Data 621 - Homework 3

Anthony Munoz

4/5/2020

Contents

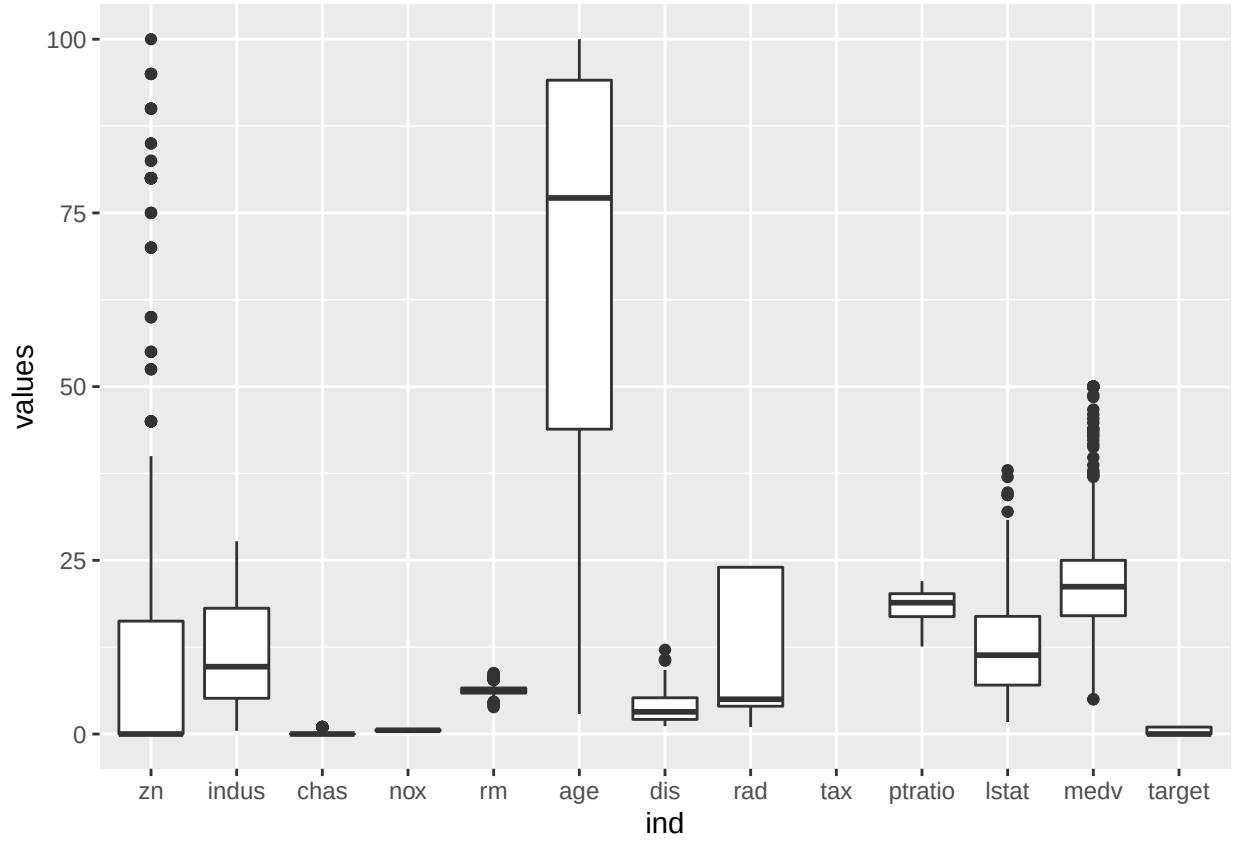
Data Exploration

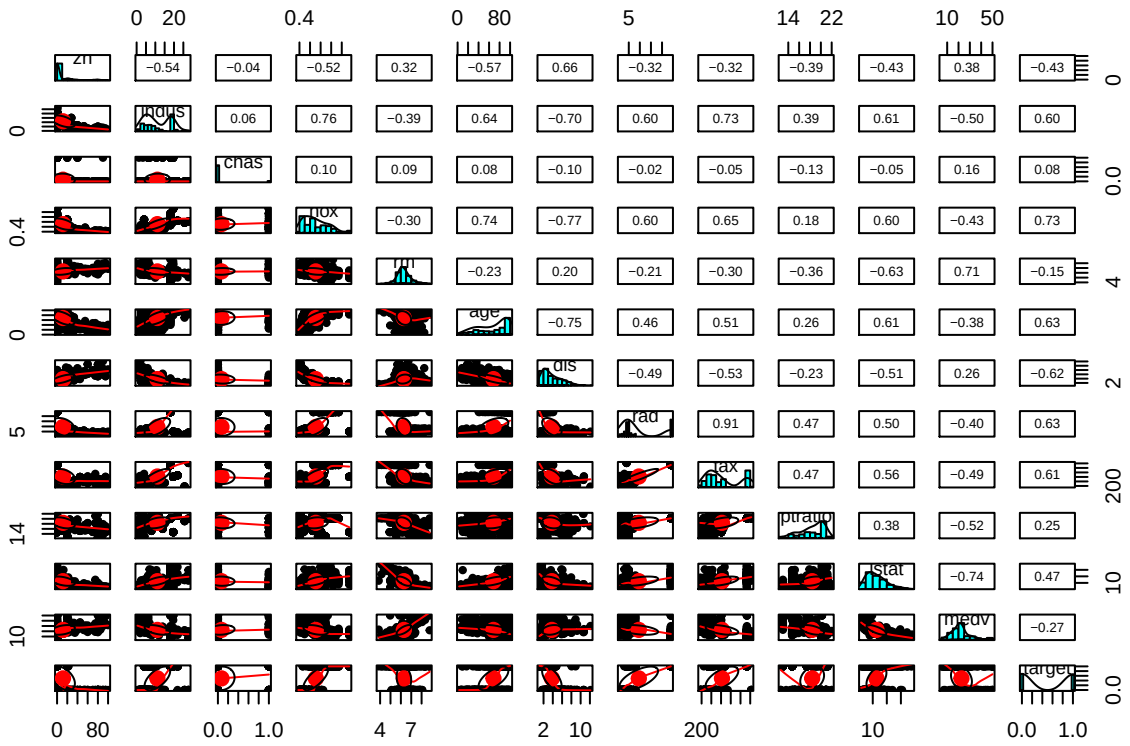
zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv	target
0	19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	3.70	50.0	1
0	19.58	1	0.871	5.403	100.0	1.3216	5	403	14.7	26.82	13.4	1
0	18.10	0	0.740	6.485	100.0	1.9784	24	666	20.2	18.85	15.4	1
30	4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	5.19	23.7	0
0	2.46	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

