KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

# 11. 진단(Diagnostics)

## 정치와 데이터분석

박상훈 (sh.park.poli@gmail.com)
강원대학교

# 오늘의 목표

## 10:05-10:45

회귀모형에 대한 진단(regression diagnostics)과 통계적 유의성 및 신뢰구간에 대해 다시 한 번 복습

`{lindia}` 패키지를 활용한 시각적 회귀 진단 방법을 배우고 기본 R의 진단 함수와 비교

가정 위반이 발견되는 경우의 해결 방법에 대한 간단한 논의

## 11:00-11:40

모델 선택에 대한 논의

## 11:55-12:35

실습과제 해설 및 질의응답

# Part I. 회귀모형에 대한 진단

# 진단(Diagnostics)

## OLS 추정의 Gauss-Markov 가정 (MLR 확장)

MLR 1. 선형성 (Linearity)

MLR 2. 무작위 표본 (Random Sample)

MLR 3. 내생성 없음 (No Endogeneity)/외생성 (Exogeneity)

MLR 4. 오차항의 동분산성 (Homoskedasticity)

MLR 5. 완벽한 다중공선성 없음 (No Perfect Multicollinearity)

⇒ **위 가정들이 충족되면 OLS 추정치는 편향되지 않고(unbiased), 효율적(최소분산)이며, Gaus-Markov 정리에 따라 BLUE**

# 진단(Diagnostics)

## 가정 위반으로 인한 문제들

회귀모델을 수립하기 전 데이터에 대한 면밀한 진단을 통해 문제 상황을 예측하고 미리 대처할 수 있음. **대표적인 문제 상황**

- 이탈치(Outliers): 다른 관측치들과 비교해 예측변수 값이 극단적인 관측치( $X$ 방향의 특이값)
- 비정규성(Non-normal errors): 잔차가 정규분포를 따르지 않는 경우
- 이분산성(Heteroscedasticity): 오차의 분산이 일정하지 않은 경우(특정 범위에서 잔차의 변동 폭 증가 등)
- 비선형성 (Non-linearity): 실제 관계가 선형 형태가 아닌 경우(모델이 구조를 놓침)
- 공선성 (Collinearity): 예측변수들 간 강한 상관관계로 계수 추정의 불안정성이 커진 경우

# 진단(Diagnostics)

## 가정 위반으로 인한 문제들

회귀모델을 수립하기 전 데이터에 대한 면밀한 진단을 통해 문제 상황을 예측하고 미리 대처할 수 있음. **대표적인 문제 상황**

이러한 문제들은 회귀분석의 가정 위반에 해당하며, 진단을 통해 발견하고 적절히 대처해야 함.

# 진단(Diagnostics)

## 가정 위반으로 인한 문제들: 이탈치

다른 데이터들과 비교하여 값이 유난히 튀는 관측값

- 단순히 종속변수 $Y$의 극단값뿐만 아니라, 다변량 공간에서의 특이점을 고려해야 함.
- 단변량 이탈치: 한 변수의 분포에서 극단치. 아래 왼쪽 그림처럼 밀도 분포의 꼬리 부분에 드문 값이 있는 경우
- 양변량/다변량 이탈치: 여러 변수 간 관계에서의 특이점. 오른쪽 그림처럼 $X$-축과 $Y$-축의 관계 상 다른 점들과 동떨어진 점이 있는 경우

단변량 분석에서의 이상치가 항상 회귀 분석에서 영향이 큰 이상치로 이어지는 것은 아님. 예측변수 공간에서의 특이점이 중요한 이유.

# 진단(Diagnostics)

## 가정 위반으로 인한 문제들: 이탈치

단변량 이탈치 | 단변량 이탈치 | 양변량 이탈치 | 양변량 이탈치

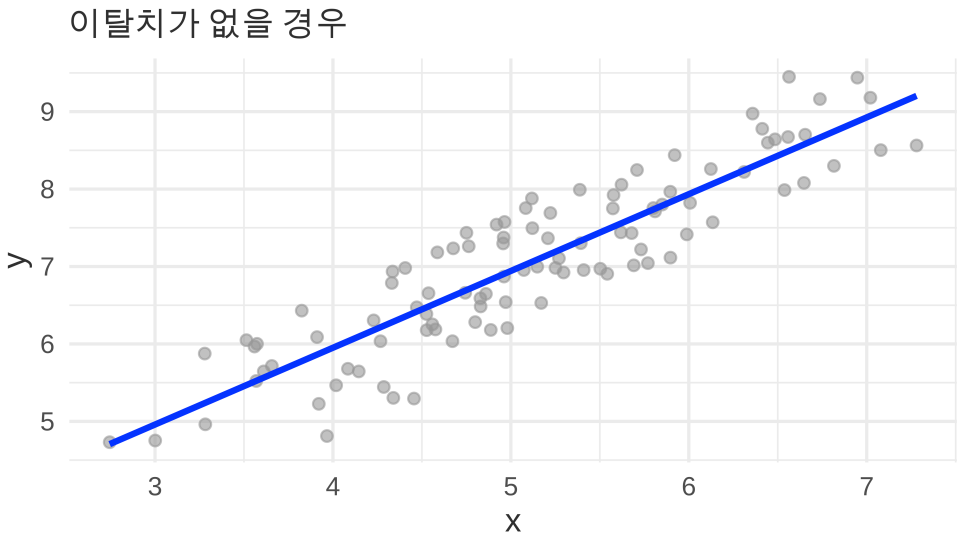
```
set.seed(123)
pool <- rnorm(100, mean = 4, sd = 1)
pool <- c(pool, 10, 11, 10, 9.5, 10.5) # X의 극단값 추가
data.frame(value = pool) |>
  ggplot(aes(x = value)) +
  geom_density(fill = "skyblue", color = "skyblue", alpha = 0.5) +
  labs(subtitle = "단변량 분석의 이탈치", x = "값", y = "밀도")
```

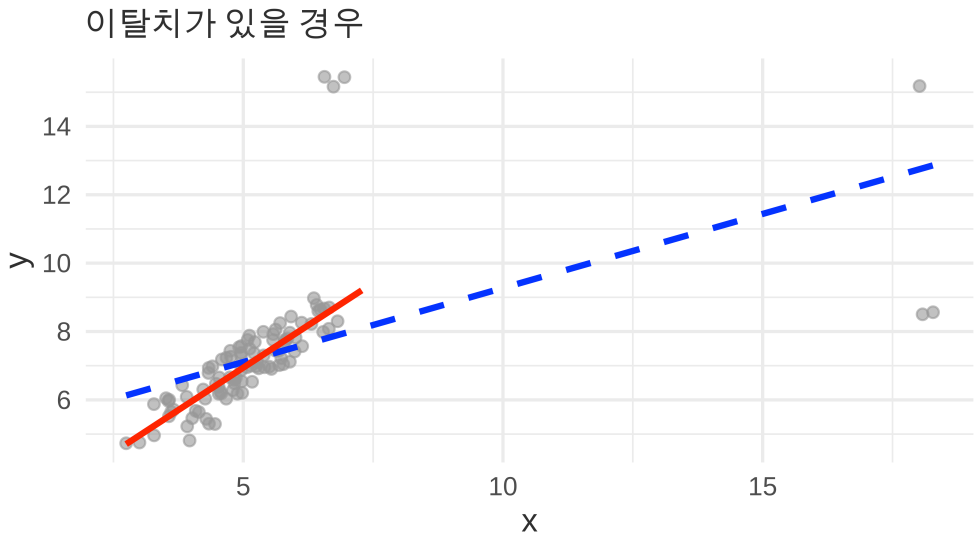
# 진단(Diagnostics)

## 가정 위반으로 인한 문제들: 이탈치

### 이탈치의 영향

회귀분석에서 이탈치는 적은 수라 할지라도 모델 적합선의 기울기와 절편 등에 큰 영향을 미칠 수 있음. 극단적인 관측치 하나가 회귀계수 추정에 큰 영향을 주는 것을 확인할 수 있음.



