BOSTON : REGRESSION ANALYSIS

STAT 350 PROJECT REPORT

Decemeber 3, 2018

Alan Bi (301253272) alanb@sfu.ca
Kevin Chen (301251095) cca220@sfu.ca

Contents

1 Introduction

The purpose of this project is to predict the level of air pollution in the Boston Metropolitan area. The 'Boston' dataset used in this study (Harrison and Rubinfeld, 1978) is available in R through the "MASS" package. This project compares a naive multiple linear regression approach to machine learning techniques such as lasso and ridge regression, and random forest regression.

Data

The Boston dataset is a compilation of data across multiple sources ranging from the U.S. Census Bureau to the FBI and Vogt, Ivers, and Associates. The dataset includes observations from 506 different census tracts in the year 1970. From Table 1, it can be noted that the dataset contains 14 variables categorized under characterisics: Structural, Neighbourhood, Accessibility, and Air Pollution.

Table 1: Boston Dataset Variables

Variable Type	**Variable**	**Description**
Air Pollution	NOX	Nitric Oxide Concentration (parts per 10 million)
Neighbourhood	MEDV	Median value of Owner-Occupied Homes (in $1000s)
	B	The Proportion of African Americans in Town
	LSTAT	Proportion of Population that is lower status
	CRIM	Per Capita Crime Rate by Town
	ZN	Proportion of Residential Land Zoned for lots over 25000sqft
	INDUS	Proportion of Non-Retail Business Acres Per Town
	TAX	Full Value Property Tax Rate (in $10000)
	PTRATIO	Pupil-Teacher Ratio in Town
	CHAS	Charles River Dummy Variable (1 if census tract bounds river, 0 otherwise)
Accessibility	DIS	Weighted Distances to Five Boston Employment
	RAD	Index of accessibility to Radial Highways
Structural	RM	Average Number of Rooms Per Dwelling
	AGE	Proportion of Owner-Occupied Units Built prior to 1940's

After some exploratory data analysis, we discover that all variables except CHAS and RAD are numerical variables. RAD is a categorical variable indicating the accessibility to radial highways on a log scale, and CHAS is a binary dummy variable indicating a census tract being near the Charles River. If we were looking to predict housing values, CHAS would seem like possible candidate during variable selection, as homes near the waterfront on average have higher values. However, CHAS and NOX have a near zero correlation. We find that INDUS and DIS have the highest correlation with NOX at 0.76 and -0.77, respectively.

Goal

The Boston dataset contains two potential response variable, housing price (MEDV) and the level of nitric oxide (NOX); where NOX is used as a proxy for the level of air pollution. Analysis of the Boston dataset is almost exclusively focused on predicting housing price using the other 13 housing market variables, however, our goal is to predict the level of air quality. And as indicated by Harrison and Rubinfeld (1978), NOX is a sufficient estimator of overall air quality. We believe that our goal in predicting the level of air quality in neighbourhoods is valid and relevant because cases of people with respiratory sensitivities (E.g. young children, the elderly, etc.) is on the rise along with air pollution and neighbourhood air quality may be a factor influencing some home buyers.

2 Analysis

I. Multiple Linear Regression

Our initial regression model using multiple linear regression includes all possible regressors in the Boston dataset.

$$NOX = \beta_0 + \beta_1 CRIM + \beta_2 ZN + \beta_3 INDUS + \beta_4 CHAS + \beta_5 RM + \beta_6 AGE + \beta_7 DIS + \beta_8 RAD + \beta_9 TAX + \beta_{10} PTRATIO + \beta_{11} BLACK + \beta_{12} LSTAT + \beta_{13} MEDV$$

By analyzing the residual plots of residuals against individual regressors, we found it necessary to apply logarithm transformation to CRIM, DIS, and LSTAT in order for the constant variance of residuals assumption to be satisfied. The improvement in constant variance can be observed in Figure 1. In addition, by examining VIF, we find that TAX (9.2) is highly correlated with other variables. Therefore, to prevent issues arising from multicollinearity, we decide to remove the TAX variable.

To check the normality and constant variance assumptions for the model, it can be seen in Figure 2 that both assumptions are violated. However, the appropriate Box-Cox transformation on our response variable NOX, ($NOX' = \frac{1}{NOX}$) solves the normality and constant variance problems in the original model.

