Data 621 Group 2 HW 3: Crime

Members: Omar Pineda, Jeff Littlejohn, Sergio Ortega Cruz, Chester Poon, Simon Ustoyev

Due: October 30, 2019

Assignment

Build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation data set using your binary logistic regression model. You can only use the variables given to you (or variables that you derive from the variables provided). Use 0.5 threshold. 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)
- indus: proportion of non-retail business acres per suburb (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)**

Write Up:

1. Data Exploration

The dataset includes information on 466 neighborhoods in the city of Boston. Despite its East Coast location and reputation as a bastion of liberalism, Boston is among the most racially segregated of American cities. Attempts to integrate the schools using busing in the 1970s led to sustained violence (https://en.wikipedia.org/wiki/Boston_desegregation_busing_crisis), including deaths. Recent scholarship has highlighted the widespread use of redlining, a process by which institutions such as banks refused to offer mortgages or other financial services to people of certain races if they wished to purchase a home in certain neighborhoods despite creditworthiness.

In short, one would probably not want to construct a model to predict crime by neighborhood that uses variables such as race without having a clear idea of the model's intended use and an ethical framework for evaluating said model. This, however, is an academic exercise, so we proceed. Let's preview the data.

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
0	18.10	0	0.693	5.453	100.0	1.4896	24	666	20.2	30.59	5.0	1
0	18.10	0	0.693	4.519	100.0	1.6582	24	666	20.2	36.98	7.0	1
0	5.19	0	0.515	6.316	38.1	6.4584	5	224	20.2	5.68	22.2	0
80	3.64	0	0.392	5.876	19.1	9.2203	1	315	16.4	9.25	20.9	0
22	5.86	0	0.431	6.438	8.9	7.3967	7	330	19.1	3.59	24.8	0
0	12.83	0	0.437	6.286	45.0	4.5026	5	398	18.7	8.94	21.4	0
0	18.10	0	0.532	7.061	77.0	3.4106	24	666	20.2	7.01	25.0	1
22	5.86	0	0.431	8.259	8.4	8.9067	7	330	19.1	3.54	42.8	1
0	2.46	0	0.488	6.153	68.8	3.2797	3	193	17.8	13.15	29.6	0

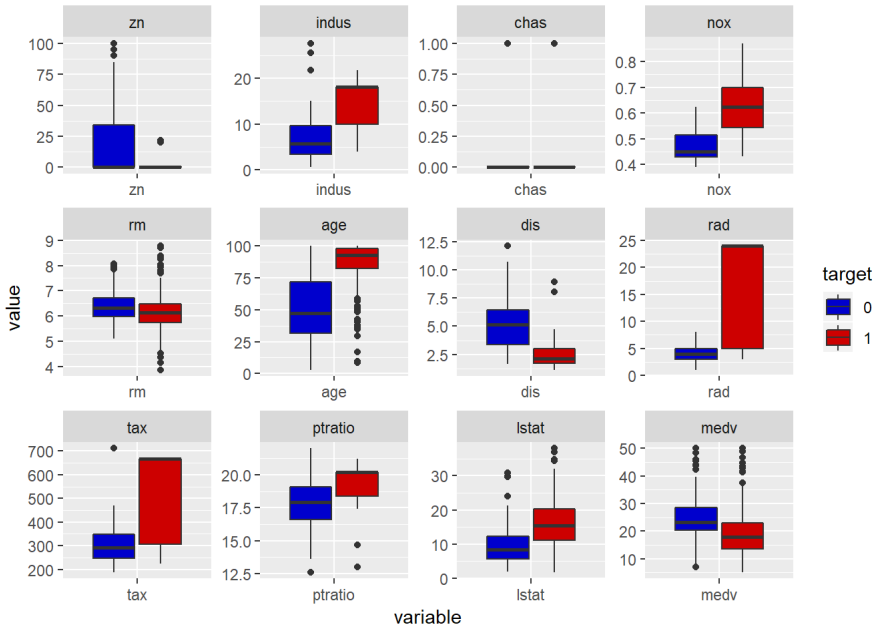
Expected variables are present. Note that, as indicated in the variables' descriptions, many of the variables have already been scaled or transformed in some way. Let's calculate summary statistics and generate a box plot for further review.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
zn	1	466	11.5772532	23.3646511	0.00000	5.3542781	0.0000000	0.0000	100.0000	100.0000	2.1768152	3.8135765

