ARE 213 PS 2a

S. Sung, H. Husain, T. Woolley, A. Watt

2021-11-08

Contents

Packages

```
library(tidyverse)
library(haven)
library(plm)
library(lmtest)
library(sandwich)
library(stargazer)
library(ggplot2)
library(gridExtra)
library(grid)
library(gtable)
library(tinytex)
library(fastDummies)
```

Problem 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.

Part (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.

Writing $\hat{\beta}_{FE}$ as the within estimator, $\hat{\beta}_{FE}$ and $\hat{\beta}_{FD}$ are given by

$$\hat{\beta}_{FD} = \left(\Delta X' \Delta X\right)^{-1} \left(\Delta X' \Delta y\right) \qquad \text{and} \qquad \hat{\beta}_{FE} = \left(\ddot{X}' \ddot{X}\right)^{-1} \left(\ddot{X}' \ddot{y}\right)$$

Expanding the inner products, we have

$$\hat{\beta}_{FD} = \left(\sum_i \sum_t \Delta X'_{it} \Delta X_{it}\right)^{-1} \left(\sum_i \sum_t \Delta X'_{it} \Delta y_{it}\right)$$

and

$$\hat{\beta}_{FE} = \left(\sum_i \sum_t \ddot{X}'_{it} \ddot{X}_{it}\right)^{-1} \left(\sum_i \sum_t \ddot{X}'_{it} \ddot{y}_{it}\right)$$

Since there are only two periods, $\hat{\beta}_{FD}$ simplifies to

$$\hat{\beta}_{FD} = \left(\sum_i \Delta X'_i \Delta X_i\right)^{-1} \left(\sum_i \Delta X'_i \Delta y_i\right)$$

where

$$\Delta X_i \equiv X_{i2} - X_{i1} \qquad \text{and} \qquad \Delta y_i \equiv y_{i2} - y_{i1}$$

Now we note that

$$\ddot{X}_{i1} = X_{i1} - \frac{1}{2}(X_{i1} + X_{i2}) = \frac{1}{2}(X_{i1} - X_{i2}) = -\frac{1}{2}\Delta X_i$$

and similiarly

$$\ddot{X}_{i2} = \frac{1}{2}\Delta X_i, \quad \ddot{y}_{i1} = -\frac{1}{2}\Delta y_i \quad \ddot{y}_{i2} = \frac{1}{2}\Delta y_i$$

Then, $\hat{\beta}_{FE}$ becomes

$$\begin{aligned}
\hat{\beta}_{FE} &= \left(\sum_i \sum_t \ddot{X}'_{it} \ddot{X}_{it} \right)^{-1} \left(\sum_i \sum_t \ddot{X}'_{it} \ddot{y}_{it} \right) \\
&= \left(\sum_i \frac{1}{4} \Delta X'_i \Delta X_i + \frac{1}{4} \Delta X'_i \Delta X_i \right)^{-1} \left(\sum_i \frac{1}{4} \Delta X'_i \Delta y_i + \frac{1}{4} \Delta X'_i \Delta y_i \right) \\
&= \left(\frac{1}{2} \sum_i \Delta X'_i \Delta X_i \right)^{-1} \left(\frac{1}{2} \sum_i \Delta X'_i \Delta y_i \right) \\
&= \left(\sum_i \Delta X'_i \Delta X_i \right)^{-1} \left(\sum_i \Delta X'_i \Delta y_i \right) \\
&= \hat{\beta}_{FD}
\end{aligned}$$

So $\hat{\beta}_{FE}$ and $\hat{\beta}_{FD}$ are numerically identical.

Part (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 freedom corrections. You are not clustering the standard errors in this problem.

The standard errors are estimates of the square root of the asymptotic variances of our estimators, so WLOG, we can compare the asymptotic variances. The asymptotic variances of our estimators are

$$\widehat{Avar(\hat{\beta}_{FE})} = \hat{\sigma}^2_{u,FE}\left(\ddot{X}'\ddot{X}\right)^{-1} \qquad \text{and} \qquad \widehat{Avar(\hat{\beta}_{FE})} = \hat{\sigma}^2_{u,FD}\left(\Delta X'\Delta X\right)^{-1}$$

where $\hat{\sigma}^2_{u,FE}$ and $\hat{\sigma}^2_{u,FD}$ are estimated from the residuals of the corresponding regressions and using the correct degrees of freedom:

$$\hat{\sigma}^2_{u,FE} = \frac{\sum_i \sum_t \hat{\ddot{u}}^2_{it}}{N(T-1)-K} = \frac{\sum_i \sum_t \hat{\ddot{u}}^2_{it}}{N-K} \qquad \text{and} \qquad \hat{\sigma}^2_{u,FD} = \frac{\sum_i \widehat{\Delta u_i}^2}{N(T-1)-K} = \frac{\sum_i \widehat{\Delta u_i}^2}{N-K}$$

Let $\hat{\beta} := \hat{\beta}_{FD} = \hat{\beta}_{FE}$. Then, from part (a), we can find the relationship between $\widehat{\Delta u_i}$ and $\hat{\ddot{u}}_{it}$:

$$\begin{aligned}
\hat{\ddot{u}}^2_{it} &= (\ddot{y}_{it} - \ddot{X}_{it}\hat{\beta})^2 \\
&= \left((-1)^t\left(\frac{1}{2}\Delta y_i - \frac{1}{2}\Delta X_i\hat{\beta}\right)\right)^2 \\
&= \frac{1}{4}\left(\Delta y_i - \Delta X_i\hat{\beta}\right)^2 \\
&= \frac{1}{4}\widehat{\Delta u_i}^2
\end{aligned}$$

So the estimated error variances are related by

$$\begin{aligned}
\hat{\sigma}^2_{u,FE} &= \frac{\sum_i \sum_t \hat{\ddot{u}}^2_{it}}{N-K} \\
&= \frac{\sum_i \sum_t \frac{1}{4}\widehat{\Delta u_i}^2}{N-K} \\
&= \frac{\sum_i \frac{1}{2}\widehat{\Delta u_i}^2}{N-K} \\
&= \frac{1}{2}\frac{\sum_i \widehat{\Delta u_i}^2}{N-K} \\
&= \frac{1}{2}\hat{\sigma}^2_{u,FD}
\end{aligned}$$

