Multiple Linear Regression (Inference and Prediction) - Lab

Blake Pappas

December 17, 2023

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)
```

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:

The code below can be used to extract these variables.

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

Exploratory Data Analysis

Numerical Summary

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

Code:

```r
summary(data)
```

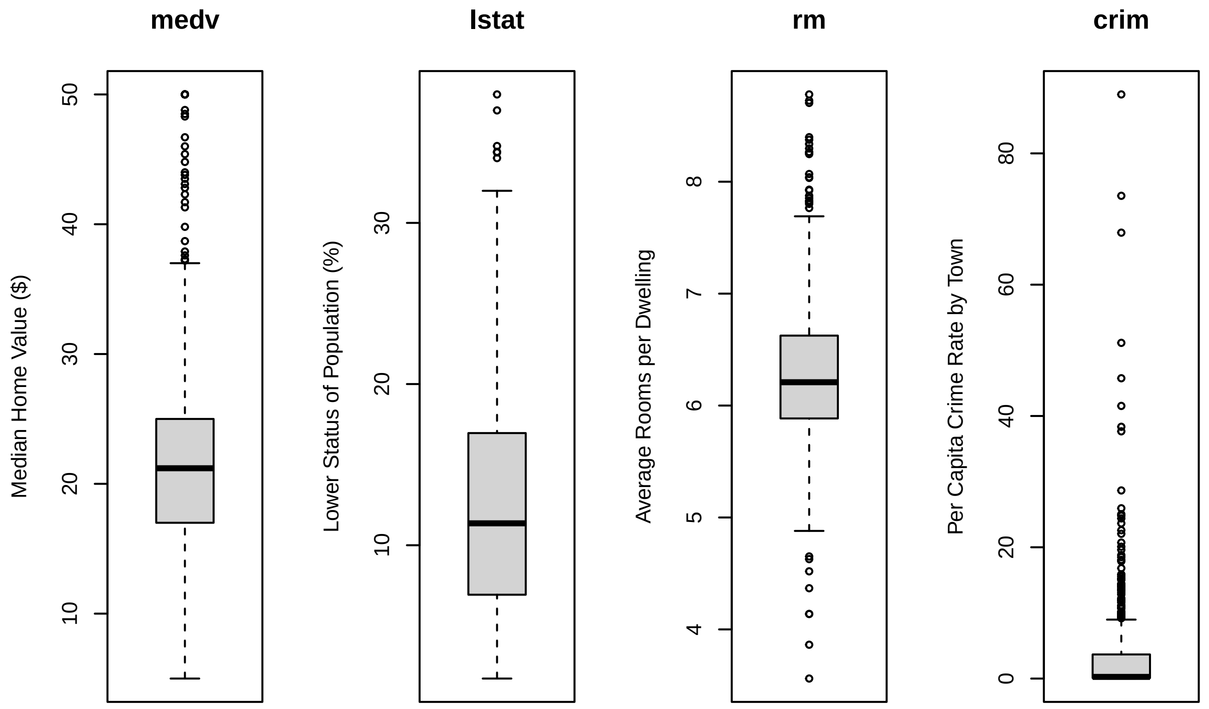
```
##       medv           lstat              rm             crim
##  Min.   : 5.00   Min.   : 1.73   Min.   :3.561   Min.   : 0.00632
##  1st Qu.:17.02   1st Qu.: 6.95   1st Qu.:5.886   1st Qu.: 0.08205
##  Median :21.20   Median :11.36   Median :6.208   Median : 0.25651
##  Mean   :22.53   Mean   :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:

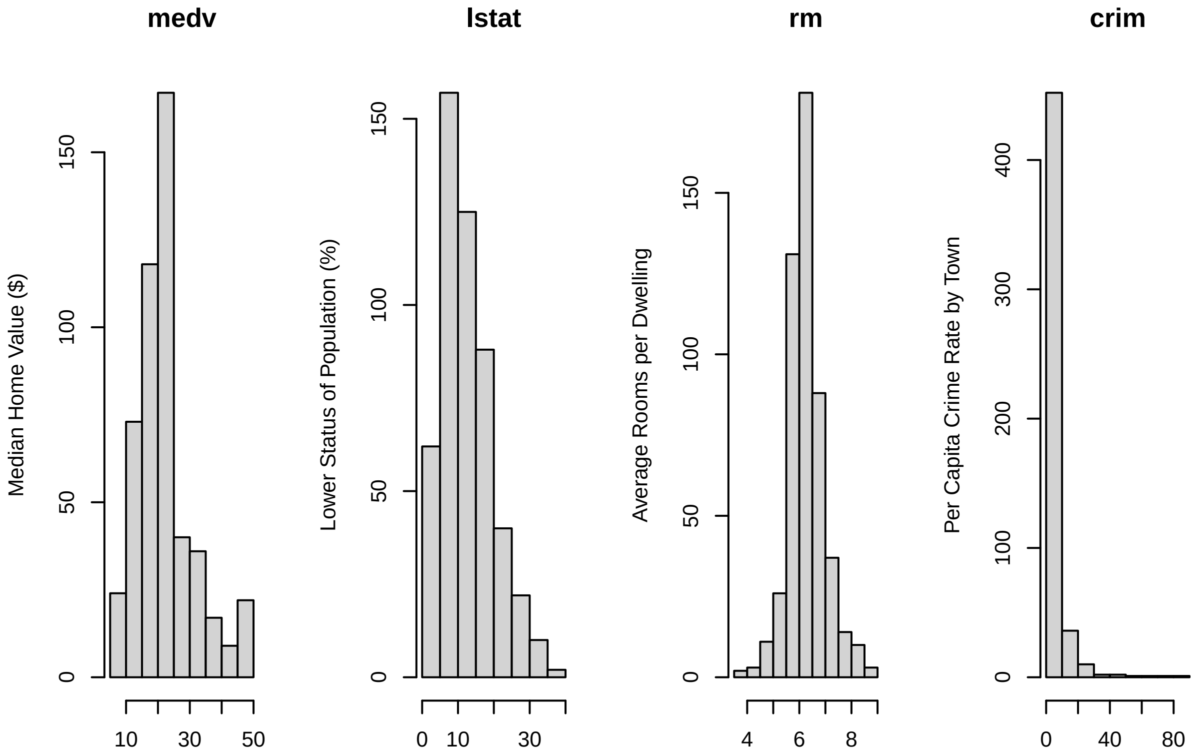
```r
par(mfrow = c(1, 4))
boxplot(data$medv, main = 'medv', ylab = 'Median Home Value ($)')
boxplot(data$lstat, main = 'lstat', ylab = 'Lower Status of Population (%)')
boxplot(data$rm, main = 'rm', ylab = 'Average Rooms per Dwelling')
boxplot(data$crim, main = 'crim', ylab = 'Per Capita Crime Rate by Town')
```



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

Code:

```
par(mfrow = c(1, 4))
hist(data$medv, main = 'medv', xlab = '', ylab = 'Median Home Value ($)')
hist(data$lstat, main = 'lstat', xlab = '', ylab = 'Lower Status of Population (%)')
hist(data$rm, main = 'rm', xlab = '', ylab = 'Average Rooms per Dwelling')
hist(data$crim, main = 'crim', xlab = '', ylab = 'Per Capita Crime Rate by Town')
```

