IDENTIFICATION OF KEY PREDICTIVE FACTORS IMPACTING MORTALITY IN HEART FAILURE (HF) PATIENTS

Chibuzo Oha

2024-11-13

```
mimicIII_data1= read.csv("C:\\Users\\CEO\\Documents\\DATA 603\\Project\\mimicIII_data1.csv")
head(mimicIII_data1, 5)
```

```
##   age_group gender_txt group     ID outcome age gender      BMI hypertensive
## 1     18-29          F     1 197509       0  25      1 57.92653            0
## 2     18-29          F     1 183860       0  19      1 29.78897            0
## 3     18-29          F     1 165367       1  28      1 80.00000            0
## 4     30-39          F     1 184941       0  39      1 29.35654            1
## 5     30-39          F     1 195748       0  38      1 27.65766            0
##   atrialfibrillation CHD_with_no_MI diabetes deficiencyanemias depression
## 1                  0              0        0                  0          0
## 2                  0              0        0                  0          0
## 3                  0              0        0                  0          0
## 4                  0              0        0                  0          0
## 5                  0              0        0                  0          0
##   Hyperlipemia Renal_failure COPD heart_rate Systolic_blood_pressure
## 1            0             0    0   90.36000                95.04545
## 2            0             0    0   90.65217                99.69565
## 3            0             0    0   87.65517               111.35294
## 4            1             0    0   63.72000               133.66667
## 5            1             0    0  110.84000               109.20833
##   Diastolic_blood_pressure Respiratory_rate temperature    SP_O2 Urine_output
## 1                 58.00000         24.72000    36.18889 96.12000         4600
## 2                 50.08696         21.69565    38.16667 96.82609         5050
## 3                 58.76471         23.45946    37.42593 91.09375         4775
## 4                 80.37500         17.27273    36.96032 95.68000         3605
## 5                 79.29167         26.00000    36.77778 99.47826         1440
##   hematocrit      RBC      MCH     MCHC      MCV      RDW Leucocyte Platelets
## 1   46.43333 5.847778 27.22222 34.24444 79.66667 16.98889  15.27778  347.4444
## 2   36.38000 4.256000 30.46000 35.60000 85.60000 12.58000  11.26000  260.2000
## 3   43.08750 4.758750 28.46250 31.46250 90.62500 14.98750   9.38750  265.7500
## 4   44.25000 5.292500 25.93750 31.02500 83.62500 14.53750   9.40000  212.0000
## 5   41.47778 4.237778 31.84444 32.52222 98.00000 15.44444  11.15556  206.2222
##   Neutrophils Basophils Lymphocyte       PT      INR NT_proBNP Creatine_kinase
## 1      78.650      0.75     14.550 17.88571 1.614286      2982           556.5
## 2      79.500      0.20     15.100 16.65397 1.490476      1968           556.5
## 3      78.700      0.20     14.000 15.42222 1.366667      2420           556.5
## 4      92.200      0.10      4.400 12.83333 1.100000      3143           142.0
```

```
## 5      80.275      0.90      9.925 19.39444 1.755556      8474           118.0
##   Creatinine Urea_nitrogen  glucose Blood_potassium Blood_sodium Blood_calcium
## 1  1.3111111      31.88889 248.4000        3.666667     131.3333      9.170588
## 2  0.8600000       8.40000  84.0000        4.200000     139.4000      8.300000
## 3  0.7545455      11.63636 113.8750        3.757143     142.4545      8.687500
## 4  1.1200000      26.50000 120.3333        4.470000     138.5000      9.066667
## 5  1.2076923      23.38462  97.5000        4.100000     135.0769      8.154545
##    Chloride Anion_gap Magnesium_ion       PH Bicarbonate Lactic_acid     PCO2
## 1  89.42857  15.33333      2.184211 7.311250    30.05556    2.800000 65.16667
## 2 103.20000  14.80000      1.950000 7.290000    25.60000    1.500000 52.00000
## 3  95.45455  11.27273      2.377778 7.332500    39.63636    1.083333 78.33333
## 4  93.60000  11.60000      2.175000 7.441000    37.80000    2.350000 60.50000
## 5  98.38462  16.61538      1.950000 7.418333    24.23077    3.512500 31.66667
##   EF outcome_txt
## 1 20       Alive
## 2 55       Alive
## 3 50        Dead
## 4 55       Alive
## 5 15       Alive
```

```
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
```

```
# Select relevant columns
selected_data <- mimicIII_data1 %>%
  select(Creatinine, age, gender, BMI, Systolic_blood_pressure, Diastolic_blood_pressure, hypertensive,

# Convert categorical variables (like gender, hypertensive, diabetes, and Renal_failure) to numeric if
selected_data$gender <- as.numeric(factor(selected_data$gender))
selected_data$hypertensive <- as.numeric(factor(selected_data$hypertensive))
selected_data$diabetes <- as.numeric(factor(selected_data$diabetes))
selected_data$Renal_failure <- as.numeric(factor(selected_data$Renal_failure))

# Calculate correlation matrix
correlation_matrix <- cor(selected_data, use = "complete.obs")

# Extract correlation of creatinine with other variables
creatinine_correlation <- correlation_matrix["Creatinine", ]

# Display the result
print(creatinine_correlation)
```

```
##               Creatinine                      age                   gender
##               1.00000000              -0.08637138              -0.13092888
##                      BMI  Systolic_blood_pressure Diastolic_blood_pressure
##               0.01722883                0.07951695                0.00566448
##             hypertensive                  diabetes            Renal_failure
##               0.07520727                0.12865092                0.44895233
```

```
summary(mimicIII_data1[c("age", "BMI", "Systolic_blood_pressure", "Diastolic_blood_pressure", "hyperten
```

```
##       age              BMI        Systolic_blood_pressure
##  Min.   :19.00   Min.   :13.35   Min.   : 75.0
##  1st Qu.:65.00   1st Qu.:25.26   1st Qu.:105.5
##  Median :77.00   Median :28.29   Median :116.0
##  Mean   :74.09   Mean   :29.82   Mean   :117.9
##  3rd Qu.:85.00   3rd Qu.:32.55   3rd Qu.:128.5
##  Max.   :99.00   Max.   :80.00   Max.   :203.0
##  Diastolic_blood_pressure  hypertensive       diabetes      Renal_failure
##  Min.   : 24.74           Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 52.28           1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 58.36           Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   : 59.50           Mean   :0.7183   Mean   :0.4213   Mean   :0.3651
##  3rd Qu.: 65.38           3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :107.00           Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
# Define the columns you want to compute quantiles for
columns <- c("BMI", "Systolic_blood_pressure", "Diastolic_blood_pressure"
             )

# Apply the quantile function to each column
quantile_results <- sapply(columns, function(col) {
  quantile(mimicIII_data1[[col]], na.rm = TRUE)
})

quantile_results
```

```
##           BMI Systolic_blood_pressure Diastolic_blood_pressure
## 0%   13.34680                 75.0000                 24.73684
## 25%  25.25823                105.4808                 52.28407
## 50%  28.28862                116.0000                 58.36364
## 75%  32.54872                128.4929                 65.38455
## 100% 80.00000                203.0000                107.00000
```

