Linear Regression

By: Udit (based on ISLR)

Setup

```
#library(MASS)
library(ISLR2)
```

Simple Linear Regression

```
names(Boston)
```

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

```
?Boston
```

```
## starting httpd help server ... done
```

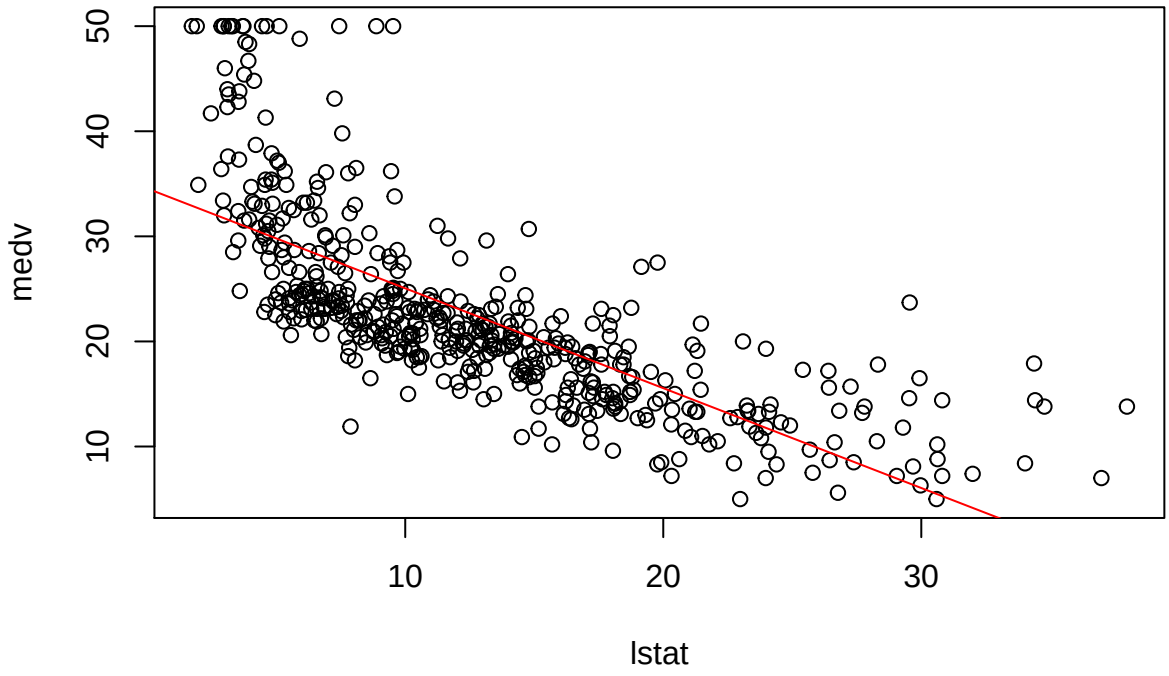
```
plot(medv~lstat, Boston)

# Linear Model
fit1 = lm(medv~lstat, data=Boston)
summary(fit1)
```

```
##
## 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
```

```
confint(fit1)
```

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

```
abline(fit1, col="red")
```




```
predict(fit1, data.frame(lstat=c(5,10,15)), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 29.80359 29.00741 30.59978
## 2 25.05335 24.47413 25.63256
## 3 20.30310 19.73159 20.87461
```

Multiple Linear Regression

```
round(cor(Boston),2)
```

```
##         crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio
## crim    1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29
```

```
## zn      -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39
## indus    0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38
## chas    -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12
## nox      0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19
## rm      -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36
## age      0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26
## dis     -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23
## rad      0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46
## tax      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46
## ptratio  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00
## lstat    0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37
## medv    -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51
##         lstat  medv
## crim     0.46 -0.39
## zn      -0.41  0.36
## indus    0.60 -0.48
## chas    -0.05  0.18
## nox      0.59 -0.43
## rm      -0.61  0.70
## age      0.60 -0.38
## dis     -0.50  0.25
## rad      0.49 -0.38
## tax      0.54 -0.47
## ptratio  0.37 -0.51
## lstat    1.00 -0.74
## medv    -0.74  1.00
```

