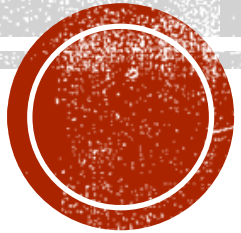
APPLIED ECONOMICS

Panel Data, Fixed Effects and Random Effects Model



Dr. Syed Hassan Raza
Assistant Professor of Economics
Quaid-i-Azam University

WHAT IS PANEL DATA?

- Panel data, also known as longitudinal data or cross-sectional time series data, is a type of dataset commonly used in various fields of research, including economics, social sciences, and epidemiology. Panel data is characterized by its unique structure, which combines both cross-sectional and time series dimensions.

ADVANTAGES OF PANEL DATA

- Panel data can model both the common and individual behaviors of groups.
- Panel data contains more information, more variability, and more efficiency than pure time series data or cross-sectional data.
- Panel data can detect and measure statistical effects that pure time series or cross-sectional data can't.
- Panel data can minimize estimation biases that may arise from aggregating groups into a single time series.

WHAT PANEL DATA LOOKS LIKE?

Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities (*i*) are observed across time (*t*).

(X_{it}, Y_{it}), $i=1,\ldots n$; $t=1,\ldots T$

These entities could be states, companies, families, individuals, countries, etc.

Entity	Year	Y	X1	X2	X3	.....
1	1	#	#	#	#	.....
1	2	#	#	#	#	.....
1	3	#	#	#	#	.....
:	:	:	:	:	:	:
2	1	#	#	#	#	.....
2	2	#	#	#	#	.....
2	3	#	#	#	#	.....
:	:	:	:	:	:	:
3	1	#	#	#	#	.....
3	2	#	#	#	#	.....
3	3	#	#	#	#	.....

NOTATIONS OF PANEL DATA

- Now it is the time to add time, which leads us to use **Panel Data**.
- A panel dataset contains observations on multiple entities, where each entity is observed at two or more points in time.
- If the data set contains observations on the variables X and Y ,then the data are denoted
  - $(X_{it}, Y_{it}),\ i = 1, \ldots n\ and\ t = 1, \ldots, T$

  - the first subscript,i refers to the entity being observed
  - the second subscript,t refers to the date at which it is observed
- Whether some observations are missing
  - **balanced** panel
  - **unbalanced** panel

DATA STRUCTURE?

TABLE 1.3 Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995

Observation Number	State	Year	Cigarette Sales (packs per capita)	Average Price per Pack (including taxes)	Total Taxes (cigarette excise tax + sales tax)
1	Alabama	1985	116.5	$1.022	$0.333
2	Arkansas	1985	128.5	1.015	0.370
3	Arizona	1985	104.5	1.086	0.362
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
47	West Virginia	1985	112.8	1.089	0.382
48	Wyoming	1985	129.4	0.935	0.240
49	Alabama	1986	117.2	1.080	0.334
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
96	Wyoming	1986	127.8	1.007	0.240
97	Alabama	1987	115.8	1.135	0.335
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

