Sepsis Logistic Regression + Best Model

Raehash Shah

2023-12-19

Overview of this file

This R markdown file uses the two datasets to identify the coefficients that are most significant in classifying if a patient has Sepsis or not. We will use this regression techniques to identify what features we will preserve when trying to create a regression for number of Sepsis patients in regions of the U.S. Consider this to be stage 1.

Load Data

```
sepsis_df <- read.csv("Data/Disease/Sepsis_Dis_indicator/patient_risk_profiles_trim.csv")
sepsis_df = subset(sepsis_df, select = -c(X))
```

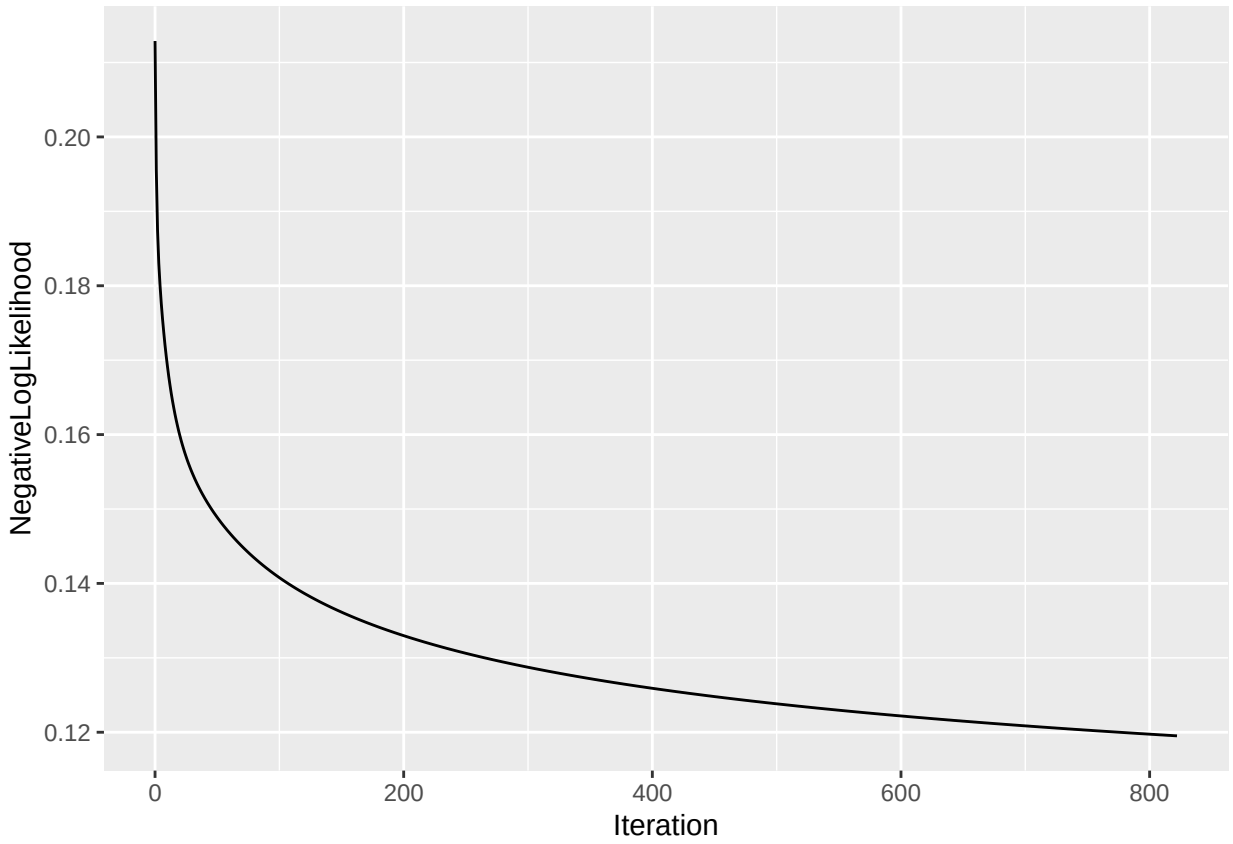
Visualize Data in 2 Dimension for Separation

Here is an exponential PCA estimation of Sepsis

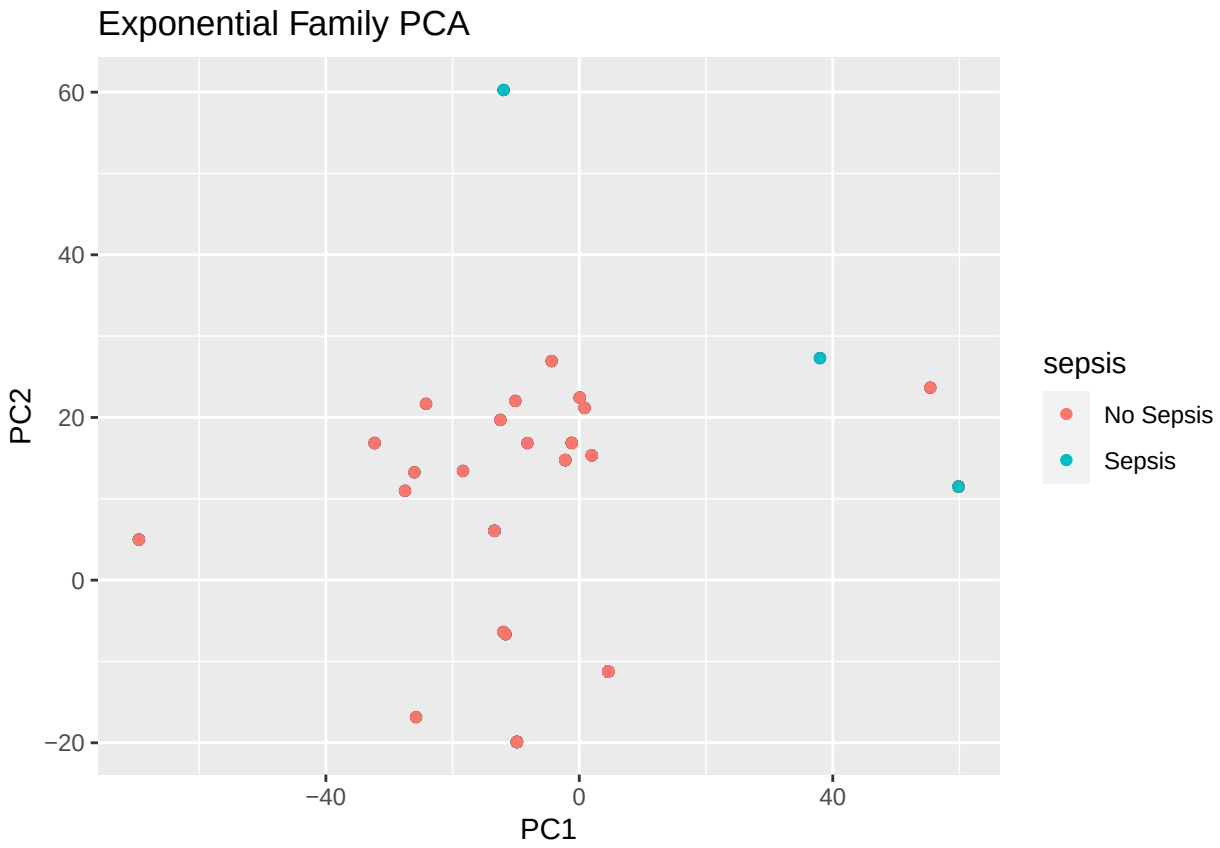
```
logsvd_model = logisticSVD(sepsis_df_sub, k = 2)
```

```
## rARPACK must be installed to use partial_decomp
```

```
plot(logsvd_model, type = "trace")
```




```
plot(logsvd_model, type = "scores") + geom_point(aes(colour = sepsis)) +
  ggtitle("Exponential Family PCA")
```

