Boston Housing Regression Project

Erick Castillo Graduate Student

Xena Adono Graduate Student

December 14, 2021

Contents

1 Introduction

The Boston data set used in our research project was found on Kaggle.com originally collected from the U.S. Census Service. The data is from 1979 and pertains to housing in the Boston Massachusetts area. It contains a total of 506 cases and 14 attributes with some missing values. We decided to use the attribute MEDV—Median value of owner-occupied homes in thousands as the response variable and all other attributes as predictor variables.

For our main research question, we wanted to produce a model that would be efficient in explaining the variability between the response variable MEDV and all other predictor variables used within the final model. That is explain the variation of median home value by taking into account a number of different attributes. For our second question, we decided on constructing a prediction interval for the median value of a home given mean values for all predictor variables except for predictors: NOX—nitric oxides concentration, CHAS—Charles River dummy variable, and TAX—full-value property-tax rate. We wanted to predict the median value given lower levels of nitric oxides, lower taxes and if the house was within close proximity to the Charles River.

2 Questions of Interest

1. How much variance can be explained by the model?

2. Obtain a 95% prediction interval for the median home value (MEDV) for a house that's by the Charles River (CHAS = 1) with minimum levels of nitric oxide (NOX) and property tax rate (TAX).

3 Regression Method

Throughout the semester, we learned how to use simple and multiple regression in order to answer different and various research questions. For our first question, we will be focusing on obtaining a reasonable R-squared greater than 0.5 in our final model. This will make our final model a good model that fits the data. Also, a high R-squared will make a good predictive model. This will help us answer our second question when

obtaining a 95% prediction interval. We maintained our questions in mind during our regression analysis and used the best methodology learned to obtain the best responses we could.

4 Regression Analysis

Before answering the above questions, a summary on how the best regression model was obtained will be provided.

4.1 Cleaning the Data

EDA revealed that the there were 20 missing values for 6 of the 11 predictors, leaving a total of 120 missing cells. Mean and mode imputation were used to fill in these gaps, leaving a complete data set ready for analysis.

4.2 Steps to Find "Best" Model

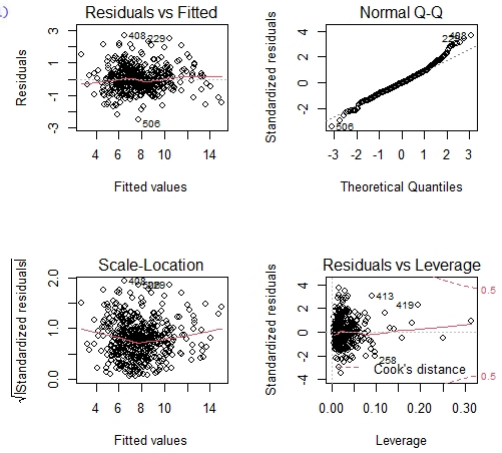
To start, a stepwise regression procedure was used to narrow down the best model in terms of AIC. An inspection of this model revealed that quadratic terms were necessary to fix the non-linearity issue present. Upon adding these quadratic terms to the model, the linearity assumption was met. A set of interaction terms were then explored, and none were added because they only marginally increased R^2_{adj} while tremendously increasing the complexity of the model.

Normality and equal variance assumptions were not met for the model, so a Box-Cox transformation was implemented to the response variable. The resulting model had variance assumptions satisfied, but lacked the normality in the residuals. Outliers were found using an Standardized Residuals (SR) vs Leverage plot, 6 observations that had a SR greater than ± 3 were dropped from the dataset.

A new model, with the same predictors as before, was fit with this reduced dataset. A new Box-Cox transformation was applied, and all assumptions were satisfied, except normality. This "best" model is what's used to answer the indicated questions.

```
> mod7 <- lm(medv2~LSTAT+LSTAT2+RM+RM2+PTRATIO+CHAS+B+DIS+NOX+CRIM+RAD+TAX, data = df1)
> summary(mod7)

Call:
lm(formula = medv2 ~ LSTAT + LSTAT2 + RM + RM2 + PTRATIO + CHAS +
    B + DIS + NOX + CRIM + RAD + TAX, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.51944 -0.46134 -0.03316  0.41103  2.73391

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.5317436  1.7564267  17.952  < 2e-16 ***
LSTAT       -0.1682116  0.0224071  -7.507 2.90e-13 ***
LSTAT2       0.0019336  0.0006063   3.189 0.001519 **
RM          -6.2227996  0.5335907 -11.662  < 2e-16 ***
RM2          0.5638993  0.0416948  13.524  < 2e-16 ***
PTRATIO     -0.1617693  0.0197985  -8.171 2.66e-15 ***
CHAS1        0.4092342  0.1434224   2.853 0.004510 **
B            0.0015035  0.0004212   3.570 0.000392 ***
DIS         -0.1514919  0.0257783  -5.877 7.77e-09 ***
NOX         -3.2032650  0.5546444  -5.775 1.37e-08 ***
CRIM        -0.0386506  0.0051200  -7.549 2.18e-13 ***
RAD          0.0379952  0.0099651   3.813 0.000155 ***
TAX         -0.0021822  0.0005206  -4.192 3.28e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.742 on 487 degrees of freedom
Multiple R-squared:  0.8708,    Adjusted R-squared:  0.8676
F-statistic: 273.5 on 12 and 487 DF,  p-value: < 2.2e-16
```

(a) Summary output



(b) Plots of interest

Figure 1: "Best" model outputs of interest.

