Factors Influencing Deaths In Heart Failure Patients

Project Groups11
Parth Deshmukh(S3825055), Subbiah Soundarapandian(S3825012), Sudershan Ravi(S3829895)

• **TABLE OF CONTENTS**

• **INTRODUCTION**

In this phase of the project we will focus majorly on the analysis of probability of mortality using logistic regression model. The binary variable “DEATH EVENT” is used as a response variable. In consideration with the previous analysis in phase-1 it is evident that there are various distinct factors impacting mortality in patients with history of heart failure like age, platelet count, serum creatinine, ejection fraction, diabetes, anaemia, blood pressure levels and habit of smoking. So, in order to extract the conclusive outcomes, we will go through various steps in order to filter out the key features for the binomial logistic regression model. The logistic model is tested for model fitting which will provide a summary to differentiate most significant features. Secondly, residual analysis is performed to make it easy to comment on the appropriateness of the model for the observed value of the dependent and predicted values. In order to be sure with the predictions we are making, visualization of the graphs is done for the response analysis. Further, we will go through the performance of the model and comment on the reliability using the goodness of the fit which would be summoned with the confidence intervals and hypothesis test to have a conclusive result. Finally, after all this analysis we will check through the nature of the model by operating it with other variables by the sensitivity analysis to see how the model react with the change in different parameters. After going through all these process we certainly found some predicted outcomes for mortality.

• **STATISTICAL MODELLING**

1. MODEL FITTING

Model Fitting of the dataset is performed in order to analyse the most significant variables which can be selected for the further analysis. As we can see that age, creatinine phosphokinase, ejection fraction, serum creatinine and serum sodium variables are significant in response to the target variable, so we will majorly focus on analyzing these variables only.
So, Finally with the help of these selected variables a new model (mod_fit1) is designed to analyze the further predictions.
Variance-Covariance matrix is obtained where all the on diagonal values gives us the variance and the values other than those interpret the covariance values.

```
setwd("C:/Users/parth/OneDrive/Desktop/SEMESTER 2/Analysis of Categorical Data/Project phase
 1")
assign_ds <- read.csv("Project Groups11_data.csv")

library(ggplot2)

mod_fit <- glm(formula = DEATH_EVENT ~ age+anaemia+creatinine_phosphokinase+diabetes+ejection
_fraction+high_blood_pressure+platelets+serum_creatinine+serum_sodium+sex+smoking,
               family = binomial(link = logit), data = assign_ds)
summary(mod_fit)
```

```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase +
##     diabetes + ejection_fraction + high_blood_pressure + platelets +
##     serum_creatinine + serum_sodium + sex + smoking, family = binomial(link = logit),
##     data = assign_ds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3184  -0.7692  -0.4436   0.8293   2.4880
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               4.964e+00  4.601e+00   1.079 0.280625
## age                       5.569e-02  1.313e-02   4.241 2.23e-05 ***
## anaemia                   4.179e-01  3.009e-01   1.389 0.164904
## creatinine_phosphokinase  2.905e-04  1.428e-04   2.034 0.041907 *
## diabetes                  1.514e-01  2.974e-01   0.509 0.610644
## ejection_fraction        -7.032e-02  1.486e-02  -4.731 2.23e-06 ***
## high_blood_pressure       4.189e-01  3.061e-01   1.369 0.171092
## platelets                -7.094e-07  1.617e-06  -0.439 0.660857
## serum_creatinine          6.619e-01  1.734e-01   3.817 0.000135 ***
## serum_sodium             -5.667e-02  3.338e-02  -1.698 0.089558 .
## sex                      -3.990e-01  3.508e-01  -1.137 0.255394
## smoking                   1.356e-01  3.486e-01   0.389 0.697300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 375.35  on 298  degrees of freedom
## Residual deviance: 294.28  on 287  degrees of freedom
## AIC: 318.28
##
## Number of Fisher Scoring iterations: 5
```

```
mod_fit1 <- glm(formula = DEATH_EVENT ~ age+creatinine_phosphokinase+ejection_fraction+serum_
creatinine+serum_sodium,
                family = binomial(link = logit), data = assign_ds)

summary(mod_fit1)
```

```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + creatinine_phosphokinase +
##     ejection_fraction + serum_creatinine + serum_sodium, family = binomial(link = logit),
##     data = assign_ds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3285  -0.7862  -0.4944   0.8743   2.4401
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               4.3687048  4.5131808   0.968    0.333
## age                       0.0534293  0.0125564   4.255 2.09e-05 ***
## creatinine_phosphokinase  0.0002091  0.0001350   1.549    0.121
## ejection_fraction        -0.0668476  0.0144514  -4.626 3.73e-06 ***
## serum_creatinine          0.6292991  0.1599542   3.934 8.35e-05 ***
## serum_sodium             -0.0514336  0.0328688  -1.565    0.118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 375.35  on 298  degrees of freedom
## Residual deviance: 300.71  on 293  degrees of freedom
## AIC: 312.71
##
## Number of Fisher Scoring iterations: 5
```

```
vcov(mod_fit1)
```

```
##                            (Intercept)           age creatinine_phosphokinase
## (Intercept)               2.036880e+01 -7.717235e-03              2.765716e-05
## age                      -7.717235e-03  1.576624e-04              2.355803e-07
## creatinine_phosphokinase  2.765716e-05  2.355803e-07              1.822494e-08
## ejection_fraction         3.904581e-03 -4.174055e-05             -4.326825e-08
## serum_creatinine         -1.199406e-01 -4.060361e-05              1.255567e-07
## serum_sodium             -1.455996e-01 -5.766839e-06             -3.854868e-07
##                           ejection_fraction serum_creatinine  serum_sodium
## (Intercept)                    3.904581e-03    -1.199406e-01 -1.455996e-01
## age                           -4.174055e-05    -4.060361e-05 -5.766839e-06
## creatinine_phosphokinase      -4.326825e-08     1.255567e-07 -3.854868e-07
## ejection_fraction              2.088425e-04    -3.300678e-04 -5.966537e-05
## serum_creatinine              -3.300678e-04     2.558533e-02  7.164803e-04
## serum_sodium                  -5.966537e-05     7.164803e-04  1.080355e-03
```

