# Contents

# 1 Introduction

Supervised learning is a branch of machine learning, which in turn is in the space of Artificial Intelligence (AI) [1]. Supervised learning involves making a prediction of a variable of interest using a set of features. The prediction and feature sets are linked by some mathematical function. Linear regression is a supervised learning technique, and vastly used. The Boston dataset (available in the MASS package in R) [1].

The dataset was collected by the U.S. Census Service concerning housing in the area of Boston, MA. Using linear regression, the aim of this project is to predict the Median value of owner-occupied homes using various features.

This report will first investigate the data by briefly describing the data, mentioning techniques used to process the data, and use exploratory data analysis (EDA). Next, this report will implement a simple multiple linear regression, regressing the median value of households against all the independent variables. Next, various techniques will be used to improve on the initial model; where improvements will be assess by the Mean Squared Error (MSE) of the test set. Finally, this report will conduct residual diagnostics of the simple MLR model, and the best model.

# 2 Data

## 2.1 Data Description

The Boston dataset contains information collected by the US census serving concerning housing in the area of Boston, MA. The dataset consists of 13 variables, where the median value of households is the target variable. Out of the 12 feature variables, two are categorical whilst the rest are classed as continuous. Table 5 displays the variables, and their description. Given that the dataset has varying units of measurements, this alludes to the fact that standardisation will be applied, so that the learning algorithm is not biased to some variables than others.

## 2.2 Data Pre-processing

With this dataset, not much preprocessing was conducted before exploring the data. The only procedure that was conducted was to manipulate the chas and rad variables into factor variables in R, as they are registered as numerical variables.

## 2.3 Exploratory Data Analysis

### 2.3.1 Distribution of the MEDV variable

Figure 1 is a visual representation of the distribution of the *medv* variable. It is quite evident that the distribution is positively skewed. Moreover, the data does exhibit outliers, and this is evident from the box plot in figure 1. From this figure, it is clear that the distribution is not normally distributed, but standardisation will be

applied as the algorithm used (i.e. linear regression) assumes the data is normally distributed.

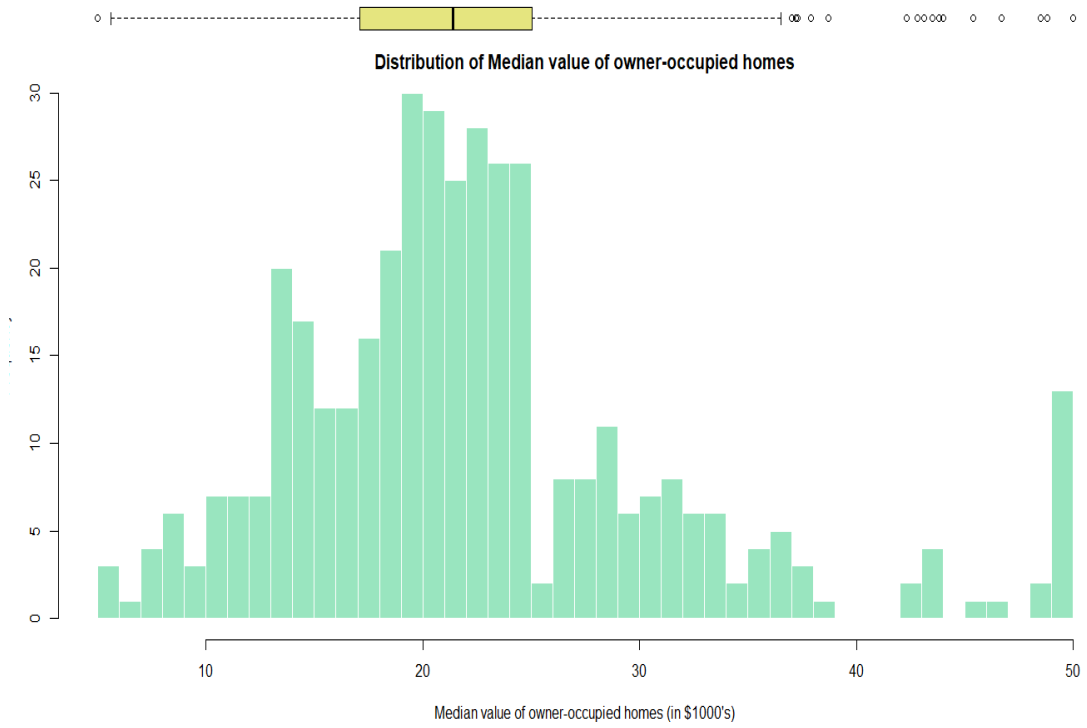


Figure 1: Histogram and Boxplot of the *medv* target variable. The y-axis is the frequency.

