# ProjectFirst

Teemu Sormunen, Abdullah Günay, Nicola Brazzale

11/18/2020

## Contents

# 1 Introduction

In this project we analyse the data from Heart desease dataset [2]. We perform a Bayesian analysis of the data using this sequence of operations: • Overview of analysis problem and of the dataset • Data preprocsessing and visualisation • Prior choice discussion • Models used in our analysis • $\hat{R}$ convergence • HMC specific convergence diagnostics • Effective sample size diagnostic (n_eff or ESS) • Model comparison • Prior sensitivity analysis • Discussion and conclusion

## 1.1 The problem

Cardiovascular Heart Disease (CHD) is the top reason causing 31% of deaths globally. Pakistan is one of the countries where CHD is increasing significantly, and previous studies do not directly apply to Pakistani area due to different diet patterns. [2]

## 1.2 The motivation

With this project we aim to estimate death events and the major risk factors for heart failure with, possibly, high accuracy [2].

## 1.3 Modeling idea

We created 3 models which are then compared based on $\hat{r}$, $n_{eff}$, using the loo package and the classification accuracy. The 1st model is the reduced model and consists in fewer varibles which are selected base on their corrolection with the death event. The 2nd model consists in all varibles except for the varibale "time" as we believe that doens't represent an important factor in the death event scenario. The 3rd model used is a hierarchical model where we treated age class patients in a group with respect to the other selected variables. The 4th model is a non-linear model, as the hierarchical one we took in considertion only the selected varibales in the first model. Modeling is done with package brms, which is a interface for non-linear multivariate multilevel models in Stan.

# 2 Dataset

## 2.1 Term explanation

Some of the terms in the dataset might not be familiar, and they are opened briefly here.

- **Creatine phosphokinase (CPK)**
  CPK is an enzyme, which helps to regulate the concentration of adenosine triphosphate (ATP) in cells. ATP is responsible for carrying energy. If the CPK level is high, it often means that there has been an injury or stress on a muscle tissue. Although CPK is one the oldest markers of heart attack, high CPK might also indicate of acute muscle injury along with acute heart problems.
  Normal level of CPK ranges from 20 to 200 IU/L [5]
- **Ejection fraction (EF)**
  EF is a measurement in percentage which describes how much blood left ventricle pumps out of heart with each contraction. Low EF might indicate potential heart issues.
  Normal EF is 50 to 70 percent, while measurement under 40 percept might be an indicator of heart failure or cardiomyopathy. [1]
- **Platelets**
  Platelets are small cell fragments which can form clots. Too many platelets can lead to clotting of blood vessels, which in turn can lead to heart attack. Too Normal range of platelets is from 150 000 to 450 000. [4]
- **Serum creatinine**
  When creatine breaks down, it forms a waste product called creatinine. Kidneys normally remove

creatinine from body. Serum creatinine measures level of creatinine in the blood, indicating the kidney health. High levels of creatinine might indicate a kidney dysfunctioning.
Normal level of creatinine range from 0.9 to 1.3 mg/dL in men and 0.6 to 1.1 mg/dL in women who are 18 to 60 years old. [6]

- **Serum sodium**
  Serum sodium measures the amount of sodium in blood. Sodium enters blood through food and drink, and leaves by urine, stool and sweat. Too much sodium can cause blood pressure, while too little sodium can cause nausea, vomiting, exhaustion or dizziness.
  Normal levels of serum sodium are 135 to 145 mEq/L, according to Mayo Clinic. There are however different interpretations of "normal".[3]

## 2.2 Dataset introduction

The dataset of 299 patients was produced as a result of study [2] from Pakistani's city Faisalabad. All of the patients were over 40 years old, each having ventricular systolic dysfunction. This means that patient has poor left ventricular ejection fraction. The dataset has 105 women, and 194 men. EF, serum creatinine and platelets are categorical variables, and age, serum sodium and CPK are continuous variables.
Statistical analysis by [2] found age, creatinine, sodium, anemia and BP as significant variables.

