# ARE 213 Problem Set 2a

*Nick Depsky, Will Gorman, Peter Worley*

*October 26, 2018*

## 1a - Show betas numerically identical

First, let's start with the fixed effects estimator which we will represent here as the within estimator. This is estimated by demeaning each equation.

To calculate the means let: $\overline{y}_i = \frac{y_{i1} + y_{i2}}{2}$, $\overline{x}_i = \frac{x_{i1} + x_{i2}}{2}$, and $\overline{\epsilon}_i = \frac{\epsilon_i 1 + \epsilon_i 2}{2}$

Then:

$$\beta_{FE} = \left[\sum_{i=1}^{N}\sum_{t=1}^{2}[x_{it} - \overline{x}_i]'[x_{it} - \overline{x}_i]\right] \sum_{i=1}^{N}\sum_{t=1}^{2}[x_{it} - \overline{x}_i]'[y_{it} - \overline{y}_i]$$

Substituting in the means for the X'X term:

$$\sum_{t=1}^{2}[x_{it} - \overline{x}_i]'[x_{it} - \overline{x}_i] = \sum_{t=1}^{2}[x_{it} - \frac{x_{i1} + x_{i2}}{2}]'[x_{it} - \frac{x_{i1} + x_{i2}}{2}] =$$

$$[x_{i1} - \frac{x_{i1} + x_{i2}}{2}]'[x_{i1} - \frac{x_{i1} + x_{i2}}{2}] + [x_{i2} - \frac{x_{i1} + x_{i2}}{2}]'[x_{i2} - \frac{x_{i1} + x_{i2}}{2}] =$$

$$[\frac{2x_{i1} - x_{i1} - x_{i2}}{2}]'[\frac{2x_{i1} - x_{i1} - x_{i2}}{2}] + [\frac{2x_{i2} - x_{i1} - x_{i2}}{2}]'[\frac{2x_{i2} - x_{i1} - x_{i2}}{2}] =$$

$$[\frac{x_{i1} - x_{i2}}{2}]'[\frac{x_{i1} - x_{i2}}{2}] + [\frac{x_{i2} - x_{i1}}{2}]'[\frac{x_{i2} - x_{i1}}{2}] =$$

$$(-1 * [\frac{x_{i2} - x_{i1}}{2}])'(-1 * [\frac{x_{i2} - x_{i1}}{2}]) + [\frac{x_{i2} - x_{i1}}{2}]'[\frac{x_{i2} - x_{i1}}{2}] =$$

$$(2 * [\frac{x_{i2} - x_{i1}}{2}]'[\frac{x_{i2} - x_{i1}}{2}] =$$

The 2s cancel.

$$[x_{i2} - x_{i1}]'[x_{i2} - x_{i1}] = [\Delta x_i' \Delta x_i]$$

Similarly for the X'Y term:

$$\sum_{t=1}^{2}[x_{it} - \overline{x}_i]'[y_{it} - \overline{y}_i] = \sum_{t=1}^{2}[x_{it} - \frac{x_{i1} + x_{i2}}{2}]'[y_{it} - \frac{y_{i1} + y_{i2}}{2}] =$$

$$[\frac{x_{i1} - x_{i2}}{2}]'[\frac{y_{i1} - y_{i2}}{2}] + [\frac{x_{i2} - x_{i1}}{2}]'[\frac{y_{i2} - y_{i1}}{2}] =$$

$$(2 * [\frac{x_{i2} - x_{i1}}{2}]'[\frac{y_{i2} - y_{i1}}{2}] =$$

$$[x_{i2} - x_{i1}]'[y_{i2} - y_{i1}] = [\Delta x_i' \Delta y_i]$$

And the fixed effect estimator is:

$$\beta_{FE} = \left[\sum_{t=1}^{N} [\Delta x_i' \Delta x_i]\right]^{-1} \sum_{t=1}^{N} [\Delta x_i' \Delta y_i]$$

Now, let's look at the difference estimator. These are calculated by using the difference across time periods to eliminate the unobservables:

$$y_{i2} - y_{i1} = [x_{i2} - x_{i1}]\beta + [\epsilon_{i2} - \epsilon_{i1}] \rightarrow \Delta y_i = \Delta x_i \beta + \Delta \epsilon_i$$

Assuming that we have strict exogeneity and the errors are orthogonal to the Xs, the difference estimate becomes the same as the fixed estimator found above:

$$\beta_{DE} = \left[\sum_{t=1}^{N} [\Delta x_i' \Delta x_i]\right]^{-1} \sum_{t=1}^{N} [\Delta x_i' \Delta y_i]$$

# 1b - Show standard errors numerically identical

Variance covariance matrix of fixed effects equals:

$$\sigma_{FE}^2 \left[\sum_{t=1}^{N} \sum_{t=1}^{2} [x_{it} - \overline{x}_i]'[x_{it} - \overline{x}_i]\right]^{-1}$$

where

$$e_{it}^{FE} = [y_{it} - \overline{y}_i] - [x_{it} - \overline{x}_i]\beta_{FE}$$

t = 1,2

so that

$$\sigma_{FE}^2 = \frac{\sum_{t=1}^{N} e_{i1}^{FE} + e_{i2}^{FE}}{N - k}$$

Furthermore, we can use the fact the the coefficients are the same for the two estimators:

$$e_{i1}^{FE} = [\frac{y_{i2} - y_{i1}}{2}] - [\frac{x_{i2} - x_{i1}}{2}]\beta_{FE} =$$

$$\frac{-\Delta y_i + \Delta x_i \beta_{DE}}{2} = \frac{e_i^{DE}}{2}$$

The same resulst happens for t = 2 and we can use these residuals to compare the variances on the estimators of the coefficients.

