Econometric\_HW2

library(wooldridge)  
data("vote1")  
library(car)

## Loading required package: carData

#### C1

1. What is the interpretation of b1? expendA가 1%만큼 증가하면, voteA가 b1/100만큼 변화한다. 단 본 문제에서 voteA의 단위가 %이기 때문에, b1/100 %포인트 만큼 변화했다고 할 수 있다.
2. In terms of the parameters, state the null hypothesis that a 1% increase in A’s ex???penditures is offset by a 1% increase in B’s expenditures.

귀무가설은 아래와 같다. H0: b1 = -b2

1. Estimate the given model using the data in VOTE1.RAW and report the results in usual form. Do A’s expenditures affect the outcome? What about B’s expendi???tures? Can you use these results to test the hypothesis in part (ii)?

model = lm(voteA~log(expendA)+log(expendB)+prtystrA, data = vote1)  
summary(model)

##   
## Call:  
## lm(formula = voteA ~ log(expendA) + log(expendB) + prtystrA,   
## data = vote1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.3968 -5.4174 -0.8679 4.9551 26.0660   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.07893 3.92631 11.48 <2e-16 \*\*\*  
## log(expendA) 6.08332 0.38215 15.92 <2e-16 \*\*\*  
## log(expendB) -6.61542 0.37882 -17.46 <2e-16 \*\*\*  
## prtystrA 0.15196 0.06202 2.45 0.0153 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.712 on 169 degrees of freedom  
## Multiple R-squared: 0.7926, Adjusted R-squared: 0.7889   
## F-statistic: 215.2 on 3 and 169 DF, p-value: < 2.2e-16

voteA = 45.08+6.083log(expendA)-6.615log(expendB)+0.152 prtystrA n = 173, R2 = 0.793

결과를 보면, 두 변수의 t-test결과 모두 유의미하게 나온다. 또한 두 변수의 부호가 반대일 뿐 유사한 값을 갖는다. 따라서 두 변수 모두 결과에 영향을 미친다고 할 수 있다. 또한 두 변수의 절대값이 비슷하기 때문에, 귀무가설을 기각하지 않을 수 있고 의심해볼 수 있다. 하지만, 이 결과를 통해서는 각각의 변수가 유의미하다는 것만 확인할 수 있을 뿐이다. 따라서 2번 문제에서 제시한 가설을 테스트하기 위해서는 다른 작업이 필요하다.

1. Estimate a model that directly gives the t statistic for testing the hypothesis in part (ii). What do you conclude? (Use a two-sided alternative.)

x1 = log(vote1$expendA)  
x2 = log(vote1$expendB)-log(vote1$expendA)  
model2 = lm(voteA~x1+x2+ prtystrA, data = vote1)  
summary(model2)

##   
## Call:  
## lm(formula = voteA ~ x1 + x2 + prtystrA, data = vote1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.3968 -5.4174 -0.8679 4.9551 26.0660   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.07893 3.92631 11.481 <2e-16 \*\*\*  
## x1 -0.53210 0.53309 -0.998 0.3196   
## x2 -6.61542 0.37882 -17.463 <2e-16 \*\*\*  
## prtystrA 0.15196 0.06202 2.450 0.0153 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.712 on 169 degrees of freedom  
## Multiple R-squared: 0.7926, Adjusted R-squared: 0.7889   
## F-statistic: 215.2 on 3 and 169 DF, p-value: < 2.2e-16

ii에서 주장한 귀무가설은 b1+b2 = 0 과 동일하다. b1+b2를 B1이라고 두고 이를 포함하는 회귀식을 만든 결과, p-value 값이 0.3이상으로 높은 수준이다. 따라서 귀무가설을 기각할 수 없기 때문에, b1 = -b2라고 주장할 수 있다.

