PS3 R Solutions

Matthew Alampay Davis

February 9, 2021

Question 1

```
gpa4 <- read.dta13("GPA4.dta")
gpa.mod1 <- lm_robust(colGPA ~ hsGPA + skipped, data = gpa4,
    se_type = "stata")
gpa.mod2 <- lm_robust(colGPA ~ hsGPA + skipped + PC, data = gpa4,
    se_type = "stata")
gpa.mod3 <- lm_robust(colGPA ~ hsGPA + skipped + PC + bgfriend +
    campus, data = gpa4, se_type = "stata")
```

Part a

```
summary(gpa.mod1)
```

```
##
## Call:
## lm_robust(formula = colGPA ~ hsGPA + skipped, data = gpa4, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  1.57917    0.32525   4.855 3.212e-06   0.9360  2.22228 138
## hsGPA        0.45880    0.09415   4.873 2.974e-06   0.2726  0.64497 138
## skipped     -0.07743    0.02539  -3.050 2.741e-03  -0.1276 -0.02724 138
##
## Multiple R-squared:  0.2227 ,    Adjusted R-squared:  0.2115
## F-statistic:  20.9 on 2 and 138 DF,  p-value: 1.181e-08
```

$$\underset{}{\widehat{\text{colGPA}} = 1.579} + \underset{(0.325)}{} \underset{}{0.458\text{hsGPA}} \underset{(0.094)}{} - \underset{(0.025)}{0.077\text{skipped}}$$

Part b

The coefficient on hsGPA is 0.458. This means that a one point increase in hsGPA will lead to an increase of 0.458 points in colGPA (college GPA). Students with higher GPAs in high school tend to have higher college GPAs.

Part c

The t-statistic for H0: skipped=0 vs H1: skipped!= 0 is -3.05. Since |-3.05|>1.96 we can reject H0 and conclude skipped is statistically different than 0. We are essentially testing if on average skipping classes would affect students' college GPA (in layman words).

Part d

```
gpa.mod1$statistic["skipped"]
```

```
##   skipped
## -3.050356
```

```
gpa.mod2$statistic["skipped"]
```

```
##   skipped
## -2.528452
```

```
gpa.mod3$statistic["skipped"]
```

```
##   skipped
## -2.736817
```

The critical value for a two-sided test at the 1% significance level is 2.58, so we reject H0 (that the coefficient on skipped is equal to zero) in Regressions 1 and 3. We cannot reject H0 in regression 2. (We may also compare the p-values with 0.01 as instructed in the problem)

Part e

```
gpa.mod3$coefficients["campus"]
```

```
##     campus
## -0.1243919
```

The coefficient on campus is -0.124. This means living on campus reduces colGPA by 0.124 points. The negative sign on the coefficient might be because students who live on campus have more distractions that they would if they lived at home. However, students who live on campus are also more able to study with one another, so it is not clear whether the sign of the coefficient should be positive or negative; it would depend on which effect was stronger. The size of the coefficient is about one-tenth of a point, which is small. Note that the coefficient is not statistically different than 0, so there is not too much we can say.

Part f

```
gpa.mod3$coefficients["bgfriend"]
```

```
##   bgfriend
## 0.08586053
```

The coefficient on bgfriend is 0.085. This means having a boyfriend or girlfriend increases GPA by 0.085. One would expect the coefficient to be positive or negative, such as: the coefficient could be positive if students who are more intelligent are more likely to be dating (here we have a correlation), alternatively, the sign could be negative if dating distracts a student from studying. However, here the magnitude of the coefficient is very small also note that the coefficient is not statistically significant. Hence, we can conclude that there is no significant effect on GPA of dating.

Question 2

```
nurse <- read.dta13("WiscoNursingHome.dta") %>% mutate(logtpy = log(tpy),
    lognumbed = log(numbed), logsqrfoot = log(sqrfoot))
nurse.2000 <- filter(nurse, cryear == 2000)
nurse.2001 <- filter(nurse, cryear == 2001)
```

Question 2.1

Part a)

```
# i)
cor(nurse.2000$tpy, nurse.2000$logtpy)
```

```
## [1] 0.9371853
```

