

Homework 3

Group 1

Contents

1	Introduction	2
2	Statement of the Problem	2
3	Data Exploration	2
3.1	Variables Explained	2
3.2	Exploration of Variables	3
3.3	Correlation Matrix	5
3.4	Outliers Treatment	6
3.5	BoxCox Transformations	7
4	Models Built	8
4.1	Model 1 - Backwards Selection Method	8
4.2	Model 2 - Forwards Selection Method	10
4.3	Model 3 - Subset Selection Method	11
5	Selected Model	13
6	Appendix A	14
6.1	Session Info	14
6.2	Data Dictionary	14
6.3	R source code	14

Prepared for:

Dr. Nathan Bastian

City University of New York, School of Professional Studies - Data 621

Prepared by:

Group 1

Senthil Dhanapal

Yadu Chittampalli

Christophe Hunt

1 Introduction

Crime has a high cost to all parts of society and it can have severe long term impact on neighborhoods. If crime rises in the neighborhood or it is invaded by criminals, then families and those with the economic means to leave for more stable areas will do so¹. Additionally, crime can even have a health cost to the community in that the perception of a dangerous neighborhood was associated with significantly lower odds of having high physical activity among both men and women². It is important to understand the propensity for crime levels of a neighborhood before investing in that neighborhood.

2 Statement of the Problem

The purpose of this report is to develop a statistical model to determine the variables that are independently associated with neighborhoods with crime rates above or below the median. Note that neighborhoods with crime rates above or below the median have already been provided in our evaluation data set.

3 Data Exploration

3.1 Variables Explained

The variables provided in our evaluation data set are explained below:

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

¹Effect of Crime on Real Estate Values. (1952). The Journal of Criminal Law, Criminology, and Police Science, 43(3), 357-357. Retrieved from [http://www.jstor.org/remote.baruch.cuny.edu/stable/1139159](http://www.jstor.org/remote/baruch.cuny.edu/stable/1139159)

²Bennett GG, McNeill LH, Wolin KY, Duncan DT, Puleo E, Emmons KM (2007) Safe To Walk? Neighborhood Safety and Physical Activity Among Public Housing Residents. PLoS Med 4(10): e306. doi:10.1371/journal.pmed.0040306

3.2 Exploration of Variables

The skewness of each input variable is shown below. The two variables with the strongest skew are the proportion of residential land zoned for large lots and the proportion of blacks by town. Respectively the magnitudes of the skewness of these two variables are 2.18 and 2.92. This indicates that the distributions for these two variables are far from symmetrical. The skewness of the dummy variable (whether the suburb borders the river or not) can be neglected because it is a binary variable. All of the other variables skewnesses that are approximately of magnitude 1 or less. This indicates that the distributions for those variables can be considered symmetric even though for three of the variables (concentration of nitrogen oxides, index of accessibility to radial highways, and median value of owner-occupied homes) are multimodal.

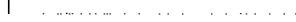

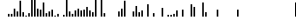

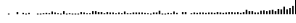





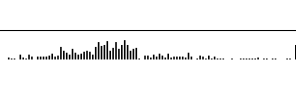
variables	skew
zn	2.1768152
indus	0.2885450
chas	3.3354899
nox	0.7463281
rm	0.4793202
age	0.5777075
dis	0.9988926
rad	1.0102788
tax	0.6593136
ptratio	0.7542681
black	2.9163108
lstat	0.9055864
medv	1.0766920
target	0.0342293

According to the standard deviations of each variable, the variable that has the highest difference from the mean is tax.

variables	sd
zn	23.3646511279634
indus	6.84585491881262
chas	0.256791996193711
nox	0.116666665669521
rm	0.704851288243787
age	28.3213784029166
dis	2.10694955535994
rad	8.68592724130043
tax	167.900088684704
ptratio	2.19684473073614
black	91.3211298387792
lstat	7.10189067779907
medv	9.23968141143397
target	0.500463581298941

Histograms of most of our variables have been plotted below so that distribution can be visualized. We have excluded target and chas due to being binary and not being well represented in the below visualization. We also excluded rad as it is an index variable and also is not best represented in the below visualization.