#### C3

1. You are interested in estimating and obtaining a confidence interval for the percent???age change in price when a 150-square-foot bedroom is added to a house. In decimal form, this is u1 5 150b1 1 b2. Use the data in HPRICE1.RAW to estimate u1.

lm1 = lm(log(hprice1$price)~hprice1$sqrft+hprice1$bdrms)  
summary(lm1)

##   
## Call:  
## lm(formula = log(hprice1$price) ~ hprice1$sqrft + hprice1$bdrms)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.75448 -0.12322 -0.01993 0.11938 0.62948   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.766e+00 9.704e-02 49.112 < 2e-16 \*\*\*  
## hprice1$sqrft 3.794e-04 4.321e-05 8.781 1.5e-13 \*\*\*  
## hprice1$bdrms 2.888e-02 2.964e-02 0.974 0.333   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1971 on 85 degrees of freedom  
## Multiple R-squared: 0.5883, Adjusted R-squared: 0.5786   
## F-statistic: 60.73 on 2 and 85 DF, p-value: < 2.2e-16

b1 = 0.000379 b2 = 0.0296 B1 = 150b1+b2이므로 B1 = 0.086 따라서 150평방미터의 침실이 추가된다면, 집가격은 8.6% 상승한다.

1. Write b2 in terms of B1 and b1 and plug this into the log(price) equation.

b2 = B1-150b1

log(price) = b0+b1(sqrft-150bdrms)+B1bdrms+u

1. Use part (ii) to obtain a standard error for ??B1 and use this standard error to construct a 95% confidence interval.

x1 = hprice1$sqrft-150\*hprice1$bdrms  
x2 = hprice1$bdrms  
  
lm2 = lm(log(hprice1$price)~x1+x2)  
summary(lm2)

##   
## Call:  
## lm(formula = log(hprice1$price) ~ x1 + x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.75448 -0.12322 -0.01993 0.11938 0.62948   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.766e+00 9.704e-02 49.112 < 2e-16 \*\*\*  
## x1 3.794e-04 4.321e-05 8.781 1.5e-13 \*\*\*  
## x2 8.580e-02 2.677e-02 3.205 0.0019 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1971 on 85 degrees of freedom  
## Multiple R-squared: 0.5883, Adjusted R-squared: 0.5786   
## F-statistic: 60.73 on 2 and 85 DF, p-value: < 2.2e-16

se(B1)은 0.0268정도다.

confint(lm2)

## 2.5 % 97.5 %  
## (Intercept) 4.5730767914 4.9589776356  
## x1 0.0002935289 0.0004653631  
## x2 0.0325803713 0.1390223615

신뢰구간은 (0.0326,0.139) 이다.

#### C5

data("mlb1")  
attach(mlb1)

Use the model estimated in equation (4.31) and drop the variable rbisyr. What happens to the statistical significance of hrunsyr? What about the size of the coefficient on hrunsyr?

lm1 = lm(log(salary)~years+gamesyr+bavg+hrunsyr+rbisyr)  
lm2 = lm(log(salary)~years+gamesyr+bavg+hrunsyr)  
  
print(summary(lm1))

##   
## Call:  
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr +   
## rbisyr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.02508 -0.45034 -0.04013 0.47014 2.68924   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.119e+01 2.888e-01 38.752 < 2e-16 \*\*\*  
## years 6.886e-02 1.211e-02 5.684 2.79e-08 \*\*\*  
## gamesyr 1.255e-02 2.647e-03 4.742 3.09e-06 \*\*\*  
## bavg 9.786e-04 1.104e-03 0.887 0.376   
## hrunsyr 1.443e-02 1.606e-02 0.899 0.369   
## rbisyr 1.077e-02 7.175e-03 1.500 0.134   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7266 on 347 degrees of freedom  
## Multiple R-squared: 0.6278, Adjusted R-squared: 0.6224   
## F-statistic: 117.1 on 5 and 347 DF, p-value: < 2.2e-16

print(summary(lm2))

##   
## Call:  
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0642 -0.4614 -0.0271 0.4654 2.7216   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.020913 0.265719 41.476 < 2e-16 \*\*\*  
## years 0.067732 0.012113 5.592 4.55e-08 \*\*\*  
## gamesyr 0.015759 0.001564 10.079 < 2e-16 \*\*\*  
## bavg 0.001419 0.001066 1.331 0.184   
## hrunsyr 0.035943 0.007241 4.964 1.08e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7279 on 348 degrees of freedom  
## Multiple R-squared: 0.6254, Adjusted R-squared: 0.6211   
## F-statistic: 145.2 on 4 and 348 DF, p-value: < 2.2e-16

rbisyr을 제외하고 회귀를 한 결과, hrunsyr의 p-value 값이 굉장히 낮아지고, 유의해졌다. 또한 회귀계수가 두배 이상 증가했다.

