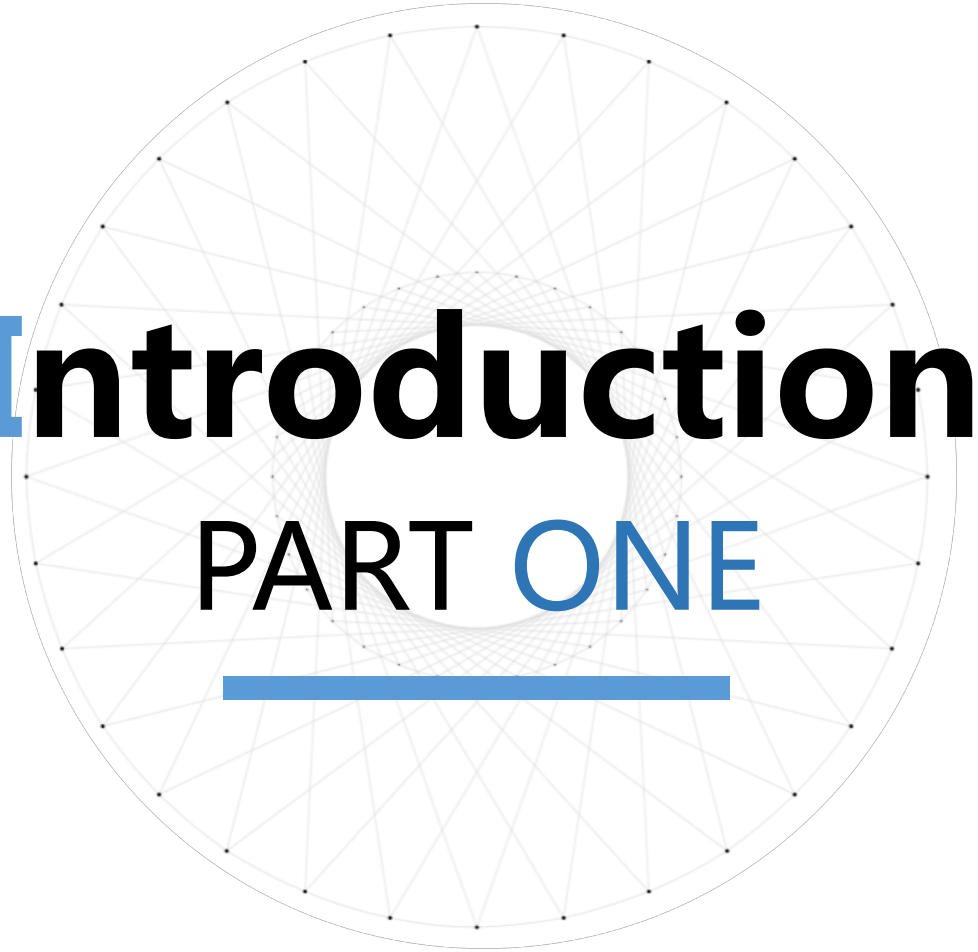
REGRESSION ANALYSIS

—GROUP PROJECT

Analysis of Boston House-Price

何钰佳　1530005008
金　澄　1530005011
唐一菡　1530012043
熊　峰　1530005039
杨　璐　1530005042
张家华　1530005048

Contents



Introduction

PART ONE

Introduction of variables

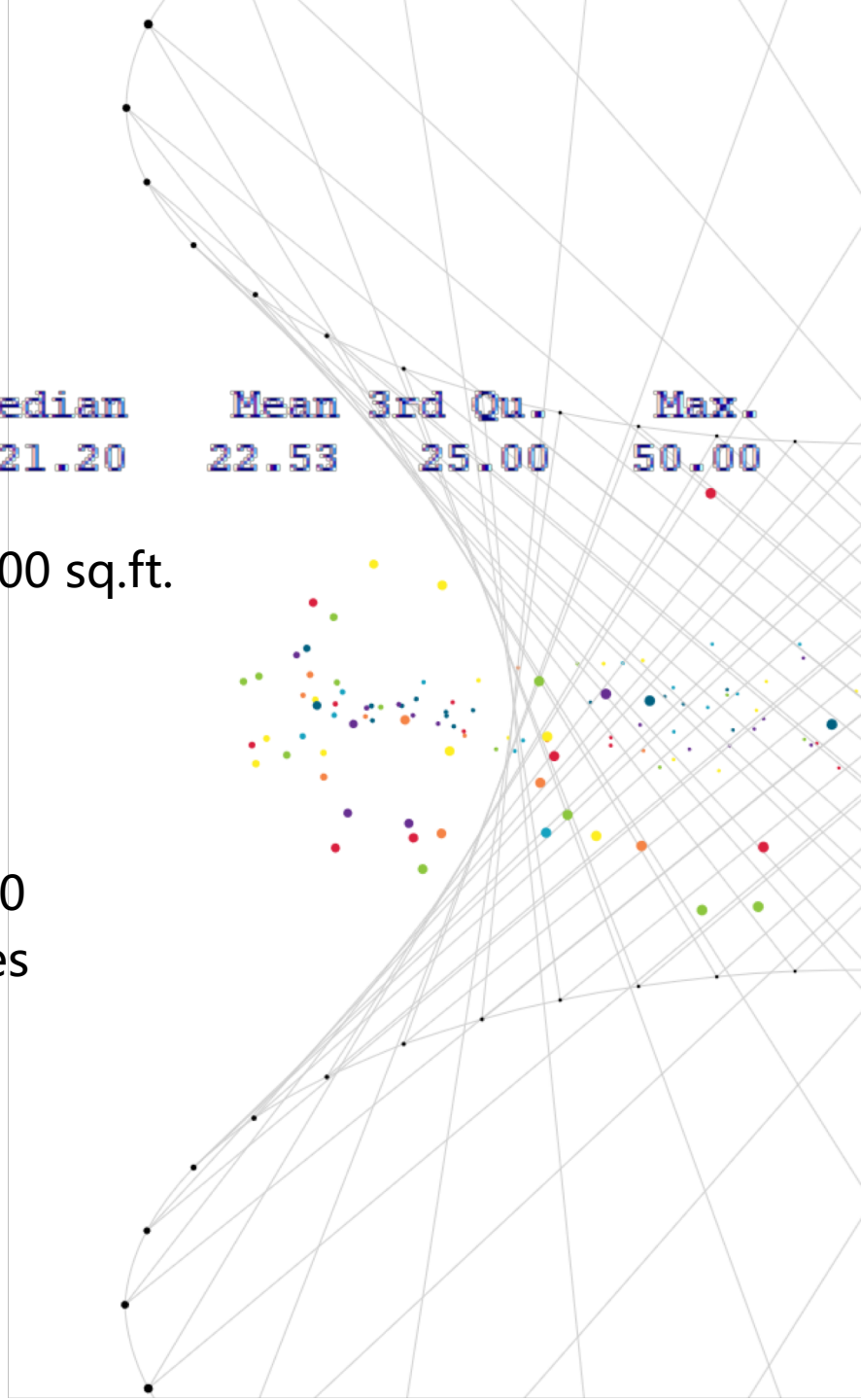
The dependent variable

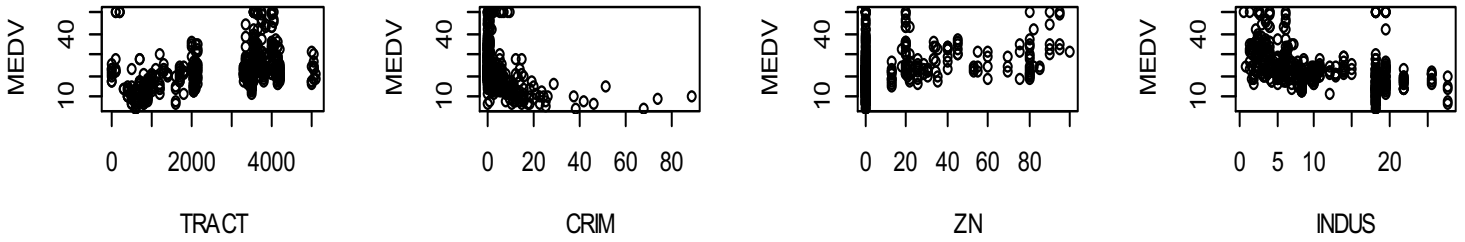
- MEDV: *median value* of owner-occupied homes in $1000’ s

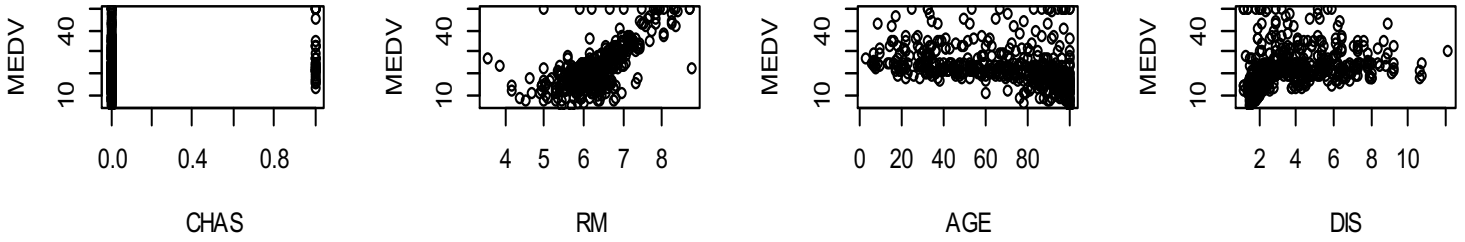
The independent variable

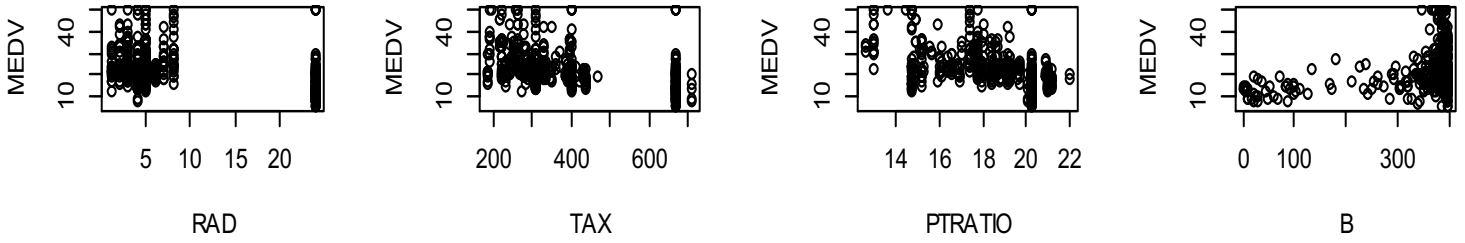
```
> summary(MEDV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   17.02   21.20   22.53   25.00   50.00
```

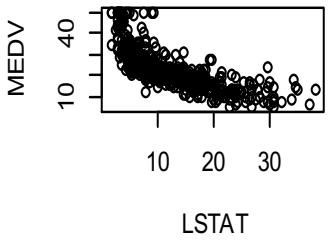
- TRACT: census of tract
- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: close to Charles river or not
- NOX: nitric oxides concentration
- RM: average number of rooms per dwelling
- 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 $10,000
- PTRATIO: pupil-teacher ratio by town
- B: proportion of blacks by town
- LSTAT: lower status of the population



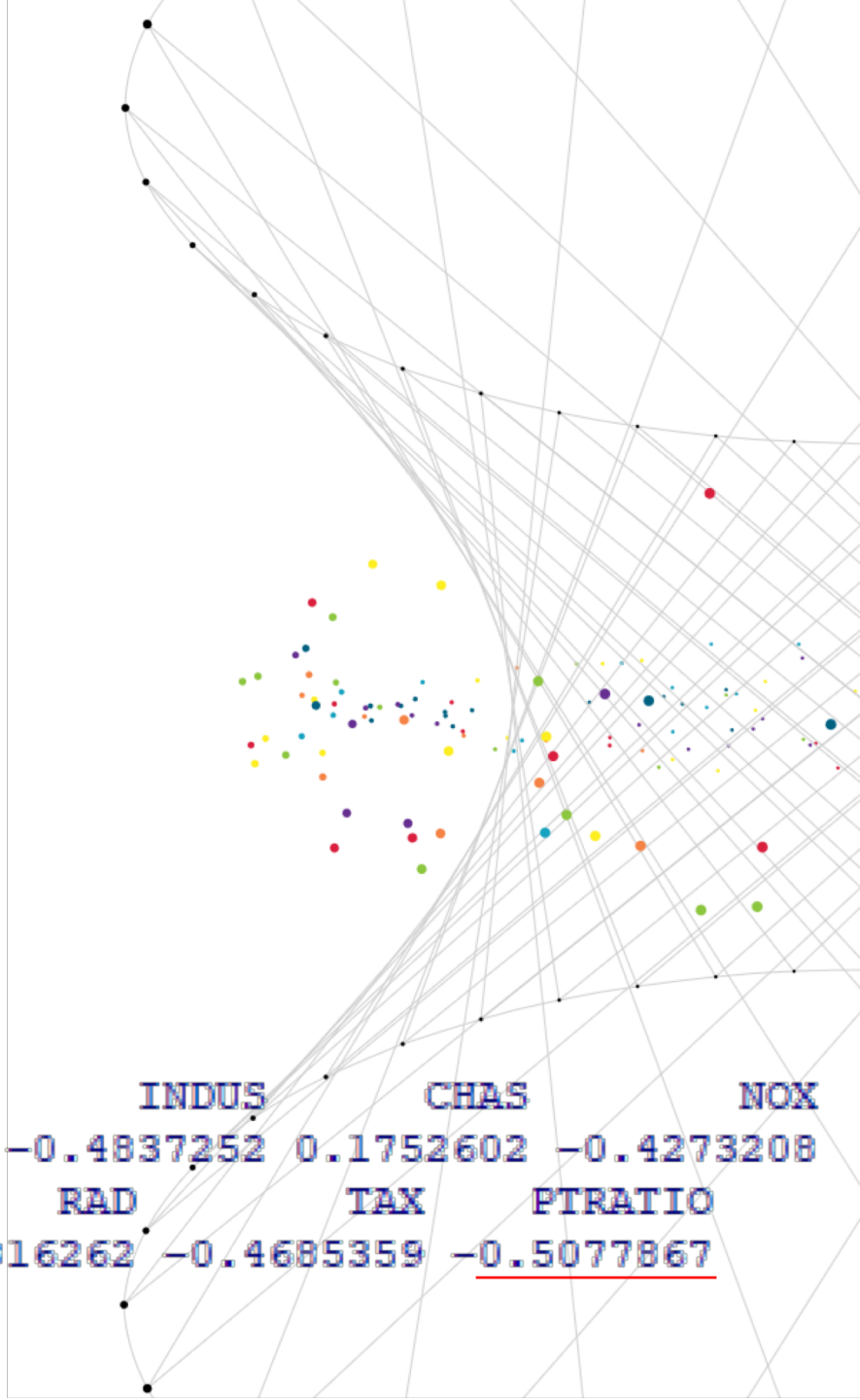









```
> cor(MEDV,a)
           TRACT MEDV       CRIM         ZN      INDUS       CHAS        NOX
[1,] 0.4263792    1 -0.3883046 0.3604453 -0.4837252 0.1752602 -0.4273208
             RM        AGE       DIS        RAD       TAX   PTRATIO
[1,] 0.6953599 -0.3769546 0.2499287 -0.3816262 -0.4685359 -0.5077867
             B      LSTAT
[1,] 0.3334608 -0.7376627
```



