e-07 ***
## ejection_fraction        -4.747e-02  9.536e-01  1.027e-02 -4.621 3.82e-06 ***
## serum_creatinine          3.139e-01  1.369e+00  6.895e-02  4.552 5.31e-06 ***
## high_blood_pressure       4.965e-01  1.643e+00  2.137e-01  2.324   0.0201 *
## anaemia                   4.460e-01  1.562e+00  2.150e-01  2.074   0.0380 *
## creatinine_phosphokinase  2.101e-04  1.000e+00  9.825e-05  2.138   0.0325 *
## serum_sodium             -4.569e-02  9.553e-01  2.336e-02 -1.956   0.0505 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## age                          1.0445     0.9574    1.0266     1.063
## ejection_fraction            0.9536     1.0486    0.9346     0.973
## serum_creatinine             1.3688     0.7306    1.1957     1.567
## high_blood_pressure          1.6430     0.6086    1.0808     2.498
## anaemia                      1.5621     0.6402    1.0249     2.381
## creatinine_phosphokinase     1.0002     0.9998    1.0000     1.000
## serum_sodium                 0.9553     1.0468    0.9126     1.000
##
## Concordance= 0.738  (se = 0.027 )
## Likelihood ratio test= 80.58  on 7 df,   p=1e-14
## Wald test            = 88.43  on 7 df,   p=3e-16
## Score (logrank) test = 87.66  on 7 df,   p=4e-16
```

We will use the following covariates: age, ejection_fraction, serum_creatinine, high_blood_pressure, anaemia, creatinine_phosphokinase, and serum_sodium.

Check Proportional Hazards Assumptions

```
res_best_model <- best_model
test.ph <- cox.zph(res_best_model)
test.ph
```

```
##                             chisq df     p
## age                       0.05920  1 0.808
## ejection_fraction         4.76495  1 0.029
## serum_creatinine          1.67518  1 0.196
## high_blood_pressure       0.00943  1 0.923
## anaemia                   0.00531  1 0.942
```