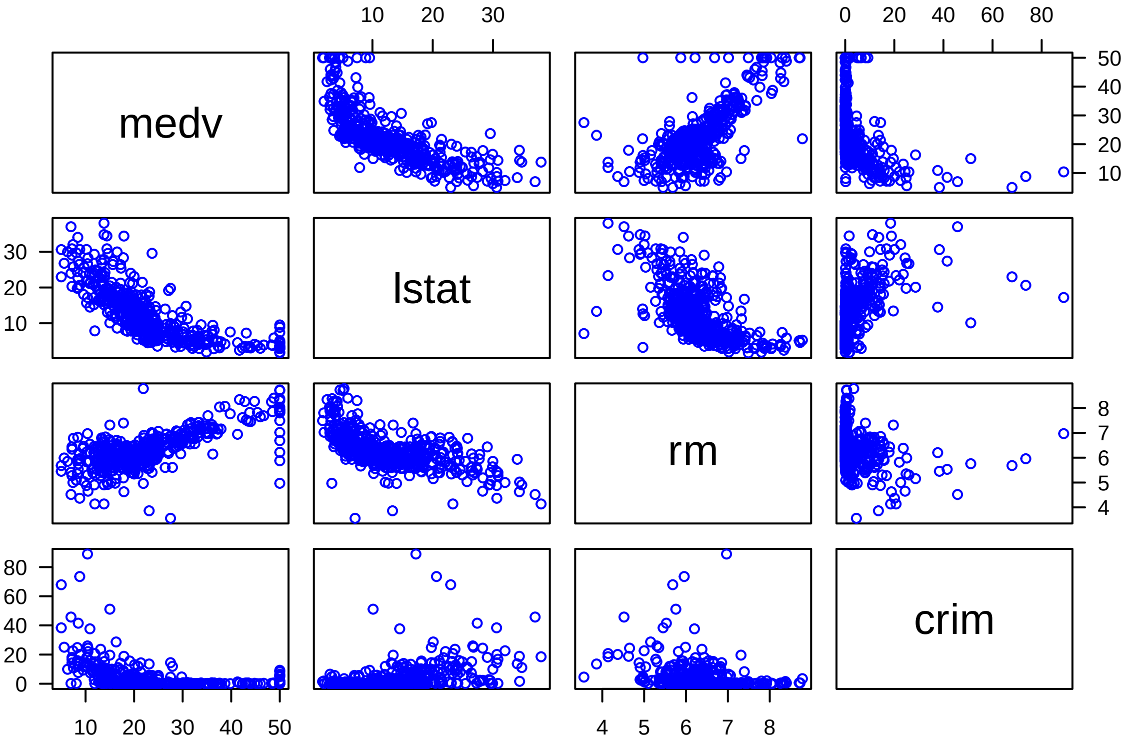


Answer: The variable `mdev` is positively skewed, asymmetric, and unimodal. The variable `lstat` is positively skewed, asymmetric, and unimodal. Although it may appear normally distributed, the variable `rm` is slightly positively skewed, asymmmetric, and unimodal. The variable `crim` is positively skewed, asymmetric and unimodal.

4. Create a scatterplot matrix to explore the interdependence between these variables.

Code:

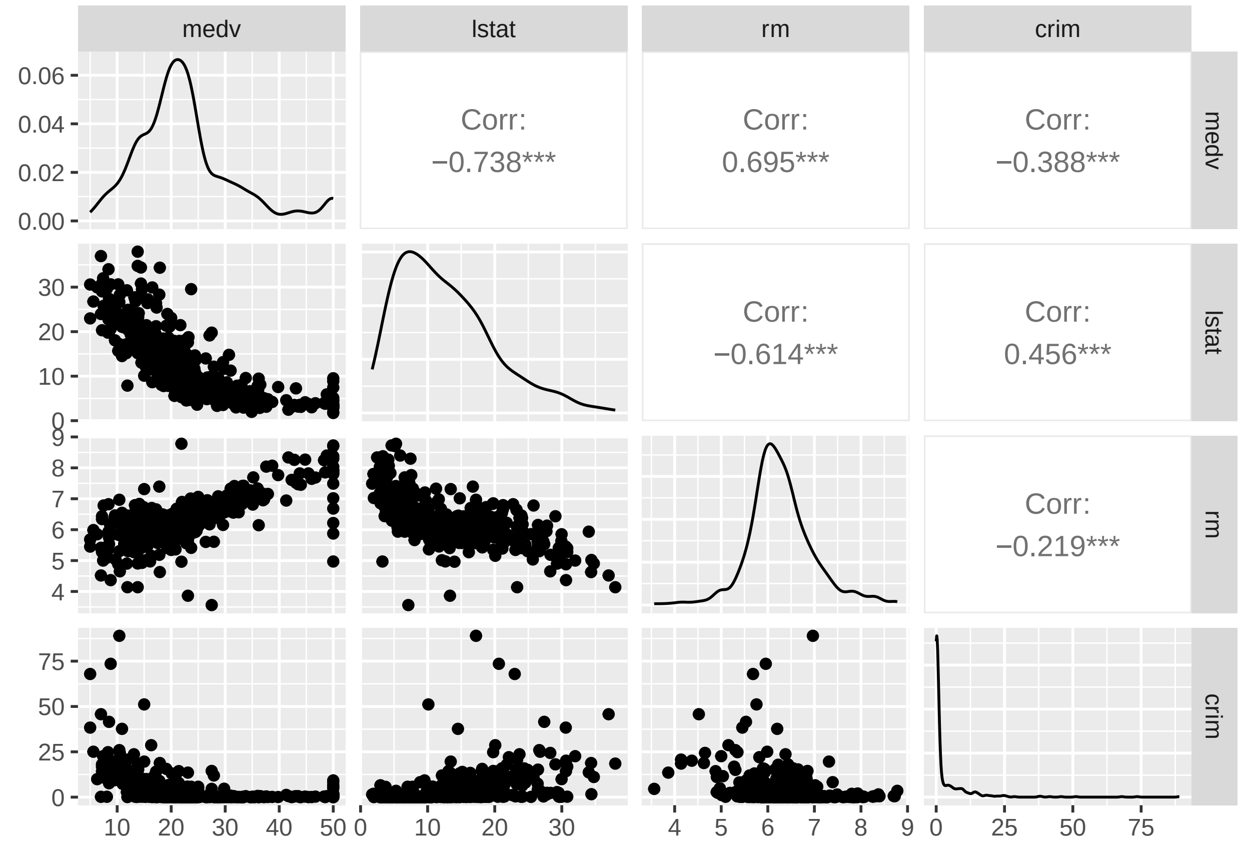
```
pairs(data, cex = 0.95, col = "blue", las = 1)
library(ggplot2)
```



```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(data)
```