‘

```
#install.packages("pheatmap")
```

```
library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 4.4.2
```

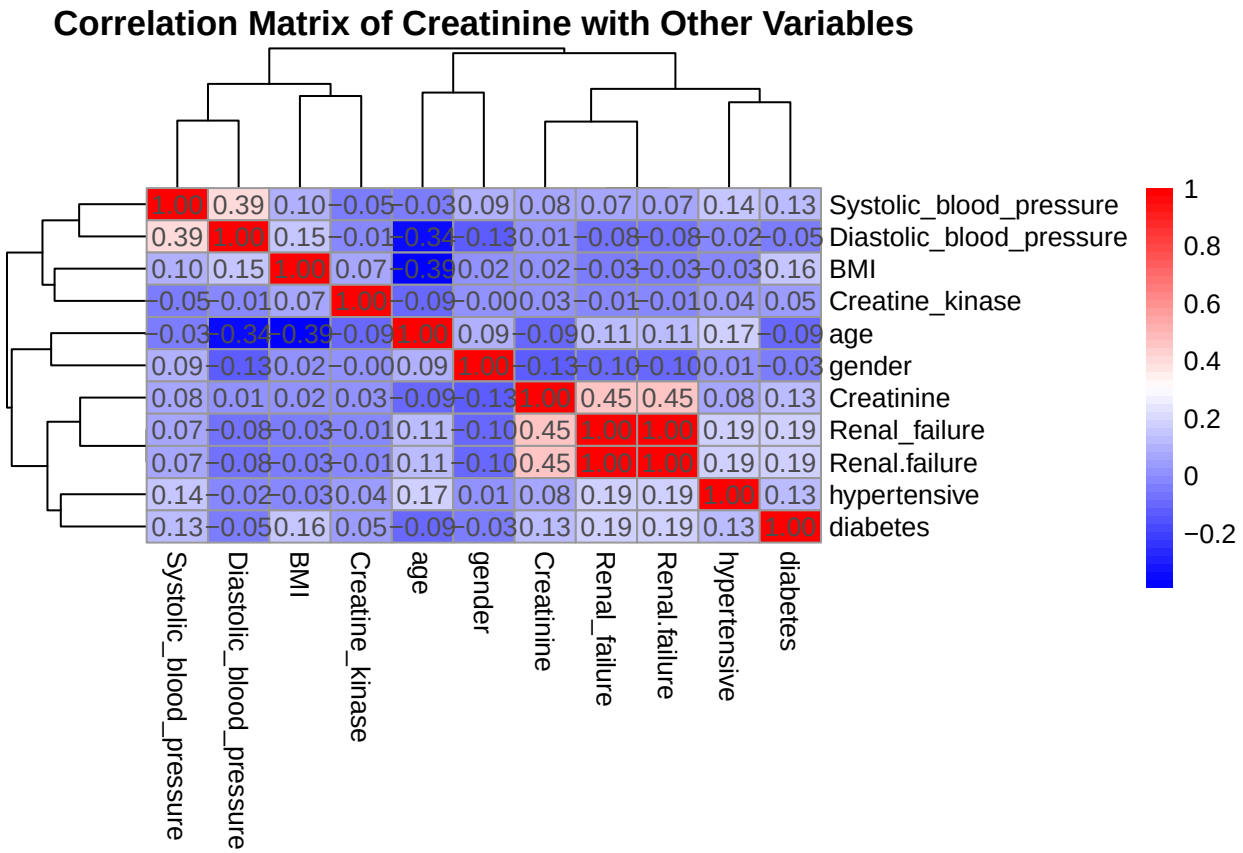
```r
# Load necessary library
library(dplyr)

# Select relevant columns (make sure the column names match your data)
selected_data <- mimicIII_data1 %>%
  select(Creatinine, age, gender, BMI, Systolic_blood_pressure, Diastolic_blood_pressure, hypertensive,

# Convert categorical variables to numeric if necessary
selected_data$gender <- as.numeric(factor(selected_data$gender))
selected_data$hypertensive <- as.numeric(factor(selected_data$hypertensive))
selected_data$diabetes <- as.numeric(factor(selected_data$diabetes))
selected_data$Renal.failure <- as.numeric(factor(selected_data$Renal_failure))

# Calculate the correlation matrix
correlation_matrix <- cor(selected_data, use = "complete.obs")

# Generate the heatmap with annotations
pheatmap(correlation_matrix,
         display_numbers = TRUE,      # Annotate with correlation values
         color = colorRampPalette(c("blue", "white", "red"))(50), # Color gradient
         main = "Correlation Matrix of Creatinine with Other Variables",
         fontsize_number = 10)        # Adjust the font size of annotations
```

