Econometric\_HW3

### C1

1. Var(u|totwrk,educ,age,yngkid,male)  
   = Var(u|male) = a0+a1male  
   만일 여성이면 male = 0이고 var(u|male=0) = a0  
   만일 남성이면 male = 1이고 var(u|male=1) = a0+a1

library(wooldridge)  
data('sleep75')

model = lm(sleep~totwrk+educ+age+agesq+yngkid+male, data = sleep75)

male = sleep75$male == 1

sd(model$fitted.values[male])

## [1] 153.951

sd(model$fitted.values[!male])

## [1] 156.3502

여성의 분산이 더 높다.

1. 분산 검정.

x = model$fitted.values[!male]   
y = model$fitted.values[male]  
  
var.test(x, y, ratio=1, alternative="two.sided", conf.level=0.95)

##   
## F test to compare two variances  
##   
## data: x and y  
## F = 1.0314, num df = 305, denom df = 399, p-value = 0.7701  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.836326 1.275846  
## sample estimates:  
## ratio of variances   
## 1.031411

등분산 검정 결과 F값이 1.03로 상당히 낮고, p-value가 0.77로 높다. 따라서 신뢰수준 95%에서 귀무가설을 기각할 충분한 증거가 없으며 두 집단의 분산은 유의미하게 다르지 않다.

### C2

data("hprice1")  
library(sandwich)  
OLS = lm(price~lotsize+sqrft+bdrms, data = hprice1)  
  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

coeftest(OLS,vcov=vcovHC(OLS,type = 'HC0'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -21.7703081 36.2843444 -0.6000 0.55013   
## lotsize 0.0020677 0.0012227 1.6912 0.09451 .   
## sqrft 0.1227782 0.0173178 7.0897 3.883e-10 \*\*\*  
## bdrms 13.8525217 8.2836880 1.6723 0.09819 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(OLS)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.1770e+01 2.9475e+01 -0.7386 0.462208   
## lotsize 2.0677e-03 6.4213e-04 3.2201 0.001823 \*\*   
## sqrft 1.2278e-01 1.3237e-02 9.2751 1.658e-14 \*\*\*  
## bdrms 1.3853e+01 9.0101e+00 1.5374 0.127945   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

일반적인 모델과 비교했을 때, 가장 큰 차이가 나는 변수는 lotsize이다. 강건한 모델에서는 lotsize의 표준오차가 거의 2배이상 증가하였고, 이로 인해 t-value값이 줄어들어 강건한 모델에서는 lotsize를 유의미하다고 보기 어려워졌다.(sqrft는 두 모델에서 유의미하고 bdrms는 두 모델에서 모두 유의미하지 않다.)

OLS = lm(log(price)~log(lotsize)+log(sqrft)+bdrms, data = hprice1)  
coeftest(OLS,vcov=vcovHC(OLS,type = 'HC0'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.297042 0.763351 -1.6991 0.09299 .   
## log(lotsize) 0.167967 0.040520 4.1453 8.067e-05 \*\*\*  
## log(sqrft) 0.700232 0.101442 6.9028 9.014e-10 \*\*\*  
## bdrms 0.036958 0.029898 1.2362 0.21984   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(OLS)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.297042 0.651284 -1.9915 0.04967 \*   
## log(lotsize) 0.167967 0.038281 4.3877 3.307e-05 \*\*\*  
## log(sqrft) 0.700232 0.092865 7.5403 5.006e-11 \*\*\*  
## bdrms 0.036958 0.027531 1.3424 0.18308   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

여전히 견고한 모델의 표준오차가 더 높은 편이지만, 그 전과 달리 log를 한 모델에서는 lotsize변수가 강건한 모델에서도 유의미하다.

1. 독립변수에 대해 적절한 변환을 사용하면 이분산성을 완화할 수 있다.

### C3

yhat = OLS$fitted  
u2 = OLS$residuals^2  
Ru2 = summary(lm(u2 ~ yhat + I(yhat^2)))$r.squared  
LM = nrow(hprice1)\*Ru2  
p.value = 1-pchisq(LM, 2)  
print(LM)

## [1] 3.447287

p.value

## [1] 0.1784149

LM통계량이 3.447287 정도로 나오며, p-value가 0.178로 homoskedasticity assumption을 기각하기 어렵다. 따라서 8.18식에 따른 모델의 경우, 오차항의 등분산성을 가정할 수 있다.

