IS621_hw4

Charley Ferrari

April 3, 2016

Data Exploration

This crime dataset includes 466 observations of 14 variables, with no missing values. Each observation is a neighborhood in Boston. I will be looking at "target" as my dependent variable. This refers to whether or not the neighborhood has a crime rate above the median (1) or below (0)

First lets look over a summary of our variables:

```
##        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         black            lstat
##  Min.   :187.0   Min.   :12.6   Min.   :  0.32   Min.   : 1.730
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.:375.61   1st Qu.: 7.043
##  Median :334.5   Median :18.9   Median :391.34   Median :11.350
##  Mean   :409.5   Mean   :18.4   Mean   :357.12   Mean   :12.631
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:396.24   3rd Qu.:16.930
##  Max.   :711.0   Max.   :22.0   Max.   :396.90   Max.   :37.970
##       medv           target
##  Min.   : 5.00   Min.   :0.0000
##  1st Qu.:17.02   1st Qu.:0.0000
##  Median :21.20   Median :0.0000
##  Mean   :22.59   Mean   :0.4914
##  3rd Qu.:25.00   3rd Qu.:1.0000
##  Max.   :50.00   Max.   :1.0000
```

Because this is logistic regression, there is no normality assumption with our variables, so we don't have to worry about power transformations.

There are still some transformations I will get to in the next section, but before we begin checking individual variables, lets look at a correlation matrix. For simplicity, I'll just display any correlations greater than 0.5 or less than -0.5 as 1, and 0 otherwise:

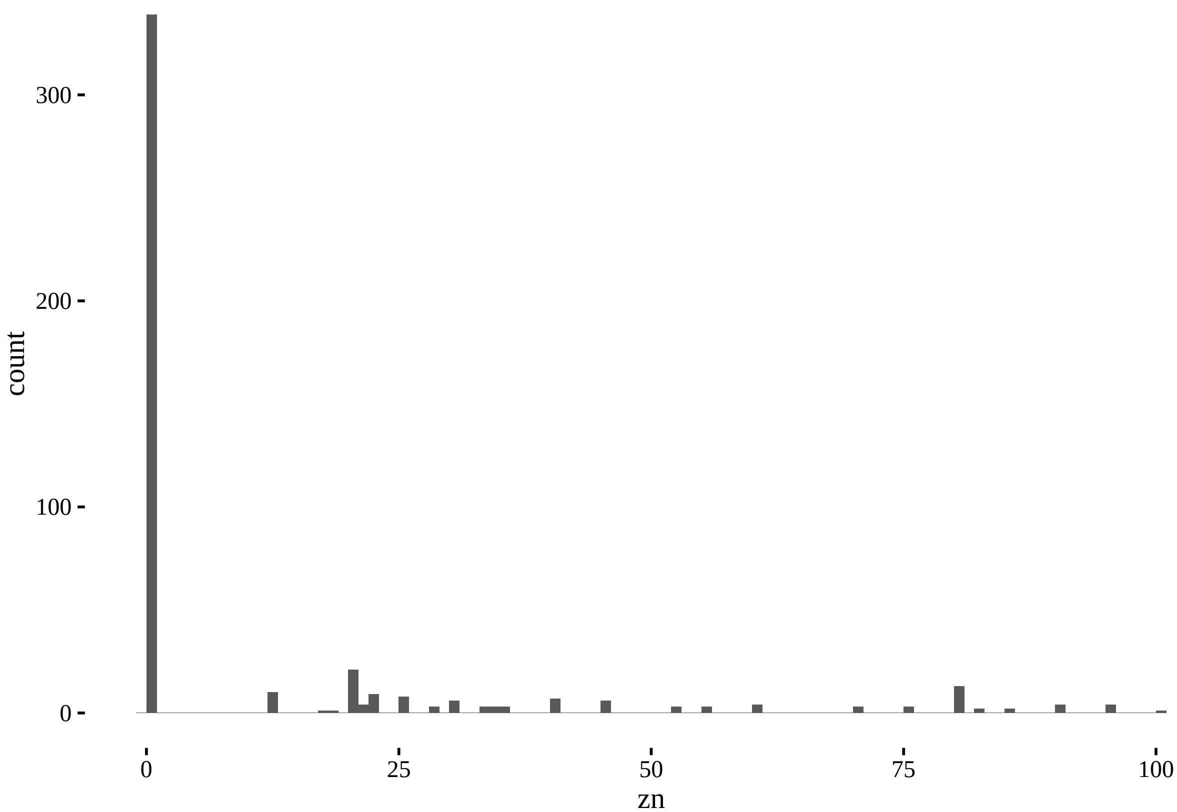
	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	target
zn	1	1	0	1	0	1	1	0	0	0	0	0	0	0
indus	1	1	0	1	0	1	1	1	1	0	0	1	0	1

	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	target
chas	0	0	1	0	0	0	0	0	0	0	0	0	0	0
nox	1	1	0	1	0	1	1	1	1	0	0	1	0	1
rm	0	0	0	0	1	0	0	0	0	0	0	1	1	0
age	1	1	0	1	0	1	1	0	1	0	0	1	0	1
dis	1	1	0	1	0	1	1	0	1	0	0	1	0	1
rad	0	1	0	1	0	0	0	1	1	0	0	1	0	1
tax	0	1	0	1	0	1	1	1	1	0	0	1	0	1
ptratio	0	0	0	0	0	0	0	0	0	1	0	0	1	0
black	0	0	0	0	0	0	0	0	0	0	1	0	0	0
lstat	0	1	0	1	1	1	1	1	1	0	0	1	1	0
medv	0	0	0	0	1	0	0	0	0	1	0	1	1	0
target	0	1	0	1	0	1	1	1	1	0	0	0	0	1

I will use this as a guideline during my model selection. Because this is a logistic regression, the VIF won't apply, so this will be my main indication of multicollinearity issues.

