# Predicting Severity and Death from Heart Failure

Using Multiple Linear and Logistic Regression Models

**Angel Silvar**

California State University - Long Beach

December 17, 2021

# 1 Introduction

This project aims to make statistical predictions of severity and death following a heart failure event. The dataset used for this project comes from a 2017 study of heart failure in patients over 40 done by researchers in Pakistan. The data may be found on the UCI Machine Learning Repository under "Heart failure clinical record." In overview, this project seeks to determine if there exists a model such that at least 1 or more of the predictors in the data set yields a significant model to predict the resultant ejection fraction (i.e. how much blood is pumped out of the heart; referred to as severity through our project) following heart failure, as well as if the heart failure leads to a death event in the follow-up period. Then, use this model to see if, given certain characteristics of a patient, predict the severity of the heart failure and determine if the patient was predicted to die.

| Name | Description |
|---|---|
| age | Age of the patient (years) |
| anaemia | Decrease of red blood cells or hemoglobin (binary) |
| high blood pressure | Patient has hypertension (binary) |
| creatinine phosphokinase (CPK) | level of the CPK enzyme in the blood (mcg/L) |
| diabetes | Patient has diabetes (binary) |
| ejection fraction | percentage of blood leaving heart at each contraction (%) |
| platelets | Platelets in the blood (kiloplatelets/mL) |
| sex | Woman or man (binary) |
| serum creatinine | Level of serum creatinine in blood (mg/dL) |
| serum sodium | Level of serum sodium in blood (mEq/L) |
| smoking | Patient smokes (binary) |
| time | Follow-up period (days) |
| death event | Patient died during follow-up period (binary) |

Table 1: Attributes of the Dataset.

# 2 Questions of Interest

- Does the regression model contain at least one predictor useful in predicting the severity of the heart failure, specifically those not related to the heart failure and part of the patient profile?
- Is there an accurate logistic model for a death based on the predictors?
- What is the probability of death for a patient with an ejection fraction one standard deviation from the mean?

# 3 Regression Method

To answer our first question, we will create a multiple linear regression model from our data, using ejection fraction (severity) as our response.

The second and third questions will be answered through the creation of a binary logistic regression model, using death as the response.

Here we will predict for certain characteristics of a patient and attempt to predict whether or not the data suggests if the person would die and determine how rough was the severity of the heart failure. Also, we will use tests to determine the significance of the final model and determine our accuracy of our model and state explicitly which model is more significant/reliable in terms of prediction.

# 4 Regression Analysis, Results and Interpretation

## 4.1 Multiple Linear Regression

Based on our research question, we have our hypotheses as follows:

$$H_0 : \beta_0 = \beta_1 = ... = \beta_{12} = 0$$

$$H_1 : \beta_k \neq 0, \text{ for at least one } k = 0, 1, ...12$$

Where $\beta_k$ is from $E(Y) = \beta_0 + \beta_1 x_1 + ... + \beta_{12} x_{12}$ a multiple linear regression model using ejection fraction (severity) as a response to the remaining data values as predictors.

First, the following correlation scatterplot (Figure 1) was used to gauge how to approach the problem and expect which variables are non correlated. The plot shows little to no correlation between any predictors and our desired severity response, lending us to believe that it is unlikely a strong linear model can be created.