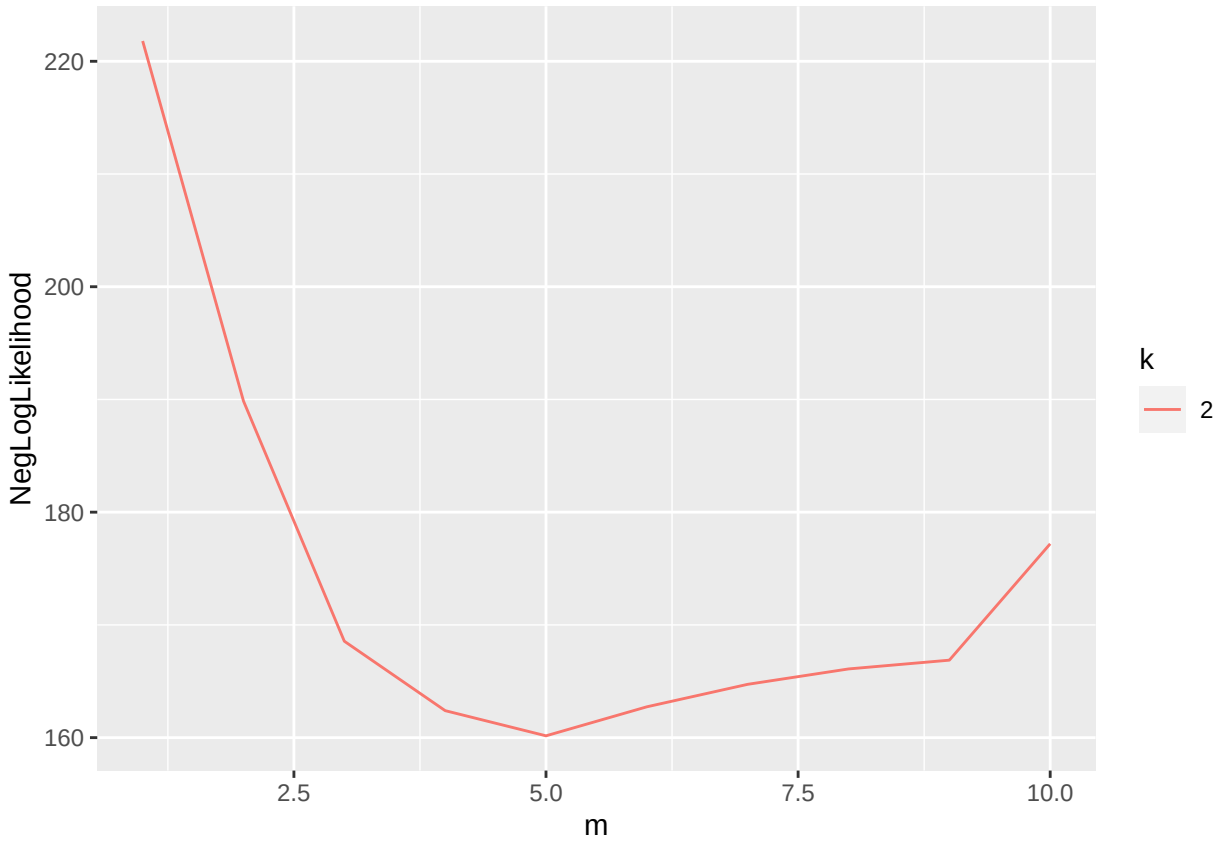


Here is a logistic PCA

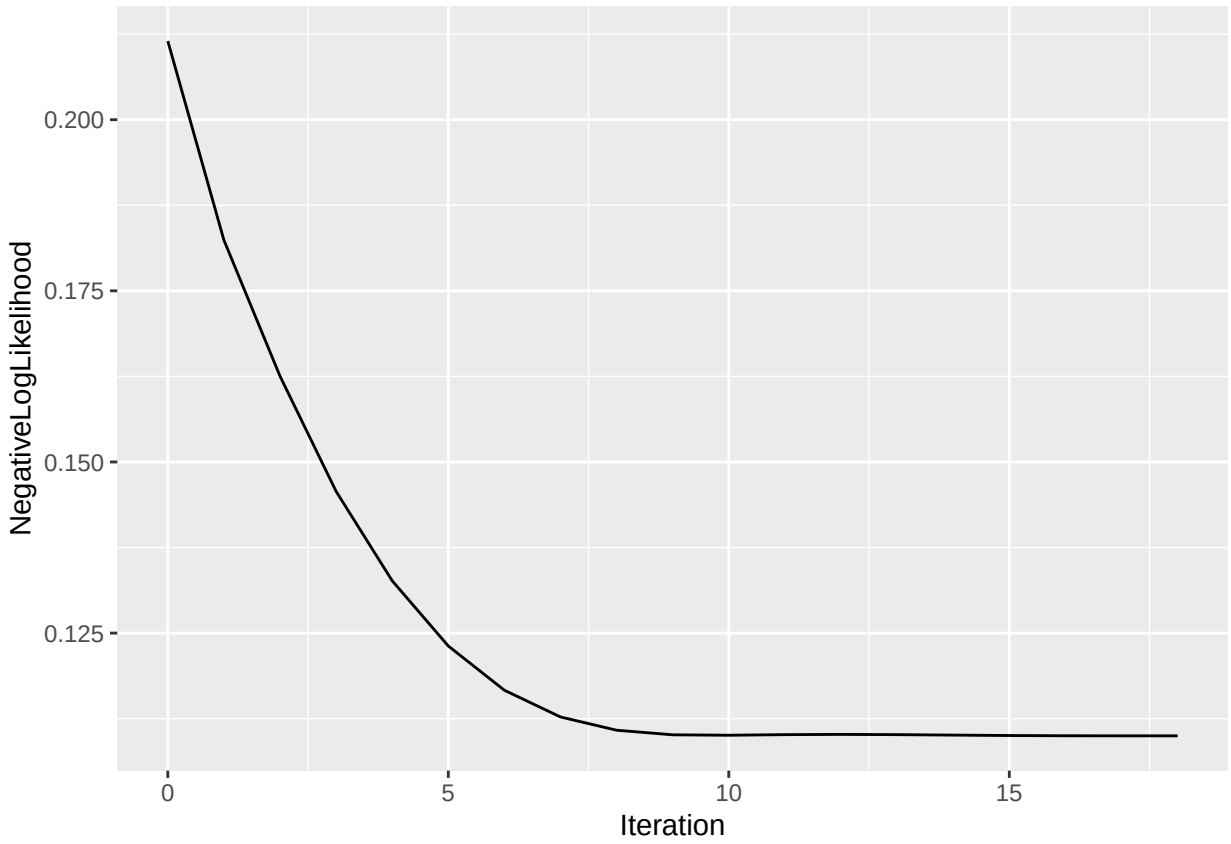
```
logpca_cv = cv.lpca(sepsis_df_sub, ks = 2, ms = 1:10)
plot(logpca_cv)
```

```
## Warning in type.convert.default(colnames(x)): 