2. RESIDUAL ANALYSIS

The plot is plotted to check for the appropriateness of the logistic regression model between the dependent and predicted variables. It can be incurred from the standardized Pearsons residuals that there are very less outliers beyond ±3. However, as the target variable is binary few outliers are bound to occur despite of high success of probability. But, In general after having a overall view for the model we can say that it seems to have some decent results.

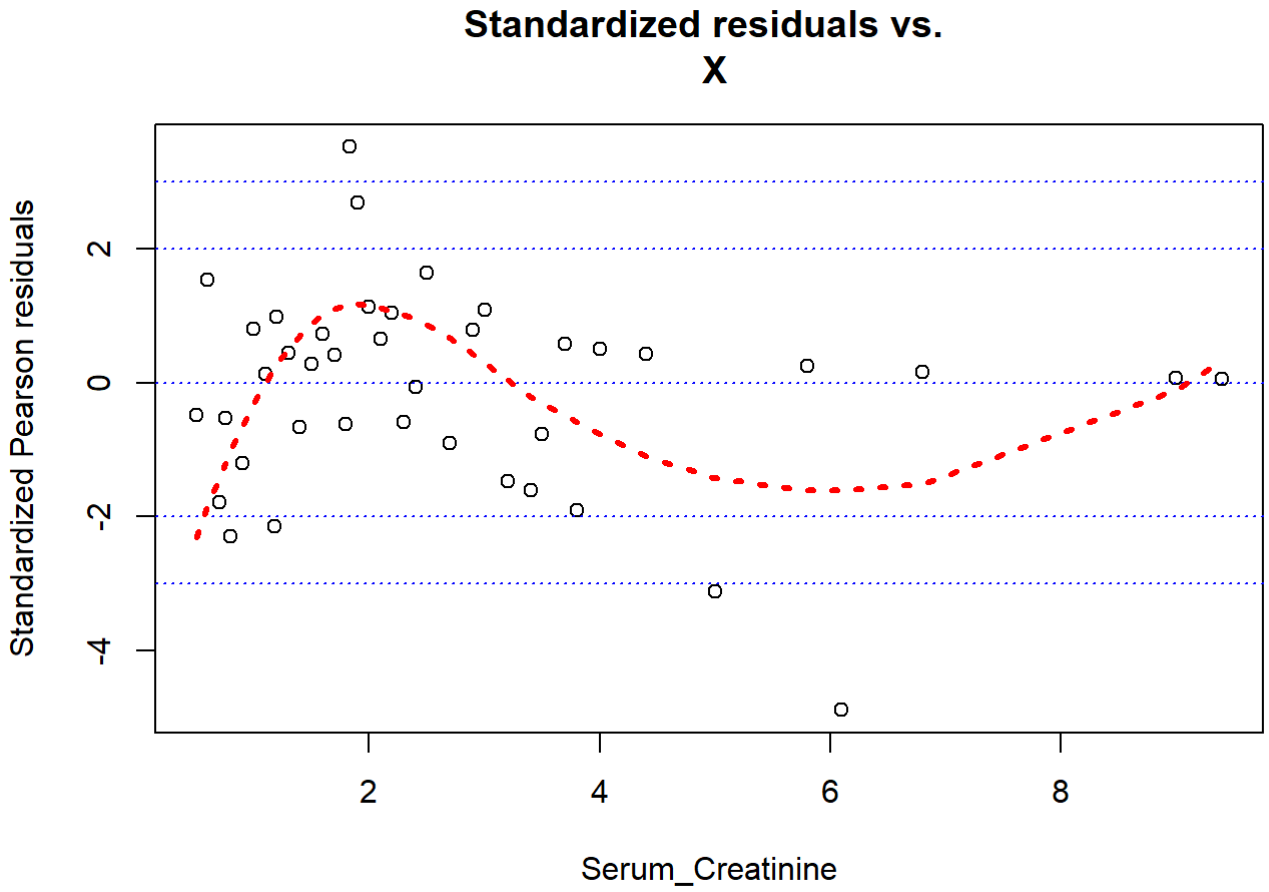
```
####
#Serum_Creatinine
w<-aggregate(formula = DEATH_EVENT ~ serum_creatinine, data = assign_ds,
             FUN = sum)
n<-aggregate(formula = DEATH_EVENT ~ serum_creatinine, data = assign_ds,
             FUN = length)
w.n<-data.frame(Serum_Creatinine = w$serum_creatinine, Death = w$DEATH_EVENT,
                trials = n$DEATH_EVENT, proportion = round(w$DEATH_EVENT/n$DEATH_EVENT,4))
head(w.n)
```

```
##   Serum_Creatinine Death trials proportion
## 1             0.50     0      1     0.0000
## 2             0.60     2      4     0.5000
## 3             0.70     1     19     0.0526
## 4             0.75     0      1     0.0000
## 5             0.80     1     24     0.0417
## 6             0.90     5     32     0.1562
```

```
####
mod_fit_group <- glm(formula = Death/trials ~ Serum_Creatinine,weights = trials,
                     family = binomial(link = logit), data = w.n)
pi.hat <- predict(mod_fit_group, type = "response")
p.res <- residuals(mod_fit_group, type = "pearson")
s.res <- rstandard(mod_fit_group, type = "pearson")
lin.pred <- mod_fit_group$linear.predictors
w.n <- data.frame(w.n, pi.hat,p.res, s.res, lin.pred)

#residuals
#continous
plot(x = w.n$Serum_Creatinine, y = w.n$s.res, xlab = "Serum_Creatinine",
     ylab = "Standardized Pearson residuals", main =
       "Standardized residuals vs. \n X")

abline(h = c(3, 2, 0, -2, -3), lty = 3, col = "blue")
# Add loess model to help visualize trend
smooth.stand <- loess(formula = s.res ~ Serum_Creatinine, data =
                        w.n, weights = trials)
# Make sure that loess estimates are ordered by "X" for
#the plots, so that they are displayed properly
order.SC <- order(w.n$Serum_Creatinine)
lines(x = w.n$Serum_Creatinine[order.SC], y =
        predict(smooth.stand)[order.SC], lty = 3, col =
        "red", lwd = 3)
```



