MATH 4130K FINAL PROJECT

HEART FAILURE SURVIVAL ANALYSIS

Dao Feng, Jinqi Guo, Shahzab Hussain

York University
Department of Mathematics
Prof. Kevin McGregor
Apr. 14th, 2023

1 Introduction

Heart failure is a medical condition characterized by the inability of the heart to pump blood efficiently, leading to symptoms such as shortness of breath, fatigue, and fluid retention. It results from various factors such as high blood pressure, diabetes, and some indicators from blood tests, among other underlying health issues. The prevalence of heart failure is high worldwide and particularly so in Ontario, Canada, where around 250,000 individuals are affected (HQO 2019). Despite treatment advancements, heart failure remains a significant public health concern as it is a leading cause of death and disability (Bragazzi, et al., 2021). Therefore, identifying the risk factors associated with mortality is critical. This study aims to comprehensively analyze various potential covariates, including demographic, clinical and lifestyle factors, to understand their impact on the hazards and survival probability of death from heart failure.

2 Data

The time-to-event dataset utilized in our research is sourced from the UCI Learning Repository and includes a total of 299 observed patients diagnosed with heart failure admitted to the Institute of Cardiology and Allied hospital in Faisalabad, Pakistan. Out of all the patients observed, 105 patients were female and the remaining 194 patients were male. The event of interest in our study is death caused by heart failure in which 96 cases of deaths were observed and 203 cases were censored.

The datasets consist of 13 variables of binary, categorical, and numerical variables that describe general information about the subjects such as age and sex, as well as clinical features that describe the patients' medical condition.

The time-to-event is the follow-up period which is defined as the time interval between the initial diagnoses (time of inclusion into the study) and the last recorded observation (death or censoring time).

For the purpose of our analysis, data wrangling techniques were employed to ensure the integrity of the analysis in which certain variables were modified. In addition, all the continuous variables of the original dataset were standardised due to the differences in their respective units.

The covariate Ejection Fraction is a continuous variable in the original dataset, measuring the percentage of blood that an individual's heart pumps each time it beats, was converted into an ordinal class ($EF \leq 40$, $40 < EF \leq 55$, $55 < EF \leq 70$, $EF > 70$) in accordance to the guidelines referenced from

Clinical Features	Description
Age	age of the patient (years)
Anaemia	decrease of red blood cells or hemoglobin (boolean)
High Blood Pressure	if the patient has hypertension (boolean)
Creatinine Phosphokinase (CPK)	level of the CPK enzyme in the blood (mcg/L)
Diabetes	if the patient has diabetes (boolean)
Ejection Fraction	percentage of blood leaving the heart at each contraction (percentage)
Platelets	platelets in the blood (kiloplatelets/mL)
Sex	woman or man (binary)
Serum Creatinine	level of serum creatinine in the blood (mg/dL)
Serum Sodium	level of serum sodium in the blood (mEq/L)
Smoking	if the patient smokes or not (boolean)
Time	follow-up period (days)
Death [Target Event]	if the patient deceased during the follow-up period (boolean)

Penn Medicine.

3 Methodology

The present analysis relies heavily on the use of Kaplan-Meier estimators to compare and visualize survival probabilities among levels within categorical covariates. To estimate the variance of the Kaplan-Meier curve, we employ Greenwood's formula, which is used to build the confidence interval at each value of t. Given that our dataset comprises 96 observed events out of 299 observations, and that it contains a high proportion of censored data, it may not be appropriate to fit any parametric models. Therefore, we employ a Cox proportional hazards regression model:

$$\lambda_i(t, x_i) = \lambda_0(t) exp\{x_i^T \beta\}$$

(where $\lambda_0(t)$ is the baseline hazard)

Which is a semi-parametric method that allows for a more reasonable analysis of this dataset. The Cox proportional hazards model assumes that the hazard ratio between two individuals with different predictors does not depend on time.

To examine this assumption, we use both log-log plots and Schoenfeld residuals in our study. We expect to observe parallel log-log plots for categorical variables and obtain large p-values for all variables in the Schoenfeld residuals test. As previously mentioned, the dataset includes 11 covariates that are all considered relevant to the survival times of heart failure. To avoid biases resulting from a lack of understanding of medical expertise, we use LASSO regression to select significant variables from these covariates.