### C4

data(vote1)  
OLS = lm(voteA~prtystrA+democA+log(expendA)+log(expendB), data = vote1)

u = OLS$residuals  
OLS2 = lm(u~prtystrA+democA+log(expendA)+log(expendB), data = vote1)  
summary(OLS2)$r.squared

## [1] 5.211387e-32

OLS를 통해 나온 잔차는 독립변수들이 span하는 열공간에 직교하기 때문에 잔차를 독립변수가 span하는 열공간에 projection하면 그 자신이 되기 때문에 설명되는 부분이 없고 R2값도 0이다. 다만 컴퓨터의 rounding error 때문에 정확히 0은 아니다.

bptest(voteA~prtystrA+democA+log(expendA)+log(expendB),~prtystrA+democA+log(expendA)+log(expendB),data=vote1)

##   
## studentized Breusch-Pagan test  
##   
## data: voteA ~ prtystrA + democA + log(expendA) + log(expendB)  
## BP = 9.0934, df = 4, p-value = 0.05881

test결과 유의확률이 0.058정도로 유의수준 5%이내에서 유의미하지 않다. 따라서 귀무가설을 기각할만한 충분한 근거고 등분산성을 가정할 수 있으나 유의확률이 유의수준에 거의 근접하기 때문에 유의수준을 어떻게 설정하는 지에 따라서 이분산성에 대한 유의미한 증거가 될 수 있다.

yhat = OLS$fitted  
u2 = OLS$residuals^2  
white = lm(u2~yhat+I(yhat^2))  
summary(white)

##   
## Call:  
## lm(formula = u2 ~ yhat + I(yhat^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -107.65 -44.80 -29.88 23.81 539.52   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 171.85840 53.14214 3.234 0.00147 \*\*  
## yhat -4.26368 2.16653 -1.968 0.05070 .   
## I(yhat^2) 0.03574 0.02124 1.682 0.09434 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 81.66 on 170 degrees of freedom  
## Multiple R-squared: 0.03173, Adjusted R-squared: 0.02034   
## F-statistic: 2.786 on 2 and 170 DF, p-value: 0.0645

F통계량(2,170)은 2.786이고 p-value는 0.0645정도로 (2)보다 유의확률이 더 높아졌다. 따라서 (2)와 마찬가지로 유의수준 5%에서는 기각이 어렵다.

### C8

data('GPA1')

## Warning in data("GPA1"): data set 'GPA1' not found

OLS = lm(colGPA~hsGPA+ACT+skipped+PC, data = gpa1)  
u = OLS$residuals  
summary(OLS)

##   
## Call:  
## lm(formula = colGPA ~ hsGPA + ACT + skipped + PC, data = gpa1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.84006 -0.20392 -0.03352 0.25346 0.74558   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.35651 0.32750 4.142 6.01e-05 \*\*\*  
## hsGPA 0.41295 0.09243 4.468 1.65e-05 \*\*\*  
## ACT 0.01334 0.01044 1.278 0.20353   
## skipped -0.07103 0.02625 -2.706 0.00768 \*\*   
## PC 0.12444 0.05731 2.171 0.03165 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3251 on 136 degrees of freedom  
## Multiple R-squared: 0.2593, Adjusted R-squared: 0.2375   
## F-statistic: 11.9 on 4 and 136 DF, p-value: 2.553e-08

yhat = OLS$fitted  
u2 = OLS$residuals^2  
white = lm(u2~yhat+I(yhat^2))  
summary(white)

##   
## Call:  
## lm(formula = u2 ~ yhat + I(yhat^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.13286 -0.07802 -0.04020 0.04954 0.60632   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.321838 2.005841 -0.160 0.873  
## yhat 0.129600 1.316763 0.098 0.922  
## I(yhat^2) 0.002946 0.215660 0.014 0.989  
##   
## Residual standard error: 0.1237 on 138 degrees of freedom  
## Multiple R-squared: 0.04934, Adjusted R-squared: 0.03557   
## F-statistic: 3.581 on 2 and 138 DF, p-value: 0.03045

F 통계량 값이 3.581 정도이고 유의확률은 0.03정도로 유의수준 5%에서 귀무가설을 기각할 수 있는 충분한 증거가 있다. 따라서 이분산성이 존재한다고 할 수 있다.

hhat = white$fitted  
  
min(hhat)

## [1] 0.02738136

최소값이 0.027이므로 모든 값이 양수이다.

