Econometrics HW 2

Tianyu Gao

2022-09-17

1.

a)

```
suicide = c(10.7, 11.1, 11.5, 11.7, 12, 12, 12.1, 12.7, 12.5, 13.3,  12.5)
unemployment = c(3.6, 3.5, 4.9, 5.9, 5.6, 4.9, 5.6, 8.5, 7.7, 7.0, 6.0)
mean_unemployment = mean(unemployment)
mean_suicide = mean(suicide)
```

For $suicide_i = \beta_0 + \beta_1 unemployment_i + u_i$, we have

$$\begin{cases} \hat{\beta}_0 = \overline{suicide} - \hat{\beta}_1 \overline{unemployment} \\ \hat{\beta}_1 = \frac{\sum_{i=1968}^{1978}(unemployment_i - \overline{unemployment})(suicide_i - \overline{suicide})}{\sum_{i=1968}^{1978}(unemployment_i - \overline{unemployment})^2} \end{cases}$$

```
cbind(mean_suicide, mean_unemployment)
```

```
##      mean_suicide mean_unemployment
## [1,]     12.00909          5.745455
```

By substitute the values of the statistics into the formula, we have

$$\begin{cases} \hat{\beta}_0 \approx 9.67 \\ \hat{\beta}_1 \approx .41 \end{cases}$$

```
beta_1 = t(unemployment - mean_unemployment) %*% t(t(suicide - mean_suicide)) / t(unemployment - mean_unemploymen
t) %*% t(t(unemployment - mean_unemployment))
beta_0 = mean_suicide - mean_unemployment * beta_1
cbind(beta_0, beta_1)
```

```
##          [,1]      [,2]
## [1,] 9.672773 0.4066376
```

b)

If we use suicide rate per 1000 population, the coefficient would be 100 times larger, and the intercept would be 100 times smaller.

c)

$$\widehat{suicide_i} = \hat{\beta}_0 + \hat{\beta}_1 unemployment_i$$

```
predicted_suicide = as.numeric(beta_0 + beta_1 * unemployment)
predicted_suicide
```

```
##  [1] 11.13667 11.09600 11.66530 12.07193 11.94994 11.66530 11.94994 13.12919
##  [9] 12.80388 12.51924 12.11260
```

```
year = c(1968:1978)
cbind(year, predicted_suicide)
```

```
##       year predicted_suicide
##  [1,] 1968          11.13667
##  [2,] 1969          11.09600
##  [3,] 1970          11.66530
##  [4,] 1971          12.07193
##  [5,] 1972          11.94994
##  [6,] 1973          11.66530
##  [7,] 1974          11.94994
##  [8,] 1975          13.12919
##  [9,] 1976          12.80388
## [10,] 1977          12.51924
## [11,] 1978          12.11260
```

$$\hat{u}_i = Y_i - \bar{Y}_i$$

```
estimated_error = suicide - predicted_suicide
cbind(year, estimated_error)
```

```
##       year estimated_error
##  [1,] 1968    -0.436668421
##  [2,] 1969     0.003995339
##  [3,] 1970    -0.165297301
##  [4,] 1971    -0.371934902
##  [5,] 1972     0.050056378
##  [6,] 1973     0.334702699
##  [7,] 1974     0.150056378
##  [8,] 1975    -0.429192663
##  [9,] 1976    -0.303882583
## [10,] 1977     0.780763738
## [11,] 1978     0.387401338
```

d)

```
mean(estimated_error)
```

```
## [1] -8.074275e-16
```

$$E(\hat{U}) = 0$$

And we observed $\bar{U} \approx 0$

This is because under the three assumptions of OLS, we have

$$E(\hat{Y}_i) = Y_i$$
$$E(U_i) = E(Y_i - \hat{Y}_i) = 0$$

e)

By substituting $unemployment = 5.8$ into $suicide = .41 unemployment + 9.67$, we can predict the suicide rate for 1979 when unemployment rate is 5.8

$$Unemployment = .41 \times 5.8 + 9.67 \approx 12.05$$

```
.41 * 5.8 + 9.67
```

```
## [1] 12.048
```