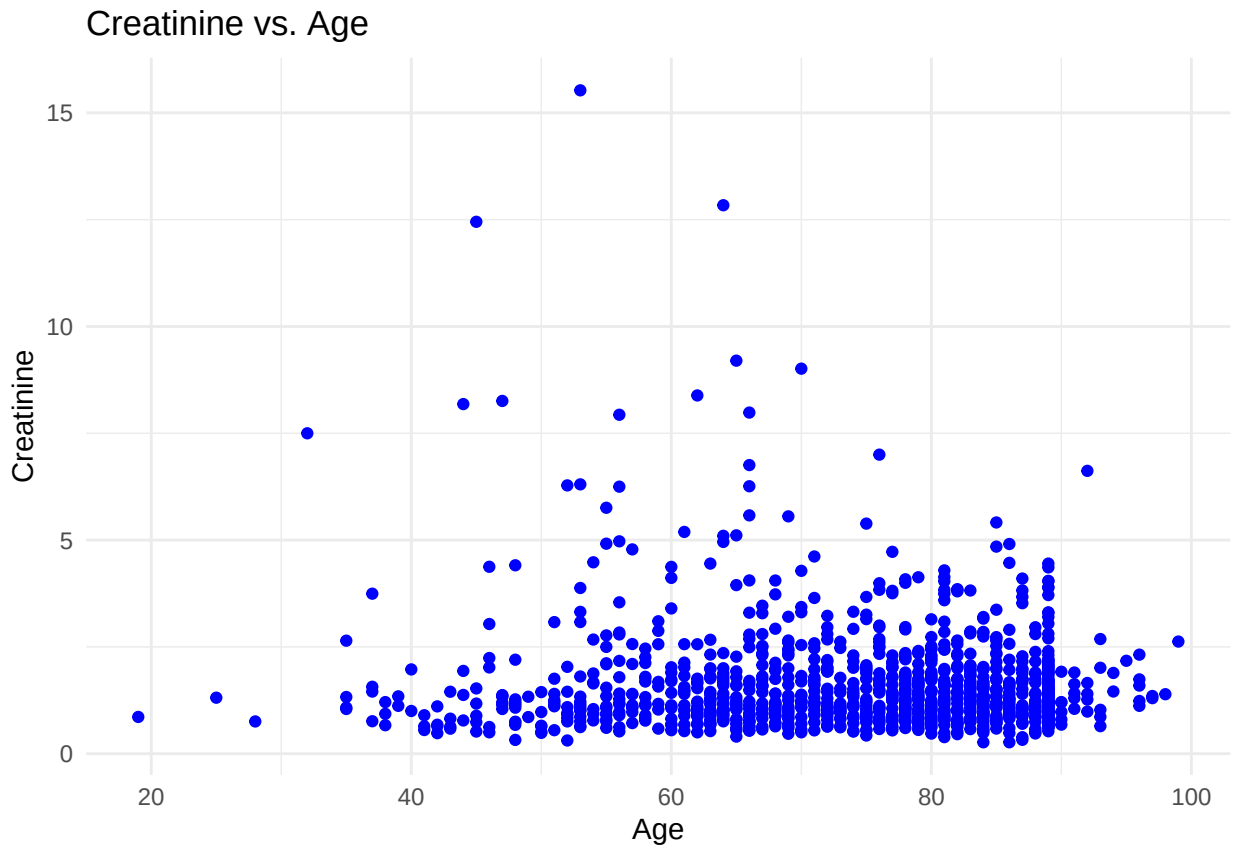



```r
fullModel1= lm(Creatinine~age+gender+BMI+Systolic_blood_pressure+Diastolic_blood_pressure+ hypertensive
summary(fullModel1)
```