WLS = lm(colGPA~hsGPA+ACT+skipped+PC, data = gpa1,,weights=(1/hhat))

summary(WLS)

##   
## Call:  
## lm(formula = colGPA ~ hsGPA + ACT + skipped + PC, data = gpa1,   
## weights = (1/hhat))  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -2.6994 -0.6892 -0.1191 0.7963 2.5098   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.401564 0.298430 4.696 6.39e-06 \*\*\*  
## hsGPA 0.402506 0.083362 4.828 3.65e-06 \*\*\*  
## ACT 0.013162 0.009827 1.339 0.182698   
## skipped -0.076365 0.022173 -3.444 0.000762 \*\*\*  
## PC 0.126005 0.056339 2.237 0.026945 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.013 on 136 degrees of freedom  
## Multiple R-squared: 0.3062, Adjusted R-squared: 0.2858   
## F-statistic: 15.01 on 4 and 136 DF, p-value: 3.488e-10

PC 변수의 회귀계수 값이 조금 높아졌지만 큰 차이는 없다. 두 모델의 t 통계량이 거의 유사하다. R2의 경우, WLS에서 0.04정도 상승해지면, 전반적으로 큰 변화는 아니어서 두 모델이 유의미하게 다르다고 말하기 어렵다.

coeftest(WLS,vcov=vcovHC(WLS,type = 'HC0'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.401564 0.305039 4.5947 9.782e-06 \*\*\*  
## hsGPA 0.402506 0.084790 4.7471 5.162e-06 \*\*\*  
## ACT 0.013162 0.010228 1.2869 0.2003075   
## skipped -0.076365 0.020830 -3.6662 0.0003521 \*\*\*  
## PC 0.126005 0.057969 2.1737 0.0314608 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(WLS)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.4015638 0.2984304 4.6965 6.394e-06 \*\*\*  
## hsGPA 0.4025059 0.0833617 4.8284 3.651e-06 \*\*\*  
## ACT 0.0131620 0.0098273 1.3393 0.1826985   
## skipped -0.0763653 0.0221727 -3.4441 0.0007621 \*\*\*  
## PC 0.1260048 0.0563391 2.2365 0.0269446 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

robust standard errors를 사용했을 때와 그렇지 않았을 때의 결과가 크게 다르지 않다. robust standard errors를 사용했을 때, 전반적으로 표준오차가 커진 경향이 있지만 매우 작아서 두 모델의 결과를 유의미하게 바꾸지 않는다.

### C9

data("smoke")  
  
OLS = lm(cigs~log(income)+log(cigpric)+educ+age+agesq+restaurn, data = smoke)  
  
summary(OLS)

##   
## Call:  
## lm(formula = cigs ~ log(income) + log(cigpric) + educ + age +   
## agesq + restaurn, data = smoke)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.819 -9.381 -5.975 7.922 70.221   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.639826 24.078659 -0.151 0.87988   
## log(income) 0.880268 0.727783 1.210 0.22682   
## log(cigpric) -0.750862 5.773342 -0.130 0.89655   
## educ -0.501498 0.167077 -3.002 0.00277 \*\*   
## age 0.770694 0.160122 4.813 1.78e-06 \*\*\*  
## agesq -0.009023 0.001743 -5.176 2.86e-07 \*\*\*  
## restaurn -2.825085 1.111794 -2.541 0.01124 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.4 on 800 degrees of freedom  
## Multiple R-squared: 0.05274, Adjusted R-squared: 0.04563   
## F-statistic: 7.423 on 6 and 800 DF, p-value: 9.499e-08

OLS 추정결과, F통계량이 유의미하고 educ,age,agesq,restaurn 변수가 유의미하게 나온다. 하지만 R2 값 자체는 매우 낮은 편이다.

