DATA621 Homework 3

Joseph Simone, Jack Russo, Javern Wilson, Paul Perez

3/22/2020

Contents

Overview

In this homework assignment, we will explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).

Our objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. We will provide classifications and probabilities for the evaluation data set using our binary logistic regression model.

We will only use the variables provided to us (or variables that we derived from the variables given). Below is a short description of the variables of interest in the data set:

- `zn`: proportion of residential land zoned for large lots (over 25000 square feet) (predictor variable)
- `chas`: a dummy var. for whether the suburb borders the Charles River (1) or not (0) (predictor variable)
- `nox`: nitrogen oxides concentration (parts per 10 million) (predictor variable)
- `rm`: average number of rooms per dwelling (predictor variable)
- `age`: proportion of owner-occupied units built prior to 1940 (predictor variable)
- `dis`: weighted mean of distances to five Boston employment centers (predictor variable)
- `rad`: index of accessibility to radial highways (predictor variable)
- `tax`: full-value property-tax rate per $10,000 (predictor variable)
- `ptratio`: pupil-teacher ratio by town (predictor variable)
- `black`: 1000(Bk - 0.63)2 where Bk is the proportion of blacks by town (predictor variable)
- `lstat`: lower status of the population (percent) (predictor variable)
- `medv`: median value of owner-occupied homes in $1000s (predictor variable)
- `target`: whether the crime rate is above the median crime rate (1) or not (0) (response variable)

Data Import and Preview

zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv	target
0	19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	3.70	50.0	1
0	19.58	1	0.871	5.403	100.0	1.3216	5	403	14.7	26.82	13.4	1
0	18.10	0	0.740	6.485	100.0	1.9784	24	666	20.2	18.85	15.4	1
30	4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	5.19	23.7	0
0	2.46	0	0.488	7.155	92.2	2.7006	3	193	17.8	4.82	37.9	0
0	8.56	0	0.520	6.781	71.3	2.8561	5	384	20.9	7.67	26.5	0

EDA

Number of Target Variables

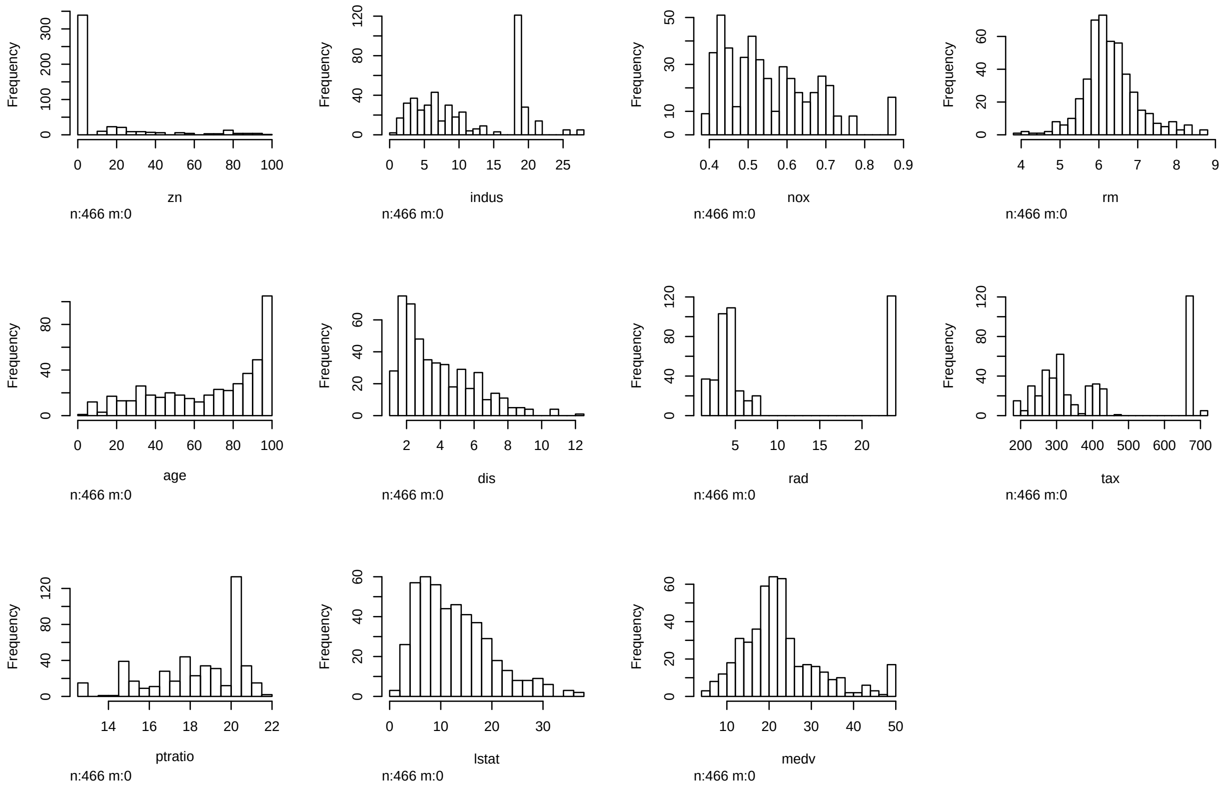
```
##
##   0   1
## 237 229
```

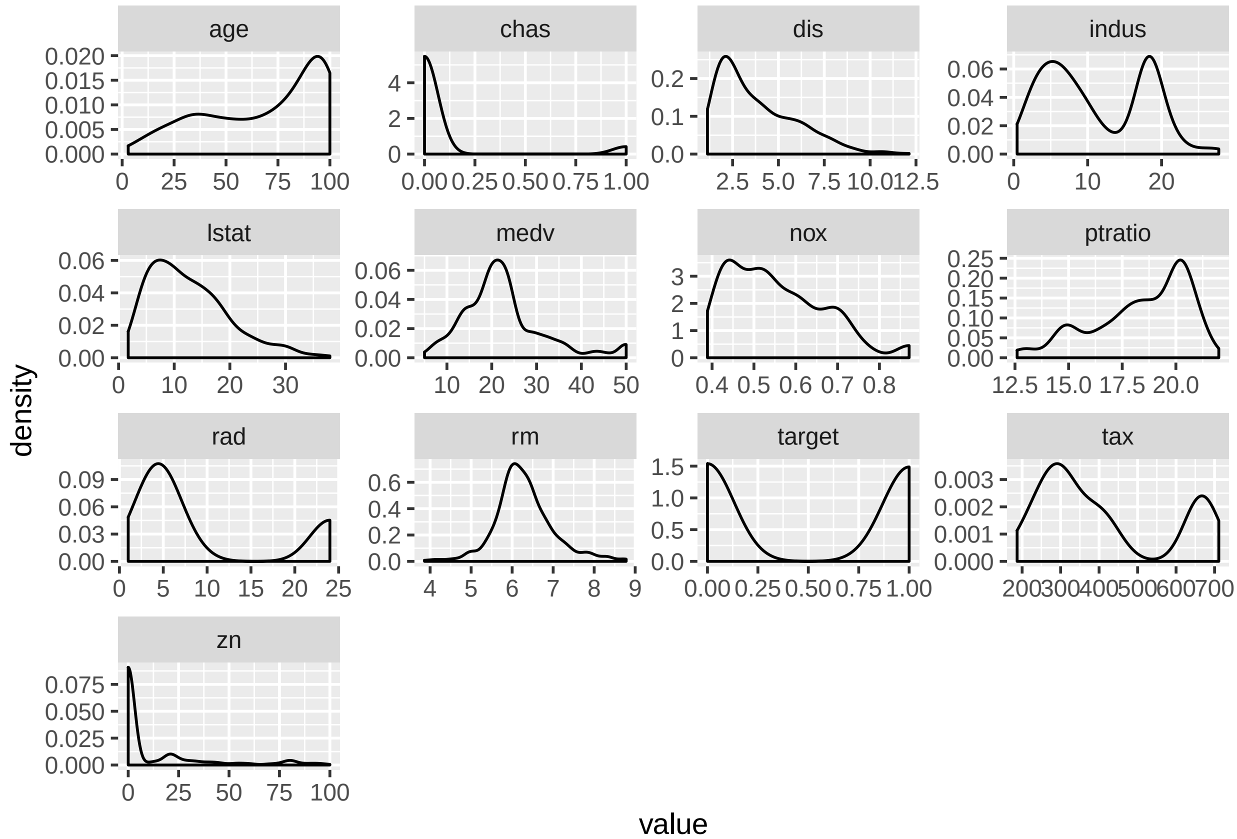
Dataset Summaries

	zn	indus	chas	nox	rm	age
	Min. : 0.00	Min. : 0.460	Min. :0.00000	Min. :0.3890	Min. :3.863	Min. : 2.90
	1st Qu.: 0.00	1st Qu.: 5.145	1st Qu.:0.00000	1st Qu.:0.4480	1st Qu.:5.887	1st Qu.: 43.88
	Median : 0.00	Median : 9.690	Median :0.00000	Median :0.5380	Median :6.210	Median : 77.15
	Mean : 11.58	Mean :11.105	Mean :0.07082	Mean :0.5543	Mean :6.291	Mean : 68.37
	3rd Qu.: 16.25	3rd Qu.:18.100	3rd Qu.:0.00000	3rd Qu.:0.6240	3rd Qu.:6.630	3rd Qu.: 94.10
	Max. :100.00	Max. :27.740	Max. :1.00000	Max. :0.8710	Max. :8.780	Max. :100.00

	dis	rad	tax	ptratio	lstat	medv
	Min. : 1.130	Min. : 1.00	Min. :187.0	Min. :12.6	Min. : 1.730	Min. : 5.00
	1st Qu.: 2.101	1st Qu.: 4.00	1st Qu.:281.0	1st Qu.:16.9	1st Qu.: 7.043	1st Qu.:17.02
	Median : 3.191	Median : 5.00	Median :334.5	Median :18.9	Median :11.350	Median :21.20
	Mean : 3.796	Mean : 9.53	Mean :409.5	Mean :18.4	Mean :12.631	Mean :22.59
	3rd Qu.: 5.215	3rd Qu.:24.00	3rd Qu.:666.0	3rd Qu.:20.2	3rd Qu.:16.930	3rd Qu.:25.00
	Max. :12.127	Max. :24.00	Max. :711.0	Max. :22.0	Max. :37.970	Max. :50.00

