Data 621 - HW 3

Murali Kunissery

April 10, 2019

Introduction

Crime has existed as long as humans have been. Various factors in combination such as deep differences in economic strata, pyschological disturbances arising out of intoxicating products , lack of culture in raising kids and multitude of other factors plays a key role. The attached spreadsheet(csv) files lists out a set of predictor variables though not outlined above and here i am attempting to create a relationship using various models that predict crime rate .

Statement of the Problem

This report presents a statistical model aimed at determining variables that are independently associated with crime rates above or below the median.

1. DATA EXPLORATION

Data Describe

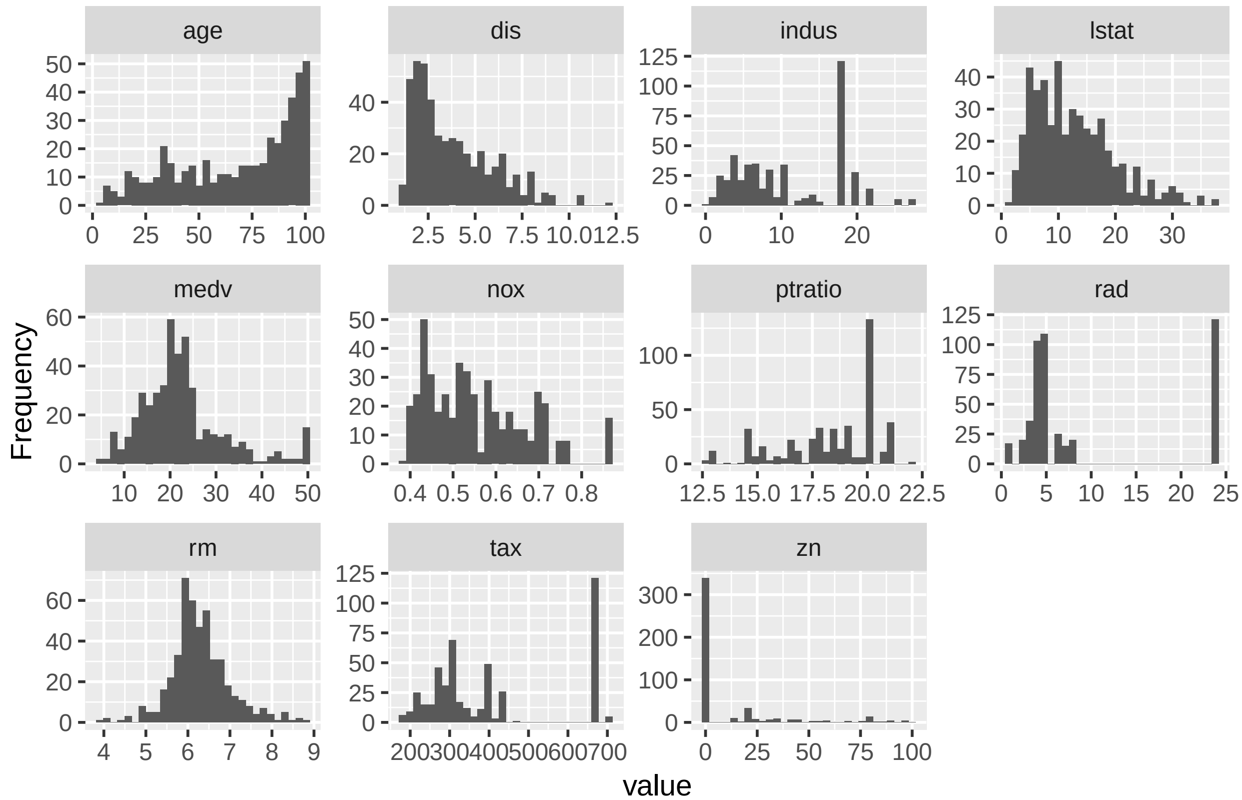
	n	mean	sd	median	min	max	skew	kurtosis
zn	466	11.5772532	23.3646511	0.00000	0.0000	100.0000	2.1768152	3.8135765
indus	466	11.1050215	6.8458549	9.69000	0.4600	27.7400	0.2885450	-1.2432132
nox	466	0.5543105	0.1166667	0.53800	0.3890	0.8710	0.7463281	-0.0357736
rm	466	6.2906738	0.7048513	6.21000	3.8630	8.7800	0.4793202	1.5424378
age	466	68.3675966	28.3213784	77.15000	2.9000	100.0000	-0.5777075	-1.0098814
dis	466	3.7956929	2.1069496	3.19095	1.1296	12.1265	0.9988926	0.4719679
rad	466	9.5300429	8.6859272	5.00000	1.0000	24.0000	1.0102788	-0.8619110
tax	466	409.5021459	167.9000887	334.50000	187.0000	711.0000	0.6593136	-1.1480456
ptratio	466	18.3984979	2.1968447	18.90000	12.6000	22.0000	-0.7542681	-0.4003627
lstat	466	12.6314592	7.1018907	11.35000	1.7300	37.9700	0.9055864	0.5033688
medv	466	22.5892704	9.2396814	21.20000	5.0000	50.0000	1.0766920	1.3737825

