Simple Linear Regression

Bella - Jiaqi Fan

Kate - Jiaqi Cai

Tracy – Xiyan Zhou

Contents

Introduction

- Data set: "BostonHousing”
- Boston census tracts from the 1970 census (Harrison and Rubinfeld)
- 506 instances & 14 attributes
- Goal: Find the best model that predict the median value of owner-occupied homes

Simple Linear Regression

$$y = \beta_0 + \beta_1 X + \epsilon$$

- Performed to determine the association between two quantitative variables.
- Can be used to determine:
  - The degree to which two variables are significantly correlated.
  - The dependent variable's value at a certain level of the independent variable.

Variables

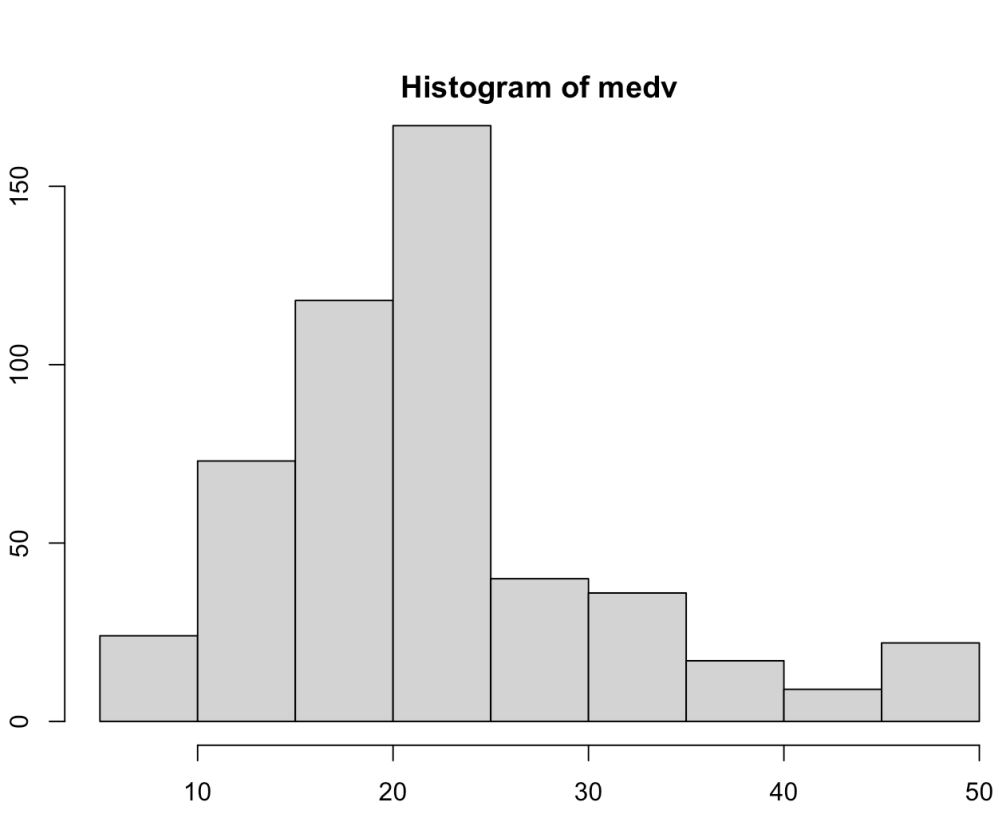
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitric oxides concentration (parts per 10 million)
rm	average number 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
b	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
lstat	percentage of lower status of the population
medv	median value of owner-occupied homes in USD 1000's