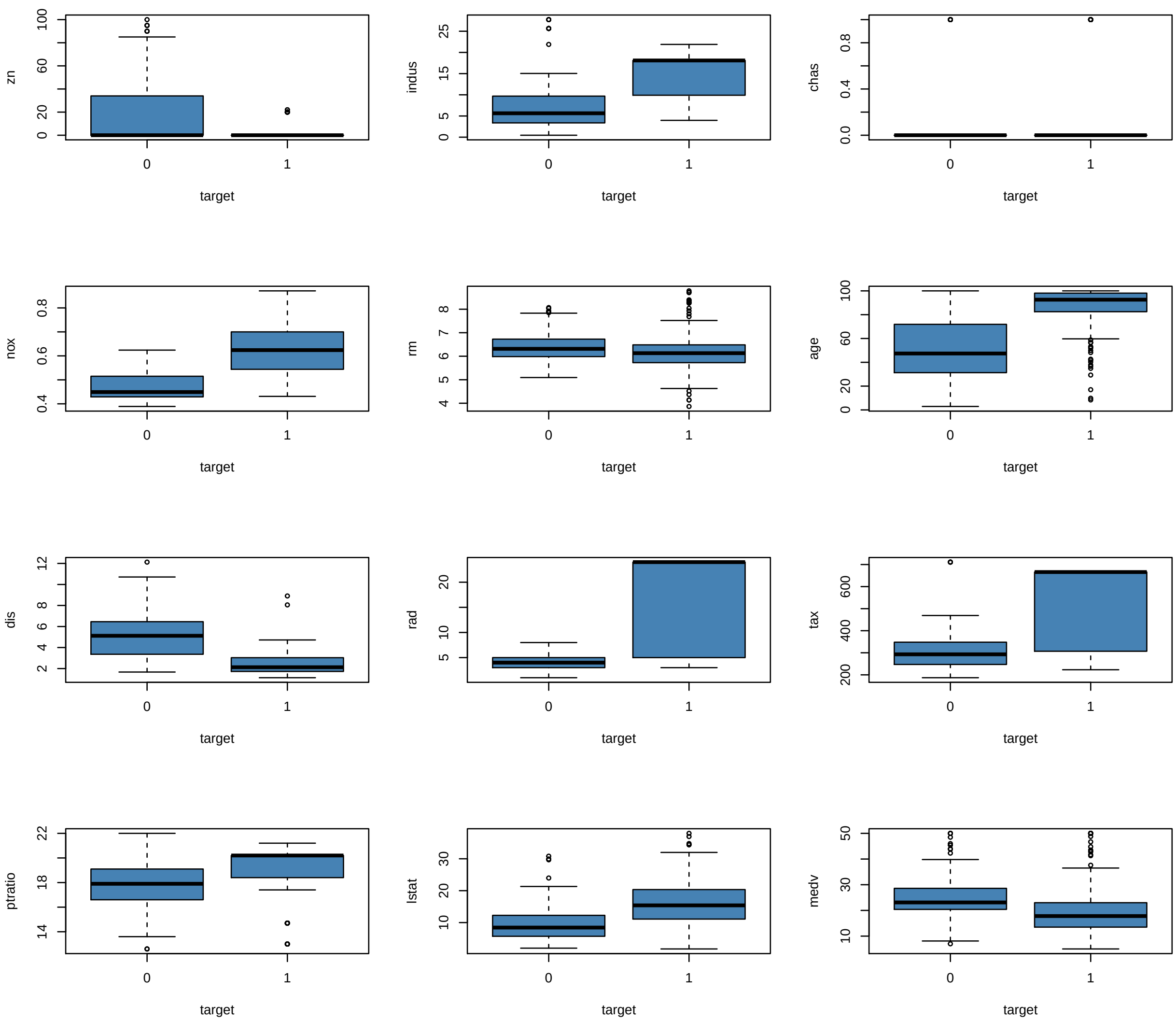
Distribution of Predictors



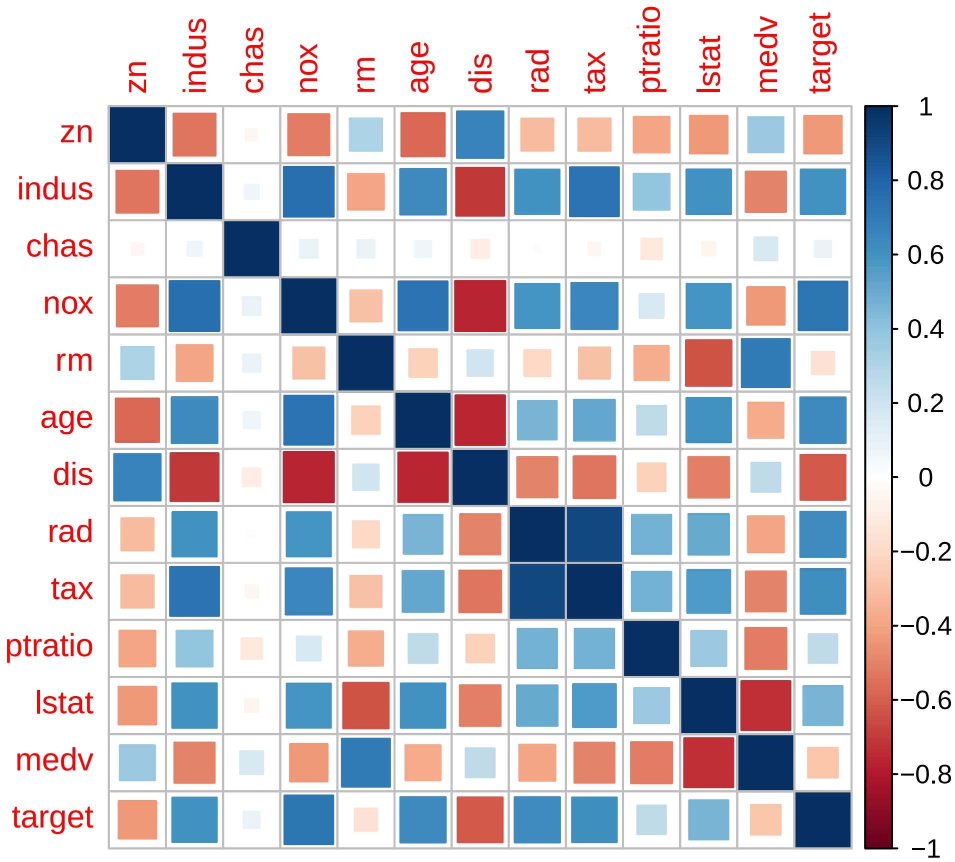


Boxplot

Of each continuous independent variable with target.



From the above plots, we can infer that the crime rate is above the median when majority of the predictors are high. For instance, have a look at `nox` (nitrogen oxide) and `tax` (property tax).




```
##
##  Pearson's product-moment correlation
##
## data:  crime$rad and crime$tax
## t = 46.239, df = 464, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8888115 0.9214292
## sample estimates:
##       cor
## 0.9064632
```

DATA PREPARATION

Missing Cases

```
## [1] FALSE
```

There does not appear to be missing cases.

Scale Data

Use scale function to scale all variables to mean and standard devatiotion of target variable.

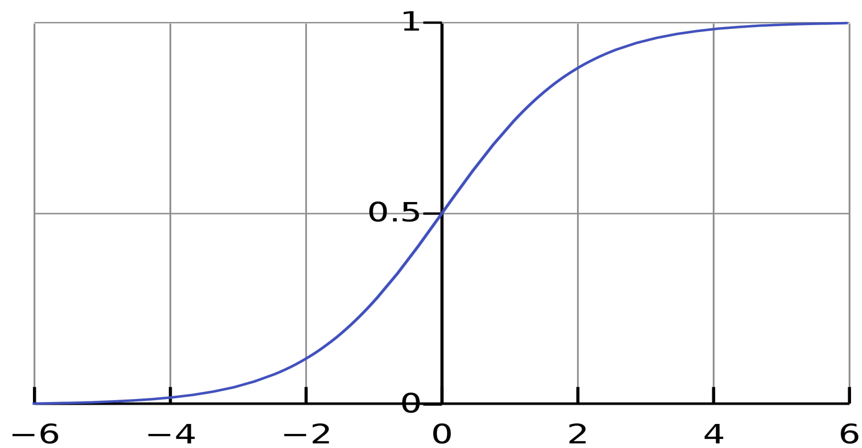
```
## [1] 0.4914163

## [1] 0.5004636

##    target        zn     indus      chas       nox        rm       age       dis
## 0.4914163 0.4914163 0.4914163 0.4914163 0.4914163 0.4914163 0.4914163 0.4914163
##       rad       tax   ptratio     lstat      medv
## 0.4914163 0.4914163 0.4914163 0.4914163 0.4914163

##    target        zn     indus      chas       nox        rm       age       dis
## 0.5004636 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
##       rad       tax   ptratio     lstat      medv
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
```

After scaling, the predictor and responce variables posses approximately equal means and standard deviations.

Sigmoid Function

We will use sigmoid function to scale all variables between zero and one.

The following transformation will map each of the values on to the Logistic curve. This will allow us to construct a linear model on with the scaled data.



```
##    target        zn     indus      chas       nox        rm       age       dis
## 0.6135460 0.6110187 0.6132653 0.6102377 0.6127209 0.6135238 0.6149403 0.6123702
##       rad       tax   ptratio     lstat      medv
## 0.6119626 0.6125669 0.6153707 0.6125611 0.6124636
```

```
##     target         zn      indus       chas        nox         rm        age
## 0.11563640 0.09545625 0.11283836 0.08254726 0.10869567 0.10784268 0.11746022
##        dis        rad        tax    ptratio      lstat       medv
## 0.10676718 0.10725490 0.11004905 0.11728960 0.10721135 0.10502008
```

After applying the sigmoid function, the predictor and responce variables still retain approximately equal means and standard deviations.

Transform Predictors

Here all variables were transformed except the target variable. The variables were transformed using Box-Cox. In addition, the variables were scaled, centered and non-zero-variance values were removed (if any).

BUILD MODELS

Model 1

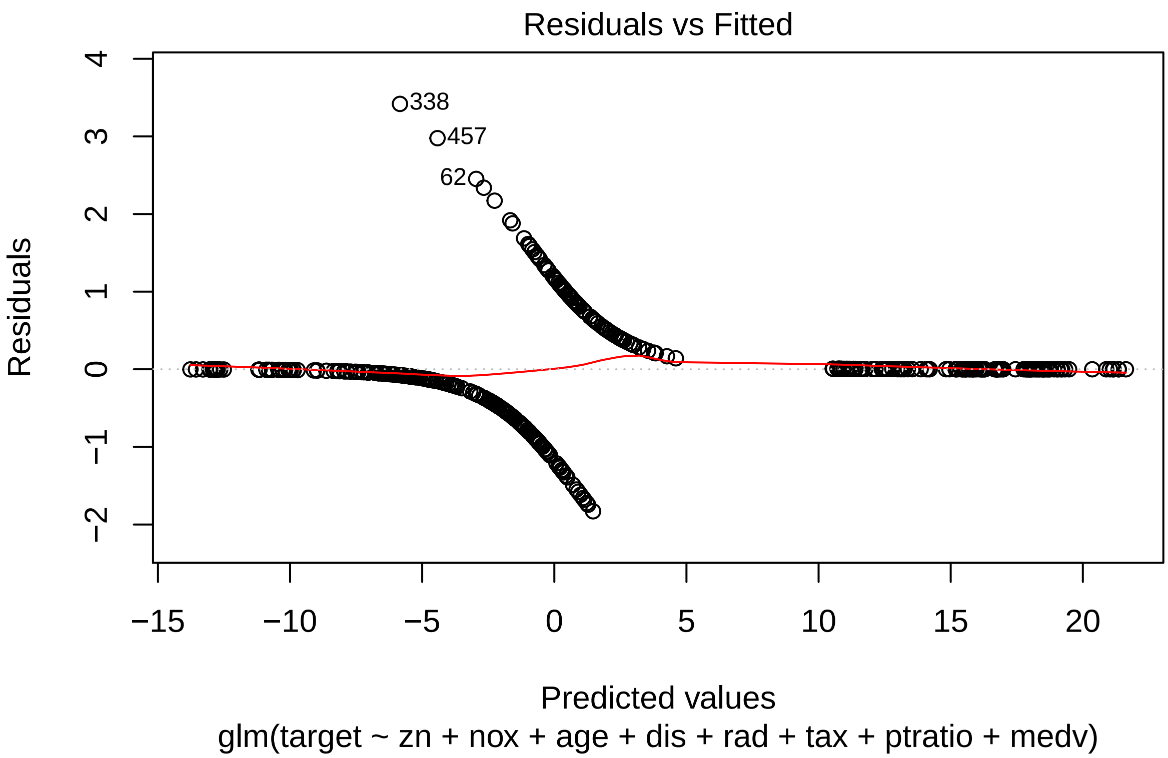
Build Binomial Regression Using The Scaled Data