Diagnostic Criteria

PART TWO

Diagnostic criteria

✓ *qq-plot:*

residual normal distribution

✓ *Influential point :*

cook distance

✓ *Independent of residual*

Durbinwaston method

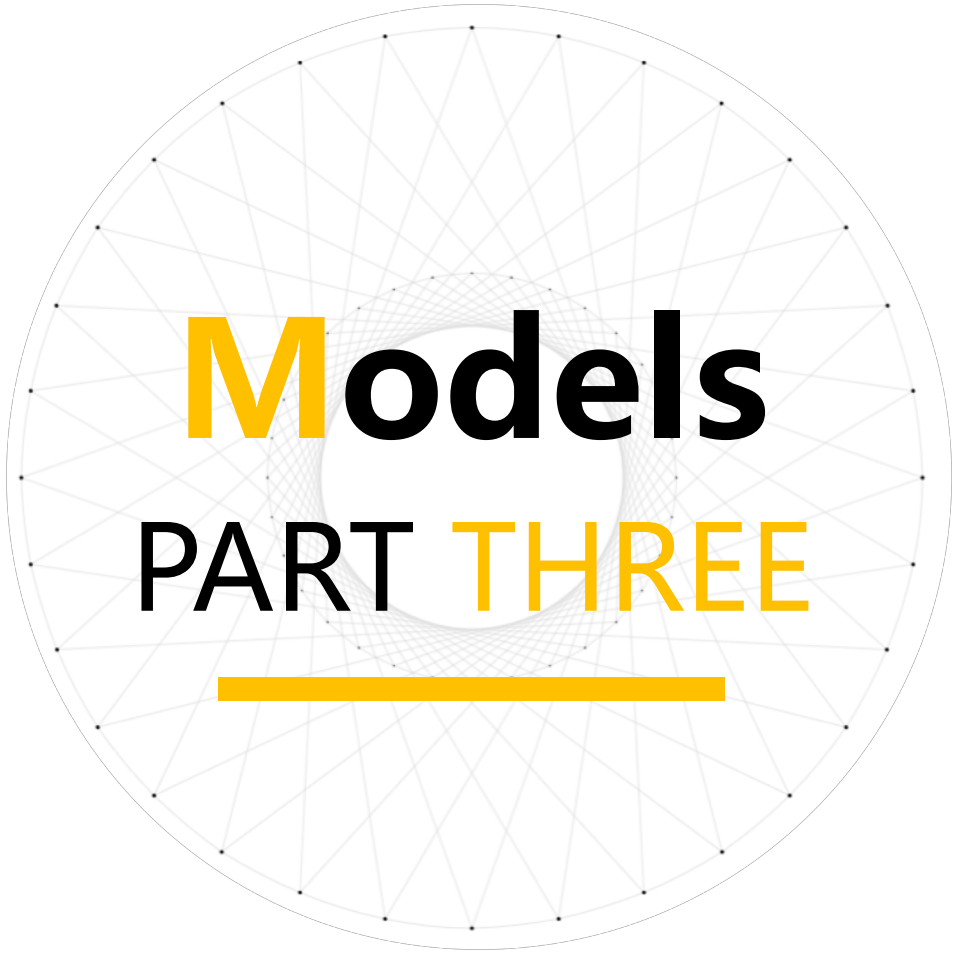
- DW=2 independent
- DW=0 positive correlation
- DW=-2 negative correlation

✓ *Variance should be constant*

(p-value>0.05)

✓ *Predict of model*

- PRESS= $\sum(y_i-\hat{y}_{i,-i})^2 = \sum(e_{i,-i})^2$
- $R_{pred}^2 = 1 - \frac{PRESS}{SST}$



Models

PART THREE

3.1 General Multiple Linear Regression

3.1 One dimensional linear regression :

MEDV~LSTAT+INDUS+RM+TAX (whole data)

R^2_{adj}=0.6464

s^2=5.469

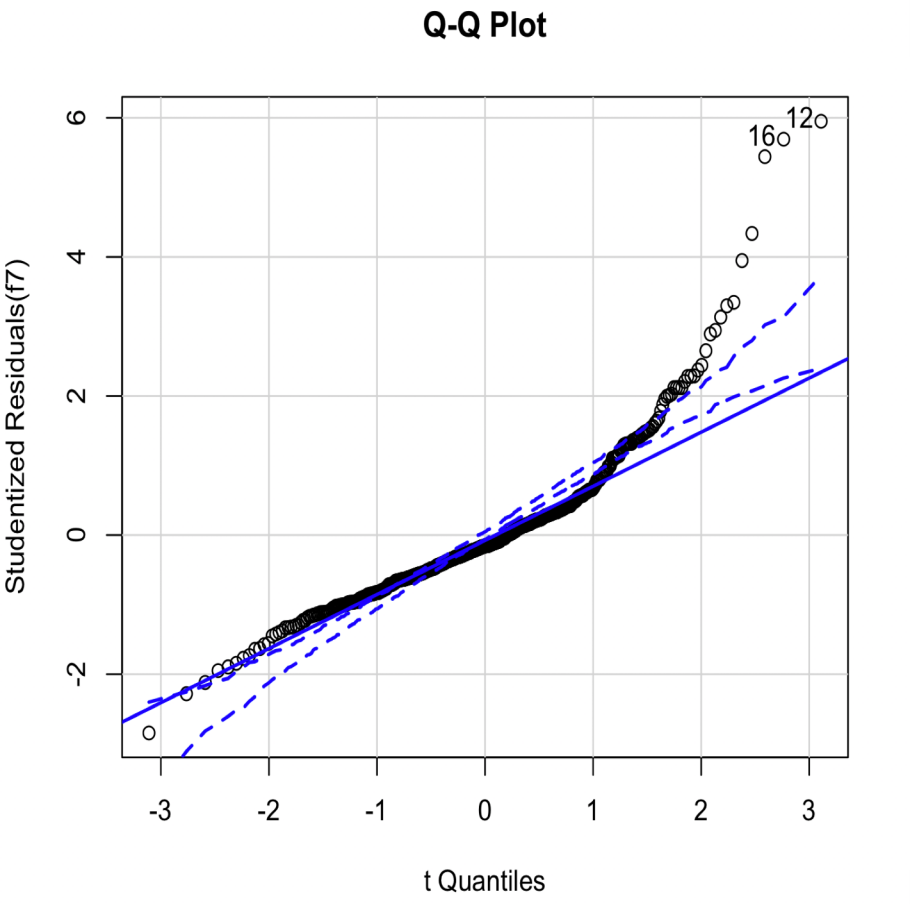
PRESS= 9491.42

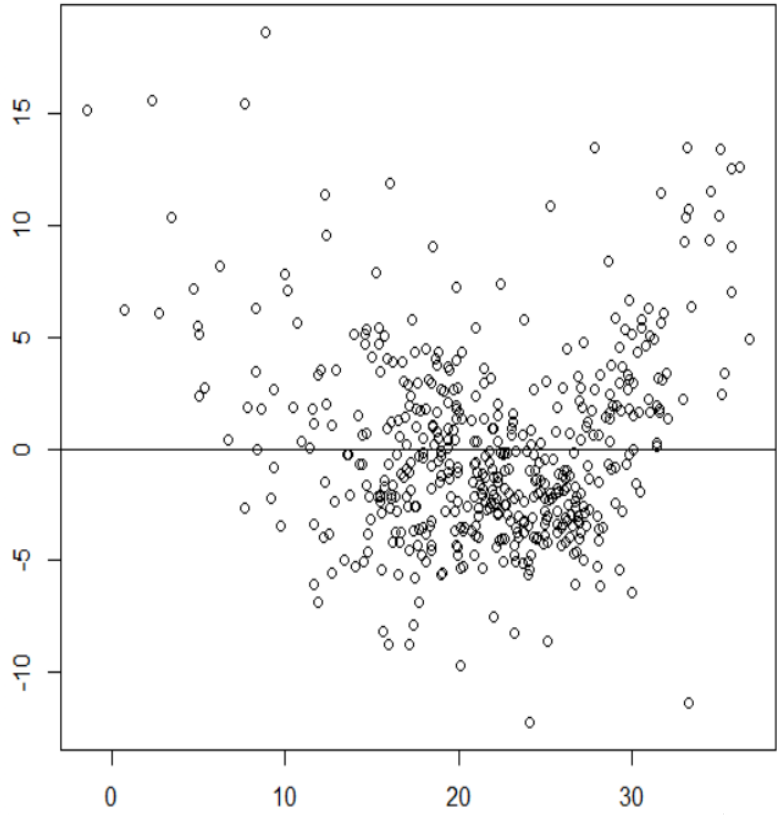
$R^2_{prediction}$ =0.6975923

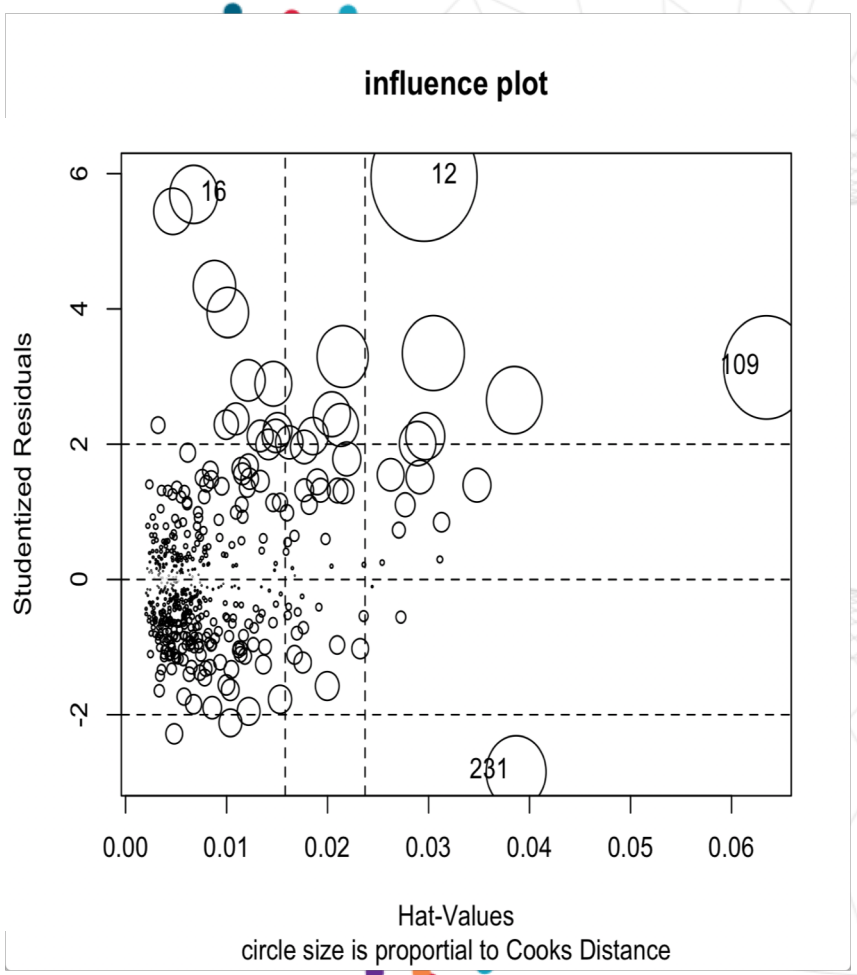
```
> reg=lm(MEDV~LSTAT+INDUS+RM+TAX,data=boston)
> summary(reg)

Call:
lm(formula = MEDV ~ LSTAT + INDUS + RM + TAX, data = boston)

Residuals:
    Min      1Q  Median      3Q     Max
-16.251  -3.459  -1.165   1.865  30.655

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.667249   3.144982  -0.212 0.832066
LSTAT       -0.565857   0.051114 -11.070  < 2e-16 ***
INDUS        0.054370   0.055156   0.986 0.324731
RM           5.237485   0.441313  11.868  < 2e-16 ***
TAX         -0.007744   0.002136  -3.625 0.000318 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.469 on 501 degrees of freedom
Multiple R-squared:  0.6492,  Adjusted R-squared:  0.6464
F-statistic: 231.8 on 4 and 501 DF,  p-value: < 2.2e-16
```