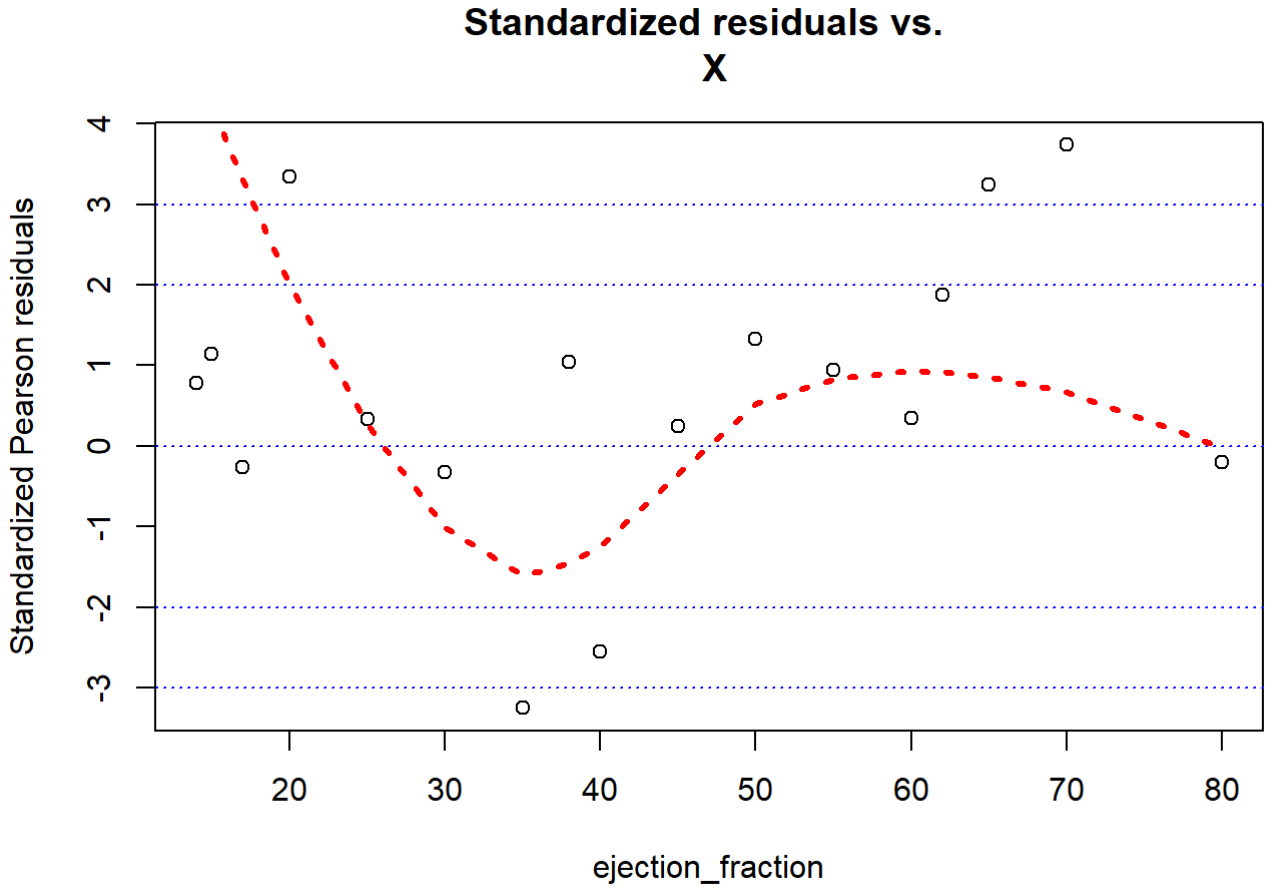

```
#ejection_fraction
w1<-aggregate(formula = DEATH_EVENT ~ ejection_fraction, data = assign_ds,
              FUN = sum)
n1<-aggregate(formula = DEATH_EVENT ~ ejection_fraction, data = assign_ds,
              FUN = length)
w.n1<-data.frame(ejection_fraction = w1$ejection_fraction, Death = w1$DEATH_EVENT,
                 trials = n1$DEATH_EVENT, proportion = round(w1$DEATH_EVENT/n1$DEATH_EVENT,4
))
head(w.n1)
```

```
##   ejection_fraction Death trials proportion
## 1                14     1      1     1.0000
## 2                15     2      2     1.0000
## 3                17     1      2     0.5000
## 4                20    16     18     0.8889
## 5                25    18     36     0.5000
## 6                30    13     34     0.3824
```

```
####
mod_fit_group1 <- glm(formula = Death/trials ~ ejection_fraction,weights = trials,
                      family = binomial(link = logit), data = w.n1)
pi.hat <- predict(mod_fit_group1, type = "response")
p.res <- residuals(mod_fit_group1, type = "pearson")
s.res <- rstandard(mod_fit_group1, type = "pearson")
lin.pred <- mod_fit_group1$linear.predictors
w.n1 <- data.frame(w.n1, pi.hat,p.res, s.res, lin.pred)

#residuals
#continous
plot(x = w.n1$ejection_fraction, y = w.n1$s.res, xlab = "ejection_fraction",
     ylab = "Standardized Pearson residuals", main =
       "Standardized residuals vs. \n X")

abline(h = c(3, 2, 0, -2, -3), lty = 3, col = "blue")
# Add loess model to help visualize trend
smooth.stand <- loess(formula = s.res ~ ejection_fraction, data =
                        w.n1, weights = trials)
# Make sure that loess estimates are ordered by "X" for
#the plots, so that they are displayed properly
order.EF <- order(w.n1$ejection_fraction)
lines(x = w.n1$ejection_fraction[order.EF], y =
        predict(smooth.stand)[order.EF], lty = 3, col =
        "red", lwd = 3)
```



3. RESPONSE ANALYSIS

After plotting the graphs it can be examined that Ejection Fraction and Creatinine Phosphokinase seems to be mostly significant as compared to the serum sodium.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

