Econometrics I

Lecture 8: Panel Data

Paul T. Scott
NYU Stern

Fall 2021

Panel Data

- Our models so far indexed observations by i:

$$y_i = \boldsymbol{\beta}\mathbf{x}_i + \varepsilon_i.$$

- Panel data adds a t subscript to the data:

$$y_{it} = \boldsymbol{\beta}\mathbf{x}_{it} + \varepsilon_{it}.$$

- Typically, i refers to individuals observed on multiple occasions over time, and t indexes the time periods. E.g., 50 states are observed in each of 8 years, and have one row of data for each state-year, so 400 observations.
- Main econometric concern: heterogeneity.

Terminology

- **Longitudinal data**: another term for panel data
- **Repeated cross section**: a data structure with multiple individuals observed in each of multiple time periods. In contrast to panel data, we don't observe the same individuals in multiple time periods.
- **Balanced panel**: each of n individuals is observed T times, usually over the same time period
- **Unbalanced panel**: at least of the individuals are not observed in every period. Sometimes unbalanced panels result from sampling designs, and sometimes they are a result of entry/exit or birth/death
- A **wide panel** has many individuals (large n); a **long panel** has many time periods (large T). The asymptotic properties of an estimator can be different when $n \to \infty$ as opposed to $T \to \infty$. Sometimes we need both.

Panel Data Model

- The typical regression equation with individual effects:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{it}.$$

  If $\mathbf{z}_i$ is observed, no problem. OLS works just fine.

- Applying standard OLS to panel data is called a **pooled regression**. In this case, there's actually nothing special about having the two indices. You're doing the same thing you would be doing if you didn't know how to group observations by individual.
- If $\mathbf{z}_i$ is unobserved and correlated with $\mathbf{x}$, pooled regression will suffer from omitted variables bias
- The $\mathbf{z}_i'\alpha$ term is known as a **fixed effect** because is is fixed across t for individual i.

Panel Data Model

- Alternatively, we might write:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma_i + \varepsilon_{it}.$$

where $\gamma_i = \mathbf{z}_i'\boldsymbol{\alpha}$ is the fixed effect.

- This version emphasizes that the fixed effect can be seen as an unobserved parameter to be estimated (one parameter for each individual i).

First Differences

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \gamma_i + \varepsilon_{it}.$$

- Note that we can eliminate the individual effects by taking first differences:
$$\Delta y_{it} = \Delta \mathbf{x}'_{it}\boldsymbol{\beta} + \Delta\varepsilon_{it}.$$
where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and similarly for other variables.
- As long as $\Delta \mathbf{x}_{it}$ is uncorrelated with $\Delta\varepsilon_{it}$, we're in business. We no longer have to worry about correlation between x and z.
- Notice that $y_{it} - y_{i,t-2}$ would also difference out the individual effects. So would $y_{it} - \frac{1}{2}y_{i,t-1} - \frac{1}{2}y_{i,t-2}$. This leads us to another estimator ...

Fixed Effects

- The **fixed effects** estimator relies on a similar idea, but we de-mean variables instead of first differencing:

$$\tilde{y}_{it} = y_{it} - \bar{y}_i$$

  where $\bar{y}_i$ is the mean value of y_{it} for individual i, and similarly for other variables.

- The fixed effects estimator amounts to applying OLS to the de-meaned variables:

$$\tilde{y}_{it} = \tilde{\mathbf{x}}'_{it}\boldsymbol{\beta} + \varepsilon_{it}$$

- Like first differences, fixed effects differences out the individual fixed effects, therefore avoiding endogeneity problems coming from correlation between the individual fixed effects and $\mathbf{x}_{it}$

Fixed Effects as Coefficients on Dummies

- Let's try to directly estimate the individual fixed effects:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \sum_{j=1}^{n} \gamma_j D\left[i == j\right] + \varepsilon_{it}.$$

  where γ_j is the fixed effect for individual j, and $D\left[i == j\right]$ is a dummy variable indicating whether the observed individual is the jth individual.