We know from part (a) that

$$\begin{aligned}
\left(\sum_i \sum_t \ddot{X}'_{it}\ddot{X}_{it}\right)^{-1} &= \left(\frac{1}{2}\sum_i \Delta X'_i \Delta X_i\right)^{-1} \\
&= 2\left(\sum_i \Delta X'_i \Delta X_i\right)^{-1}
\end{aligned}$$

And putting all these together, we have

$$\begin{aligned}
\widehat{Avar(\hat{\beta}_{FE})} &= \hat{\sigma}^2_{u,FE} \left(\ddot{X}' \ddot{X} \right)^{-1} \\
&= \frac{1}{2} \hat{\sigma}^2_{u,FD} 2 \left(\sum_i \Delta X_i' \Delta X_i \right)^{-1} \\
&= \hat{\sigma}^2_{u,FD} \left(\sum_i \Delta X_i' \Delta X_i \right)^{-1} \\
&= \widehat{Avar(\hat{\beta}_{FE})}
\end{aligned}$$

Because the estimates of the asymptotic variances are equal, the standard errors (the square roots) will be equal.

Problem 2

Question 21-3 from Cameron-Trivedi (enhanced): Consider the fixed effects, two-way error component panel data model:

$$y_{it} = \alpha + x_{it}\beta + \mu_i + \lambda_t + \epsilon_{it} \tag{2}$$

Part (a)

Show that the fixed effects estimator of β can be obtained by applying two within (one-way) transformations on this model. The first is the within transformation ignoring the time effects followed by the within transformation ignoring the individual effects. Assume the panel is balanced. (Hint: it may be easier to analyze the fixed effects regression using partitioned regression.)

We want to show that

$$\ddot{\dot{y}}_{it} = \ddot{\dot{x}}_{it}\beta + \ddot{\dot{e}}_{it}$$

Ignoring time effects, we get

$$\ddot{y}_{it} = y_{it} - \overline{y_i} = y_{it} - \frac{1}{T}\sum_t y_{it}$$

$$\implies \ddot{y}_{it} = y_{it} - \frac{1}{T}\sum_t (\alpha + x_{it}\beta + \mu_i + \lambda_t + \epsilon_{it})$$

Applying the second within transformation, we get that

$$\begin{aligned}
\ddot{\dot{y}}_{it} &= \ddot{y}_{it} - \overline{\ddot{y}_{it}} \\
&= \ddot{y}_{it} - \frac{1}{N}\sum_i \ddot{y}_{it} \\
&= y_{it} - \overline{y_i} - \overline{y_t} + \overline{y} \\
&= \alpha + X_{it}\beta + \mu_i + \lambda_t + \epsilon_{it} \\
&\quad - \left(\frac{1}{T}\sum_t (\alpha + X_{it}\beta + \mu_i + \lambda_t + \epsilon_{it})\right) \\
&\quad - \left(\frac{1}{N}\sum_i (\alpha + X_{it}\beta + \mu_i + \lambda_t + \epsilon_{it})\right) \\
&\quad + \left(\frac{1}{NT}\sum_i\sum_t (\alpha + X_{it}\beta + \mu_i + \lambda_t + \epsilon_{it})\right) \\
&= \beta\left(x_{it} - \overline{x_i} - \overline{x_t} + \overline{x}\right) + \epsilon_{it} - \overline{\epsilon_i} - \overline{\epsilon_t} + \overline{\epsilon}
\end{aligned}$$

Since

$$\ddot{\ddot{x}}_{it} = x_{it} - \overline{x_i} - \overline{x_t} + \overline{x}$$

and

$$\dddot{e}_{it} = e_{it} - \overline{e_i} - \overline{e_t} + \overline{e}$$

We get that

$$\dddot{y}_{it} = \dddot{x}_{it}\beta + \dddot{e}_{it}$$

Part (b)

Show that the order of the transformations is unimportant. Give an intuitive explanation for why.

Reversing the order, we can show that we get the same result

Again, we want to show that

$$\dddot{y}_{it} = \dddot{x}_{it}\beta + \dddot{e}_{it}$$

Ignoring individual effects, we get

$$\ddot{y}_{it} = y_{it} - \overline{y_t} = y_{it} - \frac{1}{N}\sum_i y_{it}$$

$$\implies \ddot{y}_{it} = y_{it} - \frac{1}{N}\sum_i (\alpha + x_{it}\beta + \mu_i + \lambda_t + \epsilon_{it})$$

Applying the second within transformation, we get that

$$\begin{aligned}
\dddot{y}_{it} &= \ddot{y}_{it} - \overline{\ddot{y}_{it}} \\
&= \ddot{y}_{it} - \frac{1}{T}\sum_i \ddot{y}_{it} \\
&= y_{it} - \overline{y_t} - \overline{y_i} + \overline{y} \\
&= \alpha + X_{it}\beta + \mu_i + \lambda_t + \epsilon_{it} \\
&\quad - \left(\frac{1}{N}\sum_i (\alpha + X_{it}\beta + \mu_i + \lambda_t + \epsilon_{it})\right) \\
&\quad - \left(\frac{1}{T}\sum_t (\alpha + X_{it}\beta + \mu_i + \lambda_t + \epsilon_{it})\right) \\
&\quad + \left(\frac{1}{NT}\sum_i\sum_t (\alpha + X_{it}\beta + \mu_i + \lambda_t + \epsilon_{it})\right) \\
&= \beta\left(x_{it} - \overline{x_t} - \overline{x_i} + \overline{x}\right) + \epsilon_{it} - \overline{\epsilon_t} - \overline{\epsilon_i} + \overline{\epsilon}
\end{aligned}$$

Since

$$\dddot{x}_{it} = x_{it} - \overline{x_t} - \overline{x_i} + \overline{x}$$

and

$$\ddot{\ddot{e}}_{it} = e_{it} - \overline{e_t} - \overline{e_i} + \overline{e}$$

We get that

$$\ddot{\ddot{y}}_{it} = \ddot{\ddot{x}}_{it}\beta + \ddot{\ddot{e}}_{it}$$

Intuitively, the order of the transformations is unimportant because in the end, were still manage to difference out the individual and time effects. There is nothing particular to individual or time effects that would warrant removal in any particular order.