```
## creatinine_phosphokinase  0.98930  1 0.320
## serum_sodium               0.09377  1 0.759
## GLOBAL                    10.52084  7 0.161
```

The covariate ejection_fraction has a p-value less than 0.05, which is statistically significant. Therefore, it does not meet the proportional hazards assumption. We will try again and use the same covariates from our best model, but stratify for the categorical variables of anaemia and high_blood_pressure because the data for each covariate was binary (0 or 1 values).

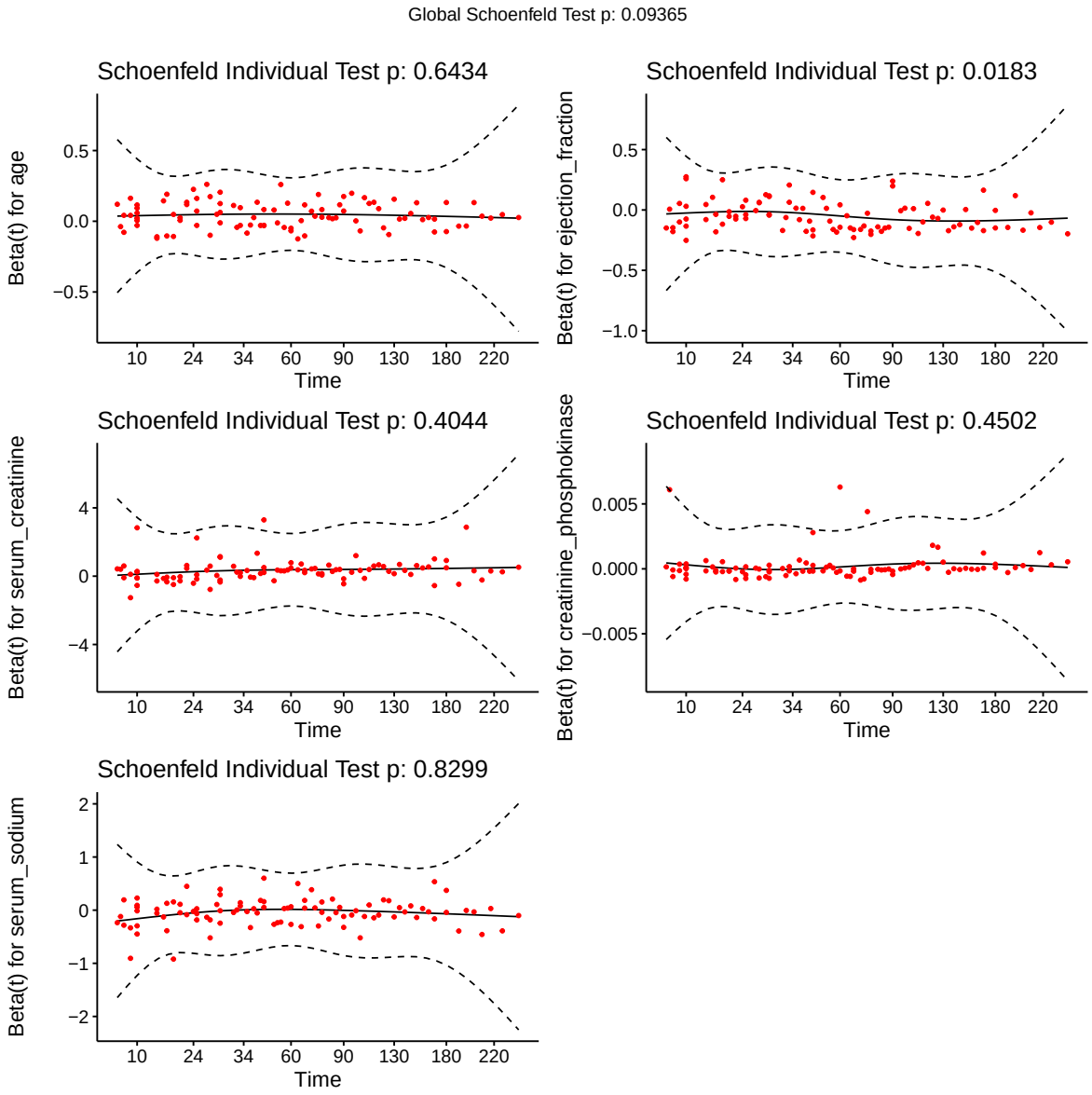
```
best_model2 <- coxph(heart_surv ~ age + ejection_fraction +
                       serum_creatinine + strata(high_blood_pressure) + strata(anaemia) +
                       creatinine_phosphokinase + serum_sodium, heart)
res_best_model2 <- best_model2
test.ph2 <- cox.zph(res_best_model2)
test.ph2
```

```
##                            chisq df     p
## age                       0.2144  1 0.643
## ejection_fraction         5.5685  1 0.018
## serum_creatinine          0.6952  1 0.404
## creatinine_phosphokinase 0.5701  1 0.450
## serum_sodium              0.0462  1 0.830
## GLOBAL                    9.4139  5 0.094
```

From the output above, the covariates are all statistically insignificant, with the exception of ejection_fraction again. The global test is also not statistically significant. Other than ejection_fraction, we can see the proportional hazards assumption is relevant.

Graphical Test of Proportional Hazards