# 진단(Diagnostics)

## 레버리지(Leverage)와 영향력(Influence)

레버리지란 한 관측치의 예측변수 $X$ 값이 전체 X 분포의 평균에서 얼마나 떨어져 있는지를 나타내는 척도

- 회귀분석에서 레버리지는 관측치가 X 공간에서 얼마나 극단적인 위치에 있는가를 의미
- 레버리지의 대표적인 지표인 모자값(hat-value) $h_i$는 관측치 $x_i$와 다른 관측치들의 평균 $\bar{x}$ 사이의 거리를 정량화한 값
  - 단순선형회귀의 경우 $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$
- $h_i$ 값이 큰 관측치일수록 해당 점은 $X$ 공간에서 멀리 떨어져 있어 높은 레버리지를 가짐. 이런 점들은 회귀식 결정에 있어 큰 잠재적 영향력을 지님.

# 진단(Diagnostics)

## 레버리지(Leverage)와 영향력(Influence)

잔차(Residual): 관측치의 실제 $Y$값과 회귀모형이 예측한 $\hat{Y}$값 간의 차이, 즉 $e_i = y_i - \hat{y}_i$

- 잔차가 크다는 것은 해당 관측치가 $Y$ 방향으로 다른 점들과 많이 동떨어져 있다(오차가 크다)는 의미: **불일치(discrepancy)**
- 회귀모형 적합 시, $X$ 공간에서 레버리지가 큰 관측치는 회귀식이 그 점을 지나도록 당기기 때문에 잔차가 작게 나타나는 경향이 있음. 반면 $X$ 영역 중심부의 관측치는 레버리지가 낮아 잔차가 크게 남을 수도 있음.

잔차는 모델 적합이 얼마나 잘 되었는지를 보여주지만, 레버리지와 함께 해석해야 이상치를 효과적으로 판단할 수 있음.

# 진단(Diagnostics)

## 레버리지(Leverage)와 영향력(Influence)

영향력(Influence): 레버리지와 잔차(불일치) 두 가지 요소의 결합

- 한 관측치의 레버리지가 높고 잔차까지 크다면 회귀계수 추정에 막대한 영향을 주게 됨.

$$\text{Influence} \propto (\text{Leverage}) \times (\text{Discrepancy})$$

- 일반적으로는 두 요소의 함수로 정의되며, 아래 쿡의 거리(Cook's Distance) 등이 그 예

# 진단(Diagnostics)

## 레버리지(Leverage)와 영향력(Influence)

표준화 잔차(Standardized residual)

- 잔차 $e_i$를 해당 관측치의 추정 표준편차로 나누어 단위를 표준화시킨 값
- $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$, 잔차의 크기를 공통 척도로 비교할 수 있게 해줌.

스튜던트화 잔차(Studentized residual)

- 표준화 잔차와 유사하지만, 해당 관측치를 제거하고 계산한 표준오차 $\hat{\sigma}(-i)$로 나눈 값
  $t_i = \frac{e_i}{\hat{\sigma}(-i)\sqrt{1-h_i}}$
- 스튜던트화 잔차는 $t$-분포를 따르므로 통계적 유의성 검정에 사용되며,
  $|\text{studentized 잔차}| > 2(\text{또는 } 3)$인 관측치는 이상치로 의심

# 진단(Diagnostics)

## 영향력 지표: 쿡의 거리(Cook's Distance)

**쿡의 거리(Cook's D)**: 특정 관측치를 제거했을 때 회귀모형의 적합 값들이 얼마나 변하는지를 나타내는 대표적인 영향력 지표

- 한 점을 제거하고 나머지 데이터로 재적합한 모델과 원래 모델의 예측값 변화량을 종합적으로 측정
- $D_i = \sum_{j=1}^{n} (\hat{y}j - \hat{y}j(-i))^2 / (p \cdot \hat{\sigma}^2)$, 여기서 $\hat{y}_{j(-i)}$는 $i$번째 관측치를 제외하고 적합한 예측값이며 $p$는 모형 모수 개수
- 값이 클수록 해당 관측치가 모형 전체에 큰 영향을 미친다는 뜻

# 진단(Diagnostics)

## 영향력 지표: 쿡의 거리(Cook's Distance)

**쿡의 거리(Cook's D)**: 특정 관측치를 제거했을 때 회귀모형의 적합 값들이 얼마나 변하는지를 나타내는 대표적인 영향력 지표

Cook's D 값이 큰 관측치를 제거하면 다른 관측치들의 예측값이 크게 변화하므로, 그 점은 회귀분석 결과에 중대한 영향을 끼치고 있다고 볼 수 있음.

일반적으로 $D_i > \frac{4}{n}$인 경우 해당 관측치는 상당한 영향력을 가진 것으로 판단(표본 크기가 충분히 큰 경우). 혹은 단순 기준으로 $D_i > 1$이면 특별히 큰 값으로 간주하기도 함.

# 진단(Diagnostics)

## 영향력 지표: DFITS와 DFBETAS

DFITS(Difference in Fits)

- $i$번째 관측치를 제거했을 때 해당 관측치의 적합값 $\hat{y}i$가 얼마나 변하는지를 나타냄.
- $DFITS_i = \frac{\hat{y}i - \hat{y}i(-i)}{\hat{\sigma}(-i)\sqrt{h_i}}$ 로 정의
  - 한 관측치가 자신의 예측값을 얼마나 끌어올렸는지를 표준화하여 보여줌.
- $|DFITS_i|$ 값이 큰 경우, 그 관측치가 자신의 예측 결과에 큰 영향. 일반적인 기준으로 $|DFITS_i| > 2\sqrt{\frac{p}{n}}$이면 영향력이 크다고 판단

# 진단(Diagnostics)

## 영향력 지표: DFITS와 DFBETAS

DFBETA/DFBETAS: 특정 관측치를 제거했을 때 회귀계수 $\beta_j$가 얼마나 변화하는지 나타냄.

- $DFBETA_{ij} = \hat{\beta}j - \hat{\beta}j(-i)$
  - $j$번째 회귀계수가 해당 관측치 $i$를 제외했을 때 얼마나 달라지는지를 의미. 이를 표준화한 것이 DFBETAS
- 만약 어떤 관측치를 제외했더니 특정 계수의 값이 크게 변한다면, 그 관측치는 해당 설명변수의 계수 추정에 큰 영향을 준 것. 일반적으로 $|DFBETAS_{ij}| > \frac{2}{\sqrt{n}}$이면 해당 관측치가 계수 $\beta_j$ 추정에 상당한 영향력이 있다고 봄.

# 진단(Diagnostics)

## 영향력 지표: DFITS와 DFBETAS

DFBETAS는 각 회귀계수마다 값을 가지므로, 영향력 있는 관측치는 어떤 계수에서는 크게 나타날 수 있음.

- 한 관측치가 특히 특정 독립변수의 계수 추정에 영향을 준다면 그 변수에 대한 DFBETAS가 높음.