Diagnostic

Normalization of residual







Diagnostic

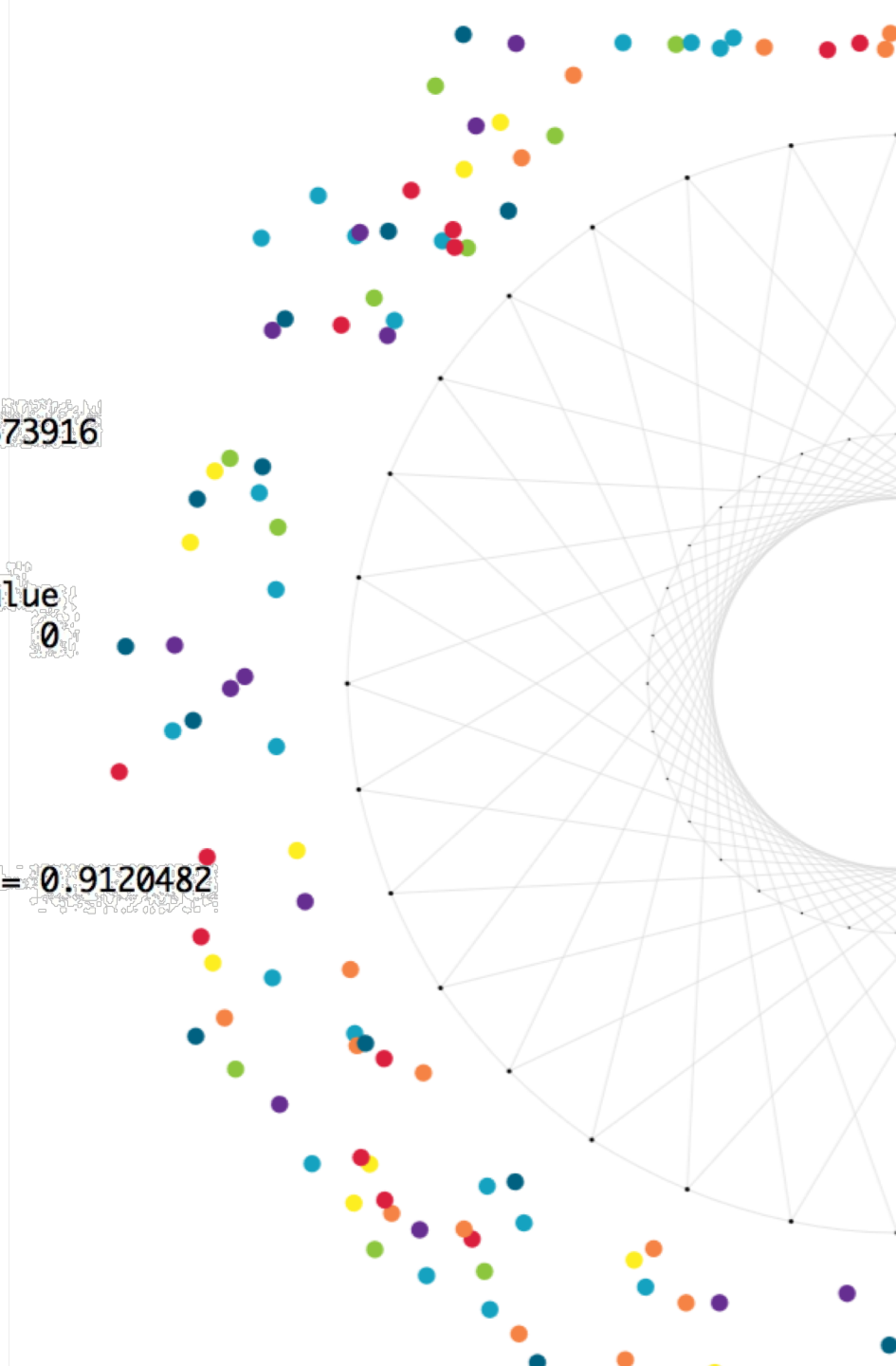
✓ Prediction Rred^2

X:Independent of residual (iid)

✓ Variance should be constant

✓ Multilinear test:

```
> shrinkage(reg)
Original R-square = 0.6491952
10 Fold Cross-Validated R-square = 0.6373916
Change = 0.01180355
>
> durbinWatsonTest(reg)
 lag Autocorrelation D-W Statistic p-value
   1       0.4700797      1.042655       0
 Alternative hypothesis: rho != 0
>
> ncvTest(reg)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.01220039    Df = 1     p = 0.9120482
>
> vif(reg)
   LSTAT    INDUS       RM      TAX
2.249472 2.417380 1.623310 2.187985
>
> sqrt(vif(reg))>2
LSTAT INDUS    RM   TAX
FALSE FALSE FALSE FALSE
```



Stepwise All Variable Into Model

MEDV~TRACT+CRIM+ZN+CHAS+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT

R^2_{adj}=0.7425

s^2=4.715

PRESS= 7258.276

$R^2_{prediction}$ =0.7296

```
> summary(f4)

Call:
lm(formula = MEDV ~ TRACT + CRIM + ZN + CHAS + NOX + RM + DIS +
    RAD + TAX + PTRATIO + B + LSTAT, data = boston)

Residuals:
     Min       1Q   Median       3Q      Max
-16.4176  -2.7873  -0.5537   1.8491  25.3075

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.223e+01  5.606e+00   7.532 2.52e-13 ***
TRACT       -7.738e-04  2.714e-04  -2.852 0.004538 **
CRIM        -1.115e-01  3.282e-02  -3.398 0.000735 ***
ZN           4.656e-02  1.352e-02   3.444 0.000624 ***
CHAS         2.643e+00  8.517e-01   3.103 0.002028 **
NOX         -1.794e+01  3.618e+00  -4.958 9.91e-07 ***
RM           3.826e+00  4.072e-01   9.395  < 2e-16 ***
DIS         -1.672e+00  1.949e-01  -8.576  < 2e-16 ***
RAD          2.684e-01  7.236e-02   3.710 0.000232 ***
TAX         -1.522e-02  3.754e-03  -4.055 5.85e-05 ***
PTRATIO     -9.902e-01  1.372e-01  -7.215 2.13e-12 ***
B            9.137e-03  2.667e-03   3.427 0.000664 ***
LSTAT       -5.516e-01  4.796e-02 -11.502  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.715 on 477 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.7488,   Adjusted R-squared:  0.7425
F-statistic: 118.5 on 12 and 477 DF,  p-value: < 2.2e-16
```

Diagnostic

✓ Prediction Rred^2

```
> shrinkage(f4)
Original R-square = 0.7487861
10 Fold Cross-Validated R-square = 0.7296037
Change = 0.01918246
```

X:Independent of residual (iid)

```
> durbinWatsonTest(f4)
 lag Autocorrelation D-W Statistic p-value
   1       0.4117052      1.159609       0
 Alternative hypothesis: rho != 0
```

X:Variance should be constant

```
> ncvTest(f4)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 14.19528    Df = 1     p = 0.0001647829
```

✓ Multilinear test:

```
> vif(f4)
    TRACT      CRIM        ZN      CHAS       NOX        RM       DIS       RAD       TAX   PTRATIO         B
3.040787  1.800978  2.240942  1.060669  3.983337  1.822050  3.768163  8.882791  8.860050  1.966482  1.338429
    LSTAT
2.606906
> sqrt(vif(f4))>2
   TRACT     CRIM       ZN     CHAS      NOX       RM      DIS      RAD      TAX  PTRATIO        B    LSTAT
   FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE     TRUE     TRUE    FALSE    FALSE    FALSE
```

The Weighted Least Square

Stepwise all variable into model

reg2_test=lm(log(resid(reg2)^2)~ CRIM + ZN + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT, data = boston)
> reg3=lm(formula = MEDV ~ CRIM + ZN + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT, weights=1/exp(fitted(reg2.test)), data = boston)

R^2_{adj}=0.7462

s^2=2.055

PRESS= 7385.819

$R^2_{prediction}$ =0.7558

```
Call:
lm(formula = MEDV ~ CRIM + ZN + NOX + RM + AGE + DIS + RAD +
    TAX + PTRATIO + B + LSTAT, data = boston, weights = 1/exp(fitted(reg2_test)))

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-7.7603 -1.3114 -0.3272  0.9169  9.6982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.836675   3.908926   6.354 4.90e-10 ***
CRIM         -0.112813   0.046154  -2.444  0.01487 *
ZN            0.033367   0.011489   2.904  0.00385 **
NOX         -11.388681   2.490633  -4.573 6.14e-06 ***
RM            4.217441   0.352214  11.974  < 2e-16 ***
AGE          -0.033759   0.008835  -3.821  0.00015 ***
DIS          -1.034363   0.135244  -7.648 1.13e-13 ***
RAD           0.226624   0.046456   4.878 1.46e-06 ***
TAX          -0.012019   0.002215  -5.425 9.21e-08 ***
PTRATIO      -0.802738   0.088782  -9.042  < 2e-16 ***
B             0.011168   0.002100   5.317 1.62e-07 ***
LSTAT        -0.254513   0.036536  -6.966 1.08e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.055 on 478 degrees of freedom
Multiple R-squared:  0.7518,    Adjusted R-squared:  0.7461
F-statistic: 131.6 on 11 and 478 DF,  p-value: < 2.2e-16
```

Diagnostic

✓ Prediction Rred^2

```
> 1-press/SST
[1] 0.7558487
```

✓ Independent of residual (iid)

```
> durbinWatsonTest(reg3)
 lag Autocorrelation D-W Statistic p-value
   1       0.3713353      1.240233       0
 Alternative hypothesis: rho != 0
```

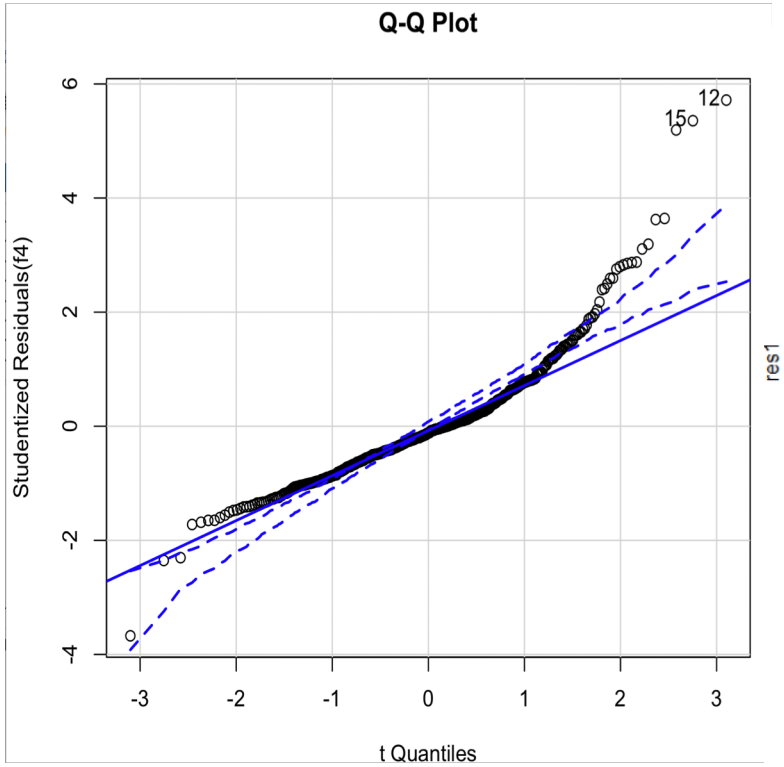
✓ Variance should be constant

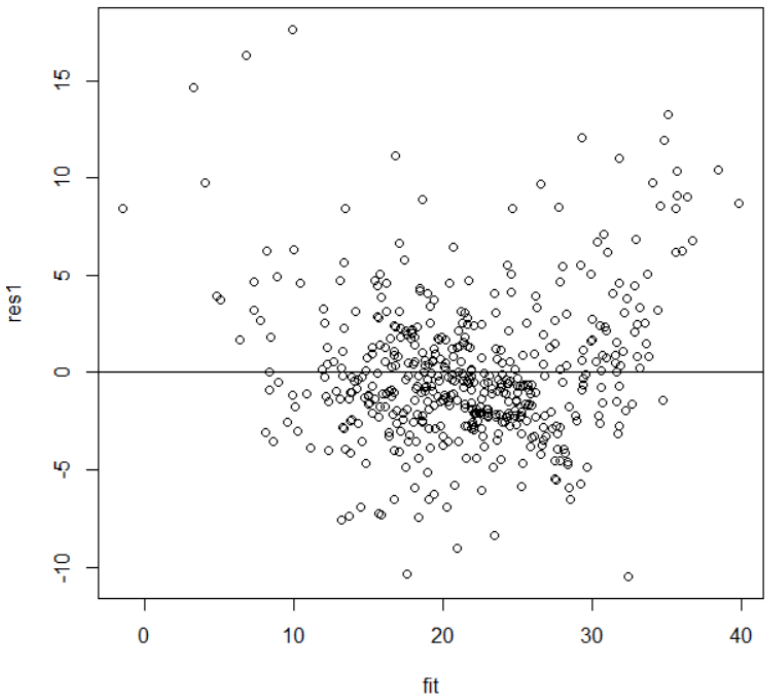
```
> ncvTest(reg3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.8321066    Df = 1     p = 0.3616641
```

