Homework 7

Gergel Anastasia

6/17/2021

Exercise 1

$$\text{(a) } ESS = \sum (\hat{y}_i - \bar{y})^2 = (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \frac{1}{n}\boldsymbol{y'i})'(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \frac{1}{n}\boldsymbol{y'i})$$

$$\text{(b) } SSR = \sum (y_i - \hat{y}_i)^2 = \sum \hat{u_i}^2 = \boldsymbol{\hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta})} = \boldsymbol{y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}}$$

$$\text{(c) } TSS = \sum (y_i - \bar{y})^2 = (\boldsymbol{y} - \frac{1}{n}\boldsymbol{y'i})'(\boldsymbol{y} - \frac{1}{n}\boldsymbol{y'i}) = (\boldsymbol{M^0 y})'(\boldsymbol{M^0 y}) = \boldsymbol{x'M^0x},$$

where $\boldsymbol{M^0} = [\boldsymbol{I} - \frac{1}{n}\boldsymbol{ii'}]$. Recall that $(\boldsymbol{M^0})' = \boldsymbol{M^0}$ and $\boldsymbol{M^0} \cdot \boldsymbol{M^0} = \boldsymbol{M^0}$

$$\text{(d) } TSS_{x_k} = \sum (x_i - \bar{x})^2 = (\boldsymbol{x_k} - \frac{1}{n}\boldsymbol{x_k'i})'(\boldsymbol{x_k} - \frac{1}{n}\boldsymbol{x_k'i}) = (\boldsymbol{M^0x_k})'(\boldsymbol{M^0x_k}) = \boldsymbol{x_k'M^0x_k}$$

Exercise 2

Show (using matrix notation) that by construction in OLS $\boldsymbol{i'\hat{u}} = 0$.

$$\boldsymbol{i'\hat{u}} = \boldsymbol{i'(y - \hat{y})} = \boldsymbol{i'(y - X\hat{\beta})}, \text{ recall that } \boldsymbol{\hat{\beta} = (X'X)^{-1}X'y}. \text{ Hence,}$$

$$\boldsymbol{i'\hat{u}} = \boldsymbol{i'(y - X(X'X)^{-1}X'y)} = \boldsymbol{i'y(i - X(X'X)^{-1}X')} = \boldsymbol{i'y(i - i)} = 0, \text{ Q.E.D.}$$

Exercise 3

Show (using matrix notation) that by construction in OLS for any $N \cdot 1$ vector of observations of the k'th regressor $\boldsymbol{x}_k$, it is the case that $cov(\boldsymbol{x_k}, \boldsymbol{\hat{u}}) = 0$.

$$cov(\boldsymbol{x_k}, \boldsymbol{\hat{u}}) = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))(\boldsymbol{\hat{u}} - \mathbb{E}(\boldsymbol{\hat{u}}))'] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))(\boldsymbol{\hat{u}} - \boldsymbol{0})']$$

We have already showed that $\boldsymbol{i'\hat{u}} = 0$. This implies that $\boldsymbol{\hat{u}'i} = 0$ and, consequently, $\boldsymbol{(\hat{u})'} = 0$ as well. Hence,

$$cov(\boldsymbol{x_k}, \boldsymbol{\hat{u}}) = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})) \cdot 0] = 0, \text{ Q.E.D.}$$

