Final Project Report

Statistical Analysis of Boston house prices

STAT 614 Applied Statistics

Utkarsh Dabholkar

Introduction

Dataset description:

The Boston Housing dataset contains information about different characteristics of houses in the Boston area and their median values. The dataset contains 505 observations (houses) and 13 variables (features) and 1 (label).

There are 13 predictor variables that describe various characteristics of the houses, the average number of rooms per dwelling (RM), the proportion of non-retail business acres per town (INDUS), and the pupil-teacher ratio by town (PTRATIO), among others. The response variable is the median value of owner-occupied homes in thousands of dollars (MEDV).

Based on the variables in the dataset, we can see that there are both continuous and categorical variables. The continuous variables include ZN, INDUS, NOX, RM, AGE, DIS, TAX, PTRATIO, B, LSTAT, and MEDV, while the categorical variables include CHAS and RAD.

Objectives:

The objectives of this report are to test and gain insights into the relationships between the variables in the Boston Housing dataset and to provide statistical evidence to support any conclusions drawn from the analysis.

Question 1:

Is there evidence to suggest that there is a significant relationship between the proportion of non-retail business acres per town and the median value of owner-occupied homes in Boston, or is there no such relationship? (Week 10 Two-Sample Hypothesis Testing)

Hypothesis:

H0: There is no significant relationship between the proportion of non-retail business acres per town and the median value of owner-occupied homes in Boston.

Ha: There is a significant relationship between the proportion of non-retail business acres per town and the median value of owner-occupied homes in Boston.

Method:

We use INDUS variable for this, we convert it into 2 groups and then test the variance of the two groups using Var-test, using that we will do the t-test to find out its relationship with the MEDV.

```
> # Subset the data into two groups based on proportion of non-retail business acres
> low_indus <- subset(df, INDUS <= median(df$INDUS))
> high_indus <- subset(df, INDUS > median(df$INDUS))
> var.test(low_indus$MEDV, high_indus$MEDV)

        F test to compare two variances

data:  low_indus$MEDV and high_indus$MEDV
F = 1.0726, num df = 257, denom df = 247, p-value = 0.579
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8371667 1.3734828
sample estimates:
ratio of variances
          1.072633

> # Perform two-sample t-test
> t.test(low_indus$MEDV, high_indus$MEDV, var.equal = TRUE)

        Two Sample t-test

data:  low_indus$MEDV and high_indus$MEDV
t = 10.367, df = 504, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.24497 9.16557
sample estimates:
mean of x mean of y
 26.30930  18.60403
```

P-value < 2.2e-16, t-stat = 10.367, alpha = 0.05.

Results and conclusion:

The hypothesis test suggests strong evidence (p-value < 2.2e-16) is smaller than 0.05 to reject the null hypothesis (H0) that there is no significant relationship between the proportion of non-retail business acres per town and the median value of owner-occupied homes in Boston. The alternative hypothesis (Ha) that there is a significant relationship between these variables is supported by the data.

Question 2:

Is there a significant difference in the median value of owner-occupied homes among the different neighborhoods in Boston, or are these median values the same across all neighborhoods? (Week 11 ANOVA)

Hypothesis:

H0: There is no significant difference in the median value of owner-occupied homes among the different neighborhoods in Boston.

Ha: There is a significant difference in the median value of owner-occupied homes among the different neighborhoods in Boston.

Method:

