ProjectFirst

Teemu Sormunen, Abdullah Günay, Nicola Brazzale

11/18/2020

Contents

1 Introduction

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

The motivation is to estimate death rates and major risk factors for heart failure to have **predictive power(?????????)**. [2]

1.3 Modeling idea

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

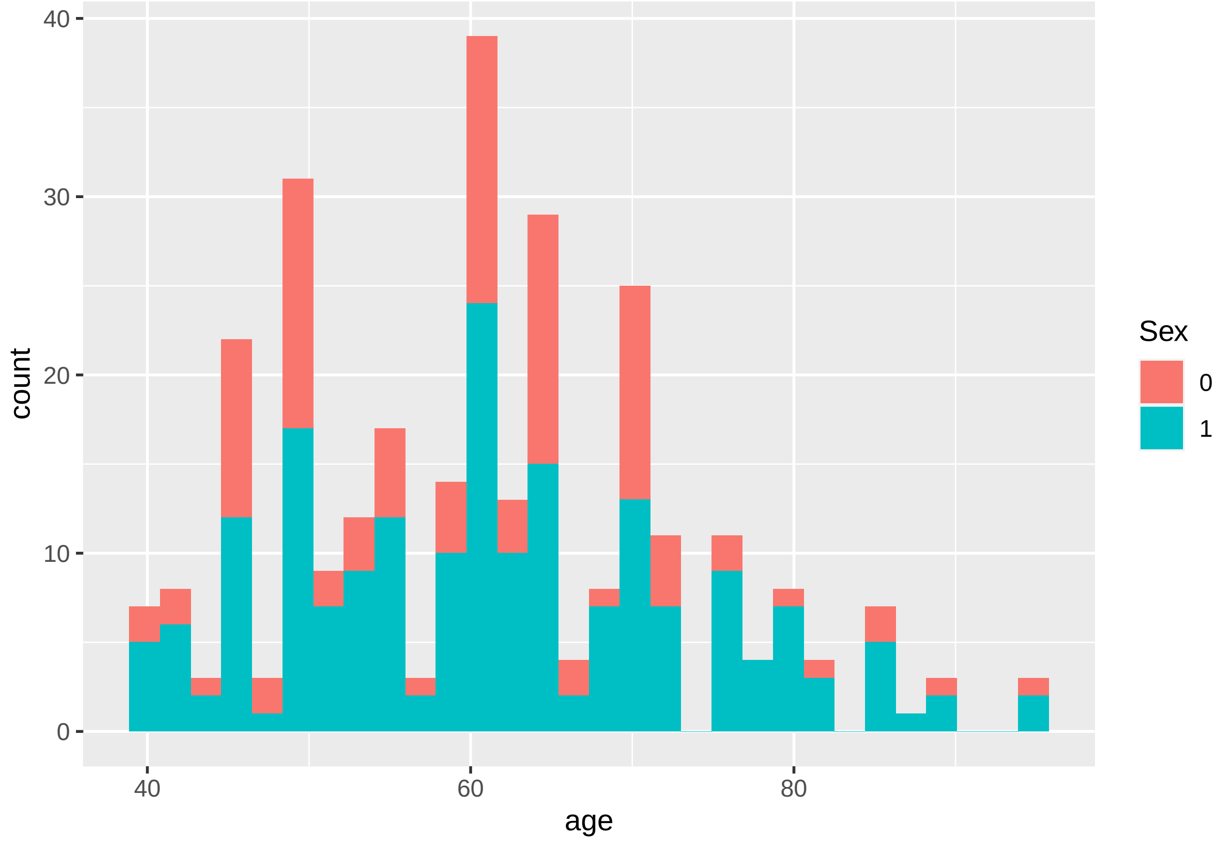
Load data

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

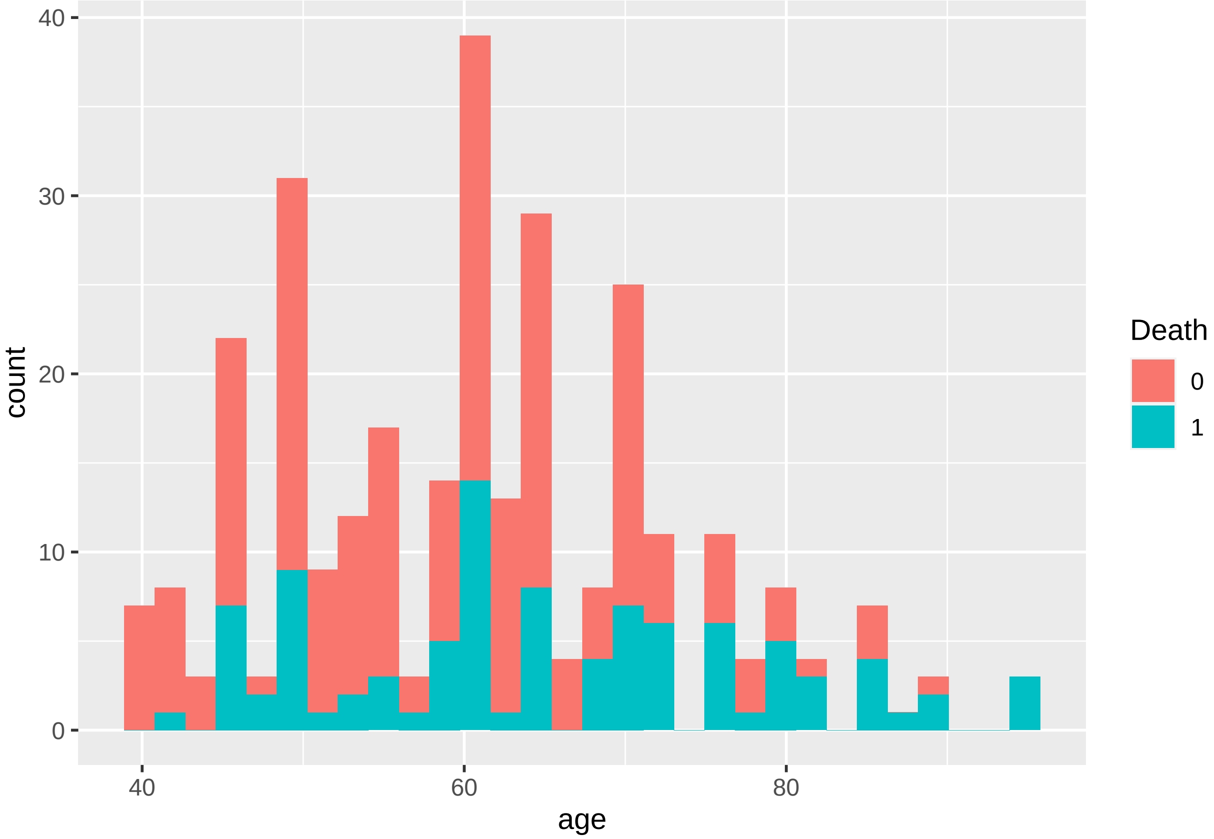
```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   anaemia = col_double(),