Exercise 4

```r
library(haven)

data <- read_dta("/Users/herrhellana/Dropbox/_NYU studies/Quant I/home assignments/HW6/gendereq.dta")

library(sandwich)
library(stargazer)

# Model I w region FEs
m_1 <- lm(data=data, femjobs ~ as.factor(region))
heterosked1 <- vcovHC(m_1, type = 'HC1')

# Model II w region FEs + religiosity, GDP, and education
m_2 <- lm(data=data, femjobs ~ relig + gdp_k + univ + as.factor(region))
heterosked2 <- vcovHC(m_2, type = 'HC1')


stargazer(m_1, m_2, type = 'text',
          dep.var.caption = 'Models I and II',
          dep.var.labels.include = F,
          digits = 3,
          df = F,
          #dep.var.labels = c("% saying religion \"very important\"",
                            # "GDP per capita ($1,000s)",
                            # '% with university degree',
                            # 'Intercept'),
          covariate.labels = c("Religiosity",
                               "GDP", "Education", "Region 2",
                               "Region 3", "Region 4",
                               "Region 5", "Region 6",
                               'Intercept'),
          se = list(sqrt(diag(heterosked1)),
                    sqrt(diag(heterosked2))))
```

```
================================================
                          Models I and II
                    ----------------------------
                         (1)            (2)
------------------------------------------------
Religiosity                           -0.155*
                                      (0.088)

GDP                                   0.326***
                                      (0.120)

Education                              0.072
                                      (0.194)

Region 2              20.514***      21.892***
                       (6.197)        (7.402)

Region 3              13.624***        3.727
                       (4.535)        (4.974)
```

```
Region 4                39.566***      30.672***
                         (7.515)        (8.276)

Region 5                44.057***      37.844***
                         (3.013)        (5.811)

Region 6                38.308***      24.106***
                         (6.198)        (7.484)

Intercept               19.864***      28.426***
                         (2.134)        (9.841)

------------------------------------------------
Observations               60             60
R2                        0.547          0.680
Adjusted R2               0.505          0.630
Residual Std. Error      15.186         13.136
F Statistic             13.050***      13.548***
================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```r
library(dplyr)
data <- data %>% mutate(region2 = relevel(as.factor(region), ref = 6))

m_1a <- lm(data=data, femjobs ~ region2)
summary(m_1a)
```

```
Call:
lm(formula = femjobs ~ region2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.092  -8.170  -0.352   7.047  42.374

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   58.172      4.384  13.270  < 2e-16 ***
region21     -38.308      6.200  -6.179 8.79e-08 ***
region22     -17.794      8.083  -2.201    0.032 *
region23     -24.684      5.545  -4.451 4.31e-05 ***
region24       1.258      8.768   0.143    0.886
region25       5.749      7.222   0.796    0.430
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.19 on 54 degrees of freedom
Multiple R-squared:  0.5472,    Adjusted R-squared:  0.5052
F-statistic: 13.05 on 5 and 54 DF,  p-value: 2.452e-08
```

- The average value of gender equality support in the baseline Region 6 is 58.2, which is significant at the 99% confidence level meaning that support for gender equality in this region is greater than 0.
- The average value of gender equality support in Region 1 is 58.2-38.3=19.9, which is significant at the 99% confidence level meaning that support for gender equality in this region is less than in the baseline region.

- The average value of gender equality support in Region 2 is 58.2-17.8=40.4, which is significant at the 95% confidence level meaning that support for gender equality in this region is less than in the baseline region.
- The average value of gender equality support in Region 3 is 58.2-24.6=33.6, which is significant at the 99% confidence level meaning that support for gender equality in this region is less than in the baseline region.
- The average value of gender equality support in Region 4 is 58.2+1.25=59.45, which is not statistically significant meaning that support for gender equality in this region is not different from the baseline region.
- The average value of gender equality support in Region 5 is 58.2+5.75=63.95, which is not statistically significant meaning that support for gender equality in this region is not different from the baseline region.

Exercise 5

Models with polynomial regressors.

```
p_1 <- lm(data=data, femjobs ~ gdp_k)
p_2 <- lm(data=data, femjobs ~ gdp_k + I(gdp_k^2))
p_3 <- lm(data=data, femjobs ~ gdp_k + I(gdp_k^2) + I(gdp_k^3))

stargazer(p_1, p_2, p_3, type = 'text',
          omit.stat=c('f', 'rsq'),
          dep.var.caption = 'Polynomial Models',
          dep.var.labels.include = F,
          digits = 2,
          df = F,
          #dep.var.labels = c("% saying religion \"very important\"",
                            # "GDP per capita ($1,000s)",
                            # '% with university degree',
                            # 'Intercept'),
          covariate.labels = c("GDP/capita",
                               "(GDP/capita)2", "(GDP/capita)3",
                               'Intercept'),
          se = list(sqrt(diag(vcovHC(p_1, type = 'HC1'))),
                    sqrt(diag(vcovHC(p_2, type = 'HC1'))),
                    sqrt(diag(vcovHC(p_3, type = 'HC1')))))
```

```
==================================================
                         Polynomial Models
                    ------------------------------
                       (1)       (2)       (3)
--------------------------------------------------
GDP/capita           0.46***   1.70***   1.79**
                     (0.17)    (0.54)    (0.91)

(GDP/capita)2                  -0.02**    -0.02
                               (0.01)    (0.04)

(GDP/capita)3                            0.0000
                                        (0.0004)
```

```
Intercept             33.81***  26.24***  25.88***
                       (3.19)    (3.89)    (4.30)

--------------------------------------------------
Observations             60        60        60
Adjusted R2            0.15      0.23      0.22
Residual Std. Error   19.94     18.96     19.12
==================================================
Note:                 *p<0.1; **p<0.05; ***p<0.01
```