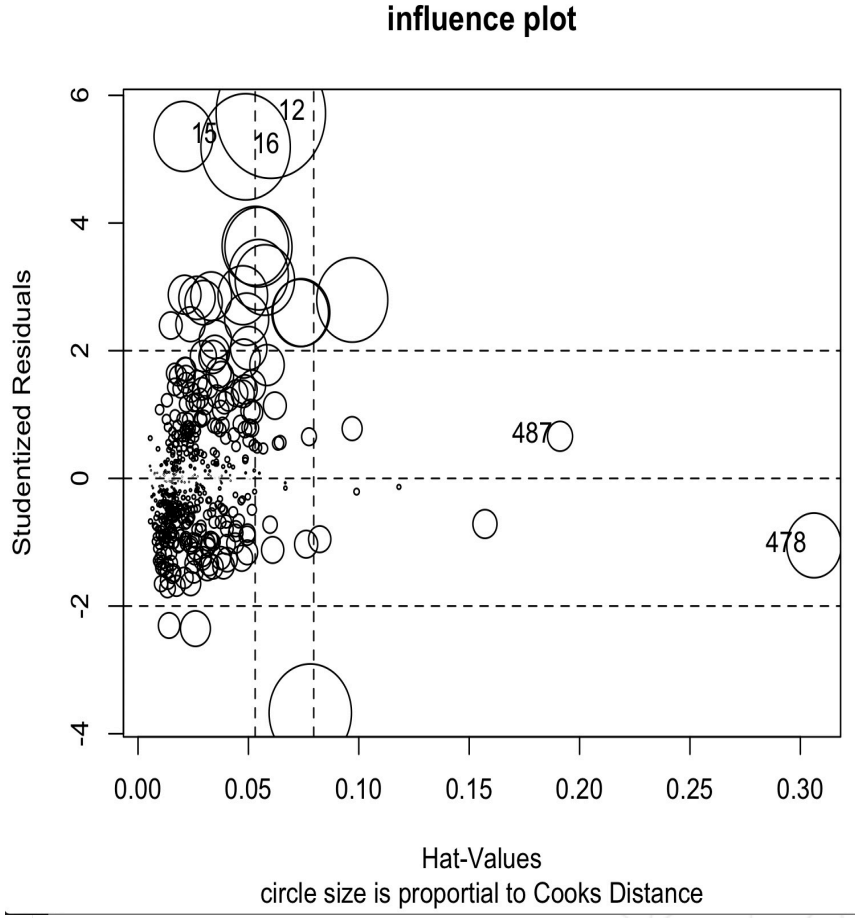
✓ Multi-linear test:

```
> vif(reg3)
    CRIM       ZN      NOX       RM      AGE      DIS      RAD      TAX
2.364084 1.908902 3.594675 1.812342 2.690959 3.458065 5.021425 4.759370
 PTRATIO        B    LSTAT
1.385226 1.443347 2.799633
> sqrt(vif(reg3))>2
   CRIM      ZN     NOX      RM     AGE     DIS     RAD     TAX PTRATIO
  FALSE   FALSE   FALSE   FALSE   FALSE   FALSE    TRUE    TRUE   FALSE
      B   LSTAT
  FALSE   FALSE
```

Diagnostic


Q-Q Plot
Studentized Residuals(f4)
t Quantiles
12
15

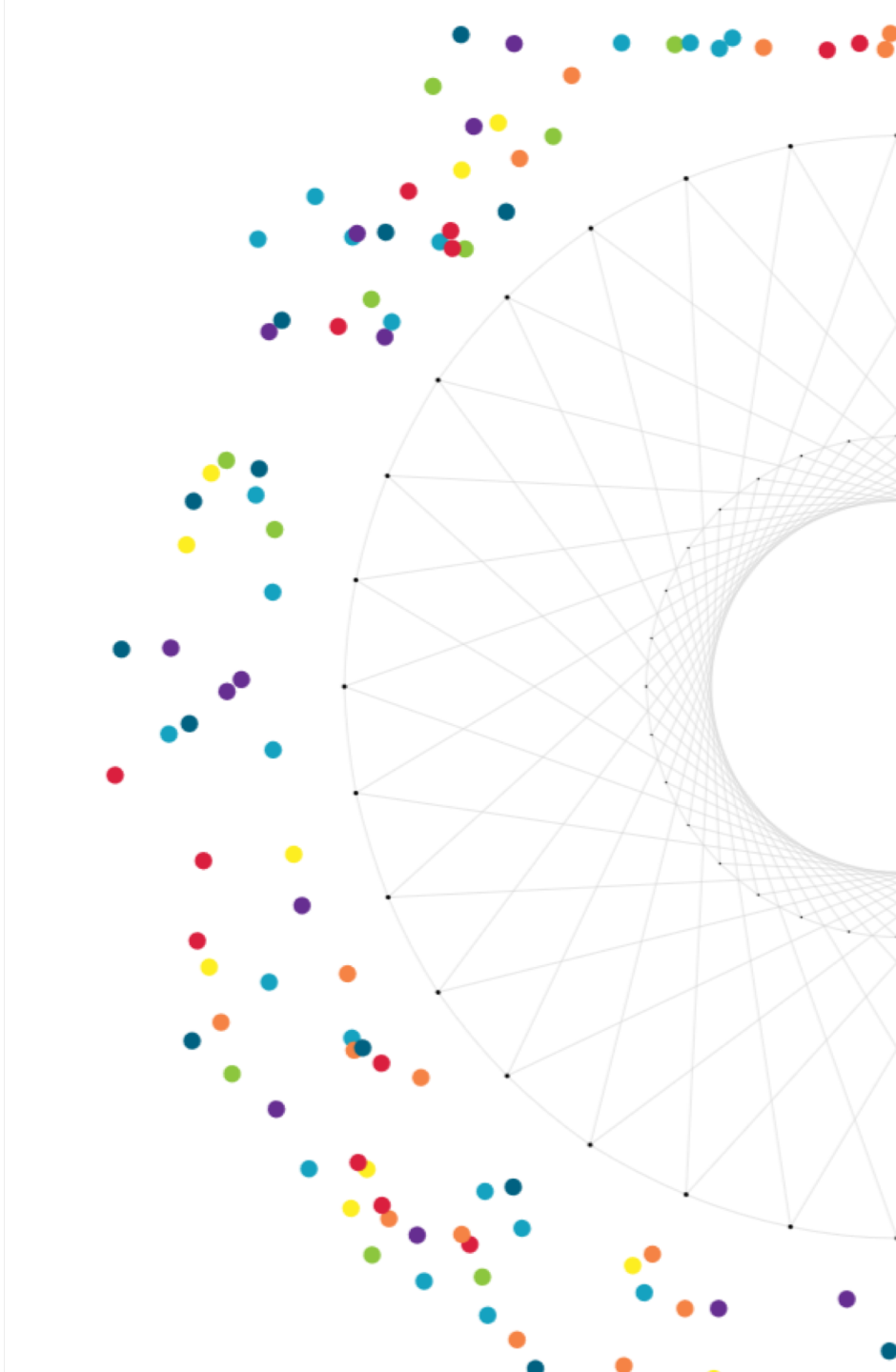

res1
fit


influence plot
Studentized Residuals
Hat-Values
circle size is proportial to Cooks Distance
12
15
16
487
478

$ Housing price
(MEDV)

GROUP A: MEDV<21

GROUP B: MEDV>=21



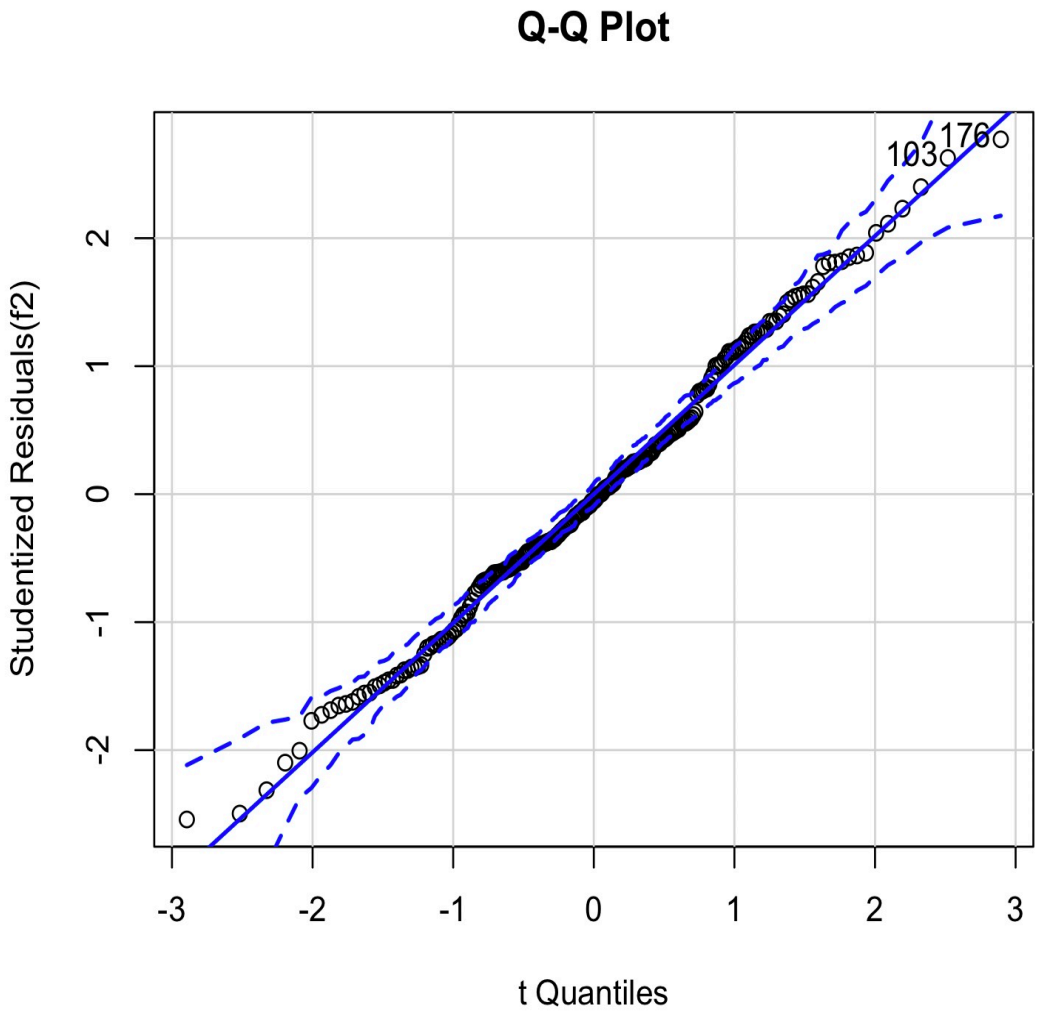
GROUP A: MEDV<21

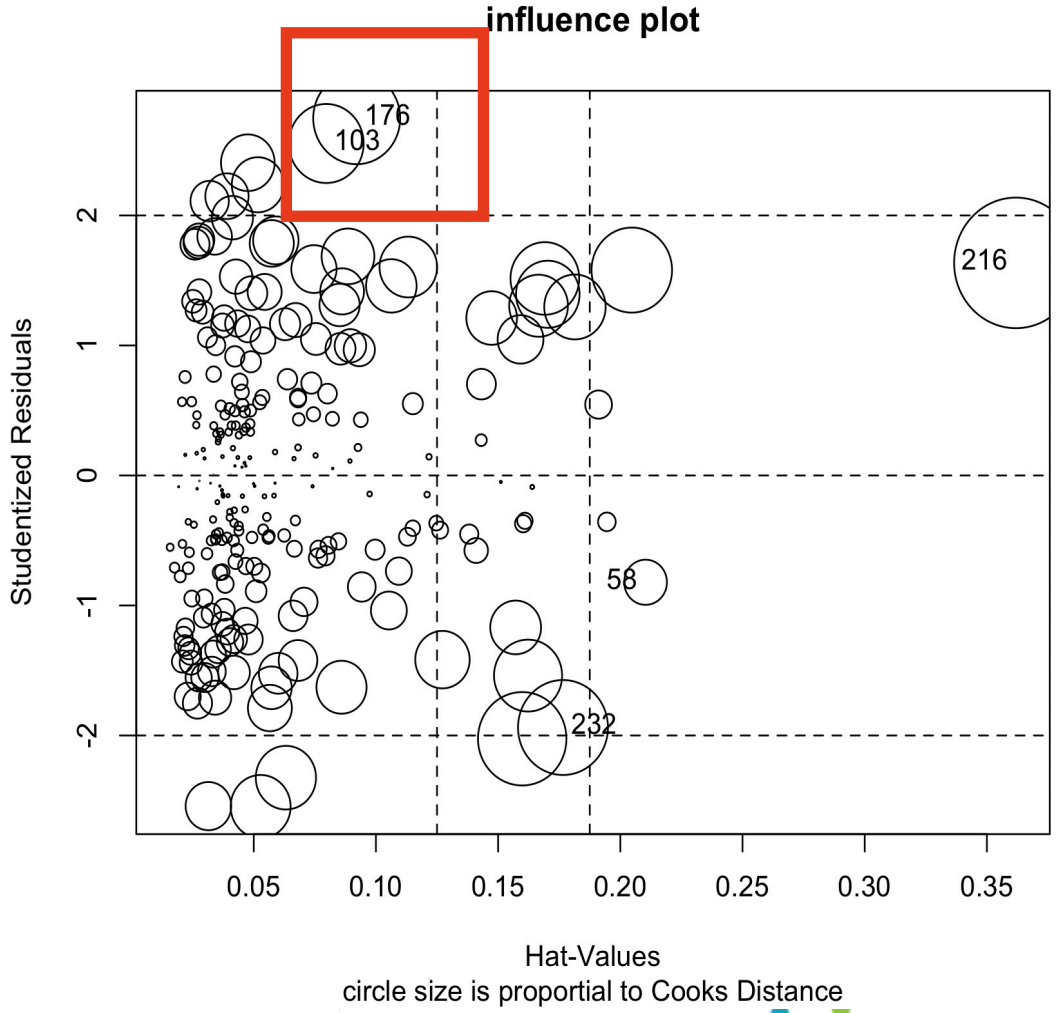
MEDV~CRIM+INDUS+CHAS+NOX+DIS+RAD+TAX+PTRATIO+B+LSTAT

```
> summary(f2)

Call:
lm(formula = MEDV ~ CRIM + INDUS + CHAS + NOX + DIS + RAD + TAX +
    PTRATIO + B + LSTAT, data = poor)

Residuals:
   Min     1Q Median     3Q    Max
-5.050 -1.221 -0.101  1.136  5.378

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.994608   2.846124  11.593  < 2e-16 ***
CRIM        -0.102329   0.014518  -7.048 2.11e-11 ***
INDUS        0.110165   0.041009   2.686 0.007753 **
CHAS         1.027109   0.732447   1.402 0.162181
NOX         -9.450145   2.176564  -4.342 2.12e-05 ***
DIS         -0.208608   0.131233  -1.590 0.113305
RAD          0.089083   0.036628   2.432 0.015780 *
TAX         -0.005964   0.002179  -2.737 0.006694 **
PTRATIO     -0.309446   0.091947  -3.365 0.000896 ***
B            0.005864   0.001212   4.838 2.41e-06 ***
LSTAT       -0.325690   0.025260 -12.894  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.033 on 229 degrees of freedom
Multiple R-squared:  0.7284,  Adjusted R-squared:  0.7165
F-statistic: 61.41 on 10 and 229 DF,  p-value: < 2.2e-16
```

```
> PRESS(f2)
[1] 1054.214
> shrinkage(f2)
Original R-square = 0.7283843
10 Fold Cross-Validated R-square = 0.7029271
Change = 0.02545726
> durbinWatsonTest(f2)
 lag Autocorrelation D-W Statistic p-value
   1       0.2914499      1.403956       0
 Alternative hypothesis: rho != 0
> ncvTest(f2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 25.22704    Df = 1     p = 5.096256e-07
> vif(f2)
    CRIM    INDUS     CHAS      NOX      DIS      RAD      TAX  PTRATIO        B    LSTAT
1.604419 3.763558 1.123934 3.731180 3.840706 7.435063 8.157019 1.562678 1.234089 1.478098
> sqrt(vif(f2))>2
   CRIM   INDUS    CHAS     NOX     DIS     RAD     TAX PTRATIO       B   LSTAT
  FALSE   FALSE   FALSE   FALSE   FALSE    TRUE    TRUE   FALSE   FALSE   FALSE
```

