**BANA 3308 Homework 4**

**Conducting a data analysis project**

Using MEDV as the outcome variable for the Boston housing analysis, several steps have been taken to choose the predictors that will help achieve the best predictive power for the outcome variable. Backward Elimination was used as a technique to select predictor variables, which began with a multiple linear regression model containing all variables and then removed those that were not useful. In this elimination step, scatterplots were employed to determine the strong correlation between each predictor and the outcome variable. The correlation between the predictors was also evaluated, revealing negative correlations between AGE and DIS, as well as DIS and NOX, while positive correlations were observed between NOX and INDUS, and TAX and RAD. Consequently, these correlated pairs of predictors were not used together in the regression model to prevent multicollinearity. Additionally, some predictors were chosen based on logical reasoning and findings from journal articles.

After completing all the steps to select predictors for optimizing the predictive power of the outcome variable MEDV in this linear regression model (utilizing linear regression due to the continuous nature of the outcome variable), the following section illustrates the impact of the chosen predictors on the outcome variable before conducting the analysis:

* **TAX (full-value property-tax rate per $10,000):** A higher property tax rate (TAX) is likely to have a negative effect on median home values (MEDV) as properties in areas with high tax rates tend to be less desirable, resulting in lower values.
* **RM (average number of rooms per dwelling):** The number of rooms (RM) is generally positively correlated with home values as more rooms often indicate larger and more valuable homes. So, a positive effect on MEDV is expected.
* **LSTAT (% of the population with low-income status):** The percentage of low-income residents (LSTAT) is expected to have a negative effect on MEDV as higher LSTAT values often indicate lower-income neighborhoods, which can lead to decreased property values.
* **CRIM (per capita crime rate by town):** It is logical to expect that a higher crime rate (CRIM) will have a negative effect on MEDV since higher crime rates can deter potential homebuyers and lead to lower property values.
* **B (proportion of black residents):** Based on Kantrowitz's research on how the Black Population faced disparities in housing costs (1979), it is logical to expect that a higher population of black residents will have a negative effect on MEDV.
* **PTRATIO (pupil-teacher ratio by town):** A higher pupil-teacher ratio (PTRATIO) might have a negative effect on MEDV as it indicates bigger class sizes and smaller teacher sizes where students may not receive better education opportunities.
* **DIS (weighted distances to five Boston employment centers):** It should have a positive effect since being close to work is a plus for home values, and shorter commutes usually mean higher home prices.
* **INDUS (proportion of non-retail business acres per town):** A higher proportion of non-retail businesses might indicate an industrial area, potentially lowering property values for residential areas and indicating a negative relationship with MEDV.

**After Running the Analysis:**

The statistical results of the linear regression analysis are illustrated below:  
A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated  
*Figure: Linear Regression Score (left); Coefficient Table (right)*

In the coefficient results, the intercept value of 19.49 units represents the estimated value of the dependent variable or the outcome (MEDV) when all the predictors or independent variables are set to zero. So, the estimated median home value (MEDV) would be $19,490.80 when all the predictor values are set to zero or have no influence. It is important to note that the intercept value only provides a mathematical reference point as having zero influence of the predictors on the median home value (MEDV) is not a reflection of an ideal and real-world interpretation.

While analyzing the coefficient values for the selected predictors, the values having negative results indicate a negative relationship and the values having positive results indicate a positive relationship with the outcome variable MEDV. In the analytical result, most of the coefficient values of the selected predictors align with the expectations before running the analysis. For instance, negative coefficients were expected for predictors like TAX, LSTAT, CRIM, PTRATIO, and INDUS, and the data supports these expectations. On the contrary, RM was expected to have a positive effect on MEDV which is also illustrated in the data. Amongst the predictors, LSTAT has a strong negative correlation, and RM has a strong positive correlation with MEDV according to the “Correlation” widget, the linear regression results, and the visual scatterplot in Orange.

A screenshot of a computer

Description automatically generated A screen shot of a computer

Description automatically generated

*Figure: Coefficient widget (left); Scatterplot for RM – MEDV correlation (right)*

A screen shot of a computer

Description automatically generated

*Figure: Scatterplot for LSTAT – MEDV correlation*

While most of the predictors were consistent with the initial predictions, there are some exceptions for the predictors DIS and B whose results didn’t match with the prediction.

The predictor variable DIS was expected to have a positive impact on housing value since proximity to work and shorter commutes typically attract buyers, leading to higher housing prices. However, the coefficient result shows a negative value. Though the concept may seem counterintuitive, the prediction and result align. It was predicted that DIS, which represents the weighted distance to five Boston employment centers, would have a positive effect, with shorter commutes resulting in higher home prices. In the regression, the variable DIS is used as a distance measure, where a higher DIS value indicates that the business centers are farther from the houses, leading to a decrease in home values when DIS increases. Conversely, a lower DIS value (shorter distance) would increase housing values. Hence, the relationship between DIS and MEDV is shown as negative, and the prediction aligns with the regression result.

While some of the predictors were chosen based on feature selection techniques, and correlations, others like CRIM, TAX, and B were chosen based on the logical and theoretical reasonings. The dataset was taken in the 1970s, and during that time according to research from Kantrowitz in the book “Racial and Ethnic Residential Segregation in Boston 1830-1970” illustrated how the Black Populations faced racial disparities in housing, including higher costs and limited access to quality housing (1979). Another journal research from Little et al. emphasizes how the economic forces and discrimination interact to shape the housing landscape for populations with low-income status and the Black population and sheds light on the complex dynamics of racial discrimination in housing markets and provides valuable insights. (1977). Based on theoretical reasoning through research, predictor B was expected to have a negative relationship with the outcome variable. However, the regression result and visualization show a different result:  
A screen shot of a computer

Description automatically generated

*Figure: Scatterplot for B – MEDV correlation*

After conducting the analysis, the statistics and visualizations reveal a complex relationship between the proportion of the Black population (B) and median housing values (MEDV). Upon examination of both the coefficient table and the scatterplot, there is no distinct evidence of a negative correlation between B and MEDV. The scatterplot indicates an absence of correlation between B and MEDV, and the coefficient for B is also near zero. One potential explanation for the complexity of the relationship between B and MEDV is that the dataset may not fully capture the intricate dynamics of racial discrimination within the housing market, as suggested by existing research. Additionally, it's essential to recognize that the sign of the coefficient results from mathematical relationships within the data and may not always directly reflect the real-world impact or historical context. **References:**  
Kantrowitz, N. (1979). Racial and ethnic residential segregation in Boston 1830-1970. *Annals of the American Academy of Political and Social Science*, *441*(1), 41–54. https://doi.org/10.1177/000271627944100105

Little, J. T., Yinger, J., Καιν, J. F., & Quigley, J. M. (1977). Housing Markets and Racial Discrimination: A Microeconomic analysis. *Land Economics*, *53*(3), 376. https://doi.org/10.2307/3146129