가중회귀분석

가중회귀 분석은 <교재 29페이지>에 나오는 내용인데 교재에 있는 예제는 중회귀모형에 대한 예제가 맞나 싶을 정도로 이해가 되지 않았습니다.

인터넷에서 적당한 예제를 찾아 교재에 있는 내용으로 실습을 해보면서 정리하였습니다.

1. 가중회귀분석을 사용하는 이유

- 회귀모형에 부여되는 가정 3가지는 아래와 같습니다.
  - ① 오차의 등분산성
  - ② 모형의 선형성
  - ③ 오차의 정규성

    이 중에 ① 오차의 등분산성 가정이 위배되는 경우에 가중회귀분석을 사용하게 됩니다.

- ① 오차의 등분산성 의 판단
  1. 산점도의 형태를 확인
     오차의 등분산성 여부를 판단하는 방법 중 가장 간단한 것으로는 잔차나 표준화 잔차를 Y축으로 하고 $\hat{Y}$를 X축으로 하는 산점도 (e, $\hat{Y}$)에서 $\hat{Y}$ 가 증가함에 따라 Y축값의 퍼빔정도가 증가 또는 감소하는 모양의 산점도는 분산이 일정하지 않음을 나타냄을 알 수 있습니다.
  2. 스코어 검정
     일반적으로 잔차산점도의 형태를 보고 등분산성의 가정에 대한 판단이 가능하지만 그 형태가 명확하지 않은 경우 주관적인 판단이 어렵기 때문에 Cook과 Weisberg(1983)가 제시한 스코어 검정 방법을 실시하여 판단할 수 있다.

2. 실습 과정 정리

- 이 실습 데이터는 https://www.youtube.com/watch?v=DTt0hLyRaTc&list=PLKmcZujz-ZJlcy1F9CUs9kCkP-DVCaQ-T&index=4 (https://www.youtube.com/watch?v=DTt0hLyRaTc&list=PLKmcZujz-ZJlcy1F9CUs9kCkP-DVCaQ-T&index=4) 에서 사용한 데이터를 사용했습니다.

예제 데이터 셋 설명

- 이 예제는 컴퓨터를 이용한 학습에 관한 데이터셋이라고 합니다.
  - x : 레슨을 완료하는 동안의 총 응답 수
  - y : 수업 중 컴퓨터 사용 시간의 비용
  - 동영상에서 사용한 예제의 원래 데이터로 스코어 검정을 하면 유의확률이 0.05보다 크게 나오기 때문에 일부 데이터를 수정하여 테스트를 하였습니다.

테스트 진행

1. 데이터를 입력한 후 모형을 적합하고 요약정보를 확인합니다.