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
indus	2	466	11.1050215	6.8458549	9.69000	10.9082353	9.3403800	0.4600	27.7400	27.2800	0.2885450	-1.2432132
chas	3	466	0.0708155	0.2567920	0.00000	0.0000000	0.0000000	0.0000	1.0000	1.0000	3.3354899	9.1451313
nox	4	466	0.5543105	0.1166667	0.53800	0.5442684	0.1334340	0.3890	0.8710	0.4820	0.7463281	-0.0357736
rm	5	466	6.2906738	0.7048513	6.21000	6.2570615	0.5166861	3.8630	8.7800	4.9170	0.4793202	1.5424378
age	6	466	68.3675966	28.3213784	77.15000	70.9553476	30.0226500	2.9000	100.0000	97.1000	-0.5777075	-1.0098814
dis	7	466	3.7956929	2.1069496	3.19095	3.5443647	1.9144814	1.1296	12.1265	10.9969	0.9988926	0.4719679
rad	8	466	9.5300429	8.6859272	5.00000	8.6978610	1.4826000	1.0000	24.0000	23.0000	1.0102788	-0.8619110
tax	9	466	409.5021459	167.9000887	334.50000	401.5080214	104.5233000	187.0000	711.0000	524.0000	0.6593136	-1.1480456
ptratio	10	466	18.3984979	2.1968447	18.90000	18.5970588	1.9273800	12.6000	22.0000	9.4000	-0.7542681	-0.4003627
lstat	11	466	12.6314592	7.1018907	11.35000	11.8809626	7.0720020	1.7300	37.9700	36.2400	0.9055864	0.5033688
medv	12	466	22.5892704	9.2396814	21.20000	21.6304813	6.0045300	5.0000	50.0000	45.0000	1.0766920	1.3737825
target	13	466	0.4914163	0.5004636	0.00000	0.4893048	0.0000000	0.0000	1.0000	1.0000	0.0342293	-2.0031131

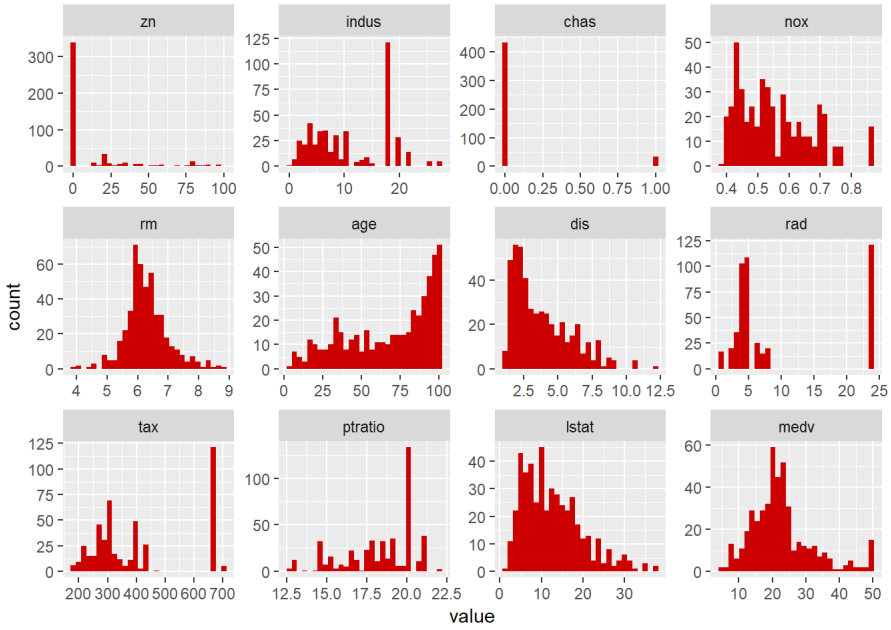
We see 466 records in our training set and no missing values for any variable. Other than that, without a specific question in mind, it's difficult to draw any conclusions from this big table of numbers. We see no missing values that would require imputation using medians or other methods.