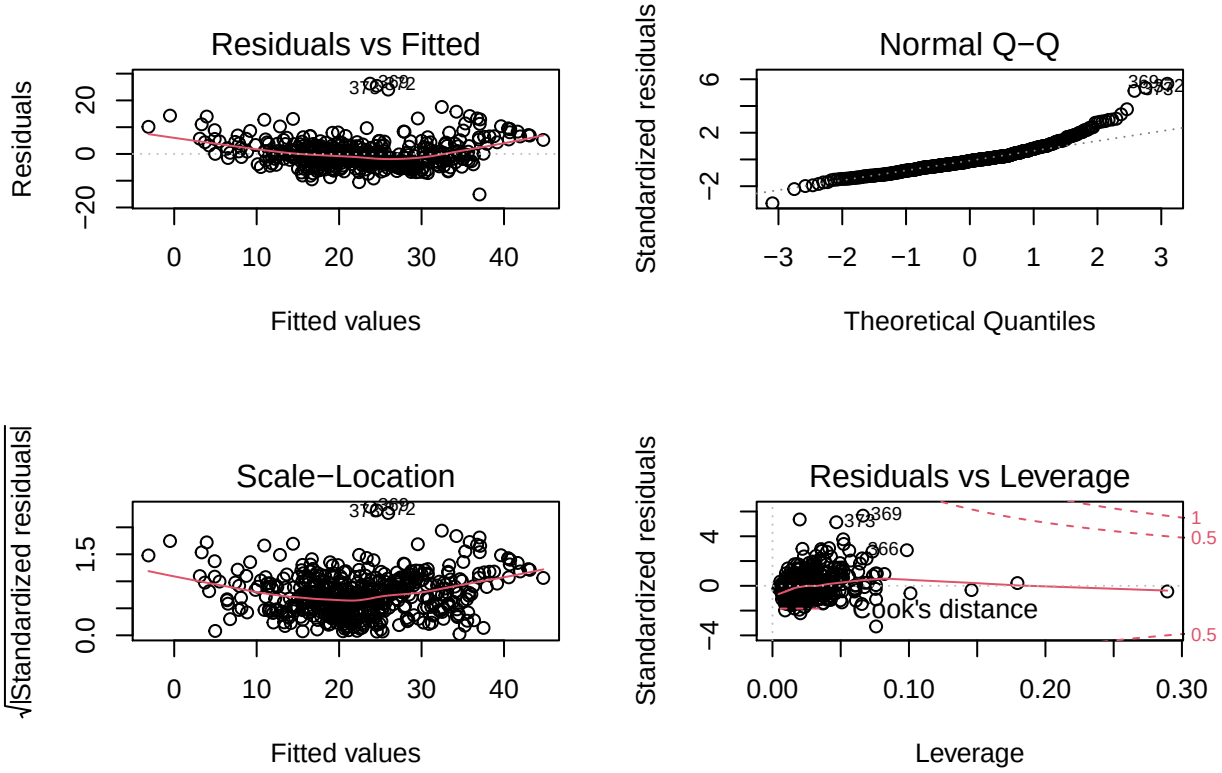
```
fit2 = lm(medv~lstat+age, data=Boston)
summary(fit2)
```

```
##
## Call:
## lm(formula = medv ~ lstat + age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.981  -3.978  -1.283   1.968  23.158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
## lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
## age          0.03454    0.01223   2.826  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.173 on 503 degrees of freedom
## Multiple R-squared:  0.5513, Adjusted R-squared:  0.5495
## F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

```
fit3 = lm(medv~., Boston)
summary(fit3)
```

```
##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1304  -2.7673  -0.5814   1.9414  26.2526
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.617270   4.936039   8.431 3.79e-16 ***
## crim         -0.121389   0.033000  -3.678 0.000261 ***
## zn            0.046963   0.013879   3.384 0.000772 ***
## indus         0.013468   0.062145   0.217 0.828520
## chas          2.839993   0.870007   3.264 0.001173 **
## nox         -18.758022   3.851355  -4.870 1.50e-06 ***
## rm            3.658119   0.420246   8.705  < 2e-16 ***
## age           0.003611   0.013329   0.271 0.786595
## dis          -1.490754   0.201623  -7.394 6.17e-13 ***
## rad           0.289405   0.066908   4.325 1.84e-05 ***
## tax          -0.012682   0.003801  -3.337 0.000912 ***
## ptratio      -0.937533   0.132206  -7.091 4.63e-12 ***
## lstat        -0.552019   0.050659 -10.897  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.798 on 493 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7278
## F-statistic: 113.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit3)
```




