대학교 2-2/회귀

5.3 연습문제

Launa 2024. 11. 28. 17:24

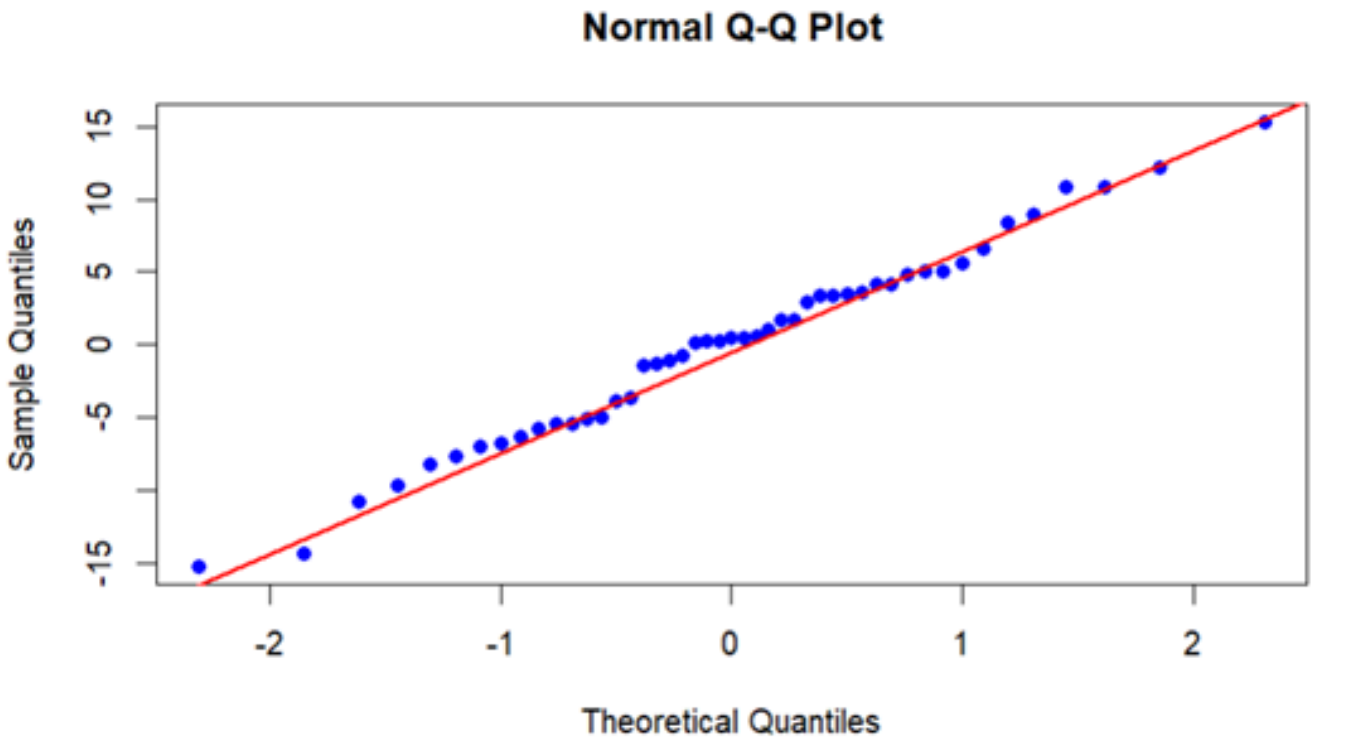
R 시스템 내장 데이터셋을로 1888년 스위스 농업과 사회경제학적데이터셋 swiss를 사용하시오. Fertility를 반응변수로 하고 나머지 변수들을 설명변수로 하여 다중선형회귀 모형을 적합하고 다음에 답하시오.