The regression model after the transformations is:

$$NOX^{-1} = \beta_0 + \beta_1 log(CRIM) + \beta_2 ZN + \beta_3 INDUS + \beta_4 CHAS + \beta_5 RM + \beta_6 AGE + \beta_7 log(DIS) + \beta_8 RAD + \beta_9 PTRATION + \beta_{10} BLACK + \beta_{11} log(LSTAT) + \beta_{12} MEDV$$

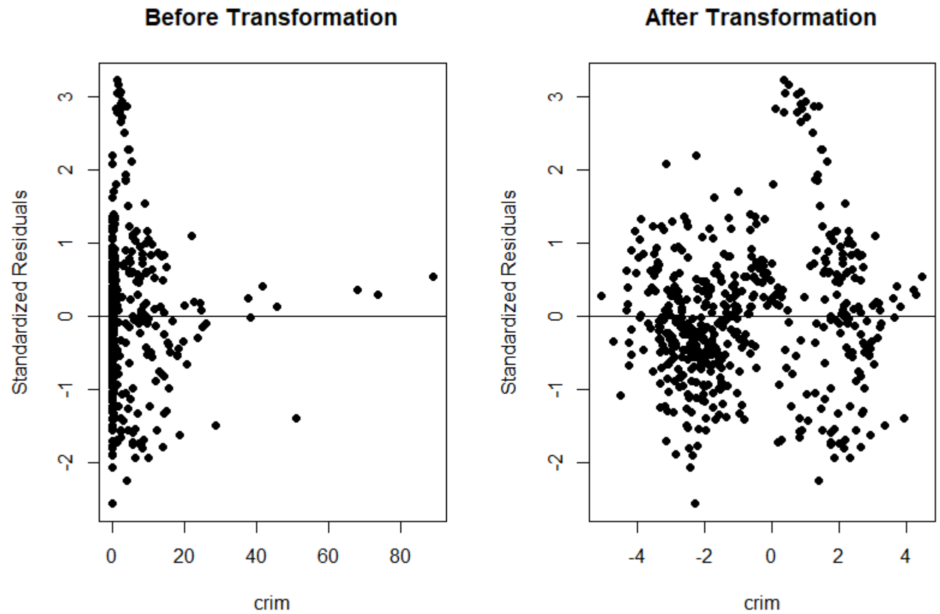

Figure 1: Residuals vs CRIM. Left plot indicates a violation of the constant variance assumption with the CRIM variable. After a log-transformation, the assumption is no longer violated, as shown on the right.

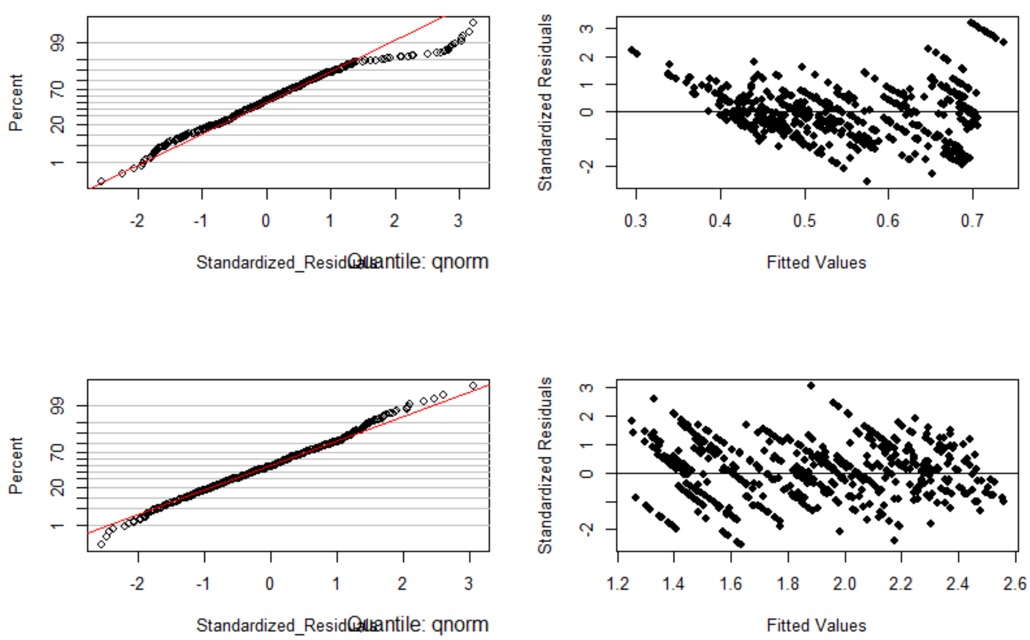

Figure 2: The plots above show that both normality assumption and constant variance assumption are violated before Box Cox transformation. The transformation improves the situation, as shown on the bottom plots.