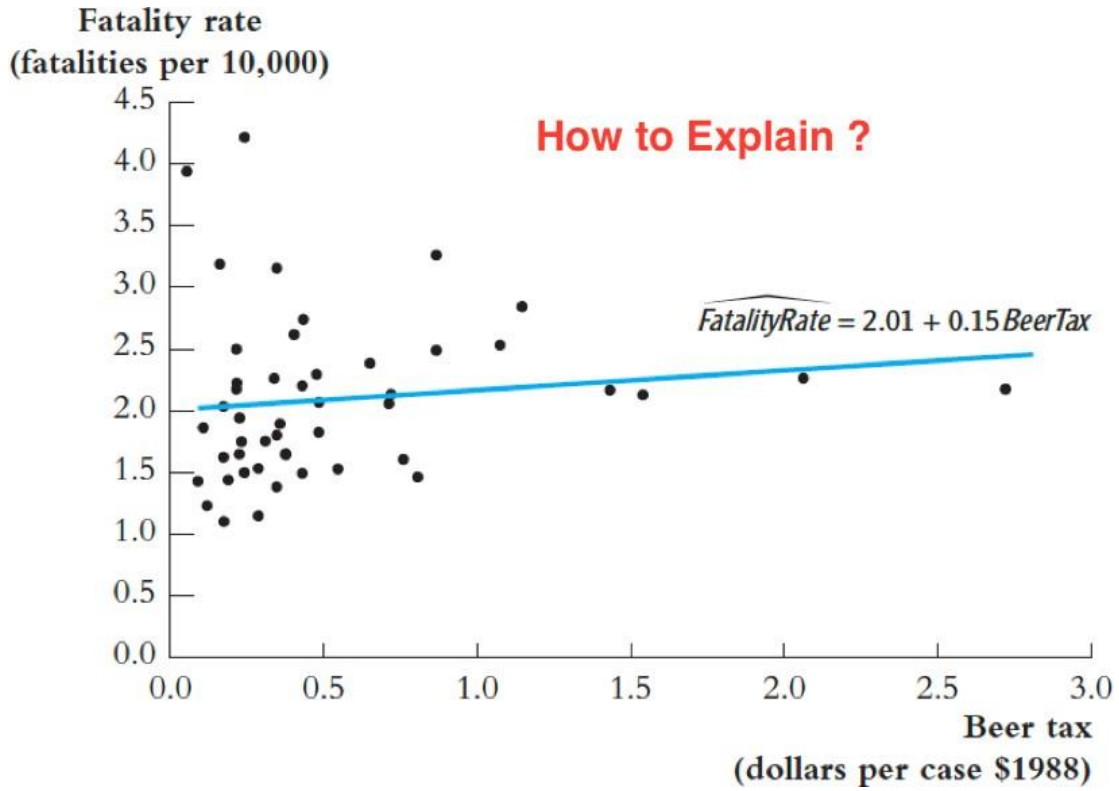
EXAMPLE: TRAFFIC DEATHS AND ALCOHOL TAXES

- Observational unit: *one* year in *one* U.S. state
- 48 U.S. states, so n = the number of entities = 48 and 7 years (1982,..., 1988),so T = # of time periods = 7. Balanced panel, so total number of observations = $7 \times 48 = 336$

- **Variables:**
  - Dependent Variable: **Traffic fatality rate** (# traffic deaths in that state in that year, per 10,000 state residents)
  - Independent Variable: **Tax on a case of Alcohol (beer)**
  - Other Controls (legal driving age, drunk driving laws, etc.)
- **A simple OLS regression model with $t = 1982, 1988$**
  - $FatalityRate_{it} = \beta_{0t} + \beta_{1t}\mathrm{Alcohol}Tax_{it} + u_{it}$