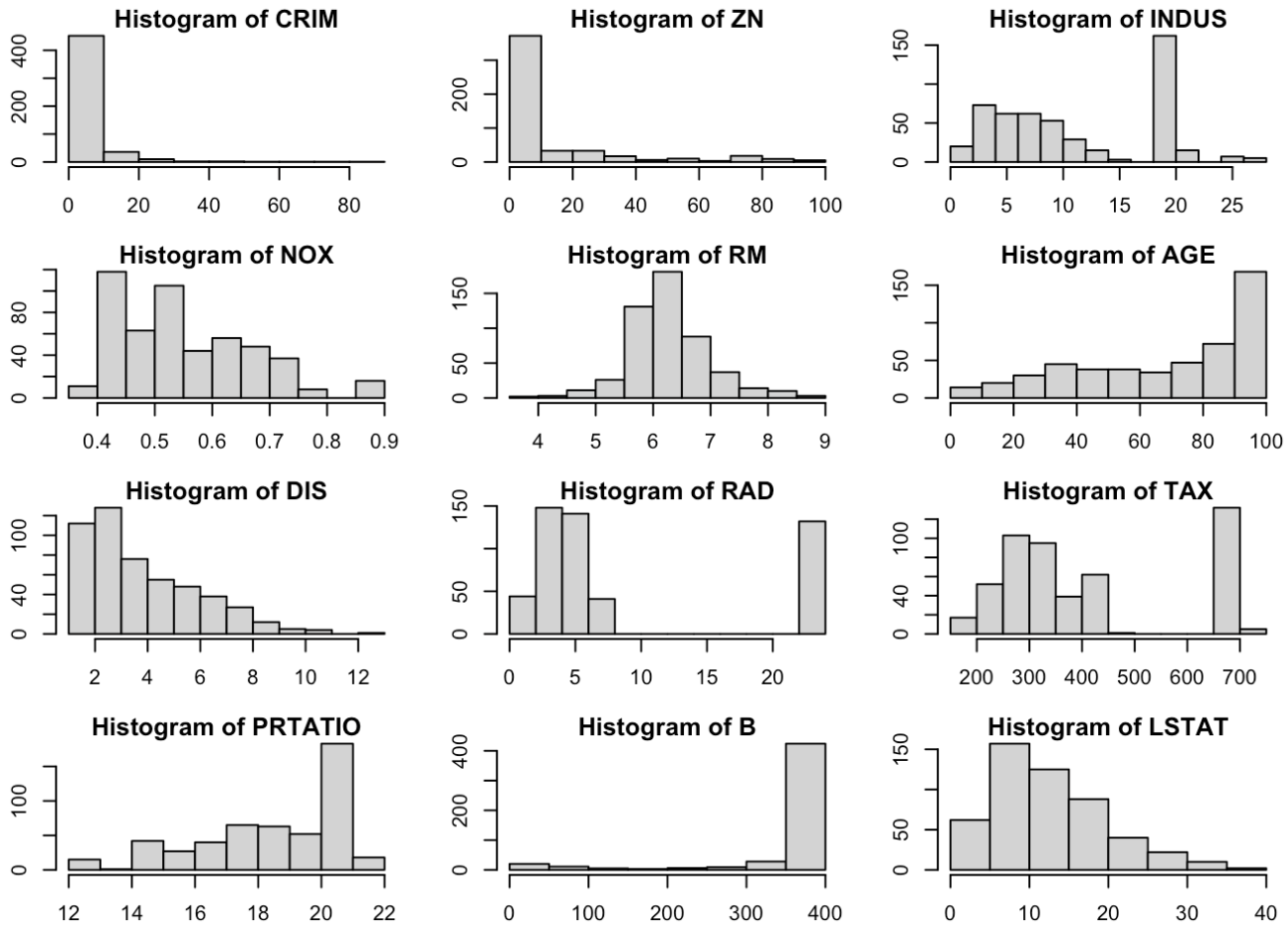
Descriptive statistical analysis

Means, Medians, Extremes and Quartiles

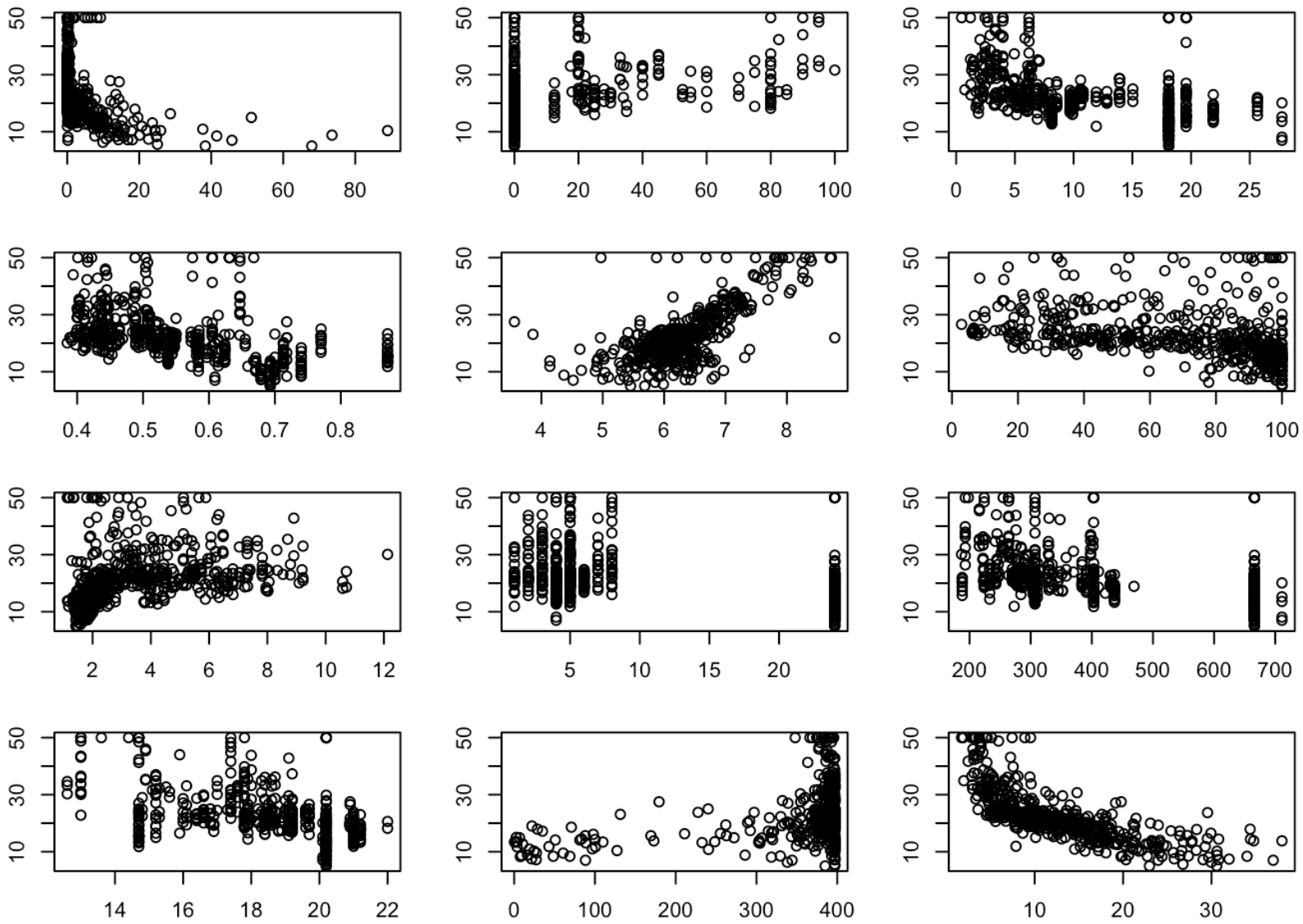
```
      crim                zn              indus            nox        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.3850  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.4490  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.5380  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.5547  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.6240  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :0.8710  
       rm              age              dis              rad        
 Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000  
 1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000  
 Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000  
 Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549  
 3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000  
 Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000  
      tax             ptratio            b               lstat       
 Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73  
 1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95  
 Median :330.0   Median :19.05   Median :391.44   Median :11.36  
 Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65  
 3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95  
 Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97  
      medv      
 Min.   : 5.00  
 1st Qu.:17.02  
 Median :21.20  
 Mean   :22.53  
 3rd Qu.:25.00  
 Max.   :50.00
```

