MultipleLinearRegression.R

vijaykalmath

2022-01-04

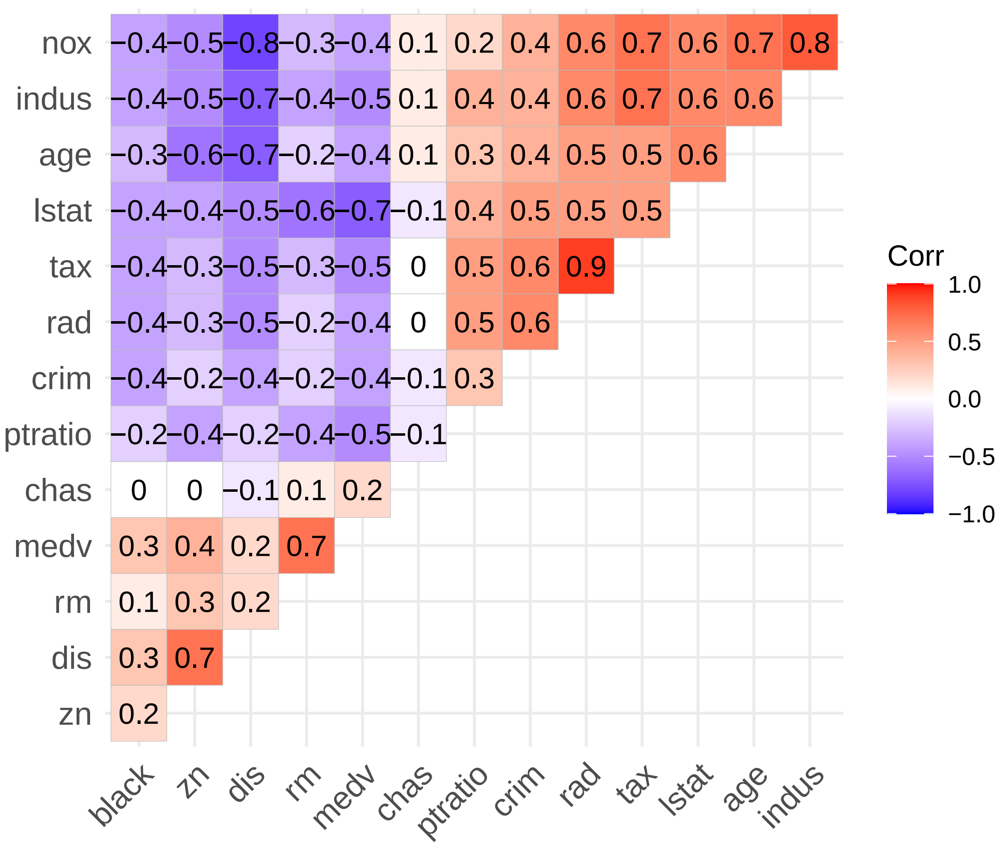
```r
# Convert R script to RmarkDown ->  Cmd + Shift + K
# Multiple Linear Regression - ISLR Lab Work

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()
```