- This is known as the **least squares with dummy variables** (LSDV) estimator.

FE and LSDV estimators

- The fixed effects estimator and LSDV are *equivalent*: both yield exactly the same estimate of $\boldsymbol{\beta}$.
- Note that with a wide panel (large n), there can be a huge number of γ_j parameters, in extreme cases making it infeasible for a computer to invert the $\mathbf{X}'\mathbf{X}$ as required by OLS. However, de-meaning variables is computationally relatively easy.

Review: Frisch-Waugh Theorem

- Separate $\mathbf{X}$ into two sub-matrices:

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2],$$

where

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Frisch-Waugh Theorem

The OLS regression of $\mathbf{y}$ on $[\mathbf{X}_1, \mathbf{X}_2]$ yields a subvector $\mathbf{b}_2$ of coefficient estimates that is the same as the result from a regression of the residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ are regressed on the residuals from a regression of $\mathbf{X}_2$ on $\mathbf{X}_1$.

Matrix of Dummies for Individuals

$$\mathbf{X}_2 = \left(\begin{array}{rl} \text{individual 1} & \left\{ \begin{array}{cccc} 1 & 0 & 0 & \dots \\ 1 & 0 & 0 & \dots \\ 1 & 0 & 0 & \dots \end{array} \right. \\ \text{individual 2} & \left\{ \begin{array}{cccc} 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & \dots \end{array} \right. \\ \text{individual 3} & \left\{ \begin{array}{cccc} 0 & 0 & 1 & \dots \\ 0 & 0 & 1 & \dots \\ 0 & 0 & 1 & \dots \end{array} \right. \\ & \vdots \end{array} \right)$$

Fixed Effects and Frisch-Waugh

- Suppose that $\mathbf{X}_1$ is a matrix of regressors, and $\mathbf{X}_2$ includes only the dummy variables for the individuals.
  - With a balanced panel, $\mathbf{X}_1$ is $nT \times K$, where K is the number of regressors,
    and $\mathbf{X}_2$ is $nT \times n$.
  - Also, $\mathbf{X}_1$ should not have a constant variable, nor any other time-invariant variables.
- For the first step of Frisch-Waugh, we regress $\mathbf{y}$ and $\mathbf{X}_1$ on $\mathbf{X}_2$, and then take the residuals. What does this amount to?
- Regressing on $\mathbf{X}_2$ and taking residuals is the same as subtracting out means by individual i.

Individual Effects Estimates

- Note that the LSDV estimator produces direct estimates of the individual effects γ_i: the coefficients on the dummy variables. How can we get them from the FE estimator?
- Obtain residuals in the usual way:

$$\mathbf{e}_{it} = y_{it} - \mathbf{x}_{it}'\mathbf{b}_{FE}.$$

- We can then estimate the individual effects as follows:

$$\hat{\gamma}_i = T^{-1}\sum_t \mathbf{e}_{it}.$$

These estimates will be the same as the coefficients on the dummies from LSDV.

Panel Data Asymptotics: A Preview

$$\hat{\gamma}_i = T^{-1} \sum_t \mathbf{e}_{it}.$$

- In a balanced panel, the full sample size is n.
- However, each fixed effect is estimated with T observations.
- In general, we might worry about having a bunch of parameters that are only informed by a small fraction of the data (see **incidental parameters**)– this is especially a concern when n is large and T is small.
- In the linear regression model, the incidental parameters don't affect the consistency of our estimate of β. This is related to the fact that the FE estimator estimates β without needing to actually estimate the fixed effects. Econometrics II will deal with this in more detail.