Histograms

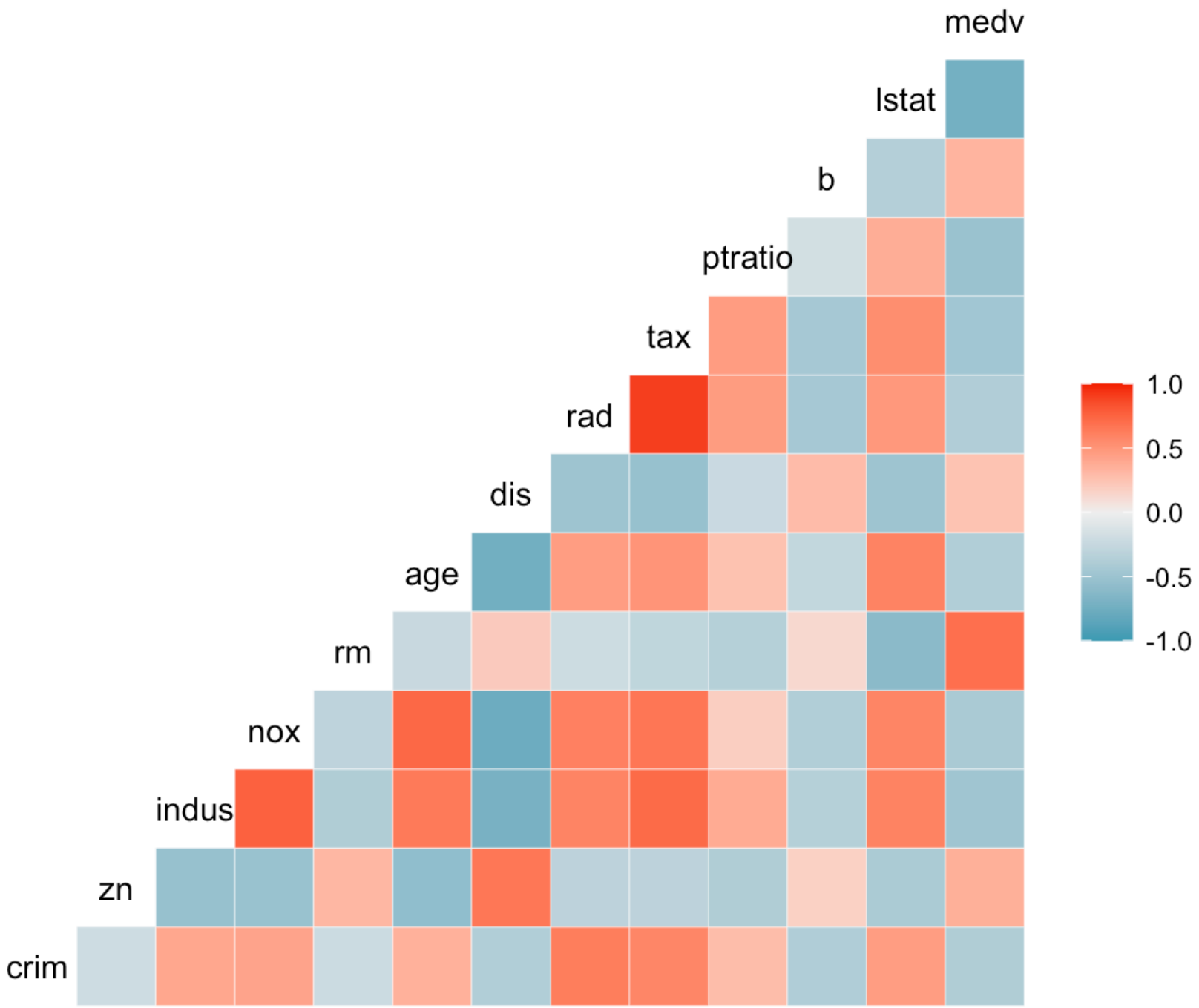

Histogram of medv
0
50
100
150
10
20
30
40
50

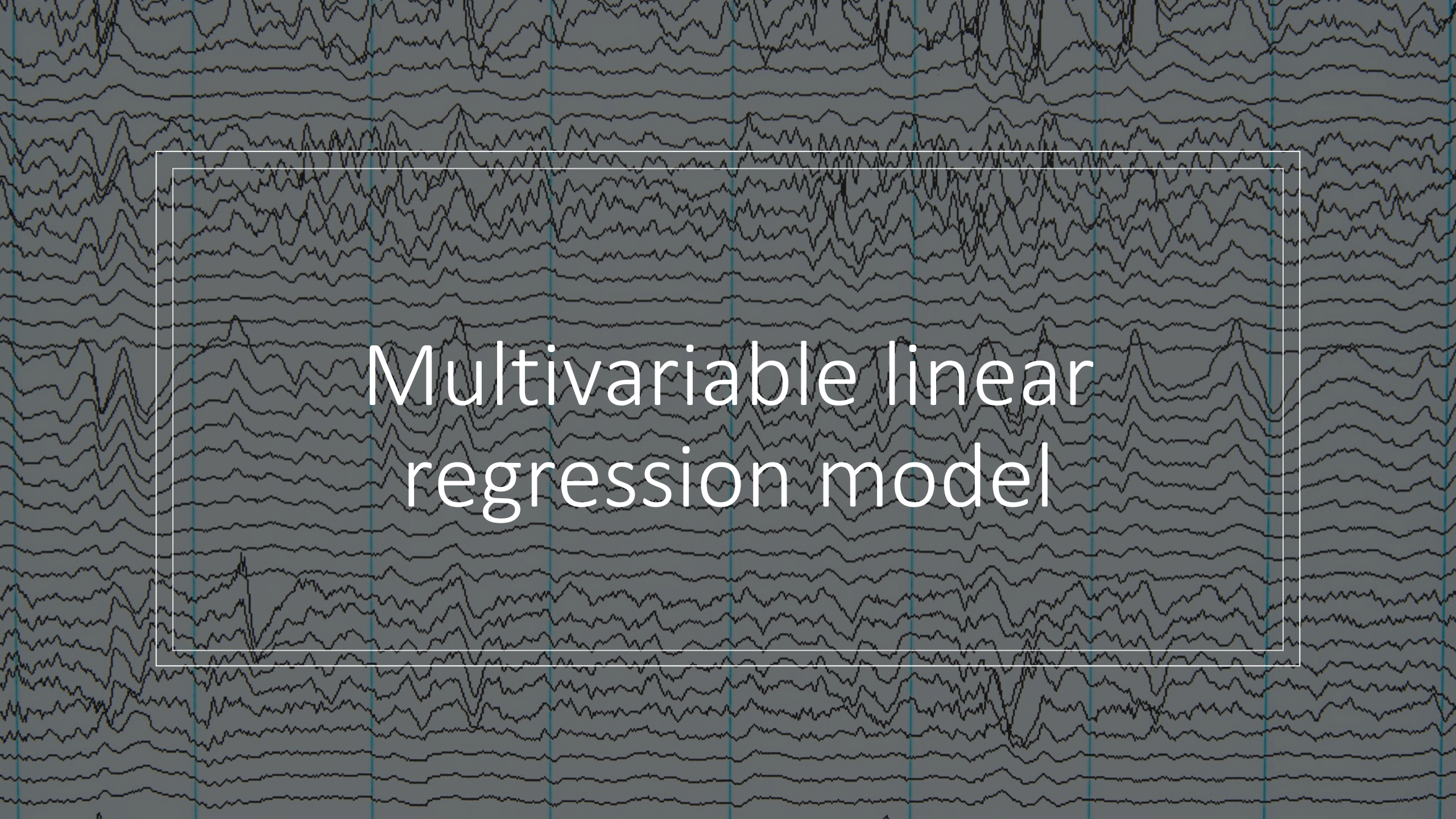

Histogram of CRIM
Histogram of ZN
Histogram of INDUS
Histogram of NOX
Histogram of RM
Histogram of AGE
Histogram of DIS
Histogram of RAD
Histogram of TAX
Histogram of PRTATIO
Histogram of B
Histogram of LSTAT

Scatterplots



Correlation coefficient matrix (Heat Map)