```
fit4 = update(fit3, ~.-age-indus)
summary(fit4)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
##     tax + ptratio + lstat, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1814  -2.7625  -0.6243   1.8448  26.3920
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.451747   4.903283   8.454 3.18e-16 ***
## crim         -0.121665   0.032919  -3.696 0.000244 ***
## zn            0.046191   0.013673   3.378 0.000787 ***
## chas          2.871873   0.862591   3.329 0.000935 ***
## nox         -18.262427   3.565247  -5.122 4.33e-07 ***
## rm            3.672957   0.409127   8.978  < 2e-16 ***
## dis          -1.515951   0.187675  -8.078 5.08e-15 ***
## rad           0.283932   0.063945   4.440 1.11e-05 ***
## tax          -0.012292   0.003407  -3.608 0.000340 ***
## ptratio      -0.930961   0.130423  -7.138 3.39e-12 ***
## lstat        -0.546509   0.047442 -11.519  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.789 on 495 degrees of freedom
## Multiple R-squared:  0.7342, Adjusted R-squared:  0.7289
## F-statistic: 136.8 on 10 and 495 DF,  p-value: < 2.2e-16
```

Interactions Terms

```
fit5 = lm(medv~lstat*age, Boston)
summary(fit5)
```

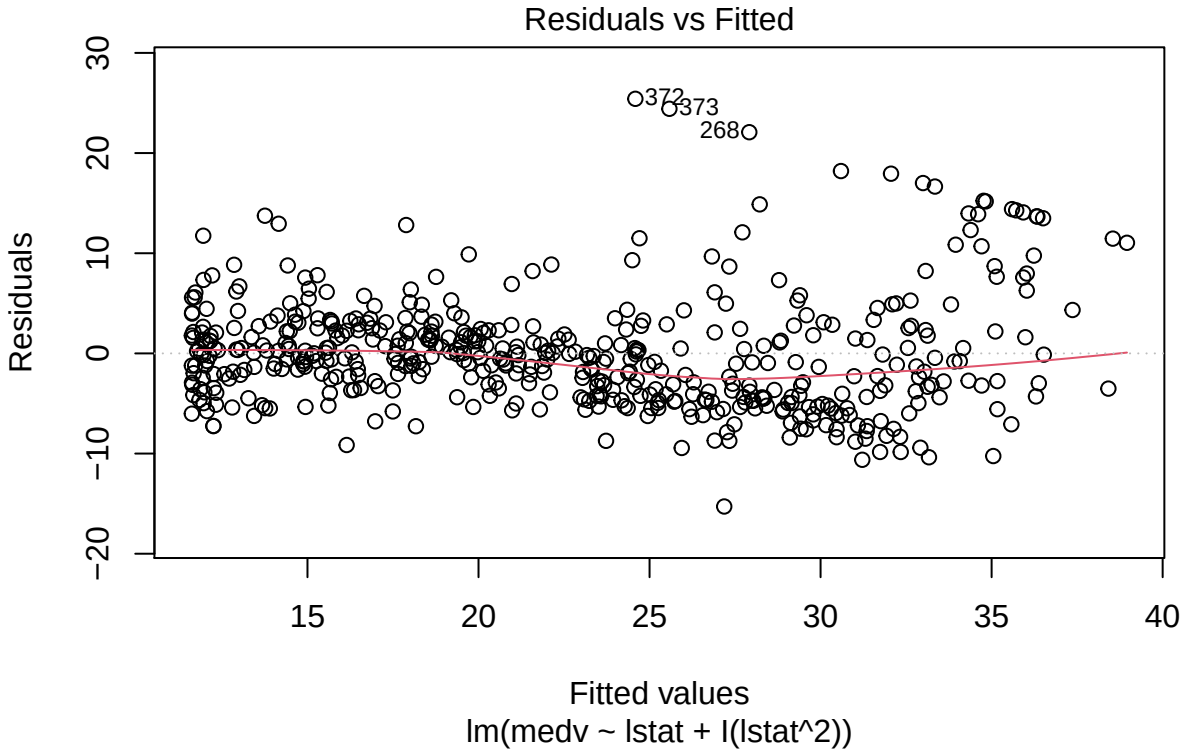
```
##
## Call:
## lm(formula = medv ~ lstat * age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.806  -4.045  -1.333   2.085  27.552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
## lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
## age         -0.0007209  0.0198792  -0.036   0.9711
## lstat:age    0.0041560  0.0018518   2.244   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.149 on 502 degrees of freedom
## Multiple R-squared:  0.5557, Adjusted R-squared:  0.5531
## F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

