DSA 8020 R Lab 2: Multiple Linear Regression I

Meredith Sliger

January 23, 2025

Contents

Housing Values in Suburbs of Boston

The Boston housing data was collected in 1978. Each of the 506 entries represents 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 $1000s;
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 the `summary` command to produce various numerical summaries of each of the 4 variables under consideration.

Code:

```
str(data)  # Structure of the data
```

```
## 'data.frame':    506 obs. of  4 variables:
##  $ medv : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
##  $ lstat: num  4.98 9.14 4.03 2.94 5.33 ...
##  $ rm   : num  6.58 6.42 7.18 7 7.15 ...
##  $ crim : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
```

```
summary(data)  # Summary statistics of the variables
```

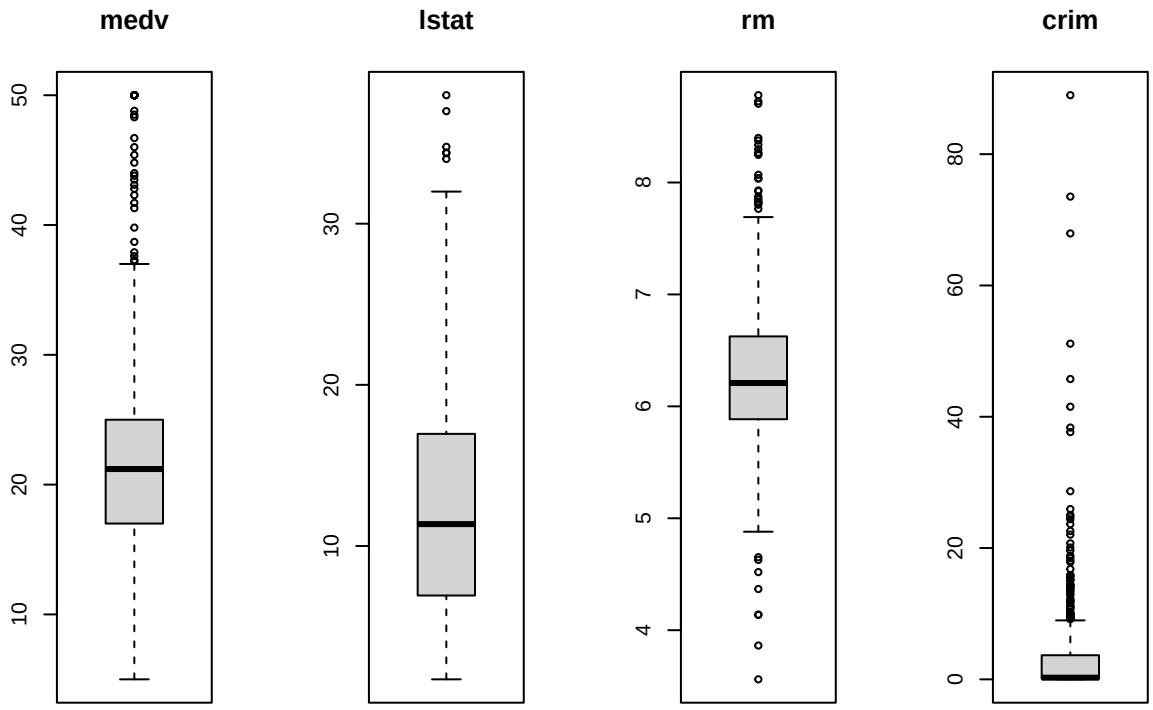
```
##       medv           lstat             rm             crim
##  Min.   : 5.00   Min.   : 1.73   Min.   :3.561   Min.   : 0.00632
##  1st Qu.:17.02   1st Qu.: 6.95   1st Qu.:5.886   1st Qu.: 0.08205
##  Median :21.20   Median :11.36   Median :6.208   Median : 0.25651
##  Mean   :22.53   Mean   :12.65   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))  # Arrange plots in a 1x4 grid
boxplot(data$medv, main = vars[1])
boxplot(data$lstat, main = vars[2])
boxplot(data$rm, main = vars[3])
boxplot(data$crim, main = vars[4])
```



