Statistical Learning Project

Ammar Hasan

11 November 2018

Contents

1 Introduction

This report summaries the steps undertaken to produce and evaluate linear regression models of the value of housing in Boston Standard Metropolitan to predict the value of logarithmic crime rate (lcrim) using other variables. The model would be built after exploratory and unsupervised statistical analysis of the data which is carried first to gain an understanding on the data characteristics and structure before hand. The models would be build using both sub-setting (best fit and step wise) and regularisation methods (LASSO and Ridge Fit) methods, these methods will be compared using k fold cross validation.

Any output (tables and plots) is placed in the Appendix at the end of the report with the R code that generated it, the description and analysis of the methods and their output is found in the document body which cross references the appendix content. Values in tables and other numerical results are corrected to 3 decimal places unless stated otherwise.

2 Exploratory Data Analysis

Before building a linear model, it is wise to first understand the overall structure of the data itself to get a feeling for the data characteristics and how the variables relate to one another - especially how they relate to the response variable (lcrime).

2.1 Data Spread and Location

To analyse the distribution of the data, the quantiles and means will be examined. In particular, this part of the report will examine outliers, scale, consistency and certainty.

2.1.1 Mean Vector

Using the col means functions a vector of mean averages can be produced for all the predictors and the response variables in the Boston data-set.

The returned table (transposed) is in table 1 in Appendix A.2, and shows that the means are well spread out from one another, which suggests a difference in the nature of the measurements.

Examining the structure of the variables in the Boston data set using the ?Boston confirms that the nature of the measurements vary from Full-value property-tax rate per $10,000 for tax to Nitrogen oxides parts per 10 million for nox for instance, which obviously suggests that the measurements cannot be directly compared scale to scale (standarisation might be required).

2.1.2 Box Plot and Quantiles

As previously stated in the previous section, the variables are of different natures and scales, hence any scale to scale comparsion needs standarisation. To standardise the data a scale transform was applied using the base R scale() function, and using the boxplot base function the plot in figure 1 is generated.

The following stands out of the plot:

- black, rm, zn and medv predictor variables have significantly more outliers than the other variables.And hence more uncertain and also their averages can get skewed.
- chas predictor variable semms to have a very tight ranges that are practicaly identical. This is because this is a binary variable.

- zn, rad and tax predictor variables have a short Q1 to Q2 range compared to the Q2 to Q3 range, suggesting that the lower values of the data are very tightly clustered. black has the opposite problem.
- rm, lstat, mdev and ptratio have long minimum and maximum ranges in comparision to their IQR ranges, and hence more exterme values. This means that rheir averages can get skewed.

2.2 Data Relationships

2.2.1 Pairs Plot and Correlation Matricies

3 Unsupervised Learning

3.1 Principle Component Analysis

3.2 Cluster Analysis

4 Supervised Learning (Linear Model)

5 Appendix

This section contains all supplementary material and is divided into three sections (Tables, Plots and Abbreviations). The code required to generate the supplementary material is also included

5.1 A.1 Abbrevations

Table 1:

	x
lcrim	-0.780
zn	11.364
indus	11.137
chas	0.069
nox	0.555
rm	6.285
age	31.425
disf	1.960
rad	9.549
tax	408.237
ptratio	18.456
black	356.674
lstat	12.653
medv	22.533

5.2 A.2 Tables

5.2.1 Code to Generate Table 1 (Transposed and Correct to 3DP)

```
table(colMeans(Boston), '')
```

Table 2: Correlation Matrix (3DP)