4 Results

The survival curve before modelling, displayed as below in Fig. 1, shows a gradual and flat trend. The curve indicates that heart failure is a condition that has high survival probability and long survival times, as inferred from the available recorded survival durations.

Kaplan-Meier survival curve as given by the formula:

$$\hat{S} = \prod_{j:t_j \leq t} (1 - d_j/n_j)$$

where d_j/n_j is the estimated hazard for j.

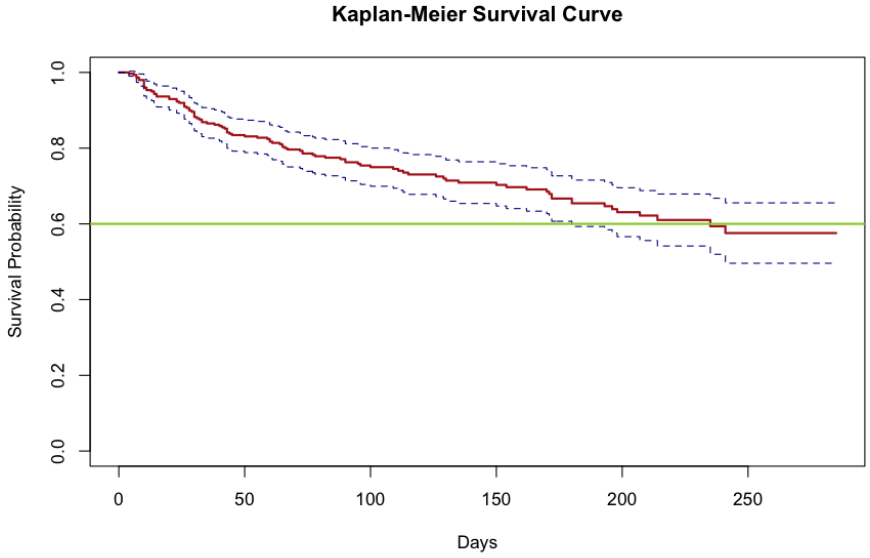


Figure 1: Overall survival curve prior to modelling

4.1 Cox-Proportional Hazard (Full Model)

$$\hat{\lambda}_i(t, x_i) = \hat{\lambda}_0(t) exp\{x_{i1}\hat{\beta}_{age} + x_{i2}\hat{\beta}_{CPK} + x_{i3}\hat{\beta}_{EF} + x_{i4}\hat{\beta}_{platelets} + x_{i5}\hat{\beta}_{SCR} + x_{i6}\hat{\beta}_{SSO} + x_{i7}\hat{\beta}_{anaemia} + x_{i8}\hat{\beta}_{diabetes} + x_{i9}\hat{\beta}_{HBP} + x_{i10}\hat{\beta}_{sex} + x_{i11}\hat{\beta}_{smoking}\}$$

The full cox proportional hazards regression model showed significant results for some covariates among the various included in the analysis. However, as the model is semi-parametric,the focus is on the parametric part, which is the hazard ratio, represented by the $exp\{x_i^T\hat{\beta}\}$. Thus, the significance of covariates is not solely based on their p-values but also on their corresponding $\hat{\beta}$ values that should make the hazard ratio meaningful.

Table 1: Full Model

Characteristic	Hazard Ratio	95% C.I.	p-value
age	1.67	(1.35, 2.07)	<0.001
CPK	1.20	(0.99, 1.46)	0.060
EF	0.69	(0.48, 0.99)	0.042
platelets	0.97	(0.78, 1.20)	0.8
SCR	1.35	(1.19, 1.55)	<0.001
SSO	0.77	(0.64, 0.93)	0.008
anaemia	1.52	(0.99, 2.33)	0.053
diabetes	1.13	(0.73, 1.75)	0.6
HBP	1.65	(1.08, 2.52)	0.022
sex	0.91	(0.56, 1.48)	0.7
smoking	1.16	(0.71, 1.90)	0.6

After the process of selection under the LASSO approach, seven covariates stand out. They are age, CPK, EF, SCR, SSO, anaemia, HBP. By examining both hazard ratio and p-values accordingly, they are remained for further analysis, and other insignificant variables are removed from the model.