The next step is to improve the model with variable selection techniques. The first technique we apply is to look at the comparison metrics RSS, $R^2_{adjusted}$, CP, and BIC. However, all four statistics suggest different "best model". Thus we cannot choose a single optimal model using this method. Since the previous technique failed, we will then continue the selection by using forward, backward and stepwise selection.

II. Forward, Backward and Stepwise Selections

The results seen in Figure 3 and Figure 4 suggest that the optimal model after using forward, backward, and stepwise selection is the model containing the 10 variables: CRIM, ZN, INDUS, RM, AGE, DIS, RAD2, RAD5, PTRATIO, and MEDV. However, since RAD is a categorical variable, we cannot include only a portion of it into our model. Hence we exclude the predictor RAD entirely. This model las the lowest RMSE among all of the models when doing 10-fold cross-validation using identical folds. Checking the significance of the model, we get a global F-statistic = 381.8 and p-value < 2.2e-16; indicating that our model is significant. The selected variables are also significant.

nvmax	RMSE
1	0.175826
2	0.15275
3	0.145732
4	0.142707
5	0.135995
6	0.134408
7	0.134542
8	0.134782
9	0.135522
10	0.134392
11	0.134409
12	0.134402
13	0.134605
14	0.134745
15	0.134722
16	0.134611
17	0.134584
18	0.134624
19	0.134486

nvmax	RMSE
1	0.175826
2	0.15275
3	0.145732
4	0.142384
5	0.136644
6	0.135025
7	0.135732
8	0.135573
9	0.13517
10	0.133548
11	0.134397
12	0.133785
13	0.134832
14	0.135006
15	0.134984
16	0.134768
17	0.134709
18	0.134585
19	0.134486

nvmax	RMSE
1	0.175826
2	0.197378
3	0.145732
4	0.142707
5	0.135995
6	0.134408
7	0.134542
8	0.135207
9	0.138174
10	0.134727
11	0.134358
12	0.133585
13	0.134823
14	0.134695
15	0.134262
16	0.13515
17	0.135406
18	0.135994
19	0.134486

Figure 3: From left to right, RMSEs for Forward Selection, Backward Elimination and Stepwise Selection are shown, respectively. We will choose the result from backward selection because it gives the smallest value of RMSE.

	crim	zn	indus	chas	rm	age	dis	rad2	rad3	rad4	rad5	rad6	rad7	rad8	rad24	ptratio	black	lstat	medv
1 (1)							*												
2 (1)	*						*												
3 (1)	*					*	*												
4 (1)	*					*	*				*								
5 (1)	*		*			*	*				*								
6 (1)	*		*			*	*				*								*
7 (1)	*		*			*	*				*					*			*
8 (1)	*	*	*			*	*				*					*			*
9 (1)	*	*	*			*	*	*			*					*			*
10 (1)	*	*	*		*	*	*	*			*					*			*

Figure 4: Variable selection with backward elimination. The above plot shows the optimal model for each number of variables, with 10 variables being the most prominent choice.

Our model after variable selection is:

$NOX^{-1} = \beta_0 + \beta_1 log(CRIM) + \beta_2 ZN + \beta_3 INDUS + \beta_4 RM + \beta_5 AGE + \beta_6 log(DIS) + \beta_7 PTRATIO + \beta_8 MEDV$

III. Ridge Regression and LASSO

Next, we apply the methods of Ridge regression and LASSO, and compare the results using RMSE obtained through 10-fold cross-validation.

Linear Regression

	coefficient
crim	-0.047421613
zn	-0.062071055
indus	-0.007859141
chas	-0.046380322
rm	-0.024344124
age	-0.002404992
dis	0.229225488
rad2	0.101593958
rad3	0.05952063
rad4	0.028384064
rad5	-0.075533846
rad6	-0.014832132
rad7	0.003681521
rad8	0.041373445
rad24	-0.007981142
ptratio	0.012036815
black	-9.65E-05
lstat	0.01431555
medv	0.005139498

Ridge Regression

	coefficients
crim	-0.047092588
zn	-0.059348826
indus	-0.007855963
chas	-0.043638073
rm	-0.022165455
age	-0.002346473
dis	0.228469039
rad2	0.094610155
rad3	0.055775782
rad4	0.024649908
rad5	-0.078947462
rad6	-0.014842368
rad7	0
rad8	0.033453707
rad24	-0.008455574
ptratio	0.010956661
black	-7.46E-05
lstat	0.003474101
medv	0.004406017

LASSO Regression

	coefficients
crim	-0.047148882
zn	-0.053187343
indus	-0.007856704
chas	-0.039512317
rm	-0.016500109
age	-0.002329125
dis	0.228924491
rad2	0.085769228
rad3	0.050857397
rad4	0.022114116
rad5	-0.080532301
rad6	-0.012468929
rad7	0
rad8	0.024886918
rad24	-0.0043713
ptratio	0.009247221
black	-3.91E-05
lstat	0
medv	0.003737041

Figure 5: Coefficient estimations for Linear Regression, Ridge regression and LASSO regression.