	lcrim	zn	indus	chas	nox	rm	age	disf	rad	tax	ptratio	black	lstat	medv
lcrim	4.674	-26.074	10.840	0.016	0.198	-0.466	-40.063	-1.348	16.066	301.796	1.823	-94.499	9.675	-9.034
zn	-26.074	543.937	-85.413	-0.253	-1.396	5.113	373.902	13.033	-63.349	-1236.454	-19.777	373.721	-68.783	77.315
indus	10.840	-85.413	47.064	0.110	0.607	-1.888	-124.514	-4.555	35.550	833.360	5.692	-223.580	29.580	-30.521
chas	0.016	-0.253	0.110	0.065	0.003	0.016	-0.619	-0.019	-0.016	-1.523	-0.067	1.131	-0.098	0.409
nox	0.198	-1.396	0.607	0.003	0.013	-0.025	-2.386	-0.082	0.617	13.046	0.047	-4.021	0.489	-0.455
rm	-0.466	5.113	-1.888	0.016	-0.025	0.494	4.752	0.137	-1.284	-34.583	-0.541	8.215	-3.080	4.493
age	-40.063	373.902	-124.514	-0.619	-2.386	4.752	792.358	19.487	-111.771	-2402.690	-15.937	702.940	-121.078	97.589
disf	-1.348	13.033	-4.555	-0.019	-0.082	0.137	19.487	0.834	-3.794	-82.595	-0.462	26.819	-3.331	2.442
rad	16.066	-63.349	35.550	-0.016	0.617	-1.284	-111.771	-3.794	75.816	1335.757	8.761	-353.276	30.385	-30.561
tax	301.796	-1236.454	833.360	-1.523	13.046	-34.583	-2402.690	-82.595	1335.757	28404.759	168.153	-6797.911	654.715	-726.256
ptratio	1.823	-19.777	5.692	-0.067	0.047	-0.541	-15.937	-0.462	8.761	168.153	4.687	-35.060	5.783	-10.111
black	-94.499	373.721	-223.580	1.131	-4.021	8.215	702.940	26.819	-353.276	-6797.911	-35.060	8334.752	-238.668	279.990
lstat	9.675	-68.783	29.580	-0.098	0.489	-3.080	-121.078	-3.331	30.385	654.715	5.783	-238.668	50.995	-48.448
medv	-9.034	77.315	-30.521	0.409	-0.455	4.493	97.589	2.442	-30.561	-726.256	-10.111	279.990	-48.448	84.587

5.2.2 Code to Generate Table 2 (Correct to 3DP)

```
table(var(Boston), 'Correlation Matrix (3DP)')
```

Table 3: PCA Summary (Contribution to Variation)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Standard deviation	2.645	1.286	1.118	0.934	0.927	0.809	0.638	0.588	0.500	0.460	0.434	0.389	0.322	0.233
Proportion of Variance	0.500	0.118	0.089	0.062	0.061	0.047	0.029	0.025	0.018	0.015	0.013	0.011	0.007	0.004
Cumulative Proportion	0.500	0.618	0.707	0.769	0.831	0.878	0.907	0.931	0.949	0.964	0.978	0.989	0.996	1.000

5.2.3 Code to Generate Table 3 (Correct to 3DP)

```
# Perform PCA based on the standardised data (means and data nature vary)
pca = prcomp(Boston, scale=TRUE)
table(summary(pca)$importance, 'PCA Summary (Contribution to Variation)')
```

5.2.4 Code to Generate Table 4 (Correct to 3DP)

```
# List PC 1, 2 and 3
table(pca$rotation[,1:3], 'PCA Components')
```

Table 4: PCA Components

	PC1	PC2	PC3
lcrim	0.341	-0.136	0.181
zn	-0.239	0.058	0.394
indus	0.324	-0.095	-0.070
chas	-0.001	-0.387	-0.255
nox	0.320	-0.227	-0.087
rm	-0.190	-0.492	0.285
age	-0.291	0.208	0.264
disf	-0.294	0.287	0.220
rad	0.295	-0.078	0.450
tax	0.315	-0.043	0.381
ptratio	0.196	0.331	0.116
black	-0.190	0.019	-0.378
lstat	0.297	0.238	-0.150
medv	-0.251	-0.475	0.095