It is noted that anaemia and HBP are both categorical with 2 levels, an interaction term between them is included as well.

4.2 Cox-Proportional Hazard (Reduced Model)

$$\hat{\lambda}_i(t, x_i) = \hat{\lambda}_0(t) exp\{x_{i1}\hat{\beta}_{age} + x_{i2}\hat{\beta}_{CPK} + x_{i3}\hat{\beta}_{EF} + x_{i4}\hat{\beta}_{SCR} + x_{i5}\hat{\beta}_{SSO} + x_{i6}\hat{\beta}_{anaemia} + x_{i7}\hat{\beta}_{HBP} + x_{i8}\hat{\beta}_{anaemia*HBP}\}$$

Table 2: Reduced Model

Characteristic	Hazard Ratio	95% C.I.	p-value
age	1.63	(1.33, 2.00)	<0.001
CPK	1.21	(0.99, 1.46)	0.057
EF	0.68	(0.48, 0.98)	0.039
SCR	1.35	(1.18, 1.54)	<0.001
SSO	0.77	(0.64, 0.92)	0.005
anaemia	1.51	(0.99, 2.30)	0.056
HBP	1.65	(1.08, 2.51)	0.020

This model provides valuable insights into the factors that influence the hazard of death for patients with heart failure. Ejection Fraction (EF), a crucial indicator of heart function, was analyzed as an ordinal variable with four levels. With each increase in EF level, the hazard of death decreased by 32%, indicating that better heart function was linked to lower death risk. The SSO, which reflects blood sodium levels, had an expected negative association with the hazard of death, with a one-unit decrease leading to a 24% decrease in hazard. Age and CPK level (an indicator for injury of the heart) are positively associated with the hazard of death, with a one-year increase in age resulting in a 64% increase in the hazard of death and a one-unit increase in CPK level leading to a 22% increase. SCR, an indicator to examine the health status for kidneys, also has a positive relationship with the hazard of death. With a one-unit increase in SCR, the hazard increases by 27%. On the other hand, patients with anaemia and high blood pressure (HBP) have significantly higher hazards of death than those without these conditions, with anaemia associated with an 81% higher hazard and HBP with a 105% higher hazard. However, the interaction between anaemia and HBP is not significant, suggesting that the effect of anaemia on the hazard of death is similar for patients with and without HBP.

The plot shown in Fig. 2 depicts the association between the categorical variables HBP and anaemia and their respective effects while maintaining a constant value for the other covariates. The reference group, comprised of individuals without high blood pressure or anaemia, exhibits the highest survival probability. Conversely, patients with either high blood pressure or anaemia exhibit lower survival rates relative to the reference group, while patients with both conditions exhibit the lowest survival probability, consistent with expectations.

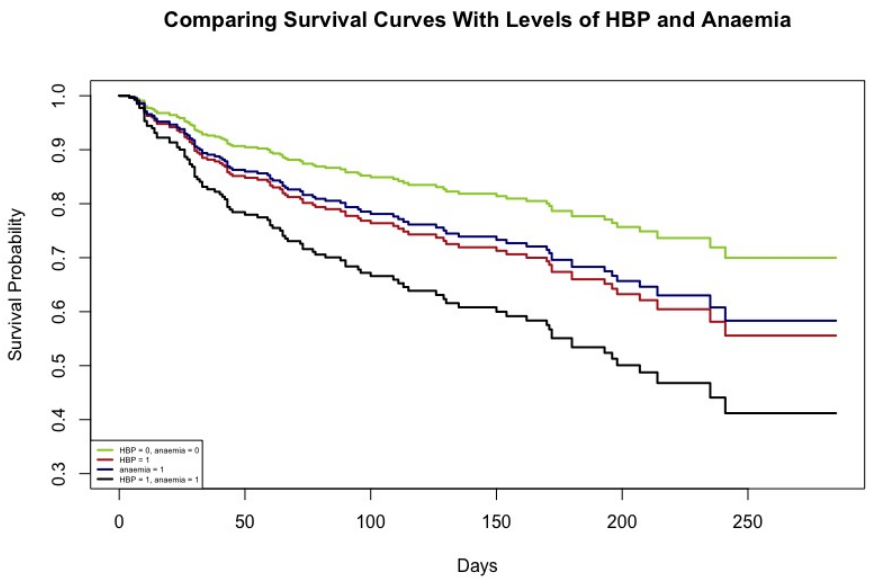


Figure 2: The survival curves with levels of HBP and anaemia

Additionally, Fig. 3 below portrays the survival curves for the ordinal covariate EF with four levels. The green curve corresponds to the normal range of EF between 55 and 70 and exhibits the highest survival probability. The red and black curves represent EF levels below the normal range, indicating previous heart damage and confirming the diagnosis of heart failure, respectively. As anticipated, survival probability decreases as the EF level decreases.

