회귀진단

회귀진단

- 회귀진단
  1) 회귀모형에 대한 진단
  2) 관찰값에 대한 진단

- 회귀모형에 대한 진단
  - 회귀모형의 가정 사항 만족 여부 확인
  - 적합 및 추론 결과의 신빙성 확보

- 관찰값에 대한 진단
  - 개별 관찰값이 모형 추정 과정에 미치는 영향력 파악

1. 회귀모형의 가정 만족 여부 확인

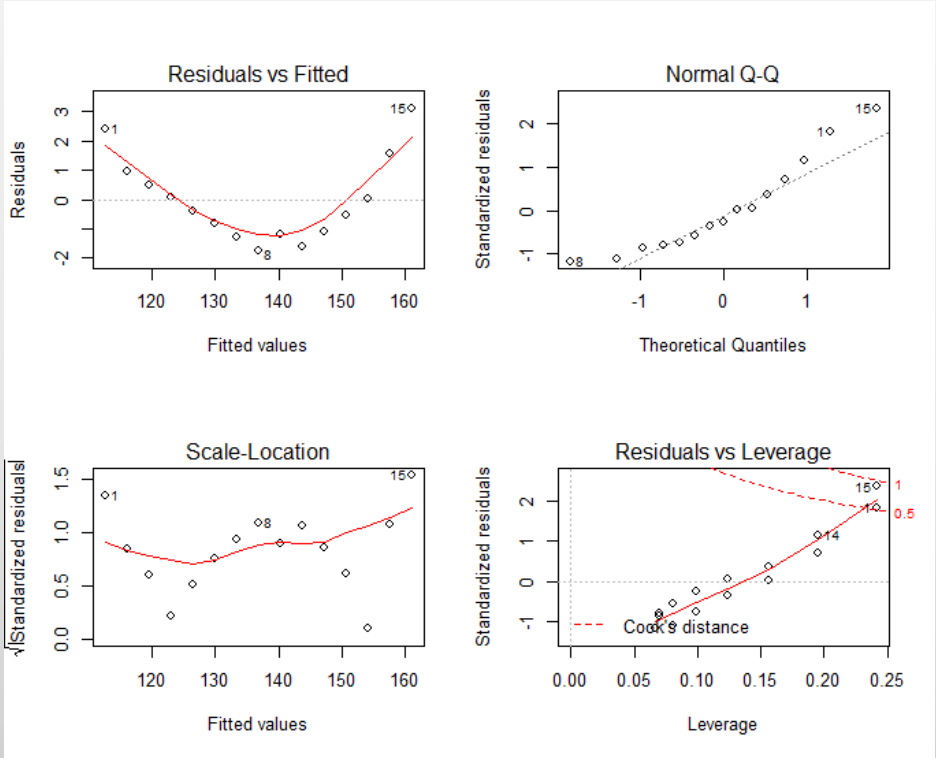
- 다중회귀모형 가정 사항

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i,\ i = 1, \ldots, n$$

1) 오차항 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$의 평균은 0, 분산은 모두 동일
2) 오차항 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$의 분포는 정규분포
3) 오차항 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$은 서로 독립
4) 반응변수와 설명변수의 관계는 선형

- 예: women의 변수 weight와 height의 회귀모형 가정 만족 여부 확인
  - 가장 기본적이 방법은 lm 객체를 함수 plot( )에 적용

```
> fit_w <- lm(weight ~ height,women)
> par(mfrow=c(2,2))
> plot(fit_w)
> par(mfrow=c(1,1))
```

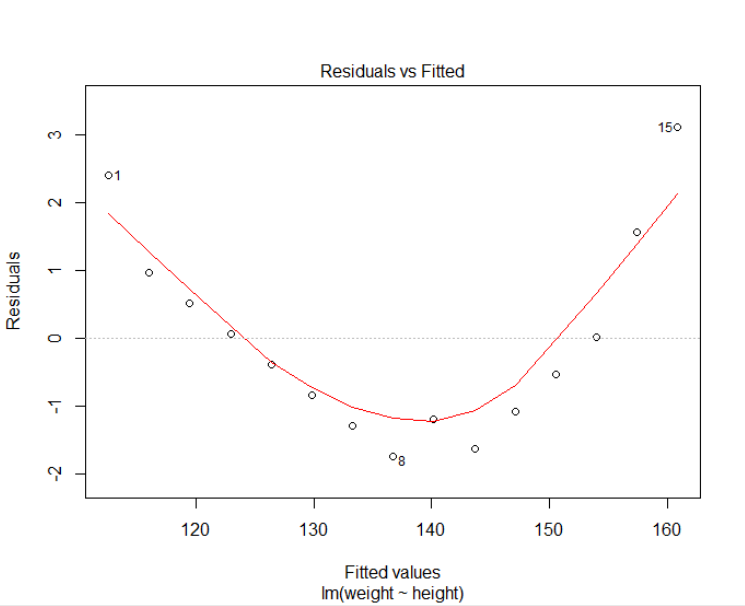


plot(lm 객체):