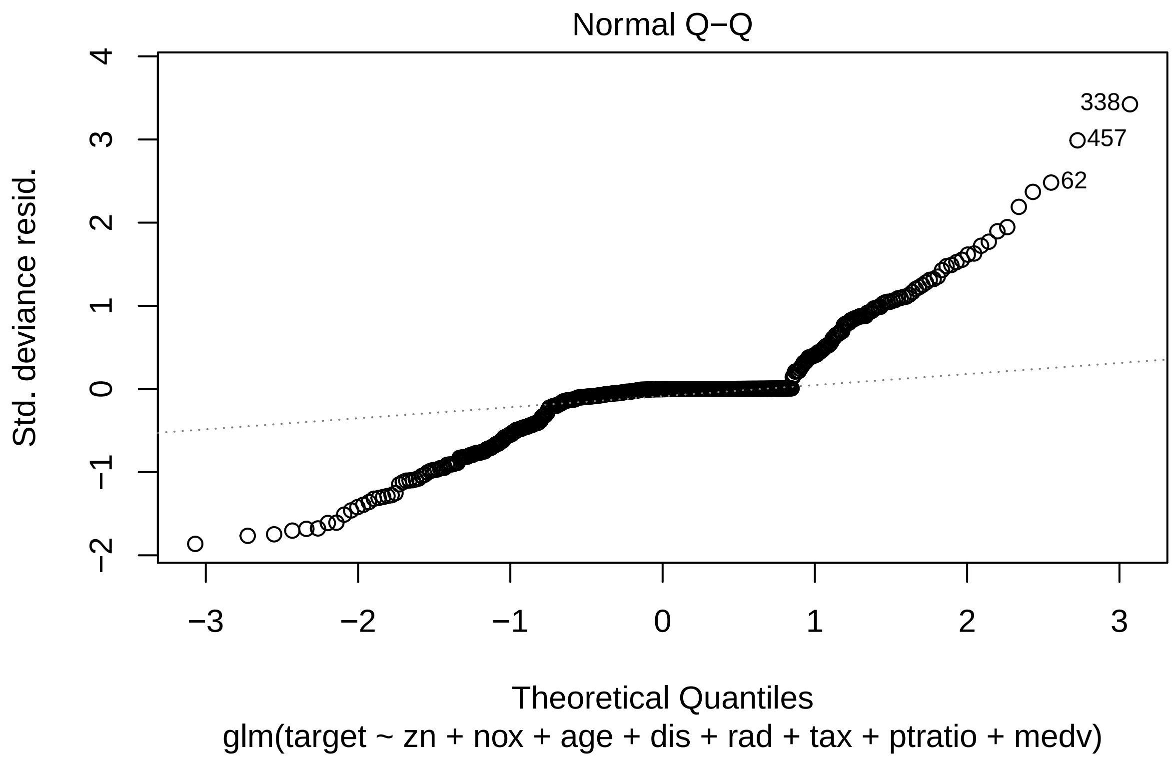
```
##
## Call:
## glm(formula = target ~ ., family = binomial, data = ntrain.scaled)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8464  -0.1445  -0.0017   0.0029   3.4665
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.1603     1.8624  -5.993 2.07e-09 ***
## zn           -3.0816     1.6195  -1.903  0.05706 .
## indus        -0.8847     0.6520  -1.357  0.17485
## chas          0.4678     0.3880   1.205  0.22803
## nox          11.4619     1.8507   6.193 5.90e-10 ***
## rm           -0.8282     1.0190  -0.813  0.41637
## age           1.9366     0.7825   2.475  0.01333 *
## dis           3.1126     0.9704   3.208  0.00134 **
## rad          11.5760     2.8343   4.084 4.42e-05 ***
## tax          -2.0724     0.9922  -2.089  0.03674 *
## ptratio       1.7687     0.5564   3.179  0.00148 **
## lstat         0.6515     0.7677   0.849  0.39608
## medv          3.3415     1.2620   2.648  0.00810 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 192.05  on 453  degrees of freedom
## AIC: 218.05
##
## Number of Fisher Scoring iterations: 9
```

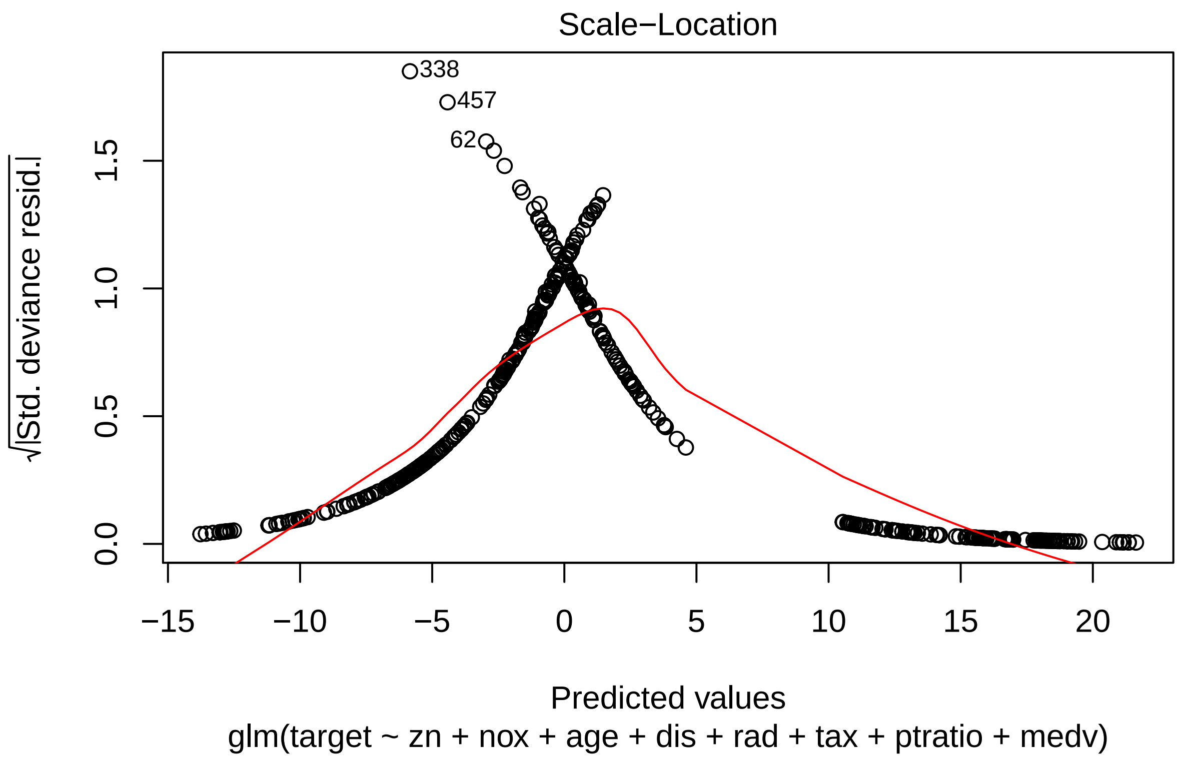
```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + tax + ptratio +
##     medv, family = binomial, data = ntrain.scaled)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8295  -0.1752  -0.0021   0.0032   3.4191
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.7770     1.5871  -6.160 7.26e-10 ***
## zn           -3.2079     1.4962  -2.144  0.03203 *
## nox           9.9885     1.5584   6.410 1.46e-10 ***
## age           1.8664     0.6203   3.009  0.00262 **
## dis           2.7597     0.9020   3.060  0.00222 **
## rad          12.5965     2.6021   4.841 1.29e-06 ***
## tax          -2.6045     0.8908  -2.924  0.00346 **
## ptratio       1.4219     0.4894   2.905  0.00367 **
## medv          2.0415     0.6550   3.117  0.00183 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 197.32  on 457  degrees of freedom
## AIC: 215.32
##
## Number of Fisher Scoring iterations: 9
```

After using stepAIC method, ‘stepAIC’ function, we are now left with eight independent variables which also resulted in having the minimum AIC value so far.

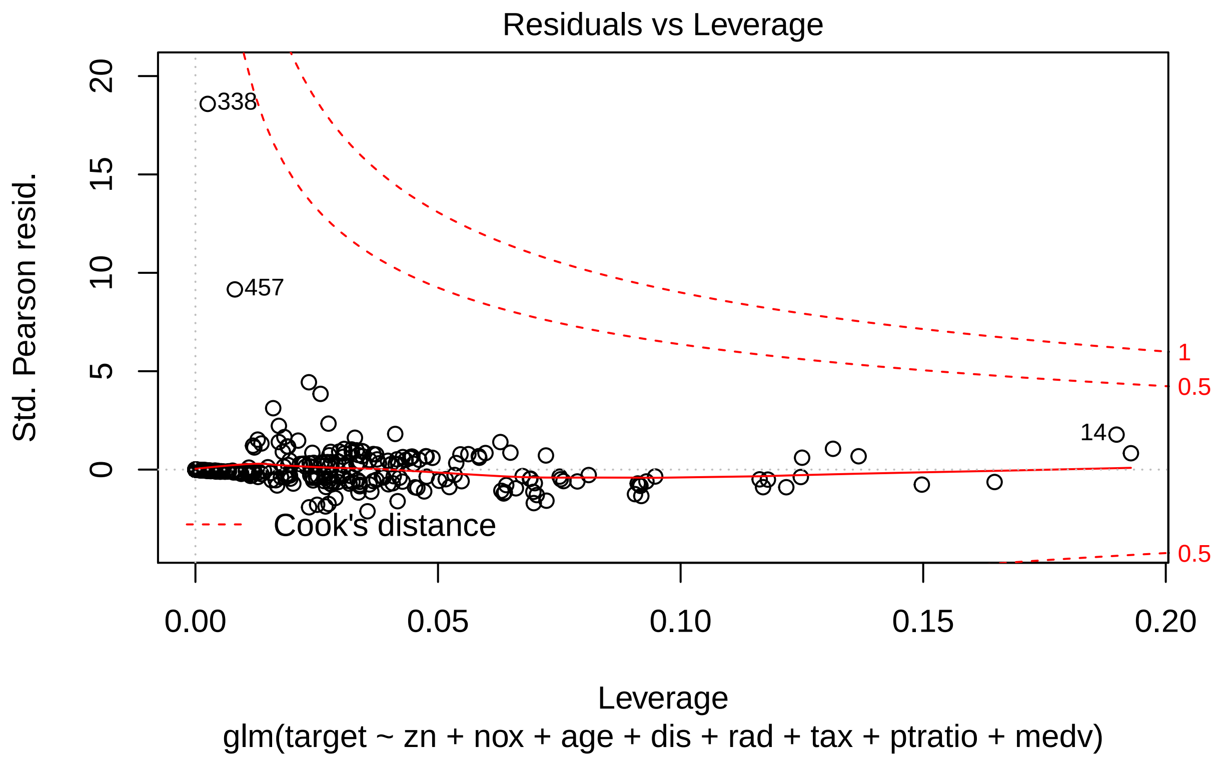

Residuals vs Fitted
Residuals
Predicted values
glm(target ~ zn + nox + age + dis + rad + tax + ptratio + medv)
338
457
62
4
3
2
1
0
−1
−2
−15
−10
−5
0
5
10
15
20