Part (c)

Does your answer to part (a) change if the panel becomes unbalanced (i.e., contains different numbers of observations for each individual i). Why or why not?

Problem 3

We now begin with an actual analysis of the data. The goal here is to determine what effect, if any, primary belt laws have on the log of traffic fatalities per capita (we log the LHS variable because we believe the effect of safety belt laws should be proportional to the overall level of fatalities per capita).

```
data = read_dta('traffic_safety2.dta') %>%
    filter(state != 99) %>%
    mutate(fatal_per_cap = fatalities / population,
           vmt_per_cap = totalvmt/population)
```

Part (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?

```
df <- data %>%
        mutate(fat_pc = fatalities/population,
               ln_fat_pc = log(fat_pc),
               ln_tvmt_pc = log(totalvmt/population))

reg3a_1 <- lm(ln_fat_pc ~ primary, data = df)
summary(reg3a_1)
```

```
##
## Call:
## lm(formula = ln_fat_pc ~ primary, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07553 -0.21571  0.03105  0.23493  1.08298
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -1.70492    0.01098 -155.249  < 2e-16 ***
## primary     -0.15490    0.02705   -5.727 1.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3335 on 1102 degrees of freedom
## Multiple R-squared:  0.0289, Adjusted R-squared:  0.02802
## F-statistic:  32.8 on 1 and 1102 DF,  p-value: 1.317e-08
```

```
stargazer(reg3a_1, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
```

```
##                              ln_fat_pc
## ---------------------------------------------
## primary                      -0.155***
##                               (0.027)
##
## Constant                     -1.705***
##                               (0.011)
##
## ---------------------------------------------
## Observations                   1,104
## R2                             0.029
## Adjusted R2                    0.028
## Residual Std. Error      0.333 (df = 1102)
## F Statistic          32.799*** (df = 1; 1102)
## =============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
reg3a_2 <- lm(ln_fat_pc ~ primary + factor(year), data = df)
stargazer(reg3a_2, type = "text", no.space = TRUE, omit.stat=c("f", "ser"))
```

```
##
## ===========================================
##                       Dependent variable:
##                   ------------------------
##                              ln_fat_pc
## -------------------------------------------
## primary                     -0.091***
##                              (0.028)
## factor(year)1982             -0.122*
##                              (0.066)
## factor(year)1983            -0.158**
##                              (0.066)
## factor(year)1984            -0.149**
##                              (0.066)
## factor(year)1985            -0.165**
##                              (0.066)
## factor(year)1986             -0.111*
##                              (0.066)
## factor(year)1987             -0.111*
##                              (0.066)
## factor(year)1988             -0.101
##                              (0.066)
## factor(year)1989            -0.160**
##                              (0.066)
## factor(year)1990            -0.182***
##                              (0.066)
## factor(year)1991            -0.255***
##                              (0.066)
## factor(year)1992            -0.313***
##                              (0.066)
## factor(year)1993            -0.311***
##                              (0.066)
## factor(year)1994            -0.317***
##                              (0.066)
```

```
## factor(year)1995          -0.293***
##                            (0.066)
## factor(year)1996          -0.304***
##                            (0.066)
## factor(year)1997          -0.291***
##                            (0.066)
## factor(year)1998          -0.308***
##                            (0.066)
## factor(year)1999          -0.313***
##                            (0.066)
## factor(year)2000          -0.333***
##                            (0.067)
## factor(year)2001          -0.329***
##                            (0.067)
## factor(year)2002          -0.318***
##                            (0.067)
## factor(year)2003          -0.329***
##                            (0.067)
## Constant                  -1.486***
##                            (0.047)
## ------------------------------------------
## Observations                1,104
## R2                          0.107
## Adjusted R2                 0.088
## ==========================================
## Note:            *p<0.1; **p<0.05; ***p<0.01
```

```
df.p <- pdata.frame(df, index = c("state", "year")) #declaring panel dataset
reg3a_2_p <- plm(ln_fat_pc ~ primary, data = df.p, effect = "time", model = "within") #checking reg3a_2
stargazer(reg3a_2_p, type = "text", no.space = TRUE, omit.stat=c("f", "ser"))
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                        ln_fat_pc
## ----------------------------------------
## primary                -0.091***
##                         (0.028)
## ----------------------------------------
## Observations             1,104
## R2                       0.010
## Adjusted R2             -0.011
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

total vehicle miles traveled per capita: logged because, after using per capita variables, we think the increase in fatalities per capita would be proportional to a percentage point increase in vehicle miles traveled per capita, not absolute levels, since the percentage point increase is a better measure of deviation from the norm, and drivers are more likely to adjust poorly to a deviation from the norm than an absolute increase in vehicle miles traveled (since an absolute increase may not be very differnt from the norm in states that have relatively large average miles traveled.)

Not inlcuding population because we are using per-capita variables

Log precip because we car about percentage-point deviation from the norm. Don't log snow because it has zeros.

When we add the covariates, the estimated effect on primary switches from significantly negative to insignificantly positive. Since we haven't yet controlled for state-level fixed effects, this could be from trying to compare states that have similar observable covariates that adopt primary safety belt laws at different times, but may have significant reasons for adopting when they did, and thus may have significant unobserved factors that are correlated with the primary safety belt law being in place.