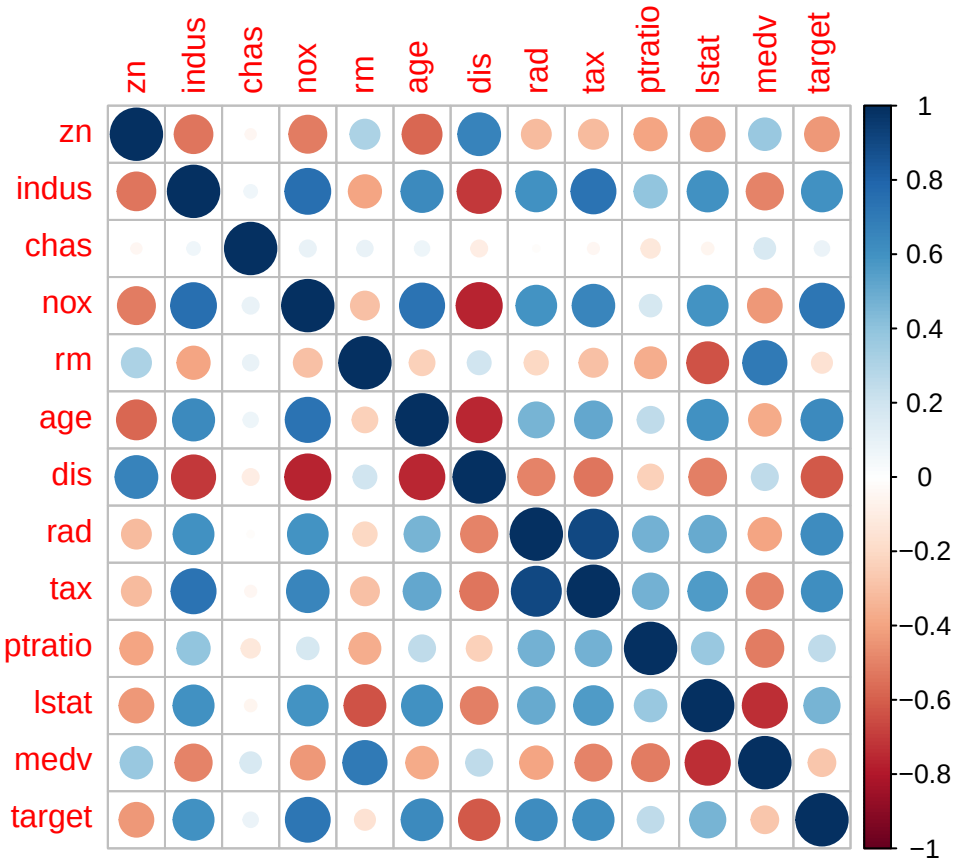
```
## 'data.frame':    466 obs. of  13 variables:
##  $ zn     : num  0 0 0 30 0 0 0 0 0 80 ...
##  $ indus  : num  19.58 19.58 18.1 4.93 2.46 ...
##  $ chas   : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.605 0.871 0.74 0.428 0.488 0.52 0.693 0.693 0.515 0.392 ...
##  $ rm     : num  7.93 5.4 6.49 6.39 7.16 ...
##  $ age    : num  96.2 100 100 7.8 92.2 71.3 100 100 38.1 19.1 ...
##  $ dis    : num  2.05 1.32 1.98 7.04 2.7 ...
##  $ rad    : int  5 5 24 6 3 5 24 24 5 1 ...
##  $ tax    : int  403 403 666 300 193 384 666 666 224 315 ...
##  $ ptratio: num  14.7 14.7 20.2 16.6 17.8 20.9 20.2 20.2 20.2 16.4 ...
##  $ lstat  : num  3.7 26.82 18.85 5.19 4.82 ...
##  $ medv   : num  50 13.4 15.4 23.7 37.9 26.5 5 7 22.2 20.9 ...
##  $ target : int  1 1 1 0 0 0 1 1 0 0 ...
```

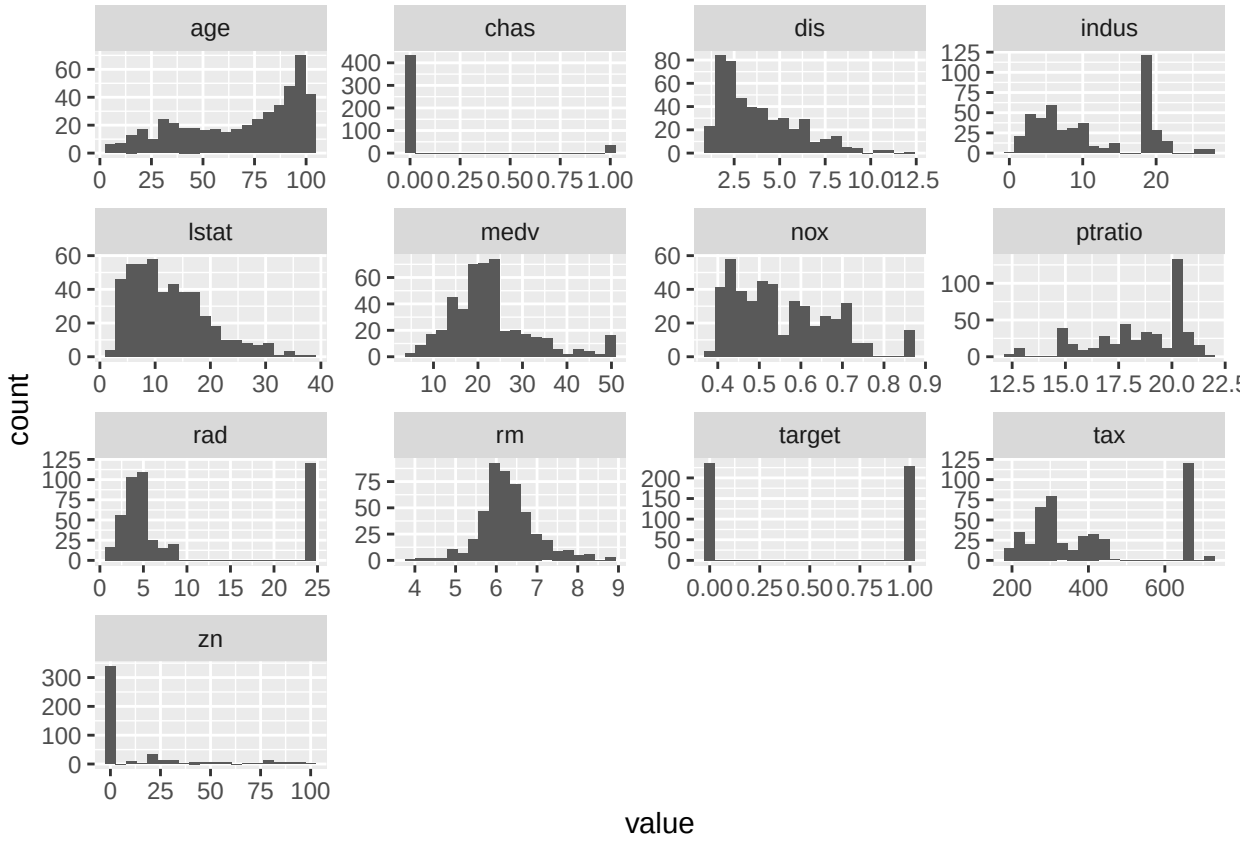
```
##        zn              indus             chas              nox        
##  Min.   :  0.00   Min.   : 0.460   Min.   :0.00000   Min.   :0.3890  
##  1st Qu.:  0.00   1st Qu.: 5.145   1st Qu.:0.00000   1st Qu.:0.4480
```

```
##  Median :  0.00   Median : 9.690   Median :0.00000   Median :0.5380
##  Mean   : 11.58   Mean   :11.105   Mean   :0.07082   Mean   :0.5543
##  3rd Qu.: 16.25   3rd Qu.:18.100   3rd Qu.:0.00000   3rd Qu.:0.6240
##  Max.   :100.00   Max.   :27.740   Max.   :1.00000   Max.   :0.8710
##        rm             age              dis              rad
##  Min.   :3.863   Min.   :  2.90   Min.   : 1.130   Min.   : 1.00
##  1st Qu.:5.887   1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00
##  Median :6.210   Median : 77.15   Median : 3.191   Median : 5.00
##  Mean   :6.291   Mean   : 68.37   Mean   : 3.796   Mean   : 9.53
##  3rd Qu.:6.630   3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.00
##       tax          ptratio          lstat            medv
##  Min.   :187.0   Min.   :12.6   Min.   : 1.730   Min.   : 5.00
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.: 7.043   1st Qu.:17.02
##  Median :334.5   Median :18.9   Median :11.350   Median :21.20
##  Mean   :409.5   Mean   :18.4   Mean   :12.631   Mean   :22.59
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:16.930   3rd Qu.:25.00
##  Max.   :711.0   Max.   :22.0   Max.   :37.970   Max.   :50.00
##      target
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4914
##  3rd Qu.:1.0000
##  Max.   :1.0000
```




zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
lstat
medv
target
−0.54
−0.04
−0.52
0.32
−0.57
0.66
−0.32
−0.32
−0.39
−0.43
0.38
−0.43
0.06
0.76
−0.39
0.64
−0.70
0.60
0.73
0.39
0.61
−0.50
0.60
0.10
0.09
0.08
−0.10
−0.02
−0.05
−0.13
−0.05
0.16
0.08
−0.30
0.74
−0.77
0.60
0.65
0.18
0.60
−0.43
0.73
−0.23
0.20
−0.21
−0.30
−0.36
−0.63
0.71
−0.15
−0.75
0.46
0.51
0.26
0.61
−0.38
0.63
−0.49
−0.53
−0.23
−0.51
0.26
−0.62
0.91
0.47
0.50
−0.40
0.63
0.47
0.56
−0.49
0.61
0.38
−0.52
0.25
−0.74
0.47
−0.27


zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
lstat
medv
target
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

