DSA 8020 R Lab 2: Multiple Linear Regression I

Whitney

February 17, 2021

Contents

Housing Values in Suburbs of Boston 1

Load the dataset 1

Exploratory Data Analysis 2

Numerical summary 2

Graphical summary 2

Model Fitting 4

Simple Linear Regression 4

Multiple Linear Regression 5

Housing Values in Suburbs of Boston

The Boston housing data was collected in 1978, each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, MA.

Data Source: Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. **J. Environ. Economics and Management** 5, 81–102.

Load the dataset

Code:

```
library(MASS)
data(Boston)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

For the purposes of this lab, we will use only the following variables for conducting data analysis:

1. `medv`: median value of owner-occupied homes in \$1000*s*;
2. `lstat`: lower status of the population (percent);
3. `rm`: average number of rooms per dwelling;
4. `crim`: per capita crime rate by town

Code:

You can use the code below to extract these variables

```
vars <- c("medv", "lstat", "rm", "crim")
data <- Boston[, vars]
```

Exploratory Data Analysis

Numerical summary

1. Use `summary` commend to produce various numerical summmaries of each of the 4 variables under consideration

Code:

```
summary(data)
```

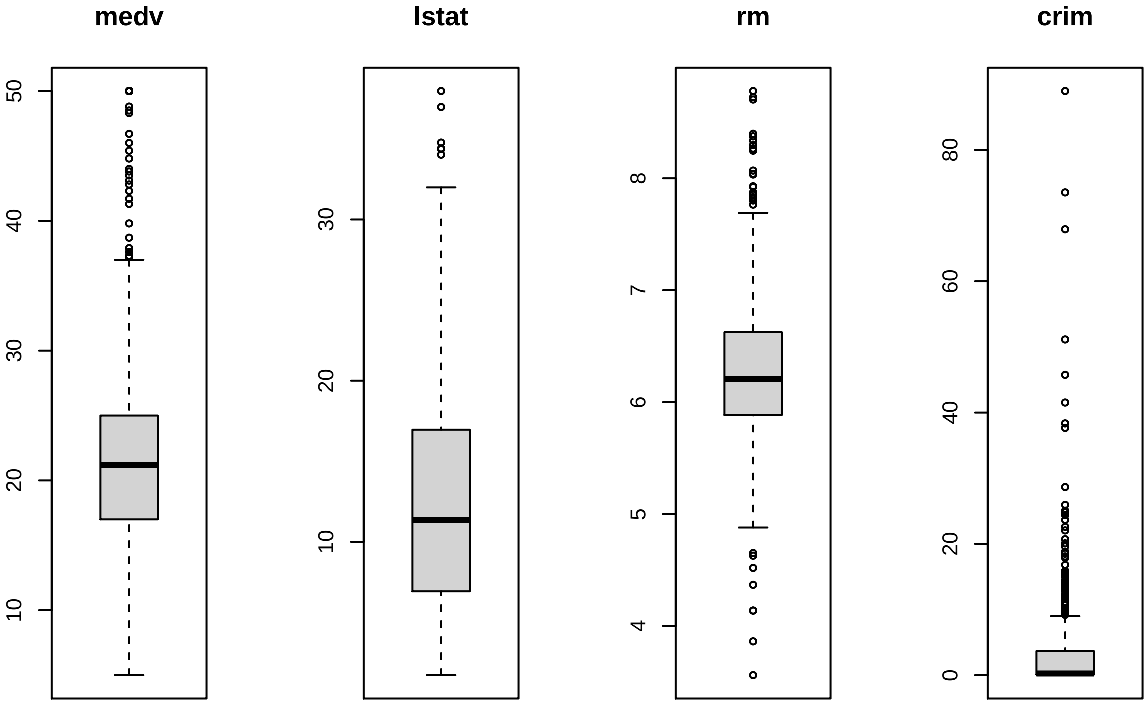
```
##       medv           lstat              rm             crim         
##  Min.   : 5.00   Min.   : 1.73   Min.   :3.561   Min.   : 0.00632  
##  1st Qu.:17.02   1st Qu.: 6.95   1st Qu.:5.886   1st Qu.: 0.08205  
##  Median :21.20   Median :11.36   Median :6.208   Median : 0.25651  
##  Mean   :22.53   Mean   :12.65   Mean   :6.285   Mean   : 3.61352  
##  3rd Qu.:25.00   3rd Qu.:16.95   3rd Qu.:6.623   3rd Qu.: 3.67708  
##  Max.   :50.00   Max.   :37.97   Max.   :8.780   Max.   :88.97620  
```

Graphical summary

2. Make a boxplot for each variable

Code:

```
par(mfrow = c(1, 4))
for (i in 1:4) boxplot(data[, i], main = vars[i])
```



3. Briefly discuss the shape of the distribution of each variable

Answer:

`medv`: the bulk distribution is approximately symmetric with many upper outliers

`lstat`: distribution appears skewed right with some upper outliers

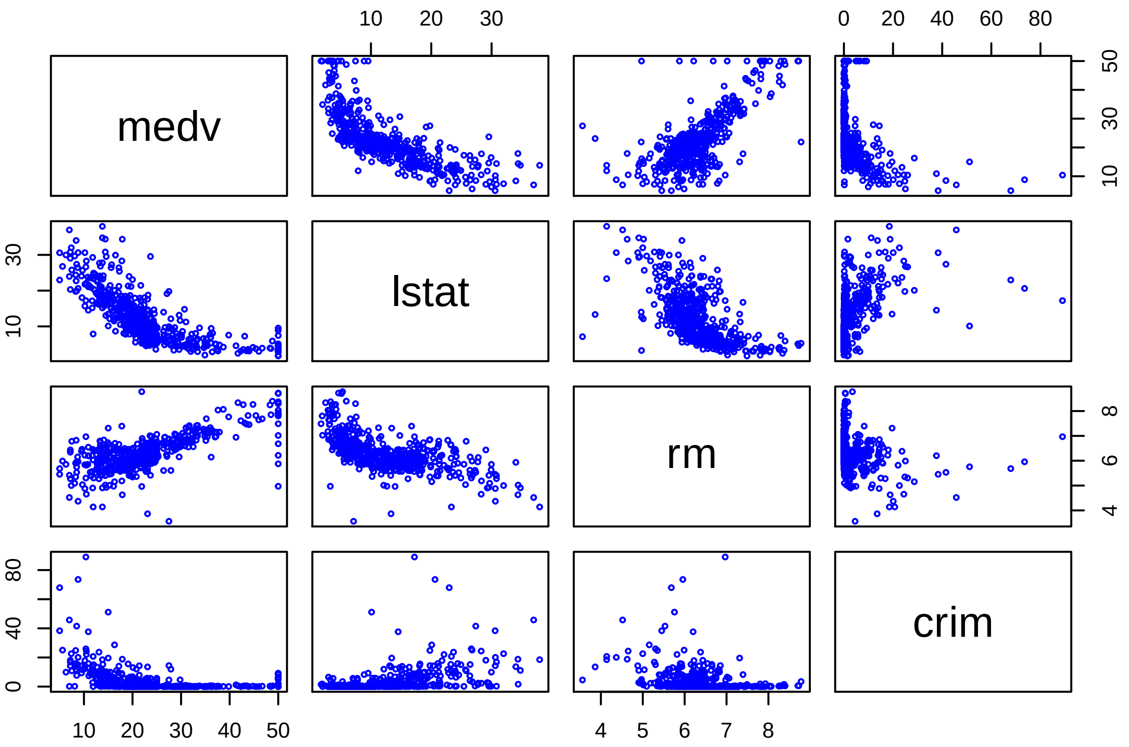
`rm`: the bulk distribution is approximately symmetric with many outliers from both directions

`crim`: istribution appears strongly skewed right with many upper outliers

4. Create a scatterplot matrix to explore the inter-dependence between these these variables

Code:

```r
pairs(data, cex = 0.5, col = "blue")
```



Model Fitting

Here we will use `medv` as the response and `lstat`, `rm`, `crim` as predictors.

Simple Linear Regression

5. Fit a simple linear regression

Code:

```r
lmfit <- lm(medv ~ lstat, data = data)
summary(lmfit)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

6. Write down the fitted linear regression equation.

Answer:

$\widehat{\texttt{medv}} = 34.55384 - 0.95005 \times \texttt{lstat}$

Multiple Linear Regression

7. Fit a multiple linear regression using all predictors

Code:

```r
lmfitFull <- lm(medv ~ ., data = data)
summary(lmfitFull)
```

```
##
## Call:
## lm(formula = medv ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.925  -3.567  -1.157   1.906  29.024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.56225    3.16602  -0.809  0.41873
## lstat       -0.57849    0.04767 -12.135  < 2e-16 ***
## rm           5.21695    0.44203  11.802  < 2e-16 ***
## crim        -0.10294    0.03202  -3.215  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.49 on 502 degrees of freedom
## Multiple R-squared:  0.6459, Adjusted R-squared:  0.6437
## F-statistic: 305.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

8. Write down the fitted linear regression equation

Answer:

$\widehat{\texttt{medv}} = -2.56225 - 0.57849 \times \texttt{lstat} + 5.21695 \times \texttt{rm} - 0.10294 \times \texttt{crim}$

9. Perform an overall F-test, state the hypotheses, test statistic, p-value, decision, and conclusion

Answer:

$H_0 : \beta_{\texttt{lstat}} = \beta_{\texttt{rm}} = \beta_{\texttt{crim}} = 0$ vs. H_a : at least one of the above regression coefficient $\neq 0$

F-statistic $= 305.2$, p-value $< 2.2 \times 10^{-16} \Rightarrow$ Reject H_0. We have sufficient evidence that at least one of $\beta_{\texttt{lstat}}, \beta_{\texttt{rm}}, \beta_{\texttt{crim}}$ is $\neq 0$.