ECON 3412 Midterm

David Chen, dc3451

October 25, 2020

Problem 1

a

5 hours is $5 * 60 = 300$ minutes, so we would expect that sleep changes by $300\beta_1$; if β_1 is negative, which it probably is, then sleep falls by $|300\beta_1|$ minutes per week.

In this case, $\beta_1 = -0.1483$, so we expect sleep to fall by 44.5 minutes a week.

R output:

```
Call:
lm(formula = formula, data = data)

Coefficients:
(Intercept)       totwrk         educ          age
  3638.2453      -0.1484     -11.1338       2.1999


[[2]]

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 3638.245312  115.082734 31.6142 < 2.2e-16 ***
totwrk        -0.148373    0.018709 -7.9305 8.609e-15 ***
educ         -11.133813    5.776658 -1.9274   0.05434 .
age            2.199885    1.436901  1.5310   0.12622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

b

No; $R^2 = 0.1134, \overline{R^2} = 0.1096$, so this regression only expains about 10% of the variation in sleep. There are a lot of other things, such as stress (probably highly correlated with work, since work is stressful!) which determine the amount of sleep someone gets (more stress = less sleep.)

Another factor might be income; higher paying jobs likely demand both more time, and increase things like stress or extracurricular things (ex: programmers often program outside of work as well in the open source domain) related to work which affect amounts of sleep than isn't captured in the totwrk, educ or age variables.

R output:

```
Residual standard error: 419.4 on 702 degrees of freedom
Multiple R-squared:  0.1134,	Adjusted R-squared:  0.1096
F-statistic: 29.92 on 3 and 702 DF,  p-value: < 2.2e-16
```

c

They are jointly significant at the 5% level, as $F = 4.4371, p = 0.01216$. Individually, neither is significant. Including them shifts β_1 from -0.1507 (without them) to -0.1483, so the effect is small, as would be expected from idividually nonsignificant results.

R output:

```
Linear hypothesis test

Hypothesis:
educ = 0
age = 0

Model 1: restricted model
Model 2: sleep ~ totwrk + educ + age

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F  Pr(>F)
1    704
2    702  2 4.4371 0.01216 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R code used:

```
library(haven)
library(lmtest)
library(sandwich)
library(car)
library(ggplot2)

robust_lm <- function(formula, data, conditions) {
  regression <- lm(formula, data = data)
  robust_coef <- coeftest(regression, vcov = vcovHC(regression, "HC1"))

  if (missing(conditions)) {
    return (list(regression, robust_coef))
  }

  f_test = linearHypothesis(regression, conditions, vcov = vcovHC(regression, "HC1"))

  return (list(regression, robust_coef, f_test))
}

sleep_dta <- read_dta("sleep75.dta ")
robust_lm(sleep ~ totwrk + educ + age, sleep_dta, c("educ = 0", "age = 0"))
robust_lm(sleep ~ totwrk, sleep_dta, c("educ = 0", "age = 0"))
```

Problem 2

a

R output:

```
Call:

lm(formula = formula, data = data)

Coefficients:
(Intercept)        black         dist   blackxdist      momcoll     incomehi
   2.608922    -0.030570    -0.004365     0.002578     0.069130     0.044019


[[2]]

t test of coefficients:

              Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  2.6089217  0.0034626 753.4486 < 2.2e-16 ***
black       -0.0305703  0.0061339  -4.9838 6.513e-07 ***
```

```
dist        -0.0043655  0.0010297  -4.2394 2.295e-05 ***
blackxdist   0.0025778  0.0027694   0.9308     0.352
momcoll      0.0691305  0.0057978  11.9235 < 2.2e-16 ***
incomehi     0.0440187  0.0047157   9.3345 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

One might want to include the interaction variable because black people might be differently impacted by being further away from school. For example, they are on balance poorer, with black households owning significantly less wealth, so transportation might be harder, since they might not have easy access to a car. Alternatively, they can be marginalized into worse developed neighborhoods since white people don’t want them in nice ones, so as a result there’s less/worse public transportation available to them.

On the other hand, we might also expect that on balance, black students are less impacted by distance to school; especially noting that if they come from a dense and underdeveloped region in a city, moving closer to campus might actually help them graduate compared to their white peers.

The coefficient is interpreted as the difference between the effect of distance on the (log of) education between white people and black people; that is, the distance effect (increase in log of education per unit increase in distance) for black people is $-0.0043 + 0.002577 = -0.0018$, compared to the effect of -0.0043 in nonblack populations.

b

We get that $F = 14.591, p = 1.902e-09 << 0.01$, so they are jointly significant (at 1% level, and also at levels much lower than that). This is expected, as black and dist are very very significant in the regression, and thus we would expect a joint hypothesis that they are both 0 to be very unlikely.

R output