##   creatinine_phosphokinase = col_double(),
##   diabetes = col_double(),
##   ejection_fraction = col_double(),
##   high_blood_pressure = col_double(),
##   platelets = col_double(),
##   serum_creatinine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

Plot histograms

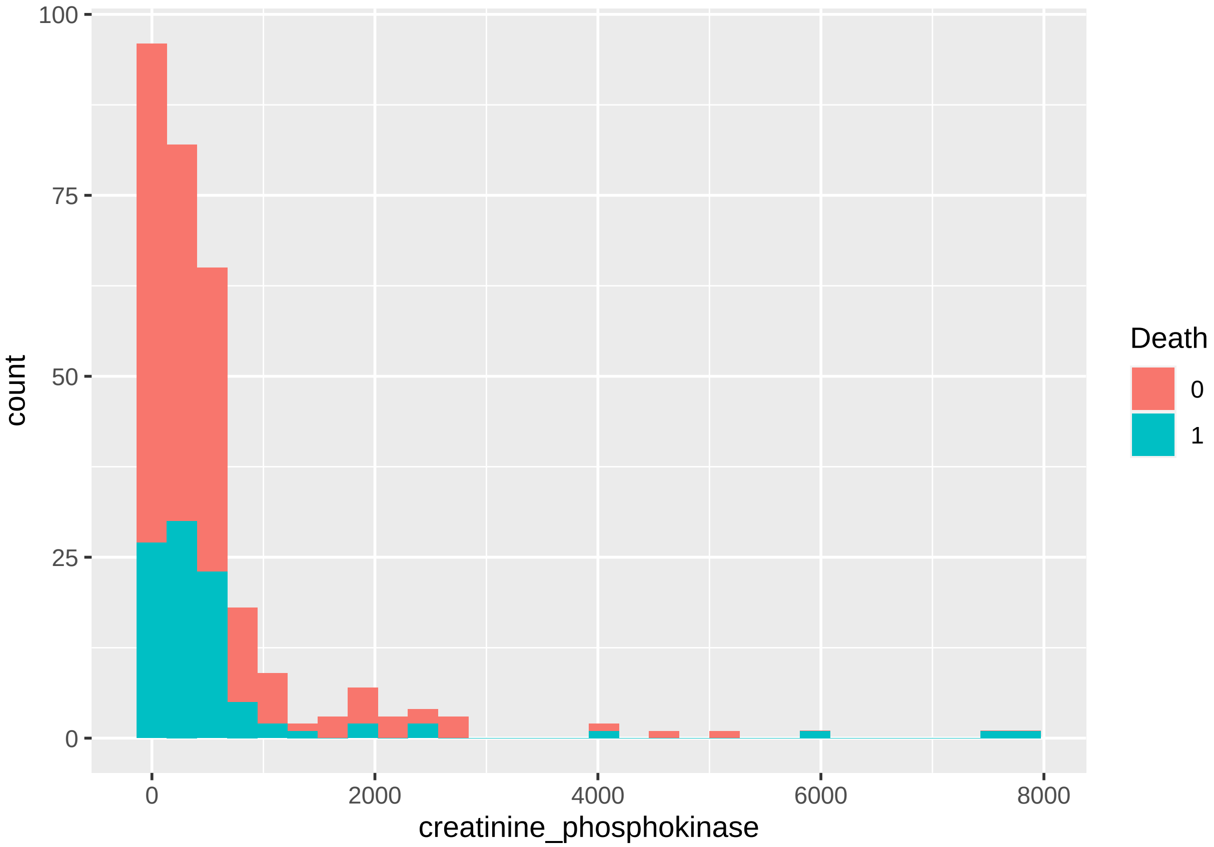
```