### 2.3.2 Continuous variable

Table 1 is a descriptive statistics table of the continuous features of the Boston dataset. For each variable, the following statistics are presented: minimum, first quartile, median, mean, third quartile, maximum value, and the standard deviation. From table 1, the huge disparities between the sample mean and median values of the: *crim*, and *zn* variables suggest the potential presence of outliers in the data. Upon further investigations, *rm* also exhibits outliers. Figure 8 in the appendix is a box-plot of various variables, with outliers represented by the small circles. Given the various unit of measurements between some features, standardising the data aims to help the machine learning algorithm, as variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias.

| **Variable** | **Min** | $\mathbf{q_1}$ | $\widetilde{\mathbf{x}}$ | $\bar{\mathbf{x}}$ | $\mathbf{q_3}$ | **Max** | **s** |
|---|---|---|---|---|---|---|---|
| crim | 0.0 | 0.1 | 0.2 | 3.4 | 3.7 | 73.5 | 8.1 |
| zn | 0.0 | 0.0 | 0.0 | 12.3 | 20.0 | 100.0 | 24.0 |
| indus | 0.5 | 4.9 | 8.6 | 10.9 | 18.1 | 27.7 | 6.8 |
| nox | 0.4 | 0.4 | 0.5 | 0.6 | 0.6 | 0.9 | 0.1 |
| rm | 3.6 | 5.9 | 6.2 | 6.3 | 6.6 | 8.8 | 0.7 |
| age | 2.9 | 44.9 | 77.5 | 68.7 | 94.0 | 100.0 | 27.9 |
| dis | 1.1 | 2.1 | 3.3 | 3.9 | 5.4 | 12.1 | 2.1 |
| tax | 187.0 | 278.5 | 330.0 | 404.5 | 666.0 | 711.0 | 166.3 |
| ptratio | 12.6 | 17.0 | 19.0 | 18.4 | 20.2 | 22.0 | 2.2 |
| lstat | 1.7 | 6.9 | 11.0 | 12.4 | 16.8 | 38.0 | 7.0 |
| medv | 5.0 | 17.1 | 21.4 | 22.6 | 25.0 | 50.0 | 9.0 |

Table 1: Descriptive stats: Continuous variables.

Figure 2 is a pairwise plot of the continuous variables. The diagonal shows distribution of each feature, the correlation matrix to the right of the diagonal, and scatter-plots of each pair of continuous variable. The asterisk represent the significance of the Pearson correlation coefficient.

Looking at the relationship between the predictor variable and the feature sets there are strong linear relationships between some of the variables. For instance, *medv* appears to have a strong negative relationship with *lstat*, whilst there is a strong positive relationship with the *rm* variable. This initial finding indicates what sign of the estimated coefficient it will take.

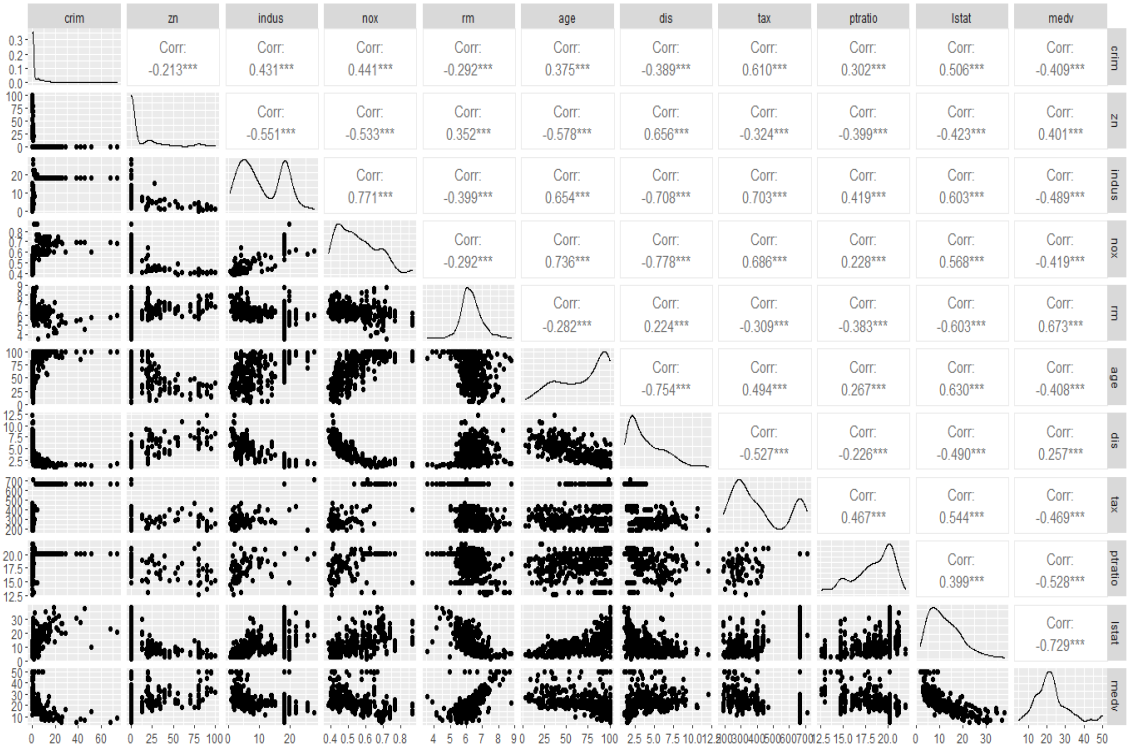


Figure 2: Pairwise plot.

