Predicting a Death Event from Ejection Fraction Values in Patients with Cardiovascular Diseases

Lindsey Hornberger

Department of Biostatistics
University of Kansas, USA
May 10, 2024

AIMS

The purpose of the study is to investigate the relationship between a "death event" and ejection fraction for patients at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan) during April–December 2015 while controlling for other potential sources of variation.

ABSTRACT

To investigate the impact of outside factors on the presence of a death event for patients at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan) during April–December 2015, a dataset is considered. The data is from the UC Irvine Machine Learning Repository. There are many different factors to consider when discussing the presence of a death event of a patient with a cardiovascular disease. The objective of this study is to determine if there is an association between the presence of a death event in a patient and their ejection fraction value and 5 different predictor variables (both continuous and binary). In this study we used logistic regression models to determine if an association between variables was present. The best logistic regression model indicates there is a correlation between the logit function and ejection fraction value, age, HBP, sex, log(serum creatinine) and an interaction term of ejection fraction and sex. This model was chosen as the best through automatic model selection and has the lowest AIC value of 308.3.

INTRODUCTION

Cardiovascular diseases are one of the leading causes of death globally taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide.[1] Cardiovascular diseases are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions.[1] Heart failure is a common event that can occur due to cardiovascular diseases. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, an unhealthy diet, obesity, physical inactivity, and harmful use of alcohol using population-wide strategies.[1] Heart failure occurs when the heart is not able to pump enough blood to the body.[1] In a clinical setting, heart failure can be classified into two different groups based on a patient's ejection fraction value or the proportion of blood pumped out of the heart during a single contraction of the heart.[1] The ejection fraction value is given as a percentage with values ranging from 50-70%. The first group is heart failure due to a reduced ejection fraction (HFrEF) also called heart failure due to left ventricular (LV) systolic dysfunction and is classified by an ejection fraction smaller than 40%. The second group is heart failure with preserved ejection fraction (HFpEF) also called diastolic heart failure or heart failure with normal ejection fraction and is classified by the left ventricle contracting normally during systole, but the ventricle is stiff and fails to relax normally during diastole and this impairs filling. Another potential cause of heart failure is high levels of serum creatine in one's bloodstream which can be caused by kidney dysfunction.[1] Serum creatine is a waste product of creatine when a muscle breaks down.[1] Another potential cause of heart failure is high blood pressure (hypertension). In this study, we will consider six potential factors that could lead to a "death event."

MATERIALS AND METHODS

Data Sources

The dataset was obtained online from Kaggle. The data is from the UC Irvine Machine Learning Repository and contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features. The data was collected by Dr. Davide Chicco and Dr. Giuseppe Jurman and was published in their paper "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." From this dataset, we have condensed the predictor variables from thirteen to six to keep our model simple. The six predictor variables are a mix of both categorical and quantitative (3 of each) and the dependent variable is categorical (binary). The dependent variable is categorical, and it is the "**death event**" or if the patient passed away during the follow-up period 0 if they died, 1 if they are still alive). The predictor variables are **age** (quantitative), **high blood pressure (HBP)** (0 if patient does not have HBP, 1 if patient does have HBP), **sex** (0 if female, 1 is male), **ejection fraction** (quantitative), **serum creatinine levels** (quantitative), and **smoking** (0 if patient does not smoke, 1 if patient does smoke).

Statistical Analysis

The data was available in .xlsx (excel) format. The data analysis is done using the statistical software R version 4.3.1 (2023-06-16). This project focuses mainly on categorical multivariate regression. All of the predictor variables are explored individually. In this dataset the sample size is 299 and there are no missing values.

Data Dictionary:

Variable Name	Data Type	Data Format	Description	Example
age	Number	123	Age of patient	75
hbp	Binary (0 or 1)	0,1	If the patient has hypertension (1) or not (0)	1
sex	Binary (0 or 1)	0,1	If patient is male (1) or female (0)	1
v	Number	123	Percentage of blood leaving the heart at each contraction (percentage)	20
serum_creatinine	Number	123	Level of serum creatinine in the blood (mg/dL)	1.90
smoking	Binary (0 or 1)	0,1	If the patient smokes (1) or not (0)	0
DEATH_EVENT	Binary (0 or 1)	0,1	If the patient is deceased during the follow-up period (1) or not (0)	1

The Preliminary Model

$$logit[\pi(x)] = a + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6}$$

A multiple logistic regression model is considered. Let

$logit[\pi(x)]$ = death event, If the patient is deceased during the follow-up period (1) or not (0) (binary)

X_{i1}= serum creatinine, the level of serum creatinine in the blood (mg/dL)

X_{i2}= age, the age of the patient

X_{i3} = HBP, If the patient has high blood pressure (hypertension) (1) or not (0) (binary)

X_{i4}= sex, the sex of the patient is male (1) or female (0) (binary)

X_{i5}= serum creatinine, the level of serum creatinine in the blood (mg/dL)

X_{i6} = smoking, If the patient smokes (1) or not (0) (binary)

i = 1,2,3, · · ·, 299.

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6,$ and σ^2 are the unknown parameters to be estimated.

RESULTS AND DISCUSSION

A. Analysis of Predictor Variables

In this section, we will examine the distribution of predictors, identify any unusually large or small values, and examine bivariate associations to identify multicollinearity between variables. A scatterplot matrix indicates positive linear associations between a death event and age and serum creatinine levels (continuous predictor variables).

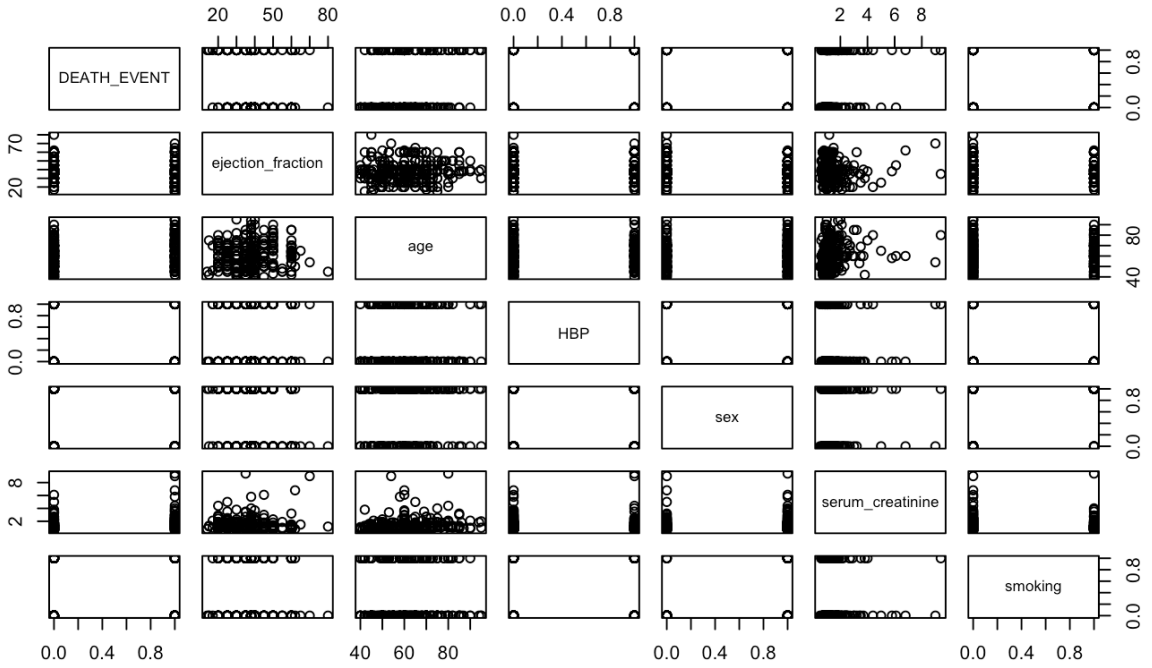


Figure 1: Scatterplot matrix for the binary independent variable and six dependent variables.