Now, we visualize using box plots. We'll separate the box plots by the target value, which signifies whether or not the neighborhood is high crime. And we'll approximate Boston Red Sox colors.



The dummy variable (chas) that represents proximity to the Charles River is not meaningful, but clear distinctions in distributions between the neighborhoods in which the crime rate is below and above the median - the target variable by which the box plots are split. We might later look at these values after transformations such as logs.

To check for skewness, let's examine the distribution of each variable independent of target variable value.

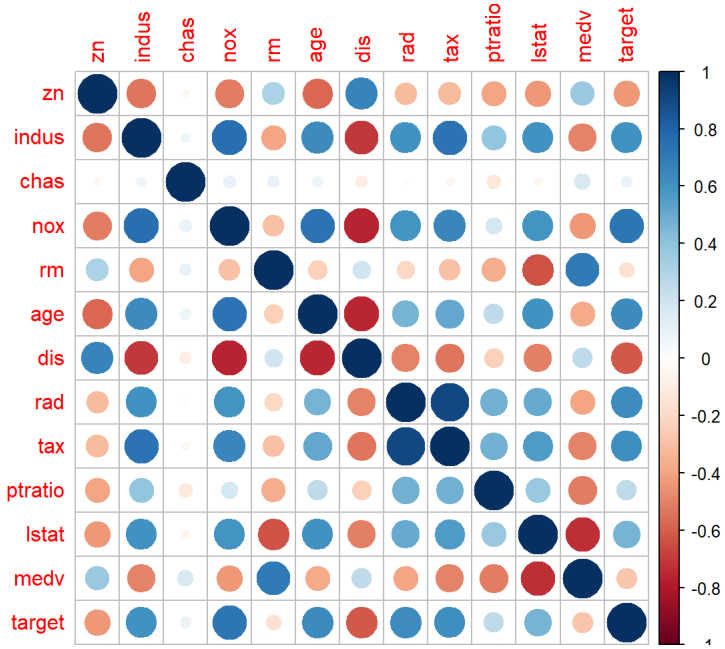


Skewness abounds. We will file this away for now and revisit in the Data Preparation part of the project. In particular, zn, nox, age, dis, ptratio, and lstat seem likely candidates for transformations.

We will now check for covariance.