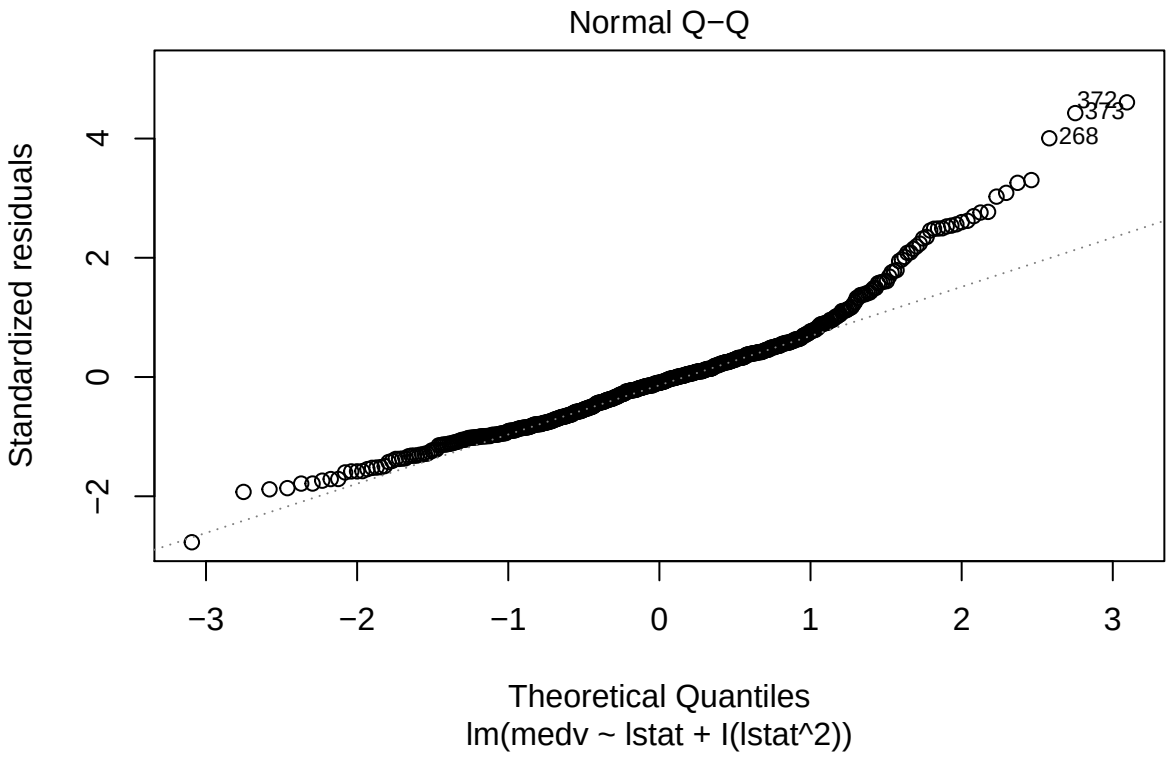
```
fit6 = lm(medv~lstat + I(lstat^2), Boston)
summary(fit6)
```