Used ANOVA-test to check if significant difference in the median value of owner-occupied homes (MEDV) among the different neighborhoods (Charles River and proportion of non-retail business acres per-town)

```
> # Perform a two-way ANOVA
> fit <- aov(MEDV ~ as.factor(CHAS) + factor(INDUS), data = df)
> summary(fit)
                  Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(CHAS)    1   1312  1312.1  29.765 8.26e-08 ***
factor(INDUS)     75  22494   299.9   6.804  < 2e-16 ***
Residuals        429  18911    44.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P-value for Chas = 8.26e-08, P-value for Indus < 2e-16, alpha= 0.05

Results and conclusion:

The results of the ANOVA test show that there is a significant difference in the median value of owner-occupied homes among the different neighborhoods in Boston. The p-value of 8.26e-08 for the "CHAS" variable, p-value of <2e-16 for the "INDUS" variable are both significantly smaller than alpha. The null hypothesis is rejected and alternative hypothesis, indicating that there is a significant difference in the median value of owner-occupied homes across the different neighborhoods in Boston.

Question 3:

Can we conclude that there exists a significant linear relationship between the predictor variables and the median value of owner-occupied homes in Boston or not? (Week 12 Regression)

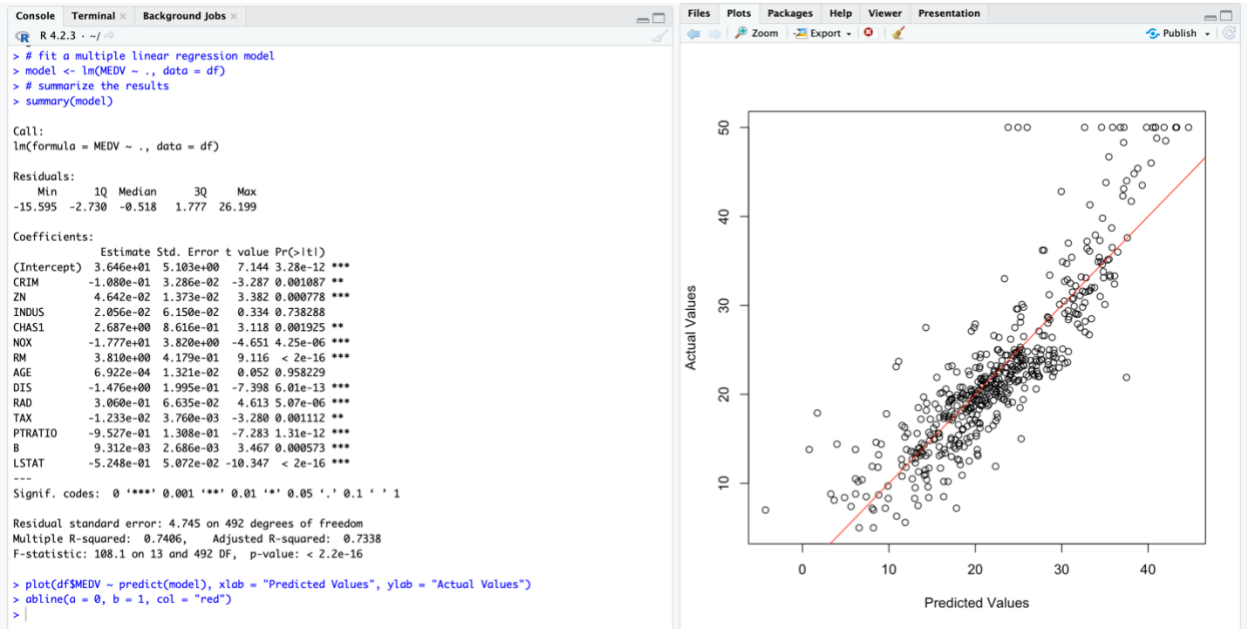
Hypothesis:

H0: There is no linear relationship between the predictor variables and the median value of owner-occupied homes in Boston.

Ha: There is a linear relationship between the predictor variables and the median value of owner-occupied homes in Boston

Method:

Used multiple liner regression to find the relationship between the predictor variables and the median value of owner-occupied homes in Boston.



```
> # fit a multiple linear regression model
> model <- lm(MEDV ~ ., data = df)
> # summarize the results
> summary(model)

Call:
lm(formula = MEDV ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
CRIM        -1.080e-01  3.286e-02  -3.287 0.001087 **
ZN           4.642e-02  1.373e-02   3.382 0.000778 ***
INDUS        2.056e-02  6.150e-02   0.334 0.738288
CHAS1        2.687e+00  8.616e-01   3.118 0.001925 **
NOX         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
AGE          6.922e-04  1.321e-02   0.052 0.958229
DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
PTRATIO     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
B            9.312e-03  2.686e-03   3.467 0.000573 ***
LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16