Looking at the histogram in figure 2, the distribution of the *rm* variable seems to be normally distributed. Moreover, the distributions of: *dis* and *lstat* are positively skewed, whilst *ptratio* variable is negatively skewed.

There are insightful patterns that can be drawn from the scatter-plot matrix. For instance, the *medv* and the *lstat* variable exhibit a non-linear relationship. The same can be said between *medv* and *nox*. Although non-linear regression is out of the scope of this report, a log transformation of the dependent variable is an appropriate technique.

Some of the independent variables exhibit high correlation, which potentially alludes to the potential presence of multicollinearity. Multicollinearity occurs when two or more explanatory variables in the MLR model are highly correlated with each other. Essentially, one (or more) independent variables can predict another explanatory variable. In the Boston dataset, there is a strong positive relationship between *age* & *nox*, *tax* & *indus*, and *dis* & *zn*. On the other hand, some independent variables have strong negative relationships such as *dis* & *age*.

### 2.3.3 Categorical variables

Figure 3 displays a grouped box-plot, specifically looking at the median household value against the Charles river dummy variable. In general, variables categorised as tract bounds river have a higher median value than homes grouped as otherwise. The dispersion in the otherwise group is greater than that of *chas* variables grouped as tract bounds river.

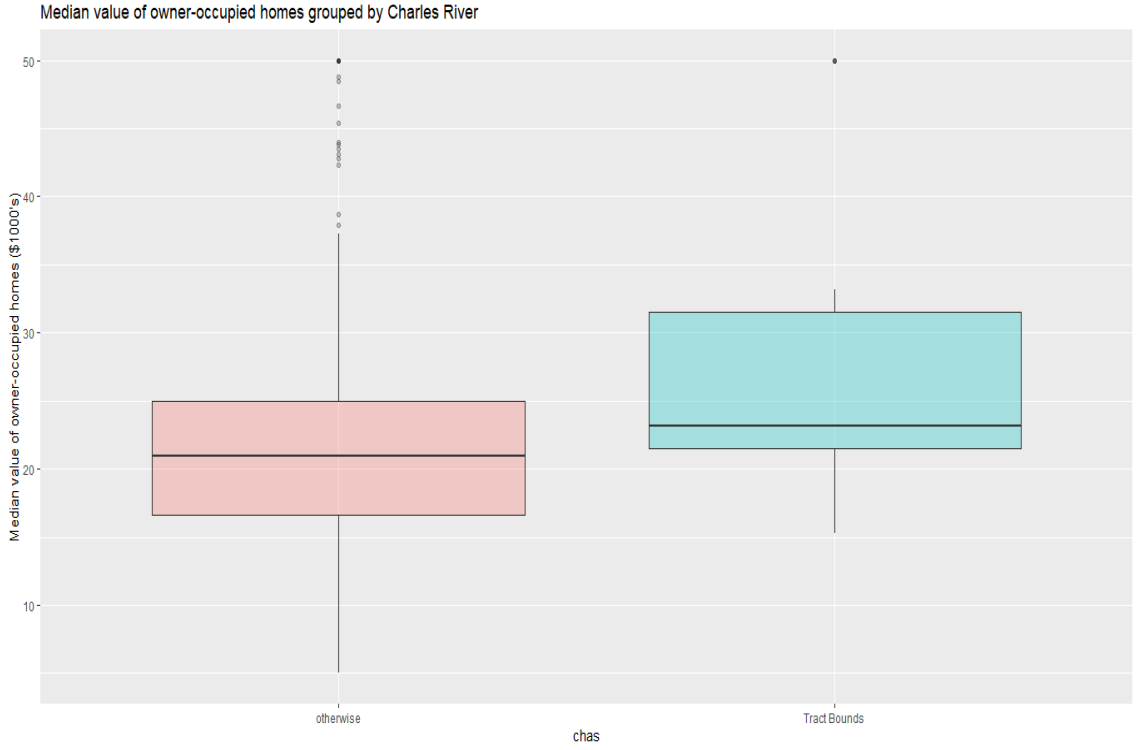


Figure 3: Boxplot: *medv* against *chas*

Figure 4 examines median value houses against the different levels of index accessibility to radial highways. Examining the box-plots, owner-occupied homes classified as level 3 have higher median value, whilst index value 24 homes have generally low median values. Apart from 24 index value, the rest of the categories fall within the same interquartile range. Moreover, some of the categories do possess outliers.