여섯 개 그래프가 작성되지만 디폴트로 네 개의 그래프가 출력됨

- 첫 번째 그래프

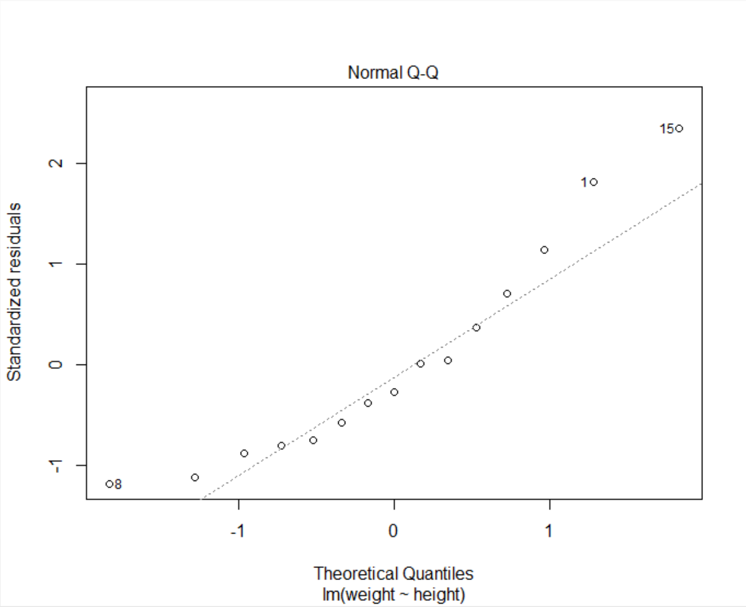
```
> plot(fit_w, which=1)
```



- 가장 일반적인 잔차 산점도
- 잔차($e_i = Y_i - \hat{Y}_i$)와 $\hat{Y}_i$의 산점도
- 동일 분산, 평균 0, 선형관계 확인
- 디폴트로 가장 극단적인 3 case의 case number가 해당 점 옆에 추가

- 두 번째 그래프

```
> plot(fit_w, which=2)
```



- 잔차의 정규성 확인
- 잔차의 정규 QQ plot:
  표준화 잔차의 표본 분위수와
  정규분포의 이론 분위수의 산점도

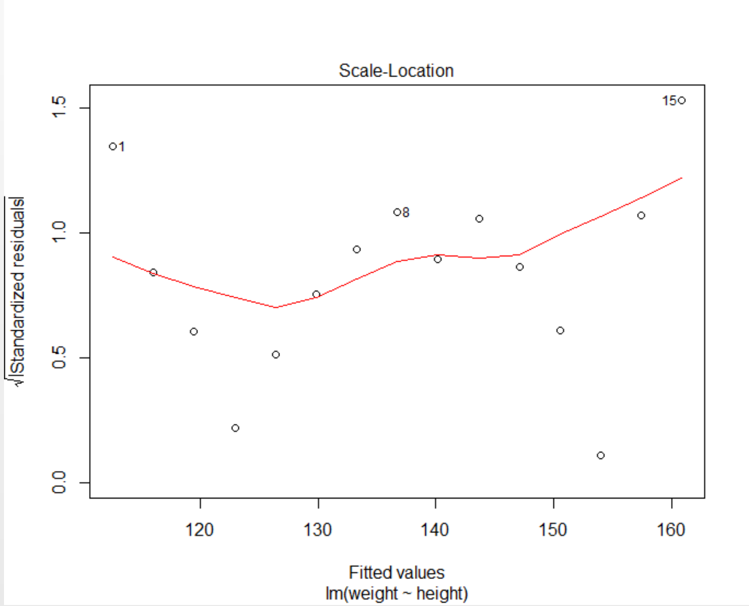
표준화 잔차(r_i)
- 잔차의 분산: $\sigma^2(1-h_i)$
- 오차의 분산과 다름
- $r_i = e_i/\sqrt{MSE(1-h_i)}$

레버리지(leverage; h_i)
- Hat matrix의 대각 원소
- i번째 관찰값이 X 변수 공간에서 자료의 중심으로 떨어진 거리

- Hat matrix: $H = X(X^TX)^{-1}X^T$
  $\hat{Y} = HY = X(X^TX)^{-1}X^TY = X\hat{\beta}$