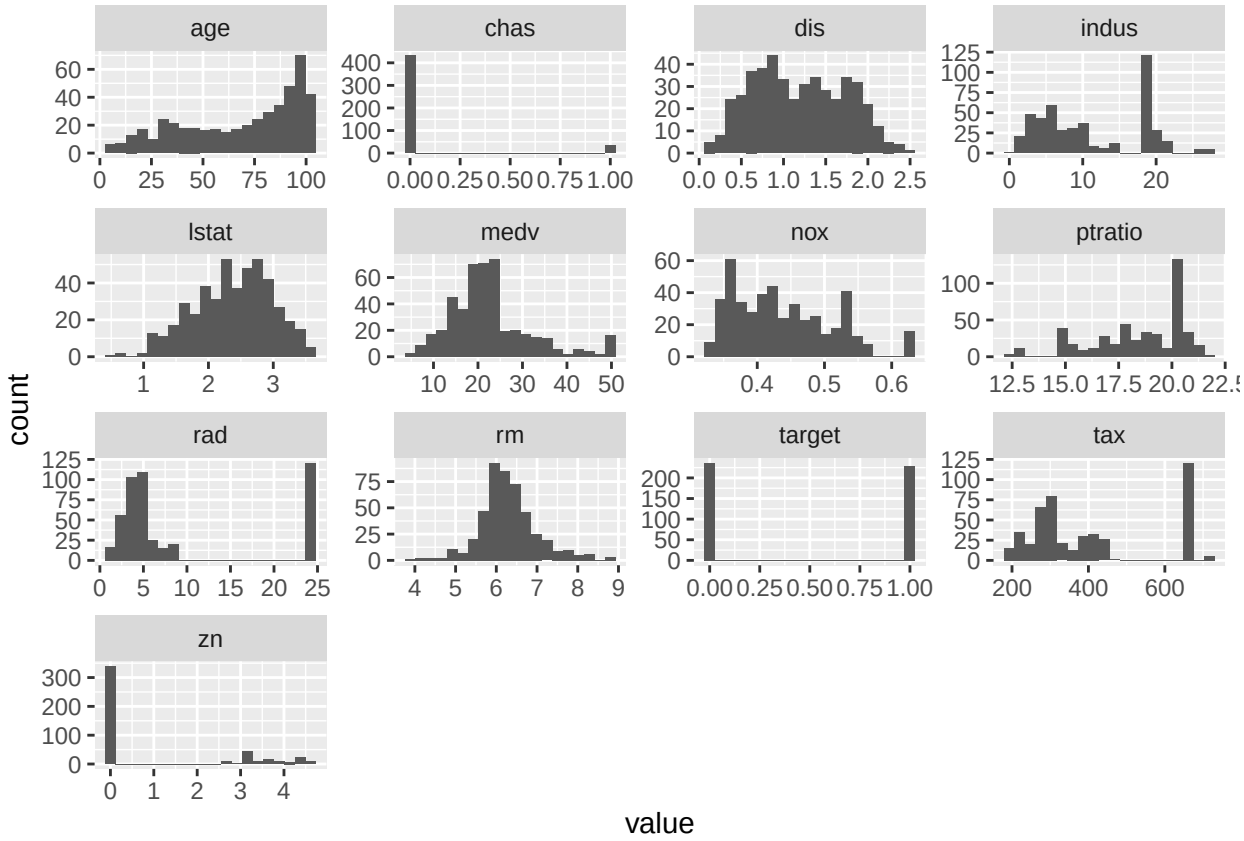


After observing the correlation plot and variables distribution we can see they are a couple of variables that need to be addressed. we don't have missing values as it shows on the data summary, but we see some high correlation between variables (rad-tax). We also can see some skewed data on some variables.

After the observation,we going to proceed with some data transformations methods and log variables in order to normalize the data for our modeling process.

Data Preparation

We decided to remove the Rad variable because it's highly correlated with the Tax variable. we proceed to do the log method to the variables(lstat,nox,dis), I tried to also work with the log method to the variable ZN but I realized it transforms some fo the observation to infinity values for which then get the drop when using for modeling. Another solution for zn variable could be to add 1 to the value of zeros.



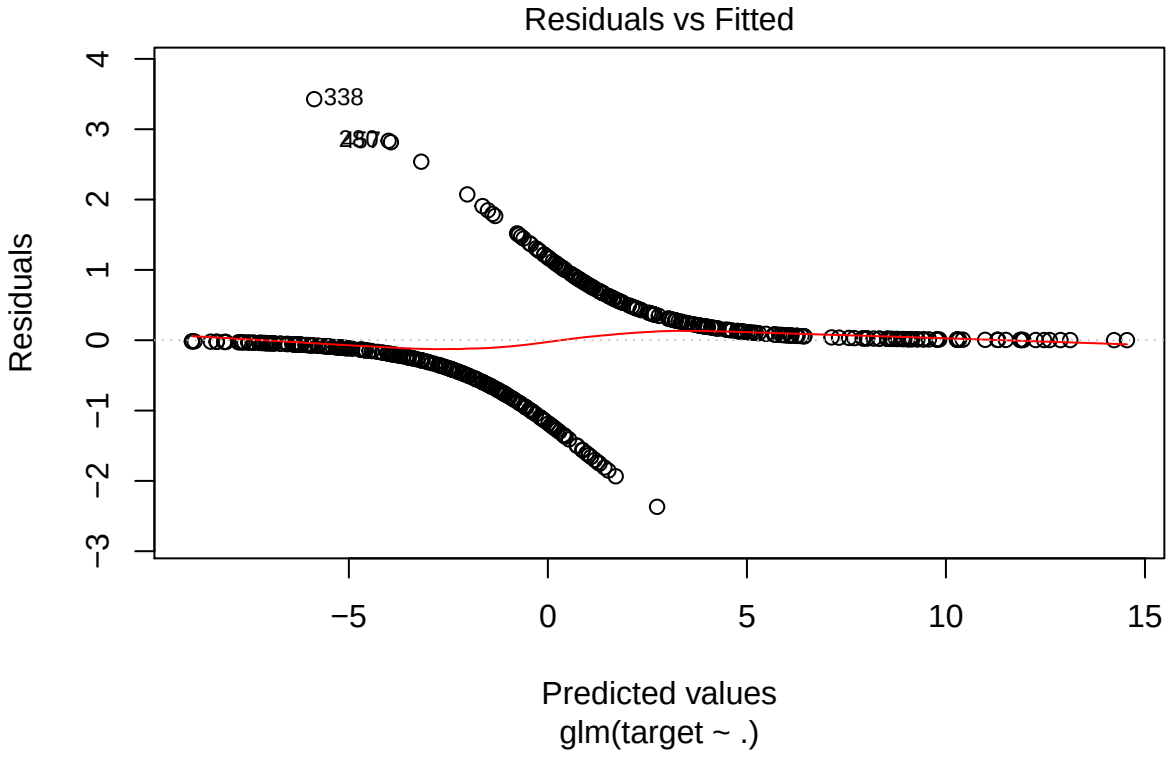
After some of the transformation, we can see that some variables seem better with the perspective of the distribution on the Histogram plot.