Normal Q−Q
Std. deviance resid.
4
3
2
1
0
−1
−2
338
457
62
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(target ~ zn + nox + age + dis + rad + tax + ptratio + medv)



Scale−Location
338
457
62
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
−15
−10
−5
0
5
10
15
20
Predicted values
glm(target ~ zn + nox + age + dis + rad + tax + ptratio + medv)




```
##       zn      nox      age      dis      rad      tax  ptratio     medv
## 1.789037 3.172660 1.701974 3.595939 1.697110 1.754274 1.865085 2.193689
```

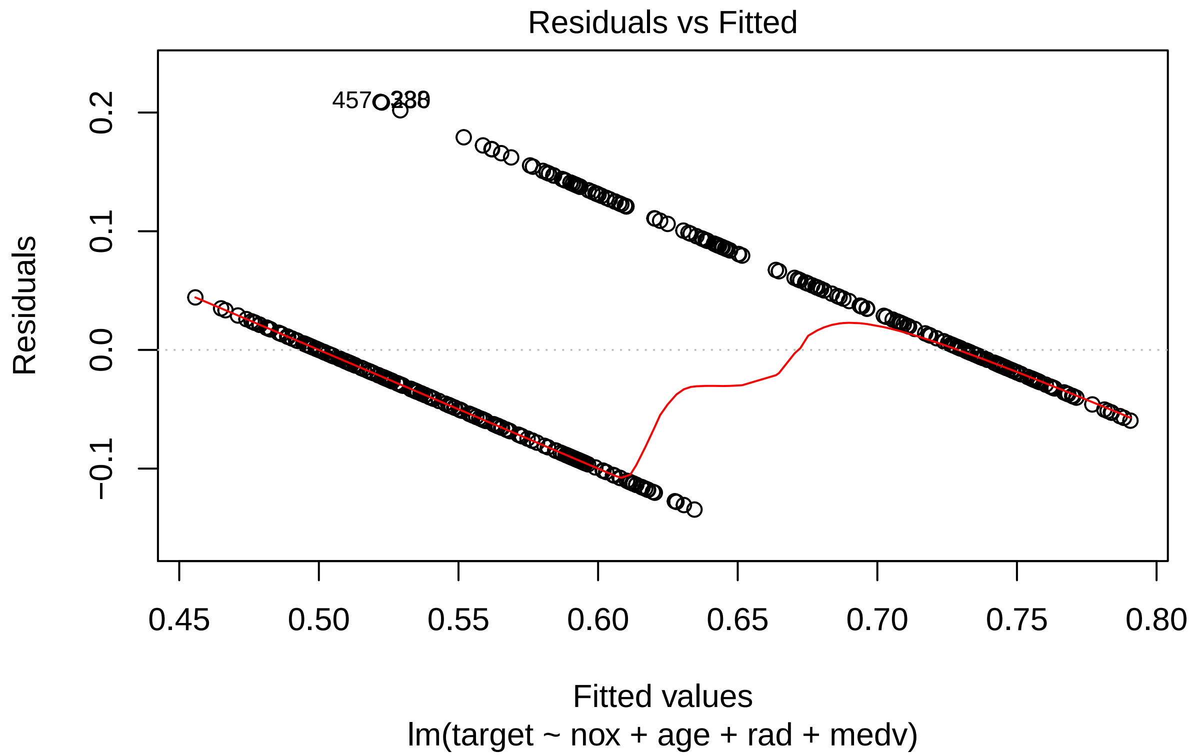
Model 2

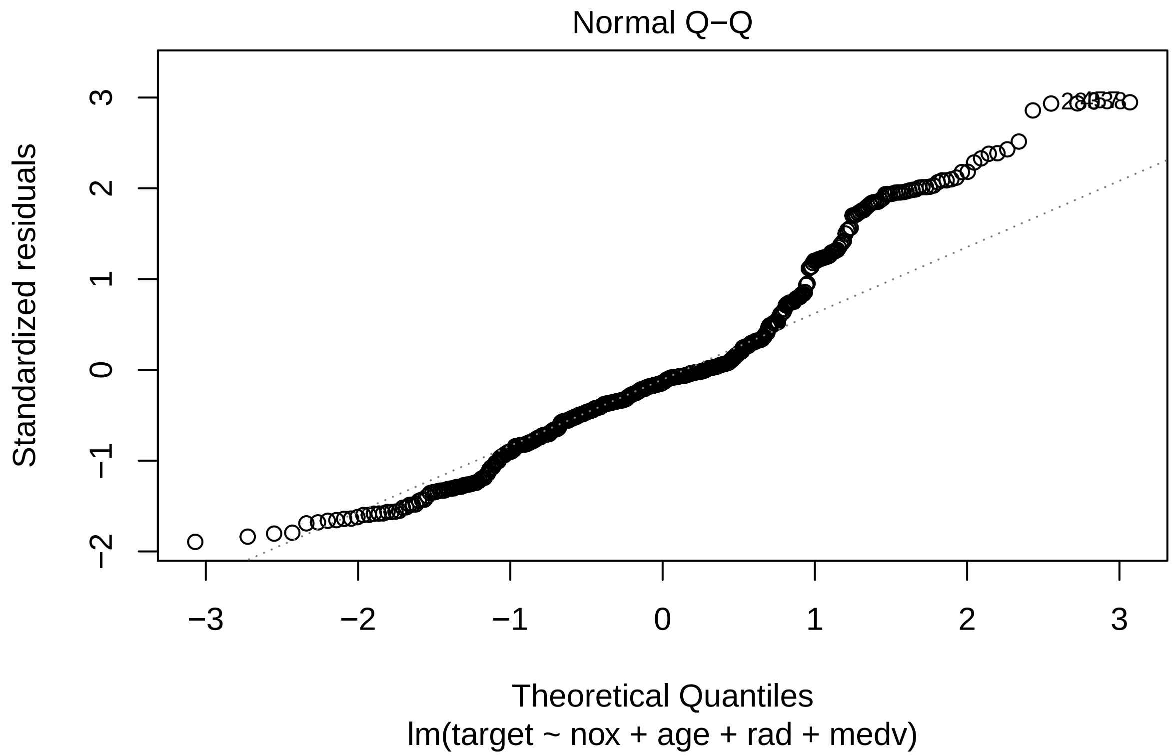
Build Linear Regression Using the Sigmoid Scaled Data.