Table 1 : Descriptive Statistics
11 Variables 466 Observations

zn												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	26	0.61	12	0	0	0	0	16	45	80	
lowest : 0 12 18 18 20, highest: 82 85 90 95 100												
indus												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	73	0.98	11	2	3	5	10	18	20	21	
lowest : 0.5 0.7 1.2 1.2 1.2, highest: 18.1 19.6 21.9 25.6 27.7												
nox												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	79	1	0.6	0.4	0.4	0.4	0.5	0.6	0.7	0.8	
lowest : 0.4 0.4 0.4 0.4 0.4, highest: 0.7 0.7 0.7 0.8 0.9												
rm												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	419	1	6	5	6	6	6	7	7	8	
lowest : 4 4 4 5 5, highest: 8 8 9 9 9												
age												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	333	1	68	18	26	44	77	94	99	100	
lowest : 3 6 6 6 7, highest: 99 99 99 99 100												
dis												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	380	1	4	1	2	2	3	5	7	8	
lowest : 1 1 1 1 1, highest: 9 9 11 11 12												
tax												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	63	0.98	410	222	233	281	334	666	666	666	
lowest : 187 188 193 198 216, highest: 432 437 469 666 711												
ptratio												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	46	0.98	18	15	15	17	19	20	21	21	
lowest : 13 13 14 14 15, highest: 21 21 21 21 22												
black												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	331	0.99	357	88	295	376	391	396	397	397	
lowest : 0.3 2.5 2.6 3.5 3.6 highest: 396.3 396.3 396.3 396.4 396.9												
lstat												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	424	1	13	4	5	7	11	17	23	27	
lowest : 2 2 2 2 3, highest: 34 34 35 37 38												
medv												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
466	0	218	1	23	10	13	17	21	25	35	43	
lowest : 5 6 6 7 7, highest: 46 47 48 49 50												

3.3 Correlation Matrix

We implement a correlation matrix to better understand the correlation between variables in the data set. The below matrix is the results and we noticed a few interesting correlations.

- High nitrogen oxides concentration (parts per 10 million) ("nox") is positively correlated with higher than median crime rates. As defined by the EPA - "NOx pollution is emitted by automobiles, trucks and various non-road vehicles (e.g., construction equipment, boats, etc.) as well as industrial sources such as power plants, industrial boilers, cement kilns, and turbines"³. It is clear to see that nox is concentrated in areas of high road traffic and possible high industrial use which would be neighborhoods of low value and may attract crime.
- The weighted mean of distances to five Boston employment centers is negatively correlated with a city with higher than median crime rate. This is intuitive in that employment centers would be more closely located in cities of high crime due to high unemployment being positively correlated with higher crimes rates⁴.
- The tax is positively correlated with higher than median crime rate which is counter intuitive because we would think as tax increases then crime would decrease (more valuable property = higher tax = less crime).
- We also see bk is negatively correlated with higher than median crime rates but it seems to be due to the transformation of $1000(Bk - 0.63)^2$. Further resources on why this type of transformation is being used were not available. It should be noted that this transformation causes a counter intuitive correlation.

