5장 Multicollinearity and Diagnostics

CONTENTS

5.1 서론

- 이 장에서는 다중선형회귀의 적합 과정에서 다중공선성 문제에 대한 처리와 잔차분석을 비롯한 회귀진단의 문제를 다룬다.
- 다중공선성은 예측변수들 간에 상관성이 높은 현상을 말하는데, 이 경우에 다중선형회귀에서는 문제가 될 수 있다. 이 장에서는 다중공선성의 문제를 살펴보고 이에 대한 해결책을 제시한다.
- 회귀진단은 적합된 다중회귀모형에 대한 진단을 수행한다. 다중선형회귀를 수행한 뒤 잔차분석을 통해 적합모형에 대한 타당성을 체크하고, 문제점에 대한 해결(개선) 방안을 제시한다.

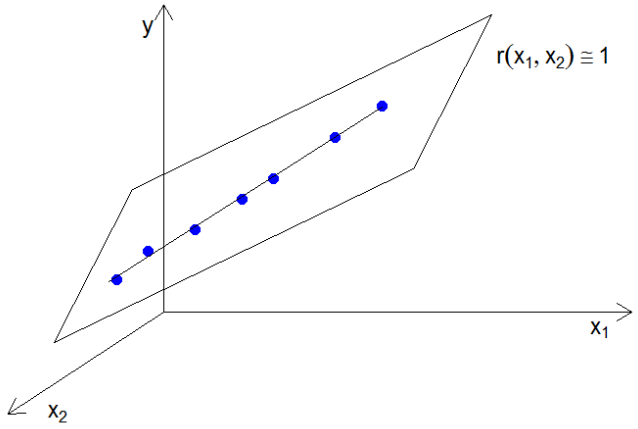
5.2 다중공선성과 해결방법

5.2.1 다중공선성 문제

- 다중선형회귀에서는 자료의 수가 모수의 수 보다 많아야 하며($n \geq p+1$), 회귀계수의 추정에 $(X'X)^{-1}$의 계산이 요구된다
- (두 개 또는 그 이상의) 예측변수들 간에 상관성이 높은 현상을 다중공선성(multicollinearity)이라고 하는데, 완전한 다중공선성이(독립변수들 간에 정확한 직선관계가) 존재하는 경우에는 역행렬 $(X'X)^{-1}$의 계산이 불가능해진다.
- 따라서 강한 상관은 계산의 불안정성을 초래하며, OLS 추정치가 더 이상 BLUE(Best Linear Unbiased Estimator)가 되지 않는다.

5.2 다중공선성과 해결방법

- 다음의 [그림 5.1] 에서와 같이 예측변수들 간에 높은 상관이 존재하는 경우를 생각해 보자.



[그림 5.1] 다중공선성이 존재하는 경우

- 이 경우에는 초평면이 아닌 직선식으로도 y 값에 추정이 가능해지며, 이를 초평면으로 추정하게 되면 자료의 작은 변화에도 추정결과가(추정된 초평면이) 크게 달라질 수 있다(불안정한 추정이 이루어짐).

5.2 다중공선성과 해결방법

- 다중공선성이 발생되는 원인으로는 1.(데이터 수집) 데이터가 매우 좁은 독립변수의 영역에서 수집되었거나 2.(물리적 제약) 많은 제조 또는 서비스 공정에서 독립변수들은 물리적, 정치적, 법적으로 범위에 제한을 받거나 3.(과-정의된 모형, over-defined model) 변수의 수가 자료의 수보다 많거나 4.(모형 선택 또는 지정) 원 변수의 멱(power) 또는 교호(interaction) 변수를 예측변수로 사용하거나, 5.(이상치) 예측변수-공간에서의 이상치 등이 있다.

5.2 다중공선성과 해결방법

- 다중회귀에서 다중공선성이 존재할 때의 징후는 다음과 같다.
  - 반응변수와의 상관이 높을 것으로 (이론적으로) 생각되는 변수임에도 불구하고 회귀계수가 유의하지 않을 때
  - X 변수를 추가하거나 제외했을 때, 회귀계수의 변화가 심하게 일어날 때
  - X 에 따라 반응변수가 증가해야 함에도 회귀계수가 음의 값을 가질 때
  - X 의 증가에 따라 반응변수가 감소해야 하는 상황에서 양의 회귀계수를 가질 때
  - X 변수들 간에 높은 상관계수가 존재할 때 등이다.

5.2 다중공선성과 해결방법

- 다중공선성에 대한 측도로는 분산팽창요인(variance inflation factor, 이하 VIF)과 조건지수(condition number, k)가 대표적이다. VIF와 k의 정의는 다음과 같다.

$$VIF_j = \frac{1}{1-R_j^2} \quad \& \quad k = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

- 위의 식에서 $R_j{}^2$는 예측변수(x_j)를 나머지 $p-1$개의 예측변수로 회귀를 수행하여 구해지는 결정계수를 나타내며, λ_{max}와 λ_{min}는 $X'X$ 행렬의 고유치의 최댓값과 최솟값을 나타낸다.