```
ggcoxzph(test.ph2)
```

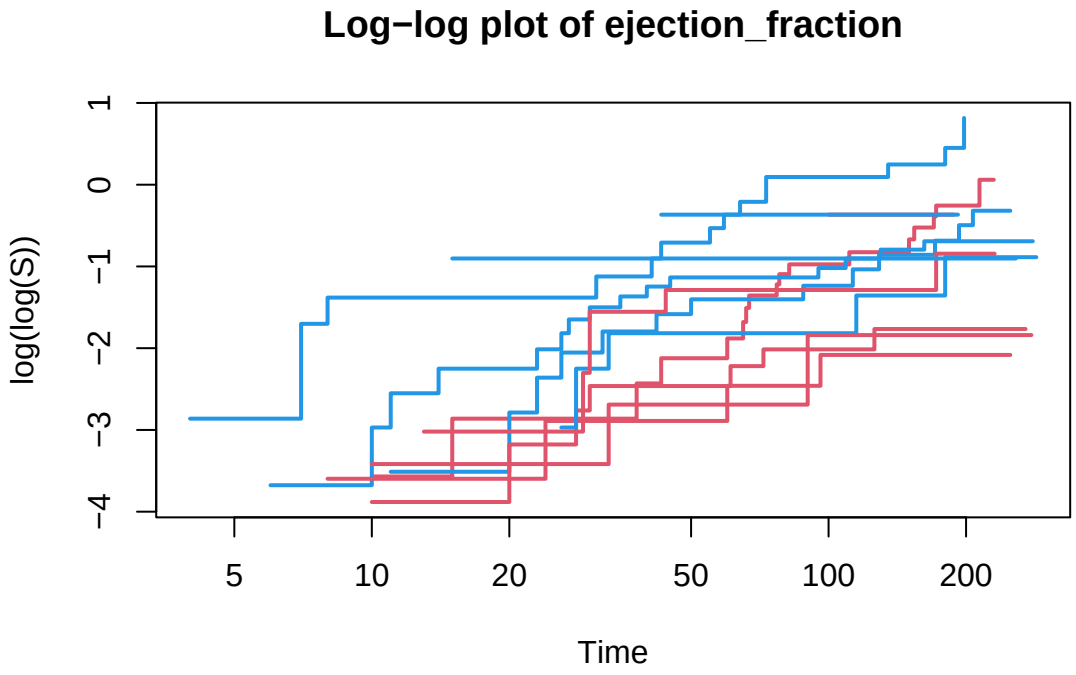


From the Schoenfeld Residual test, we see the residuals (red dots) deviate from a horizontal line centered at Y = 0. We also see that the residuals are centered at 0 and for the most part, constant with time. The residuals are random, and there is no pattern with time.

Log-Log Plot for Ejection Fraction

Since ejection_fraction is significant, we will create a log-log plot to examine the issue.

```
plot(survfit(Surv(time, DEATH_EVENT) ~ factor(ejection_fraction), data = heart), fun = "cloglog",
     lwd = 2, col =c(2, 4), xlab="Time", ylab = "log(log(S))",
     main = "Log-log plot of ejection_fraction")
```



We can see that the proportional hazards assumptions for ejection_fraction is not met. However, based from the Schoenfeld Residual test for ejection_fraction, we see its residuals are centered at 0 and for the most part, constant with time. Since it is better not to remove this covariate from the model, we will just keep ejection_fraction in the model, but not include it in the conclusion.

Summary of Model

```
summary(res_best_model2)
```

```
## Call:
## coxph(formula = heart_surv ~ age + ejection_fraction + serum_creatinine +
##     strata(high_blood_pressure) + strata(anaemia) + creatinine_phosphokinase +
##     serum_sodium, data = heart)
##
##   n= 299, number of events= 96
##
##                              coef  exp(coef)   se(coef)      z Pr(>|z|)
## age                      4.437e-02  1.045e+00  8.899e-03  4.986 6.17e-07 ***
## ejection_fraction       -4.828e-02  9.529e-01  1.042e-02 -4.631 3.64e-06 ***
## serum_creatinine         3.271e-01  1.387e+00  7.368e-02  4.439 9.02e-06 ***
## creatinine_phosphokinase 2.032e-04  1.000e+00  9.688e-05  2.097   0.0360 *
## serum_sodium            -4.689e-02  9.542e-01  2.368e-02 -1.980   0.0477 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                          exp(coef) exp(-coef) lower .95 upper .95
## age                         1.0454     0.9566    1.0273    1.0638
## ejection_fraction           0.9529     1.0495    0.9336    0.9725
## serum_creatinine            1.3869     0.7210    1.2004    1.6024
```