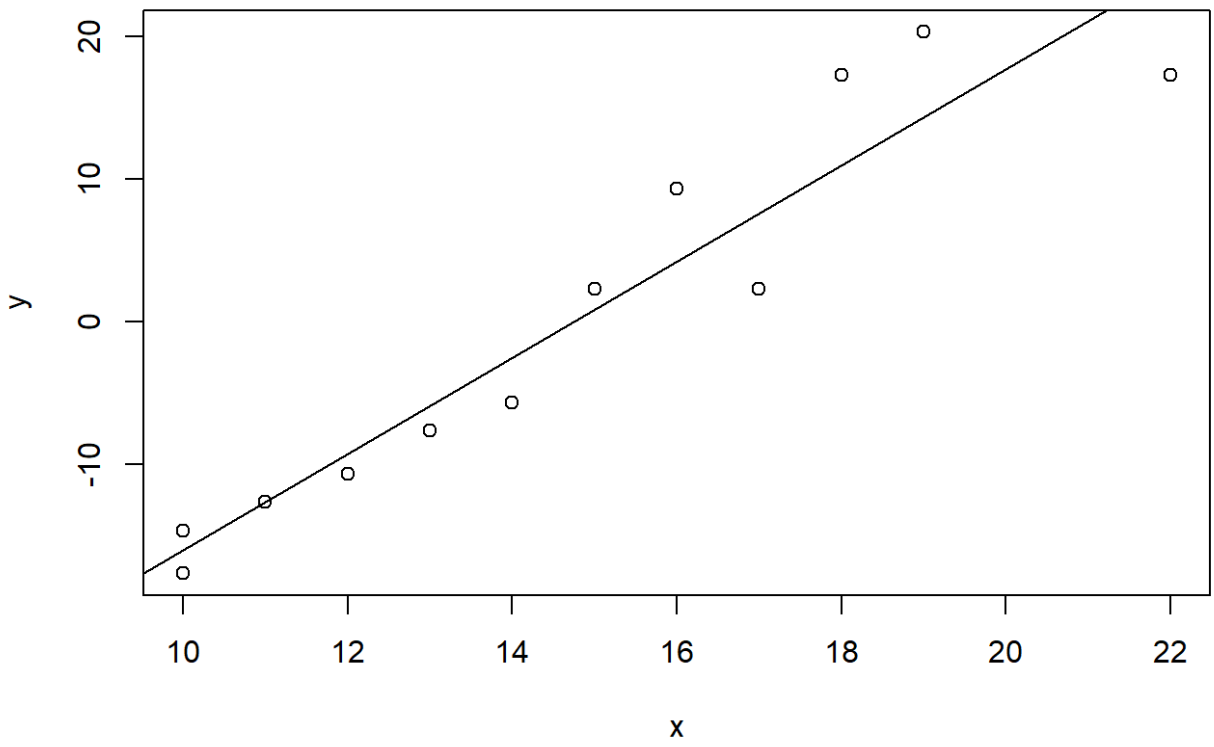
```
x = c(16,15,22,10,14,17,10,13,19,12,18,11)
y2 = c(77,70,85,50,62,70,53,60,88,57,85,55)
y = y2 - mean(y2)
olsfit=lm(y~x)
summary(olsfit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1311 -2.1051 -0.6998  2.3961  6.3665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -49.7725     5.4370  -9.154 3.55e-06 ***
## x             3.3744     0.3579   9.428 2.72e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.503 on 10 degrees of freedom
## Multiple R-squared:  0.8989, Adjusted R-squared:  0.8888
## F-statistic: 88.88 on 1 and 10 DF,  p-value: 2.721e-06
```

위에서 적합한 회귀선은 $\hat{y} = 17.8942 + 3.3744 * x$ 입니다.