Multivariable linear regression model

When there are two or more independent variables and one dependent variable, multiple linear regression is used to determine the connection between them.

```
##
## Call:
## lm(formula = medv ~ . - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8433  -2.5888  -0.5865   1.5341  30.3980
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## crim     -0.098157   0.034691  -2.829 0.004853 **
## zn        0.049416   0.014532   3.401 0.000727 ***
## indus     0.016628   0.064685   0.257 0.797243
## nox      -2.255775   3.383425  -0.667 0.505267
## rm        5.998046   0.311101  19.280  < 2e-16 ***
## age      -0.005146   0.013923  -0.370 0.711815
## dis      -0.972546   0.197394  -4.927 1.14e-06 ***
## rad       0.193115   0.066989   2.883 0.004113 **
## tax      -0.010870   0.003930  -2.766 0.005894 **
## ptratio  -0.425742   0.110342  -3.858 0.000129 ***
## b         0.015433   0.002716   5.683 2.27e-08 ***
## lstat    -0.424929   0.051171  -8.304 9.72e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.025 on 494 degrees of freedom
## Multiple R-squared:  0.9584, Adjusted R-squared:  0.9574
## F-statistic: 947.6 on 12 and 494 DF,  p-value: < 2.2e-16
```

Many coefficients of the variables are not significant

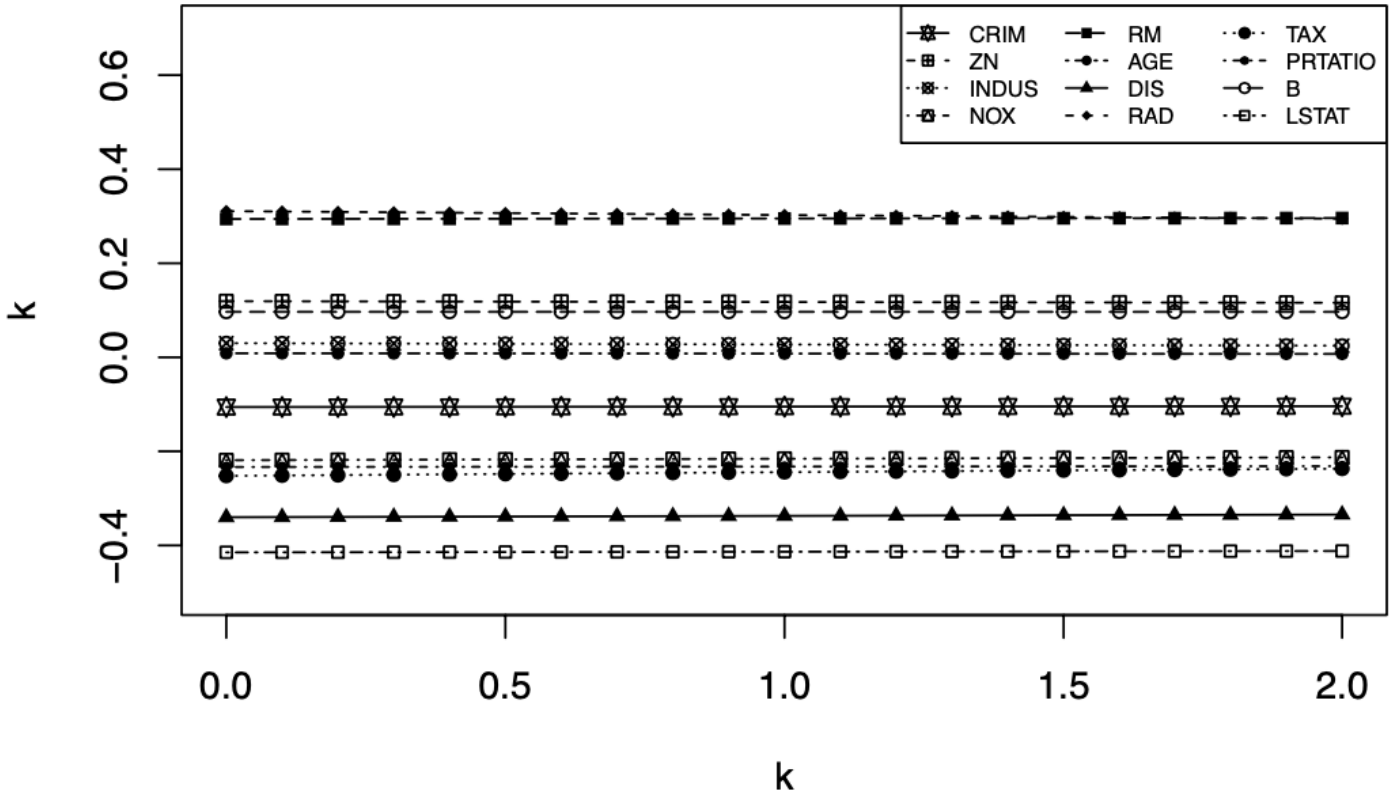
→ Variable selection needed

Variable selection

- Condition number = 94.77388 (close to 100) → Multicollinearity exists

Ridge Regression

- Ridge regression coefficients are stable → Bad model → Cannot use it to select variables



Stepwise regression

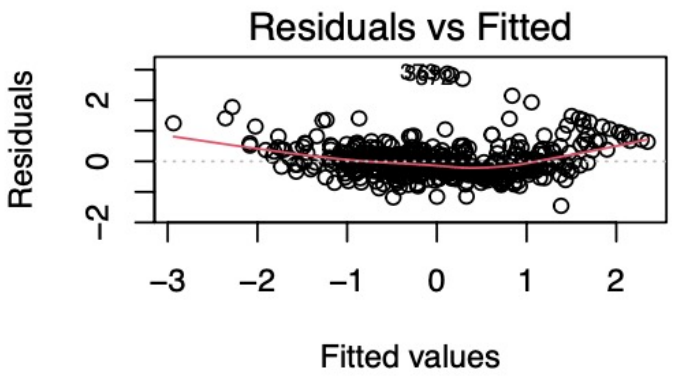
```
## Start:  AIC=1645.71
## medv ~ (crim + zn + indus + nox + rm + age + dis + rad + tax +
##     ptratio + b + lstat) - 1
##
##           Df Sum of Sq   RSS    AIC
## - indus    1       1.7 12477 1643.8
## - age      1       3.5 12479 1643.8
## - nox      1      11.2 12486 1644.2
## <none>                  12475 1645.7
## - tax      1     193.2 12668 1651.5
## - crim     1     202.2 12678 1651.8
## - rad      1     209.9 12685 1652.2
## - zn       1     292.0 12767 1655.4
## - ptratio  1     376.0 12851 1658.7
## - dis      1     613.0 13088 1668.0
## - b        1     815.6 13291 1675.8
## - lstat    1    1741.4 14217 1709.8
## - rm       1    9387.3 21863 1927.6

## Step:  AIC=1643.78
## medv ~ crim + zn + nox + rm + age + dis + rad + tax + ptratio +
##     b + lstat - 1
##
##           Df Sum of Sq   RSS    AIC
## - age      1       3.4 12480 1641.9
## - nox      1       9.7 12487 1642.2
## <none>                  12477 1643.8
## + indus    1       1.7 12475 1645.7
## - crim     1     204.3 12681 1650.0
## - rad      1     215.3 12692 1650.4
## - tax      1     218.1 12695 1650.5
## - zn       1     290.6 12768 1653.4
## - ptratio  1     375.1 12852 1656.8
## - dis      1     670.9 13148 1668.3
## - b        1     813.9 13291 1673.8
## - lstat    1    1741.0 14218 1707.9
## - rm       1    9659.4 22136 1931.9

## Step:  AIC=1641.92
## medv ~ crim + zn + nox + rm + dis + rad + tax + ptratio + b +
##     lstat - 1
##
##           Df Sum of Sq   RSS    AIC
## - nox      1      14.3 12495 1640.5
## <none>                  12480 1641.9
## + age      1       3.4 12477 1643.8
## + indus    1       1.6 12479 1643.8
## - crim     1     204.0 12684 1648.1
## - rad      1     218.9 12699 1648.7
## - tax      1     219.4 12700 1648.7
## - zn       1     303.2 12784 1652.1
## - ptratio  1     377.7 12858 1655.0
## - dis      1     731.7 13212 1668.8
## - b        1     810.9 13291 1671.8
## - lstat    1    2013.9 14494 1715.6
## - rm       1   10046.5 22527 1938.7

## Step:  AIC=1640.5
## medv ~ crim + zn + rm + dis + rad + tax + ptratio + b + lstat -
##     1
##
##           Df Sum of Sq   RSS    AIC
## <none>                  12495 1640.5
## + nox      1      14.3 12480 1641.9
## + age      1       8.0 12487 1642.2
## + indus    1       0.0 12495 1642.5
## - crim     1     199.1 12694 1646.5
## - rad      1     234.9 12730 1647.9
## - tax      1     293.7 12788 1650.2
## - zn       1     319.7 12814 1651.3
## - ptratio  1     383.3 12878 1653.8
## - dis      1     748.7 13243 1668.0
## - b        1     796.8 13291 1669.8
## - lstat    1    2737.3 15232 1738.7
## - rm       1   14397.2 26892 2026.4
```