- 세 번째 그래프

```
> plot(fit_w, which=3)
```

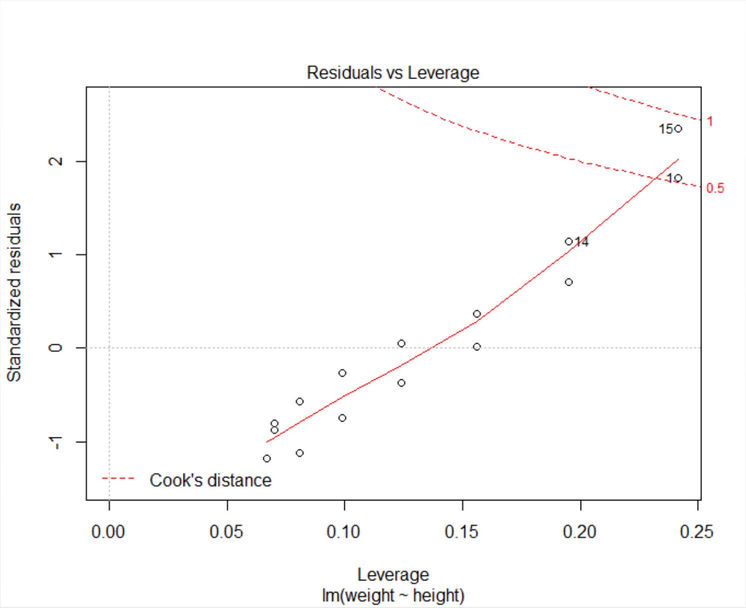


scale-location plot

- 동일 분산 만족 여부 확인
- $\sqrt{|r_i|}$와 $\hat{Y}_i$의 산점도
- 점들이 체계적으로 증가하거나 감소하는 경향이 있는지 확인

- 네 번째 그래프

```
> plot(fit_w, which=5)
```



- 관찰값 진단에 사용되는 그래프

1) 오차항의 동일 분산 가정

- 확인 방법

  ① 함수 plot( )으로 생성된 그래프 중 옵션 which=1과 which=3의 그래프

  ② 패키지 car의 함수 spreadLevelPlot( )에 의해 생성되는 그래프

  ③ 패키지 car의 함수 ncvTest( )로 실행되는 score 검정

- 패키지 car의 함수 spreadLevelPlot( )에 의한 확인

  - 실행 결과
    - 스튜던트화(Studentized) 잔차 t_i와 $\hat{Y}_i$의 산점도
    - 분산 안정화를 위한 반응변수의 변환 지수, 즉 Y^p의 p 값 제안

  - Studentized residual, t_i

$$t_i = \left(Y_i - \hat{Y}_{(i)}\right)/SE\left(Y_i - \hat{Y}_{(i)}\right)$$

    - 단, $\hat{Y}_{(i)}$는 i번째 관찰값을 제외한 나머지 $n-1$개의 자료로 수립한 회귀모형으로 제외된 Y_i를 예측한 통계량

    - 스튜던트화 잔차가 크다는 것은 i번째 관찰값이 이상값으로 분류될 가능성이 높다는 의미

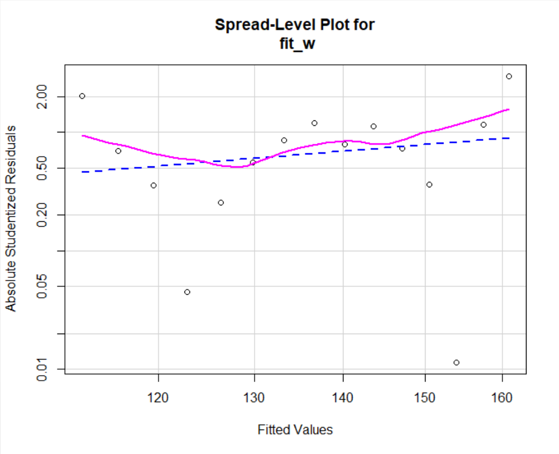
- 패키지 car의 함수 ncvTest( )에 의한 확인

  - Non Constant error Variance에 대한 Breusch-Pagan 검정 실시

  - 귀무가설: 오차의 분산이 일정
  - 대립가설: 오차의 분산이 $\hat{Y}_i$의 수준에 따라 변한다.

- 예: women 자료 회귀모형의 동일 분산 가정 확인

```
> fit_w <- lm(weight ~ height, women)

> library(car)
> spreadLevelPlot(fit_w)

Suggested power transformation:  -0.8985826
```



- 점들의 체계적인 증가 혹은 감소 여부

  파란 직선: 로버스트 회귀 직선
  빨간 곡선: 국소다항회귀 곡선

- 제안된 지수가 1과 큰 차이가 있는가?