Data Preparation

There are no missing values, so we don't have to worry about the treatment of them. There are still some problematic issues with our variables however. First, the "zn" variable, proportion of residential land zoned for large lots, seems very heavily skewed:



A large number of neighborhoods, 73% of them, have no residential land zoned for large lots. It might make sense to make this a binary variable: neighborhoods with no large lots and neighborhoods with large lot zoning.

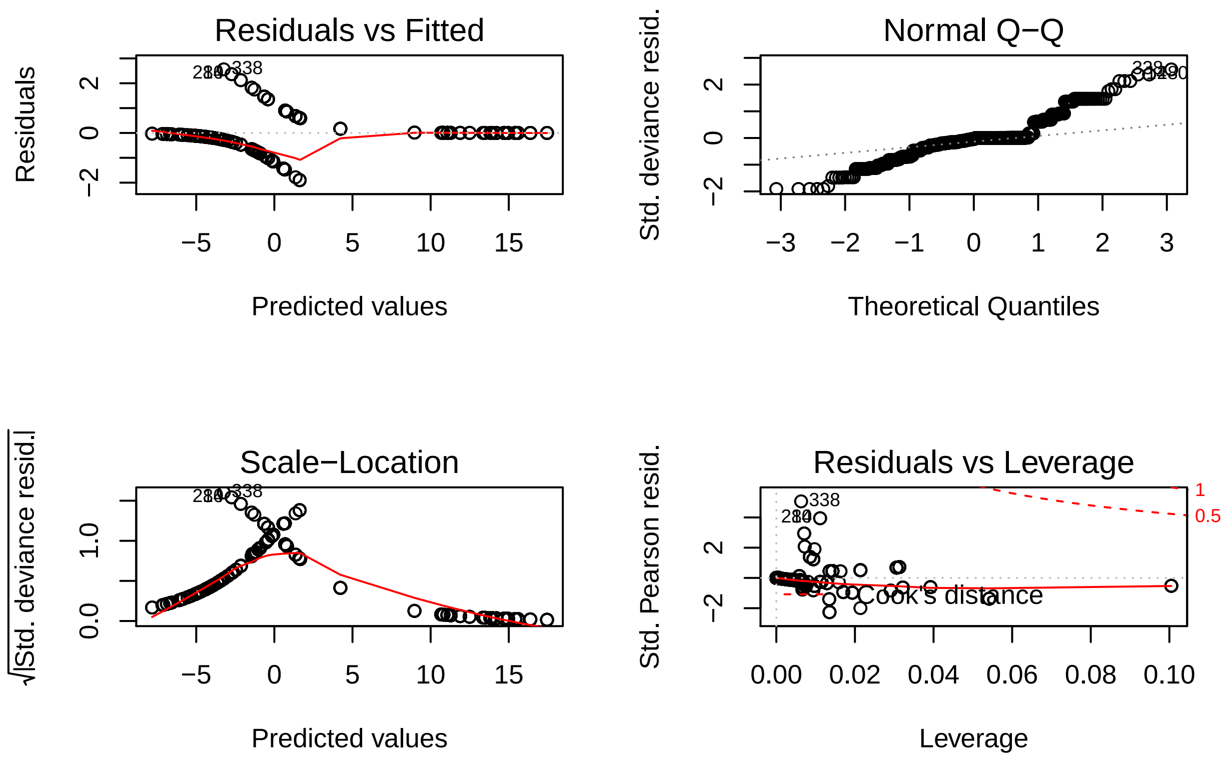
There are a few other weird distributions, proportion of non-retail business acreage and property tax rates for example have suspicious spikes, and the age of buildings variable seems to cap at 100 (suggesting any building greater than 100 years old are just listed as 100), but there aren't any clear ways to deal with these variables. I'll try for a best subset with what I have, and just remove these variables if the results aren't what I expect.

Build Models

For my first model, I'll use the bestglm function (from the bestglm library) to find the best subset out of all of my variables using the Bayesian Information Criterion.

