Asteriod_Regression

Madeline Pfister

2023-04-26

```
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(mlbench)
library(rattle)
```

```
## Loading required package: tibble

## Loading required package: bitops

## Rattle: A free graphical interface for data science with R.
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:rattle':
##
##     importance

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(ROCR)
```

```
nasa <- read.csv("nasa_v2.csv")
```

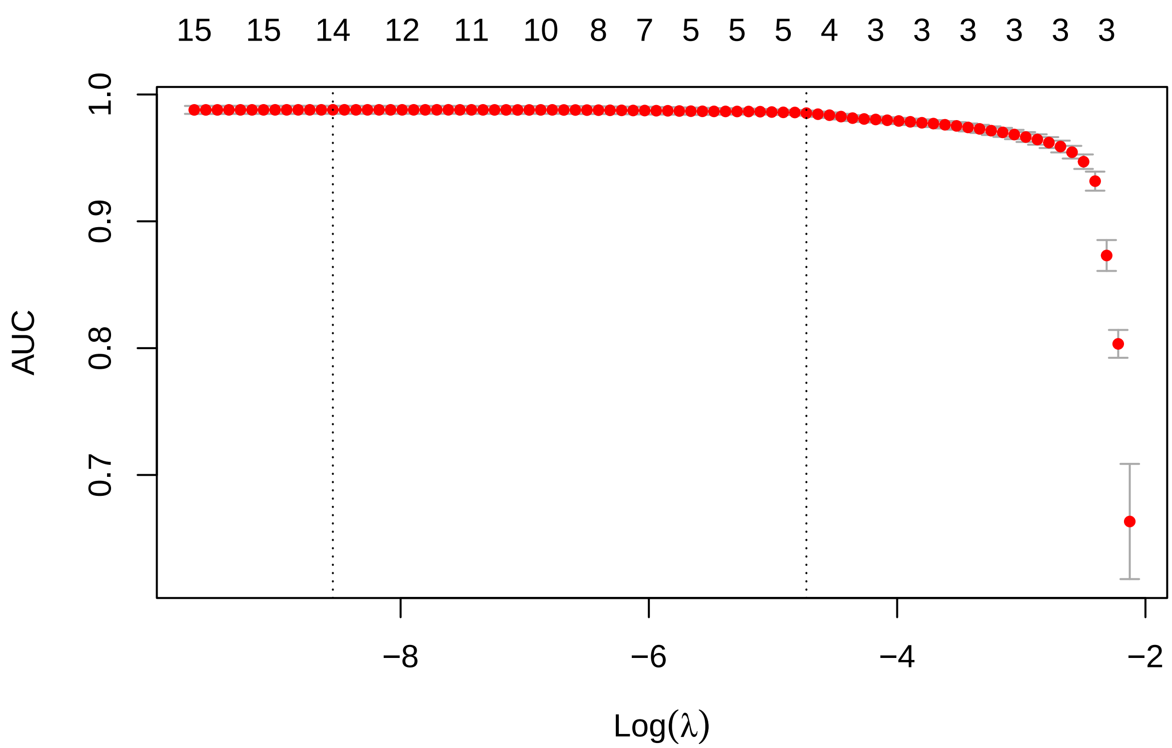
```
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:bitops':
##
##     %&%

## Loaded glmnet 4.1-7
```

```
#create the design matrix
X = model.matrix(Hazardous ~ ., data=nasa)
Y = as.numeric(nasa$Hazardous=="True")

#conduct the cross-validation
set.seed(1)
cvfit = cv.glmnet(x=X[,-1], y=Y, family="binomial", type.measure="auc")
cvfit
```

```
##
## Call:  cv.glmnet(x = X[, -1], y = Y, type.measure = "auc", family = "binomial")
##
## Measure: AUC
##
##       Lambda Index Measure       SE Nonzero
## min 0.000194    70  0.9879 0.003095      14
## 1se 0.008816    29  0.9853 0.001998       5
```

```
plot(cvfit)
```




```
#Variables selected using lambda.1se
sel.vars <- which(coef(cvfit, s=cvfit$lambda.1se)!=0)[-1]-1
sel.names <- colnames(nasa)[sel.vars]
sel.names
```

```
## [1] "Absolute.Magnitude"          "Est.Dia.in.KM.min."
## [3] "Orbit.Uncertainity"          "Minimum.Orbit.Intersection"
## [5] "Jupiter.Tisserand.Invariant"
```

```
#fit a lasso model using the selected variables
fit.lasso <- glm(as.factor(Hazardous) ~ Absolute.Magnitude + Est.Dia.in.KM.min. + Orbit.Uncertainity +
                   Minimum.Orbit.Intersection + Jupiter.Tisserand.Invariant,
```

