Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

15

회귀모델의 설명력

경북대학교 배준현 교수
(joonion@knu.ac.kr)

15. 회귀모델의 설명력

- mtcars 데이터셋
  - 자동차의 연비에 대한 데이터셋(1074년): 변수 11개, 관측값 32개
    - mpg: 연비 (miles/gallon)
    - cyl: 실린더 개수
    - disp: 배기량 (cu. in.)
    - hp: 마력
    - drat: 기어비(후방 차축 비율)
    - wt: 중량 (1000 lbs)
    - qsec: ¼ 마일 시간
    - vs: 엔진 (0=V-모양, 1=straight)
    - am: 트랜스미션 (0=자동, 1=수동)
    - gear: 전방 기어의 개수
    - carb: 기화기의 개수

15. 회귀모델의 설명력

```
> str(mtcars)
'data.frame':  32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

15. 회귀모델의 설명력

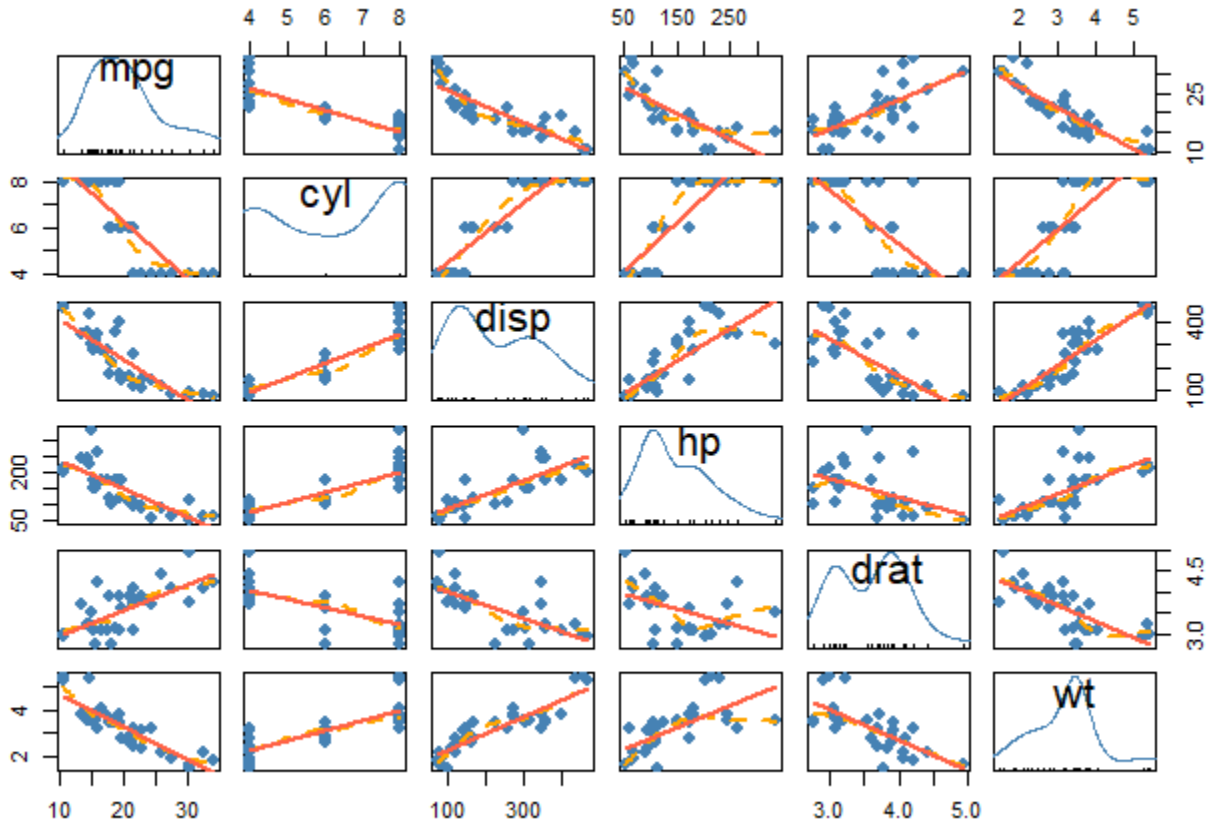
```
> str(mtcars)
> df <- subset(mtcars, select = 1:6)
> str(df)
'data.frame':  32 obs. of  6 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
```

15. 회귀모델의 설명력

