SimpleLinearRegression.R

vijaykalmath

2022-01-04

```r
# Convert R script to RmarkDown ->  Cmd + Shift + K
# Simple Linear Regression - ISLR Lab Work

library(MASS)
library(ISLR)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```r
library(ggcorrplot)


# Boston Data - Housing Data in suburbs of Boston - 500 Observations and 14 Variables

colnames(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

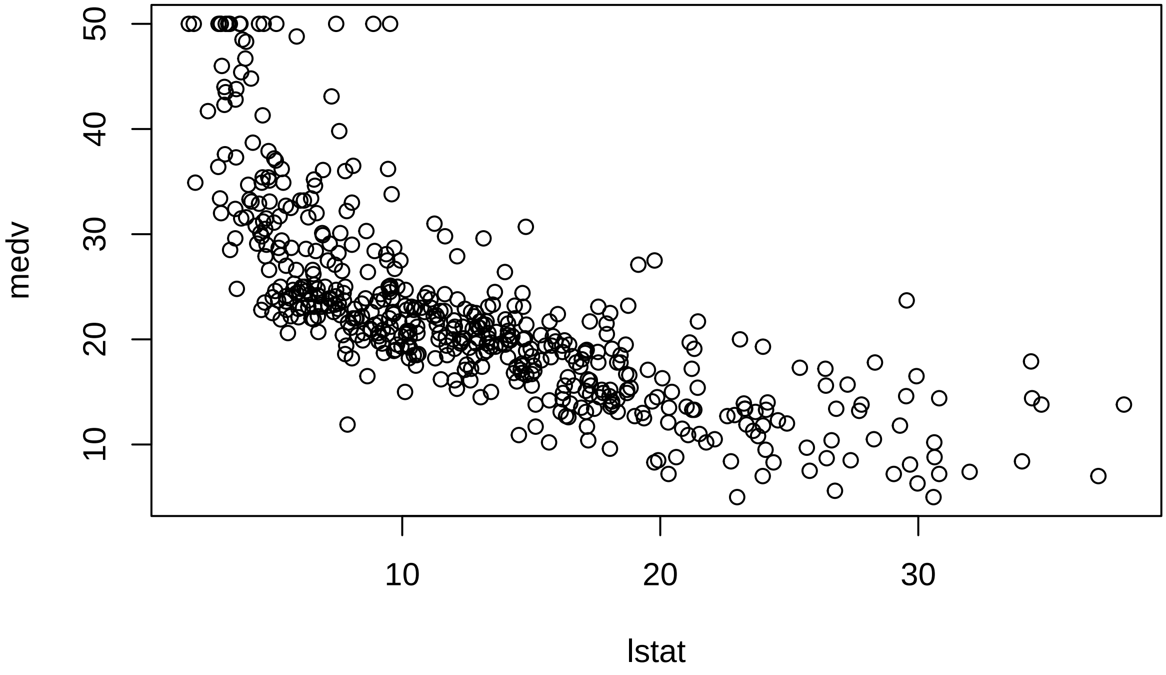
```r
str(Boston)
```