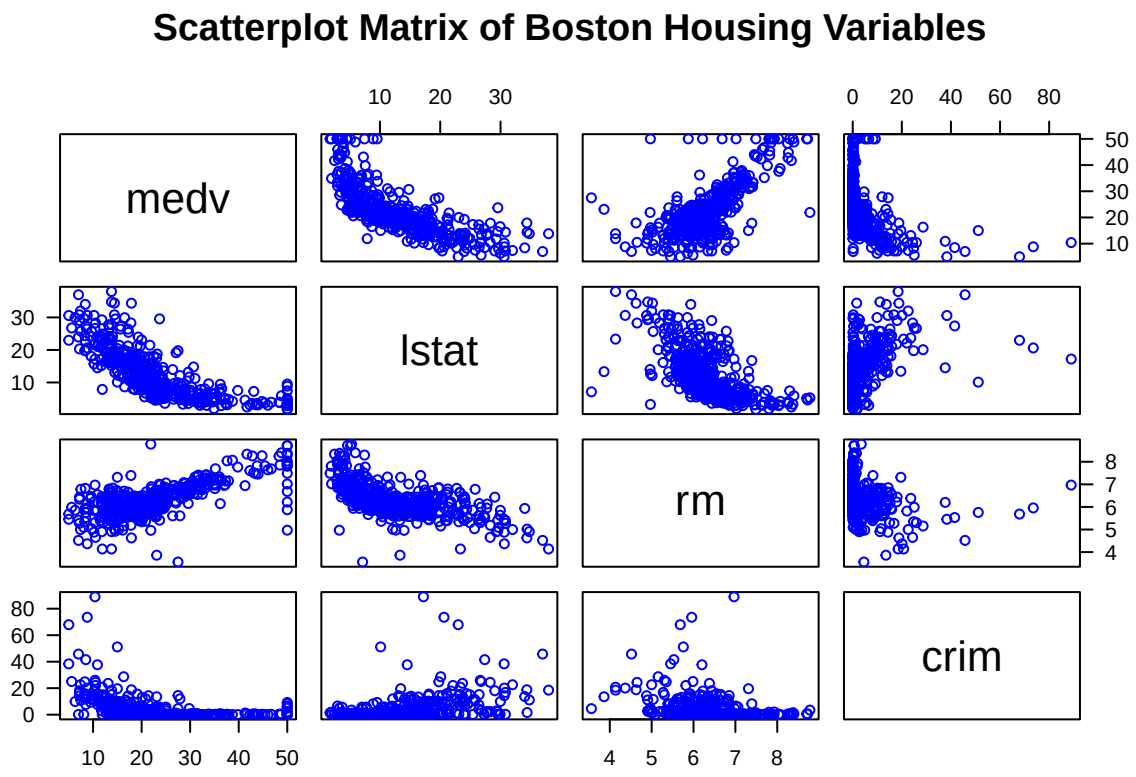
3. Briefly discuss the shape of the distribution of each variable

Answer: The distribution of the median home value **`medv`** appears fairly symmetric, with the median centered within the interquartile range and a few high-value outliers suggesting some expensive homes. In contrast, the lower status population **`lstat`** shows a right-skewed distribution, with most data concentrated at the lower end and several high-value outliers indicating areas with a significantly higher lower-status population. The average number of rooms per dwelling **`rm`** exhibits a slightly right-skewed distribution, with the median closer to the lower quartile and a few outliers representing houses with an unusually high number of rooms. The per capita crime rate **`crim`** is highly right-skewed, with most values clustered near the lower end and a long tail extending to the right, indicating that while crime rates are low in most areas, a few neighborhoods experience significantly higher crime levels.

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

Code:

```
pairs(data,
      cex = 0.95,        # Adjust point size
      col = "blue",      # Set point color to blue
      las = 1,           # Make axis labels horizontal
      main = "Scatterplot Matrix of Boston Housing Variables")
```



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.

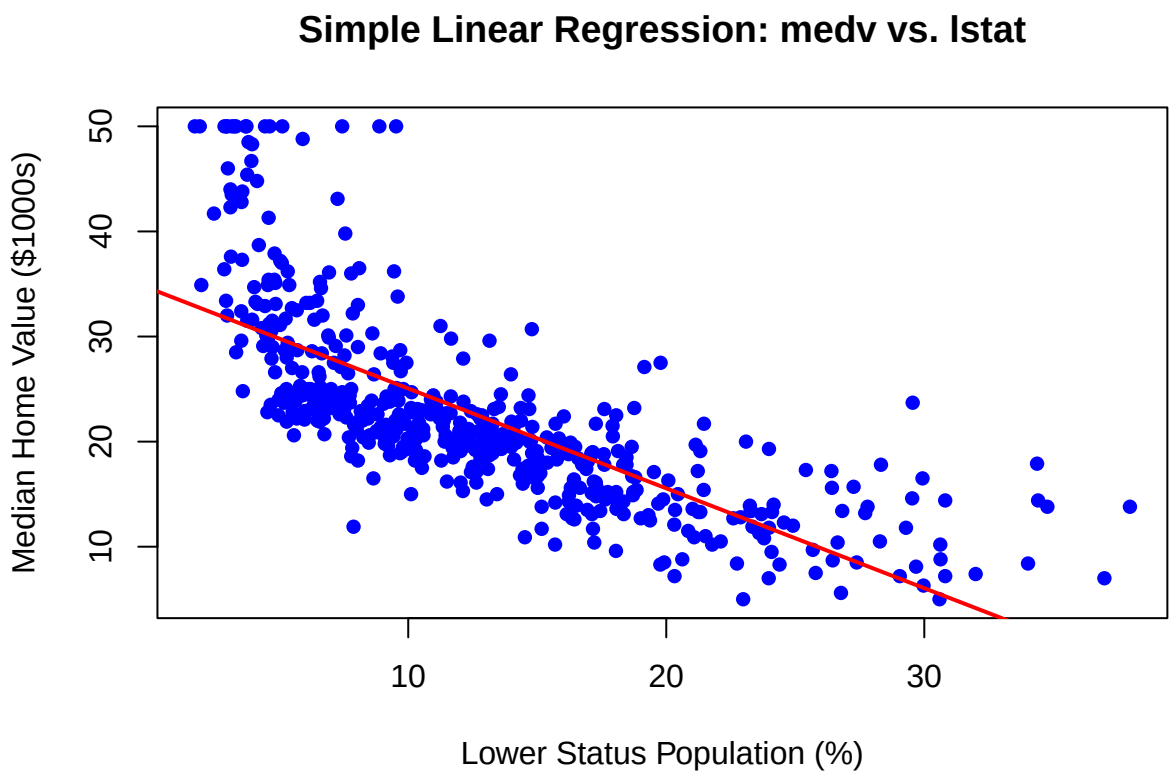
Here we use *lstat* as the predictor as it has the highest correlation with *medv*.