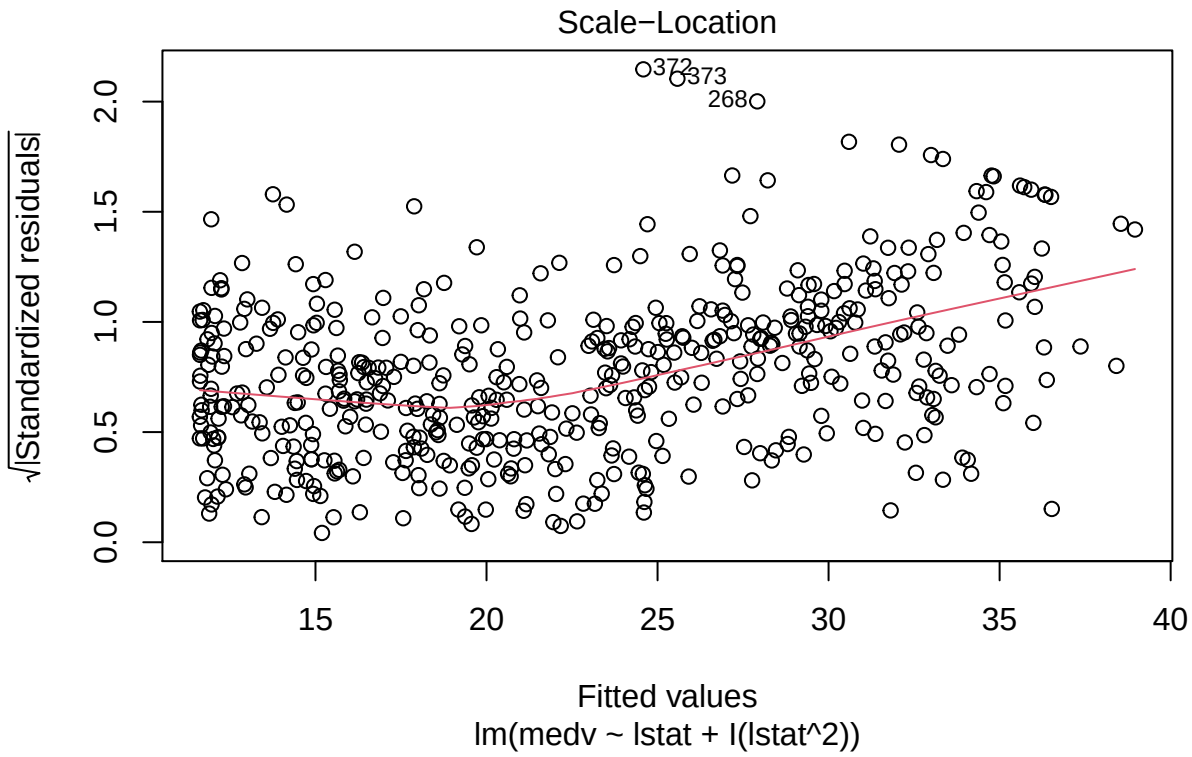
```
##
## Call:
## lm(formula = medv ~ lstat + I(lstat^2), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2834  -3.8313  -0.5295   2.3095  25.4148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.862007   0.872084   49.15   <2e-16 ***
## lstat       -2.332821   0.123803  -18.84   <2e-16 ***
## I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

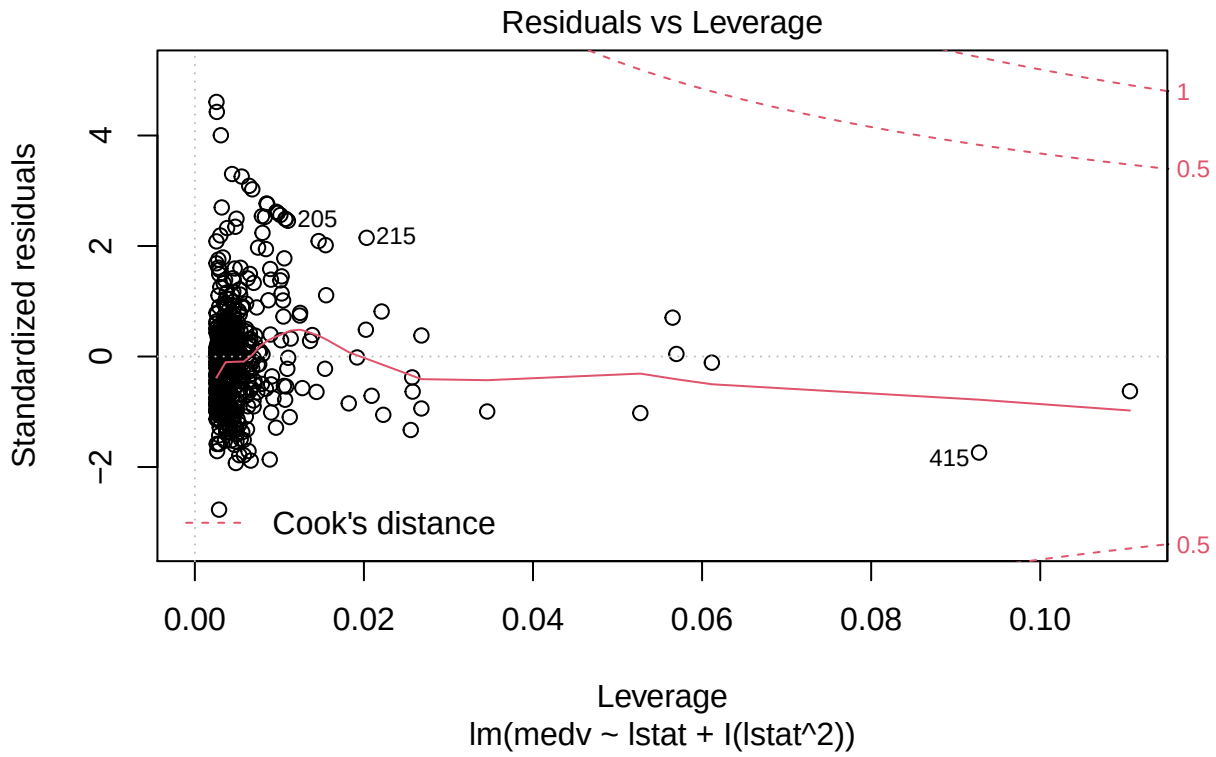
```
