**Homework 2. Multiple Linear Regression**

Author: Rohit V. Akole

UConn ID: 3133252

University of Connecticut

OPIM5604 Predictive Modeling

Professor: Pankaj Prakash

Abstract

This homework is about Boston Housing Dataset and Freshman College Data. SAS JMP is used as the statistical analytics software.

**Definitions** of the datasets of the housing in Boston Massachusetts are as follows:

**CRIM** - per capita crime rate by town

**ZN** - proportion of residential land zoned for lots over 25,000 sq.ft.

**INDUS** - proportion of non-retail business acres per town.

**CHAS** - Charles River dummy variable (1 if tract bounds river; 0 otherwise)

**NOX** - nitric oxides concentration (parts per 10 million)

**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

**PTRATIO** - pupil-teacher ratio by town

**LSTAT** - % of lower economic status population

**MEDV** - Median value of owner-occupied homes in $1000's

**CAT.MEDV** - Derived variable.Median value of owner-occupied homes is greater than $30,000 when CAT.MEDV = 1 and less than $30,000 when CAT.MEDV = 0

Table of Contents

[Question 1 4](#_Toc145084227)

[a. All iterations of the Boston Housing Model 4](#_Toc145084228)

[b. Final Model Interpretation of relation between DIS vs. MEDV 7](#_Toc145084229)

[Question 2 8](#_Toc145084230)

[a. Inter Quartile Range (IQR) for Test 1 and Test 2 8](#_Toc145084231)

[For Test 1: 8](#_Toc145084232)

[For Test 2: 8](#_Toc145084233)

[b. Takeaway by comparing box plots 8](#_Toc145084234)

[Question 3 9](#_Toc145084235)

[a. Screenshots of the Multiple Linear Regression Model 9](#_Toc145084236)

[b. Best predictor 11](#_Toc145084237)

[c. Relationship between freshman coll gpa (best predictor) vs. first-year retention rate 11](#_Toc145084238)

# Question 1

Q1. Using the dataset Boston Housing.JMP (provided with Homework 1), build a linear regression model. The outcome variable to be modelled is median house value (MEDV).

## All iterations of the Boston Housing Model

|  |  |
| --- | --- |
| Figure 1. Multiple Linear Regression Part 1 | Figure 2. Multiple Linear Regression Part 2 |

|  |  |
| --- | --- |
| Figure 3. Multiple Linear Regression Part 3 | Figure 4. Multiple Linear Regression Part 4 |