5.3 A.3 Plots

5.3.1 Code to Generate Figure 1

```
# scale transforms to deal with the variation in the nature of the measurements
boxplot(scale(Boston), cex.axis=0.6)
```

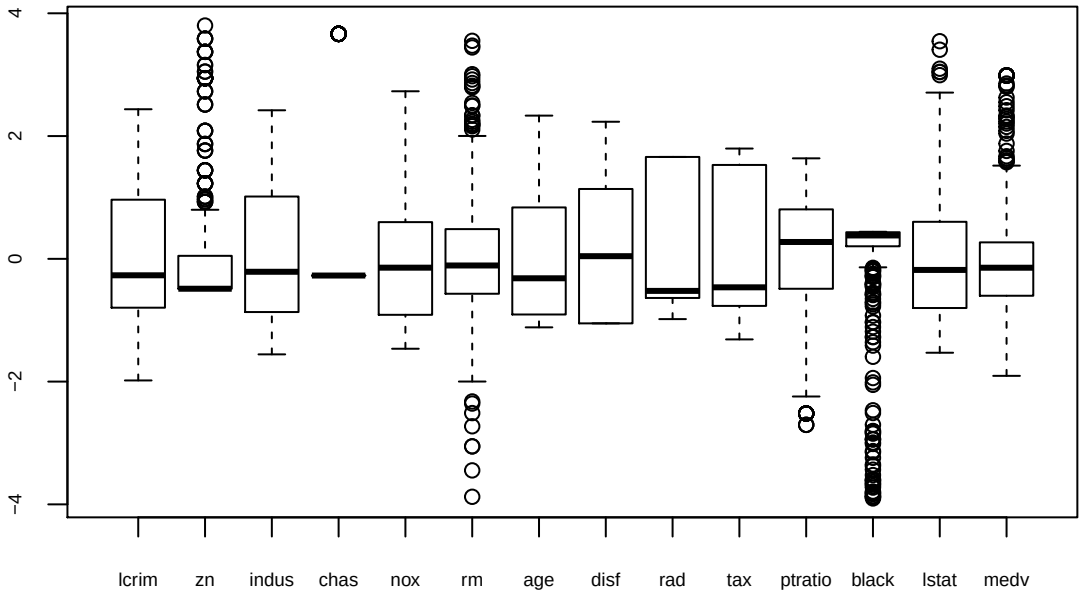

Figure 1: Box Plot

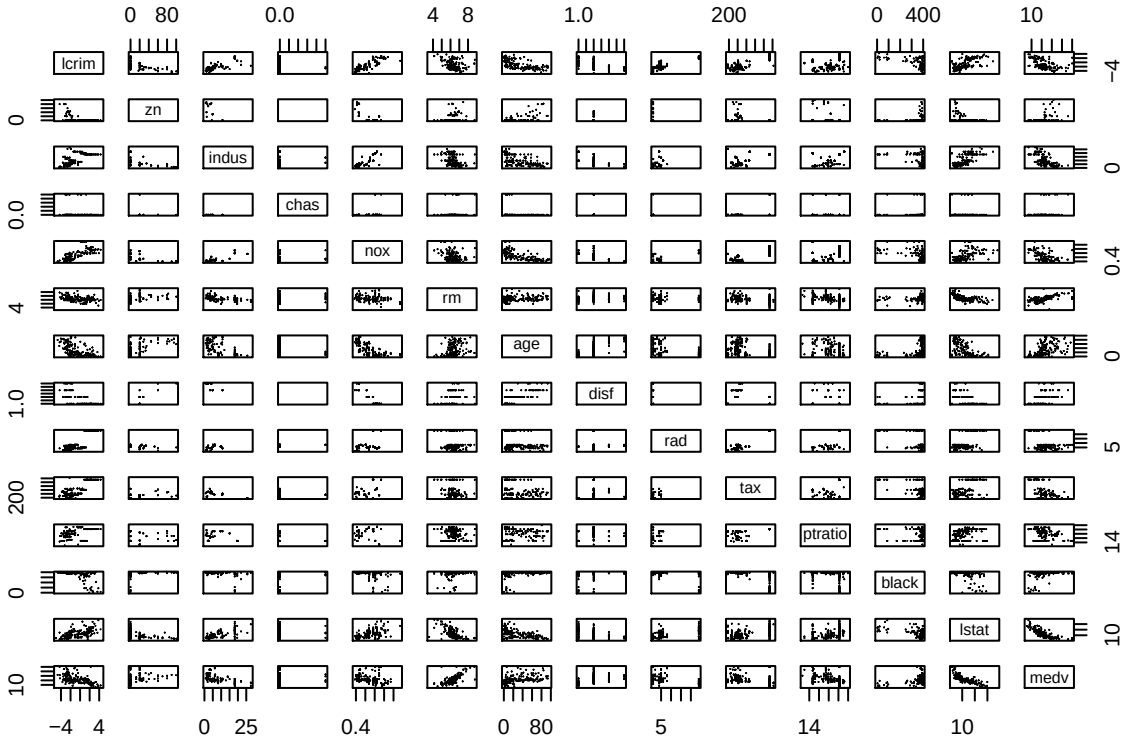

Figure 2: Pairs Plot

5.3.2 Code to Generate Figure 2

```
pairs(Boston, cex=0.0005)
```

5.3.3 Code to Generate Figure 3

```
plot(pca, type='l', main='Scree Plot for Boston Housing Values')
title(xlab='Principle Component number')
```

5.3.4 Code to Generate Figure 4

```
# Plot PCA 1 against PCA 2
plot(pca$x[,1], pca$x[,2], main ="Principle Component 1 vs 2 for Boston Housing Values",
     xlab="Component 1", ylab="Component 2")
```