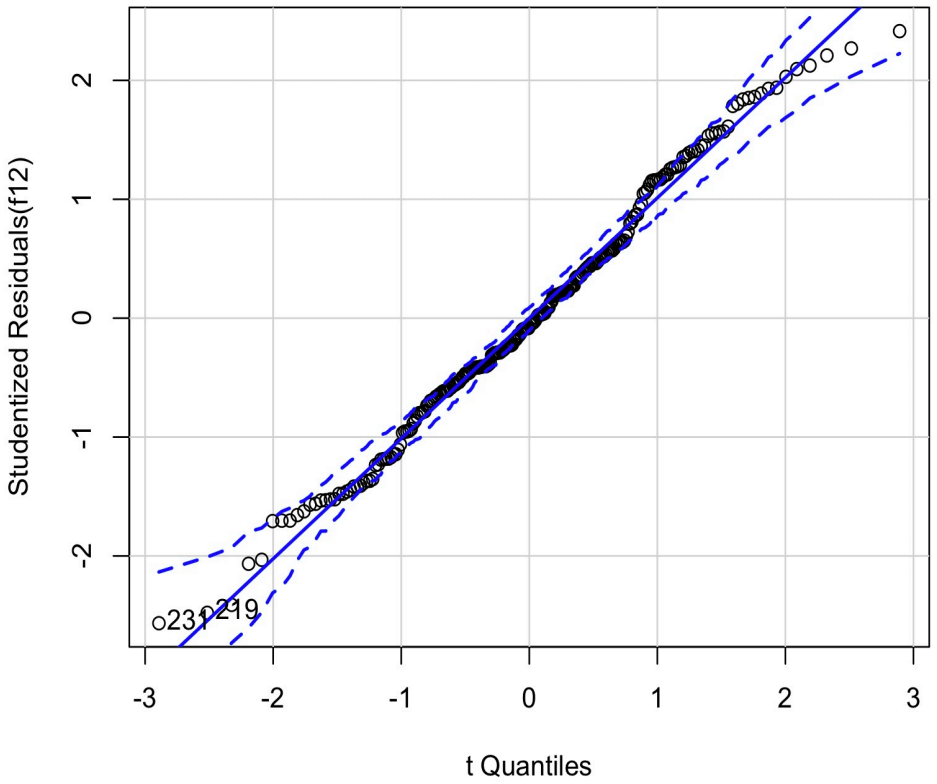
Diagnostic


Q-Q Plot
Studentized Residuals(f2)
t Quantiles
176
103
-3
-2
-1
0
1
2
3


influence plot
Studentized Residuals
Hat-Values
circle size is proportial to Cooks Distance
176
103
216
58
232
0.05
0.10
0.15
0.20
0.25
0.30
0.35

Delete outliers (103 , 176)

```
> summary(f12)

Call:
lm(formula = poor1$MEDV ~ CRIM + INDUS + CHAS + NOX + DIS + RAD +
    TAX + PTRATIO + B + LSTAT, data = poor1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9527 -1.1955 -0.1309  1.1913  4.6242

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.177319   2.767188  11.990  < 2e-16 ***
CRIM        -0.101660   0.014114  -7.203 8.61e-12 ***
INDUS        0.118112   0.039920   2.959  0.00342 **
CHAS         1.206124   0.715044   1.687  0.09302 .
NOX         -9.989697   2.138412  -4.672 5.12e-06 ***
DIS         -0.209264   0.127672  -1.639  0.10258
RAD          0.091557   0.035646   2.569  0.01085 *
TAX         -0.006026   0.002119  -2.844  0.00486 **
PTRATIO     -0.285528   0.089727  -3.182  0.00167 **
B            0.005490   0.001183   4.642 5.84e-06 ***
LSTAT       -0.346868   0.025224 -13.751  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.977 on 227 degrees of freedom
Multiple R-squared:  0.745,    Adjusted R-squared:  0.7337
F-statistic: 66.31 on 10 and 227 DF,  p-value: < 2.2e-16
```

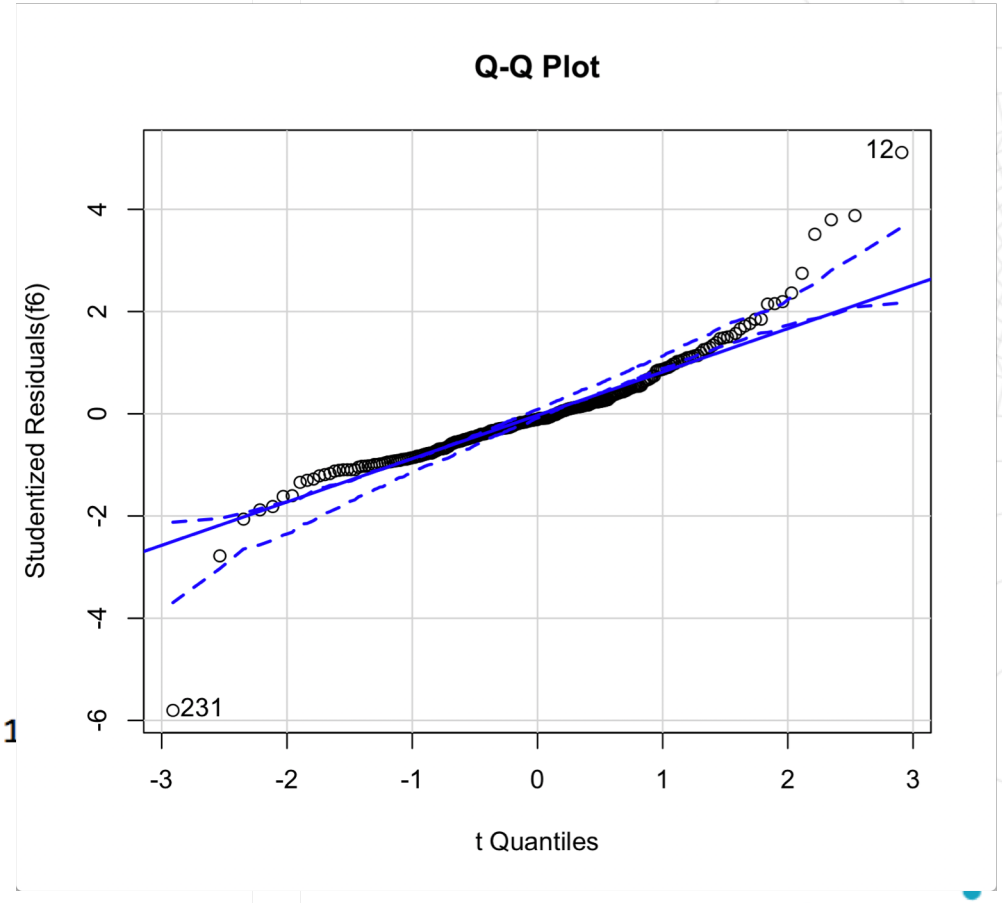


GROUP B: MEDV>=21

MEDV~TRACT+CRIM+ZN+CHAS+NOX+RM+DIS+RAD+TAX+PTRATIO+LSTAT

```
> summary(f6)

Call:
lm(formula = MEDV ~ TRACT + CRIM + ZN + CHAS + NOX + RM + DIS +
    RAD + TAX + PTRATIO + LSTAT, data = rich)

Residuals:
     Min       1Q   Median       3Q      Max
-24.4150  -2.7307  -0.5211   2.0321  22.1999

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.165e+01  8.383e+00   3.776 0.000201 ***
TRACT       -1.028e-03  3.871e-04  -2.655 0.008456 **
CRIM         7.005e-01  2.720e-01   2.576 0.010607 *
ZN           3.262e-02  1.691e-02   1.929 0.054906 .
CHAS         1.549e+00  1.051e+00   1.473 0.142096
NOX         -1.432e+01  6.607e+00  -2.168 0.031157 *
RM           5.595e+00  5.653e-01   9.897  < 2e-16 ***
DIS         -1.510e+00  2.794e-01  -5.404 1.56e-07 ***
RAD          3.480e-01  1.350e-01   2.578 0.010528 *
TAX         -1.948e-02  5.857e-03  -3.325 0.001021 **
PTRATIO     -8.292e-01  2.126e-01  -3.901 0.000124 ***
LSTAT       -6.613e-01  1.017e-01  -6.504 4.48e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.864 on 241 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:   0.67    Adjusted R-squared:  0.6549
F-statistic: 44.48 on 11 and 241 DF,  p-value: < 2.2e-16
```

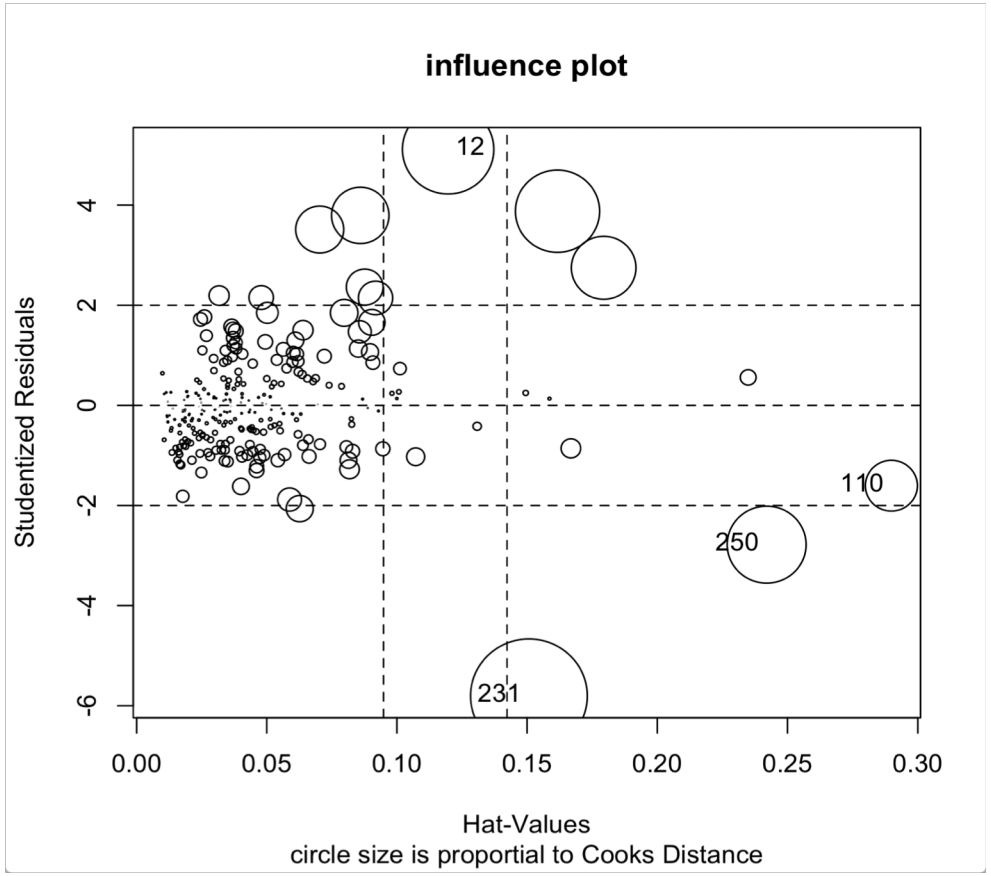
```
> shrinkage(f6)
Original R-square =  0.669997
10 Fold Cross-Validated R-square = 0.5700064
Change = 0.09999061
```