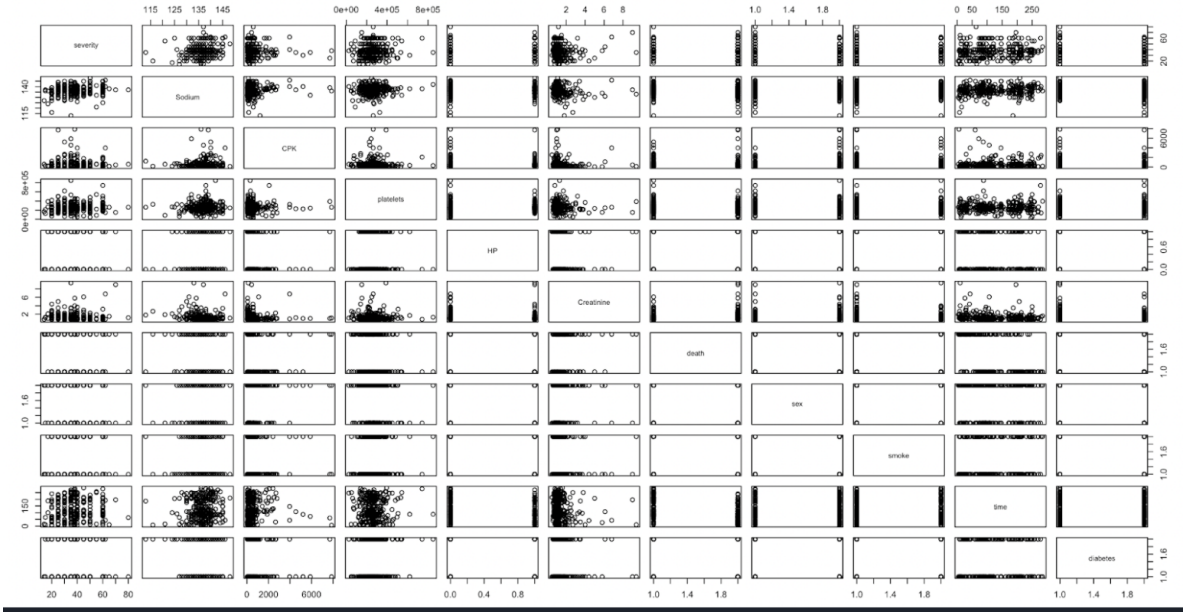



Figure 1: Correlation Scatterplot for all values in dataset

To start with variable selection, we performed a step-wise regression of all predictors using AIC as our criteria. After much iteration and double checking our results by manually testing the last few models that the step wise method gave, we concluded that the following model would be the best for the data set. Ultimately, this process gave us a model with six distinct predictors: death, sex, age, serum sodium, time, and serum creatinine.

```
Step:  AIC=1443.31
severity ~ death + sex + age + Sodium + time + Creatinine

             Df Sum of Sq   RSS    AIC
<none>                    35623 1443.3
- Creatinine  1     259.1 35882 1443.5
+ platelets   1      60.8 35562 1444.8
- time        1     456.9 36080 1445.1
+ smoke       1       1.5 35622 1445.3
+ CPK         1       1.5 35622 1445.3
+ HP          1       0.6 35623 1445.3
+ diabetes    1       0.3 35623 1445.3
+ anaemia     1       0.2 35623 1445.3
- age         1     625.9 36249 1446.5
- Sodium      1     692.4 36316 1447.1
- sex         1    1025.6 36649 1449.8
- death       1    3626.0 39249 1470.3

Call:
lm(formula = severity ~ death + sex + age + Sodium + time + Creatinine)

Coefficients:
(Intercept)       death1         sex1          age       Sodium         time   Creatinine
  -11.69061     -9.25724     -3.89133      0.12758      0.35599     -0.01889      0.95671
```

```
Call:
lm(formula = severity ~ death + sex + age + Sodium + time + Creatinine)

Residuals:
    Min      1Q  Median      3Q     Max
-22.168  -7.635  -1.561   6.230  37.240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.690615  20.983266  -0.557  0.57786
death1       -9.257241   1.698012  -5.452 1.06e-07 ***
sex1         -3.891335   1.342081  -2.899  0.00402 **
age           0.127580   0.056324   2.265  0.02424 *
Sodium        0.355993   0.149434   2.382  0.01785 *
time         -0.018894   0.009763  -1.935  0.05391 .
Creatinine    0.956714   0.656535   1.457  0.14613
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.05 on 292 degrees of freedom
Multiple R-squared:  0.1465,    Adjusted R-squared:  0.129
F-statistic: 8.355 on 6 and 292 DF,  p-value: 2.273e-08
```

Figure 2: RStudio output for AIC model and a Summary table for the best model the AIC suggested.

However, a small adjusted $R^2$ value of 0.129 showed that only around 13% of the variation in severity could be explained by our model. This strong variation of the suggests that transformation is

needed. The Normal Q-Q plot suggested that normality of the errors are not met, hence, rending this model to transformations before interpretation.

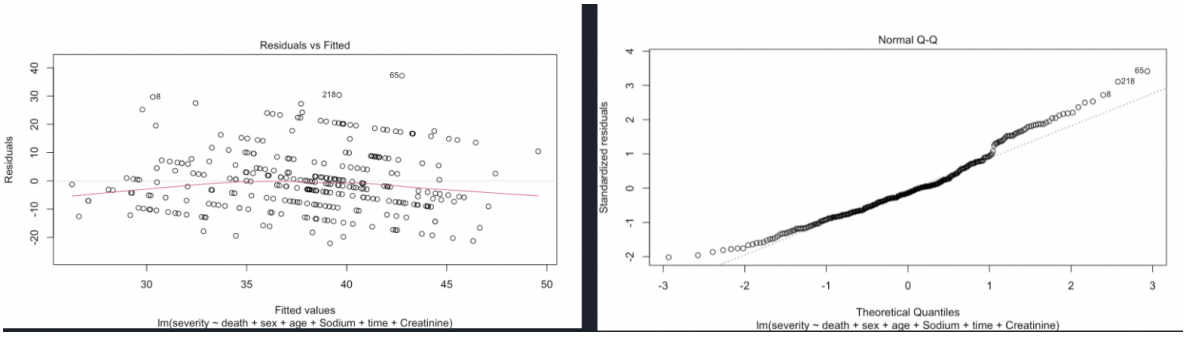
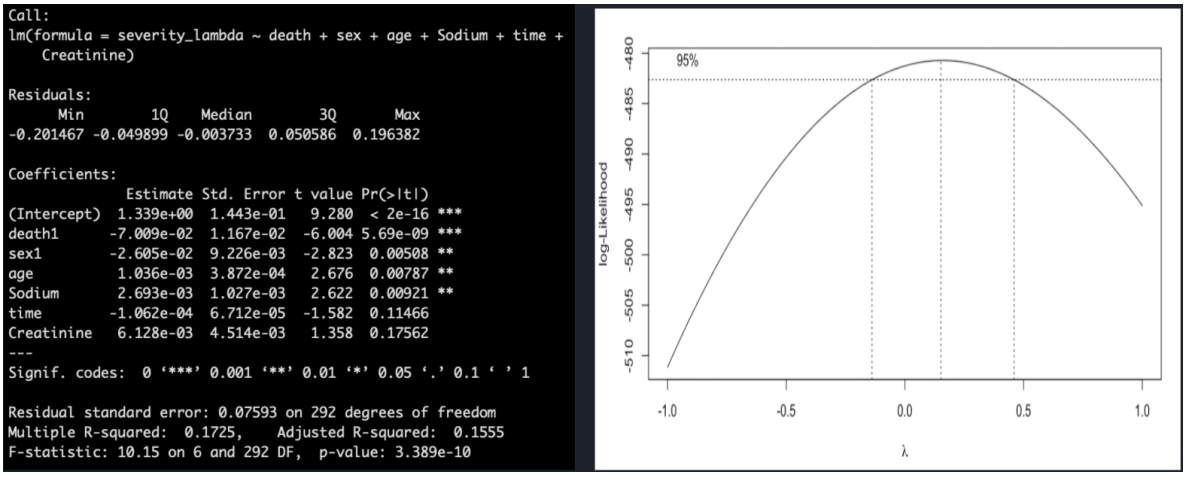



Figure 3: Residual vs. Fitted Values and Normal Q-Q plots for initial linear model

The Box-Cox method yielded a $\lambda$ value in between of roughly [-0.25 , 0.40]. Seeing that $\lambda = 0$ is an element of this interval we decided to log-transform the Y response variable, This value meets the $\alpha = 0.05$ significance threshold.



```
Call:
lm(formula = severity_lambda ~ death + sex + age + Sodium + time +
    Creatinine)

Residuals:
      Min        1Q    Median        3Q       Max
-0.201467 -0.049899 -0.003733  0.050586  0.196382

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.339e+00  1.443e-01   9.280  < 2e-16 ***
death1      -7.009e-02  1.167e-02  -6.004 5.69e-09 ***
sex1        -2.605e-02  9.226e-03  -2.823  0.00508 **
age          1.036e-03  3.872e-04   2.676  0.00787 **
Sodium       2.693e-03  1.027e-03   2.622  0.00921 **
time        -1.062e-04  6.712e-05  -1.582  0.11466
Creatinine   6.128e-03  4.514e-03   1.358  0.17562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07593 on 292 degrees of freedom
Multiple R-squared:  0.1725,    Adjusted R-squared:  0.1555
F-statistic: 10.15 on 6 and 292 DF,  p-value: 3.389e-10
```



Figure 4: R-Studio output for log-transformed model and its associated Box-Cox lambda plot

This transformation fixed the normality issue. The Normal Q-Q plot depicts a plot of the errors that are close to the diagonal line. Hence, normality condition was met. The Shapiro-Wilk test gave us a p-value of .6345, which supports a failure to reject the the null hypothesis that the residuals are normality distributed.