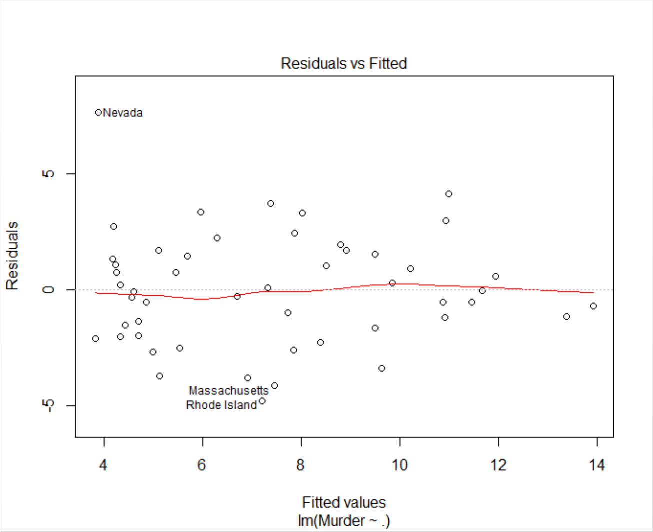
```
> ncvTest(fit_w)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.8052115, Df = 1, p = 0.36954
```

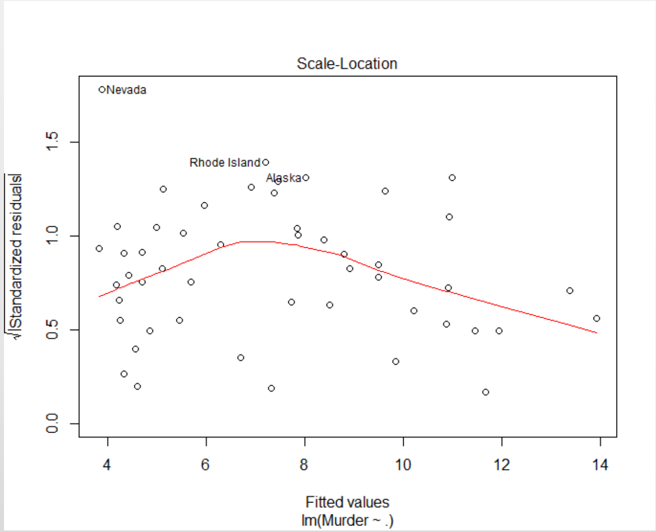
- 귀무가설 기각 못함

- 예: states 회귀모형의 동일 분산 가정 확인

```
> library(dplyr)
> states <- as.data.frame(state.x77)
> states <- select(states, Murder, Population,
                  Illiteracy, Income, Frost)
> fit_s <- lm(Murder ~ ., states)
```

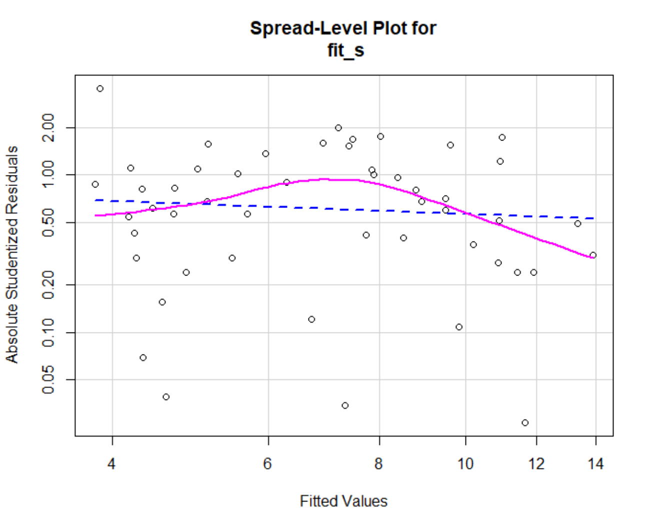
```
> plot(fit_s, which=1)
> plot(fit_s, which=3)
```





- states 회귀모형의 동일 분산 가정 확인

```
> spreadLevelPlot(fit_s)

Suggested power transformation:  1.209626
```




```
> ncvTest(fit_s)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.746514, Df = 1, p = 0.18632
```

2) 오차항의 정규분포 가정

- 확인 방법
  - 그래프에 의한 확인: QQ plot 작성
    ① 함수 plot( )에서 which=2로 생성되는 그래프: 표준화 잔차의 정규 분위수-분위수 그래프