# 3 Packages

Load data

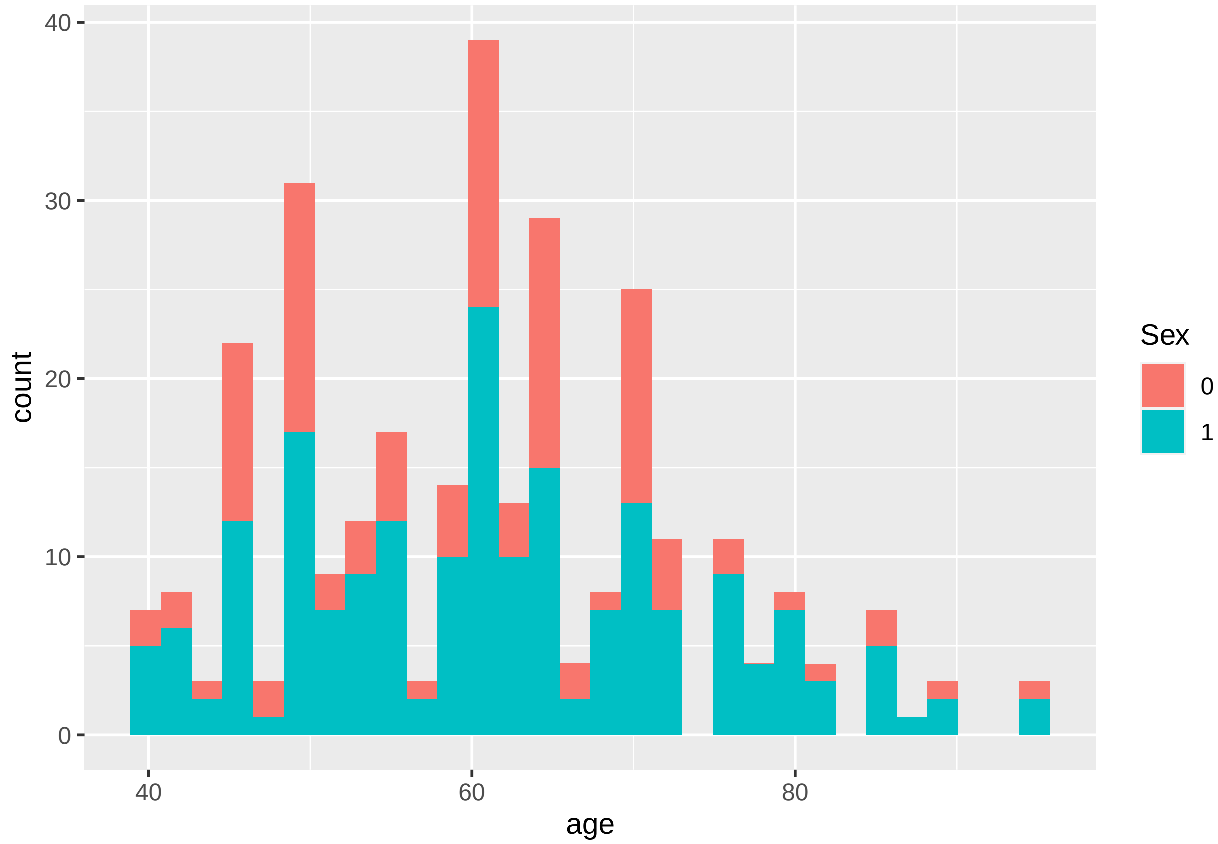
```
file.name <- './data/heart_failure_clinical_records_dataset.csv'
heart <- read_csv(file.name)
```