```
> cor(df)
             mpg        cyl       disp         hp       drat         wt
mpg    1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
cyl   -0.8521620  1.0000000  0.9020329  0.8324475 -0.6999381  0.7824958
disp  -0.8475514  0.9020329  1.0000000  0.7909486 -0.7102139  0.8879799
hp    -0.7761684  0.8324475  0.7909486  1.0000000 -0.4487591  0.6587479
drat   0.6811719 -0.6999381 -0.7102139 -0.4487591  1.0000000 -0.7124406
wt    -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406  1.0000000
```

15. 회귀모델의 설명력

```
> library(car)
> scatterplotMatrix(df, pch = 19, col = "steelblue", cex = 1.2,
                    regLine = list(method = lm, lwd = 2, col = "tomato"),
                    smooth = list(smoother = loessLine, spread = FALSE,
                                  lwd.smooth = 2, col.smooth = "orange"))
```



15. 회귀모델의 설명력

- **결정계수:** *coefficient of determination*
  - R^2(*R-squared*): **선형 회귀식의 설명력 지표**
    - $R^2 = \frac{SSE(Explained\ Sum\ of\ Squares)}{SST(Total\ Sum\ of\ Squares)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
    - $R^2 = 0$: **독립변수와 종속변수 간의 선형 관계가 존재하지 않음**
    - $R^2 = 1$: **독립변수와 종속변수 간에는 완전한 선형 관계가 존재함**
  - *Adjusted* R^2: **다중 독립변수의 영향을 줄여줌**
    - R^2**는 독립변수의 개수가 증가하면 항상 값이 증가함**
    - **독립변수의 개수가 많아지면 페널티를 부과하여 설명력을 보정**
    - **과적합**(*overfitting*)**에 대한 고려**

15. 회귀모델의 설명력

```
> model <- lm(mpg ~ cyl + disp + hp + drat + wt, data = df)
> summary(model)
Call:
lm(formula = mpg ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7014 -1.6850 -0.4226  1.1681  5.7263

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.00836    7.57144   4.756  6.4e-05 ***
cyl         -1.10749    0.71588  -1.547  0.13394
disp         0.01236    0.01190   1.039  0.30845
hp          -0.02402    0.01328  -1.809  0.08208 .
drat         0.95221    1.39085   0.685  0.49964
wt          -3.67329    1.05900  -3.469  0.00184 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.538 on 26 degrees of freedom
Multiple R-squared:  0.8513,     Adjusted R-squared:  0.8227
F-statistic: 29.77 on 5 and 26 DF,  p-value: 5.618e-10
```

15. 회귀모델의 설명력

```
> model <- lm(mpg ~ hp + wt, data = df)
> summary(model)

Call:
lm(formula = mpg ~ hp + wt, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,     Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