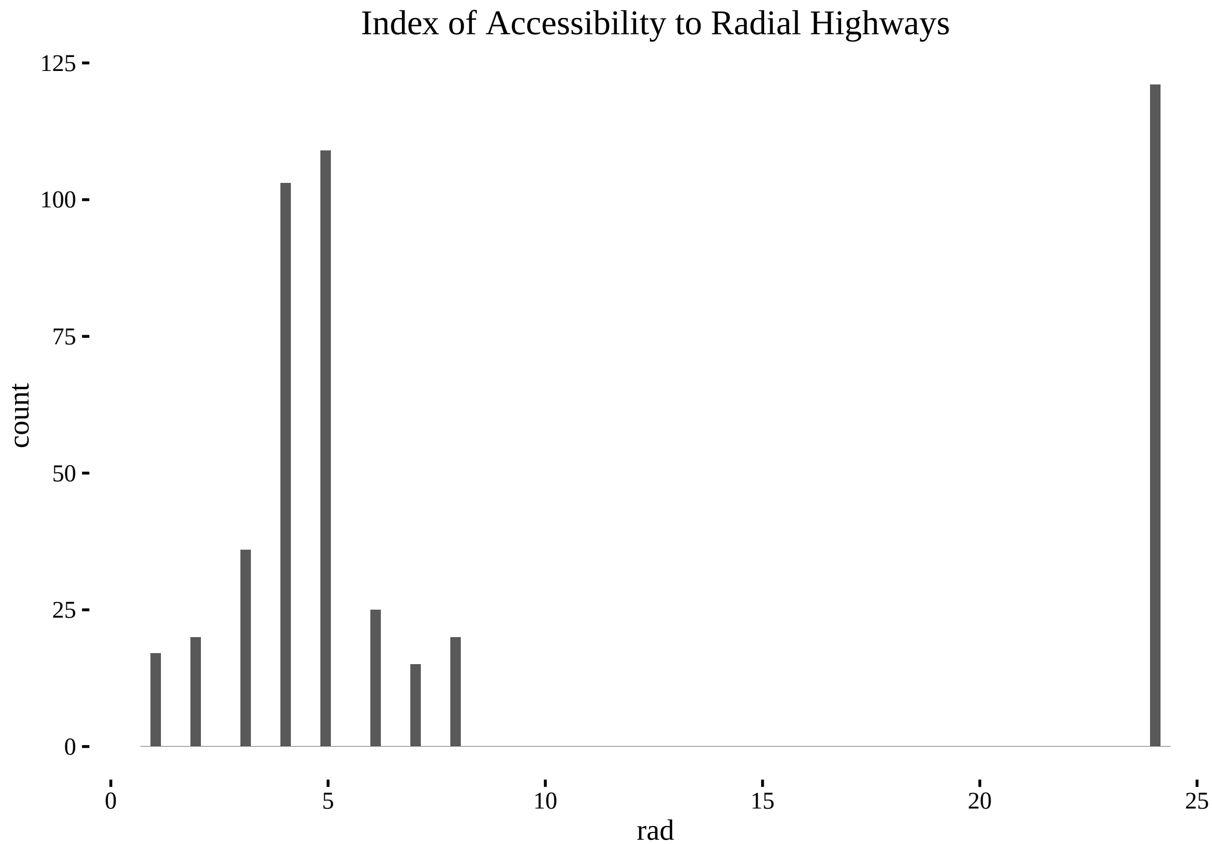
```
## Morgan-Tatar search since family is non-gaussian.
```

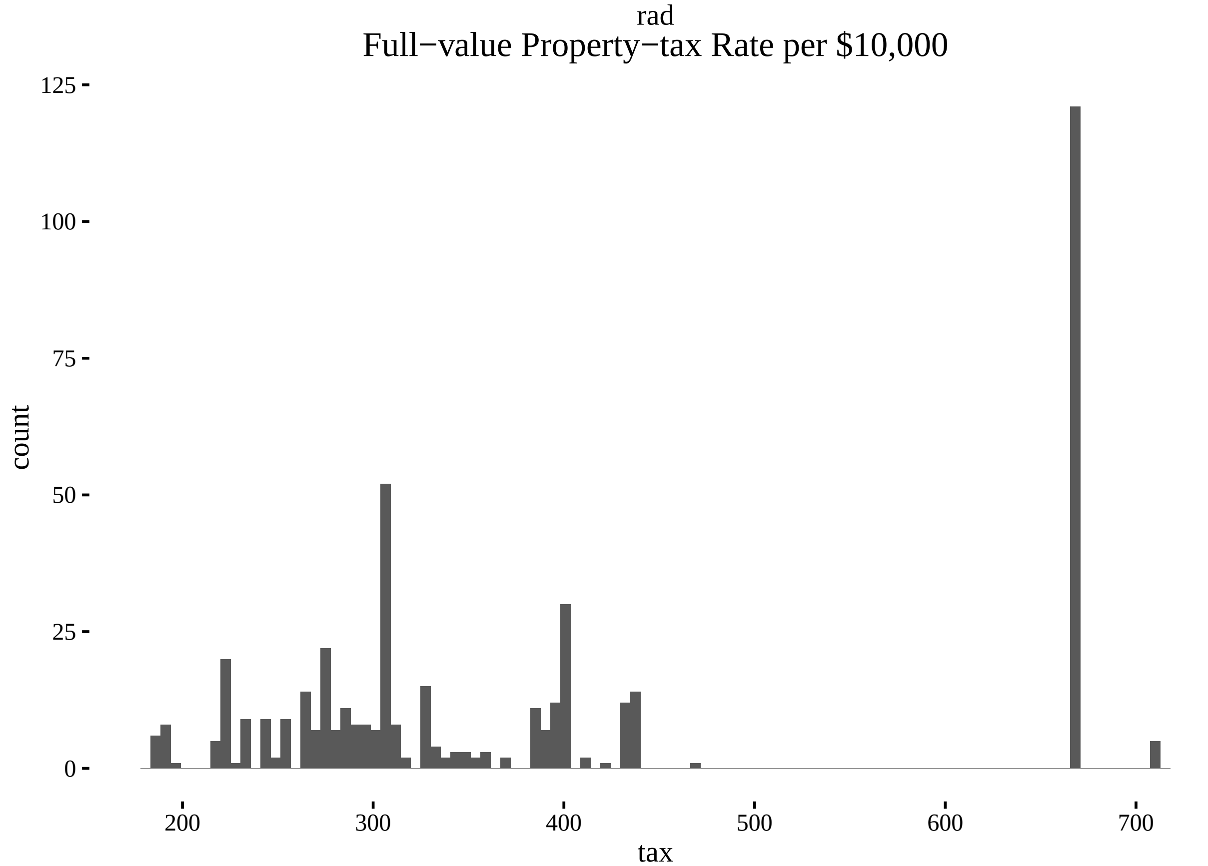
```
##
## Call:
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.89721  -0.27798  -0.03997   0.00557   2.55954
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.867422   2.368325  -8.389  < 2e-16 ***
## nox          35.633515   4.523677   7.877 3.35e-15 ***
## rad           0.637643   0.119444   5.338 9.38e-08 ***
## tax          -0.008146   0.002332  -3.493 0.000478 ***
## ---
## 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: 224.47  on 462  degrees of freedom
## AIC: 232.47
##
## Number of Fisher Scoring iterations: 8
```