For our next analysis, we attempt to see if it were possible to reduce the number of variables in the model. First we started with the matrix correlation with the selected variables suggested by the AIC Step Wise Method.

Investigating our correlation of severity and our newly selected predictors for our model, shown in Figure 6, we can see that the lack of spread of sodium and creatinine may be amenable to logarithmic transformations.

Following their transformations, the newly created model (Figure 7) shows a slight increase in adjusted $R^2$ to 0.15; however, the p-value for creatinine raised considerably to 0.92479, making it no longer statistically significant.

The lack of significance in creatinine brings up further need for analysis in our variable selection. As a check for our step-wise selection based on AIC values, we will also perform a regression subset analysis. In Figure 8, the results of the regression subset analysis can be seen. Subset 6 shows the same model found with our step-wise variable selection, while subset 5 does not include creatinine.

In Table 2, the adjusted $R^2$ values are listed for the regression subsets. Subset 6 (the subset that matches our variables from step-wise selection) has the highest value of 0.1290, while subset 5 (without

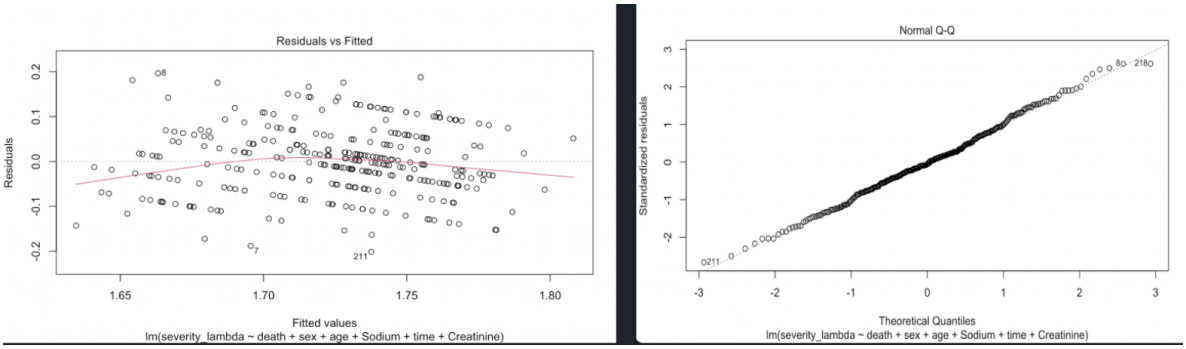



Figure 5: Residual vs. Fitted Values and Normal Q-Q for log-transformed model

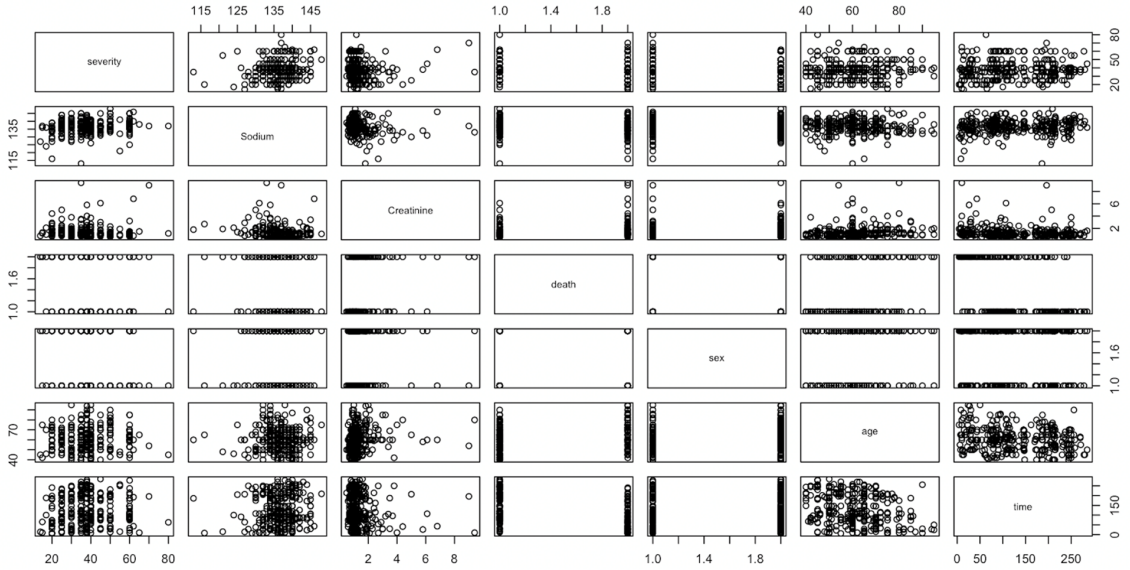
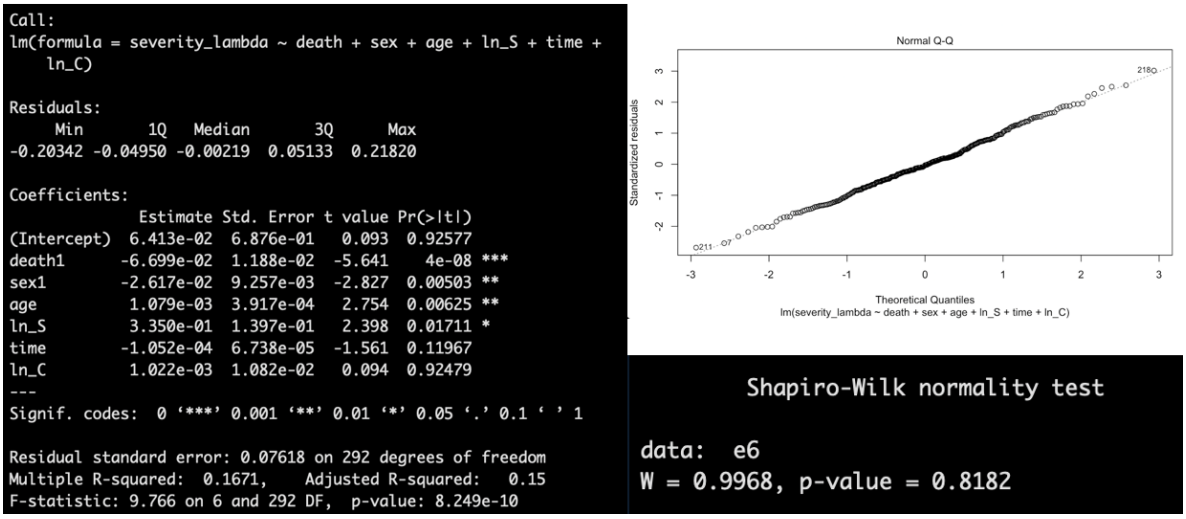