2

a)

$$\beta_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \approx 6.11$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} = \frac{1}{110}\sum_{i=1}^{110} Y_i - \frac{\beta_1}{110}\sum_{i=1}^{110} X_i \approx -267.83$$

```
7625.9/1248.9
```

```
## [1] 6.106093
```

```
(17375 - 7665.5 * 6.11) / 110
```

```
## [1] -267.8291
```

The slope is 6.11 and the intercept is -267.83, which means with one more inch height, the students tend to weigh 6.11 pounds more. The intercept has no economic meaning because no one in real life has negative weight.

b)

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=i}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}(\hat{\beta}_1 X_i + \hat{\beta}_0 - \hat{\beta}_1 \bar{X} - \hat{\beta}_0)^2}{\sum_{i=1}^{n} y_i^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} \approx .49$$

```
6.11^2 * 1248.9 / 94228.8
```

```
## [1] 0.4947963
```

c)

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2} = \sqrt{\frac{1}{108}SSR}$$
$$= \sqrt{\frac{1}{108}(TSS - ESS)} = \sqrt{\frac{1}{108}[\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2]}$$
$$= \sqrt{\frac{1}{108}(\sum_{i=1}^{110} y_i^2 - \hat{\beta}_1^2 \sum_{i=1}^{110} x_i^2)} = 20.99$$

```
sqrt((94228.8 - 6.11^2 * 1248.9)/108)
```

```
## [1] 20.99487
```

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2}{[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2]^2}}$$

d)

$$H_A : \hat{\beta}_1 \neq 0$$

e)

The calories taken in by individuals will also have influence on the weight.

4

a)

It does not make sense to hold Sleep, Work and Leisure fixed while changing Study. This is because the time in one day is spent on either Sleep, Work, Leisure and Study. When all three variables are fixed the fourth variable, Study, is also fixed.

b)

The sum of hours in the four activities is 168 because any activity is put into one of the four categories. Mathematically,

$$Study_i + Sleep_i + Work_i + Leisure_i = 168$$
$$\Rightarrow Leisure_i = 168 - (Study_i + Sleep_i + Work_i)$$

which means one of the variables can be expressed into a linear combination of other three variables.

c)

We delete one variable from the model.

$$GPA_i = \beta_0 + \beta_1 Study_i + \beta_2 Sleep_i + \beta_3 Work_i + u_i$$

where β_0 expresses the effect of the leisure time.

4

a)

Yes. Consumption of alcohol is what we are interested in. Holding other variables fixed, when consumption of alcohol increases 1 unit, the college GPA increases $\beta_{alcohol}$ units.

b)

Only one of them should be included as explanatory variables. SAT score is highly correlated with hsGPA, including both of them will introduce the imperfect multi-colinearity.

5

a)

```
library(tidyverse)
library(haven)
setwd("/Users/kevintsukuyo/Documents/Course Files/2022F/Applied Econometrics/HW2")
colGPA = read_dta('college.dta')
```

```
colnames(colGPA)
```

```
##  [1] "age"          "drinks"       "male"         "freshman"     "sophomore"
##  [6] "junior"       "senior"       "fraternity"   "abstainer"    "lightdrinker"
## [11] "moddrinker"   "heavydrinker" "marijuana"    "gpa"          "work"
```

i)

```
colGPA %>% summarize(mean_GPA = mean(gpa, na.rm = TRUE))
```

```
## # A tibble: 1 × 1
##   mean_GPA
##      <dbl>
## 1     3.24
```

After removing null values, the mean GPA in the sample is 3.2388

ii)

```
colGPA %>% summarize(Number_of_Men = sum(male == 1),
                     Number_of_Women = sum(male == 0))
```

```
## # A tibble: 1 × 2
##   Number_of_Men Number_of_Women
##           <int>           <int>
## 1          3467            6423
```

There are 3467 men and 6423 women in the dataset.

iii)

```
unique(colGPA$fraternity)
```

```
## [1] 0 1
```

```
colGPA %>% summarize(Fraternity_And_Sorority_Share = sum(fraternity == 1) / dim(colGPA)[1])
```

```
## # A tibble: 1 × 1
##   Fraternity_And_Sorority_Share
##                           <dbl>
## 1                         0.123
```