ggplot(heart, aes(x=age)) + geom_histogram(aes(fill=as.character(sex)), bins = 30) + labs(fill = "Sex")
```



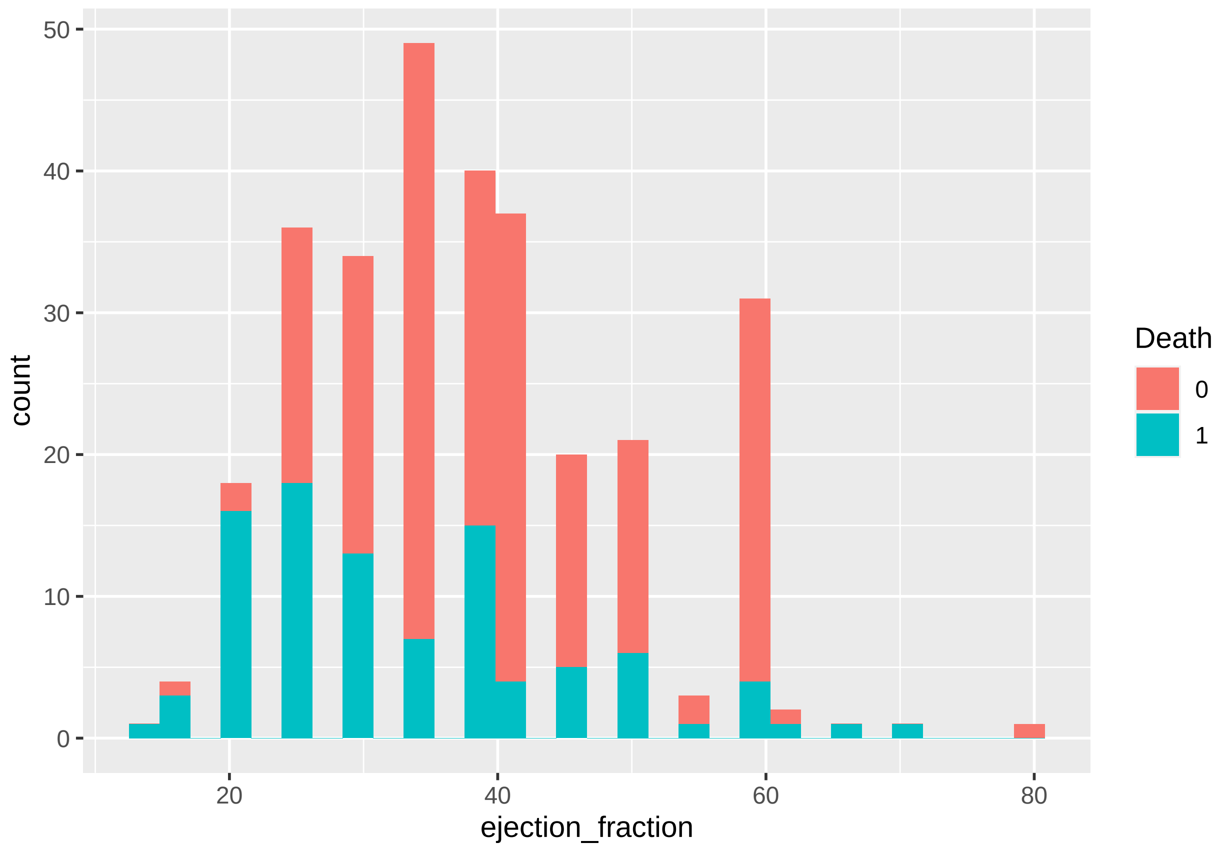
```
ggplot(heart, aes(x=age)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bins = 30) + labs(fill =
```



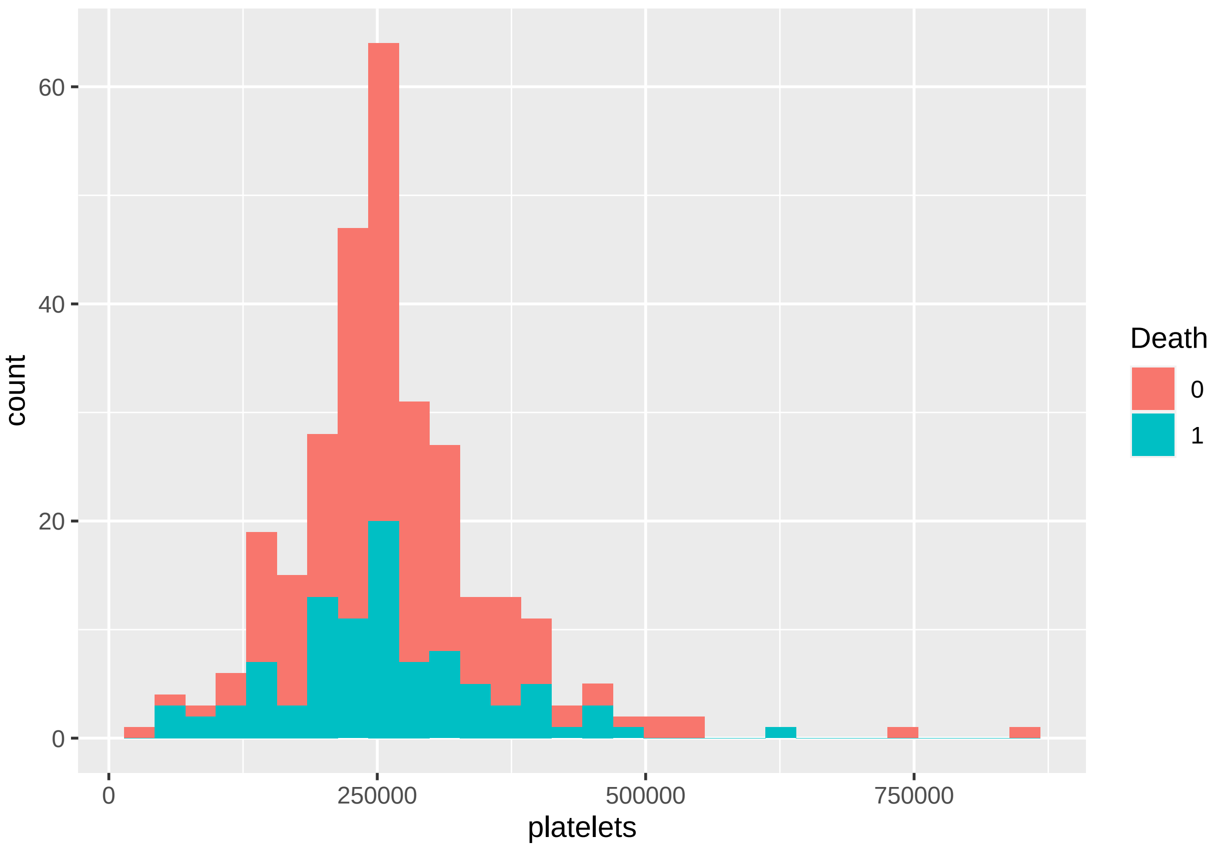
```