When returning to the squared sigmas we can now relate them between fixed effects and the differences estimator by plugging in what we just solved for to get:

$$\sigma_{FE}^2 = \frac{1}{2}\sigma_{DE}^2$$

Then, we can show that the variances of the fixed effects estimator are the same to difference estimator:

$$V(\beta_{FE}) = \sigma^2_{DE}\left[\sum_{t=1}^{N}[\Delta x_i'\Delta x_i]\right]^{-1} =$$

$$\frac{\sigma^2_{DE}}{2}\left[\sum_{t=1}^{N}[\frac{\Delta x_i'\Delta x_i}{2}]\right]^{-1} =$$

$$\sigma^2_{FE}\left[\sum_{t=1}^{N}\sum_{t=1}^{2}[x_{it}-\overline{x}_i]'[x_{it}-\overline{x}_i]\right]^{-1} =$$

One notices that this final equation is the Variance of the coefficient on the fixed effects, proving that the variances are the same between the estimators.

## 2a - Show fixed effect estimator with transformations

First, we will show the result of the fixed effects estimator:

We define $\overline{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$, $\overline{y}_t = N^{-1}\sum_{i=1}^{N} y_{it}$, and $\overline{\overline{y}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$.

Then we use equivalent definitions for $\overline{\mathbf{x}}_\mathbf{i}$, $\overline{\mathbf{x}}_\mathbf{t}$, $\overline{\overline{\mathbf{x}}}$ as well as $\overline{\epsilon}_i$, $\overline{\epsilon}_t$, $\overline{\overline{\epsilon}}$.

Using the above definitions, we know the fixed effects two-way within model yields:

$$y_{it} - \overline{y}_i - \overline{y}_t + \overline{\overline{y}} = (\mathbf{x_{it}} - \overline{\mathbf{x}}_\mathbf{i} - \overline{\mathbf{x}}_\mathbf{t} + \overline{\overline{\mathbf{x}}})\beta + (\epsilon_{it} - \overline{\epsilon}_i - \overline{\epsilon}_t + \overline{\overline{\epsilon}})$$

Now, we want to show that this same two-way model can be obtained through two within one-way transformations. If we assume the first transformation uses the time-averaged model, then:

$$y_{it} - \overline{y}_i = (\mathbf{x_{it}} - \overline{\mathbf{x}}_\mathbf{i})\beta + (\lambda_t - \overline{\lambda}_t) + (\epsilon_{it} - \overline{\epsilon}_i)$$

and the fixed effect $\mu_i$ is eliminated. We'll define $y_{it} - \overline{y}_i = z_i$ then run a transformation using an individual-averaged model defined by:

$$z_t = \overline{y}_t - \overline{\overline{y}}$$

such that:

$$z_t = \overline{y}_t - \overline{\overline{y}} = (\overline{\mathbf{x_t}} - \overline{\overline{\mathbf{x}}})\beta + (\lambda_t - \overline{\lambda}_t) + (\overline{\epsilon_t} - \overline{\overline{\epsilon}})$$

and, again, the fixed effect $\mu_i$ is eliminated. Subtracting $z_t$ from $z_i$ yields:

$$y_{it} - \overline{y}_i - \overline{y}_t + \overline{\overline{y}} = (\mathbf{x_{it}} - \overline{\mathbf{x}}_\mathbf{i} - \overline{\mathbf{x}}_\mathbf{t} + \overline{\overline{\mathbf{x}}})\beta + (\epsilon_{it} - \overline{\epsilon}_i - \overline{\epsilon}_t + \overline{\overline{\epsilon}})$$

which is the same outcome as the two-way within model approach listed above.

## 2b - Show order of operation is important and explain why

If we were to instead run the first order transformation on the individual fxed effects averaged model, and use time-averaged values in place of individual fixed effects averages in the second transformation, rather than obtaining $\lambda_t - \overline{\lambda}_t$ in both one-way transformations, we would obtain $\mu_i - \overline{\mu}_i$ in each transformation, but they would still fall out in the last step and yield the same outcome. Intuitively, it makes sense that we should see the same outcome, whether we control for time or individual effects first or second.

## 2c - Change with imbalance?

Generally speaking, fixed effects models can handle missing observations suggested by an unbalanced panel. However, if the panel becomes unbalanced for a reason correlated to the variables of interest, then there may be some sample selection bias that arises.

Also, it may be possible that the $\overline{\mu}_i$ terms or the $\overline{\lambda}_t$ terms will not be equivalent form the first transformation to the second in an unbalanced panel, which would yield a remainder term that makes the two within estimator approach unequal to the two-way error component model.