There are 12% students are menbers of fraternity or sorority.

iv)

```
unique(colGPA$work)
```

```
## [1] 0 1
```

```
colGPA %>% summarize(Work_Percentage = sum(work == 1) / dim(colGPA)[1])
```

```
## # A tibble: 1 × 1
##   Work_Percentage
##             <dbl>
## 1           0.647
```

64% students have work experience.

v)

```
colGPA %>% summarize(Marijuana_Percentage = sum(marijuana == 1)/dim(colGPA)[1])
```

```
## # A tibble: 1 × 1
##   Marijuana_Percentage
##                  <dbl>
## 1                0.164
```

16% students are reported to use marijuana in the past 30 days.

b)

```
colGPA %>% summarize(Not_Report = sum(is.na(gpa) == TRUE) / dim(colGPA)[1])
```

```
## # A tibble: 1 × 1
##   Not_Report
##        <dbl>
## 1     0.0146
```

1.4% students did not report their GPA.

```
colGPA %>%
  filter(is.na(gpa))
```

```
## # A tibble: 144 × 15
##      age drinks       male fresh…¹ sopho…² junior senior frate…³ absta…⁴ light…⁵
##    <dbl> <dbl+lbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
##  1    20 0 [none pa…     1       0       0      1      0       0       1       0
##  2    21 0 [none pa…     1       0       0      0      1       0       1       0
##  3    19 5 [5 (5)]       1       1       0      0      0       0       0       0
##  4    21 3 [3 (3)]       0       0       1      0      0       0       0       0
##  5    21 3 [3 (3)]       0       0       0      1      0       0       0       1
##  6    18 0 [none pa…     0       1       0      0      0       0       1       0
##  7    22 2 [2 (2)]       0       1       0      0      0       0       0       1
##  8    19 0 [none pa…     0       1       0      0      0       0       1       0
##  9    18 3 [3 (3)]       0       1       0      0      0       0       0       1
## 10    18 0 [none pa…     1       1       0      0      0       0       1       0
## # … with 134 more rows, 5 more variables: moddrinker <dbl>, heavydrinker <dbl>,
## #   marijuana <dbl>, gpa <dbl>, work <dbl>, and abbreviated variable names
## #   ¹freshman, ²sophomore, ³fraternity, ⁴abstainer, ⁵lightdrinker
```

After glancing at the table, I find that the value of drinks tends to be bigger than 0. Then we compare the value of variable drinks for student reported gpa and those did not.

```
library(magrittr)
library(ggplot2)
colGPA %<>% mutate(recorded = as.factor(!is.na(gpa)))
```

```r
colGPA %>%
  group_by(recorded)%>%
  summarise(mean_drinks = mean(drinks), )
```

```
## # A tibble: 2 × 2
##   recorded mean_drinks
##   <fct>          <dbl>
## 1 FALSE           1.98
## 2 TRUE            2.56
```

In the session above I set recorded to be true if people recorded their GPA, and it seems that people who record their GPA will drink 25% more than those don't, which means my assumption is wrong. Therefore I suggest the non-reporting is random.

c)

```r
summary(lm(gpa ~ male + work, data = colGPA))
```

```
##
## Call:
## lm(formula = gpa ~ male + work, data = colGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61588 -0.28258  0.06195  0.39535  0.83251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.28258    0.01062 309.194   <2e-16 ***
## male        -0.10386    0.01190  -8.730   <2e-16 ***
## work        -0.01123    0.01189  -0.945    0.345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5594 on 9743 degrees of freedom
##   (144 observations deleted due to missingness)
## Multiple R-squared:  0.00778,	Adjusted R-squared:  0.007576
## F-statistic:  38.2 on 2 and 9743 DF,  p-value: < 2.2e-16
```

The regression model is：

$$\widehat{gpa_i} = 3.29 - .10male_i - .01work_i$$

Interpretation:

For women who have not worked, the average gpa is 3.29; for men who have not worked, the average gpa is 3.29 - .10 = 3.19; for men who have worked, the average gpa is 3.19 - .01 = 3.18; fro women working, the average gpa is 3.28

```r
library(lmtest)
library(sandwich)
lc = lm(gpa ~ male + work, data = colGPA)
coeftest(lc, vcov = vcovHC(lc, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  3.282580   0.010606 309.4918   <2e-16 ***
## male        -0.103857   0.012078  -8.5992   <2e-16 ***
## work        -0.011230   0.011965  -0.9386    0.348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The hetroskedatistic model is quite similar to the homoskedastistic one, which means the data is homoskedastisitc.

d)

```r
summary(lm(gpa~freshman+sophomore+junior+senior, data = colGPA[colGPA$male == 1, ]))
```

```
##
## Call:
## lm(formula = gpa ~ freshman + sophomore + junior + senior, data = colGPA[colGPA$male ==
##     1, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58672 -0.44183  0.07988  0.48726  0.89147
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.25342    0.02003 162.445  < 2e-16 ***
## freshman    -0.14489    0.02844  -5.095 3.67e-07 ***
## sophomore   -0.10974    0.02842  -3.862 0.000115 ***
## junior      -0.07398    0.02739  -2.701 0.006956 **
## senior            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.577 on 3415 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.008289,   Adjusted R-squared:  0.007418
## F-statistic: 9.515 on 3 and 3415 DF,  p-value: 2.951e-06
```

If we try to include all four coefficients for freshman, sophomore, junior and senior, we can see that the the coefficient for senior is not a number. This is because a student has to belong to either of four categories above, leading to $freshman_i + sophomore_i + junior_i + senior_i = 1$, so anyone of the four variables can be expressed as the linear combination of other three variables, which is called multicollinearity. The solution is to either drop one variable from the model or omit the intercept term.

```
ld = lm(gpa~freshman+sophomore+junior+senior, data = colGPA[colGPA$male == 1, ])
coeftest(ld, vcov = vcovHC(ld, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  3.253416   0.018768 173.3506 < 2.2e-16 ***
## freshman    -0.144888   0.028973  -5.0008 5.997e-07 ***
## sophomore   -0.109744   0.027508  -3.9895 6.760e-05 ***
## junior      -0.073981   0.026107  -2.8338  0.004627 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The hetroskedatistic model is quite similar to the homoskedastistic one, which means the data is homoskedastisitc.

e)

```
summary(lm(gpa~ sophomore+junior+senior, data = colGPA[colGPA$male == 1, ]))
```

```
##
## Call:
## lm(formula = gpa ~ sophomore + junior + senior, data = colGPA[colGPA$male ==
##     1, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58672 -0.44183  0.07988  0.48726  0.89147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.10853    0.02019 153.990  < 2e-16 ***
## sophomore    0.03514    0.02853   1.232  0.21811
## junior       0.07091    0.02751   2.577  0.00999 **
## senior       0.14489    0.02844   5.095 3.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.577 on 3415 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.008289,   Adjusted R-squared:  0.007418
## F-statistic: 9.515 on 3 and 3415 DF,  p-value: 2.951e-06
```

```
le = lm(gpa~ sophomore+junior+senior, data = colGPA[colGPA$male == 1, ])
coeftest(le, vcov = vcovHC(le, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 3.108528   0.022073 140.8314 < 2.2e-16 ***
## sophomore   0.035144   0.029861   1.1769   0.23930
## junior      0.070907   0.028575   2.4814   0.01313 *
## senior      0.144888   0.028973   5.0008 5.997e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model is

$$\widehat{gpa_i^{male}} = 3.11 + .04sophomore_i^{male} + .07junior_i^{male} + .14senior_i^{male}$$

Interpretation: for male freshman student, their average gpa is 3.11, for male sophomore student, their average gpa is 3.11+.04 = 3.15, for male junior students, their average gpa is 3.11 + .07 = 3.18, for male senior students, their average gpa is 3.11 + .14 = 3.25

f)

In the test, q = 1