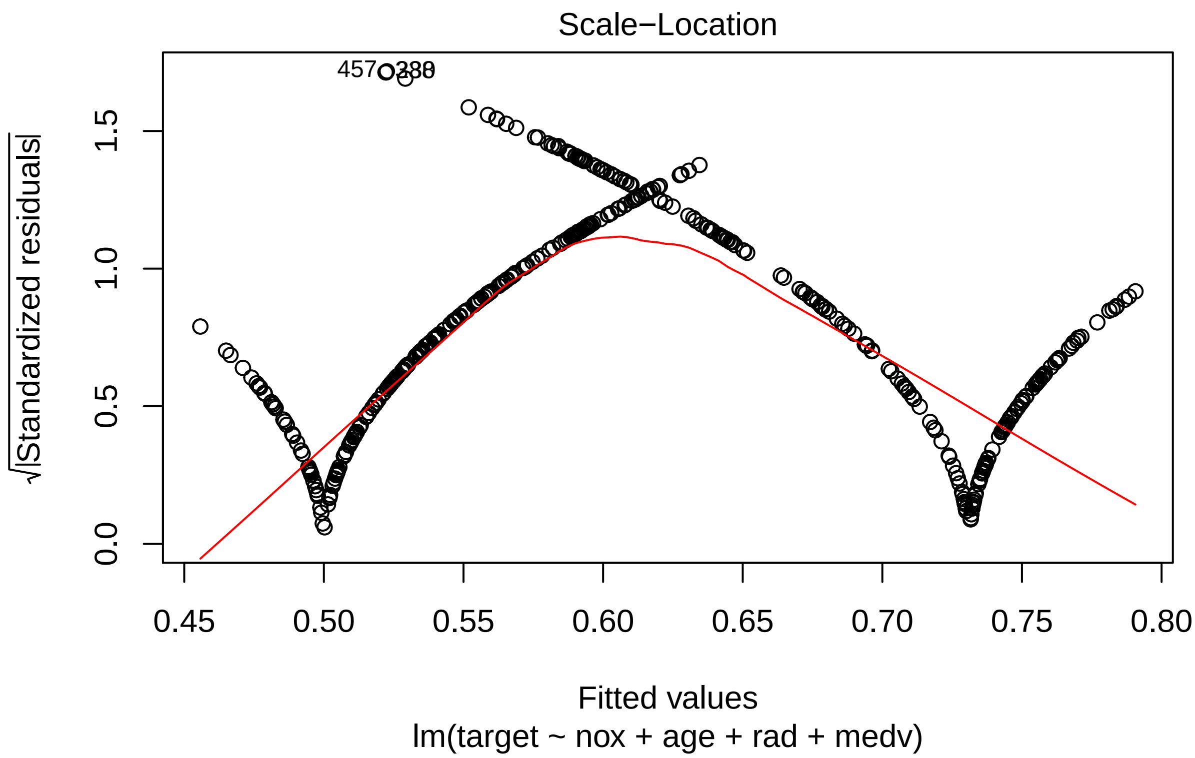
```
##
## Call:
## lm(formula = target ~ nox + age + rad + medv, data = ntrain.scaled.sigmoid)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.134511 -0.042567 -0.009906  0.027400  0.208979
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08854    0.04178  -2.120 0.034580 *
## nox          0.52443    0.05396   9.718  < 2e-16 ***
## age          0.16209    0.04349   3.727 0.000218 ***
## rad          0.32503    0.04035   8.056 6.83e-15 ***
## medv         0.13418    0.03697   3.629 0.000316 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

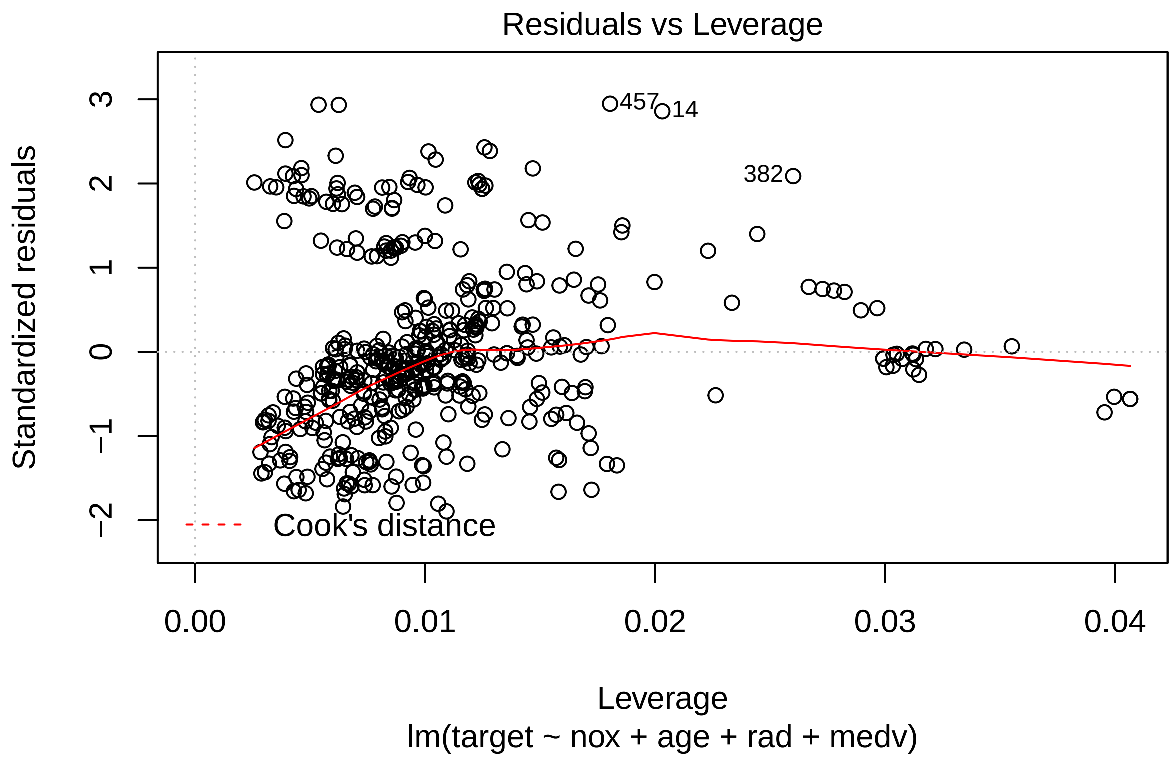
```
## Residual standard error: 0.07138 on 461 degrees of freedom
## Multiple R-squared:  0.6223, Adjusted R-squared:  0.619
## F-statistic: 189.9 on 4 and 461 DF,  p-value: < 2.2e-16
```

Using the sigmoid scaled data in a linear model has further contrained the relevant predictor variables and minimized the intercept (when compared to the binomial model). This suggests these adjusted variables contain most of the relevant information about this system.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target ~ nox + age + rad + medv)
−3
−2
−1
0
1
2
3


Scale−Location
457
380
289
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
Fitted values
lm(target ~ nox + age + rad + medv)




```
##      nox      age      rad     medv
## 3.140318 2.382030 1.709335 1.376141
```

Model 3

Build Linear Regression Using transformed Predictors.

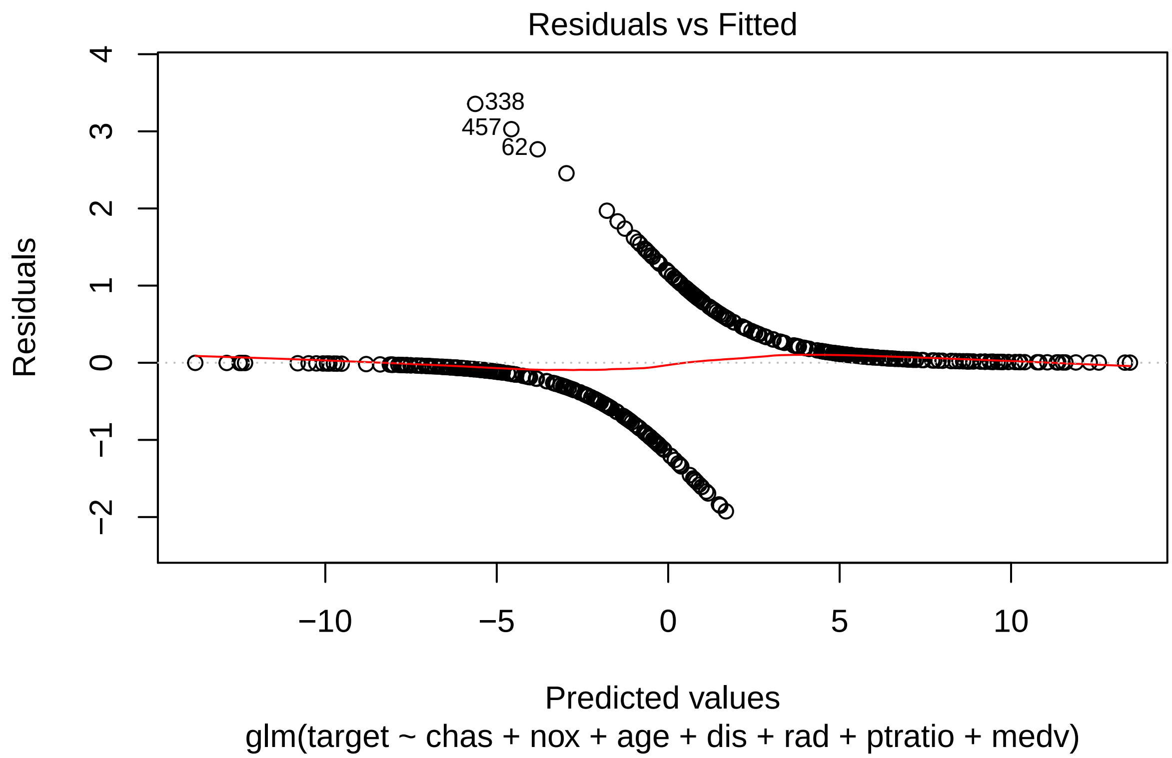
```
##
## Call:
## glm(formula = target ~ ., family = binomial, data = crime_trans)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9381  -0.1116  -0.0010   0.1137   3.4325
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.082410   0.316075  -0.261 0.794301
## zn          -0.519732   0.627379  -0.828 0.407433
## indus       -0.003522   0.379970  -0.009 0.992603
## chas         0.242925   0.195625   1.242 0.214316
## nox          4.889837   0.772981   6.326 2.52e-10 ***
## rm          -0.375393   0.453263  -0.828 0.407556
## age          1.154904   0.373442   3.093 0.001984 **
## dis          1.834862   0.469904   3.905 9.43e-05 ***
```

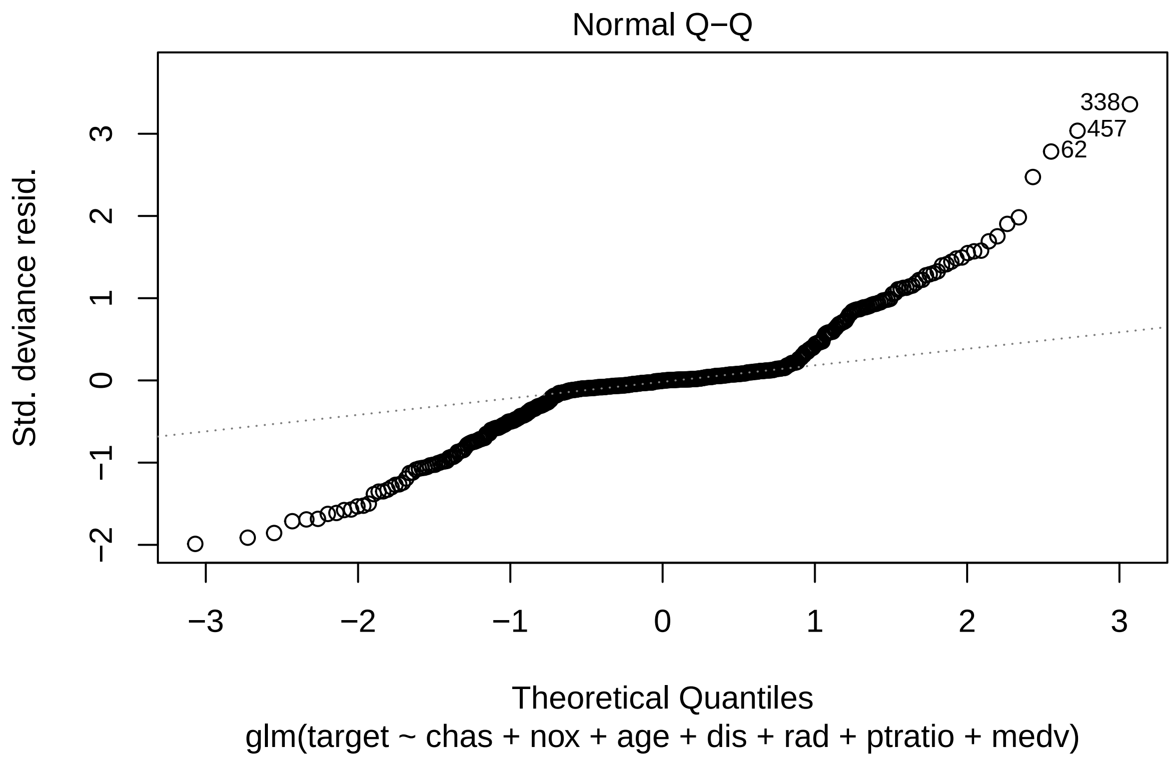
```
## rad          2.728346   0.634460   4.300 1.71e-05 ***
## tax         -0.322880   0.408076  -0.791 0.428812
## ptratio      0.979115   0.277238   3.532 0.000413 ***
## lstat       -0.048931   0.424222  -0.115 0.908173
## medv         1.845058   0.642208   2.873 0.004066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 196.79  on 453  degrees of freedom
## AIC: 222.79
##
## Number of Fisher Scoring iterations: 8
```

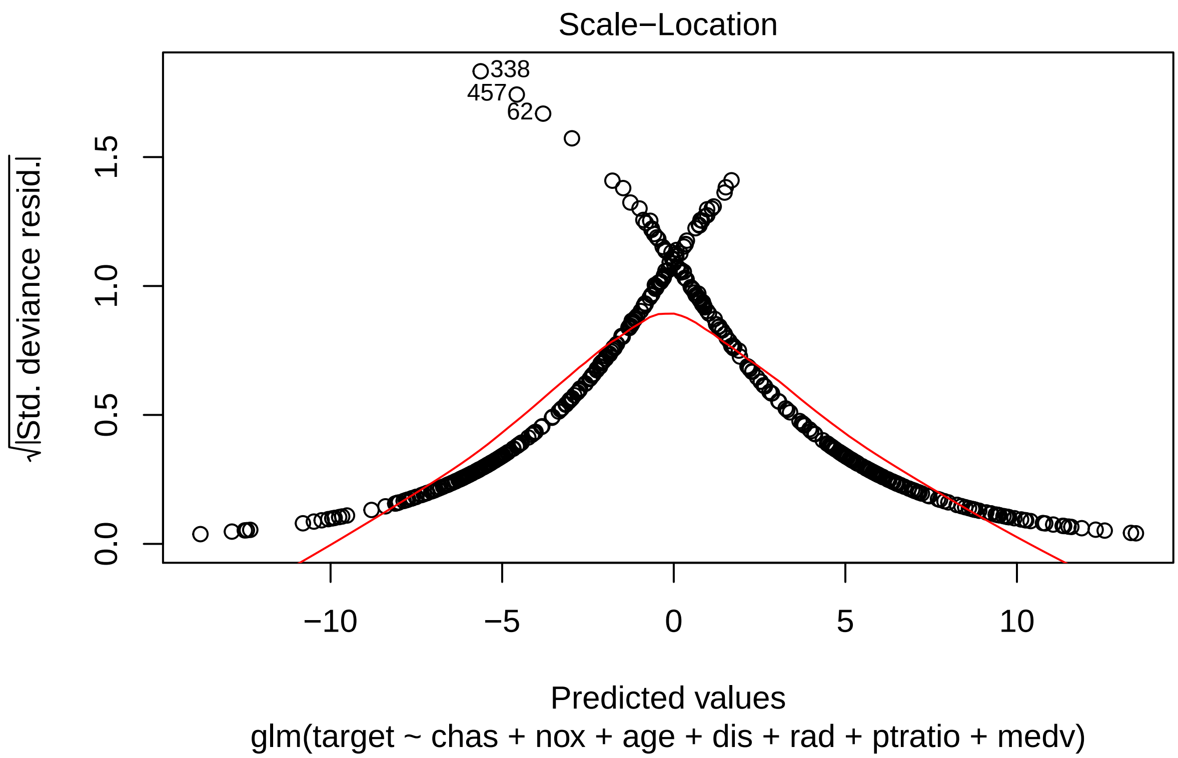
```
##
## Call:
## glm(formula = target ~ chas + nox + age + dis + rad + ptratio +
##     medv, family = binomial, data = crime_trans)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9253  -0.1524  -0.0029   0.1179   3.3556
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1509     0.2265   0.666  0.50533
## chas          0.3094     0.1779   1.739  0.08200 .
## nox           4.8640     0.7230   6.728 1.73e-11 ***
## age           1.0055     0.3060   3.286  0.00102 **
## dis           1.8161     0.4323   4.201 2.66e-05 ***
## rad           2.2998     0.4433   5.188 2.12e-07 ***
## ptratio       0.9617     0.2432   3.954 7.67e-05 ***
## medv          1.5353     0.3609   4.254 2.10e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 199.53  on 458  degrees of freedom
## AIC: 215.53
##
## Number of Fisher Scoring iterations: 8
```

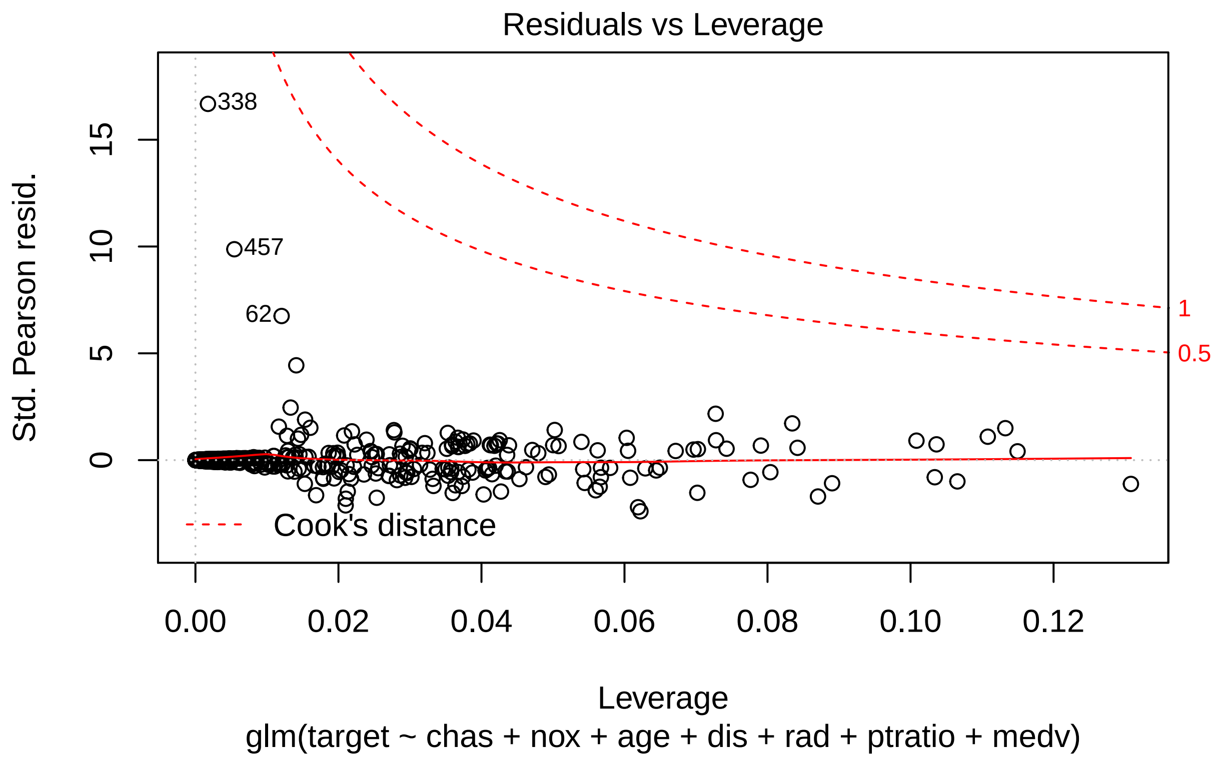
Results show that model did not improve with just transforming the predictors, even after using the stepAIC function. However, this model would be considered less complex as it has 1 less predictor than model 1 and model 2.