Terminology: Within Estimator

- The FE estimator is sometimes called the **within estimator**. This is because we first transform the data to eliminate differences in means between the individuals, and then use variation *within* the individuals to estimate $\boldsymbol{\beta}$.
- Note: after de-meaning, every i has a zero mean:

  $$E\left[y_{it} - \bar{y}_i\right] = 0,$$

  and the same is true of each regressor.

Terminology: Between Estimator

- There's also a **between estimator** that runs a regression with the *n* aggregate observations after taking group means:

$$\bar{y}_i = \bar{\mathbf{x}}_i' \beta + \varepsilon_i^*.$$

- Note that the individual effects will be part of the within-group error term ε_i^*.
- The between estimator doesn't come up much. It eliminates some of the variation in the data (and reduces the number of observations) without dealing with the potential correlation between **x** and the individual effects.

R^2 for Panel Data Models

- There is the R^2 of the LSDV regression (no special name).
- The **within** R^2 is the R^2 of the FE regression run with the de-meaned data, or the squared correlation between $\hat{y}_{it} - \hat{y}_i$ and $y_{it} - \bar{y}_i$, where

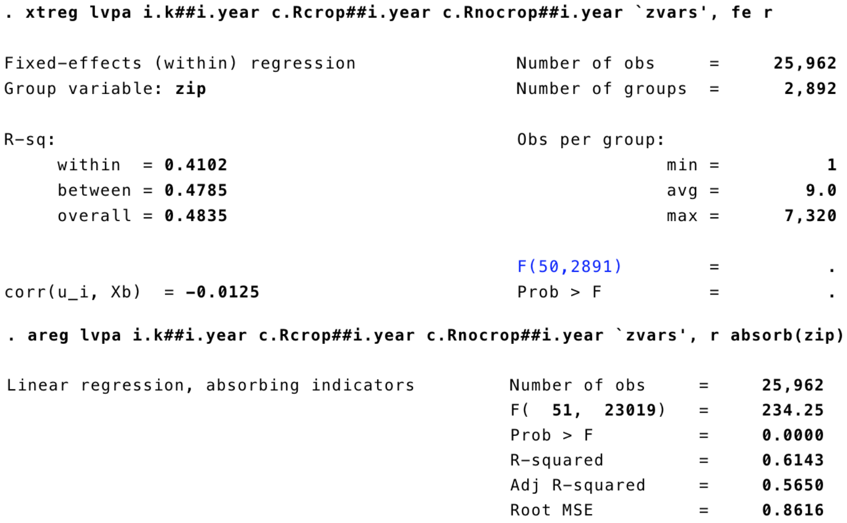
$$\begin{aligned} \hat{y}_{it} &= \mathbf{x}_{it}' \mathbf{b}_{FE}, \\ \hat{y}_i &= \bar{\mathbf{x}}_i' \mathbf{b}_{FE}. \end{aligned}$$

- The **between** R^2 is the squared correlation between $\hat{y}_i$ and $\bar{y}_i$. It's similar to the R^2 from the between regression, but we use the FE estimate of β for $\hat{y}_i$, not the between estimator.
- The **overall** R^2 is the squared correlation between y_{it} and $\hat{y}_{it}$.
  - This is different from the R^2 of the LSDV regression! Note that the $\hat{y}_{it}$ predictions don't involve the individual effects.

Goodness of Fit with Fixed Effects

- The overall R^2 tells us how much of the variation in the data is explained by data $\mathbf{x}$
  - It's different from the within R^2 because it's assessing how much of the variation gets explained without subtracting group means.
  - It's different from the LSDV R^2 because it doesn't use the fixed effects to predict $\hat{y}_{it}$.
- Note that R^2 from the LSDV regression can be misleading if we're interested in understanding how well x explains y. This R^2 could be high if the fixed effects themselves explain a lot of the variation in y even if $\mathbf{b}_{FE} = \mathbf{0}$.