1. Add the variables runsyr (runs per year), fldperc (fielding percentage), and sbasesyr (stolen bases per year) to the model from part (i). Which of these factors are individually significant?

lm3 = lm(log(salary)~years+gamesyr+bavg+hrunsyr+fldperc+runsyr+sbasesyr)  
  
summary(lm3)

##   
## Call:  
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr +   
## fldperc + runsyr + sbasesyr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.11554 -0.44557 -0.08808 0.48731 2.57872   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.4082678 2.0032546 5.196 3.50e-07 \*\*\*  
## years 0.0699848 0.0119756 5.844 1.18e-08 \*\*\*  
## gamesyr 0.0078995 0.0026775 2.950 0.003391 \*\*   
## bavg 0.0005296 0.0011038 0.480 0.631656   
## hrunsyr 0.0232106 0.0086392 2.687 0.007566 \*\*   
## fldperc 0.0010351 0.0020046 0.516 0.605936   
## runsyr 0.0173922 0.0050641 3.434 0.000666 \*\*\*  
## sbasesyr -0.0064191 0.0051842 -1.238 0.216479   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7176 on 345 degrees of freedom  
## Multiple R-squared: 0.639, Adjusted R-squared: 0.6317   
## F-statistic: 87.25 on 7 and 345 DF, p-value: < 2.2e-16

새롭게 넣은 3개의 변수중에, runsyr만 충분히 큰 t 값을 갖고 유의미하다. 나머지 두 개는 귀무가설을 기각하기에 충분하지 않다.

In the model from part (ii), test the joint significance of bavg, fldperc, and sbasesyr.

lm4 = lm(log(salary)~years+gamesyr+hrunsyr+runsyr)  
anova(lm4,lm3)

## Analysis of Variance Table  
##   
## Model 1: log(salary) ~ years + gamesyr + hrunsyr + runsyr  
## Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + fldperc + runsyr +   
## sbasesyr  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 348 178.72   
## 2 345 177.66 3 1.0583 0.685 0.5617

detach(mlb1)

t-test값을 통해 확인 할 수 있듯이, 세 변수는 개별적으로 유의미하지 않다. 또한 세변수의 joint된 영향은 anova 결과에서 알 수 있듯이, p-value 값이 0.56으로 굉장히 높다.따라서 귀무가설을 기각하기 어렵고, 세 변수의 joint가 유의미한 영향을 미친다고 할 수 없다.

#### C6

1. Consider the standard wage equation log(wage) = b0 + b1educ + b2exper + b3tenure + u.

State the null hypothesis that another year of general workforce experience has the same effect on log(wage) as another year of tenure with the current employer.

귀무가설은 b2 = b3이다.

Test the null hypothesis in part (i) against a two-sided alternative, at the 5% signifi???cance level, by constructing a 95% confidence interval. What do you conclude?

B1 = b2-b3로 두고 회귀식을 정리하면, 아래와 같다.  
log(wage) = b0 + b1educ+B1exper+b3(exper+tenure)+u

attach(wage2)  
library(car)  
lmR = lm(log(wage)~educ+exper+I(exper+tenure))  
confint(lmR)

## 2.5 % 97.5 %  
## (Intercept) 5.279782328 5.71360900  
## educ 0.062082990 0.08764456  
## exper -0.007355358 0.01126271  
## I(exper + tenure) 0.008297389 0.01845221

B1의 95% 신뢰구간은 (-0.007355358,0.01126271)인데 신뢰구간에 0이 포함되므로 귀무가설을 기각할 수 없다. 따라서 b2와 b3는 유의수준 5%에서 같다고 할 수 있다.

#### C7

Refer to the example used in Section 4.4. You will use the data set TWOYEAR.RAW. (i)

The variable phsrank is the person’s high school percentile. (A higher number is better. For example, 90 means you are ranked better than 90 percent of your graduating class.) Find the smallest, largest, and average phsrank in the sample.

data('twoyear')  
summary(twoyear$phsrank)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 44.00 50.00 56.16 76.00 99.00

최소값은 0이고 최대값은 99이고 평균은 56.16이다.