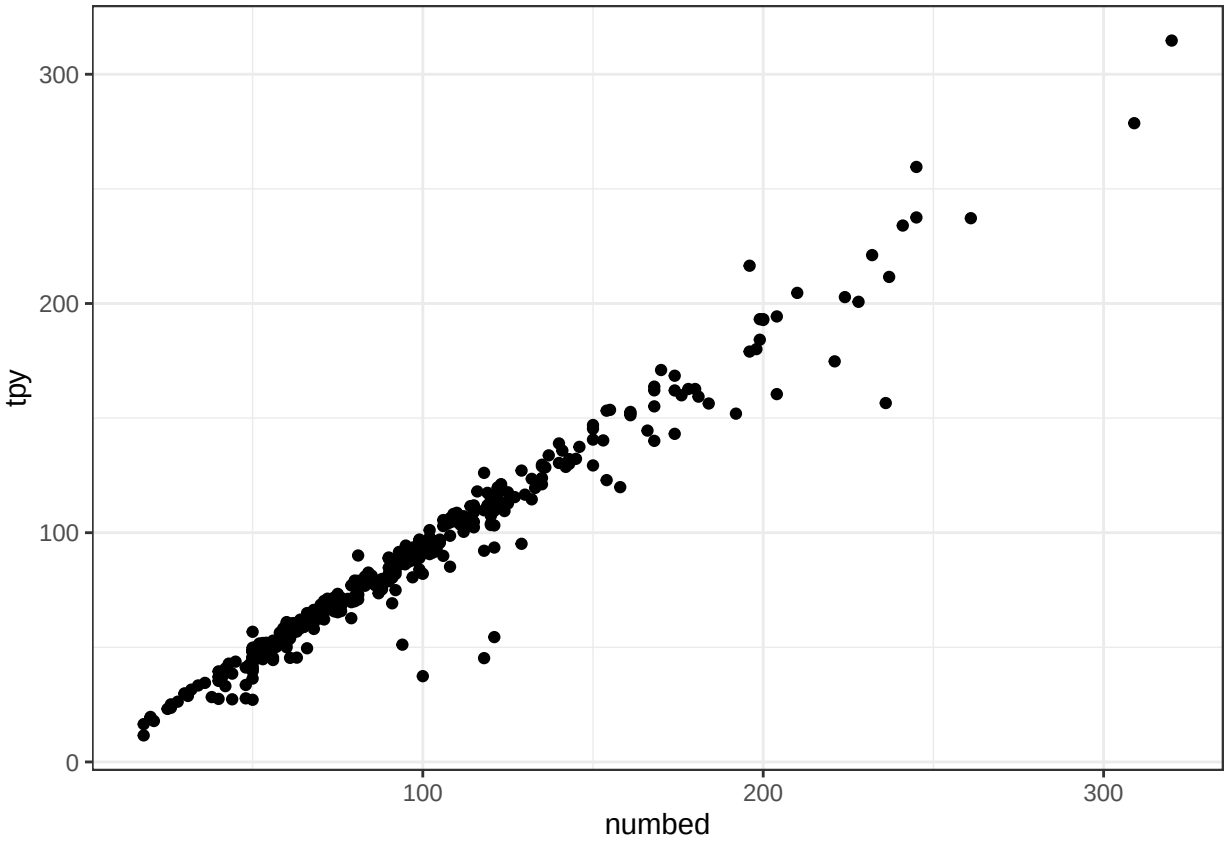
```
# ii)
nurse.2000[, c("tpy", "logtpy", "sqrfoot")] %>% na.omit() %>%
    cor
```

```
##               tpy    logtpy   sqrfoot
## tpy     1.0000000 0.9367097 0.8244198
## logtpy  0.9367097 1.0000000 0.7361852
## sqrfoot 0.8244198 0.7361852 1.0000000
```

The correlation between TOY and LOG(TPY) is very strong and close to 1 (i.e., 94%). The correlation coefficients among these three variables appear to be highly correlated. The lowest is 82 % and the highest is 97% correlation.

Part b)