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   anaemia = col_double(),
##   creatinine_phosphokinase = col_double(),
##   diabetes = col_double(),
##   ejection_fraction = col_double(),
##   high_blood_pressure = col_double(),
##   platelets = col_double(),
##   serum_creatinine = col_double(),
##   serum_sodium = col_double(),
##   sex = col_double(),
##   smoking = col_double(),
##   time = col_double(),
##   DEATH_EVENT = col_double()
## )
```

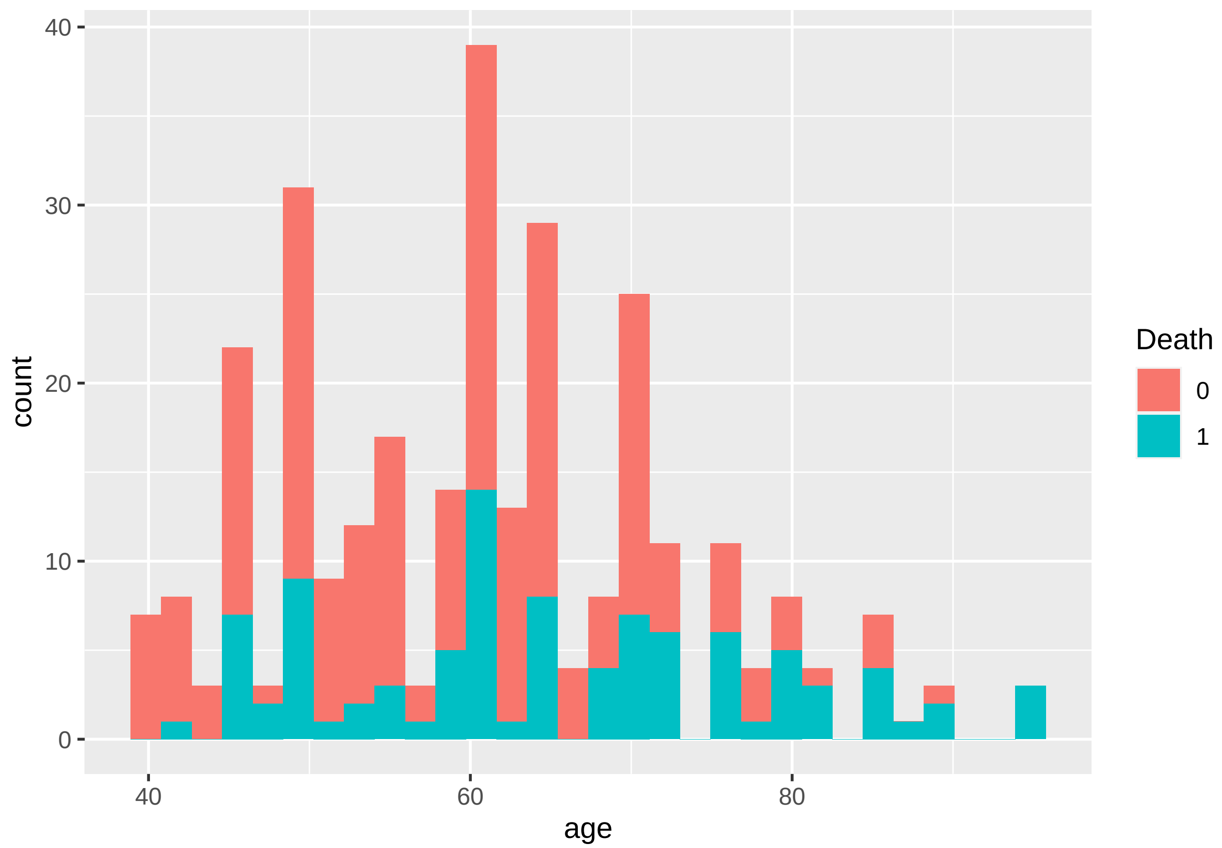
# 4 Data preprocessing and visualization

## 4.1 Plot histograms

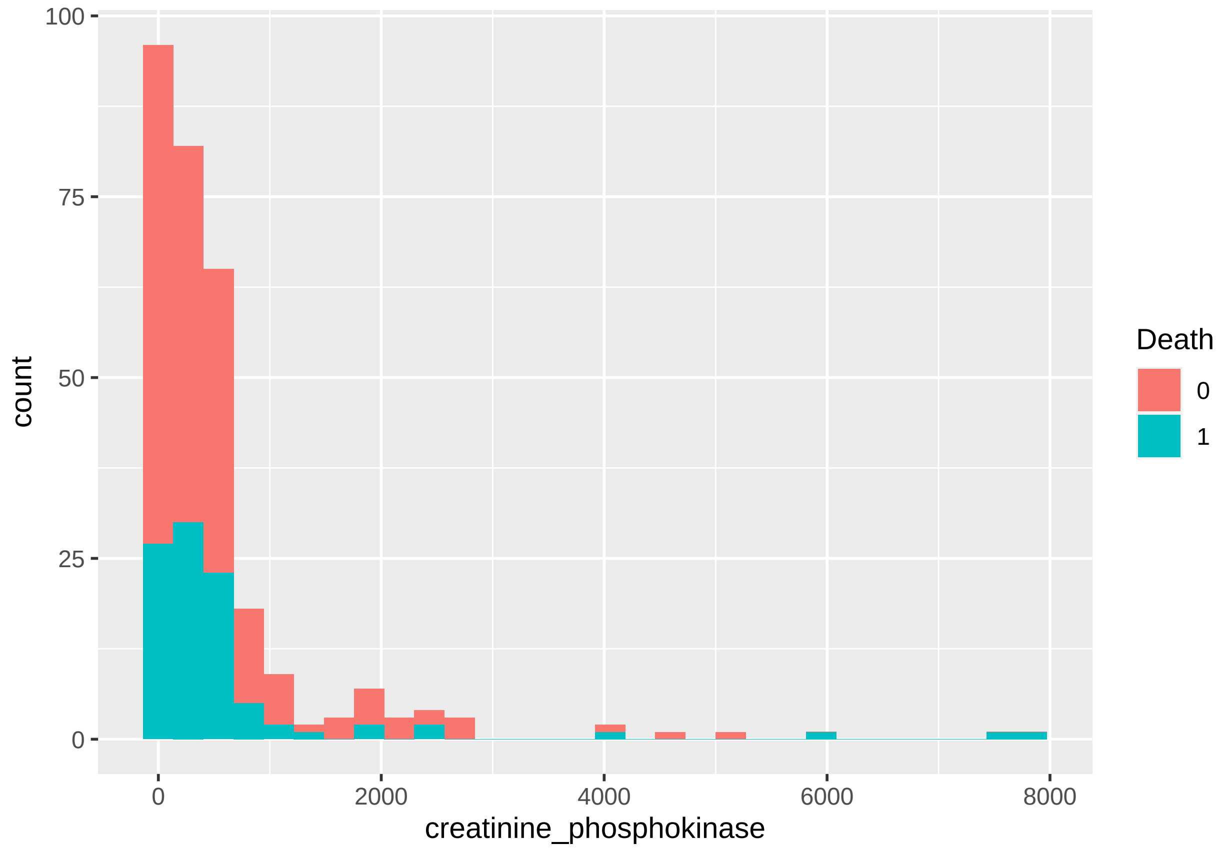
```
ggplot(heart, aes(x=age)) + geom_histogram(aes(fill=as.character(sex)), bins = 30) + labs(fill = "Sex")
```



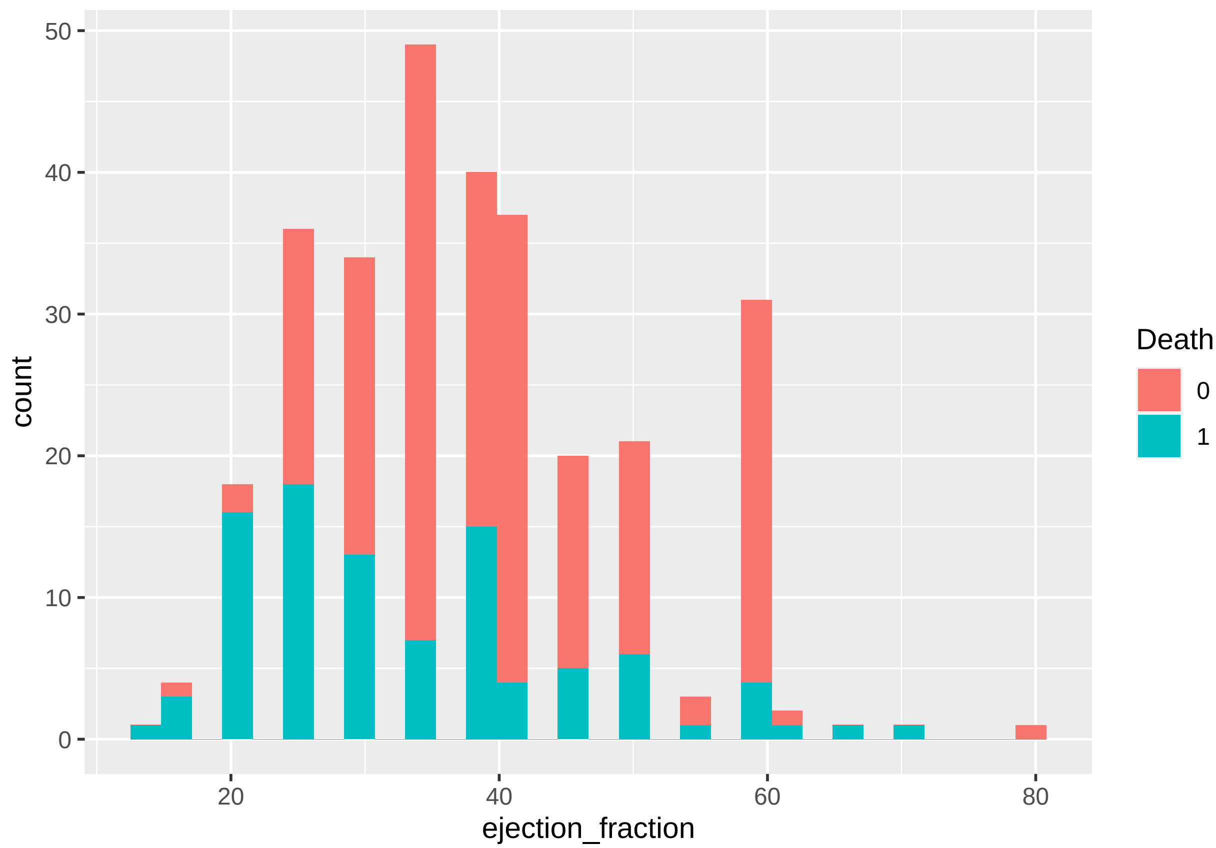

