다변량분석

회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 데이터 소개

- 보스턴 지역 주택 가격 데이터 (MASS 패키지의 Boston 데이터)

```
library(MASS)
data <- Boston
```

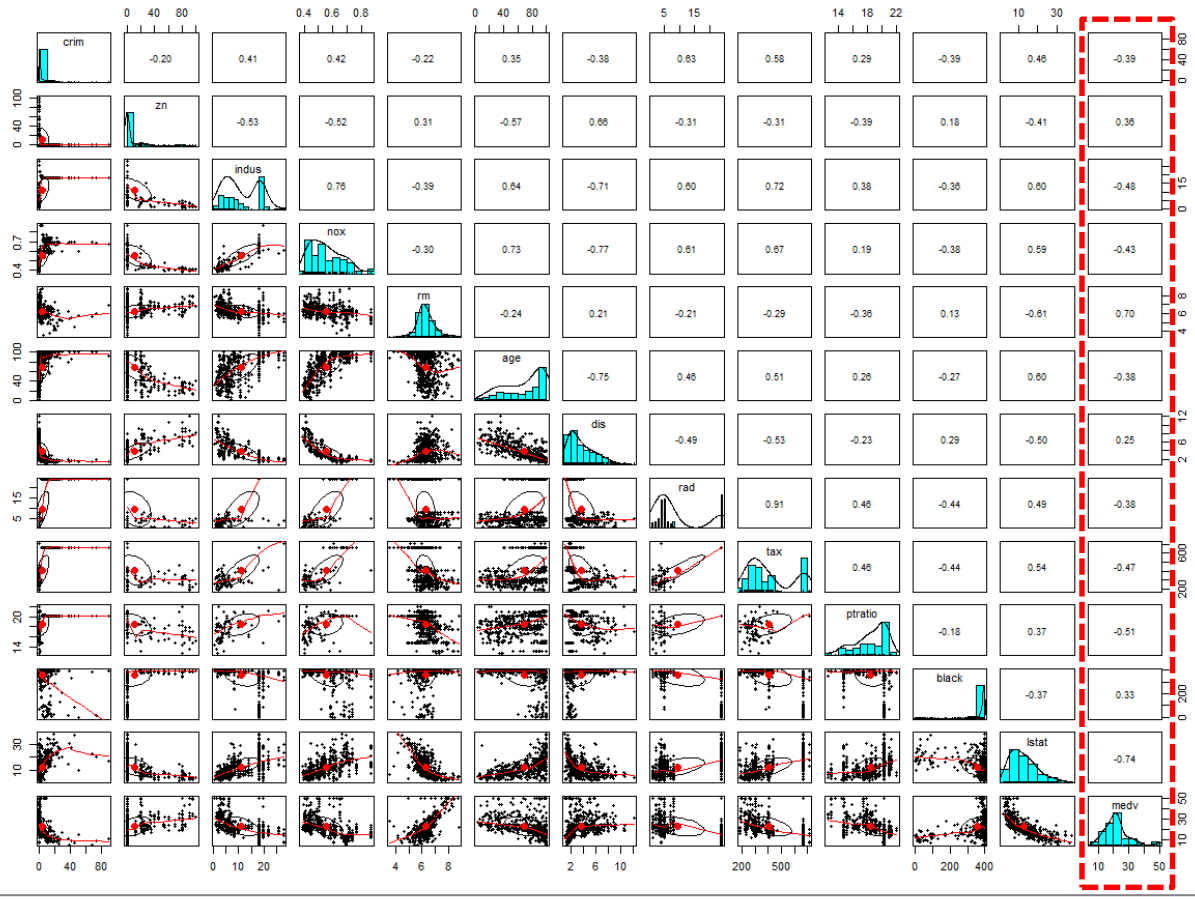
- 변수
  - crim: 범죄발생률
  - zn: 주거지 중 2500 ft^2 이상 크기의 대형주택이 차지하는 비율
  - indus: 소매상 이외의 상업지구의 면접 비율
  - chas: 찰스강과 접한 지역은 1, 아니면 0인 더미변수
  - nox: 산화질소 오염도
  - rm: 주거지당 평균 방 개수
  - age: 소유자 주거지 중 1940년 이전에 지어진 집들의 비율
  - dis: 보스턴의 5대 고용중심으로부터의 가중 평균 거리
  - rad: 도시 순환 고속도로에의 접근 용이 지수
  - tax: 만달러당 주택 재산세율
  - ptratio: 학생-선생 비율
  - black: 흑인(BK) 인구 비율이 0.64과 다른 정도의 제곱 $(1000(BK - 0.63)^2$
  - lstat : 저소득 주민들의 비율 퍼센트
  - medv: 소유자 주거지 주택 가격