```
## creatinine_phosphokinase    1.0002     0.9998    1.0000    1.0004
## serum_sodium                0.9542     1.0480    0.9109    0.9995
##
## Concordance= 0.729  (se = 0.03 )
## Likelihood ratio test= 73.41  on 5 df,   p=2e-14
## Wald test            = 74.21  on 5 df,   p=1e-14
## Score (logrank) test = 78.98  on 5 df,   p=1e-15
```

Conclusion

We attempted to figure out which covariates associated with blood sampling were most significant with time. The first thing we did was to check the effect gender had on CVDs. We determined that men and women had similar survival functions, but did not meet the proportional hazards assumption due to the log-log plot. We then used stepwise AIC in order to determine which covariates were the best to be used in our model, and found those covariates to be age, ejection_fraction, serum_creatinine, high_blood_pressure, anaemia, creatinine_phosphokinase, and serum_sodium. Next, using the cox.zph() function, we observed which covariates in our model met the proportional hazards assumption, and we found that ejection_fraction (with a p-value of 0.029), did not meet or satisfy the proportional hazards assumption. We then created another model that stratified for categorical variables, such as high_blood_pressure and anaemia. We fit an effect from ejection_fraction but found it did not have a corresponding parameter estimate. After observing the residuals for each covariate, we then decided to keep the covariate ejection_fraction in our model. To sum up, of the list of covariates associated with blood measurements, we found serum_creatinine, high_blood_pressure, anaemia, creatinine_phosphokinase, and serum_sodium to be most important and most significant with time.

Interpretation of each covariate in 95% confidence intervals

Every time the serum_creatinine increases by a unit of 1, the hazard ratio increases by 38.69 percent, meaning the survival ratio went down 38.69 percent. The 95 percent confidence interval for the hazard ratio for the covariate serum_creatinine is (0.9336, 0.9725).

Every time the creatinine_phosphokinase increases by a unit of 1, the hazard ratio increases by 0.02 percent, meaning the survival ratio went down 0.02 percent. The 95 percent confidence interval for the hazard ratio for the covariate creatinine_phosphokinase is (1.0000, 1.0004).

Every time the serum_sodium increases by a unit of 1, the hazard ratio decreases by 0.046 percent, meaning the survival ratio increases by 0.046 percent. The 95 percent confidence interval for the hazard ratio for the covariate serum_sodium is (0.91, 0.99).

Note: confidence intervals for high_blood_pressure and anaemia were omitted because they were stratified and considered categorical variables.

Advanced Methods

Median: Day 115

We will use the time-splitting method. Here, we will find the times in quantiles.

```
heart.new <- heart
heart.new$time2 <- heart$time + (heart$time == 0)/2
quantile(heart.new$time2)
```

```
##   0%  25%  50%  75% 100%
##    4   73  115  203  285
```

Then, we will apply the cut and create a new dataframe. Day 115 is the median for time so we will use that for the cut.

```
split.heart <- survSplit(Surv(time2, DEATH_EVENT) ~ age + ejection_fraction
  + serum_creatinine + strata(high_blood_pressure) + strata(anaemia) +
  creatinine_phosphokinase + serum_sodium, data = heart.new, cut = 115,
  start = "start", episode = "Epi", end = "stop", id = "subject")
```

Test Significance for serum_creatinine

We will create a model to find if serum_creatinine is significantly different before and after the 115th follow-up day.

```
cox.split.serum_creatinine <- coxph(Surv(start, stop, DEATH_EVENT) ~
                                      serum_creatinine:strata(Epi),
                                    data = split.heart)

summary(cox.split.serum_creatinine)
```

```
## Call:
## coxph(formula = Surv(start, stop, DEATH_EVENT) ~ serum_creatinine:strata(Epi),
##     data = split.heart)
##
##   n= 447, number of events= 96
##
##                                       coef exp(coef) se(coef)     z Pr(>|z|)
## serum_creatinine:strata(Epi)Epi=1 0.28455   1.32916  0.06293 4.522 6.13e-06 ***
## serum_creatinine:strata(Epi)Epi=2 0.30642   1.35855  0.11043 2.775  0.00552 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                   exp(coef) exp(-coef) lower .95 upper .95
## serum_creatinine:strata(Epi)Epi=1     1.329     0.7524     1.175     1.504
## serum_creatinine:strata(Epi)Epi=2     1.359     0.7361     1.094     1.687
##
## Concordance= 0.667  (se = 0.029 )
## Likelihood ratio test= 17.97  on 2 df,   p=1e-04
## Wald test            = 28.15  on 2 df,   p=8e-07
## Score (logrank) test = 32.44  on 2 df,   p=9e-08
```

Looking at the summary we can see that serum_creatinine is significant both before and after the 115th day of the follow-up period. The p-value before the 115th day is 6.13e-06, and p-value after the 115th day is 0.00552. Both p-values are less than the significance level of 0.05.

Test Significance for creatinine_phosphokinase

We will create another model to find out if creatinine phosphokinase is significant before and after the 115th follow-up day.