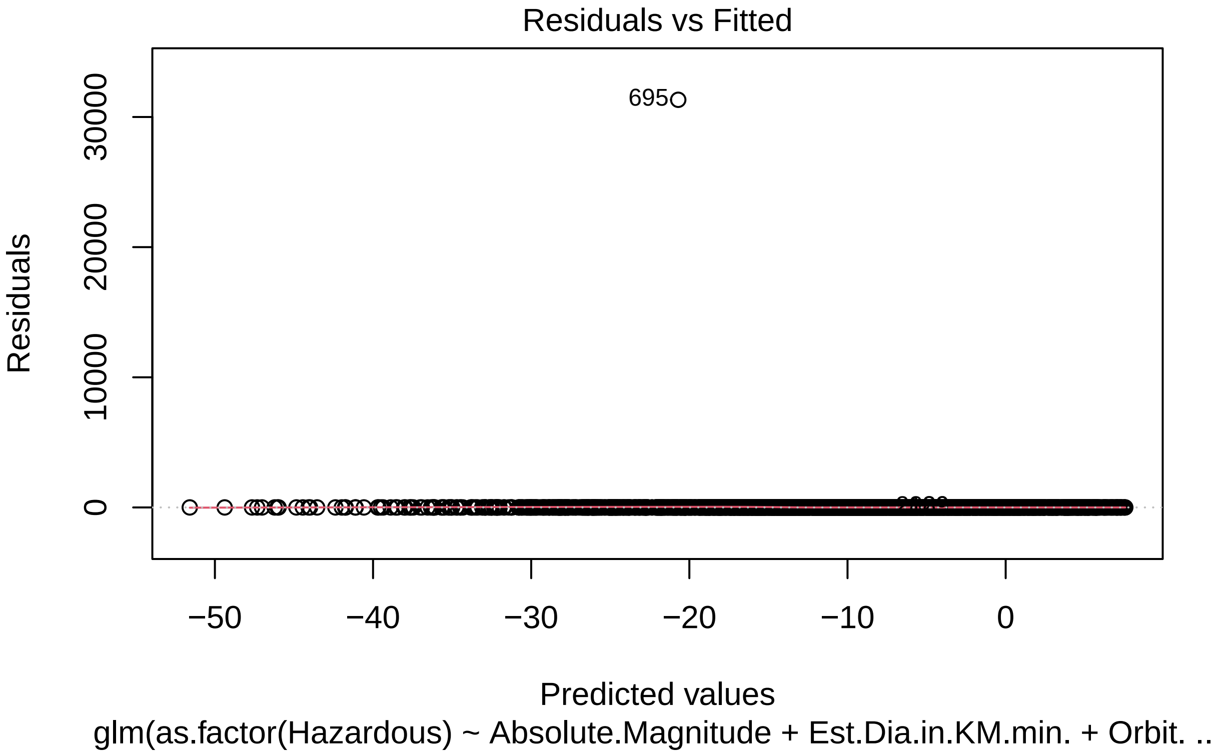
```r
                 family="binomial", data=nasa)
```

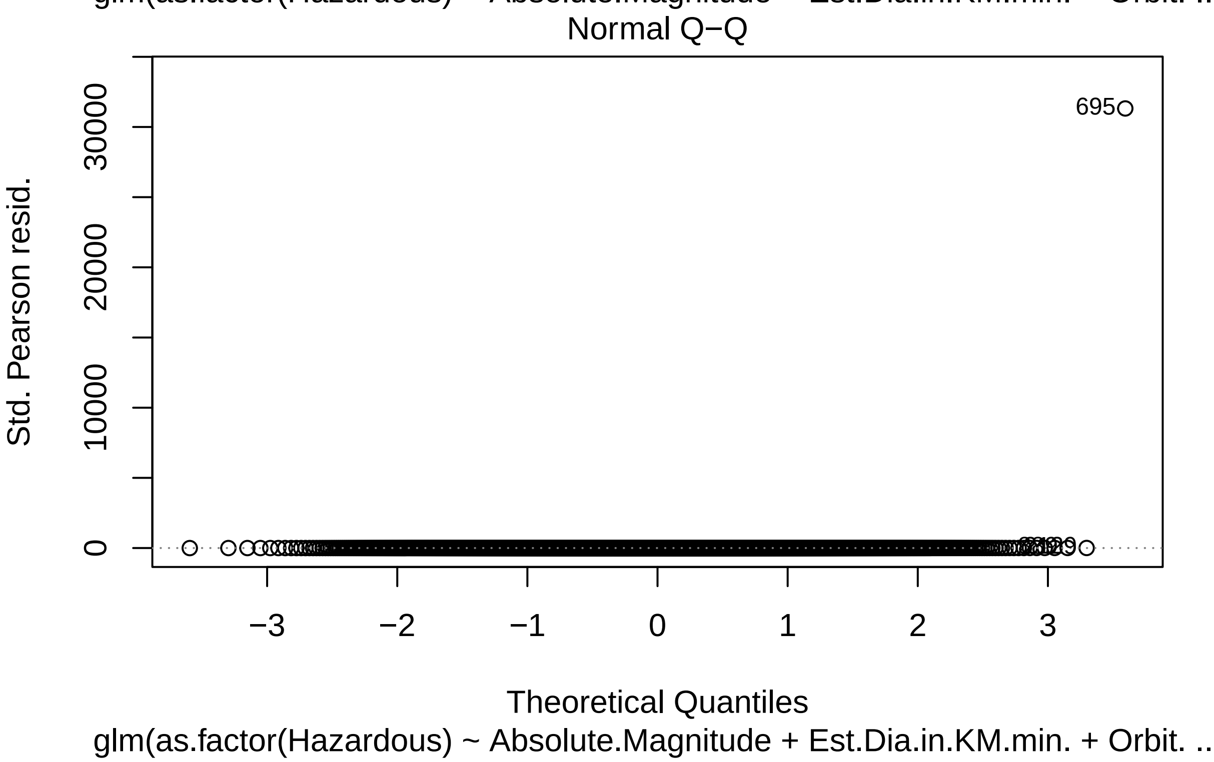
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