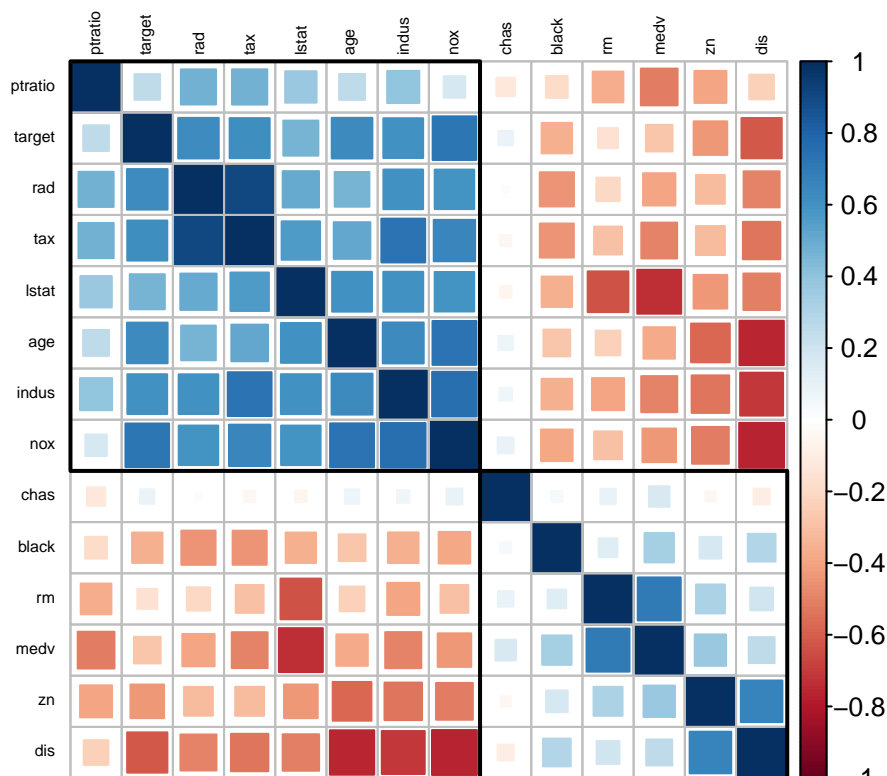


Figure 1: Correlation Plot of Training Data Set

³"Nitrogen Oxides Control Regulations | Ground-level Ozone | New England | US EPA." EPA. Environmental Protection Agency, n.d. Web. 22 Oct. 2016.

⁴Ajimoto, S., Haskins, A., & Wade, Z. (2015). The Effects of Unemployment on Crime Rates in the US.

3.4 Outliers Treatment

We chose winsorizing as the method to address outliers. Instead of trimming values, winsorizing uses the interquantile range to replace values that are above or below the interquantile range multiplied by a factor. Those values above or below the range multiplied by the factor are then replaced with max and min value of the interquantile range. Using the factor 2.2 for winsorizing outliers is a method developed by Hoaglin and Iglewicz and published in the Journal of American Statistical Association in 1987⁵.

The below table is the summary results of the winsorizing of the data.

Table 4:

Statistic	N	Mean	St. Dev.	Min	Max
zn	466	8.739	15.567	0.000	45.000
indus	466	11.105	6.846	0.460	27.740
chas	466	0.071	0.257	0	1
nox	466	0.554	0.117	0.389	0.871
rm	466	6.289	0.686	4.368	8.259
age	466	68.368	28.321	2.900	100.000
dis	466	3.793	2.096	1.130	10.710
rad	466	9.530	8.686	1	24
tax	466	409.502	167.900	187	711
ptratio	466	18.398	2.197	12.600	22.000
black	466	380.268	22.690	331.290	396.900
lstat	466	12.631	7.102	1.730	37.970
medv	466	22.273	8.399	5.000	42.300
target	466	0.491	0.500	0	1

⁵Hoaglin, D. C., and Iglewicz, B. (1987), Fine tuning some resistant rules for outlier labeling, Journal of American Statistical Association, 82, 1147-1149.

3.5 BoxCox Transformations

Using the `BoxCox.lambda` function from the `forecast` package we are able to determine our necessary transformations to our independent variables.

λ	Variables
0.1396180	zn
-0.0877933	indus
0.4722021	chas
-0.9999242	nox
0.0389955	rm
1.9999242	age
-0.6099464	dis
-0.3353947	rad
-0.9999242	tax
1.9999242	ptratio
1.9999242	black
-0.1792021	lstat
0.1044075	medv

Utilizing the below table of common transformations based on the lambda value of the BoxCox we further transform our independent variables.

Common Box-Cox Transformations⁶

λ	Y'
-2	$Y^{-2} = \frac{1}{Y^2}$
-1	$Y^{-1} = \frac{1}{Y}$
-0.5	$Y^{-0.5} = \frac{1}{\sqrt{Y}}$
0	$\log(Y)$
0.5	$Y^{0.5} = \sqrt{Y}$
1	$Y^1 = Y$
2	Y^2

Lambda values that did not fall in the proximity of common transformations were ignored. All other Lambda values were truncated to the nearest tenth that match a common transformation as per the below table.

variable	variable transformation
indus	$\log \text{indus}$
chas	$\sqrt{\text{chas}}$
nox	nox^{-1}
rm	$\log \text{rm}$
age	age^2
dis	$\text{dis}^{-.5}$
tax	tax^{-1}
ptratio	ptratio^2
black	black^2

⁶By Understanding Both the Concept of Transformation and the Box-Cox Method, Practitioners Will Be Better Prepared to Work with Non-normal Data. . "Making Data Normal Using Box-Cox Power Transformation." ISixSigma. N.p., n.d. Web. 29 Oct. 2016.