	age	ejection_fraction	HBP	serum_creatinine	sex	smoking	DEATH_EVENT
age	1.00000000	0.06009836	0.093288685	0.159187133	0.065429524	0.01866787	0.253728543
ejection_fraction	0.06009836	1.00000000	0.024444731	-0.011302475	-0.148385965	-0.06731457	-0.268603312
HBP	0.09328868	0.02444473	1.000000000	-0.004934525	-0.104614629	-0.05571137	0.079351058
serum_creatinine	0.15918713	-0.01130247	-0.004934525	1.000000000	0.006969778	-0.02741414	0.294277561
sex	0.06542952	-0.14838597	-0.104614629	0.006969778	1.000000000	0.44589171	-0.004316376
smoking	0.01866787	-0.06731457	-0.055711369	-0.027414135	0.445891712	1.00000000	-0.012623153
DEATH_EVENT	0.25372854	-0.26860331	0.079351058	0.294277561	-0.004316376	-0.01262315	1.000000000

Figure 2: Correlation coefficient matrix for the binary independent variable and six dependent variables.

Multicollinearity occurs when two or more predictor variables are highly correlated. Values of r close to ±1 are considered to be highly correlated. The largest Pearson correlation coefficient value is between smoking and sex +0.445891712 which indicates moderate correlation. All other correlation coefficients are below 0.29 which indicate weak correlation. While this value is not too close to ±1, we can check for collinearity by calculating the VIF (variation inflation factor). If VIF exceeds 10, multicollinearity is present.

ejection_fraction	age	HBP	sex	serum_creatinine	smoking
8.810010	6.647187	6.134143	7.671497	8.765550	7.481240

Figure 3: VIF values for all predictor variables.

Due to the variance inflation factors (VIFs) of the response variables, each value is below 10. Based on the maximum VIF value of 8.810010 for ejection fraction, there do not appear to be any issues that need remediation.

Strip plots and **box plots** for each continuous variable were made to determine if the data is skewed.

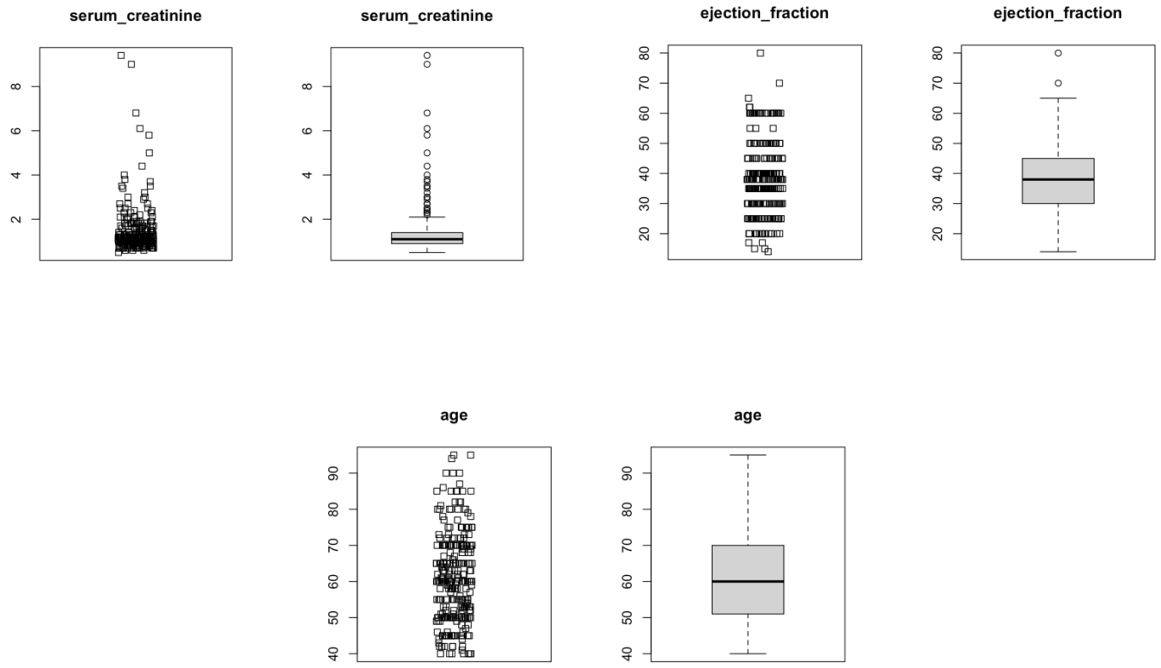


Figure 4(a,b,c): Box plots and strip plots for the three continuous variables: serum creatinine, ejection fraction, and age.

For the three continuous variables: ejection fraction, age, and serum creatinine, box plots and strip plots were made. The variables age and ejection fraction appear to be evenly distributed with ejection fraction having a couple data points outside of the boxplot standard deviation but is not skewed enough to consider performing a log transformation. However, the serum creatinine data is positively skewed with most observations clustered together and a few observations at much higher levels. Due to this, we could consider a log transformation.

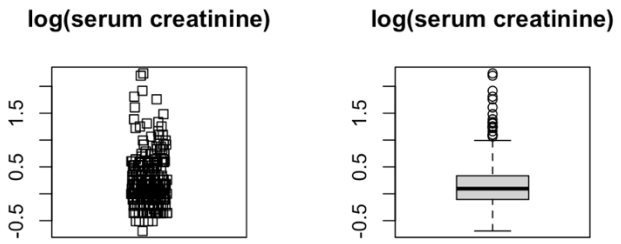


Figure 5: Strip plot of log(serum creatinine).

The log transformation of the predictor variable serum creatinine was taken, and the data does appear to be more evenly distributed with less data points outside of the standard deviation.

For the three continuous predictor variables, a curve of the **generalized additive model** smoothing fit and the logistic regression models are plotted.