(a) 가정에 대한 검토로 잔차를 이용하여 등분산성을 점검하시오

```
> data(swiss)
> swiss.lm = lm(Fertility ~ ., data = swiss)
#데이터의 모든 나머지 변수들을 Fertility 설명하는는 설명변수로 사용
> plot(swiss.lm$fitted.values, resid(swiss.lm), col=4)
> abline(h = 0, col = "red")
```

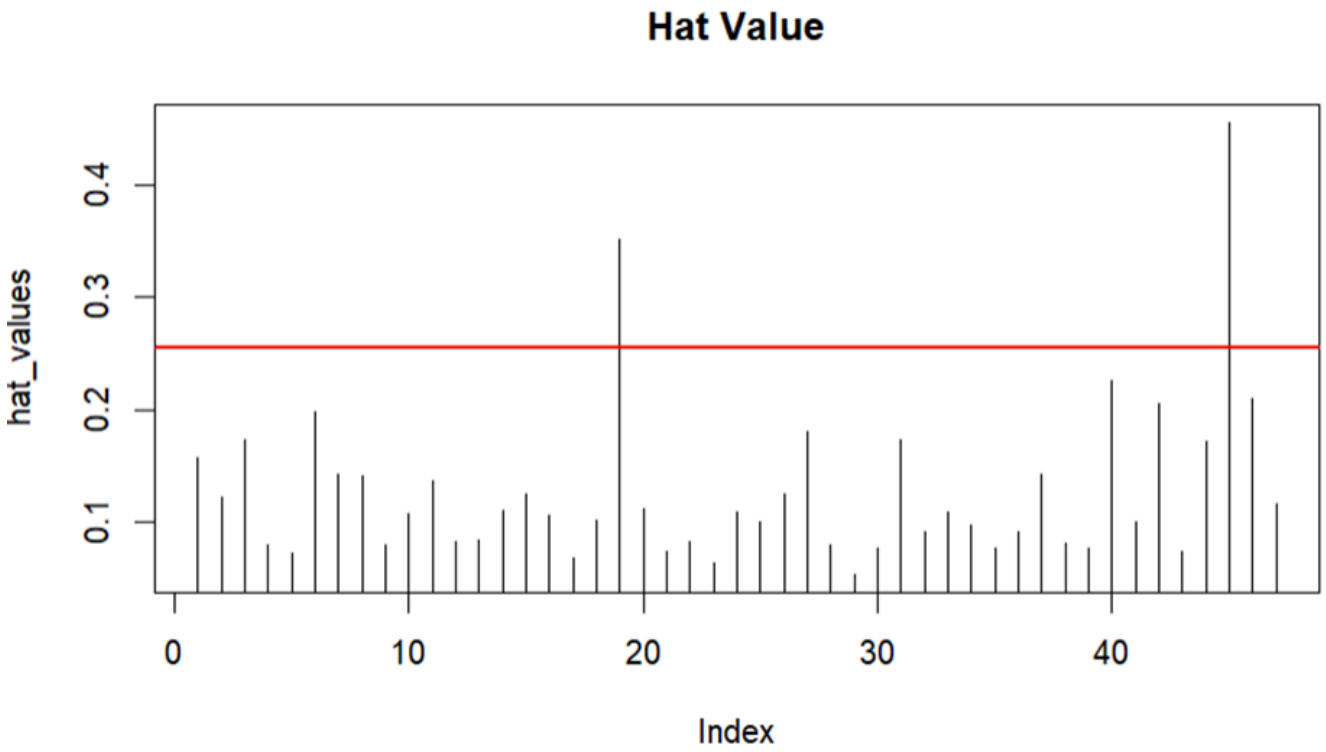
(b) 가정에 대한 검토로 잔차를 이용하여 정규성을 점검하시오

```
> residuals=residuals(swiss.lm)
> qqnorm(residuals,pch = 19, col = "blue")
> qqline(residuals, col = "red", lwd = 2)
```



(c) 지렛대점이 있는지 점검하시오

```
> hat_values = hatvalues(swiss.lm)
> n=nrow(swiss)
#관측치 개수
> p=length(coef(swiss.lm))-1
# 설명변수 개수
> threshold = 2 * (p + 1) / n
# 임계값
> plot(hat_values, type = "h", main = "Hat Value")
> abline(h = threshold, col = "red", lwd = 2)
# 임계값 표시
```