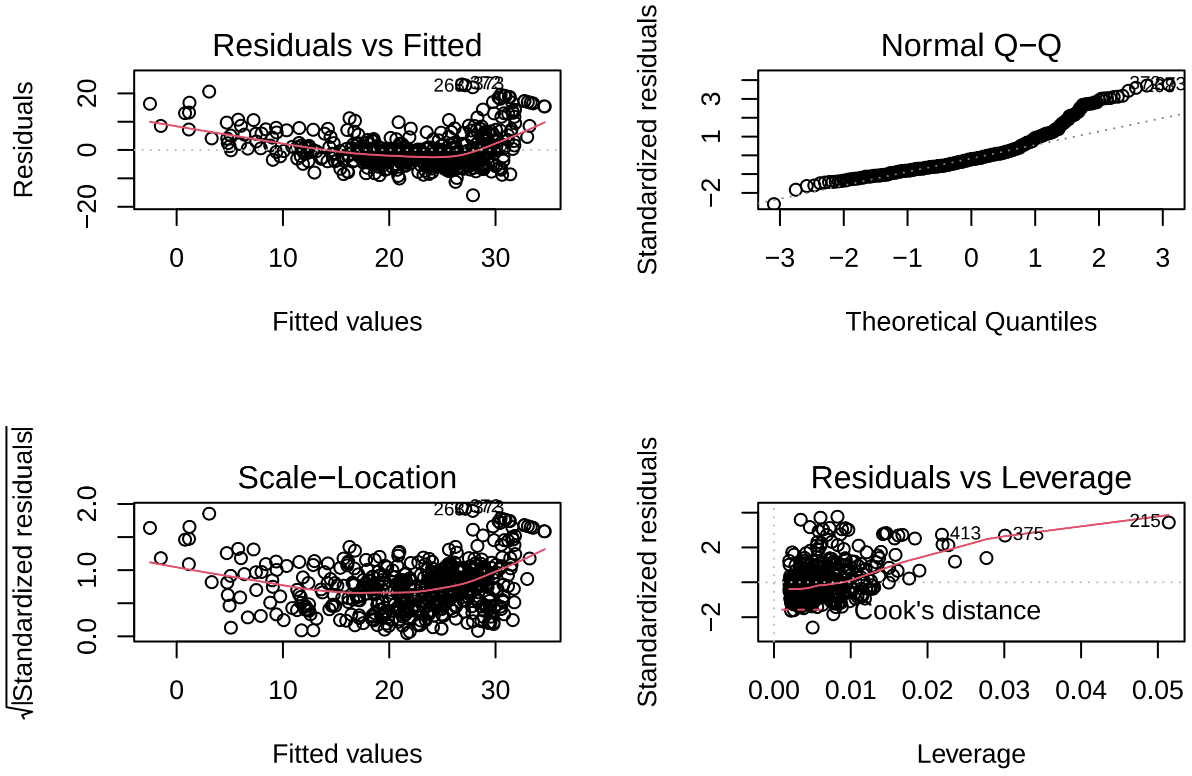
```r
library(ggcorrplot)


# 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

mlr.fit  = lm(medv~lstat + age,data = Boston)

summary(mlr.fit)
```

```
##
## 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
```

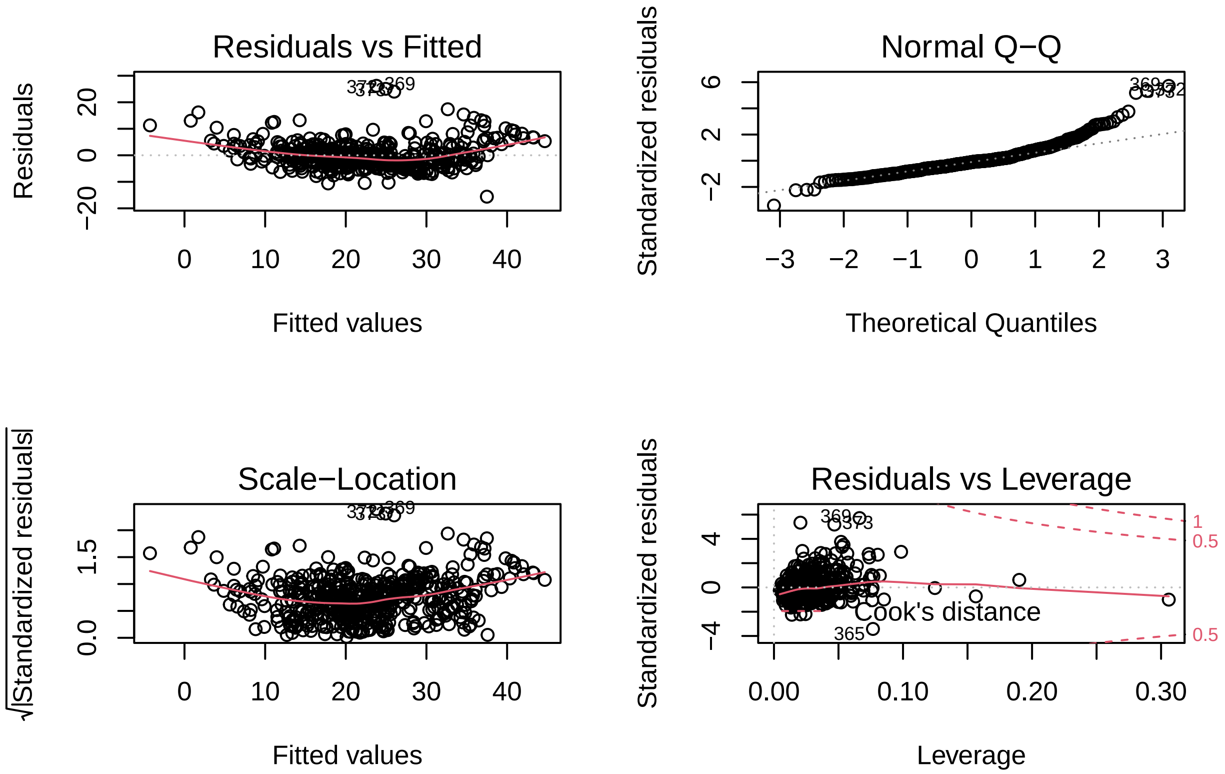
```
par(mfrow=c(2,2));plot(mlr.fit)
```



```
# Linear Regression with all terms

mlr.fit1 = lm(medv~.,data=Boston)

summary(mlr.fit1)
```

```
##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2));plot(mlr.fit1)
```