'as.is' should be specified by
## the caller; using TRUE
```

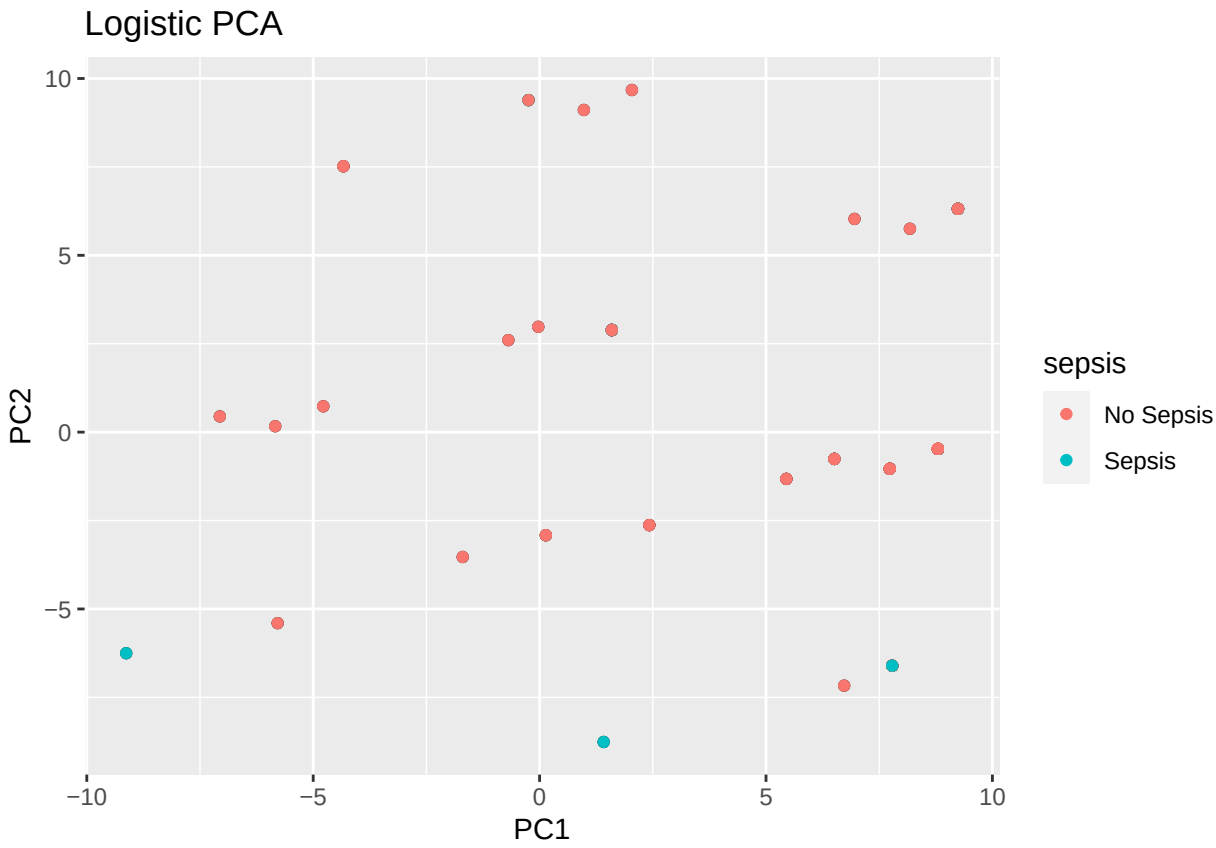
```
## Warning in type.convert.default(rownames(x)): 'as.is' should be specified by
## the caller; using TRUE
```