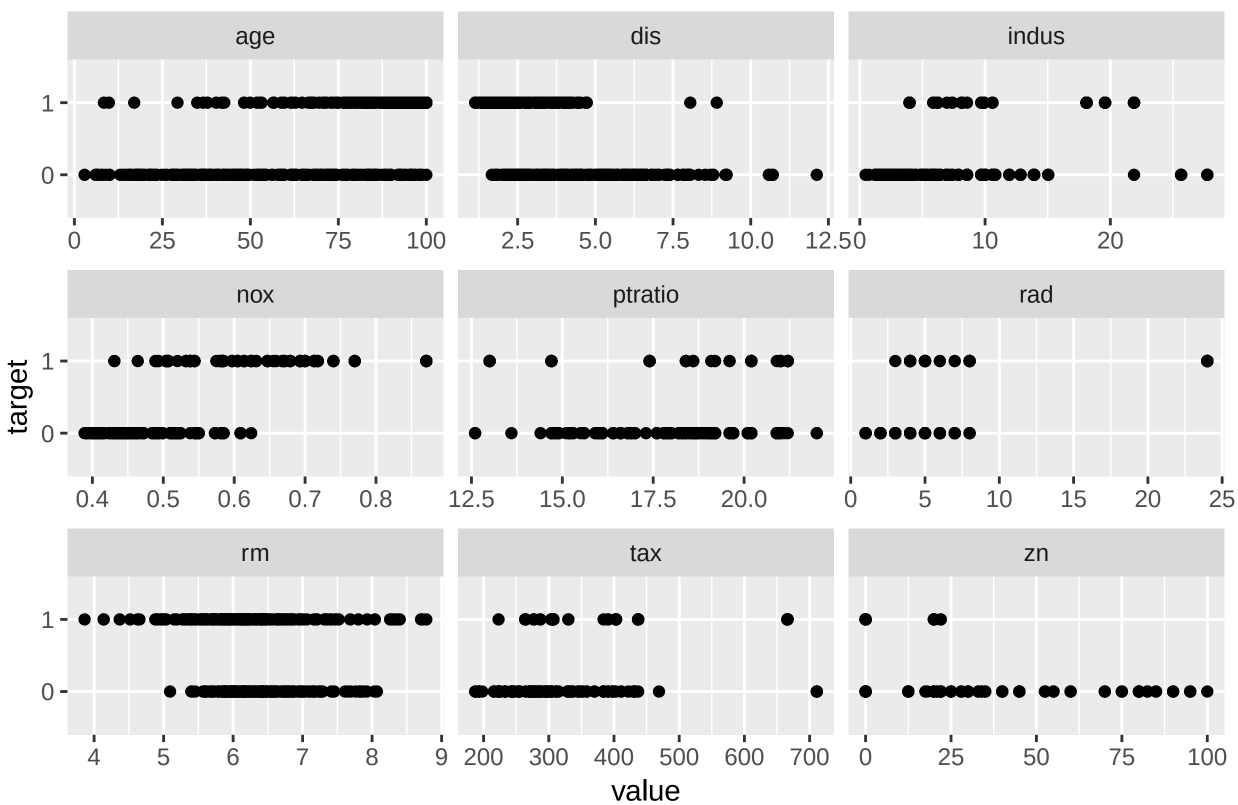
Check distribution

Check distribution of the target variable in our training data.