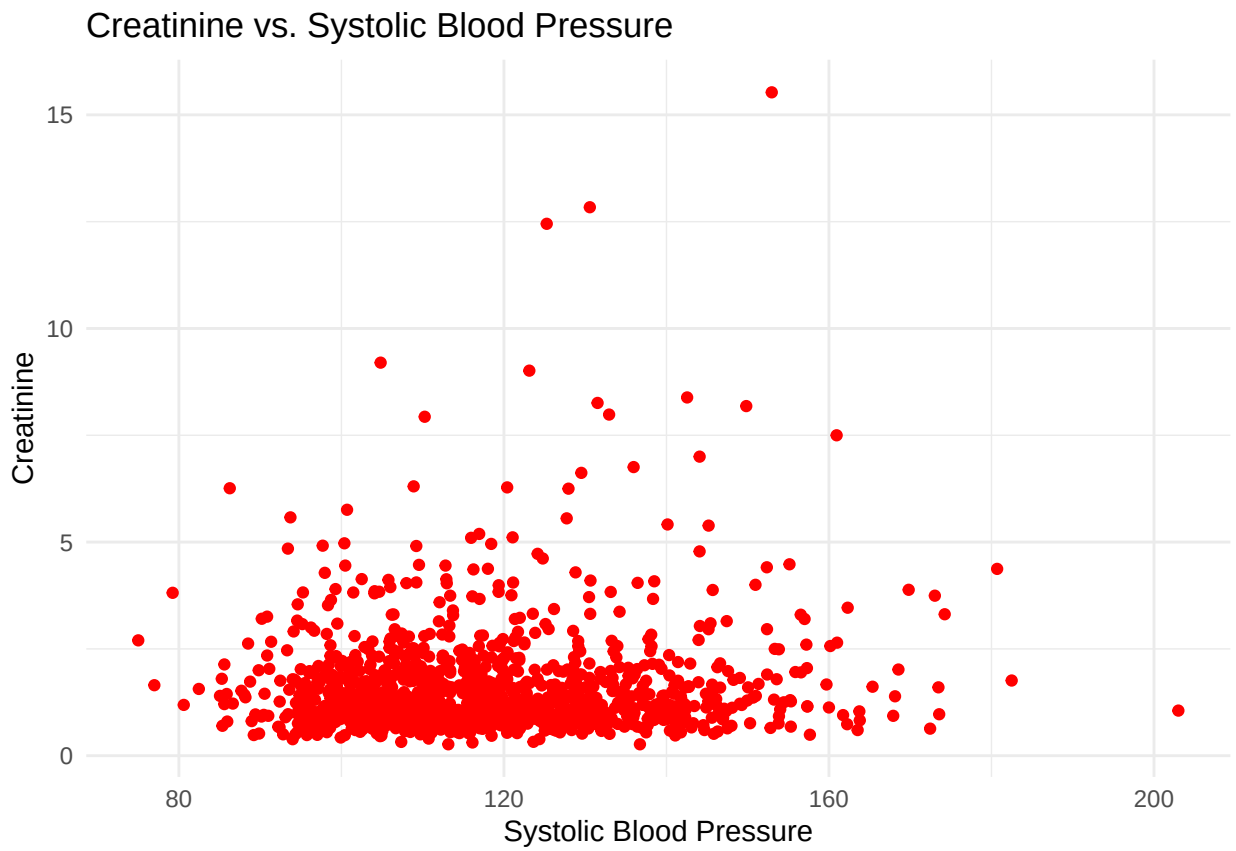
```
##
## Call:
## lm(formula = Creatinine ~ age + gender + BMI + Systolic_blood_pressure +
##     Diastolic_blood_pressure + hypertensive + diabetes + Renal_failure,
##     data = mimicIII_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9055 -0.5792 -0.2166  0.2360 13.9176
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.411907   0.407628   5.917 4.31e-09 ***
## age                       -0.014213   0.002906  -4.890 1.15e-06 ***
## gender                    -0.216546   0.067846  -3.192  0.00145 **
## BMI                       -0.004286   0.004369  -0.981  0.32686
## Systolic_blood_pressure    0.004739   0.002168   2.186  0.02899 *
## Diastolic_blood_pressure -0.004678   0.003712  -1.260  0.20785
## hypertensive               0.008698   0.076412   0.114  0.90939
## diabetes                   0.058274   0.070304   0.829  0.40734
## Renal_failure              1.176358   0.071520  16.448  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.125 on 1166 degrees of freedom
## Multiple R-squared:  0.2305, Adjusted R-squared:  0.2252
## F-statistic: 43.65 on 8 and 1166 DF,  p-value: < 2.2e-16
```

```r
library(ggplot2)

# Scatter Plot
ggplot(mimicIII_data1, aes(x = age, y = Creatinine)) +
  geom_point(color = "blue") +
  labs(title = "Creatinine vs. Age", x = "Age", y = "Creatinine") +
  theme_minimal()
```




```
ggplot(mimicIII_data1, aes(x = Systolic_blood_pressure, y = Creatinine)) +
  geom_point(color = "red") +
  labs(title = "Creatinine vs. Systolic Blood Pressure",
       x = "Systolic Blood Pressure", y = "Creatinine") +
  theme_minimal()
```




```
reducedModel= lm(Creatinine~age+gender+Systolic_blood_pressure+Renal_failure, data= mimicIII_data1)
summary(reducedModel)
```

```
##
## Call:
## lm(formula = Creatinine ~ age + gender + Systolic_blood_pressure +
##     Renal_failure, data = mimicIII_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7741 -0.5761 -0.2302  0.2479 14.0357
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.993144   0.302217   6.595 6.42e-11 ***
## age                      -0.012209   0.002489  -4.905 1.07e-06 ***
## gender                   -0.205994   0.066649  -3.091  0.00204 **
## Systolic_blood_pressure   0.003645   0.001916   1.902  0.05745 .
## Renal_failure             1.198152   0.069228  17.307  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.125 on 1170 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2256
## F-statistic: 86.48 on 4 and 1170 DF,  p-value: < 2.2e-16
```