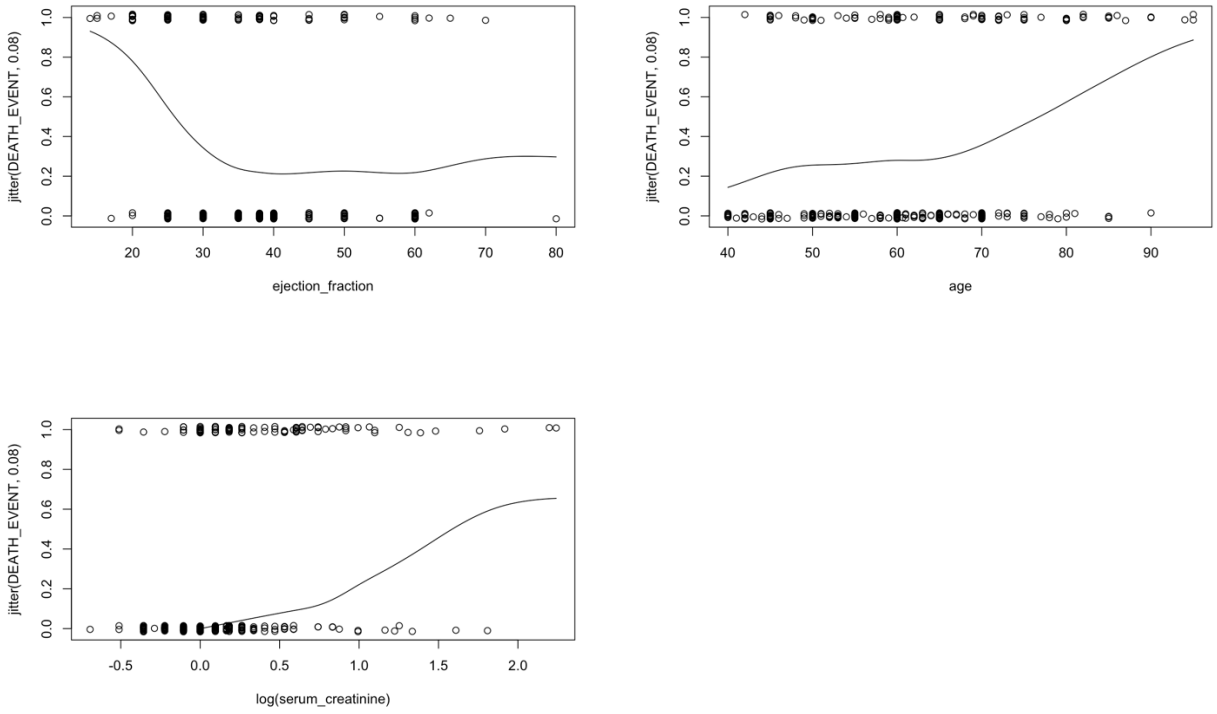


Figure 6 (a,b,c): A generalized additive model smoothing fit was performed for each continuous variable (ejection fraction, age, and log(serum creatinine)).

From these generalized additive model smoothing fit plots, we can see that as ejection fraction increases, the logit(π). As age increases, the logit(π) of the likelihood of a death event decreases. As serum creatinine levels increase, the logit(π) of the likelihood of a death even also increases.

Partial residual plots for each of the covariates are displayed. Partial residual plots attempt to show the relationship between a given independent variable and the response variable given that other independent variables are also in the model. They also display the nature of the relationship between the covariate and the outcome (i.e., linear, curvilinear, transformation necessary, etc.) and any problematic data points with respect to the predictor. It is important to note that a log transformation of serum creatinine was done prior due to the significant skewedness data points within the variable.

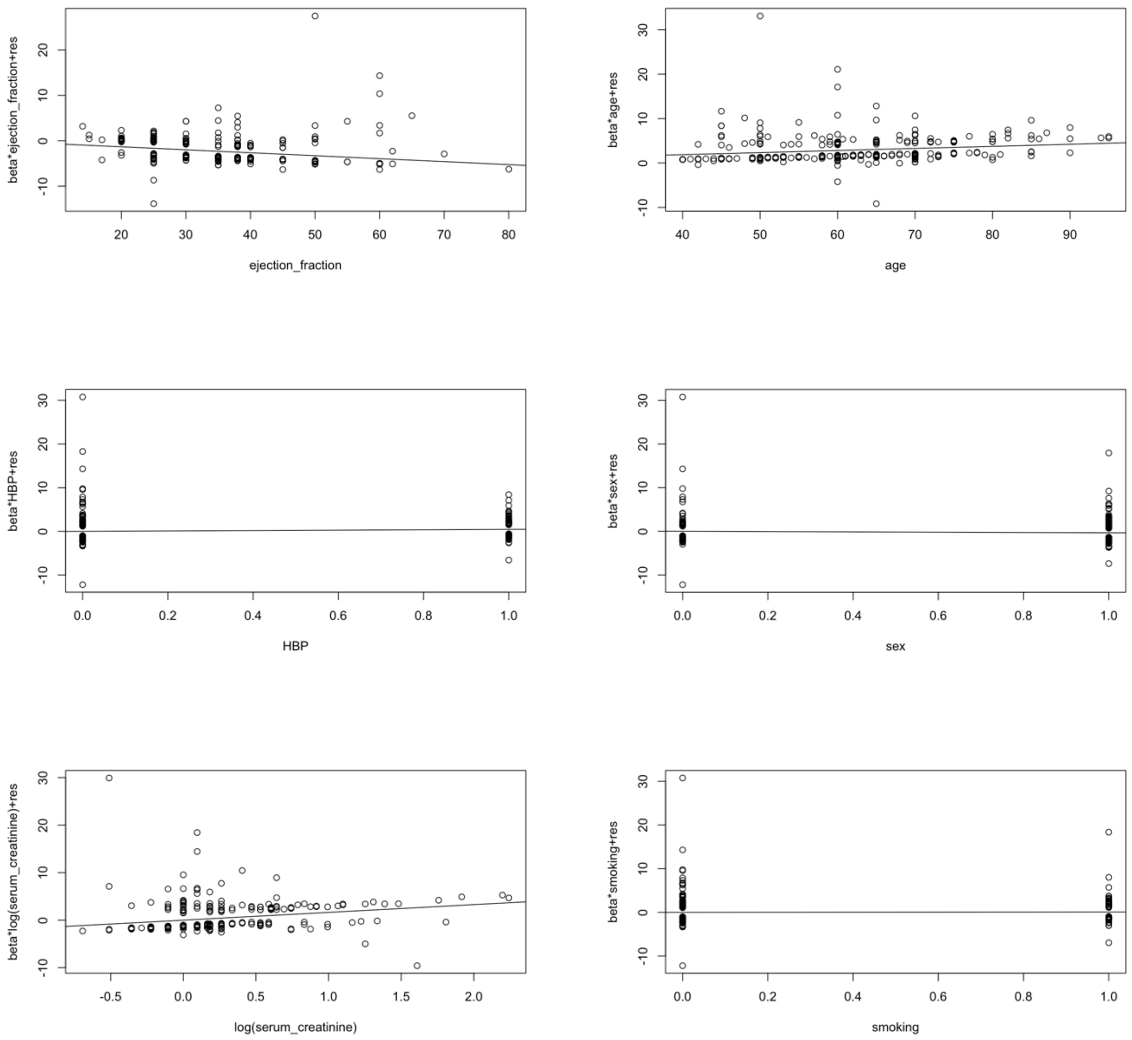


Figure 7(a,b,c,d,e,f): Partial Residual plots for all predictor variables.