```
##     chas      nox      age      dis      rad  ptratio     medv
## 1.114122 3.738991 1.862143 3.338080 1.113786 1.889139 2.520938
```


Residuals vs Fitted
Residuals
Predicted values
glm(target ~ chas + nox + age + dis + rad + ptratio + medv)
4
3
2
1
0
−1
−2
−10
−5
0
5
10
338
457
62


Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
338
457
62
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(target ~ chas + nox + age + dis + rad + ptratio + medv)


Scale−Location
338
457
62
$\sqrt{|\text{Std. deviance resid.}|}$
0.0
0.5
1.0
1.5
−10
−5
0
5
10
Predicted values
glm(target ~ chas + nox + age + dis + rad + ptratio + medv)



Model 4

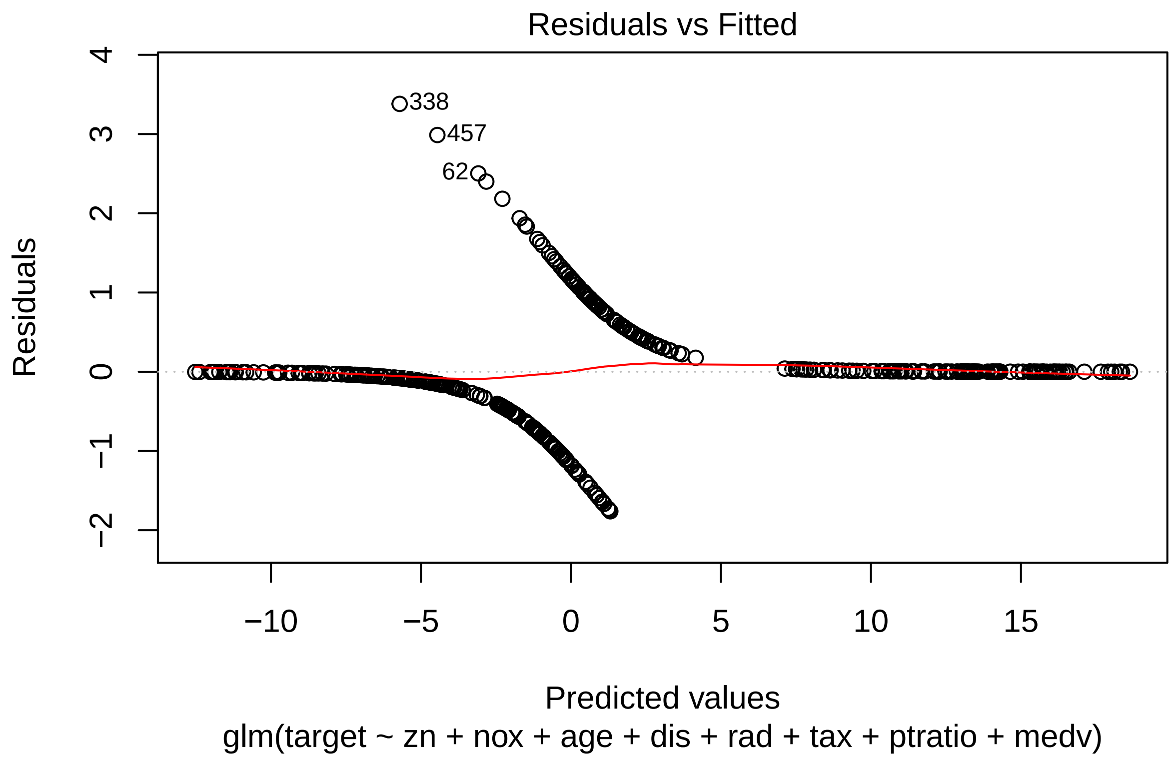
Build Generalized Linear Regression Using Sigmoid Scaled Data.