```
# Bivariate with lm
reg_a_bivariate_lm = lm(log(fatal_per_cap) ~ primary , data = data)
# Bivariate with plm
reg_a_bivariate_plm = plm(log(fatal_per_cap) ~ primary,
                      data = data,
                      model = "pooling")
# FE with lm
reg_a_yfe_lm = lm(log(fatal_per_cap) ~ primary + factor(year), data = data)
# FE with plm
reg_a_yfe_plm = plm(log(fatal_per_cap) ~ primary,
                data = data,
                index = c("year"),
                model = "within",
                effect = "individual")
# FE + covars with lm
reg_a_full_lm = lm(log(fatal_per_cap) ~ primary + factor(year) + college + beer
                   + secondary + unemploy + log(vmt_per_cap) + log(precip) + snow32
                   + log(rural_speed) + log(urban_speed), data = data)
# FE + covars with plm
reg_a_full_plm = plm(log(fatal_per_cap) ~ primary + factor(year) + college + beer
                     + secondary + unemploy + log(vmt_per_cap) + log(precip) + snow32
                     + log(rural_speed) + log(urban_speed),
                     data = data,
                     index = c("state", "year"),  # order matters: unit-var, time-var
                     model = "within",
                     effect = "time")  # only do time


# Checking in-state variation in rural_speed
# d = data %>% group_by(state) %>% mutate(rural_speed_dev = abs(rural_speed - mean(rural_speed)), rural
# plot(d$state, d$rural_speed_dev_mean)
```

```
stargazer(reg_a_bivariate_plm, reg_a_yfe_plm, reg_a_full_plm, type='text')
```

```
==========================================================================
Dependent variable:
-------------------------- log(fatal_per_cap)
(1) (2) (3)
--------------------------------------------------------------------------
primary -0.155*** -0.091*** 0.031
(0.027) (0.028) (0.020)

college -2.387***
(0.146)

beer 0.176***
(0.025)
```

secondary 0.038**
(0.017)

unemploy 0.017***
(0.004)

log(vmt_per_cap) 1.154***
(0.041)

log(precip) -0.051***
(0.012)

snow32 -0.171***
(0.014)

log(rural_speed) 0.554***
(0.127)

log(urban_speed) 0.182**
(0.091)

Constant -1.705***
(0.011)

Observations 1,104 1,104 1,104 R2 0.029 0.010 0.744
Adjusted R2 0.028 -0.011 0.736 F Statistic 32.799*** (df
= 1; 1102) 10.848*** (df = 1; 1080) 311.366*** (df = 10; 1071)
==========
Note: *p<0.1;* ***p<0.05;*** p<0.01

---

primary 0.033 (0.021) secondary 0.041 **(0.017) college**
-2.246 *(0.147) beer 0.199* **(0.024) unemploy 0.017**
(0.004) ln_tvmt_pc 1.161 **(0.041) precip -0.006 (0.005)**
snow32 -0.174 *(0.014) rural_speed 0.011*** (0.002)
urban_speed 0.003* (0.001)

Observations 1,104
R2 0.741
Adjusted R2 0.734
========== Note: *p<0.1;* ***p<0.05;*** p<0.01

```
reg3a_3.1 <- plm(ln_fat_pc ~ primary + college + beer + unemploy
                 + ln_tvmt_pc + precip + snow32 + rural_speed + urban_speed,
                 data = df.p, effect = "time", model = "within")
stargazer(reg3a_3.1, type = "text", no.space = TRUE, omit.stat=c("f", "ser"))
```

========== Dependent variable:
——————— ln_fat_pc
———————————- primary -0.001
(0.015)
college -2.230***
(0.148)
beer 0.196***
(0.024)
unemploy 0.017***
(0.004)
ln_tvmt_pc 1.148***

(0.041)
precip -0.006
(0.005)
snow32 -0.180***
(0.014)
rural_speed 0.012***
(0.002)
urban_speed 0.002*
(0.001)
————————————- Observations 1,104
R2 0.740
Adjusted R2 0.732
========== Note: *p<0.1;* ***p<0.05;*** p<0.01

```
# with state fixed effects as well
reg3a_3.2 <- plm(ln_fat_pc ~ primary + secondary + college + beer + unemploy
                 + ln_tvmt_pc + precip + snow32 + rural_speed + urban_speed + factor(state),
                 data = df.p, effect = "time", model = "within")
stargazer(reg3a_3.2, type = "text", no.space = TRUE, omit.stat=c("f", "ser"))
```

========== Dependent variable:
———————— ln_fat_pc
————————————- primary -0.083***
(0.014)
secondary -0.017*
(0.010)
college -0.109
(0.223)
beer 0.572***
(0.040)
unemploy -0.025***
(0.003)
ln_tvmt_pc 0.422***
(0.054)
precip -0.021***
(0.006)
snow32 0.008
(0.013)
rural_speed 0.002**
(0.001)
urban_speed 0.002*
(0.001)
factor(state)2 0.020
(0.026)
factor(state)3 -0.369***
(0.043)
factor(state)4 -0.519***
(0.039)
factor(state)5 -0.609***
(0.050)
factor(state)6 -0.635***
(0.042)
factor(state)7 -0.549***
(0.032)

factor(state)8	-0.312***
	(0.032)
factor(state)9	-0.245***
	(0.027)
factor(state)10	-0.493***
	(0.035)
factor(state)11	-0.222***
	(0.035)
factor(state)12	-0.673***
	(0.039)
factor(state)13	-0.453***
	(0.028)
factor(state)14	-0.374***
	(0.036)
factor(state)15	-0.112***
	(0.026)
factor(state)16	-0.220***
	(0.033)
factor(state)17	-1.012***
	(0.044)
factor(state)18	-0.571***
	(0.040)
factor(state)19	-0.522***
	(0.035)
factor(state)20	-0.523***
	(0.034)
factor(state)21	-0.783***
	(0.038)
factor(state)22	-0.383***
	(0.030)
factor(state)23	0.090***
	(0.026)
factor(state)24	-0.302***
	(0.041)
factor(state)25	-0.174***
	(0.028)
factor(state)26	-0.784***
	(0.037)
factor(state)27	-0.647***
	(0.037)
factor(state)28	-1.093***
	(0.049)
factor(state)29	-0.710***
	(0.038)
factor(state)30	-0.140***
	(0.041)
factor(state)31	-0.633***
	(0.052)
factor(state)32	-0.616***
	(0.047)
factor(state)33	-0.669***
	(0.033)
factor(state)34	-0.217***
	(0.029)

factor(state)35 -0.375***
(0.035)
factor(state)36 -0.608***
(0.038)
factor(state)37 -0.999***
(0.040)
factor(state)38 -0.117***
(0.028)
factor(state)39 -0.432***
(0.035)
factor(state)40 -0.116***
(0.025)
factor(state)41 -0.515***
(0.039)
factor(state)42 -0.291***
(0.044)
factor(state)43 -0.635***
(0.035)
factor(state)44 -0.609***
(0.039)
factor(state)45 -0.593***
(0.036)
factor(state)46 -0.857***
(0.040)
factor(state)47 -0.018
(0.030)
factor(state)48 -0.203***
(0.039)
———————————————- Observations 1,104
R2 0.941
Adjusted R2 0.936
== Note: *p<0.1;* ***p<0.05;***
p<0.01

