**Case Study 2:**

**Multiple Linear Regression in Housing Values**

Case Study by: Diyang Yu & Elif Yurtseven

Professor: Vinayak Elangovan

Class: CSCI 300

Date: 3/15/17

CONTENTS

Introduction………………………………………………………………………………………...……..2

Aim………………………………………...……………………………………………………...………2

Variables ………………………………………...…………………………………………………...……2

Rationale of selecting certain variables…………………...………………………………………………3

Graph 1: Scatter plot of each variable…………...…………………………………………………4

Table 1: Correlation table of MEDV with each predictor.…………………………………………5

Modeling and Results………………………………………...……………………………………………..6

Conclusion………………………………………...………………………………………………………..7

Appendix A: Complete Correlation Table………………………………………………………………….8

Appendix B: R code………………………………………...……………………………………………..10

Introduction

Regression is a form of analysis that predicts a model that investigates the relationship between a dependent and independent variables. With this technique, we can understand how the typical value of the dependent variable changes when any of the independent variables is varied as the independent variables stay fixed. Regression indicates strength of impact of multiple independent variables on a dependent variable. Linear regression establishes a relationship between one independent variable and one dependent variable using a straight fit line. In this case study, we will analyze the impact of multiple independent concerns about housing in the suburbs of Boston.

Aim

In this case study, we have chosen the domain about Boston housing. The data offers 13 continuous attributes and 1 binary attribute. We are going to choose the appropriate attributes out of these 14, as some of them will probably not be as relevant to computing linear regression as others. After deciding which attributes are more valuable for the desired regression and justifying it, we will perform multivariate regression on the chosen attributes such that we can find the equation of best fit and the coefficient of determination. We are expecting that certain variables such as the number of rooms per house and the number of lower status people in the neighbourhood will affect the median value of owner occupied homes in $1000s more than other variables. In addition, we will try and understand which independent variable is the most relevant to the decision making about the housing data.

Variables

The following link gives us access to the data that we used for this case study:

<http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data>

Below, we represent the attribute information for the complete list of variables obtained from the data folder:

1. CRIM: per capita crime rate by town

2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.

3. INDUS: proportion of non-retail business acres per town

4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

5. NOX: nitric oxides concentration (parts per 10 million)

6. RM: average number of rooms per dwelling

7. AGE: proportion of owner-occupied units built prior to 1940

8. DIS: weighted distances to five Boston employment centres

9. RAD: index of accessibility to radial highways

10. TAX: full-value property-tax rate per $10,000

11. PTRATIO: pupil-teacher ratio by town

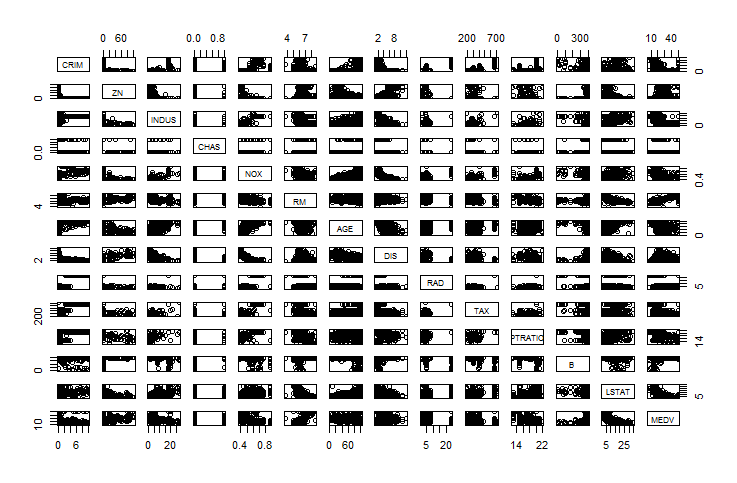
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

13. LSTAT: % lower status of the population

14. MEDV: Median value of owner-occupied homes in $1000's

Rationale of selecting certain variables

To choose the appropriate attributes for independent variables x, the first thing we created is the scatterplot of each variable against another. The complete graph can be seen below.