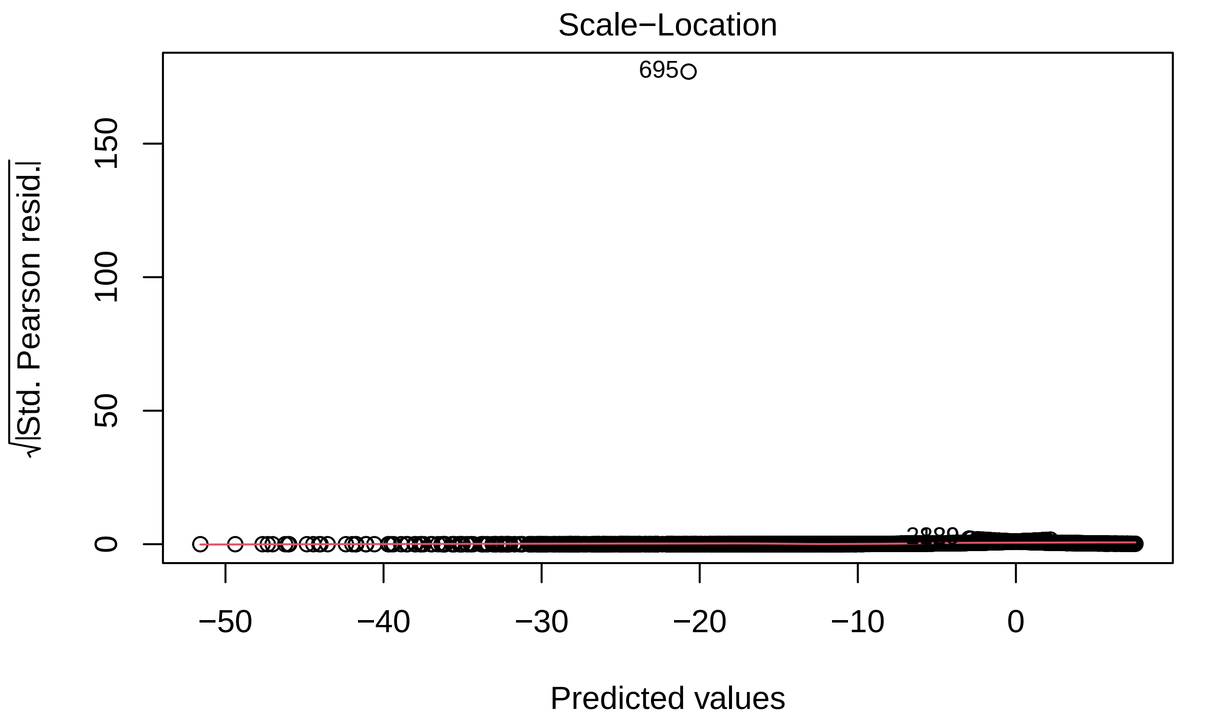
```r
summary(fit.lasso)
```

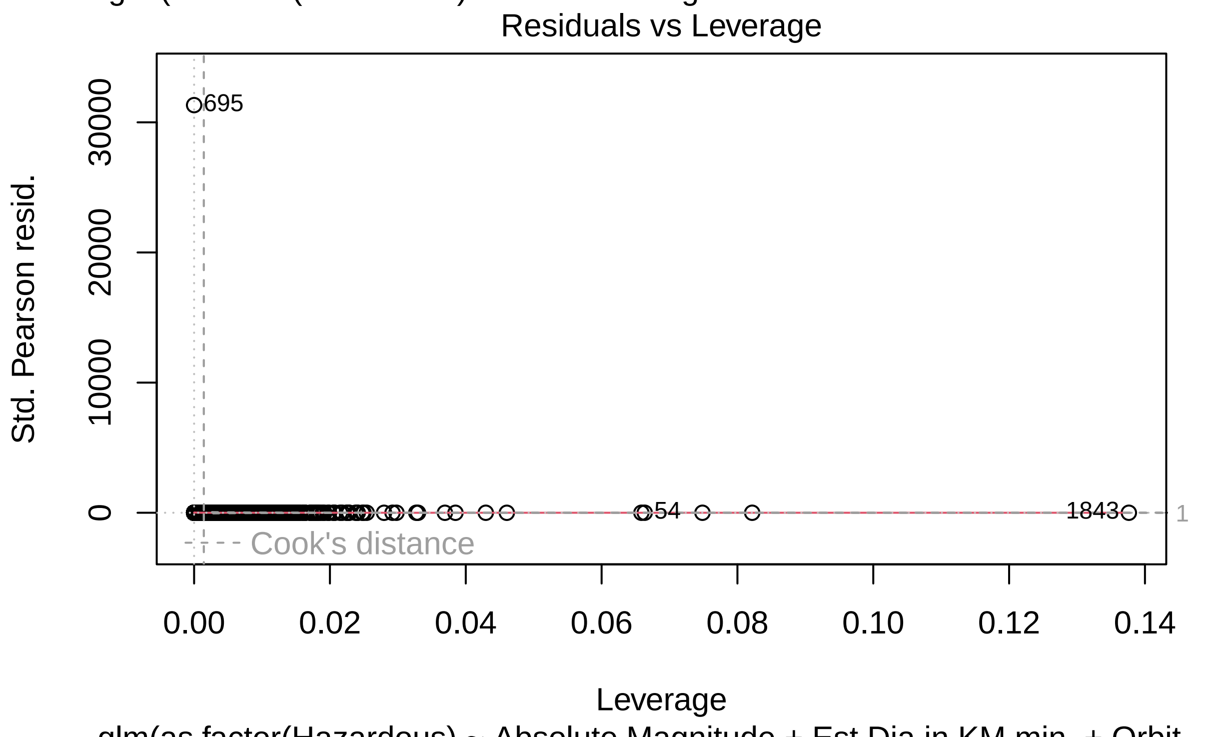
```
##
## Call:
## glm(formula = as.factor(Hazardous) ~ Absolute.Magnitude + Est.Dia.in.KM.min. +
##     Orbit.Uncertainity + Minimum.Orbit.Intersection + Jupiter.Tisserand.Invariant,
##     family = "binomial", data = nasa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1427  -0.0381  -0.0029   0.0000   6.4349
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  69.42299    5.04357  13.765  < 2e-16 ***
## Absolute.Magnitude           -2.94351    0.22007 -13.375  < 2e-16 ***
## Est.Dia.in.KM.min.          -12.52488    1.33533  -9.380  < 2e-16 ***
## Orbit.Uncertainity           -0.13951    0.04682  -2.980  0.00288 **
## Minimum.Orbit.Intersection -119.42430    7.97567 -14.974  < 2e-16 ***
## Jupiter.Tisserand.Invariant   -0.13442    0.08851  -1.519  0.12887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2508.03  on 3078  degrees of freedom
## Residual deviance:  557.22  on 3073  degrees of freedom
## AIC: 569.22
##
## Number of Fisher Scoring iterations: 9
```

```r
#address the model assumptions of the lasso model
plot(fit.lasso)
```


Residuals vs Fitted
Residuals
0
10000
20000
30000
695
−50
−40
−30
−20
−10
0
Predicted values
glm(as.factor(Hazardous) ~ Absolute.Magnitude + Est.Dia.in.KM.min. + Orbit. ...


Normal Q−Q
Std. Pearson resid.
0
10000
20000
30000
695
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(as.factor(Hazardous) ~ Absolute.Magnitude + Est.Dia.in.KM.min. + Orbit. ...