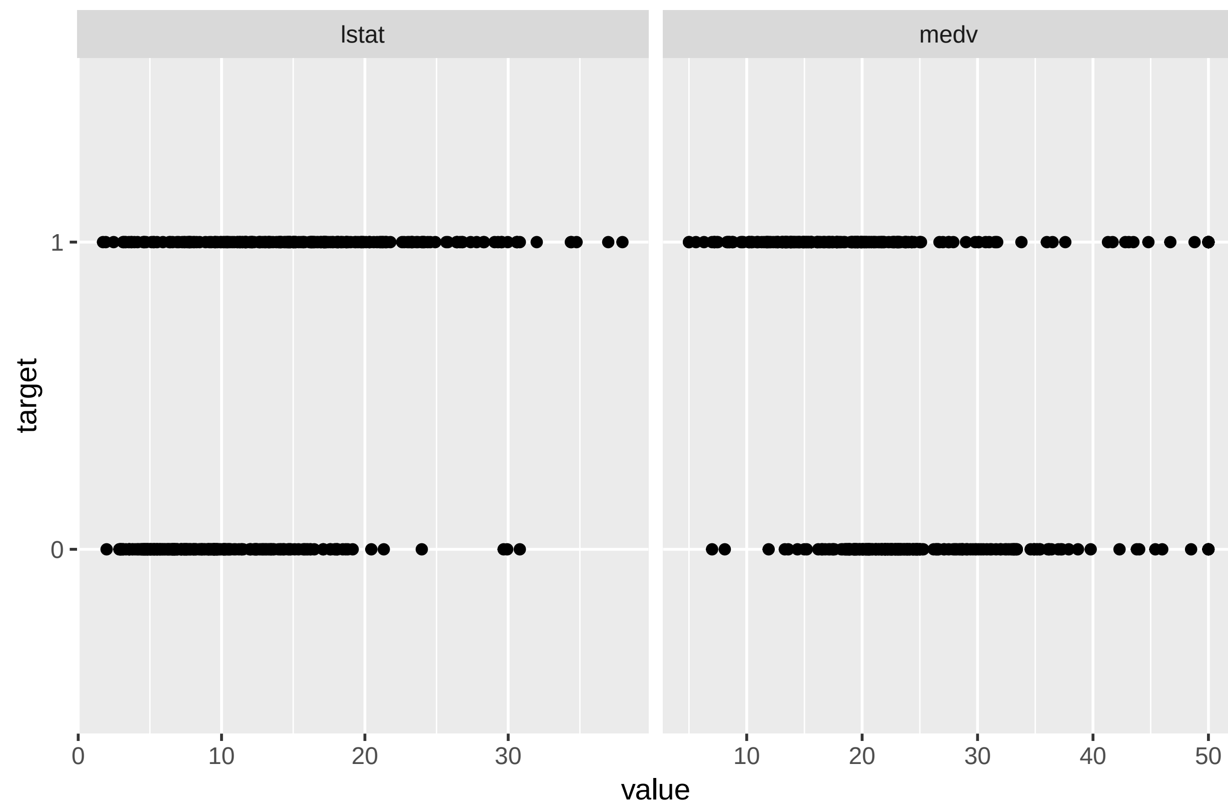
Var1	Freq
0	237
1	229

Histogram of Variables





Page 1



Page 2

2. DATA CLEANSING

We can see from our visualizations a few variables with some issues. We'll modify `rad` so that the value for 24 is now sequential and 9. We'll also center and scale our data based on the mean and standard deviation of each variable during the model building step. We'll otherwise avoid binning.

3. BUILD MODELS

Because we have a small number of observations to train over, we'll use k-fold Cross Validation to train, with k = 10. We'll hold out 15% of the data for validation while doing initial modeling, but once we select our model, we'll retrain over the full training set.

Each of our logistic regression models will use bionomial regression with a logit link function.

Model 1

The first model fits includes all the variables.