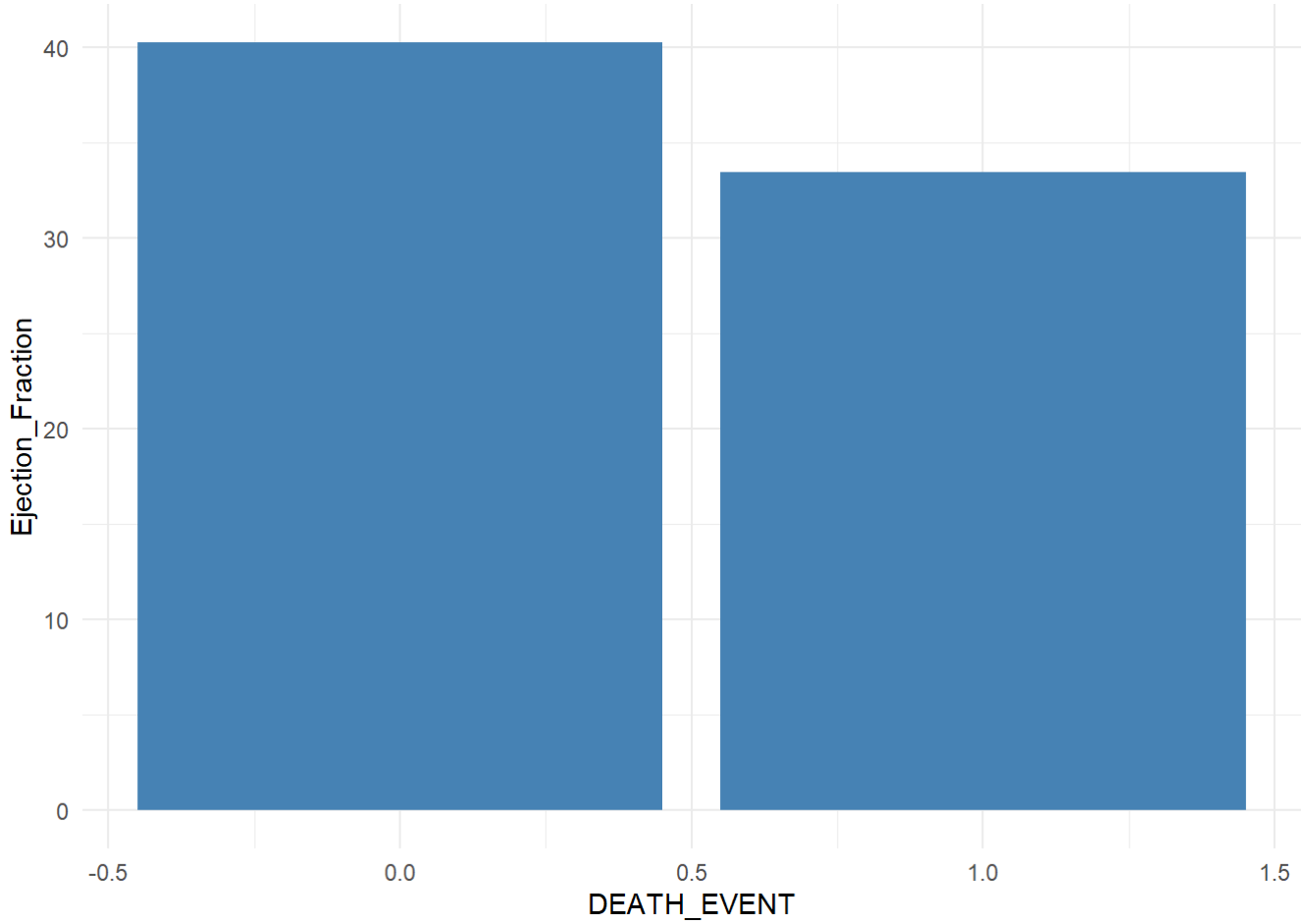
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#Response Analysis
EF <- assign_ds %>% group_by(DEATH_EVENT) %>% summarise(Ejection_Fraction = mean(ejection_fra
ction))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