```
ggplot(heart, aes(x=age)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bins = 30) + labs(fill =
```



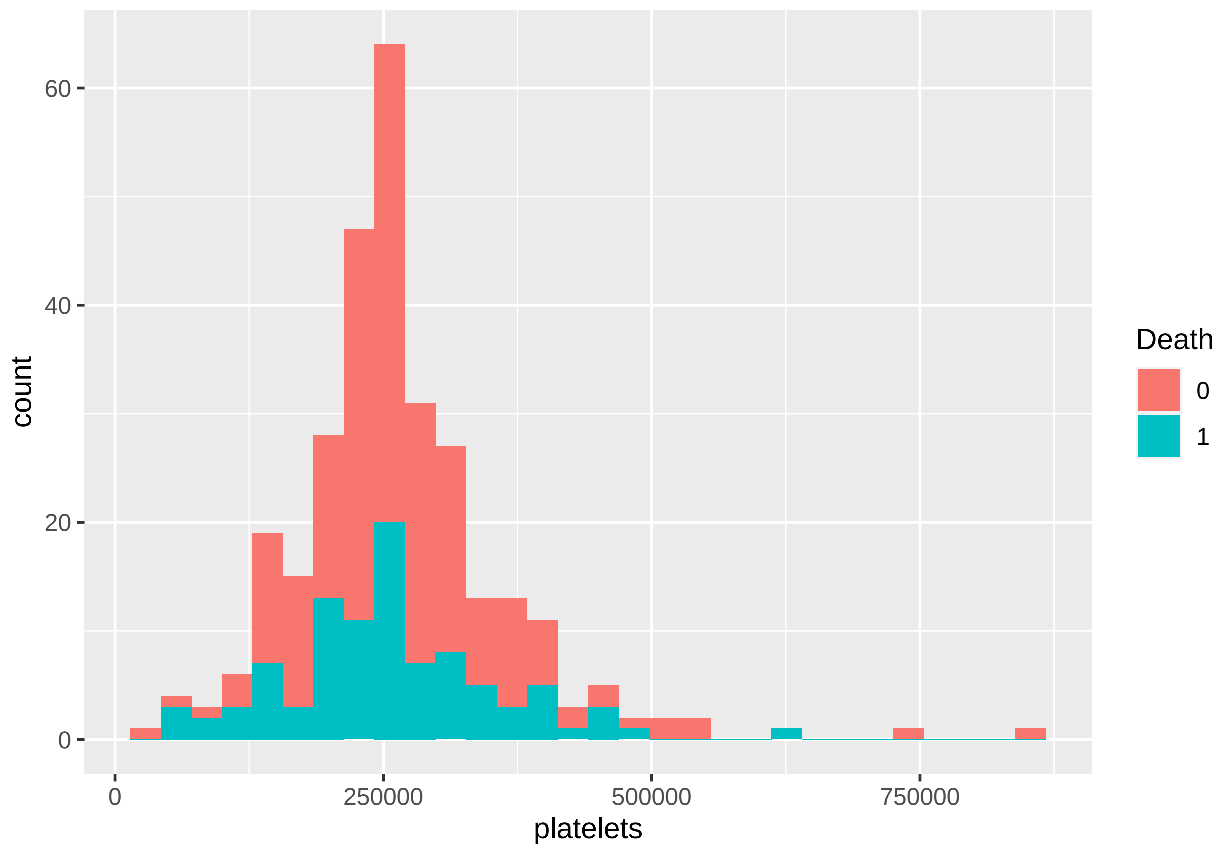