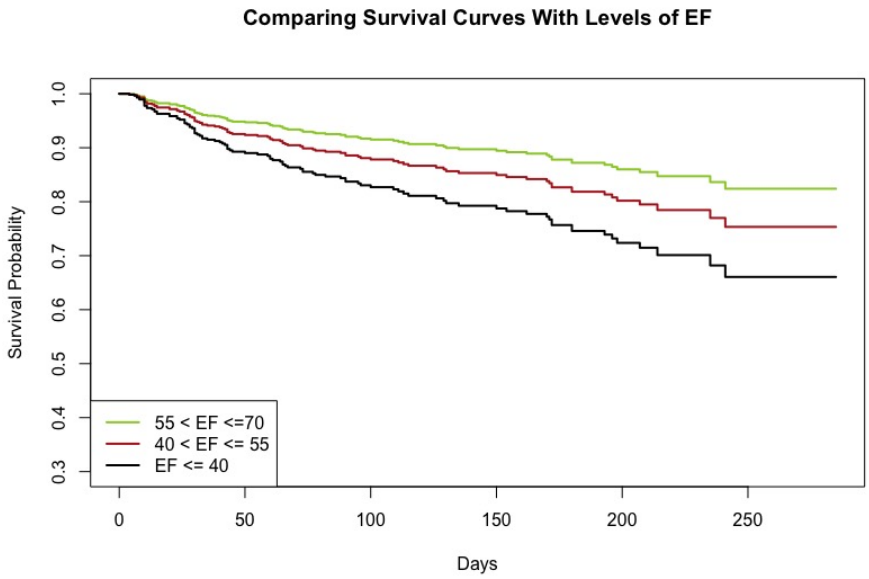


Figure 3: The survival curves with levels of HBP and anaemia

4.3 Checking Model Assumptions

Table 3: χ^2 Test

	χ^2	degrees of freedom	p-value
age	0.0763	1	0.78
CPK	0.8376	1	0.36
EF	0.0850	1	0.77
SCR	0.7984	1	0.37
SSO	0.0686	1	0.79
anaemia	0.0933	1	0.76
HBP	0.1485	1	0.70
GLOBAL	2.7444	7	0.91

Table 3 shows the chi-squared test for each feature in our reduced model.

Log-log plots were created for the two categorical covariates high blood pressure and anaemia using KM estimates. It is observed that both plots show a proportional pattern (refer to Appendix A) after approximately day 20. However, it is noted that the dataset analyzed for this study is small and not well distributed, which could be a reason for the non-parallel before day 20.

The Schoenfeld residuals test is designed to examine whether the proportional hazards assumption of the model is met. The null hypothesis of this test is that the proportional hazards assumption holds, and thus, large p-values are expected for all variables. The test statistics reveal that we have evidence supporting the null hypothesis (refer to Appendix A).

5 Limitations

In terms of limitations to our proposed model, the overall accuracy and statistical confidence of our inference is limited by the data set by factors of experimental design and the chronic nature of the disease which introduces variations in our estimations of survival probability and hazard levels

5.1 Experimental Design Limitations

We must first recognize that the sample size of the dataset is relatively small with only 299 observations, of which a large portion of it (203 observations) were censored. Furthermore, given that this is an observational study, there is a degree of sampling bias in which the data is not representative of the overall population.

Ideally, it is desired to have data from multiple hospitals in multiple countries, spanning multiple continents, amongst different demographics to be able to capture the variations between the overall population. However, the data used in this model was solely collected at one hospital, which limits our interpretation and inference in terms of losing generality to other geographical locations.