Null Hypothesis(H_0): The predictor variables removed has no significant effect on the affects Creatinine.

Alternative Hypothesis(H_a): The predictor variables removed has a significant effect on the affects Creatinine.

$$H_0 : \beta_k = 0$$
$$H_a : \beta_k \neq 0 \quad (k = 1, 2, \ldots, p)$$

```
anova(reducedModel,fullModel1)
```

```
## Analysis of Variance Table
##
## Model 1: Creatinine ~ age + gender + Systolic_blood_pressure + Renal_failure
## Model 2: Creatinine ~ age + gender + BMI + Systolic_blood_pressure + Diastolic_blood_pressure +
##     hypertensive + diabetes + Renal_failure
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1170 1480.1
## 2   1166 1475.7  4    4.3443 0.8581 0.4885
```

At p-value > 0.05, we fail to reject the null hypothesis;

```
reducedModel1= lm(Creatinine~age+gender+Renal_failure, data= mimicIII_data1)
summary(reducedModel1)
```

```
##
## Call:
## lm(formula = Creatinine ~ age + gender + Renal_failure, data = mimicIII_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7647 -0.5646 -0.2239  0.2250 14.1569
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.416941   0.204370  11.826  < 2e-16 ***
## age           -0.012438   0.002489  -4.997 6.72e-07 ***
## gender        -0.193681   0.066408  -2.917  0.00361 **
## Renal_failure  1.209703   0.069038  17.522  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 1171 degrees of freedom
## Multiple R-squared:  0.2258, Adjusted R-squared:  0.2238
## F-statistic: 113.8 on 3 and 1171 DF,  p-value: < 2.2e-16
```

We choose reducedModel due to reducedModel having higher $R^2_{adj} = 0.2256$ and lower RSE 1.125 compared to reducedModel1

```
InteractReducedModel= lm(Creatinine~(age+gender+Systolic_blood_pressure+Renal_failure)^2, data= mimicII
summary(InteractReducedModel)
```

```
##
## Call:
## lm(formula = Creatinine ~ (age + gender + Systolic_blood_pressure +
##     Renal_failure)^2, data = mimicIII_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1460 -0.5090 -0.2148  0.2194 14.0328
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -0.5202731  1.3788044  -0.377  0.70599
## age                                 0.0430143  0.0169161   2.543  0.01113
## gender                             -0.9130136  0.5874129  -1.554  0.12039
## Systolic_blood_pressure             0.0175249  0.0110161   1.591  0.11191
## Renal_failure                       3.5532837  0.6584807   5.396 8.24e-08
## age:gender                         -0.0003066  0.0048859  -0.063  0.94998
## age:Systolic_blood_pressure        -0.0003662  0.0001336  -2.740  0.00623
## age:Renal_failure                  -0.0386148  0.0055270  -6.987 4.73e-12
## gender:Systolic_blood_pressure      0.0061507  0.0038091   1.615  0.10664
## gender:Renal_failure               -0.0407685  0.1364415  -0.299  0.76515
## Systolic_blood_pressure:Renal_failure  0.0049952  0.0038601   1.294  0.19590
##
## (Intercept)
## age                                *
## gender
## Systolic_blood_pressure
## Renal_failure                      ***
## age:gender
## age:Systolic_blood_pressure        **
## age:Renal_failure                  ***
## gender:Systolic_blood_pressure
## gender:Renal_failure
## Systolic_blood_pressure:Renal_failure
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.094 on 1164 degrees of freedom
## Multiple R-squared:  0.2737, Adjusted R-squared:  0.2675
## F-statistic: 43.87 on 10 and 1164 DF,  p-value: < 2.2e-16
```

```
InteractModelBest= lm(Creatinine~age+Systolic_blood_pressure+Renal_failure+age*Systolic_blood_pressure+a
summary(InteractModelBest)
```

```
##
## Call:
## lm(formula = Creatinine ~ age + Systolic_blood_pressure + Renal_failure +
##     age * Systolic_blood_pressure + age * Renal_failure, data = mimicIII_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1149 -0.4931 -0.2158  0.1975 13.9799
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.8227156  1.1518748  -1.582  0.11383
## age                          0.0386174  0.0155804   2.479  0.01333 *
## Systolic_blood_pressure      0.0267917  0.0097852   2.738  0.00628 **
## Renal_failure                4.0849817  0.4183529   9.764  < 2e-16 ***
## age:Systolic_blood_pressure -0.0003423  0.0001323  -2.586  0.00982 **
## age:Renal_failure           -0.0381749  0.0054792  -6.967 5.39e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.099 on 1169 degrees of freedom
## Multiple R-squared:  0.264,  Adjusted R-squared:  0.2609
## F-statistic: 83.86 on 5 and 1169 DF,  p-value: < 2.2e-16
```

```
#summary(reducedModel)
```