Add phsrank to equation (4.26) and report the OLS estimates in the usual form. Is phsrank statistically significant? How much is 10 percentage points of high school rank worth in terms of wage?

twoyear = na.omit(twoyear)  
lm1 = lm(lwage~jc+totcoll+exper+phsrank, data = twoyear)  
summary(lm1)

##   
## Call:  
## lm(formula = lwage ~ jc + totcoll + exper + phsrank, data = twoyear)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.09049 -0.28135 0.00538 0.28543 1.79060   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.4587472 0.0236211 61.756 <2e-16 \*\*\*  
## jc -0.0093108 0.0069693 -1.336 0.182   
## totcoll 0.0754756 0.0025588 29.496 <2e-16 \*\*\*  
## exper 0.0049396 0.0001575 31.360 <2e-16 \*\*\*  
## phsrank 0.0003032 0.0002389 1.269 0.204   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4301 on 6758 degrees of freedom  
## Multiple R-squared: 0.2226, Adjusted R-squared: 0.2222   
## F-statistic: 483.8 on 4 and 6758 DF, p-value: < 2.2e-16

log(wage) = 1.46 - 0.0093jc + 0.0754totcoll + 0.00493exper + 0.0003032phsrank  
n = 6763, R^2 = 0.223  
phrank의 t 값(1.269)이 충분히 낮은 p-value을 가질 만큼 유의미하지 않다. 만약 phrank가 10 증가한다면(단위가 %이므로 10%), 종속변수는 0.003증가하는데, 종속변수가 log scale되어 있으므로 wage가 0.3% 증가한다고 할 수 있다.

Does adding phsrank to (4.26) substantively change the conclusions on the returns to two- and four-year colleges? Explain.

lm2 = lm(lwage~jc+totcoll+exper, data = twoyear)  
  
summary(lm2)

##   
## Call:  
## lm(formula = lwage ~ jc + totcoll + exper, data = twoyear)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.10362 -0.28132 0.00551 0.28518 1.78167   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.4723256 0.0210602 69.910 <2e-16 \*\*\*  
## jc -0.0101795 0.0069359 -1.468 0.142   
## totcoll 0.0768762 0.0023087 33.298 <2e-16 \*\*\*  
## exper 0.0049442 0.0001575 31.397 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4301 on 6759 degrees of freedom  
## Multiple R-squared: 0.2224, Adjusted R-squared: 0.2221   
## F-statistic: 644.5 on 3 and 6759 DF, p-value: < 2.2e-16

summary(lm1)

##   
## Call:  
## lm(formula = lwage ~ jc + totcoll + exper + phsrank, data = twoyear)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.09049 -0.28135 0.00538 0.28543 1.79060   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.4587472 0.0236211 61.756 <2e-16 \*\*\*  
## jc -0.0093108 0.0069693 -1.336 0.182   
## totcoll 0.0754756 0.0025588 29.496 <2e-16 \*\*\*  
## exper 0.0049396 0.0001575 31.360 <2e-16 \*\*\*  
## phsrank 0.0003032 0.0002389 1.269 0.204   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4301 on 6758 degrees of freedom  
## Multiple R-squared: 0.2226, Adjusted R-squared: 0.2222   
## F-statistic: 483.8 on 4 and 6758 DF, p-value: < 2.2e-16

phrank를 추가하므로써 다른 변수들의 회귀계수 값이나 p-value가 바꾼 부분도 있으나 검정 결과를 바꿀 만큼 큰 변화는 없다.

The data set contains a variable called id. Explain why if you add id to equation (4.17) or (4.26) you expect it to be statistically insignificant. What is the two-sided p-value?

lm3 = lm(lwage~jc+totcoll+exper+id, data = twoyear)  
summary(lm3)$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.467533e+00 2.283058e-02 64.279273 0.000000e+00  
## jc -1.021802e-02 6.936629e-03 -1.473053 1.407833e-01  
## totcoll 7.688128e-02 2.308868e-03 33.298260 2.956450e-225  
## exper 4.945596e-03 1.575019e-04 31.400223 3.809016e-202  
## id 1.139082e-07 2.094424e-07 0.543864 5.865530e-01

id는 개인을 구별하기위해 랜덤하게 주어진 인식번호이므로 어떠한 변수와도 상관이 없다. p-value 값(0.586)도 상당히 높은 편이므로 유의미하지 않다는 것을 쉽게 확인할 수 있다.