- VIF는 변수가 상관 될 때 분산의 추정된 회귀계수의 분산이 얼마나 커지는지를 나타내는 값으로 이 값이 매우 크거나 (5 또는 10 이상), 독립변수들의 상관행렬의 고유치가 0에 가까운 값을 가질 때 즉, 조건지수가 15(또는 30)보다 크면 다중공선성이 존재하는 것으로 생각할 수 있다.

5.2 다중공선성과 해결방법

- R의 vif{car} 함수는 선형모형과 일반화 선형모형에서 (일반화) 분산팽창요인(VIF) 값을 제공한다.

```
vif(mod, …)
```

  - mod는 lm 또는 glm 객체임

5.2 다중공선성과 해결방법

5.2.2 해결방법

- 이 절에서는 다중공선성과 관련된 몇 가지 질문과 함께, 이에 대한 올바른 이해와 해결방법을 제시한다.
- 회귀모형의 적합 시에 다중공선성은 문제가 된다. 다중공선성은 예측변수가 다른 예측변수들과 상관이 된 경우를 말한다.

5.2 다중공선성과 해결방법

(a) 다중공선성이 왜 문제가 되는가?

- 적당한 다중공선성은 문제가 되지 않는다. 그러나, 심각한 다중공선성은 문제가 된다
- 그 이유는 계수 추정치의 분산을 증가시킬 수 있고, 모형의 작은 변화에도 추정치가 매우 민감하게 반응한다. 그 결과로 추정치가 불안정하고 해석이 어렵다.
- 다중공선성은 통계적 분석력을 약화시키고, 계수 추정치의 부호를 변화시킬 수 있으며, 이로 인해 올바른 모형 선택을 어렵게 한다.

5.2 다중공선성과 해결방법

(b) 다중공선성을 제거해야만 하는가?

- 다중공선성의 증상은 매우 심각하게 들리지만, 치료의 필요성은 분석의 목적에 따라 다르다. 요약하면, 다중공선성은
  - 모형에 포함될 올바른 예측변수의 선택을 보다 어렵게 할 수 있으며
  - 각 예측변수의 정확한 효과를 파악하는데 장애가 된다. 그러나,
  - 모형의 전체적인 적합에는 영향을 미치지 않으며, 나쁜 예측을 제공하지도 않는다.
- 분석의 목적에 따라 다중공선성은 항상 문제가 되는 것은 아니다. 그러나, 심각한 다중공선성이 존재할 때 올바른 모형 선택에 어려움이 있으므로 항상 관심을 가질 필요가 있다.

5.2 다중공선성과 해결방법

(c) 다중공선성에 대한 올바른 생각

- 다중공선성은 여러 가지 문제를 야기한다. 여러 개의 예측변수를 가지는 모형에서 다중공선성으로 인해 추정치의 부호가 바뀌거나 부정확한 p –값이 제공되면 올바른 모형 선택이 어렵게 된다(이 경우 단계별선택법 등의 변수선택법도 제대로 작동하지 않는다).
- 그러나, 다중공선성이 모형이 얼마나 잘 적합 되는 지에는 영향을 미치지 않는다. 만약 모형이 잔차의 가정을 잘 만족시키고, 다중 상관계수(R^2)가 충분히 크다면, 심각한 다중공선성을 가지는 경우에도 훌륭한 예측을 제공할 수 있다.
- 또한 높은 상관을 가지는 모든 한 쌍의 예측변수들에 대해서는 걱정할 필요가 없다. 예를 들어, 두 예측변수 간의 상관계수가 0.85인 경우(매우 높음)에도 VIF의 값은 3,6 정도에 불과하다. 따라서 상관된 예측변수를 포함하는데 두려워 할 필요가 없다(-VIF를 체크하는 것은 명심할 것).

5.2 다중공선성과 해결방법

(d) 다중공선성에 대한 해결책

- 다중선형회귀에서 다중공선성의 문제를 해결하는 방법에는 다음과 같은 방법들이 있다.

- **상관된 예측변수를 제거**: 변수선택법, 최적부분회귀 등의 방법으로 상관성이 높은 변수를 제거해 나가되, R^2가 높은 모형을 선택한다.
- **벌점회귀**(penalized regression)**를 적용** : 능형(ridge) 회귀, 라소(lasso) 회귀, 주성분회귀(PCR) 또는 부분최소제곱회귀(PLSR)를 수행한다.
- **통계적 학습 회귀**(statistical learning regression)**의 적용** : 회귀나무, 베깅회귀, 랜덤포리스트, 신경망, 지지벡터회귀를 수행한다.