Null Hypothesis ((H_0)): The reduced model is a best model

Alternative Hypothesis ((H_a)): The reduced model is not a best model

```
reducedModel1= lm(Creatinine~age+gender+Renal_failure, data= mimicIII_data1)
InteractModelBest= lm(Creatinine~age+Systolic_blood_pressure+Renal_failure+age*Systolic_blood_pressure+a
anova(InteractModelBest, reducedModel )
```

```
## Analysis of Variance Table
##
## Model 1: Creatinine ~ age + Systolic_blood_pressure + Renal_failure +
##     age * Systolic_blood_pressure + age * Renal_failure
## Model 2: Creatinine ~ age + gender + Systolic_blood_pressure + Renal_failure
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1169 1411.4
## 2   1170 1480.1 -1   -68.668 56.874 9.296e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The InteractModelBest model has higher $R^2_{adj} = 0.2609$ and lower RSE of 1.099 meaning that InteractModelBest model is better whe compared to the Reduced Model with $R^2_{adj} = 2256$ and RSE =1125

From the With a p-value of 9.296e-14< 0.05 significance, we reject the null hypothesis and accept the alternate hypothesis meaning that InteractModelBest model is the best fit model to be used to prediction

The best fit model

$$\hat{Creatinine} = -1.8227156 + 0.0386174age + 0.0267917Systolic_blood_pressure + 4.0849817Renal_failure - 0.0003423age * Systolic_blood_pressure - 0.0381749age * Renal_failure$$

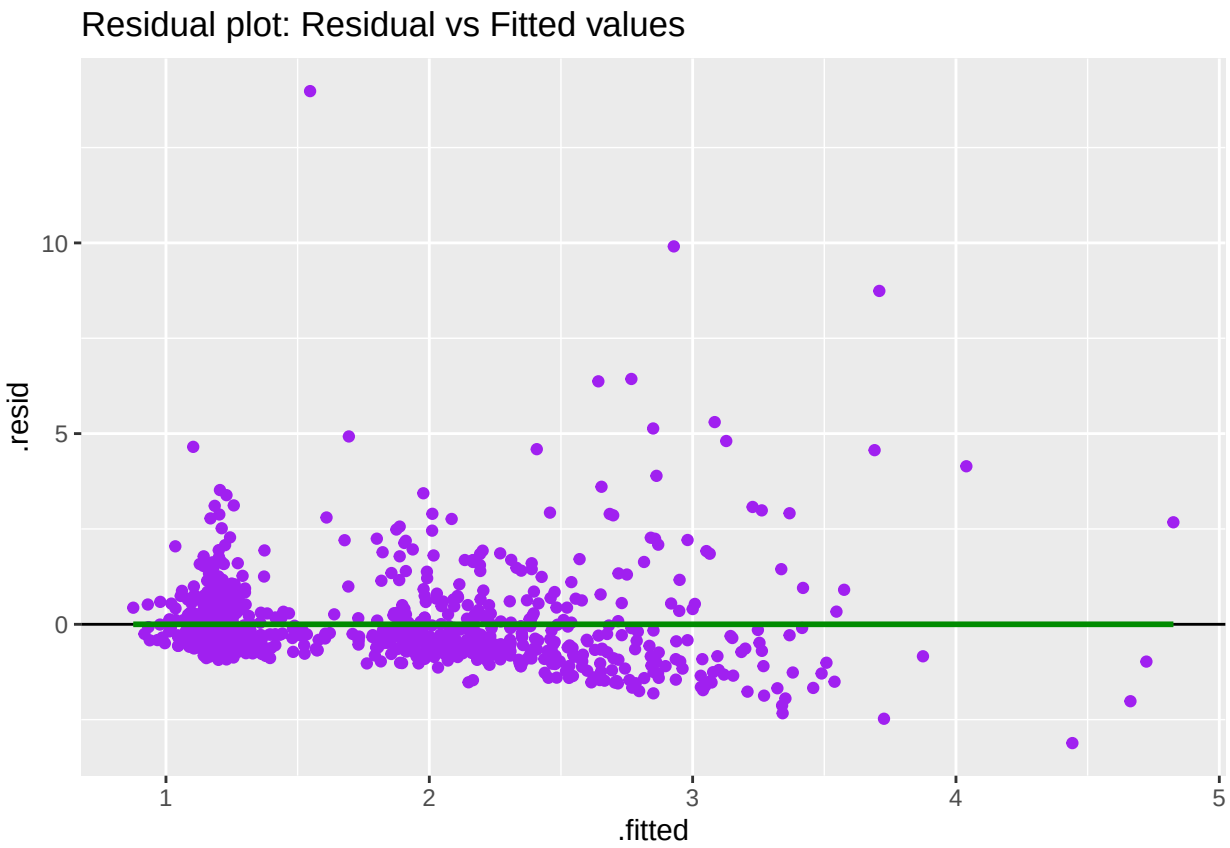
```
newData1 = data.frame(age =72 ,Systolic_blood_pressure=155.8667, Renal_failure= 1 )
predict(InteractModelBest, newData1,interval = "predict")
```