The partial residual plots for each predictor variable appear to be evenly distributed. These plots also indicate that each predictor variable provides some added value to a model that already includes all other covariates because the slopes of the linear relationships are all appear to be non-zero.

B. Residual Diagnostics

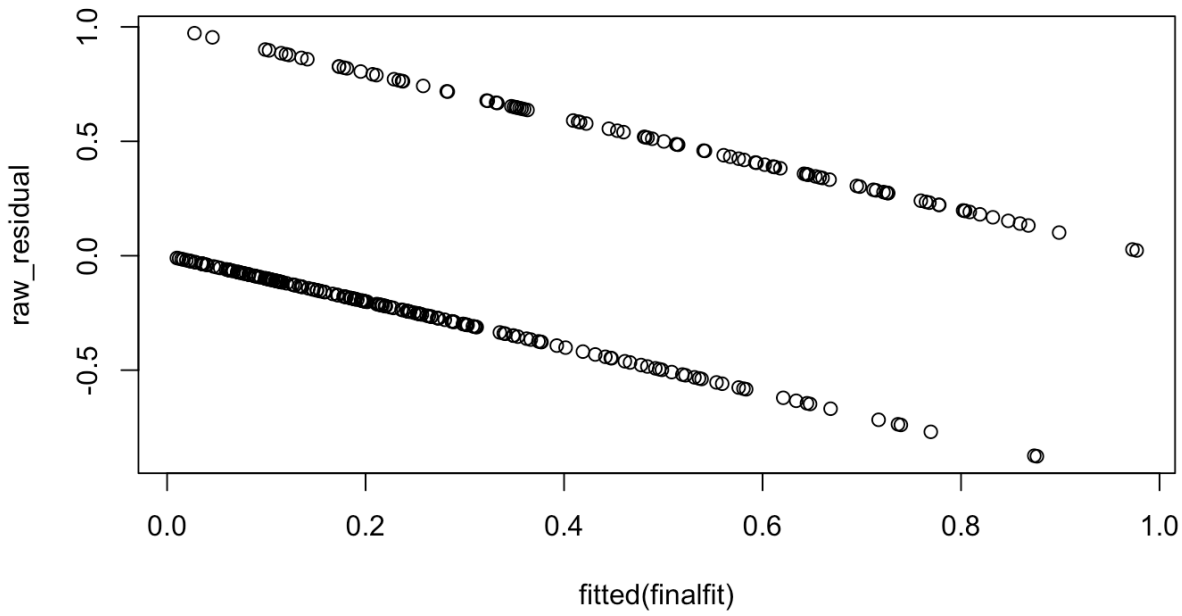
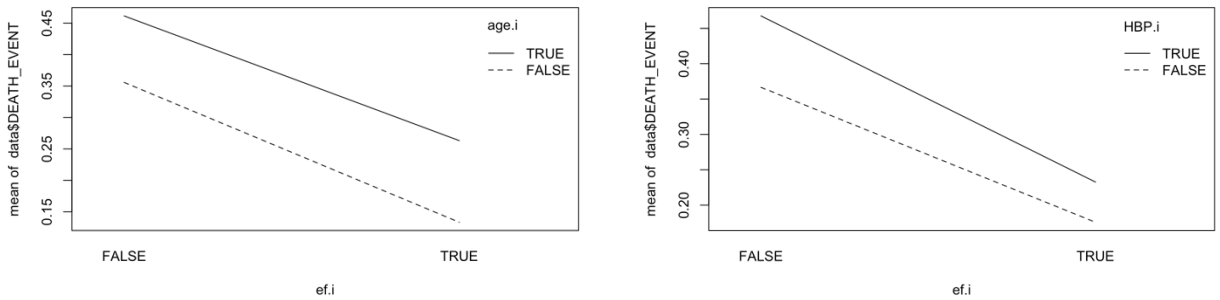
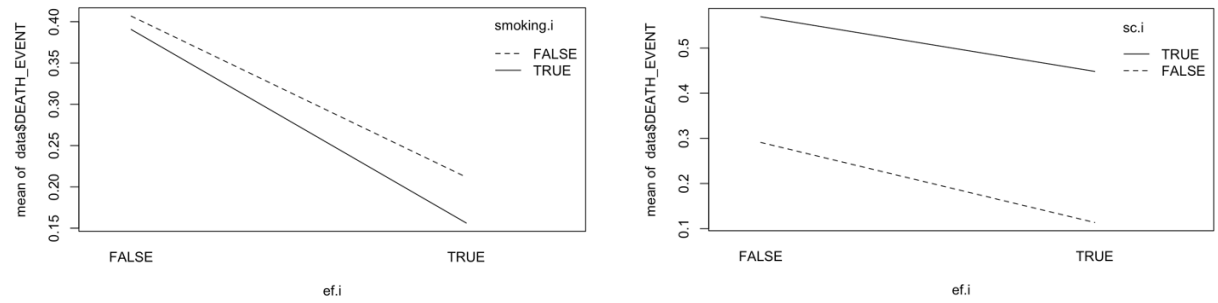


Figure 8: Plot of the fitted values vs. the residuals of the full model.

The fitted-versus-residual plot looks like noise. This plot supports normality and constant variance of the residuals.

Interaction plots test for a significant interaction between ejection fraction and the other predictor variables using the general linear f-test.





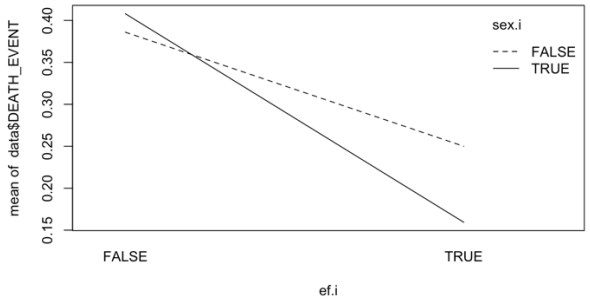


Figure 9(a,b,c,d,e): Interaction plots of ejection fraction vs. the five other predictor variables.

Based on the interaction plots for each of the predictor variables against ejection fraction, there appears to be an interaction between ejection fraction and sex. The addition of an interaction term between ejection fraction and sex could benefit the fit of the model.