회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

```
library(psych)
pairs.panels(data[,-4])
```

➢ 상관분석 (수치형 변수만)



lstat 와 rm 변수가
높은 상관을 가짐

회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 데이터 나누기

- set.seed(1606)
- n <- nrow(data)
- idx <- 1:n
- training_idx <- sample(idx, n * .60)
- idx <- setdiff(idx, training_idx)
- validate_idx <- sample(idx, n * .20)
- test_idx <- setdiff(idx, validate_idx)

- training <- data[training_idx,]
- validation <- data[validate_idx,]
- test <- data[test_idx,]

회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 수치형 변수로만 회귀모형 구축하기 (1차항만 고려)

- data_lm_full <- lm(medv ~ ., data=training[,-4])
- summary(data_lm_full)

계수 개수 확인
length(coef(data_lm_full))

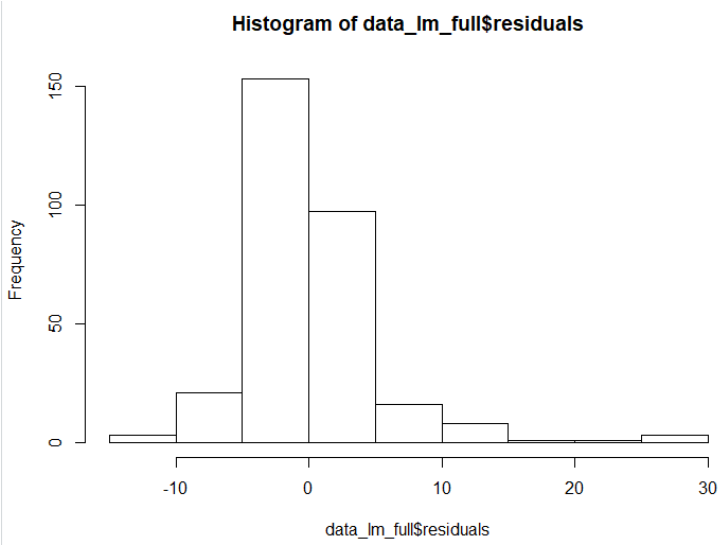
```
> summary(data_lm_full)

Call:
lm(formula = medv ~ ., data = training[, -4])

Residuals:
     Min       1Q   Median       3Q      Max
-11.4521  -2.8810  -0.5993   1.5620  26.2954

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.686657   7.007530   5.521 7.50e-08 ***
crim         -0.088227   0.049860  -1.770  0.07786 .
zn            0.052560   0.018504   2.840  0.00482 **
indus        -0.033040   0.083686  -0.395  0.69327
nox         -15.301392   5.070699  -3.018  0.00277 **
rm            3.187607   0.589729   5.405 1.35e-07 ***
age          -0.013116   0.018068  -0.726  0.46849
dis          -1.607115   0.278257  -5.776 1.97e-08 ***
rad           0.288908   0.089987   3.211  0.00147 **
tax          -0.011779   0.004990  -2.361  0.01890 *
ptratio      -0.857479   0.182124  -4.708 3.88e-06 ***
black         0.010240   0.003705   2.764  0.00608 **
lstat        -0.506730   0.069001  -7.344 2.11e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.077 on 290 degrees of freedom
Multiple R-squared:  0.6944,    Adjusted R-squared:  0.6817
F-statistic: 54.91 on 12 and 290 DF,  p-value: < 2.2e-16
```