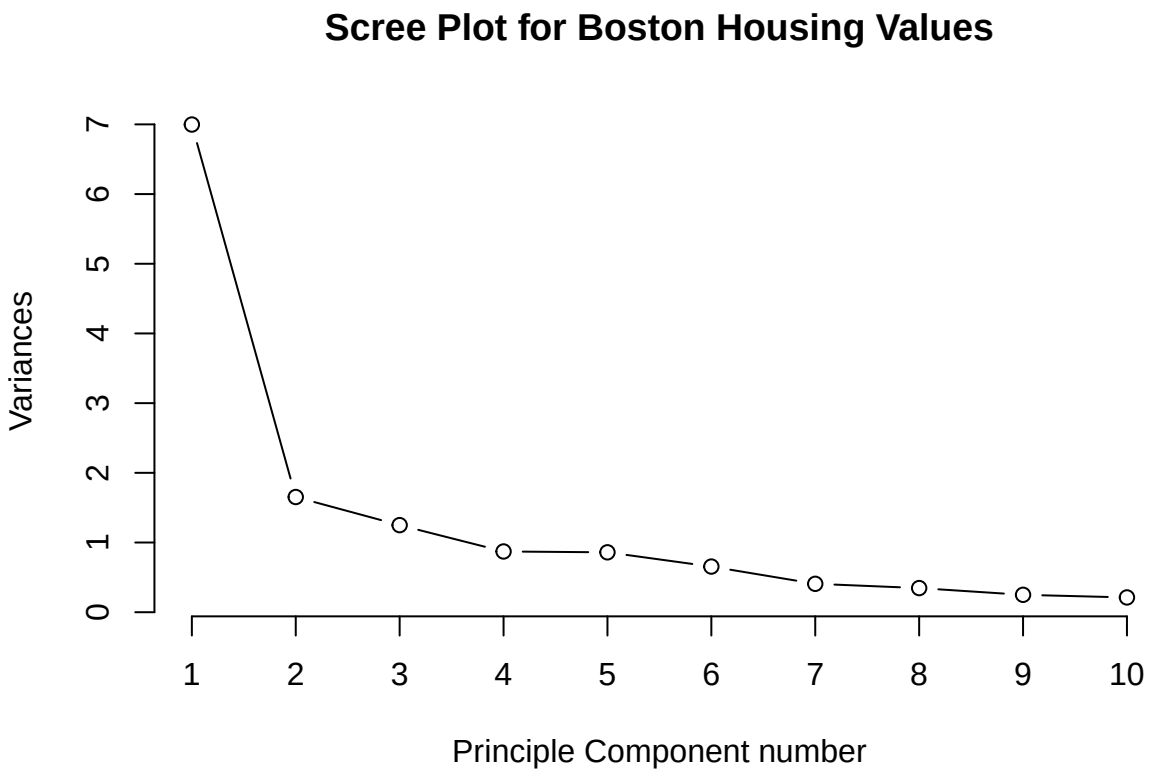

Figure 3: Scree Plot