## 3 - Regression analysis

Import Data

```
setwd("~/Dropbox/Berkeley_tings/Fall 2018/ARE213/Problem Sets/SharedFiles/are213/PS2a")
#setwd("C:\\Users\\will-\\Desktop\\are213\\PS1b")
#setwd("C:\\Users\\Will\\Desktop\\are213\\PS2a")

dat <- read.dta("traffic_safety2.dta")
```

## 3a - Pooled bivariate OLS

The below regression shows that the existence of a primary belt law has a result of decreasing per capita fatalities by 14%.

```
# as.factor command on fixed effects
dat$state <- as.factor(dat$state)
dat$year <- as.factor(dat$year)

dat$fatal_pc <- log(dat$fatalities/dat$population)

# fix the attribute labels
attributes(dat)$var.labels = c("state", "year", attributes(dat)$var.labels[-(1:2)],
                               "log of traffic fatalities per capita")

# pooled bivariate OLS
reg1 <- lm(fatal_pc ~ primary, data = dat)
summary(reg1)
```

```
##
## Call:
```

```
## lm(formula = fatal_pc ~ primary, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07731 -0.21175  0.03609  0.23078  1.08120
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -1.70315    0.01089 -156.465  < 2e-16 ***
## primary     -0.14388    0.02584   -5.568 3.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3314 on 1125 degrees of freedom
## Multiple R-squared:  0.02682,    Adjusted R-squared:  0.02596
## F-statistic: 31.01 on 1 and 1125 DF,  p-value: 3.214e-08
```

When including time trends, the effect is reduced to a 7.4% reduction in per capita fatalities.

```
dat$time <- as.numeric(as.character(dat$year))-1981
dat$time_sq <- (dat$time)^2

# pooled bivariate OLS
reg2 <- lm(fatal_pc ~ primary + dat$time_sq, data = dat)
summary(reg2)
```

```
##
## Call:
## lm(formula = fatal_pc ~ primary + dat$time_sq, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06421 -0.21520  0.02451  0.22688  1.00424
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -1.626e+00  1.430e-02 -113.749  < 2e-16 ***
## primary     -8.208e-02  2.630e-02   -3.121  0.00185 **
## dat$time_sq -5.329e-04  6.649e-05   -8.015 2.74e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3225 on 1124 degrees of freedom
## Multiple R-squared:  0.07943,    Adjusted R-squared:  0.0778
## F-statistic: 48.49 on 2 and 1124 DF,  p-value: < 2.2e-16
```

When including relevant covariates, the effect is also reduced to a 9% (rather than 14%) reduction in per capita fatalities.

```
# pooled bivariate OLS
reg3 <- lm(fatal_pc ~ primary + college + beer + secondary + population + unemploy + totalvmt + precip +
             snow32 + rural_speed + urban_speed, data = dat)
```

```r
summary(reg3)
```

```
##
## Call:
## lm(formula = fatal_pc ~ primary + college + beer + secondary +
##     population + unemploy + totalvmt + precip + snow32 + rural_speed +
##     urban_speed, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93468 -0.13302  0.00395  0.14086  0.84815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.422e+00  1.302e-01 -10.915  < 2e-16 ***
## primary     -9.026e-02  2.428e-02  -3.717 0.000212 ***
## college     -3.247e+00  1.768e-01 -18.360  < 2e-16 ***
## beer         2.754e-01  3.137e-02   8.778  < 2e-16 ***
## secondary   -8.310e-02  1.965e-02  -4.229 2.54e-05 ***
## population  -4.235e-05  6.063e-06  -6.984 4.91e-12 ***
## unemploy     9.386e-03  4.039e-03   2.324 0.020320 *
## totalvmt     3.750e-06  7.293e-07   5.143 3.20e-07 ***
## precip      -3.540e-02  5.881e-03  -6.019 2.38e-09 ***
## snow32      -2.750e-01  1.848e-02 -14.885  < 2e-16 ***
## rural_speed -4.046e-04  1.783e-03  -0.227 0.820570
## urban_speed  5.836e-03  1.663e-03   3.509 0.000467 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2176 on 1115 degrees of freedom
## Multiple R-squared:  0.5841, Adjusted R-squared:   0.58
## F-statistic: 142.3 on 11 and 1115 DF,  p-value: < 2.2e-16
```

Adding covariates leads to the reduction of coefficient of interest (the effect of the primary belt laws). This result makes sense as predetermined covariates which are correlated with seat belt laws in some way might also explain the variation in per capita fatalities. For this reason, including the covariates gives us a better esimate as we including relevant variables.