Figure 6: Correlation Scatterplot for selected variables



```
Call:
lm(formula = severity_lambda ~ death + sex + age + ln_S + time +
    ln_C)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20342 -0.04950 -0.00219  0.05133  0.21820

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.413e-02  6.876e-01   0.093  0.92577
death1      -6.699e-02  1.188e-02  -5.641    4e-08 ***
sex1        -2.617e-02  9.257e-03  -2.827  0.00503 **
age          1.079e-03  3.917e-04   2.754  0.00625 **
ln_S         3.350e-01  1.397e-01   2.398  0.01711 *
time        -1.052e-04  6.738e-05  -1.561  0.11967
ln_C         1.022e-03  1.082e-02   0.094  0.92479
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07618 on 292 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:   0.15
F-statistic: 9.766 on 6 and 292 DF,  p-value: 8.249e-10
```



```
        Shapiro-Wilk normality test

data:  e6
W = 0.9968, p-value = 0.8182
```

Figure 7: RStudio output for log-transformed sodium and creatinine

creatinine) is close in value with 0.1256. Figure 9 shows the plot of the subsets' Mallows's $C_p$, and both

| | (Intercept) | age | anaemia | diabetes | CPK | platelets | HBP | Creatinine | Sodium | death | sex | smoke | time |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| 2 | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE |
| 3 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE |
| 4 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | FALSE |
| 5 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | TRUE |
| 6 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | FALSE | TRUE |
| 7 | TRUE | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE | FALSE | TRUE |
| 8 | TRUE | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |

Figure 8: RStudio output for Regression Subset analysis

subsets 5 and 6 can be seen as viable options, as their $C_p$ values fall far below the desired line. Based on this analysis, and in search of a more parsimonious model, subset 5 would be a strong candidate for our final model.

| Adjusted $R^2$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0.0690 | 0.0884 | 0.1060 | 0.1178 | 0.1256 | 0.1290 | 0.1275 | 0.1246 |

Table 2: Adjusted $R^2$ Values for regression subsets

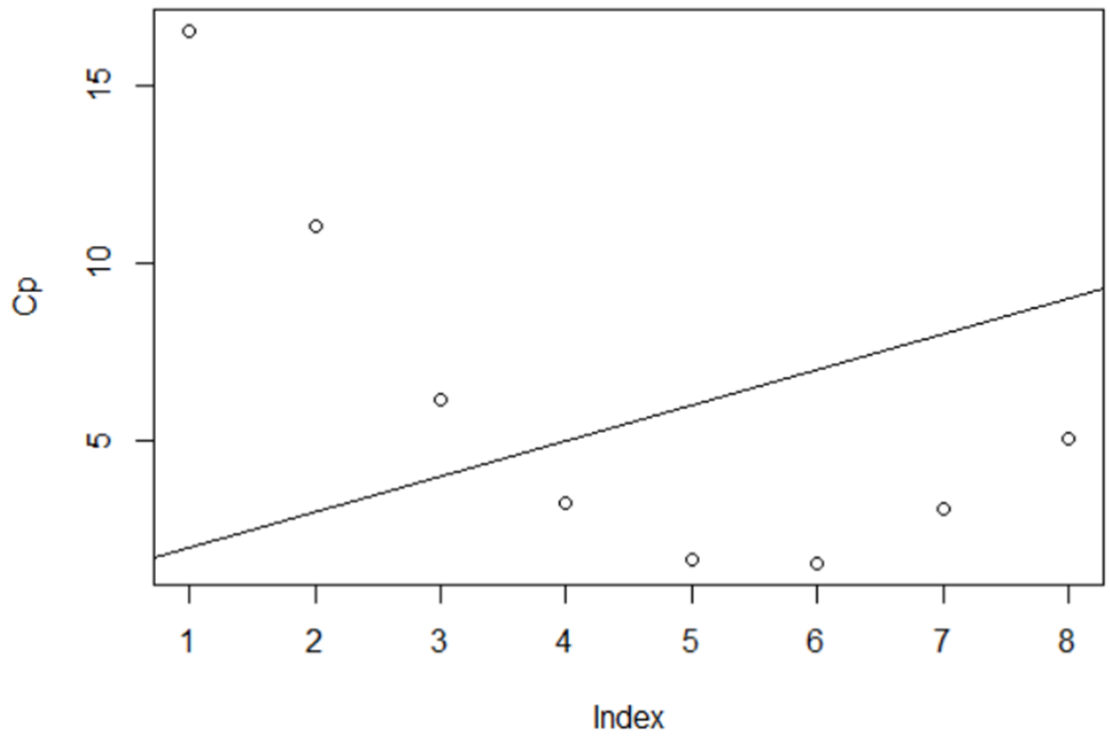



Figure 9: Mallows's $C_p$ for regression subsets

The next step was the play around with the variables and determine the best model.

Here we transformed both Sodium and Severity and it appears that normality is the same. We have excluded most of the analysis for space, but we have transformed both sodium and creatinine to see if it the 2nd variable was significant. The test failed. Then, we test each individual variable along

with log transforming the variable to see which one was more significant. Creatinine resulted to be the most insignificant variable of the two. Hence, we decided to just keep the log sodium transformation since it both was the more significant variable and the log sodium also gave us higher normality. In other words, normality condition went from p value 0.6 to 0.8 as shown in Figure 7.

From our previous regression subset analysis, it is possible to justify the removal of creatinine from our model. Figure 10 shows the model used to conduct a Box Cox for our 5 variable model that keeps our log-transformed sodium variable.