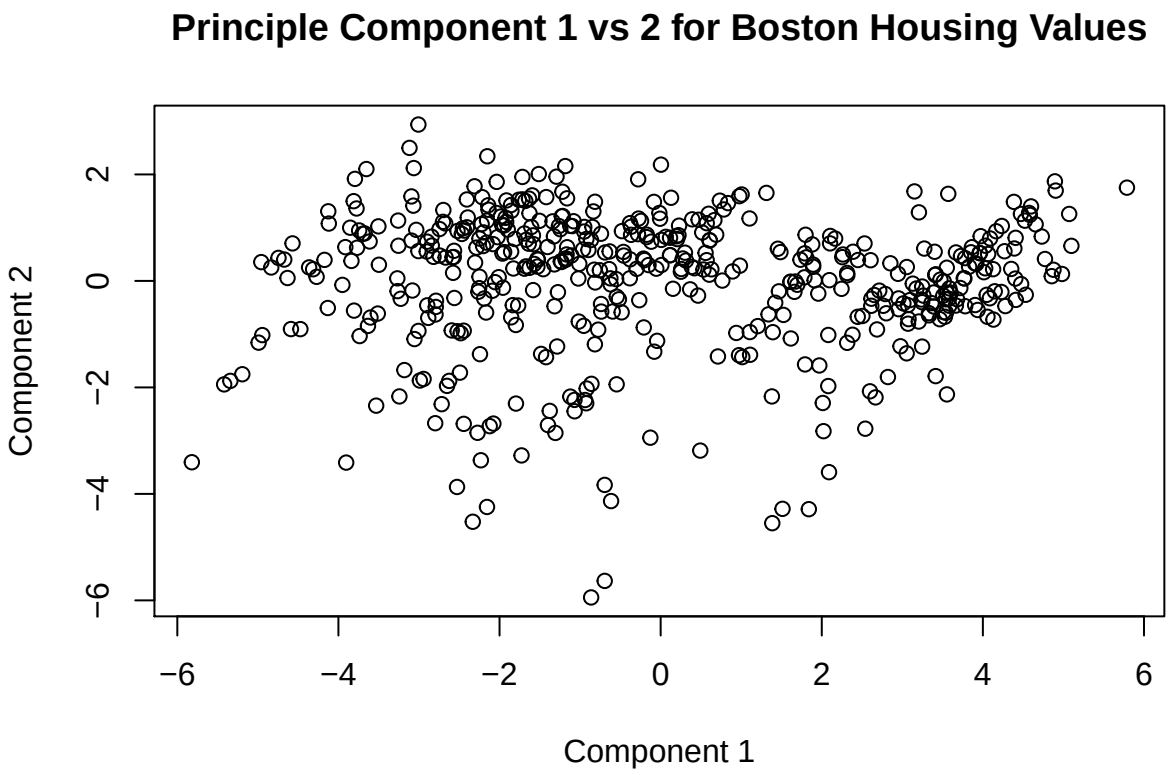

Figure 4: PCA Plot