Maya Hirsch

Applying a Multiple Linear Regression Model in order to understand the relationship between predictors and the median housing prices

1. Introduction

In contemporary society, being able to predict housing price is a major asset. Individuals, families, and businesses constantly look at properties to see whether this is an investment they would like to take on. However, an issue arises when trying to forecast the value of the property. There are many factors which contribute to the price. Location, schools in the area, weather, safety etc. Understanding the relationships between variables provides important insights for predicting future housing prices. I will be applying multiple linear regression in order to understand the relationship between predictors and the median housing prices using the Boston Housing Dataset.

2. The Dataset

The Boston housing dataset comes from UCI Machine Learning Repository and MASS package in R, which contains information on houses in Boston from the 1970s. There are 506 samples and 14 predictors in the dataset, with no missing values. In this study, I aim to understand the relationship between five predictors and the response variable, the median housing prices in Boston, MEDV (Table 1). These predictors were chosen because of higher correlations with MEDV ($r > |0.3|$). Indus and Nox appeared highly correlated, so Nox was removed. Furthermore, the correlation coefficients showed high collinearity between tax and indus with an $r > 0.7$ (Figure 1). The two predictors that I predict to negatively impact median home price are indus and lstat, and the three predictors I predict positively impact home price are rm, ptratio, and tax.

Table 1. *Variables with descriptions for the Boston Housing Dataset.*

Variable Name	Description
indus	Proportion of non-retail business acres per town
rm	Average number of rooms
ptratio	Pupil-teacher ratio
tax	Full-value property-tax rate per $10,000
lstat	Percentage of lower status of the population
medv	Median value of owner-occupied homes in thousands USD

3. Descriptive Statistics

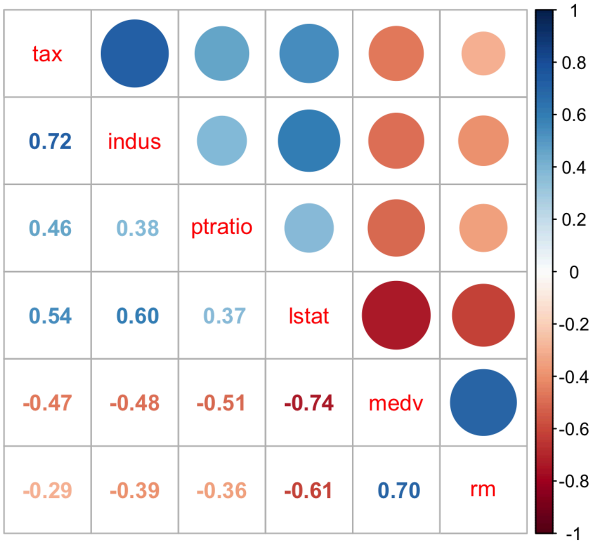
For the response variable, the average median housing price is $22.53 thousand with a standard deviation of 9.20 and a median of 21.20, which is very similar to the mean (Table 2). With the exception of tax, all other predictors have medians that are approximately similar to their means. Through the coefficients of the model, we can understand the importance of each predictor in influencing median housing prices.

Table 2. *Summary statistics for medv, and the five predictors.*

Variable	Mean	Variance	Std. Dev.	Median
medv	22.53	84.64	9.20	21.20
indus	11.14	48.44	6.86	9.69

rm	6.28	0.49	0.70	6.21
ptratio	18.46	4.67	2.16	19.05
tax	408.24	28405.73	168.54	330.00
lstat	12.65	160.02	7.14	11.36

Figure 1. Correlogram with pairwise Pearson correlation coefficients



4. Methodology and Diagnostics

An initial model was first constructed as follows:

$$\hat{y} = 17.52 + 0.06(indus) + 0.46(rm) - 0.004(tax) - 0.88(ptratio) - 0.56(lstat)$$

To evaluate this model, I assessed whether the assumptions for multiple linear regression appeared reasonable under our model. The key assumptions for multiple linear regression are (1) linearity of dependent and independent variables, (2) normally distributed error terms, (3) constant error terms.

Checking for linearity in the original full model, the plot of residual vs fitted showed some curvature. This indicated there is a pattern in response which is unexplained by the current model. To check the constant variance assumption, the scale-location plot has curvature and is not a flat line at 0 which suggests there is not constant variance. The Q-Q residuals plot is not a straight line, indicating that there are

several error terms which do not follow a normal distribution. Further, the residuals vs leverage plot show influential points with high leverage and some outliers. Thus, all assumptions seem to be violated in the original full model.