```
## ’data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

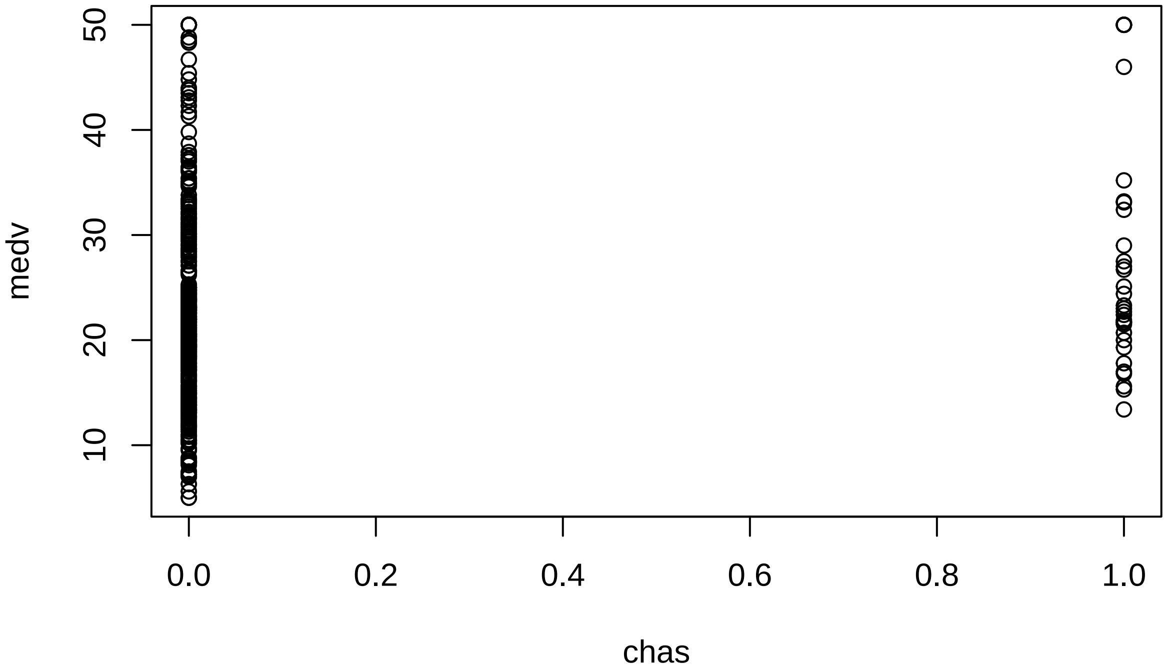
```
# All are Numerical Values.

# medv is the Y - Variable that we need to fit the rest of the variables to.

plot(medv~lstat,Boston)
```

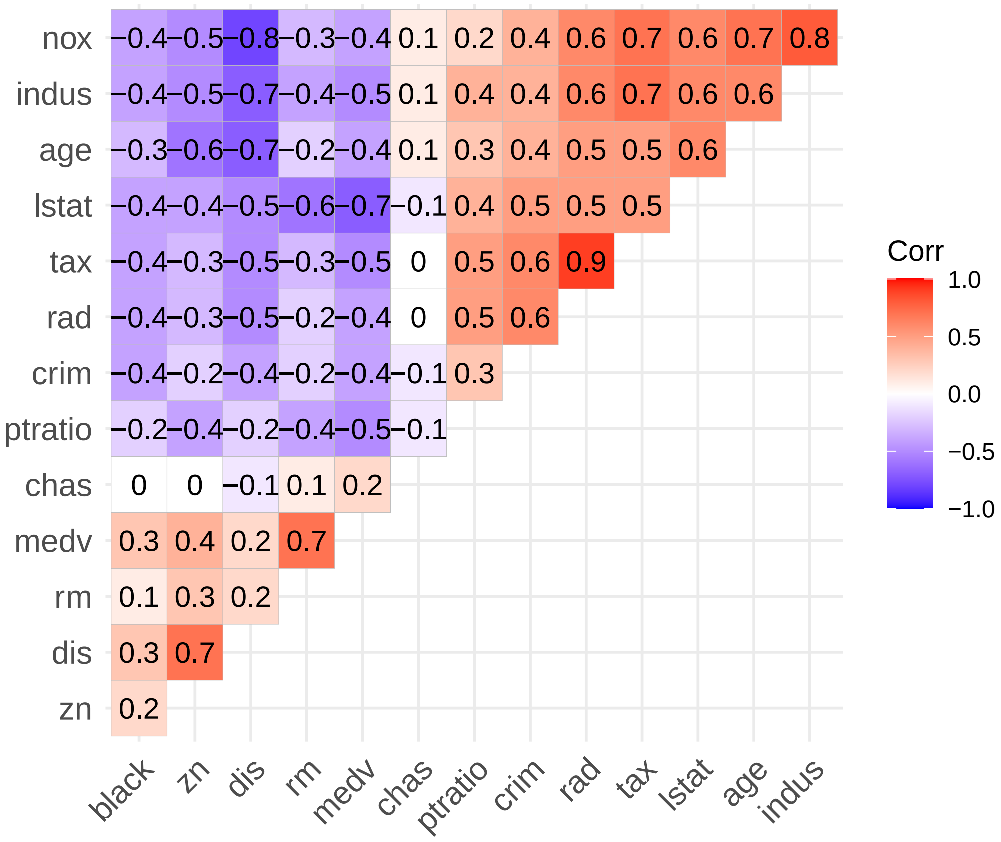


```
plot(medv~chas,Boston)
```



```
# Plotting Correlation Matrix

corr <- round(cor(Boston), 1)
ggcorrplot(corr, hc.order = TRUE, type = "upper",lab=TRUE)
```



```
# Tax and Rad seem to be highly correlated -> 0.9
# dis and Zn seem to be highly correlated as well -> 0.7

lm.fit  = lm(medv~lstat,data = Boston)
lm.fit
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Coefficients:
## (Intercept)        lstat
##       34.55        -0.95
```

```
# Summary of lm.fit

summary(lm.fit)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## 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
```

```
# Parts in lm.fit

names(lm.fit)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
# Coeefficient of lm.fit
coef(lm.fit)
```

```
## (Intercept)       lstat
##  34.5538409  -0.9500494
```