```
reg3a_3.3 <- plm(ln_fat_pc ~ primary + college + beer + unemploy
                 + ln_tvmt_pc + precip + snow32 + rural_speed + urban_speed + factor(state),
                 data = df.p, effect = "time", model = "within")
stargazer(reg3a_3.3, type = "text", no.space = TRUE, omit.stat=c("f", "ser"))
```

== Dependent variable:
———————— ln_fat_pc
———————————————- primary -0.067***
(0.011)
college -0.091
(0.223)
beer 0.567***
(0.040)
unemploy -0.024***
(0.003)
ln_tvmt_pc 0.421***
(0.054)
precip -0.021***
(0.006)
snow32 0.007

(0.013)
rural_speed 0.002**
(0.001)
urban_speed 0.002**
(0.001)
factor(state)2 0.020
(0.026)
factor(state)3 -0.369***
(0.043)
factor(state)4 -0.524***
(0.039)
factor(state)5 -0.612***
(0.050)
factor(state)6 -0.641***
(0.042)
factor(state)7 -0.548***
(0.032)
factor(state)8 -0.316***
(0.032)
factor(state)9 -0.247***
(0.027)
factor(state)10 -0.495***
(0.035)
factor(state)11 -0.225***
(0.035)
factor(state)12 -0.678***
(0.039)
factor(state)13 -0.455***
(0.028)
factor(state)14 -0.378***
(0.036)
factor(state)15 -0.110***
(0.026)
factor(state)16 -0.224***
(0.033)
factor(state)17 -1.012***
(0.044)
factor(state)18 -0.576***
(0.040)
factor(state)19 -0.519***
(0.035)
factor(state)20 -0.527***
(0.034)
factor(state)21 -0.787***
(0.038)
factor(state)22 -0.388***
(0.030)
factor(state)23 0.088***
(0.026)
factor(state)24 -0.303***
(0.041)
factor(state)25 -0.178***
(0.028)
factor(state)26 -0.781***

(0.037)
factor(state)27 -0.645***
(0.037)
factor(state)28 -1.082***
(0.049)
factor(state)29 -0.717***
(0.038)
factor(state)30 -0.142***
(0.041)
factor(state)31 -0.633***
(0.052)
factor(state)32 -0.623***
(0.047)
factor(state)33 -0.672***
(0.033)
factor(state)34 -0.213***
(0.029)
factor(state)35 -0.376***
(0.036)
factor(state)36 -0.611***
(0.038)
factor(state)37 -1.000***
(0.040)
factor(state)38 -0.118***
(0.028)
factor(state)39 -0.428***
(0.035)
factor(state)40 -0.121***
(0.025)
factor(state)41 -0.517***
(0.039)
factor(state)42 -0.298***
(0.044)
factor(state)43 -0.639***
(0.035)
factor(state)44 -0.607***
(0.039)
factor(state)45 -0.599***
(0.036)
factor(state)46 -0.858***
(0.040)
factor(state)47 -0.017
(0.030)
factor(state)48 -0.203***
(0.039)
———————————————- Observations 1,104
R2 0.940
Adjusted R2 0.936
== Note: *p<0.1;* ***p<0.05;*** p<0.01

We included covariates that we believe to be correlated (both statistically and theoretically) with both the outcome (log fatalities per capita) and the treatment variable (primary). This includes weather variables, education, beer, unemployment, speed limits, and tvmt. Since tvmt is a level variable, we thought it best to

divide it by population and take its log. We included "secondary" in one of the regressions to identify the marginal effect of having a primary seatbelt law in addition to the secondary one where that is already in place.

Including these covariates drastically changes our point estimate. When the covariates other than secondary are include, the effect essentially disappears though remains negative. This is due to the fact that variation in fatalities per capita that we previously attributed to "primary" in the OLS regression without covariates is actually attributable to these other characteristics. Furthermore, once secondary is included, the effect remains insignificant but turns positive. If this is actually a true zero, then this could be because the secondary policy already prevents fatalities. If this is a true positive (we are less convinced), then it could be that the primary policy causes people to take less responsibility of one's own actions and depend on law enforcement. Laws can cause rebellion from teens and also unintentiolly shift responsibility from the citizen to the government.

In addition, it is worth noting that when adding state fixed effects, the coefficient on primary becomes negative and significant, even with and without controling for "secondary." In fact, it is even more negative when controlling for "secondary." This, to us, seems to identify the treatment effect since it controls for any intrinsic state characteristics that were not included and could have cause omitted variable bias.

Part (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 (e.g., ", robust"). Do they change much? Compute the clustered standard errors that are robust to within-state correlation (e.g., ", cluster(state)"). 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.

Standard errors from specifications in part (a) is likely incorrect given we have potential correlation across observations from same state.