## 3b - Standard errors for bivariate OLS

No, the above standard errors are likely not correct due to serial correlation across time within the cross-sectional units. Importantly, this is not solved by using the Huber-White Heteroskedastic robust standard errors because that design is not meant to deal with serial correlation but rather heterogenous error terms that are correlated with the covariates.

```r
#OLS coefficients and regular standard errors
round(coeftest(reg3),4)
```

```
##
## t test of coefficients:
##
```

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -1.4216     0.1302 -10.9153   <2e-16 ***
## primary      -0.0903     0.0243  -3.7167   0.0002 ***
## college      -3.2465     0.1768 -18.3599   <2e-16 ***
## beer          0.2754     0.0314   8.7784   <2e-16 ***
## secondary    -0.0831     0.0197  -4.2288   <2e-16 ***
## population    0.0000     0.0000  -6.9843   <2e-16 ***
## unemploy      0.0094     0.0040   2.3237   0.0203 *
## totalvmt      0.0000     0.0000   5.1425   <2e-16 ***
## precip       -0.0354     0.0059  -6.0192   <2e-16 ***
## snow32       -0.2750     0.0185 -14.8852   <2e-16 ***
## rural_speed  -0.0004     0.0018  -0.2269   0.8206
## urban_speed   0.0058     0.0017   3.5092   0.0005 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#OLS coefficients and white standard errors
round(coeftest(reg3, vcov = vcovHC(reg3, type = "HC1")),4)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -1.4216     0.1270 -11.1913   <2e-16 ***
## primary      -0.0903     0.0234  -3.8577   0.0001 ***
## college      -3.2465     0.1705 -19.0374   <2e-16 ***
## beer          0.2754     0.0299   9.2214   <2e-16 ***
## secondary    -0.0831     0.0209  -3.9672   0.0001 ***
## population    0.0000     0.0000  -7.2072   <2e-16 ***
## unemploy      0.0094     0.0041   2.2695   0.0234 *
## totalvmt      0.0000     0.0000   5.5640   <2e-16 ***
## precip       -0.0354     0.0059  -5.9829   <2e-16 ***
## snow32       -0.2750     0.0196 -14.0107   <2e-16 ***
## rural_speed  -0.0004     0.0019  -0.2115   0.8325
## urban_speed   0.0058     0.0015   3.8221   0.0001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that the standard errors in the standard case (0.0243) are not much different from the standard errors from Hubert-White (0.0234).

```
formula <- 'fatal_pc ~ primary + college + beer + primary + secondary + population + unemploy +
totalvmt + precip + snow32 + rural_speed + urban_speed'


felm_formula_clust <- paste(formula,'| 0 | 0 | state', sep = '') %>%
  as.formula()
felm.clust <- felm( felm_formula_clust,
                    dat)
# the estimates don't change, but the standard errors do
#OLS coefficients and regular standard errors
round(coeftest(felm.clust),4)
```

```
##
```

```
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.4216     0.3101 -4.5839   <2e-16 ***
## primary      -0.0903     0.0569 -1.5874   0.1127
## college      -3.2465     0.4933 -6.5815   <2e-16 ***
## beer          0.2754     0.0914  3.0118   0.0027 **
## secondary    -0.0831     0.0400 -2.0753   0.0382 *
## population    0.0000     0.0000 -2.6848   0.0074 **
## unemploy      0.0094     0.0096  0.9749   0.3298
## totalvmt      0.0000     0.0000  2.1183   0.0344 *
## precip       -0.0354     0.0206 -1.7182   0.0860 .
## snow32       -0.2750     0.0553 -4.9697   <2e-16 ***
## rural_speed  -0.0004     0.0033 -0.1230   0.9021
## urban_speed   0.0058     0.0028  2.0820   0.0376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, while the coefficient estimates remain the same, the standard errors are doubled to 0.0557 and the test for significance did not pass. We were not surprised of the change since clustered standard errors do have a significant effect, though we were more surprised by the fact that it was a large enough change that the significance changed.

```
# Function to convert tibble, data.frame, or tbl_df to matrix
to_matrix <- function(the_df, vars) {
  # Create a matrix from variables in var
  new_mat <- the_df %>%
    # Select the columns given in 'vars'
    select_(.dots = vars) %>%
    # Convert to matrix
    as.matrix()
  # Return 'new_mat'
  return(new_mat)
}

#OLS
b_ols <- function(y, X) {
  # Calculate beta hat
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
  # Return beta_hat
  return(beta_hat)
}

#create function to calculate vcov matrix
vcov_cluster <- function(data, y_var, X_vars,
  cluster_var, intercept = T) {
    # Turn data into matrices
    y <- to_matrix(data, y_var)
    X <- to_matrix(data, X_vars)
    # Add intercept
    if (intercept == T) X <- cbind(1, X)
    # Calculate n and k for degrees of freedom
    n <- nrow(X)
```

```
    k <- ncol(X)
    # Estimate coefficients
    b <- b_ols(y, X)
    # Update names
    if (intercept == T) rownames(b)[1] <- "Intercept"
    # Calculate OLS residuals
    e <- y - X %*% b
    # Inverse of X'X
    XX_inv <- solve(t(X) %*% X)
    # Find the levels of the variable on which we are clustering
    cl_levels <- data[, cluster_var] %>% unique() %>% unlist()
    # Calculate the meat, iterating over the clusters
    meat_hat <- lapply(X = cl_levels, FUN = function(g) {
      # Find the row indices for the current cluster
      indices <- which(unlist(data[, cluster_var]) == g)
      # Grab the current cluster's rows from X and e
      X_g <- X[indices,]
      e_g <- e[indices] %>% matrix(ncol = 1)
      # Calculate this cluster's part of the meat estimate
      return(t(X_g) %*% e_g %*% t(e_g) %*% X_g)
      }) %>% Reduce(f = "+", x = .) / n
    # Find the number of clusters
    G <- length(cl_levels)
    # Degrees-of-freedom correction
    df_c <- G/(G-1) * (n-1)/(n-k)
    # Return the results
    return(df_c * n * XX_inv %*% meat_hat %*% XX_inv)
  }