I end up with three variables, all of which are highly significant according to their p-values. Index of accessibility to radial highways, property tax rate, and nitrogen oxide concentration.

Neighborhoods with large lots were already signaled as potentially problematic and accounted for by converting to a binary variable, but property tax rates and indexof acccessibility also had odd distributions:





No information is given about the index of accessibility for radial highways. They seem to have a distribution around 1 to 8 in whole numbers, with a plurality of values equal to 24. Tax rates have a similar suspicious spike, which is a bit easier to explain (perhaps there is a regional organization that sets taxes for multiple neighborhoods, or this represents a large number of neighborhoods in the city limits of Boston, verus in the surrounding towns.) If indices of acessibility are calculated in a similar way (based on municipality), then it can also be suspect.

If I knew more about these variables, I would treat them in a similar way to large lot zoning. It might make sense to add a binary variable based on whether or not the neighborhood is located in the City of Boston for example.

There is also higher than normal correlations between these three variables, which may indicate multicollinearity:

	nox	tax	rad
nox	1.0000000	0.6538780	0.5958298
tax	0.6538780	1.0000000	0.9064632
rad	0.5958298	0.9064632	1.0000000

The nitrogen oxides concentration variable is the least problematic, and according to the anova table it reduces the deviance the most:

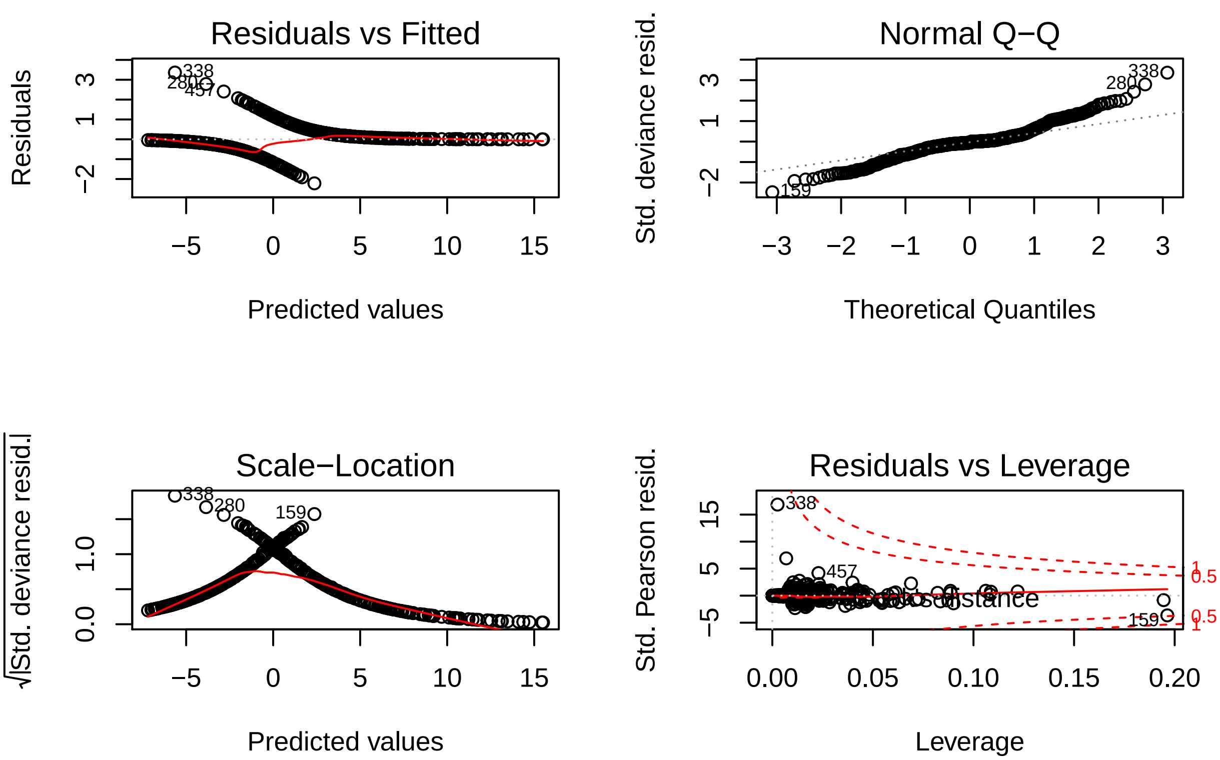
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                   465     645.88
## nox   1   353.86       464     292.01 < 2.2e-16 ***
## rad   1    52.50       463     239.51   4.3e-13 ***
## tax   1    15.04       462     224.47 0.0001053 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, I will recalculate my best subset model with the radial index and tax variables removed. Below is the new summary of this second best fit model:

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

```
##
## Call:
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2174  -0.3289  -0.0455   0.2651   3.3615
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -24.943055   4.357913  -5.724 1.04e-08 ***
## nox          41.551868   5.444047   7.633 2.30e-14 ***
## age           0.023928   0.009275   2.580 0.009885 **
## dis           0.813869   0.192498   4.228 2.36e-05 ***
## black        -0.012444   0.005061  -2.459 0.013943 *
## medv          0.125447   0.028186   4.451 8.56e-06 ***
## zn01         -2.405854   0.671090  -3.585 0.000337 ***
## ---
## 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: 251.35  on 459  degrees of freedom
## AIC: 265.35
##
## Number of Fisher Scoring iterations: 7
```



This new model improves our Q-Q plot in the residuals. We have more variables, but according to the below matrix, some of them are still highly correlated:

	nox	age	dis	black	medv	zn01
nox	1.0000000	0.7351278	-0.7688840	-0.3801549	-0.4301227	-0.5258884
age	0.7351278	1.0000000	-0.7508976	-0.2734677	-0.3781560	-0.5452811
dis	-0.7688840	-0.7508976	1.0000000	0.2938441	0.2566948	0.6600268
black	-0.3801549	-0.2734677	0.2938441	1.0000000	0.3300286	0.2225509
medv	-0.4301227	-0.3781560	0.2566948	0.3300286	1.0000000	0.3901422
zn01	-0.5258884	-0.5452811	0.6600268	0.2225509	0.3901422	1.0000000

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    465     645.88
## nox    1   353.86       464     292.01 < 2.2e-16 ***
## age    1     1.39       463     290.63 0.2388976
## dis    1     1.94       462     288.68 0.1635830
## black  1     7.66       461     281.02 0.0056382 **
## medv   1    14.64       460     266.39 0.0001303 ***
## zn01   1    15.04       459     251.35 0.0001054 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

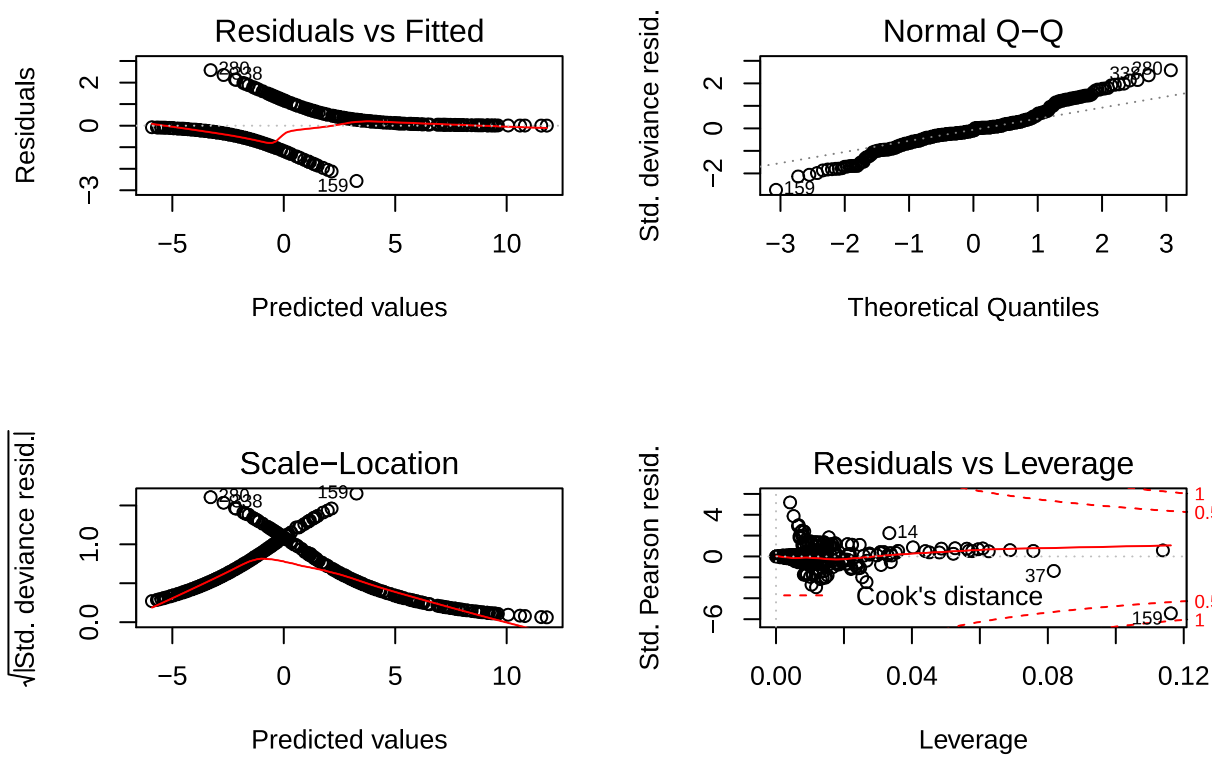
The transformed proportion of black residents and median value of owner occupied homes seem to be the least problematic in terms of correlation. According to our anova table, the proportion of Nitrous Oxide is adding the largest amount to the deviance (part of this is due to its placement as first in the anova table, but when placed last it still adds 108, which shows how significant it is.) Age, weighted distance to employment centers, and the binary variable of large lot zoning are highly correlated with eachother. Out of these three, the large lot zoning seems to lead to the most change in deviance. Age is also a problematic variable (it is capped at 100, which implies that value is given to any home with ages greater than 100.)