#### C8

1. How many single-person households are there in the data set?

data(k401ksubs)  
head(k401ksubs)

## e401k inc marr male age fsize nettfa p401k pira incsq agesq  
## 1 0 13.170 0 0 40 1 4.575 0 1 173.4489 1600  
## 2 1 61.230 0 1 35 1 154.000 1 0 3749.1128 1225  
## 3 0 12.858 1 0 44 2 0.000 0 0 165.3282 1936  
## 4 0 98.880 1 1 44 2 21.800 0 0 9777.2539 1936  
## 5 0 22.614 0 0 53 1 18.450 0 0 511.3930 2809  
## 6 0 15.000 1 0 60 3 0.000 0 0 225.0000 3600

single = k401ksubs[k401ksubs$fsize == 1,]  
dim(single)

## [1] 2017 11

1인가구는 2017개 있다.

1. Use OLS to estimate the model nettfa = b0 + b1inc + b2age + u,

and report the results using the usual format. Be sure to use only the single-person households in the sample. Interpret the slope coefficients. Are there any surprises in the slope estimates?

lm1 = lm(nettfa~inc+age, data = single)  
coef(lm1)

## (Intercept) inc age   
## -43.0398119 0.7993167 0.8426563

분석 결과에 따르면, 나이가 고정된 상태에서 income이 1달러 늘어나면, netfa는 0.8달러 증가한다.(단위가 동일하므로) 그리고 income이 고정된 상태에서 1살이 증가하면 netfa는 843달러 증가한다. 나이가 많고 수입이 많을수록 총자산이 늘어나는 것은 자명한 결과로 놀랍지 않다.

Does the intercept from the regression in part (ii) have an interesting meaning? Explain.

나이가 0이고 수입이 0인 사람의 총자산의 평균이다. 이런 경우는 샘플에 없기 때문에 흥미로울 바 없다.

Find the p-value for the test H0: b2 = 1 against H1: b2 >  1. Do you reject H0 at the 1% significance level?

B1 = b2-1로 두고 회귀를 하면 다음과 같다.

summary(lm1)

##   
## Call:  
## lm(formula = nettfa ~ inc + age, data = single)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -179.95 -14.16 -3.42 6.03 1113.94   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -43.03981 4.08039 -10.548 <2e-16 \*\*\*  
## inc 0.79932 0.05973 13.382 <2e-16 \*\*\*  
## age 0.84266 0.09202 9.158 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 44.68 on 2014 degrees of freedom  
## Multiple R-squared: 0.1193, Adjusted R-squared: 0.1185   
## F-statistic: 136.5 on 2 and 2014 DF, p-value: < 2.2e-16

t 통계량은[( ^b2-1) - (b2-1)]/se(b2-1)로 정의된다. 귀무가설에서 b2-1 = 0으로 주장하고 있고, 분산의 성질에 의해 se(b2-1) = se(b2)임으로 t 값은

(0.84266-1)/0.09202

## [1] -1.709846

-1.70정도가 되는데, n이 충분히 크므로 표준정규분포로 근사된다고 할 때, 유의수준 5%에서 기각역 -Z0.05 = -1.645 정도이고 -1.7은 기각역에 포함되기 때문에 귀무가설을 가각할 수 있고 b2 = 1이 아니다.

If you do a simple regression of nettfa on inc, is the estimated coefficient on inc much different from the estimate in part (ii)? Why or why not?

summary(lm(nettfa~inc, data = single))

##   
## Call:  
## lm(formula = nettfa ~ inc, data = single)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -185.12 -12.85 -4.85 1.78 1112.66   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -10.5709 2.0607 -5.13 3.18e-07 \*\*\*  
## inc 0.8207 0.0609 13.48 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 45.59 on 2015 degrees of freedom  
## Multiple R-squared: 0.08267, Adjusted R-squared: 0.08222   
## F-statistic: 181.6 on 1 and 2015 DF, p-value: < 2.2e-16