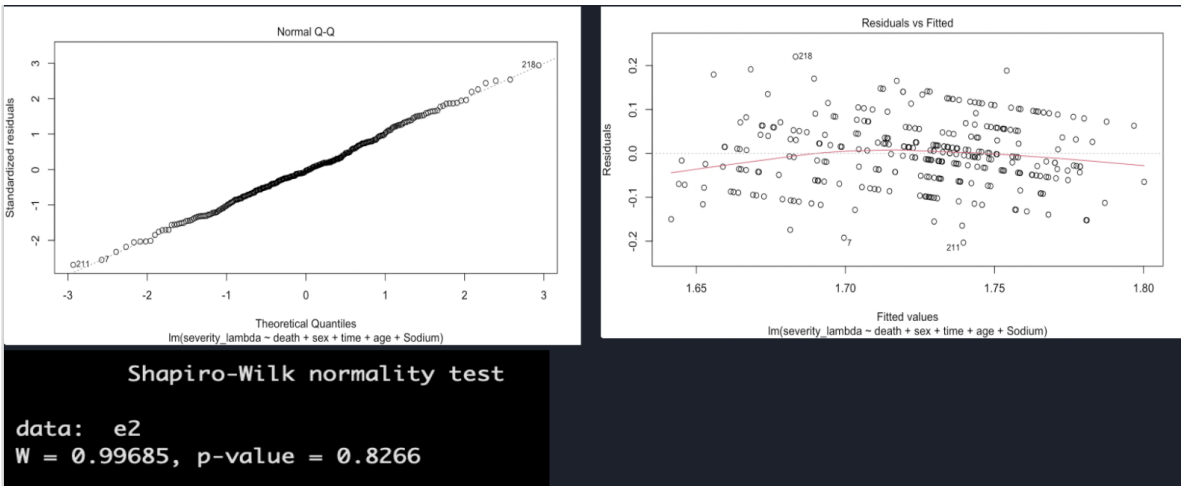



Figure 10: Q-Q Plot and Residual vs Fit plots for 5-variable model

Our final model can be seen in Figure 11.

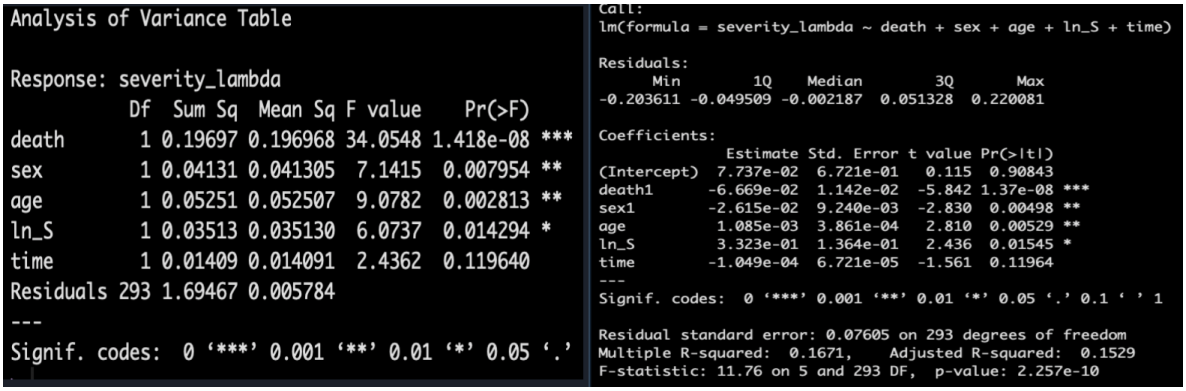



Figure 11: RStudio output for Final Linear Regression Model

The next step was to see if we could include an interaction term. Since this variable is separate from the AIC & Cp model, we sought to run the test shown in Figure 12.

The interaction test shows that none of the interaction terms are significant. The table on the left shows that the p value of death and age is insignificant. We tested this to double check significance and integrity of the model. Overall, we did not include any interaction terms and stuck with the final model mentioned earlier.

### 4.1.1 Predictions Using the Multiple Linear Model

The last portion of our MLR analysis is prediction. This was conducted just for fun, since we know that given any result, our extreme variation in our model is too high and does not lend itself to any predictions without any introduction of newer variables to the dataset.

The following figure will show a prediction that we made in order to satisfy this piece of our research question. The values were generated by the the mean values of each predictor (if continuous

```
Single term additions

Model:
severity_lambda ~ death + sex + age + ln_S + time
          Df Sum of Sq    RSS     AIC F value  Pr(>F)
<none>                  1.6947 -1534.7
death:ln_S  1 0.0045240 1.6902 -1533.5  0.7816 0.37738
death:time  1 0.0011741 1.6935 -1532.9  0.2024 0.65309
death:age   1 0.0160406 1.6786 -1535.6  2.7903 0.09591
sex:age     1 0.0031582 1.6915 -1533.3  0.5452 0.46088
sex:ln_S    1 0.0034216 1.6912 -1533.3  0.5907 0.44275
sex:time    1 0.0002452 1.6944 -1532.8  0.0423 0.83727
```

```
Call:
lm(formula = severity_lambda ~ death + sex + age + ln_S + time +
    death * age)

Residuals:
      Min        1Q    Median        3Q       Max
-0.198996 -0.050575 -0.006536  0.049814  0.227453

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.448e-01  6.712e-01   0.216  0.82935
death1      -1.475e-01  4.970e-02  -2.968  0.00325 **
sex1        -2.722e-02  9.234e-03  -2.948  0.00346 **
age          5.454e-04  5.024e-04   1.086  0.27856
ln_S         3.250e-01  1.361e-01   2.389  0.01754 *
time        -1.008e-04  6.705e-05  -1.503  0.13389
death1:age   1.296e-03  7.761e-04   1.670  0.09591 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07582 on 292 degrees of freedom
Multiple R-squared:  0.175,     Adjusted R-squared:  0.158
F-statistic: 10.32 on 6 and 292 DF,  p-value: 2.258e-10
```

Figure 12: RStudio output for testing interaction terms

, not binary). For example, the average age of the patient in the data set is 60years. This effectively tells us: given the average values of each continuous predictor and if the patient is a male and death was the outcome, we have the following interval.

```
      fit      lwr      upr
1.668301 1.517451 1.819152
```

Figure 13: RStudio output for prediction interval

## 4.2 Binary Logistic Regression

Note: The graphs on the bottom was also used to the MLR section. To construct a binary logistic regression model, we again perform a step-wise regression of all predictors using AIC as our criteria. Following this, we are left with a model that includes time, ejection fraction, serum creatinine, age, and serum sodium as its predictors (see RStudio output Figure 14).

All predictors shown have very low p-values associated with their estimated coefficients. Only the p-value for serum sodium falls outside of the 5%-alpha window, but is still acceptable at 0.09254.

To assess the model's accuracy, predictions are made based on the entire dataset using the predict() function, and those are compared with an actual death event. The result is in the table below (Table 3), with those values in the diagonal being the correct predictions and those in the off-diagonal being incorrect.

| | Death | |
|---|---|---|
| | No | Yes |
| Predict No | 184 | 30 |
| Predict Yes | 19 | 66 |

Table 3: True-False Prediction Table for initial logistic regression

Already the model is quite accurate, with the model correctly predicting a death event (or lack thereof) 83.6% of the time. From here, we assess the residuals to ensure that the model maintains linearity and does not have any error variance issues.

The Residual vs Fit plot for this initial model (Figure 15) shows that there are no apparent linearity or non-constant variance issues, and all residuals fall within 3 and -3, as is desirable for a logistic model. However, there is some slight bunching of residuals for higher index values. In exploration of this, we will analyze the individual predictors with our fit.