5.2 다중공선성과 해결방법

- 원 변수로부터의 파생변수(제곱항 또는 교호효과를 모형에 추가할 경우 등)에 기인하는 구조적인 다중공선성 문제는 위에서 제시한 여러 가지 해결책들을 고려하기 전에 변수의 중심화(또는 표준화)를 수행하는 것이 쉬운 해결책이 될 수 있다.
- 어떠한 해결책을 고려하더라도, 모든 치료법은 잠재적으로 약점을 가진다는 점을 명심할 필요가 있다. 만약 덜 정확한 계수추정치 또는 덜 유의하나 높은 R^2를 가지는 모형에 만족한다면, 아무런 조치를 하지 않는 것이 올바른 결정이 될 수도 있다(적합에는 영향을 미치지 않으므로).

5.3* 회귀진단

- 회귀진단(regression diagnostics)은 다중회귀에서 적합된 회귀모형이 타당한지에 대한 다음의 진단을 수행한다. 이 가운데 다중공선성의 진단에 대해서는 앞 절에서 다룬 바 있다.
  - 이상치와 영향점 진단
  - 오차에 대한 진단(정규성, 등분산성, 독립성)
  - 다중공선성 진단

5.3* 회귀진단

- 회귀진단에 등장하는 주요 용어를 소개(정의)하면 다음과 같다.

- **잔차**(residual): (추정된 회귀식에 기초한) 예측값과 실제 관측값과의 차이를 말한다.
- **이상치**(outlier): (다른 관측값들이 모형을 잘 따르는 반면) 모형을 잘 따르지 않는 관측값($y, x_1, \ldots, x_p$)을 말한다.
  - **예측변수에서 이상치**: 다른 예측변수들의 모임으로부터 벗어난 예측변수값을 말한다. 이 값은 모형을 부적절하게 적합할 수 있다. h_{ii}는 i-번째 예측변수값이 얼마나 벗어나 있는가를 나타낸다. h_{ii}는 hat 행렬(H)의 i-번째 대각원소로 다음과 같이 정의되며, 이 값을 흔히 i-번째 자료의 지레(leverage)라고 한다.

$$leverage_i = h_{ii} = (X(X'X)^{-1}X')_{ii}$$

  - **반응변수에서 이상치**: 적합모형으로부터 멀리 떨어진 반응변수값을 말한다. 표준화 잔차의 값이 클수록 이상치에 가깝다.

5.3* 회귀진단

- **지레점**(leverage point): (예측변수의) 평균으로부터 멀리 떨어진 예측변수에서 측정된 관측값을 말한다. 이 값은 회귀계수의 추정에 큰 영향을 미칠 수 있다.
- **영향점**(influential point): 회귀식에 크게 영향을 미치는 관측값을 말하며, 이 값이 제거될 경우 회귀계수의 추정치가 크게 변하게 된다. 영향점은 지렛점과 이상치의 산물로 생각될 수 있다.
- **쿡의 거리**(Cook's D): i – 번째 관측값이 제거될 때 회귀함수가 얼마나 크게 변하는 가를 측정하는 것으로, 관측치의 지레(leverage)와 잔차(residual)의 정보를 결합한 측도이다.
- R의 {stats} 패키지에서는 제공하는 회귀진단 함수는 [표 5.1]과 같다(자세한 내용은 관련 서적을 참고할 것). 또한, R의 {car} 패키지는 회귀진단을 포함한 회귀모형에서 유용한 다양한 고급기법을 제공한다.

5.3* 회귀진단

[표 5.1] R의 회귀진단 기본 함수({stats} 패키지 제공)