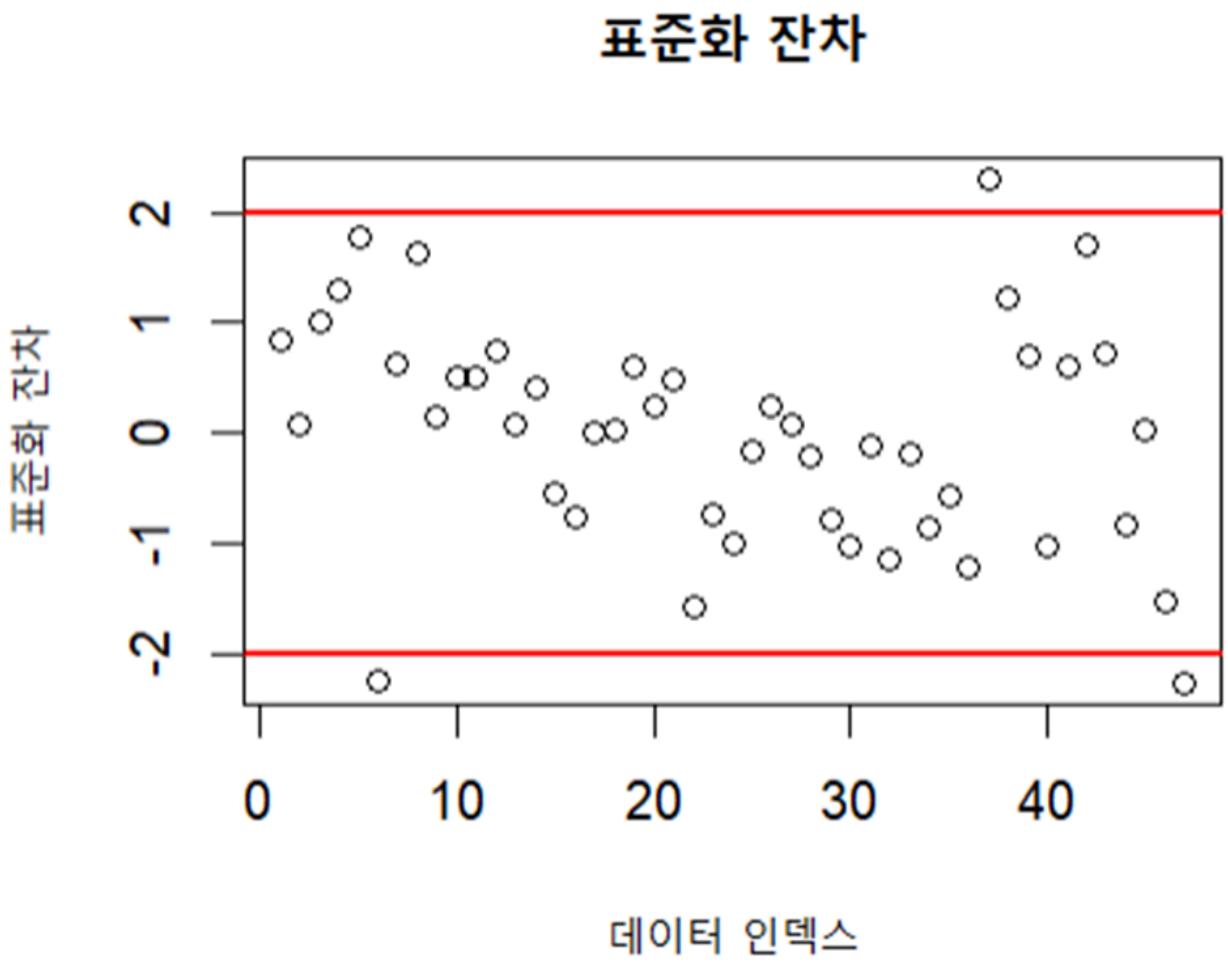
Hat 값: 지렛대점을 확인하는 가장 간단한 지표. 각 관측치가 회귀선에 미치는 영향을 나타냄.
임계값: 임계값=2(p+1)/n
p: 설명변수의 개수 (독립 변수의 수)
n: 관측치의 개수 (데이터의 행 수)
관측치의 Hat 값이 이 임계값보다 크면 지렛대점으로 간주