회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 수치형 변수로만 회귀모형 구축하기 (잔차 확인)

- hist(data_lm_full$residuals)
- plot(data_lm_full, which=1)
- plot(data_lm_full, which=2)

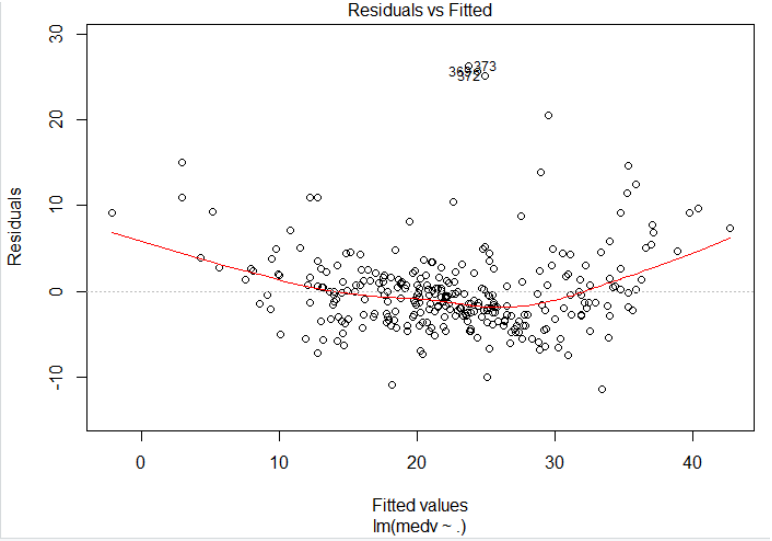


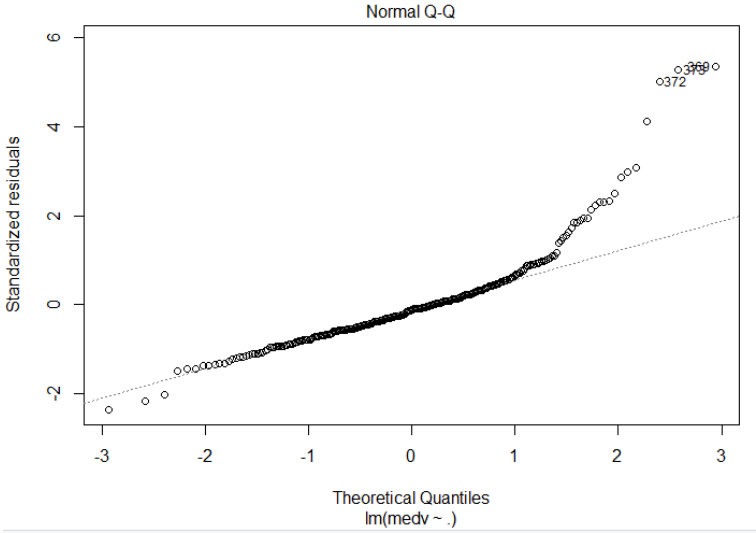
회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 수치형 변수로만 회귀모형 구축하기 (잔차 확인)

- hist(data_lm_full$residuals)
- plot(data_lm_full, which=1)
- plot(data_lm_full, which=2)





회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 수치형 변수로만 회귀모형 구축하기 (2차 상호작용까지 고려)