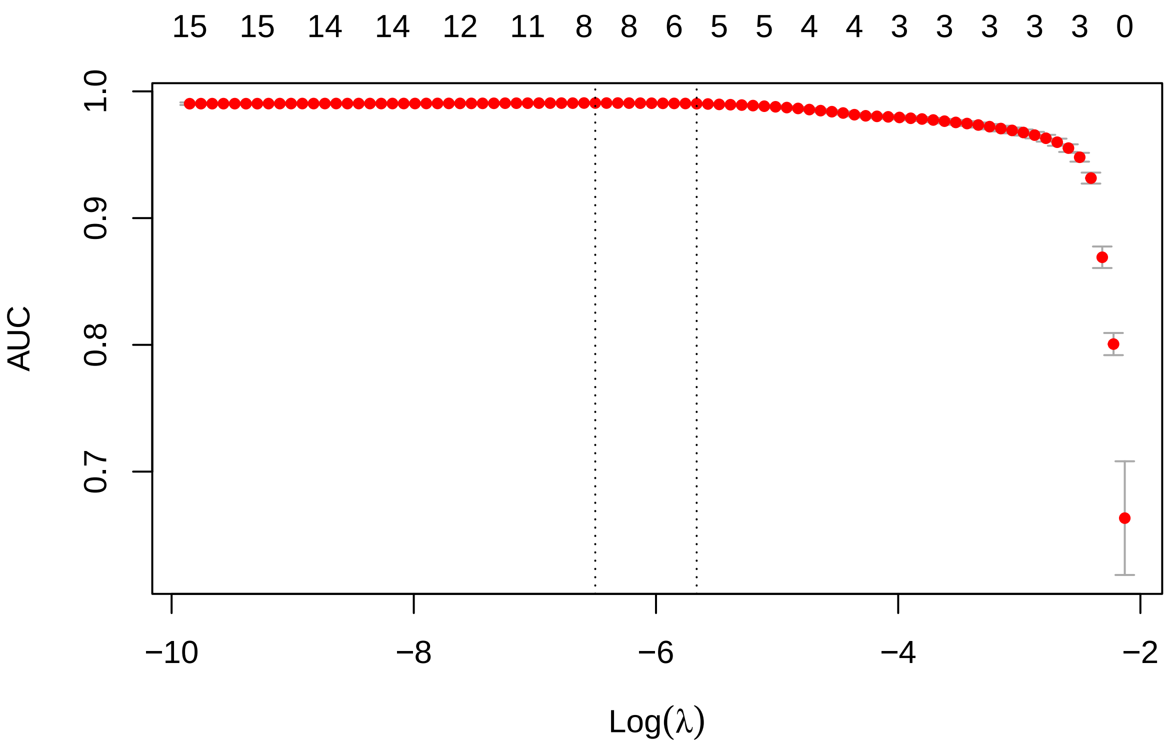
Due to a large outlier (observation 695), none of the assumptions for regression are met.

```r
#removing the large outlier to satisfy model assumptions
nasa2 <- nasa[-695,]

#create the design matrix
X2 = model.matrix(Hazardous ~ ., data=nasa2)
```

```r
Y2 = as.numeric(nasa2$Hazardous=="True")

#conduct the cross-validation
set.seed(1)
cvfit2 = cv.glmnet(x=X2[,-1], y=Y2, family="binomial", type.measure="auc")
cvfit2
```

```
##
## Call:  cv.glmnet(x = X2[, -1], y = Y2, type.measure = "auc", family = "binomial")
##
## Measure: AUC
##
##        Lambda Index Measure        SE Nonzero
## min 0.001501    48   0.9908 0.0007380       8
## 1se 0.003468    39   0.9902 0.0008217       5
```

```r
plot(cvfit2)
```




```r
#Variables selected using lambda.1se
sel.vars2 <- which(coef(cvfit2, s=cvfit2$lambda.1se)!=0)[-1]-1
sel.names2 <- colnames(nasa2)[sel.vars2]
sel.names2
```

```
## [1] "Absolute.Magnitude"         "Est.Dia.in.KM.min."
## [3] "Orbit.Uncertainity"         "Minimum.Orbit.Intersection"
## [5] "Mean.Motion"
```

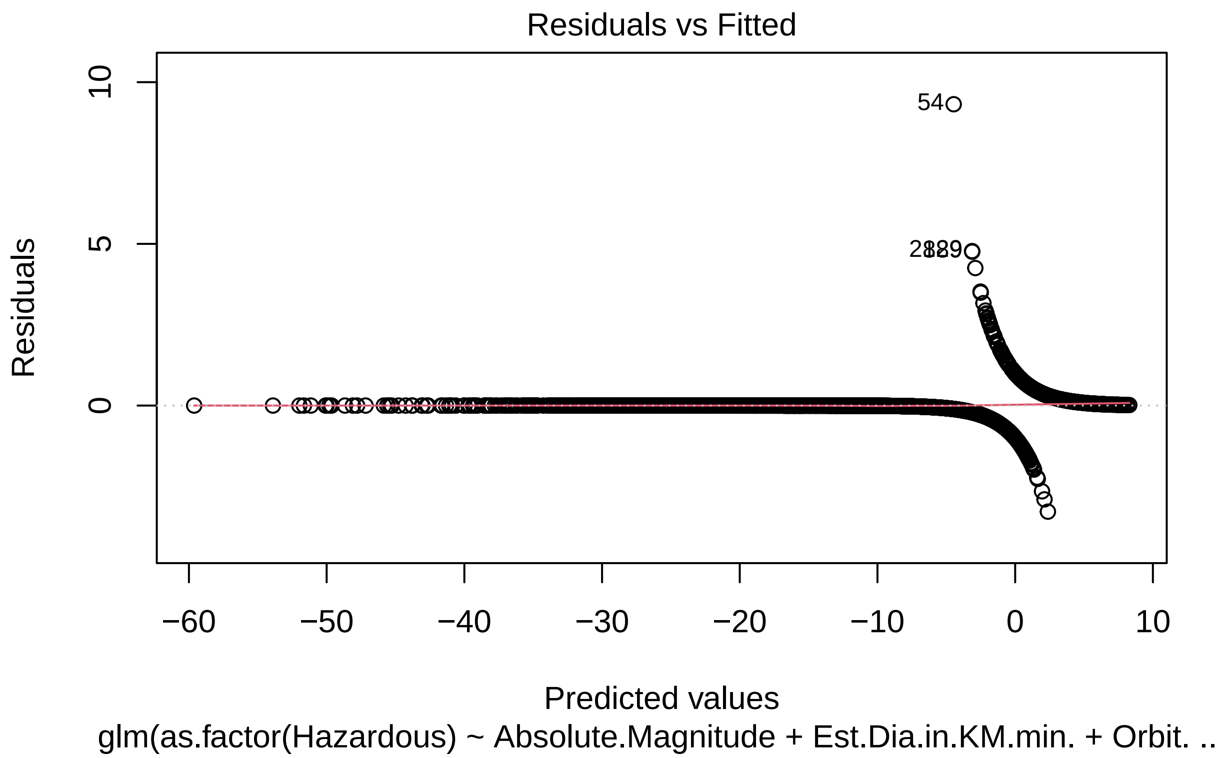
```r
#create a new model
fit.lasso2 <- glm(as.factor(Hazardous) ~ Absolute.Magnitude + Est.Dia.in.KM.min. +
                    Orbit.Uncertainity + Minimum.Orbit.Intersection + Mean.Motion,
                family="binomial", data=nasa2)
```