C. Model Selection and Validation

Automatic Variable Selection Methods helps to eliminate redundant variables from the dataset. For this, the "MASS" package is used and more specifically the "stepAIC" function from the package. The "best" model will have the smallest AIC value.

```
## stepAIC(fit)
Start:  AIC=310.28
DEATH_EVENT ~ ejection_fraction + age + HBP + sex + smoking +
    log(serum_creatinine) + ejection_fraction * sex

                        Df Deviance    AIC
- smoking                1   294.30 308.30
<none>                       294.28 310.28
- HBP                    1   297.11 311.11
- ejection_fraction:sex  1   298.04 312.04
```

```
- age                    1   309.58 323.58
- log(serum_creatinine)  1   319.64 333.64

Step:  AIC=308.3
DEATH_EVENT ~ ejection_fraction + age + HBP + sex + log(serum_creatinine) +
    ejection_fraction:sex

                       Df Deviance    AIC
<none>                      294.30 308.30
- HBP                   1   297.14 309.14
- ejection_fraction:sex  1   298.08 310.08
- age                   1   309.61 321.61
- log(serum_creatinine)  1   319.70 331.70

Call:  glm(formula = DEATH_EVENT ~ ejection_fraction + age + HBP + sex +
    log(serum_creatinine) + ejection_fraction:sex, family = binomial,
    data = data)

Coefficients:
     (Intercept)    ejection_fraction      age         HBP          sex
      -2.84478          -0.03653         0.04869     0.51826     1.67142
log(serum_creatinine)  ejection_fraction:sex
       1.57109               -0.05449

Degrees of Freedom: 298 Total (i.e. Null);  292 Residual
Null Deviance:     375.3
Residual Deviance: 294.3      AIC: 308.3
```

Figure 10: R output for step(AIC) for the preliminary model + interaction term.

```
## 311.1383
```

Figure 11: PRESS statistic for final model.

In this case, the best model includes ejection fraction, age, HBP, sex, log(serum creatinine), and an interaction term between ejection fraction and sex. This model has the lowest AIC at a value of 308.3 and a low PRESS statistic of 311.1383.

A **receiver operating characteristic (ROC) curve** plots sensitivity on the vertical axis versus (1 − specificity) on the horizontal axis.[1] The ROC area under the curve summarizes the predictive power for all possible π_0. The greater the area under the curve, the better the predictive power of the model.

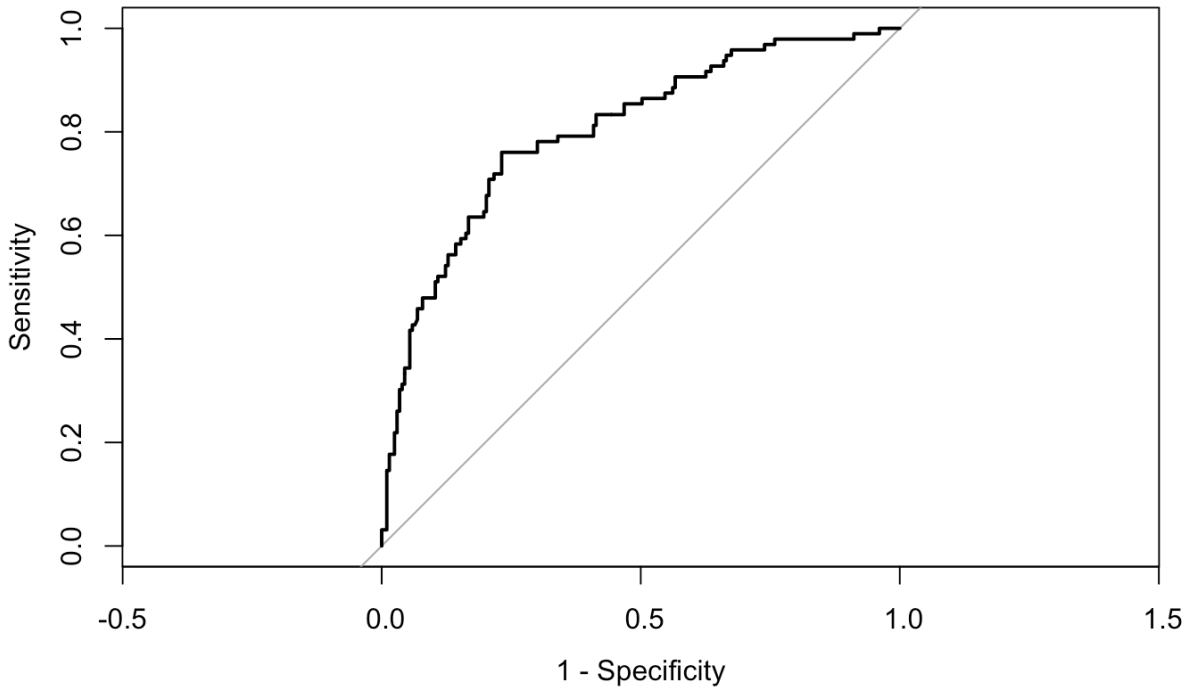


Figure 12: ROC curve for logistic regression model with final model.

An AUC value of 0.5 indicates that the model is no better than random guessing. The AUC for the final model is 0.8024, and this indicates that the model has a strong predictive power.

D. Outliers

To detect outliers in this study, the studentized residual plot is used. Based on this plot, at least 90% of the data points should be within the range of ±3 and this is found to be true.

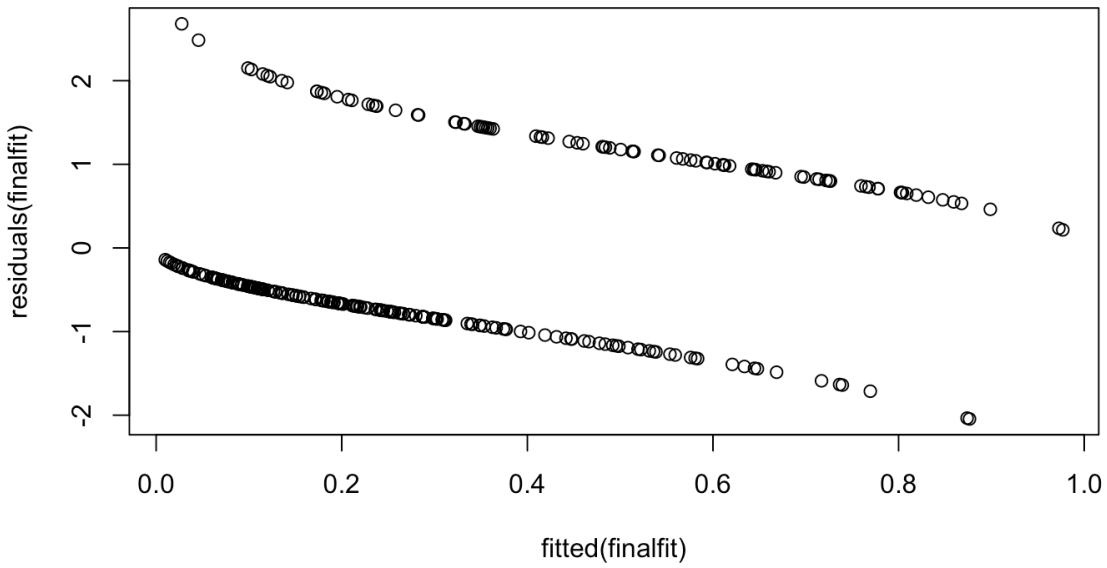


Figure 13: Studentized residual plot for the final model.

Another way to detect outliers is to analyze the plot of Cook's distance.

