Multiple Linear Regression Analysis: Boston Housing Prices

Yeachan Park
University of Amsterdam
yeachan.park@student.uva.nl

Natalie Glomsda
University of Amsterdam
nglomsda@gmail.com

Abstract

Using multiple linear regression as the baseline model to predict the Boston housing prices (including 13 features), we improved the model fit by excluding features that did not linearly correlate with the targets, removing any outliers and normalising the data. CHAS and ZN features were excluded from the analysis, and data transformations were applied to 'RM' and 'LSTAT.' Monte Carlo cross validation was then used to train the data with a 70/30 training/validating split. The training data set resulted with R^2 = .80 - .86 (depending on the trials), while the test data set achieved the R^2 = .70 - .80.

1 Introduction

1.1 Linear Regression

Regression analysis aims to construct mathematical models that explain the relationship between two or more variables. The simplest case of linear regression with when dealing with two variables. With n pairs of observation, a scatter plot diagram allows insight into the relationship between the two variables, where:

$$y = b_0 + b_1 x$$

where b_0 refers to the intercept and b_1 refers to the slope in which y is expected to change depending on x.

Multiple linear regression is an extension of simple regression, which attempts to model the relationship between one dependent variable and multiple independent (predicting) variables. The regression line for the explanatory variables is defined as:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_i x_i$$

$$i = 1, 2, ...n$$

This predicts how the mean response (y) changes with the predicting variables (x_i), where values for y are assumed to have the same standard deviations. The fitted values b_0, b_1 , ... , b_i estimate the parameters or coefficient for the regression line, where each predicting variables may have different coefficients/weight that contribute to the change in y given that there is a change in x. To minimise the cost function, Normal Equation will be used to transpose x and y vectors in order to determine the Θ or b_i values, as follows:

$$\Theta = (X^T X)^{-1} X^T y$$

Since multiple regression entails analysing features on different scales, normalisation (or feature-scaling) can be implemented to allow easier interpretation of the regression coefficients, as shown below:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

After we train the training data, we obtain the coefficients which is then used to predict the training data again or the test data. The agreement between the predicted and observed results can be calculated using: (1) the Mean Squared Errors (MSE) and (2) the coefficient of determination (R^2). The MSE measures the average of squares of the difference between the predicted and actual values, and can be calculated by:

$$MSE = \frac{1}{n} \sum_{n}^{i=1} (\widehat{x}_i - x_i)^2$$

R^2 refers to the proportion of variance in the dependent variable that is predictable from the independent variables. The better the linear regression fits the data, the closer the R^2 value is to 1. R^2 can

be computed by:

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

R^2 allows us to determine the fit of the multiple regression model, and thus will be used as the indicator of model fit.

1.2 Data Analysis

This project aims to evaluate the Boston housing dataset, which represents the housing prices in Boston given 13 features (see Table 1 below). The goal is to improve the fit of the linear regression model, thus achieving a high R^2 value. The current baseline regression model is a multiple linear regression, combining all the possible predictors for the housing prices.

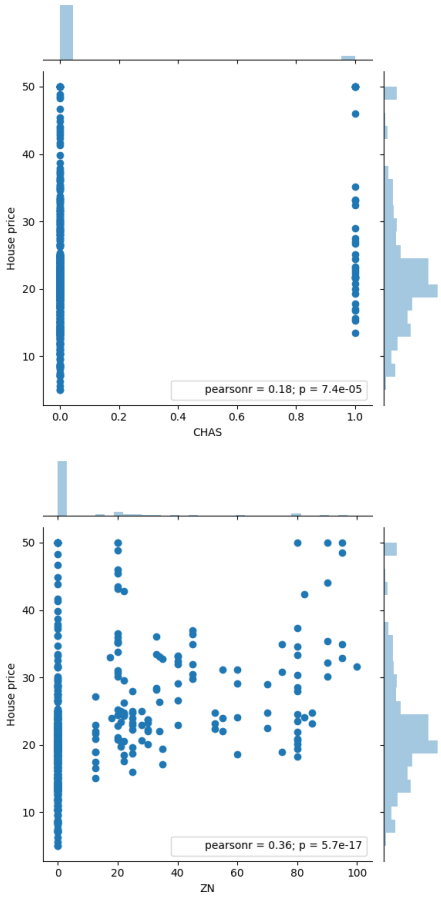
2 Improvements

2.1 Data Processing

Prior to the analysis, we proceeded by removing any data points with MEDV value of 50.00, as it is assumed to be a mistake in the data. Other improvements to the initial dataset was implemented as following:

- Linearity assumption: Checking the correlation between each feature and target after data transformation to determine whether the data is linearly correlated to the target. We modelled this by adding high order polynomials (e.g. $LSTAT^3$), and logarithmally scaled (e.g. MEDV and RM). Features that are not are excluded from further analysis, as they do not meet the linearity assumption for multiple regression, including: CHAS and ZN (see Graphs 1 and 2 below).