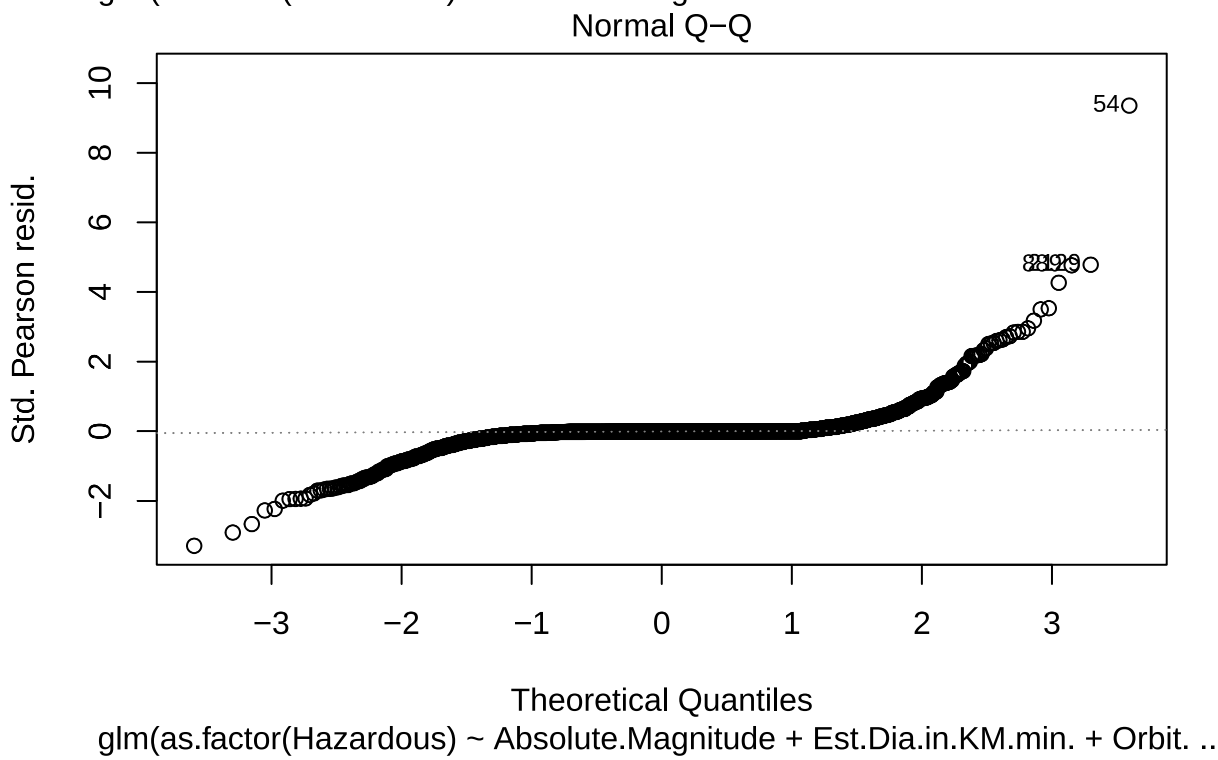
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