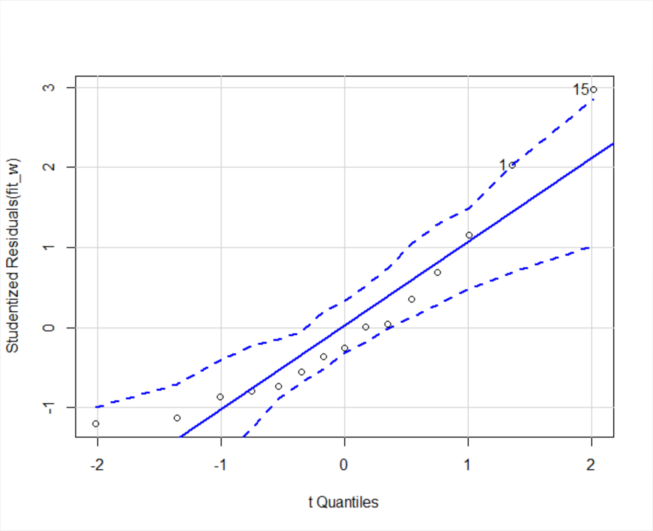
    ② 패키지 car의 함수 qqPlot( ): 스튜던트화 잔차의 t-분포 분위수-분위수 그래프. 가정이 만족되면 $t_i \sim t(n-k-2)$

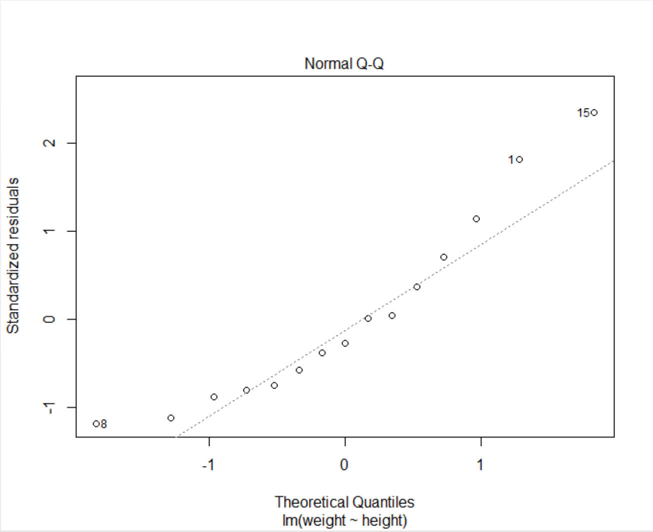
  - 검정에 의한 확인: 함수 shapiro.test()에 의한 Shapiro_Wilk 검정
    - 귀무가설: 정규 분포

- 예: women 자료 회귀모형의 정규 분포 가정 확인
  - QQ plot 작성

```
> qqPlot(fit_w)
[1]  1 15
```

```
> plot(fit_w, which=2)
```





점선: 모수적 Bootstrap에 의한 95% 신뢰영역

- Shapiro-Wilk 검정

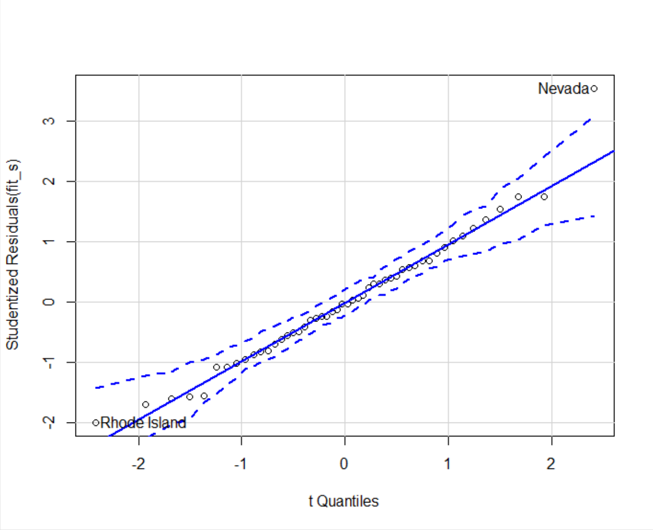
```
> shapiro.test(residuals(fit_w))

	Shapiro-Wilk normality test

data:  residuals(fit_w)
W = 0.91909, p-value = 0.1866
```