	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
zn	1.0000000	-0.5382664	-0.0401620	-0.5170452	0.3198141	-0.5725805	0.6601243	-0.3154812	-0.3192841	-0.3910357	-0.4329925	0.376717
indus	-0.5382664	1.0000000	0.0611832	0.7596301	-0.3927118	0.6395818	-0.7036189	0.6006284	0.7322292	0.3946898	0.6071102	-0.496174
chas	-0.0401620	0.0611832	1.0000000	0.0974558	0.0905098	0.0788837	-0.0965771	-0.0159004	-0.0467648	-0.1286606	-0.0514232	0.161565
nox	-0.5170452	0.7596301	0.0974558	1.0000000	-0.2954897	0.7351278	-0.7688840	0.5958298	0.6538780	0.1762687	0.5962426	-0.430122
rm	0.3198141	-0.3927118	0.0905098	-0.2954897	1.0000000	-0.2328125	0.1990158	-0.2084457	-0.2969343	-0.3603471	-0.6320245	0.705336
age	-0.5725805	0.6395818	0.0788837	0.7351278	-0.2328125	1.0000000	-0.7508976	0.4603143	0.5121245	0.2554479	0.6056200	-0.378156
dis	0.6601243	-0.7036189	-0.0965771	-0.7688840	0.1990158	-0.7508976	1.0000000	-0.4949919	-0.5342546	-0.2333394	-0.5075280	0.256694
rad	-0.3154812	0.6006284	-0.0159004	0.5958298	-0.2084457	0.4603143	-0.4949919	1.0000000	0.9064632	0.4714516	0.5031013	-0.397668
tax	-0.3192841	0.7322292	-0.0467648	0.6538780	-0.2969343	0.5121245	-0.5342546	0.9064632	1.0000000	0.4744223	0.5641886	-0.490032
ptratio	-0.3910357	0.3946898	-0.1286606	0.1762687	-0.3603471	0.2554479	-0.2333394	0.4714516	0.4744223	1.0000000	0.3773560	-0.515915
lstat	-0.4329925	0.6071102	-0.0514232	0.5962426	-0.6320245	0.6056200	-0.5075280	0.5031013	0.5641886	0.3773560	1.0000000	-0.735800
medv	0.3767171	-0.4961743	0.1615653	-0.4301227	0.7053368	-0.3781560	0.2566948	-0.3976683	-0.4900329	-0.5159153	-0.7358008	1.000000
target	-0.4316818	0.6048507	0.0800419	0.7261062	-0.1525533	0.6301062	-0.6186731	0.6281049	0.6111133	0.2508489	0.4691270	-0.270550

We see some very high positive and negative correlations between variables. Let's construct a more effective visualization.



We see candidates for combination due to covariance.

As a final step, let's look just a correlation between the independent variables and the target variables.

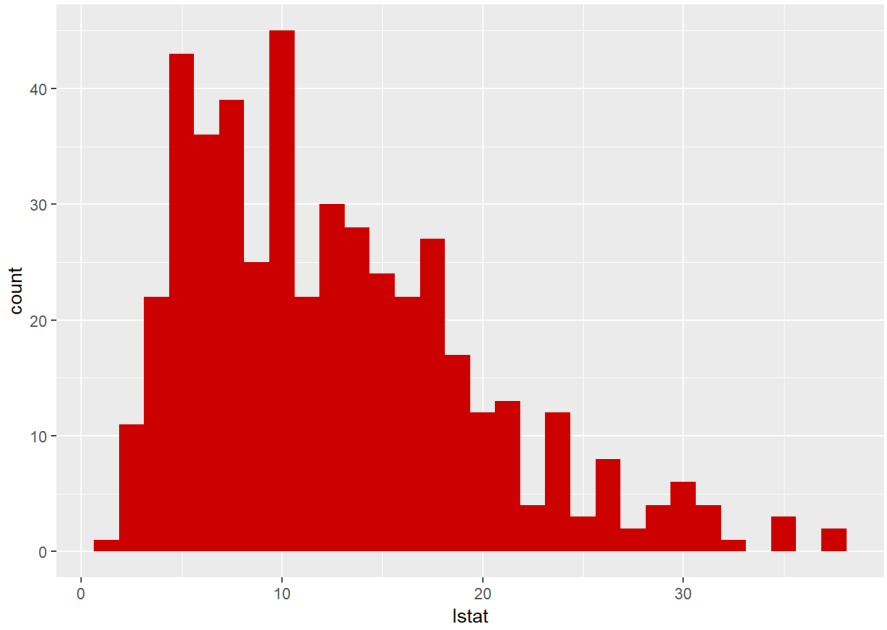
	target
zn	-0.4316818
indus	0.6048507
chas	0.0800419
nox	0.7261062
rm	-0.1525533
age	0.6301062
dis	-0.6186731
rad	0.6281049
tax	0.6111133
ptratio	0.2508489
lstat	0.4691270
medv	-0.2705507
target	1.0000000

Nox, or the concentration of nitrogen oxide, a significant pollutant that's harmful to human health, in a neighborhood, shows the closest correlation with the target variable at .73. Next, age, rad, tax, and indus all correlate with the target value just above .6. Zn showed the largest negative correlation with the target at -.43. Zn represents the percentage of residential lots zoned for large lots, which may be an indicator large rental housing - apartments.