회귀계수의 값이 다소 증가했다. 그러나 큰 변화는 없다. inc값이 변화하는 이유는 age가 변해서 inc가 변해서 nettfa를 설명하는 부분이 다중에서는 분리가 되었다가 단순에서는 합쳐지면서 변화한 것이다. 다만, 두 상관관계가 양수이기 때문에 단순에서 inc의 회귀계수가 증가했다. 다만 증가량이 매우 작기 때문에 상관관계가 아주 작은 양의 값임을 추론할 수 있다. 실제로 값을 구해보면 0.04정도가 나온다.

cor(single$inc, single$age)

## [1] 0.03905864

#### C11

Use the data in HTV.RAW to answer this question. See also Computer Exercise C10 in Chapter 3.

1. Estimate the regression model by OLS and report the results in the usual form. Test the null hypothesis that educ is linearly related to abil against the alternative that the relationship is quadratic.

data(htv)  
lm1 = lm(educ~motheduc+fatheduc+abil+I(abil\*abil), data = htv)  
summary(lm1)

##   
## Call:  
## lm(formula = educ ~ motheduc + fatheduc + abil + I(abil \* abil),   
## data = htv)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2506 -1.1274 -0.1355 1.0223 7.0482   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.240226 0.287410 28.671 < 2e-16 \*\*\*  
## motheduc 0.190126 0.028096 6.767 2.03e-11 \*\*\*  
## fatheduc 0.108939 0.019601 5.558 3.35e-08 \*\*\*  
## abil 0.401462 0.030288 13.255 < 2e-16 \*\*\*  
## I(abil \* abil) 0.050599 0.008304 6.093 1.48e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.758 on 1225 degrees of freedom  
## Multiple R-squared: 0.4444, Adjusted R-squared: 0.4425   
## F-statistic: 244.9 on 4 and 1225 DF, p-value: < 2.2e-16

educ = 8.24 + 0.19motheduc + 0.108fatheduc + 0.4abil + 0.5abil^2

n = 1230 R2 = 0.4444

H0: b4 = 0 H1: b4 != 0

lmH0 = lm(educ~motheduc+fatheduc+abil, data = htv)  
lmH1 = lm(educ~motheduc+fatheduc+abil+I(abil\*abil), data = htv)  
anova(lmH0,lmH1)

## Analysis of Variance Table  
##   
## Model 1: educ ~ motheduc + fatheduc + abil  
## Model 2: educ ~ motheduc + fatheduc + abil + I(abil \* abil)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1226 3900.0   
## 2 1225 3785.2 1 114.73 37.13 1.478e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

f 통계량값이 충분히 크므로 유의확률이 굉장히 작은 값을 갖는다. 따라서 귀무가설을 기각할 만한 유의미한 증거를 확인했다고 할 수 있고 abil이 edu에 주는 영향을 선형이 아니라 quadratic으로 주장할 수 있다.

Using the equation in part (i), test H0:b1 = b2 against a two-sided alternative. What is the p-value of the test?

lm1 = lm(educ~motheduc+fatheduc+abil+I(abil\*abil), data = htv)

H0: b1=b2를 검정하는 것은 B1 = 0, (B1 = b1-b2)의 유의성 검정과 동일하다. B1=b1-b2으로 두고,i의 회귀식을 이에 맞게 정리하면 다음과 같다.

lm2 = lm(educ~motheduc+I(motheduc+fatheduc)+abil+I(abil\*abil), data = htv)  
summary(lm2)

##   
## Call:  
## lm(formula = educ ~ motheduc + I(motheduc + fatheduc) + abil +   
## I(abil \* abil), data = htv)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2506 -1.1274 -0.1355 1.0223 7.0482   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.240226 0.287410 28.671 < 2e-16 \*\*\*  
## motheduc 0.081187 0.041943 1.936 0.0531 .   
## I(motheduc + fatheduc) 0.108939 0.019601 5.558 3.35e-08 \*\*\*  
## abil 0.401462 0.030288 13.255 < 2e-16 \*\*\*  
## I(abil \* abil) 0.050599 0.008304 6.093 1.48e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.758 on 1225 degrees of freedom  
## Multiple R-squared: 0.4444, Adjusted R-squared: 0.4425   
## F-statistic: 244.9 on 4 and 1225 DF, p-value: < 2.2e-16

B1의 p-value가 0.0531로 일반적인 유의수준 5%보다 높다. 따라서 귀무가설을 기각할 수 있는 충분한 증거가 있다고 보기 어렵고, 유의수준 5%이내에서 b1,b2이 동일하다고 할 수 있다. 다만 유의확률이 0.0531로 0.05와 큰 차이가 없기 때문에 표본의 수를 늘려서 유의확률을 낮춘다면 귀무가설을 기각하고 b1과 b2가 다르다고 주장할 수도 있을 것이다.