Code:

```
slr <- lm(medv ~ lstat, data = data)

summary(slr)
```

```
##
## 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
```

```r
plot(data$lstat, data$medv,
     xlab = "Lower Status Population (%)",
     ylab = "Median Home Value ($1000s)",
     main = "Simple Linear Regression: medv vs. lstat",
     col = "blue", pch = 16)

abline(slr, col = "red", lwd = "2") # improving visual clarity of the regression line
```



6. Write down the fitted linear regression equation.

Answer: medv = 34.55384 - 0.95005 * lstat

Multiple Linear Regression

7. Fit a multiple linear regression using all predictors

Code:

```
mlr <- lm(medv ~ lstat + rm + crim, data = data)

summary(mlr)
```

```
##
## Call:
## lm(formula = medv ~ lstat + rm + crim, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.925  -3.566  -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
```

```
# Load necessary libraries
library(fields)
```

```
## Warning: package 'fields' was built under R version 4.4.2
```

```
## Loading required package: spam
```

```
## Warning: package 'spam' was built under R version 4.4.2
```

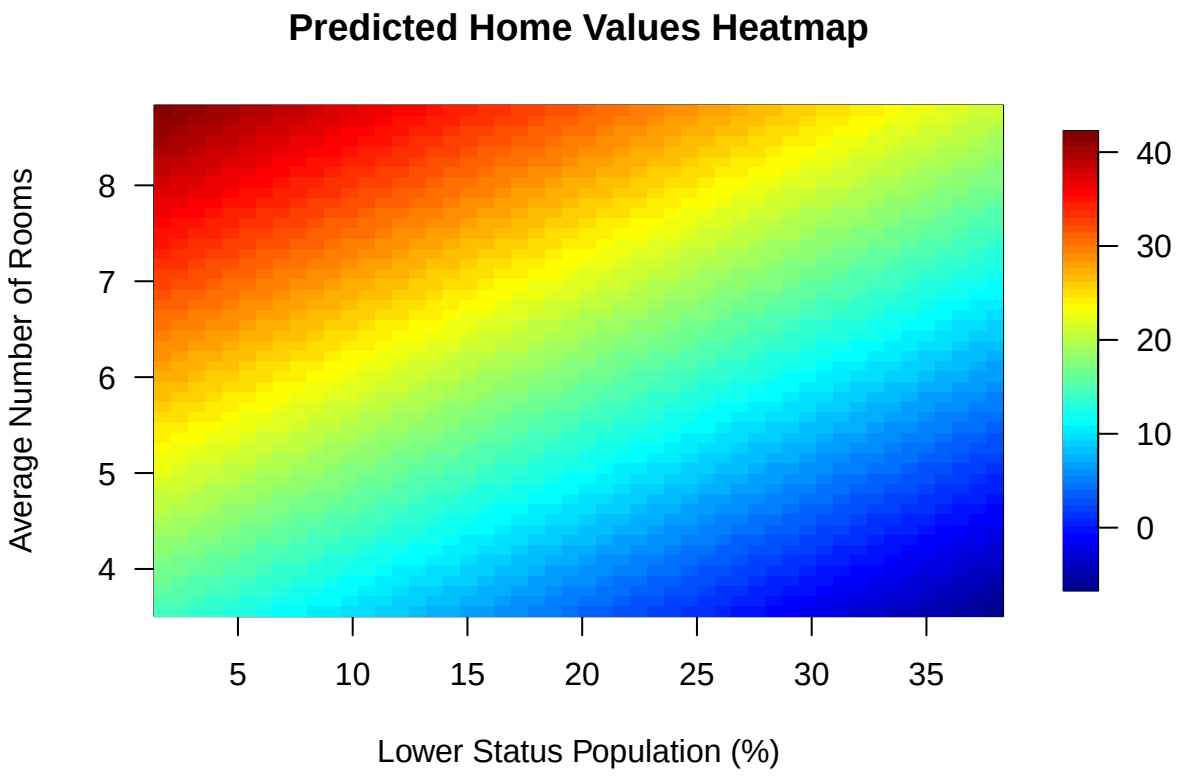
```
## Spam version 2.11-1 (2025-01-20) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.
```

```
##
## Attaching package: 'spam'

## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve

## Loading required package: viridisLite

##
## Try help(fields) to get started.
```