4 Models Built

4.1 Model 1 - Backwards Selection Method

Call: glm(formula = step(fullModel, direction = "backward", trace = F))

Deviance Residuals: Min 1Q Median 3Q Max
-0.72541 -0.19825 -0.02076 0.14252 0.97959

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.288e+00 2.923e-01 7.827 3.52e-14 ***l(nox^-1) -8.299e-01 9.026e-02 -9.195 < 2e-16*** ***l(age^2)***
2.828e-05 7.524e-06 3.759 0.000193 ***l(dis^-0.5) -7.566e-01 2.100e-01 -3.602 0.000350*** rad 1.063e-02
2.857e-03 3.723 0.000222 ***l(tax^-1) -6.682e+01 2.347e+01 -2.847 0.004608*** ***l(ptratio^2) 3.532e-04 2.310e-***
04 1.529 0.126849
l(black^2) -2.870e-06 9.661e-07 -2.971 0.003123 ***lstat 6.299e-03 3.728e-03 1.690 0.091768***.
medv 1.553e-02 3.087e-03 5.029 7.10e-07 — Signif. codes: 0 "0.001" 0.01 "0.05" 0.1 "1"

(Dispersion parameter for gaussian family taken to be 0.0885353)

Null deviance: 116.466 on 465 degrees of freedom

Residual deviance: 40.372 on 456 degrees of freedom AIC: 204.59

Number of Fisher Scoring iterations: 2

Call: glm(formula = step(fullModel, direction = "backward", trace = F))

Deviance Residuals: Min 1Q Median 3Q Max
-0.55263 -0.20834 -0.04561 0.12313 0.88637

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.4696221 0.3668166 -1.280 0.20110
nox 1.9142262 0.2214486 8.644 < 2e-16 ***age 0.0036605 0.0007597 4.818 1.97e-06*** rad 0.0203953 0.0040123
5.083 5.42e-07 ***tax -0.0003312 0.0002254 -1.470 0.14236***
ptratio 0.0162088 0.0086484 1.874 0.06154.
black -0.0023900 0.0007342 -3.255 0.00122 medv 0.0090505 0.0023199 3.901 0.00011 ** — Signif. codes:
0 "0.001" 0.01 "0.05" 0.1 "1"

(Dispersion parameter for gaussian family taken to be 0.09500619)

Null deviance: 116.466 on 465 degrees of freedom

Residual deviance: 43.513 on 458 degrees of freedom AIC: 235.5

Number of Fisher Scoring iterations: 2

We will use a value of 5 as our threshold for multicollinearity of our variables⁷. Here in our backwards selection model we find that rad exceeds our pre-established threshold.

variables	VIF
nox	3.266936
age	2.265818
rad	5.944666
tax	7.008013
ptratio	1.766752
black	1.358375
medv	1.858471

⁷"Variance Inflation Factor (VIF)." How2stats. N.p., n.d. Web. 27 Oct. 2016.

Table 6:

<i>Dependent variable:</i>	
	fullModel
$l(\text{nox}^{-1})$	−0.830*** (0.090)
$l(\text{age}^2)$	0.00003*** (0.00001)
$l(\text{dis}^{-0.5})$	−0.757*** (0.210)
rad	0.011*** (0.003)
$l(\text{tax}^{-1})$	−66.824*** (23.469)
$l(\text{ptratio}^2)$	0.0004 (0.0002)
$l(\text{black}^2)$	−0.00000*** (0.00000)
lstat	0.006* (0.004)
medv	0.016*** (0.003)
Constant	2.288*** (0.292)
Observations	466
Log Likelihood	−92.296
Akaike Inf. Crit.	204.593
Note: *p<0.1; **p<0.05; ***p<0.01	

4.2 Model 2 - Forwards Selection Method

4.3 Model 3 - Subset Selection Method

Using the `leaps` package and the `regsubsets` function we are able to subset our independent variables by looking at the best model for each predictor. Our final model will use variables as indicated in line 8 of the below table which we will further implement into our subset selection model.