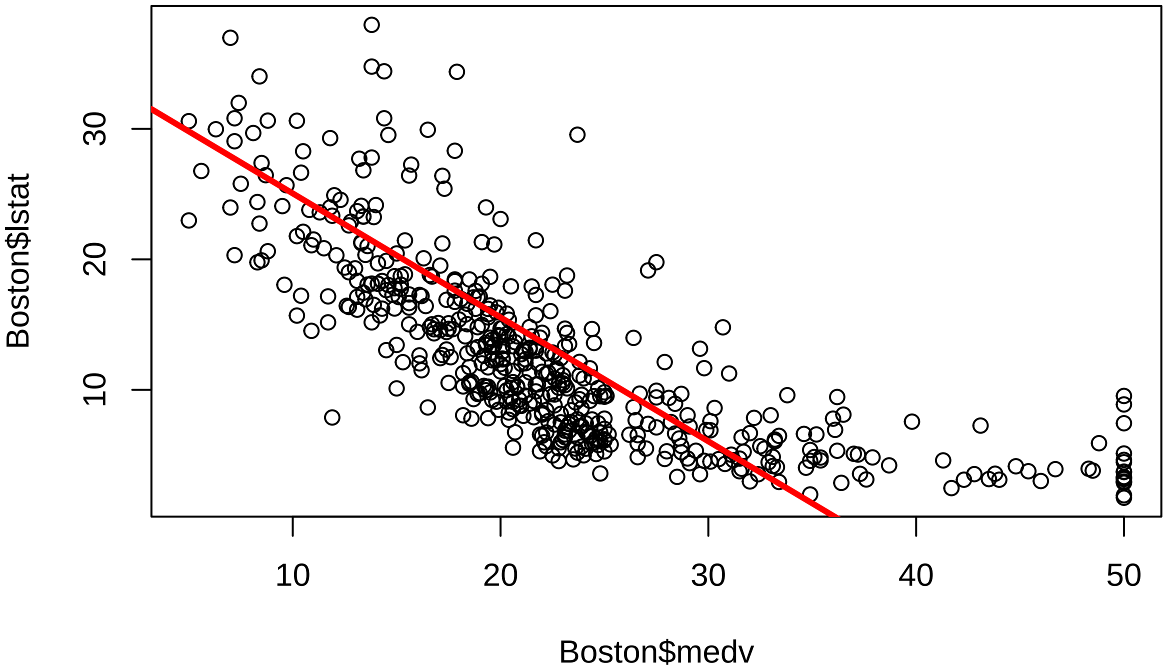
```
# Confidence Interval
confint(lm.fit)
```

```
##                 2.5 %     97.5 %
## (Intercept) 33.448457 35.6592247
## lstat       -1.026148 -0.8739505
```