# get the vcov matrix
y <- 'fatal_pc'
x <- c('primary','college', 'beer', 'secondary' , 'population' , 'unemploy' , 'totalvmt' , 'precip' ,
       'snow32' , 'rural_speed' ,'urban_speed')
clus <- 'state'
vcov.lm.clust = vcov_cluster(dat,y,x,clus)

# show results
round(coeftest(reg3, vcov = vcov.lm.clust),4)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## primary      -0.0903     0.0569 -1.5874   0.1127
## college      -3.2465     0.4933 -6.5815   <2e-16 ***
## beer          0.2754     0.0914  3.0118   0.0027 **
## secondary    -0.0831     0.0400 -2.0753   0.0382 *
## population    0.0000     0.0000 -2.6848   0.0074 **
## unemploy      0.0094     0.0096  0.9749   0.3298
## totalvmt      0.0000     0.0000  2.1183   0.0344 *
## precip       -0.0354     0.0206 -1.7182   0.0860 .
## snow32       -0.2750     0.0553 -4.9697   <2e-16 ***
## rural_speed  -0.0004     0.0033 -0.1230   0.9021
## urban_speed   0.0058     0.0028  2.0820   0.0376 *
```

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

See the manual clustering strategy also resulted in a std error of 0.0569.

## 3c - between estimator

```
between_nocov <- plm(fatal_pc ~ primary, data = dat, model = "between")
summary(between_nocov)
```

```
## Oneway (individual) effect Between Model
##
## Call:
## plm(formula = fatal_pc ~ primary, data = dat, model = "between")
##
## Balanced Panel: n = 49, T = 23, N = 1127
## Observations used in estimation: 49
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.7175456 -0.1826791 -0.0082988  0.2073385  0.5772611
##
## Coefficients:
##              Estimate Std. Error  t-value Pr(>|t|)
## (Intercept) -1.716042   0.051659 -33.2185   <2e-16 ***
## primary     -0.071216   0.154766  -0.4602   0.6475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    4.4286
## Residual Sum of Squares: 4.4087
## R-Squared:      0.004485
## Adj. R-Squared: -0.016696
## F-statistic: 0.211743 on 1 and 47 DF, p-value: 0.64753
```

```
between_cov <- plm(fatal_pc ~ primary + college + beer + secondary + population + unemploy +
                     totalvmt + precip + snow32 + rural_speed + urban_speed,
                   data = dat, model = "between")
summary(between_cov)
```

```
## Oneway (individual) effect Between Model
##
## Call:
## plm(formula = fatal_pc ~ primary + college + beer + secondary +
##     population + unemploy + totalvmt + precip + snow32 + rural_speed +
##     urban_speed, data = dat, model = "between")
##
## Balanced Panel: n = 49, T = 23, N = 1127
## Observations used in estimation: 49
##
## Residuals:
```

```
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.3801926 -0.0789188  0.0018059  0.0911217  0.3713152
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept) -5.8632e+00  1.2658e+00 -4.6319 4.371e-05 ***
## primary      1.7063e-01  1.8063e-01  0.9446   0.35098
## college     -1.9846e+00  8.0437e-01 -2.4672   0.01837 *
## beer         1.5904e-01  1.2357e-01  1.2870   0.20609
## secondary    3.0933e-02  1.7189e-01  0.1800   0.85817
## population  -2.0382e-05  2.9347e-05 -0.6945   0.49168
## unemploy     2.3279e-02  2.6365e-02  0.8829   0.38297
## totalvmt     6.6292e-07  3.5306e-06  0.1878   0.85209
## precip       2.8798e-02  3.2177e-02  0.8950   0.37659
## snow32      -2.3712e-01  9.1485e-02 -2.5919   0.01358 *
## rural_speed  6.5087e-02  2.0680e-02  3.1473   0.00325 **
## urban_speed  3.0317e-03  1.4462e-02  0.2096   0.83510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    4.4286
## Residual Sum of Squares: 0.96387
## R-Squared:      0.78235
## Adj. R-Squared: 0.71765
## F-statistic: 12.091 on 11 and 37 DF, p-value: 3.5301e-09
```

Between estimator without covariates: We see a 7.12 percent reduction in fatalities with a standard error that results in an insignificant result.

Including covariates we see a 17.06 percent increase in fatalities but again the standard error suggests that these results are not statistically significant.

If the state specific term is correlated with the covariates we expect the estimates to be biased. Here, we do expect a correlation. For example: unemployment, speed limits, snow, college are all expected to be corelated to state.

Between estimator only uses between individual variation rather than the time variation. In other words, it discards all the information due to intertemporal variability. Since we can see some degree of intertemporal variability in our data, we expect biased results. If there wasn't any time period based variation, we could expect this to give less biased estimates.

For the same reasons, we are also concerned about the standard errors here. Furthermore, we are not clustering again here which would lead to inaccurate standard errors.

## 3d - random effects