Furthermore, given that it is a retrospective study and there is no randomization in the sampling process, therefore our data is prone to selection bias introduced by the specific inclusion criteria in the study design which only include patients who are:

1. of age 40 years and older,
2. has Left Ventricular Systolic Dysfunction of NYHA class III or class IV.

The New York Heart Heart Association (NYHA) classifies the stages of heart failure into four classes with class IV being the most serious.

With this in mind, by the nature of the design of this study and how the dataset was created, we can not infer our model results and findings to patients outside of this inclusion criteria.

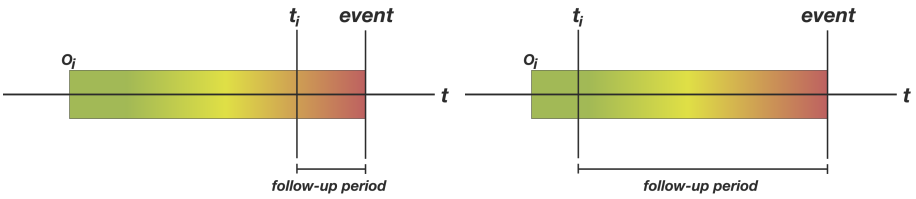
There is a loss of generality in terms of Heart Failure Survival for patients who do not meet this criteria - meaning patients who are younger than 40 years old and have Left Ventricular systolic dysfunction of (Class I or Class II).

5.2 Lead Time / Onset Time Bias

Unlike infectuous diseases where the exact time of contraction and disease onset can be observed, heart failure is a chronic disease such that the symptoms are developed gradually over time. This nature of the disease creates difficulties in determining the precise onset time.Therefore, in our study, we do not have a variable that measures the seriousness of the condition relative to the time of initial diagnoses. Although we know that patients included in the study are either of the NYHA Class III or IV, it is not clear which stage these patients belong to at their initial diagnoses. This poses a challenge by introducing an uncertain variability to our model.

Given that the patient has an event (dies of heart failure), the patients who are introduced to the study at end-stage heart-failure (left diagram) will have a shorter follow-up period and therefore have an decreasing effect on the overall survival probability and increase the hazard. Conversely, if the patient was introduced to the study in an early/healthier stage (right diagram), then they will have a longer follow-up period in which they increase the overall survival of heart failure and decrease the hazard.

Due of this confounding variation in the follow-up period, the accuracy of our survival and hazard models are affected. This is a problem that is common in modeling chronic diseases due to the challenges in measuring exact onset time, however, the inclusion of a potential variable or measurement of the seriousness of the patients illness at the initial time of diagnoses can alleviate this limitation.



6 Conclusions

As we have presented in our research findings, our cox-proportional hazard model identified high blood pressure and anaemia as two of the most significant risk factors associated with increased hazards of mortality from heart failure. Individuals with high blood pressure and anaemia have been found to have lower survival probabilities than those without these conditions. Additionally, age is also a significant covariate that contributes to the death from heart failure, as it is a natural cause of mortality. Another important feature is ejection fraction such that low ejection fraction below 40 percent suggest relatively dangerous risks to heart failure and can significantly increase the hazards of death from heart failure.

Given the limitations to our dataset, a more comprehensive list of features should be collected and improvements to the experimental design aspects such as wider inclusion criterions and bigger samples randomized to the general population should be utilized in effort to produce models without the loss of generality. Furthermore, in consideration to the difficulties of capturing exact disease onset times in a chronic illness like heart failure, a more comprehensive and informative dataset should be explored in effort to produce more statistically accurate models in regards to overall survival.

References

[1] Bragazzi, N. L., Zhong, W., Shu, J., Abu Much, A., Lotan, D., Grupper, A., Younis, A., Dai, H. (2021). Burden of heart failure and underlying causes in 195 countries and territories from 1990 to 2017. European Journal of Preventive Cardiology, 28(15), 1682-1690. https://doi.org/10.1093/eurjpc/zwaa147

[2] Ejection Fraction: What the Numbers Mean. (n.d.). Penn Medicine. Retrieved April 15, 2023, from https://www.pennmedicine.org/updates/blogs/heart-and-vascular-blog/2022/april/ejection-fraction-what-the-numbers-mean