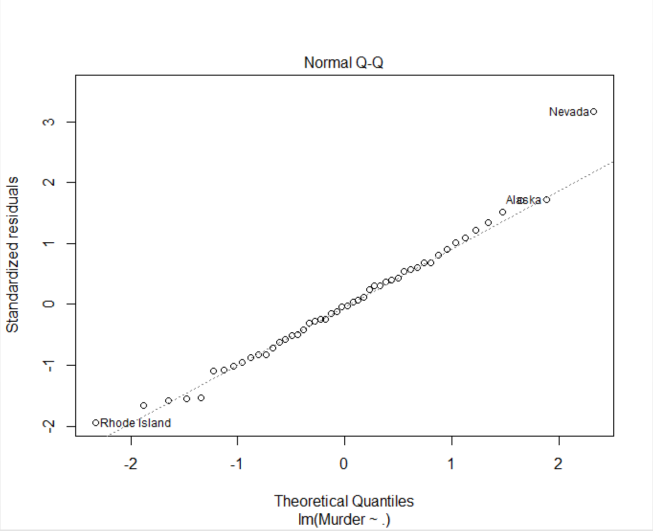
귀무가설 기각 못함

- 예: states 자료 회귀모형의 정규 분포 가정 확인
  - QQ plot 작성

```
> qqPlot(fit_s)
      Nevada Rhode Island
          28           39
```

```
> plot(fit_s, which=2)
```





전반적으로 문제는 없으나 Nevada는 주의해야 할 자료

- Shapiro-Wilk 검정

```
> shapiro.test(residuals(fit_s))

        Shapiro-Wilk normality test

data:  residuals(fit_s)
W = 0.98264, p-value = 0.6672
```

3) 오차항의 독립성 가정

- 독립성 가정을 확인해야 하는 경우
  1) 시간의 흐름에 따라 관측된 시계열 자료
  2) 공간에 따라 관측된 공간 자료

- 독립성 가정의 위반 형태: 매우 다양함

  - k차 자기상관 계수: $\rho_k = \mathrm{corr}(\varepsilon_i, \varepsilon_{i-k}),\ k = 1, 2, \ldots$

  - Durbin-Watson 검정의 귀무가설 $H_0\colon \rho_1 = 0$

  - Breusch-Godfrey 검정의 귀무가설 $H_0\colon \rho_1 = \cdots = \rho_K = 0$

- 독립성 가정 위반의 예: p차 자기회귀오차 회귀모형

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

$$\varepsilon_i = \phi_1 \varepsilon_{i-1} + \cdots + \phi_p \varepsilon_{i-p} + a_i, \qquad a_i \ \ iid \ \ N(0, \sigma^2)$$

- 오차항의 1차 자기상관 관계 여부 확인

  - 패키지 car의 함수 durbinWatsonTest( )

- 예제: 패키지 carData의 Hartnagel

  - 1931년부터 1968년까지 캐나다 범죄율 자료
  - 반응변수: 여성 범죄율(fconvict)
  - 설명변수: 출산율(tfr), 여성 고용률(partic)
  - 회귀모형의 독립성 가정 확인

```
> library(car)
> fit_h <- lm(fconvict ~ tfr + partic, Hartnagel)

> durbinWatsonTest(fit_h)
 lag Autocorrelation D-W Statistic p-value
   1        0.714117     0.5453534       0
 Alternative hypothesis: rho != 0
```

- 유의한 1차 자기상관계수
- 독립성 가정 만족 못함

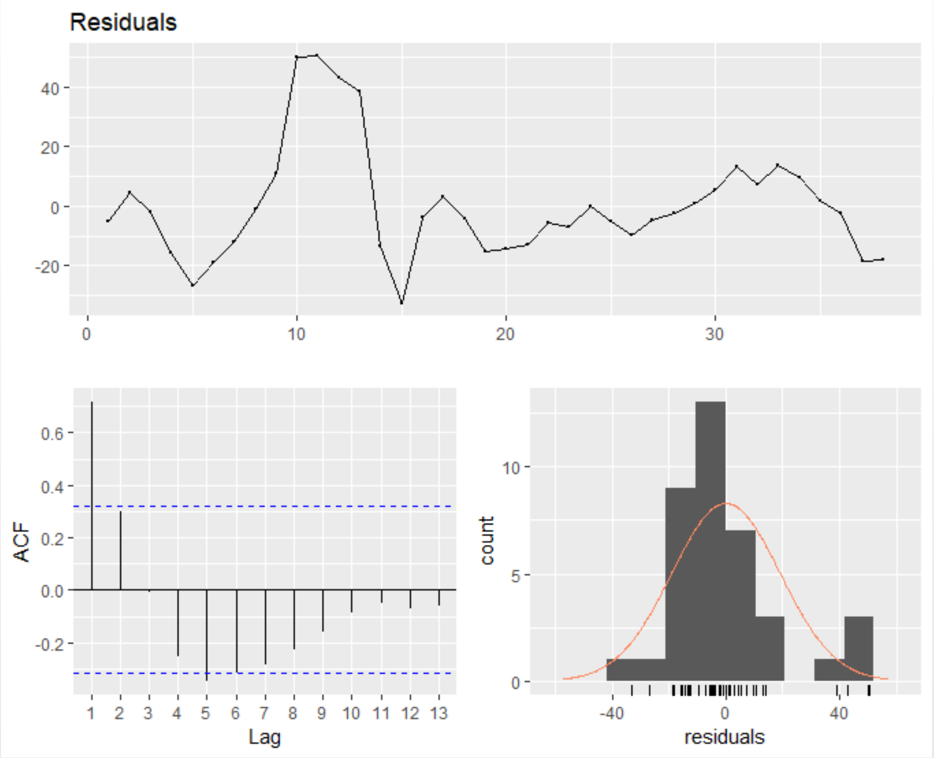
- H_0: $\rho_1 = \cdots = \rho_K = 0$의 검정