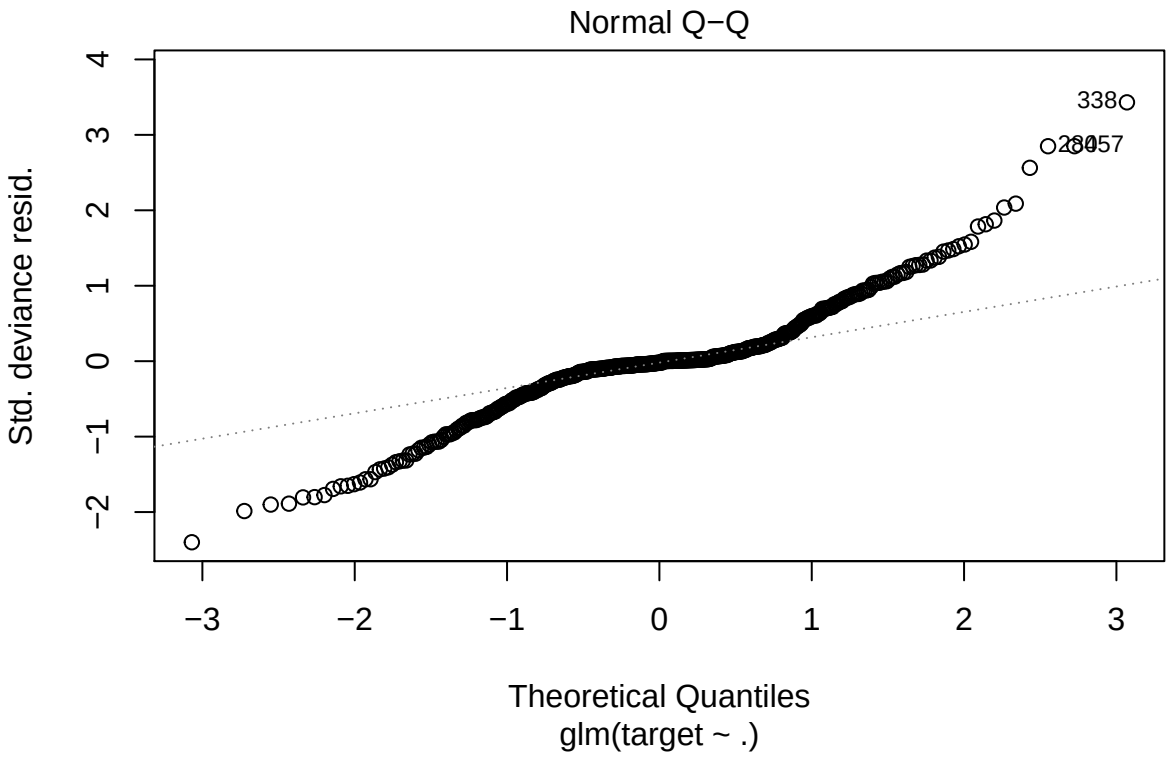
Build Models

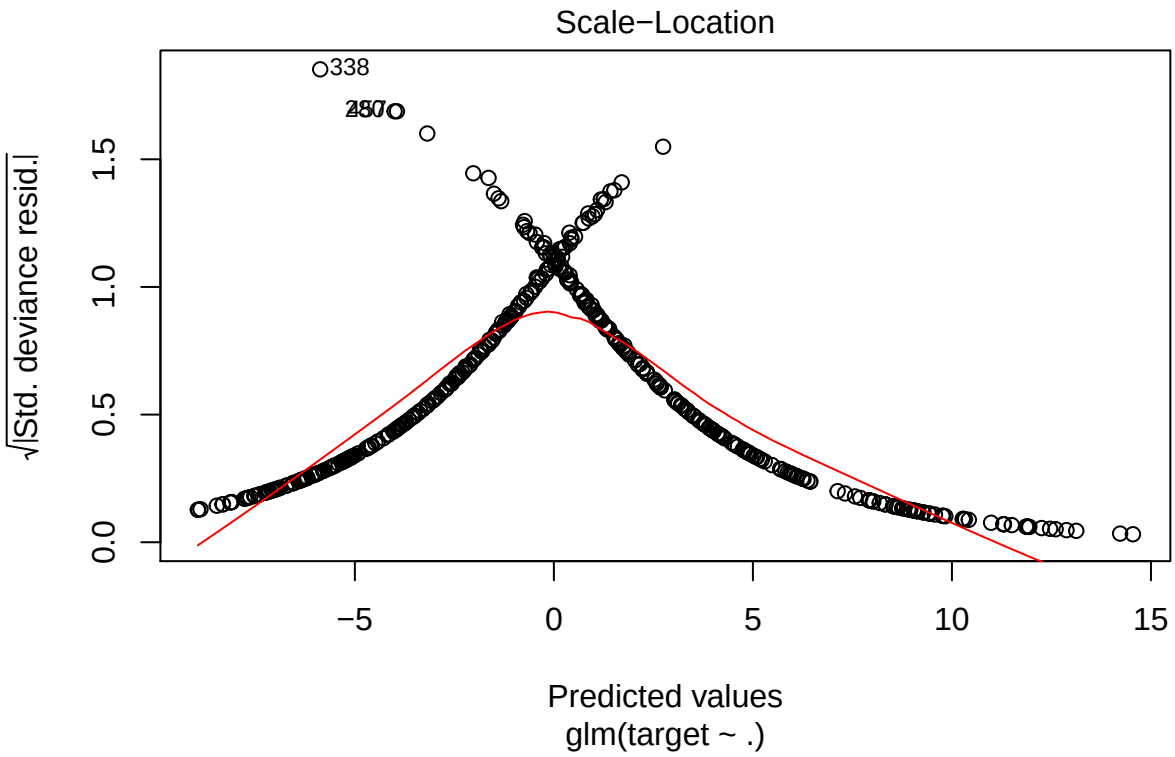
Model 1

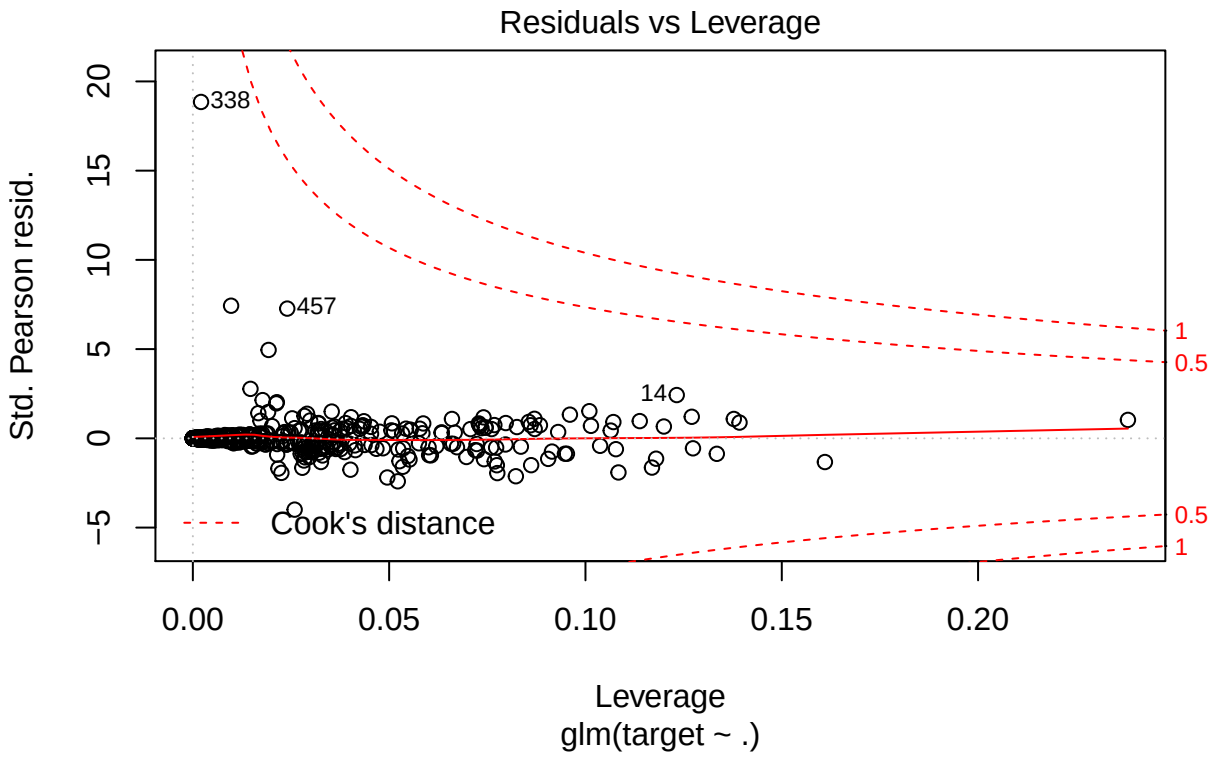
```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3688  -0.2434  -0.0213   0.2062   3.4274
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -53.323010   7.471072  -7.137 9.52e-13 ***
## zn           -0.470767   0.218692  -2.153 0.031346 *
## indus        -0.126789   0.050114  -2.530 0.011406 *
## chas          1.713163   0.665191   2.575 0.010011 *
## nox          80.892405  11.065437   7.310 2.66e-13 ***
## rm           -0.279025   0.560608  -0.498 0.618682
## age           0.030450   0.012026   2.532 0.011339 *
## dis           4.076773   0.868246   4.695 2.66e-06 ***
## tax           0.007532   0.002043   3.686 0.000227 ***
```

```
## ptratio       0.327166   0.113049   2.894 0.003804 **
## lstat         0.445131   0.638304   0.697 0.485574
## medv          0.204381   0.054933   3.721 0.000199 ***
## ---
## 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: 226.91  on 454  degrees of freedom
## AIC: 250.91
##
## Number of Fisher Scoring iterations: 7
```




Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(target ~ .)
338
2057
−3
−2
−1
0
1
2
3
4


Scale−Location
338
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
−5
0
5
10
15
Predicted values
glm(target ~ .)