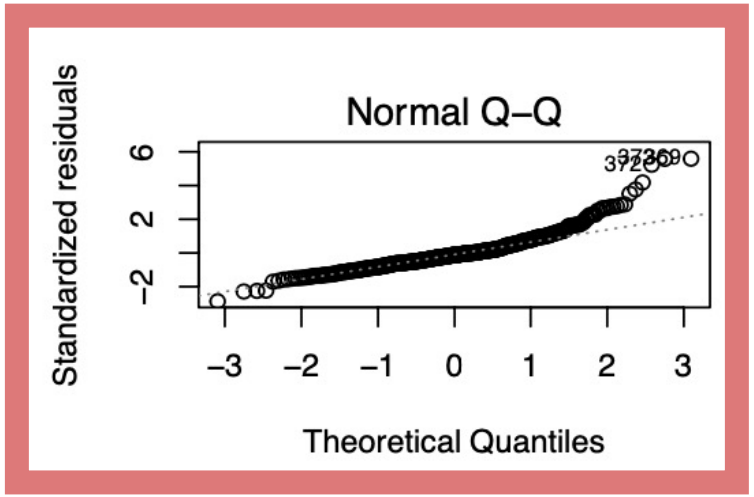
Lasso regression

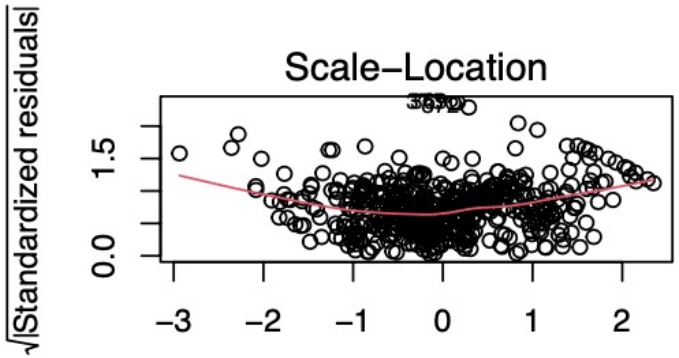
```
## LARS/LASSO
## Call: lars(x = as.matrix(x), y = as.matrix(y), type = "lasso")
##     Df    Rss       Cp
## 0    1 42716 1360.011
## 1    2 36326 1083.143
## 2    3 21335  431.009
## 3    4 14960  154.804
## 4    5 13867  109.115
## 5    6 13720  104.718
## 6    7 13262   86.725
## 7    8 12658   62.354
## 8    9 12186   43.749
## 9   10 12091   41.625
## 10 11 11328   10.335
## 11 12 11310   11.514
## 12 13 11298   13.000
```

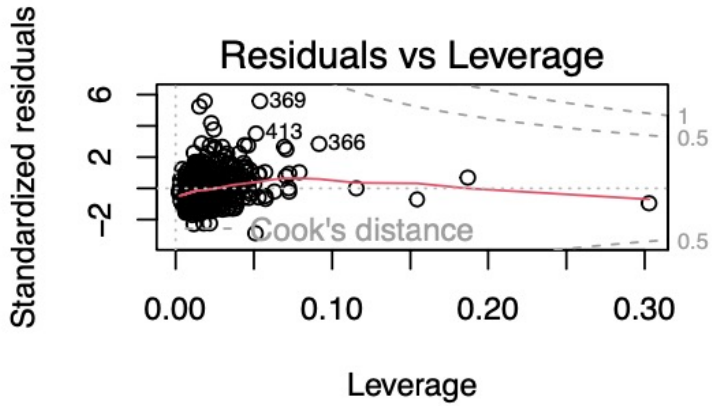
```
##
## Call:
## lars(x = as.matrix(x), y = as.matrix(y), type = "lasso")
## R-squared: 0.736
## Sequence of LASSO moves:
##      lstat rm ptratio  b crim dis nox zn rad tax indus age
## Var     12  5      10 11    1   7   4  2   8   9     3   6
## Step     1  2       3  4    5   6   7  8   9  10    11  12
```

```
##
## Call:
## lm(formula = medv ~ . - 1, data = data.2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45389 -0.30382 -0.05989  0.20596  2.87027
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## crim    -0.10667    0.03086  -3.456 0.000595 ***
## zn       0.11600    0.03457   3.355 0.000854 ***
## nox     -0.20750    0.04476  -4.636 4.55e-06 ***
## rm       0.29371    0.03128   9.391  < 2e-16 ***
## dis     -0.34941    0.04280  -8.163 2.71e-15 ***
## rad      0.29873    0.06033   4.952 1.01e-06 ***
## tax     -0.23226    0.06208  -3.741 0.000205 ***
## ptratio -0.23032    0.03054  -7.542 2.22e-13 ***
## b        0.09658    0.02672   3.614 0.000332 ***
## lstat   -0.41004    0.03710 -11.053  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5192 on 496 degrees of freedom
## Multiple R-squared:  0.7353, Adjusted R-squared:  0.7299
## F-statistic: 137.8 on 10 and 496 DF,  p-value: < 2.2e-16
```









New model after deleting “indus” and “age”