##
## Residual standard error: 5.524 on 503 degrees of freedom
## Multiple R-squared:  0.6407, Adjusted R-squared:  0.6393
## F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

```
plot(fit6)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(medv ~ lstat + I(lstat^2))
372
373
268
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
372
373
268
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
15
20
25
30
35
40
Fitted values
lm(medv ~ lstat + I(lstat^2))

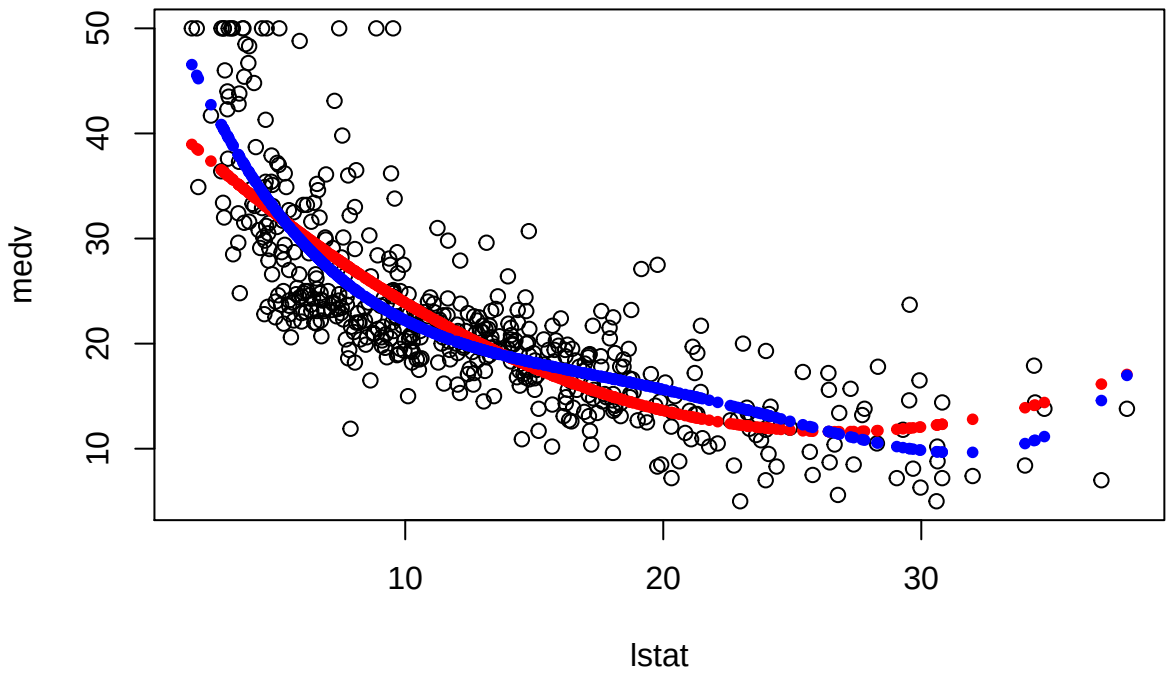