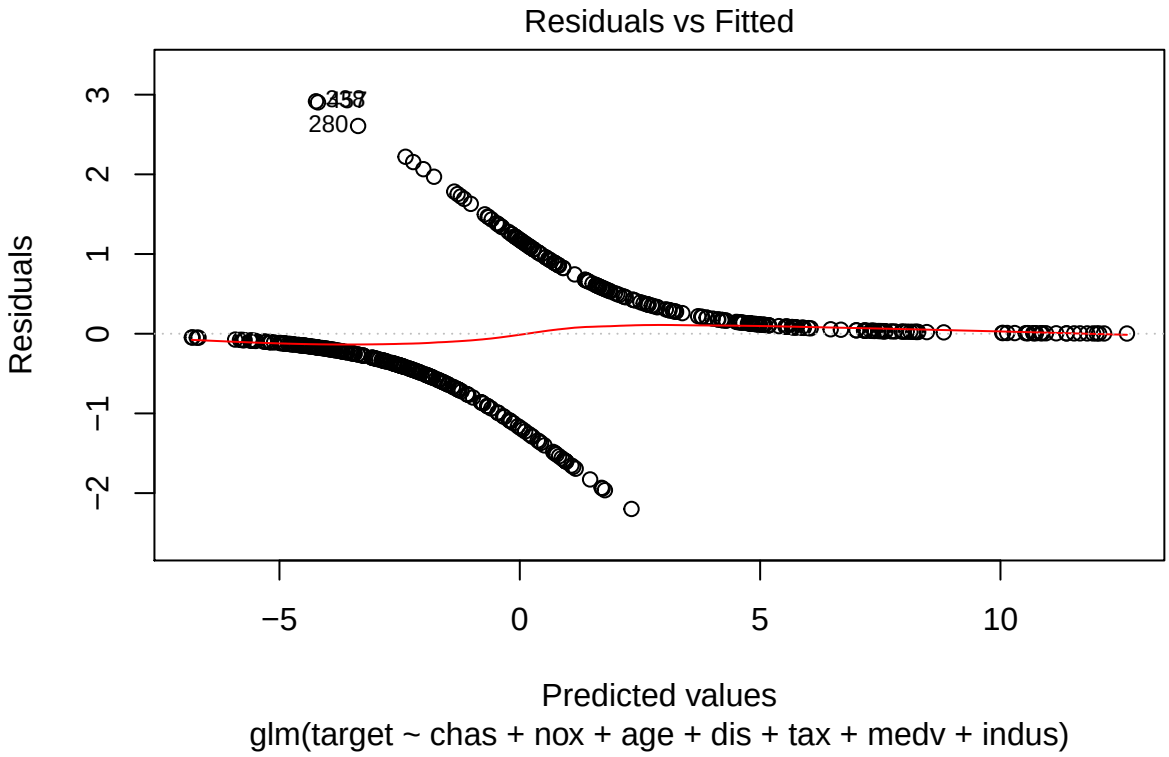


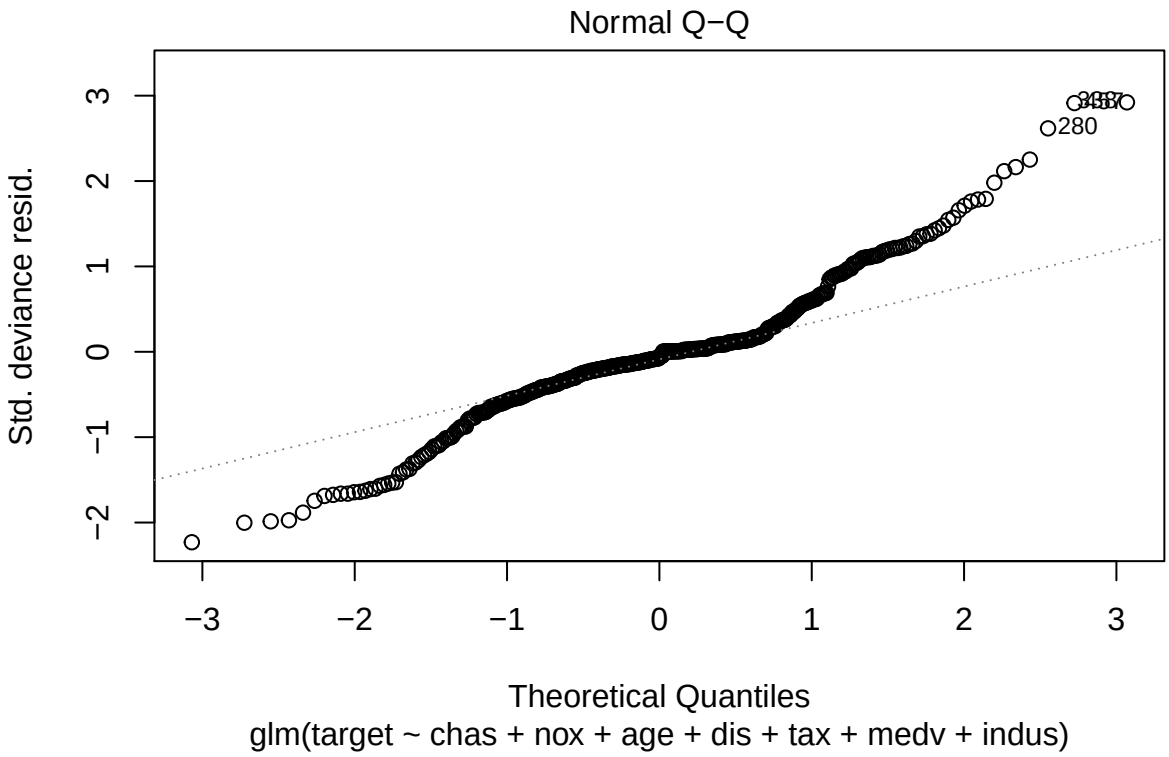

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data.new$target, fitted(model1)
## X-squared = 6.82, df = 8, p-value = 0.5562
```

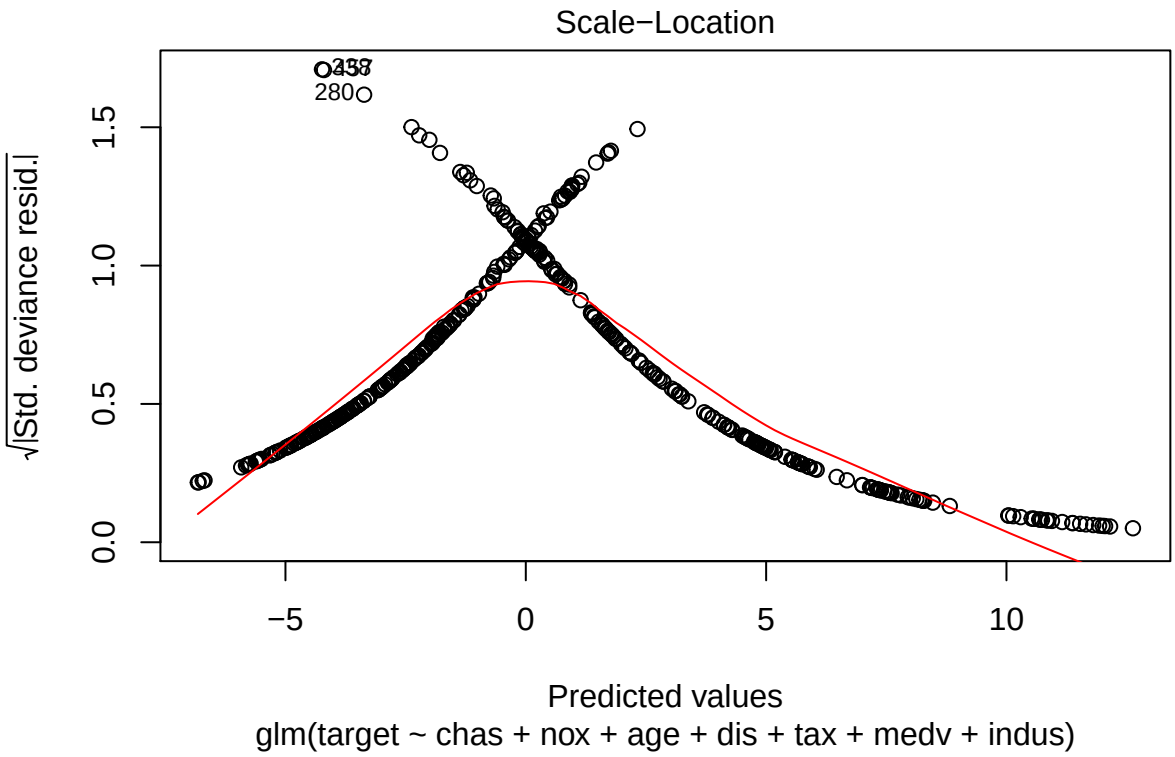
Model 2

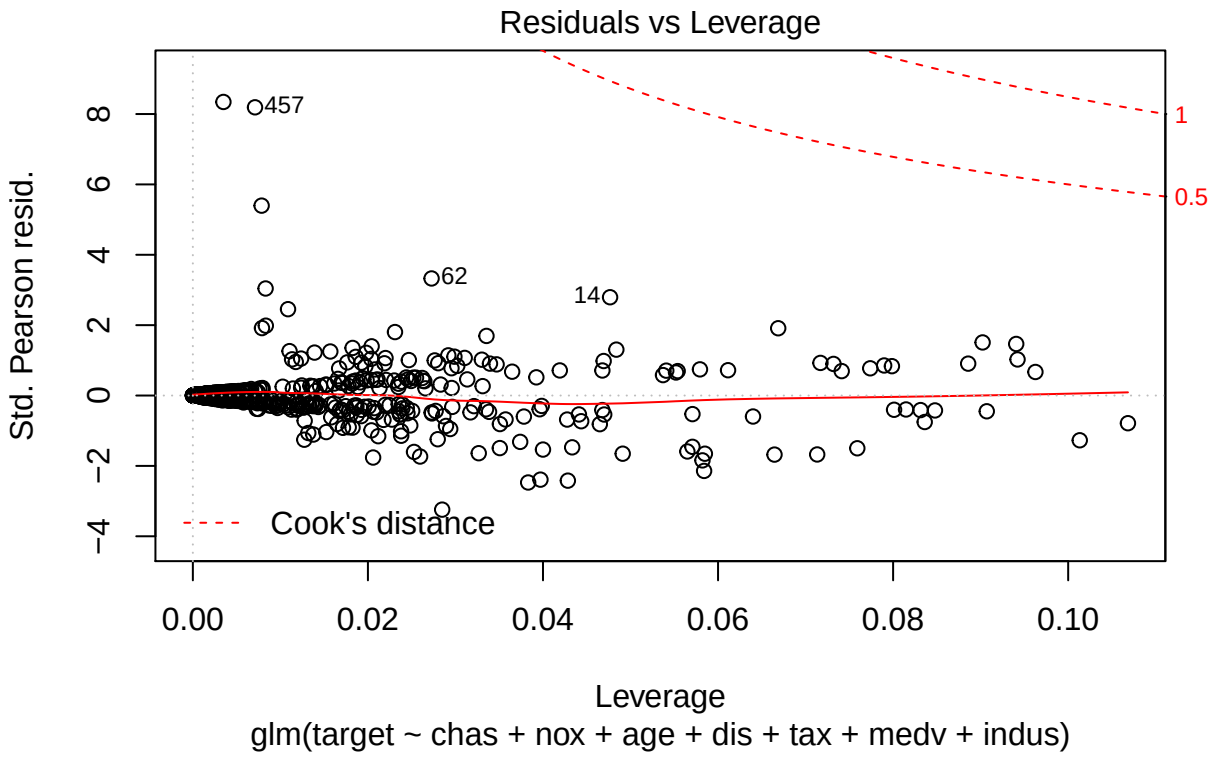
```
##
## Call:
## glm(formula = target ~ chas + nox + age + dis + tax + medv +
##     indus, family = "binomial", data = train.df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.19873  -0.37393  -0.06161   0.19783   2.91677
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -38.507342   5.275757  -7.299 2.90e-13 ***
## chas          1.354735   0.609202   2.224 0.026163 *
## nox          69.799609  10.263424   6.801 1.04e-11 ***
## age           0.025125   0.009629   2.609 0.009073 **
## dis           2.621239   0.724793   3.617 0.000299 ***
## tax           0.006223   0.001680   3.704 0.000212 ***
```

```
## medv          0.095171   0.026791   3.552 0.000382 ***
## indus        -0.099453   0.047973  -2.073 0.038161 *
## ---
## 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: 250.77  on 458  degrees of freedom
## AIC: 266.77
##
## Number of Fisher Scoring iterations: 7
```




Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(target ~ chas + nox + age + dis + tax + medv + indus)
280


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
280
−5
0
5
10
Predicted values
glm(target ~ chas + nox + age + dis + tax + medv + indus)