2. x와 y의 산점도와 적합된 회귀선 확인하기

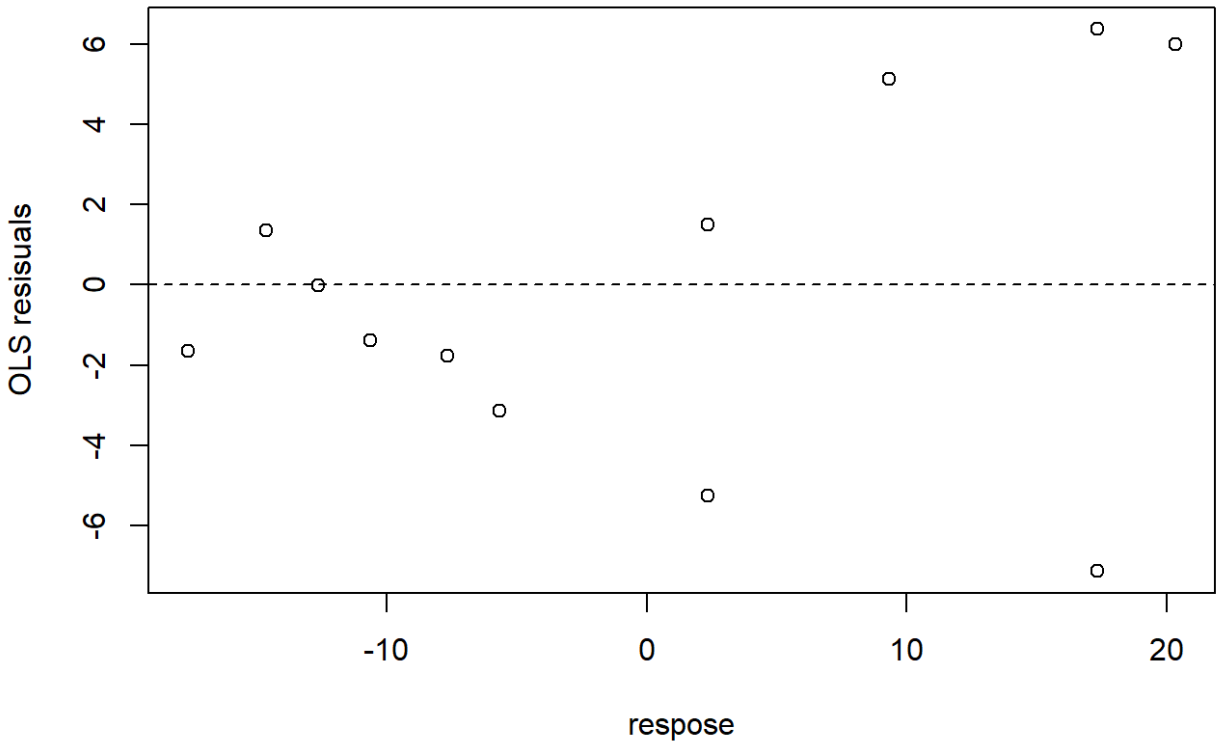
```
plot(x, y)
abline(olsfit)
```



x와 y의 산점도와 적합한 회귀선을 확인해보면 데이터와 회귀선이 잘 맞는것으로 보여집니다.

3. y와 잔차의 산점도 확인하기

```
plot(y, olsfit$residuals, xlab="respose", ylab="OLS resisuals")
abline(h=0,lty=2)
```



반응변수 y와 회귀모형의 잔차의 산점도를 그려보면 y가 0보다 작을때 잔차는 0에 가까이 있는게 보여지지만 0보다 큰 경우에는 잔차가 0과 멀어지져서 점점넓어지는 < 와 비슷한 모형으로 볼 수도 있을 것 같습니다. 즉 등분산 가정을 만족시키지 않는다고 볼 수 있습니다.

4. 스코어 검정으로 확인하기

```
library(car)
```

```
## Loading required package: carData
```

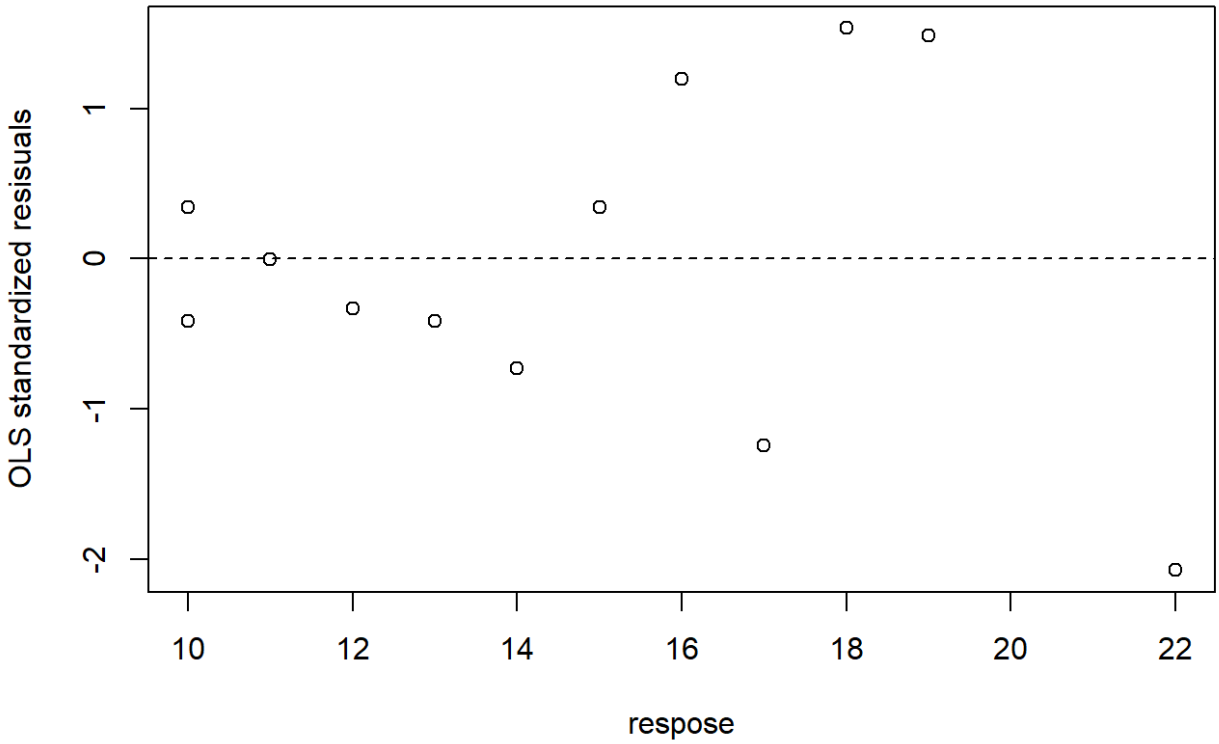
```
ncvTest(olsfit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 5.562254, Df = 1, p = 0.018352
```

Chisquare = 5.562254 이고, p = 0.018352 로 0.05보다 작기 때문에 등분산 가정을 기각하게 됩니다.