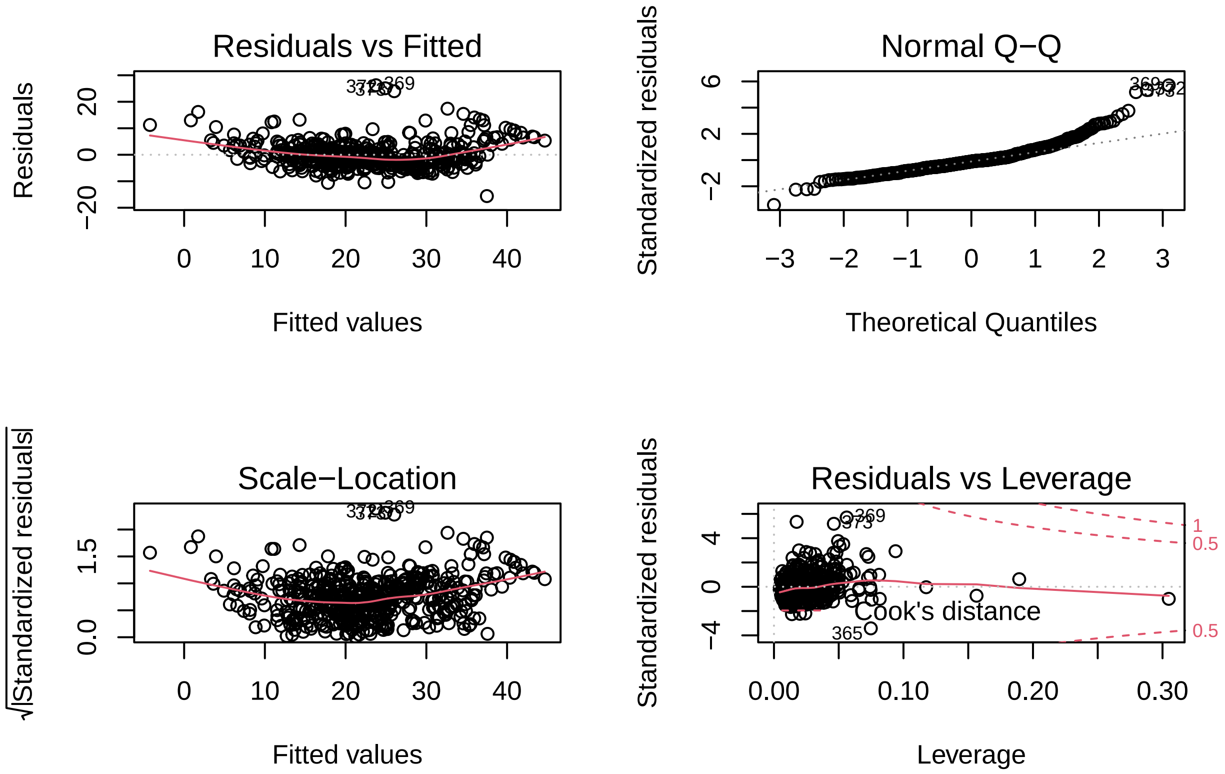


```
# Indus, age have huge P values for Beta=0 Hypothesis test, therefore they can be removed from the line


mlr.fit2 = lm(medv ~ . -indus -age,data=Boston)

summary(mlr.fit2)
```

```
##
## Call:
```

```
## lm(formula = medv ~ . - indus - age, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## crim         -0.108413   0.032779  -3.307 0.001010 **
## zn            0.045845   0.013523   3.390 0.000754 ***
## chas          2.718716   0.854240   3.183 0.001551 **
## nox         -17.376023   3.535243  -4.915 1.21e-06 ***
## rm            3.801579   0.406316   9.356  < 2e-16 ***
## dis          -1.492711   0.185731  -8.037 6.84e-15 ***
## rad           0.299608   0.063402   4.726 3.00e-06 ***
## tax          -0.011778   0.003372  -3.493 0.000521 ***
## ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
## black         0.009291   0.002674   3.475 0.000557 ***
## lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

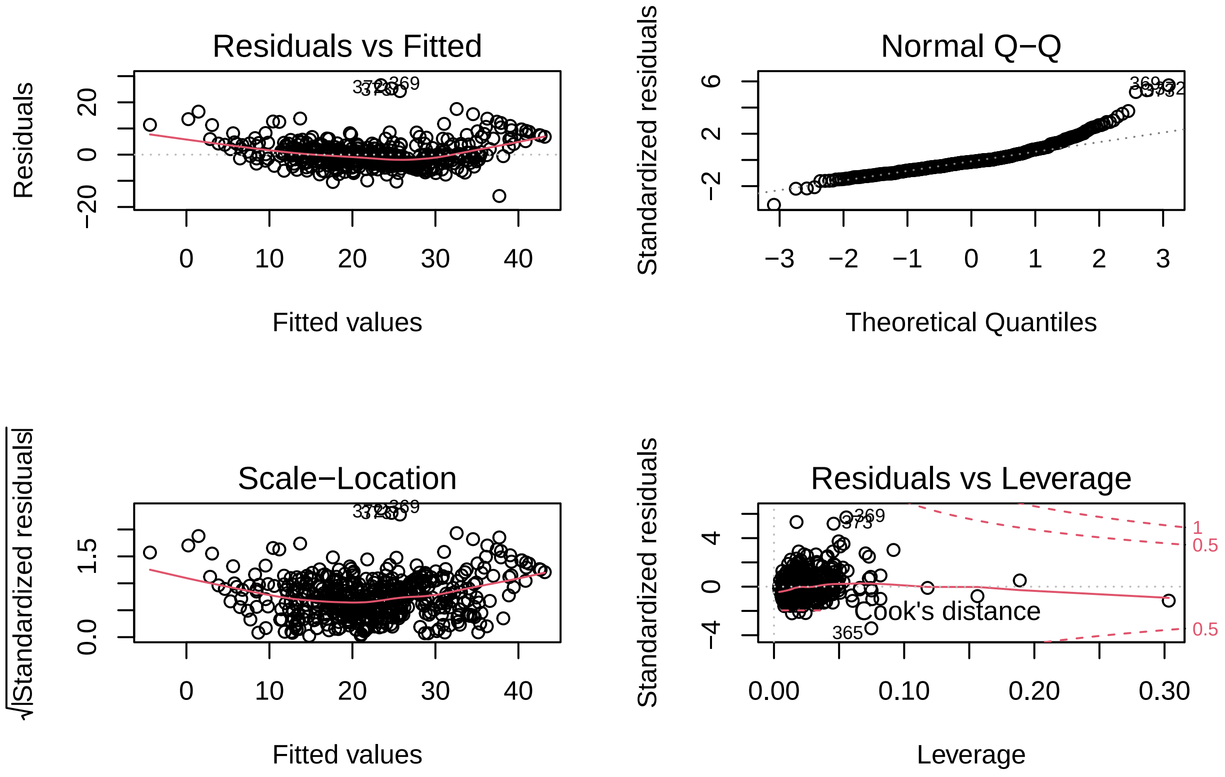
```
par(mfrow=c(2,2));plot(mlr.fit2)
```