ggplot(heart, aes(x=creatinine_phosphokinase)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bi
```



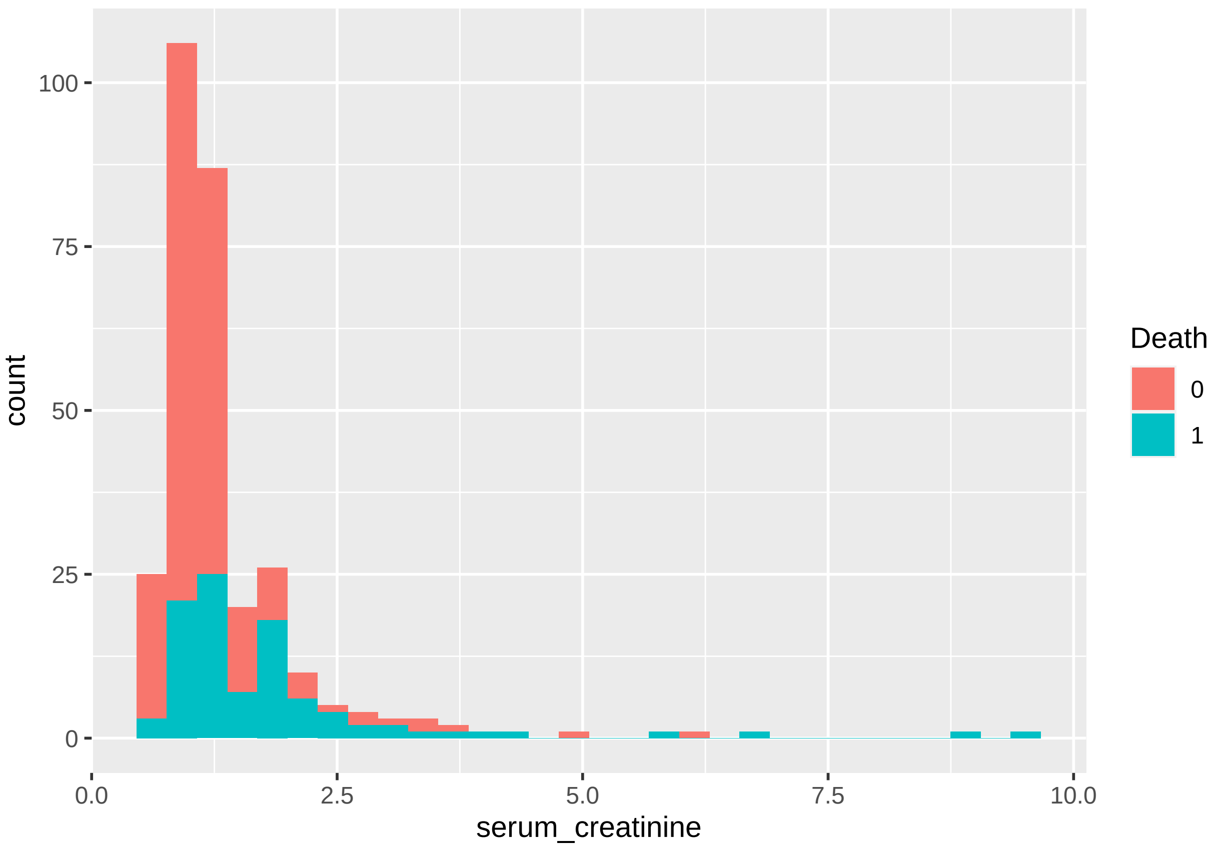
```
ggplot(heart, aes(x=ejection_fraction)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bins = 30)
```



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

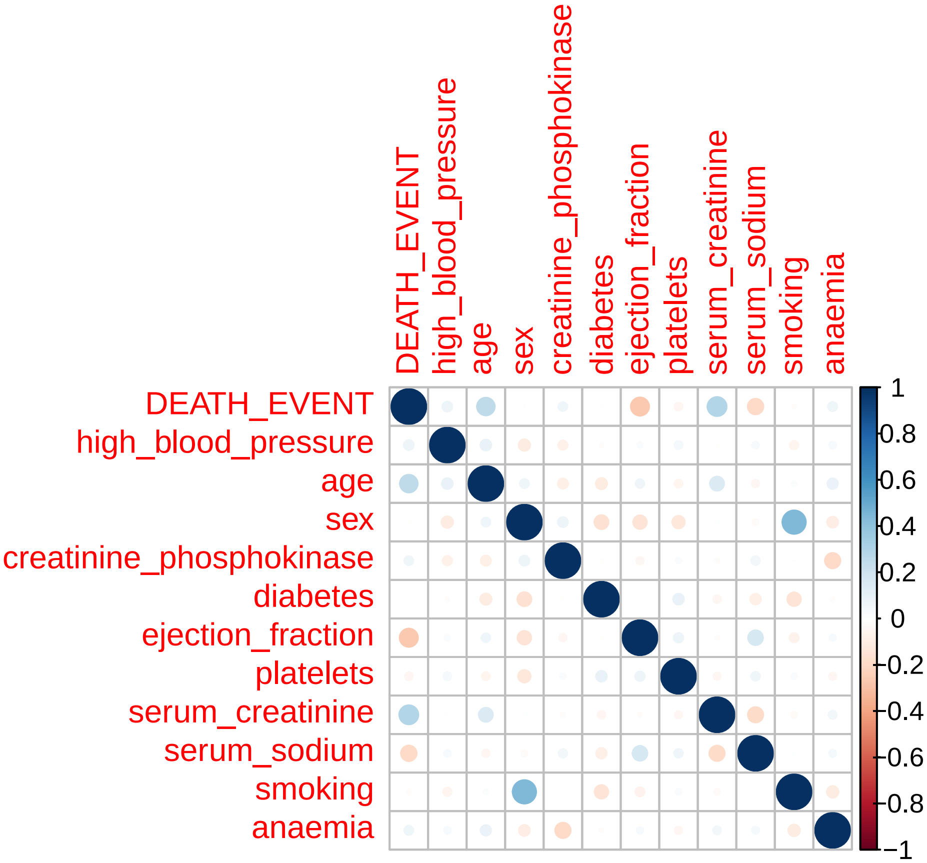


