ARE 213 Problem Set 2A

Becky Cardinali, Yuen Ho, Sara Johns, and Jacob Lefler

Due 10/26/2020

Question 1

Question 10.3 from Wooldridge: For $T = 2$ consider the standard unobserved effects model:

$$y_{it} = \alpha + x_{it}\beta + c_i + u_{it} \tag{1}$$

Let $\hat{\beta}_{FE}$ and $\hat{\beta}_{FD}$ represent the fixed effects and first differences estimators respectively.

(a) Show that $\hat{\beta}_{FE}$ and $\hat{\beta}_{FD}$ are numerically identical. Hint: it may be easier to write $\hat{\beta}_{FE}$ as the "within estimator" rather than the fixed effects estimator.

(b) Show that the standard errors of $\hat{\beta}_{FE}$ and $\hat{\beta}_{FD}$ are numerically identical. If you wish, you may assume that x_{it} is a scalar (i.e. there is only one regressor) and ignore any degree of freedome corrections. You are not clustering the standard errors in this problem.

Question 2

Question 3

a -

Run pooled bivariate OLS. Interpret. Add year fixed effects. Interpret. Add all covariates that you believe are appropriate. Think carefully about which covariates should be log transformed and which should enter in levels. What happens when you add these covariates? Why?

```
# a- Pooled bivariate OLS , yr FE, All covariates

# create y variable
traffic[, ln_fat_pc := log((fatalities/population))]
# log covariates
traffic[,ln_unemploy := log(unemploy)]
traffic[,ln_totalvmt := log(totalvmt)]
traffic[,ln_precip := log(precip)]
traffic[,ln_snow := log(snow32+0.01)] # to avoid NA from zeroes
# create dummies for FEs (to be used later)
traffic <- dummy_cols(traffic, select_columns = c("year", "state"))

# bivariate OLS
biv <- feols(ln_fat_pc ~ primary + secondary, data=traffic)
summary(biv, se="standard")
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
```

```
## Standard-errors: Standard
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) -1.625000   0.016033 -101.3500 < 2.2e-16 ***
## primary     -0.222018   0.028046   -7.9162  5.84e-15 ***
## secondary   -0.140641   0.021510   -6.5383  9.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -332.47   Adj. R2: 0.06081
```

```
biv_yfe <- feols(ln_fat_pc ~ primary + secondary, fixef = "year", data=traffic)
summary(biv_yfe, se = "standard")
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
## Fixed-effects: year: 23
## Standard-errors: Standard
##            Estimate Std. Error   t value Pr(>|t|)
## primary   -0.086378   0.037159 -2.324500 0.020278 *
## secondary -0.008271   0.032443 -0.254946 0.798812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -304.06   Adj. R2: 0.08915
##                         R2-Within: 0.00834
```

```
biv_yfe_cov <- feols(ln_fat_pc ~ primary + college +
                       beer + ln_unemploy + ln_totalvmt + ln_precip +
                       ln_snow + rural_speed + urban_speed, fixef = "year", data=traffic)
summary(biv_yfe_cov, se = "standard")
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
## Fixed-effects: year: 23
## Standard-errors: Standard
##              Estimate Std. Error    t value  Pr(>|t|)
## primary     -0.007105   0.017712  -0.401160  0.688381
## college     -3.005600   0.175954 -17.082000 < 2.2e-16 ***
## beer         0.192174   0.029576   6.497700  1.24e-10 ***
## ln_unemploy -0.022254   0.027020  -0.823601  0.410345
## ln_totalvmt -0.063080   0.007658  -8.236900  4.99e-16 ***
## ln_precip   -0.089407   0.015351  -5.824100  7.54e-09 ***
## ln_snow     -0.069665   0.003610 -19.296000 < 2.2e-16 ***
## rural_speed  0.020363   0.002432   8.372100 < 2.2e-16 ***
## urban_speed  0.001837   0.001794   1.024100  0.305995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: 219.80   Adj. R2: 0.63821
##                        R2-Within: 0.60861
```