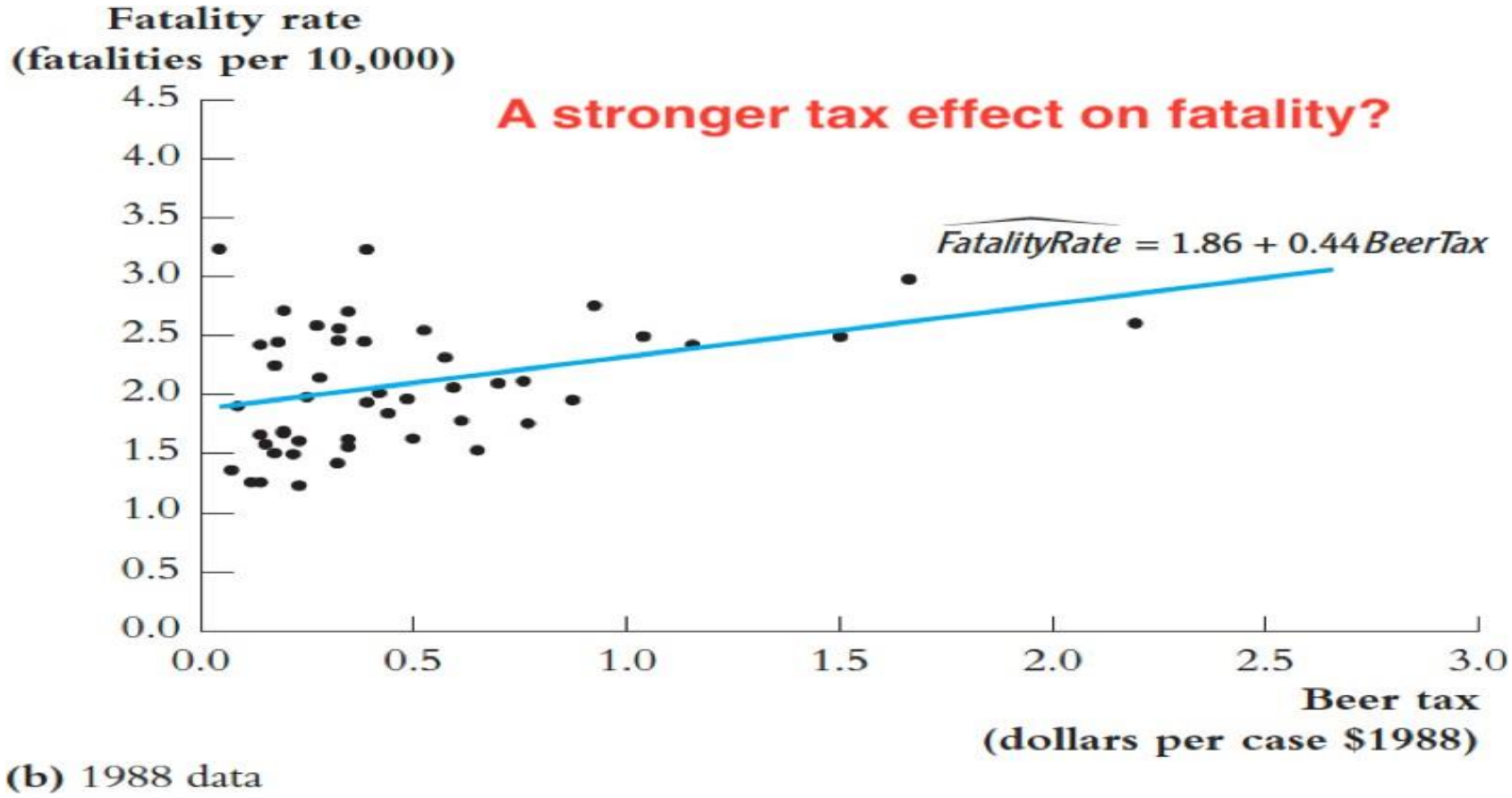
U.S. TRAFFIC DEATH DATA FOR 1982

- Higher alcohol taxes, more traffic deaths



U.S. TRAFFIC DEATH DATA FOR 1988

- Still higher alcohol taxes, more traffic deaths



(b) 1988 data

SIMPLE CASE: PANEL DATA WITH TWO TIME PERIODS

- Let adjust our model with some unobservables
  - $FatalityRate_{it} = \beta_0 + \beta_1 \mathrm{Alcohol}Tax_{it} + \beta_2 Z_i + u_{it}$
- where u_{it} is the error term and $i = 1, \ldots n$ *and* $t = 1, \ldots, T$
- Z_i is the unobservable factor that determines the fatality rate in the i state but **does not change over time**.

  eg. local cultural attitude toward drinking and driving.

- The omission of Z_i might cause omitted variable bias but we don't have data on Z_i.
- The key idea: *Any change* in the fatality rate from 1982 to 1988 cannot be caused by Z_i, because Z_i (by assumption) *does not change* between 1982 and 1988.

PANEL DATA WITH TWO TIME PERIODS: BEFORE AND AFTER MODEL

- Consider the regressions for 1982 and 1988…

  - $FatalityRate_{i1988} = \beta_0 + \beta_1 \text{Alcohol}Tax_{i1988} + \beta_2 Z_i + u_{i1988}$
  - $FatalityRate_{i1982} = \beta_0 + \beta_1 \text{Alcohol}Tax_{i1982} + \beta_2 Z_i + u_{i1982}$

- **Then make a difference**

- $FatalityRate_{i1988} - FatalityRate_{i1982} = \beta_1(\text{Alcohol}Tax_{i1988} - \text{Alcohol}Tax_{i1982}) + (u_{i1988} - u_{i1982})$

- F

PANEL DATA WITH TWO TIME PERIODS

- Assumption: if $E(u_{it}|\text{Alcohol}Tax_{it}, Z_{it}) = 0$, then $(u_{i1988} - u_{i1982})$ is uncorrelated with $(\text{Alcohol } Tax_{i1988} - \text{Alcohol } Tax_{i1982})$

- Then this "difference" equation can be estimated by OLS, even though Z_i isn't observed.

- Because the omitted variable Z_i doesn't change, it cannot be a determinant of the change in Y .

CASE: TRAFFIC DEATHS AND ALCOHOL TAXES

- 1982 data
- $\widehat{fatalityrate} = \underset{(0.11)}{1.82} + \underset{(0.13)}{0.44Alcohol}$ **(n=48)**