It can be seen in Figure 16 that there is a logarithmic shape to the creatinine predictor values versus the log-odds. Otherwise, the remaining predictors follow a linear pattern. To continue, we will apply a logarithmic transformation to creatinine.

```
Call:
glm(formula = death ~ time + severity + Creatinine + age + Sodium,
    family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1590  -0.5888  -0.2281   0.5144   2.7959

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.493034   5.405768   1.756  0.07907 .
time        -0.020895   0.002916  -7.166 7.74e-13 ***
severity    -0.073430   0.015785  -4.652 3.29e-06 ***
Creatinine   0.685990   0.174044   3.941 8.10e-05 ***
age          0.042466   0.015030   2.825  0.00472 **
Sodium      -0.064557   0.038377  -1.682  0.09254 .
---
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: 223.49  on 293  degrees of freedom
AIC: 235.49

Number of Fisher Scoring iterations: 6
```

Figure 14: RStudio output for initial Logistic Regression Model

As shown in Figure 17, the logarithmic transformation of serum creatinine sufficiently linearizes the predictor values vs the model log-odds. Also, as a result of this transformation, the p-value associated with the $\beta$ estimate for serum sodium increased further to 0.1719, far outside of a desired value less than 0.05. Therefore, in search of a parsimonious model with only significant predictors, it is acceptable to remove serum sodium altogether as a predictor.

Our final model is shown in Figure 18 as an output from RStudio. All predictors are now statistically significant, with associated p-values much less than 0.05. The final residual versus fit plot (Figure 19) shows no linearity or residual variance issues. Also, as one final check, a plot of Cook's distance for all observations (Figure 20) reveals no values greater than 0.5, indicating that there are no outliers/high leverage points.

Assessing the final model, we again determine the accuracy as before, using the table below (Table 4).

| | Death | |
|---|---|---|
| | No | Yes |
| Predict No | 182 | 29 |
| Predict Yes | 21 | 67 |

Table 4: True-False Prediction Table for final logistic regression

Little has changed for the accuracy of our model, now correctly predicting death at a rate of 83.3%. In all, this was the goal for our second research question posed. We now have a binary logistic model that can correctly predict death following heart failure based on time, ejection fraction, serum creatinine and age, while also maintaining our initial assumptions on linearity and error variance.

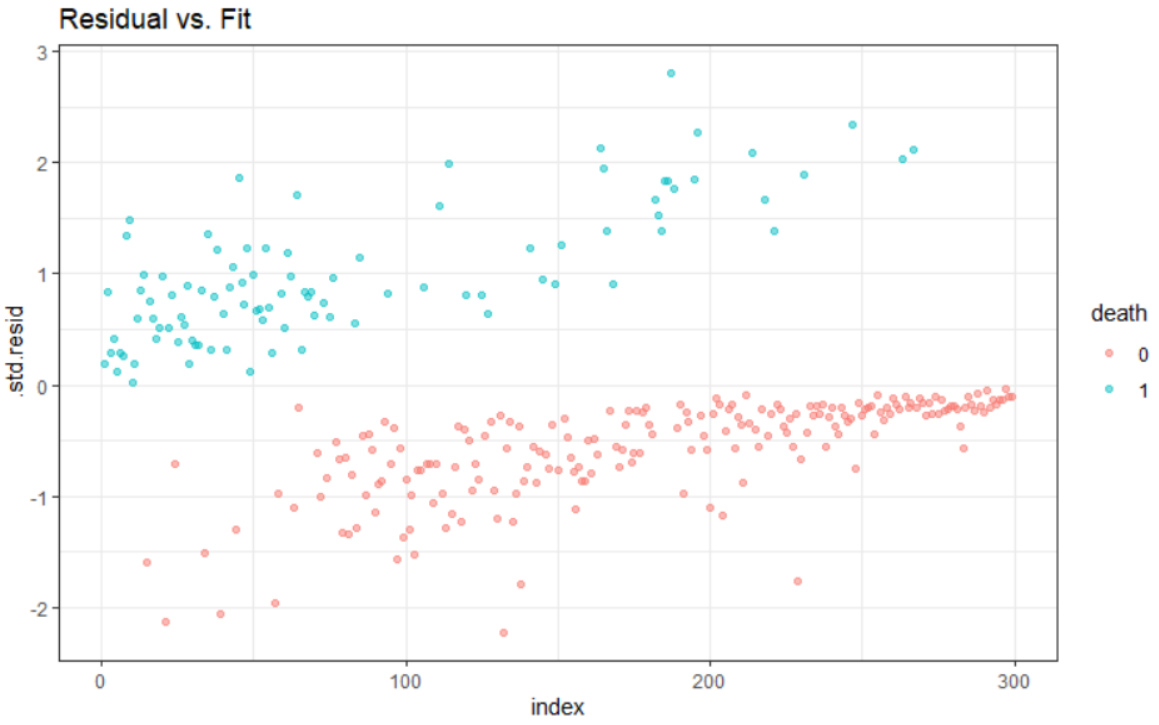



Figure 15: Residual vs. Fit for initial Logistic Regression Model

### 4.2.1 Predictions Using the Logistic Model

With our model, it is now possible to make predictions from new inputs. With all inputs as their mean value, the model provides a prediction of 0.254, indicating that a death would not occur.

To answer our final research question, with all predictor values set as mean value except for ejection fraction set to be one standard deviation lower from the mean (indicating less blood being pumped from the left ventricle), the model provides a prediction of 0.4327, again indicating that a death would not occur. (Note: there are no confidence intervals associated to these predictions as with a binary output, one cannot have a half-death.)

So, even with a statistically severe heart failure, our model predicts that the patient would survive.