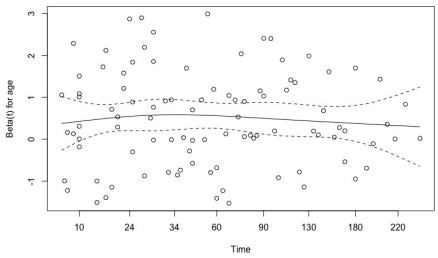
[3] Table: New York heart association (NYHA) classification of heart failure. (n.d.). Merck Manuals Professional Edition. Retrieved April 15, 2023, from https://www.merckmanuals.com/en-ca/professional/multimedia/table/new-york-heart-association-nyha-classification-of-heart-failure

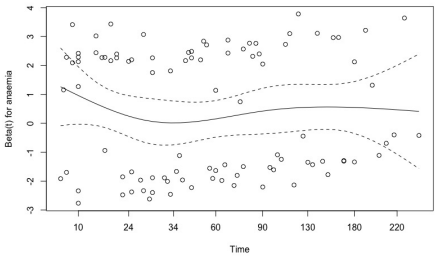
[4] UCI Machine Learning Repository: Heart failure clinical records Data Set. (n.d.). Uci.edu. Retrieved April 15, 2023, from https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

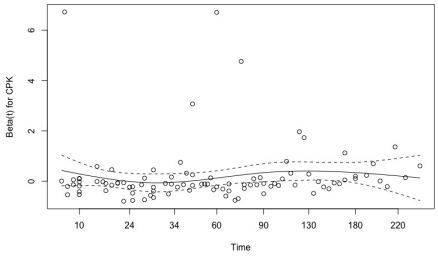
[5] Health Quality Ontario (2019). Heart Failure Care in the Community for Adults. Available from: https://www.hqontario.ca/Portals/0/documents/evidence/quality-standards/qsheart-failure-quality-standard-en.pdf.