Ridge regression has shrunk one variable, RAD, close to zero. However, due to the nature of Ridge regression, the coefficients cannot be shrunk to zero. Thus we will not exclude that variable. LASSO Regression has shrunk two variables, RAD and LSTAT, close to zero. Unlike Ridge regression, the coefficients that are reduced to zero by LASSO will be taken out. For LASSO, choosing between the two lambdas, lambda.1se and lambda.min, was not difficult as they both resulted in the same number of variables and thus only differed by RMSE; we chose to use lambda.min in order to minimize the RMSE. LASSO is a technique that can be used in the presence of high degrees of multicollinearity. However, when used with the Boston data, only two variables are removed as there is not a lot of multicollinearity between variables.

Through cross-validation and after the removal of unwanted variables , we obtain the prediction RMSE of 0.009479611 for Multiple Linear regression, 0.009477958 for Ridge regression, and 0.008178095 for LASSO. The prediction RMSE for LASSO is lower than both Ridge regression and Multiple Linear regression.

The model that LASSO suggests is:

$NOX^{-1} = \beta_0 + \beta_1 log(CRIM) + \beta_2 ZN + \beta_3 INDUS + \beta_4 CHAS + \beta_5 RM + \beta_6 AGE + \beta_7 log(DIS) + \beta_8 PTRATION + \beta_9 BLACK + \beta_{10} MEDV$

IV. Random Forest

The last technique we used was random forest. One helpful function of random forest is its ability to determine variable importance. From Figure 6, it can be noted that DIS and INDUS have the largest effect on MSE, which also happen to be the two variables mentioned in our exploratory data analysis to be the most correlated with NOX.

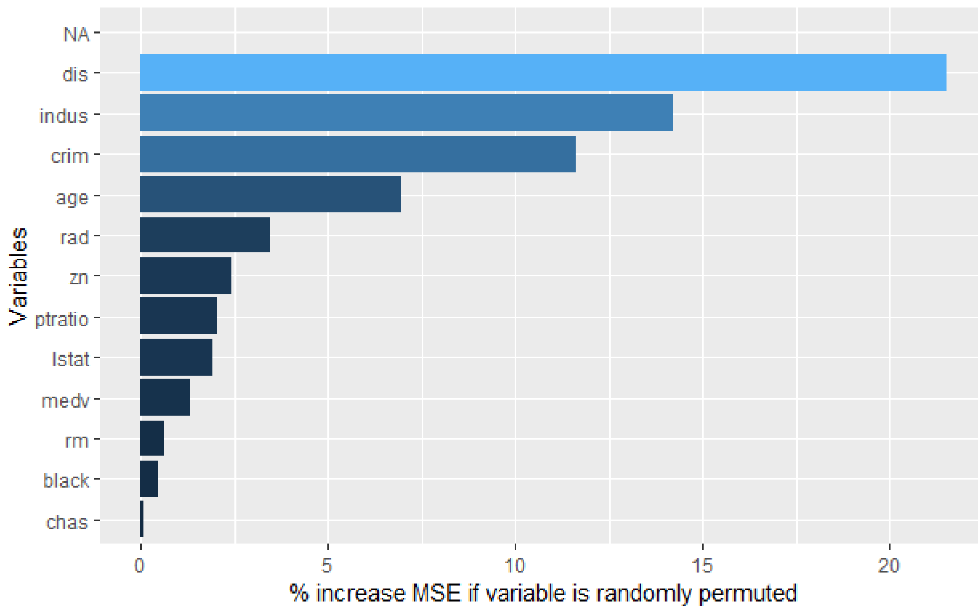

Figure 6: Feature significance according to random forest

Random forest is an ensembling technique similar to bagging. Due to the nature of random forest, we do not end up with a model the same way as with LASSO, Multiple Linear and Ridge regression. Some advantages to this non-linear modeling approach include: preventing overfitting since the model generalizes with experience from multiple samples of data, decorrelating trees since it selects random subsets of variables at each node split, and reducing variance by averaging outputs. However, random forest is slow for large datasets with lots of variables and it is a predictive modeling tool but not descriptive, so we do not know the coefficients like in the previously mentioned techniques. Though random forest provides us with a ranked list of variable importance in relation to MSE. When we consider all variables as inputs, 500 bootstrapped samples, and a subset of four randomly selected variables at each decision node, random forest obtains an RMSE of 0.007153467.