# 진단(Diagnostics)

## 영향력 지표: COVRATIO

$i$번째 관측치를 제거했을 때 회귀계수 추정치들의 공분산 행렬의 변동을 나타내는 지표

- 해당 관측치가 빠졌을 때 모형의 전반적인 추정 안정성이 어떻게 변하는지를 평가
- $\mathrm{COVRATIO}_i = \frac{\det(\mathbf{Cov}(-i))}{\det(\mathbf{Cov})}$, 여기서 $\mathbf{Cov}$는 전체 데이터로 적합한 모형의 공분산 행렬, $\mathbf{Cov}(-i)$는 $i$를 제외하고 적합한 경우의 공분산 행렬
- $\mathrm{COVRATIO}_i$가 1에 가까우면 해당 관측치의 제거가 분산-공분산 구조에 큰 변화가 없음을 의미
  - 1에서 크게 벗어나면 그 관측치로 인해 모형의 전반적 추정 분산이 많이 변함을 의미

# 진단(Diagnostics)

## 영향력 지표: COVRATIO

일반적으로 $\mathrm{COVRATIO}_i$가 $1 \pm \frac{3p}{n}$ 범위를 벗어나면 이상치로 간주

- 예를 들어 $\mathrm{COVRATIO}_i < 1 - \frac{3p}{n}$이면 해당 관측치 제거 시 분산이 크게 감소(모형이 더 정확해짐)하거나, 반대로 $> 1 + \frac{3p}{n}$이면 제거 시 분산이 크게 증가하는 관측치
- 이러한 관측치는 모형의 안정성에 영향을 미치는 점으로 볼 수 있음.

# 진단(Diagnostics)

## 기본 R 진단 vs. {lindia} 진단

R의 기본 함수인 plot(lm)은 네 가지 주요 진단 플롯을 제공: 잔차 vs 적합값, 정규 Q-Q, Scale-Location, 잔차 vs 레버리지.

mtcars 데이터의 예시 모형으로 lm(mpg ~ wt + hp, data = mtcars)을 적합하여, 기본 플롯과 {lindia} 패키지 함수의 결과를 비교

```
# mtcars 데이터로 회귀모형 적합
lm_model <- lm(mpg ~ wt + hp, data = mtcars)
get_regression_table(lm_model)
```

```
## # A tibble: 3 × 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   37.2       1.60      23.3    0        34.0    40.5
## 2 wt          -3.88      0.633     -6.13   0        -5.17   -2.58
## 3 hp          -0.032     0.009     -3.52   0.001    -0.05   -0.013
```

# 진단(Diagnostics)

## 잔차 대 적합값 플롯(Residuals vs Fitted)

이 플롯은 모형의 선형성 및 등분산성 가정을 진단하는데 사용

- 예측값 ($\hat{Y}$)에 대한 잔차를 산점도로 그리고, 잔차의 평균이 0을 중심으로 특정 패턴 없이 무작위로 분포하는지 확인

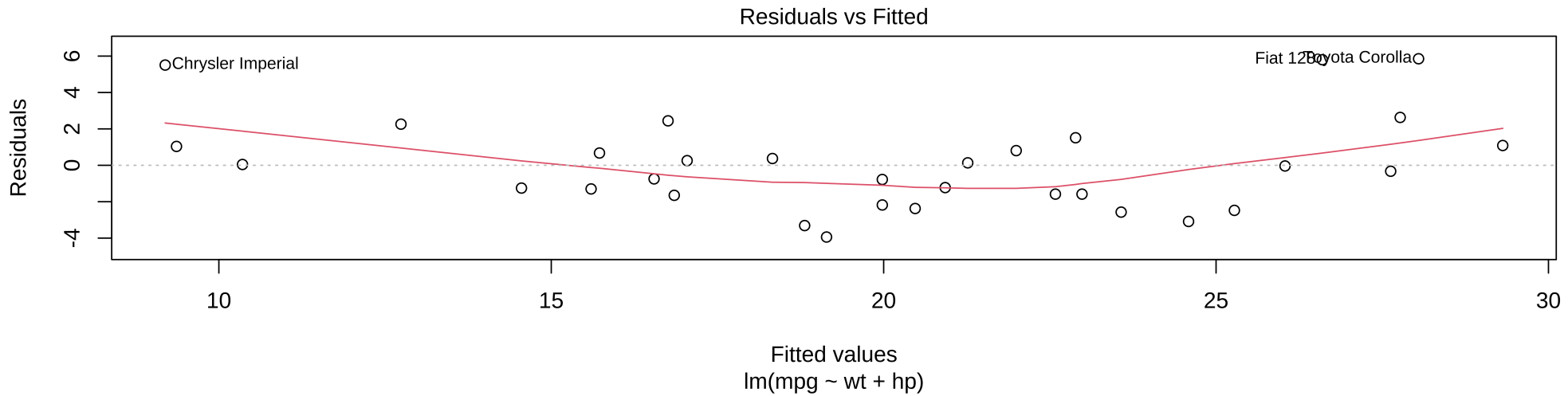
**이상적인 경우**: 잔차들이 0선을 기준으로 랜덤하게 흩어진 패턴(아무 곡선형이나 뚜렷한 구조 없음). 또한 잔차의 흩어지는 폭이 일정해야 등분산성을 만족

**패턴이 보이는 경우**: 예를 들어 잔차들이 곡선을 그리면 비선형성을 의심할 수 있고, 예측값이 커짐에 따라 잔차의 산포가 점점 커지면 이분산성 문제가 존재할 수 있음.

# 진단(Diagnostics)

## 잔차 대 적합값 플롯(Residuals vs Fitted)

```
plot(lm_model, which = 1)
```



# 진단(Diagnostics)

## 잔차 대 적합값 플롯(Residuals vs Fitted)

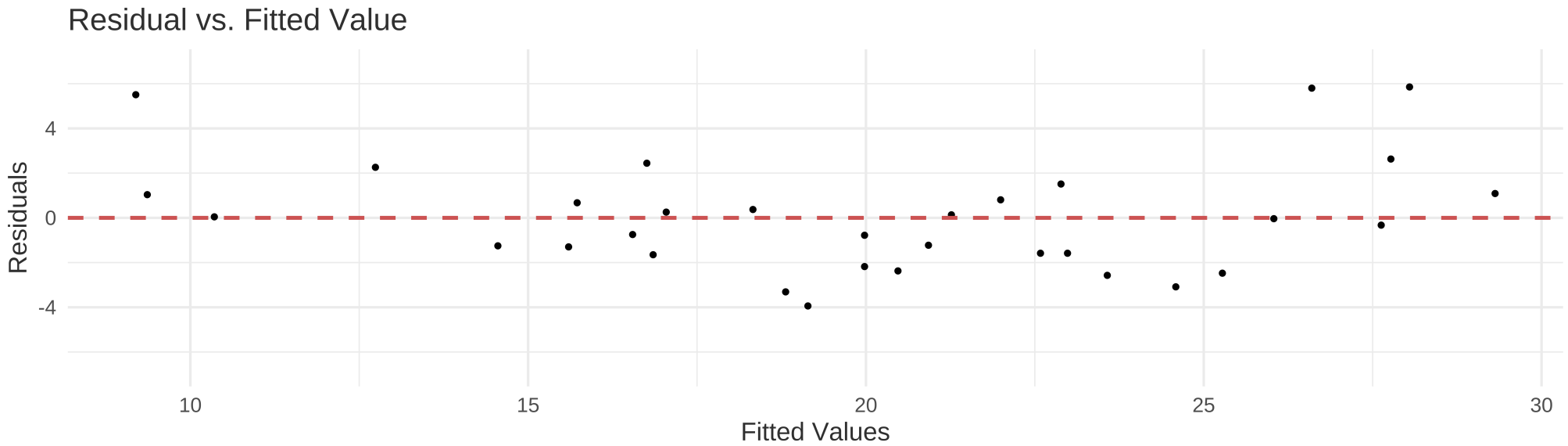
붉은 실선은 loess 추세선: 잔차들의 비선형 경향

- 이상적으로 이 선이 수평선에 가까워야 선형성 가정을 충족
- 약간 굽은 형태의 추세가 보인다면, 모형에 다항식 또는 누락 변수 고려 등 비선형 패턴 보완이 필요함을 시사
- 또한 잔차의 산포가 예측값 구간에 따라 달라진다면 이분산성 의심

# 진단(Diagnostics)

## 잔차 대 적합값 플롯(Residuals vs Fitted)

```
lindia::gg_resfitted(lm_model)
```



# 진단(Diagnostics)

## 잔차 대 적합값 플롯(Residuals vs Fitted)

`{lindia}` 패키지의 `residual_plot()`: 잔차 대 적합값의 산점도와 함께 추세를 보조적으로 표시. 기본 플롯과 유사하게 잔차들이 0을 중심으로 랜덤하게 분포하는지 보여주며, 추세선 또는 구간을 통해 패턴을 파악하기 쉬움.