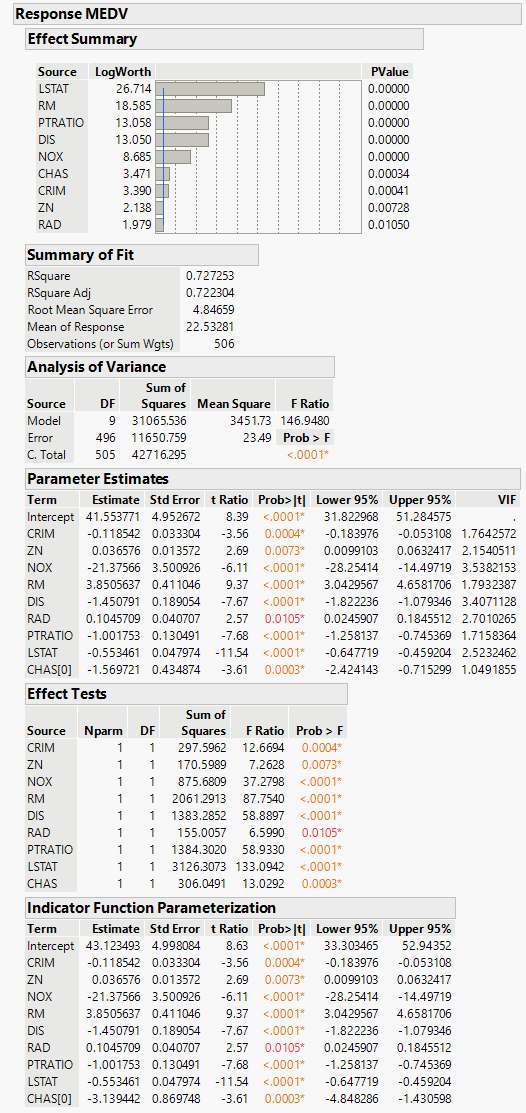


Figure 5. Multiple Linear Regression Part 5

## Final Model Interpretation of relation between DIS vs. MEDV

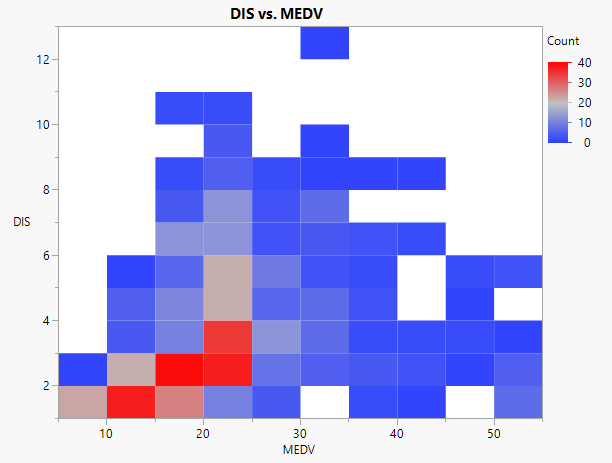


Figure 6. DIS vs. MEDV Heatmap

According to the heatmap, we can say that the largest cluster is in between MEDV 10 to 25 and DIS 0 to 4. The rest of the graph is spread all over. We can see by the Indicator Function Parameterization, for every DIS increase, the MEDV value is decreasing by around $1,450. The log worth of DIS is 13.050, t Ratio is -7.67, and p-value is <0.0001. It shows that the DIS is statistically significant for our model.

# Question 2

Q2. The box plots below compare the final exam (test #2) and mid-term exam (test#1) scores

(on a scale of 0 to 24) for 100 students in the Statistics 101 class.

## Inter Quartile Range (IQR) for Test 1 and Test 2

Formula:

IQR = Quartile 3 – Quartile 1

Minimum = Quartile 1 – (1.5 \* IQR)

Maximum = Quartile 3 + (1.5 \* IQR)

### For Test 1:

IQR = [16 – 6] = 10

### For Test 2:

IQR = [16 – 9] = 7

## Takeaway by comparing box plots

The box plot of Test 1 is skewed left (Negative). But the box plot of Test 2 is skewed right (positive). This means, students did well in Final exams as compared to Midterm exams. The size of the boxplot in Test 1 was more spread as compared to Test 2. Score towards Quartile 3 remained constant throughout the course but scores towards quartile 1 improved significantly.

# Question 3

Q3. a school wants to explain the first-year retention rate (% of students who make it to the second year). Build a Linear Multiple Regression Model to help explain the factors that may or may not significantly affect retention? The outcome variable to be modelled is the 1st year retention rate.

## Screenshots of the Multiple Linear Regression Model

|  |  |
| --- | --- |
| Figure 7. Multiple Linear Regression Part 1 | Figure 8. Multiple Linear Regression Part 2 |

For the Freshman college Dataset, In Figure 7, all the predictors are available. Since college is just a random number and the p-value of the college is >0.05, we cannot use it as a predictor since it won’t give correct data. So in Figure 8, I dropped college first. At this point, the RSquare didn’t change much. But it changed from 0.938384 to 0.937094 and Adjusted RSquare from 0.928382 to 0.928817. After dropping college, the p-value of 3 predictors are still more than 0.05(HS GPA, top 10%, top 20%).

|  |  |
| --- | --- |
| Figure 9. Multiple Linear Regression Part 3 | Figure 10. Multiple Linear Regression Part 4 |

Since the p-value of % 10% was the highest, I dropped in as seen in Figure 9. It changed the RSquare from 0.937094 to 0.93709 and Adjusted RSquare from 0.928817 to 0.930638. But still the p-value for % top 20% and HS GPA is higher than the 0.05. So in Figure 10, we dropped % top 20% and it changed the RSquare from 0.93709 to 0.934939 and Adjusted RSquare from 0.930638 to 0.93006. Till now, we dropped all the statistically insignificant predictors such as college, % top 10%, % top 20%. But still the p-value for HS GPA is greater than 0.05. So let’s drop that as well.

|  |  |
| --- | --- |
| Figure 11. Multiple Linear Regression Part 5 | Figure 12. Multiple Linear Regression Part 6 |

In Figure 11, we dropped HS GPA, now p-value for both Avg ACT and freshman coll gpa are under 0.05. While looking at VIF though, VIF is more than 4 for both. But it’s the same for both at 12.99039. Since the LogWorth of freshman coll gpa is higher than the Avg ACT and the p-value is also smaller for freshman coll gpa as compared to Avg ACT. So, for the final predictor in Figure 12, we dropped Avg ACT and now we have freshman coll gpa with LogWorth of 24.002.

## Best predictor

For the Freshman College Dataset, freshman coll gpa is the best predictor with the LogWorth of 24.002.

## Relationship between freshman coll gpa (best predictor) vs. first-year retention rate

We had the RSqure of 0.938384 intially, and at the final model, the RSquare changed to 0.920625. Similarly, Adjusted RSquare was initially 0.928382, down to 0.918735 in the final model. The final model states than for ever point in freshman coll gpa, the 1st year retention rate will increase by 0.1567436. Similarly, as per Figure 11, for every point increase in Avg ACT score, the 1st year retention rate is increasing by 0.011651. Overall, freshman coll gpa shows upward trend with 1st year retention rate in colleges.