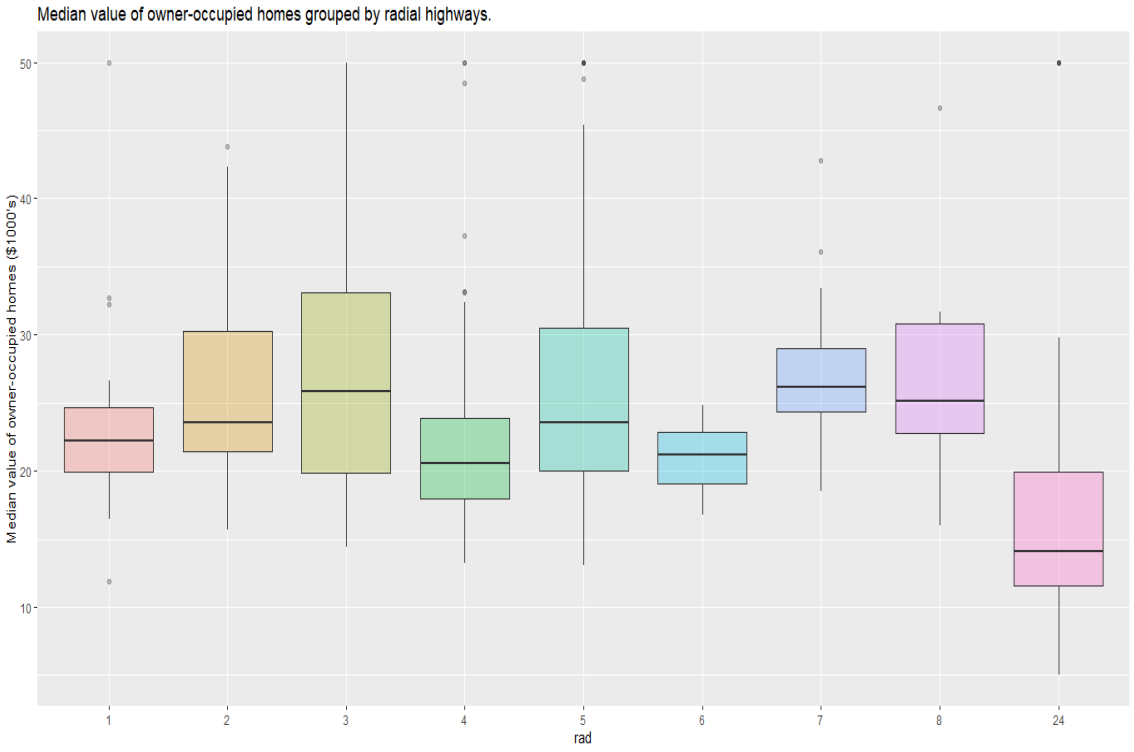


Figure 4: Boxplot: *medv* against *rad*

# 3 Methodology

## 3.1 Data processing

After gaining some insight from EDA, standardisation will be applied before training the data. Next, the data will be split, where 80% of the data is the training set, whilst 20% consists of the testing set. In order to assess model performance, the testing Mean Squared Error (MSE) (i.e. equation 1) will be used.

$$MSE = \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N} \tag{1}$$

## 3.2 Multiple Linear Regression

MLR models are an extension of the simple linear regression model, where more than one independent variable is used to model the dependent variable [2]. In order for MLR estimates to be BLUE (Best Linear Unbiased Estimates), they have to satisfy the Gauss-Markov assumptions such as: the error term has a population mean of

zero, and constant variance, the independent variables are uncorrelated with the error term, and the regression models are linear in its parameters [2].

In this report, a simple MLR model will be trained on the training data, where *medv* is regressed against all the explanatory variable. Using this model, the MSE of the test set will be used as the baseline to measure performance of later models explored in this report. Moreover, the report will check the residuals to verify if the Gauss-Markov assumptions do hold.

### 3.3 Best Subset Selection

There are various selection techniques one can explore such as: forward selection, backward selection, and step-wise selection technique. For the purposes of this report, the Best Subset selection method will be used in order to improve on the baseline model. The algorithm of this technique is as follows [1]:

1. Start with a Null model, which is the dependent variable regressed against $\beta_0$ (i.e,$M_0$). That is, the prediction of this model is the average of the $y_i$ variables.
2. For k = 1,2,..,p; where p equals the number of predictors.
   (a) Fit all $\binom{p}{k}$ models that contains exactly k models. For instance, if $k = 3$, various combinations of models that only contain 3 predictor variables.
   (b) Pick the best model amongst the $\binom{p}{k}$ various models (and label it $M_k$). The best model is defined as having a small Residual Sum of Squares (RSS) (or largest $R^2$).
3. Now you have p models $(M_0, M_1, ..., M_p)$. To select the best model, this report will compare the following statistics: Mallow's $C_p$, Akaike Information Criterion (AIC),Bayesian Information Criterion (BIC), and the adjusted $R^2$.