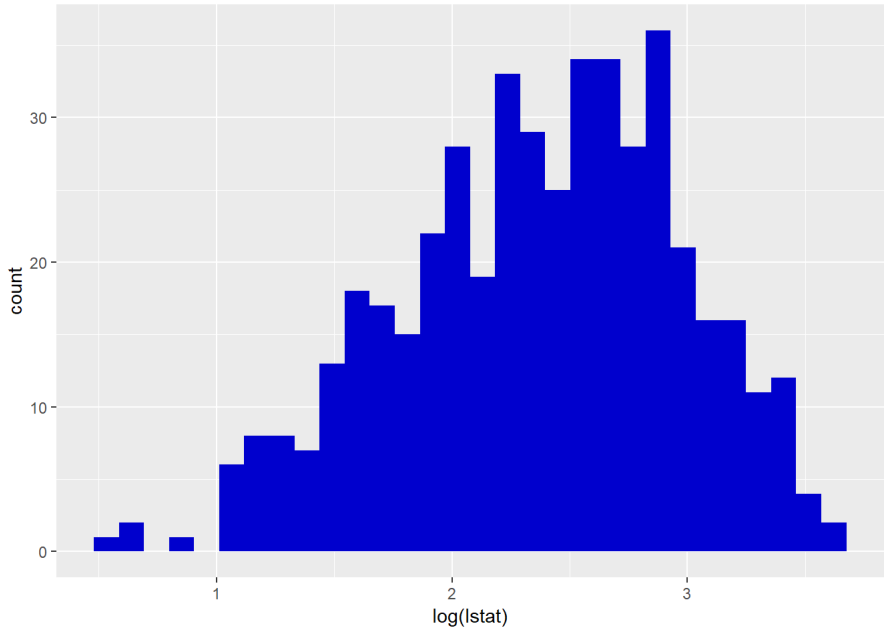
2. Data Preparation

Prior to modelling the training data set, we must prepare the data. We do not have any missing values, so imputation is not required. We will probably actually start with a model that uses all variables regardless of skew or covariance. However, we will definitely progress to using transformations and will also combine variables due to covariance in seeking the construction of accurate and valid models.

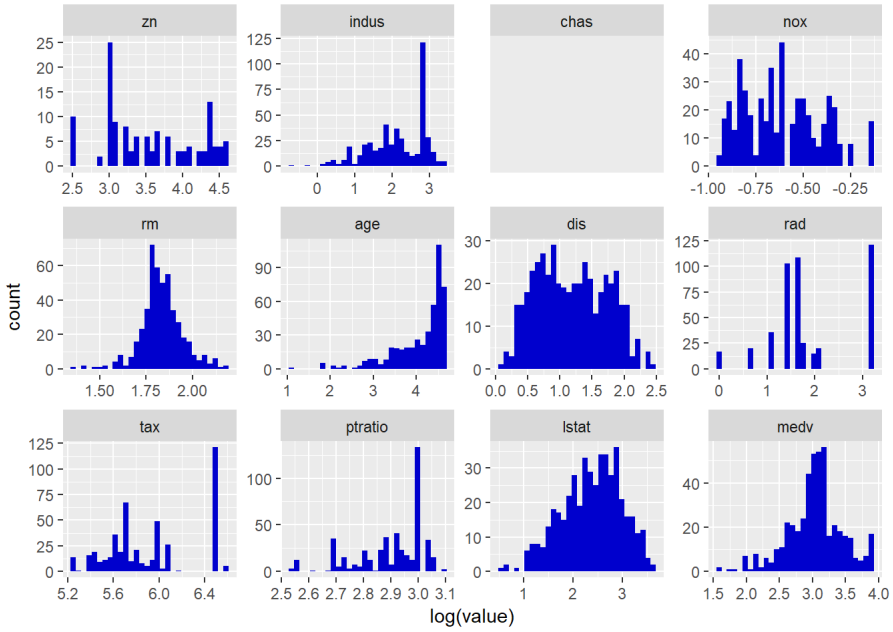
Let's look at how transformations might solve distribution issues with some of our variables. Earlier, we saw a strong right skew in the distribution of the variable lstat, which tracks the "lower status" of a neighborhood's population. Probably not the best phrasing.