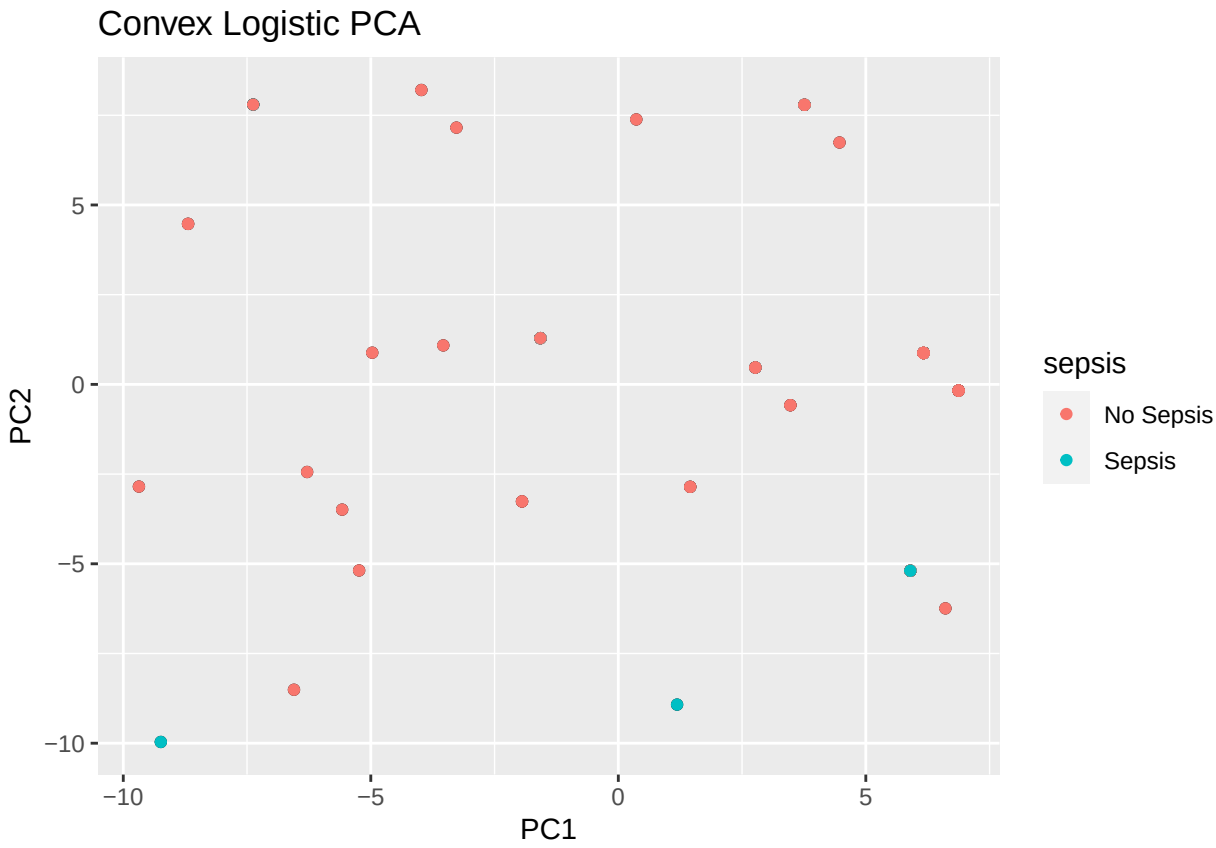

```
logpca_model = logisticPCA(sepsis_df_sub, k = 2, m = which.min(logpca_cv))
clogpca_model = convexLogisticPCA(sepsis_df_sub, k = 2, m = which.min(logpca_cv))
plot(clogpca_model, type = "trace")
```




```
plot(logpca_model, type = "scores") + geom_point(aes(colour = sepsis)) + ggtitle("Logistic PCA")
```

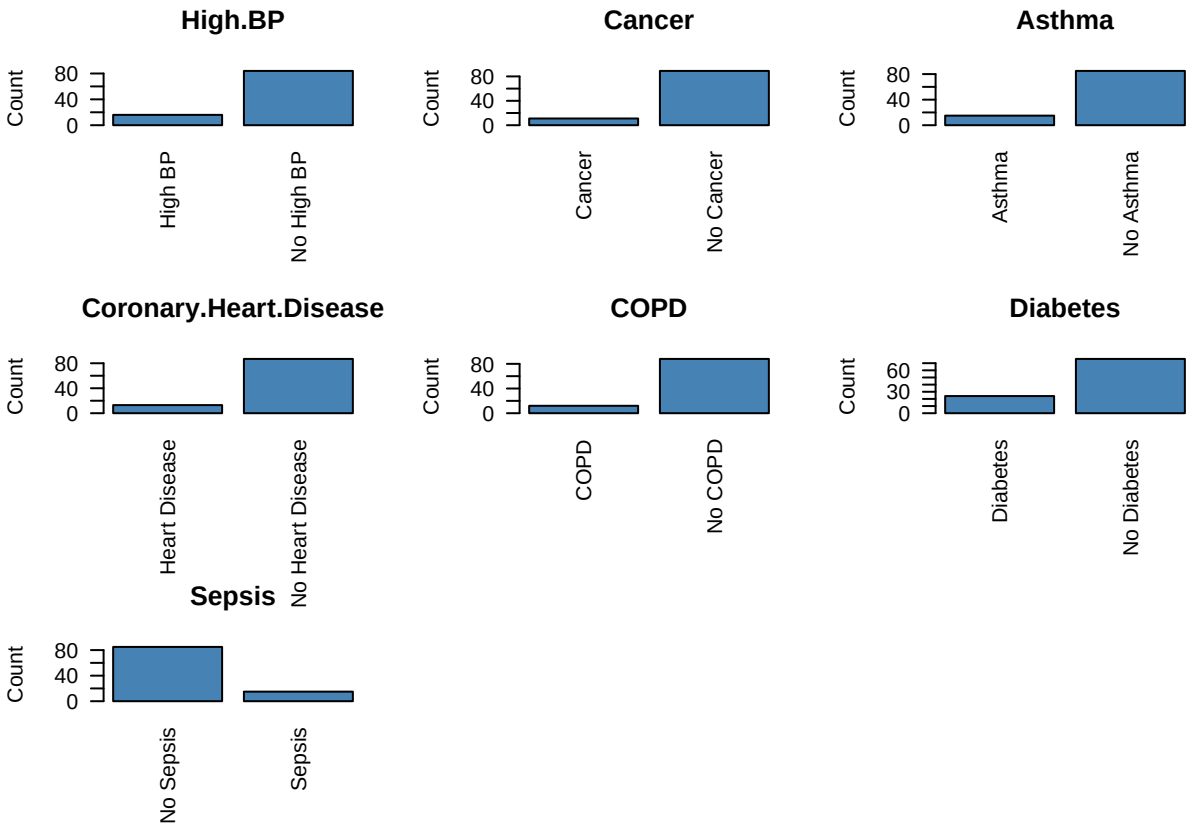



```
plot(clogpca_model, type = "scores") + geom_point(aes(colour = sepsis)) + ggtitle("Convex Logistic PCA")
```