잔차에 뚜렷한 곡선 경향이 있다면 선형성 가정 위배로 볼 수 있음.

# 진단(Diagnostics)

## 정규 Q-Q 플롯(Normal Q-Q Plot)

잔차가 정규분포를 따르는지를 시각적으로 점검

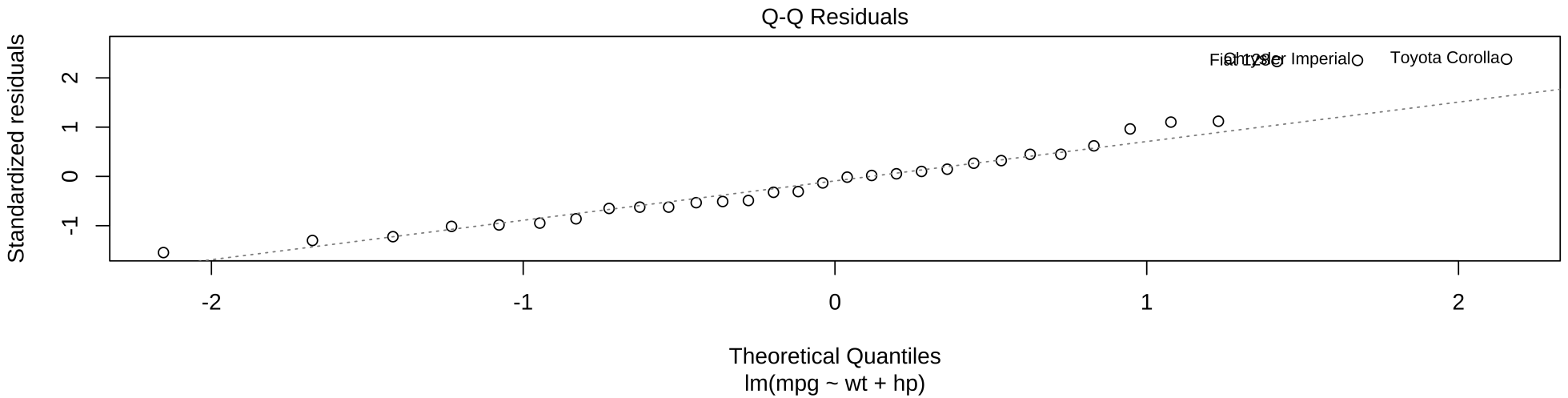
- 잔차들의 정규 Quantile-Quantile 그래프에서 점들이 대각선 상에 놓여 있으면 정규성 가정에 부합
- **이상적인 경우**: 모든 점들이 대체로 직선(45도 대각선) 부근에 놓임 → 잔차의 분포가 정규분포에 가깝다.
- **문제가 있는 경우**: 직선에서 크게 벗어나는 패턴이 보임. 예를 들어 S자 곡선 모양이면 꼬리가 두꺼운 분포이거나, 극단치들이 정규 예측보다 더 멀리 있음을 의미. 여러 집단의 혼합 분포일 가능성도 있음.

만약 중간 부분부터 직선을 벗어나 점들이 치우치는 경우, 잔차의 정규성에 의심이 가며, 모형에 누락된 변수나 형태 오특정(misspecification)이 있는지 고려(예: 오차 분포가 여러 봉우리를 가지면 설명변수 누락 가능성)

# 진단(Diagnostics)

## 정규 Q-Q 플롯(Normal Q-Q Plot)

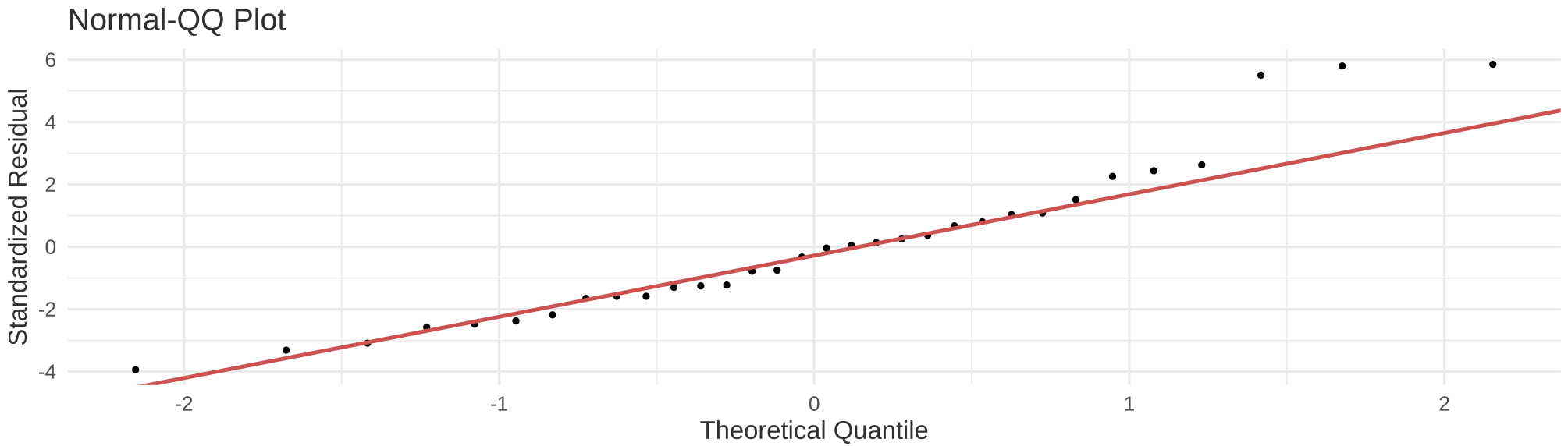
```
plot(lm_model, which = 2)
```



# 진단(Diagnostics)

## 정규 Q-Q 플롯(Normal Q-Q Plot)

```
lindia::gg_qqplot(lm_model)
```



# 진단(Diagnostics)

## 정규 Q-Q 플롯(Normal Q-Q Plot)

대각선은 이론적 정규분포선

- `mtcars` 예시에서 점들이 비교적 선을 따르지만 끝부분에서 약간 벗어난다면, 잔차 정규성에 경미한 위배가 있을 수 있음.
- 다만 표본 크기가 크다면 중심극한정리에 의해 잔차 정규성 가정이 다소 완화될 수 있음.
- 정규성 위배의 영향은 주로 신뢰구간과 $p-$값 추정의 정확성에 나타남.

잔차가 크게 비정규적이면 비모수적 추정이나 **변환**을 고려할 수 있음.

