Section 4: Panel Data

1 Introduction

In the first part of this course, we looked at *cross-sectional* data sets. The linear model that we wanted to estimate had one observation for each unit i and a total of N observations:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + u_i,\ i = 1, \ldots, N \tag{1}$$

In contrast, in a *panel* data set we have observations $(Y_{it}, X_{1it}, ..., X_{Kit})$ at $T \geq 2$ time periods for each unit i for a total of $N \cdot T$ observations:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \ldots + \beta_K X_{Kit} + u_{it},\ i = 1, \ldots, N;\ t = 1, \ldots, T \tag{2}$$

For this section, we are going to use a panel dataset *alcohol.dta* which contains vehicle fatality, beer tax and other covariates in 48 states between 1982 and 1988. We are going to explore the question whether vehicle fatality responds to state taxes on beer.

Q1: What does i represent for this dataset? What does t represent? Is this a balanced panel or an unbalanced panel?

2 Fixed Effects

2.1 Entity Fixed Effects:

Some variables do not vary over time but only between entities. Let's call those W_j's. There could be many W_j's. The grouping j of the entity may be different from the unit i of the panel. For our panel data example, we will explore 1) when j is defined as the census region the state belongs to, and 2) when j is defined as the state.

Our dependent variable is $vfrall$, the number of traffic deaths in a given state in a given year per 10,000 people living in that state in that year, and we have one **time-varying** independent variable *beertax*, the state excise tax on a case of beer. Let's suppose we also have one or more **time-constant** independent variables $W's$:

$$vfrall_{it} = \beta_0 + \beta_1 beertax_{it} + \gamma_1 W_{1j} + \gamma_2 W_{2j} + u_{it} \tag{3}$$

We can re-write all of an entity' time-constant variables as a single time-constant variable α_j. This is the **entity fixed effect**.

Q2: How do we implement entity fixed effects?

Q3: What is α_j equal to for Equation (3)?

Q4: How would you interpret α_j?

Q5: What standard errors would you use?

Q6: What are some examples of time-invariant variables?

Regression1: controlling for region fixed effects (i.e. j denotes region of state i).

$$vfrall_{it} = \beta_0 + \beta_1 beertax_{it} + \alpha_j + u_{it}$$

```
. regress vfrall beertax i.region_numeric, vce(cluster state_fips)

Linear regression                               Number of obs     =        336
                                                F(4, 47)          =      21.06
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4940
                                                Root MSE          =     .40803

                                  (Std. Err. adjusted for 48 clusters in state_fips)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |   .2948787   .1161945     2.54   0.015     .0611255     .528632
             |
region_numeric |
           2 |   .1854124   .1051148     1.76   0.084    -.0260512    .3968761
           3 |   .6248305   .1282668     4.87   0.000     .3667909    .8828701
           4 |   .9878274   .2089068     4.73   0.000     .5675612    1.408094
             |
       _cons |   1.408088   .0944773    14.90   0.000     1.218025    1.598152
------------------------------------------------------------------------------
```

Regression 2: controlling for state fixed effects (i.e. $j = i$ denotes the state itself):

$$vfrall_{it} = \beta_0 + \beta_1 beertax_{it} + \tilde{\alpha}_i + u_{it}$$

```
. reghdfe vfrall beertax i.region_numeric, absorb(state_fips) vce(cluster state_fips)
note: 2bn.region_numeric is probably collinear with the fixed effects (all partialled-out values are close to zero; tol = 1.0e-09)
note: 3bn.region_numeric is probably collinear with the fixed effects (all partialled-out values are close to zero; tol = 1.0e-09)
note: 4bn.region_numeric is probably collinear with the fixed effects (all partialled-out values are close to zero; tol = 1.0e-09)
(MWFE estimator converged in 1 iterations)
note: 2.region_numeric omitted because of collinearity
note: 3.region_numeric omitted because of collinearity
note: 4.region_numeric omitted because of collinearity

HDFE Linear regression                            Number of obs   =        336
Absorbing 1 HDFE group                            F(   1,     47) =       5.05
Statistics robust to heteroskedasticity           Prob > F        =     0.0294
                                                  R-squared       =     0.9050
                                                  Adj R-squared   =     0.8891
                                                  Within R-sq.    =     0.0407
Number of clusters (state_fips) =         48      Root MSE        =     0.1899

                              (Std. Err. adjusted for 48 clusters in state_fips)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.6558736   .2918556    -2.25   0.029    -1.243011   -.0687358
             |
region_numeric |
           2 |          0  (omitted)
           3 |          0  (omitted)
           4 |          0  (omitted)
             |
       _cons |   2.377075   .1497966    15.87   0.000     2.075723    2.678427
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
  state_fips |         48          48           0    *|
-----------------------------------------------------+
* = FE nested within cluster; treated as redundant for DoF computation
```

Q7: What happens to the region fixed effects when you add state fixed effects?