15. 회귀모델의 설명력

```
> model <- lm(mpg ~ wt, data = df)
> summary(model)

Call:
lm(formula = mpg ~ wt, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,     Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

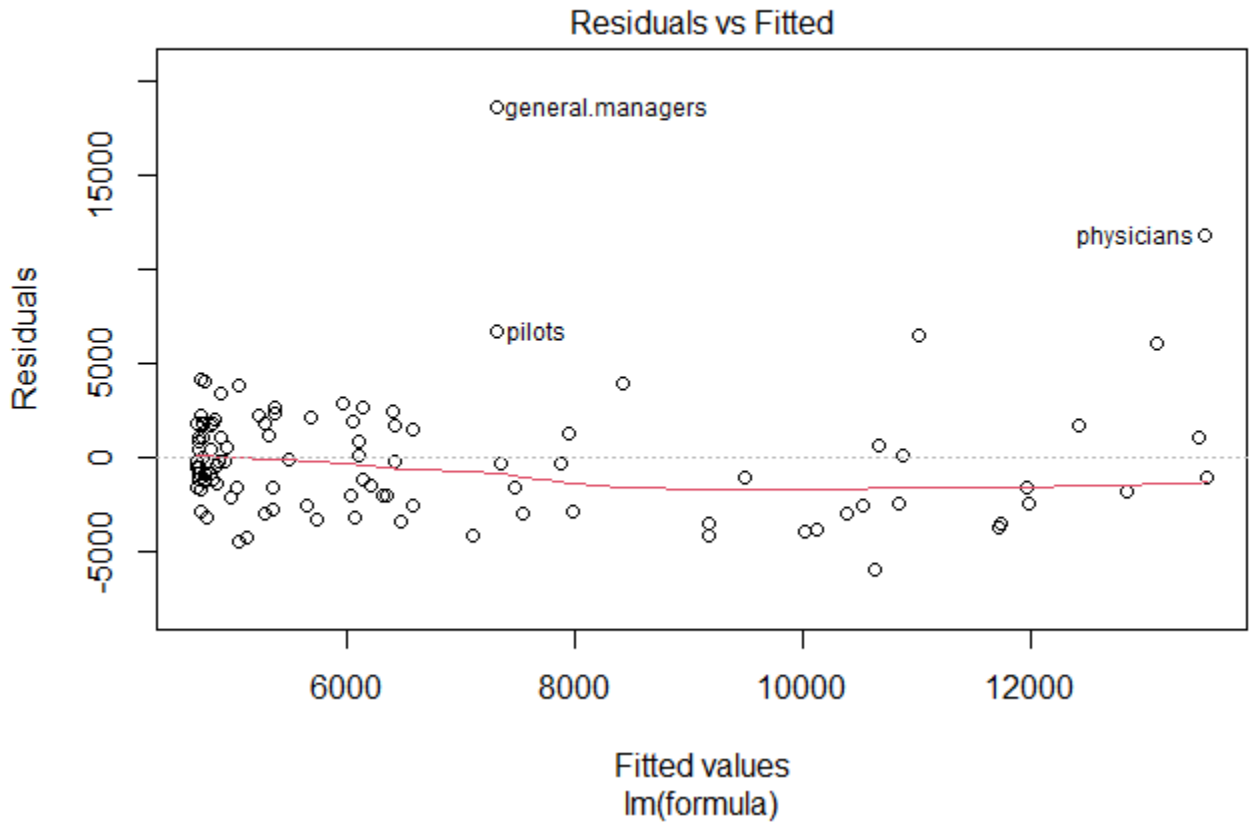
15. 회귀모델의 설명력

- **선형회귀 모델을 적용하기 위한 전제 조건:**
  - **선형성**: *linearity*
    - 독립변수와 종속변수 간의 선형적 관계가 존재한다.
  - **정규성**: *normality*
    - 종속변수의 값들이 정규분포를 가진다.
  - **등분산성**: *homoscedasticity*, *homogeneity of variance*
    - 종속변수 값들의 분포는 모두 동일한 분산을 가진다.
  - **독립성**: *independence*
    - 모든 독립변수의 관측값들은 서로 독립이다.

15. 회귀모델의 설명력

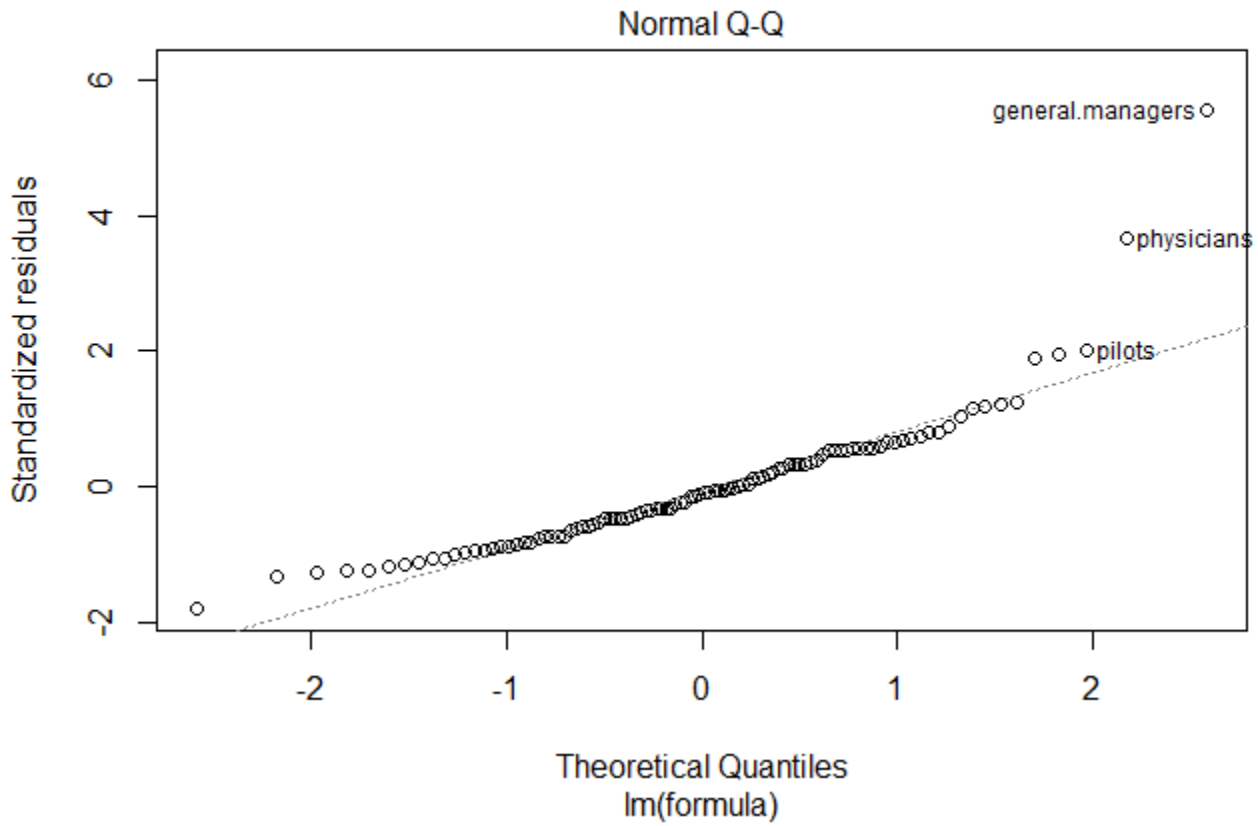
- **선형성 진단**: Residuals .vs. Fitted