- data_lm_full_2 <- lm(medv ~ .^2, data=training[,-4])
- summary(data_lm_full_2)

```
> summary(data_lm_full_2)

Call:
lm(formula = medv ~ .^2, data = training[, -4])

Residuals:
     Min       1Q   Median       3Q      Max
-10.5484  -1.5449  -0.1696   1.4307  13.7241

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.413e+02  9.975e+01  -2.419  0.01638 *
crim          -1.732e+01  9.814e+00  -1.765  0.07900 .
zn             4.338e-01  6.842e-01   0.634  0.52670
                    ...

tax:ptratio    1.155e-02  3.895e-03   2.966  0.00334 **
tax:black     -1.422e-04  2.965e-04  -0.479  0.63209
tax:lstat     -5.936e-04  1.641e-03  -0.362  0.71790
ptratio:black  5.490e-03  4.607e-03   1.192  0.23461
ptratio:lstat -6.282e-02  4.580e-02  -1.371  0.17160
black:lstat   -1.467e-03  5.620e-04  -2.610  0.00967 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.095 on 224 degrees of freedom
Multiple R-squared:  0.9122,    Adjusted R-squared:  0.8817
F-statistic: 29.85 on 78 and 224 DF,  p-value: < 2.2e-16
```

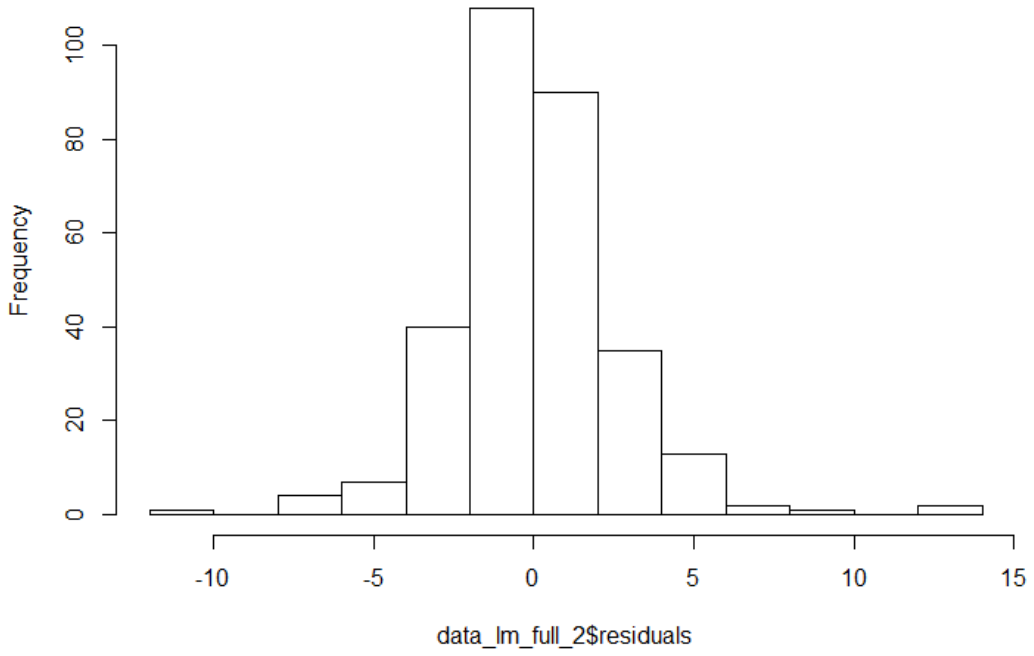
계수 개수 확인
length(coef(data_lm_full_2))

회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 수치형 변수로만 회귀모형 구축하기 (잔차 확인)

- hist(data_lm_full_2$residuals)
- plot(data_lm_full_2, which=1)
- plot(data_lm_full_2, which=2)