Graph 1: Scatter plot of each variable

From the plot, we can see average number of rooms per dwelling (RM) and percentage of lower status of the population (LSTAT) are highly correlated to the housing values (MEDV). As RM goes up, MEDV increases as well; while as LSTAT increases, MEDV goes down. Although LSTAT might not be a linear relationship with MEDV since the scatter plot shows some curves in there, it does have an influence on MEDV. These findings meet our expectation, because it is very clear that the increase of number of room per house will raise housing prices. At the same time, with more lower status people in the neighborhood, the housing prices are definitely cheaper than rich area. Therefore, we will definitely include LSTAT and RM in our model. Besides those two variables, we also see some patterns in other variables, such as ZN, INDUS, NOX, AGE, DIS, TAX and PTRATIO. We delete CHAS and RAD that are the variables that indicate the location of the house regarding its closeness to the Charles River and highways. CHAS is a binary attribute hence it is not valuable as an independent variable as it it not continuous. We do not see a difference in the scatter plot, so we decide to exclude those two variables. We also want to omit CRIM (crime rate by town) and B (proportion of blacks) because the scatter plot shows a right-angle pattern. This means that the data are vertical at first and after certain point, they become horizontal. At the same time, we cross out TAX (tax rate) because there is a huge pag between data, which violates the normality and equal variance assumption. We don’t think adding these attributes might improve the model so we delete these three variables.

Then we look at the correlation table and decide to include which variable further. Below is the partial correlation table since we care more about the correlation between each predictor response variable (MEDV). The complete correlation table is in the appendix.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | MEDV |
| MEDV | -0.286245 | 0.33156988 | -0.4119145 | 0.154408725 | -0.3327782 | 0.7401808 | -0.2998932 | 0.1387984 | -0.2179021 | -0.34589757 | -0.46121356 | 0.26479723 | -0.706255059 | 1 |

Table 1: Correlation table of MEDV with each predictor.

The correlation table also justifies our thoughts that RM and LSTAT are highly correlated with MEDV with correlation of 0.74 and -0.71 respectively. By applying the common sense, we remove AGE (proportion of owner-occupied units built prior to 1940) and NOX (nitric oxides concentration) because the influence of these values on the model will be small. For other variables, both scatter plot and correlation table convey that they are influential, and the real-world meanings make sense. Thus finally we decide to build a linear model using ZN, INDUS, RM, DIS, PTRATIO and LSTAT to predict MEDV.

Modeling and Results

We used language R to calculate slopes and the coefficient of determination (r^2). The complete code is in the appendix, and below are the results.

> beta

[,1]

15.2745261

ZN 0.0280086

INDUS -0.1060614

RM 5.2482780

DIS -0.9762066

PTRATIO -0.7168135

LSTAT -0.6164145

> r.squared

[1] 0.7081114

If we write the equation out, it would be:

We can interpret the intercept in the following way. When proportion of residential land, proportion of non-retail business acres, average number of rooms per dwelling, weighted distances to five Boston employment centers, pupil-teacher ratio and percent of lower status population are all 0, the median value of houses in Boston would be $15,275. To interpret slopes, we pick RM as an example. While holding other variables constant, as the mean number of rooms per dwelling increases by 1, the median value of houses in Boston would increase by $5,258. Then, we can interpret r^2 as 70.8% of the variability explained by the model. The results we got in the model were consistent with correlation and scatter plot; r^2 also shows that the model does pretty well in predicting the median value of houses in Boston.

Conclusion

The aim of this case study was to be able to perform linear regression on a chosen dataset of interest. We did this by computing the effect of each independent variable be got from the dataset on the dependent variable which was MEDV (median value of owner occupied homes). We computed the correlation of all variables using both a scatter plot and a correlation table. As we slowly omitted certain variable due to lack of value affecting the dependent variable, we computed a reduced correlation table which allowed us to write a best fit equation with all most relevant data which were ZN (proportion of residential land zoned for lots over 25,000 sq.ft.), RM (average number of rooms per dwelling), DIS (weighted distances to five Boston employment centres), PTRATIO (pupil-teacher ratio by town) and LSTAT (percentage lower status of the population). All in all, we can say that we were able to compute such correlation; though we found the two most relevant attributes to be LSTAT and RM. Our value for r^2 was 0.708 which is a relatively good value for the coefficient of determination of a dataset that initially had 14 variables, one of which was irrelevantly binary and one of which was the dependent one. One way we could increase the value of r^2 in similar future studies is by adding more variables or trying the plotting more times or by comparing the residual errors to decide what modifications we want to make.