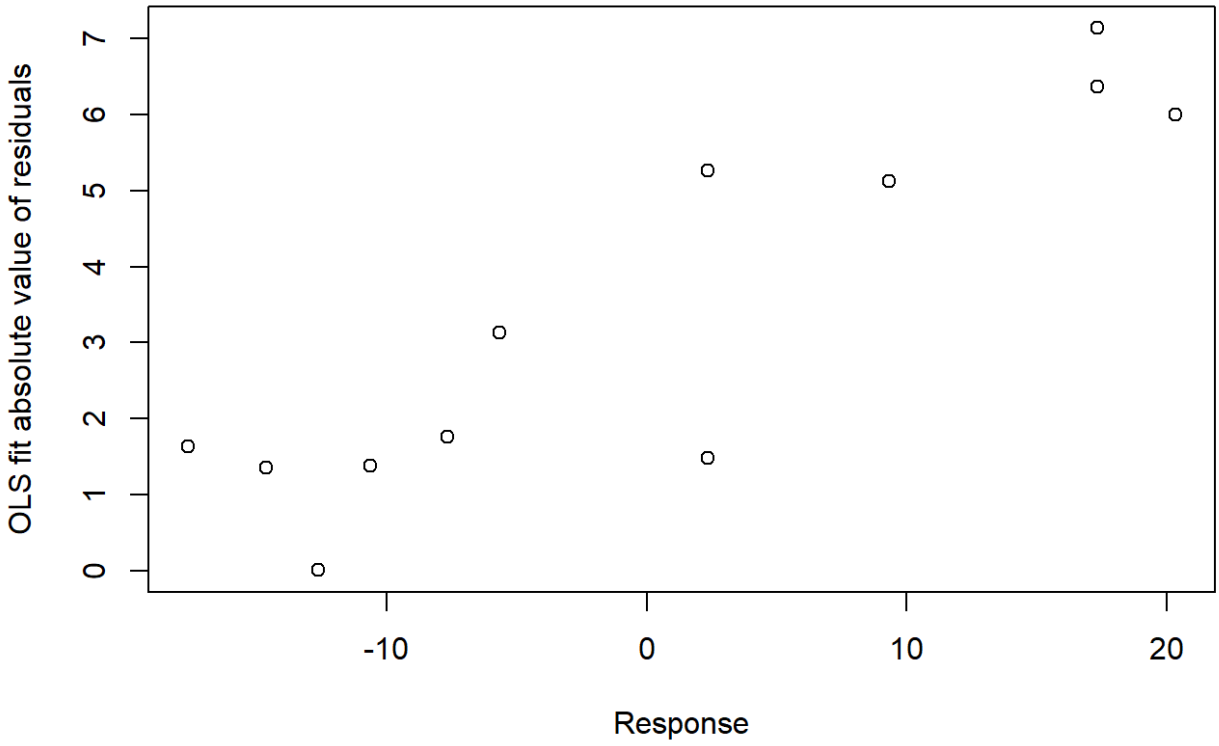
5. 표준화 잔차도의 산점도를 그려보았습니다.
    - 가중회귀선을 반응변수와 잔차와의 산점도를 그려보면 확인이 어려울 수있지만 표준화 잔차로 산점도를 그려보면 차이가 나는것을 확인할 수도 있어서 한번 출력해 봅니다.

```
plot(x, rstandard(olsfit), xlab="respose", ylab="OLS standardized resisuals")
abline(h=0,lty=2)
```



6. 아래는 적합한 회귀선의 잔차를 절대값으로 출력해 보았습니다.
   잔차는 0보다 크기 때문에 잔차가 0을 기준으로 위쪽으로 산점도가 겹쳐집니다.

```
plot(y, abs(olsfit$residuals), xlab="Response", ylab="OLS fit absolute value of residuals")
```



위의 산점도를 보면 전체적인 잔차의 분포가 설명변수가 증가하면서 점차 커지는 것을 확인할 수 있습니다.