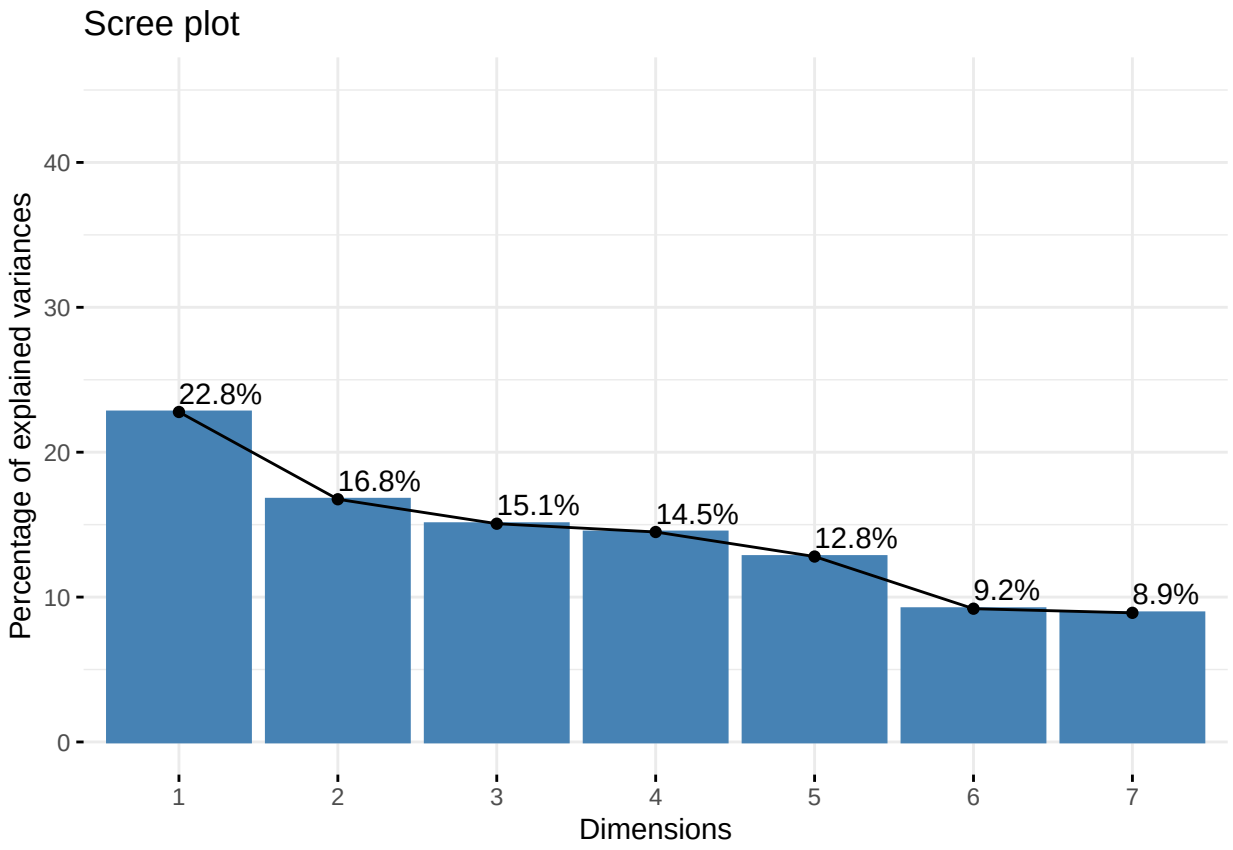


Multiple Correspondence Analysis

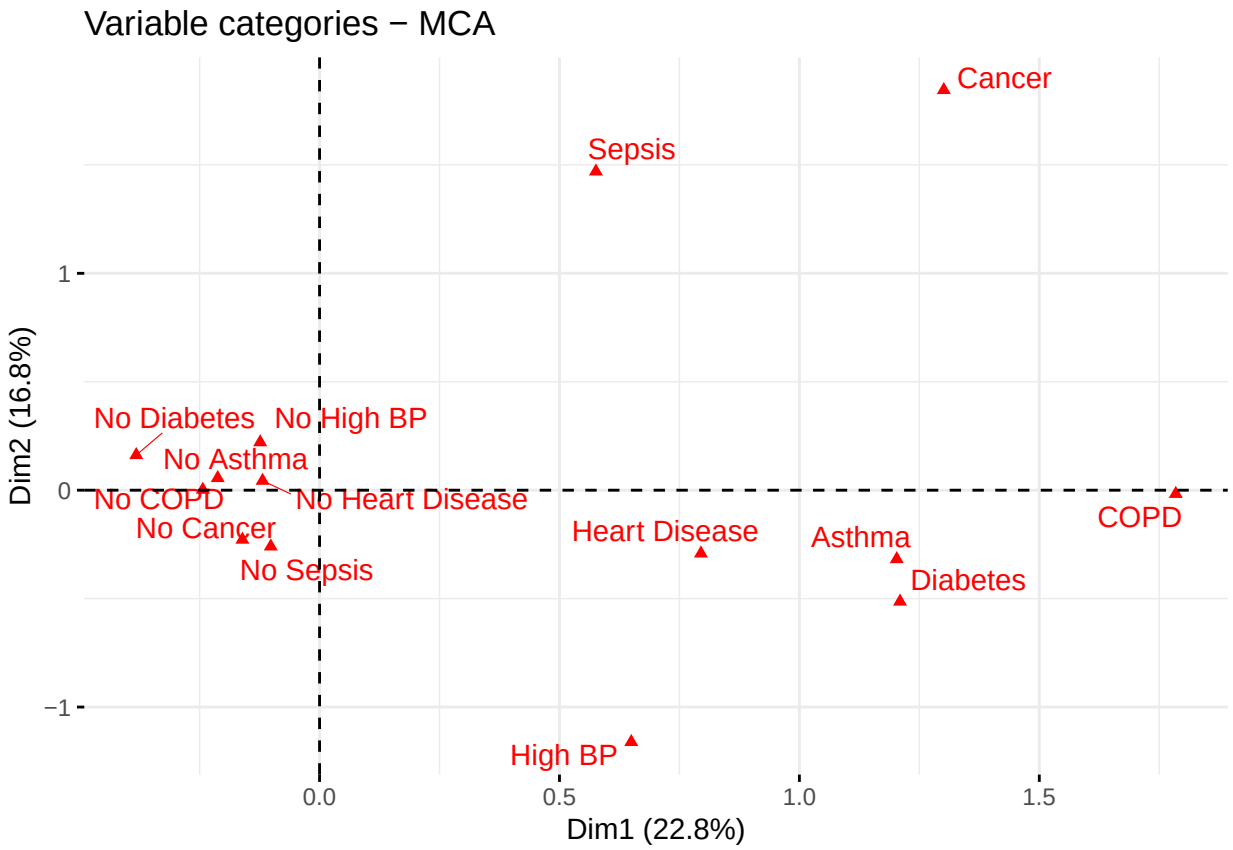
```
par(mfrow = c(3,3))
for (i in 1:7){
  plot(sepsis_df_new[,i], main = colnames(sepsis_df_new)[i], ylab = "Count", col = "steelblue", las = 2