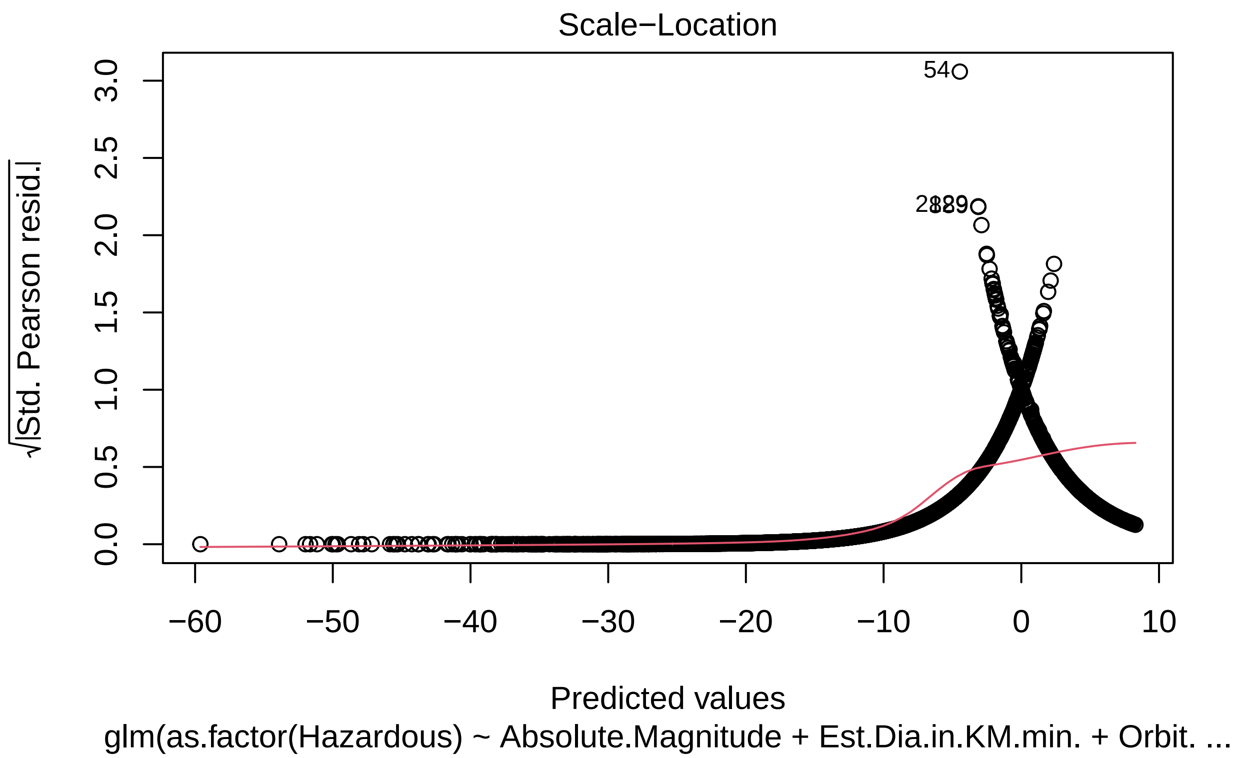
```r
summary(fit.lasso2)
```

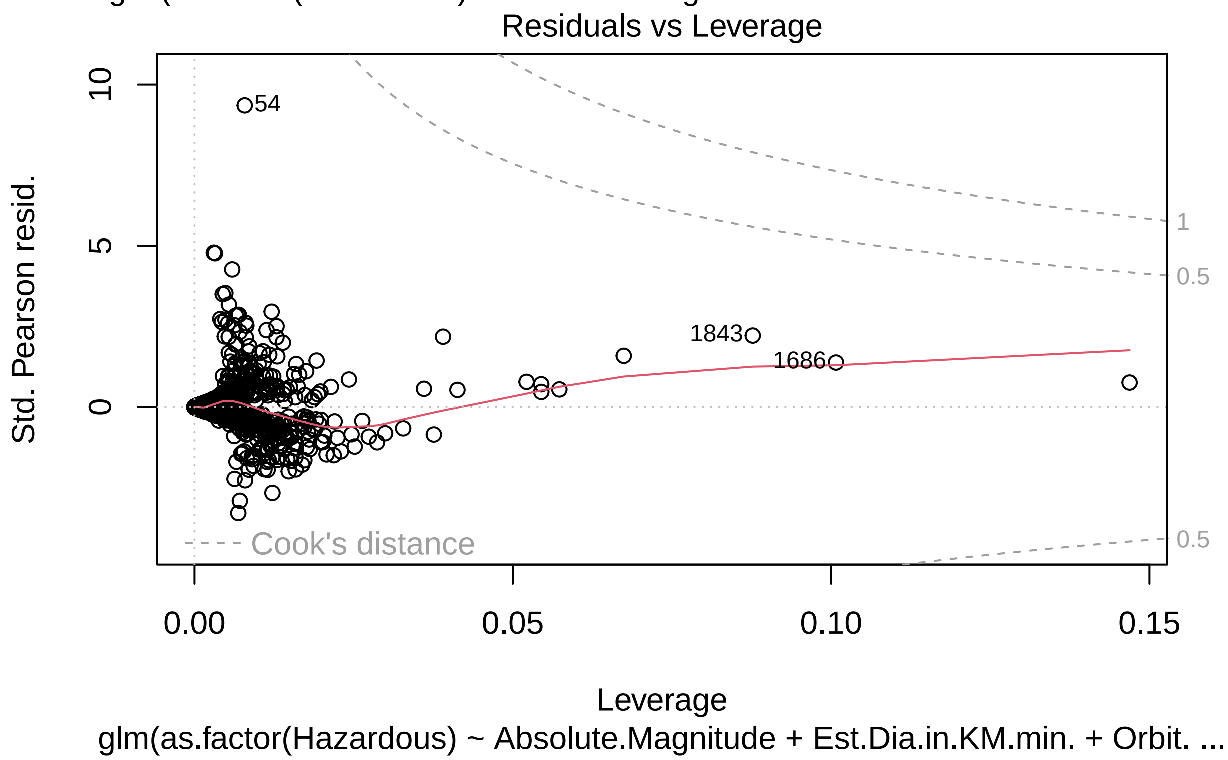
```
##
## Call:
## glm(formula = as.factor(Hazardous) ~ Absolute.Magnitude + Est.Dia.in.KM.min. +
##     Orbit.Uncertainity + Minimum.Orbit.Intersection + Mean.Motion,
##     family = "binomial", data = nasa2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.22020  -0.02263  -0.00116  -0.00001   2.99174
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  83.83471    6.10198  13.739  < 2e-16 ***
## Absolute.Magnitude           -3.57931    0.26542 -13.486  < 2e-16 ***
## Est.Dia.in.KM.min.          -17.12997    1.54537 -11.085  < 2e-16 ***
## Orbit.Uncertainity           -0.13717    0.04917  -2.790  0.00528 **
## Minimum.Orbit.Intersection -129.95192    8.97244 -14.483  < 2e-16 ***
## Mean.Motion                  -0.49633    0.33275  -1.492  0.13580
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2504.11  on 3077  degrees of freedom
## Residual deviance:  504.81  on 3072  degrees of freedom
## AIC: 516.81
##
## Number of Fisher Scoring iterations: 10
```

```r
#assess model assumptions
plot(fit.lasso2)
```


Residuals vs Fitted
Residuals
10
5
0
54
2189
−60
−50
−40
−30
−20
−10
0
10
Predicted values
glm(as.factor(Hazardous) ~ Absolute.Magnitude + Est.Dia.in.KM.min. + Orbit. ...


Normal Q−Q
Std. Pearson resid.
10
8
6
4
2
0
−2
54
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(as.factor(Hazardous) ~ Absolute.Magnitude + Est.Dia.in.KM.min. + Orbit. ...