What would a log transformation do to this distribution?

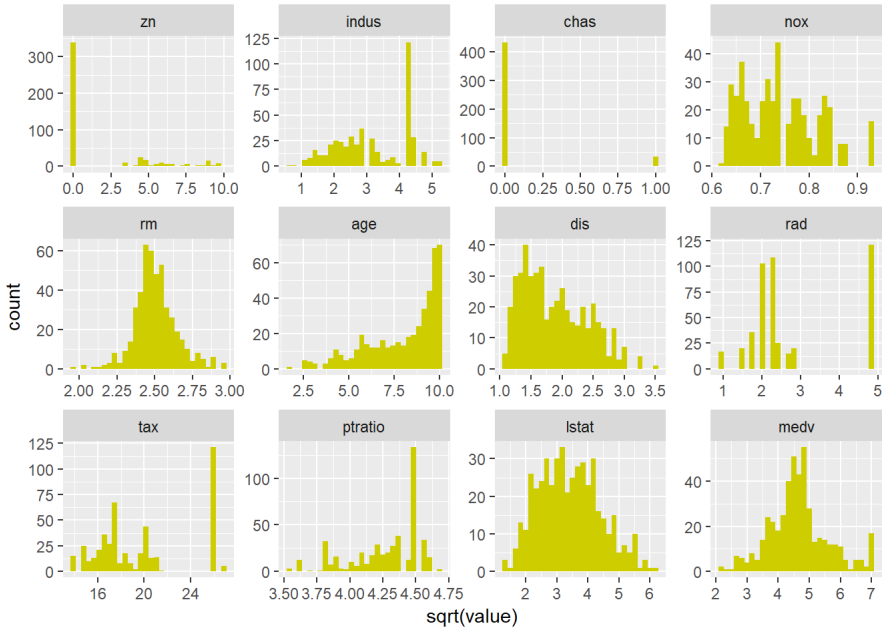


Looks slightly better. Let's generate log transformations for all variables in the dataset.



Medv looks slightly better. However, age remains strongly left skewed. Dis is now bimodal.

What about other transformations such as quadratic ones?



Nope. Not a lot of improvement.

In Part 1, we saw high covariances among variables such as rad and tax (.91). To build the best models, we'll likely want to examine combining some of these variables that are correlated to each other, which tends to increase standard errors. This can lead to overfitting and inefficient models. We will not combine variables here but instead revisit this concept in part 3 when evaluating our models.

Our textbooks have also discussed the possibility of creating bins for continuous variables. For example, dis, the weighted distance of means of distances from a neighborhood to five Boston job centers, might be better suited to fall into three categories than to remain a continuous variable for performance reasons. For time's sake, we will not explore this in this assignment.