### 3.4 Lasso Regularization

In supervised learning, over-fitting is major issue. Overfitting is when the model does very well on the raining set, but performs poorly on the testing set. In other words, the model has low bias, but high variance. One way of fixing this issue is by applying regularization to the algorithm. Regularization is a form of regression, where it constraints (i.e. shrinks) the beta estimates towards zero. What regularization does is that is simply penalises a more complex/flexible model. The aim is to avoid the risks of having a model that overfits.

Equation 2 represents lasso regularization where the part to the right of the addition is the penalty factor. As the model sets higher values of $\lambda$, the heavier the penalisation, meaning that the algorithm is restricting the model from being complex. However, lower values of $\lambda$ encourages the use of more flexible models. Lasso makes use of $|\beta_j|$ as its penalty, which is referred to as the L1 norm.

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

### 3.5 Cross Validation

Cross-validation is an important component of machine learning. This technique enables use to construct multiple training and test sets. In other words, cross-validation is a general procedure for estimating the out-of-sample performance of various models. For a K-fold cross validation (where $K = 1, 2, .., N$), the data is first divided into K-groups. At each iteration, the model is trained on the data except on the $K^{th}$ fold, where prediction and error scores are calculated on the fold. For the purposes of this report, a 10-fold cross validation method is when conducting Lasso regularization [1].

### 3.6 Interaction terms

An interaction effect exists if the effect of a predictor variable on a the target variable changes, depending on the value(s) (or categories) of one or more other independent variables. Equation 3 is an example of a MLR model with an interaction term [1].

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_1 + \beta_3 X_1 * X_2 + \epsilon \tag{3}$$

For the purposes of this report, various continuous predictor variables will be tested with the *chas* predictor to investigate if there are any effects of interaction terms.

## 4 Results & Discussion

### 4.1 MSE Test Scores

Table 2 represents the different models built in this report, whilst later sections of the results goes into detail about the different methods used. The baseline model is the model to 'beat'. That is, using the different techniques outlined in the methodology, 4 other models were built, and the test set was used to compare the MSE scores. From table 2, the best model was the one that included the interaction term. More specifically, the interaction term was between the variables *chas* and *nox*.

| Model | Description | MSE Test score |
|---|---|---|
| Baseline | Simple MLR | 0.4265 |
| Model 1 | Best subset method: 8 predictors | 0.4074 |
| Model 2 | Best subset method: 12 predictors | 0.41714 |
| Model 3 | Lasso Regularization: 10-fold CV | 0.4210 |
| Model 4 | Interaction Term | 0.397 |

Table 2: MSE scores of different models.

### 4.2 Simple MLR results

Before delving into variable significance, it is important to discuss the model fit. The simple MLR model reported an F-statistic of 42.64 and a p-value of, essentially, 0. These statistics are a result of a hypothesis test that looks at the overall fitness

of the model. With such a small p-value, there is very strong evidence to reject the null hypothesis (that states there is no linear relationship between the predictors and target variable), and conclude that there is a linear relationship. The model has an $R^2$ of 72.97%. That is 72.97% of the variation in the median value of owner-occupied homes can be explained by the MLR model.

| Predictor variable | Beta Estimate | P-value |
|---|---|---|
| (Intercept) | -0.597 | 0.000 *** |
| crim | -0.090 | 0.018 * |
| zn | 0.172 | 0.000 *** |
| indus | 0.054 | 0.349 |
| nox | -0.249 | 0.000 *** |
| rm | 0.188 | 0.000 *** |
| age | 0.034 | 0.465 |
| dis | -0.317 | 0.000 *** |
| tax | -0.183 | 0.055 * |
| ptratio | -0.213 | 0.000 *** |
| lstat | -0.491 | 0.000 *** |
| boston.data.chasTract Bounds | 0.409 | 0.002 *** |
| boston.data.rad2 | 0.277 | 0.180 |
| boston.data.rad3 | 0.777 | 0.000 *** |
| boston.data.rad4 | 0.393 | 0.018* |
| boston.data.rad5 | 0.467 | 0.006** |
| boston.data.rad6 | 0.305 | 0.129 |
| boston.data.rad7 | 0.692 | 0.001** |
| boston.data.rad8 | 0.285 | 0.218 |
| boston.data.rad24 | 0.922 | 0.000*** |

Table 3: Simple MLR coefficient estimates and p-values.

Table 3 represents the estimates of the predictor variables alongside their respective p-values to indicate statistical significant. The asterisks represents the level of significance, where *, **, and *** represents a variables that is significant at the 10%, 5%, and 1% respectively. The variables *indus* and *age* do not seem to possess any statistical significant, indicating that the model would perhaps do better if these variables were ommitted. The variables: *zn*, *nox*, *rm*, *dis*, *ptratio*, and *chas* possess strong statistical significant relative to the other coefficients.