```
rand_eff <- plm(fatal_pc ~ primary + college + beer + secondary + population +
                  unemploy + totalvmt + precip + snow32 + rural_speed + urban_speed,
                data = dat, index =c("state","year"),model = "random")
summary(rand_eff)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
```

```
## plm(formula = fatal_pc ~ primary + college + beer + secondary +
##     population + unemploy + totalvmt + precip + snow32 + rural_speed +
##     urban_speed, data = dat, model = "random", index = c("state",
##     "year"))
##
## Balanced Panel: n = 49, T = 23, N = 1127
##
## Effects:
##                    var  std.dev share
## idiosyncratic 0.008464 0.092002 0.248
## individual    0.025683 0.160258 0.752
## theta: 0.8811
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.3566848 -0.0629151  0.0062161  0.0632908  0.3280764
##
## Coefficients:
##                Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept) -1.8375e+00  9.6238e-02 -19.0932 < 2.2e-16 ***
## primary     -1.4270e-01  1.5109e-02  -9.4449 < 2.2e-16 ***
## college     -1.5044e+00  1.7297e-01  -8.6975 < 2.2e-16 ***
## beer         7.5255e-01  3.8094e-02  19.7553 < 2.2e-16 ***
## secondary   -6.6898e-02  1.0234e-02  -6.5368 9.543e-11 ***
## population  -1.8536e-05  7.7910e-06  -2.3791 0.0175217 *
## unemploy    -2.2706e-02  2.3677e-03  -9.5902 < 2.2e-16 ***
## totalvmt     4.7723e-07  6.7126e-07   0.7109 0.4772650
## precip      -2.4487e-02  6.1188e-03  -4.0019 6.697e-05 ***
## snow32      -1.6055e-02  1.4386e-02  -1.1161 0.2646334
## rural_speed -5.8437e-03  9.2654e-04  -6.3070 4.094e-10 ***
## urban_speed  3.0360e-03  8.5194e-04   3.5636 0.0003812 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    26.553
## Residual Sum of Squares: 10.449
## R-Squared:      0.60648
## Adj. R-Squared: 0.6026
## F-statistic: 156.217 on 11 and 1115 DF, p-value: < 2.22e-16
```

The RE estimator shows that states that have primary seatbelt laws see a decrease of 14 percent in fatalities per capita as compared to states that do not have this primary law. This result is significant. The RE estimate is different from the OLS estimate. The OLS SE is greater than the RE SE since we expect the RE estimator to give us more efficient estimates, i.e., more time invariant explanatory variables are accounted for in the RE regression, improving the predictive capacity of the model and reducing its overall variance. The RE estimator will give us unbiased estimates of the effect of primary seat bet laws on fatalities per capita if the state specific (unobserved) effects are uncorrelated with the explanatory variables.

## 3e - standard errors from RE

```
C <- length(unique(dat$state))
N <- length(dat$state)
```

```
K <- 11 +                                  # all the other covariates
     1 +                                   # the intercept
     length(unique(dat$state)) - 1 +       # the state dummies - ommitted state = 1
     length(unique(dat$year)) - 1          # the year dummies - ommitted year = 1981

adjustment <- (C/(C - 1)) * (N - 1)/(N - K)
remove(C, N, K)

# when we cluster within group, we need to use the arellano method
#   when we cluster within time, we need to use the white method (see help file)
vcov.plm.clust = vcovHC(rand_eff,
                        method = "arellano",
                        cluster = "group") * adjustment

# save results

round(coeftest(rand_eff, vcov = vcov.plm.clust),4)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -1.8375     0.1659 -11.0766   <2e-16 ***
## primary      -0.1427     0.0313  -4.5619   <2e-16 ***
## college      -1.5044     0.3055  -4.9246   <2e-16 ***
## beer          0.7526     0.0703  10.7109   <2e-16 ***
## secondary    -0.0669     0.0185  -3.6126   0.0003 ***
## population    0.0000     0.0000  -1.5182   0.1293
## unemploy     -0.0227     0.0029  -7.7623   <2e-16 ***
## totalvmt      0.0000     0.0000   0.4241   0.6716
## precip       -0.0245     0.0067  -3.6315   0.0003 ***
## snow32       -0.0161     0.0205  -0.7826   0.4340
## rural_speed  -0.0058     0.0016  -3.6010   0.0003 ***
## urban_speed   0.0030     0.0015   2.0262   0.0430 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is potential serial correlation between the composite error terms across each time period in the RE model, resulting in biased standard errors. The clustered standard errors are almost double the normal standard errors indicating that there is more variability at the cluster level (by state) than there is at the individual state level. Further, the number of cluster observations are fewer than individual state level observations, increasing the variance in clustered SEs.

This implies that the correlation error structure assumed in RE is incorrect and FE would be a better method in this case.