A review of the VIF (Variance Inflation Factors) output of the model suggests some points that are highly colinear and a number of variables that may not be necessary. Model 1 uses the formula:

target ~ .

	x
zn	278.69427

	x
indus	51.51210
nox	347.41191
rm	111.56756
age	62.25171
dis	95.88056
rad	64.76307
tax	85.16809
ptratio	34.04320
lstat	69.48539
medv	194.75857

Model 2

Our second model ignores the colinear issues, but removes models that seemed unnecessary in Model #1. Model 2 uses the formula:

target ~ zn + nox + age + dis + rad + ptratio + medv

	x
zn	228.41925
nox	207.13059
age	40.21003
dis	84.13470
rad	35.21336
ptratio	24.54924
medv	51.75492

Model #3

Model #3 removes the variables with the 2 highest VIF values from model1. The model formula is:

target ~ indus + rm + age + dis + tax + ptratio + lstat + medv

	x
indus	22.90998
rm	35.83250
age	28.92772
dis	32.52778
tax	27.67659
ptratio	12.05431
lstat	32.66123
medv	56.37327

Model #4

Model #4 takes the advances in model #3 and removes those values shown to be poor predictors.

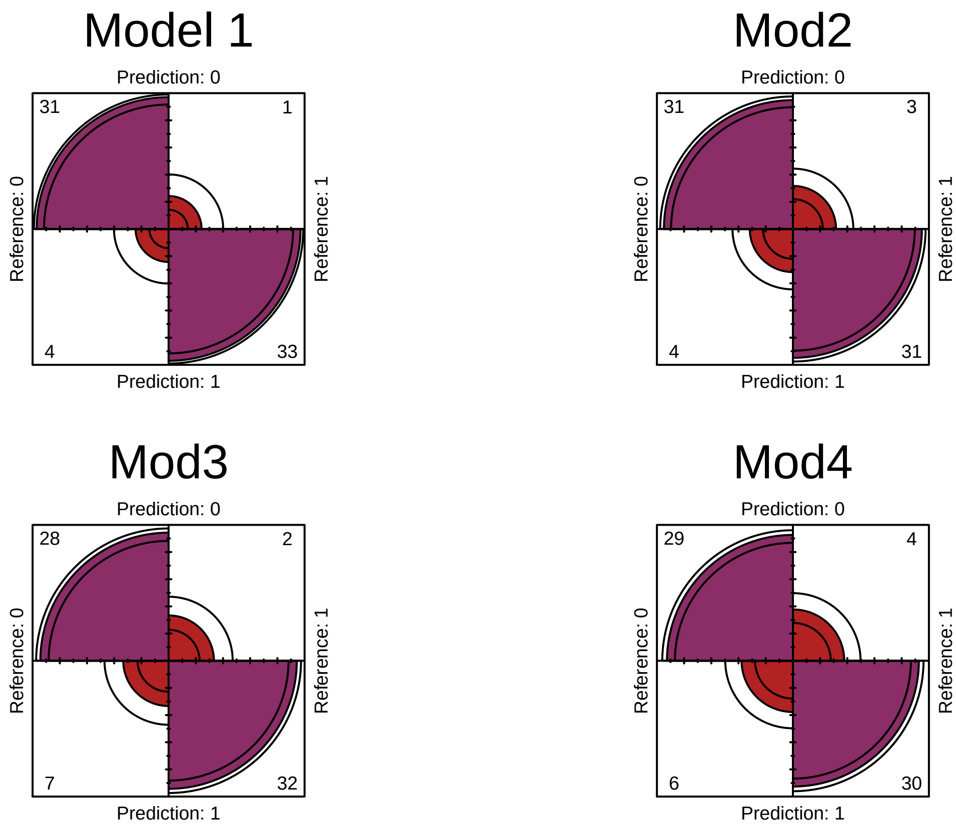
target ~ age + dis + tax + medv

	x
age	24.78884
dis	28.73403
tax	20.52367
medv	14.44172

4. SELECT MODELS

To help aid in model selection, we'll review their accuracy by making predictions on our holdout validation set, and comparing their performance using a variety of confusion matrix adjacent functions like fourfold plots, summary statistics, and ROC / AUC plots.