# 진단(Diagnostics)

## 등분산성 진단 플롯(Scale-Location Plot)

Scale-Location (또는 Spread-Location) 플롯

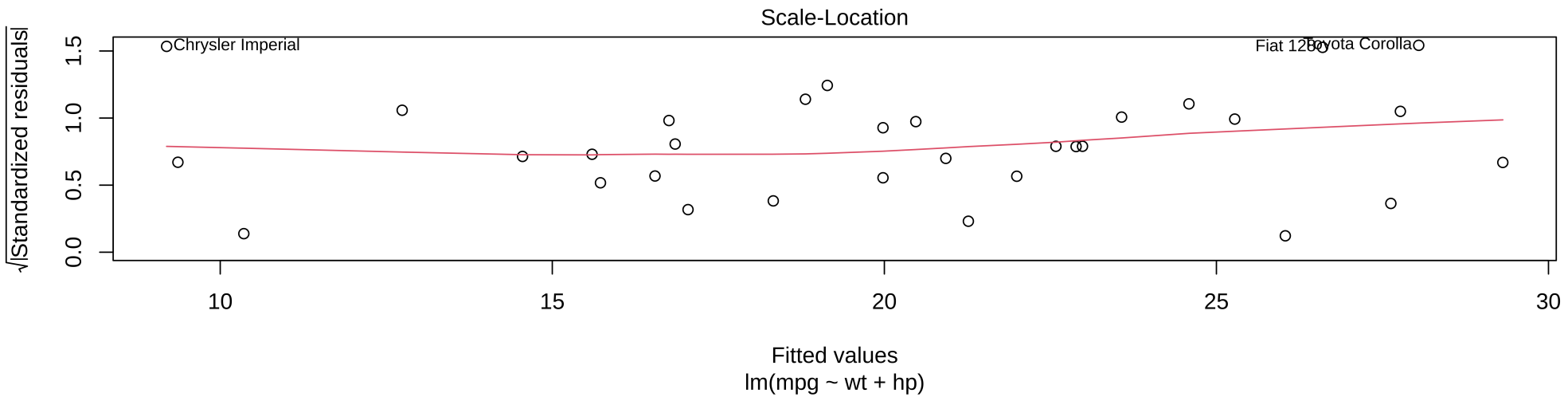
- 예측값에 따른 잔차의 표준편차 변화를 보여주어 등분산성 가정을 진단
- 일반적으로 $\sqrt{|e_i|}$ (잔차 절댓값의 제곱근) vs **$₩hat{y}_i$**를 산점도로 그리고 추세선을 확인
- **이상적인 경우**: 추세선이 수평으로 유지되고, 점들의 산포 범위가 예측값 전체 구간에 걸쳐 일정함 $\rightarrow$ 등분산성 만족
- **문제 상황**: 추세선이 기울어지거나 곡선 형태를 보이면 분산이 시스템적으로 변화함을 의미. 예를 들어 오른쪽으로 갈수록 위로 상승하는 형태라면 예측값이 클수록 잔차의 분산이 커지는 이분산성이 존재

# 진단(Diagnostics)

## 등분산성 진단 플롯(Scale-Location Plot)

mtcars 예시의 경우, 만약 약간의 상승 추세가 보인다면 연비가 높아질수록 오차 분산이 증가하는 경향이 있을 수 있음.

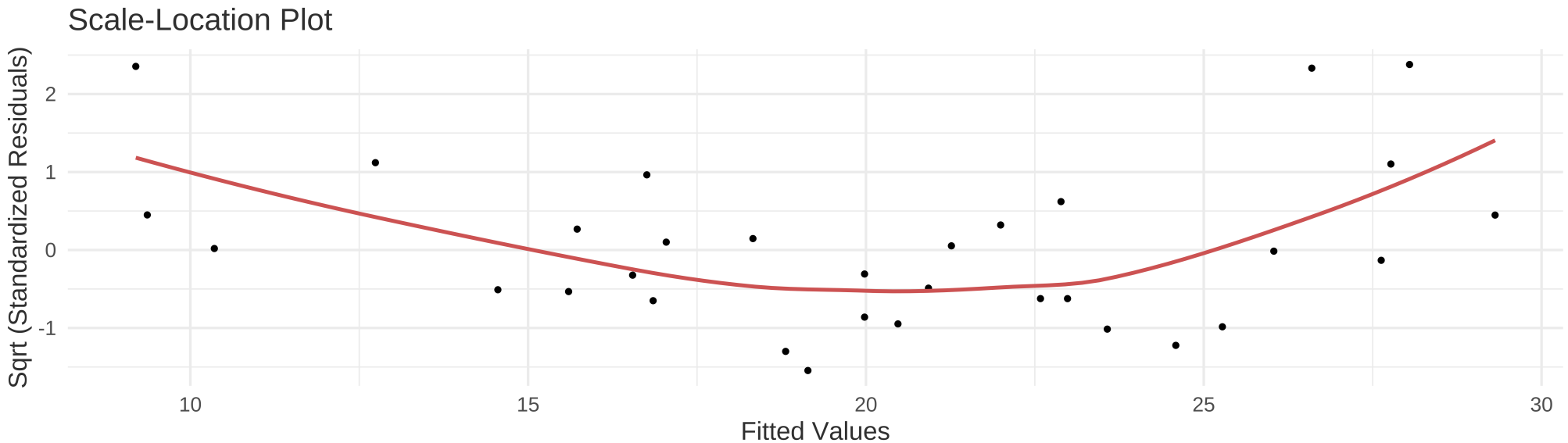
```
plot(lm_model, which = 3)
```



# 진단(Diagnostics)

## 등분산성 진단 플롯(Scale-Location Plot)

```
lindia::gg_scalelocation(lm_model)
```



# 진단(Diagnostics)

## 등분산성 진단 플롯(Scale-Location Plot)

그래프에서 빨간 추세선이 완만한 수평선을 유지하면 등분산성을 만족한 것.

- 그러나 추세선이 경향을 보이거나 점들이 깔때기 모양으로 퍼져 나가면 이분산성 문제가 의심
- 이분산성이 확인되면 **로버스트 표준오차(heteroskedasticity-consistent SE)**를 활용하여 추정 결과의 표준오차를 보정할 수 있음.
  - 로버스트 표준오차는 가정된 등분산성과 달리, 분산이 다른 경우에도 유효한 표준오차를 제공
  - R에서는 `{estimatr}` 패키지의 `lm_robust()` 함수를 사용하거나, `{sandwich}` 패키지와 `{lmtest}` 패키지로 Breusch-Pagan 테스트 후 보정할 수 있음.