RegModel = cigs~log(income)+log(cigpric)+educ+age+agesq+restaurn  
OLS = lm(RegModel,data = smoke)  
smoke$e = OLS$residuals  
AUX = lm(update(OLS,I(log(e^2))~.), data = smoke)  
smoke$ghat = AUX$fitted  
smoke$hhat = exp(smoke$ghat)  
FGLS = lm(RegModel,data=smoke,weights = 1/hhat)  
summary(FGLS)

##   
## Call:  
## lm(formula = RegModel, data = smoke, weights = 1/hhat)  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -1.9036 -0.9532 -0.8099 0.8415 9.8556   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.6354627 17.8031394 0.317 0.751673   
## log(income) 1.2952393 0.4370117 2.964 0.003128 \*\*   
## log(cigpric) -2.9403117 4.4601446 -0.659 0.509931   
## educ -0.4634464 0.1201587 -3.857 0.000124 \*\*\*  
## age 0.4819480 0.0968082 4.978 7.86e-07 \*\*\*  
## agesq -0.0056272 0.0009395 -5.990 3.17e-09 \*\*\*  
## restaurn -3.4610640 0.7955050 -4.351 1.53e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.579 on 800 degrees of freedom  
## Multiple R-squared: 0.1134, Adjusted R-squared: 0.1068   
## F-statistic: 17.06 on 6 and 800 DF, p-value: < 2.2e-16

head(FGLS$residuals)

## 1 2 3 4 5 6   
## -9.246794 -9.914233 -7.527828 -9.667242 -8.440092 -2.048961

head(OLS$residuals)

## 1 2 3 4 5 6   
## -10.332992 -10.754790 -7.719894 -10.260642 -7.504153 2.439415

두 잔차간의 차이가 있긴 하지만, 그렇게 크진 않다.

uhat = FGLS$residuals  
yhat = FGLS$fitted

u2 = uhat/sqrt(hhat)

## Warning in uhat/sqrt(hhat): 두 객체의 길이가 서로 배수관계에 있지 않습니다

y2 = yhat/sqrt(hhat)

## Warning in yhat/sqrt(hhat): 두 객체의 길이가 서로 배수관계에 있지 않습니다

white = lm((u2)^2~y2+I(y2^2))  
summary(white)

##   
## Call:  
## lm(formula = (u2)^2 ~ y2 + I(y2^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7361 -1548 -968 -546 62142   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 682.677 890.185 0.767 0.443   
## y2 -35.737 63.975 -0.559 0.577   
## I(y2^2) 2.942 1.140 2.581 0.010 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5087 on 804 degrees of freedom  
## Multiple R-squared: 0.05745, Adjusted R-squared: 0.05511   
## F-statistic: 24.5 on 2 and 804 DF, p-value: 4.675e-11

R2 값은 0.027로 낮은 수치지만, F-통계량이 상당히 높고 p-value가 0에 수렴한다. 특히 I(y2^2)변수가 유의미한 값이 나온다. 이를 통해 오차와 독립변수의 함수관계가 없다는 귀무가설을 기각할 만한 충분한 증거를 확인할 수 있으며, 오차와 독립변수 사이에 함수관계가 있다고 볼 수 있다. 또한 오차와 독립변수 사이에 함수 관계가 있다면 이를 이분산성으로 간즈 할 수 있다.

3번 문제에서 확인할 수 있듯이 FGSL가 이분산성을 제거하지 않는다는 것을 알 수 있다.따라서 일반적인 검정에 사용되는 통계량을 그대로 사용할 수 없고 지금 출력된 그 결과도 유효하지 않다. (검정통계량들이 등분산을 가정하고 만들어졌기 때문에 이분산성이 있다면 이를 조정한 통계량을 사용해야한다.)