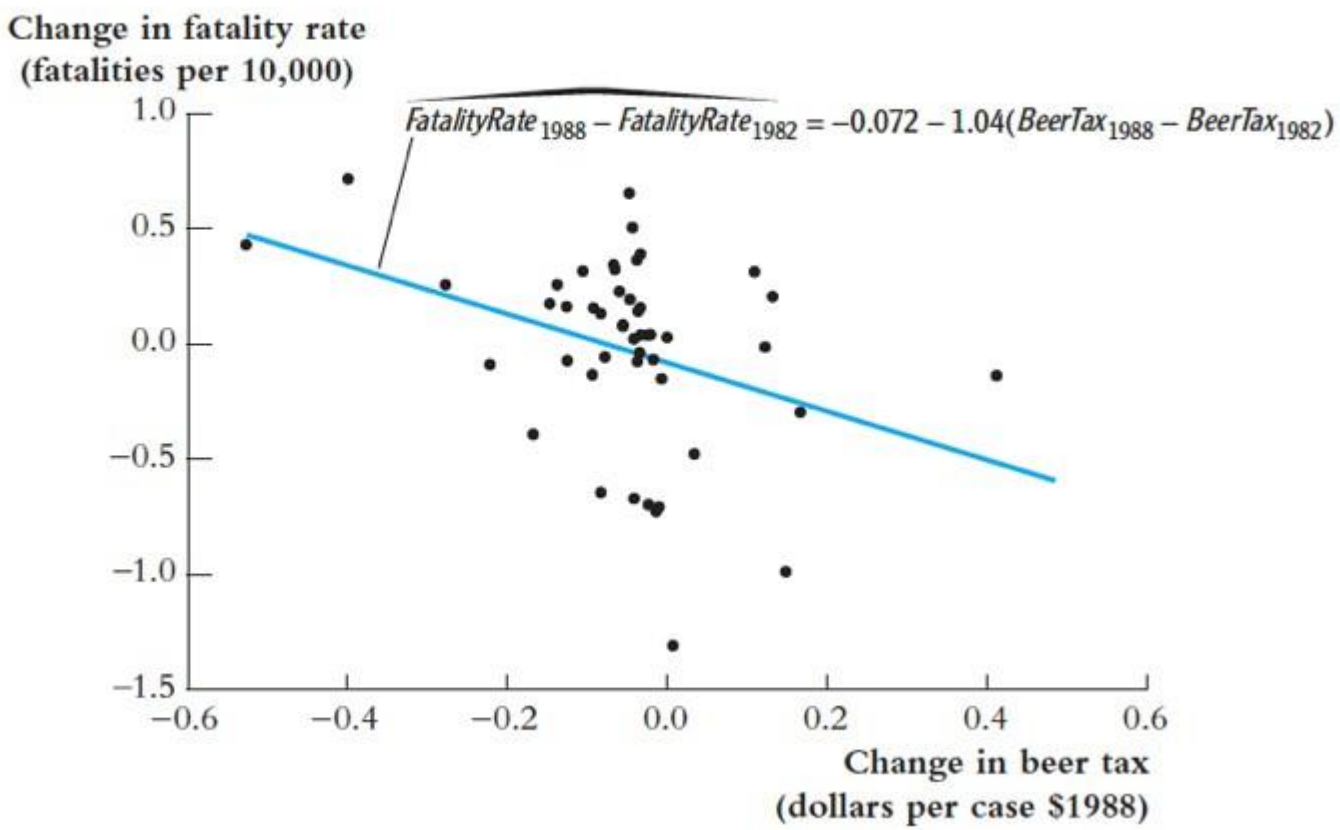
- 1988 data
- $\widehat{fatalityrate} = \underset{(0.51)}{2.01} + \underset{(0.13)}{0.15Alcohol}$ **(n=48)**

- Difference Regression (n=48)
- $\widehat{FR_{1988} - FR_{1982}} = \underset{(0.065)}{-0.72} - \underset{(0.31)}{1.04(Alcohol_{1988} - Alcohol_{1988})}$ **(n=48)**

CHANGE IN TRAFFIC DEATHS AND CHANGE IN ALCOHOL TAXES

FIGURE 10.2 Changes in Fatality Rates and Beer Taxes, 1982–1988

This is a scatterplot of the *change* in the traffic fatality rate and the *change* in real beer taxes between 1982 and 1988 for 48 states. There is a negative relationship between changes in the fatality rate and changes in the beer tax.



WRAP UP

- In contrast to the cross-sectional regression results, the estimated effect of a change in the real Alcohol tax is **negative**, as predicted by economic theory.

- By examining changes in the fatality rate over time, the regression *controls for fixed factors* such as cultural attitudes toward drinking and driving.

- But there are many factors that influence traffic safety, and if they change over time and are correlated with the real alcohol tax, then their omission will produce omitted variable bias.

WRAP UP

- This "before and after" analysis works *when the data are observed in* ***two*** *different years*.

- Our data set, however, contains observations for **seven** different years, and it seems foolish to discard those potentially useful additional data.

- But the "before and after" method does not apply directly when

$T > 2$. To analyze all the observations in our panel data set, we use a more general regression setting: **fixed effects**

FIXED EFFECT

- Fixed effects regression is a method for controlling for omitted variables in panel data when *the omitted variables vary across entities (states) but do not change over time*.

- Unlike the “before and after” comparisons, fixed effects regression can be used when there are **two or more time** observations for each entity.

