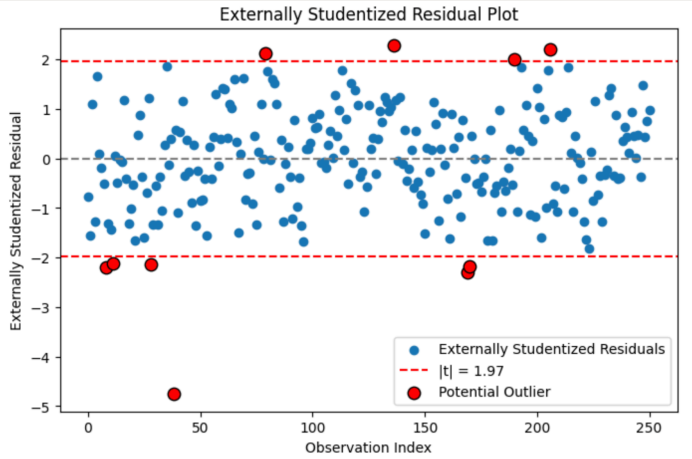
# 회귀분석2_과제3 (보고서)

# 1-(a)

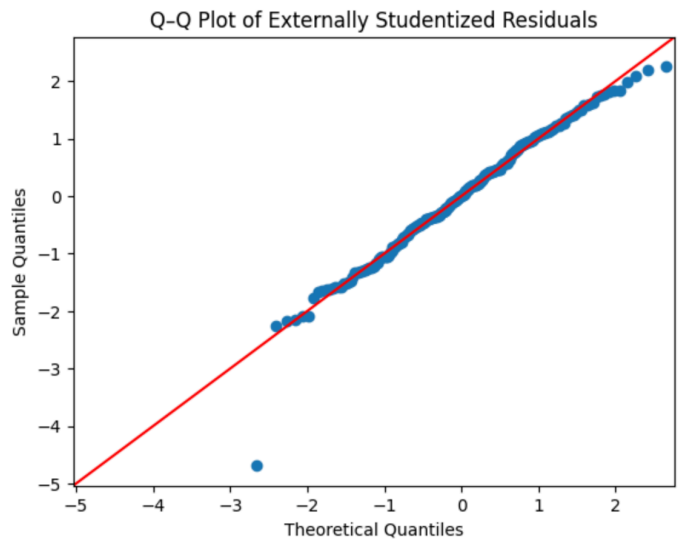
## 1. 특이값(outlier)

- Externally Studentized Residual Plot




⇒ 대부분의 관측치는 외표준화 잔차가 ±1.97 범위 내에 분포하여 95% 신뢰구간을 만족한다. 그러나 총 7개의 관측치가 이 범위를 벗어나 통계적으로 유의한 특이값(Potential Outlier)으로 식별되었다. 특히, 좌측 하단의 기준선과 가장 멀리 떨어진 관측치의 외부표준화 잔차는 약 -4.8로, 모델 예측에서 가장 크게 벗어난 특이값(Outlier)으로 판단된다.

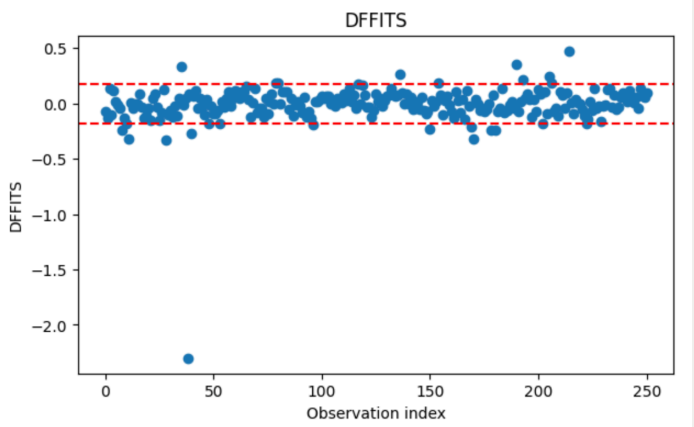
- Q-Q Plot of Residuals




⇒ 대부분의 관측치는 정규분포 선을 따라 분포하고 있으나, 좌측 하단 하나의 관측치가 정규분포 선을 크게 벗어나, 뚜렷한 특이값(outlier)로 판단된다.

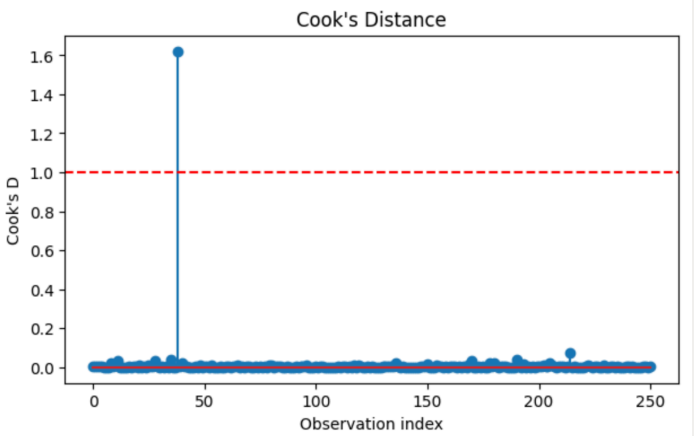
## 2. 영향력이 큰 관측값(influential point)