	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1 (1)				*									
2 (1)				*				*					
3 (1)				*		*		*					
4 (1)				*		*		*					*
5 (1)				*		*		*			*		*
6 (1)	*			*		*		*			*		*
7 (1)				*		*		*	*	*	*		*
8 (1)	*			*		*		*	*	*	*		*

Here in our subset selection model we find that no variable exceeds our pre-established threshold of 5 for multicollinearity.

variables	VIF
nox	1.968453
age	1.444190
rad	1.568994
tax	1.976171
ptratio	1.872547
black	1.162274
lstat	2.155295
medv	2.847963

Table 9:

<i>Dependent variable:</i>	
	target
nox	33.570*** (5.186)
age	0.020** (0.010)
rad	0.746*** (0.146)
tax	−0.011*** (0.003)
ptratio	0.410*** (0.118)
black	−0.040*** (0.012)
lstat	0.046 (0.047)
medv	0.108*** (0.042)
Constant	−15.047** (6.129)
Observations	466
Log Likelihood	−97.485
Akaike Inf. Crit.	212.971

Note: *p<0.1; **p<0.05; ***p<0.01

5 Selected Model

6 Appendix A

6.1 Session Info

- R version 3.3.1 (2016-06-21), x86_64-w64-mingw32
- Locale: LC_COLLATE=English_United States.1252, LC_CTYPE=English_United States.1252, LC_MONETARY=English_United States.1252, LC_NUMERIC=C, LC_TIME=English_United States.1252
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, utils
- Other packages: abc 2.1, abc.data 1.0, bibtex 0.4.0, car 2.1-3, corrplot 0.77, data.table 1.9.6, doParallel 1.0.10, dplyr 0.5.0, e1071 1.6-7, foreach 1.4.3, forecast 7.3, Formula 1.2-1, ggplot2 2.1.0, glmulti 1.0.7, highlight 0.4.7, Hmisc 3.17-4, iterators 1.0.8, itertools 0.1-3, knitr 1.14, lattice 0.20-34, leaps 2.9, locfit 1.5-9.1, magrittr 1.5, MASS 7.3-45, matrixStats 0.51.0, missForest 1.4, nnet 7.3-12, pacman 0.4.1, purrr 0.2.2, quantreg 5.29, randomForest 4.6-12, readr 1.0.0, rJava 0.9-8, scales 0.4.0, SparseM 1.72, stargazer 5.2, stringr 1.1.0, survival 2.39-5, tibble 1.2, tidyr 0.6.0, tidyverse 1.0.0, timeDate 3012.100, xlsx 0.5.7, xlsxjars 0.6.1, xtable 1.8-2, zoo 1.7-13
- Loaded via a namespace (and not attached): acepack 1.4.1, assertthat 0.1, bitops 1.0-6, chron 2.3-47, class 7.3-14, cluster 2.0.5, codetools 0.2-15, colorspace 1.2-7, DBI 0.5-1, digest 0.6.10, evaluate 0.10, foreign 0.8-67, formatR 1.4, fracdiff 1.4-2, grid 3.3.1, gridExtra 2.2.1, gtable 0.2.0, highr 0.6, htmltools 0.3.5, http 1.2.1, latticeExtra 0.6-28, lazyeval 0.2.0, lme4 1.1-12, lubridate 1.6.0, Matrix 1.2-7.1, MatrixModels 0.4-1, mgcv 1.8-15, minqa 1.2.4, munsell 0.4.3, nlme 3.1-128, nloptr 1.0.4, pbkrtest 0.4-6, plyr 1.8.4, quadprog 1.5-5, R6 2.2.0, RColorBrewer 1.1-2, Rcpp 0.12.7, RCurl 1.95-4.8, RefManager 0.11.0, RJSONIO 1.3-0, rmarkdown 1.1, rpart 4.1-10, splines 3.3.1, stringi 1.1.2, tools 3.3.1, tseries 0.10-35, XML 3.98-1.4, yaml 2.1.13

6.2 Data Dictionary

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

6.3 R source code

Please see Homework 3.rmd on GitHub for source code.

<https://github.com/ChristopheHunt/DATA-621-Group-1/blob/master/Homework%203/Homework%203.Rmd>