b -

Ignore omitted variables bias issues for the moment. Do you think the standard errors from above are right? Compute the Huber-White heteroskedasticity robust standard errors. Do they change much? Compute the clustered standard errors that are robust to within-state correlation. Do this using both the canned command and manually using the formulas we learned in class. Do the standard errors change much? Are you surprised? Interpret.

```r
# b - white robust and clustered

# package command - heteroskedastic
summary(biv, se = "white")
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
## Standard-errors: White
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) -1.625000   0.015258 -106.5000 < 2.2e-16 ***
## primary     -0.222018   0.028474   -7.7972  1.44e-14 ***
## secondary   -0.140641   0.021134   -6.6546  4.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -332.47   Adj. R2: 0.06081
```

```r
summary(biv_yfe, se = "white")
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
## Fixed-effects: year: 23
## Standard-errors: White
##            Estimate Std. Error   t value Pr(>|t|)
## primary   -0.086378   0.038634 -2.235800 0.025563 *
## secondary -0.008271   0.031955 -0.258838 0.795808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -304.06   Adj. R2: 0.08915
##                         R2-Within: 0.00834
```

```r
summary(biv_yfe_cov, se = "white")
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
## Fixed-effects: year: 23
## Standard-errors: White
##              Estimate Std. Error    t value  Pr(>|t|)
## primary     -0.007105   0.014468  -0.491105   0.62345
## college     -3.005600   0.170652 -17.612000 < 2.2e-16 ***
## beer         0.192174   0.027053   7.103700  2.18e-12 ***
## ln_unemploy -0.022254   0.027104  -0.821045    0.4118
## ln_totalvmt -0.063080   0.009123  -6.914200  7.98e-12 ***
## ln_precip   -0.089407   0.016890  -5.293600  1.45e-07 ***
## ln_snow     -0.069665   0.003296 -21.135000 < 2.2e-16 ***
## rural_speed  0.020363   0.002409   8.453000 < 2.2e-16 ***
## urban_speed  0.001837   0.001796   1.023000  0.306553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: 219.80   Adj. R2: 0.63821
##                         R2-Within: 0.60861
```

```r
# package command - cluster
summary(biv, cluster  = traffic$state)
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
```

```
## Standard-errors: Clustered
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -1.625000   0.046411 -35.0140 < 2.2e-16 ***
## primary     -0.222018   0.090393  -2.4561  0.014195 *
## secondary   -0.140641   0.035964  -3.9107   9.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -332.47   Adj. R2: 0.06081
```

```
summary(biv_yfe, cluster = traffic$state)
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
## Fixed-effects: year: 23
## Standard-errors: Clustered
##             Estimate Std. Error   t value Pr(>|t|)
## primary   -0.086378   0.134724 -0.641150 0.521559
## secondary -0.008271   0.079979 -0.103418 0.917650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -304.06   Adj. R2: 0.08915
##                          R2-Within: 0.00834
```

```
summary(biv_yfe_cov, cluster = traffic$state)
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 1,127
## Fixed-effects: year: 23
## Standard-errors: Clustered
##               Estimate Std. Error   t value  Pr(>|t|)
## primary     -0.007105   0.033285 -0.213469  0.831001
## college     -3.005600   0.508200 -5.914200  4.45e-09 ***
## beer         0.192174   0.081483  2.358500  0.018526 *
## ln_unemploy -0.022254   0.068644 -0.324187  0.745858
## ln_totalvmt -0.063080   0.035634 -1.770200  0.076967 .
## ln_precip   -0.089407   0.058891 -1.518200  0.129261
## ln_snow     -0.069665   0.008290 -8.402900 < 2.2e-16 ***
## rural_speed  0.020363   0.004489  4.536500  6.35e-06 ***
## urban_speed  0.001837   0.003479  0.528160  0.597496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: 219.80   Adj. R2: 0.63821
##                         R2-Within: 0.60861
```

```