Example: Some Hedonic Regressions



```
. xtreg lvpa i.k##i.year c.Rcrop##i.year c.Rnocrop##i.year `zvars', fe r

Fixed-effects (within) regression               Number of obs     =     25,962
Group variable: zip                             Number of groups  =      2,892

R-sq:                                           Obs per group:
     within  = 0.4102                                         min =          1
     between = 0.4785                                         avg =        9.0
     overall = 0.4835                                         max =      7,320

                                                F(50,2891)        =          .
corr(u_i, Xb)  = -0.0125                        Prob > F          =          .

. areg lvpa i.k##i.year c.Rcrop##i.year c.Rnocrop##i.year `zvars', r absorb(zip)

Linear regression, absorbing indicators         Number of obs     =     25,962
                                                F(  51,  23019)   =     234.25
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6143
                                                Adj R-squared     =     0.5650
                                                Root MSE          =     0.8616
```

Random Effects Model

- Consider again the standard panel data regression equation:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \gamma_i + \varepsilon_{it},$$

  but now suppose that $\mathbf{x}'_{it}$ and γ_i are uncorrelated so that the individual effect does not create an endogeneity problem.

- Assuming that $\mathbf{x}'_{it}$ and ε_{it} are uncorrelated (which is also needed for the fixed effects estimator to be unbiased), OLS without fixed effects will deliver unbiased estimates here.

- But will OLS be efficient?
  - No – γ_i as part of the error term implies that the error term is correlated across t for a given i, violating the heteroscedasticity assumption.

Generalized Least Squares

- Think back to the standard linear regression framework, but now assume that

  $$Var\left(\varepsilon|\mathbf{X}\right) = \mathbf{\Omega},$$

  where $\mathbf{\Omega}$ need not be diagonal, correlation in the error terms is allowed.

- Then, the **generalized least squares** estimator is efficient, defined as

  $$\begin{aligned} \mathbf{b}_{GLS} &= \arg\min_{\mathbf{b}} \left(\mathbf{y}-\mathbf{Xb}\right)' \mathbf{\Omega}^{-1} \left(\mathbf{y}-\mathbf{Xb}\right) \\ &= \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y} \end{aligned}$$

- Note that if $\mathbf{\Omega}$ were the identity matrix, we would just have OLS.

- OLS here is still unbiased, but not BLUE. Now, GLS is BLUE.

The Random Effects Estimator

- Note that the covariance structure $\mathbf{\Omega}$ is typically not known *ex ante*. **Feasible generalized least squares** (FGLS) refers to two-step estimators in which
  1. We first estimate the model (typically with OLS) so that we can use the residuals to estimate $\mathbf{\Omega}$,
  2. Then we estimate the model again using GLS with our estimate of $\mathbf{\Omega}$.
- The FGLS estimator for the standard panel data model is known as the **random effects estimator**.

FE vs RE

- If the fixed effects are not correlated with regressors, then all three (OLS, FE, RE) are unbiased but RE will have the lowest variance of the three.
- **NB**: if the fixed effects are correlated with regressors, the FE is unbiased but OLS and RE are biased. *This limits the practical applicability of the random effects estimator* – in panel data contexts, we are often worried about endogeneity.

Testing Endogeneity

- Much as the Wu-Hausman test allowed us to use an IV estimator to test the OLS exogeneity assumption, the FE estimator can be used to test the RE exogeneity assumption.
- Intuition: see if RE and FE parameter estimates are significantly different from each other. If the γ_i terms are uncorrelated with the regressors, the two estimates should not be very different.
- This is known as the **Hausman test**. Link

Time Fixed Effects

- For some applications, we're concerned about unobservables that are changing over time but affect individuals in similar ways:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \gamma_t + \varepsilon_{it}.$$

- Mathematically, these time fixed effects can be dealt with in the same way as individual effects. That is, we can subtract group means by time period rather than by individual.