Because of this, I will recalculate a new best subset, this time removing age distance to employment centers along with property tax rates and index of accessibility to radial highways.

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

```
##
## Call:
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.56909  -0.39456  -0.09267   0.26649   2.57629
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.496957   3.755499  -5.192 2.09e-07 ***
## nox          30.692915   3.189908   9.622  < 2e-16 ***
## ptratio       0.280234   0.095455   2.936 0.003327 **
## black        -0.012083   0.005723  -2.111 0.034735 *
## medv          0.096654   0.025503   3.790 0.000151 ***
## ---
## 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: 267.11  on 461  degrees of freedom
## AIC: 277.11
##
## Number of Fisher Scoring iterations: 7
```



This model does lead to a further reduction in the Normal Q-Q plot, and has fewer variables leading to a more parsimonious solution. The correlation matrix also shows variables that are less correlated with eachother than in the previous two models:

	nox	ptratio	black	medv
nox	1.0000000	0.1762687	-0.3801549	-0.4301227
ptratio	0.1762687	1.0000000	-0.1816395	-0.5159153
black	-0.3801549	-0.1816395	1.0000000	0.3300286
medv	-0.4301227	-0.5159153	0.3300286	1.0000000

For one last sanity check, lets analyze whether or not the effect of the variables make intuitive sense in each model.

For model 1, NOX seems to have a high positive effect on crime, while the index of accessibility to radial highways has a smaller positive effect, and tax has a negative effect (meaning the higher the property tax rate, the lower the crime.) Nitrous Oxide seems like the most useful variable here (perhaps this chemical could have a similar effect to lead on crime.) One could imagine that access to radial highways would increase access to a neighborhood, therefore making it a target to crimes committed by people from other areas of the city. Tax doesn't really have an intuitive link to crime in my opinion, but out of these three models this one is the weakest due to the problematic variable distributions.

For model 2, nitrous oxide still has a heavy positive effect on crime. The age of buildings, distance from employment centers, and median home values have positive effects on crime, while proportion of black residents and large zoning lots have negative effects on crime.

The most problematic finding here is the difference in effect between median home values and large zoning lots. Intuitively, I would think these two variables would be highly correlated, and that they would have similar effects on crime. Their correlation is only 0.39, meaning they're not correlated, and thus the fact that they have different effects is not problematic. I would single this relationship out for further research however.

For model 3, Nitrous Oxide once again has the largest effect, with small effects caused by the proportion of black residents, pupil teacher ratios, and median home value. The pupil teacher ratio has a positive effect on crime, which means higher crime neighborhoods tend to have more students per classroom. The sign of all of these effects remained the same from model 2.