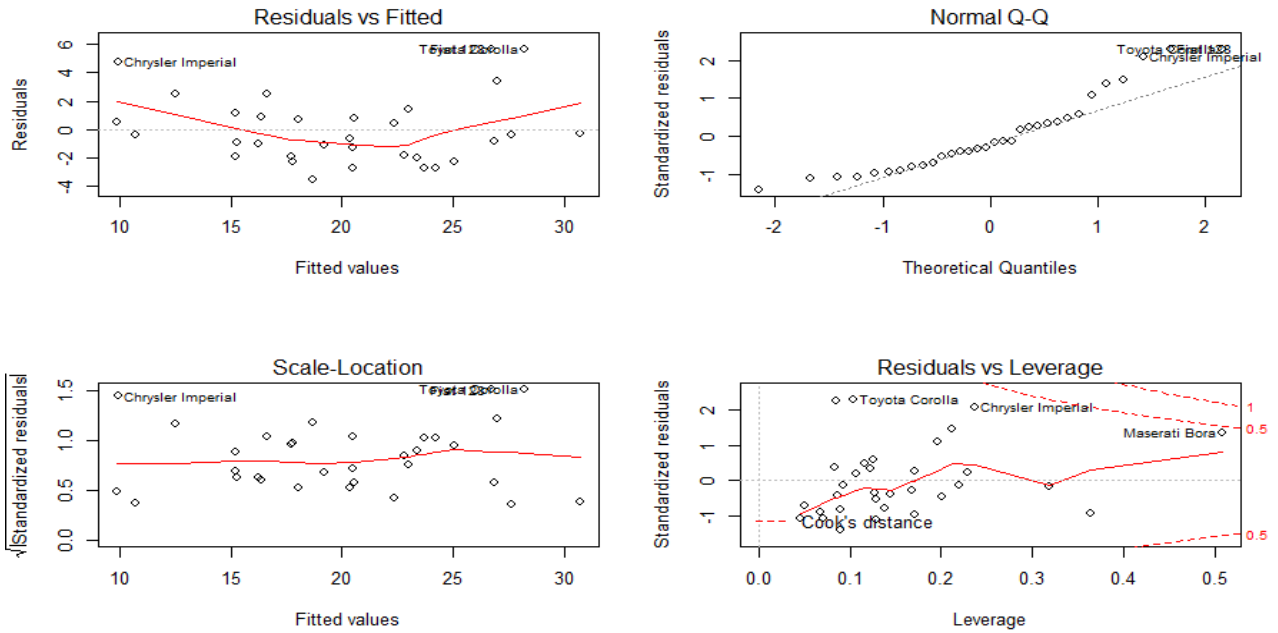
R 함수	기능	기준(또는 정의)
resid()	원 잔차	·
fitted()	적합값	·
rstandard()	스튜던트화 잔차	$r_i = e_i / s(e_i) = e_i / \hat{\sigma}\sqrt{1-h_{ii}}$
rstudent()	externally 스튜던트화 잔차(aka detected t 잔차): 이상치 진단(검정)	$t_i = e_i / \hat{\sigma}_{(i)}\sqrt{1-h_{ii}} \sim t_{n-p-2}$
dffits()	영향점 진단	$DFFITS_i > 3\sqrt{\frac{p+1}{n-p-1}}$
dfbetas()	영향점 진단	$DFBETAS_{k(i)} > 1$
cooks.distance()	영향점 진단	$D_i >$ median of $F_{p+1,n-p-1}$
hatvalues()	영향점 진단	$h_i > \frac{3(p+1)}{n}$
covratio()	영향점 진단	$\lvert 1 - COVR_i \rvert > \frac{3(p+1)}{n-p-1}$, $COVR_i = \left(\frac{\hat{\sigma}^2_{(i)}}{\hat{\sigma}^2}\right)^{p+1} \frac{1}{1-h_i}$
influence.measures()	영향점 진단 측도 5개	

5.3* 회귀진단

예제 1 mtcars 자료를 이용하여 회귀진단을 수행한다. 여기서는 진단의 방법만을 제시하고, 진단 결과에 의한 모형의 개선은 실시하지 않는다.

```
> data(mtcars)
> fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
> ## 표준 진단 그림: 종합적인 진단 결과 제공
> par(mfrow=c(2,2))
> plot(fit); par(mfrow=c(1,1))
```

5.3* 회귀진단



해 석

(좌상)은 적합 값에 대한 잔차 그림(residual plot)으로, 영(0)을 중심으로 골고루 퍼져있어 대 체로 무난하나 약간의 이차 곡선의 경향이 보인다. **(우상)**은 잔차에 대한 정규확률그림으로 기울 기가 1인 직선을 벗어나는 점들이 많으며, 따라서 정규성 가정을 잘 따르지 않는다(즉, 모형의 개선이 요구됨). **(좌하)**는 적합값에 따른 표준화 잔차(standardized residuals) 그림으로 y의 제곱값(표준화 잔차)이 표준정규분포의 가정을 크게 벗어나지 않으나, 라벨이 붙은 몇 개의 점은 이상치 여부를 살펴볼 필요가 있다. **(우하)**는 지렛값(leverage)에 대한 표준화 잔차의 그림으로, 라벨이 붙은 점들에 대해서는 영향점 또는 이상치 여부에 대한 검토가 필요하다. 아울러 그림4 는 영향점 진단을 위한 쿡의 거리도 등고선으로 제시하고 있다.

5.3* 회귀진단

- 위의 표준 진단 그림을 {car} 패키지를 이용하여 보다 자세히 살펴보면 다음의 [예제 2] ~ [예제 4]와 같다.
- R의 {car} 패키지는 Fox와 Weisberg(2011)의 "An R Companion to Applied Regression"에 소개된 다양한 회귀분석 관련 함수를 지원한다.