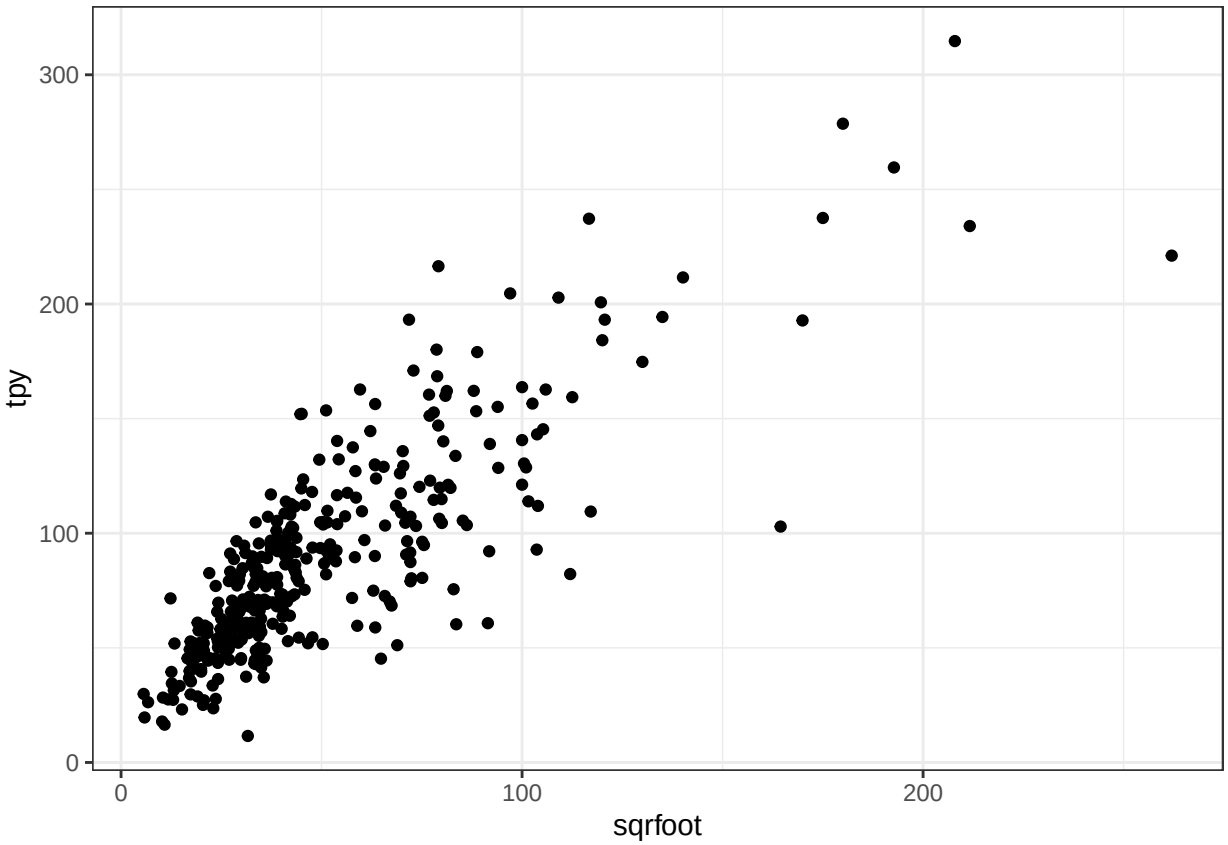
```
ggplot(nurse.2000, aes(x = numbed, y = tpy)) + theme_bw() + geom_point()
```



Based on visual inspection of this plot there is evidence for positive correlation between these two variables and the association seems to be precise.

```
ggplot(nurse.2000, aes(x = sqrfoot, y = tpy)) + theme_bw() +
    geom_point()
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



Based on visual inspection of this plot like the previous one there is evidence for positive correlation between these two variables but the association seems to be less precise

Part c)

```
nurse.mod1 <- lm_robust(tpy ~ numbed, data = nurse.2000)
nurse.mod2 <- lm_robust(tpy ~ sqrfoot, data = nurse.2000)
nurse.mod3 <- lm_robust(tpy ~ lognumbed, data = nurse.2000)
nurse.mod4 <- lm_robust(tpy ~ logsqrfoot, data = nurse.2000)

nurse.mod1$r.squared
```

```
## [1] 0.9586405
```

```
nurse.mod1$statistic[2]
```

```
##   numbed
## 56.29556
```

```
nurse.mod2$r.squared
```

```
## [1] 0.6796681
```

```
nurse.mod2$statistic[2]
```

```
##  sqrfoot
## 14.39876
```

```
nurse.mod3$r.squared
```

```
## [1] 0.8519338
```

```
nurse.mod3$statistic[2]
```

```
## lognumbed
##  22.45329
```

```
nurse.mod4$r.squared
```

```
## [1] 0.6500417
```

```
nurse.mod4$statistic[2]
```

```
## logsqrfoot
##   17.14457
```

Question 2.2

Same thing just replace the filter with 2001

Question 3

```
wage <- read.dta13("WAGE1.dta")
```

Part a

For the omitted variable X_2 to cause omitted variable bias (OVB), it should satisfy the following to conditions:

i) Years of potential experience X_2 should be a determinant factor for average hourly earnings/wage Y). That is, $Y = f(X_2)$ so X_2 is part of the error term u.

ii) Years of potential experience X_2 should be correlated with years of education X_1). Mathematically, it means that their correlation is non-zero2) != 0. Intuitively, this implies that more years of experience is correlated or sometimes affects years of education. The more you spent your years in acquiring work experience, the less time you are left with to spend in (formal) education or the number of years of education would increase as you might be taking few courses (i.e., you are part time student) as you are on the job.

Math is in the official solutions

Part b