Based on visual analysis, the model assumptions appear to be satisfied?

```
#assess model assumptions -- multicollinearity
library(car)
```

```
## Loading required package: carData
```

```r
vif(fit.lasso2)
```

```
##        Absolute.Magnitude         Est.Dia.in.KM.min.
##                 13.374033                   8.520165
##        Orbit.Uncertainity Minimum.Orbit.Intersection
##                  1.365700                   3.027934
##               Mean.Motion
##                  1.130061
```

There are multicollinearity with Absolute.Magnitude and Est.Dia.in.KM.min. as they both have a vif > 5.

```r
#create a new model removing Absolute.Magnitude as it has the largest VIF
fit.lasso3 <- glm(as.factor(Hazardous) ~ Est.Dia.in.KM.min. + Orbit.Uncertainity +
              Minimum.Orbit.Intersection + Mean.Motion,family="binomial", data=nasa2)
summary(fit.lasso3)
```

```
##
## Call:
## glm(formula = as.factor(Hazardous) ~ Est.Dia.in.KM.min. + Orbit.Uncertainity +
##     Minimum.Orbit.Intersection + Mean.Motion, family = "binomial",
##     data = nasa2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1854  -0.3306  -0.0984  -0.0036   3.1953
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  2.34246    0.27245   8.598  < 2e-16 ***
## Est.Dia.in.KM.min.           4.65616    0.51422   9.055  < 2e-16 ***
## Orbit.Uncertainity          -0.51699    0.03237 -15.974  < 2e-16 ***
## Minimum.Orbit.Intersection -62.10827    4.01705 -15.461  < 2e-16 ***
## Mean.Motion                 -1.41537    0.23374  -6.055  1.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2504.1  on 3077  degrees of freedom
## Residual deviance: 1189.8  on 3073  degrees of freedom
## AIC: 1199.8
##
## Number of Fisher Scoring iterations: 8
```

```r
#test the multicollinearity of the adjusted model
vif(fit.lasso3)
```

```
##         Est.Dia.in.KM.min.         Orbit.Uncertainity
##                   2.155630                   1.479609
## Minimum.Orbit.Intersection                Mean.Motion
##                   2.135797                   1.080829
```

There are no multicollinearity issues in the adjusted model.

$$H_0\text{: the data fits the model well}$$

$$H_1\text{: the data does not fit the model well}$$

$$\alpha = 0.05$$

```r
#test the goodness of fit of the model
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```r
hoslem.test(fit.lasso3$y, fit.lasso3$fitted.values)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  fit.lasso3$y, fit.lasso3$fitted.values
## X-squared = 9.4684, df = 8, p-value = 0.3043
```

There is statistically sufficient evidence (p = 0.3043, df = 8) to reject the null hypothesis and conclude that the model fits the data well.

```r
#create a new model keeping Absolute.Magnitude and removing Est.Dia.in.KM.min
fit.lasso4 <- glm(as.factor(Hazardous) ~ Absolute.Magnitude + Orbit.Uncertainity +
              Minimum.Orbit.Intersection + Mean.Motion,family="binomial", data=nasa2)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(fit.lasso4)
```

```
##
## Call:
## glm(formula = as.factor(Hazardous) ~ Absolute.Magnitude + Orbit.Uncertainity +
##     Minimum.Orbit.Intersection + Mean.Motion, family = "binomial",
##     data = nasa2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.84132  -0.10369  -0.01797  -0.00008   2.56744
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  35.50759    2.16373  16.410  < 2e-16 ***
## Absolute.Magnitude           -1.49105    0.09507 -15.684  < 2e-16 ***
## Orbit.Uncertainity           -0.16626    0.04068  -4.087 4.37e-05 ***
## Minimum.Orbit.Intersection -109.77605    6.87597 -15.965  < 2e-16 ***
## Mean.Motion                  -0.53944    0.28509  -1.892   0.0585 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2504.11  on 3077  degrees of freedom
## Residual deviance:  690.15  on 3073  degrees of freedom
## AIC: 700.15
##
## Number of Fisher Scoring iterations: 9
```

```r
#test the multicollinearity of the adjusted model
vif(fit.lasso4)
```

```
##         Absolute.Magnitude         Orbit.Uncertainity
```

```
##                    3.045975                         1.406947
## Minimum.Orbit.Intersection                    Mean.Motion
##                    2.861872                         1.094045
```

There are no multicollinearity issues in the adjusted model.

$$H_0\text{: the data fits the model well}$$

$$H_1\text{: the data does not fit the model well}$$

$$\alpha = 0.05$$

```
#test the goodness of fit of the model
library(ResourceSelection)
hoslem.test(fit.lasso4$y, fit.lasso4$fitted.values)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  fit.lasso4$y, fit.lasso4$fitted.values
## X-squared = 7.0768, df = 8, p-value = 0.5284
```

There is statistically sufficient evidence (p = 0.5284, df = 8) to reject the null hypothesis and conclude that the model fits the data well.

fit.lasso4 is the best fit for the data as it has the highest p-value in the Hosmer and Lemeshow goodness of fit test.

```
#create the cross-validated model the best fitting lasso model
library(caret)
set.seed(1)
fit.cv <- train(as.factor(Hazardous) ~ Absolute.Magnitude + Orbit.Uncertainity +
                   Minimum.Orbit.Intersection + Mean.Motion ,
                   method = "glm", family = "binomial", trControl = trainControl(method="cv", number=5,
                                                          savePredictions = TRUE, classProbs = TRUE),data=nasa2)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
fit.cv
```

```
## Generalized Linear Model
##
## 3078 samples
##    4 predictor
##    2 classes: 'False', 'True'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 2462, 2463, 2463, 2462, 2462
## Resampling results:
```

```
##
##   Accuracy   Kappa
##   0.9470499  0.7763952
```

```r
#determine the final model
summary(fit.cv$finalModel)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.84132  -0.10369  -0.01797  -0.00008   2.56744
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  35.50759    2.16373  16.410  < 2e-16 ***
## Absolute.Magnitude           -1.49105    0.09507 -15.684  < 2e-16 ***
## Orbit.Uncertainity           -0.16626    0.04068  -4.087 4.37e-05 ***
## Minimum.Orbit.Intersection -109.77605    6.87597 -15.965  < 2e-16 ***
## Mean.Motion                  -0.53944    0.28509  -1.892   0.0585 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2504.11  on 3077  degrees of freedom
## Residual deviance:  690.15  on 3073  degrees of freedom
## AIC: 700.15
##
## Number of Fisher Scoring iterations: 9
```

```r
#verify the multicollinearity
vif(fit.cv$finalModel)
```

```
##         Absolute.Magnitude         Orbit.Uncertainity
##                   3.045975                   1.406947
## Minimum.Orbit.Intersection                Mean.Motion
##                   2.861872                   1.094045
```

```r
#verify the goodness of fit
hoslem.test(fit.cv$finalModel$y, fit.cv$finalModel$fitted.values)
```