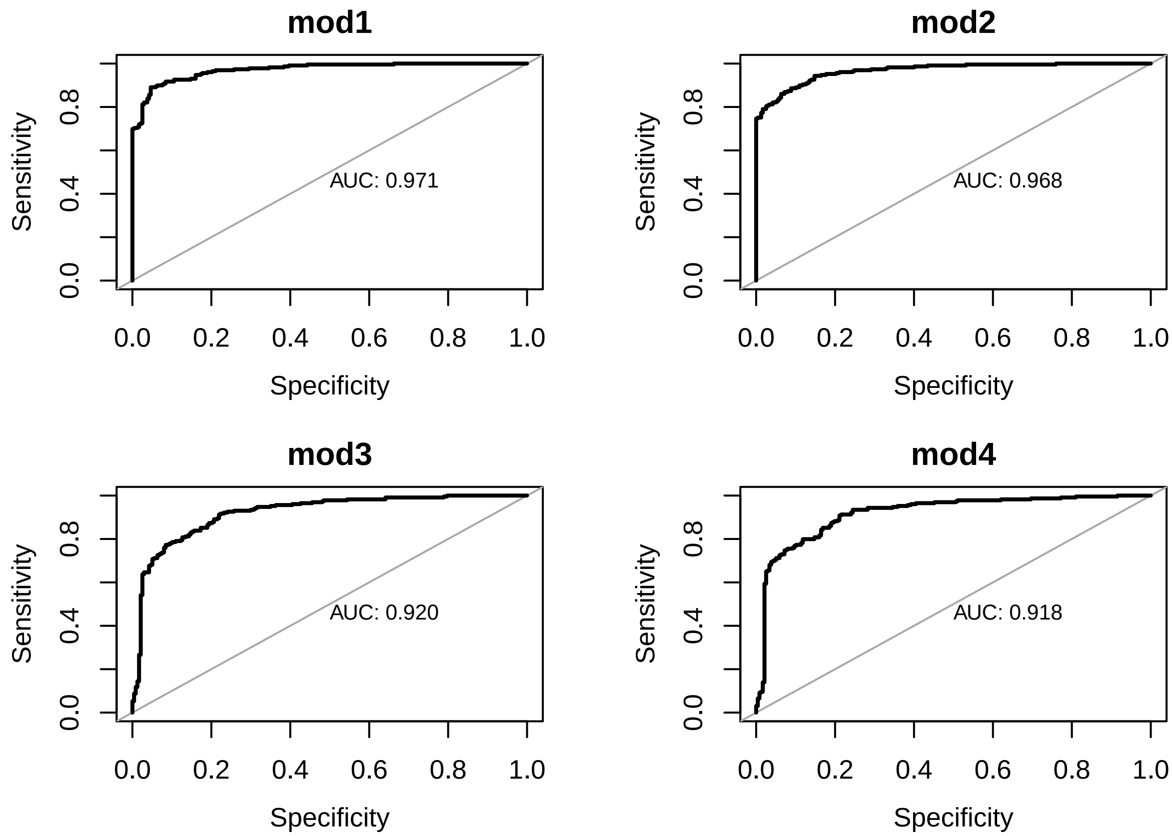
Fourfold Plots



Summary Statistics

	Sensitivity	Specificity	Precision	Recall	F1
Model1	0.8857143	0.9705882	0.9687500	0.8857143	0.9253731
Model2	0.8857143	0.9117647	0.9117647	0.8857143	0.8985507
Model3	0.8000000	0.9411765	0.9333333	0.8000000	0.8615385
Model4	0.8285714	0.8823529	0.8787879	0.8285714	0.8529412