3. Build Models

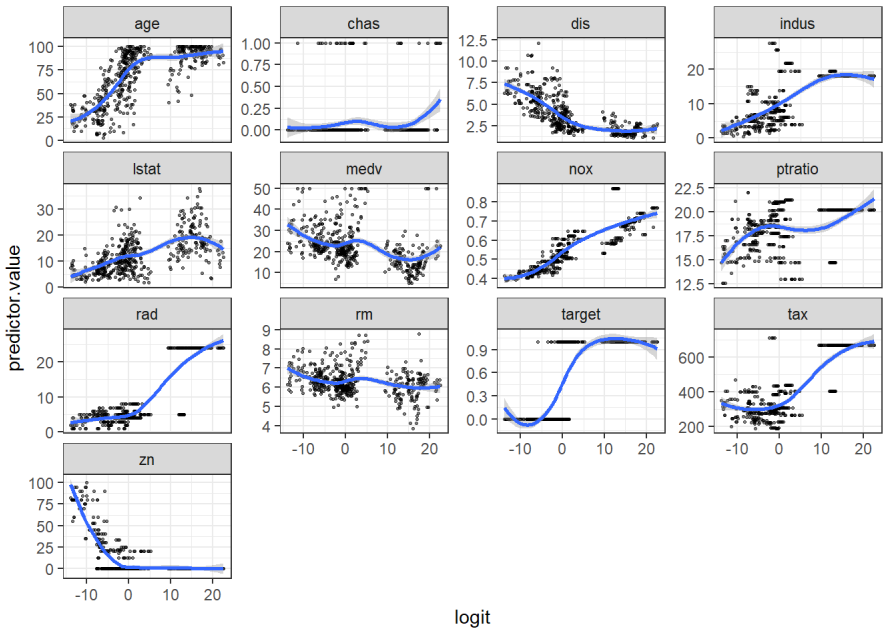
Following convention, we will start with a model consisting of all variables, none of which have been transformed. While we've moved on to Part 3, where we will construct the models, the boundary between data preparation and model building is grey. We will to explore transformations and collinearity.

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = crime_train_data)
##
## 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) -40.822934   6.632913  -6.155 7.53e-10 ***
## zn           -0.065946   0.034656  -1.903  0.05706 .
## indus        -0.064614   0.047622  -1.357  0.17485
## chas          0.910765   0.755546   1.205  0.22803
## nox          49.122297   7.931706   6.193 5.90e-10 ***
## rm           -0.587488   0.722847  -0.813  0.41637
## age           0.034189   0.013814   2.475  0.01333 *
## dis           0.738660   0.230275   3.208  0.00134 **
## rad           0.666366   0.163152   4.084 4.42e-05 ***
## tax          -0.006171   0.002955  -2.089  0.03674 *
## ptratio       0.402566   0.126627   3.179  0.00148 **
## lstat         0.045869   0.054049   0.849  0.39608
## medv          0.180824   0.068294   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
```

Our most significant variables generally tie to the variables we saw have the highest correlations with the target value earlier. We have an AIC of 218.05 and a residual deviance of 192.05.

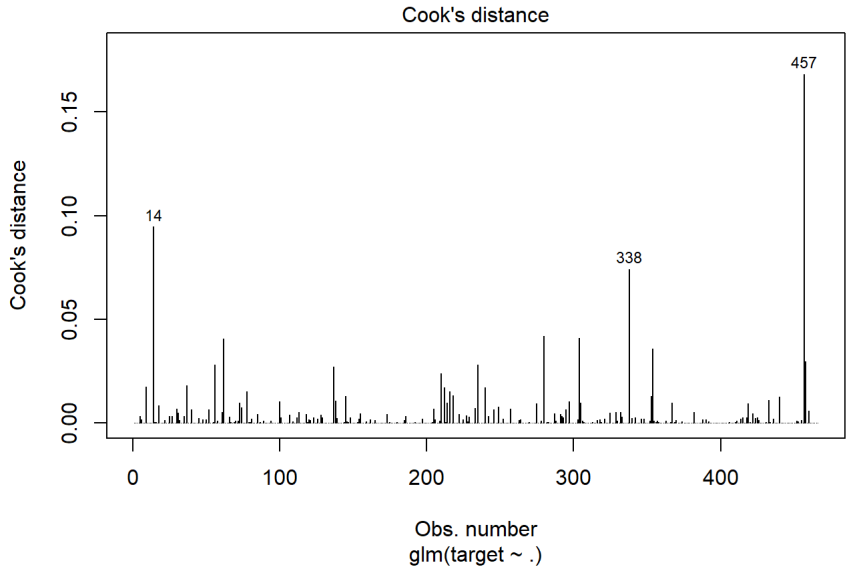
Let's run further diagnostics on the model. We will set a probability of .5 as being the cutoff for determining if a neighborhood will be high crime. Here, we check the relationship between the logit of the outcome and each predictive variable. (Target and the binary dummy variable chas should be ignored.) Again, these steps also could be labelled as data preparation.