```
ggplot(heart, aes(x=serum_creatinine)) + geom_histogram(aes(fill=as.character(DEATH_EVENT)), bins = 30)
```



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)
```



# BRMS modeling

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.

```
# 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,]

# Fit model based on age
fit <- brm(formula = DEATH_EVENT ~ age,
           data = train.data,
           family = bernoulli(),
           prior = c(set_prior("normal(50,50)", coef="age")),
           refresh=0
           )
```

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

```
## Start sampling
```

Analyze stan code

```
summary(fit)
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: DEATH_EVENT ~ age
##    Data: train.data (Number of observations: 209)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -3.94      0.85    -5.64    -2.29 1.00     3329     2429
## age           0.05      0.01     0.03     0.08 1.00     3527     2547
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
stancode(fit)
```

```
## // generated with brms 2.14.4
## functions {
## }
## data {
##   int<lower=1> N;  // total number of observations
##   int Y[N];  // response variable
##   int<lower=1> K;  // number of population-level effects
##   matrix[N, K] X;  // population-level design matrix
##   int prior_only;  // should the likelihood be ignored?
## }
## transformed data {
##   int Kc = K - 1;
##   matrix[N, Kc] Xc;  // centered version of X without an intercept
##   vector[Kc] means_X;  // column means of X before centering
##   for (i in 2:K) {
##     means_X[i - 1] = mean(X[, i]);
##     Xc[, i - 1] = X[, i] - means_X[i - 1];
##   }
## }
## parameters {
##   vector[Kc] b;  // population-level effects
##   real Intercept;  // temporary intercept for centered predictors
## }
## transformed parameters {
## }
## model {
##   // likelihood including all constants
##   if (!prior_only) {
##     target += bernoulli_logit_glm_lpmf(Y | Xc, Intercept, b);
##   }
##   // priors including all constants
##   target += normal_lpdf(b[1] | 50,50);
##   target += student_t_lpdf(Intercept | 3, 0, 2.5);
## }
## generated quantities {
##   // actual population-level intercept
##   real b_Intercept = Intercept - dot_product(means_X, b);
## }
```

Predict survival

```
preds <- round(predict(fit, newdata = test.data))[1]
pred.corr <- preds == test.data$DEATH_EVENT

acc <- length(pred.corr[pred.corr == TRUE])/nrow(test.data)
acc
```

```
## [1] 0.7333333
```

```
#PRIORS ?
# 1 LINEAR MODEL WITH VARIABLE SELECTION
# 2 LINEAR MODEL WITH ALL VARIABLES
# HIERARCHICAL - in the Titanic one a hier. model have been used so we can use that one for reference

# All models with bernoulli outcome (1-0, death or not)
```

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.