ROC / AUC



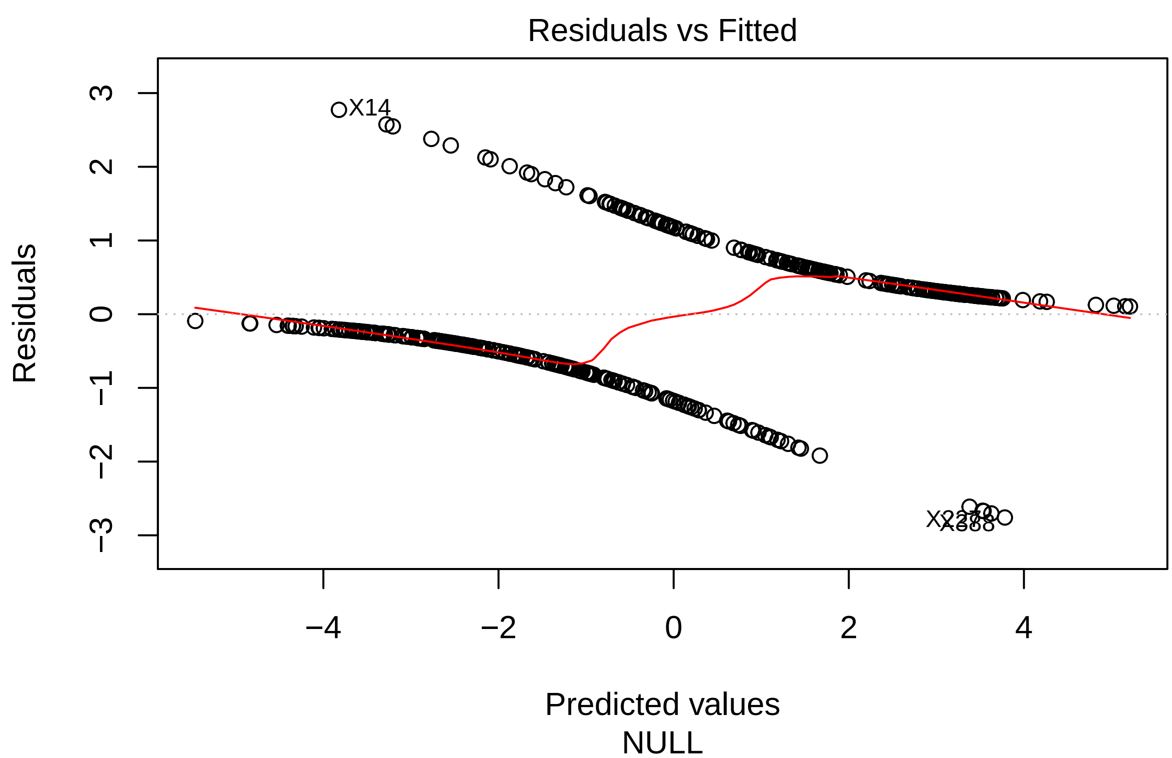
Model Selection

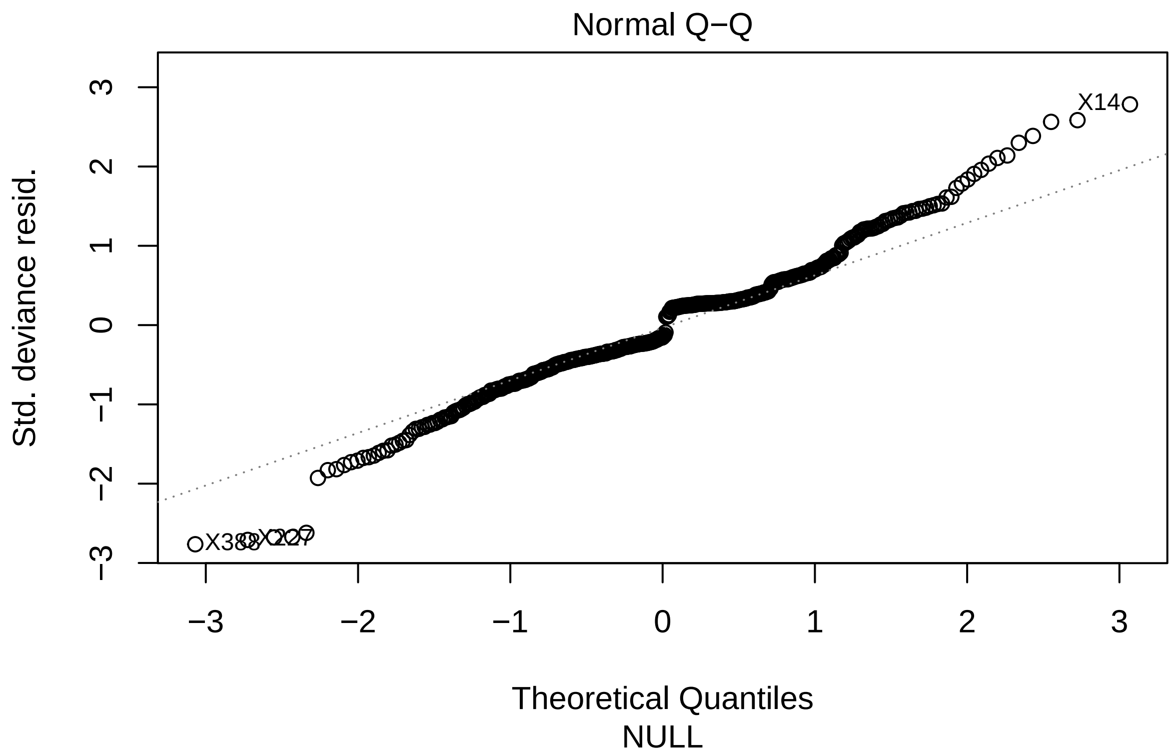
While the first 2 models may have the most information, they also suffer from so co-linearity issues as shown by the variance VIF output. Model #3 performs well, but has some additional variables that may be poor predictors of whether a neighborhood will be above or below the median crime rate. Instead, while stripped out, we'll use Model #4 with only age, dis, tax and medv as predictors.