- Dffits



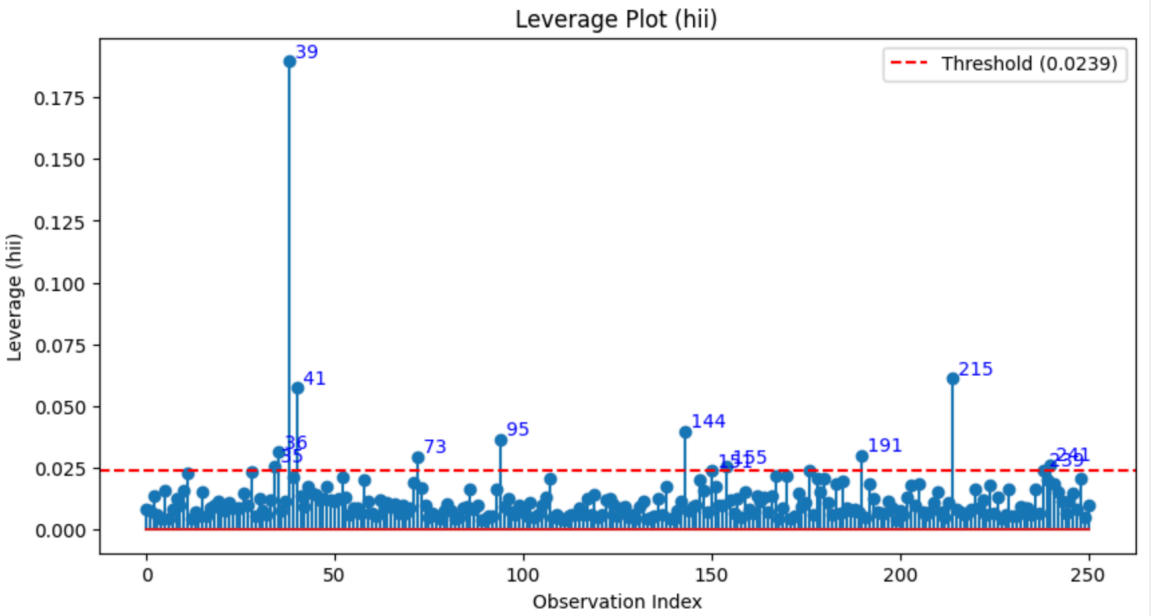

⇒ 대부분의 관측치는 0 근처에 분포하지만, 하나의 관측치의 Diffits 값이 약 -2.2로 크게 벗어나, 회귀선을 강하게 아래쪽으로 끌어당기는 영향력이 큰 관측값으로 판단된다.

- Cook’s Distance




⇒ 대부분 관측치의 Cook’s Distance는 0 근처에 분포하지만, 하나의 관측치는 1.0을 크게 넘어서고 있어, 모델 전체에 큰 영향을 미치는 영향력이 큰 관측값으로 판단된다.

# 1-(c)




- **진단 과정**
  Hat Matrix의 대각성분인 $h_{ii}$를 이용해 각 관측치에 대한 Leverage 값을 계산하고, Leverage Plot으로 시각화하였다. $\frac{2(k+1)}{n} = 0.024$를 기준선으로 표시하고, 이 기준선을 초과하는 관측치를 설명변수의 값이 크게 다른 특이값으로 판단하였다.
- **진단 결과**
  대부분의 관측치는 기준선보다 낮아 설명변수 값이 비교적 균일하게 분포하였다. 그러나 일부 관측치는 기준선을 초과하였으며, 특히 39번째 관측치는 이를 크게 기준선을 매우 크게 벗어나고 있으므로 설명변수의 값이 크게 다른 특이값으로 판단된다.

## 1-(e)

```python
# Cook's Distance
influence = lm.get_influence()
Dcooks = influence.cooks_distance[0]
n = len(df)
threshold = 1
influential_idx = np.where(Dcooks > threshold)[0]

print("Cook's D 기준 영향점 관측치:", df.index[influential_idx].tolist())
print("Cook's D", Dcooks[38])
```

```
Cook's D 기준 영향점 관측치: [39]
Cook's D 1.6213851067281648
```

- **진단 과정**
  (d)에서 구한 Cook’s Distance 값을 분석하여, 1을 초과하는 관측치와 해당 관측치의 Cook’s Distance 값을 출력하였다.
- **진단 결과**
  데이터프레임 기준 39번 관측치는 Cook’s Distance 값이 1을 넘어 모델 전체 회귀계수에 큰 영향을 미치는 Influential point로 확인되었다.

## 2-(b)

| | feature | VIF |
|---|---|---|
| 0 | age | 2.250580 |
| 1 | weight | 44.692519 |
| 2 | height | 2.938055 |
| 3 | neck | 4.447757 |
| 4 | chest | 10.226609 |
| 5 | abd | 12.768313 |
| 6 | hip | 14.649188 |
| 7 | thigh | 7.862916 |
| 8 | knee | 4.783489 |
| 9 | ankle | 1.923161 |
| 10 | biceps | 3.669970 |
| 11 | forearm | 2.194400 |
| 12 | wrist | 3.368475 |

- 진단 과정
  다중공선성 여부를 진단하기 위해 각 설명변수에 대한 VIF를 계산하였으며, VIF가 10을 초과하는 경우, 다중공선성 가능성이 있다고 판단하였다.
- 진단 결과
  총 4개의 변수 weight, chest, abd, hip은 VIF 기준치 10을 초과하였다. 특히, weight 변수의 VIF는 44로 매우 높게 나타나, 강한 다중공선성 가능성이 있을 것으로 판단된다.