5.3* 회귀진단

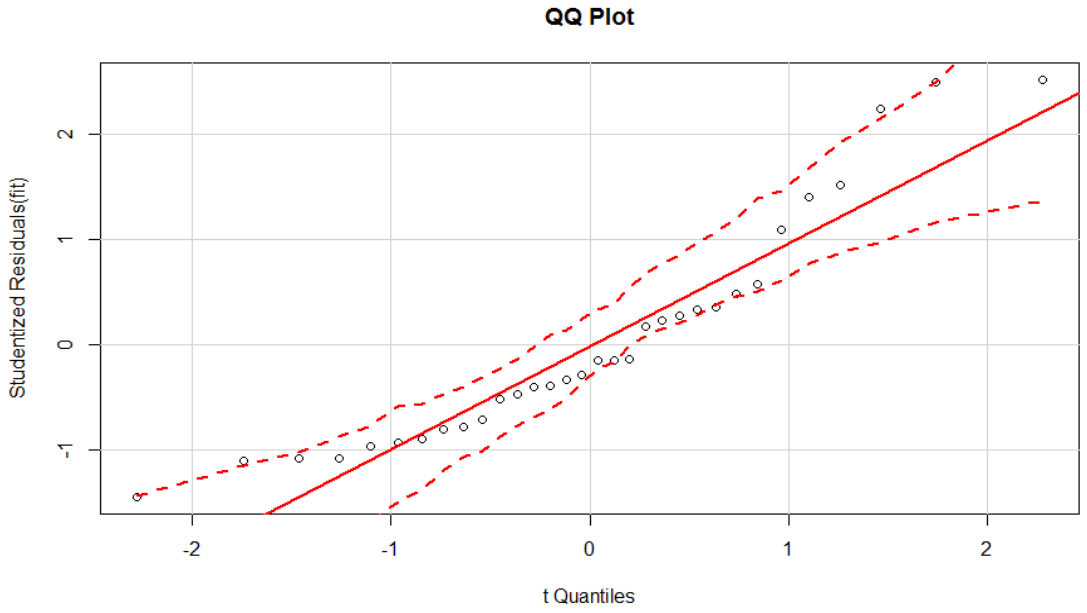
예제 2 ([예제1]의 계속) 이상치와 영향점 진단: {car} 패키지 이용

```
> ## 이상치 진단
> # 이상치 검정
> library(car)
> outlierTest(fit)     # 가장 극단치에 대한 Bonferonni p-검정

No Studentized residuals with Bonferonni p  < 0.05
Largest |rstudent|:
               rstudent unadjusted p-value  Bonferonni p
Toyota Corolla  2.51597            0.01838       0.58816
```

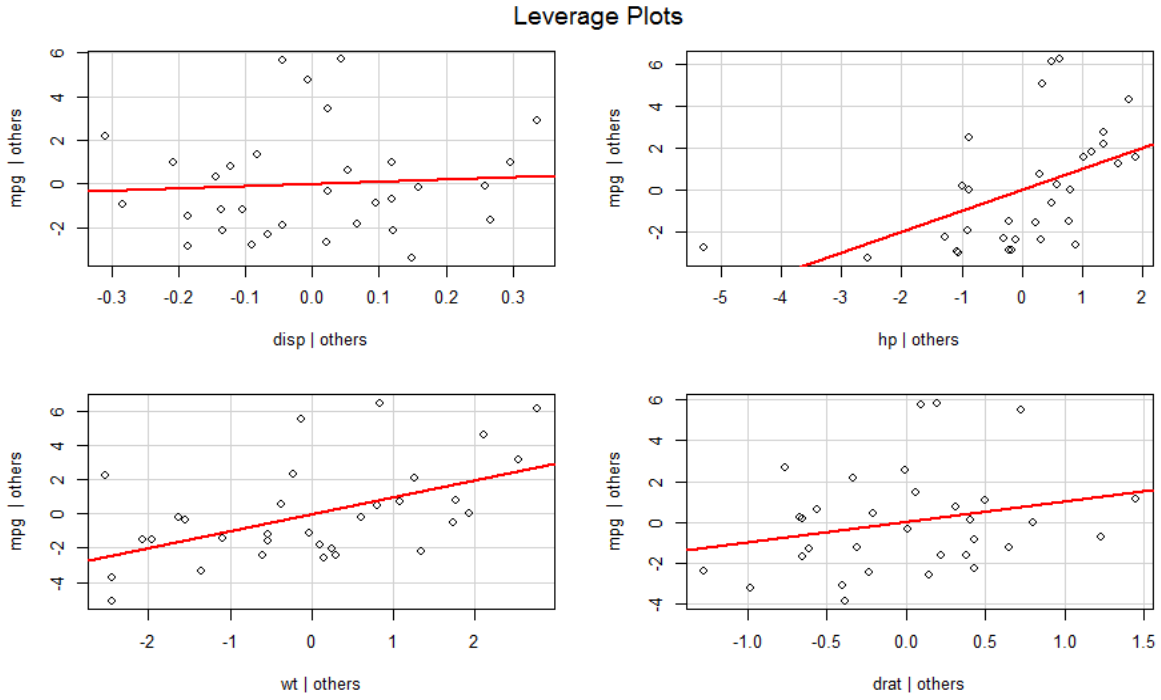
5.3* 회귀진단

```
> # 스튜던트화 잔차에 대한 qq-플롯
> qqPlot(fit, main="QQ Plot")
> # 적절한 t-분포에 대해 그려지며, 신뢰대는 (각 점별로) 붓스트랩 방법으로 구해짐
```