```
library(car)
lf = lm(gpa~ sophomore+junior+senior, data = colGPA[colGPA$male == 1, ])
linearHypothesis(lf, c('sophomore = junior'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## sophomore - junior = 0
##
## Model 1: restricted model
## Model 2: gpa ~ sophomore + junior + senior
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   3416 1137.5
## 2   3415 1136.9  1   0.56336 1.6922 0.1934
```

Therefore we fail to reject the null hypothesis at the 5% significant level.

g)

Significance level: .05, Critical value: $F_{3,3415}(0.95) = 2.61 < 9.95$

```
qf(.95, 3, 3415)
```

```
## [1] 2.60751
```

Therefore we reject the null hypothesis.

h)

```
summary(lm(gpa ~ age + sophomore + junior + senior, data = colGPA[colGPA$male == 1,]))
```

```
##
## Call:
## lm(formula = gpa ~ age + sophomore + junior + senior, data = colGPA[colGPA$male ==
##     1, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61488 -0.44215  0.08411  0.49217  0.91275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.903715   0.126469  22.960  < 2e-16 ***
## age         0.010796   0.006581   1.640  0.10099
## sophomore   0.021988   0.029630   0.742  0.45809
## junior      0.044092   0.031995   1.378  0.16826
## senior      0.107955   0.036264   2.977  0.00293 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5769 on 3414 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.00907,    Adjusted R-squared:  0.007909
## F-statistic: 7.812 on 4 and 3414 DF,  p-value: 2.898e-06
```

The coefficients of grades turn to be less significant. This is because as age increases linearly, the grade individual is in increases linearly as well, which leads to imperfect multicollinearity.

i)

```
summary(lm(gpa ~ male + work + marijuana + lightdrinker + moddrinker +heavydrinker + sophomore + junior + senior,
data = colGPA))
```

```
##
## Call:
## lm(formula = gpa ~ male + work + marijuana + lightdrinker + moddrinker +
##     heavydrinker + sophomore + junior + senior, data = colGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75810 -0.33835  0.04664  0.42588  1.16512
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.28713    0.01650 199.238  < 2e-16 ***
## male         -0.09273    0.01188  -7.809 6.37e-15 ***
## work         -0.03459    0.01196  -2.893  0.00382 **
## marijuana    -0.08909    0.01621  -5.496 3.97e-08 ***
## lightdrinker -0.05647    0.01433  -3.940 8.19e-05 ***
## moddrinker   -0.11258    0.01736  -6.486 9.22e-11 ***
## heavydrinker -0.23584    0.03572  -6.603 4.23e-11 ***
## sophomore     0.06627    0.01631   4.064 4.87e-05 ***
## junior        0.09059    0.01585   5.717 1.12e-08 ***
## senior        0.17226    0.01627  10.590  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5525 on 9736 degrees of freedom
##   (144 observations deleted due to missingness)
## Multiple R-squared:  0.0328, Adjusted R-squared:  0.0319
## F-statistic: 36.68 on 9 and 9736 DF,  p-value: < 2.2e-16
```

The model is

$$gpa_i = 3.287 - .093male_i - .035work_i - .089marijuana_i - .056lightdrinker_i - .112moddrinker_i - .236heavydrinker_i + .066sophomore_i + .091junior$$

For the individual being discussed, predicted GPA is 3.287 - .035 - .112 + .172 = 3.312

```
3.287-.035-.112+.172
```

```
## [1] 3.312
```

j)

```
lj = lm(gpa ~ male + work + marijuana + lightdrinker + moddrinker +heavydrinker + sophomore + junior + senior, da
ta = colGPA)
linearHypothesis(lj, c('sophomore = 0'), white.adjust = 'hc1')
```

```
## Linear hypothesis test
##
## Hypothesis:
## sophomore = 0
##
## Model 1: restricted model
## Model 2: gpa ~ male + work + marijuana + lightdrinker + moddrinker + heavydrinker +
##     sophomore + junior + senior
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df     F    Pr(>F)
## 1   9737
## 2   9736  1 15.52 8.223e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\because p - value = 4.867e - 05 < 0.01 \Rightarrow H_0 \; is \; rejected \; under \; significance \; level \; of \; 99\%$$

k)

$ R^2_{adj} = .0319$ means 3.2 variation in the data can be explained by the regression model.

The coefficent between $sophomore$ increased a lot, which indicates there is an omitted variable bias.