> plot(df$MEDV ~ predict(model), xlab = "Predicted Values", ylab = "Actual Values")
> abline(a = 0, b = 1, col = "red")
>
```


Results and conclusion:

Based on these findings, we may conclude that the predictor factors and the median value of owner-occupied residences in Boston have a statistically significant linear connection. The F-statistic's p-value is less than 2.2e-16, indicating that at least one of the predictor variables and the response variable are significantly correlated. According to the modified R-squared value of 0.7338, the predictor variables can account for around 73.4% of the variation in the median price of owner-occupied homes in Boston. Many of the predictor variables are significantly correlated with the response variable, such as RM (average number of rooms per dwelling), DIS (weighted distances to five Boston employment centers), and variables like CRIM (per capita crime rate by town) and LSTAT (percent lower status of the population) have negative coefficients, which means that as these predictor variables rise, the median value of owner-occupied homes tends to fall.

In conclusion, we reject the null hypothesis and conclude that there is a significant linear relationship between the predictor variables and the median value of owner-occupied homes in Boston.

Question 4:

Is there evidence to suggest that the median value of owner-occupied homes in Boston follows a normal distribution, or does the distribution of these median values differ significantly from a normal distribution? (Week 13 Goodness of Fit and Independence)

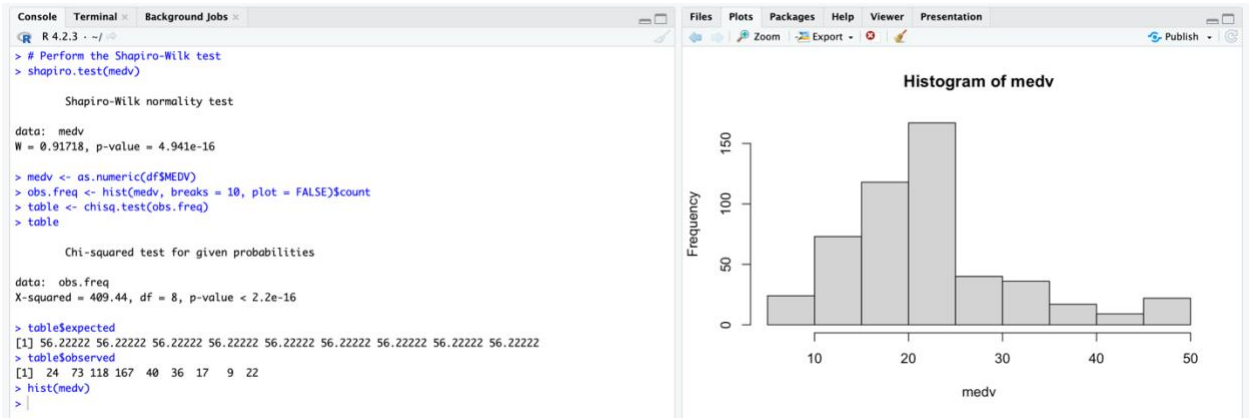
Hypothesis:

H0: The median value of owner-occupied homes in Boston follows a normal distribution.

Ha: The median value of owner-occupied homes in Boston does not follow a normal distribution.

Method:

For this Question we perform two test chi-squared goodness-of-fit test and Shapiro-Wilk test. We break MEDV variable into 10 groups and remove its expected vs actual values for frequency using chi-square test which will help find if median value of owner-occupied homes in Boston follows a normal distribution and Shapiro-Wilk test to check it too.



Alpha = 0.5

Results and conclusion:

Based on the results of the chi-squared test and the Shapiro-Wilk normality test, we can conclude that there is evidence to suggest that the median value of owner-occupied homes in Boston does not follow a normal distribution. The p-value of the chi-squared test is less than 0.05, indicating that we reject the null hypothesis and accept the alternative hypothesis that the median value of owner-occupied homes in Boston does not follow a normal distribution. The p-value of the Shapiro-Wilk test is also less than 0.05, further supporting this conclusion. Therefore, we can conclude that the distribution of these median values differs significantly from a normal distribution.