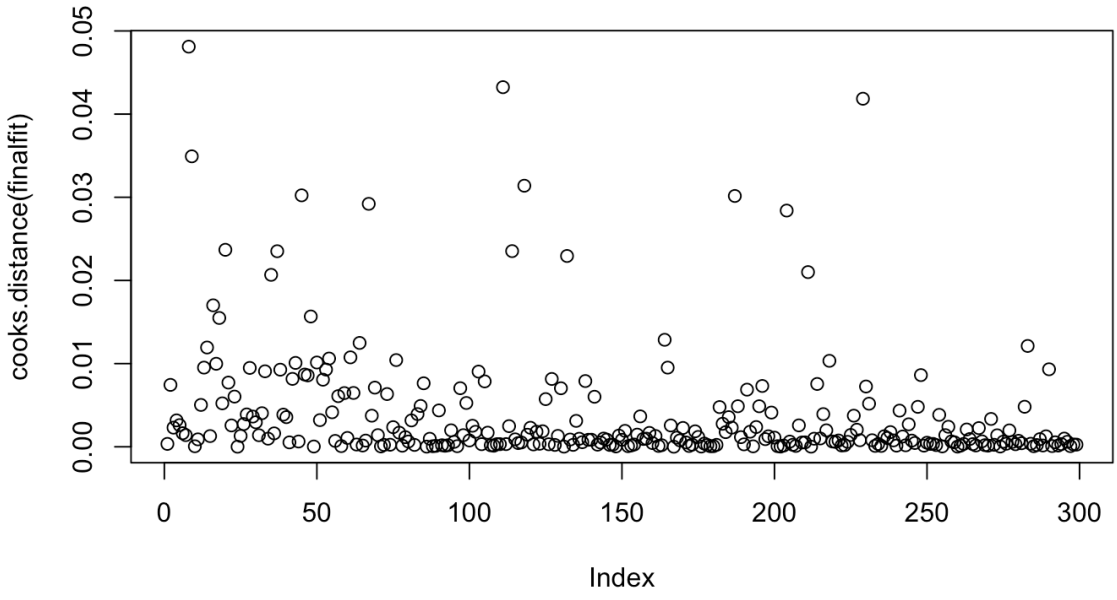


Figure 14: Plot of Cook's distance for the final model.

We compare the cook's distances to 2/sqrt(n) for large sample sizes, and this is a large sample size with 299 observations. The threshold is 2/sqrt(299) = 0.115663. All of the Cook's

distance values are less than the threshold which indicates that there are no outliers that are significant. No outliers are removed from the dataset.

E. Testing hypotheses

For each of the predictor variables, we test if there is a linear association between the logit(π) and the predictor variables. The Wald test and the Likelihood Ratio Test are run under the following conditions:

* **Null hypothesis** : H_0 :: $\beta_x = 0$
* **Alternative hypothesis** :: $H_1 : \beta_x \neq 0$.

(where x = 1, 2, 3, 4, 5, 6 ,7 for each predictor variable)

For large samples, the **Wald test** statistic z = hat(β)/SE has a standard normal distribution when β = 0.[2] Equivalently, for the two-sided Ha: β = 0, z^2 = (hat(β)/SE)^2 has a large-sample chi-squared null distribution with df = 1.[2]

```
Call:
glm(formula = DEATH_EVENT ~ ejection_fraction + age + HBP + sex +
    log(serum_creatinine) + ejection_fraction * sex, family = binomial,
    data = data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.84478    1.08369  -2.625 0.008663 **
ejection_fraction     -0.03653    0.01988  -1.838 0.066107 .
age                    0.04869    0.01285   3.788 0.000152 ***
HBP                    0.51826    0.30852   1.680 0.092985 .
sex                    1.67142    1.06357   1.572 0.116062
log(serum_creatinine)  1.57109    0.33205   4.732 2.23e-06 ***
ejection_fraction:sex -0.05449    0.02814  -1.936 0.052833 .
---
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: 294.30  on 292  degrees of freedom
AIC: 308.3
```

```
Number of Fisher Scoring iterations: 5
```

Figure 15: Summary table of the final model.

The Wald test shows strong evidence that there is a positive effect of age and log(serum creatine) on the presence of a death event at the 95% significant level (we reject H_0) and a positive effect of ejection fraction and the interaction term between ejection fraction and sex on the presence of a death event at the 90% significance level(we reject H_0).

We can also conduct the **Likelihood Ratio Test**. Using the Likelihood Ratio test, we can compute the test statistic $2(L_1 – L_0)$ where L_0 denotes the maximum of the log-likelihood function when $\beta = 0$, which is the null model, containing only an intercept term and let L_0 denote the maximum log-likelihood for unrestricted β.

```
Analysis of Deviance Table (Type II tests)


Response: DEATH_EVENT
                      LR Chisq Df Pr(>Chisq)
ejection_fraction      25.2203  1  5.114e-07 ***
age                    15.3085  1  9.131e-05 ***
HBP                     2.8316  1    0.09242 .
sex                     0.9868  1    0.32052
log(serum_creatinine)  25.3947  1  4.672e-07 ***
ejection_fraction:sex   3.7792  1    0.05189 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16: ANOVA table for final model.

The Likelihood Ratio test shows that ejection fraction, age, and log(serum creatine) have a strong effect on the presence of a death event and all three have p-values < 0.001 which indicates that we reject the null hypothesis that $\beta = 0$. For the predictor variables HBP and the interaction term ejection fraction*sex, the p-value < 01 but does not show significance at the 95% significance level.

CONCLUSIONS
In real terms, the final model can be expressed as:

$$logit[\pi(x)] = -2.84478 - 0.03653X_{i1} + 0.04869X_{i2} + 0.51826X_{i3} + 1.67142X_{i4} + 1.67142X_{i5} - 0.05449 X_{i1} * X_{i4}$$

where

$logit[\pi(x)]$ = death event, If the patient is deceased during the follow-up period (1) or not (0) (binary)

X_{i1}= serum creatinine, the level of serum creatinine in the blood (mg/dL)

X_{i2}= age, the age of the patient

X_{i3} = HBP, If the patient has high blood pressure (hypertension) (1) or not (0) (binary)

X_{i4}= sex, the sex of the patient is male (1) or female (0) (binary)

X_{i5}= serum creatinine, the level of serum creatinine in the blood (mg/dL)

X_{i6} = smoking, If the patient smokes (1) or not (0) (binary)

$X_{i1} * X_{i4}$ = interaction term, $X_{i1} * X_{i4}$ (ejection fraction * sex).

i = 1,2,3, · · ·, 299.

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6,$ and σ^2 are the unknown parameters to be estimated.

This study shows that a patient's average ejection fraction value, age, whether or not they have high blood pressure, the sex of the patient, the serum creatine values and an interaction term between ejection fraction and sex can predict the whether or not a death event will occur in a patient. All statistical analysis was conducted at a 95% confidence interval and at 0.05 significance level (α). A linear association was found between 5 of the 6 original predictor variables with an additional linear association from an interaction term.

Dependent variables present in model.	AIC	AUC	PRESS
Ejection fraction	355.968	0.6761	356.826
ejection fraction + age	333.401	0.7418	334.443
ejection fraction + age + HBP	334.10	0.7475	335.522
ejection fraction + age + HBP + sex	335.117	0.7493	337.037
ejection fraction + age + HBP + sex + log(serum creatinine)	310.083	0.798	312.344
ejection fraction + age + HBP + sex + log(serum creatinine) + smoking	312.036	0.7976	314.728
ejection fraction + age + HBP + sex + log(serum creatinine) + smoking + ejection fraction * sex (interaction term)	310.275	0.8031	313.574
ejection fraction + age + HBP + sex + log(serum creatinine) + ejection fraction * sex (interaction term)	**308.303**	**0.8024**	**311.138**

Figure 17: Automatic variable selection methods statistics.

The final model was selected based on the model's low AIC value, high ROC AUC value, and low PRESS value as illustrated by Figure 13. There did not appear to be a presence of outliers in the data with no datapoints with Cook's distances above the threshold. The bold model in Figure 17 above is the best model and possesses the strongest predictive power over all of the other models tested in this study.

REFERENCES

1. Davide Chicco, Giuseppe Jurman: "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". BMC Medical Informatics and Decision Making 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5
2. Agresti, Alan. *An Introduction to Categorical Data Analysis*. Wiley, 2019.
3. *Kaggle*, www.kaggle.com/. Accessed 10 May 2024.

CODE APPENDIX