The sign in front of the beta estimates indicates the relationship between a particular predictor and the target variable. Some of the relationships between the independent and dependent variables align with initial findings when EDA was conducted. For instance, the variables *zn* and *rm* have respective positive relationships with the *medv* variable. This result is similar to what was found in EDA. However, this is not the case for all fetaures. For instance, the correlation between *medv* and *age* was negative, however the MLR model suggests that there is a positive relationship.

As mentioned earlier, the Simple MLR model serves as a baseline for following

models. The MSE of the test set is recorded in table 2. On its own, the test MSE value is arbitrary, but in this particular case it serves as a reference point, with the goal to improve on the initial model.

## 4.3 Best subset Method

For the best subset method, p was set to 12 (i.e, all the features were used). Figure 5 plots the best chosen model in each sub-category, against the following statistics: Residual Sum of Squares, Adjusted $R^2$, Marllow's $C_p$, and the Akaike Information Criterion (AIC). For RSS, $C_p$, and AIC lower values indicate beter models, whilst the opposite holds for the Adjusted $R^2$ statistic.

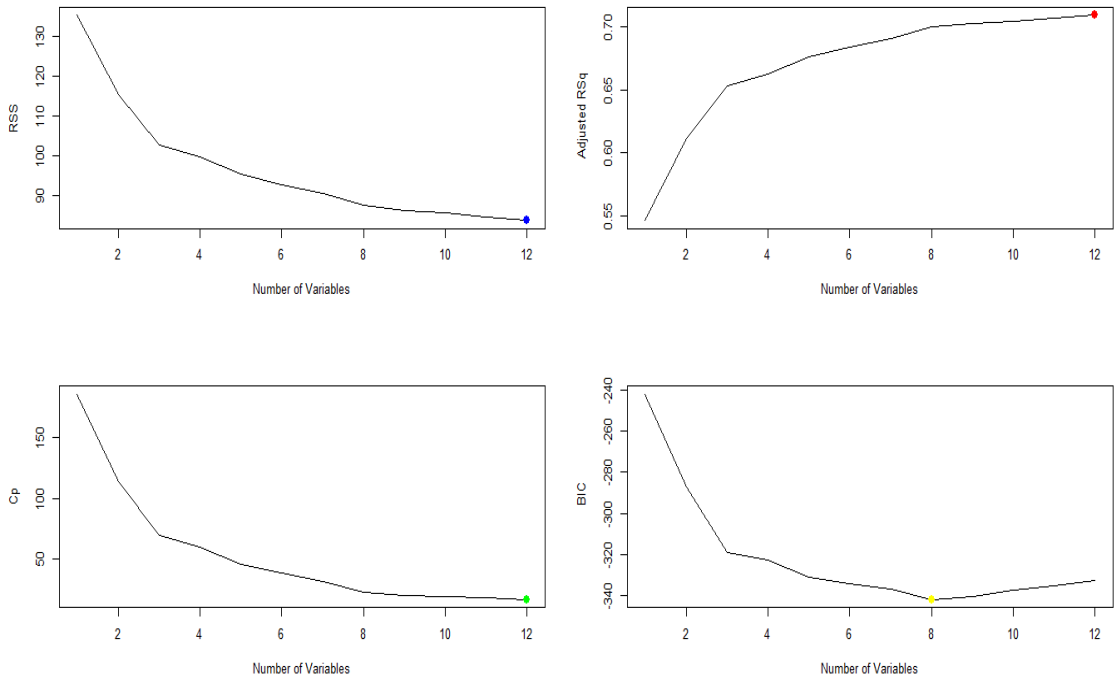


Figure 5: Best Subset Method: Results

Based on the RSS,Adjusted $R^2$, Marllow's $C_p$; the algorithm chooses $M_{12}$ as the optimal model; that is the model that makes use of all the feature variables. Base don the AIC, model $M_8$ is optimal. Due to these conflicting findings, predictions were made using both $M_8$ and $M_{12}$, and the MSE of the test sets referred to as Model 1 and Model 2 in table 2.

## 4.4 Interaction term

In order to investigate interaction terms, 7 out of the 8 predictors that were used in model $M_8$ were used. The categorical variable *chas* was used to test for interactions between the other continuous variables, hence 6 models were trained on the training data.

| Model | Interaction term | Beta Estimate | P-value | Adj $R^2$ | Test MSE |
|---|---|---|---|---|---|
| A | lstat:chas | -0.117 | 0.431 | 0.6903 | 0.399 |
| B | zn:chas | -0.166 | 0.620 | 0.69 | 0.406 |
| C | nox:chas | 0.022 | 0.867 | 0.6897 | 0.397 |
| D | rm:chas | -0.181 | 0.088 | 0.6926 | 0.446 |
| E | ptratio:chas | 0.465 | 0.002 | 0.6992 | 0.420 |
| F | dis:chas | -0.823 | 0.004 | 0.698 | 0.446 |

Table 4: Results for the interaction terms included in the different models

Table 4 are results of 6 various model where each model included an interaction term. The table displays the estimate coefficient of the interaction term and its corresponding p-value, the adjusted $R^2$ of the model in its entirety, and the testing MSE. The results are interesting as the best model (in terms of testing MSE) is model C. This is interesting as the interaction term in this model is statistically insignificant. Model F performs poorly (in terms of adjusted $R^2$), yet the interaction term is statistically significant.