```
par(mfrow=c(1,1))
plot(medv~lstat, Boston)
points(Boston$lstat, fitted(fit6), col="red", pch=20)

fit7 = lm(medv~poly(lstat,4), Boston)
summary(fit7)
```

```
##
## Call:
## lm(formula = medv ~ poly(lstat, 4), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.563  -3.180  -0.632   2.283  27.181
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         22.5328     0.2347  95.995  < 2e-16 ***
## poly(lstat, 4)1 -152.4595     5.2801 -28.874  < 2e-16 ***
## poly(lstat, 4)2   64.2272     5.2801  12.164  < 2e-16 ***
## poly(lstat, 4)3  -27.0511     5.2801  -5.123 4.29e-07 ***
## poly(lstat, 4)4   25.4517     5.2801   4.820 1.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5.28 on 501 degrees of freedom
## Multiple R-squared:  0.673,  Adjusted R-squared:  0.6704
## F-statistic: 257.8 on 4 and 501 DF,  p-value: < 2.2e-16
```

```
points(Boston$lstat, fitted(fit7), col="blue", pch=20)
```




```
#fix(Boston) #to view and edit data
```

Qualitative Predictors

```
names(Carseats)
```

```
##  [1] "Sales"       "CompPrice"   "Income"      "Advertising" "Population"
##  [6] "Price"       "ShelveLoc"   "Age"         "Education"   "Urban"
## [11] "US"
```

```
summary(Carseats)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
```

```
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

```
fit1 = lm(Sales~.+Income:Advertising+Age:Price, Carseats)
summary(fit1)
```

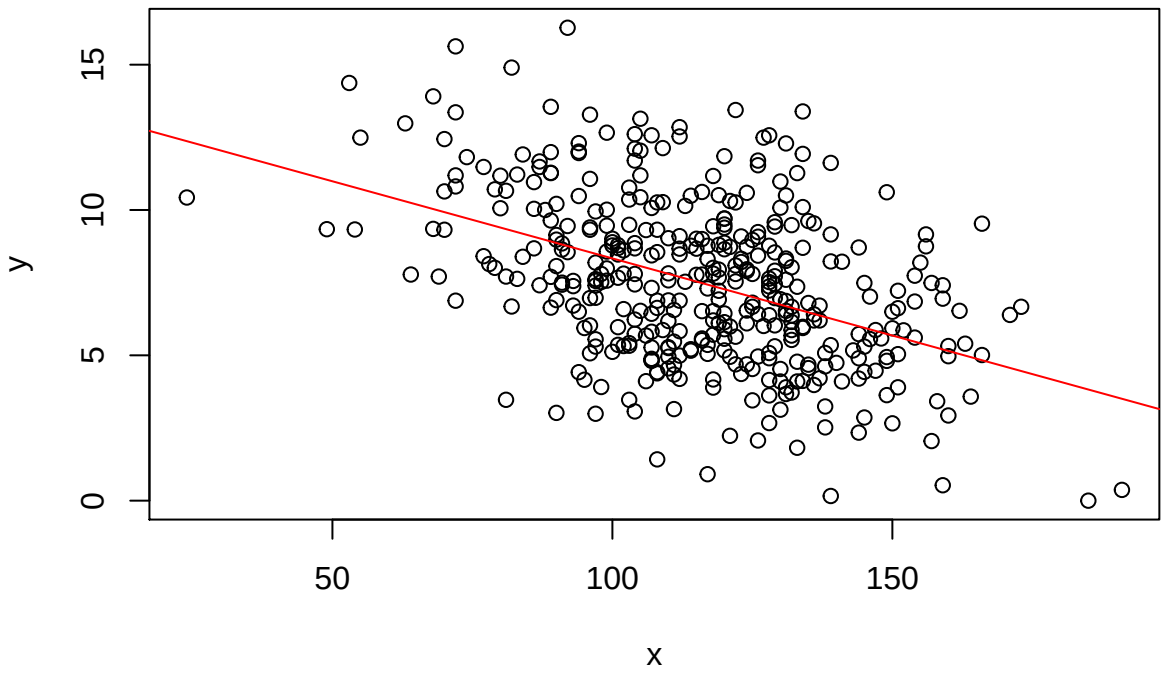
```
##
## Call:
## lm(formula = Sales ~ . + Income:Advertising + Age:Price, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9208 -0.7503  0.0177  0.6754  3.3413
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.5755654  1.0087470   6.519 2.22e-10 ***
## CompPrice           0.0929371  0.0041183  22.567  < 2e-16 ***
## Income              0.0108940  0.0026044   4.183 3.57e-05 ***
## Advertising         0.0702462  0.0226091   3.107 0.002030 **
## Population          0.0001592  0.0003679   0.433 0.665330
## Price              -0.1008064  0.0074399 -13.549  < 2e-16 ***
## ShelveLocGood       4.8486762  0.1528378  31.724  < 2e-16 ***
## ShelveLocMedium     1.9532620  0.1257682  15.531  < 2e-16 ***
## Age                -0.0579466  0.0159506  -3.633 0.000318 ***
## Education          -0.0208525  0.0196131  -1.063 0.288361
## UrbanYes            0.1401597  0.1124019   1.247 0.213171
## USYes              -0.1575571  0.1489234  -1.058 0.290729
## Income:Advertising  0.0007510  0.0002784   2.698 0.007290 **
## Price:Age           0.0001068  0.0001333   0.801 0.423812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.011 on 386 degrees of freedom
## Multiple R-squared:  0.8761, Adjusted R-squared:  0.8719
## F-statistic:   210 on 13 and 386 DF,  p-value: < 2.2e-16
```

```
contrasts(Carseats$ShelveLoc)
```