```
##     fit       lwr      upr
## 1 2.629 0.4664461 4.791554
```

The response value is between the lower and upper limit of the predicted value, which signifies that the model predicted correctly and can be used for further prediction Cretinine in indiviuals.

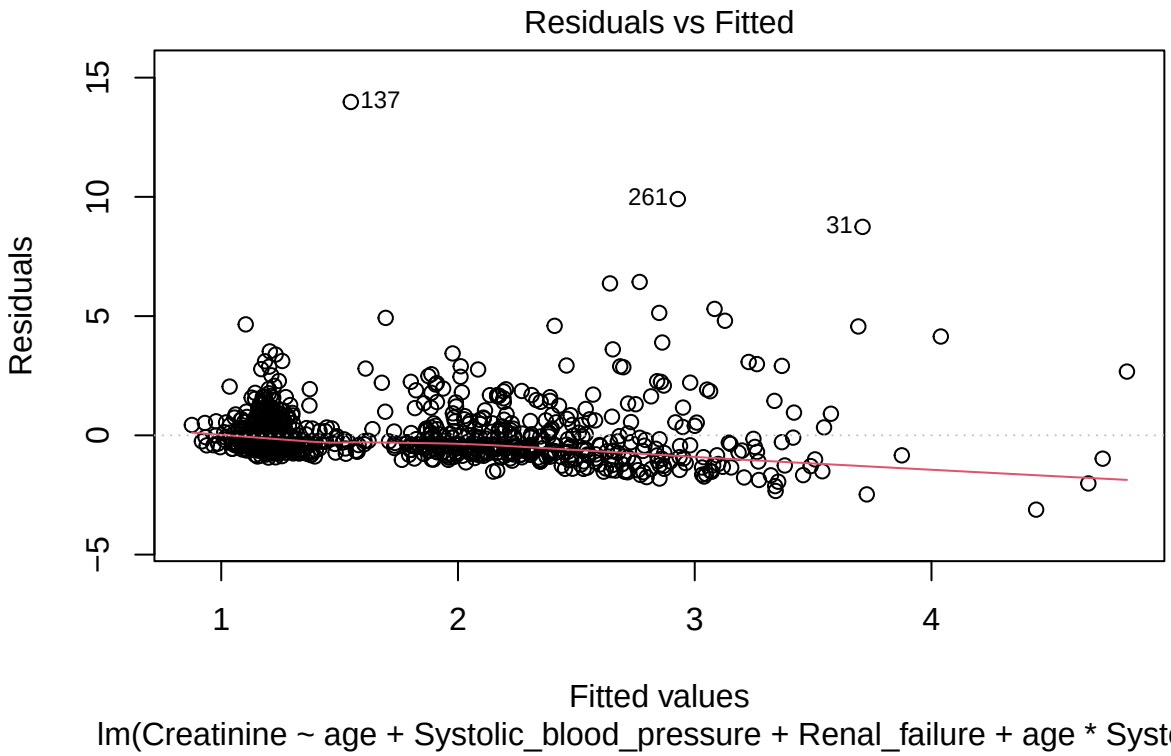
```
library(ggplot2)
ggplot(InteractModelBest, aes(x=.fitted, y=.resid)) +
  geom_point(colour = "purple") +
  geom_hline(yintercept = 0) +
  geom_smooth(colour = "green4")+
  ggtitle("Residual plot: Residual vs Fitted values")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



This is a perfect model as there is no patterns, the plot shows no funneling, which means there is no problem with linearity assumption.

```
plot(InteractModelBest, which=1)
```



This is a perfect model as there is no patterns, the plot shows no funneling, which means there is no problem with linearity assumption.