```
reg3b_1_robust <- coeftest(reg3a_1, vcov = vcovHC(reg3a_1, method = "white1", type = "HC0"))
reg3b_2_robust <- coeftest(reg3a_2_p, vcov = vcovHC(reg3a_2_p, method = "white1", type = "HC0"))
reg3b_3_robust <- coeftest(reg3a_3, vcov = vcovHC(reg3a_3, method = "white1", type = "HC0"))

stargazer(reg3a_1, reg3b_1_robust, reg3a_2_p, reg3b_2_robust, reg3a_3, reg3b_3_robust,
          title = "Regression with Robust SE",
          dep.var.caption = "Log(Fatality per Population)",
          dep.var.labels.include = FALSE, model.names = FALSE,
          column.labels = c("Biv", "Biv(Robust)", "Yr FE", "Yr FE(Robust)", "Ctr/Yr FE", "Ctr/Yr FE(Robu
          keep = "primary",
          font.size = "footnotesize", column.sep.width = "1pt", no.space = TRUE, omit.stat=c("f", "ser"
          type = "text", digits = 4)
```

```
##
## Regression with Robust SE
## ================================================================================================
##                                      Log(Fatality per Population)
##              ------------------------------------------------------------------------------
##                 Biv      Biv(Robust)    Yr FE     Yr FE(Robust) Ctr/Yr FE Ctr/Yr FE(Robust)
##                 (1)          (2)         (3)           (4)          (5)           (6)
## ------------------------------------------------------------------------------------------------
## primary      -0.1549*** -0.1549***  -0.0906***  -0.0906***     0.0331        0.0331*
##               (0.0270)   (0.0283)    (0.0275)    (0.0296)    (0.0205)      (0.0192)
```

```
## ------------------------------------------------------------------------------------------
## Observations   1,104                  1,104                   1,104
## R2             0.0289                 0.0099                 0.7413
## Adjusted R2    0.0280                -0.0111                 0.7336
## ==========================================================================================
## Note:                                                       *p<0.1; **p<0.05; ***p<0.01
```

We observe that for simple bivariate regression and spepcification with only year fixed effects, robust standards are slightly larger. In model with controls and time fixed effects, robust standard errors are slightly smaller.

```
reg3b_1_cluster <- coeftest(reg3a_1, vcov = vcovCL, cluster = ~state)
reg3b_2_cluster <- coeftest(reg3a_2_p, vcov = vcovHC(reg3a_2_p, type = "HC0", cluster = "group"))
reg3b_3_cluster <- coeftest(reg3a_3, vcov = vcovHC(reg3a_3, type = "HC0", cluster = "group"))

stargazer(reg3a_1, reg3b_1_cluster, reg3a_2_p, reg3b_2_cluster, reg3a_3, reg3b_3_cluster,
          title = "Regression with Clustered SE",
          dep.var.caption = "Log(Fatality per Population)",
          dep.var.labels.include = FALSE, model.names = FALSE,
          column.labels = c("Biv", "Biv(Cluster)", "Yr FE", "Yr FE(Cluster)", "Ctr/Yr FE", "Ctr/Yr FE(C
          keep = "primary",
          font.size = "footnotesize", column.sep.width = "1pt", no.space = TRUE, omit.stat=c("f", "ser"
          type = "text", digits = 4)
```

```
##
## Regression with Clustered SE
## ==========================================================================================
##                                    Log(Fatality per Population)
##               ---------------------------------------------------------------------------
##                  Biv     Biv(Cluster)   Yr FE    Yr FE(Cluster) Ctr/Yr FE Ctr/Yr FE(Cluster)
##                  (1)         (2)         (3)          (4)          (5)           (6)
## ------------------------------------------------------------------------------------------
## primary      -0.1549***   -0.1549    -0.0906***    -0.0906      0.0331         0.0331
##               (0.0270)    (0.1011)    (0.0275)     (0.1119)    (0.0205)       (0.0491)
## ------------------------------------------------------------------------------------------
## Observations   1,104                  1,104                   1,104
## R2             0.0289                 0.0099                 0.7413
## Adjusted R2    0.0280                -0.0111                 0.7336
## ==========================================================================================
## Note:                                                       *p<0.1; **p<0.05; ***p<0.01
```

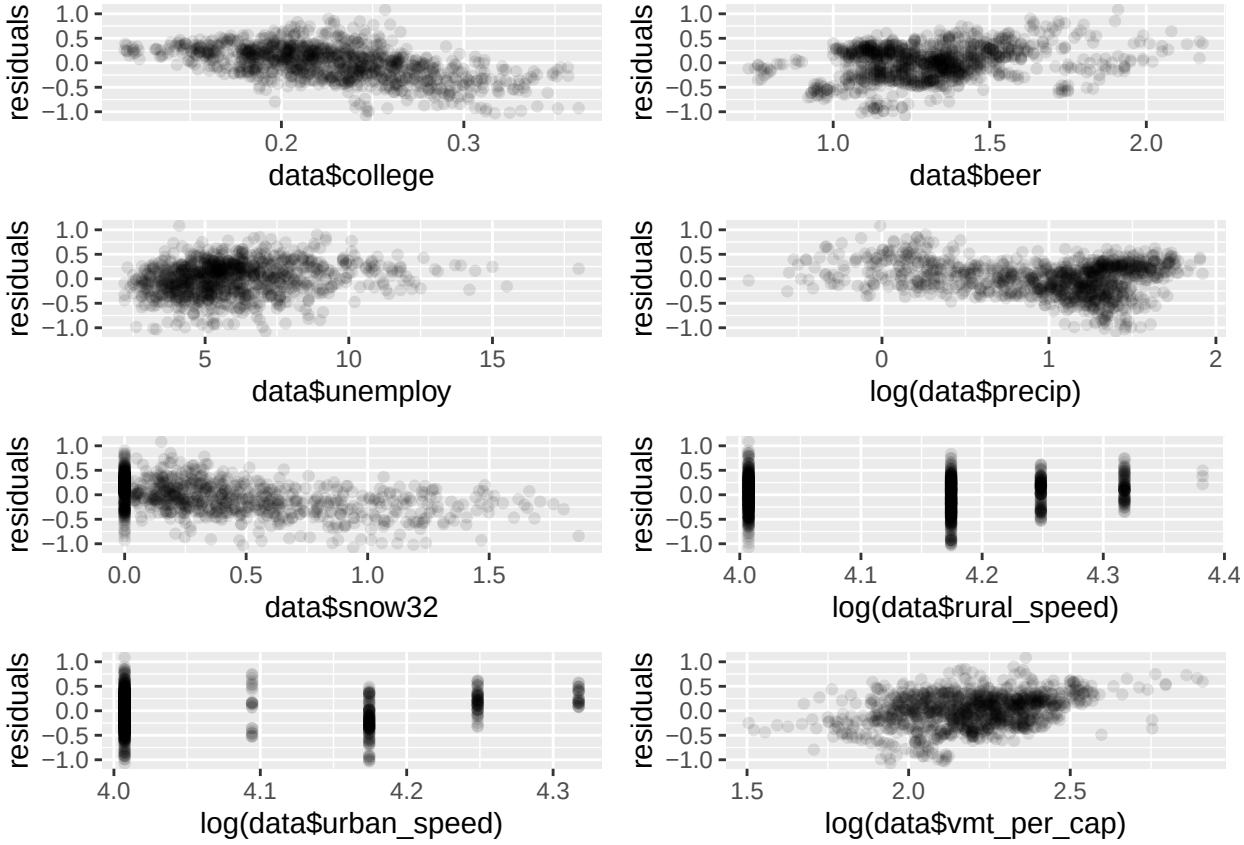
```
# Manually Calculate Robust SE for Bivariate OLS
Y_a <- df[,"ln_fat_pc"] %>%
          as.matrix() #1104 x 1