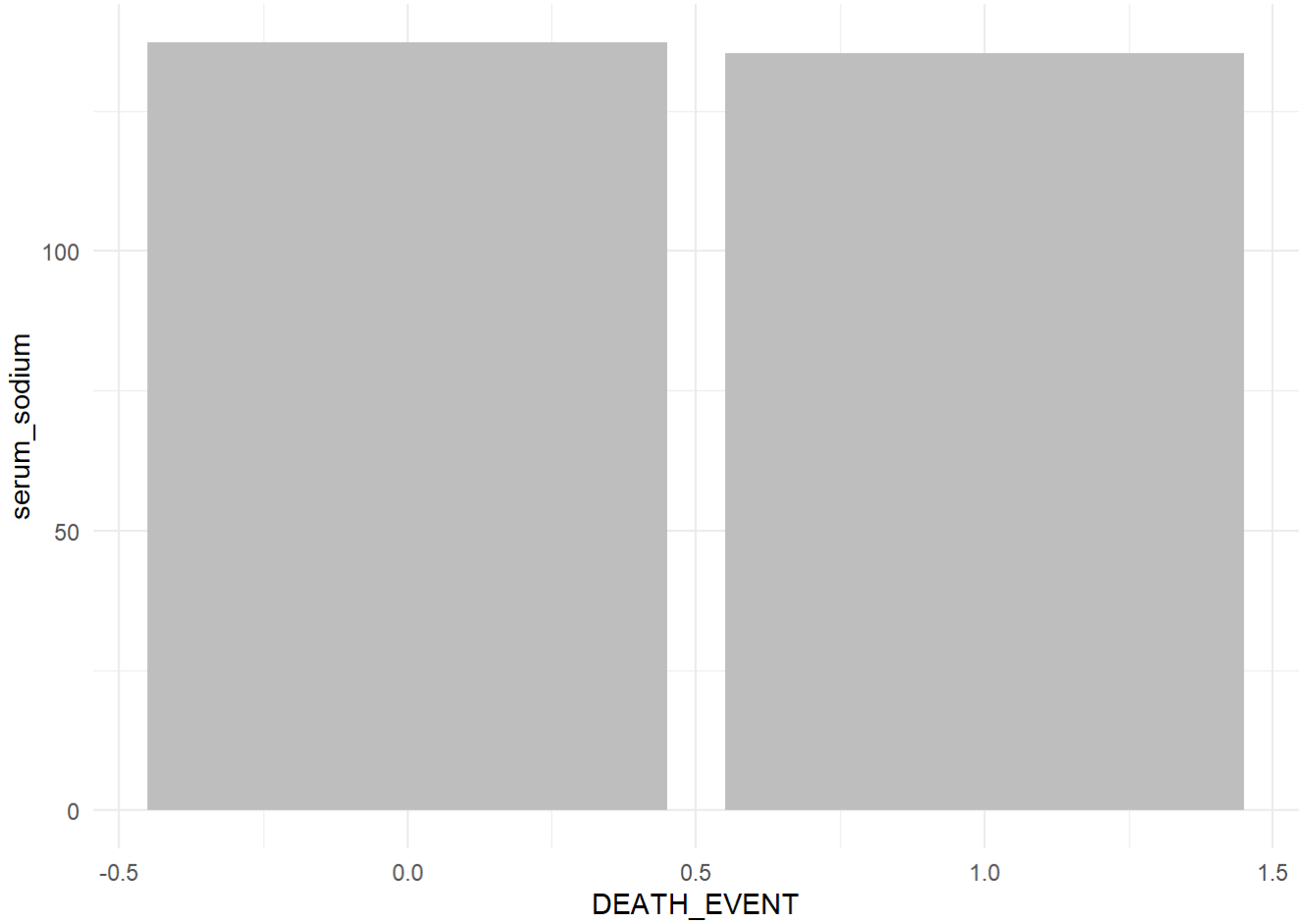
```
p1 <- ggplot(data=EF, aes(x=DEATH_EVENT, y=Ejection_Fraction)) +
  geom_bar(stat="identity", fill = 'Steel Blue')+
  theme_minimal()
p1
```



```
SS <- assign_ds %>% group_by(DEATH_EVENT) %>% summarise(serum_sodium = mean(serum_sodium))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

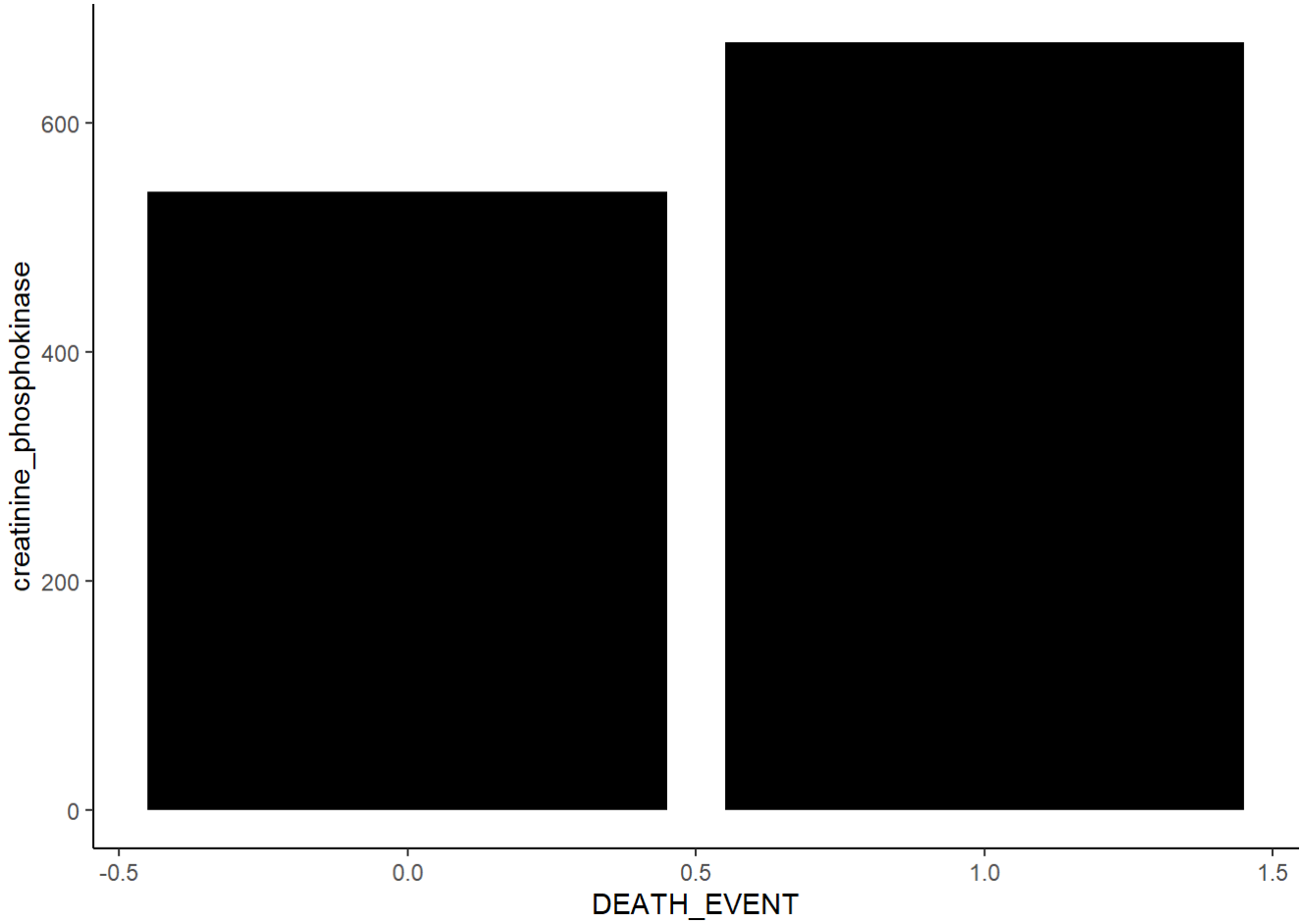
```
p2 <- ggplot(data=SS, aes(x=DEATH_EVENT, y=serum_sodium)) +
  geom_bar(stat="identity", fill = 'Grey')+
  theme_minimal()
p2
```