  - 패키지 forecast의 함수 checkresiduals( )

- 예제: 패키지 carData의 Hartnagel

```
> library(forecast)
> checkresiduals(fit_h)

Breusch-Godfrey test for serial correlation of order up to 7

data:  Residuals
LM test = 25.083, df = 7, p-value = 0.0007336
```

- 함수 checkresiduals( )로 작성된 그래프



- 잔차 시계열 그래프
- 잔차의 각 시차(lag)별 표본 자기상관도표(표본 자기상관계수)
- 잔차의 히스토그램

- 파란 점선: 각 시차별 자기상관계수의 95% 신뢰구간
- 각 시차별 자기상관계수의 유의성 검정
- 다중검정에 주의

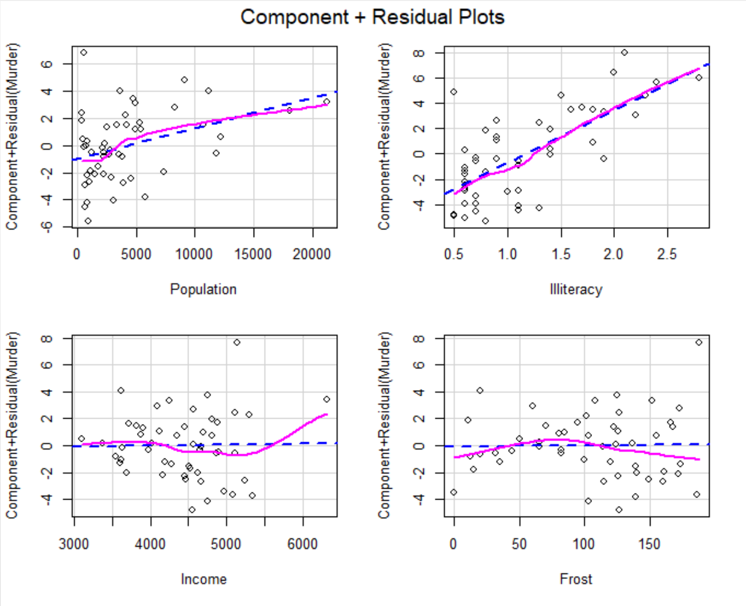
4) 선형관계

- 반응변수와 설명변수의 선형관계 확인
  1) 단순회귀: 두 변수의 산점도로 간단하게 확인 가능
  2) 다중회귀: 다른 변수의 영향력으로 인하여 (X_i, Y) 혹은 (X_i, e_i)의 산점도는 큰 의미가 없음
- 변수 X_i의 부분 잔차(partial residual)
  - 모형에 포함된 다른 설명변수의 영향력이 제거된 잔차
  - $Y - \sum_{j \neq i} \hat{\beta}_j X_j = \hat{Y} + e - \sum_{j \neq i} \hat{\beta}_j X_j = e + \hat{\beta}_i X_i$
- 다중회귀모형에서 X_i와 Y의 선형관계 확인:
  - X_i와 부분 잔차 $e + \hat{\beta}_i X_i$의 산점도 작성

- 선형관계 확인
  - 부분 잔차 산점도 작성: 패키지 car의 함수 crPlots( )
  - Curvature test: 패키지 car의 함수 residualPlots( )