Outliers and influential points

```
##
## Call:
## lm(formula = medv ~ . - 1, data = data.3.1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42082 -0.28477 -0.09735  0.15682  1.89798
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## crim    -0.09034    0.02617  -3.452 0.000605 ***
## zn       0.09529    0.02933   3.249 0.001236 **
## nox     -0.17295    0.03820  -4.528 7.49e-06 ***
## rm       0.35772    0.02708  13.212  < 2e-16 ***
## dis     -0.27299    0.03667  -7.444 4.42e-13 ***
## rad      0.22330    0.05139   4.345 1.69e-05 ***
## tax     -0.24620    0.05263  -4.678 3.76e-06 ***
## ptratio -0.23076    0.02590  -8.908  < 2e-16 ***
## b        0.09488    0.02288   4.146 3.99e-05 ***
## lstat   -0.30966    0.03267  -9.479  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4398 on 490 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.7898
## F-statistic: 188.9 on 10 and 490 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = medv ~ . - 1, data = data.3.2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19716 -0.29289 -0.08712  0.14470  1.43534
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## crim    -0.08759    0.02426  -3.610 0.000338 ***
## zn       0.08628    0.02725   3.167 0.001638 **
## nox     -0.15990    0.03550  -4.504 8.35e-06 ***
## rm       0.43552    0.02754  15.813  < 2e-16 ***
## dis     -0.23585    0.03422  -6.893 1.71e-11 ***
## rad      0.19338    0.04795   4.033 6.40e-05 ***
## tax     -0.23213    0.04890  -4.747 2.72e-06 ***
## ptratio -0.21597    0.02408  -8.970  < 2e-16 ***
## b        0.11292    0.02139   5.279 1.96e-07 ***
## lstat   -0.23797    0.03172  -7.503 3.02e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4067 on 484 degrees of freedom
## Multiple R-squared:  0.8181, Adjusted R-squared:  0.8144
## F-statistic: 217.7 on 10 and 484 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = medv ~ . - 1, data = data.3.3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19465 -0.28626 -0.08509  0.13752  1.42596
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## crim    -0.09475    0.02416  -3.921 0.000101 ***
## zn       0.09091    0.02708   3.357 0.000850 ***
## nox     -0.15753    0.03541  -4.449 1.07e-05 ***
## rm       0.43489    0.02787  15.607  < 2e-16 ***
## dis     -0.23345    0.03404  -6.859 2.16e-11 ***
## rad      0.20424    0.04775   4.277 2.29e-05 ***
## tax     -0.24374    0.04866  -5.009 7.72e-07 ***
## ptratio -0.21005    0.02401  -8.748  < 2e-16 ***
## b        0.10613    0.02183   4.861 1.59e-06 ***
## lstat   -0.23211    0.03171  -7.319 1.06e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4029 on 478 degrees of freedom
## Multiple R-squared:  0.8169, Adjusted R-squared:  0.8131
## F-statistic: 213.3 on 10 and 478 DF,  p-value: < 2.2e-16
```

Heteroscedasticity diagnose

```
## [[1]]
##
##  Spearman's rank correlation rho
##
## data:  data.3.3[, i] and abse
## S = 18229252, p-value = 0.1944
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.05884218
##
##
## [[2]]
##
##  Spearman's rank correlation rho
##
## data:  data.3.3[, i] and abse
## S = 17767511, p-value = 0.06801
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.08268141
```

```
cor.spearman
```

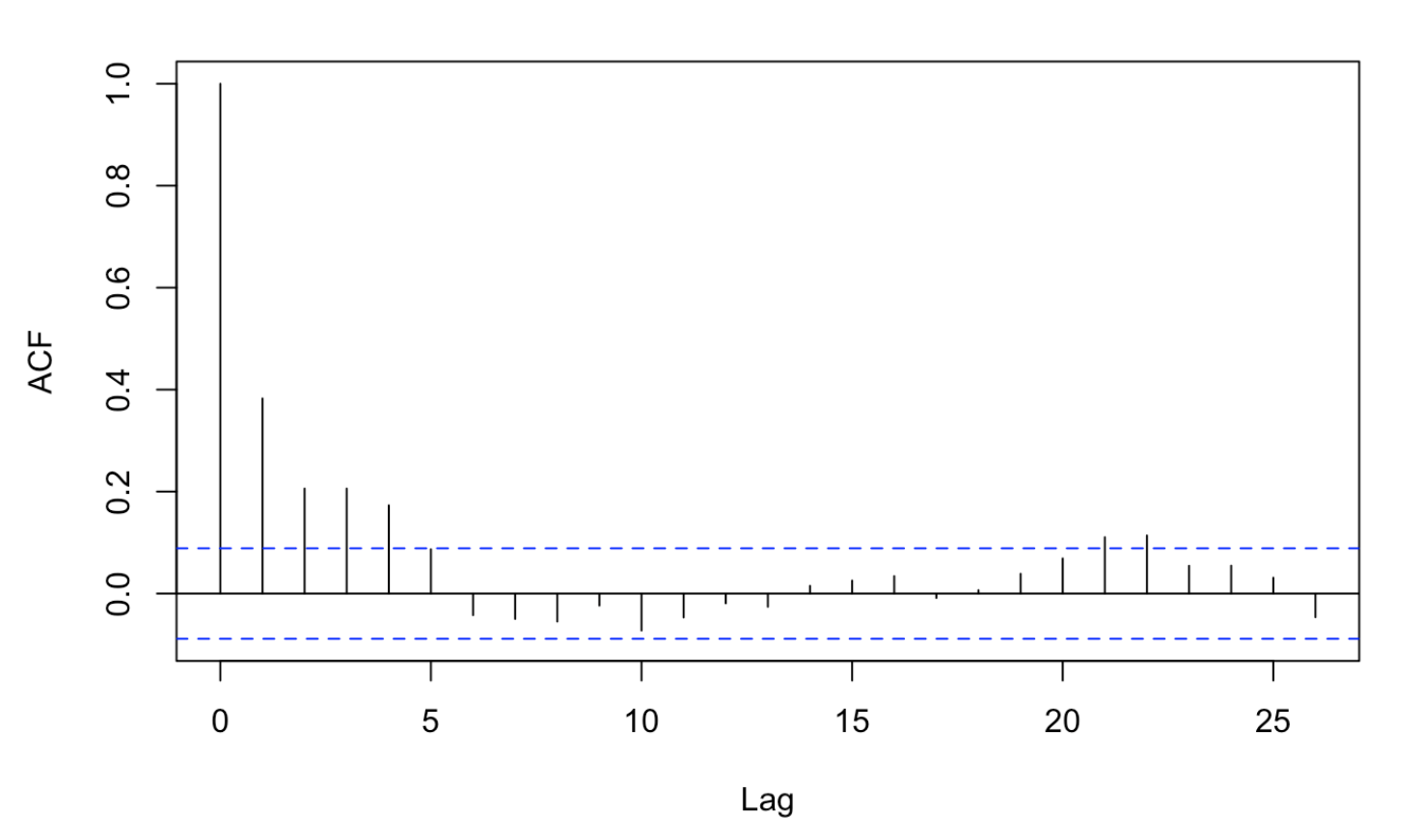
```
##  [1] 1.944051e-01 6.801105e-02 7.082236e-02 5.448079e-05 2.030279e-03
##  [6] 9.229305e-03 6.731165e-02 8.128597e-02 3.863399e-01 8.679934e-01
```

```
names(data.3.3[,-(11)])[cor.spearman<0.05]
```

```
## [1] "rm"  "dis" "rad"
```