```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data.new$target, fitted(model2)
## X-squared = 10.029, df = 8, p-value = 0.263
```

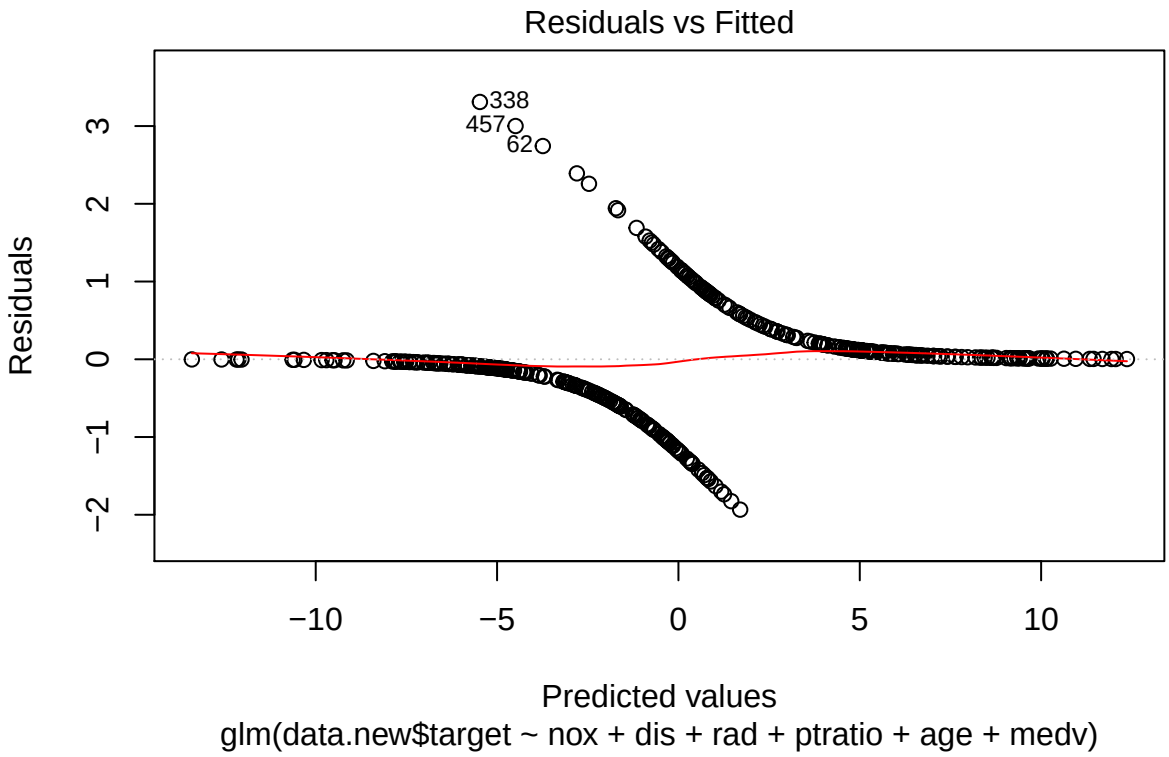
Model 3

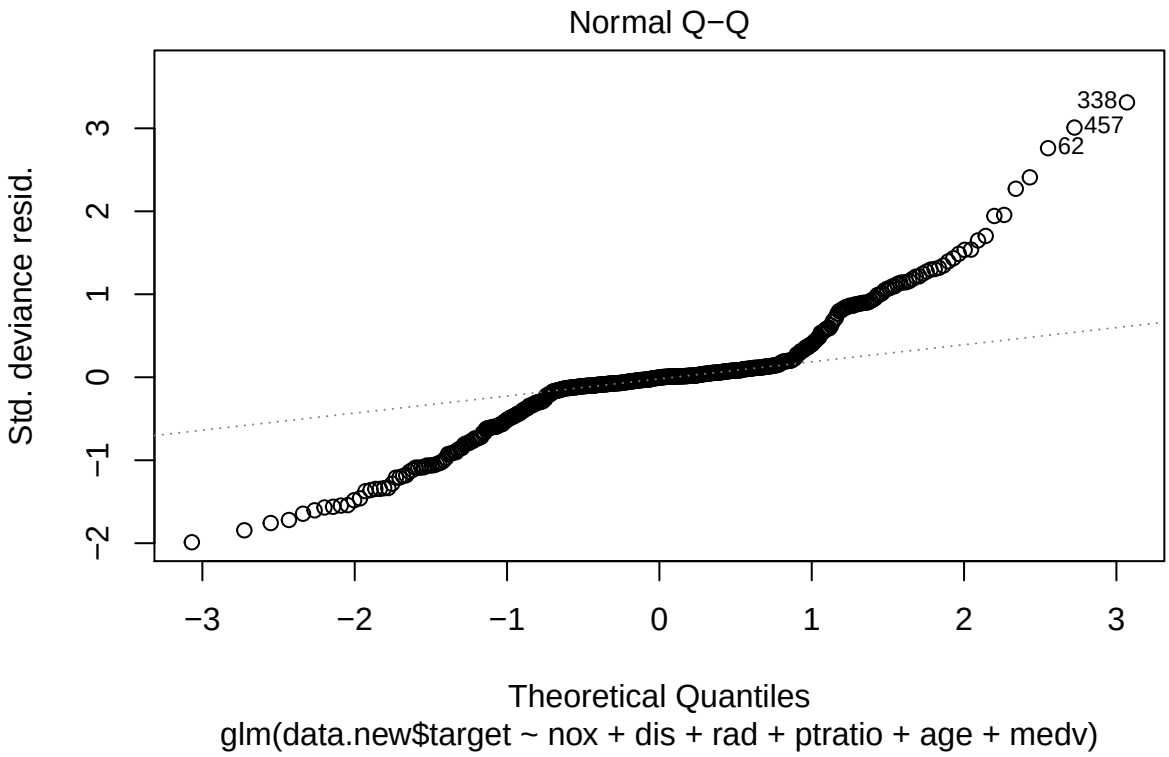
```
##
## Call:
## glm(formula = data.new$target ~ ., family = "binomial", data = df.tranformed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9381  -0.1116  -0.0010   0.1137   3.4325
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  21.134102  38.495758   0.549 0.583007
## zn           -0.022244   0.026852  -0.828 0.407433
## indus        -0.002008   0.216566  -0.009 0.992603
## chas          0.945998   0.761805   1.242 0.214316
## nox          14.172248   2.240335   6.326 2.52e-10 ***
## rm           -2.330063   2.813401  -0.828 0.407556
## age           0.012105   0.003914   3.093 0.001984 **
```

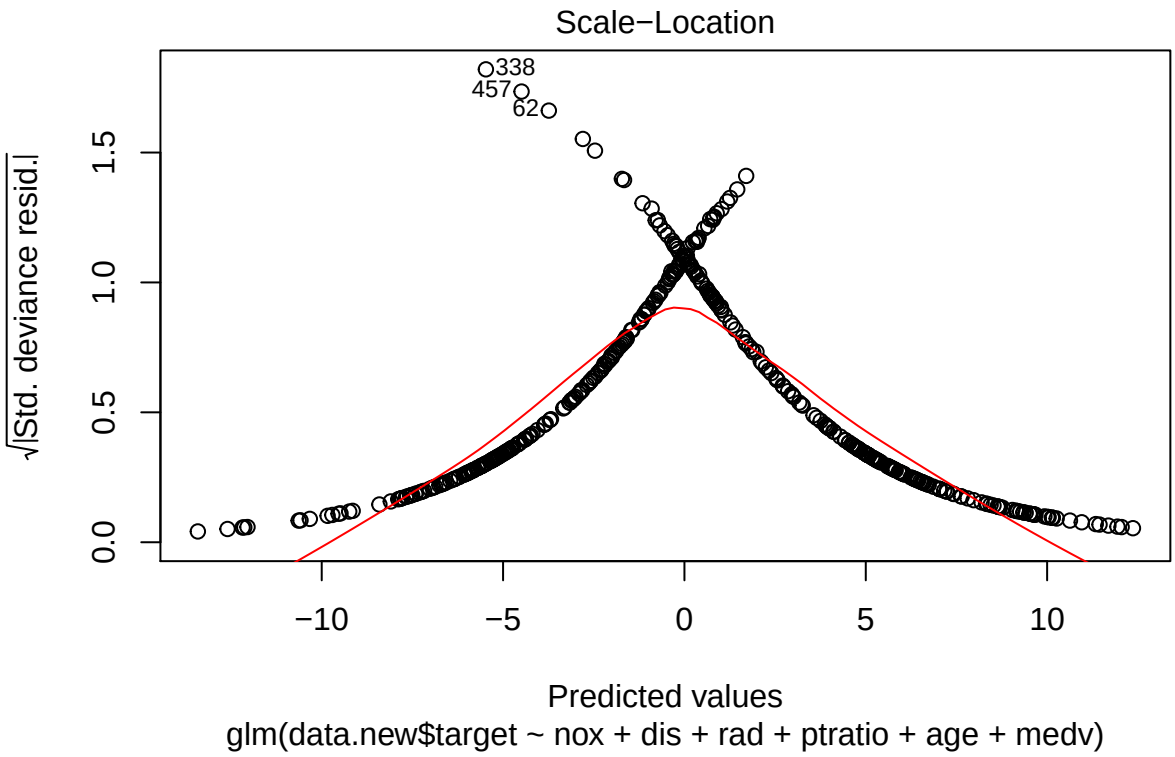
```
## dis           3.390172   0.868215   3.905 9.43e-05 ***
## rad           3.152839   0.733173   4.300 1.71e-05 ***
## tax         -16.176693  20.445106  -0.791 0.428812
## ptratio       0.025318   0.007169   3.532 0.000413 ***
## lstat        -0.051425   0.445840  -0.115 0.908173
## medv          2.461332   0.856713   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