Exercise 6

```r
cps <- read_dta("/Users/herrhellana/Dropbox/_NYU studies/Quant I/home assignments/HW6/cpsnov2018abr.dta

cps <- cps %>% mutate(voted01 = case_when(voted == 2 ~ 1,
                                          voted == 1 ~ 0))

cps <- cps %>% mutate(faminc_new = case_when(faminc == 100 ~ 2500,
                                             faminc == 110 ~ 500,
                                             faminc == 111 ~ 250,
                                             faminc == 112 ~ 750,
                                             faminc == 120 ~ 1500,
                                             faminc == 121 ~ 1250,
                                             faminc == 122 ~ 1750,
                                             faminc == 130 ~ 2500,
                                             faminc == 131 ~ 2250,
                                             faminc == 132 ~ 2750,
                                             faminc == 140 ~ 3500,
                                             faminc == 141 ~ 3250,
                                             faminc == 142 ~ 3750,
                                             faminc == 150 ~ 4500,
                                             faminc == 200 ~ 6500,
                                             faminc == 210 ~ 6250,
                                             faminc == 220 ~ 5500,
                                             faminc == 230 ~ 7000,
                                             faminc == 231 ~ 6750,
                                             faminc == 232 ~ 6500,
                                             faminc == 233 ~ 7250,
                                             faminc == 234 ~ 7500,
                                             faminc == 300 ~ 8750,
                                             faminc == 310 ~ 7750,
                                             faminc == 320 ~ 8250,
                                             faminc == 330 ~ 8750,
                                             faminc == 340 ~ 8500,
                                             faminc == 350 ~ 9500,
                                             faminc == 400 ~ 12500,
                                             faminc == 410 ~ 10500,
                                             faminc == 420 ~ 11500,
                                             faminc == 430 ~ 11250,
                                             faminc == 440 ~ 11000,
                                             faminc == 450 ~ 12500,
```

```r
                                              faminc == 460 ~ 13500,
                                              faminc == 470 ~ 13750,
                                              faminc == 480 ~ 13500,
                                              faminc == 490 ~ 14500,
                                              faminc == 500 ~ 17500,
                                              faminc == 510 ~ 15500,
                                              faminc == 520 ~ 16500,
                                              faminc == 530 ~ 17500,
                                              faminc == 540 ~ 16250,
                                              faminc == 550 ~ 18750,
                                              faminc == 560 ~ 19000,
                                              faminc == 600 ~ 22500,
                                              faminc == 700 ~ 37500,
                                              faminc == 710 ~ 27500,
                                              faminc == 720 ~ 32500,
                                              faminc == 730 ~ 37500,
                                              faminc == 740 ~ 45000,
                                              faminc == 800 ~ 75000,
                                              faminc == 810 ~ 62500,
                                              faminc == 820 ~ 55000,
                                              faminc == 830 ~ 67500,
                                              faminc == 840 ~ 112500,
                                              faminc == 841 ~ 87500,
                                              faminc == 842 ~ 125000,
                                              faminc == 843 ~ 225000))



m_voted1 <- lm(data=cps, voted01 ~ faminc_new)
m_voted2 <- lm(data=cps, voted01 ~ log(faminc_new))
m_voted3 <- lm(data=cps, voted01 ~ faminc_new + I(faminc_new^2))
m_voted4 <- lm(data=cps, voted01 ~ faminc_new + I(faminc_new^2) + I(faminc_new^3))


stargazer(m_voted1, m_voted2, m_voted3, m_voted4,
          type = 'text', dep.var.caption = 'Voted',
          omit.stat = c('f', 'rsq'),
          digits = 3,
          dep.var.labels.include = F,
          covariate.labels = c('Income',
                               'Income (log)',
                               '(Income)2',
                               '(Income)3',
                               'Intercept'),
          se = list(sqrt(diag(vcovHC(m_voted1, type = 'HC1'))),
                    sqrt(diag(vcovHC(m_voted2, type = 'HC1'))),
                    sqrt(diag(vcovHC(m_voted3, type = 'HC1'))),
                    sqrt(diag(vcovHC(m_voted4, type = 'HC1')))))
```