Two-Way Fixed Effects

- Sometimes we'd like to allow for time-specific effects as well as individual effects:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \gamma_i + \gamma_t + \varepsilon_{it}.$$

- This can be estimated using a LSDV strategy or by differencing with respect to one dimension and including dummy variables for the other. Also see two-way within estimators, or the documentation in R or Stata, for how to do two-way fixed effects with only differencing.

R Implementation

```
> pm1 <- plm(mrall~beertax, data=fatality, index = c("state","year"),
  effect = c("twoways"))
> coeftest(pm1, vcov = vcovHC(pm1, type = "HC1", cluster = "group"))

t test of coefficients:

          Estimate  Std. Error t value Pr(>|t|)
beertax -6.3998e-05  3.5015e-05 -1.8277  0.06865 .
---
    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

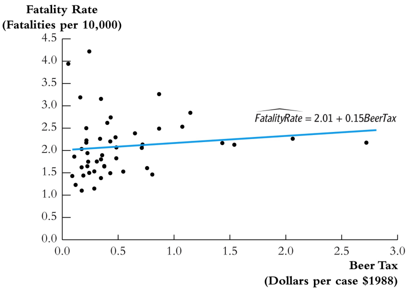
- The effect can be “individual”, “time” or “twoways”

Example: Traffic Deaths and Alcohol Taxes

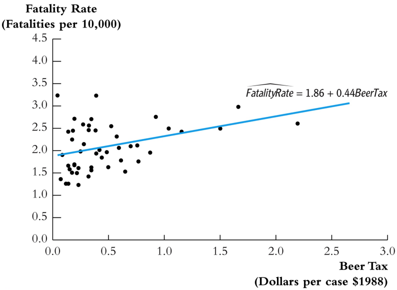
- $n = 48$ US states
- $T = 7$ years from 1982-1988
- Balanced panel
- Variables:
  - Traffic fatality rate (per 10000 residents)
  - Tax on a case of beer
  - Other controls including driving age, drunk driving laws
- Example borrowed from Stock & Watson.

FIGURE 10.1 **The Traffic Fatality Rate and the Tax on Beer**

Panel a is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Panel b shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.



(a) 1982 data



(b) 1988 data

Example: Traffic Deaths and Alcohol Taxes

- Why might there be more traffic deaths in states with higher alcohol taxes?
- There are several potential omitted variables:
  - Quality of automobiles
  - Quality of roads
  - Rate of driving vs public transit usage
  - Density of cars on the road
  - Drinking and driving culture

OVB Stories I

- Some potential sources of OVB:
  1. Western states have lower traffic density and lower alcohol taxes (due to their history as frontier states populated by outlaws and cowboys).
  2. Cultural attitudes that are relatively more critical of drinking and driving might lead to higher alcohol taxes and lead people to avoid drinking and driving.
- Panel data will allow us to avoid OVB bias from these sources if the omitted variables stay constant over time within each state. That is, if the sources of OVB are all included within the γ_i term:

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \mathbf{x}_{it}\boldsymbol{\beta}_2 + \gamma_i + \varepsilon_{it},$$

and we can use the standard fixed effects estimator (or LSDV).

OVB Stories II

- We might also worry about variation over time:
  1. Cars are getting safer (air bags)
  2. Changes in national laws about drinking, driving, or drinking and driving
- If these factors are shifting fatality rates for different states in the same way, we could avoid OVB bias from these factors. Let

  $$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \mathbf{x}_{it}\boldsymbol{\beta}_2 + \gamma_t + \varepsilon_{it},$$

  We can deal with the time fixed effects by subtracting means by time periods, or with LSDV with dummy variables for time periods.