```
> plot(model)
```



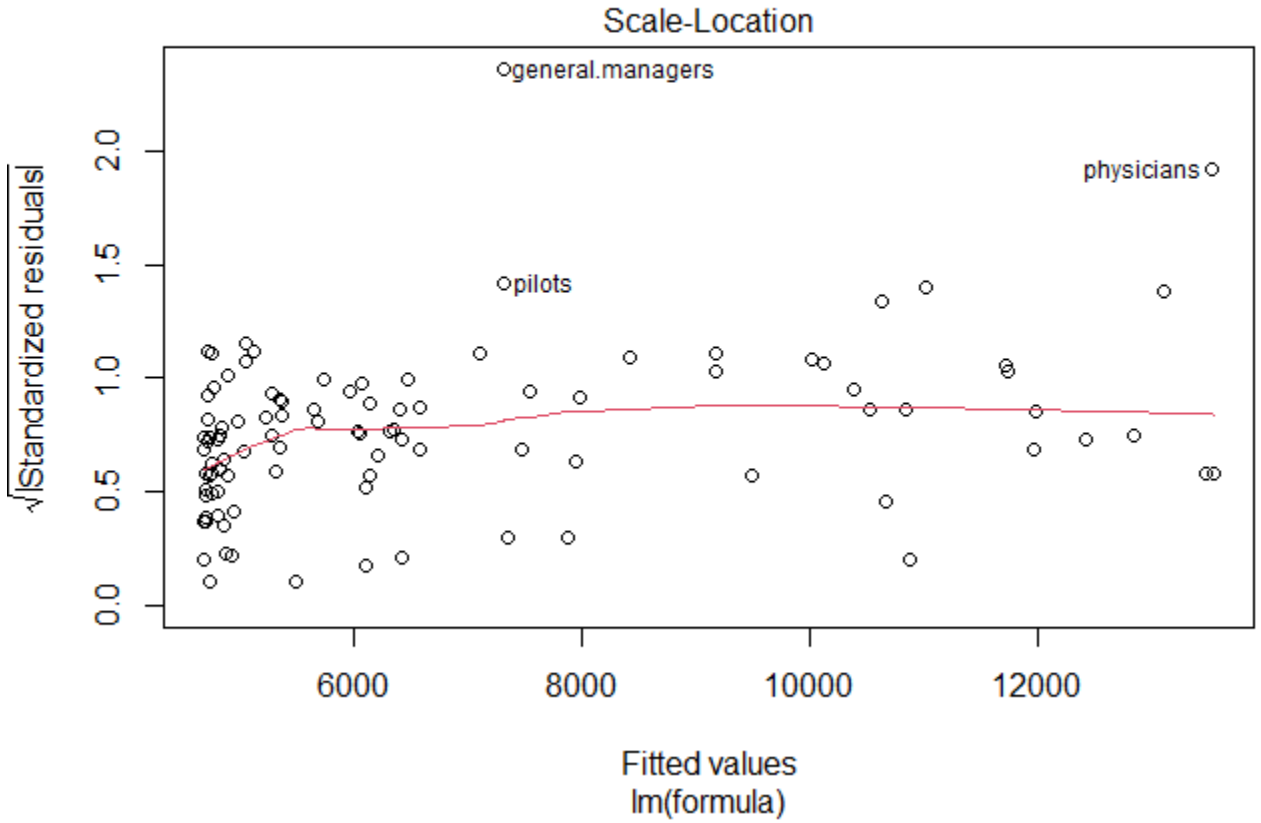
15. 회귀모델의 설명력

- **정규성 진단**: Normal Q-Q Plot



15. 회귀모델의 설명력

- **등분산성 진단:** Scale-Location Plot

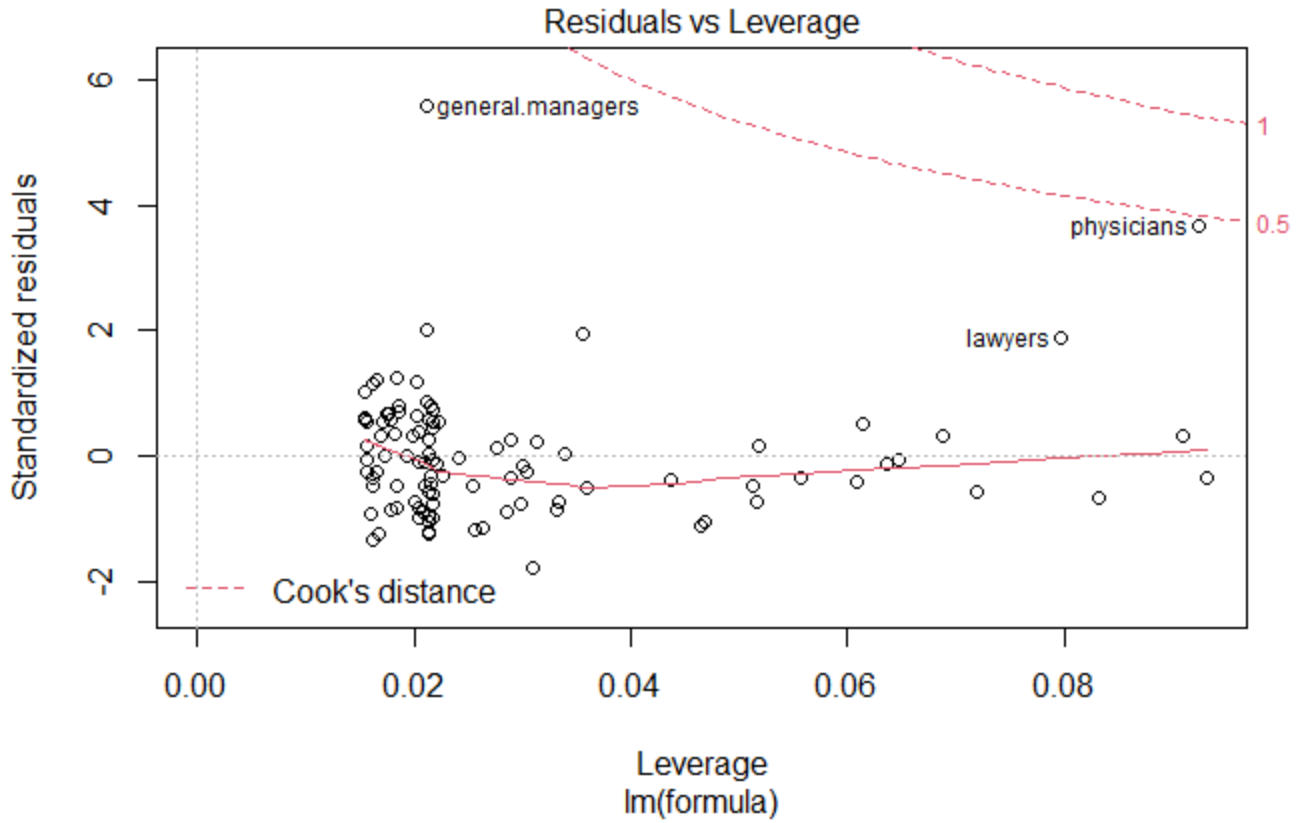


15. 회귀모델의 설명력

- **독립성 진단:**
  - 독립성은 model의 plot으로는 확인이 불가함
    - 독립성 여부는 연구자의 데이터에 대한 이해를 바탕으로 판단

Residual .vs. Leverage:

레버리지가 높은 지점 확인(이상치 등)



15. 회귀모델의 설명력

- **회귀 모델의 선택:**
  - 모델의 예측 정확도와 간명도를 높이기 위한 독립변수의 선택
    - 예측 정확도 (predictive accuracy): 모델의 데이터 적합도
    - 간명도 (parsimony): 모델의 간결함과 재현가능성
  - 독립변수의 숫자에 따른 회귀 모델의 특성
    - 많은 변수를 사용하면 데이터 적합도는 증가하지만 간명도가 떨어짐
    - 최소한의 예측변수를 사용하면 재현성이 증가함
    - 일반적으로 가능한 단순한 모델을 선택하는 것이 바람직

15. 회귀모델의 설명력

- 마력(hp)과 무게(wt)가 포함된 모델과 배기량(disp)과 기어비(drat)가 포함된 모델의 적합도 비교