check the p-value and the correlation

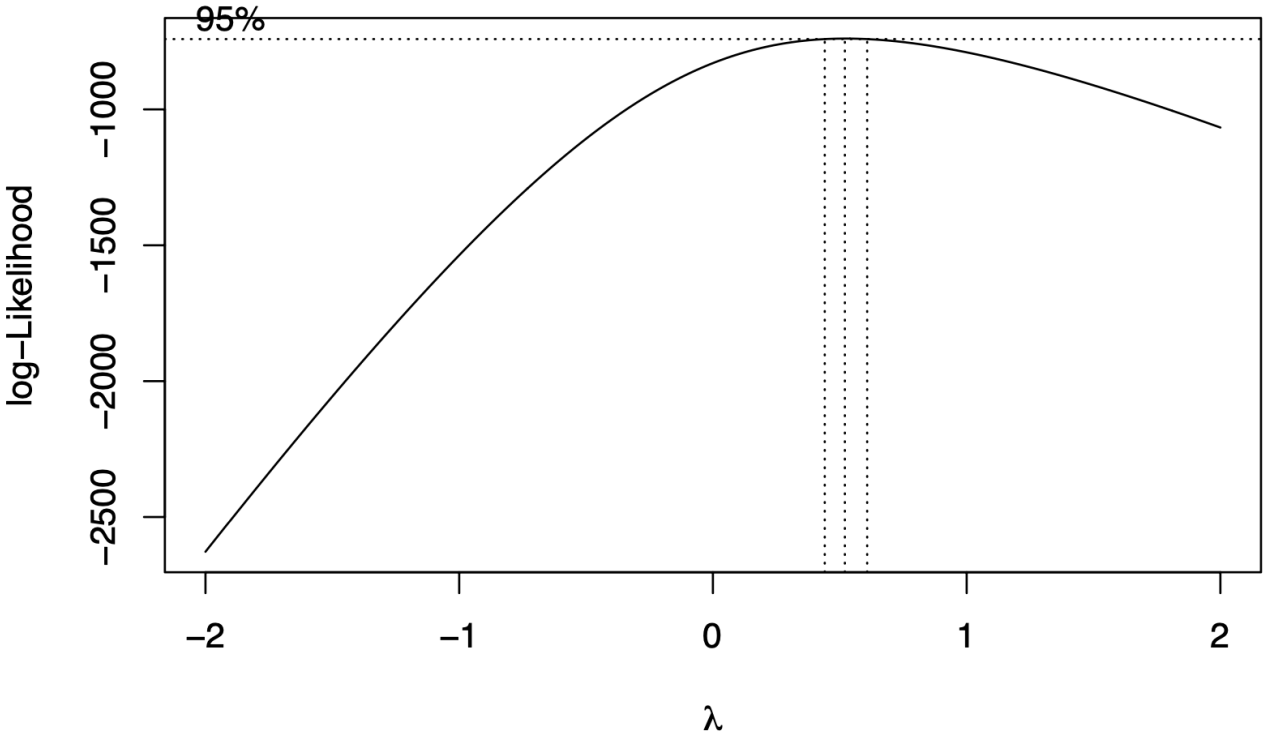
Autocorrelation diagnose



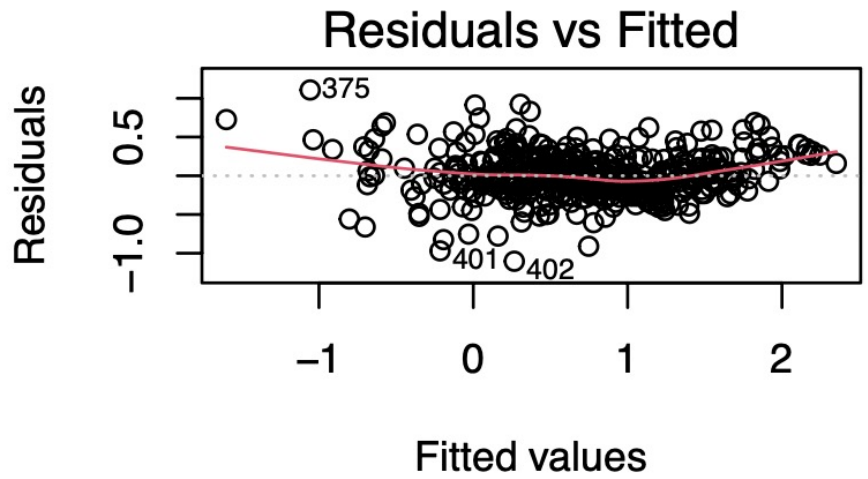
"Autocorrelation is the degree of correlation between values taken at different times from the same time series."

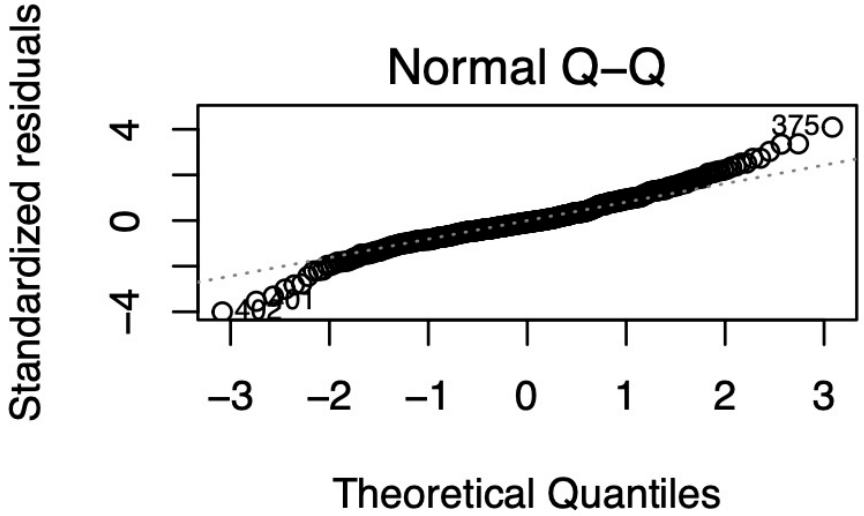
As shown in the graph, when the lag number is 0, the correlation coefficient is 1, and as the lag number increases, the correlation coefficient gradually decreases and becomes stable.