# 진단(Diagnostics)

## 잔차 vs 레버리지 플롯(Residuals vs Leverage Plot)

극단적인 관측치들이 얼마나 모형에 영향력을 가지는지 확인하는 데 사용

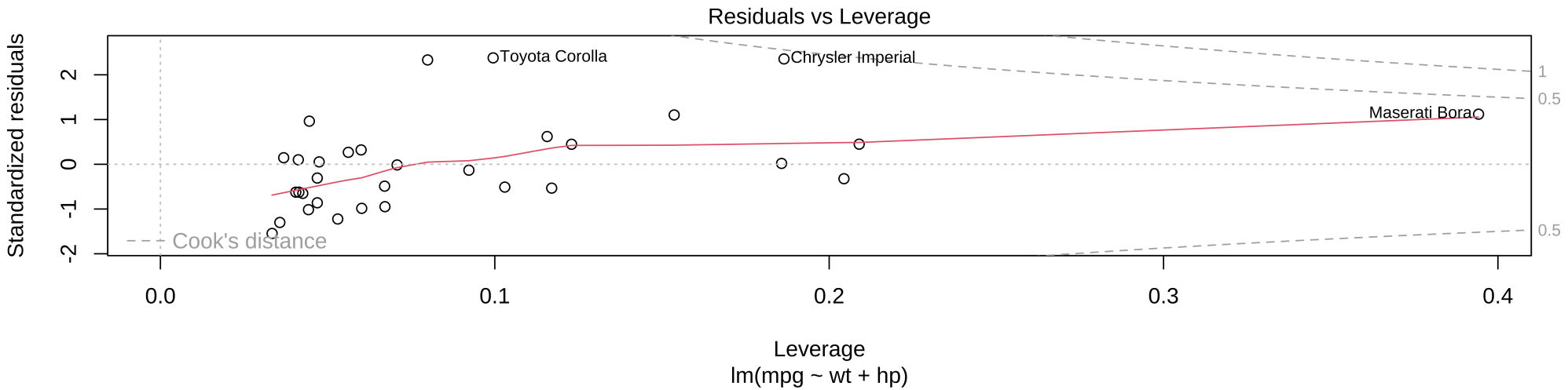
- $X$ 축은 레버리지($h_i$), $Y$ 축은 스튜던트화 잔차(또는 표준화 잔차)를 표시하며, 그래프 상에 쿡의 거리 기준선이 함께 그려짐.
- 레버리지가 매우 큰 관측치들($X$ 방향 극단치). 빨간 점선으로 표시된 쿡의 거리 기준선을 넘어서는 점이 있다면 그 점은 영향력이 큰 관측치로 간주

`mtcars` 예에서 우측 극단에 몇 개 점이 보이면, 해당 차량들이 특히 독특한 $X$ (중량, 마력 조합) 값을 가져 모형에 큰 영향력을 행사하고 있을 수 있음.

# 진단(Diagnostics)

## 잔차 vs 레버리지 플롯(Residuals vs Leverage Plot)

```
plot(lm_model, which = 5)
```



# 진단(Diagnostics)

## 잔차 vs 레버리지 플롯(Residuals vs Leverage Plot)

그래프의 오른쪽 끝부분에 위치한 점들이 높은 레버리지 관측치

- 만약 이 점들의 스튜던트화 잔차도 크다면 그래프 상단이나 하단에 위치

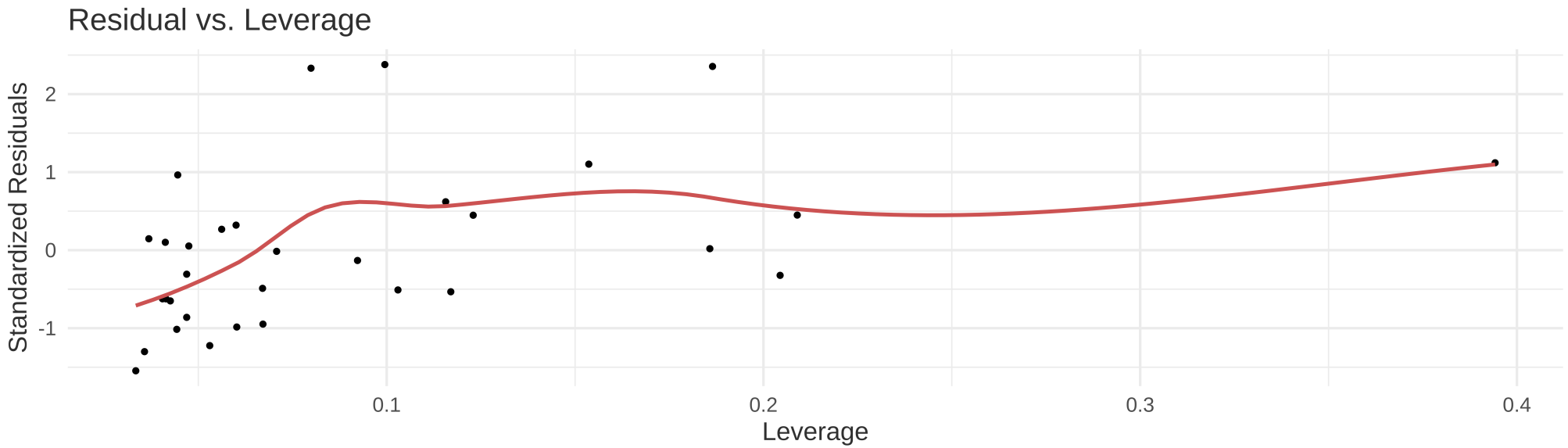
빨간 점선은 일정한 Cook's D 값을 나타내는 등고선

- 이 선을 넘는 점(그래프 상에 번호 표시됨)은 Cook's D가 큰 관측치로, 회귀 결과에 큰 영향력이 있다고 볼 수 있음.

# 진단(Diagnostics)

## 잔차 vs 레버리지 플롯: 영향력 플롯(Influence Plot)

```
lindia::gg_resleverage(lm_model)
```

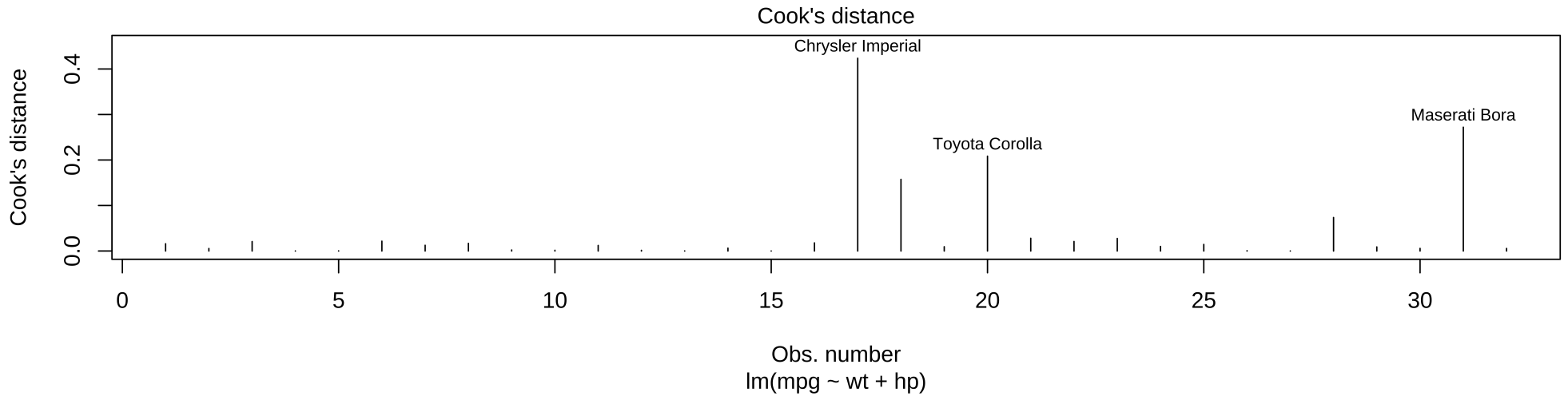


# 진단(Diagnostics)

## 쿡의 거리 플롯(Cook's Distance Plot)

쿡의 거리 플롯은 각 관측치별 Cook’s D 값을 나열한 그래프. 관측치의 인덱스(또는 이름) vs Cook’s D를 그려 어느 관측치의 Cook's D가 큰지 직접 보여줌.

```
plot(lm_model, which = 4)
```