(d) 이상점이 있는지 검토하시오

```
# 표준화 잔차 계산
> standardized_residuals = rstandard(swiss.lm)
# 표준화 잔차가 ±2를 넘는 이상점 찾기
> outliers_residuals = which(abs(standardized_residuals) > 2)
> outliers_residuals
```

```
Porrentruy      Sierre Rive Gauche
         6         37         47
```

```
>plot(standardized_residuals, main = "표준화 잔차",  ylab = "표준화 잔차", xlab = "데이터 인덱스")
> abline(h = c(-2, 2), col = "red", lwd = 2)  # ±2 기준선
```



이상점: 예측한 값과 실제 값이 많이 차이가 나는 점들

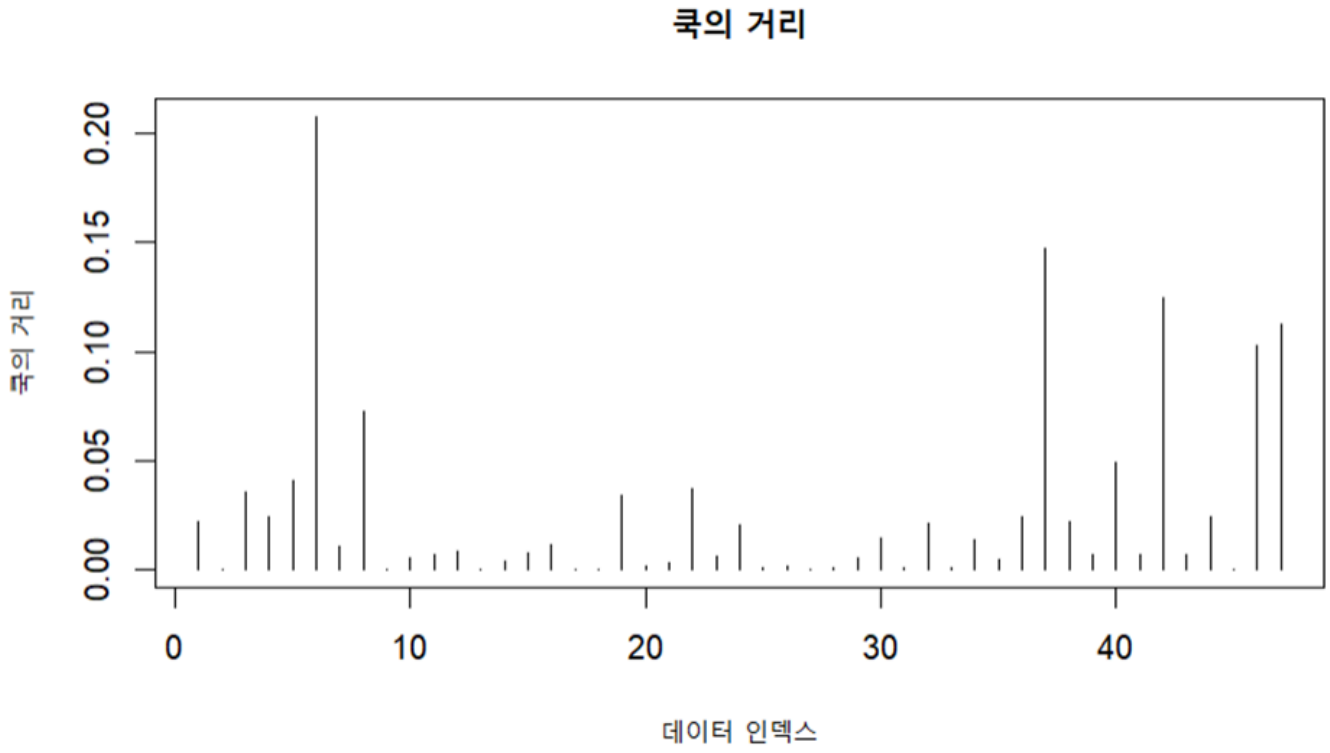
표준화 잔차: 잔차(예측 값과 실제 값의 차이)를 표준화한 값/ 표준화 잔차가 ±2를 넘는 값은 이상점으로 간주
쿡의 거리: 각 점이 모델에 미치는 영향/ 쿡의 거리가 0.5를 넘으면 이상점일 가능성 O

