Multiple Linear Regression on Boston Housing Dataset

2025-03-10

1 Introduction

Introduce the model, data background, related knowledge, and R packages that you have used and result you have gotten in this essay.

2 Data description

```r
housing.df <- read.csv("boston_house_prices.csv")

nrow(housing.df) # number of rows
```

```
## [1] 506
```

```r
sum(is.na(housing.df)) # missing data
```

```
## [1] 0
```

2.1 Dataset

This dataset contains information about 506 neighborhoods in Boston, collected by the U.S Census Service in 1970 census. There are 506 records and 13 variables in the dataset. From an initial analysis, there were no missing data points.

2.2 Variables

- crim: per capita crime rate by town
- zn: proportion of residential land zoned for lots over 25,000 sq.ft
- indus: proportion of non-retail business acres per town
- chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- nox: nitric oxides concentration (parts per 10 million)
- rm: average number of rooms per dwelling
- age: proportion of owner-occupied units built prior to 1940
- dis: weighted distances to five Boston employment centres
- rad: index of accessibility to radial highways
- tax: full-value property-tax rate per USD 10,000
- ptratio: pupil-teacher ratio by town
- lstat: percentage of lower status of the population
- medv: median value of owner-occupied homes in USD 1000's

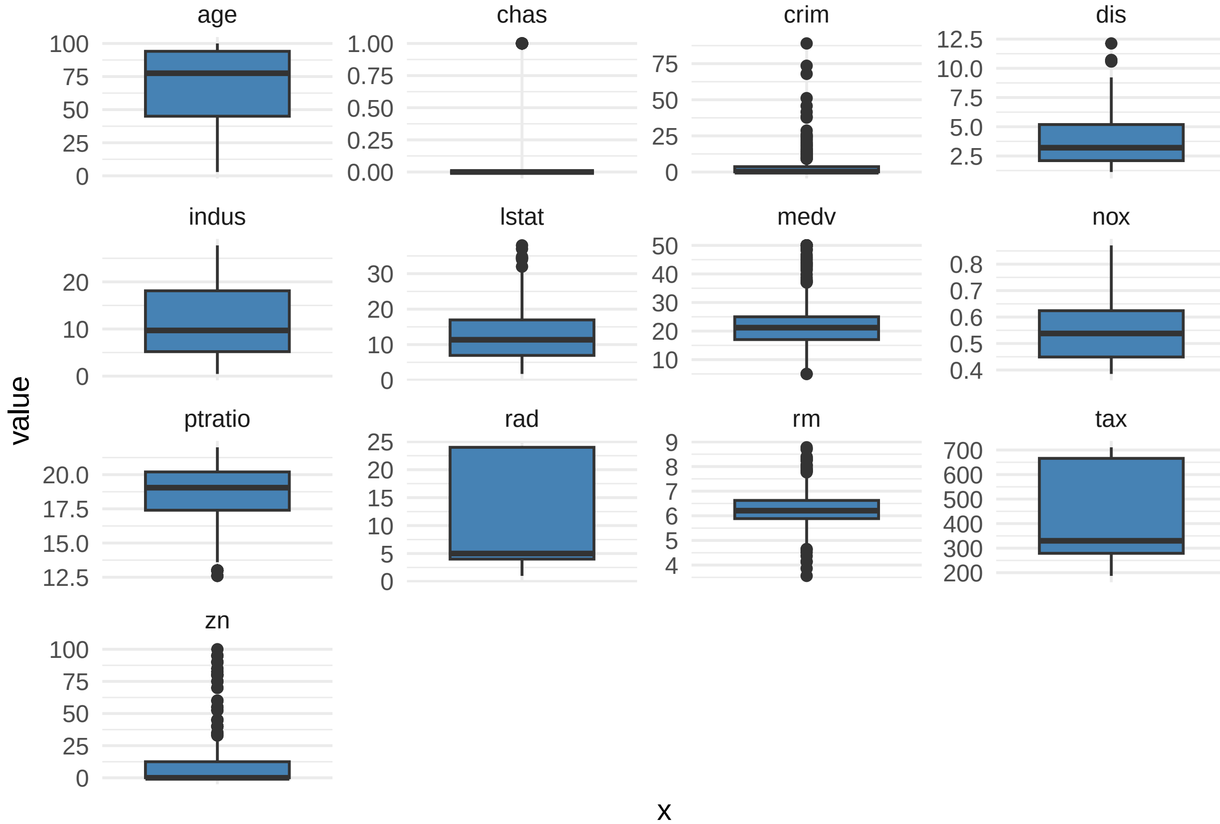
2.3 Outlier Detection

```
summary(housing.df)
```

```
##       crim                zn             indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm             age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          lstat
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97
##       medv
##  Min.   : 5.00
##  1st Qu.:17.02
##  Median :21.20
##  Mean   :22.53
##  3rd Qu.:25.00
##  Max.   :50.00
```

```
housing_box <- housing.df %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value")

ggplot(housing_box, aes(x = "", y = value)) +
  geom_boxplot(fill = "steelblue") +
  theme_minimal() +
  facet_wrap(~ variable, scales = "free_y") +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```