# write own commands

# will need to get beta matrix manually
calc.beta <- function(xmat, ymat) {
  (solve(t(xmat)%*%xmat)) %*% (t(xmat)%*%ymat)
}

white_middle <- function(xmat, ymat, beta) {
  residsq <- diag(as.vector((ymat - xmat %*% beta)^2))
  mid <- (t(xmat)%*%residsq%*%xmat)
  return(mid)
}
```

```
robust.se <- function(xmat, middle) {

  var.robust <- solve(t(xmat)%*%xmat) %*% middle %*% solve(t(xmat)%*%xmat)

  se <- sqrt(diag(var.robust))

  return(se)
}

cluster_middle <- function(i, beta, DT, yvar, xvars) {

  state.xmat <- as.matrix(cbind(1,select(DT[state == i,], xvars)))
  state.ymat <- as.matrix(select(DT[state == i,], yvar))

  resid <- as.vector(state.ymat - state.xmat %*% beta)

  middle.term <- t(state.xmat) %*% resid %*% t(resid) %*% state.xmat

  return(middle.term)
}

# List of our variables for the three regressions
biv_var <- c("primary", "secondary")
biv_yfe_var <- c("primary", "secondary", colnames(traffic[,year_1982:year_2003]))
biv_yfe_cov_var <- c("primary", "secondary", "college", "beer",
                     "ln_unemploy", "ln_totalvmt", "ln_precip",
                     "ln_snow", "rural_speed", "urban_speed", colnames(traffic[,year_1982:year_2003]))

# Run regression
xmat_biv <- as.matrix(cbind(1,select(traffic, all_of(biv_var))))
xmat_biv_yfe <- as.matrix(cbind(1, select(traffic, all_of(biv_yfe_var))))
xmat_biv_yfe_cov <- as.matrix(cbind(1, select(traffic, all_of(biv_yfe_cov_var))))
ymat <- as.matrix(select(traffic, ln_fat_pc))

beta_biv <- calc.beta(xmat_biv, ymat)
beta_biv_yfe <- calc.beta(xmat_biv_yfe, ymat)
beta_biv_yfe_cov <- calc.beta(xmat_biv_yfe_cov, ymat)

# White robust
# get middle terms
w_mid_biv <- white_middle(xmat_biv, ymat, beta_biv)
w_mid_biv_yfe <- white_middle(xmat_biv_yfe, ymat, beta_biv_yfe)
w_mid_biv_yfe_cov <- white_middle(xmat_biv_yfe_cov, ymat, beta_biv_yfe_cov)
# get standard errors
white_biv <- robust.se(xmat_biv, w_mid_biv)
white_biv
```

```
##         V1    primary  secondary
## 0.01523731 0.02843620 0.02110615
```

```
white_biv_yfe <- robust.se(xmat_biv_yfe, w_mid_biv_yfe)
white_biv_yfe[1:3]
```

```
##         V1    primary  secondary
## 0.04384825 0.03820280 0.03159891
```

```r
white_biv_yfe_cov <- robust.se(xmat_biv_yfe_cov, w_mid_biv_yfe_cov)
white_biv_yfe_cov[1:11]
```

```
##          V1     primary   secondary     college        beer ln_unemploy
## 0.167454386 0.022443850 0.019771083 0.168086198 0.026255527 0.026718615
## ln_totalvmt   ln_precip     ln_snow rural_speed urban_speed
## 0.009568140 0.016899622 0.003248055 0.002362976 0.001757033
```

```r
# Clustered by state
states <- as.vector(unique(traffic[,state]))

cl_mid_biv_terms <- mclapply(states, cluster_middle, beta = beta_biv, DT = traffic,
                             yvar="ln_fat_pc", xvars=biv_var, mc.cores = core.num)
cl_mid_biv <- Reduce('+', cl_mid_biv_terms)

cl_mid_biv_yfe_terms <- mclapply(states, cluster_middle, beta = beta_biv_yfe, DT = traffic,
                                 yvar="ln_fat_pc", xvars=biv_yfe_var, mc.cores = core.num)
cl_mid_biv_yfe <- Reduce('+', cl_mid_biv_yfe_terms)

cl_mid_biv_yfe_cov_terms <- mclapply(states, cluster_middle, beta = beta_biv_yfe_cov, DT = traffic,
                                     yvar="ln_fat_pc", xvars=biv_yfe_cov_var, mc.cores = core.num)
cl_mid_biv_yfe_cov <- Reduce('+', cl_mid_biv_yfe_cov_terms)

cl_biv <- robust.se(xmat_biv, cl_mid_biv)
cl_biv
```

```
##         V1    primary  secondary
## 0.04589383 0.08938682 0.03556304
```

```r
cl_biv_yfe <- robust.se(xmat_biv_yfe, cl_mid_biv_yfe)
cl_biv_yfe[1:3]
```

```
##         V1    primary  secondary
## 0.04384825 0.13191304 0.07831071
```

```r
cl_biv_yfe_cov <- robust.se(xmat_biv_yfe_cov, cl_mid_biv_yfe_cov)
cl_biv_yfe_cov[1:11]
```

```
##          V1     primary   secondary     college        beer ln_unemploy
## 0.502704959 0.044755718 0.038382908 0.492029109 0.077636743 0.066674947
## ln_totalvmt   ln_precip     ln_snow rural_speed urban_speed
## 0.036514564 0.057562915 0.008064637 0.004362628 0.003323011
```