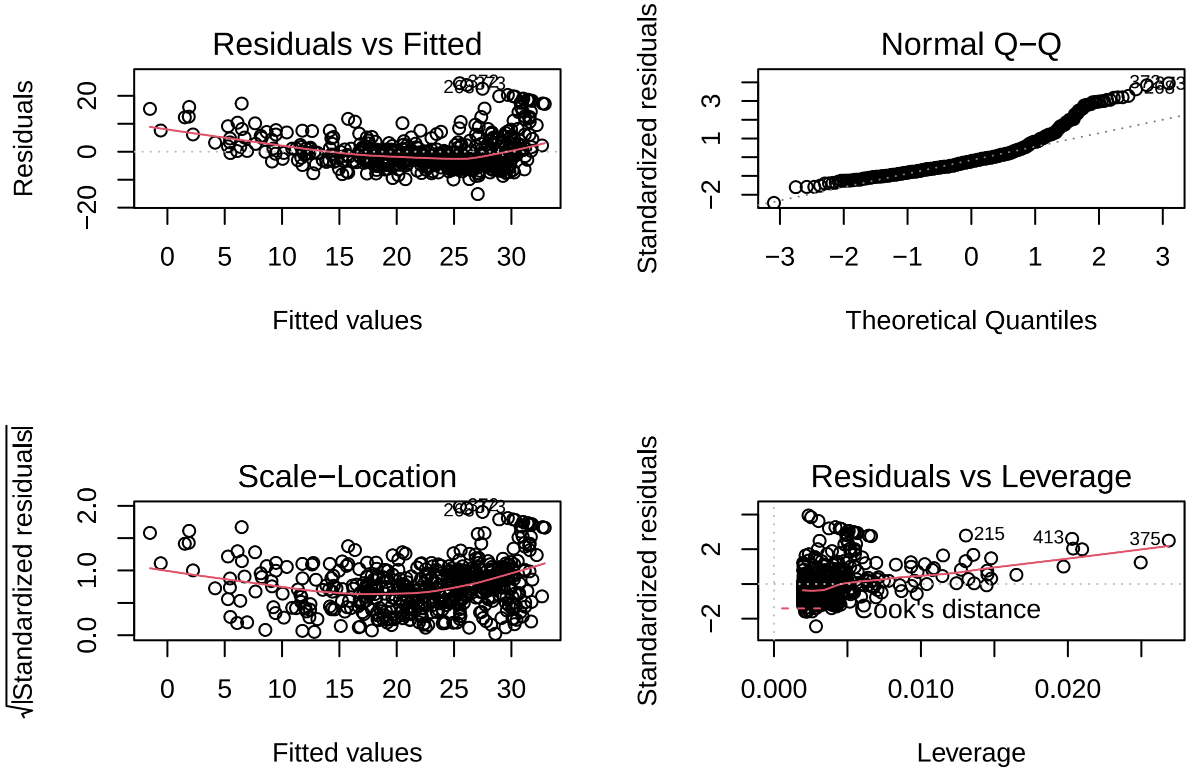
```
# Predict function
predict(lm.fit,data.frame(lstat=c(1)))
```

```
##        1
## 33.60379
```

```
plot(Boston$medv,Boston$lstat)
abline(lm.fit,col="red",lwd=3)
```



```
par(mfrow=c(2,2))
plot(lm.fit)
```


Residuals vs Fitted
Residuals
−20
0
20
0
5
10
15
20
25
30
Fitted values
Normal Q−Q
Standardized residuals
−2
1
3
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
0
5
10
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
−2
2
215
413
375
Cook's distance
0.000
0.010
0.020
Leverage