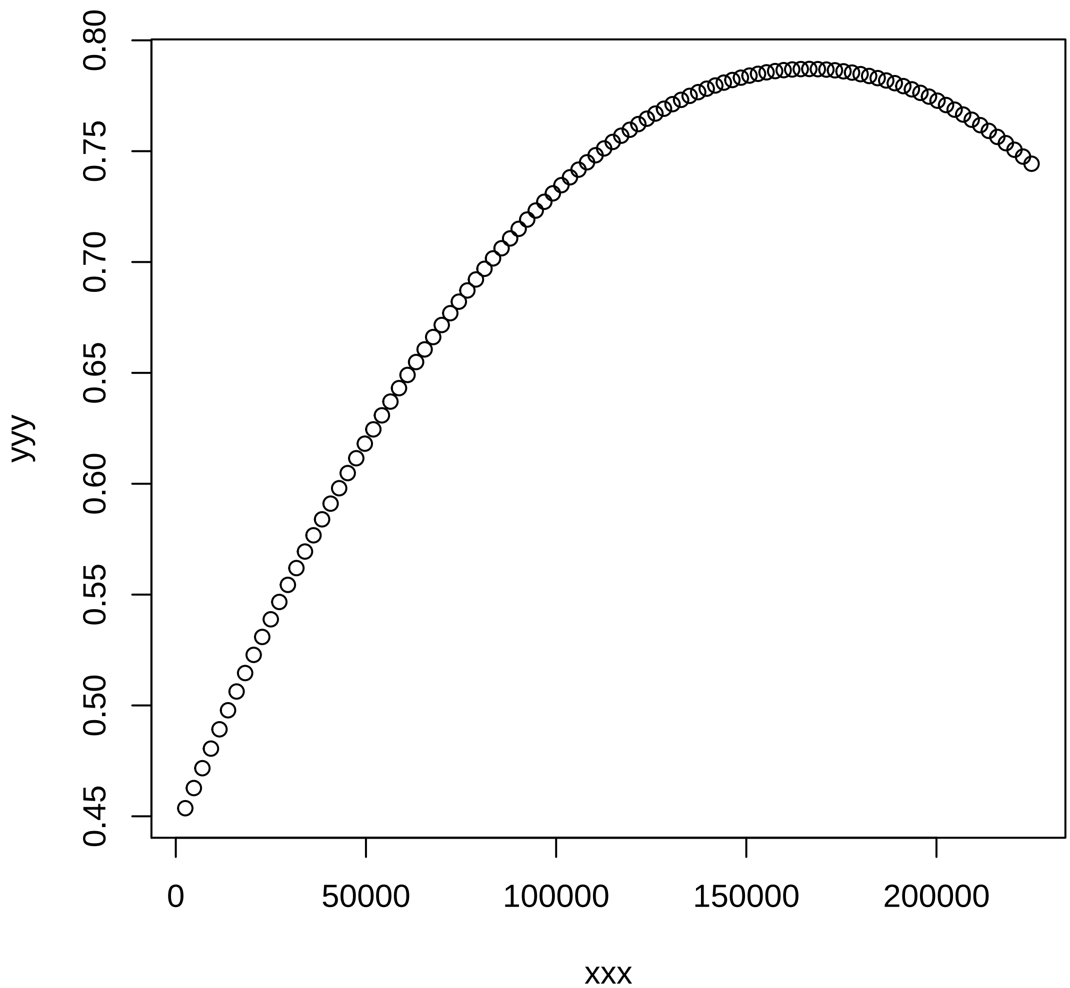
```
========================================================================================
                                                 Voted
                    --------------------------------------------------------------------
                          (1)              (2)              (3)              (4)
```

```
-----------------------------------------------------------------------------------
Income                0.00000***                     0.00000***        0.00000
                      (0.00000)                      (0.00000)        (0.00000)

Income (log)                           0.096***
                                       (0.019)

(Income)2                                             -0.000***         -0.000
                                                       (0.000)          (0.000)

(Income)3                                                               -0.000
                                                                        (0.000)

Intercept              0.550***         -0.409*        0.443***         0.451***
                       (0.030)          (0.213)        (0.047)          (0.075)

-----------------------------------------------------------------------------------
Observations             702              702            702              702
Adjusted R2             0.026            0.035          0.037            0.036
Residual Std. Error 0.471 (df = 700) 0.469 (df = 700) 0.469 (df = 699) 0.469 (df = 698)
===================================================================================
Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

There may be a non-linear relationship between income and turning out to vote: logarithmic or quadratic. Non-linear logarithmic relationship makes sense here because of the diminishing marginal return: each additional increase in income makes less difference in the economic status of a person. The effect in the logarithmic model is statistically significant at the 99% confidence level and the largest among the estimated models.

The inverse-U curve (see the plot below), i.e. the quadratic polynom: this relationship also makes substantive sense, given the positive FOC and the negative SOC that we got. We observe the majority of citizens in poor countries (esp. autocracies) being apolitical as they struggle financially on a daily basis and do not have time to care about politics or educate themselves about electoral process. As financial security of a person grows, the person can engage more and more in politics to fight for their rights and/or better quality of life (because they've got a privilege of having free of work time). However, extremely rich people have extremely privileged ways of living, these people may be just delighted about the social conditions they find themselves under. Hence, they can be socially detached and alienated from the majority of other citizens. Why care about politics and participate in elections (which is costly per se) if you are fully satisfied with your class position? (#EatTheRich)

```r
xxx <- seq(min(cps$faminc_new), max(cps$faminc_new), length.out = 100)
yyy <- 4.434e-01 + 4.132e-06*xxx + -1.242e-11*(xxx)^2

plot(xxx, yyy)
```



The cube relationship neither produces statistically significant results nor has it a meaningful and intuitive interpretation. In social sciences, it usually just overfits the data.