```
cox.split.creatinine_phosphokinase <- coxph(Surv(start, stop, DEATH_EVENT) ~
                                        creatinine_phosphokinase:strata(Epi),
                                          data = split.heart)

summary(cox.split.creatinine_phosphokinase)
```

```
## Call:
## coxph(formula = Surv(start, stop, DEATH_EVENT) ~ creatinine_phosphokinase:strata(Epi),
##     data = split.heart)
##
##   n= 447, number of events= 96
##
##                                                 coef exp(coef)  se(coef)     z
## creatinine_phosphokinase:strata(Epi)Epi=1 8.272e-05 1.000e+00 1.146e-04 0.722
## creatinine_phosphokinase:strata(Epi)Epi=2 3.497e-04 1.000e+00 2.808e-04 1.245
##                                           Pr(>|z|)
## creatinine_phosphokinase:strata(Epi)Epi=1    0.470
## creatinine_phosphokinase:strata(Epi)Epi=2    0.213
##
##                                           exp(coef) exp(-coef) lower .95
## creatinine_phosphokinase:strata(Epi)Epi=1         1     0.9999    0.9999
## creatinine_phosphokinase:strata(Epi)Epi=2         1     0.9997    0.9998
##                                           upper .95
## creatinine_phosphokinase:strata(Epi)Epi=1     1.000
## creatinine_phosphokinase:strata(Epi)Epi=2     1.001
##
## Concordance= 0.477  (se = 0.03 )
## Likelihood ratio test= 1.82  on 2 df,   p=0.4
## Wald test            = 2.07  on 2 df,   p=0.4
## Score (logrank) test = 2.11  on 2 df,   p=0.3
```

Looking at the summary we can see creatinine phosphokinase is not significant for both time intervals (p-value before: 0.470 & p-value after: 0.213).

Test Significance for for serum_sodium

We will create another model to find if serum_sodium is significant before and after the 115th follow-up day.

```
cox.split.serum_sodium <- coxph(Surv(start, stop, DEATH_EVENT) ~
                                  serum_sodium:strata(Epi),
                                data = split.heart)

summary(cox.split.serum_sodium)
```

```
## Call:
## coxph(formula = Surv(start, stop, DEATH_EVENT) ~ serum_sodium:strata(Epi),
##     data = split.heart)
##
##   n= 447, number of events= 96
##
##                                    coef exp(coef) se(coef)      z Pr(>|z|)
## serum_sodium:strata(Epi)Epi=1 -0.07266   0.92992  0.02164 -3.358 0.000786 ***
## serum_sodium:strata(Epi)Epi=2 -0.04837   0.95278  0.04455 -1.086 0.277572
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                exp(coef) exp(-coef) lower .95 upper .95
## serum_sodium:strata(Epi)Epi=1     0.9299      1.075    0.8913    0.9702
## serum_sodium:strata(Epi)Epi=2     0.9528      1.050    0.8731    1.0397
##
## Concordance= 0.596  (se = 0.032 )
## Likelihood ratio test= 10.48  on 2 df,   p=0.005
## Wald test            = 12.45  on 2 df,   p=0.002
## Score (logrank) test = 11.94  on 2 df,   p=0.003
```

Looking at the summary we can see that serum_sodium is significant before the 115th day (p-value = 0.00078) but insignficant after the 115th day (p-value = 0.27757).

Test Significance for all variables at Day 115

We will create a final model that combines all variables together to find out which is significant before and after the 115th follow-up day.