## 3f - FE estimator

```
#non-clustered results
formula <- 'fatal_pc ~ primary + time_sq'
fixed_eff <- felm( as.formula( paste( formula, '+ year | state | 0 | 0' ) ),
                   dat)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
summary(fixed_eff)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
##
## Call:
##    felm(formula = as.formula(paste(formula, "+ year | state | 0 | 0")),      data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34264 -0.05903 -0.00108  0.06826  0.34006
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## primary  -9.262e-02  1.377e-02  -6.726 2.86e-11 ***
## time_sq  -6.872e-04  4.546e-05 -15.117  < 2e-16 ***
## year1982 -1.210e-01  2.132e-02  -5.676 1.78e-08 ***
## year1983 -1.570e-01  2.126e-02  -7.384 3.12e-13 ***
## year1984 -1.443e-01  2.115e-02  -6.823 1.50e-11 ***
## year1985 -1.572e-01  2.100e-02  -7.484 1.51e-13 ***
## year1986 -9.714e-02  2.086e-02  -4.656 3.64e-06 ***
## year1987 -9.048e-02  2.066e-02  -4.380 1.30e-05 ***
## year1988 -7.115e-02  2.040e-02  -3.488 0.000507 ***
## year1989 -1.196e-01  2.012e-02  -5.945 3.76e-09 ***
## year1990 -1.305e-01  1.983e-02  -6.582 7.31e-11 ***
## year1991 -1.900e-01  1.955e-02  -9.720  < 2e-16 ***
## year1992 -2.329e-01  1.926e-02 -12.089  < 2e-16 ***
## year1993 -2.161e-01  1.901e-02 -11.367  < 2e-16 ***
## year1994 -2.040e-01  1.878e-02 -10.864  < 2e-16 ***
## year1995 -1.627e-01  1.860e-02  -8.746  < 2e-16 ***
## year1996 -1.502e-01  1.851e-02  -8.113 1.37e-15 ***
## year1997 -1.182e-01  1.850e-02  -6.388 2.51e-10 ***
## year1998 -1.120e-01  1.862e-02  -6.018 2.44e-09 ***
## year1999 -9.350e-02  1.884e-02  -4.963 8.09e-07 ***
## year2000 -8.685e-02  1.923e-02  -4.516 7.01e-06 ***
## year2001 -5.652e-02  1.976e-02  -2.861 0.004312 **
## year2002 -1.780e-02  2.046e-02  -0.870 0.384620
## year2003         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1057 on 1055 degrees of freedom
## Multiple R-squared(full model): 0.9072   Adjusted R-squared: 0.901
## Multiple R-squared(proj model): 0.5311   Adjusted R-squared: 0.4995
## F-statistic(full model):145.3 on 71 and 1055 DF, p-value: < 2.2e-16
## F-statistic(proj model): 15.92 on 24 and 1055 DF, p-value: < 2.2e-16
```

```
#non-clustered results
formula <- 'fatal_pc ~ primary + time_sq'
fixed_eff <- felm( as.formula( paste( formula, '+ year | state | 0 | state' ) ),
                   dat)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
summary(fixed_eff)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
##
## Call:
##    felm(formula = as.formula(paste(formula, "+ year | state | 0 | state")),      data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34264 -0.05903 -0.00108  0.06826  0.34006
##
## Coefficients:
##            Estimate Cluster s.e. t value Pr(>|t|)
## primary  -9.262e-02    3.302e-02  -2.805 0.005124 **
## time_sq  -6.872e-04    7.351e-05  -9.349  < 2e-16 ***
## year1982 -1.210e-01    1.647e-02  -7.348 4.01e-13 ***
## year1983 -1.570e-01    1.903e-02  -8.250 4.69e-16 ***
## year1984 -1.443e-01    2.188e-02  -6.594 6.76e-11 ***
## year1985 -1.572e-01    2.291e-02  -6.862 1.16e-11 ***
## year1986 -9.714e-02    2.522e-02  -3.852 0.000124 ***
## year1987 -9.048e-02    2.915e-02  -3.104 0.001962 **
## year1988 -7.115e-02    2.741e-02  -2.596 0.009557 **
## year1989 -1.196e-01    2.716e-02  -4.405 1.17e-05 ***
## year1990 -1.305e-01    2.372e-02  -5.503 4.69e-08 ***
## year1991 -1.900e-01    2.292e-02  -8.292 3.37e-16 ***
## year1992 -2.329e-01    2.348e-02  -9.915  < 2e-16 ***
## year1993 -2.161e-01    2.225e-02  -9.710  < 2e-16 ***
## year1994 -2.040e-01    1.942e-02 -10.504  < 2e-16 ***
## year1995 -1.627e-01    2.107e-02  -7.723 2.64e-14 ***
## year1996 -1.502e-01    1.824e-02  -8.235 5.25e-16 ***
## year1997 -1.182e-01    1.655e-02  -7.141 1.72e-12 ***
## year1998 -1.120e-01    1.682e-02  -6.660 4.40e-11 ***
## year1999 -9.350e-02    1.697e-02  -5.508 4.55e-08 ***
## year2000 -8.685e-02    1.444e-02  -6.015 2.48e-09 ***
## year2001 -5.652e-02    1.246e-02  -4.536 6.38e-06 ***
## year2002 -1.780e-02    1.081e-02  -1.647 0.099905 .
## year2003         NA    0.000e+00      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1057 on 1055 degrees of freedom
## Multiple R-squared(full model): 0.9072   Adjusted R-squared: 0.901
## Multiple R-squared(proj model): 0.5311   Adjusted R-squared: 0.4995
## F-statistic(full model, *iid*):145.3 on 71 and 1055 DF, p-value: < 2.2e-16
## F-statistic(proj model): 21.86 on 24 and 48 DF, p-value: < 2.2e-16
```