## 4.5 Residual Diagnostics

In this section of the report, a residual diagnostic check was conducted on the best model, based on the testing MSE, being the model containing the interaction term.

The plot to the left of figure 6 displays the residuals against fitted values (a), whilst the plot to the right on figure 6 displays a normal Q-Q plot (b). From plot (a) it is evident that the mean of the residuals is not zero, and the residuals do not have a constant variance. Plot (b) shows there are several points that severely deviate from the 45 degree line, indicating that the residuals are not normally distributed. Overall, the Gauss-Markov assumptions are violated in this case when the simple MLR model is used.

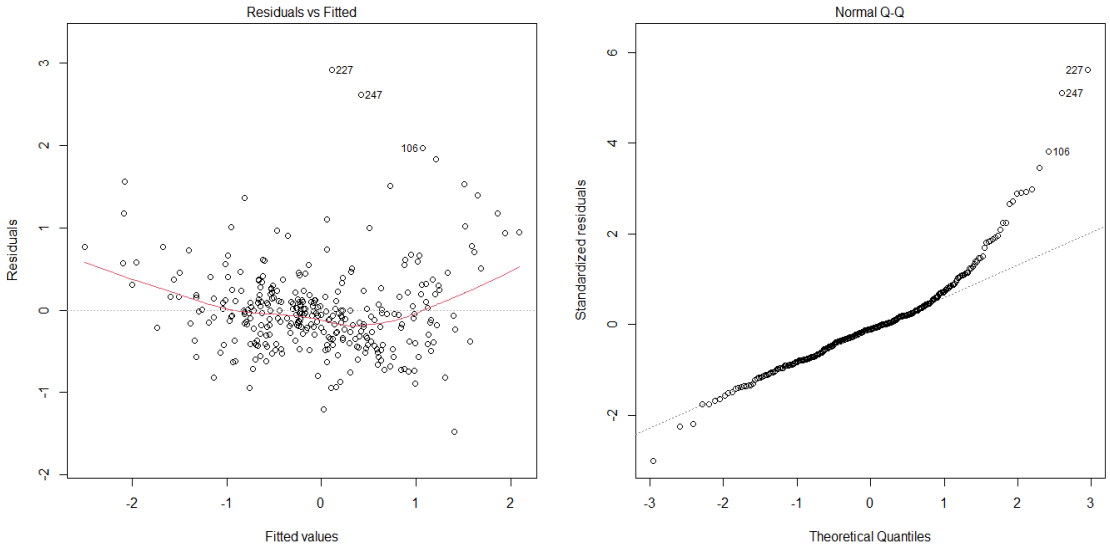


Figure 6: Residual diagnostics: (a) Residuals vs Fitted values, (b) Normal Q-Q plot

Figure 7 is the distribution of the residuals from the simple MLR model, and it is evident that the residuals are not normally distributed, rather the distribution is skewed to the right.

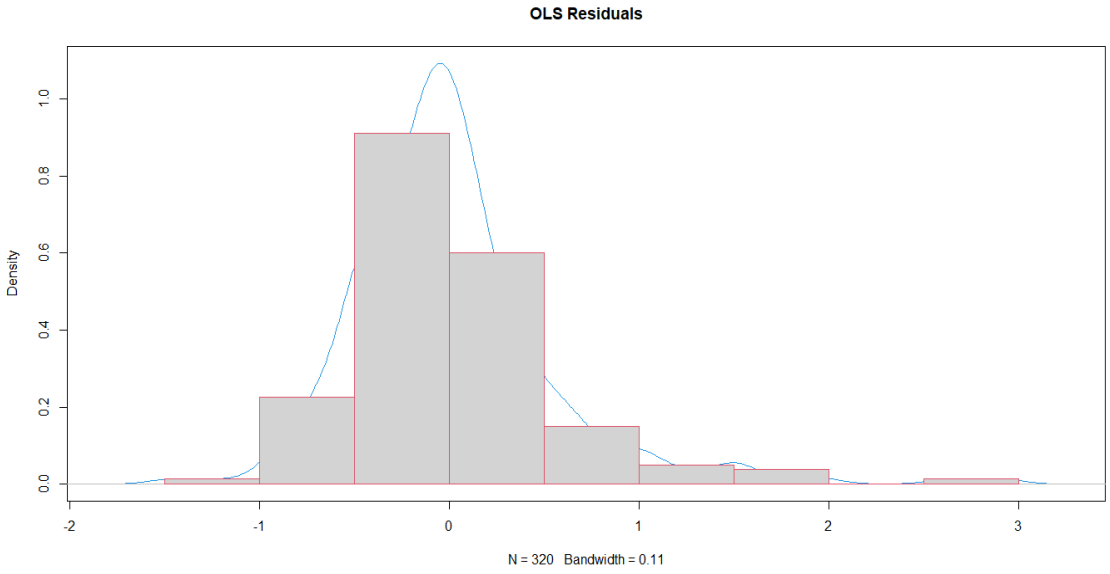


Figure 7: Distribution of the residuals.