```r
mlr.fit3 = lm(medv ~ . -indus -age -zn, - rad,data=Boston)

summary(mlr.fit3)
```

```
##
## Call:
## lm(formula = medv ~ . - indus - age - zn, data = Boston, subset = -rad)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8016  -2.8416  -0.6879   1.8522  26.5601
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.944368   5.144916   7.375 7.14e-13 ***
## crim          -0.098243   0.033044  -2.973 0.003094 **
## chas           2.732615   0.864866   3.160 0.001678 **
## nox          -18.972509   3.563913  -5.324 1.56e-07 ***
## rm             3.936449   0.407843   9.652  < 2e-16 ***
## dis           -1.208523   0.163336  -7.399 6.08e-13 ***
## rad            0.278760   0.064060   4.352 1.65e-05 ***
## tax           -0.009043   0.003357  -2.694 0.007313 **
## ptratio       -1.110758   0.124043  -8.955  < 2e-16 ***
## black          0.009283   0.002706   3.430 0.000655 ***
## lstat         -0.528039   0.048357 -10.920  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.791 on 486 degrees of freedom
## Multiple R-squared:  0.7353, Adjusted R-squared:  0.7298
## F-statistic:   135 on 10 and 486 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2));plot(mlr.fit3)
```



```
# Since Residuals plot has a curve is Adding Non-Linear Transformations

mlr.fit4 = lm(medv ~ . + I(lstat^3) -indus -age -zn, - rad ,data=Boston)

summary(mlr.fit4)
```

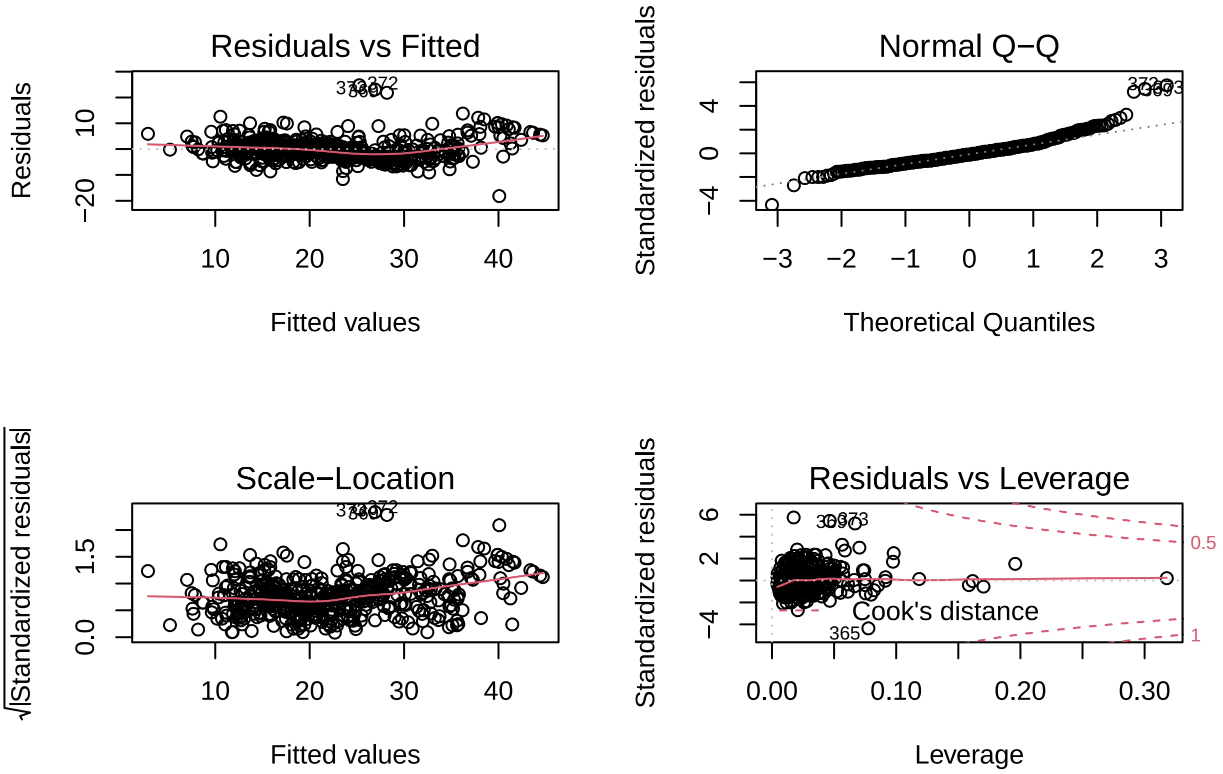
```
##
## Call:
## lm(formula = medv ~ . + I(lstat^3) - indus - age - zn, data = Boston,
##     subset = -rad)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1843  -2.7248  -0.4241   2.0495  24.7402
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.930e+01  4.671e+00   8.414 4.48e-16 ***
## crim        -1.410e-01  3.027e-02  -4.658 4.12e-06 ***
## chas         2.684e+00  7.848e-01   3.419  0.00068 ***
## nox         -1.412e+01  3.269e+00  -4.319 1.90e-05 ***
## rm           3.522e+00  3.723e-01   9.460  < 2e-16 ***
## dis         -1.232e+00  1.482e-01  -8.309 9.71e-16 ***
## rad          2.551e-01  5.818e-02   4.386 1.42e-05 ***
## tax         -7.784e-03  3.049e-03  -2.553  0.01098 *
## ptratio     -8.673e-01  1.150e-01  -7.539 2.34e-13 ***
## black        7.987e-03  2.459e-03   3.248  0.00124 **
## lstat       -1.171e+00  7.649e-02 -15.304  < 2e-16 ***
## I(lstat^3)   6.106e-04  5.953e-05  10.256  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.347 on 485 degrees of freedom
## Multiple R-squared:  0.7824, Adjusted R-squared:  0.7775
## F-statistic: 158.6 on 11 and 485 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2));plot(mlr.fit4)