```
cox.split.all <- coxph(Surv(start, stop, DEATH_EVENT) ~
                       serum_creatinine:strata(Epi) +
                       creatinine_phosphokinase:strata(Epi) +
                       serum_sodium:strata(Epi), data = split.heart)

summary(cox.split.all)
```

```
## Call:
## coxph(formula = Surv(start, stop, DEATH_EVENT) ~ serum_creatinine:strata(Epi) +
##     creatinine_phosphokinase:strata(Epi) + serum_sodium:strata(Epi),
##     data = split.heart)
##
##   n= 447, number of events= 96
##
##                                              coef  exp(coef)   se(coef)
## serum_creatinine:strata(Epi)Epi=1          0.2656845  1.3043234  0.0665983
## serum_creatinine:strata(Epi)Epi=2          0.3210845  1.3786220  0.1137338
## strata(Epi)Epi=1:creatinine_phosphokinase  0.0001112  1.0001112  0.0001164
## strata(Epi)Epi=2:creatinine_phosphokinase  0.0004796  1.0004797  0.0002918
## strata(Epi)Epi=1:serum_sodium             -0.0628642  0.9390710  0.0219045
## strata(Epi)Epi=2:serum_sodium             -0.0422870  0.9585946  0.0439960
##                                                z Pr(>|z|)
## serum_creatinine:strata(Epi)Epi=1          3.989 6.63e-05 ***
## serum_creatinine:strata(Epi)Epi=2          2.823  0.00476 **
## strata(Epi)Epi=1:creatinine_phosphokinase  0.955  0.33944
## strata(Epi)Epi=2:creatinine_phosphokinase  1.643  0.10030
## strata(Epi)Epi=1:serum_sodium             -2.870  0.00411 **
## strata(Epi)Epi=2:serum_sodium             -0.961  0.33647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                           exp(coef) exp(-coef) lower .95
## serum_creatinine:strata(Epi)Epi=1            1.3043     0.7667    1.1447
## serum_creatinine:strata(Epi)Epi=2            1.3786     0.7254    1.1032
## strata(Epi)Epi=1:creatinine_phosphokinase    1.0001     0.9999    0.9999
```

```
## strata(Epi)Epi=2:creatinine_phosphokinase    1.0005     0.9995    0.9999
## strata(Epi)Epi=1:serum_sodium                0.9391     1.0649    0.8996
## strata(Epi)Epi=2:serum_sodium                0.9586     1.0432    0.8794
##                                            upper .95
## serum_creatinine:strata(Epi)Epi=1             1.4862
## serum_creatinine:strata(Epi)Epi=2             1.7229
## strata(Epi)Epi=1:creatinine_phosphokinase     1.0003
## strata(Epi)Epi=2:creatinine_phosphokinase     1.0011
## strata(Epi)Epi=1:serum_sodium                 0.9803
## strata(Epi)Epi=2:serum_sodium                 1.0449
##
## Concordance= 0.646  (se = 0.032 )
## Likelihood ratio test= 28.3  on 6 df,   p=8e-05
## Wald test            = 37.49  on 6 df,   p=1e-06
## Score (logrank) test = 42.32  on 6 df,   p=2e-07
```

From the output, we see the variables that are significant before day 115 are serum_creatinine and serum_sodium. Variables that are significant after the 115th day is only serum_creatinine.

Cut Off at Different Time Intervals

In order to do more exploration of the data, we are going to see if the variables are significant at different cuts. Referring to the quantile code above the 1st quarter ends at 73 and the 3rd quarter ends at 203. We will test two additional models, one that cuts off at 73 and one that cuts off at 203.

First Quantile: Day 73

Create the new dataframe:

```
split.heart.first_quantile <- survSplit(Surv(time2, DEATH_EVENT) ~ age + ejection_fraction
  + serum_creatinine + strata(high_blood_pressure) + strata(anaemia) +
  creatinine_phosphokinase + serum_sodium, data = heart.new, cut = 73,
  start = "start", episode = "Epi", end = "stop", id = "subject")
```

Create the new model:

```
cox.split.first_quantile <- coxph(Surv(start, stop, DEATH_EVENT) ~
                      serum_creatinine:strata(Epi) +
                      creatinine_phosphokinase:strata(Epi) +
                      serum_sodium:strata(Epi), data = split.heart.first_quantile)
summary(cox.split.first_quantile)
```