```
# Load packages
library(tidyverse)
library(caret)
library(asbio)
library(olsrr)
library(xtable)
library(shiny)
library(knitr)
library(DT)
require(scatterplot3d)
require(Hmisc)
require(rgl)
require(faraway)
library(car)
library(gam)
# my data
data <- read_csv("~/Desktop/STAT 835/Final Project
/heart_failure_clinical_records_dataset.csv",
          col_types = cols(anaemia = col_skip(),
                           creatinine_phosphokinase = col_skip(),
```

```
                    diabetes = col_skip(), platelets = col_skip(),
                    serum_sodium = col_skip(), time = col_skip()))
data <- data %>% rename(HBP = high_blood_pressure)

```

III. Results

Fit the initial model:
```{r}
fit <- glm(DEATH_EVENT~ejection_fraction+age+HBP+sex+serum_creatinine+smoking, family
= binomial(link = logit), data)
```

```{r}
summary(fit)
Anova(fit)
confint(fit)

```
The ML fit of the logistic regression model is
logit[Pˆ(y = 1)] = -2.28558 - 0.07249x1 + 0.05231x2 + 0.38794x3 - 0.30699x4 + 0.68922 + 0.03840. Since we have a large n value (299 with 6 predictor variables and degrees of freedom equal to 1), the Wald test was used at the 95% confidence level to determine if H0:βx = 0. The ejection fraction effect βˆ = -0.07249 and SE = 0.0145. The z-value = -4.990 with a p-value = 6.04e-07. Based on this, we have strong evidence to reject the null hypothesis and this provides very strong evidence that as ejection fraction decreases, tthe likelihood of a death event occurring increases. The age effect βˆ = 0.05231 and SE = 0.01256. The z-value = 4.166 with a p-value = 3.10e-05. Based on this, we have strong evidence to reject the null hypothesis and this provides very strong evidence that as age increases, the likelihood of a death event occurring increases. The serum creatinine effect βˆ = 0.68922 and SE = 0.16579. The z-value = 4.157 with a p-value = 3.22e-05. Based on this, we have strong evidence to reject the null hypothesis and this provides very strong evidence that as serum creatinine levels increase, the likelihood of a death event occurring increases. The other three predictor variables: high blood pressure, sex, and smoking do not appear to be significant at the 95% confidence interval given this analysis.

A. Diagnostics for Predictors.

In order to determine the presence of any bivariate associations between predictor variables, we will compute a scatter plot matrix and Pearson correlation coefficient matrix. These two tests will help to determine if there is any multicollinearity between variables in the model.
```{r}
pairs(DEATH_EVENT~ejection_fraction+age+HBP+sex+
    serum_creatinine+smoking, data)
```

```{r}
# The Pearson correlation coefficients for all pairwise association
cor(data)
```
The largest Pearson correlation coefficient value is between smoking and sex +0.445891712 which indicates moderate correlation. All other correlation coefficients are below 0.29 which indicate weak correlation.

Strip plots for each continuous variable
```{r}
for (i in 1:6){
  par(mfrow=c(1,2))
  stripchart(data[,i], main = names(data)[i],
          vertical = T, method = "jitter")
  boxplot(data[,i], main = names(data)[i])
  par(mfrow=c(1,1))
}

stripchart(log(data$serum_creatinine), main = "log(serum creatinine)", vertical = T, method = "jitter")
boxplot(log(data$serum_creatinine), main = "log(serum creatinine)")
```
For the three continuous variables: ejection fraction, age, and serum creatinine, strip plots were made. The variables age and ejection fraction appear to be evenly distributed with ejection fraction having a couple data points outside of the boxplot standard deviation. However, the serum creatinine data is positively skewed with most observations clustered together and a few observations at much higher levels. Due to this, we could consider a log transformation

Log transform serum creatinine variable
```{r}
fit2 <- glm(DEATH_EVENT~ejection_fraction+age+HBP+sex+log(serum_creatinine)+smoking, family = binomial(link = logit), data)
```

C. Screening of Predictors

1. **Added variable plots** Partial Residual plots
```{r}
prplot(fit2,1)
prplot(fit2,2)
prplot(fit2,3)
```

```
prplot(fit2,4)
prplot(fit2,5)
prplot(fit2,6)
```
Scatterplot of y by different predictor variables
```{r}
# ejection fraction
plot(jitter(DEATH_EVENT, 0.08) ~ ejection_fraction, data=data)
gam.fit1 <- gam(DEATH_EVENT ~ s(ejection_fraction), family=binomial, data=data)
curve(predict(gam.fit, data.frame(ejection_fraction=x), type="resp"), add=TRUE)
# age
plot(jitter(DEATH_EVENT, 0.08) ~ age, data=data)
gam.fit2 <- gam(DEATH_EVENT ~ s(age), family=binomial, data=data)
curve(predict(gam.fit2, data.frame(age=x), type="resp"), add=TRUE)
# log(serum_creatinine)
plot(jitter(DEATH_EVENT, 0.08) ~ log(serum_creatinine), data=data)
gam.fit5 <- gam(DEATH_EVENT ~ s(log(serum_creatinine)), family=binomial, data=data)
curve(predict(gam.fit5, data.frame(serum_creatinine=x), type="resp"), add=TRUE)

fitten <- glm(DEATH_EVENT ~ ejection_fraction, family=binomial, data=data)
predict(fitten, data.frame(ejection_fraction = mean(data$ejection_fraction)),
type="response")
 # estimated probability of a death event at the mean ejection fraction 0.3039407
```
For the three continuous predictor variables, a curve of the generalized additive model smoothing fit and the logistic regression models are plotted.
Caption: Whether a death event occurs (y = 1, yes; y = 0, no), by x = ejection fraction of the patient, and generalized additive model smoothing fit.
Repeat for each variable.

Fitted values and confidence intervals for probabilities (page 111)
```{r}
pred.prob <- fitted(fitten) # ML fitted value estimate of P(Y=1)
lp <- predict(fitten, se.fit=TRUE) # linear predictor
LB <- lp$fitten - 1.96*lp$se.fit # confidence bounds for linear predictor
UB <- lp$fitten + 1.96*lp$se.fit # better: use qnorm(0.975) instead of 1.96
LB.p <- exp(LB)/(1 + exp(LB)) # confidence bounds for P(Y=1)
UB.p <- exp(UB)/(1 + exp(UB))
cbind(pred.prob, LB.p, UB.p)
```

2. Look at VIF values
```{r}
vif(fit)
```