Table 3. *Coefficients for the three models run with Variance Inflation Factors (VIF)*

Predictor	**β** **(intial)**	**β** **(w/ BoxCox transform)**	**β** **(weighted least squares/ final model)**	**VIF** **(BoxCox transform)**
indus	0.057	0.0064	-	2.42
rm	4.63***	0.38***	0.44***	1.69
tax	-0.004	-.001**	-	2.38
ptratio	-0.88***	-0.082***	-0.09***	1.36
lstat	-0.56***	-.066***	-0.07***	2.25
Adjusted R^2	0.677	0.713	0.733	-

P-value significance: ***** $p < 0.01$, *** $p < 0.001$***

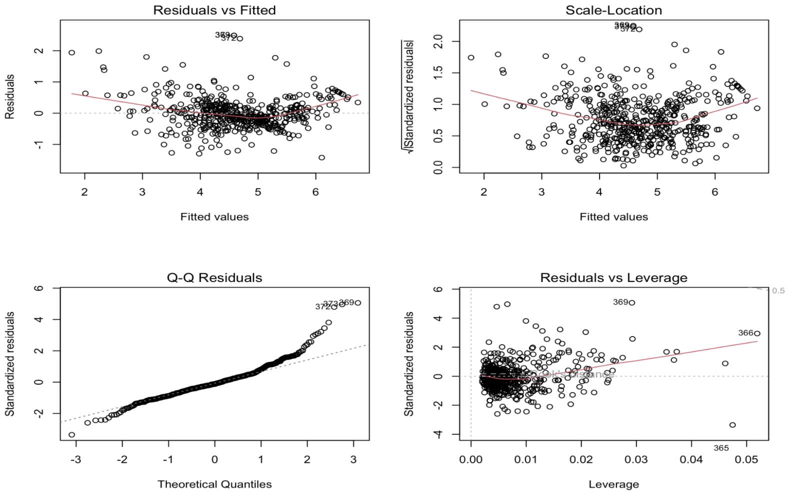
To combat the violation of the linearity assumption, I applied a box-cox transformation. Looking at the diagnostic plots, the transformation improved the linearity, constant variance, and normality assumptions. However, there was still heteroscedasticity. As such, I next investigated influential points, leverages, and outliers to see whether outliers cause this. After removing outliers, the linearity assumption was violated, so the outliers remained in the model. I also performed a Breusch-Pagan test; the results indicated significant heteroscedasticity. Consequently, I applied weighted least squares (WLS) as well as box-cox transform on the reduced model (Table 3). The final model was as follows:

$$\hat{y} = \beta_0 + \beta_1(\text{rm}) + \beta_2(\text{ptratio}) + \beta_3(\text{lstat})$$

$$\hat{y} = 4.46 + 0.44(\text{rm}) - 0.09(\text{ptratio}) - 0.07(\text{lstat})$$

The evaluation metric I used to assess fit was the adjusted R^2 which had been 0.7126 after box-cox transform was now 0.733 (Table 3), suggesting improvement with WLS. Thus, in the final model, 73.3% of the variation in median housing price was explained by the combination of the variables. The assumptions of the final model are overall met, although there are still significant unequal variances as shown by the curvature in the scale-location plot and the Breusch-Pagan test (Figure 2). Although there are outliers, the Q-Q plot shows residuals as largely on the reference line between the quantiles of -2 and +2, suggesting that there is an approximately normal distribution of error terms. There is some violation of linearity, as seen by the curvature of the residuals vs. fitted plot (Figure 2), and this is even after both box-cox transform and WLS, which may need further correction.

Figure 2. Residuals diagnostic plots for the final WLS model with box-cox transform.



5. Interpretation

In the study, we analyzed whether or not the predictors in the model were significant predictors of the response variable MEDV using techniques including hypothesis testing, F-tests, partial F-tests, and t-tests. For the variable indus, a partial F-test showed p > 0.05, indicating it did not significantly contribute in explaining the variance of MEDV. Therefore, I removed this predictor and adopted a reduced model. Tax was also removed due to high VIF and collinearity with lstat (Figure 1). Thus, the final weighted least squares contained rm, ptratio, and lstat as the independent variables and MEDV as the dependent variable (Table 3). Because the square root of

MEDV was taken for the box-cox transform and WLS models, coefficients will need to be back-transformed for interpretation.

6. Conclusions

The final model contained three significant independent predictors of median house value: number of rooms (rm), pupil-teacher ratio (ptratio), and percent lower status (lstat). I found the only positive predictor of MEDV to be RM, which suggests that for every additional room in a home, on average, after controlling for all other variables there is an average increase in median house price of $194. For the predictor of ptratio, the effect was negative, such that for every one additional person per teacher the average median house price falls by $81. For the predictor of lstat, the effect was negative which suggests that for an additional percentage of lower status of the population, the price falls by $49.

Building and evaluating multiple linear regression models helps us understand the importance of each predictor in influencing median housing prices. These insights can help Individuals, families, and businesses looking to make real-estate and lifestyle decisions. Through modeling relationships between variables we gain valuable insights and exciting opportunities for many possible applications.