```
## Call:
## coxph(formula = Surv(start, stop, DEATH_EVENT) ~ serum_creatinine:strata(Epi) +
##     creatinine_phosphokinase:strata(Epi) + serum_sodium:strata(Epi),
##     data = split.heart.first_quantile)
##
##   n= 522, number of events= 96
##
##                                              coef  exp(coef)   se(coef)
## serum_creatinine:strata(Epi)Epi=1          0.2617176  1.2991595  0.0738450
## serum_creatinine:strata(Epi)Epi=2          0.2965001  1.3451427  0.0918937
## strata(Epi)Epi=1:creatinine_phosphokinase  0.0001476  1.0001476  0.0001132
```

```
## strata(Epi)Epi=2:creatinine_phosphokinase  0.0001735  1.0001736  0.0002535
## strata(Epi)Epi=1:serum_sodium             -0.0592255  0.9424942  0.0245090
## strata(Epi)Epi=2:serum_sodium             -0.0556964  0.9458262  0.0328517
##                                                z Pr(>|z|)
## serum_creatinine:strata(Epi)Epi=1          3.544 0.000394 ***
## serum_creatinine:strata(Epi)Epi=2          3.227 0.001253 **
## strata(Epi)Epi=1:creatinine_phosphokinase  1.304 0.192161
## strata(Epi)Epi=2:creatinine_phosphokinase  0.684 0.493670
## strata(Epi)Epi=1:serum_sodium             -2.416 0.015671 *
## strata(Epi)Epi=2:serum_sodium             -1.695 0.090001 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                           exp(coef) exp(-coef) lower .95
## serum_creatinine:strata(Epi)Epi=1            1.2992     0.7697    1.1241
## serum_creatinine:strata(Epi)Epi=2            1.3451     0.7434    1.1234
## strata(Epi)Epi=1:creatinine_phosphokinase    1.0001     0.9999    0.9999
## strata(Epi)Epi=2:creatinine_phosphokinase    1.0002     0.9998    0.9997
## strata(Epi)Epi=1:serum_sodium                0.9425     1.0610    0.8983
## strata(Epi)Epi=2:serum_sodium                0.9458     1.0573    0.8868
##                                           upper .95
## serum_creatinine:strata(Epi)Epi=1            1.5015
## serum_creatinine:strata(Epi)Epi=2            1.6106
## strata(Epi)Epi=1:creatinine_phosphokinase    1.0004
## strata(Epi)Epi=2:creatinine_phosphokinase    1.0007
## strata(Epi)Epi=1:serum_sodium                0.9889
## strata(Epi)Epi=2:serum_sodium                1.0087
##
## Concordance= 0.638  (se = 0.032 )
## Likelihood ratio test= 26.79  on 6 df,   p=2e-04
## Wald test            = 35.78  on 6 df,   p=3e-06
## Score (logrank) test = 40.51  on 6 df,   p=4e-07
```

Considering only the data before Day 73, the significant variables are: serum_creatinine and serum_sodium. These were also the same variables that were significant in the model when we cut at the 115th day. If we only considered data after the 73rd day, the significant variables is: serum_creatinine. This is also the same variable that is significant from the model cut at 115. Thus, we can tell that conclude that cutting at the first quantile and second quantile yield the same results.

Third Quantile: Day 203

Create the new dataframe:

```
split.heart.third_quantile <- survSplit(Surv(time2, DEATH_EVENT) ~ age + ejection_fraction
  + serum_creatinine + strata(high_blood_pressure) + strata(anaemia) +
  creatinine_phosphokinase + serum_sodium, data = heart.new, cut = 203,
  start = "start", episode = "Epi", end = "stop", id = "subject")
```

Create the new model:

```
cox.split.third_quantile <- coxph(Surv(start, stop, DEATH_EVENT) ~
                      serum_creatinine:strata(Epi) +
                      creatinine_phosphokinase:strata(Epi) +
                      serum_sodium:strata(Epi), data = split.heart.third_quantile)
summary(cox.split.third_quantile)
```