FIXED EFFECTS REGRESSION MODEL

- The **dependent variable** (Fatality Rate) and **independent variable**
- (Alcohol Tax) denoted as Y_{it} and X_{it}, respectively. Then our model is
  - $Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$ (1.1)

- Where Z_i is an **unobserved variable** that varies from one state to the next but **does not change over time**
  - eg. Z_i can still represent cultural attitudes toward drinking and driving.
- We want to estimate β_1, the effect on Y of X holding constant the unobserved state characteristics Z.

FIXED EFFECTS REGRESSION MODEL

- Because Z_i varies from one state to the next but is constant over time, then let $\boldsymbol{\alpha_i = \beta_0 + \beta_2 Z_i}$,the Equation becomes
  - $Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$ (1.2)

- This is the **fixed effects regression model**, in which α_i are treated as *unknown intercepts* to be estimated, one for each state. The interpretation of α_i as a *state-specific intercept* in Equation (1.2).

- Because the intercept α_i can be thought of as the "effect" of being in entity i (in the current application, entities are states, the terms α_i, known as **entity fixed effects**.

- The variation in the entity fixed effects comes from omitted variables that, like Z_i in Equation (1.1), vary across entities but not over time.

ALTERNATIVE : FIXED EFFECTS BY USING BINARY VARIABLES

- **How to estimate these parameters α_i.**

- To develop the fixed effects regression model using binary variables, let $D1_i$ be a binary variable that equals 1 when i = 1 and equals 0 otherwise, let $D2_i$ equal 1 when i = 2 and equal 0 otherwise, and so on.

- Arbitrarily omit the binary variable $D1_i$ for the first group. Accordingly, the fixed effects regression model in Equation (1.2) can be written equivalently as

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \ldots + \gamma_n Dn_i + u_{it} \quad (1.3)$$

- Thus there are two equivalent ways to write the fixed effects regression model, Equations (1.2) and (1.3).

- In both formulations, the slope coefficient on X is the same from one state to the next.

ESTIMATION AND INFERENCE

- In principle the binary variable specification of the fixed effects regression model can be estimated by OLS.

- But it is tedious to estimate so many fixed effects. If $n = 1000$, then you have to estimate $1000 - 1 = 999$ fixed effects.

- There are some special routines, which are equivalent to using OLS on the full binary variable regression, are *faster* because they employ some *mathematical simplifications* that arise in the algebra of fixed effects regression.

ESTIMATION: THE "ENTITY-DEMEANED"

- Computes the OLS fixed effects estimator in two steps The **first** step:
  - take the average across times t of both sides of Equation (1.2);

$$\bar{Y}_i = \beta_1 \bar{X}_i + \alpha_i + \bar{u}_t \qquad \textbf{(1.4)}$$

- demeaned: let Equation(1.2) minus (1.4)

$$Y_{it} - \bar{Y}_i = \beta_1 X_{it} - \bar{X}_i + (\alpha_i - \alpha_i) + u_{it} - \bar{u}_i$$

ESTIMATION: THE "ENTITY-DEMEANED"

- Let

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$$
$$\tilde{X}_{it} = X_{it} - \bar{X}_i$$
$$\tilde{u}_{it} = u_{it} - \bar{u}_i$$

- Then the **second** step: accordingly, estimate

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \qquad (1.5)$$

- Then the estimator is known as the **within estimator**. Because it matters not if a unit has consistently high or low values of Y and X. All that matters is how the variations around those mean values are correlated.
- In fact, this estimator is identical to the OLS estimator of β_1 without intercept obtained by estimation of the fixed effects model in Equation (1.3)

OLS ESTIMATOR WITHOUT INTERCEPT

- OLS estimator without intercept
  - $Y_i = \beta_1 X_i + u_i$
- The least squared term

$$\min_{b_1} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - b_1 X_i)^2$$

- F.O.C, thus differentiating with respect to β_1, we get
- At last,

$$\sum_{i=1}^{n} 2(Y_i - b_1 X_i) X_i = 0$$

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

FIXED EFFECTS ESTIMATOR(I)

- The second step:

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \qquad \textbf{(1.4)}$$

- Then the fixed effects estimator can be obtained based on OLS estimator without intercept

$$\hat{\beta}_{fe} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{Y}_{it} \tilde{X}_{it}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{X}_{it}^2}$$

FIXED EFFECT ESTIMATOR(II)

- The fixed effects model is

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \quad (1.2)$$

- Equivalence to

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \ldots + \gamma_n Dn_i + u_{it} \quad (1.3)$$

- Then we can think of α_i as fixed effects or "nuisance parameters" to be estimated, thus yields

$$(\hat{\beta}, \hat{\alpha}_1, \ldots, \hat{\alpha}_n) = \underset{b, a_1, \ldots, a_n}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{t=1}^{T} (Y_{it} - bX_{it} - a_i)^2$$

- this amounts to including $n = n + 1 - 1$ dummies in regression of Y_{it} on X_{it}