coeftest(FGLS, vcov = vcovHC(FGLS,type = 'HC0'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.6354627 37.1611619 0.1516 0.879502   
## log(income) 1.2952393 0.5327686 2.4311 0.015270 \*   
## log(cigpric) -2.9403117 8.9314572 -0.3292 0.742084   
## educ -0.4634464 0.1484143 -3.1227 0.001857 \*\*   
## age 0.4819480 0.1144917 4.2095 2.850e-05 \*\*\*  
## agesq -0.0056272 0.0011719 -4.8018 1.877e-06 \*\*\*  
## restaurn -3.4610640 0.7127922 -4.8556 1.443e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(FGLS)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.63546270 17.80313936 0.3165 0.751673   
## log(income) 1.29523934 0.43701172 2.9639 0.003128 \*\*   
## log(cigpric) -2.94031167 4.46014462 -0.6592 0.509931   
## educ -0.46344636 0.12015869 -3.8570 0.000124 \*\*\*  
## age 0.48194797 0.09680824 4.9784 7.856e-07 \*\*\*  
## agesq -0.00562721 0.00093948 -5.9897 3.175e-09 \*\*\*  
## restaurn -3.46106399 0.79550504 -4.3508 1.532e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

통계량을 보정한 결과, 해석을 유의미하게 바꿀 변화는 관측되지 않는다. 하지만 restaurn를 제외하고 모든 추정량의 표준오차가 늘어난 것을 확인할 수 있다. 특히 log(cigpric)은 두배 가까이 표준오차가 증가했다.

### C12

data("meap00\_01")  
OLS = lm(math4~lunch+log(enroll)+log(exppp), data = meap00\_01)  
coeftest(OLS)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.932404 19.961698 4.6054 4.424e-06 \*\*\*  
## lunch -0.448743 0.014642 -30.6476 < 2.2e-16 \*\*\*  
## log(enroll) -5.399152 0.940412 -5.7413 1.112e-08 \*\*\*  
## log(exppp) 3.524751 2.097846 1.6802 0.09311 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(OLS, vcov = vcovHC(OLS,type = 'HC0'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.932404 23.059797 3.9867 6.986e-05 \*\*\*  
## lunch -0.448743 0.016565 -27.0903 < 2.2e-16 \*\*\*  
## log(enroll) -5.399152 1.129837 -4.7787 1.917e-06 \*\*\*  
## log(exppp) 3.524751 2.350953 1.4993 0.134   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(OLS, vcov = vcovHC(OLS,type = 'HC1'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.932404 23.087103 3.9820 7.125e-05 \*\*\*  
## lunch -0.448743 0.016584 -27.0583 < 2.2e-16 \*\*\*  
## log(enroll) -5.399152 1.131175 -4.7730 1.971e-06 \*\*\*  
## log(exppp) 3.524751 2.353737 1.4975 0.1344   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(OLS, vcov = vcovHC(OLS,type = 'HC2'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.932404 23.131338 3.9744 7.354e-05 \*\*\*  
## lunch -0.448743 0.016601 -27.0318 < 2.2e-16 \*\*\*  
## log(enroll) -5.399152 1.132304 -4.7683 2.018e-06 \*\*\*  
## log(exppp) 3.524751 2.359103 1.4941 0.1353   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(OLS, vcov = vcovHC(OLS,type = 'HC3'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.932404 23.203501 3.9620 7.742e-05 \*\*\*  
## lunch -0.448743 0.016637 -26.9732 < 2.2e-16 \*\*\*  
## log(enroll) -5.399152 1.134779 -4.7579 2.123e-06 \*\*\*  
## log(exppp) 3.524751 2.367336 1.4889 0.1367   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(OLS, vcov = vcovHC(OLS,type = 'HC4'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.932404 23.268244 3.9510 8.104e-05 \*\*\*  
## lunch -0.448743 0.016646 -26.9588 < 2.2e-16 \*\*\*  
## log(enroll) -5.399152 1.135214 -4.7561 2.142e-06 \*\*\*  
## log(exppp) 3.524751 2.376165 1.4834 0.1382   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

어떠한 경우라도 heteroskedasticity-robust한 모델의 표준오차가 OLS 모델보다 언제나 다소 크다.

yhat = OLS$fitted  
u2 = OLS$residuals^2  
white = lm(u2~yhat+I(yhat^2))  
summary(white)