# 5 Conclusion

The aim of this report was to build a linear model for predicting the median value (MEDV) of owner occupied homes (in $1000s) for suburbs in Boston, based on 12

explanatory variables. The report first described the data, then proceeded to do some explanatory data analysis, in order to get a better understanding of the data. Next a simple MLR model was trained on the training set, and the test set MSE was computed. Using techniques such as: best subset method, lasso regularization, cross validation, and interactions term, four different models were constructed. Looking at the various models built, the best model (in terms of low test set MSE) was the model that included an interaction term. Residual diagnostics was done on the best model, and it was found that the errors did not conform to the Gauss-Markov assumptions.

# References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[2] Jeffrey M Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2016.

# A Data Description

| Variable | Description | Type |
|---|---|---|
| CRIM | per capita crime rate by town | Continuous |
| ZN | proportion of residential land zoned for lots over 25,000 sq.ft. | Continuous |
| INDUS | proportion of non-retail business acres per town. | Continuous |
| CHAS | Charles River dummy variable (1 if tract bounds river; 0 otherwise) | Categorical |
| NOX | nitric oxides concentration (parts per 10 million). | Continuous |
| RM | Average number of rooms per dwelling | Continuous |
| AGE | proportion of owner-occupied units built prior to 1940. | Continuous |
| DIS | weighted distances to five Boston employment centres. | Continuous |
| RAD | index of accessibility to radial highways. | Categorical |
| TAX | full-value property-tax rate per $10,000. | Continuous |
| PTRATIO | pupil-teacher ratio by town. | Continous |
| LSTAT | Percentage of lower status of the population. | Continous |
| MEDV | Median value of owner-occupied homes in $1000's. | Dependant |

Table 5: Data descriprion

# B Box-plots

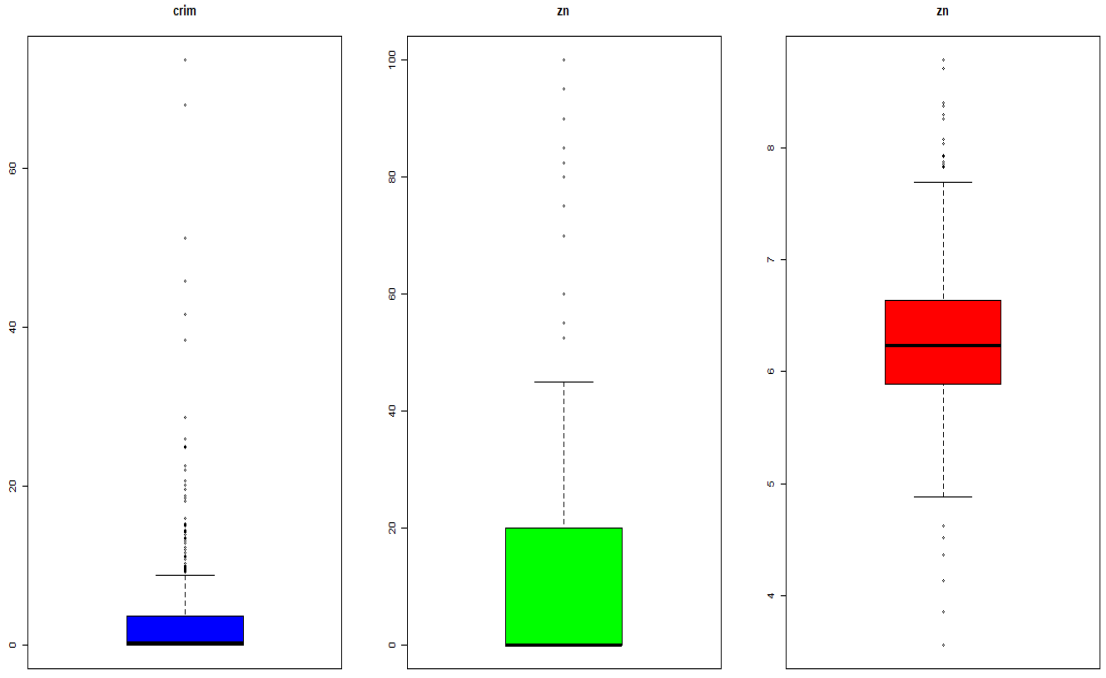


Figure 8: Boxplot of continuous variables.

# C R Code