Test for heteroscedasticity (non constant variance)

H_0: heteroscedasticity is not present (homoscedasticity) H_a: heteroscedasticity is present

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.4.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.4.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(InteractModelBest)
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  InteractModelBest
## BP = 49.282, df = 5, p-value = 1.943e-09
```

We reject the null hypothesis (p-value < 0.05), so we conclude we do have heteroscedasticity

```
library(mctest)
nteractModelBest= lm(Creatinine~age+Systolic_blood_pressure+Renal_failure+age*Systolic_blood_pressure+ag
imcdiag(InteractModelBest, method="VIF")
```

```
##
## Call:
## imcdiag(mod = InteractModelBest, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                                 VIF detection
## age                           42.0464         1
## Systolic_blood_pressure       27.7674         1
## Renal_failure                 39.4828         1
## age:Systolic_blood_pressure 68.0319         1
## age:Renal_failure             40.6422         1
##
## Multicollinearity may be due to age Systolic_blood_pressure Renal_failure age:Systolic_blood_pressur
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

check for normality test

H_0: We have normality H_a: We do not have normality

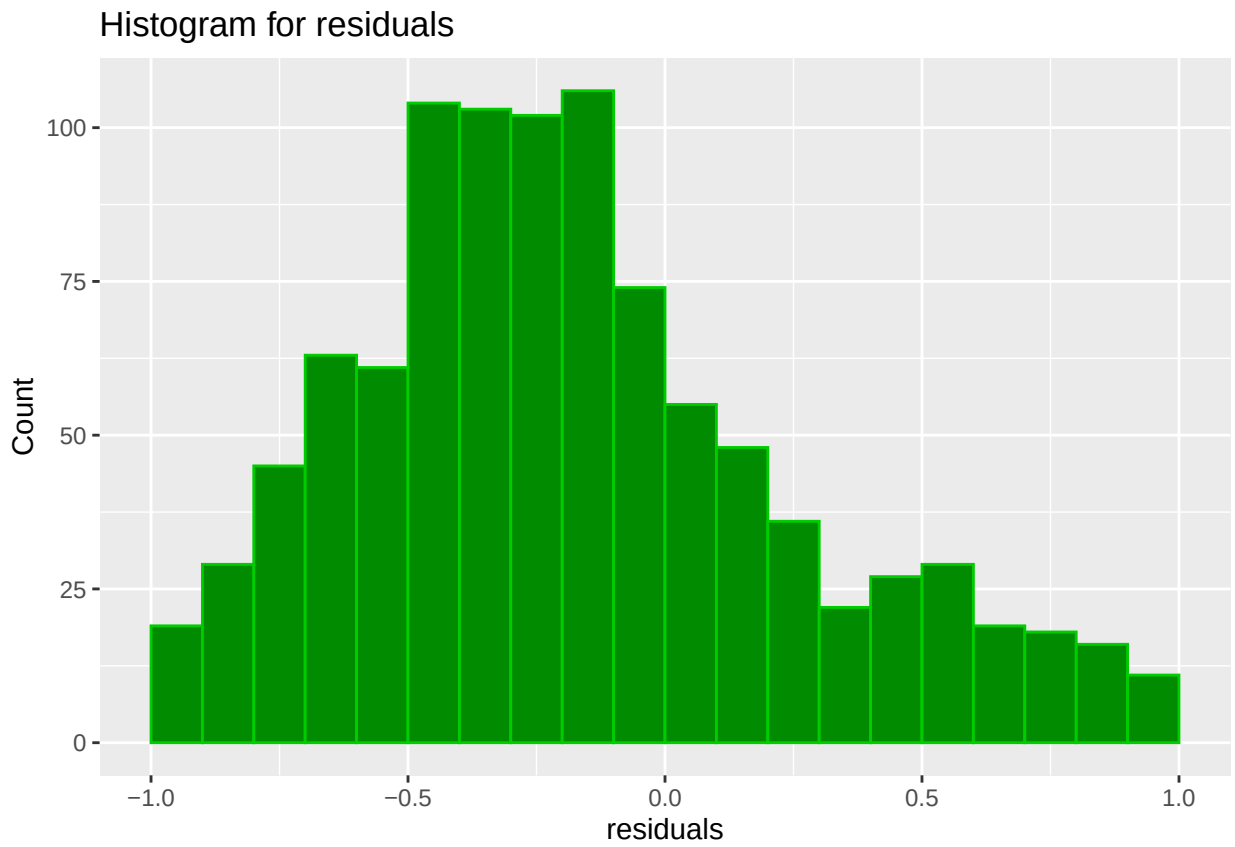
```
shapiro.test(residuals(InteractModelBest))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(InteractModelBest)
## W = 0.70846, p-value < 2.2e-16
```

Shapiro-Wilk normality test confirms that the residuals are not normally distributed as the p- value= 2.2e-16 <0.05 From the Shapiro-Wilk analysis with p-value= 2.2e-16 <0.05 we reject the null hypothesis

H_0 : Sample data are significantly normally distributed H_a : Sample data are not significantly normally distributed

```
ggplot(data=mimicIII_data1, aes(residuals(InteractModelBest))) +
geom_histogram(breaks = seq(-1,1,by=0.1), col="green3", fill="green4") +
labs(title="Histogram for residuals") +
labs(x="residuals", y="Count")
```



p-value =2.2e-16 $<$ o.05, We reject the null hypothesis, therefore the Sample data are not significantly normally distributed