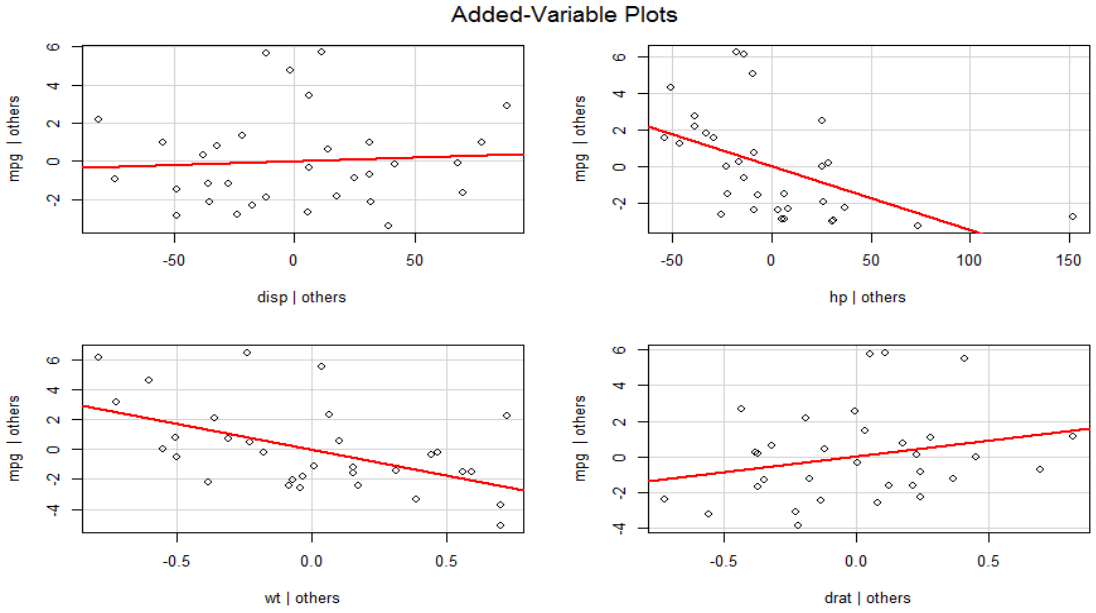
5.3* 회귀진단

```
>  # leverage 플롯
> leveragePlots(fit)
```



5.3* 회귀진단

```
> ##  영향점 진단
> #  추가된 변수(added variable) 그림
> #  특정 변수의 비선형성과 진단과 함께 영향점과 이상치를 찾는데 유용
> avPlots(fit)
```