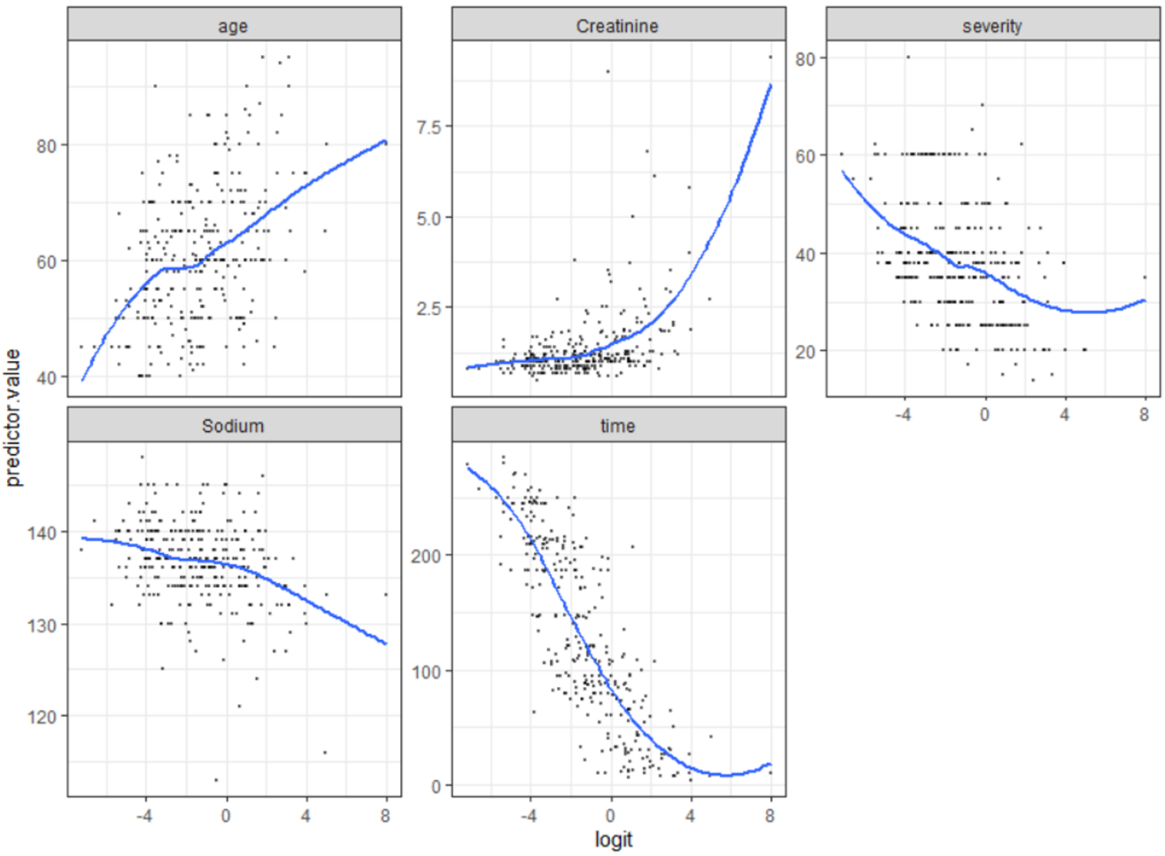



Figure 16: Predictor Values vs. Log-Odds for initial Logistic Regression Model

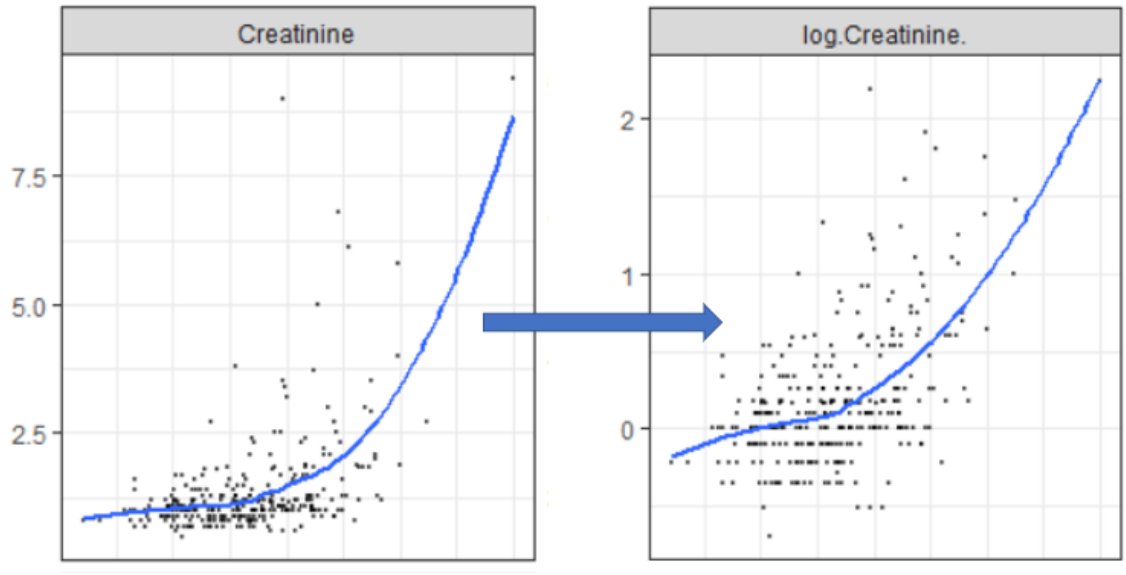



Figure 17: Effect of log-transforming creatinine

```
Call:
glm(formula = death ~ time + severity + log(Creatinine) + age,
    family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2102  -0.5712  -0.2436   0.4956   3.0334

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.340250   1.040497   1.288   0.1977
time            -0.020476   0.002861  -7.157 8.23e-13 ***
severity        -0.068162   0.015427  -4.418 9.94e-06 ***
log(Creatinine)  1.709369   0.397945   4.295 1.74e-05 ***
age              0.037445   0.014983   2.499   0.0125 *
---
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: 223.58  on 294  degrees of freedom
AIC: 233.58

Number of Fisher Scoring iterations: 5
```

Figure 18: RStudio output for final Logistic Regression Model

# 5 Conclusion

While the linear regression model does not greatly explain the variance of severity from heart failure, it did show some significance in various predictors that are representative of the patient's overall health, such as serum sodium levels. Perhaps with more data that documented other aspects of the patient's overall health, the model could more accurately predict the severity of heart failure. In doing so, we could determine changes to diet and lifestyle that could influence and reduce severity, and thereby death of a patient. One factor of interest could be kidney health. While serum creatinine levels shows current kidney function, it is possible that previous renal issues lead to increased risk for severe heart failure. In other words, if we had more information regarding patient's health history and lifestyle, we believe this MLR would be more accurate i.e. variation would be less severe. For the sake of interpretation, 15% of the variation of Severity of the heart failure is explained by the variation of the predictor variables. Ultimately, we were able to answer our research question: does there exist a model such that, for any predictor variables in the data set, are we able to make a model. The drawback is that, despite the Box-Cox method and Log Transformations, the variation is severe and this model will not be used to predict any values. The positive is that we do have normality conditions met by the model. Finally, we also have randomness of the errors. i.e. constant variation is met.