```
> mtcars.lm1 <- lm(mpg ~ hp + wt, data = mtcars)
> mtcars.lm2 <- lm(mpg ~ hp + wt + disp + drat, data=mtcars)
> anova(mtcars.lm1, mtcars.lm2)
Analysis of Variance Table

Model 1: mpg ~ hp + wt
Model 2: mpg ~ hp + wt + disp + drat
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     29 195.05
2     27 182.84  2     12.21 0.9016 0.4178
```

15. 회귀모델의 설명력

- AIC 지표:
  - AIC: *Akaike Information Criterion*
    - 모델의 적합도와 파라미터의 개수를 함께 고려한 정보량의 척도
      - $AIC = 2\,(log\text{-}likelihood) + 2k$
    - 일반적으로 AIC 값이 작을수록
      - 더 적은 개수의 파라미터로 적절한 적합도를 달성하고 있음

```
> AIC(mtcars.lm1, mtcars.lm2)
           df      AIC
mtcars.lm1  4 156.6523
mtcars.lm2  6 158.5837
```

15. 회귀모델의 설명력

- **다중회귀분석시 독립변수를 선택하는 방법:**
  - **전진선택법**: *forward* selection
    - 상수항을 갖는 모델에서 시작해서 단계별로 한 개씩 독립변수를 추가
  - **후진선택법**: *backward* selection
    - 모든 독립변수가 포함된 모델에서 시작해서 단계별로 독립변수를 제거
  - **단계선택법**: *stepwise* selection
    - 전진선택법과 후진선택법을 혼합하여 수행

15. 회귀모델의 설명력

- mtcars 데이터셋에서 후진선택법으로 회귀모델 구축

```
> mtcars.lm <- lm(mpg ~ hp + wt + disp + drat, data=mtcars)
> step(mtcars.lm, direction="backward")
Start:  AIC=65.77
mpg ~ hp + wt + disp + drat