```
cps$age_new <- ifelse(cps$age >= 18 & cps$age <= 85, cps$age, NA)


m_voted1a <- lm(data=cps, voted01 ~ faminc_new + as.factor(age_new))
m_voted2a <- lm(data=cps, voted01 ~ log(faminc_new) + as.factor(age_new))
m_voted3a <- lm(data=cps, voted01 ~ faminc_new + I(faminc_new^2) + as.factor(age_new))
m_voted4a <- lm(data=cps, voted01 ~ faminc_new +
                  I(faminc_new^2) + I(faminc_new^3) +
                  as.factor(age_new))


stargazer(m_voted1a, m_voted2a, m_voted3a, m_voted4a,
          type = 'text', dep.var.caption = 'Voted',
          omit = "age_new",
          omit.stat=c('f', 'rsq'),
          digits = 3,
          dep.var.labels.include = F,
          covariate.labels = c('Income',
                               'Income (log)',
                               '(Income)2',
```

```r
                              '(Income)3',
                              'Intercept'),
          add.lines = list(c('Age fixed effects',
                            'Yes', 'Yes', 'Yes', 'Yes')),
          se = list(sqrt(diag(vcovHC(m_voted1a, type = 'HC1'))),
                    sqrt(diag(vcovHC(m_voted2a, type = 'HC1'))),
                    sqrt(diag(vcovHC(m_voted3a, type = 'HC1'))),
                    sqrt(diag(vcovHC(m_voted4a, type = 'HC1')))))
```

```
=======================================================================================
                                               Voted
                    -------------------------------------------------------------------
                           (1)               (2)               (3)               (4)
---------------------------------------------------------------------------------------
Income                 0.00000***                          0.00000***          0.00000
                       (0.00000)                           (0.00000)          (0.00000)

Income (log)                               0.104***
                                           (0.019)

(Income)2                                                  -0.000***           0.000
                                                            (0.000)           (0.000)

(Income)3                                                                      -0.000
                                                                              (0.000)

Intercept                0.215            -0.839***          0.082             0.107
                        (0.154)            (0.265)          (0.163)           (0.177)

---------------------------------------------------------------------------------------
Age fixed effects         Yes               Yes               Yes               Yes
Observations              702               702               702               702
Adjusted R2              0.093             0.101             0.105             0.104
Residual Std. Error 0.455 (df = 637) 0.453 (df = 637) 0.452 (df = 636) 0.452 (df = 635)
=======================================================================================
Note:                                                       *p<0.1; **p<0.05; ***p<0.01
```

After including fixed effects, intercepts' coefficients and their significance have changed. However, all signs and significance of independent variables coefficients remained unchanged. The magnitude of independent variables coefficients have only slightly changed – see the output:

```r
coef(m_voted1)[2]
```

```
  faminc_new
1.161331e-06
```

```r
coef(m_voted1a)[2]
```

```
  faminc_new
1.299348e-06
```

```r
coef(m_voted2)[2]
```

```
log(faminc_new)
     0.09625976
```

```r
coef(m_voted2a)[2]
```

```
log(faminc_new)
      0.1043129
```

```r
coef(m_voted3)[2:3]
```

```
     faminc_new I(faminc_new^2)
   4.132387e-06   -1.242198e-11
```

```r
coef(m_voted3a)[2:3]
```

```
     faminc_new I(faminc_new^2)
   4.394392e-06   -1.287592e-11
```

```r
coef(m_voted4)[2:4]
```

```
     faminc_new I(faminc_new^2) I(faminc_new^3)
   3.740047e-06   -7.886827e-12   -1.312436e-17
```

```r
coef(m_voted4a)[2:4]
```

```
     faminc_new I(faminc_new^2) I(faminc_new^3)
   3.151140e-06    1.524268e-12   -4.171328e-17
```

Coefficients for $\texttt{faminc}^2$ and $\texttt{faminc}^3$ have changed but they remained statistically not significant – so we do not really care about transformations in this model.

Overall, we do not observe any meaningful changes after including the fixed effects.