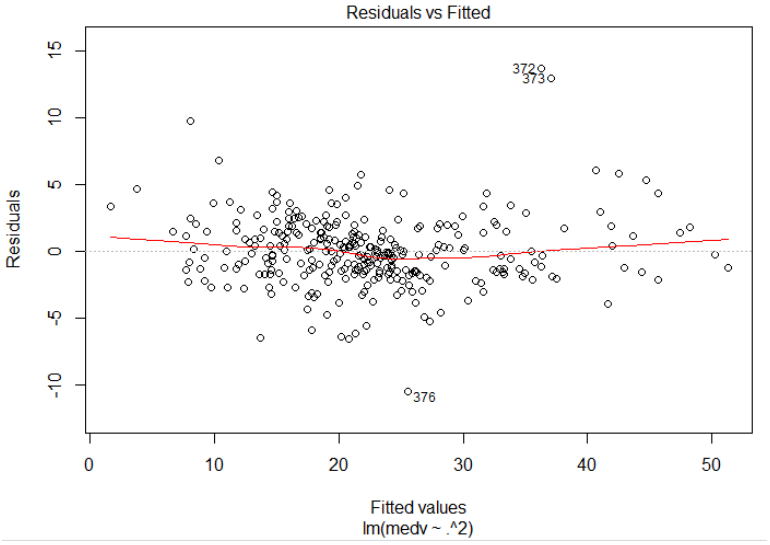


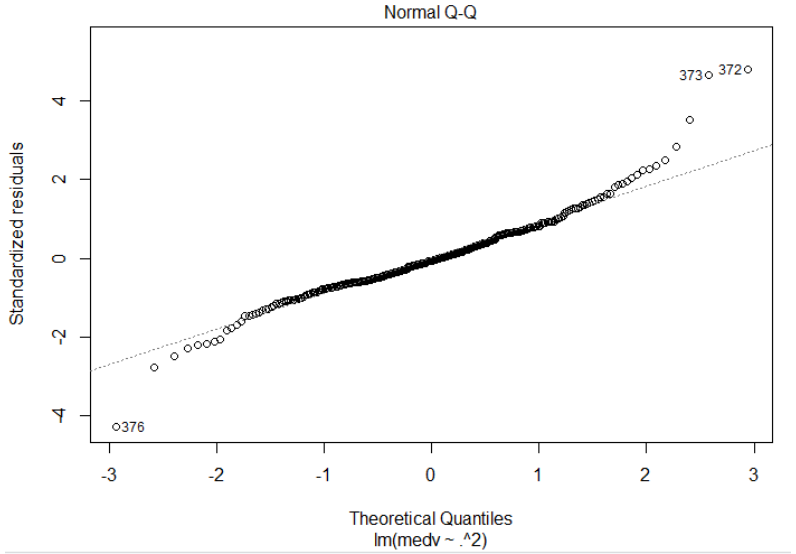
회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 수치형 변수로만 회귀모형 구축하기 (잔차 확인)

- hist(data_lm_full$residuals)
- plot(data_lm_full, which=1)
- plot(data_lm_full, which=2)





회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 변수선택법 활용하기

- stepAIC 함수 이용하기

stepAIC {MASS}

Choose a model by AIC in a Stepwise Algorithm

Description

Performs stepwise model selection by AIC.

Usage

```
stepAIC(object, scope, scale = 0,
        direction = c("both", "backward", "forward"),
        trace = 1, keep = NULL, steps = 1000, use.start = FALSE,
        k = 2, ...)
```

- Direction
  - both : 전진선택법과 후진제거법을 번갈아가면서 수행
  - backward (후진제거법): 모든 변수가 포함된 모형에서 유의하지 않은 변수를 제거
  - Forward (전진선택법): 가장 유의한 변수들부터 하나씩 추가

회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 변수선택법 활용하기 (stepAIC 함수 이용하기 (both 방법))

- library(MASS)
- data_step_both <- stepAIC(data_lm_full, direction = "both",
  scope = list(upper = ~ .^2, lower = ~1))
- summary(data_step_both)

```
> summary(data_step_both)

Call:
lm(formula = medv ~ crim + zn + indus + nox + rm + age + dis +
    rad + tax + ptratio + black + lstat + rm:lstat + rad:lstat +
    rm:rad + dis:rad + black:lstat + crim:rm + crim:lstat + age:black +
    age:lstat + rm:age + rm:ptratio + rm:black + crim:nox + crim:ptratio +
    tax:ptratio + indus:dis + nox:rm + zn:lstat + dis:lstat +
    age:dis + nox:age + age:rad + age:tax + rad:tax + indus:ptratio +
    age:ptratio, data = training[, -4])

Residuals:
     Min       1Q   Median       3Q      Max
-10.2560  -1.7125  -0.1914   1.4723  14.6563

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.809e+02  3.619e+01  -5.000 1.05e-06 ***
crim         4.026e+00  2.983e+00   1.350 0.178235
zn           4.522e-02  2.308e-02   1.959 0.051140 .
```

계수 개수 확인
length(coef(data_step_both))

...