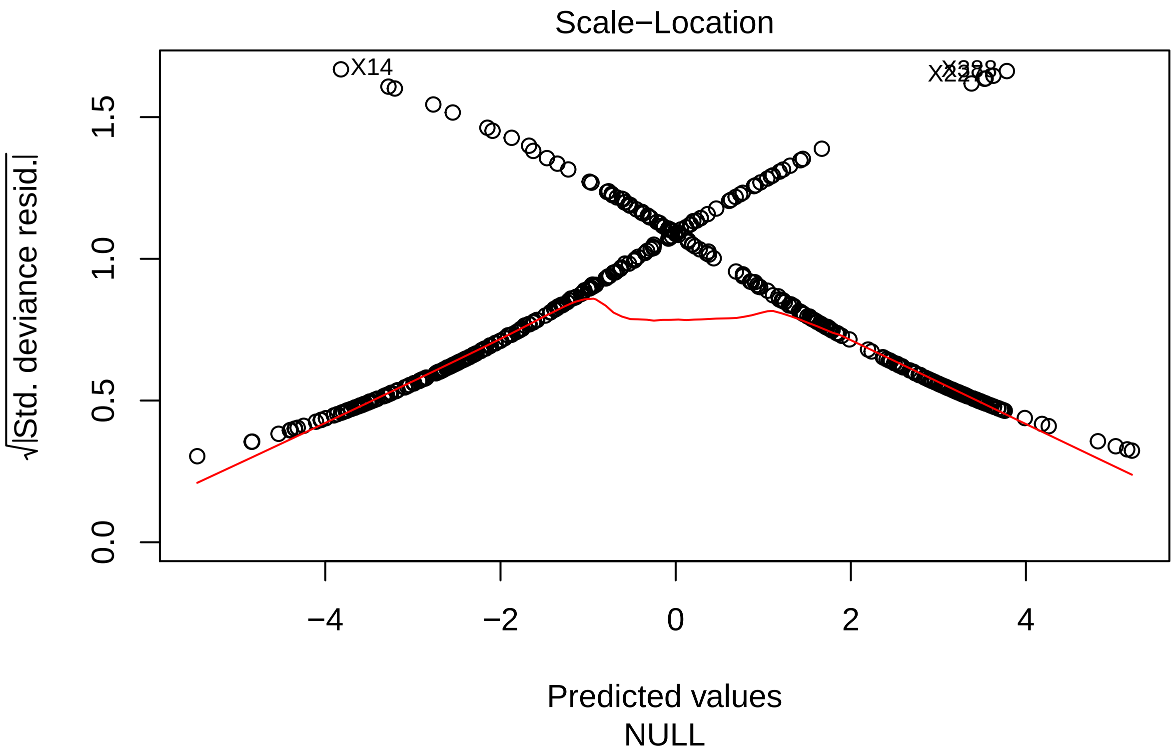
Before we make predictions, let's run this final model over our full dataset, and review some summary diagnostic plots and output.