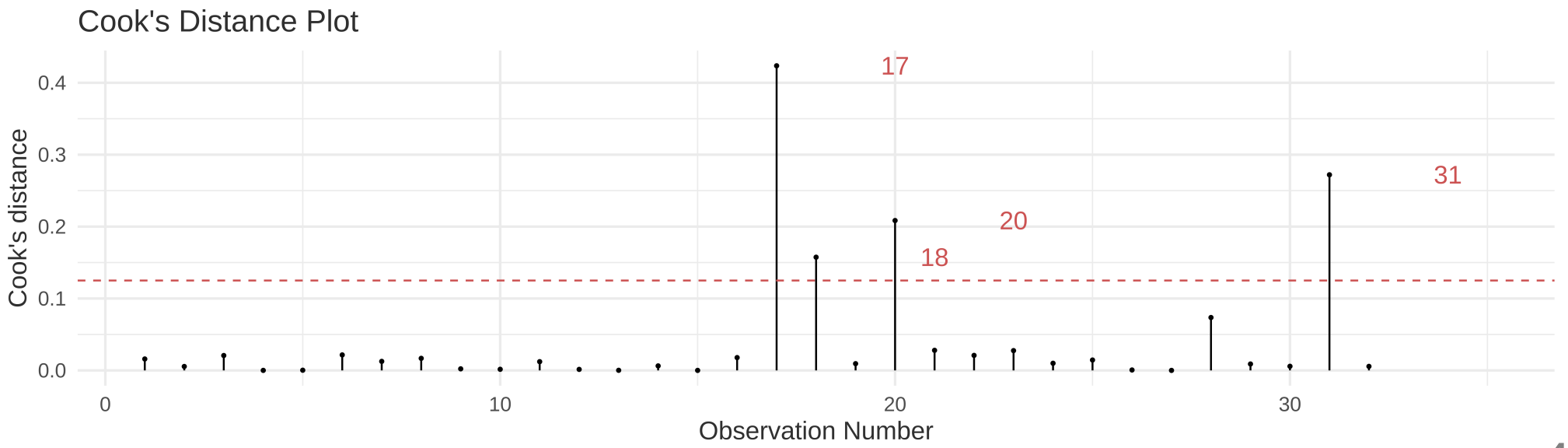


# 진단(Diagnostics)

## 쿡의 거리 플롯(Cook's Distance Plot)

쿡의 거리 플롯은 각 관측치별 Cook’s D 값을 나열한 그래프. 관측치의 인덱스(또는 이름) vs Cook’s D를 그려 어느 관측치의 Cook's D가 큰지 직접 보여줌.

```
lindia::gg_cooksd(lm_model)
```



# 진단(Diagnostics)

## 쿡의 거리 플롯(Cook's Distance Plot)

$X$ 축은 관측치 번호 (또는 이름), $Y$ 축은 Cook's distance 값

- 값이 큰 관측치에는 번호가 표시
- 만약 특정 몇 개의 관측치가 다른 점들에 비해 월등히 높은 Cook’s D를 가진다면, 그 점들은 제거 시 회귀계수들이 상당히 변할 가능성이 있는 영향치

그렇다고 이러한 영향력이 큰 관측치들을 '제외'해야 할까?

# 진단(Diagnostics)

## 계수영향지표 플롯(DFBETAS Plot)

기본 R의 `dfbetas()` 함수: 각 관측치가 회귀계수들에 미치는 영향을 시각화

```
dfbetas(lm_model) |> head()
```

```
##                    (Intercept)             wt            hp
## Mazda RX4         -0.161347204  0.0639304305  0.032966471
## Mazda RX4 Wag     -0.069324050 -0.0004066495  0.045785122
## Datsun 710        -0.211199646  0.0972314374  0.043374926
## Hornet 4 Drive     0.002672687  0.0044886906 -0.006839301
## Hornet Sportabout  0.001784844 -0.0015536931  0.009208434
## Valiant           -0.005985946 -0.1516565139  0.180374447
```

절대값이 큰 DFBETAS (특히 기준인 $2/\sqrt{n}$ 초과)가 있는지 확인하여, 특정 관측치가 어떤 계수에 강하게 영향을 미치는지 파악

# 진단(Diagnostics)

## 계수영향지표 플롯(DFBETAS Plot)

위 결과는 `Intercept`, `wt`, `hp` 세 계수 각각에 대한 DFBETAS 값을 보여줌. 만약 어떤 점이 `wt` 계수의 DFBETAS에서 현저히 높다면, 그 관측치를 제거하면 `wt`의 추정치가 크게 변한다는 의미

- `mtcars` 데이터의 경우, 극단적으로 무거운 차 혹은 고출력 차가 `wt` 또는 `hp` 계수 추정에 큰 영향력을 행사할 수 있음.
- 여러 계수에 동시에 큰 영향을 주는 관측치는 전반적으로 데이터에 특이한 패턴을 가진 경우일 수 있으므로 추가 조사가 필요

# 진단(Diagnostics)

## 다중공선성 진단(Multicollinearity)

다중공선성은 예측변수들 간에 강한 상관관계가 있어 회귀계수의 추정 불안정성을 초래하는 현상

- 공선성이 심하면 계수의 표준오차가 커지고 (신뢰구간이 넓어지고, $p-$값이 커짐), 변수 중요도의 해석이 어려워짐.
    - 가정 위배는 아니지만, 모형의 효율성 저하 문제

# 진단(Diagnostics)

## 다중공선성 진단(Multicollinearity)

분산 팽창 요인(Variance Inflation Factor, VIF): 각 회귀계수의 분산이 공선성 때문에 몇 배로 증가했는지를 나타내는 지표

- 일반적으로 VIF > 10이면 심각한 다중공선성을 의심(VIF > 5 정도를 경고 수준으로 보기도 함).
    - `mtcars` 데이터의 `wt`와 `hp`는 어느 정도 상관관계가 있어 VIF 값이 4~5 정도 될 수 있음. 이는 공선성이 있지만 아직 치명적이지는 않은 수준

R$^2$ 값이 0.9 (VIF≈10)에 달하지 않는 한 추정치의 정확성이 크게 훼손되지는 않는다고 봄. 따라서 VIF가 10 미만인 공선성은 우려를 가지되, 어느 정도 허용 가능

# 진단(Diagnostics)

## 다중공선성 진단(Multicollinearity)

진단 방법:

- `cor()` 함수로 예측변수들 간 상관계수를 탐색
- `{car}` 패키지의 `vif()` 함수를 사용하여 VIF 값을 계산

대처: 공선성이 큰 변수들은 변수 선택이나 차원 축소(PCA), 또는 상호작용 항 추가/제거, 중심화(centering) 등을 통해 완화. 경우에 따라 변수를 결합하거나 표준화하여 해석을 용이하게 만들기도 함.

# 진단(Diagnostics)

## 진단에 따른 처치

비선형성 발견 시: 모델에 다항식 항(예: $X^2$)이나 변환 항(로그, 제곱근 등) 또는 스플라인(spline) 등을 추가하여 비선형 관계를 모형에 반영. 혹은 완전히 다른 모형(예: 나무모형, GAM)을 고려

이분산성 발견 시: 모형 적합에는 문제가 없으나 표준 오차 추정이 잘못되므로, 로버스트 표준오차를 사용해 회귀계수의 표준오차와 $p-$값을 보정. 또한 가법적인 이분산 구조가 의심되면 **가중회귀(Weighted LS)**를 사용하거나, 종속변수 변환(예: Box-Cox 변환)으로 분산 안정화를 시도

정규성 위배 시: 표본이 크다면 크게 문제되지 않을 수 있지만, 심한 경우 정규성을 만족하도록 데이터 변환(예: 응답변수의 로그 변환)이나 이상치 제거를 고려. 또는 비모수적 부트스트랩을 통해 신뢰구간을 추정하는 등 분포가정을 완화한 방법을 사용

# 진단(Diagnostics)

## 진단에 따른 처치

이상치(Outliers) 존재 시: 데이터 오류 여부를 확인하고 필요하면 해당 관측치를 제외하거나 별도로 분석. 하지만 분석 목적상 중요한 값이라면 강건(robust) 회귀 방법으로 이상치에 덜 민감한 모형을 적합. 이상치를 자동으로 제거하기보다, 분석 목적에 맞게 처리(단순 제거는 신중해야 함).