(e) 영향점이 있는지 검토하시오

```
# 쿡의 거리 계산
> cooks_distance = cooks.distance(swiss.lm)
# 쿡의 거리가 0.5를 넘는 이상점 찾기
> influential_points = which(cooks_distance > 0.5)
> outliers_cooks
named integer(0)

# 쿡의 거리 시각화
```

```
> plot(cooks_distance, type = "h", main = "쿡의 거리",xlab = "데이터 인덱스", ylab = "쿡의 거리")
> abline(h = 0.5, col = "red", lty = 2)  # 0.5 기준선
```



영향점 확인을 위해 주로 쿡의 거리(Cook's Distance) 사용
쿡의 거리는 각 데이터 점이 회귀 분석 결과에 얼마나 영향을 미치는지 나타내는 지표
쿡의 거리가 0.5 이상이면 영향점으로 의심 O / 쿡의 거리가 1 이상이면 강력한 영향점으로 간주할 수 O

(f) 설명변수들과 반응변수 간의 회귀식이 적합한지 진단하시오

```
> summary(swiss.lm)
```

```
Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07
Agriculture      -0.17211    0.07030  -2.448  0.01873
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05
Catholic          0.10412    0.03526   2.953  0.00519
Infant.Mortality  1.07705    0.38172   2.822  0.00734

(Intercept)      ***
Agriculture      *
Examination
Education        ***
Catholic         **
Infant.Mortality **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,     Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

잔차 분석을 통해:
정규성: 잔차가 정규분포를 따른다면 회귀식이 적합하다고 볼 수 있음
잔차가 정규분포를 따르는지 Q-Q 플롯과 샤피로-윌크 검정으로 확인할 수 있음
등분산성: 잔차가 예측값에 대해 고르게 분포하면 적합한 모델
잔차가 등분산성을 만족하는지, 즉 잔차의 분산이 일정한지 확인하려면 잔차 대 예측값의 산점도를 확인
선형성: 잔차와 설명변수 간에 선형 관계가 있으면 적합한 모델
잔차가 선형성을 만족하는지 확인하려면 잔차 대 설명변수의 산점도를 확인

모델 적합도:
R^2 값이 크고, F-검정에서 p-value가 0.05 이하라면 모델이 잘 적합된 것
R^2 (결정계수)는 모델이 설명할 수 있는 변동성의 비율을 나타냄/ 값이 1에 가까울수록 모델이 잘 적합된 것

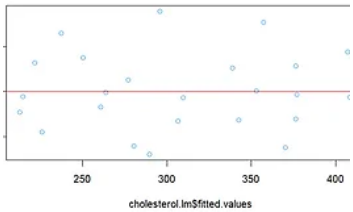
F-검정은 전체 모델이 유의미한지 확인하는 검정/ p-value가 0.05보다 작으면 모델이 유의미하다고 할 수 있음

공감

'대학교 2-2 > 회귀' 카테고리의 다른 글

5.4 연습문제 (0) 2024.11.28

'대학교 2-2/회귀' Related Articles



5.4 연습문제

초보 개발자의 성장기
Launa 님의 블로그입니다.

댓글 0

Launa
내용을 입력하세요.

등록