}
```



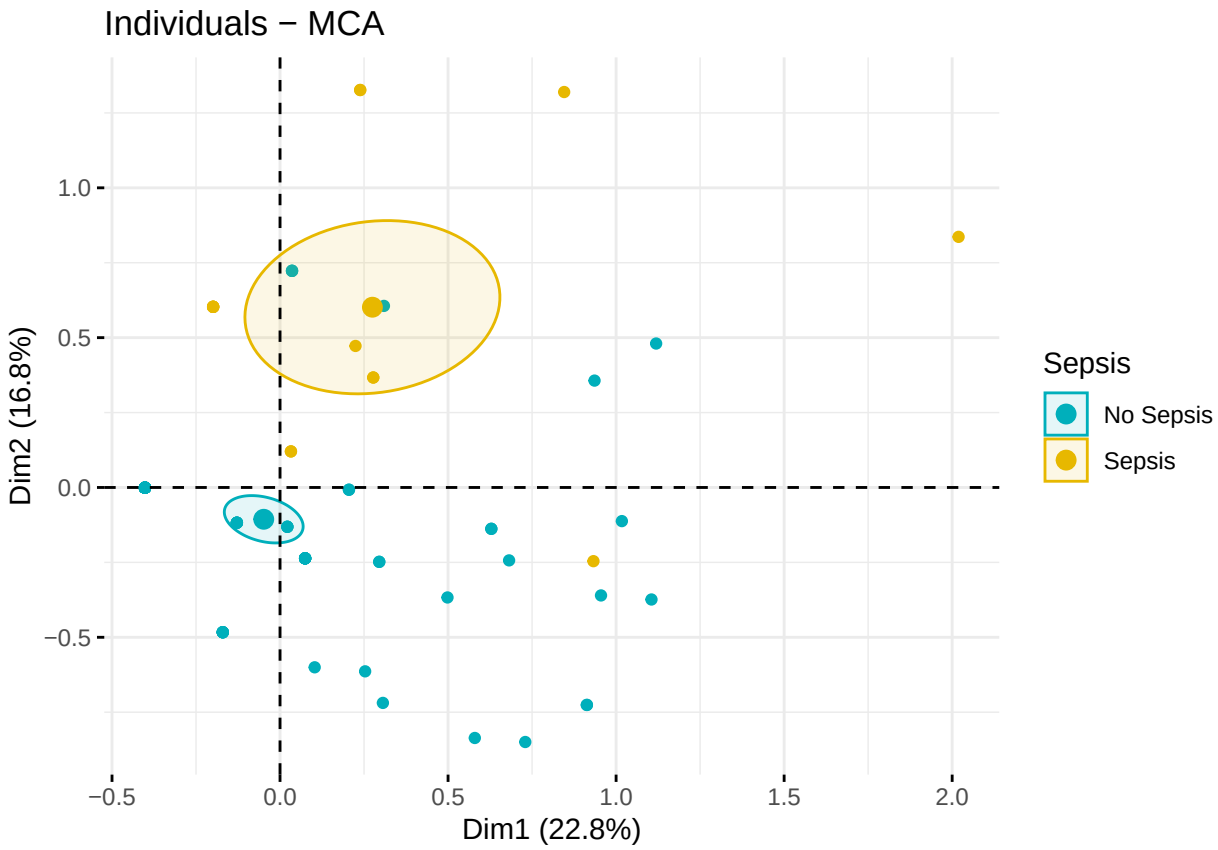

```
res.mca <- MCA(sepsis_df_new, graph = FALSE)
fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0, 45))
```




```
fviz_mca_var(res.mca,
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())
```