The logistic regression model provides an accurate prediction of death of a patient. Based on a patient's age, how well their blood is pumping (ejection fraction), how well their kidneys are performing (levels of serum creatinine), and the number of days for their follow-up period, we can predict their survival. While macabre, this information may be useful for a doctor who needs to frankly discuss a patient's situation with them. It would be advisable to show doctors this study so that they can make the best decisions in care for their heart failure patients.

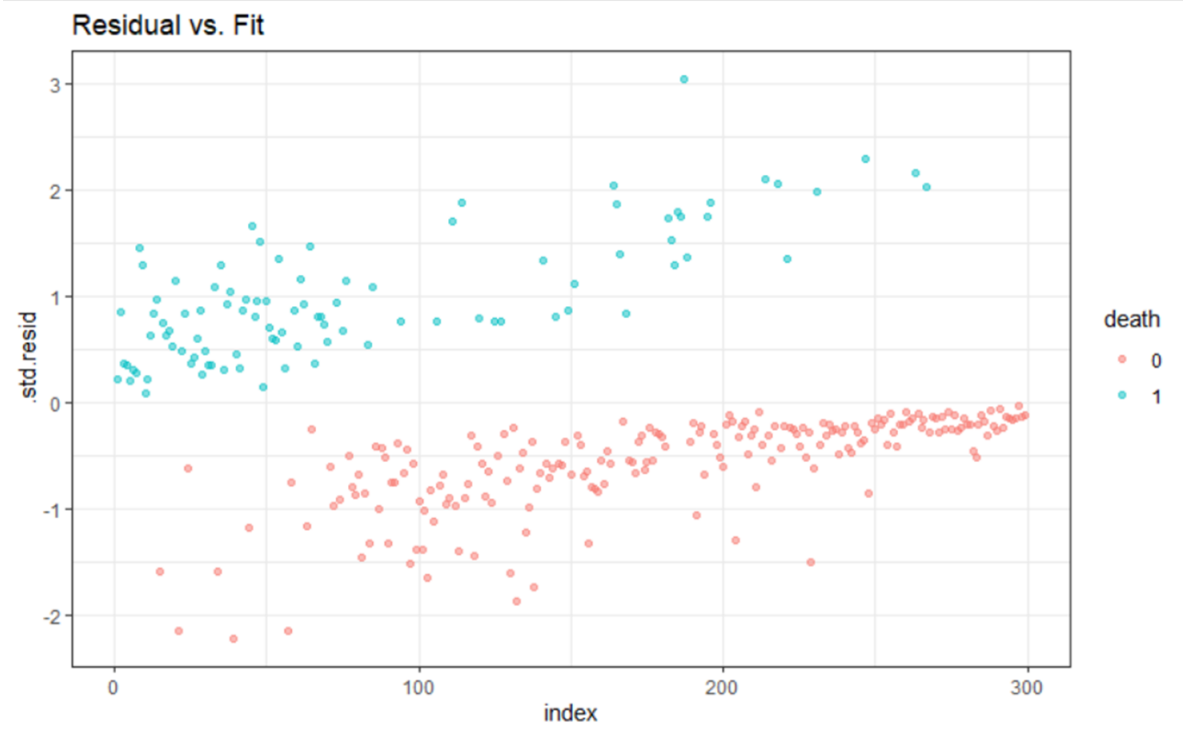



Figure 19: Residual vs. Fit for final Logistic Regression Model

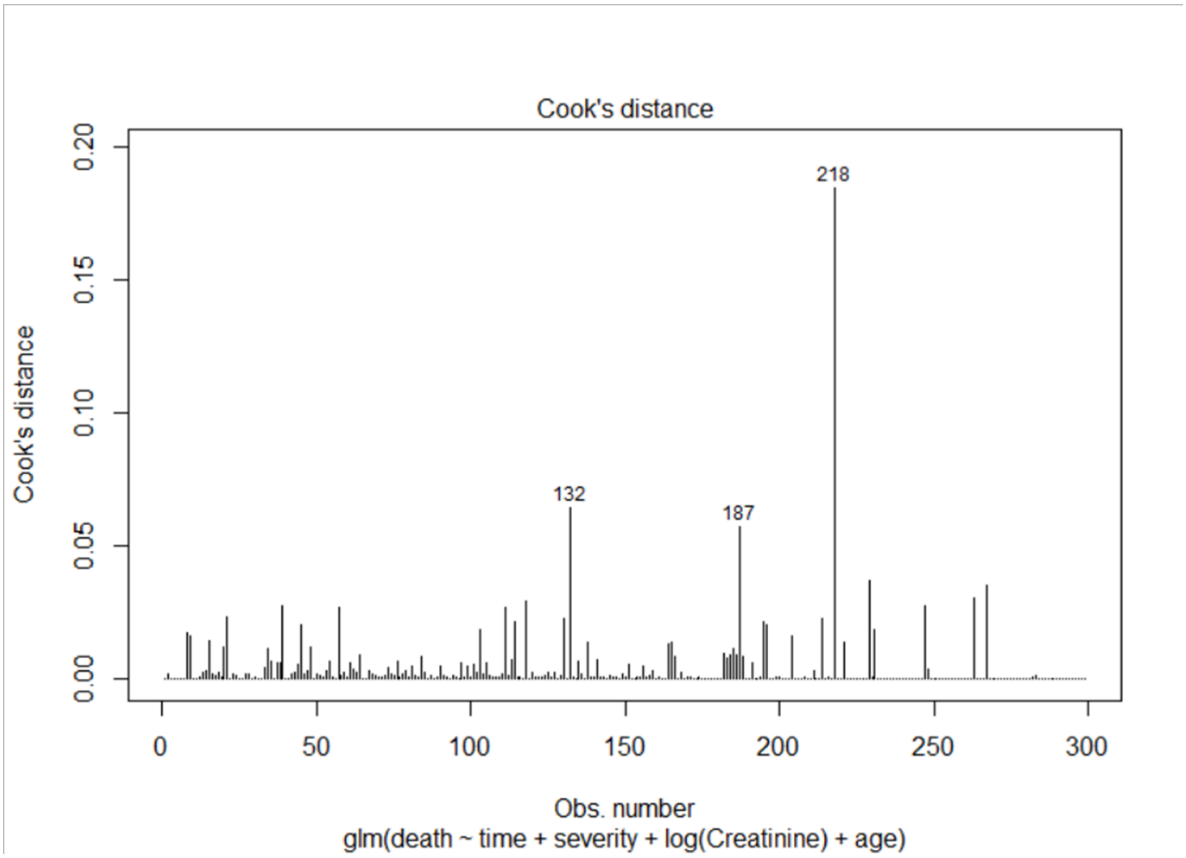



Figure 20: Cook's Distance for final Logistic Regression Model