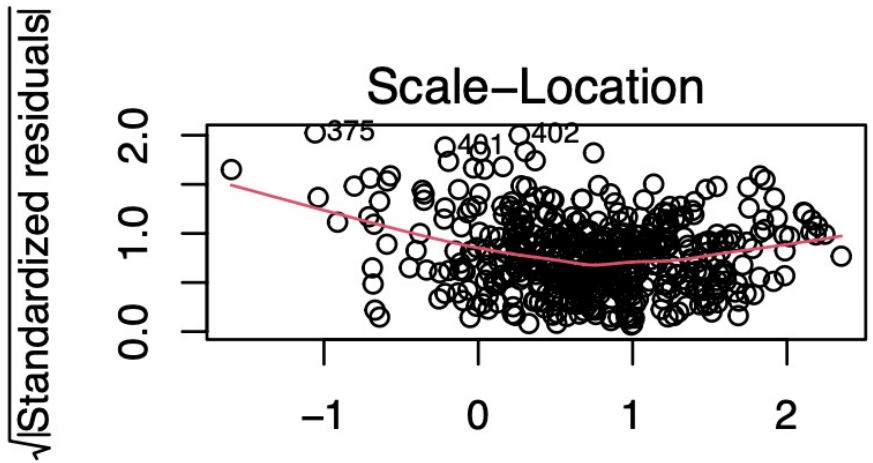
Solution to heteroscedasticity & autocorrelation: Box-Cox

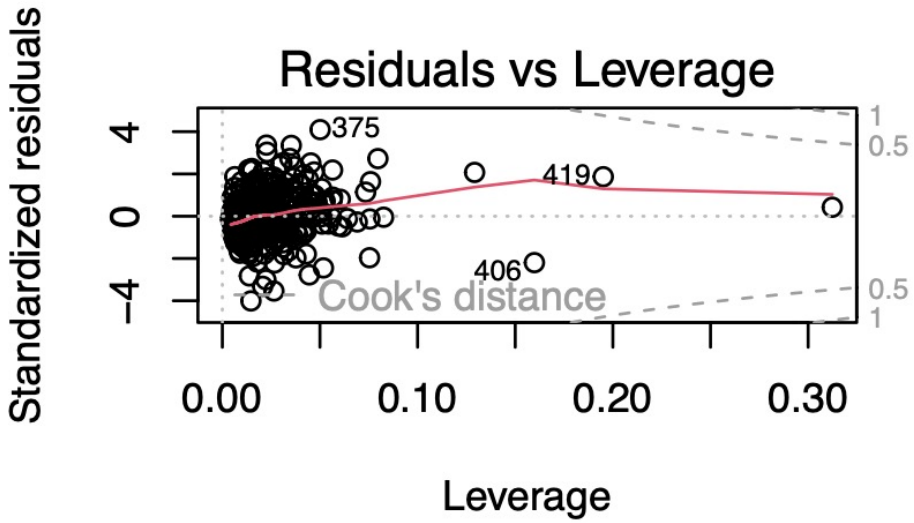



```
## Call:
## lm(formula = medv_bc ~ . - medv, data = data.4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10338 -0.14968 -0.02337  0.14951  1.10706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.72469    0.01259  57.567  < 2e-16 ***
## crim        -0.12156    0.01663  -7.308 1.15e-12 ***
## zn           0.04399    0.01864   2.360   0.0187 *
## nox         -0.10894    0.02438  -4.469 9.83e-06 ***
## rm           0.24044    0.01919  12.528  < 2e-16 ***
## dis         -0.14616    0.02344  -6.235 9.98e-10 ***
## rad          0.15773    0.03288   4.798 2.15e-06 ***
## tax         -0.17683    0.03350  -5.279 1.98e-07 ***
## ptratio     -0.13990    0.01653  -8.463 3.22e-16 ***
## b            0.07696    0.01503   5.119 4.45e-07 ***
## lstat       -0.22734    0.02185 -10.404  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2774 on 477 degrees of freedom
## Multiple R-squared:  0.8326, Adjusted R-squared:  0.8291
## F-statistic: 237.2 on 10 and 477 DF,  p-value: < 2.2e-16
```

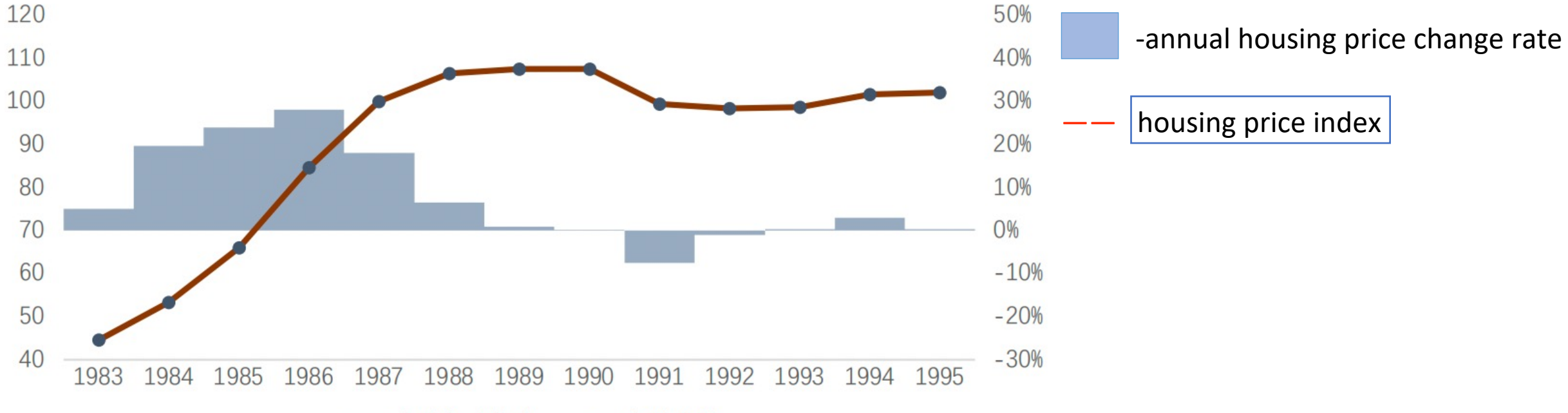

Residuals vs Fitted
Residuals
Fitted values
375
401
402


Normal Q–Q
Standardized residuals
Theoretical Quantiles
375
402
401


Scale–Location
√|Standardized residuals|
Fitted values
375
401
402


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
375
419
406

CONCLUSION



when the real estate overheated

1. Adjusting the property tax rate
2. Restrict environmental indicators such as nitrogen oxide content through administrative measures
3.Promote low-rent housing and public housing

When housing prices go down

1. Boston government could increase police security spending .
2.Strengthen infrastructure and improve education by increasing the number of schools and bringing in high-level teachers.
3.Simultaneously improve institutional protection for the mobile population through development of social welfare system.

Thank you

-Any questions?