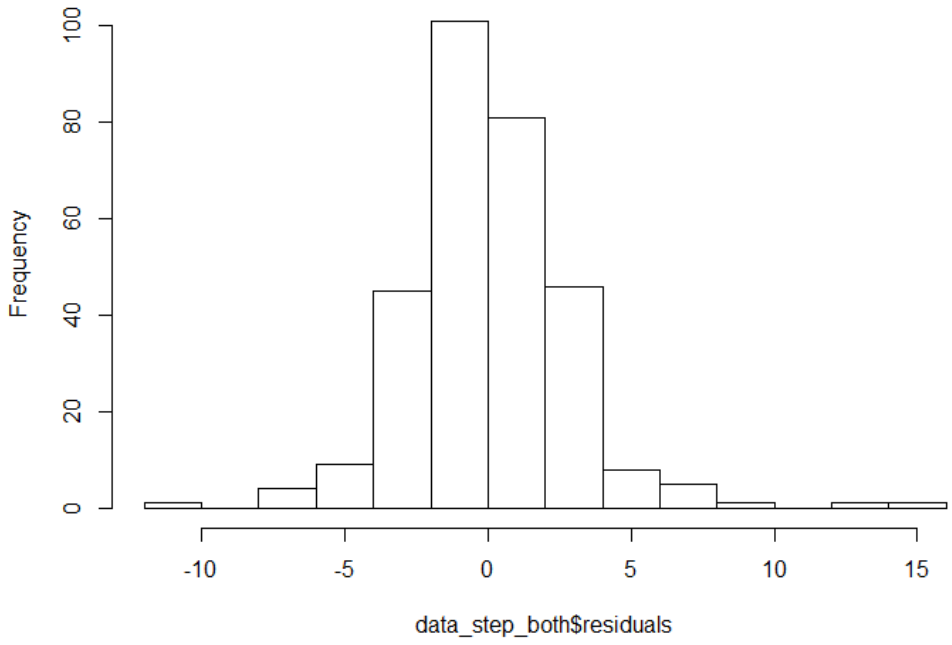
```
rad:tax       -1.242e-03  4.761e-04  -2.608 0.009616 **
indus:ptratio -6.277e-02  3.064e-02  -2.049 0.041476 *
age:ptratio   -7.158e-03  5.299e-03  -1.351 0.177878
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.996 on 264 degrees of freedom
Multiple R-squared:  0.9031,    Adjusted R-squared:  0.8892
F-statistic: 64.78 on 38 and 264 DF,  p-value: < 2.2e-16
```

회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 변수선택법 활용하기(잔차분석) <stepAIC 함수 이용하기 (both 방법)>

- hist(data_step_both$residuals)
- plot(data_step_both,which=1)
- plot(data_step_both,which=2)