```r
head(housing.df, 5)
```

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

3 Analysis

You should cut your data set into train set and test set and tell us the data size for train set and data size for test set. You must use the diagonal plot to verify the assumptions, and show us the coding for computation, Pictures, tables, and Interpretation.

```r
set.seed(123)

split <- 0.7

trainIndex <- createDataPartition(housing.df$medv, p = split)
trainIndex <- unlist(trainIndex)

train <- housing.df[trainIndex, ]
test <- housing.df[-trainIndex, ]
```

```r
nrow(train)
```

```
## [1] 356
```

```r
nrow(test)
```

```
## [1] 150
```

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

```
##
## Call:
## lm(formula = medv ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9718  -2.7582  -0.5824   2.1372  24.6614
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.708122   5.520024   7.012 1.25e-11 ***
## crim         -0.104485   0.038876  -2.688 0.007546 **
## zn            0.030259   0.016408   1.844 0.066027 .
## indus        -0.047212   0.071656  -0.659 0.510418
## chas          3.017559   1.008072   2.993 0.002959 **
## nox         -17.124116   4.508089  -3.799 0.000172 ***
## rm            3.811233   0.460112   8.283 2.72e-15 ***
## age           0.003945   0.015726   0.251 0.802086
## dis          -1.290967   0.233049  -5.539 6.05e-08 ***
## rad           0.257805   0.074778   3.448 0.000636 ***
## tax          -0.011159   0.004196  -2.660 0.008191 **
## ptratio      -0.926238   0.153703  -6.026 4.33e-09 ***
## lstat        -0.513649   0.061922  -8.295 2.51e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.662 on 343 degrees of freedom
## Multiple R-squared:  0.7419, Adjusted R-squared:  0.7329
## F-statistic: 82.16 on 12 and 343 DF,  p-value: < 2.2e-16
```

```r
model_significant <- lm(medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + lstat, data =
summary(model_significant)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
##     tax + ptratio + lstat, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -15.0279  -2.8343  -0.5875   2.0912  24.7694
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.846064   5.451581   7.126 6.09e-12 ***
## crim         -0.104102   0.038775  -2.685 0.007609 **
## zn            0.031257   0.016237   1.925 0.055042 .
## chas          2.950082   0.998579   2.954 0.003349 **
## nox         -17.656374   4.128291  -4.277 2.46e-05 ***
## rm            3.869482   0.443515   8.725  < 2e-16 ***
## dis          -1.279680   0.217777  -5.876 9.90e-09 ***
## rad           0.268497   0.071958   3.731 0.000223 ***
## tax          -0.012304   0.003805  -3.234 0.001339 **
## ptratio      -0.936412   0.150593  -6.218 1.45e-09 ***
## lstat        -0.510371   0.056674  -9.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.651 on 345 degrees of freedom
## Multiple R-squared:  0.7415, Adjusted R-squared:  0.734
## F-statistic: 98.97 on 10 and 345 DF,  p-value: < 2.2e-16
```

4 Model Evaluation and Prediction

You should base on the training set you should hav gotten a model in (3), then you must use subset selection and all necessary test to verify the final model that you get is the best model (Model assessment and model accuracy) and use it to make a prediction

```r
predictions <- predict(model_significant, newdata = test)
actual <- test$medv

mae <- mean(abs(predictions - actual)) # MAE
mse <- mean((predictions - actual)^2) # MSE
rmse <- sqrt(mse) # RMSE

mae
```

```
## [1] 3.309707
```

```r
mse
```

```
## [1] 26.35976
```

```r
rmse
```

```
## [1] 5.134176
```

5 Conclusion

The summary your all your work and results in this part and point out positive side of your model, negative side of your model and possible future work or any factors that affect your model accuracy.

6 References

- https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html
- https://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/