FIXED EFFECT ESTIMATOR(II)

- The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\hat{\beta}X_{it}-\hat{\alpha}_i)X_{it}=0$$

- And

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\hat{\beta}X_{it}-\hat{\alpha}_i)=0$$

FIXED EFFECT ESTIMATOR(II)

- Therefore, for $i = 1, \ldots, N$,

$$\hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \hat{\beta}X_{it}) = \overline{Y}_i - \overline{X}_i\hat{\beta},$$

- where

$$\bar{X}_i \equiv \frac{1}{T}\sum_{t=1}^{T} X_{it}; \bar{Y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} Y_{it}$$

FIXED EFFECT ESTIMATOR(II)

- Plug this result into the first FOC to obtain:

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\hat{\beta}X_{it}-\hat{\alpha}_i)X_{it}=\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-X_{it}\hat{\beta}-\overline{Y}_i+\overline{X}_i\hat{\beta})X_{it}$$

$$=\left(\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\overline{Y}_i)X_{it}\right)$$

$$-\hat{\beta}\left(\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it}-\bar{X}_i)X_{it}\right)=0$$

FIXED EFFECT ESTIMATOR(II)

- Then we could obtain

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it}-\bar{X}_i)(X_{it}-\bar{X}_i)}{\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\bar{Y})(X_{it}-\bar{X}_i)}$$

$$= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{Y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2}$$

with time-demeaned variables $\tilde{X}_{it} \equiv X_{it} - \bar{X}$, $Y_{it} \equiv Y_{it} - Y_i$

- which is same as we obtained in demeaned method.

THE FIXED EFFECTS REGRESSION ASSUMPTIONS

- The simple fixed effect model

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}, i = 1, \ldots n \; t = 1, \ldots, T$$

- **Assumption 1:** u_{it} has conditional mean zero with $X_i t$, or X_i at any time and α_i

$$E(u_{it} | X_{i1}, X_{i2}, \ldots, X_{iT}, \alpha_i) = 0$$

- **Assumption 2:** $(X_{i1}, X_{i2}, \ldots, X_{iT}, u_{i1}, u_{i2}, \ldots, u_{iT}), i = 1, 2, \ldots, n$ are $i.i.d$

- **Assumption 3:** Large outliers are unlikely. Assumption 4: There is no perfect multicollinearity.

- For multiple regressors, X_{it} should be replaced by the full list
- $X_{1,it}, X_{2,it}, \ldots, X_{k,it}$

STATISTICAL PROPERTIES

- Unbiasedness and Consistency

$$\begin{aligned}\hat{\beta}_{fe} &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{Y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2} \\ &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}(\beta_1\tilde{X}_{it}+\tilde{u}_{it})}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2}\end{aligned}$$

- It is very familiar: paralleling the derivation of OLS estimator, we could prove the estimator of fixed effects model is **unbiased** and **consistent**.

STATISTICAL PROPERTIES

- Similarly, in panel data, if the fixed effects regression assumptions—holds, then the sampling distribution of the fixed effects OLS estimator is normal in large samples,
- Then the variance of that distribution can be estimated from the data, the square root of that estimator is the standard error,
- And the standard error can be used to construct t-statistics and confidence intervals.
- Statistical inference—testing hypotheses (including joint hypotheses using F-statistics) and constructing confidence intervals—proceeds in exactly the same way as in multiple regression with cross-sectional data.

FIXED EFFECTS: EXTENSION TO MULTIPLE X'S.

- The fixed effects regression model is

$$Y_{it} = \beta_1 X_{1,it} + \ldots + \beta_k X_{k,it} + \alpha_i + u_{it}$$

- Equivalently, the fixed effects regression can be expressed in terms of a common intercept
  - $Y_{it} = \beta_0 + \beta_1 X_{1,it} + \ldots + \beta_k X_{k,it} + \gamma_2 D2_i + \gamma_3 D3_i + \ldots + \gamma_n Dn_i + u_{it}$

APPLICATION TO TRAFFIC DEATHS

- The OLS estimate of the fixed effects regression based on all 7 years of data (336 observations), is
  - $\widehat{FatalityRate} = -0.66 \text{AlcohalTax} + StateFixedEffects$ (0.29)

- The estimated state fixed intercepts are not listed to save space and because they are not of primary interest.
- As predicted by economic theory, higher real Alcohol taxes are associated with fewer traffic deaths, which is the opposite of what we found in the initial cross-sectional regressions.

APPLICATION TO TRAFFIC DEATHS

- Recall: The result in Before-After Model is
- $\widehat{FR_{1988} - FR_{1982}} = \underset{(0.065)}{-0.72} - \underset{(0.31)}{1.04(Alcohol_{1988} - Alcohol_{1988})}$ **(n=48)**

- The magnitudes of estimate coefficients are not identical, because they use different data.
- And because of the additional observations, the standard error now is also smaller than before-after model.