The FE estimate using only primary and quadratic year covariates shows that fatalities per capita reduce by 9.3 percent on average in states that have a primary seatbelt law. The normal robust standard error is less than half of the clustered standard error indicating that correlation within states over time is biasing the normal se.

The normal and clustered standard errors are different due to the serial correlation between the error terms of the states over time.

## 3g - stability of FE

```
formula <- 'fatal_pc ~ primary + college + beer + primary + secondary + population + unemploy +
totalvmt + precip + snow32 + rural_speed + urban_speed'
fixed_eff <- felm( as.formula( paste( formula, '+ year | state | 0 | 0' ) ),
                    dat)
summary(fixed_eff)
```

```
##
## Call:
##    felm(formula = as.formula(paste(formula, "+ year | state | 0 | 0")),      data = dat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.288886 -0.049080  0.001332  0.048947  0.312285
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## primary     -9.116e-02  1.413e-02  -6.452 1.69e-10 ***
## college     -4.935e-01  2.311e-01  -2.136 0.032930 *
## beer         6.219e-01  3.947e-02  15.758  < 2e-16 ***
## secondary   -1.836e-02  1.007e-02  -1.823 0.068516 .
## population  -5.952e-05  1.160e-05  -5.132 3.41e-07 ***
## unemploy    -2.523e-02  2.704e-03  -9.329  < 2e-16 ***
## totalvmt     3.491e-06  7.973e-07   4.378 1.32e-05 ***
## precip      -2.146e-02  5.960e-03  -3.601 0.000331 ***
## snow32       8.582e-03  1.332e-02   0.644 0.519414
## rural_speed  3.065e-03  1.191e-03   2.573 0.010220 *
## urban_speed  1.400e-03  8.550e-04   1.637 0.101946
## year1982    -5.366e-02  1.760e-02  -3.050 0.002350 **
## year1983    -8.016e-02  1.789e-02  -4.480 8.28e-06 ***
## year1984    -1.114e-01  1.702e-02  -6.546 9.23e-11 ***
## year1985    -1.295e-01  1.725e-02  -7.511 1.26e-13 ***
## year1986    -7.942e-02  1.764e-02  -4.501 7.51e-06 ***
## year1987    -9.110e-02  1.862e-02  -4.893 1.15e-06 ***
## year1988    -1.331e-01  2.212e-02  -6.017 2.45e-09 ***
## year1989    -1.849e-01  2.244e-02  -8.239 5.17e-16 ***
## year1990    -2.174e-01  2.211e-02  -9.830  < 2e-16 ***
## year1991    -2.404e-01  2.240e-02 -10.733  < 2e-16 ***
## year1992    -2.796e-01  2.308e-02 -12.114  < 2e-16 ***
## year1993    -2.817e-01  2.358e-02 -11.945  < 2e-16 ***
## year1994    -2.965e-01  2.443e-02 -12.135  < 2e-16 ***
## year1995    -2.720e-01  2.552e-02 -10.657  < 2e-16 ***
## year1996    -2.923e-01  2.860e-02 -10.222  < 2e-16 ***
## year1997    -2.946e-01  3.013e-02  -9.778  < 2e-16 ***
## year1998    -3.203e-01  3.173e-02 -10.094  < 2e-16 ***
## year1999    -3.483e-01  3.228e-02 -10.788  < 2e-16 ***
## year2000    -3.719e-01  3.313e-02 -11.226  < 2e-16 ***
## year2001    -3.511e-01  3.318e-02 -10.581  < 2e-16 ***
```

```
## year2002    -3.234e-01  3.343e-02  -9.675  < 2e-16 ***
## year2003    -3.096e-01  3.420e-02  -9.053  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08188 on 1045 degrees of freedom
## Multiple R-squared(full model): 0.9448   Adjusted R-squared: 0.9405
## Multiple R-squared(proj model): 0.721   Adjusted R-squared: 0.6994
## F-statistic(full model):220.9 on 81 and 1045 DF, p-value: < 2.2e-16
## F-statistic(proj model): 81.84 on 33 and 1045 DF, p-value: < 2.2e-16
```

The FE estimate gives us an ATE of -9 percent which is the same as the OLS estimator while the standard errors are lower for FE estimates (both without clustering). The FE estimates are more stable than the pooled OLS estimates since they account for time invariant state-specific effects which are usually unobserved, and hence cannot be explicitly included in the pooled OLS regression.