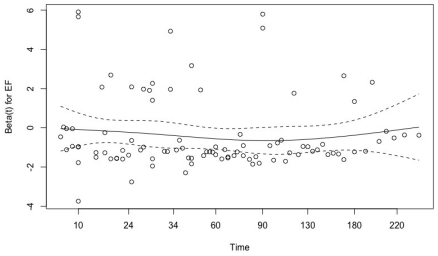
7 Appendix A

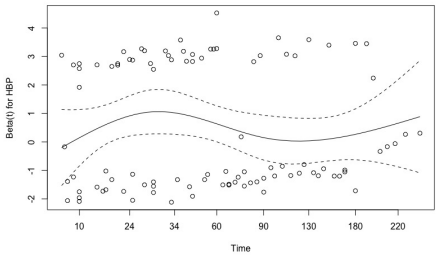
Schoenfeld Residual Plots

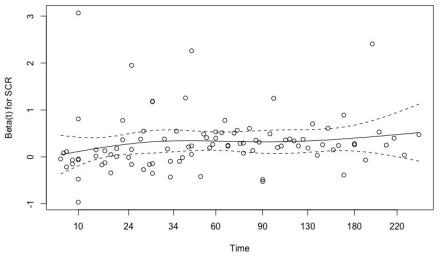


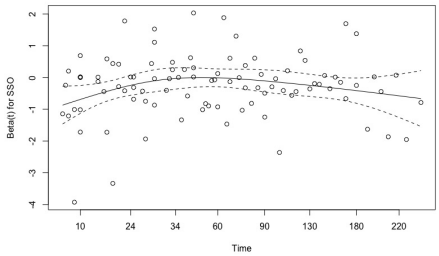




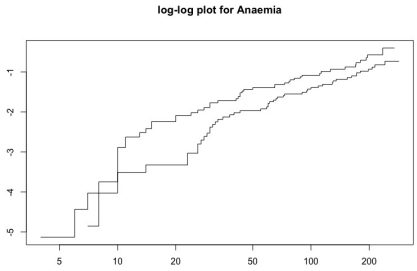


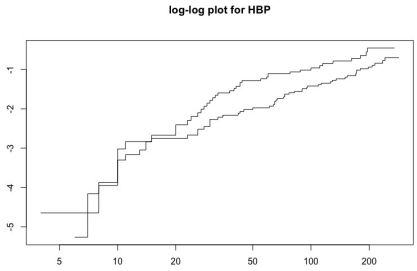






Log-Log Plots





8 Appendix B

```
### MATH 4130k Project ###

heart.data <-read.csv("heart_failure_clinical_records_dataset.csv")
# Rename some variables.
names(heart.data)[names(heart.data) == "creatinine_phosphokinase"] <- "CPK"
names(heart.data)[names(heart.data) == "ejection_fraction"] <- "EF"
names(heart.data)[names(heart.data) == "high_blood_pressure"] <- "HBP"
names(heart.data)[names(heart.data) == "serum_creatinine"] <- "SCR"
names(heart.data)[names(heart.data) == "serum_sodium"] <- "SSO"
names(heart.data)[names(heart.data) == "DEATH_EVENT"] <- "status"

# See how many categorical variables
colSums(heart.data==0)

# Turn variable 'EF' into categorical variable
library(dplyr)
heart.data$EF <- cut(heart.data$EF, breaks=c(10, 40, 55, 70, 80),
                     labels=c('1','2','3','4'))
heart.data$EF <- unclass(heart.data$EF)

# Standardize all the continuous variables
heart.std <- heart.data %>% mutate_at(c('age','CPK','platelets','SCR','SSO'),
                                      ~(scale(.) %>% as.vector))
head(heart.std)

# The overall survival curve by KM estimators
library(survival)
fit <- survfit(Surv(heart.std$time, heart.std$status)~1)
nj <- fit$n.risk
dj <- fit$n.event
## Kaplan-Meier estimator ##
km <- cumprod(1-dj/nj)
v.km <- km^2*cumsum(dj/(nj*(nj-dj))) # Estimating the variance
km <- c(1, km) # Concatenating a 1 before the first event
# Event times
etimes <- c(0, fit$time)
v.km <- c(0, v.km)
# Plotting Kaplan-Meier curve --> an overall picture
plot(etimes, km, type="s", ylim=c(0,1),
     xlab="Days", ylab="Survival_Probability",
     col="firebrick", lwd=2)
# Adding confidence intervals
lines(etimes, km-1.96*sqrt(v.km), type="s", lty=2, col="darkblue")
lines(etimes, km+1.96*sqrt(v.km), type="s", lty=2, col="darkblue")
```

```
abline(h=0.6, col="yellowgreen", lwd=2)
title(main="Kaplan-Meier Survival Curve")

# Cox proportional hazards model (overall)
cph.fit <- coxph(Surv(time,status) ~
                   age+CPK+EF+platelets+SCR+SSO+anaemia
                 +diabetes+HBP+sex+smoking, data=heart.std)
cph.fit
#plot(survfit(cph.fit), ylim=c(0,1), xlab="Days", ylab="S(t)")
#abline(h=0.6, col="green")


library(gtsummary)  #install.packages("gtsummary")
tbl_regression(cph.fit, exponentiate = TRUE)

# Applying LASSO regression to select significant variables
library(glmnet)
X <- model.matrix(cph.fit)
y <- Surv(heart.std$time, heart.std$status)
set.seed(20230414)
cv.lasso <- cv.glmnet(X, y, alpha=1, family="cox", nfolds=10)
best.lambda <- cv.lasso$lambda.min
coef <- coef(cv.lasso, s=best.lambda)
coef

# Reduced model: only includes covariates
# 'age', 'CPK', 'EF', 'SCR', 'SSO', 'anaemia' and 'HBP'
cph.fit2 <- coxph(Surv(time, status) ~ age + CPK + EF + SCR + SSO
                  + anaemia + HBP,
                  data = heart.std)
cph.fit2
tbl_regression(cph.fit2, exponentiate = TRUE)

# Create a survival object for HBP = 0 and anaemia = 0
sf00 <- survfit(cph.fit2, newdata = data.frame(HBP = 0, anaemia = 0,
                                               age = mean(heart.std$age),
                                               CPK = mean(heart.std$CPK),
                                               EF = mean(heart.std$EF),
                                               SCR = mean(heart.std$SCR),
                                               SSO = mean(heart.std$SSO)))
# Create a survival object for HBP = 1 and anaemia = 0
sf10 <- survfit(cph.fit2, newdata = data.frame(HBP = 1, anaemia = 0,
                                               age = mean(heart.std$age),
                                               CPK = mean(heart.std$CPK),
                                               EF = mean(heart.std$EF),
                                               SCR = mean(heart.std$SCR),
```

```
                                                    SSO = mean(heart.std$SSO)))