```
##     1     2     3     4     5     6
## "pos" "pos" "pos" "neg" "neg" "neg"
```



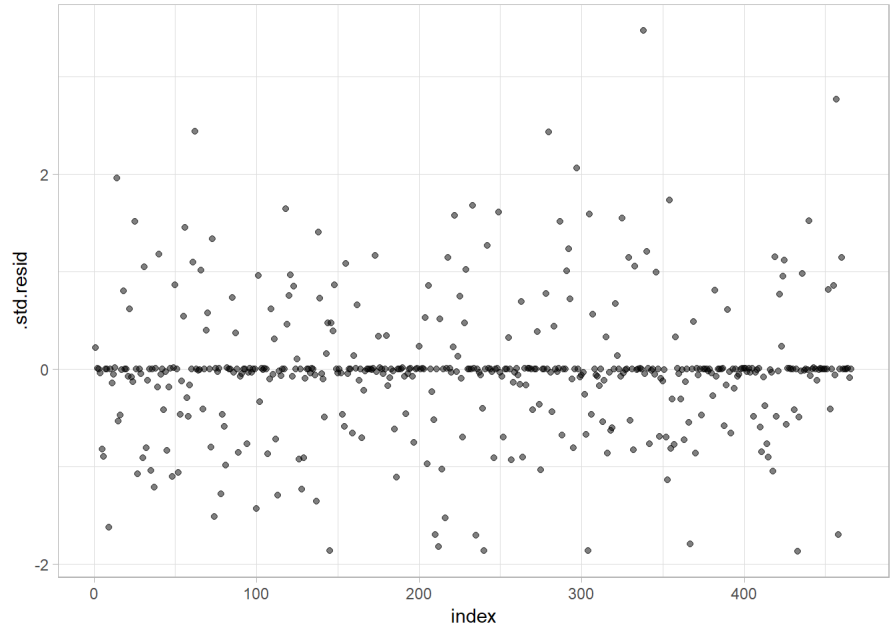
Tax and zn do not show linear associations with the outcomes in logit scale. Along with the previously discussed lstat, they might benefit from transformations.

Let's use Cook's Distance to check for outliers.



target	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd
1	22	5.86	0	0.431	8.259	8.4	8.9067	7	330	19.1	3.54	42.8	-1.247878	1.1180742	1.732213	0.2166514	0.6452970	0.0945920
1	20	6.96	0	0.464	5.856	42.1	4.4290	3	223	18.6	13.00	21.1	-6.005994	0.9824904	3.466541	0.0023667	0.6310574	0.0742398
1	0	10.59	0	0.489	5.412	9.8	3.5875	4	277	18.6	29.55	23.7	-3.596230	1.4378546	2.691946	0.0537163	0.6387020	0.1682452

An outlier is not necessarily influential. Let's check for that.



Let's pull that point that's above 3 standardized residuals from 0.

Observation 338 is an influential outlier.

Next, we check multicollinearity.

```
##       zn    indus     chas      nox       rm      age      dis      rad
## 1.823146 2.682271 1.241479 4.160497 5.813851 2.569961 3.887981 1.942967
##      tax  ptratio    lstat     medv
## 2.144040 2.275557 2.642656 8.122037
```

The rule of thumb is that vif scores above 5 should be judged as having a high amount of multicollinearity. So rm and medv have issues in this regard.

In summary, we have:

1. Multiple predictors that do not have linear relationships with the logit of the outcome variable.
2. One influential outlier - index 338.

3. Two predictors with potentially problematically high multicollinearity.

The above are among many methods to check assumptions and diagnostics of logistic regression models. We will not repeat these steps - other than the summary diagnostics - for our additional attempts at constructing a model to predict high-crime neighborhoods.

4. Select Models

References

Appendix