```
## Call:
## coxph(formula = Surv(start, stop, DEATH_EVENT) ~ serum_creatinine:strata(Epi) +
##     creatinine_phosphokinase:strata(Epi) + serum_sodium:strata(Epi),
##     data = split.heart.third_quantile)
##
##   n= 374, number of events= 96
##
##                                                coef  exp(coef)   se(coef)
## serum_creatinine:strata(Epi)Epi=1          0.2750525  1.3165998  0.0575413
## serum_creatinine:strata(Epi)Epi=2         -0.4443545  0.6412380  0.7622609
## strata(Epi)Epi=1:creatinine_phosphokinase  0.0001350  1.0001350  0.0001062
## strata(Epi)Epi=2:creatinine_phosphokinase  0.0020592  1.0020613  0.0010985
## strata(Epi)Epi=1:serum_sodium             -0.0511914  0.9500968  0.0204756
## strata(Epi)Epi=2:serum_sodium             -0.6747305  0.5092937  0.3036023
##                                                z Pr(>|z|)
## serum_creatinine:strata(Epi)Epi=1          4.780 1.75e-06 ***
## serum_creatinine:strata(Epi)Epi=2         -0.583   0.5599
## strata(Epi)Epi=1:creatinine_phosphokinase  1.270   0.2039
## strata(Epi)Epi=2:creatinine_phosphokinase  1.875   0.0608 .
## strata(Epi)Epi=1:serum_sodium             -2.500   0.0124 *
## strata(Epi)Epi=2:serum_sodium             -2.222   0.0263 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                           exp(coef) exp(-coef) lower .95
## serum_creatinine:strata(Epi)Epi=1            1.3166     0.7595    1.1762
## serum_creatinine:strata(Epi)Epi=2            0.6412     1.5595    0.1439
## strata(Epi)Epi=1:creatinine_phosphokinase    1.0001     0.9999    0.9999
## strata(Epi)Epi=2:creatinine_phosphokinase    1.0021     0.9979    0.9999
## strata(Epi)Epi=1:serum_sodium                0.9501     1.0525    0.9127
## strata(Epi)Epi=2:serum_sodium                0.5093     1.9635    0.2809
##                                           upper .95
## serum_creatinine:strata(Epi)Epi=1            1.4738
## serum_creatinine:strata(Epi)Epi=2            2.8566
## strata(Epi)Epi=1:creatinine_phosphokinase    1.0003
## strata(Epi)Epi=2:creatinine_phosphokinase    1.0042
## strata(Epi)Epi=1:serum_sodium                0.9890
## strata(Epi)Epi=2:serum_sodium                0.9234
##
## Concordance= 0.639  (se = 0.032 )
## Likelihood ratio test= 35.3  on 6 df,   p=4e-06
## Wald test            = 37.42  on 6 df,   p=1e-06
## Score (logrank) test = 45.65  on 6 df,   p=3e-08
```

Using only data before the 203rd day, the significant variables are: serum_creatinine and serum_sodium. These variables are the same from the previous models. Using only data after the 203rd day, the only significant variable is serum_sodium. Unlike the other previous models, serum_sodium is now significant after the intended splitting time. The previous models showed serum_creatinine to be significant after the intended splitting time.

Summary of Advanced Methods

We first decided to split the data at time 115 because that was the median time for the data set. Through these splits, we determined that serum creatinine and serum_sodium were statistically significant before the 115 day cut off. After the 115th day, only serum_creatinine was statistically

significant. We can conclude that serum_creatinine was the only significant variable before and after splitting at Day 115.

After we decided to split at the median, we also wanted to perform two other splits: one at the first quantile at Day 73 and another at the third quantile at Day 203. In splitting at Day 73, we found the same results as when we split at the median. Serum creatinine and serum_sodium were statistically significant before the cut off, while only serum_creatinine was statistically significant after the cut off. Therefore, we can conclude that cutting at the first quantile and second quantile produced the same results.

When we split at the third quantile, we saw serum_creatinine and serum_sodium to be significant before the cut off. This was something familiar because at all three splits before the intended time, serum_creatinine and serum_sodium were significant. However, when splitting at the third quantile, the only significant variable after the split was serum_sodium. This differed from the two previous models.

We can conclude that for all splits, the significant variables before each intended time split were serum creatinine and serum_sodium. The significant variable after the split in the first and second quantile was serum creatinine, and in the third quantile, serum_sodium. Depending on when the data is split, some variables will become insignificant and others will become significant.

References

Larxel. "Heart Failure Prediction." *Kaggle*, 20 June 2020, https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data.

UCI Machine Learning Repository: Heart Failure Clinical Records Data Set, https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.