       Df Sum of Sq    RSS    AIC
- disp  1     0.844 183.68 63.919
<none>              182.84 65.772
- drat  1    12.153 194.99 65.831
- hp    1    60.916 243.75 72.974
- wt    1    70.508 253.35 74.209

Step:  AIC=63.92
mpg ~ hp + wt + drat

       Df Sum of Sq    RSS    AIC
- drat  1    11.366 195.05 63.840
<none>              183.68 63.919
- hp    1    85.559 269.24 74.156
- wt    1   107.771 291.45 76.693
```

15. 회귀모델의 설명력

```
Step:  AIC=63.84
mpg ~ hp + wt

       Df Sum of Sq    RSS    AIC
<none>              195.05 63.840
- hp    1    83.274 278.32 73.217
- wt    1   252.627 447.67 88.427

Call:
lm(formula = mpg ~ hp + wt, data = mtcars)

Coefficients:
(Intercept)           hp           wt
   37.22727     -0.03177     -3.87783
```

- 더미변수를 이용한 회귀분석
  - 회귀분석을 위한 변수가 연속형 변수가 아닐 때
    - 더미변수로 변환하여 회귀분석을 할 수 있음
  - 더미변수: dummy variable
    - 어떤 속성(또는 사건)이 존재할 경우 값을 1로, 존재하지 않으면 0으로 인코딩

15. 회귀모델의 설명력

- 계절(봄, 여름, 가을, 겨울) 변수를 봄을 기준변수로 회귀분석의 독립변수로 이용하고자 할 때

계절 변수	변수값	D1	D2	D3
봄	1	0	0	0
여름	2	1	0	0
가을	3	0	1	
겨울	4	0	0	1

15. 회귀모델의 설명력

- lm() 함수는 독립변수가 범주형 변수이면 자동으로 더미 변수로 변환한 후 회귀분석을 수행함

```
> str(InsectSprays)
'data.frame':  72 obs. of  2 variables:
 $ count: num  10 7 20 14 14 12 10 23 17 20 ...
 $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...

> levels(InsectSprays$spray)
[1] "A" "B" "C" "D" "E" "F“

> tapply(InsectSprays$count, InsectSprays$spray, mean)
        A         B         C         D         E         F
14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

15. 회귀모델의 설명력

```
> sprays.lm <- lm(count ~ spray, data=InsectSprays)
> summary(sprays.lm)
Call:
lm(formula = count ~ spray, data = InsectSprays)

Residuals:
   Min     1Q Median     3Q    Max
-8.333 -1.958 -0.500  1.667  9.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5000     1.1322  12.807  < 2e-16 ***
sprayB        0.8333     1.6011   0.520    0.604
sprayC      -12.4167     1.6011  -7.755 7.27e-11 ***
sprayD       -9.5833     1.6011  -5.985 9.82e-08 ***
sprayE      -11.0000     1.6011  -6.870 2.75e-09 ***
sprayF        2.1667     1.6011   1.353    0.181
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.922 on 66 degrees of freedom
Multiple R-squared:  0.7244,         Adjusted R-squared:  0.7036
F-statistic:  34.7 on 5 and 66 DF,  p-value: < 2.2e-16
```

15. 회귀모델의 설명력

- contrasts() 함수를 이용하여 더미변수의 코딩 구조를 확인 가능함

```
> contrasts(InsectSprays$spray)
  B C D E F
A 0 0 0 0 0
B 1 0 0 0 0
C 0 1 0 0 0
D 0 0 1 0 0
E 0 0 0 1 0
F 0 0 0 0 1
```

15. 회귀모델의 설명력

- 기준 범주를 변경하고자 할 때는 relevel() 함수를 이용

```
> respray <- relevel(InsectSprays$spray, ref=6)
> sprays.lm <- lm(count ~ respray, data=InsectSprays)
> summary(sprays.lm)

> contrasts(relevel(InsectSprays$spray, ref=6))
  A B C D E
F 0 0 0 0 0
A 1 0 0 0 0
B 0 1 0 0 0
C 0 0 1 0 0
D 0 0 0 1 0
E 0 0 0 0 1
```

Any Questions?