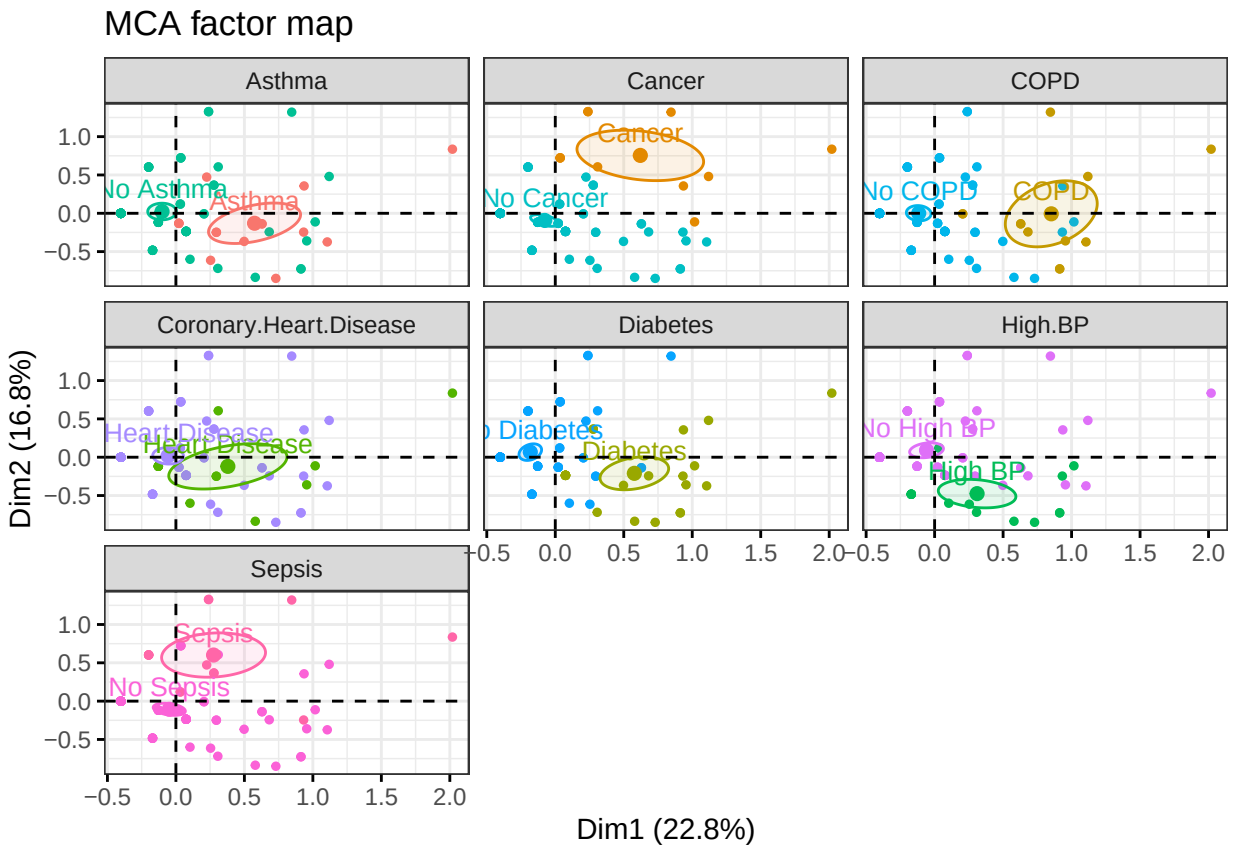



```
fviz_mca_ind(res.mca,
             label = "none", # hide individual labels
             habillage = "Sepsis", # color by groups
             palette = c("#00AFBB", "#E7B800"),
             addEllipses = TRUE, ellipse.type = "confidence",
             ggtheme = theme_minimal())
```




```
fviz_ellipses(res.mca, c("High.BP", "Cancer", "Asthma", "Coronary.Heart.Disease", "COPD", "Diabetes", "S
              geom = "point")
```

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## i Please use `gather()` instead.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Apply Logistic Regression Models

```
full_model <- glm(Sepsis ~ ., data = sepsis_df, family = "binomial")
summary(full_model)
```

```
##
## Call:
## glm(formula = Sepsis ~ ., family = "binomial", data = sepsis_df)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.03257    0.40305  -5.043 4.58e-07 ***
## High.BP1                  0.57964    0.76121   0.761   0.4464
## Cancer1                   1.72714    0.77260   2.235   0.0254 *
## Asthma1                   0.50683    0.77652   0.653   0.5140
## Coronary.Heart.Disease1 -1.36858    1.17342  -1.166   0.2435
## COPD1                   -0.21022    0.91897  -0.229   0.8191
## Diabetes1               -0.03531    0.72998  -0.048   0.9614
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 84.542  on 99  degrees of freedom
```

```
## Residual deviance: 78.519  on 93  degrees of freedom
## AIC: 92.519
##
## Number of Fisher Scoring iterations: 5
```

```
step_model <- step(full_model, direction = "both")
```

```
## Start:  AIC=92.52
## Sepsis ~ High.BP + Cancer + Asthma + Coronary.Heart.Disease +
##     COPD + Diabetes
##
##                           Df Deviance    AIC
## - Diabetes                 1   78.521 90.521
## - COPD                     1   78.572 90.572
## - Asthma                   1   78.924 90.924
## - High.BP                  1   79.065 91.065
## - Coronary.Heart.Disease   1   80.296 92.296
## <none>                         78.519 92.519
## - Cancer                   1   83.165 95.165
##
## Step:  AIC=90.52
## Sepsis ~ High.BP + Cancer + Asthma + Coronary.Heart.Disease +
##     COPD
##
##                           Df Deviance    AIC
## - COPD                     1   78.584 88.584
## - Asthma                   1   78.930 88.930
## - High.BP                  1   79.075 89.075
## - Coronary.Heart.Disease   1   80.316 90.316
## <none>                         78.521 90.521
## + Diabetes                 1   78.519 92.519
## - Cancer                   1   83.176 93.176
##
## Step:  AIC=88.58
## Sepsis ~ High.BP + Cancer + Asthma + Coronary.Heart.Disease
##
##                           Df Deviance    AIC
## - Asthma                   1   78.954 86.954
## - High.BP                  1   79.145 87.145
## - Coronary.Heart.Disease   1   80.386 88.386
## <none>                         78.584 88.584
## + COPD                     1   78.521 90.521
## + Diabetes                 1   78.572 90.572
## - Cancer                   1   83.184 91.184
##
## Step:  AIC=86.95
## Sepsis ~ High.BP + Cancer + Coronary.Heart.Disease
##
##                           Df Deviance    AIC
## - High.BP                  1   79.573 85.573
## - Coronary.Heart.Disease   1   80.631 86.631
## <none>                         78.954 86.954
## + Asthma                   1   78.584 88.584
## + COPD                     1   78.930 88.930
```