4.3 Variance Explained by the Model

To find how much variation is explained by the model, we study the summary output of our best model. The $R^2_{adj} \approx 0.8676$. This means that about 87% of the variability in the response variable MEDV is explained by the 12 predictors present in the model. It is important to be aware that the normality assumption is not met. This results in predictors that have incredibly low p-values, that don't significantly increase the R^2_{adj}. This means that the model isn't as parsimonious as it could be, and hypothesis testing on the $\hat{\beta}_i$'s is unreliable. Nonetheless, with an R^2 this high, predictions and both prediction and confidence intervals will be relatively accurate.

4.4 Prediction Intervals

A prediction interval setting minimum levels of nitrous oxide (NOX), minimum property taxes (TAX), location by the Charles River while holding all other variables constant was generated, with the following output:

```
> newdata = data.frame(LSTAT = mean(df1$LSTAT), LSTAT2 = mean(LSTAT2), RM = mean(df1$RM), RM2 = mean(RM2),
+                      PTRATIO = mean(df1$PTRATIO), CHAS = as.factor(1), B = mean(df1$B), DIS = mean(df1$DIS),
+                      NOX = min(df1$NOX), CRIM = mean(df1$CRIM), RAD = mean(df1$RAD), TAX = min(df1$TAX))
> predict(mod7, newdata, interval = 'predict')^(3/2)
       fit      lwr      upr
1 27.82487 21.28298 34.92643
```

Figure 2: Prediction interval

This output indicates that the price of a house with these set conditions would be between $21,000 and $35,000, which is on the lower end of the presented response variables, as the original data had houses priced as high as $150,000.

5 Conclusion

With the proposed question answered, it's best to take a step back and see the results for what they are. Due to the high R^2_{adj}, this model has very high predictive power, while preserving a decent amount of parsimony. Though it was not performed in this study, a train-test split with cross-validation can be performed in the future to examine if this model is as strong as it hypothetically seems. A major point of interest for future work would be to find how to enforce the normality assumption on the model. This would hopefully lead to an even more parsimonious model that would be much more easier to interpret.

On a final note, the importance of this model is negligible. To reiterate the source of this data, it's from the housing scene in Boston from 1979, meaning that it has very little widespread application. One interesting fact that we were able to observe from this work is that specific detriments in a neighborhood decreased the price of housing. Examples of these detriments includes things like high levels of NOX and high crime rates.

6 Acknowledgments

We would like to thank Dr. Liao for a great semester. We learned a lot about the fundamentals of regression analysis, and we look forward to applying our knowledge again when the moment presents itself.

7 Appendix

Our R Code for the analysis. Note that because this paper was prepared in LaTeX, a lot of the carets (^) and comments (#) had to be dropped so that the file could compile properly. The following code is not complete for these reasons.

```
library(faraway)
library(ISLR)
library(alr4)
library(leaps)
library(corrplot)
df <- read.csv("C:/Users/casti/Downloads/boston_comp1.csv")
df$CHAS <- factor(df$CHAS)
cont.df <- subset(df, select = -c(CHAS))
corrplot(cor(cont.df), method = 'color')
mod1.redu <- lm(MEDV 1, data = df)
mod1.full <- lm(MEDV ., data = df)
step(mod1.redu, scope = list(lower = mod1.redu, upper = mod1.full))
df.reg <- subset(df, select = -c(MEDV, CHAS))
mod2 <- lm(MEDV   LSTAT + RM + PTRATIO + CHAS + B + DIS + NOX + ZN + CRIM + RAD
+ TAX, data = df)
summary(mod2)
anova(mod2, mod1.full)
par(mfrow=c(2,2))
plot(mod2)
plot(lm(MEDV LSTAT, data = df))
plot(lm(MEDV RM, data = df))
mod3 <- update(mod2, . .+I(LSTAT2) + I(RM2))
```

```
summary(mod3)
mod4 <- update(mod3, . . - ZN)
summary(mod4)
add1(mod4, .+LSTAT*RM + CRIM*RM + CRIM*DIS + CRIM*B + NOX*B + NOX*RM + RM*DIS
+ DIS*B + RAD*DIS + RAD*B + TAX*RM + TAX*DIS + TAX*B + PTRATIO*RM + B*RM, data =
df, test ='F')
plot(mod4)
vif(mod4.1)
mod.boxcox = boxCox(mod4, lambda = seq(0, 0.5, length = 10))
medv1 <- df$MEDV(1/3)
mod5 <- lm(medv1 LSTAT+I(LSTAT2)+RM+I(RM2)+PTRATIO+CHAS+B+DIS+NOX+CRIM+RAD+TAX,
data = df)
summary(mod5)
plot(mod5)
df1 <- df[-c(365,369,370,371,372,373), ]
mod6 <- lm(MEDV LSTAT+I(LSTAT2)+RM+I(RM2)+PTRATIO+CHAS+B+DIS+NOX+CRIM+RAD+TAX,
data = df1)
summary(mod6)
mod.boxcox = boxCox(mod6, lambda = seq(0, 1, length = 10))
medv2 <- df1$MEDV(2/3)
LSTAT2 <- df1$LSTAT2
RM2 <- df1$RM2
mod7 <- lm(medv2 LSTAT+LSTAT2+RM+RM2+PTRATIO+CHAS+B+DIS+NOX+CRIM+RAD+TAX,
data = df1)
summary(mod7)
newdata = data.frame(LSTAT = mean(df1$LSTAT), LSTAT2 = mean(LSTAT2), RM = mean(df1$RM),
```

RM2 = mean(RM2), PTRATIO = mean(df1$PTRATIO), CHAS = as.factor(1), B = mean(df1$B), DIS = mean(df1$DIS), NOX = min(df1$NOX), CRIM = mean(df1$CRIM), RAD = mean(df1$RAD), TAX = min(df1$TAX))

predict(mod7, newdata, interval = 'predict')