```
ggplot(heart, aes(x=creatinine_phosphokinase)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bi
```



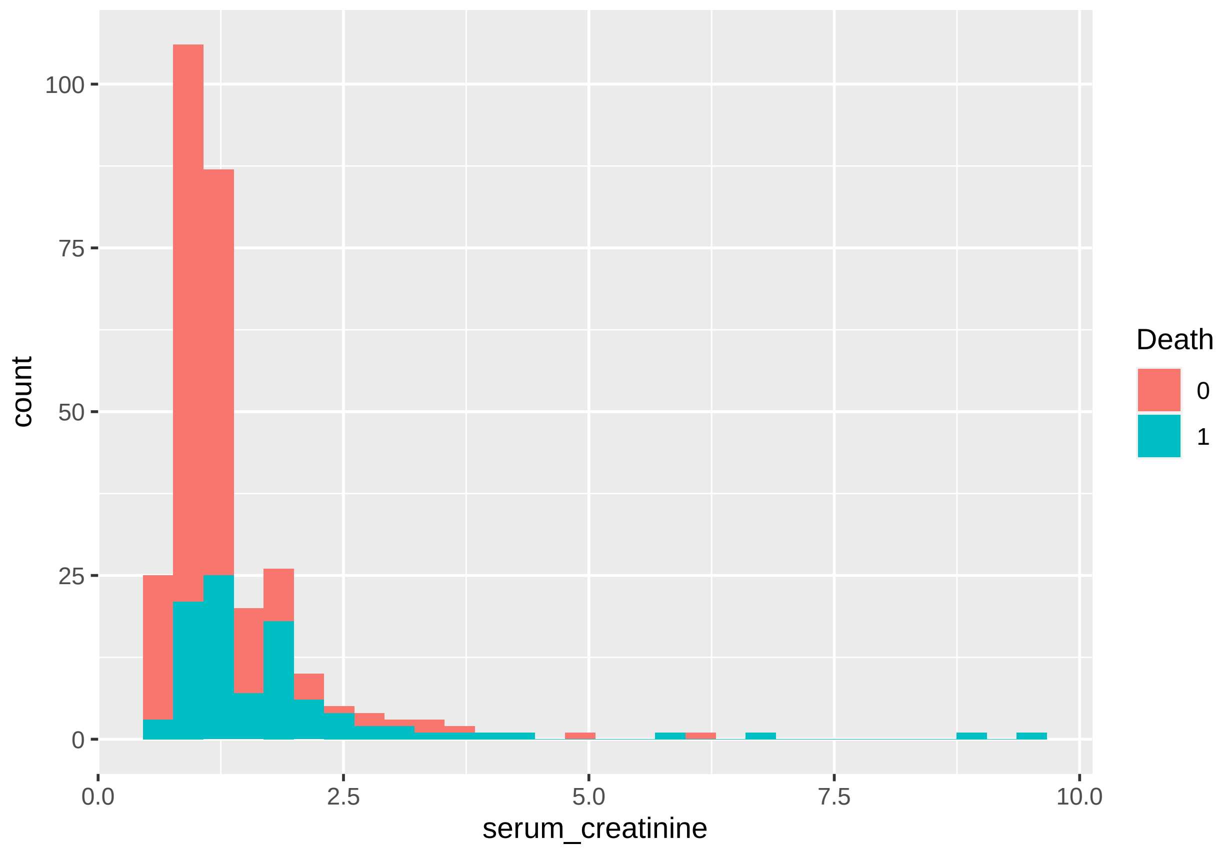

```
ggplot(heart, aes(x=ejection_fraction)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bins = 30)
```




```
ggplot(heart, aes(x=platelets)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bins = 30) + labs
```

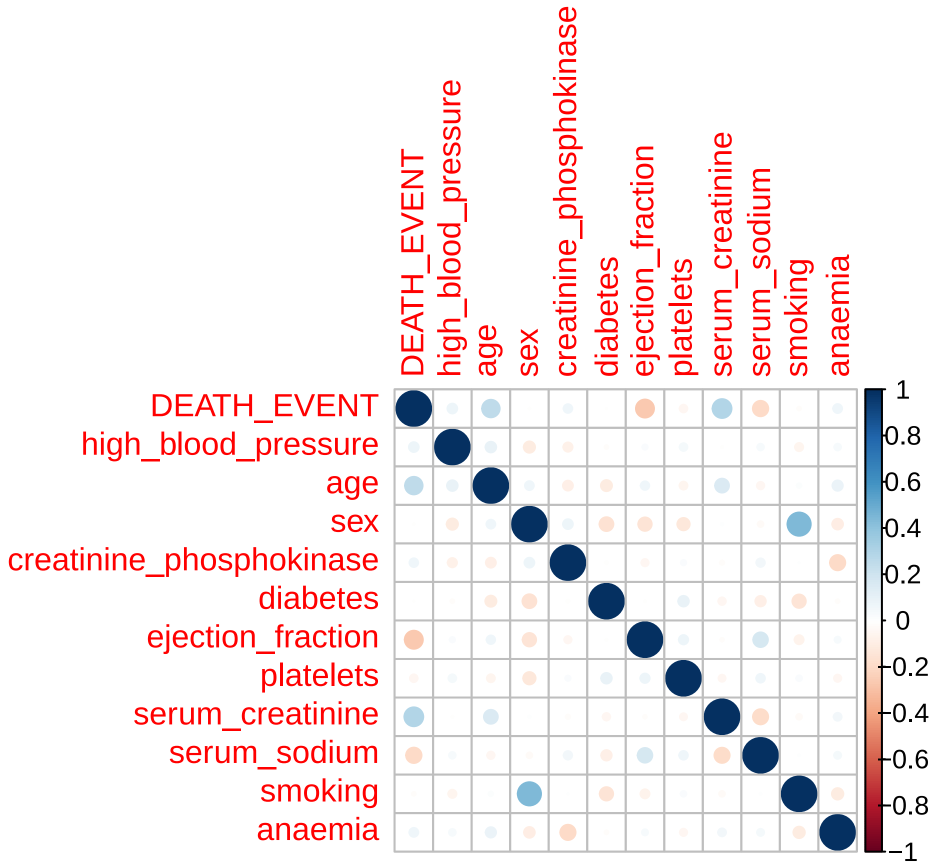