Add the two college tuition variables to the regression from part (i) and determine whether they are jointly statistically significant.

lmU = lm(educ~motheduc+fatheduc+abil+I(abil\*abil)+tuit17+tuit18, data = htv)  
lmR = lm(educ~motheduc+fatheduc+abil+I(abil\*abil), data = htv)  
  
anova(lmR,lmU)

## Analysis of Variance Table  
##   
## Model 1: educ ~ motheduc + fatheduc + abil + I(abil \* abil)  
## Model 2: educ ~ motheduc + fatheduc + abil + I(abil \* abil) + tuit17 +   
## tuit18  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 1225 3785.2   
## 2 1223 3780.1 2 5.1884 0.8393 0.4322

H0: b5=b6=0

f 검정결과, p-value 값이 상당히 높다. 따라서 귀무가설을 기각할 수 없고 b5,b6은 jointly insignificant하다.

What is the correlation between tuit17 and tuit18? Explain why using the average of the tuition over the two years might be preferred to adding each separately. What happens when you do use the average?

cor(htv$tuit17,htv$tuit18)

## [1] 0.9808333

두 독립변수가 높은 상관관계를 갖는다.(0.9808), 평균 수업료를 변수로 사용하는 이유는 서로 상관관계가 높은 변수를 모델에 함께 사용할 경우 회귀계수 추정량의 분산이 커지기 때문이다. 분산이 작을 수록 좋은 추정량이므로 두 변수를 각각 넣는 것 보다 평균을 해서 하나의 변수로 사용하는 것이 선호된다. 평균을 사용하면 두 변수를 하나의 변수로 줄일 수 있다.

tuit = (htv$tuit17+htv$tuit18)/2  
lmavg = lm(educ~motheduc+fatheduc+abil+I(abil\*abil)+tuit, data = htv)  
  
summary(lmavg)

##   
## Call:  
## lm(formula = educ ~ motheduc + fatheduc + abil + I(abil \* abil) +   
## tuit, data = htv)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1469 -1.1591 -0.1132 1.0312 7.0709   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.081339 0.312614 25.851 < 2e-16 \*\*\*  
## motheduc 0.192863 0.028168 6.847 1.19e-11 \*\*\*  
## fatheduc 0.108368 0.019601 5.529 3.94e-08 \*\*\*  
## abil 0.399081 0.030336 13.156 < 2e-16 \*\*\*  
## I(abil \* abil) 0.050599 0.008302 6.095 1.46e-09 \*\*\*  
## tuit 0.015963 0.012373 1.290 0.197   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.757 on 1224 degrees of freedom  
## Multiple R-squared: 0.4451, Adjusted R-squared: 0.4428   
## F-statistic: 196.4 on 5 and 1224 DF, p-value: < 2.2e-16

1. Do the findings for the average tuition variable in part (iv) make sense when interpreted causally? What might be going on?

수업료를 인과관계로 생각할 수 없다. 수업료의 회귀계수가 작을 뿐더러 유의미하지 않다. 수업료와 높은 성적이 유의미한 관계가 있다고 주장할 수 있는 근거가 데이터 상에 존재하지 않는다. 또한 회귀분석은 두 변수 사이의 상관관계를 설명할 뿐이며 상관관계가 꼭 인과관계가 되는 것은 아니다. 일반적으로 교육에 많은 지출을 하면 좋은 성적을 받는다고 기대한다. 하지만 데이터 상에서는 좋은 성적을 위해서는 부모님의 성적이 더 중요하다. 즉 사교육을 많이 시켜도 유전적인 영향이 더 중요하다는 뜻으로 해석할 수 있다.