Fixed effects regression using xtreg, fe

$$Y_{it} = \alpha_i + \beta X_{it} + u_i + e_{it}$$

Outcome

Predictor(s)

Fixed effects option

Controlling for heteroskedasticity

Total number of cases (rows)

Total number of entities (*i*)

```
. xtreg ln_gdppc ln_trade ln_labor, fe robust

Fixed-effects (within) regression               Number of obs      =       2,772
Group variable: country1                        Number of groups   =         126

R-squared:                                      Obs per group:
     Within  = 0.6267                                         min =          22
     Between = 0.3872                                         avg =        22.0
     Overall = 0.3906                                         max =          22

                                                F(2,125)           =       87.57
corr(u_i, Xb) = 0.1067                          Prob > F           =      0.0000

                                    (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
   ln_gdppc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   ln_trade |   .3603947   .0737076     4.89   0.000     .2145182    .5062712
   ln_labor |    .053167   .1608747     0.33   0.742     -.265224     .371558
      _cons |  -.9384681   1.075791    -0.87   0.385    -3.067592    1.190656
-------------+----------------------------------------------------------------
     sigma_u |  1.1155513
     sigma_e |  .10989953
         rho |  .99038791   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The within entity errors u_i are correlated with the regressors in the fixed effects model.

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time. In this example, all the variables are log-transformed, the interpretation is: when the predictor increases 1% over time, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value).
A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

Intraclass correlation (rho), shows how much of the variance in the output is explained by the difference across entities. In this example is 99%.

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

sigma_u = sd of residuals within groups u_i
sigma_e = sd of residuals (overall error term) e_{it}

REGRESSION WITH TIME FIXED EFFECTS

- Just as fixed effects for each entity can control for variables that are constant over time but differ across entities, so can **time fixed effects** control for variables that *are constant across entities but evolve over time*.
  - Like *safety improvements in new cars* as an **omitted variable** that changes over time but has the same value for all states.
- Now our regression model with **time fixed effects**
  - $Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_3 S_t + u_{it}$

- where S_t is *unobserved* and where the single t subscript emphasizes that safety changes over time but is constant across states. Because $\beta_3 S_3$ represents variables that determine Y_{it}, if S_t is correlated with X_{it}, then omitting S_t from the regression leads to omitted variable bias.

TIME EFFECTS ONLY

- Although S_t is unobserved, its influence can be eliminated because it varies over time but not across states, just as it is possible to eliminate the effect of Z_i, which varies across states but not over time.
- Similarly, the presence of S_t leads to a regression model in which each time period has its own intercept, thus

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$$

- This model has a different intercept, λ_t, for each time period, which are known as **time fixed effects**. The variation in the time fixed effects comes from omitted variables that vary over time but not across entities.

WRAP UP

- We've showed that how panel data can be used to control for unobserved omitted variables that differ across entities but are constant over time.
- The key insight is that if the unobserved variable does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics.
- Double fixed Effects model, thus both entity and time fixed effects can be included in the regression to control for variables that vary across entities but are constant over time and for variables that vary over time but are constant across entities.
- Despite these virtues, entity and time fixed effects regression *cannot* control for *omitted variables* that *vary both across entities and over time*. There remains a need for a new method that can eliminate the influence of unobserved omitted variables.

Entity and time fixed effects regression using xtreg, fe

$$Y_{it} = \alpha_i + \beta X_{it} + \delta_t + u_i + e_{it}$$

Time fixed effects

Fixed effects option

Total number of cases (rows)

Outcome

Predictor(s)

Controlling for heteroskedasticity

Total number of entities (*i*)

```
. xtreg ln_gdppc ln_trade ln_labor i.year, fe robust

Fixed-effects (within) regression               Number of obs      =       2,772
Group variable: country1                        Number of groups   =         126

R-squared:                                      Obs per group:
     Within  = 0.7083                                         min =          22
     Between = 0.7977                                         avg =        22.0
     Overall = 0.7581                                         max =          22

                                                F(23,125)          =       34.28
corr(u_i, Xb) = 0.7525                          Prob > F           =      0.0000

                                    (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ln_trade |   .2401329   .0695213     3.45   0.001     .1025416    .3777242
    ln_labor |  -.2958837    .081081    -3.65   0.000     -.456353   -.1354145
             |
        year |
        2001 |   .0119809   .0042779     2.80   0.006     .0035144    .0204475
         ... |      ...        ...      ...    ...          ...         ...
         ... |      ...        ...      ...    ...          ...         ...
        2021 |   .2878247   .0705454     4.08   0.000     .1482065    .4274428
             |
       _cons |   7.213881   1.961627     3.68   0.000     3.331578    11.09619
-------------+----------------------------------------------------------------
     sigma_u |  1.0561892
     sigma_e |  .09753735
         rho |  .99154389   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The within entity errors u_i are correlated with the regressors in the fixed effects model.