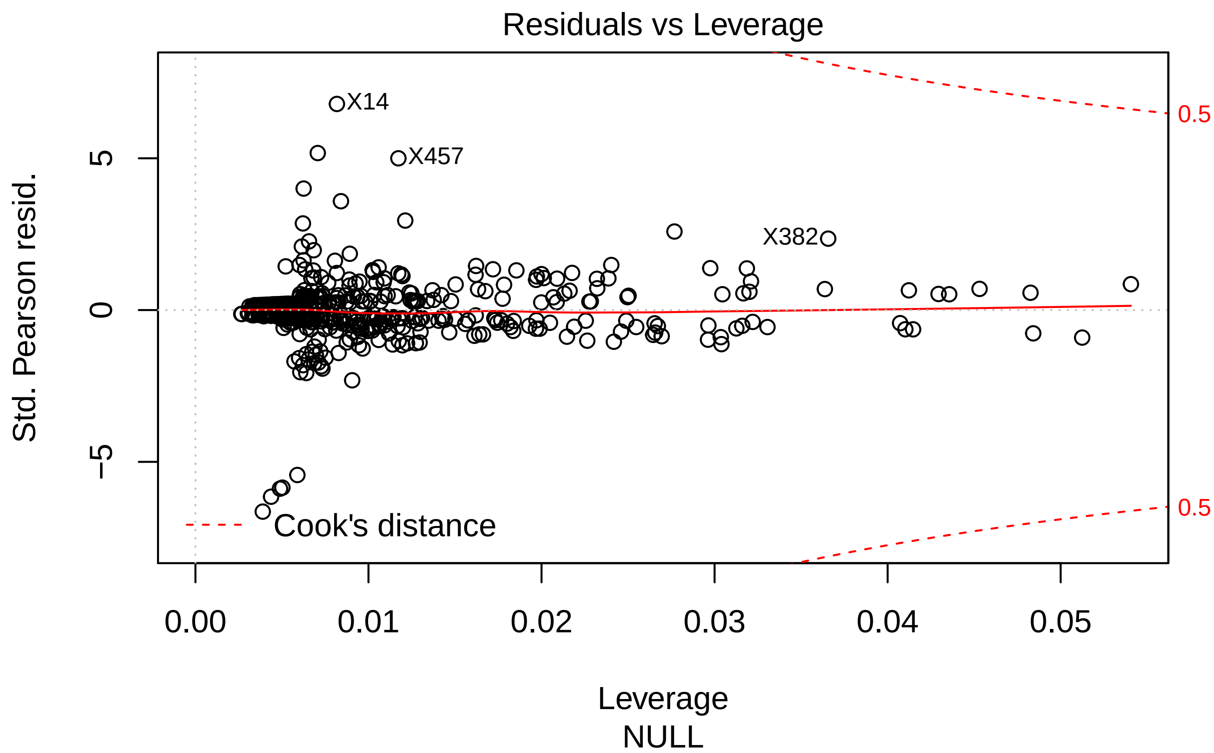
```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7585  -0.4775  -0.1510   0.4092   2.7727
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.02129    0.15913  -0.134  0.89357
## age          1.09457    0.23318   4.694 2.68e-06 ***
## dis         -0.69898    0.24751  -2.824  0.00474 **
## tax          1.31365    0.21438   6.128 8.91e-10 ***
```

```
## medv         0.38137    0.17572   2.170  0.02999 *
## ---
## 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: 333.92  on 461  degrees of freedom
## AIC: 343.92
##
## Number of Fisher Scoring iterations: 5
```




Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
NULL
X14
X38
X227
3
2
1
0
−1
−2
−3


Scale−Location
X14
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
−4
−2
0
2
4
Predicted values
NULL



Odds Ratio

We'll also create a table of the Odds Ratio for our final model beside the 95% confidence interval of those boundaries.

	OddsRatio	2.5 %	97.5 %
(Intercept)	0.979	0.717	1.342
age	2.988	1.915	4.790
dis	0.497	0.299	0.793
tax	3.720	2.493	5.802
medv	1.464	1.046	2.087

So we can now say that with a one unit increase in the scaled age variable, the odds of the neighborhood being below the median crime rate increase by 2.988%.

All that is left is to use our final to make predictions over the test dataset.

Make Predictions

We make our final predictions, create a dataframe with the prediction and the predicted probabilities. We can see from the head of our final dataframe and the table output of our predicted variable class that the prediction distribution seems similar to our initial test distribution.

0	1	prediction
0.8177905	0.1822095	0
0.6461040	0.3538960	0
0.5519857	0.4480143	0
0.6788301	0.3211699	0
0.8995530	0.1004470	0
0.9157008	0.0842992	0

Var1	Freq
0	23
1	17

R source code

Please see HW3.rmd on GitHub for source code.
Prediction File Predictions.csv