```
CP <- assign_ds %>% group_by(DEATH_EVENT) %>% summarise(creatinine_phosphokinase = mean(creat
inine_phosphokinase))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
p3 <- ggplot(data=CP, aes(x=DEATH_EVENT, y=creatinine_phosphokinase)) +
  geom_bar(stat="identity", fill = 'black')+
  theme_classic()
p3
```



4. GOODNESS OF FIT

The goodness of fit test is performed using “Hoslem and Lemeshow goodness of fit” to check whether the sample data fits a distribution from a population with a normal distribution. We used this test as it is most efficient way of testing the goodness of fit for the model. It can be observed that we get p-value as 0.2711 which seems to be moderatley good fit.However, good fit would have a p-value around 1.

```
library("ResourceSelection")
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(mod_fit1$y,mod_fit1$fitted.values)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  mod_fit1$y, mod_fit1$fitted.values
## X-squared = 9.9144, df = 8, p-value = 0.2711
```

5. CONFIDENCE INTERVAL

CI for all the regression parameters of the newly fitted model is shown below.

```
confint(mod_fit1)
```

```
## Waiting for profiling to be done...
```

```
##                                 2.5 %        97.5 %
## (Intercept)              -4.416289e+00 13.3825771287
## age                       2.938756e-02  0.0787866606
## creatinine_phosphokinase -5.822699e-05  0.0004873637
## ejection_fraction        -9.638203e-02 -0.0395594844
## serum_creatinine          3.380334e-01  0.9814098149
## serum_sodium             -1.172304e-01  0.0123820640
```

6. HYPOTHESIS TEST

   The hypothesis test is performed using ANOVA with Likelihood Ratio test. The hypothesis test that we are going to perform here is for the “diabetes” variable to check whether it is significant or not.

   H0 : β3 = 0
   H1 : β3 ≠ 0

   The p-value for the Chi-square test performed for “diabetes” variable seems to be greater than the significance value of 0.05. Hence, we can say that null hypothesis cannot be rejected as we do not have any significant evidence.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
Anova(mod_fit,test= "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: DEATH_EVENT
##                           LR Chisq Df Pr(>Chisq)
## age                        19.7820  1  8.680e-06 ***
## anaemia                     1.9397  1    0.16371
## creatinine_phosphokinase    4.1776  1    0.04096 *
## diabetes                    0.2592  1    0.61064
## ejection_fraction          26.6816  1  2.399e-07 ***
## high_blood_pressure         1.8703  1    0.17144
## platelets                   0.1940  1    0.65963
## serum_creatinine           19.7575  1  8.792e-06 ***
## serum_sodium                2.9308  1    0.08691 .
## sex                         1.3018  1    0.25389
## smoking                     0.1513  1    0.69728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. SENSITIVITY ANALYSIS

   The ODDs Ratio test is performed to analyze the effect of odds ratio due to effect of unit increase in the

regression parameters.
It can be said that the estimated ODDs of mortality increase by 4.3% for every 5mg/dl increase in serum creatinine.

```
exp(mod_fit1$coefficients[5])
```

```
## serum_creatinine
##         1.876295
```

```
1/exp(5*mod_fit1$coefficients[5])
```

```
## serum_creatinine
##       0.04300257
```

• FINAL MODEL

Logit(Death_Event) = 4.368 + 0.05Age - 0.07Ejection_fraction + 0.63Serum_creatinine - 0.05Serum_sodium

• CRITIQUE AND LIMITATIONS

This report initially explores all the variables available in the dataset along with the significance of the most impactful influencers of the target variable, thoroughly.