c -

Compute the between estimator, both with and without covariates. Under what conditions will this give an unbiased estimate of the effect of primary seat belt laws on fatalities per capita? Do you believe those conditions are met? Are you concerned about the standard errors in this case?

```r
# c - between estimator with and without covariates
traffic_bet <- traffic[, lapply(.SD, mean), by = "state"] # get means by state

between <- feols(ln_fat_pc ~ primary + secondary, data=traffic_bet)
summary(between, se="standard")
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 49
## Standard-errors: Standard
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) -1.674000   0.171207 -9.777500 8.29000e-13 ***
## primary     -0.125092   0.260865 -0.479530 6.33834e-01
## secondary   -0.071135   0.275735 -0.257982 7.97571e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: -10.49   Adj. R2: -0.0373
```

```
between_cov <- feols(ln_fat_pc ~ primary + secondary + college +
                       beer + ln_unemploy + ln_totalvmt + ln_precip +
                       ln_snow + rural_speed + urban_speed, data=traffic_bet)
summary(between_cov, se = "standard")
```

```
## OLS estimation, Dep. Var.: ln_fat_pc
## Observations: 49
## Standard-errors: Standard
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) -4.809700   1.116600 -4.307500 0.000112 ***
## primary      0.321236   0.174393  1.842000 0.073285 .
## secondary    0.233456   0.164156  1.422200 0.163135
## college     -2.097700   0.718585 -2.919200 0.005869 **
## beer         0.112346   0.111615  1.006500 0.320518
## ln_unemploy  0.116680   0.134577  0.867012 0.391378
## ln_totalvmt -0.122812   0.032914 -3.731300 0.000621 ***
## ln_precip    0.058401   0.077797  0.750683 0.457467
## ln_snow     -0.069262   0.016324 -4.242900 0.000136 ***
## rural_speed  0.067399   0.017554  3.839500 0.000453 ***
## urban_speed -0.002351   0.012843 -0.183063 0.855722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-likelihood: 30.89   Adj. R2: 0.76812
```

d -

Compute the RE estimator (including covariates). Under what conditions will this give an unbiased estimate of the effect of primary seat belt laws on fatalities per capita? What are its advantages or disadvantages as compared to pooled OLS?

```
# d - random effects estimator
random <- plm(ln_fat_pc ~ primary + secondary + college +
                beer + ln_unemploy + ln_totalvmt + ln_precip +
                ln_snow + rural_speed + urban_speed, data=traffic, model="random")
summary(random)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = ln_fat_pc ~ primary + secondary + college + beer +
##     ln_unemploy + ln_totalvmt + ln_precip + ln_snow + rural_speed +
##     urban_speed, data = traffic, model = "random")
##
## Balanced Panel: n = 49, T = 23, N = 1127
##
## Effects:
```

```
##                     var  std.dev share
## idiosyncratic 0.008127 0.090151 0.279
## individual    0.021041 0.145054 0.721
## theta: 0.8715
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.3721113 -0.0612727  0.0058319  0.0665328  0.3316814
##
## Coefficients:
##                Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept) -1.11123728  0.20996847  -5.2924 1.207e-07 ***
## primary     -0.13636132  0.01488872  -9.1587 < 2.2e-16 ***
## secondary   -0.05599048  0.01043423  -5.3660 8.048e-08 ***
## college     -1.49468923  0.17199787  -8.6902 < 2.2e-16 ***
## beer         0.76046185  0.03768054  20.1818 < 2.2e-16 ***
## ln_unemploy -0.16002704  0.01453137 -11.0125 < 2.2e-16 ***
## ln_totalvmt -0.06931491  0.02031308  -3.4123 0.0006441 ***
## ln_precip   -0.06959186  0.01745034  -3.9880 6.663e-05 ***
## ln_snow     -0.00533264  0.00297623  -1.7917 0.0731739 .
## rural_speed -0.00528299  0.00096709  -5.4628 4.687e-08 ***
## urban_speed  0.00263347  0.00085253   3.0890 0.0020084 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    26.797
## Residual Sum of Squares: 10.631
## R-Squared:      0.60326
## Adj. R-Squared: 0.5997
## Chisq: 1696.92 on 10 DF, p-value: < 2.22e-16
```

e -

Do you think the standard errors from RE are right? Compute the clustered standard errors. Are they substantially different? If so, why? (i.e., what assumption(s) are being violated?)

```r
# e - clustered SEs
coeftest(random, vcovHC(random, type="sss", cluster="group"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -1.1112373  0.3822252  -2.9073 0.0037179 **
## primary     -0.1363613  0.0284185  -4.7983 1.817e-06 ***
## secondary   -0.0559905  0.0181115  -3.0914 0.0020413 **
## college     -1.4946892  0.2992997  -4.9940 6.860e-07 ***
## beer         0.7604618  0.0653529  11.6362 < 2.2e-16 ***
## ln_unemploy -0.1600270  0.0140974 -11.3516 < 2.2e-16 ***
## ln_totalvmt -0.0693149  0.0343015  -2.0208 0.0435436 *
## ln_precip   -0.0695919  0.0232735  -2.9902 0.0028493 **
## ln_snow     -0.0053326  0.0039183  -1.3610 0.1737976
## rural_speed -0.0052830  0.0014609  -3.6162 0.0003124 ***
## urban_speed  0.0026335  0.0013984   1.8832 0.0599368 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```