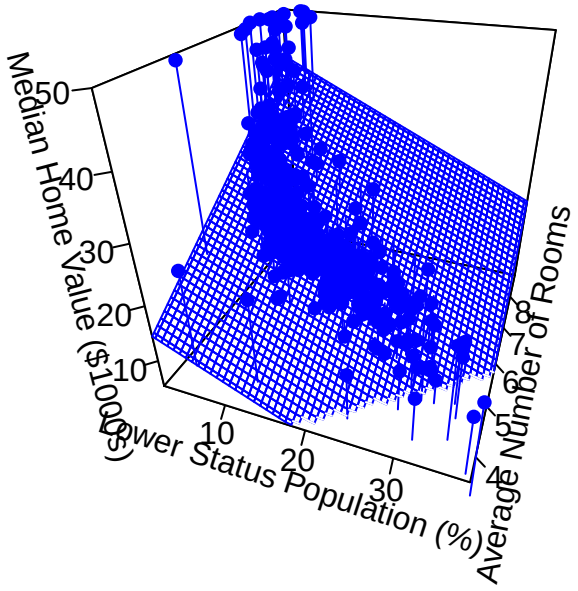
```r
library(plot3D)
```

```
## Warning: package 'plot3D' was built under R version 4.4.2
```

```r
# Create grids for lstat and rm
lstat_grid <- seq(min(data$lstat), max(data$lstat), length.out = 50)
rm_grid <- seq(min(data$rm), max(data$rm), length.out = 50)
temp <- expand.grid(lstat_grid, rm_grid)

# Prepare new data for prediction, fixing crim at its mean value
x_new <- data.frame(lstat = temp$Var1, rm = temp$Var2, crim = mean(data$crim))

# Fit the multiple linear regression model
mlr <- lm(medv ~ lstat + rm + crim, data = data)

# Predict median home values based on the grid
y_pred <- matrix(predict(mlr, newdata = x_new), nrow = length(lstat_grid))

# Generate a 2D heatmap of predicted values
image.plot(lstat_grid, rm_grid, y_pred, las = 1,
           xlab = "Lower Status Population (%)",
           ylab = "Average Number of Rooms",
           main = "Predicted Home Values Heatmap")
```




```r
# Get predicted fitted values for scatter plot
fitpoints <- predict(mlr)

# Create 3D scatter plot with regression plane
scatter3D(x = data$lstat,
          y = data$rm,
          z = data$medv,
          pch = 16, cex = 1, colkey = FALSE, col = "blue",
          xlab = "Lower Status Population (%)",
          ylab = "Average Number of Rooms",
          zlab = "Median Home Value ($1000s)",
          theta = 20, phi = 30, ticktype = "detailed",
          surf = list(x = lstat_grid, y = rm_grid, z = y_pred, facets = NA, fit = fitpoints))
```



8. Write down the fitted linear regression equation

Answer: medv = -2.56225 - 0.57849 * lstat + 5.21695 * rm - 0.10294 * crim

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

Code:

```
anova(mlr)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## lstat       1 23243.9 23243.9 771.320 < 2e-16 ***
## rm          1  4033.1  4033.1 133.832 < 2e-16 ***
## crim        1   311.4   311.4  10.334 0.00139 **
## Residuals 502 15127.9    30.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

H_0 : lstat = rm = crim = 0

vs.

H_a : at least one of the above regression coefficient $\neq 0$

F-statistic = 305.2, p-value = $<$ 2.2e-16;

Decision: Reject H_0

Conclusion: The null hypothesis (H_0) assumes that all regression coefficients are equal to zero, meaning that none of the predictors (`lstat`, `rm`, `crim`) have an impact on medv. The alternative hypothesis (H_a) suggests that at least one predictor significantly affects the median home value.

There is sufficient evidence that at least one of the predictors (`lstat`, `rm`, `crim`) is $\neq 0$. This means that at least one of the predictors helps explain the median home value responses.