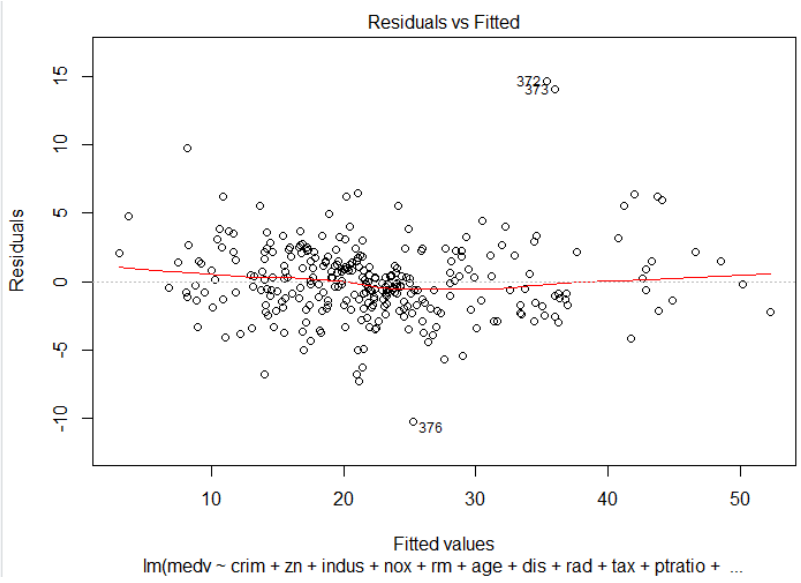


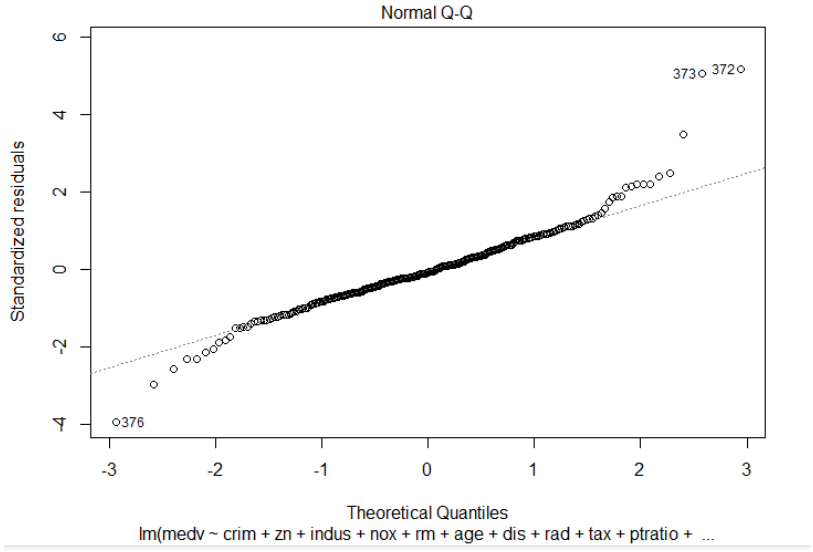
회귀분석 실습

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 변수선택법 활용하기(잔차분석) <stepAIC 함수 이용하기 (both 방법)>

- hist(data_step_both$residuals)
- plot(data_step_both,which=1)
- plot(data_step_both,which=2)





회귀분석 D.I.Y #1

❖ 모형 평가 (학습데이터)

- 계수 개수
- 결정계수
- 수정 결정계수
- RMSE 학습데이터
- RMSE 검증데이터

- 비교모형
  - 선형회귀모형
  - 2차 상호작용까지 포함한 선형회귀모형
  - 변수선택을 고려한 선형회귀모형 (stepwise)
  - 변수선택을 고려한 선형회귀모형 (forward)
  - 변수선택을 고려한 선형회귀모형 (backward)

- 각 모형별 잔차분석
  - 각 학습과 검증 데이터의 잔차 분석 (히스토그램, Q-Q plot, 잔차와 적합된값 산점도)

회귀분석 D.I.Y #2

❖ 다중선형회귀분석

➢ 데이터

- library(carData)의 Prestige 데이터 (총 6개의 변수로 구성)
- 설명력이 높은 모형 구축하기 (변수선택방법 적용하여 비교하기)
- 종속변수는 income, 독립변수는 education, women, prestige, census

Format

This data frame contains the following columns:

education

Average education of occupational incumbents, years, in 1971.

income

Average income of incumbents, dollars, in 1971.

women

Percentage of incumbents who are women.

prestige

Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.

census

Canadian Census occupational code.

type

Type of occupation. A factor with levels (note: out of order): `bc`, Blue Collar; `prof`, Professional, Managerial, and Technical; `wc`, White Collar.