##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data.new$target, fitted(model3)
## X-squared = 31.722, df = 8, p-value = 0.0001045
```

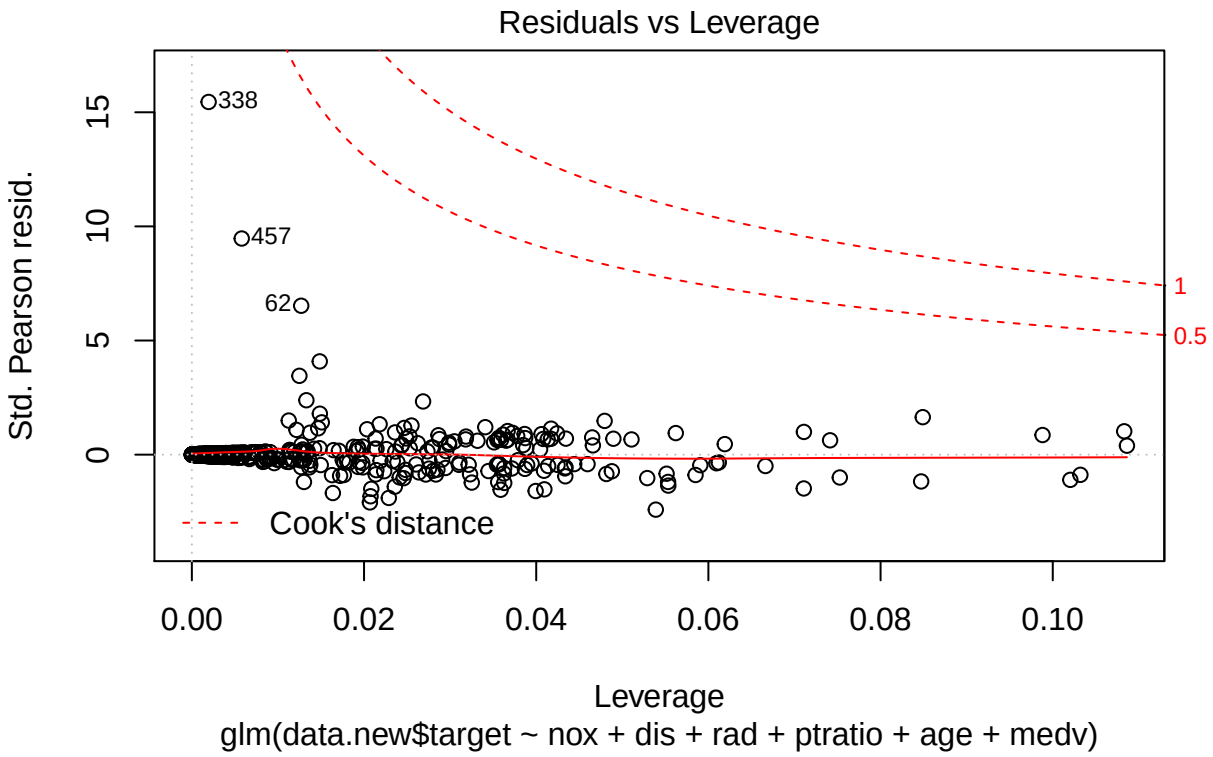
Model 4

```
##
## Call:
## glm(formula = data.new$target ~ nox + dis + rad + ptratio + age +
##     medv, family = "binomial", data = df.tranformed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9339  -0.1585  -0.0034   0.1187   3.3099
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.881152   3.271336  -3.632 0.000281 ***
## nox          13.437240   1.999435   6.721 1.81e-11 ***
## dis           3.203674   0.780114   4.107 4.01e-05 ***
## rad           2.703999   0.516073   5.240 1.61e-07 ***
## ptratio       0.022804   0.006089   3.745 0.000180 ***
## age           0.011040   0.003200   3.450 0.000561 ***
## medv          2.027111   0.480943   4.215 2.50e-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: 202.63  on 459  degrees of freedom
## AIC: 216.63
##
## Number of Fisher Scoring iterations: 7
```


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


Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(data.new$target ~ nox + dis + rad + ptratio + age + 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
Predicted values
glm(data.new$target ~ nox + dis + rad + ptratio + age + medv)




```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data.new$target, fitted(model4)
## X-squared = 9.119, df = 8, p-value = 0.3324
```

Select Model

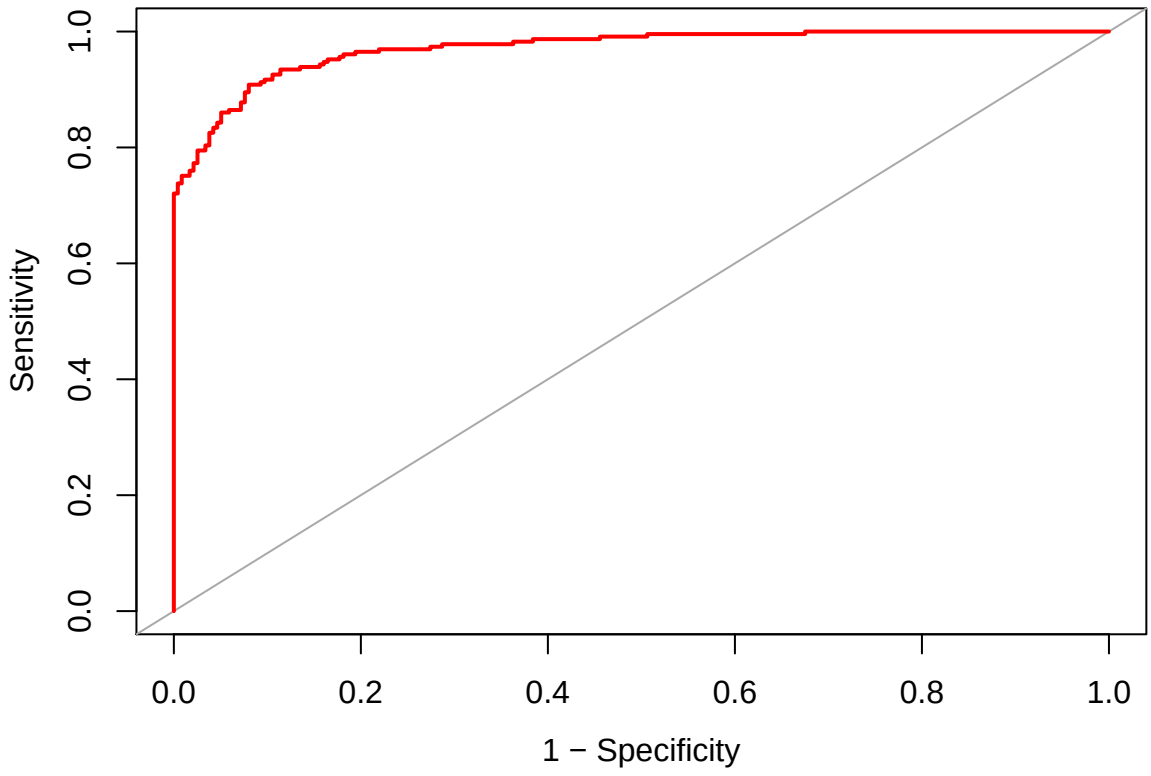
After working with different model solutions for our data, we going to select model # 4 taking into consideration it has the lowest AIC value of 216. Model number #4 is created using the Boxcox transformation and also selecting the most significant variables.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 225  36
##          1  12 193
##
##                Accuracy : 0.897
##                  95% CI : (0.8658, 0.9231)
##     No Information Rate : 0.5086
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7936
```

```
##
##  Mcnemar's Test P-Value : 0.0009009
##
##             Sensitivity : 0.9494
##             Specificity : 0.8428
##          Pos Pred Value : 0.8621
##          Neg Pred Value : 0.9415
##              Prevalence : 0.5086
##          Detection Rate : 0.4828
##    Detection Prevalence : 0.5601
##       Balanced Accuracy : 0.8961
##
##        'Positive' Class : 0
##
```

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

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




```
## Warning in coords.roc(roc.value, "best"): The 'transpose' argument to FALSE
## by default since pROC 1.16. Set transpose = TRUE explicitly to revert to
## the previous behavior, or transpose = TRUE to silence this warning. Type
## help(coords_transpose) for additional information.
```

threshold	specificity	sensitivity
0.0521061	0.9198312	0.9082969

zn	indus	chas	nox	rm	age	dis	rad	tax	p
-0.3864190	-0.6244111	-0.2265299	-0.8386288	1.4261128	-0.3764993	0.5569890	-0.8579380	-0.8543307	-0.809
-0.3864190	-0.4738235	-0.2265299	-0.1969278	-0.1737603	0.5143080	0.3184479	-0.6372466	-0.4877862	1.153
-0.3864190	-0.4738235	-0.2265299	-0.1969278	0.4124191	0.8911880	0.3150483	-0.6372466	-0.4877862	1.153
-0.3864190	-0.4738235	-0.2265299	-0.1969278	-0.3882520	0.4191362	0.0956301	-0.6372466	-0.4877862	1.153
-0.3864190	-0.7806283	-0.2265299	-0.5596284	-0.5351642	-1.1226456	0.0692829	-0.5269009	-0.6456823	0.049
0.7020852	-0.8974393	-0.2265299	-0.9874290	-0.6952984	-0.1823490	1.6232942	-0.1958637	-0.6174866	0.355

Appendix

Rcode: Github

GitHub CSV file: Github