```
Linear hypothesis test

Hypothesis:
black = 0
dist = 0
blackxdist = 0

Model 1: restricted model
Model 2: ln_ed ~ black + dist + blackxdist + momcoll + incomehi

Note: Coefficient covariance matrix supplied.
```

```
  Res.Df Df      F    Pr(>F)
1   3793
2   3790  3 14.591 1.902e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c

R output:

```
Call:
lm(formula = formula, data = data)

Coefficients:
   (Intercept)           black            dist         momcoll        incomehi
      2.607350       -0.022287       -0.004074        0.068960        0.047416
blackxincomehi
     -0.024795


[[2]]

t test of coefficients:

                 Estimate Std. Error  t value  Pr(>|t|)
(Intercept)     2.6073500  0.0034403 757.8885 < 2.2e-16 ***
black          -0.0222868  0.0053183  -4.1906 2.846e-05 ***
dist           -0.0040737  0.0009546  -4.2675 2.025e-05 ***
momcoll         0.0689598  0.0058062  11.8770 < 2.2e-16 ***
incomehi        0.0474156  0.0050663   9.3591 < 2.2e-16 ***
blackxincomehi -0.0247948  0.0127776  -1.9405   0.05239 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

One might want to include the interaction variable, as we have that income effects might be different between black people and nonblack people. For example, the lower frequency of college graduates in black populations might cause a lower willingness to go to college than a white person, even given the same income. Similarly, racism in the process (such as the *terrible* Columbia (undergrad) application, which is so skewed towards white ideals of good art) might also discouage disproportionately black students.

This is also what we see in the regression; the income effect on black people is lower by 0.024, so we expect it to be about half of the income effect of nonblack people.

d

The null is

$$H_0 : blackxincomehi_i + black_i = 0$$

and alternative:

$$H_1 : blackxincomehi_i + black_i \neq 0$$

R output:

```
Linear hypothesis test

Hypothesis:
black  + blackxincomehi = 0

Model 1: restricted model
Model 2: ln_ed ~ black + dist + momcoll + incomehi + blackxincomehi

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   3791
2   3790  1 16.323 5.447e-05 ***
---
```

This gives that high income black people are different in education from high income nonblack people.

e

There's probably omitted variable bias here: for example, *dadcoll* is not included, but this is probably correlated with *momcoll* since people often marry to similar individuals to themselves, and probably also has an important effect as well, especially if the father is primary breadwinner, which is often the case.

We may also be missing important nonlinearity here: for example, the relationship between distance and (log of) education might be nonlinear. In particular, being too close to the college might result in being too tied up with school, increasing stress, and thus likelihood of dropping out (alternatively, might signal too poor to live outside of campus, and being poorer makes you less likely to finish college, or losing guaranteed housing). Being too far might have a similar, but opposite effect: not invsested enough in school, and dropping out because of that.

R code used for this problem:

```
college_dta <- read_dta("CollegeDistance_for Midterm.dta")
robust_lm(ln_ed ~ black + dist + blackxdist + momcoll + incomehi, college_dta)
robust_lm(ln_ed ~ black + dist + blackxdist + momcoll + incomehi, college_dta,
          c("black = 0", "dist = 0", "blackxdist = 0"))
robust_lm(ln_ed ~ black + dist + momcoll + incomehi + blackxincomehi, college_dta)
robust_lm(ln_ed ~ black + dist + momcoll + incomehi + blackxincomehi, college_dta,
          "blackxincomehi = -black")
```

Problem 3

a

Under classical measurement error, $expend_s$ is biased towards 0, such that $truespend_s > expend_s$ and you are underestimating the effect of expenditure.

We can write

$$expend_s = \frac{\sigma^2_{expend} - \sigma^2_v}{\sigma^2_{expend}} truespend_s = \frac{0.6 - 0.15}{0.6} truespend_s$$

so $truespend = 2.456/0.75 = 3.274$.

b

No; adding the squared term does not largely increase $\overline{R^2}$, so it doesn't really explain any more of the variance from before; however, the coefficient on the squared term isn't significat, as $t = 1.01/0.64 = 1.578$, and $2(1 - \Phi(1.578)) = 0.114$, so it isn't significant at a 10% level. I would discard the quadratic term.

c

Yes; the only thing we need for the OLS estimator to be unbiased are that $E[u \mid X] = 0$. In particular, this variance, i.e. $E[u^2 \mid X] - (E[u \mid X])^2$ can vary (i.e. in heteroskedastic data), so long as the expected values of the residuals vanish.

d

We would expect that older buildings were built with different space constraints than new buildings, and thus the effect size has on the rent of older buildings is different from the effect size has on the rent of newer buildings. For example, in New York everything is very dense; then, new construction places size at a premium, and thus size has a larger effect on rent than in older buildings (i.e. the increase in size of a new building is associated with more rent due than the same increase in an older building).