##   
## Call:  
## lm(formula = u2 ~ yhat + I(yhat^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -593.52 -163.92 -72.42 52.12 2735.87   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1664.12800 315.21436 5.279 1.46e-07 \*\*\*  
## yhat -28.29903 9.21685 -3.070 0.00217 \*\*   
## I(yhat^2) 0.11553 0.06591 1.753 0.07982 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 365.1 on 1689 degrees of freedom  
## Multiple R-squared: 0.1358, Adjusted R-squared: 0.1348   
## F-statistic: 132.7 on 2 and 1689 DF, p-value: < 2.2e-16

F 통계량이 132.7로 상당히 높고 유의확률이 0에 수렴한다. 따라서 이분산성에 대한 강한 증거가 존재한다고 볼 수 있다.

math4hat = OLS$fitted  
uhat = OLS$residuals  
  
ghat = lm(log(uhat^2)~math4hat+I(math4hat^2))$fitted  
hhat =exp(ghat)  
  
WLS = lm(math4~lunch+log(enroll)+log(exppp),data = meap00\_01,weights = 1/hhat)  
coeftest(WLS)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 50.478152 16.510316 3.0574 0.0022679 \*\*   
## lunch -0.448592 0.014613 -30.6971 < 2.2e-16 \*\*\*  
## log(enroll) -2.647282 0.835936 -3.1668 0.0015686 \*\*   
## log(exppp) 6.474200 1.685877 3.8403 0.0001274 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(OLS)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.932404 19.961698 4.6054 4.424e-06 \*\*\*  
## lunch -0.448743 0.014642 -30.6476 < 2.2e-16 \*\*\*  
## log(enroll) -5.399152 0.940412 -5.7413 1.112e-08 \*\*\*  
## log(exppp) 3.524751 2.097846 1.6802 0.09311 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lunch 변수는 두 모델 모두 유의미하고 계수의 크기도 거의 유사하다. log(enroll)은 OLS,WLS모두 유의미하지만 WLS 2.7 정도 증가했다. 가장 큰 변화는 log(exppp)인데, OLS 추정에는 유의미하지 않지만, WLS 추정에서는 유의미해진다.

coeftest(WLS, vcov = vcovHC(WLS,type = 'HC0'))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 50.478152 18.902796 2.6704 0.0076487 \*\*   
## lunch -0.448592 0.014237 -31.5084 < 2.2e-16 \*\*\*  
## log(enroll) -2.647282 1.053361 -2.5132 0.0120574 \*   
## log(exppp) 6.474200 1.810724 3.5755 0.0003594 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(3)에서 WLS추정에서 유의미한 결과를 강건한 표준오차를 기반한 통계량으로 다시 계산해볼 필요가 있다. 가장 특징적인 변화를 보였던 log(exppp)변수의 경우, 표준오차값이 다소 증가했지만 여전히 유의미하다. 다른 변수도 표준오차 값이 전보다 증가하거나 비슷하지만 여전히 유의미하다.

1. WLS 추정이 OLS보다 정확하다. WLS 추정의 log(exppp)의 강건한 표준오차는 1.8로 OLS의 경우(2.09) 보다 작다. 추정량의 표준오차가 더 작을수록 effiecncy가 좋은 추정이라고 할 수 있으며, t값도 훨씬 줄어들기 때문에 보다 유의미해진다.

### CHAPTER 10

### C7 - (1)

data("consump")  
  
OLS = lm(consump$gc~consump$gy)  
summary(OLS)

##   
## Call:  
## lm(formula = consump$gc ~ consump$gy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0140496 -0.0035407 -0.0005813 0.0044080 0.0116890   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.008079 0.001899 4.254 0.000155 \*\*\*  
## consump$gy 0.570781 0.067354 8.474 6.75e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.007268 on 34 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.6787, Adjusted R-squared: 0.6692   
## F-statistic: 71.81 on 1 and 34 DF, p-value: 6.754e-10

gc = 0.008079 + 0.570781gy n = 35 R2 = 0.6787  
회귀 분석 결과, 소득의 상승률이 1퍼센트 포인트 오르다면 소비 상승률이 0.571퍼센트 포인트 가량 오른다고 볼 수 있다.  
gy변수의 t값이 8.5정도롤 높기 떄문에 추정량이 유의미하다.

### CHAPTER 12

### C6 - (1)