```
wage.mod <- lm_robust(wage ~ female, data = wage, se_type = "stata")
summary(wage.mod)
```

```
##
## Call:
## lm_robust(formula = wage ~ female, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    7.099     0.2514   28.24 2.453e-107    6.606    7.593 524
## female        -2.512     0.2976   -8.44  3.125e-16   -3.097   -1.927 524
##
```

```
## Multiple R-squared:  0.1157 ,    Adjusted R-squared:  0.114
## F-statistic: 71.23 on 1 and 524 DF,  p-value: 3.125e-16
```

Since X_2 = gender dummy (binary) variable that takes the value of 1 if female and 0 otherwise, the slope coefficient is interpreted as the difference-in-group mean. That is, average hourly earnings declines by $2.51 if the individual is female. Mathematically,

$$\beta_4 = E[Y_i|X_4 = 1] - E[Y_i|X - 4 = 0] = -2.512$$

Part c

```
wage %<>% mutate(D = 1 - female)
```

Since female = 0 are male individuals, this generate command would give you D = 1 for male and D = 0 for female. In other words, D and female are dummy variables that takes opposite values. D = 1 is the same as female = 0.

```
wage.mod2 <- lm_robust(wage ~ educ + female + D, se_type = "stata",
    data = wage)
summary(wage.mod2)
```

```
## 1 coefficient  not defined because the design matrix is rank deficient
```

```
##
## Call:
## lm_robust(formula = wage ~ educ + female + D, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients: (1 not defined because the design matrix is rank deficient)
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)       NA         NA      NA        NA       NA       NA  NA
## educ          0.5065     0.0599  8.4556 2.784e-16   0.3888   0.6241 523
## female       -1.6505     0.7161 -2.3050 2.156e-02  -3.0573  -0.2438 523
## D             0.6228     0.7287  0.8547 3.931e-01  -0.8087   2.0543 523
##
## Multiple R-squared:  0.2588 ,    Adjusted R-squared:  0.256
## F-statistic: 31.95 on 2 and 523 DF,  p-value: 8.072e-14
```

The error tells us there is multicollinearity. In Stata, this is displayed by having the variable D drop out of the regression. Here, it removes the intercept. Both are ways of removing multicollinearity.

See math in the official solutions.

Part d

```
wage.mod3 <- lm_robust(wage ~ educ, data = wage, se_type = "stata")
summary(wage.mod3)
```

```
##
## Call:
## lm_robust(formula = wage ~ educ, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
```

```
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  -0.9049    0.72548  -1.247 2.129e-01   -2.330   0.5204 524
## educ          0.5414    0.06126   8.837 1.489e-17    0.421   0.6617 524
##
## Multiple R-squared:  0.1648 ,    Adjusted R-squared:  0.1632
## F-statistic: 78.09 on 1 and 524 DF,  p-value: < 2.2e-16
```

```
wage.mod4 <- lm_robust(wage ~ educ + exper, data = wage, se_type = "stata")
summary(wage.mod4)
```

```
##
## Call:
## lm_robust(formula = wage ~ educ + exper, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  -3.3905    0.86487  -3.920 1.002e-04  -5.0896 -1.69148 523
## educ          0.6443    0.06519   9.883 3.108e-21   0.5162  0.77233 523
## exper         0.0701    0.01099   6.376 4.008e-10   0.0485  0.09169 523
##
## Multiple R-squared:  0.2252 ,    Adjusted R-squared:  0.2222
## F-statistic: 50.32 on 2 and 523 DF,  p-value: < 2.2e-16
```

As can be seen from the above two tables, the coefficient on education has increased from 0.54 to 0.64. The reason for this increment is the addition of one of the omitted variable, namely, experience. The fact that it is also statistically significant suggests that it is one of the determinant variable for our dependent variable (condition #1). This result is similar to the test score example that we are using in the text that when we add percentage of English language learner in the model, the coefficient on class size has changed.

Part e

```
wage.mod5 <- lm(wage ~ educ + exper + tenure + female + nonwhite,
    data = wage)
wage.mod6 <- lm_robust(wage ~ educ + exper + tenure + female +
    nonwhite, data = wage, se_type = "stata")

summary(wage.mod5)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure + female + nonwhite,
##     data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6623 -1.7842 -0.4355  1.0810 13.9945
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.54030    0.73231  -2.103   0.0359 *
## educ         0.57034    0.04957  11.507  < 2e-16 ***
## exper        0.02534    0.01158   2.188   0.0291 *
## tenure       0.14107    0.02118   6.660 6.98e-11 ***
```