3 Results

A summary of the coefficient estimates for linear and ridge regression and LASSO are shown in Figure 5 located section 2.III. One should note that the response variable used is $NOX' = \frac{1}{NOX}$. As an example, when looking at the variable DIS, a one percent increase in DIS leads to roughly a 0.23 increase in $\frac{1}{NOX}$ on average, which is actually a decrease in the NOX. This result intuitively translates to neighbourhoods that are farther away from a Boston employment center have lower levels of air pollution. As employment centers are generally located within the city, this result may indicate that areas outside of the dense city may have better air quality. However, because we use the log distance, the positive effect on air quality diminishes quickly the further a neighbourhood is from an employment centers. Conversely, a variable with a negative coefficient estimate like that of CHAS on average results in an increase between 0.052 to 0.062 in NOX. Recall that CHAS is a dummy variable to indicate whether a census tract is bounded by the Charles river and contain waterfront properties. The Charles River borders downtown Boston and since Boston is a major port city, there is both higher density in the urban areas along the Charles River as well commercial industry shipping. This may be the reason our models predict that census tracts bound by the Charles River have higher levels of air pollution on average.

There is a summary of the models used in this project found in Table 2, including results from our initial naive approach to our use of Ridge regression, LASSO, and random forest. We are interested in the predictive powers of our model and as such we compare each of the model's root mean squared error (RMSE), where a lower RMSE indicates better predictive power. Therefore, random forest has the greatest predictive power, followed by LASSO, ridge regression, our backward selection model, and finally our original naive model. This is somewhat expected since we are trading off interpretability for predictive power when using random forest. However, LASSO if able to provide coefficient estimates for linear model interpretation.

4 Conclusions

In conclusion, out of the models we used, random forest model has the best predictive power and is able to provide insight into variable importance, which is somewhat consistent with the results of variable selection techniques like forward, backwards, and stepwise selection that look at improved AIC. DIS, INDUS, and CRIM have been consistently important for each of our models. However, one downside to random forest variable importance calculation is that has trouble indicating which variables would result in issues involving multicollinearity as RAD is ranked as fifth most important in regards to MSE. From our results, it seems that good predictors of NOX, or the general overall level of air pollution, are related to being further out from the city. On average, we can expect lower levels of air pollution the further we are from Boston's city centers. However, air pollution is higher in sparse residential areas with larger properties, areas with higher crime rates, and older neighbourhoods. Downtown cores are generally one of the older neighbourhoods in the metropolitan area and suffer from higher crime rates, as is the case in Boston, so it is not surprising that higher levels of both of those predictors suggest higher levels of air pollution. However, something that was

Table 2: Summary of models and comparing their RMSE

Model	**RMSE** (Cross-validated)	**Description**
Naive	0.0554	Multiple Linear Regression using all the variables No transformations Contains multicollinearity OLS assumptions violated
Variable Selection	0.009480	10 variable model Transformed variables Contains no multicollinearity OLS assumptions satisfied
Ridge Regression	0.009478	RAD is removed Transformed variables Contains no multicollinearity
LASSO	0.008178	Lambda.min is used RAD and LSTAT are removed Transformed variables Contains no multicollinearity
Random Forest	0.006951	500 trees generated Subset of 4 variables randomly chosen at each node no estimations for coefficients for a linear model

unexpected is that neighbourhoods with larger zoned residential properties seems to also have higher levels of air pollution, and we are uncertain as to why that seems to be the case. This may be a resulting outcome due to the nature of the data, since the data deals mostly with homeowners and somewhat ignores renters. In summary, for anyone that finds air quality important when searching for a home, the results from this project suggest that you purchase a home outside of downtown Boston in a neighbourhood with moderate sized lots, in which most suburbs seem to satisfy the criteria.

References

[1] Friedman, J., Hastie, T. Tibshirani, R (2010). glmnet: Lasso and Elastic-Net Regulaized Generalized Linear Models. R package version 2.0-16. https://cran.r-project.org/web/packages/glmnet/

[2] Harrison, D. Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management, 5(1)*, 81-102.

[3] Liaw, A. Wiener, M. (2002) Classification and Regression by randomForest. R package version 7.3-51.1. https://cran.r-project.org/web/packages/randomForest/

[4] Kuhn, M. (2008). caret: Classification and Regression Training. R package version 6.0-81. https://cran.r-project.org/web/packages/caret/

[5] Venables, W. N. Ripley, B. D. (2002) Modern Applied Statistics with S. R package version 7.3-51.1. https://cran.r-project.org/web/packages/MASS/