However, as in every report there are limitations in this analytical report as well. First, the p-value of the Holsem and Lameshow Goodness of Fit Test used to check the model's fit is calculated to be 0.2711. Though this proves the model is a modest fit as the obtained value is significantly higher than the required p-value of 0.05, the ideal p-value of goodness of fit test would be much closer to 1,if not equal to it.
Second, the significant effect of influencing variables such as Age, Serum Creatinine Levels, Creatinine Phosphokinase and Ejection Fraction were clear on target variable, Event of Death. The significance of the influencing variables on the target variables is noted to decrease when combined with each other. This phenomenon causes difficulties to predict the outcome of the target variable,Event of Death which in turn, hinders the ability of the analysis to estimate the mortality of the heart patients with an history of heart disease. These were the limitations observed in the current analysis in addition to the common limitations of analytical reports such as limited access to data, time constraints and sample size.

• SUMMARY AND CONCLUSIONS

The first phase of the analysis explored in detail, the major factors impacting fatality rate inpatients with an history of heart failure. The principal contributors to the patients' mortality were examined along with the intensity of each of these variables. On studying the results of the first phase, it was evident that the influence of variables such as Age, Serum Creatinine Levels, Creatinine Phosphokinase and Ejection Fraction were significantly higher than that of other variables.
In the second and final phase of the project, detailed statistical procedures involving binomial logistic regression were performed on the dataset. The statistical procedures included Model Fitting, Residual Analysis, Response Analysis, Goodness of Fit Test involving Hosmer and Lemeshow goodness of fit (GOF) test, Confidence Intervals Calculations, Hypothesis Testing involving Analysis of Variance and Sensitivity Analysis with the help of Odds Ratio Comparison.

The first procedure, model fitting revealed that Age, Ejection Fraction and Serum Creatine Levels were the most variables as the p-value is significantly lower than 0.05. While it is noted that Creatinine Phosphokinase Levels were moderately significant as the p-value is slightly lesser than 0.05. The appropriateness of the logistic regression model between the dependant and predicted variables is investigated by analysing the residuals. It is evident from the Standardized Residuals plot that the number of outliers beyond ±3 was very less as the cluster was predominantly in the range of ±2. This indicates that the number deviations from the expected values are minimal. Though the probability of success is discovered to be high, the number of

outliers is expected to be significantly low as the target variable is binary. On further analysis of the responses and the resultant plot, it is learned that Ejection Fraction and Creatinine Phosphokinase Levels are the most significant variables as the difference in the average count of fatality in the case of the two variables is significantly high when compared to Serum Sodium Levels as the difference in the averages were virtually non-existent.

• **REFERENCES**

1] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics And Decision Making, 20(1). doi: 10.1186/s12911-020-1023-5

2] Heart Failure Prediction. (2020). Retrieved 31 October 2020, from https://www.kaggle.com/andrewmvd/heart-failure-clinical-data (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data)