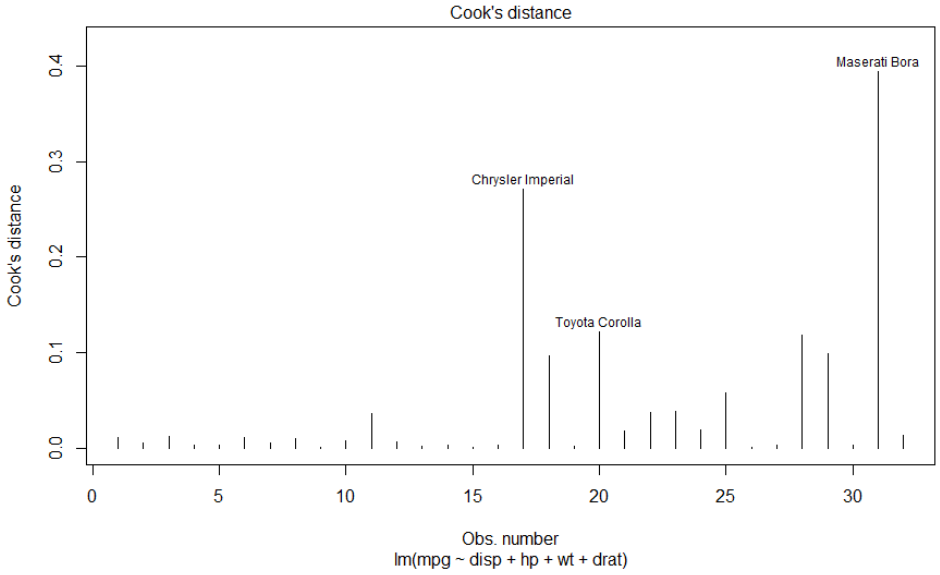
5.3* 회귀진단

해 석

(disp|others, mpg|others) 그림은 예측변수 disp와 반응변수 mpg를 각각 (disp를 제외한) 나머지 예측변수들로 회귀(부분 회귀, partial regression)를 수행한 후 잔차들 간의 관계를 그린 것이다. 즉, 나머지 변수들의 영향력을 제거한 후 잔차들 간의 관계를 그린 것으로, 두 변수 disp와 mpg 간의 관계가 직선적이면 이 그림은 직선으로 적합된다. 따라서 직선식으로부터 벗어난 자료 값들은 영향점(또는 이상치)의 가능성이 높다고 할 수 있다.

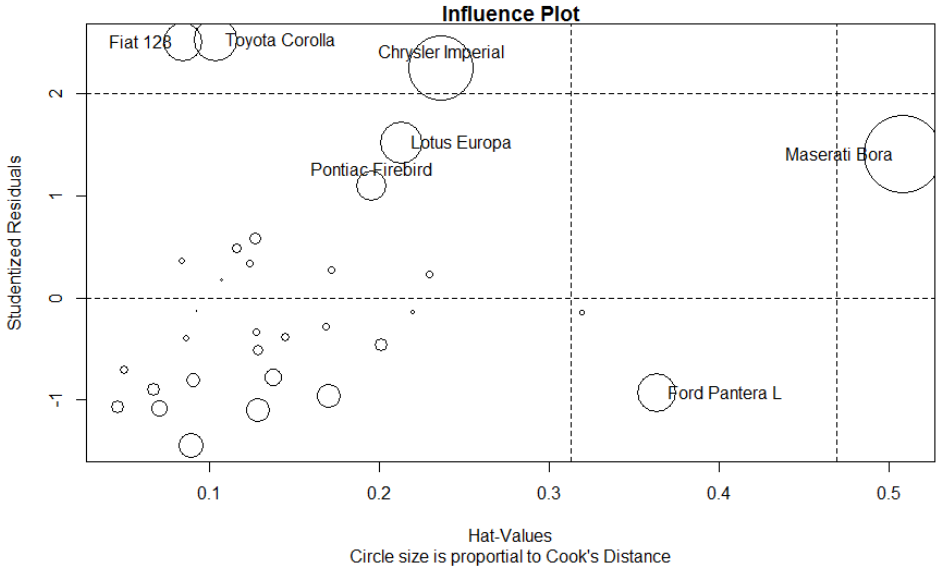
5.3* 회귀진단

```
> # Cook의 거리(D) 그림
> # 기준값인 "4/(n-k-1)" 보다 큰 D를 식별
> cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
> plot(fit, which=4, cook.levels=cutoff)
```



5.3* 회귀진단

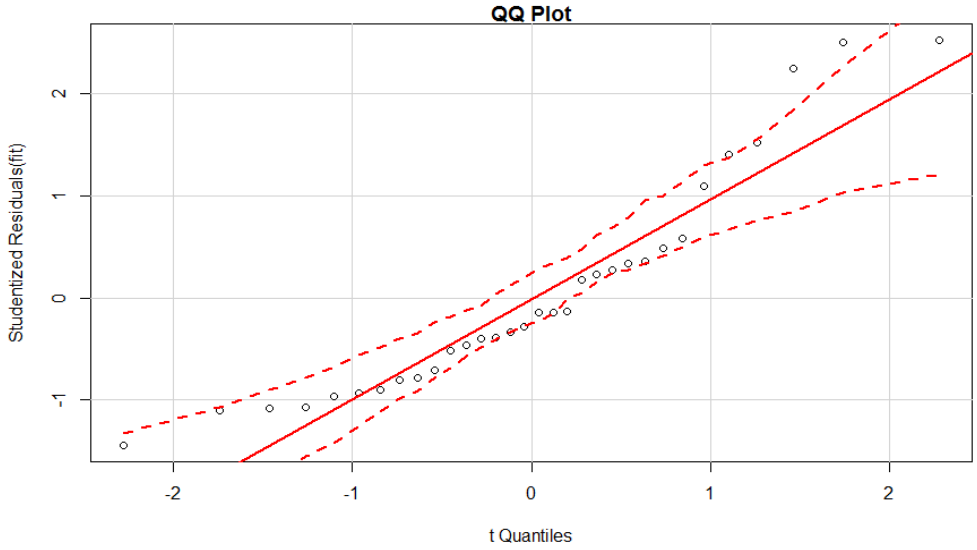
```
> # 회귀 영향(regression influence) 그림
> influencePlot(fit, id.method="identify", main="Influence Plot",
                sub="Circle size is proportial to Cook's Distance" )
> # 실행 후 마우스 대기 상태: 원하는 지점 클릭
```