Diagnostic

```
> durbinWatsonTest(f6)
 lag Autocorrelation D-W Statistic p-value
   1       0.3417846       1.298136       0
 Alternative hypothesis: rho != 0
> ncvTest(f6)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 21.05538    Df = 1     p = 4.46199e-06
> vif(f6)
    TRACT      CRIM        ZN      CHAS       NOX        RM       DIS       RAD       TAX   PTRATIO     LSTAT
2.081978  3.835879  2.389638  1.090067  3.695313  1.712441  3.498378  7.114647  5.631840  2.049461  1.688264
> sqrt(vif(f6))>2
   TRACT     CRIM       ZN     CHAS      NOX       RM      DIS      RAD      TAX  PTRATIO    LSTAT
   FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE     TRUE     TRUE    FALSE    FALSE
```

```
> PRESS(f6)
[1] 6874.42
```



Delete outliers (12 , 231)

```
> summary(f10)

Call:
lm(formula = rich1$MEDV ~ TRACT + CRIM + ZN + CHAS + RM + AGE +
    DIS + RAD + TAX + PTRATIO + B + LSTAT, data = rich1)

Residuals:
   Min     1Q Median     3Q    Max
-9.784 -2.524 -0.441  1.918 18.153

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.4914727  7.2622884   2.409 0.016778 *
TRACT       -0.0011427  0.0003394  -3.366 0.000888 ***
CRIM         0.4624392  0.2857050   1.619 0.106860
ZN           0.0325554  0.0148310   2.195 0.029124 *
CHAS         2.7594806  0.9359546   2.948 0.003513 **
RM           7.4470069  0.5460063  13.639  < 2e-16 ***
AGE         -0.0400236  0.0158516  -2.525 0.012224 *
DIS         -1.4028202  0.2377095  -5.901 1.23e-08 ***
RAD          0.3056681  0.1220024   2.505 0.012900 *
TAX         -0.0188882  0.0050880  -3.712 0.000256 ***
PTRATIO     -0.5898539  0.1826811  -3.229 0.001418 **
B           -0.0221388  0.0128093  -1.728 0.085225 .
LSTAT       -0.4951010  0.0937103  -5.283 2.86e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.269 on 238 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.7416,  Adjusted R-squared:  0.7285
F-statistic: 56.91 on 12 and 238 DF,  p-value: < 2.2e-16
```

```
> PRESS(f10)
[1] 5387.718

> shrinkage(f10)
Original R-square = 0.7415788
10 Fold Cross-Validated R-square = 0.6794631
Change = 0.06211567
>
> durbinWatsonTest(f10)
 lag Autocorrelation D-W Statistic p-value
   1        0.259537      1.455725       0
 Alternative hypothesis: rho != 0
>
> ncvTest(f10)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.221792    Df = 1     p = 0.2690093
```

BEST SUBSET

Number of Predictors	R-sq	R-sq(adj)	R-sq(pred)	Mallows' Cp	S	TRACT	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	RAD	TAX	B	LSTAT
1	55.0	54.9	54.5	267.0	6.242												X
1	48.5	48.4	47.7	375.6	6.676						X						
2	64.4	64.3	63.5	110.4	5.551						X						X
2	57.1	57.0	56.4	232.4	6.094								X				X
3	65.6	65.4	64.6	92.4	5.462						X					X	X
3	65.5	65.3	64.2	95.2	5.476						X				X		X
4	67.7	67.4	66.4	60.0	5.302						X		X		X		X
4	67.0	66.7	65.8	71.7	5.359						X		X			X	X
5	69.5	69.1	68.1	32.7	5.161			X			X		X		X		X
5	68.4	68.1	67.0	50.1	5.248						X		X		X	X	X
6	70.2	69.8	68.7	22.5	5.105			X			X		X		X	X	X
6	69.9	69.5	68.4	27.4	5.130			X			X		X	X	X		X
7	70.7	70.3	69.1	15.8	5.065			X			X		X	X	X	X	X
7	70.5	70.1	69.0	19.2	5.083		X	X			X		X	X	X		X
8	71.2	70.7	69.5	10.0	5.030		X	X			X		X	X	X	X	X
8	70.9	70.4	69.2	14.5	5.053			X		X	X		X	X	X	X	X
9	71.4	70.9	69.6	7.8	5.013		X	X		X	X		X	X	X	X	X
9	71.2	70.7	69.4	11.5	5.033		X	X	X		X		X	X	X	X	X
10	71.5	70.9	69.6	9.0	5.014	X	X	X		X	X		X	X	X	X	X
10	71.4	70.8	69.5	9.8	5.018		X	X	X	X	X		X	X	X	X	X
11	71.5	70.8	69.5	11.0	5.019	X	X	X	X	X	X		X	X	X	X	X
11	71.5	70.8	69.4	11.0	5.019	X	X	X		X	X	X	X	X	X	X	X
12	71.5	70.7	69.3	13.0	5.025	X	X	X	X	X	X	X	X	X	X	X	X

Minitab(12 variables) :

TRACT

CRIM

ZN

INDUS

CHAS

NOX

RM

AGE

DIS

RAD

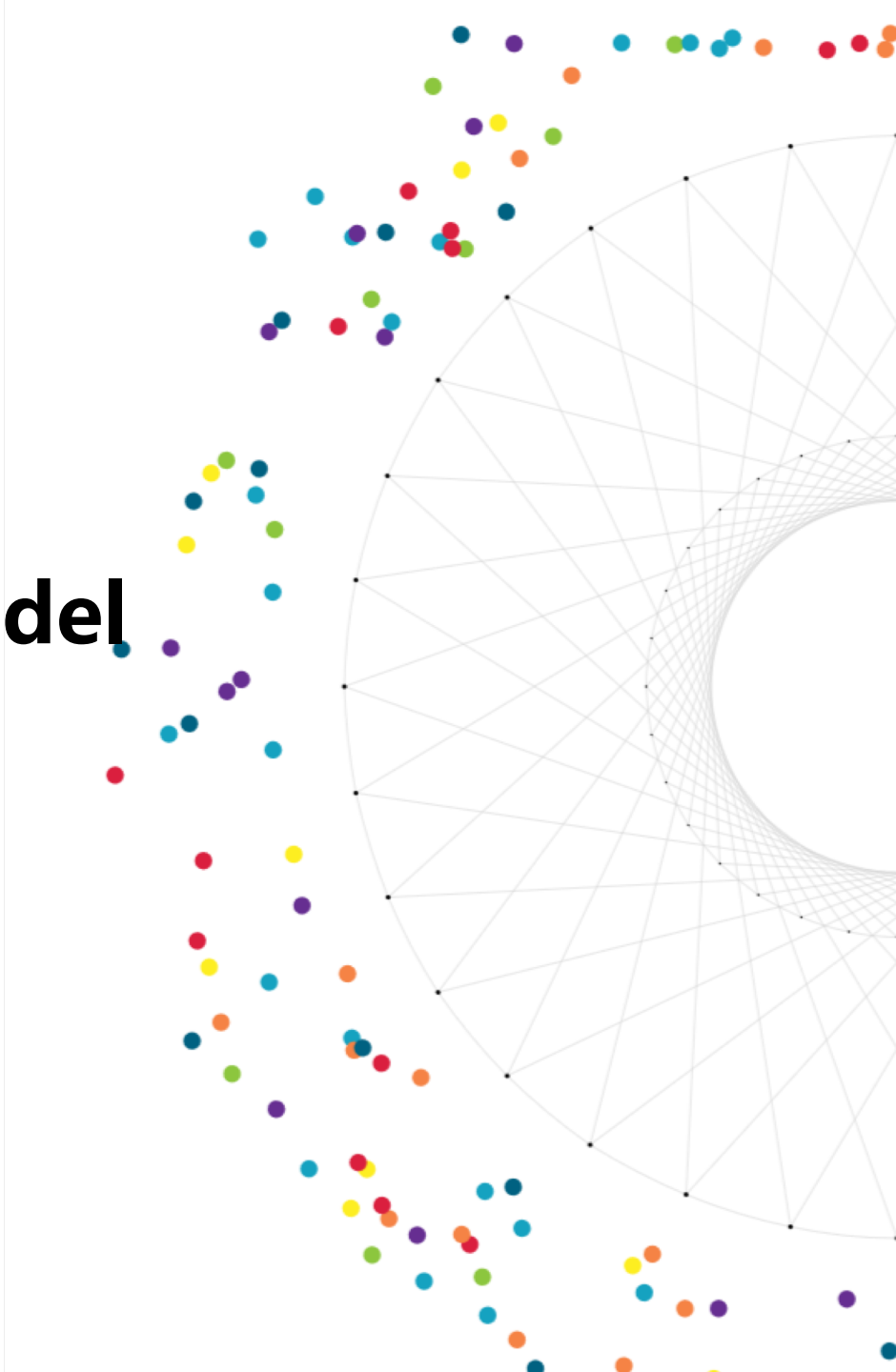
TAX

PTRATIO

B

LSTAT

3.2 Quadratic Model



Multiple linear regression

Housing price
(MEDV)

POOR (MEDV<=21)

RICH (MEDV>21)

POOR (MEDV<=21)