- 예: states 자료 회귀모형의 선형관계 확인

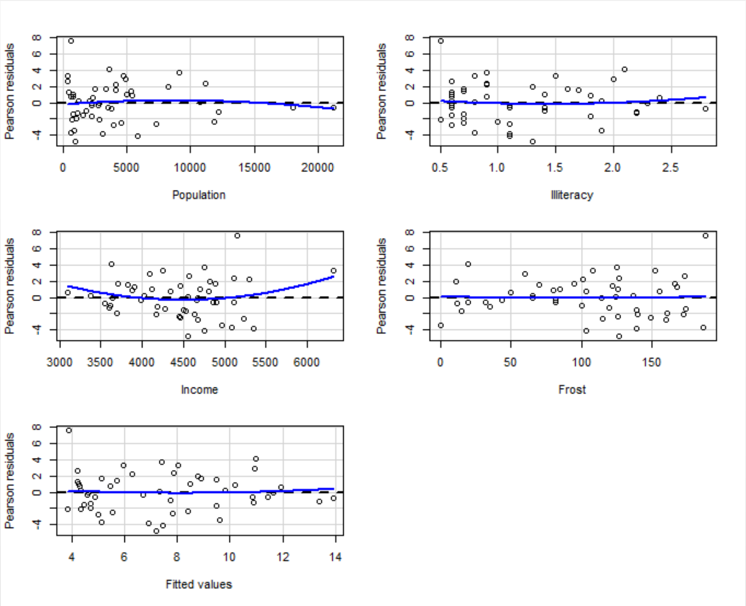
```
> crPlots(fit_s)
```



- 각 변수의 부분 잔차에 대한 회귀직선과 국소다항회귀곡선 추가
- 선형 관계에는 문제가 없음
- 변수 Income, Frost의 영향력이 매우 적음을 확인

- Curvature test: 모형에 하나씩 포함된 각 변수의 제곱항의 유의성 검정

```
> residualPlots(fit_s)
           Test stat Pr(>|Test stat|)
Population   -0.5218          0.60446
Illiteracy    0.4625          0.64601
Income        1.8355          0.07319 .
Frost         0.0507          0.95981
Tukey test    0.3190          0.74971
```



5) 다중공선성

- 설명변수 사이에 강한 상관관계가 존재하는 경우
- 회귀모형의 가정과 직접적인 연관은 없음
- 추정량의 분산이 크게 증가: 회귀모형 추정에 영향을 줄 수 있음

• VIF에 의한 다중공선성 존재 여부 확인

  - 분산팽창계수(VIF): 변수 X_j의 VIF는 $1/(1-R_j^2)$, 단 R_j^2는 X_j를 종속변수, 나머지 설명변수를 독립변수로 하는 회귀모형의 결정계수
  - $Var(\hat{\beta}_j) = \sigma^2/(1-R_j^2)$: VIF 값이 크면 $\hat{\beta}_j$의 분산이 커짐
  - $R_j^2 = 0.75$ → VIF=4 / $R_j^2 = 0.8$ → VIF=5 / $R_j^2 = 0.9$ → VIF=10

- 분산팽창계수 계산
  - 패키지 car의 함수 vif( )
  - 패키지 faraway의 함수 vif( )
- 예: states 자료 회귀모형의 다중공선성 확인

```
> library(car)

> vif(fit_s)
Population Illiteracy     Income     Frost
  1.245282   2.165848   1.345822   2.082547
```

```
> faraway::vif(fit_s)
Population Illiteracy     Income     Frost
  1.245282   2.165848   1.345822   2.082547
```

- 예: 데이터 프레임 Duncan에서 반응변수 prestige, 설명변수 income, education, type의 회귀 모형 설정

```
> fit_1 <- lm(prestige ~ income + education + type, data=Duncan)
> summary(fit_1)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.18503    3.71377  -0.050  0.96051    
income        0.59755    0.08936   6.687 5.12e-08 ***
education     0.34532    0.11361   3.040  0.00416 ** 
typeprof     16.65751    6.99301   2.382  0.02206 *  
typewc      -14.66113    6.10877  -2.400  0.02114 *  
---

Residual standard error: 9.744 on 40 degrees of freedom
Multiple R-squared:  0.9131,	Adjusted R-squared:  0.9044 
F-statistic:   105 on 4 and 40 DF,  p-value: < 2.2e-16
```

- VIF로 모형의 다중공선성 확인

1) 패키지 faraway의 함수 vif( )

```
> faraway::vif(fit_1)
   income education  typeprof    typewc
 2.209178  5.297584  5.562395  2.043721
```

- 개별 dummy variable에 대한 VIF 값 계산은 큰 의미가 없음
- 범주형 변수 type에 대한 VIF가 더 의미가 있음

2) 패키지 car의 함수 vif( )

```
> vif(fit_1)
              GVIF Df GVIF^(1/(2*Df))
income    2.209178  1        1.486330
education 5.297584  1        2.301648
type      5.098592  2        1.502666
```

- DF: 각 변수에 대한 자유도
- 범주형 변수 type을 표현하기 위해 2개의 가변수 사용

- GVIF(Generalized VIF): 범주형 변수에도 적용 가능한 VIF
- GVIF^(1/2*DF))의 제곱 값: 연속형 변수의 VIF에 대한 기준 값으로 비교 가능