```
##
## Call:
## glm(formula = target ~ ., family = binomial, data = ntrain.scaled.sigmoid2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8536  -0.1490  -0.0032   0.0128   3.4196
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -73.414     12.477  -5.884 4.00e-09 ***
## zn             -10.849      6.394  -1.697 0.089724 .
## indus           -3.328      3.023  -1.101 0.270913
## chas             2.711      2.353   1.152 0.249269
## nox             49.003      7.861   6.234 4.55e-10 ***
## rm              -4.482      4.403  -1.018 0.308709
## age              8.887      3.337   2.664 0.007732 **
## dis             15.312      4.424   3.461 0.000537 ***
## rad             45.586     11.386   4.004 6.24e-05 ***
## tax             -7.362      4.204  -1.751 0.079923 .
## ptratio          7.755      2.445   3.172 0.001512 **
```

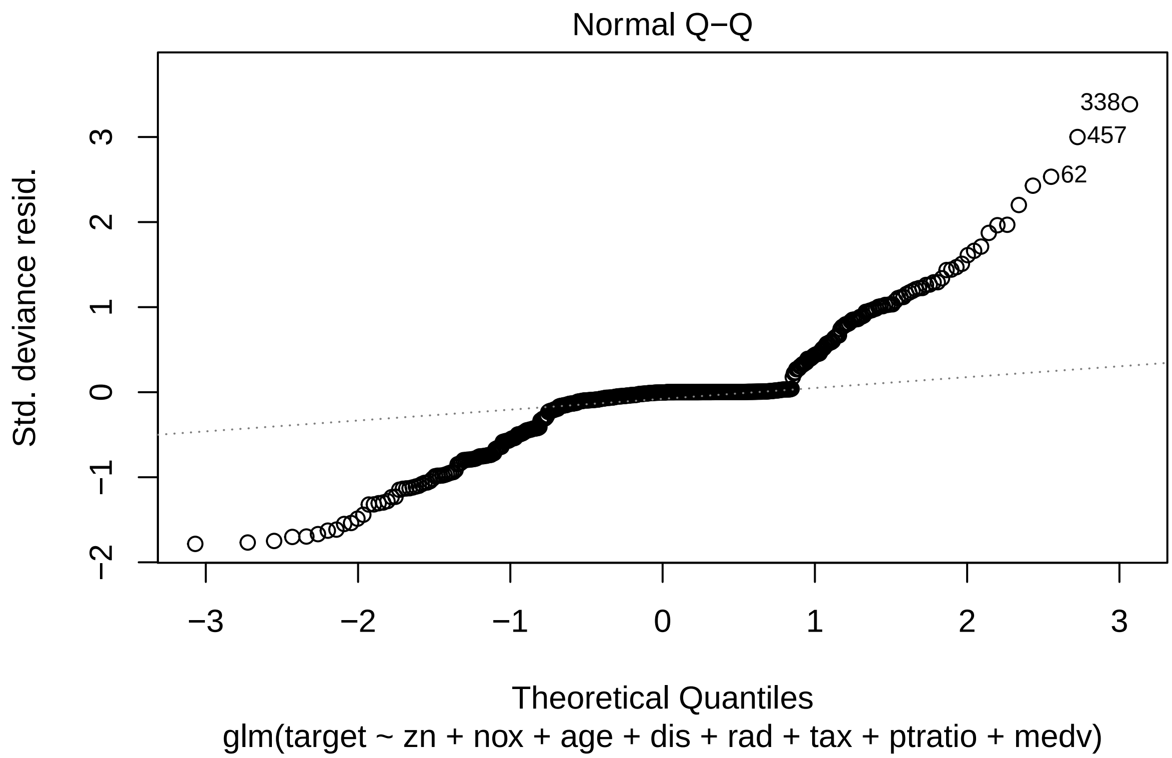
```
## lstat          2.348      3.636   0.646 0.518439
## medv          16.961      5.923   2.863 0.004192 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 195.93  on 453  degrees of freedom
## AIC: 221.93
##
## Number of Fisher Scoring iterations: 9
```

```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + tax + ptratio +
##     medv, family = binomial, data = ntrain.scaled.sigmoid2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7611  -0.1632  -0.0036   0.0080   3.3806
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -64.719     10.770  -6.009 1.87e-09 ***
## zn           -11.852      5.947  -1.993  0.04627 *
## nox           43.260      6.654   6.502 7.95e-11 ***
## age            8.042      2.684   2.996  0.00273 **
## dis           13.442      4.135   3.251  0.00115 **
## rad           49.279     10.579   4.658 3.19e-06 ***
## tax           -9.616      3.811  -2.523  0.01162 *
## ptratio        6.000      2.139   2.806  0.00502 **
## medv          10.007      3.292   3.040  0.00236 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 200.56  on 457  degrees of freedom
## AIC: 218.56
##
## Number of Fisher Scoring iterations: 8
```

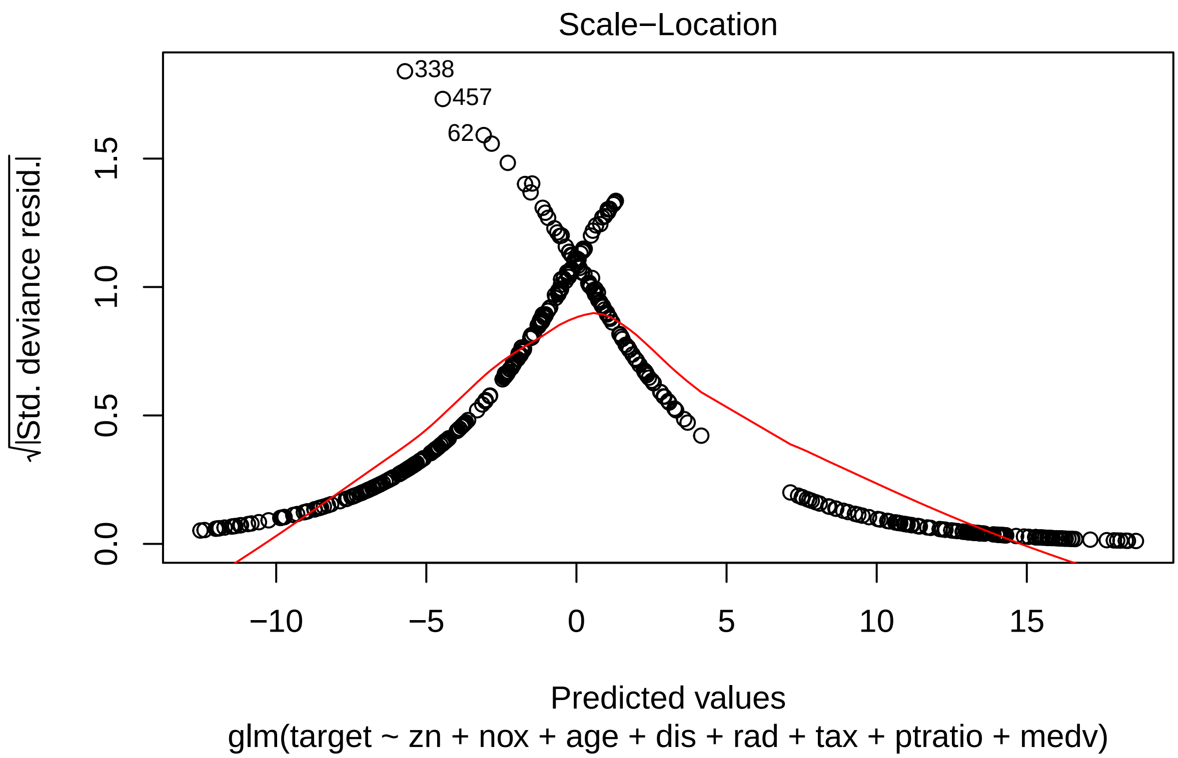
```
##       zn      nox      age      dis      rad      tax  ptratio     medv
## 1.612196 3.354050 1.746804 3.611908 1.709746 1.875620 1.983801 2.466023
```

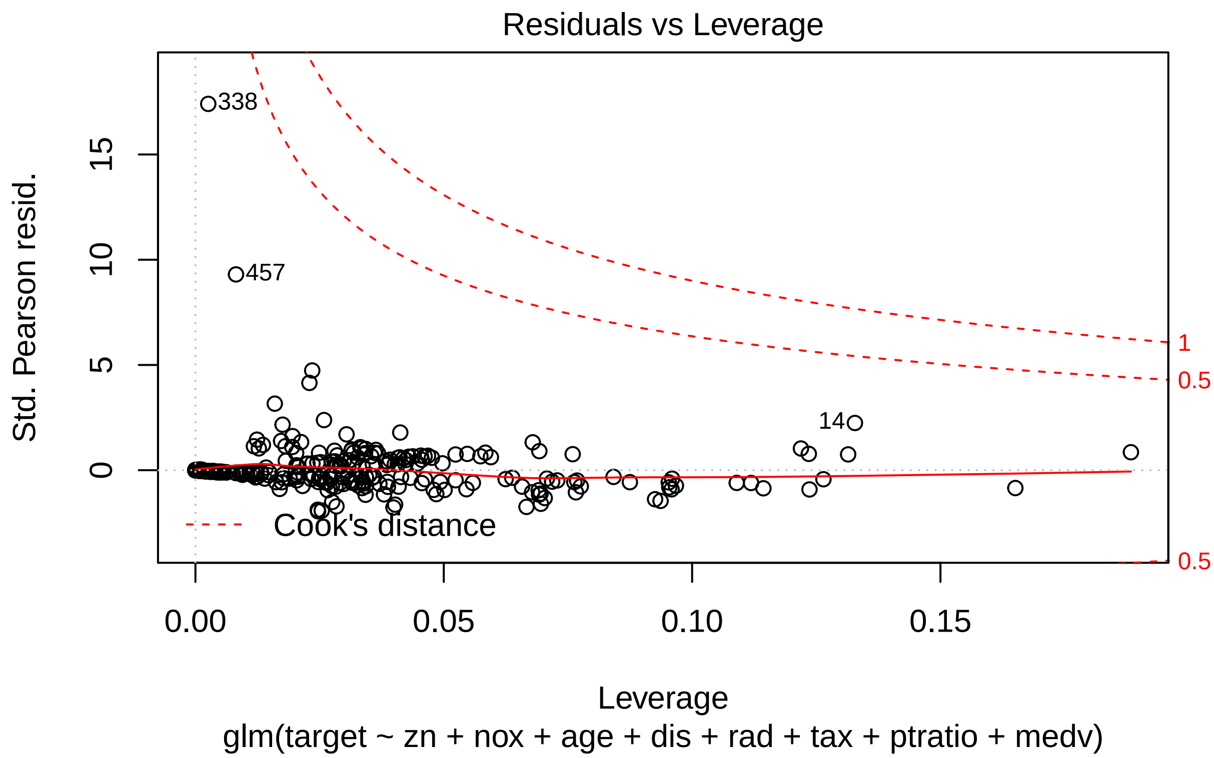


Residuals vs Fitted
338
457
62
Residuals
Predicted values
glm(target ~ zn + nox + age + dis + rad + tax + ptratio + medv)
4
3
2
1
0
−1
−2
−10
−5
0
5
10
15



Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(target ~ zn + nox + age + dis + rad + tax + ptratio + medv)
338
457
62
−3
−2
−1
0
1
2
3


Scale−Location
338
457
62
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
−10
−5
0
5
10
15
Predicted values
glm(target ~ zn + nox + age + dis + rad + tax + ptratio + medv)



ANOVA

Model 2 is built on a different scale than the other 3 models, so it cannot be compared with them under the anova function. However we'll show what it looks like in that case.

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      465     645.88
## zn       1  127.411       464     518.46 < 2.2e-16 ***
## nox      1  230.177       463     288.29 < 2.2e-16 ***
## age      1    0.767       462     287.52 0.3810001
## dis      1    4.296       461     283.22 0.0382133 *
## rad      1   55.953       460     227.27 7.423e-14 ***
## tax      1   15.916       459     211.35 6.620e-05 ***
## ptratio  1    2.706       458     208.65 0.0999454 .
## medv     1   11.326       457     197.32 0.0007644 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      465     645.88
## chas     1     3.02       464     642.86   0.08244 .
## nox      1   354.53       463     288.33 < 2.2e-16 ***
## age      1     0.62       462     287.71   0.43151
## dis      1     6.40       461     281.31   0.01141 *
## rad      1    56.08       460     225.24 6.971e-14 ***
## ptratio  1     2.71       459     222.53   0.09999 .
## medv     1    23.00       458     199.53 1.620e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      465     645.88
## zn       1  126.485       464     519.39 < 2.2e-16 ***
## nox      1  230.091       463     289.30 < 2.2e-16 ***
## age      1    0.651       462     288.65 0.4195877
## dis      1    4.994       461     283.65 0.0254332 *
## rad      1   56.916       460     226.74 4.549e-14 ***
## tax      1   13.462       459     213.28 0.0002434 ***
## ptratio  1    2.195       458     211.08 0.1384305
## medv     1   10.516       457     200.56 0.0011832 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA (Comparing the other 3 models)