7. $\hat{y}^2$ 값을 가중값으로 모형 적합
    - $\hat{y}^2$ 값을 가중값으로 모형을 적합한 후 기존 모형과 가중모형의 요약정보를 확인해 봅니다.

```
wghts = 1/((olsfit$fitted.values)^2)
wlsfit = lm(y~x, weights = wghts)
summary(wlsfit)
```

```
##
## Call:
## lm(formula = y ~ x, weights = wghts)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2642 -0.2555  0.1248  0.2802  0.6756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -65.1234     9.5978  -6.785 4.83e-05 ***
## x             4.4754     0.6432   6.959 3.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6476 on 10 degrees of freedom
## Multiple R-squared:  0.8288, Adjusted R-squared:  0.8117
## F-statistic: 48.42 on 1 and 10 DF,  p-value: 3.906e-05
```

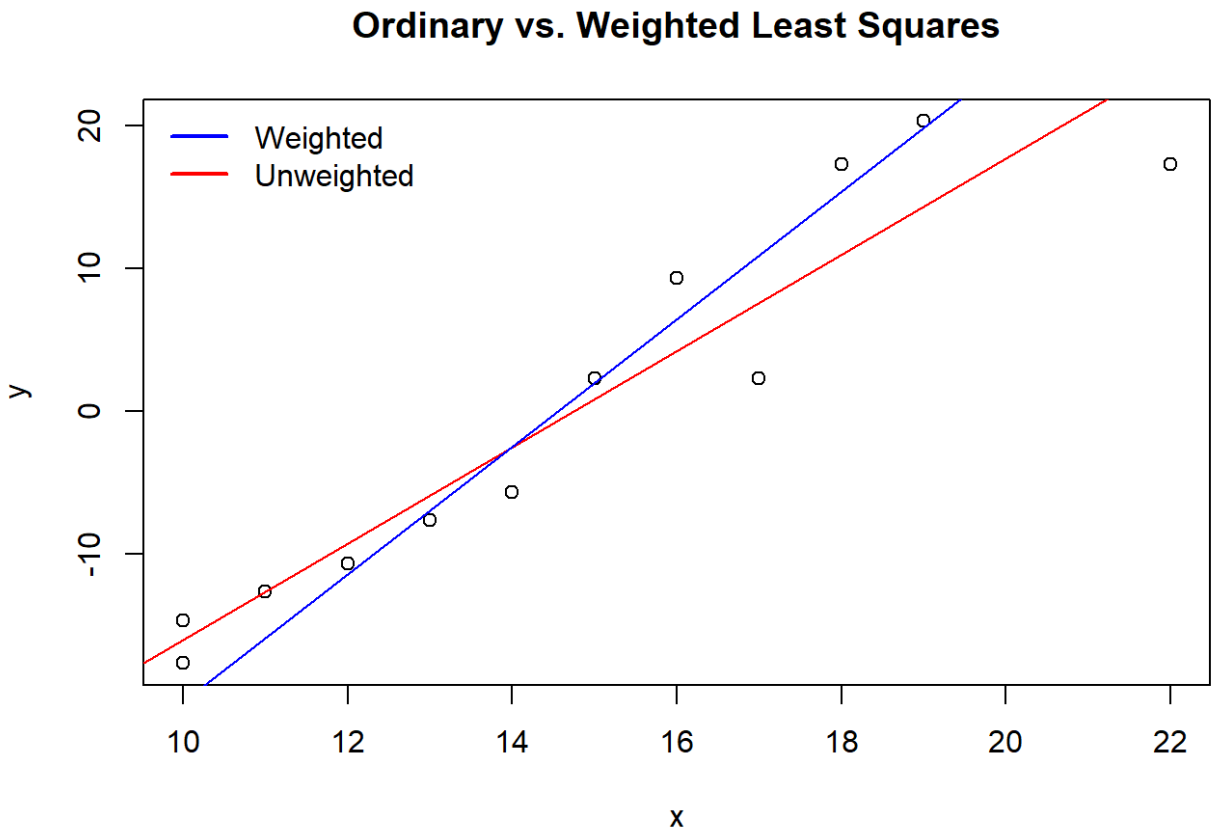
```
summary(olsfit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1311 -2.1051 -0.6998  2.3961  6.3665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -49.7725     5.4370  -9.154 3.55e-06 ***
## x             3.3744     0.3579   9.428 2.72e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.503 on 10 degrees of freedom
## Multiple R-squared:  0.8989, Adjusted R-squared:  0.8888
## F-statistic: 88.88 on 1 and 10 DF,  p-value: 2.721e-06
```