영향치(Influential points) 존재 시: Cook's D 등의 지표가 큰 관측치는 결과 해석 시 특별히 언급하거나, 모형 적합시 해당 관측치를 제거한 대안 모형도 함께 적합해 보는 것이 좋음. 두 모형의 결과 차이를 비교해 영향치의 효과를 평가

다중공선성 심각 시: 공선성이 높은 변수 중 하나를 제거하거나, 둘을 합친 새로운 지표를 사용. 예를 들어 `wt`와 `hp` 둘 다 차량의 크기와 성능을 나타낸다면 이들의 지수 결합이나 PCA로 만든 하나의 요인으로 대체.

# Part II. 모형의 선택

# 진단(Diagnostics)

## 모델 선택의 필요성

하나의 **정답**인 모형은 **존재하지 않으며**, 고려할 수 있는 설명 변수가 매우 많음. 따라서 적절한 변수 선별과 모형 선택이 필요

- 변수를 너무 많이 포함하면 훈련 데이터에 과도하게 맞춰져 새로운 데이터에 성능이 떨어지는 과적합 위험이 있고, 너무 적게 포함하면 중요한 관계를 놓치는 과소적합 문제가 발생할 수 있음.
- 모형 선택을 통해 모형의 복잡도(변수 개수)를 관리하여 예측 정확도와 모형의 간명성(단순성) 사이의 균형을 맞춰야 함.

# 진단(Diagnostics)

## 과적합 vs. 과소적합

과적합(Overfitting)

- 모형이 지나치게 복잡하여 훈련 데이터의 노이즈까지 설명하는 상태
- 이 경우 훈련 데이터 적합도는 높지만 새로운 데이터에 대한 예측 성능은 저하

과소적합(Underfitting)

- 모형이 너무 단순하여 데이터의 중요한 패턴을 놓치는 상태
- 이 경우 훈련 데이터에 대해서도 오차가 크며, 더 복잡한 모형이 필요

# 진단(Diagnostics)

## 모델 선택의 기준: 선형회귀 수준에서

결정계수 $R^2$ (설명력 지표): 선형 회귀모형의 적합도

- 종속변수 변동 중에서 모형이 설명하는 비율을 나타냄.

$$R^2 = 1 - \frac{SSE}{SST}$$

- $R^2$ 값은 0부터 1까지이며, 1에 가까울수록 모형이 데이터를 잘 설명함을 의미
- $R^2$는 변수를 추가하면 감소하지 않고 늘어나기만 하기 때문에, 변수 개수가 다른 모형 간 직접 비교 지표로 부적절

# 진단(Diagnostics)

## 모델 선택의 기준: 선형회귀 수준에서

조정된 $R^2$ (Adjusted $R^2$, $R^2_{\text{adj.}}$): 설명 변수 개수에 따른 $R^2$ 보정

- 설명변수의 개수에 따라 $R^2$를 보정하여 모형 복잡도에 대한 페널티를 부여한 지표
    - 여기서 $n$은 표본 크기, $k$는 예측 변수 개수

$R^2_{\text{adj.}}$는 불필요한 변수를 추가하면 오히려 감소할 수 있으므로, 최고의 조정 $R^2$을 주는 모형을 선택함으로써 과적합을 방지할 수 있음(값이 클수록 좋은 모형).

# 진단(Diagnostics)

## 모델을 조정하는 선택 전략들

전진 선택법(Forward Selection): 변수를 하나씩 추가하는 방법

후진 제거법(Backward Elimination): 변수를 하나씩 제거하는 방법

단계적 선택법(Stepwise Selection): 전진/후진 결합한 혼합 접근

릿지 회귀(Ridge Regression): 계수의 크기를 줄이는 정규화 기법

라쏘 회귀(Lasso Regression): 계수를 0으로 만들어 변수 선택까지 수행하는 정규화 기법

# 진단(Diagnostics)

## 모델을 조정하는 선택 전략들

### 전진 선택법

전진 선택은 시작 단계에서 모든 변수가 제외된 상태에서 출발

- 가장 유의미한 변수를 하나씩 추가하여 모형을 확대
- 추가할 변수는 보통 **유의확률(p-value)**이 가장 낮거나 AIC 등의 개선을 가장 크게 하는 변수를 선택

더 이상 추가할 변수가 없거나 기준에 부합하는 변수가 없으면 알고리즘을 종료

- 초기에는 상수항만 있는 모형에서 시작하여, 가장 설명력이 높은 변수를 첫 번째로 추가하는 식으로 진행

# 진단(Diagnostics)

## 모델을 조정하는 선택 전략들

### 후진 제거법

후진 제거는 모든 후보 변수를 포함한 최대 모형에서 시작

- 기여도가 낮은 변수를 하나씩 제거하여 모형을 축소
- 제거할 변수는 보통 유의확률이 가장 높은 변수(혹은 기준에 따라 기여 미미한 변수)를 선택

더 이상 제거할 변수가 없거나 제거 기준에 맞는 변수가 없으면 종료

- 전진/후진 방법은 탐욕적(greedy) 방법이므로, 최적의 변수 조합을 보장하지는 않지만 계산 효율이 높음.

# 진단(Diagnostics)

## 모델을 조정하는 선택 전략들

### 단계적 선택법

단계적 선택은 전진법과 후진법을 절충한 알고리즘

- 변수를 추가할 때마다 불필요해진 변수가 있으면 제거하는 단계를 병행
- 우선 전진 선택으로 변수 추가를 시도하고, 각 단계마다 모형 내 기존 변수들의 유의성 검토를 통해 유의미하지 않은 변수는 제거
- 이러한 추가/제거를 교차적으로 수행하여 최종 모형을 선정

일반적으로 전진/후진보다 더 안정적인 변수집합을 찾는 경향이 있지만, 여전히 단순 탐욕 알고리즘의 한계에서 벗어나지 못함.

# 진단(Diagnostics)

## 선형 모델에서의 선택 비교

평가 지표: 결정계수 $R^2$ 및 조정 $R^2$를 주로 사용

모형 비교: F-검정 등을 통해 유의미한 변수 추가 여부를 검정하거나, $R^2_{\text{adj.}}$ 최대화 모형을 선택

잔차의 분산으로 모형 성능 평가(RMSE 감소, $R^2$ 증가). $R^2$는 변수 추가 시 단조증가하므로, 과적합 방지를 위해 $R^2_{\text{adj.}}$나 AIC/BIC로 복잡도 페널티를 고려

# 진단(Diagnostics)

## 모델 선택 시 유의할 점

단일 최적 모형은 존재하지 않음: 분석 목적에 따라 다른 모형이 더 유용할 수 있으며, 여러 기준을 종합적으로 고려해야 함.

- 통계 지표뿐 아니라 변수의 이론적 중요성을 고려해야 함. 자동적으로 변수를 선택해도 도메인 지식 검증이 필요

과적합 검사: 선정된 모형의 성능은 교차검증 등으로 검증하여, 데이터에 과하게 특화되지 않았는지 확인

결과 민감도: 변수 선택에 따라 결과 및 해석이 달라질 수 있으므로, 핵심 결론이 특정 변수 조합에 의존하지 않는지 확인하는 것이 좋음.

KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

# 감사합니다!

## 궁금한 것이 있으면 언제든 연락하세요.

강사 연락처

| 연락처 | 박상훈 |
| --- | --- |
| | sh.park.poli@gmail.com |
| | sanghoon-park.com/ |
| | 영상바이오관 405 |