2.2 Time Fixed Effects:

Similarly, some variables do not vary over entities but only over time, let us summarize those in a **time fixed effect** μ_t.

Note: μ_t isn't indexed over entities, since it doesn't vary between different entities.

Q8: What are examples of relevant variables do not vary between states, but only vary over time?

Regression 3: controlling for state fixed effects and year fixed effects:

$$vfrall_{it} = \beta_0 + \beta_1 beertax_{it} + \tilde{\alpha}_i + \mu_t + u_{it}$$

```
. reghdfe vfrall beertax, absorb(state_fips year) vce(cluster state_fips)
(MWFE estimator converged in 2 iterations)

HDFE Linear regression                            Number of obs   =        336
Absorbing 2 HDFE groups                           F(   1,     47) =       3.21
Statistics robust to heteroskedasticity           Prob > F        =     0.0795
                                                  R-squared       =     0.9089
                                                  Adj R-squared   =     0.8910
                                                  Within R-sq.    =     0.0361
Number of clusters (state_fips) =         48      Root MSE        =     0.1882

                          (Std. Err. adjusted for 48 clusters in state_fips)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |   -.6399799   .3570783    -1.79   0.080    -1.358329    .0783691
       _cons |    2.368917   .1832726    12.93   0.000      2.00022    2.737614
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
  state_fips |         48          48           0    *|
        year |          7           0           7     |
-----------------------------------------------------+
* = FE nested within cluster; treated as redundant for DoF computation
```

2.3 Time Fixed Effects x Other Fixed Effects:

Including entity and time fixed effects may not eliminate OVB if the omitted variables vary over time within entities. One way to control for some of these omitted variables is to include interaction terms. If you observe multiple people in a state, you can include a state by time fixed effect to control for factors that are common to everyone in a state that vary over time. If you observe multiple states within a Census region, you can include a region by time fixed effect to control for factors common to all states in the region that vary over time. If you observe multiple companies within an industry, you can include industry by time fixed effects to control for factors common to all companies in the industry that vary over time.

Regression 4: including the interaction between region and year fixed effects (let's denote this ϕ_{jt}), thereby controlling for all variables that vary over time within a region:

$$vfrall_{it} = \beta_0 + \beta_1 beertax_{it} + \phi_{jt} + u_{it}$$

```
. reghdfe vfrall beertax, absorb(state_fips year year#region_numeric) vce(cluster state_fips)
(MWFE estimator converged in 2 iterations)

HDFE Linear regression                            Number of obs   =        336
Absorbing 3 HDFE groups                           F(   1,     47) =       0.95
Statistics robust to heteroskedasticity           Prob > F        =     0.3342
                                                  R-squared       =     0.9143
                                                  Adj R-squared   =     0.8891
                                                  Within R-sq.    =     0.0163
Number of clusters (state_fips) =         48      Root MSE        =     0.1899

                          (Std. Err. adjusted for 48 clusters in state_fips)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |   -.469676   .4813167    -0.98   0.334    -1.43796    .4986084
       _cons |   2.281508   .2470387     9.24   0.000     1.78453    2.778486
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------------------------------
          Absorbed FE | Categories  - Redundant  = Num. Coefs   |
----------------------+------------------------------------------+
            state_fips |         48          48           0    *|
                  year |          7           0           7     |
   year#region_numeric |         28           7          21     |
-----------------------------------------------------------------------------
* = FE nested within cluster; treated as redundant for DoF computation
```

Note that ϕ_{jt} contains all factors included in both α_j and μ_t (and more).

Q9: Why do we not include the interaction between state and year fixed effects?

2.4 Omitted Variable Bias and Other Key Insights:

Panel data allows us to control for entity and time fixed effects even if we do not observe them.

What type of variation is each using?

- **Entity fixed effect regressions exploit the variation *within* an entity over time periods - by taking away the part of the entity that is common in multiple periods (α_j).**
- **Time fixed effect regressions exploit the variation *between* entities within a time period - by taking away the part of a time period that is common in multiple entities (μ_t).**
- **Both types of fixed effects can be combined to exploit variation *between* entities within a time period ("between" variation) and to exploit variation *within* entities over time ("within" variation).**
- **Thus we can effectively control for omitted variables that vary over time (and not over entities) or over entities (and not over time) in panel data.**

Q10: What type of omitted variable does Regression 3 takes care of as compared to Regression 2?