```
vif(fit2)
vif(finalfit)
```
3. **Automatic variable selection methods**
```{r}
#Using AIC
library(MASS)
stepAIC(fit2)

PRESS(fit2)

# trying something
fit3 <- glm(DEATH_EVENT ~ ejection_fraction + age + HBP + sex + smoking+
  log(serum_creatinine) + ejection_fraction*sex, family = binomial, data = data)
stepAIC(fit3)

# for chart
glm1 <- glm(DEATH_EVENT~ejection_fraction, family = binomial, data = data)
AIC(glm1)
glm2 <- glm(DEATH_EVENT~ejection_fraction+age, family = binomial, data = data)
AIC(glm2)
glm3 <- glm(DEATH_EVENT~ejection_fraction+age+HBP, family = binomial, data = data)
AIC(glm3)
glm4 <- glm(DEATH_EVENT~ejection_fraction+age+HBP+sex, family = binomial, data = data)
AIC(glm4)
glm5 <- glm(DEATH_EVENT~ejection_fraction+age+HBP+sex+log(serum_creatinine), family =
binomial, data = data)
AIC(glm5)
glm6 <-
glm(DEATH_EVENT~ejection_fraction+age+HBP+sex+log(serum_creatinine)+smoking, family
= binomial, data = data)
AIC(glm6)
glm7 <- glm(DEATH_EVENT~ejection_fraction+age+HBP+sex+log(serum_creatinine)+smoking
+ejection_fraction*sex,family = binomial, data = data)
AIC(glm7)
glm8 <-
glm(DEATH_EVENT~ejection_fraction+age+HBP+sex+log(serum_creatinine)+ejection_fractio
n*sex, family = binomial, data = data)
AIC(glm8)

prediction <- predict(finalfit, data,
                type="response")
# create roc curve
prediction1 <- predict(glm1, data, type="response")
roc_object1 <- roc(data$DEATH_EVENT, prediction1)
auc(roc_object1)
```

```
prediction2 <- predict(glm2, data, type="response")
roc_object2 <- roc(data$DEATH_EVENT, prediction2)
auc(roc_object2)
prediction3 <- predict(glm3, data, type="response")
roc_object3 <- roc(data$DEATH_EVENT, prediction3)
auc(roc_object3)
prediction4 <- predict(glm4, data, type="response")
roc_object4 <- roc(data$DEATH_EVENT, prediction4)
auc(roc_object4)
prediction5 <- predict(glm5, data, type="response")
roc_object5 <- roc(data$DEATH_EVENT, prediction5)
auc(roc_object5)
prediction6 <- predict(glm6, data, type="response")
roc_object6 <- roc(data$DEATH_EVENT, prediction6)
auc(roc_object6)
prediction7 <- predict(glm7, data, type="response")
roc_object7 <- roc(data$DEATH_EVENT, prediction7)
auc(roc_object7)
prediction8 <- predict(glm8, data, type="response")
roc_object8 <- roc(data$DEATH_EVENT, prediction8)
auc(roc_object8)
# for each in chart!!!!!! dont forget to add those other plots
# PRESS VALUES
PRESS(glm1)
PRESS(glm2)
PRESS(glm3)
PRESS(glm4)
PRESS(glm5)
PRESS(glm6)
PRESS(glm7)
PRESS(glm8)
```
ROC curve
```{r}
finalfit <- glm(DEATH_EVENT ~ ejection_fraction + age + HBP + sex +
  log(serum_creatinine) + ejection_fraction*sex, family = binomial, data = data)

library(pROC)

prediction <- predict(finalfit, data,
                      type="response")
# create roc curve
roc_object <- roc(data$DEATH_EVENT, prediction)
# area under the curve = 0.8024
auc(roc_object)
#ROC curve for fit0
```

```
rocplot0 <- roc(DEATH_EVENT~fitted(fit0), data)
plot.roc(rocplot0, legacy.axes = TRUE)

abline(v=1)
abline(h=1)

rocplot <- roc(DEATH_EVENT~fitted(fit), data)
plot.roc(rocplot, legacy.axes = TRUE)

abline(v=1)
abline(h=1)

rocplot2 <- roc(DEATH_EVENT~fitted(finalfit), data)
plot.roc(rocplot2, legacy.axes = TRUE)

abline(v=1)
abline(h=1)

# R value
cor(data$DEATH_EVENT, fitted(finalfit))
```
Caption: ROC curve for logistic regression model with final model.
A receiver operating characteristic (ROC) curve plots sensitivity on the vertical axis versus (1 – specificity) on the horizontal axis. reference for textbook!! It summarizes the predictive power for all possible π0. The greater the area under the curve, the better the predictive power of the model.

Residual plots
```{r}
# Residual plots
raw_residual=data$DEATH_EVENT-fitted(finalfit)
plot(raw_residual~fitted(finalfit))

plot(residuals(finalfit)~fitted(finalfit))
```
Interaction plots
```{r}

age.i <- data$age > mean(data$age)
ef.i <- data$ejection_fraction > mean(data$ejection_fraction)
HBP.i <- data$HBP > mean(data$HBP)
sc.i <- log(data$serum_creatinine) > mean(log(data$serum_creatinine))
smoking.i <- data$smoking > mean(data$smoking)
sex.i <- data$sex > mean(data$sex)
```

```
interaction.plot(ef.i,age.i,data$DEATH_EVENT)
interaction.plot(ef.i,HBP.i,data$DEATH_EVENT) # Look at that! Is it significant?
interaction.plot(ef.i,sc.i,data$DEATH_EVENT)
interaction.plot(ef.i,smoking.i,data$DEATH_EVENT)
interaction.plot(ef.i,sex.i,data$DEATH_EVENT)

interactionfit <- glm(DEATH_EVENT ~ ejection_fraction + age + HBP +
   log(serum_creatinine) + ejection_fraction*sex, family = binomial, data = data)
Anova(interactionfit)
# Does not appear to be significant at the 95% level therefore, we will not include
# the interaction term in our final model.
```

2. **Outliers**

```{r}
plot(residuals(finalfit)~fitted(finalfit))

plot(cooks.distance(finalfit)) # Compare percentile F(p,n-p) to 10th or 20th
which(cooks.distance(finalfit)>0.115663) #2/sqrt(n)

```

```{r}
summary(finalfit)
Anova(finalfit)
```