가중값을 구해서 적합한 모형과 원래의 모형의 요약정보를 비교해보면 가중값을 적용한 가중회귀모형의 회귀계수의 값이 조금 커진것을 확인할 수 있습니다.

8. 두모형을 산점도와 함께 각각의 회귀선을 그려봅니다.

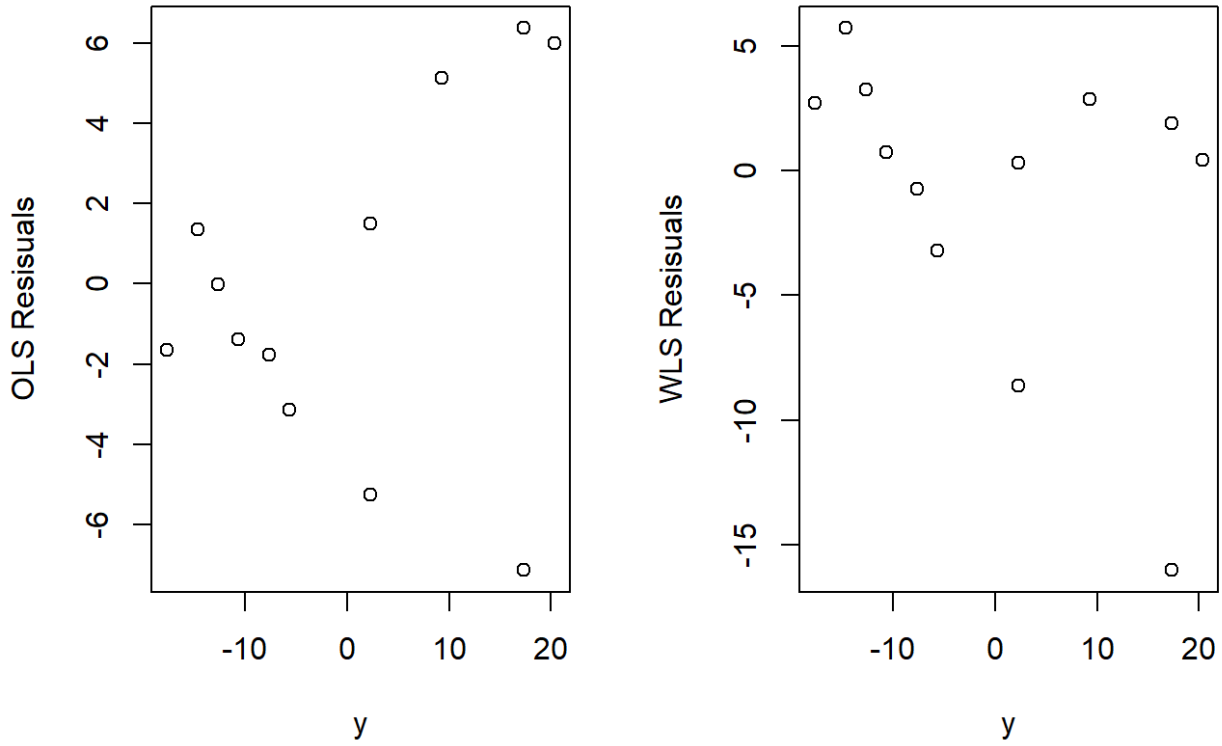
```
plot(x, y, main="Ordinary vs. Weighted Least Squares")
abline(olsfit, col="red")
abline(wlsfit, col="blue")
legend("topleft", c("Weighted", "Unweighted"), lty=c(1,1), lwd=c(2,2), col=c("blue", "red"), bty="n")
```



가중회귀선의 기울기가 더 커진것을 확인할 수 있습니다.