# Create a survival object for HBP = 0 and anaemia = 1
sf01 <- survfit(cph.fit2, newdata = data.frame(HBP = 0, anaemia = 1,
                                                    age = mean(heart.std$age),
                                                    CPK = mean(heart.std$CPK),
                                                    EF = mean(heart.std$EF),
                                                    SCR = mean(heart.std$SCR),
                                                    SSO = mean(heart.std$SSO)))
# Create a survival object for HBP = 1 and anaemia = 0
sf11 <- survfit(cph.fit2, newdata = data.frame(HBP = 1, anaemia = 1,
                                                    age = mean(heart.std$age),
                                                    CPK = mean(heart.std$CPK),
                                                    EF = mean(heart.std$EF),
                                                    SCR = mean(heart.std$SCR),
                                                    SSO = mean(heart.std$SSO)))
# Plot the survival curves
plot(sf00, col = "yellowgreen", lty = 1, ylim = c(0.3, 1), xlab = "Days",
     ylab = "Survival Probability", lwd=2, conf.int=FALSE)
lines(sf10, col = "firebrick", lty = 1, lwd=2, conf.int=FALSE)
lines(sf01, col = "navy", lty = 1, lwd=2, conf.int=FALSE)
lines(sf11, col = "black", lty = 1, lwd=2, conf.int=FALSE)
legend("bottomleft", legend = c("HBP = 0, anaemia = 0", "HBP = 1",
                                "anaemia = 1",
                                "HBP = 1, anaemia = 1"),
       col = c("yellowgreen", "firebrick", "navy", "black"),
       lty = 1, cex=0.45, bg="white",
       lwd=2)
title(main="Comparing Survival Curves With Levels of HBP and Anaemia",
      line=2.5)



# Create a survival object for EF = 3
sf1 <- survfit(cph.fit2, newdata = data.frame( EF = 3,
                                                    HBP = 0, anaemia = 0,
                                                    age = mean(heart.std$age),
                                                    CPK = mean(heart.std$CPK),
                                                    SCR = mean(heart.std$SCR),
                                                    SSO = mean(heart.std$SSO)))
# Create a survival object for EF = 2
sf2 <- survfit(cph.fit2, newdata = data.frame( EF = 2,
                                                    HBP = 0, anaemia = 0,
                                                    age = mean(heart.std$age),
                                                    CPK = mean(heart.std$CPK),
                                                    SCR = mean(heart.std$SCR),
                                                    SSO = mean(heart.std$SSO)))
```

```
# Create a survival object for EF = 1
sf3 <- survfit(cph.fit2, newdata = data.frame( EF = 1,
                                               HBP = 0, anaemia = 0,
                                               age = mean(heart.std$age),
                                               CPK = mean(heart.std$CPK),
                                               SCR = mean(heart.std$SCR),
                                               SSO = mean(heart.std$SSO)))
# Plot the survival curves
plot(sf1, col = "yellowgreen", lty = 1, ylim = c(0.3, 1), xlab = "Days",
     ylab = "Survival Probability", lwd=2, conf.int=FALSE)
lines(sf2, col = "firebrick", lty = 1, lwd=2, conf.int=FALSE)
lines(sf3, col = "black", lty = 1, lwd=2, conf.int=FALSE)
legend("bottomleft",
       legend = c("55 < EF <=70", "40 < EF <= 55", "EF <= 40"),
       col = c("yellowgreen", "firebrick", "black"),
       lty = 1, cex=1, bg="white",
       lwd=2)
title(main="Comparing Survival Curves With Levels of EF", line=2.5)



# Testing proportional hazards assumption
cz <- cox.zph(cph.fit2)
cz
plot(cz)


# Check assumption for high blood pressure
# (by create KM curves separately for HBP)

km.hbp <- survfit(Surv(time, status)~HBP, data=heart.std)
plot(km.hbp, fun="cloglog")
title(main="log-log plot for HBP", line=2.5)

# Check assumption for anaemia
km.anaemia <- survfit(Surv(time, status)~anaemia, data=heart.std)
plot(km.anaemia, fun="cloglog")
title(main="log-log plot for Anaemia", line=2.5)
```