X_a <- df[,"primary"] %>%
          mutate(cons = 1) %>%
          select(cons, primary) %>%
          as.matrix() #1104 x 2

XX_inv <- solve(t(X_a) %*% X_a) #2 x 2

B_a = XX_inv %*% (t(X_a)%*%Y_a) #2 x 1 # This matches reg3a_1$coefficients
```

```
Resid_a <- Y_a - (X_a %*% B_a) #1104 x 1 # This matches reg3a_1$residuals

v2_a <- diag(diag(Resid_a %*% t(Resid_a))) #1104 x 1104

vcov_cluster <- (XX_inv) %*% (t(X_a) %*% v2_a %*% X_a) %*% XX_inv

vcov_cluster[2,2]^0.5
```

```
## [1] 0.02829574
```

We observe that clustered SE are significantly higher than the original as well as robust SE. This suggest that correlation within state over time is seriously biasing our conventional standard errors.

From the below plots of the residuals, we can see that the variance of the residuals is not constant along many of our covariates – especially speed limits, unemployment rates, and snow levels. Heteroskedasticity is an issue we need to correct for.




```
coeftest(reg_a_bivariate_lm, vcov = vcovHC(reg_a_bivariate_lm, method="white1", type="HC0"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) -1.704923   0.010841 -157.2664 < 2.2e-16 ***
## primary     -0.154902   0.028296   -5.4744 5.435e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(reg_a_bivariate_plm, vcov = vcovHC(reg_a_bivariate_plm, method="white1", type="HC0"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) -1.704923   0.010841 -157.2664 < 2.2e-16 ***
## primary     -0.154902   0.028296   -5.4744 5.435e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# by hand
Sigma = diag(resid(reg_a_bivariate_plm)^2)
X = model.matrix(~ primary, data = data)
XpX = t(X)%*%X
XpXinv = solve(XpX)
XpSigmaX = t(X)%*%Sigma%*%X
var_b = XpXinv %*% XpSigmaX %*% XpXinv
SE_b = sqrt(diag(var_b))
SE_b
```

```
## (Intercept)     primary
##  0.01084099  0.02829574
```

```
library(lmtest)
library(sandwich)
coeftest(reg_a_yfe_lm, vcov = vcovHC(reg_a_yfe_lm, method="white1", type="HC0"))[[2,2]]
```

```
## [1] 0.02958757
```

```
coeftest(reg_a_yfe_plm, vcov = vcovHC(reg_a_yfe_plm, method="white1", type="HC0"))
```

```
##
## t test of coefficients:
##
##          Estimate Std. Error t value Pr(>|t|)
## primary -0.090637   0.029588 -3.0633 0.002243 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# by hand
d_temp = data %>% mutate(resids = resid(reg_a_bivariate_plm))
meat = matrix(0, nrow=1, ncol=1)
# sum over all years
# for each year, subset just that year and calc X'(ee')X
for (y in unique(data$year)) {
    X = model.matrix(~ primary - 1, data = filter(d_temp, year==y))
    e = as.matrix(filter(d_temp, year==y) %>% select(resids))
```

```
    meat = meat + t(X)%*% (e%*%t(e)) %*% X
}

X = as.matrix(d_temp %>% select(primary))
XpX = t(X)%*%X
XpXinv = solve(t(X)%*%X)
var_b = XpXinv %*% meat %*% XpXinv / (dim(X)[1] - dim(X)[2])
SE_b = sqrt(diag(var_b))
rm(d_temp)
SE_b
```

```
##     primary
## 0.000628117
```

```
coeftest(reg_a_yfe_lm, vcov = vcovHC(reg_a_yfe_lm, method="white1", type="HC0"))[[2,2]]
```

```
## [1] 0.02958757
```

```
coeftest(reg_a_yfe_plm, vcov = vcovHC(reg_a_yfe_plm, method="white1", type="HC0"))
```