9. 두 회귀모형의 반응변수 y와 잔차의 산점도

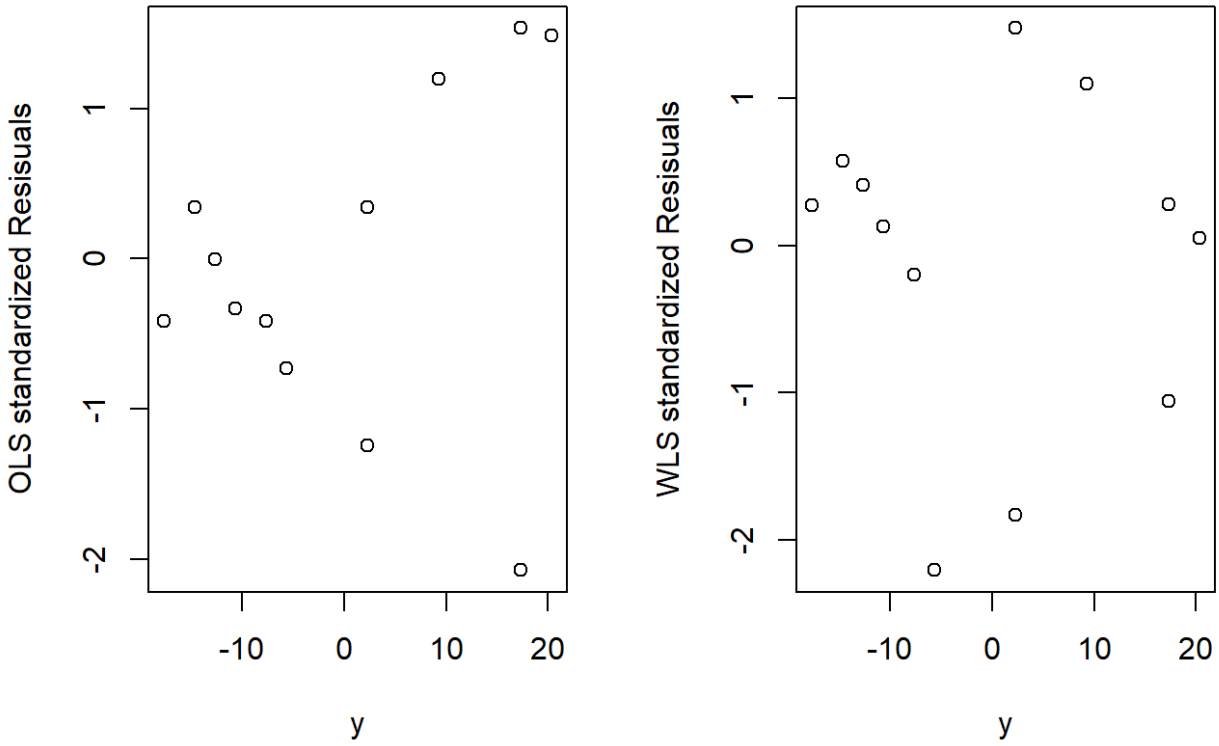
```
par(mfrow=c(1,2))
plot(y, olsfit$resid, ylab="OLS Resisuals")
plot(y, wlsfit$resid, ylab="WLS Resisuals")
```



두 회귀모형의 반응변수 y와 잔차의 산점도를 각각 그려보면 중회귀모형 좀 더 괜찮아 보이긴 합니다.

10. 두 회귀모형의 반응변수 y와 표준화 잔차의 산점도

```
par(mfrow=c(1,2))
plot(y, rstandard(olsfit), ylab="OLS standardized Resisuals")
plot(y, rstandard(wlsfit), ylab="WLS standardized Resisuals")
```



표준화 잔차의 산점도에서는 중회귀 모형의 잔차의 분포가 좀 더 개선된 것을 확인할 수 있습니다.

11. 가중회귀 모형을 스코어 검정으로 확인하기

```
library(car)
ncvTest(wlsfit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.5218856, Df = 1, p = 0.47004
```

Chisquare 값이 5.562254에서 0.5218856 많이 낮아 졌고
p-값은 0.018352에서 0.47004 으로 0.05보다 많이 커졌기 때문에 등분산이라고 할 수 있습니다.

결론

오차의 등분산성은 가중회귀모형으로 개선이 될 수 있음을 확인하였습니다.

하지만 테스트 과정에서 데이터가 테스트용으로 유효한 것인지 알기가 어려웠습니다.