```
## Analysis of Deviance Table
##
## Model 1: target ~ zn + nox + age + dis + rad + tax + ptratio + medv
## Model 2: target ~ chas + nox + age + dis + rad + ptratio + medv
## Model 3: target ~ zn + nox + age + dis + rad + tax + ptratio + medv
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       457     197.32
## 2       458     199.53 -1  -2.2080   0.1373
## 3       457     200.56  1  -1.0329
```

Model one does seem to perform better than the other three models.

SELECT MODELS

```
##            Actual
## Predictions   0   1
##           0 218  22
##           1  19 207
```

Metrics

ACCURACY Accuracy can be defined as the fraction of predicitons our model got right. Also known as the error rate, the accuracy rate makes no distinction about the type of error being made.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

CLASSIFICATION ERROR RATE The Classification Error Rate calculates the number of incorrect predictions out of the total number of predictions in the dataset.

$$\text{Classification Error Rate} = \frac{FP + FN}{TP + FP + TN + FN}$$

PRECISION This is the positive value or the fraction of the positive predictions that are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

SENSITIVITY The sensitivity is sometimes considered the true positive rate since it measures the accuracy in the event population.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

SPECIFICITY This is the true negatitive rate or the proportion of negatives that are correctly identified.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

F1 SCORE OF PREDICTIONS The F1 Score of Predictions measures the test's accuracy, on a scale of 0 to 1 where a value of 1 is the most accurate and the value of 0 is the least accurate.

$$\text{F1 Score} = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$$

F1 SCORE BOUNDS

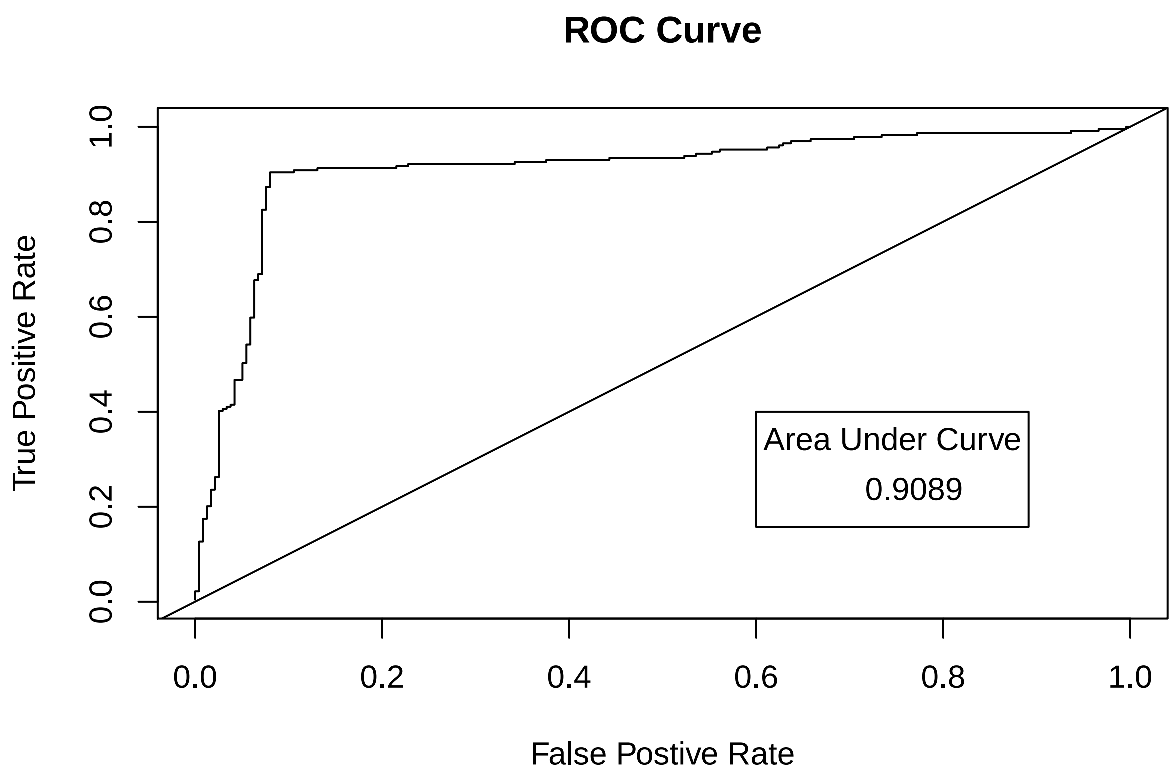
```
## [1] 0
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00658 0.23356 0.39955 0.39468 0.57036 0.91824
```

ROC CURVE Shows how the true positive rate against the false positive rate at various threshold settings. The AUC (Area Under Curve) tells how much model is capable of distinguishing between classes. Higher the AUC is better, that is, how well the model is at predicting 0s as 0s and 1s as 1s.

Creating an ROC Function

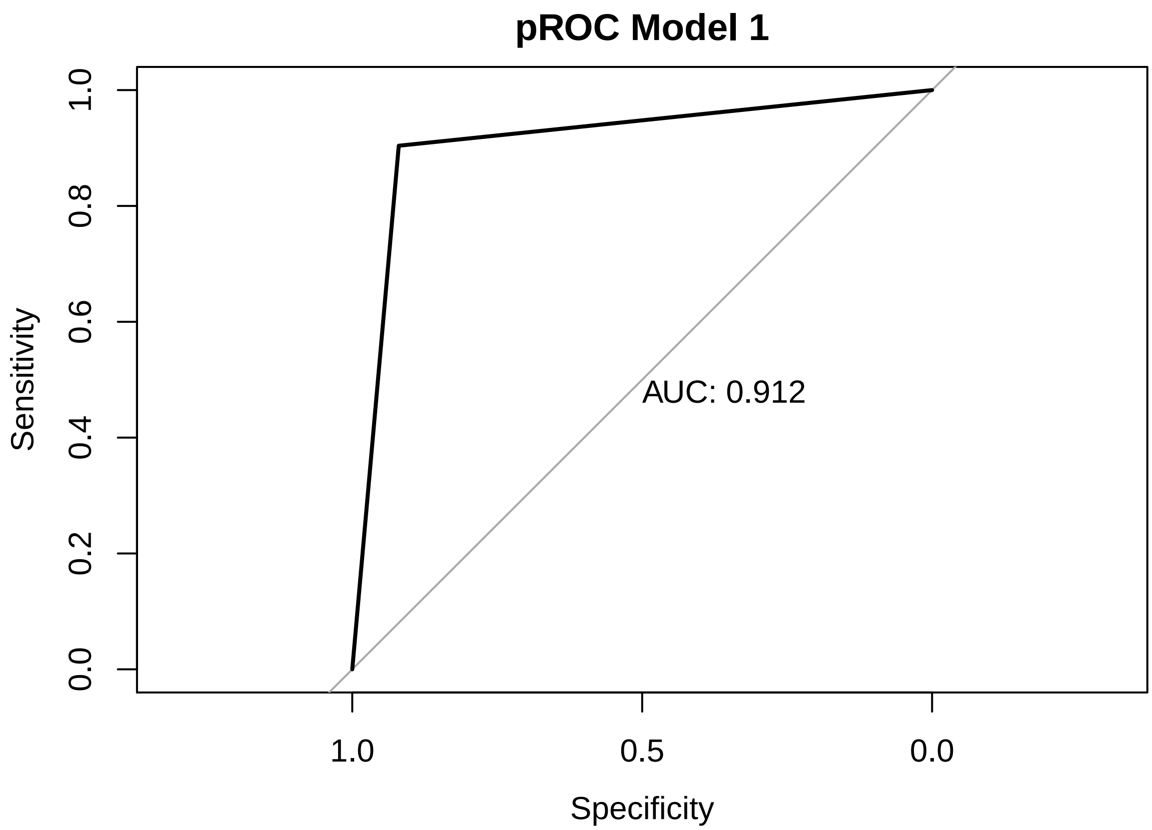



```
## $rect
## $rect$w
## [1] 0.2913194
##
## $rect$h
## [1] 0.2425452
##
## $rect$left
## [1] 0.6
##
```

```
## $rect$top
## [1] 0.4
##
##
## $text
## $text$x
## [1] 0.7164354
##
## $text$y
## [1] 0.2383032
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Despite the custom and built-in functions for both ROC curves are slightly different, the measure is still the same rounded off to the nearest tenths (`0.91`). However, the second ROC curve is more accurate.

RESULTS Listed below are the results of metrics done for the classification model that was chosen (Model 1).

Metric	Score
Accuracy	0.912
Classification Error Rate	0.088
Precision	0.9159
Sensitivity	0.9039
Specificity	0.9198
F1 Score	0.9099

CONFUSION MATRIX

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 218  19
##          1  22 207
##
##                Accuracy : 0.912
##                  95% CI : (0.8825, 0.9361)
##     No Information Rate : 0.515
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8239
##
##  Mcnemar's Test P-Value : 0.7548
##
##             Sensitivity : 0.9159
##             Specificity : 0.9083
##          Pos Pred Value : 0.9039
##          Neg Pred Value : 0.9198
##              Prevalence : 0.4850
##          Detection Rate : 0.4442
##    Detection Prevalence : 0.4914
##       Balanced Accuracy : 0.9121
##
##        'Positive' Class : 1
##
```

Compared with the custom functions used, the results are swapped.

Predict on Test Data –> Results

Appendix –> Find Source Code on GITHUB