**Appendix A: Complete Correlation Table**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | MEDV |
| CRIM | 1 | -0.28123893 | 0.5738223 | 0.050065901 | 0.6369411 | -0.14245773 | 0.4476638 | -0.4619674 | 0.89798831 | 0.82566775 | 0.3194701 | -0.41301437 | 0.424788624 | -0.286245 |
| ZN | -0.2812389 | 1 | -0.5142733 | -0.059696546 | -0.5007285 | 0.30662341 | -0.5556789 | 0.6561521 | -0.26660593 | -0.26944592 | -0.36412219 | 0.15037968 | -0.411165213 | 0.3315699 |
| INDUS | 0.5738223 | -0.51427332 | 1 | 0.103016306 | 0.7385173 | -0.36489467 | 0.6059816 | -0.6693615 | 0.51330603 | 0.67331226 | 0.31733613 | -0.31675222 | 0.565402489 | -0.4119145 |
| CHAS | 0.0500659 | -0.05969655 | 0.1030163 | 1 | 0.1340642 | 0.07657959 | 0.1231407 | -0.1406621 | 0.0573367 | 0.01689409 | -0.09962273 | 0.0134152 | -0.009429928 | 0.1544087 |
| NOX | 0.6369411 | -0.50072854 | 0.7385173 | 0.134064175 | 1 | -0.26459441 | 0.7071478 | -0.7458124 | 0.54249915 | 0.61511982 | 0.10346419 | -0.35843313 | 0.536824184 | -0.3327782 |
| RM | -0.1424577 | 0.30662341 | -0.3648947 | 0.076579589 | -0.2645944 | 1 | -0.1878709 | 0.1387741 | -0.09593148 | -0.21494783 | -0.33416415 | 0.10835237 | -0.607288917 | 0.7401808 |
| AGE | 0.4476638 | -0.55567886 | 0.6059816 | 0.123140665 | 0.7071478 | -0.18787087 | 1 | -0.7203343 | 0.35932632 | 0.42709472 | 0.19295579 | -0.22376517 | 0.573266278 | -0.2998932 |
| DIS | -0.4619674 | 0.65615208 | -0.6693615 | -0.140662124 | -0.7458124 | 0.13877413 | -0.7203343 | 1 | -0.38838521 | -0.44413069 | -0.15225328 | 0.2344494 | -0.423724689 | 0.1387984 |
| RAD | 0.8979883 | -0.26660593 | 0.513306 | 0.057336696 | 0.5424992 | -0.09593148 | 0.3593263 | -0.3883852 | 1 | 0.87287642 | 0.38748427 | -0.35258569 | 0.309787921 | -0.2179021 |
| TAX | 0.8256677 | -0.26944592 | 0.6733123 | 0.016894085 | 0.6151198 | -0.21494783 | 0.4270947 | -0.4441307 | 0.87287642 | 1 | 0.38451066 | -0.36708374 | 0.410926755 | -0.3458976 |
| PTRATIO | 0.3194701 | -0.36412219 | 0.3173361 | -0.099622729 | 0.1034642 | -0.33416415 | 0.1929558 | -0.1522533 | 0.38748427 | 0.38451066 | 1 | -0.08960928 | 0.303043086 | -0.4612136 |
| B | -0.4130144 | 0.15037968 | -0.3167522 | 0.013415202 | -0.3584331 | 0.10835237 | -0.2237652 | 0.2344494 | -0.35258569 | -0.36708374 | -0.08960928 | 1 | -0.291093833 | 0.2647972 |
| LSTAT | 0.4247886 | -0.41116521 | 0.5654025 | -0.009429928 | 0.5368242 | -0.60728892 | 0.5732663 | -0.4237247 | 0.30978792 | 0.41092675 | 0.30304309 | -0.29109383 | 1 | -0.7062551 |
| MEDV | -0.286245 | 0.33156988 | -0.4119145 | 0.154408725 | -0.3327782 | 0.7401808 | -0.2998932 | 0.1387984 | -0.2179021 | -0.34589757 | -0.46121356 | 0.26479723 | -0.706255059 | 1 |

**Appendix B: R code**

housing <- read.csv("housing.csv")

housing <- na.omit(housing)

plot(housing)

cor(housing,use="complete")

x <- cbind(rep(1,nrow(housing)),as.matrix(housing[,c(2,3,6,8,11,13)]))

y <- as.matrix(housing[,14])

beta <- solve(t(x) %\*% x) %\*% t(x) %\*% y

yhat <- x %\*% beta

epsilon <- y - yhat

r.squared <- 1-sum(epsilon\*epsilon)/sum((y-mean(y))^2)