Interaction items

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.505e+01  4.372e+01   1.945 0.053140 .
TRACT       -6.468e-03  2.152e-03  -3.006 0.002983 **
CRIM         4.549e-01  2.549e-01   1.784 0.075869 .
ZN           2.610e-02  1.508e-02   1.730 0.085115 .
INDUS       -1.625e+00  5.128e-01  -3.169 0.001768 **
CHAS        -1.191e+00  9.591e-01  -1.242 0.215838
NOX         -1.216e+02  3.131e+01  -3.883 0.000140 ***
RM          -1.999e+01  7.244e+00  -2.760 0.006322 **
AGE          6.099e-01  1.142e-01   5.338 2.52e-07 ***
DIS          3.382e+00  1.975e+00   1.713 0.088314 .
RAD          4.308e-01  1.757e-01   2.452 0.015067 *
B            6.615e-02  1.308e-02   5.059 9.48e-07 ***
LSTAT        1.968e+00  3.654e-01   5.386 2.00e-07 ***
TAX          2.086e-02  8.298e-03   2.514 0.012724 *
PTRATIO     -1.737e+00  1.600e+00  -1.086 0.278936
TRACT:RM     1.062e-03  3.290e-04   3.227 0.001461 **
TRACT:TAX   -3.946e-06  1.736e-06  -2.272 0.024126 *
TRACT:LSTAT  1.225e-04  2.202e-05   5.563 8.40e-08 ***
CRIM:AGE    -3.485e-03  1.966e-03  -1.772 0.077839 .
CRIM:DIS    -9.638e-02  5.704e-02  -1.690 0.092614 .
CRIM:B      -2.230e-04  9.123e-05  -2.444 0.015375 *
INDUS:NOX    3.457e+00  9.111e-01   3.795 0.000196 ***
INDUS:DIS    1.259e-01  3.393e-02   3.710 0.000268 ***
INDUS:TAX   -7.159e-04  2.360e-04  -3.034 0.002734 **
INDUS:LSTAT -2.345e-02  7.390e-03  -3.173 0.001744 **
CHAS:RAD     1.428e-01  7.381e-02   1.935 0.054403 .
NOX:RM       1.200e+01  4.014e+00   2.991 0.003129 **
NOX:DIS      5.459e+00  2.317e+00   2.357 0.019407 *
NOX:RAD     -1.082e+00  2.781e-01  -3.890 0.000136 ***
NOX:LSTAT   -1.002e+00  3.220e-01  -3.111 0.002137 **
RM:PTRATIO   7.268e-01  2.558e-01   2.842 0.004952 **
RM:LSTAT    -1.717e-01  3.885e-02  -4.419 1.62e-05 ***
AGE:RAD      2.885e-03  1.387e-03   2.081 0.038732 *
AGE:PTRATIO -2.581e-02  5.489e-03  -4.702 4.78e-06 ***
AGE:B       -4.286e-04  1.130e-04  -3.794 0.000196 ***
DIS:PTRATIO -1.777e-01  6.049e-02  -2.938 0.003685 **
DIS:B       -5.995e-03  2.347e-03  -2.554 0.011386 *
DIS:LSTAT   -1.314e-01  3.070e-02  -4.280 2.90e-05 ***
B:LSTAT     -3.882e-04  1.982e-04  -1.958 0.051618 .
---
```

POOR (MEDV<=21)

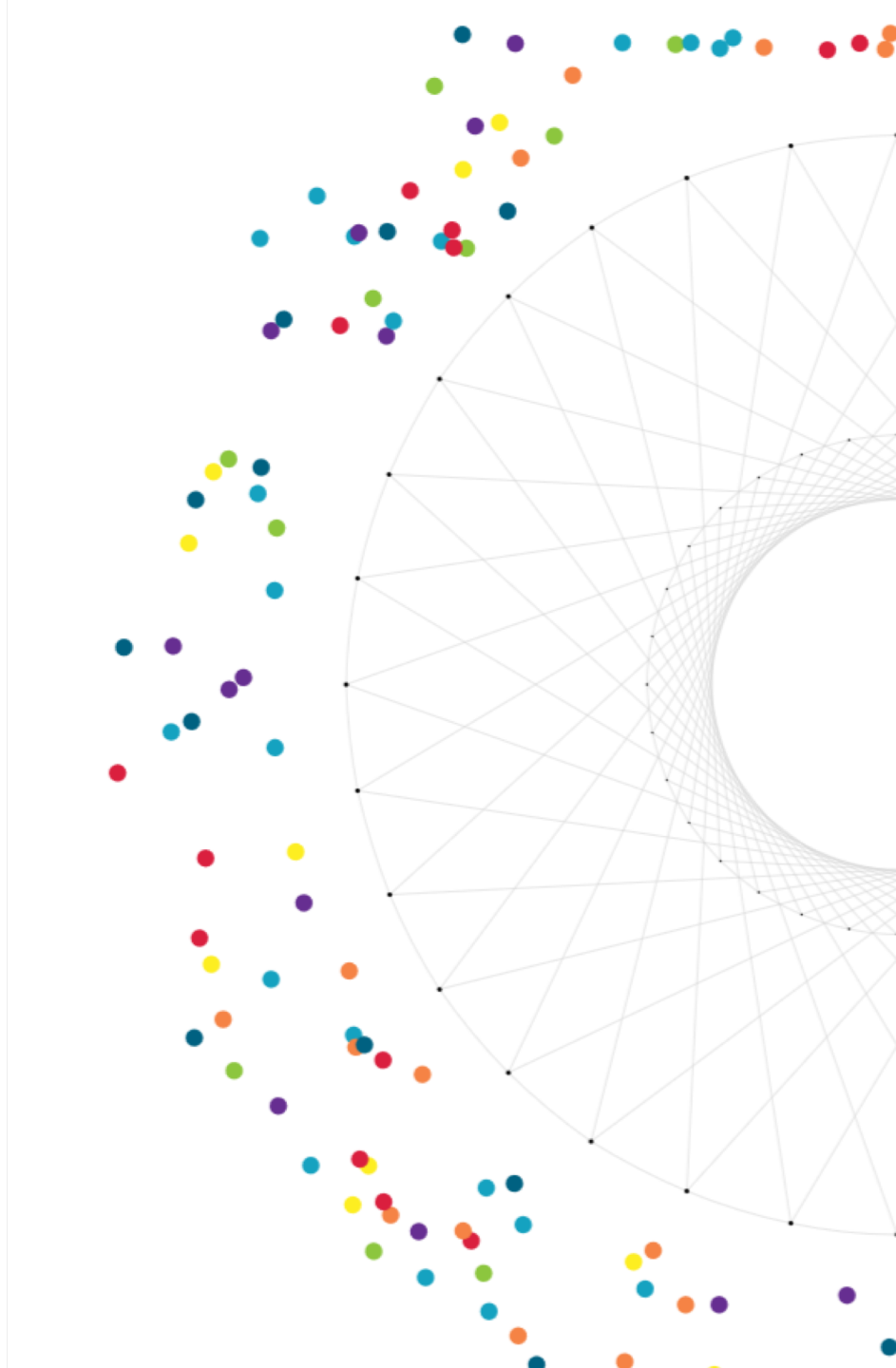
Interaction items

R^2_{adj}=0.8226

s^2=2.271752

PRESS=788.6653

$R^2_{prediction}$ =0.7737665



POOR (MEDV<=21)

Interaction items

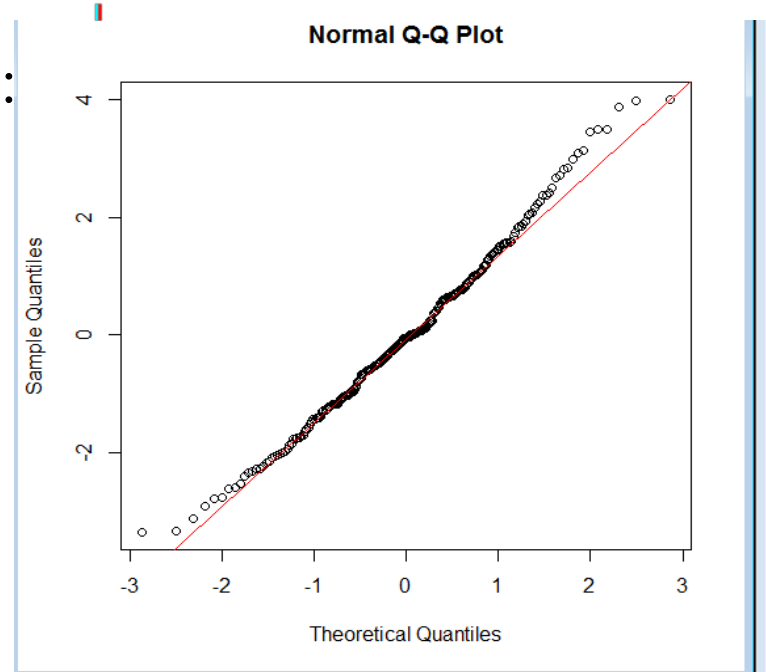
✓independent :

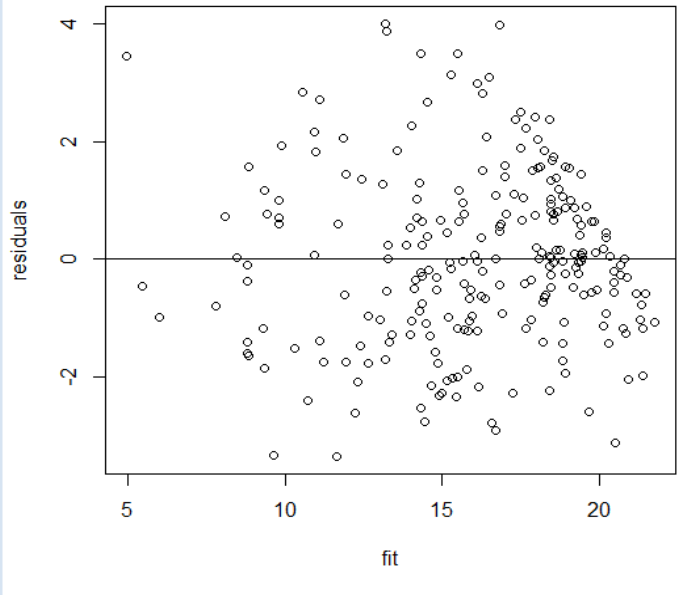
```
> durbinWatsonTest(m5)
 lag Autocorrelation D-W Statistic p-value
   1       0.09369881       1.810084    0.09
 Alternative hypothesis: rho != 0
```

✓same variance:

```
> ncvTest(m5)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 10.23603    Df = 1     p = 0.001377239
```

✓normal distribution:





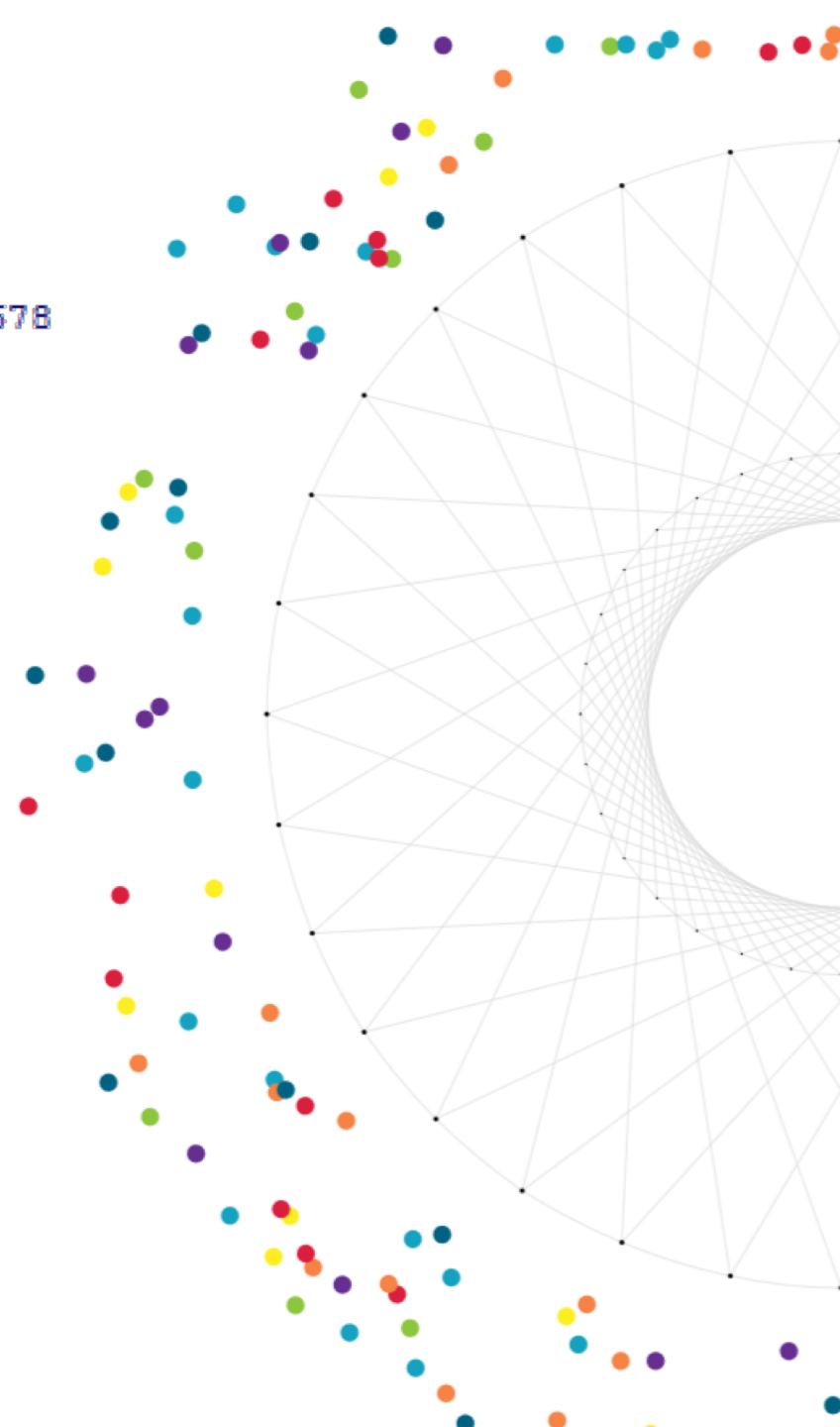
The Weighted Least Square

```
Residual standard error: 2.083 on 197 degrees of freedom
Multiple R-squared:  0.891,     Adjusted R-squared:  0.8678
F-statistic: 38.34 on 42 and 197 DF,  p-value: < 2.2e-16

> press=PRESS(lm.test3)
> SST=sum((poor[,1]-mean(poor[,1]))^2)
> pRsquare=1-press/SST
> pRsquare
[1] 0.77946
```

S^2=2.091565

```
> ncvTest(m5)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 10.23603    Df = 1     p = 0.001377239
```

```
> ncvTest(lm.test3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1608822    Df = 1     p = 0.6883456
```



Rich (MEDV>21)

Interaction items

R^2_{adj}=0.8002

s^2=2.646

PRESS=2192.921

$R^2_{prediction}$ =0.7271517

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.795e+02  1.213e+02  -1.479  0.14235
TRACT        1.010e-03  7.748e-04   1.303  0.19567
CRIM        -6.417e+00  2.996e+00  -2.142  0.03476 *
INDUS       -4.358e-02  3.116e-01  -0.140  0.88907
RM           3.444e+01  1.755e+01   1.962  0.05267 .
AGE          6.794e-01  1.388e-01   4.895 4.07e-06 ***
DIS         -9.778e-01  2.228e-01  -4.389 2.97e-05 ***
RAD          2.085e-01  1.732e-01   1.204  0.23174
TAX         -2.153e-02  7.167e-03  -3.004  0.00342 **
PTRATIO      8.003e+00  1.850e+00   4.325 3.79e-05 ***
B           -3.192e-02  2.607e-01  -0.122  0.90282
LSTAT        6.419e-01  3.830e-01   1.676  0.09705 .
TRACT:LSTAT -3.076e-04  1.141e-04  -2.697  0.00830 **
CRIM:INDUS   3.090e-01  1.698e-01   1.819  0.07202 .
INDUS:TAX    2.290e-04  9.463e-04   0.242  0.80932
RM:AGE      -1.026e-01  1.996e-02  -5.140 1.49e-06 ***
RM:PTRATIO  -1.251e+00  2.594e-01  -4.823 5.43e-06 ***
RM:B         2.933e-03  3.774e-02   0.078  0.93822
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.646 on 94 degrees of freedom
Multiple R-squared:  0.8308,    Adjusted R-squared:  0.8002
F-statistic: 27.15 on 17 and 94 DF,  p-value: < 2.2e-16
```

Rich (MEDV>21)

Interaction items

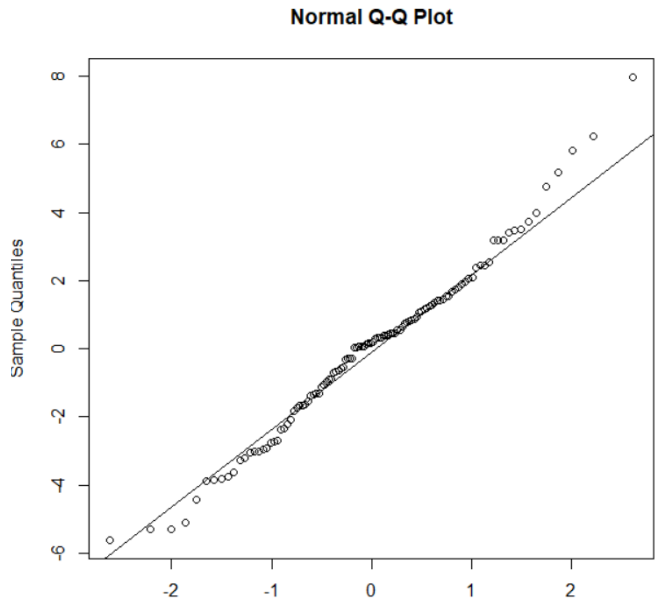
✓independent :