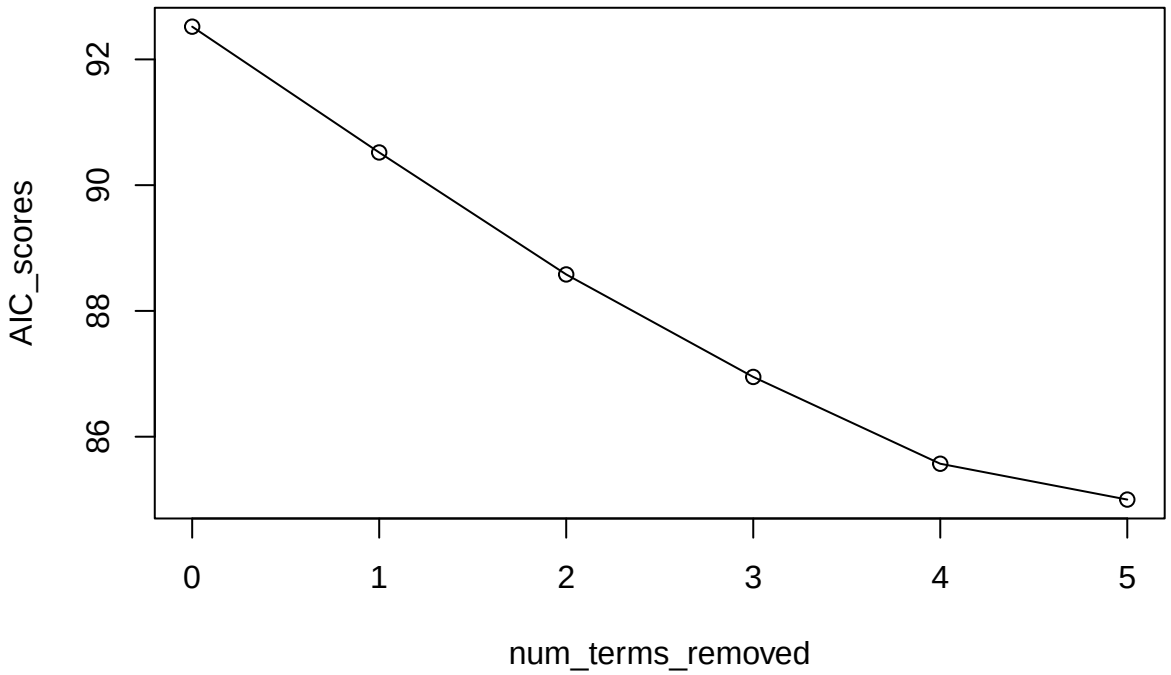
```
## + Diabetes                1   78.953 88.953
## - Cancer                  1   83.557 89.557
##
## Step:  AIC=85.57
## Sepsis ~ Cancer + Coronary.Heart.Disease
##
##                           Df Deviance    AIC
## - Coronary.Heart.Disease  1   80.998 84.998
## <none>                        79.573 85.573
## + High.BP                 1   78.954 86.954
## + Asthma                  1   79.145 87.145
## + Diabetes                1   79.539 87.539
## + COPD                    1   79.553 87.553
## - Cancer                  1   83.818 87.818
##
## Step:  AIC=85
## Sepsis ~ Cancer
##
##                           Df Deviance    AIC
## <none>                        80.998 84.998
## + Coronary.Heart.Disease  1   79.573 85.573
## - Cancer                  1   84.542 86.542
## + High.BP                 1   80.631 86.631
## + Asthma                  1   80.732 86.732
## + COPD                    1   80.967 86.967
## + Diabetes                1   80.995 86.995
```

```
glm_model <- bestglm(sepsis_df, IC = "AIC", method = "exhaustive", family = binomial)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
glm_model
```

```
## AIC
## BICq equivalent for q in (0.629594182610971, 0.830630784263591)
## Best Model:
##              Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -1.958814  0.3220693 -6.081964 1.187194e-09
## Cancer1      1.399198  0.7046884  1.985555 4.708272e-02
```



Here the goal was to apply a logistic regression to identify the features of our dataset that are most likely to correspond to a Sepsis diagnosis. Once we had that first model, we wanted to identify which variables were the most significant in our dataset in classifying whether a patient had Sepsis. Therefore by minimizing AIC (Akaike Information Criterion), a score that compares how much the model deviates from the data, we got the best logistic regression model. We approached this in two ways, `step` and `bestglm`. The objective function for both is to minimize AIC, however, step performs the estimation in an iterative manner (the iterative process and decrease in AIC as we remove variables are shown in the graph above) while bestglm considers all possible subsets of the variables and identifies the model that has the minimal AIC. Using both techniques we got the following variables to be the most statistically significant in classifying Sepsis and to also have the same coefficients and intercept as shown in the table below.

	Step Coefficients	GLM Coefficients	Pr(>|t|)
(Intercept)	-1.959	-1.959	0.0000000012
Cancer	1.399	1.399	0.0470827200

Table 1: Coefficients and P value of Remaining Variables after Step and Exhaustive Approach to finding significant coefficients