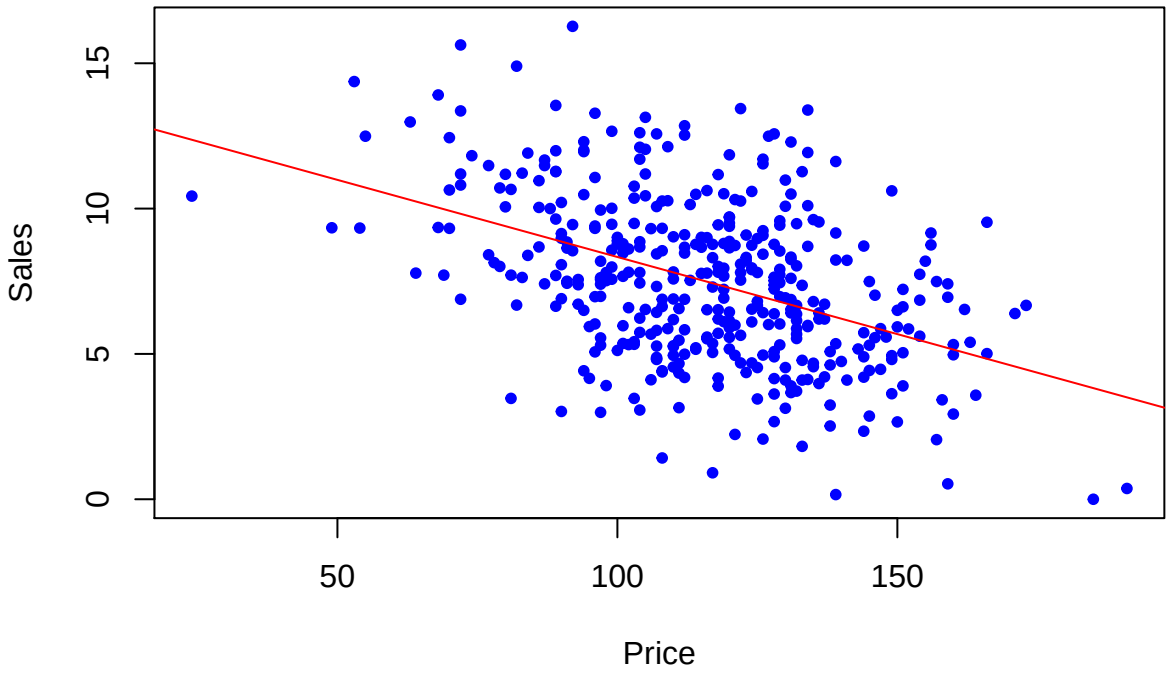
```
##        Good Medium
## Bad       0      0
## Good      1      0
## Medium    0      1
```

Writing R function

```
regplot=function(x,y,...){
  fit = lm(y~x)
  plot(x,y,...)
  abline(fit, col="red")
}
regplot(Carseats$Price, Carseats$Sales)
```




```
regplot(Carseats$Price, Carseats$Sales, xlab="Price", ylab="Sales", col="blue", pch=20)
```


Sales
0
5
10
15
50
100
150
Price