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time. In this example, all the variables are log-transformed, the interpretation is: when the predictor increases 1% over time, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value).
A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

Intraclass correlation (rho), shows how much of the variance in the output is explained by the difference across entities. In this example is 99%.

sigma_u = sd of residuals within groups u_i
sigma_e = sd of residuals (overall error term) e_{it}

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

RANDOM EFFECTS

- The rationale behind random effects model is that, unlike the fixed effects model, the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model:

"…the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not" [Green, 2008, p.183]

RANDOM EFFECTS

- If you have reason to believe that differences across entities have some influence on your dependent variable but are not correlated with the predictors then you should use random effects. An advantage of random effects is that you can include time invariant variables (i.e. gender). In the fixed effects model these variables are absorbed by the intercept.

THE RANDOM EFFECTS IDEA

- Random effects assume that the entity's error term is not correlated with the predictors which allows for time invariant variables to play a role as explanatory variables. In random-effects you need to specify those individual characteristics that may or may not influence the predictor variables.

- The problem with this is that some variables may not be available therefore leading to omitted variable bias in the model. RE allows to generalize the inferences beyond the sample used in the model.

DERIVATION

- Here is regression model for panel data with random effects

$$Y_{it} = \alpha + X_{it}\beta + \mu_i + \epsilon_{it}$$

- Y_{it} is the dependent variable for entity i at time t.
- X_{it} is the independent variable(s) for entity i at time t.
- α is the intercept.
- β is the coefficient vector for the independent variables.
- μ_i represents the unobserved entity-specific effect (the random effect).
- ϵ_{it} is the error term.

- Now assuming that μ_i follows a random effect model
- $\mu_i = \gamma + u_i$

γ is the population mean of the random effects ($E(u_i) = 0$).

u_i is the individual-specific random effect with $E(u_i) = 0$ and $\mathrm{Var}(u_i) = \sigma_u^2$.

DERIVATION

Substituting this into the original regression equation:

$$Y_{it} = \alpha + X_{it}\beta + (\gamma + u_i) + \epsilon_{it}$$

Rearrange terms:

$$Y_{it} = \alpha + \gamma + X_{it}\beta + u_i + \epsilon_{it}$$

Define a new error term that combines the individual-specific random effect and the original error term:

$$\tilde{\epsilon}_{it} = u_i + \epsilon_{it}$$

The equation now becomes:

$$Y_{it} = \alpha + \gamma + X_{it}\beta + \tilde{\epsilon}_{it}$$

In the random effects model, we assume that $\tilde{\epsilon}_{it}$ is independently and identically distributed (i.i.d.) with $E(\tilde{\epsilon}_{it}) = 0$ and $\mathrm{Var}(\tilde{\epsilon}_{it}) = \sigma_u^2 + \sigma_\epsilon^2$, where σ_u^2 is the variance of the individual-specific random effects, and σ_ϵ^2 is the variance of the original error term.

Random effects regression using xtreg, re

$$Y_{it} = \alpha_i + \beta X_{it} + \gamma Z_i + e_{it}$$

Random effects option

Outcome

Predictor(s)

Controlling for heteroskedasticity

Total number of cases (rows)

Total number of entities (*i*)

```
. xtreg ln_gdppc ln_trade ln_labor, re robust

Random-effects GLS regression                   Number of obs     =      2,772
Group variable: country1                        Number of groups  =        126

R-squared:                                      Obs per group:
     Within  = 0.6110                                         min =         22
     Between = 0.7295                                         avg =       22.0
     Overall = 0.7212                                         max =         22

                                                Wald chi2(2)      =     192.71
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

                                    (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ln_trade |   .4175909   .0760404     5.49   0.000     .2685543    .5666274
    ln_labor |  -.1597685   .1312262    -1.22   0.223    -.4169671    .0974302
       _cons |   .9295612   .6361615     1.46   0.144    -.3172923    2.176415
-------------+----------------------------------------------------------------
     sigma_u |  .41594682
     sigma_e |  .10989953
         rho |  .93474564   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The between entity errors u_{it} are uncorrelated with the regressors in the random effects model.

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time and across entities (average effect). In this example, all the variables are log-transformed, the interpretation is: when the predictor increases, on average, 1%, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value).
A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

Intraclass correlation (rho), shows how much of the variance in the output is explained by the difference across entities. In this example is 99%.

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

sigma_u = sd of residuals within groups u_i
sigma_e = sd of residuals (overall error term) e_{it}

THANK YOU