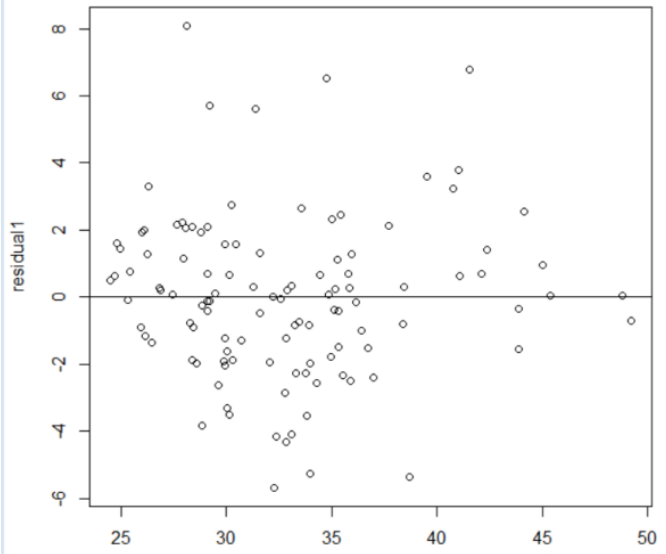
```
> durbinWatsonTest(model1)
 lag Autocorrelation D-W Statistic p-value
   1       0.2297439       1.540131   0.012
 Alternative hypothesis: rho != 0
>
```

✓same variance:

```
> ncvTest(model1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.2429704    Df = 1     p = 0.6220689
```

✓normal distribution:





Rich (MEDV>21)

The Weighted Least Square

R^2_{adj}=0.9137

s^2=1.889

PRESS=1117.117

$R^2_{prediction}$ =0.7022604

```
Weighted Residuals:
    Min      1Q  Median      3Q     Max
-4.3990 -1.0965 -0.0795  0.9095  4.6883

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.969e+02  1.124e+02  -1.751 0.083306 .
TRACT        4.747e-04  6.545e-04   0.725 0.470182
CRIM        -5.640e+00  2.222e+00  -2.539 0.012862 *
INDUS       -6.465e-01  5.408e-01  -1.195 0.235072
RM           3.948e+01  1.667e+01   2.368 0.020032 *
AGE          5.988e-01  1.040e-01   5.755 1.21e-07 ***
DIS         -1.031e+00  1.395e-01  -7.391 7.59e-11 ***
RAD          2.894e-01  1.381e-01   2.096 0.038908 *
TAX         -2.999e-02  8.764e-03  -3.422 0.000942 ***
PTRATIO      7.415e+00  1.429e+00   5.190 1.31e-06 ***
B            4.683e-02  2.583e-01   0.181 0.856574
LSTAT        7.402e-01  5.098e-01   1.452 0.150006
TRACT:LSTAT -2.999e-04  1.332e-04  -2.252 0.026805 *
CRIM:INDUS   1.094e-01  1.851e-01   0.591 0.556213
INDUS:TAX    2.504e-03  1.925e-03   1.301 0.196755
RM:AGE      -9.410e-02  1.489e-02  -6.321 1.01e-08 ***
RM:PTRATIO  -1.167e+00  2.007e-01  -5.817 9.27e-08 ***
RM:B        -1.325e-02  3.833e-02  -0.346 0.730357
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.889 on 89 degrees of freedom
Multiple R-squared:  0.9276,    Adjusted R-squared:  0.9137
F-statistic: 67.06 on 17 and 89 DF,  p-value: < 2.2e-16
```

Rich (MEDV>21)

The Weighted Least Square

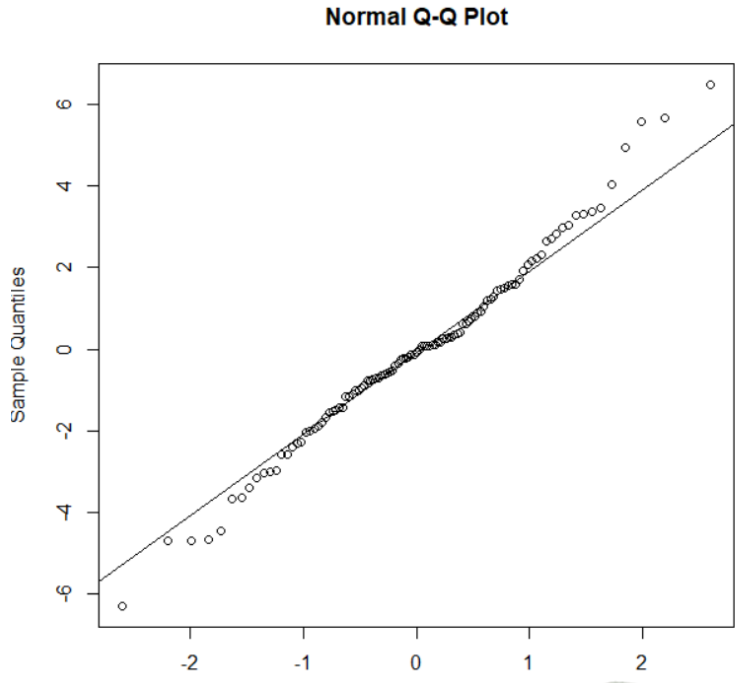
✓independent :

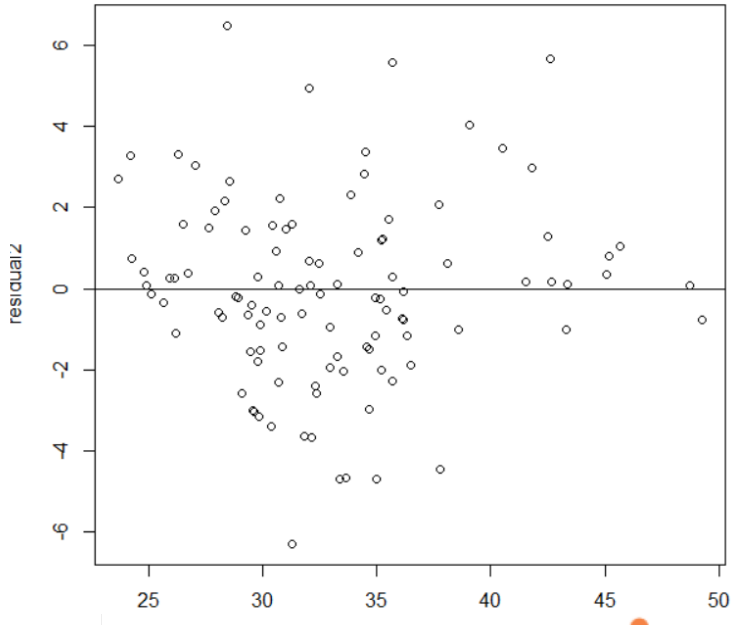
```
> durbinWatsonTest(lm.test4)
 lag Autocorrelation D-W Statistic p-value
   1       0.1857541       1.627525   0.052
 Alternative hypothesis: rho != 0
```

✓same variance:

```
> ncvTest(lm.test4)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.3213137    Df = 1     p = 0.5708193
>
```

✓normal distribution:





Conclusion

PART FOUR

	models	R^2 adj	s^2	PRESS	R^2 pred	P-value (Independent test of residual)	P-value (non-constant variance test)
Complete data	Stepwise: MEDV~TRACT+CRIM+ZN+CHAS+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT	0.745	4.71	7258.276	0.7296	0 reject	0.00016 reject
	Weight least squares model	0.7462	2.055	7385.819	0.7558	0 reject	0.912 accept
poor	stepwise: MEDV~CRIM+INDUS+CHAS+NOX+DIS+RAD+TAX+PTRATIO+B+LSTAT	0.7337	1.977	987.4527	0.7113	0 reject	0.0000005 reject
	Quadratic model	0.8226	2.271752	788.6653	0.7737665	0.09 accept	0.0013 reject
	Weighted least squares model	0.8678	2.091565	812,233	0.77946	0.215 accept	0.6883456 accept
rich	Stepwise: MEDV~TRACT+CRIM+ZN+CHAS+NOX+RM+DIS+RAD+TAX+PTRATIO+LSTAT	0.7285	4.269	5387.718	0.679	0 reject	0.269 accept
	Quadratic model	0.8002	2.646	2192.921	0.7271517	0.012 reject	0.622 accept
	Weighted least squares model	0.9137	1.889	1117.117	0.7022604	0.052 accept	0.5708193 accept

POOR (MEDV<=21)

Interaction items

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.505e+01  4.372e+01   1.945 0.053140 .
TRACT       -6.468e-03  2.152e-03  -3.006 0.002983 **
CRIM         4.549e-01  2.549e-01   1.784 0.075869 .
ZN           2.610e-02  1.508e-02   1.730 0.085115 .
INDUS       -1.625e+00  5.128e-01  -3.169 0.001768 **
CHAS        -1.191e+00  9.591e-01  -1.242 0.215838
NOX         -1.216e+02  3.131e+01  -3.883 0.000140 ***
RM          -1.999e+01  7.244e+00  -2.760 0.006322 **
AGE          6.099e-01  1.142e-01   5.338 2.52e-07 ***
DIS          3.382e+00  1.975e+00   1.713 0.088314 .
RAD          4.308e-01  1.757e-01   2.452 0.015067 *
B            6.615e-02  1.308e-02   5.059 9.48e-07 ***
LSTAT        1.968e+00  3.654e-01   5.386 2.00e-07 ***
TAX          2.086e-02  8.298e-03   2.514 0.012724 *
PTRATIO     -1.737e+00  1.600e+00  -1.086 0.278936
TRACT:RM     1.062e-03  3.290e-04   3.227 0.001461 **
TRACT:TAX   -3.946e-06  1.736e-06  -2.272 0.024126 *
TRACT:LSTAT  1.225e-04  2.202e-05   5.563 8.40e-08 ***
CRIM:AGE    -3.485e-03  1.966e-03  -1.772 0.077839 .
CRIM:DIS    -9.638e-02  5.704e-02  -1.690 0.092614 .
CRIM:B      -2.230e-04  9.123e-05  -2.444 0.015375 *
INDUS:NOX    3.457e+00  9.111e-01   3.795 0.000196 ***
INDUS:DIS    1.259e-01  3.393e-02   3.710 0.000268 ***
INDUS:TAX   -7.159e-04  2.360e-04  -3.034 0.002734 **
INDUS:LSTAT -2.345e-02  7.390e-03  -3.173 0.001744 **
CHAS:RAD     1.428e-01  7.381e-02   1.935 0.054403 .
NOX:RM       1.200e+01  4.014e+00   2.991 0.003129 **
NOX:DIS      5.459e+00  2.317e+00   2.357 0.019407 *
NOX:RAD     -1.082e+00  2.781e-01  -3.890 0.000136 ***
NOX:LSTAT   -1.002e+00  3.220e-01  -3.111 0.002137 **
RM:PTRATIO   7.268e-01  2.558e-01   2.842 0.004952 **
RM:LSTAT    -1.717e-01  3.885e-02  -4.419 1.62e-05 ***
AGE:RAD      2.885e-03  1.387e-03   2.081 0.038732 *
AGE:PTRATIO -2.581e-02  5.489e-03  -4.702 4.78e-06 ***
AGE:B       -4.286e-04  1.130e-04  -3.794 0.000196 ***
DIS:PTRATIO -1.777e-01  6.049e-02  -2.938 0.003685 **
DIS:B       -5.995e-03  2.347e-03  -2.554 0.011386 *
DIS:LSTAT   -1.314e-01  3.070e-02  -4.280 2.90e-05 ***
B:LSTAT     -3.882e-04  1.982e-04  -1.958 0.051618 .
---
```

Rich (MEDV>21)

The Weighted Least Square

R^2_{adj}=0.9137

s^2=1.889

PRESS=1117.117

$R^2_{prediction}$ =0.7022604

```
Weighted Residuals:
    Min      1Q  Median      3Q     Max
-4.3990 -1.0965 -0.0795  0.9095  4.6883

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.969e+02  1.124e+02  -1.751 0.083306 .
TRACT        4.747e-04  6.545e-04   0.725 0.470182
CRIM        -5.640e+00  2.222e+00  -2.539 0.012862 *
INDUS       -6.465e-01  5.408e-01  -1.195 0.235072
RM           3.948e+01  1.667e+01   2.368 0.020032 *
AGE          5.988e-01  1.040e-01   5.755 1.21e-07 ***
DIS         -1.031e+00  1.395e-01  -7.391 7.59e-11 ***
RAD          2.894e-01  1.381e-01   2.096 0.038908 *
TAX         -2.999e-02  8.764e-03  -3.422 0.000942 ***
PTRATIO      7.415e+00  1.429e+00   5.190 1.31e-06 ***
B            4.683e-02  2.583e-01   0.181 0.856574
LSTAT        7.402e-01  5.098e-01   1.452 0.150006
TRACT:LSTAT -2.999e-04  1.332e-04  -2.252 0.026805 *
CRIM:INDUS   1.094e-01  1.851e-01   0.591 0.556213
INDUS:TAX    2.504e-03  1.925e-03   1.301 0.196755
RM:AGE      -9.410e-02  1.489e-02  -6.321 1.01e-08 ***
RM:PTRATIO  -1.167e+00  2.007e-01  -5.817 9.27e-08 ***
RM:B        -1.325e-02  3.833e-02  -0.346 0.730357
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.889 on 89 degrees of freedom
Multiple R-squared:  0.9276,    Adjusted R-squared:  0.9137
F-statistic: 67.06 on 17 and 89 DF,  p-value: < 2.2e-16
```

Problem?

Variance not constant

Autocorrelation

Further improvement

Weighted least squares

Difference transformation method

Reference

PART FIVE

张淑玲，庞进丽，最小二乘法原理在计量测试中的应用[J]，商丘职业技术学院学报，2008(05)

代大山，测量系统的线性分析[J]，电子质量.，2004(06)

王敏，残差分析在统计中的应用[J]，江苏统计， 2000(08)

沈其君，SAS统计分析[M]，东南大学出版社，2001

张宇山，多元线性回归分析的实例研究[J]，科技信息,2009(09):54-56.

王学仁,王松桂，实用多元统计分析[M]. 上海科学技术出版社，1990

回归分析及其试验设计[M]. 上海教育出版社 ，上海师范大学数学系概率统计教研组 编, 1978

赵玉新.多元线性回归分析中自变量的筛选[J].中国城市经济,2011(27):31-32+34.

数据来源： ‘Hedonic prices and the demand for clean air’, J. Environ. Economics & Management, vol.5, 81-102, 1978.