```
ggplot(heart, aes(x=serum_creatinine)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bins = 30)
```




### 4.2 Correlation matrix

```
pred <- c("high_blood_pressure", "age", "sex", "creatinine_phosphokinase", "diabetes", "ejection_fracti
target <- c("DEATH_EVENT")
#formula <- paste("DEATH_EVENT ~", paste(pred, collapse = "+"))
p <- length(pred)
n <- nrow(heart)
x = cor(heart[, c(target,pred)])
corrplot(x)
```




### 4.3 General functions

For testing purposes

```
predict.pointwise.accuracy <- function(fitted.model, test.data) {
  preds <- round(predict(fitted.model, newdata = test.data)[,1])
  preds.correct <- preds == test.data$DEATH_EVENT

  pointwise.accuracy <- length(preds.correct[preds.correct == TRUE])/nrow(test.data)

  return(pointwise.accuracy)
}
```

## 5 Models

We chose to use BRMS for modeling, due to . . .

In BRMS modeling, the parameters are said to either be population level or group level. Population level probably means the same thing as regular parameters in our course, and group level equals hyper parameters (?????????????????)

Brms example, investigate results based on age. **Family argument** specifies the distribution family of the output.

**Prior argument** for each of the parameters, in this case only age. One can set different priors for each population level parameter, or group level parameter.

```r
# Split test and train data
test.size <- 0.3
train.indice <- sample(nrow(heart), nrow(heart)*(1-test.size))
train.data <- heart[train.indice,]
test.data <- heart[-train.indice,]
```

## 5.1 Prior choices

The types of priors we considered in this project are uniformative priors and regularizing priors. Surely this dataset is not the only one about heart failure but since no prior knowledge are provided in the original papers we opted for using uninformative priors.

## 5.2 Model fitting

### 5.2.1 WIP: Full model

Full model includes all the parameters that are specified in the dataset.

```r
fit.full <- brm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + serum_sodium + hig
           data = train.data,
           family = bernoulli(),
           refresh=0
)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

### 5.2.2 WIP: Feature selected model

In feature selected model, we hand pick the features that seems to be the most promising with regards to fitting the model. As described above, we saw that ejection_fraction, serum_creatinine, serum_sodium and age were correlating highly to death.

```r
fit.feature_selected <- brm(formula = DEATH_EVENT ~ ejection_fraction + serum_creatinine + serum_sodium
                            data = train.data,
                            family = bernoulli(),
                            refresh = 0)
```

```
## Compiling Stan program...
```

```
## recompiling to avoid crashing R session
```

```
## Start sampling
```

### 5.2.3 WIP: Hierarchical model

In hierarchical model, we choose age and sex as hyperparameters.

First we will discretize age data in to 3 equal depth bins. Data preprocessing - Age Bins

```r
heart$age[heart$age<=55] = 0
heart$age[heart$age>55 & heart$age<70] = 1
heart$age[heart$age>=70] = 2
hist(heart$age)
```


Then we can fit the model

```
fit.hierarchical <- brm(formula = DEATH_EVENT ~ ejection_fraction + serum_creatinine + serum_sodium + (
           data = train.data,
           family = bernoulli(),
           prior = set_prior('normal(0, 1000)'),
           refresh=0,
           control = list(adapt_delta = 0.99),
           save_all_pars=T,
           save_pars=save_pars(all = T)
)
```

```
## Warning: Argument 'save_all_pars' is deprecated. Please use argument 'all' in
## function 'save_pars()' instead.
```

```
## Compiling Stan program...
```

```
## Start sampling
```