```
## female      -1.81204    0.26510  -6.835 2.30e-11 ***
## nonwhite    -0.11587    0.42692  -0.271   0.7862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.96 on 520 degrees of freedom
## Multiple R-squared:  0.3636, Adjusted R-squared:  0.3575
## F-statistic: 59.43 on 5 and 520 DF,  p-value: < 2.2e-16
```

```
summary(wage.mod5)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure + female + nonwhite,
##     data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6623 -1.7842 -0.4355  1.0810 13.9945
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.54030    0.73231  -2.103   0.0359 *
## educ         0.57034    0.04957  11.507  < 2e-16 ***
## exper        0.02534    0.01158   2.188   0.0291 *
## tenure       0.14107    0.02118   6.660 6.98e-11 ***
## female      -1.81204    0.26510  -6.835 2.30e-11 ***
## nonwhite    -0.11587    0.42692  -0.271   0.7862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.96 on 520 degrees of freedom
## Multiple R-squared:  0.3636, Adjusted R-squared:  0.3575
## F-statistic: 59.43 on 5 and 520 DF,  p-value: < 2.2e-16
```

Here the first table provides a regression result based on homoscedasticity-only standard error and the second one is based on heteroscedasticity-robust standard errors. As it can be seen from these two tables, the coefficients are the same in both cases but the corresponding standard errors are different for each coefficient. Since, the remaining t- statistics, p-values, and the resulting confidence intervals in the two tables are different as all of them are dependent of the standard errors. The interpretation will proceed as usual. We care about the presence of heteroscedasticity in the data because, if indeed there is the problem of heteroscedasticity, the homoscedasticity-only standard errors will be wrong. As mentioned above, if the standard errors are wrong, then everything else that depends on these wrong standard errors will result in misleading and incorrect statistical inference. It is advisable to use heteroscedasticity-robust standard errors whenever possible even if there is no heteroscedasticity. This is because, if there is no heteroscedasticity in the data, both will give us the correct standard errors. (see page 163 of the text on this issue.)

Part f

i) Testing the null hypothesis for single coefficients being equal to zero

For individual null hypothesis of these coefficients, yo u can directly use the reported t-statistics and the corresponding p-values and confidence intervals.

```
wage.mod6$statistic
```

```
## (Intercept)        educ       exper      tenure      female    nonwhite
```

```
##   -1.854303    9.299842    2.582175    5.040128   -7.118957   -0.295107
```

```
wage.mod6$p.value
```

```
##  (Intercept)         educ        exper       tenure       female
## 6.426184e-02 3.882121e-19 1.009032e-02 6.432218e-07 3.632659e-12
##     nonwhite
## 7.680300e-01
```

```
confint(wage.mod6)
```

```
##                    2.5 %      97.5 %
## (Intercept) -3.172163262  0.09156705
## educ         0.449860762  0.69082358
## exper        0.006061879  0.04462419
## tenure       0.086083712  0.19605578
## female      -2.312091563 -1.31199432
## nonwhite    -0.887251178  0.65550311
```

ii) Testing joint hypotheses

```
linearHypothesis(wage.mod6, c("educ = 0", "exper = 0", "tenure = 0",
    "female = 0", "nonwhite = 0"), test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 0
## exper = 0
## tenure = 0
## female = 0
## nonwhite = 0
##
## Model 1: restricted model
## Model 2: wage ~ educ + exper + tenure + female + nonwhite
##
##   Res.Df Df      F    Pr(>F)
## 1    525
## 2    520  5 35.616 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Note we can also do the heteroskedasticity-robust test
# using the non-robust model wage.mod5
linearHypothesis(wage.mod5, c("educ = 0", "exper = 0", "tenure = 0",
    "female = 0", "nonwhite = 0"), white.adjust = "hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 0
## exper = 0
## tenure = 0
## female = 0
## nonwhite = 0
##
## Model 1: restricted model
```

```
## Model 2: wage ~ educ + exper + tenure + female + nonwhite
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1    525
## 2    520  5 35.616 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see, the computed F-stat is 35.62 and from the F-distribution table we know that the 1%, 5%, and 10% critical values for q =5 are 3.02, 2.21 and 1.85, respectively. This implies that we can reject the null hypothesis of all slope coefficients are zero. In fact, STATA has already computed the p-value for you and it is Prob $>$ F = 0.0000. This implies that we can reject the null at 1% significance level.