# Using Poly

mlr.fit5 = lm(medv ~ . + poly(lstat,5) -indus -age -zn, - rad ,data=Boston)

summary(mlr.fit5)
```

```
##
## Call:
## lm(formula = medv ~ . + poly(lstat, 5) - indus - age - zn, data = Boston,
##     subset = -rad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.996  -2.250  -0.174   1.650  26.442
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      40.292185   4.546113   8.863  < 2e-16 ***
## crim             -0.138039   0.028744  -4.802 2.10e-06 ***
## chas              1.976932   0.747907   2.643 0.008478 **
## nox             -14.296902   3.093314  -4.622 4.89e-06 ***
## rm                2.789861   0.381076   7.321 1.04e-12 ***
## dis              -1.183509   0.140861  -8.402 4.96e-16 ***
## rad               0.278803   0.055122   5.058 6.03e-07 ***
## tax              -0.010390   0.002904  -3.578 0.000381 ***
## ptratio          -0.855066   0.109173  -7.832 3.08e-14 ***
## black             0.008271   0.002328   3.552 0.000419 ***
## lstat            -0.640497   0.043090 -14.864  < 2e-16 ***
## poly(lstat, 5)1         NA         NA      NA       NA
## poly(lstat, 5)2  51.425795   4.492945  11.446  < 2e-16 ***
## poly(lstat, 5)3 -12.912844   4.505097  -2.866 0.004335 **
## poly(lstat, 5)4  22.079550   4.279381   5.160 3.62e-07 ***
## poly(lstat, 5)5 -15.791612   4.208092  -3.753 0.000196 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.111 on 482 degrees of freedom
## Multiple R-squared:  0.8066, Adjusted R-squared:  0.801
## F-statistic: 143.6 on 14 and 482 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2));plot(mlr.fit4)
```



Residuals vs Fitted
Residuals
Fitted values
−20
10
10
20
30
40
373
369
372
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−4
0
4
−3
−2
−1
0
1
2
3
372
369
373
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.5
10
20
30
40
373
369
372
Residuals vs Leverage
Standardized residuals
Leverage
−4
2
6
0.00
0.10
0.20
0.30
369
373
365
Cook's distance
0.5
1