Q11: Can there still be omitted variable bias in Regression 3 and Regression 4?

Q12: Is there a downside to fixed effects regressions?

3 Heterogeneity

Q13: How could we modify Regression 4 to test whether the effect of beer tax depends also on whether the minimum legal drinking age (MLDA) is 21 or under 21?

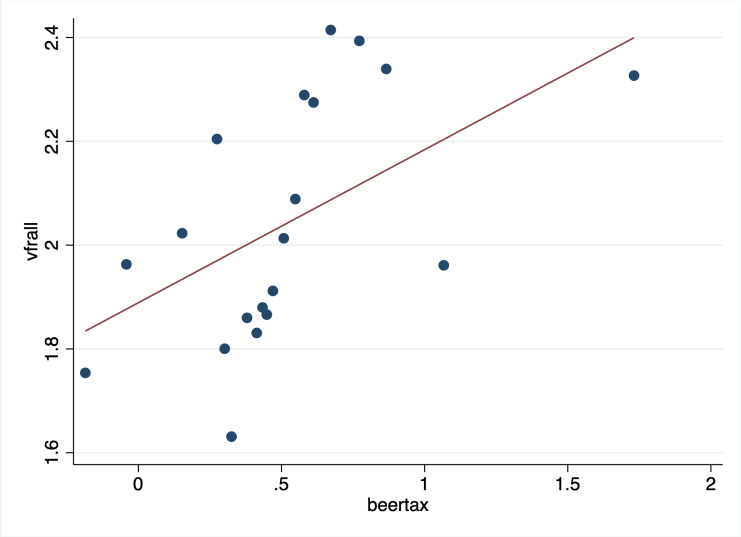
```
gen MLDA_21 = (mlda == 21)
gen beertax_MLDA_21 = MLDA_21 * beertax

reghdfe vfrall beertax beertax_MLDA_21 MLDA_21 , absorb(state_fips year year#region_numeric) vce(cluster state_fips)
```

```
. reghdfe vfrall beertax beertax_MLDA_21 MLDA_21 , absorb(state_fips year year#region_numeric) vce(cluster state_fips)
(MWFE estimator converged in 2 iterations)

HDFE Linear regression                            Number of obs   =        336
Absorbing 3 HDFE groups                           F(   3,     47) =       0.61
Statistics robust to heteroskedasticity           Prob > F        =     0.6088
                                                  R-squared       =     0.9148
                                                  Adj R-squared   =     0.8890
                                                  Within R-sq.    =     0.0227
Number of clusters (state_fips) =         48      Root MSE        =     0.1900

                            (Std. Err. adjusted for 48 clusters in state_fips)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |   -.512686   .4984715    -1.03   0.309    -1.515481    .4901094
beertax_MLDA_21 |  -.0010957   .0844259    -0.01   0.990    -.1709387    .1687473
     MLDA_21 |   .0580559   .0873248     0.66   0.509    -.1176189    .2337307
       _cons |   2.263872   .2504731     9.04   0.000     1.759985    2.767758
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------------------------------
          Absorbed FE | Categories  - Redundant  = Num. Coefs   |
----------------------+-----------------------------------------|
            state_fips |         48          48           0    *|
                  year |          7           0           7     |
   year#region_numeric |         28           7          21     |
-----------------------------------------------------------------------------
* = FE nested within cluster; treated as redundant for DoF computation
```

4 Appendix: Visualization

In this part of the section, we present two ways to visualize fixed-effects regression results by residualizing the outcome variable $vfrall$ and the explanatory variable of interest $beertax$. The first method uses the $,controls()$ option with $binscatter$.

```
* Installing binscatter
ssc install binscatter

* Creating a binscatter plot for Regression 1
* the ,controls() option residualizes vfrall and beertax on i.region_numeric before plotting
binscatter vfrall beertax, controls( i.region_numeric)
graph export reg1.png, replace
```



The second method residualizes the outcome variable $vfrall$ and the explanatory variable of interest $beertax$ manually.

```
* Creating a binscatter plot for Regression 2
*residualize fataility rate and beer tax first
reghdfe vfrall , absorb(state_fips year) vce(cluster state_fips) residuals(vfrall_res2)
reghdfe beertax, absorb(state_fips year) vce(cluster state_fips) residuals(beertax_res2)

*Now plot data
*Note that slope of this line exactly equals regression 2 above
binscatter vfrall_res2 beertax_res2
graph export reg2.png, replace
```