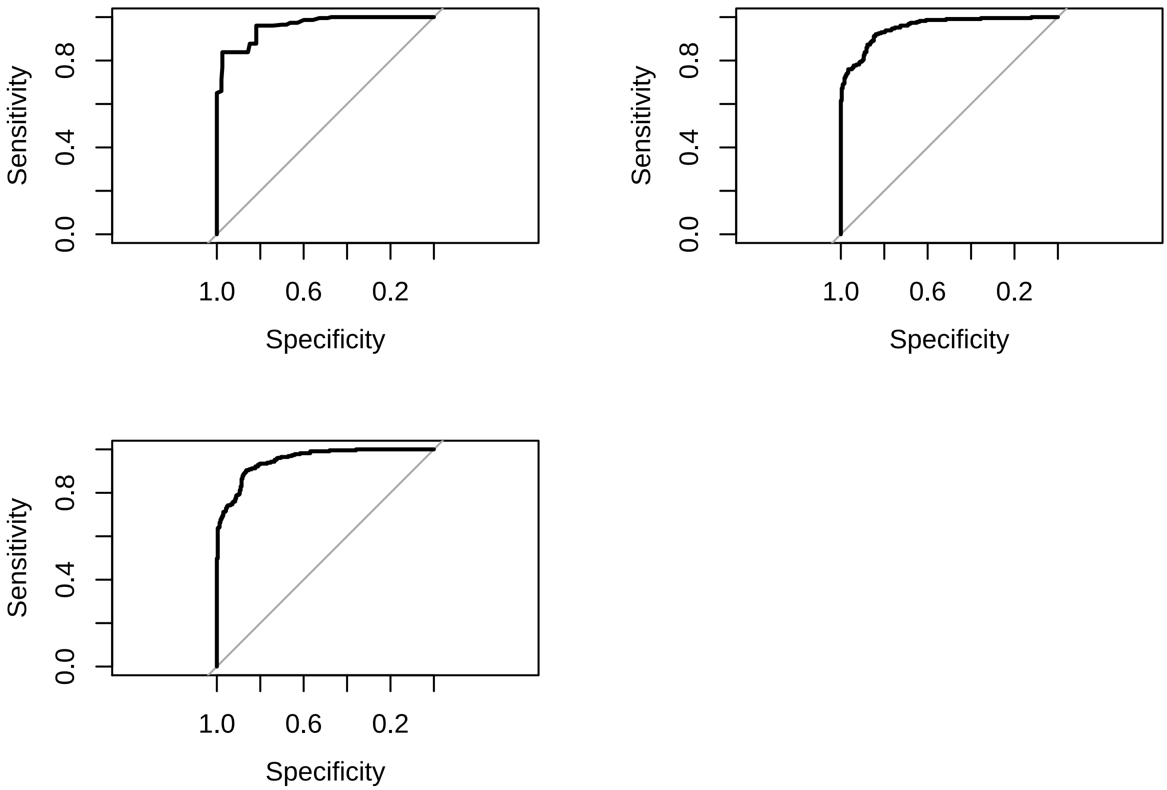
Select Models

The main method I'll use to select my models will be the ROC curves. Below you'll see plots of the three curves, along with the areas under the curves, or AUC:

```
##
## Call:
## roc.formula(formula = factor(target) ~ model1p, data = crime)
##
## Data: model1p in 237 controls (factor(target) 0) < 229 cases (factor(target) 1).
## Area under the curve: 0.9594
```

```
##
## Call:
## roc.formula(formula = factor(target) ~ model2p, data = crime)
##
## Data: model2p in 237 controls (factor(target) 0) < 229 cases (factor(target) 1).
## Area under the curve: 0.9531
```

```
##
## Call:
## roc.formula(formula = factor(target) ~ model3p, data = crime)
##
## Data: model3p in 237 controls (factor(target) 0) < 229 cases (factor(target) 1).
## Area under the curve: 0.951
```



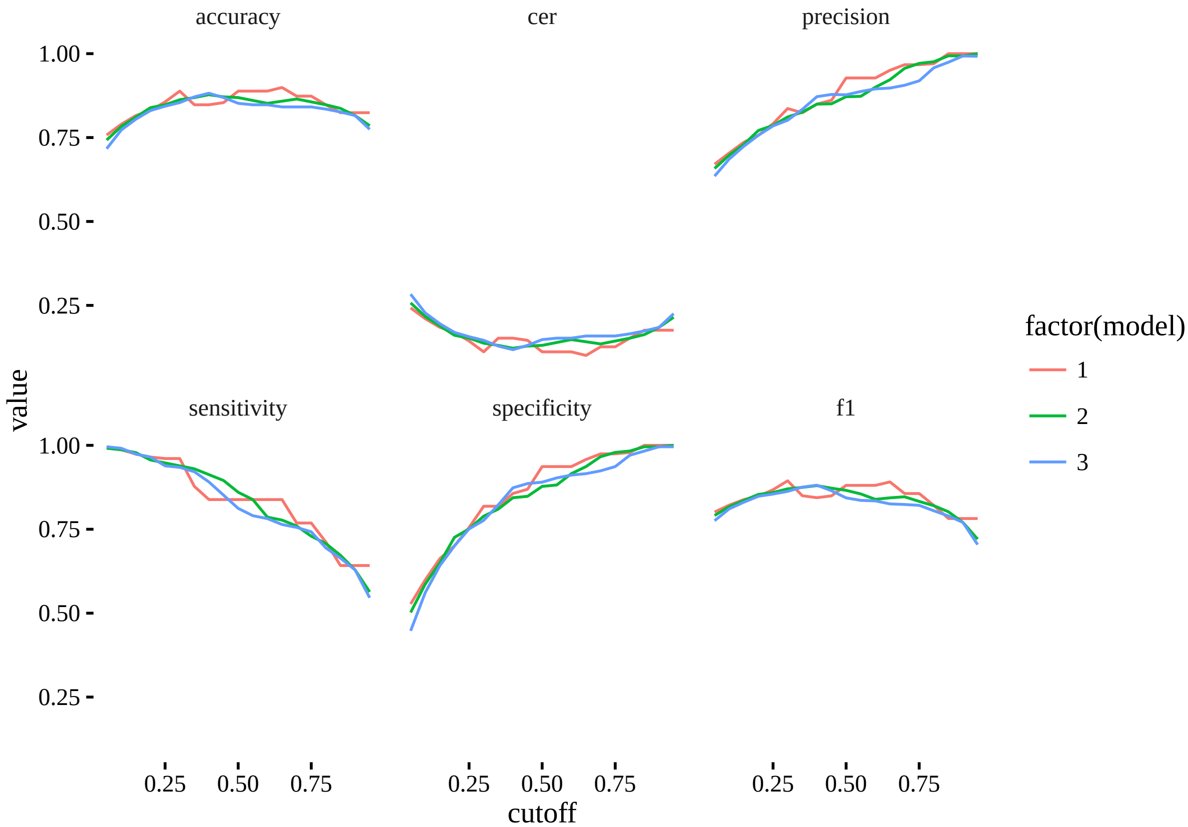

```
## [1] "Model 1 Area under the curve: 0.959445764929154"

## [1] "Model 2 Area under the curve: 0.953070587584987"

## [1] "Model 3 Area under the curve: 0.951043797099847"
```