- We can deal with both types of omitted variables using two-way fixed effects.

Fixed effects and Identifying Variation

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \mathbf{x}_{it}\boldsymbol{\beta}_2 + \gamma_i + \gamma_t + \varepsilon_{it},$$

- With individual fixed effects, recall that $\mathbf{x}_{it}$ cannot include any variables that are fixed by individual. Such variables would have no variation after subtracting group means. Equivalently, the variables would be colinear with the individual dummies.
- Similarly with time fixed effects, $\mathbf{x}_{it}$ cannot include any variables that are fixed within time period.
- With two-way fixed effects, to estimate the coefficient on a variable, that variable has to change in different ways for different individuals. Thus, to estimate β_1, it can't be the case that all states change their alcohol taxes in the same way each year.

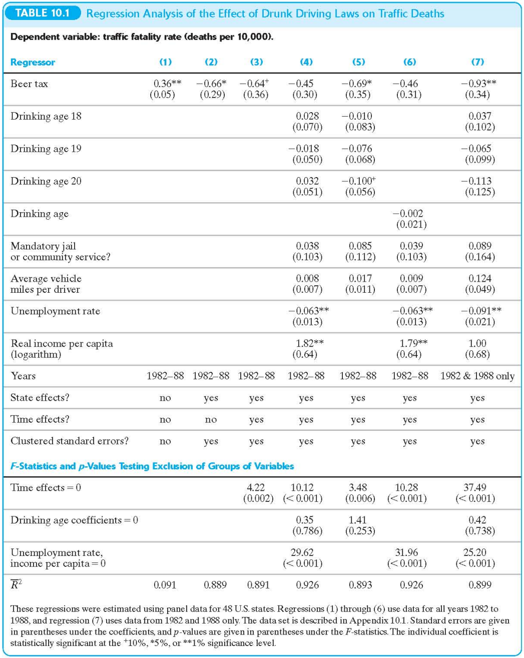


TABLE 10.1 Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent variable: traffic fatality rate (deaths per 10,000).

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36** (0.05)	−0.66* (0.29)	−0.64$^+$ (0.36)	−0.45 (0.30)	−0.69* (0.35)	−0.46 (0.31)	−0.93** (0.34)
Drinking age 18				0.028 (0.070)	−0.010 (0.083)		0.037 (0.102)
Drinking age 19				−0.018 (0.050)	−0.076 (0.068)		−0.065 (0.099)
Drinking age 20				0.032 (0.051)	−0.100$^+$ (0.056)		−0.113 (0.125)
Drinking age						−0.002 (0.021)	
Mandatory jail or community service?				0.038 (0.103)	0.085 (0.112)	0.039 (0.103)	0.089 (0.164)
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				−0.063** (0.013)		−0.063** (0.013)	−0.091** (0.021)
Real income per capita (logarithm)				1.82** (0.64)		1.79** (0.64)	1.00 (0.68)
Years	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes
***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**							
Time effects = 0			4.22 (0.002)	10.12 (< 0.001)	3.48 (0.006)	10.28 (< 0.001)	37.49 (< 0.001)
Drinking age coefficients = 0				0.35 (0.786)	1.41 (0.253)		0.42 (0.738)
Unemployment rate, income per capita = 0				29.62 (< 0.001)		31.96 (< 0.001)	25.20 (< 0.001)
$\overline{R}^2$	0.091	0.889	0.891	0.926	0.893	0.926	0.899

These regressions were estimated using panel data for 48 U.S. states. Regressions (1) through (6) use data for all years 1982 to 1988, and regression (7) uses data from 1982 and 1988 only. The data set is described in Appendix 10.1. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. The individual coefficient is statistically significant at the $^+$10%, *5%, or **1% significance level.

Two-way fixed effects and DiD

- To estimate a coefficient on a regressor with two-way fixed effects, $\Delta x_{it} = x_{i,t} - x_{i,t-1}$ must vary across individuals. That means that different individuals must experience *different changes* in x over time.
- If $T = 2$ and $x_{it} \in \{0, 1\}$, then two-way fixed effects becomes differences-in-differences. Recall that to estimate a DID model we need a treatment group that becomes treated within the sample ($\Delta x_{it} = 1$) and a control group that remains untreated ($\Delta x_{it} = 0$).