5.3* 회귀진단

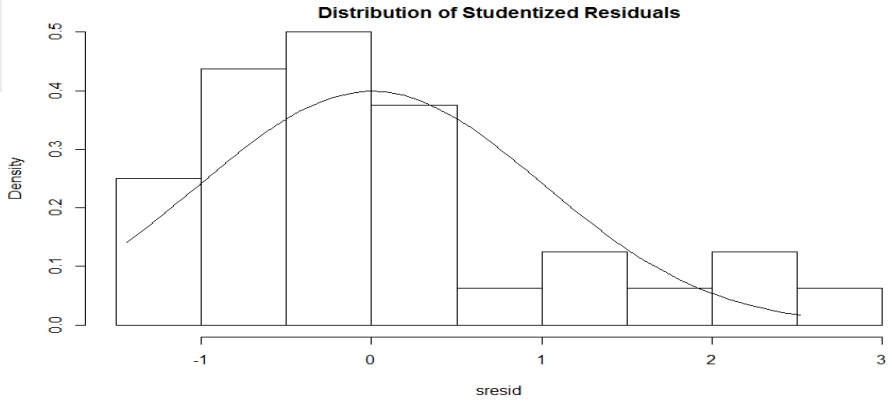
예제 3 ([예제 2]의 계속) 오차에 대한 가정(정규성, 등분산성, 독립성) 진단

```
> ##  잔차의 정규성 검토
> #  스튜던트화 잔차에 대한 qq-플롯
> qqPlot(fit, main="QQ Plot")
```



5.3* 회귀진단

```
> # 스튜던트화 잔차의 분포(히스토그램)
> library(MASS)
> sresid <- studres(fit)
> hist(sresid, freq=FALSE,
      main="Distribution of Studentized Residuals")
> xfit<-seq(min(sresid),max(sresid),length=40)
> yfit<-dnorm(xfit)
> lines(xfit, yfit)
```



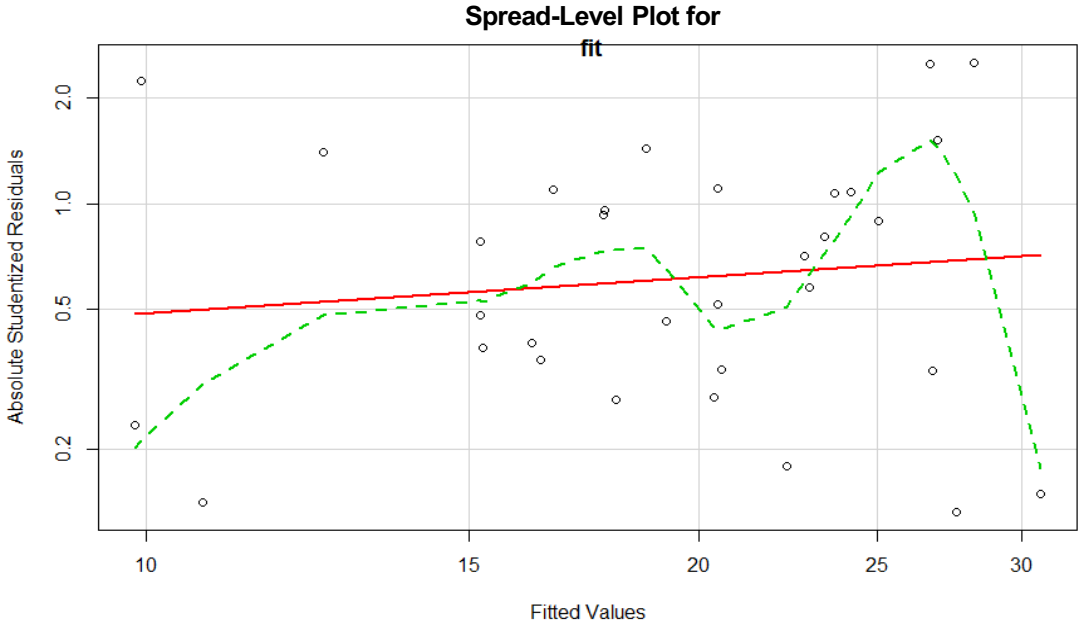
5.3* 회귀진단

```
> ## 등분산성 검토
> # ncvTest(): “non-constant error variance test”의 약어
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.429672    Df = 1     p = 0.231818
```

```
> # 스튜던트화 잔차 vs 적합값 그림
> spreadLevelPlot(fit)

Suggested power transformation:  0.6616338
```

5.3* 회귀진단


Spread-Level Plot for
fit
Absolute Studentized Residuals
0.2
0.5
1.0
2.0
10
15
20
25
30
Fitted Values