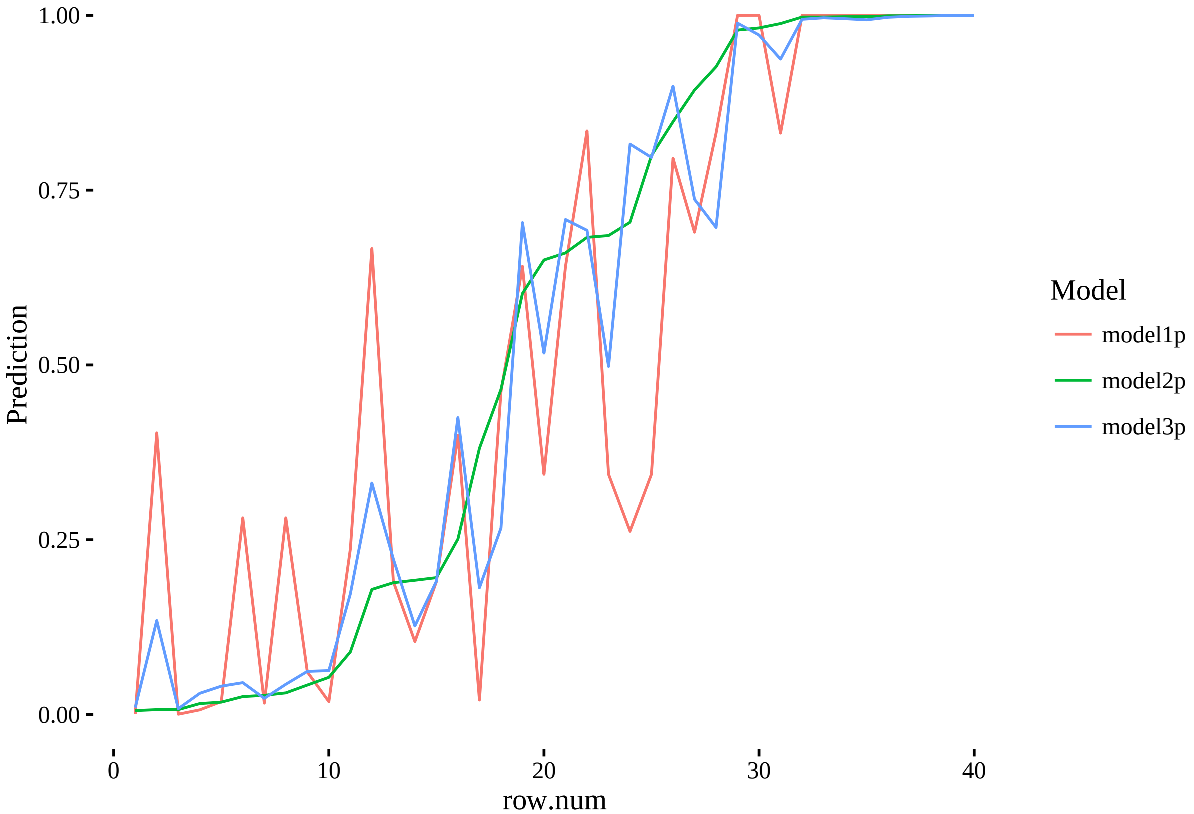
These are all good measures, around 0.95 for the AUC. Model 1 has the highest AUC, and it was further analysis that led us to analyze further models.

Accuracy, classification error rate, precision, sensitivity, specificity, and f1 score depend on the cutoff number we use. I visualized these measures below for my three models.



These graphs show that the models have very similar measures at different cutoff points. Model 1 appears more jagged, which can be expected given the distributions of the underlying data. Without further study of the variables, it would be tough to say which model is best. The measures such as tax and index of accessibility performed well, but if more was known about their derivation other transformations could have been performed (similar to what we did with the large lot zoning variable.)

Lastly, we could use our models to calculate probabilities for our evaluation dataset:



I decided to visualize this as well. I sorted my probabilities by model 2, and then see the differences between that and what models 1 and 3 predicted. I chose to sort by model 2 because the results of models 2 and 3 were the closest, while 1 was more different from both.

The fact that models 2 and 3 were so similar on the evaluation set lead me to trust their predictive capability over model 1. The parsimony of model 3, along with the fact that it includes the most problematic variables taken away, would lead me to trust it for this data.

Interestingly, Models 1 and 3 tend to depart from model 2 in similar directions. If one looks at the spikes of model 1, one can see less pronounced spikes in model 3. This leads me to believe model 3 captures some of the effects of model 1, without being too influenced by them.

While model 3 would be my final model choice, it's obvious the percentage of Nitrous Oxide is describing the most about crime levels. I would