## 5.3 $\hat{R}$ convergence

## 5.4 HMC specific convergence diagnostics (divergences, tree depth) with interpretation of the results

## 5.5 Effective sample size diagnostic (n_eff or ESS) and an interpretation of the results

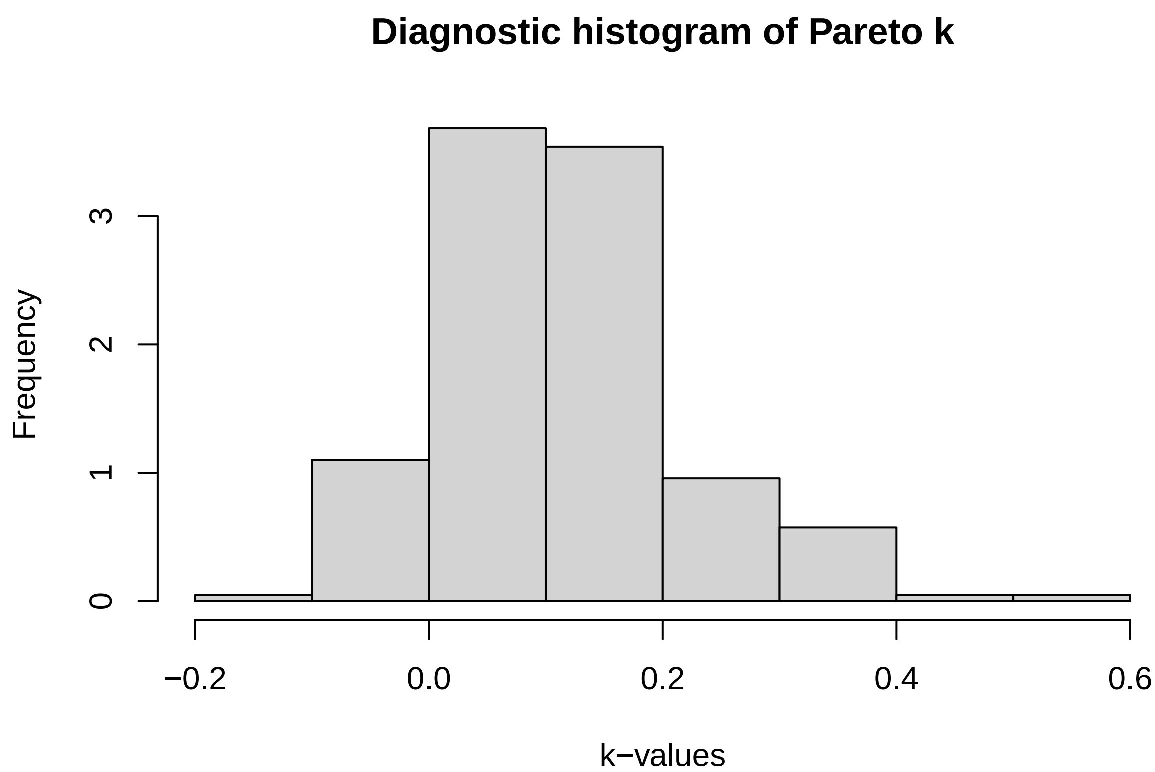
## 5.6 Posterior predictive checking and interpretation of the results

## 5.7 Model comparison and interpretation of the results

## 5.8 WIP: Predictive performance assessment (classification)

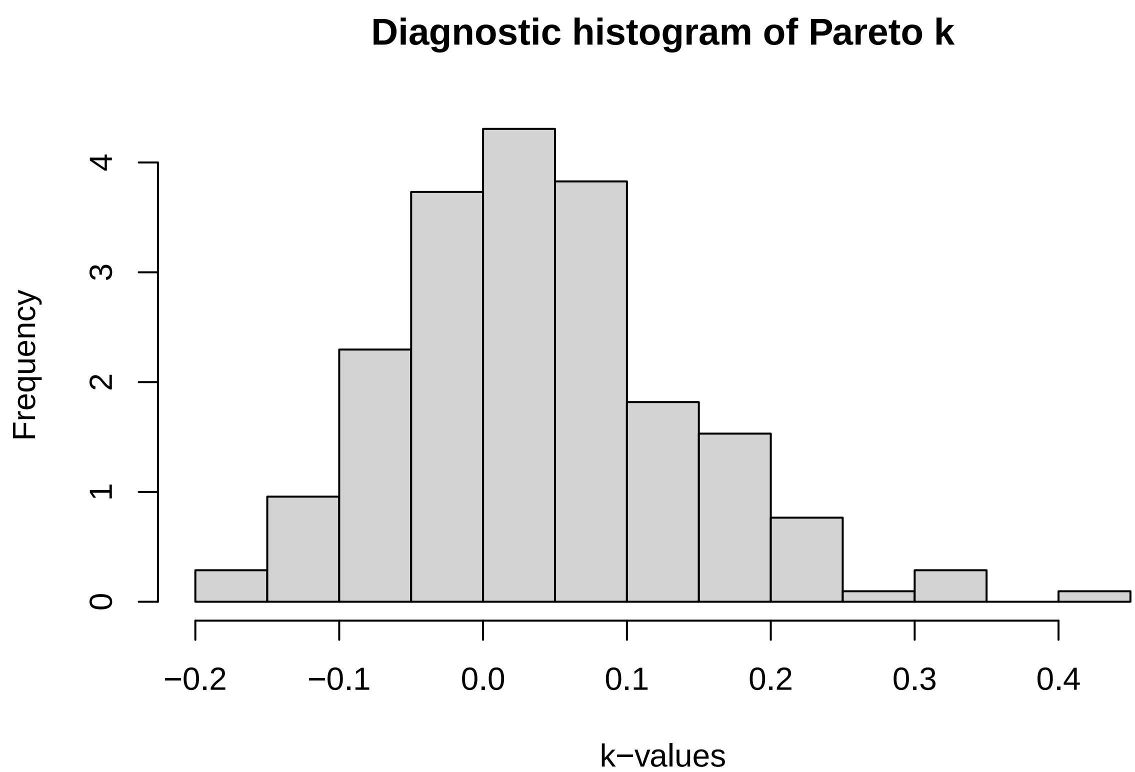
### 5.8.1 WIP: Full model

```
loo.full <-loo(fit.full)

hist(loo.full$diagnostics$pareto_k, main = "Diagnostic histogram of Pareto k",  xlab = "k-values",
     ylab = "Frequency", freq = FALSE)
```