Model Fitting

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

Simple Linear Regression

5. Fit a simple linear regression.

Code:

```
# Simple Linear Regression Using mdev as the Response and rm as the Predictor
slr <- lm(medv ~ rm, data = data) # This will be the reduce model for the F-test
summary(slr)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825 
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

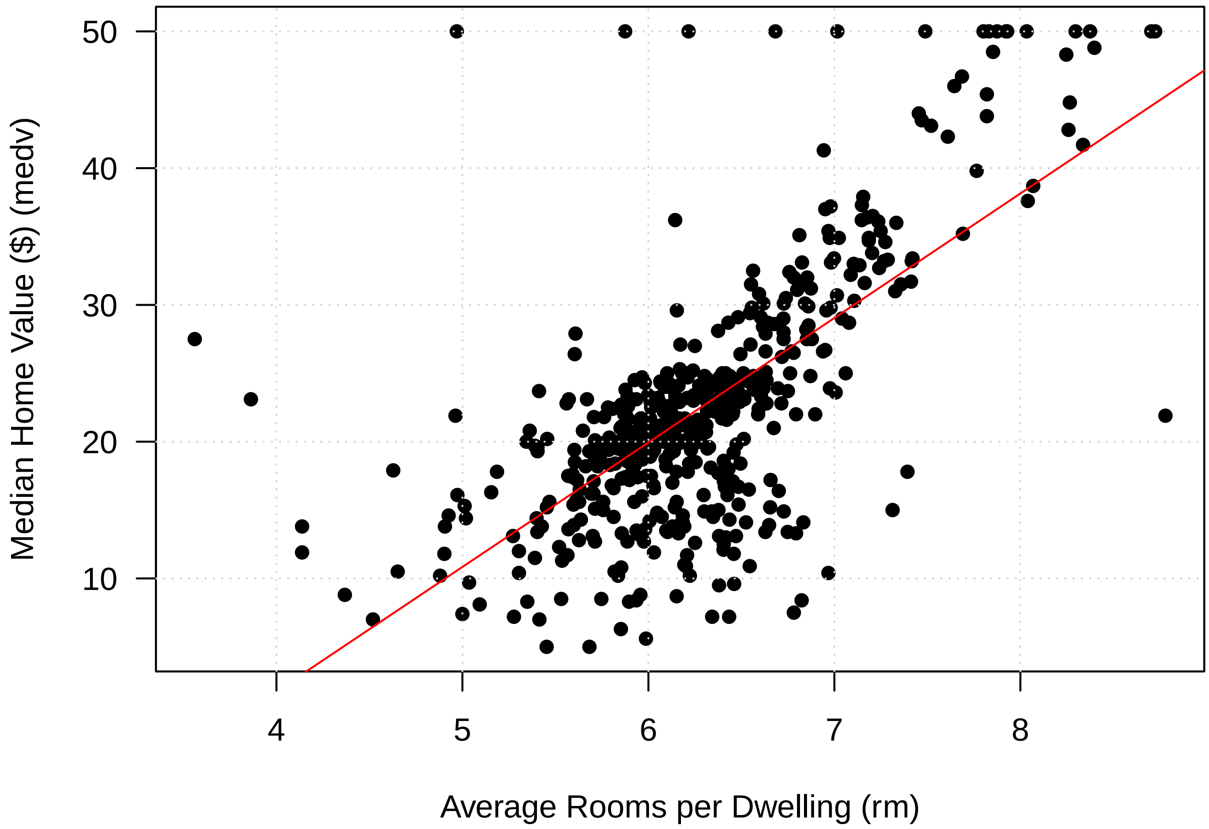
```
y <- data$medv; x <- data$rm
y_diff <- y - mean(y)
x_diff <- x - mean(x)
beta_1 <- sum(y_diff * x_diff) / sum((x_diff)^2)
beta_1
```

```
## [1] 9.102109
```

```
beta_0 <- mean(y) - mean(x) * beta_1
beta_0
```

```
## [1] -34.67062
```

```
par(las = 1, mar = c(4.1, 4.1, 1.1, 1.1))
plot(x, y, pch = 16, xlab = "Average Rooms per Dwelling (rm)", ylab = "Median Home Value ($) (medv)")
grid()
abline(a = beta_0, b = beta_1, col = "red")
```



6. Write down the fitted linear regression equation.

Answer: `medv` **=** **-34.671 + 9.102 x** `rm` **+** ϵ**, where Y =** `medv`**, X =** `rm`**,** $\hat{\beta}_0$ **= -34.671,** $\hat{\beta}_1$ **= 9.102, and** ϵ **= stochastic error.**

Multiple Linear Regression

7. Fit a multiple linear regression using all predictors.

Code:

```
mlr <- lm(medv ~ ., data = data) # This will be the full model for the F-test
summary(mlr)
```

```
##
## Call:
## lm(formula = medv ~ ., 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
```

8. Write down the fitted linear regression equation.

Answer: `medv` = -2.56225 - 0.57849 x `lstat` + 5.21695 x `rm` - 0.10294 x `crim` + ϵ

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

Code:

```
anova(slr, mlr)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ rm
## Model 2: medv ~ lstat + rm + crim
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    504 22062
## 2    502 15128  2      6934 115.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Answer: The null hypothesis was H0 : β**lstat =** β**rm =** β**crim = 0. The alternative hypothesis was HA : at least one of the three coefficients** $\neq$ **0. The test statistic was 115.05, the p-value was 2.2e-16, and conclusion was to reject the null hypothesis.**