- Outliers: Any extreme outliers (z-scores $>$ 10) are also then excluded from the dataset.



2.2 Improving Learning Algorithm

Once the data is ready for training, the dataset is then split into 70/20 training/validating. We train our model purely based on the train data and targets. The **Monte Carlo cross validation** technique allows us to test as many data points as possible in order to ensure that the coefficients obtained will more a more accurate representation of the weights (removes random noise when we fit the data).

3 Experiment

70/30 Testing/Validating We assigned each index/row of the data a value between 0-1 from a uniform distribution. Values (rows) under 0.7 was then selected for training, and those over 0.7 were selected for validating. This ensures a random 70/30 data split. Training will proceed purely on the training data and training targets. Each index of a row is randomly assigned a number from a uniform distribution. For example, a 70/30 split would split the data randomly into a 70/30 train/test set. We obtain model coefficients based on this new training data. This process is repeated iteratively for n times, the coefficients summed

Features	Explanation
CRIM	crime rate
ZN	proportion of residential lots $>$ 25,000 sq.ft.
INDUS	proportion of non-retail business acres/town
NOX	nitric oxides concentration
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per 10,000 dollars
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	percentage lower status of the population

Table 1: Features in the Boston housing dataset

and a mean is obtained for the coefficients.

In our regression model, we found that 'CHAS' and 'ZN' were not very good predictors of the log median price and were removed. 'AGE' was also not a good predictor by itself and was removed, however we modelled an interaction effect of AGE and $LSTAT^2$ and this was a good predictor of log house price. We also noted that adding in $LSTAT^2$ and $LSTAT^3$ allowed the model to achieve a higher R^2 value. Logarithmically scaling RM also increased R^2.Overall, after normalisation, the model we obtained in one attempt was: $log(y) = 3.069 + (CRIM * -0.532) + (INDUS * 0.072) + (NOX * -0.061) + (RM * 4.398) + (log(RM) * -3.961) + (DIS * -0.235) + (RAD * 0.464) + (TAX * -0.310) + (PTRATIO * -0.263) + (B * 0.177) + (LSTAT * -1.209) + (LSTAT^2 * 1.821) + (LSTAT^3 * -0.393) + (-0.79 * (LSTAT^2 * AGE))$. However, note that the coefficients will change every time you run the regression as the train/split data that is used is different each time. In this model, we see an increase in crime, $NOX, log(RM), DIS, TAX, PTRATIO, LSTAT, LSTAT^3$ and $LSTAT^2 * AGE$ all result in a decrease in $log('MEDV')$. Conversely, an increase in $INDUS, RM, RAD, B$ and $LSTAT^2$ all predict an increase in $log('MEDV')$.

Using the training data coefficients to predict new data, we can see that the adjusted R^2 value is around 0.80 - 0.84. This value changes for each trial, as our original train/test split algorithm is random, as is the cross validation in to a small degree. However, the latter's effect is minimised over multiple iterations.The adjusted R^2 value for the test data is usually around 0.7.

3.1 Post-Prediction Checks

This step checks for (1) first order autocorrelations between residuals, (2) residual distribution using Kolmogorov-Smirnov, (3) residual histogram, and (4) residual outliers. Figure X below shows a plot with loess line, which checks for residual homoscedasticity.

4 Conclusion

In conclusion, multiple linear regression was employed to model the Boston housing dataset. We found that by excluding $CHAS$ and ZN from the analysis as they do not linearly correlate with the target, as well as completing log(RM), $LSTAT^2$, $LSTAT^3$, and through k-fold cross validation we are able to obtain the R^2 of .80 - .86 for the training data, and R^2 = .70 for the test data.

Minimising the cost function by using normal equation allows us to determine the coefficients without having to choose a learning rate or iteration. However, a limitation of using this method instead of gradient descent is that if the features have a higher nrow, it becomes problematic when cross validating.

Future improvements for the Boston housing data analysis include assessing the multicollinearity assumption within the features, performing regularisation to avoid overfitting.