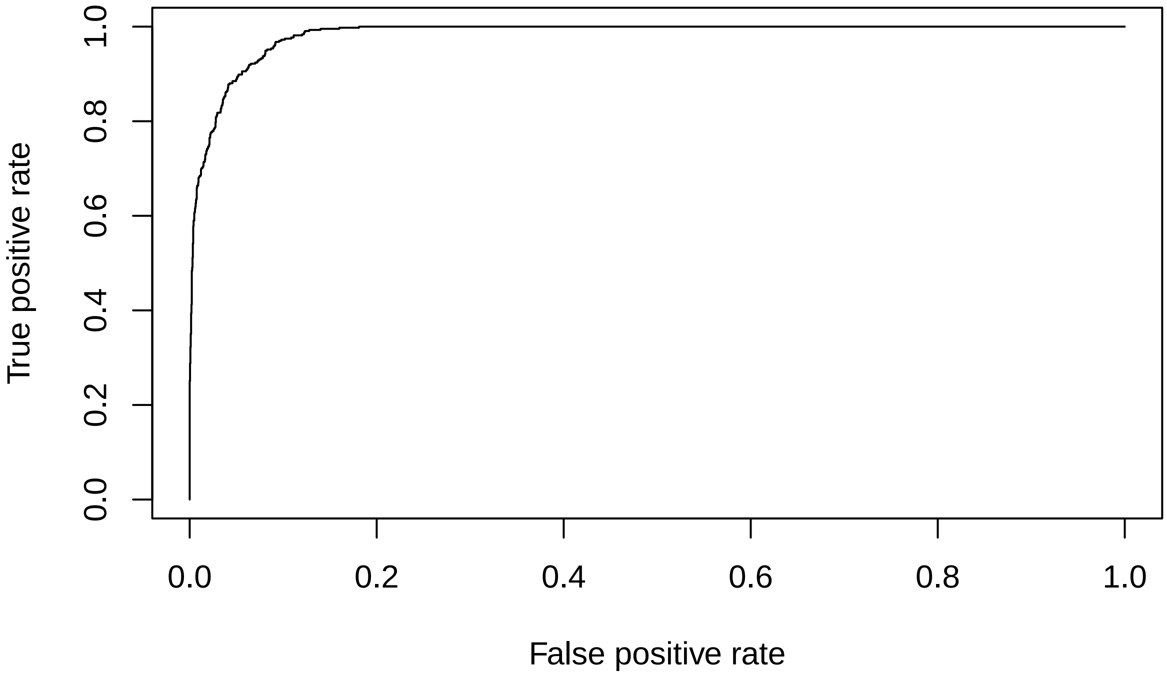
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  fit.cv$finalModel$y, fit.cv$finalModel$fitted.values
## X-squared = 7.0768, df = 8, p-value = 0.5284
```

```r
#assess the predictive performance using the predictive model
pihatcv <- fit.cv$pred
head(cbind(nasa$Hazardous[pihatcv$rowIndex], pihatcv))
```

```
##   nasa$Hazardous[pihatcv$rowIndex]  pred   obs     False         True rowIndex
## 1                             True False  True 0.7412096 2.587904e-01        5
## 2                            False False False 1.0000000 1.617148e-08       11
```

```
## 3                              False False False 0.9871608 1.283920e-02       12
## 4                              False False False 0.9946017 5.398327e-03       13
## 5                              False False False 0.9957621 4.237882e-03       25
## 6                              False False False 0.9999993 7.158642e-07       30
##   parameter Resample
## 1      none    Fold1
## 2      none    Fold1
## 3      none    Fold1
## 4      none    Fold1
## 5      none    Fold1
## 6      none    Fold1
```

```
predcv <- prediction(pihatcv$"True", pihatcv$obs)
perfcv <- performance(predcv, "tpr", "fpr")
plot(perfcv)
```




```
auccv <- performance(predcv, "auc")@y.values
auccv
```

```
## [[1]]
## [1] 0.9834971
```

```
get_stats <- function(CM) {
  TP <- CM[2,2]
  FP <- CM[1,2]
  TN <- CM[1,1]
  FN <- CM[2,1]

  acc <- (TP+TN) / (TP+TN+FN+FP)
  err <- (FP+FN) / (TP+TN+FN+FP)
  pre <- (TP) / (TP+FP)
  sen <- (TP) / (TP+FN)
  spe <- (TN) / (TN+FP)
  fme <- (2*pre*sen) / (pre+sen)
```

```r
  mcc_denom <- sqrt(TP+FP)*sqrt(TP+FN)*sqrt(TN+FP)*sqrt(TN+FN)
  mcc <- (TP*TN - FP*FN) / mcc_denom

  name <- c("accuracy", "error rate", "precision", "sensitivity", "specificity", "F-measure", "Matthew's
  value <- c(acc, err, pre, sen, spe, fme, mcc)
  stats <- data.frame(name, value)

  return (stats)
}
```

```r
#evaluate Matthew's Correlation Coefficient using Hannah's stats equation
confMat <- table(pihatcv$obs, pihatcv$pred)
confMat
```

```
##
##         False True
##   False  2574   70
##   True     93  341
```

```r
rf.stats <- get_stats(confMat)
rf.stats
```

```
##           name      value
## 1     accuracy 0.94704353
## 2   error rate 0.05295647
## 3    precision 0.82968370
## 4  sensitivity 0.78571429
## 5  specificity 0.97352496
## 6    F-measure 0.80710059
## 7 Matthew's CC 0.77682255
```