### 5.8.2 WIP: Feature selected model

```
loo.feature_selected <-loo(fit.feature_selected)

hist(loo.feature_selected$diagnostics$pareto_k, main = "Diagnostic histogram of Pareto k",  xlab = "k-va
     ylab = "Frequency", freq = FALSE)
```

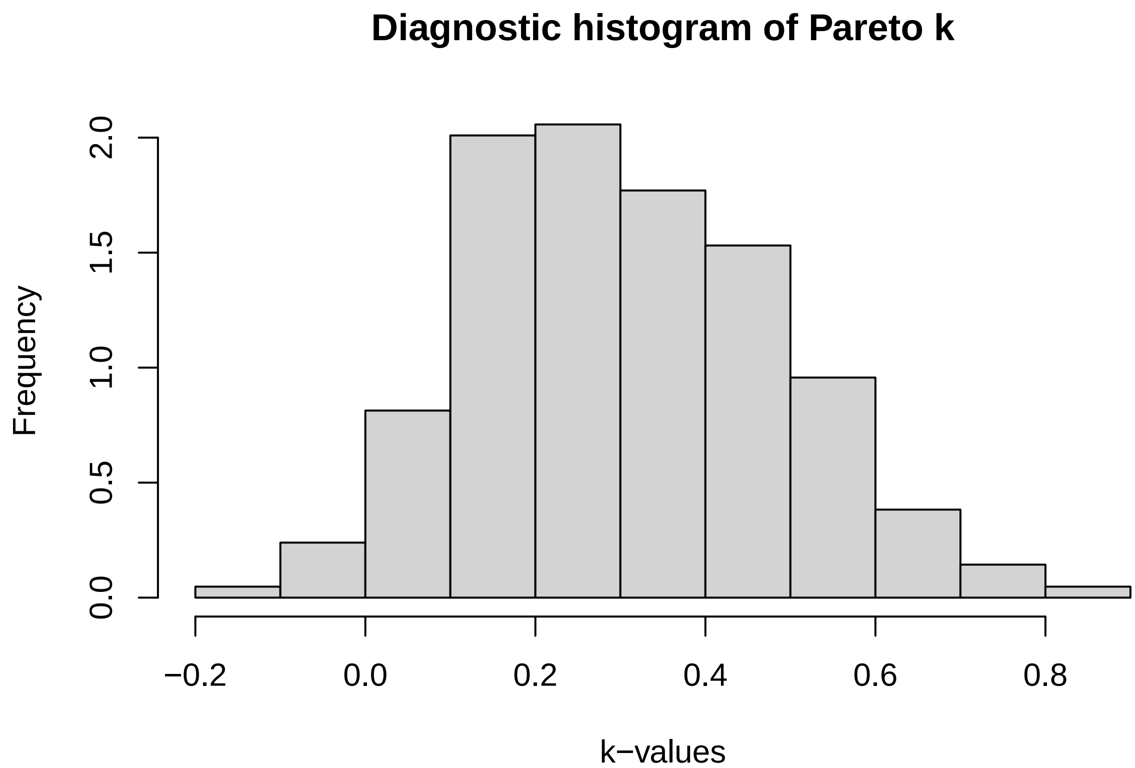



#### 5.8.3 WIP: Hierarchical model

```
loo.hier <- loo(fit.hierarchical)
```

```
## Warning: Found 4 observations with a pareto_k > 0.7 in model 'fit.hierarchical'.
## It is recommended to set 'moment_match = TRUE' in order to perform moment
## matching for problematic observations.
```

```
hist(loo.hier$diagnostics$pareto_k, main = "Diagnostic histogram of Pareto k",  xlab = "k-values",
     ylab = "Frequency",
     freq = FALSE)
```




### 5.9 Prior sensitivity analysis (alternative prior tested)

### 5.10 Discussion of problems and further improvements

NON-LINEAR SCRAP CODE:

```
#NON LINEAR
fitNonLinear <- brm(formula = DEATH_EVENT ~ s(ejection_fraction) + s(serum_creatinine) + s(serum_sodium)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

```
## Warning: There were 42 divergent transitions after warmup. See
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

## 6 Conclusion

## References

[1] Ejection fraction heart failure measurement, 2017.

[2] Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: A case study. 2017. doi: https://doi.org/10.1371/journal.pone.0181001.

[3] Christine Case-Lo. Blood sodium test, 2018. URL https://www.healthline.com/health/sodium-blood.

[4] Gregg D, Goldschmidt-Clermont P. J. Platelets and cardiovascular disease. *Journal of the American Heart Association*, 108, 2003. doi: https://doi.org/10.1161/01.CIR.0000086897.15588.4B.

[5] Roshan Patel Ravinder S. Aujla. Creatine phosphokinase. *StatPearls*, 2020. URL https://www.ncbi.nlm.nih.gov/books/NBK546624/.

[6] Roth Erica. Creatinine blood test, 2019. URL https://www.healthline.com/health/creatinine-blood.