```
##
## t test of coefficients:
##
##          Estimate Std. Error t value Pr(>|t|)
## primary -0.090637   0.029588 -3.0633 0.002243 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# by hand
d_temp = data %>%
    mutate(resids = resid(reg_a_bivariate_plm)) %>%
    dummy_cols(., select_columns = 'year') %>%
    select(primary, resids, contains('year'))
k = ncol(d_temp) - 2
meat = matrix(0, nrow=k, ncol=k)
# sum over all years
# for each year, subset just that year and calc X'(ee')X
for (y in unique(data$year)) {
    X = as.matrix(d_temp %>% filter(year==y) %>% select(-year, -resids))
    e = as.matrix(d_temp %>% filter(year==y) %>% select(resids))
    meat = meat + t(X)%*% (e%*%t(e)) %*% X
}
# X'X on full data
X = as.matrix(d_temp %>% select(-year, -resids))
XpX = t(X)%*%X
XpXinv = solve(t(X)%*%X)
N = length(unique(data$state))
T_ = length(unique(data$year))
var_b = XpXinv %*% meat %*% XpXinv  * (N*T_) / (N*(T_ - 1) - k)
SE_b = sqrt(diag(var_b))
rm(d_temp)
SE_b
```

```
##    primary  year_1981  year_1982  year_1983  year_1984  year_1985  year_1986
## 0.02031169 0.22616781 0.10038824 0.06285054 0.07391261 0.05676555 0.11784721
##  year_1987  year_1988  year_1989  year_1990  year_1991  year_1992  year_1993
## 0.11949776 0.12971298 0.06807772 0.04561553 0.02848324 0.08811091 0.08541361
##  year_1994  year_1995  year_1996  year_1997  year_1998  year_1999  year_2000
## 0.09114166 0.06686852 0.07591656 0.06251720 0.07659113 0.08106425 0.09773802
##  year_2001  year_2002  year_2003
## 0.09396715 0.07873012 0.08960850
```

Part (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?

In order for between estimators to be unbiased, individual state time-in-varying effects, which now rests within error term, should not correlated with included 'primary' or other co-variates. I do not think this "uncorrelated effects" condition would hold, given each states' time-invariant characteristics regarding average driving styles or fatality rate would likely impact whether state implements primary and secondary laws as well as driving conditions captured by covariates. With the between estimator, I am worried about SE, given we are down to 48 observations, one for each state, by averaging each state's observation over time.

```
df_within <- df %>% select(-c("state", "year"))
df_within <- aggregate(df_within, list(df$state), mean)

sapply(df_within, typeof)
```

```
##       Group.1       college          beer       primary     secondary
##      "double"      "double"      "double"      "double"      "double"
##    population      unemploy    fatalities      totalvmt        precip
##      "double"      "double"      "double"      "double"      "double"
##        snow32   rural_speed   urban_speed fatal_per_cap   vmt_per_cap
##      "double"      "double"      "double"      "double"      "double"
##        fat_pc     ln_fat_pc    ln_tvmt_pc
##      "double"      "double"      "double"
```

```
is.numeric(df_within$college)
```

```
## [1] TRUE
```

```
reg_3c_btw_nocov <- plm(ln_fat_pc ~ primary,
                        data = df_within,
                        model = "between")


reg_3c_btw_nocov
```

```
##
## Model Formula: ln_fat_pc ~ primary
## <environment: 0x55fb986a4200>
```

```
##
## Coefficients:
## (Intercept)     primary
##   -1.715025   -0.093623
```

```
# Is between estimator same as simply including time fixed effects
#reg_3c_time <- plm(ln_fat_pc ~ primary,
#                               data = df,
#                               index = c("state", "year"),
#                               model = "between",
#                               effects = "time")

reg_3c_btw_cov <- plm(ln_fat_pc ~ primary + secondary + beer + as.numeric(college) + unemploy
                              + ln_tvmt_pc + precip + snow32 + rural_speed + urban_speed,
                              data = df_within,
                              model = "between")

reg_3c_btw_cov
```

```
##
## Model Formula: ln_fat_pc ~ primary + secondary + beer + as.numeric(college) +
##     unemploy + ln_tvmt_pc + precip + snow32 + rural_speed + urban_speed
## <environment: 0x55fb986a4200>
##
## Coefficients:
##         (Intercept)             primary           secondary                beer
##           -6.4646311            0.0839521            0.0518547            0.1395967
## as.numeric(college)             unemploy           ln_tvmt_pc              precip
##           -0.0055123            0.0439015            1.1552620            0.0004868
##               snow32         rural_speed         urban_speed
##           -0.1700806            0.0275599            0.0030704
```

Part (d)

Compute the Random Effects 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?

```
reg_3d_btw_nocov <- plm(ln_fat_pc ~ primary + secondary + college + beer + unemploy
                              + ln_tvmt_pc + precip + snow32 + rural_speed + urban_speed,
                              data = df,
                              model = "random")

reg_3d_btw_nocov
```

```
##
## Model Formula: ln_fat_pc ~ primary + secondary + college + beer + unemploy +
##     ln_tvmt_pc + precip + snow32 + rural_speed + urban_speed
## <environment: 0x55fb986a4200>
##
## Coefficients:
## (Intercept)     primary   secondary     college        beer    unemploy
```

```
##  -1.9352982  -0.1533228  -0.0705016  -1.7487048   0.7460317  -0.0224230
##  ln_tvmt_pc      precip      snow32 rural_speed urban_speed
##   0.0570305  -0.0228768  -0.0261057  -0.0065531   0.0030721
```

Part (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?)

Part (f)

Compute the FE estimator using only primary and year fixed effects as the covariates. Compute the normal standard errors and the clustered standard errors. If they are different, why?

Part (g)

Add the same range of covariates to the FE estimator that you did to the OLS estimator. Are the FE estimates more or less stable than the OLS estimates? Why?

Part (h)

Estimate a first-differences estimator, a 5-year differences estimator, and a long differences estimator, including year fixed effects (when feasible) and the appropriate covariates in each case. Briefly describe the pattern that emerges from the three differencing estimates. Where does the FE estimate fall in this pattern? Are you surprised?

Part (i)

Make the case that the first-differences estimate is superior to the 5-year or long differences estimates.

Part (j)

Make the case that the 5-year or long differences estimates are superior to the first-differences estimate.