# Lecture 11: Panel Data and Difference in Differences

*Introduction ot Econometrics,Fall 2017*

**Zhaopeng Qu**

**Nanjing University**

*12/18/2017*

1. Panel Data: What and Why
2. Fixed Effect Model
3. Extension: Regression with Time Fixed Effects
4. The Fixed Effects Regression Assumptions and Standard Errors

# Panel Data: What and Why

# Introduction

- A panel dataset contains observations on multiple entities, where each entity is observed at two or more points in time.

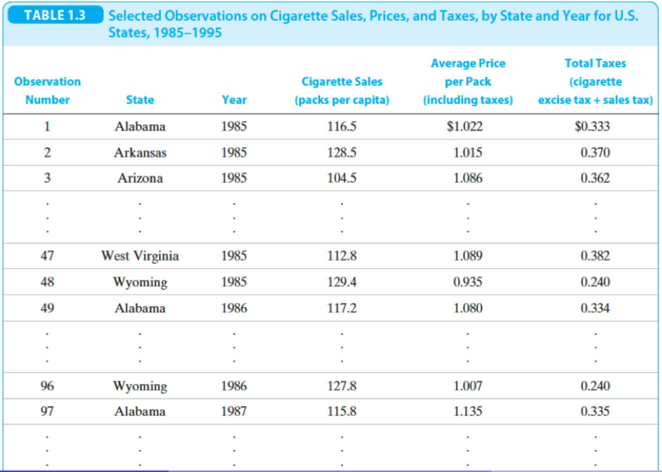


**TABLE 1.3** Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995

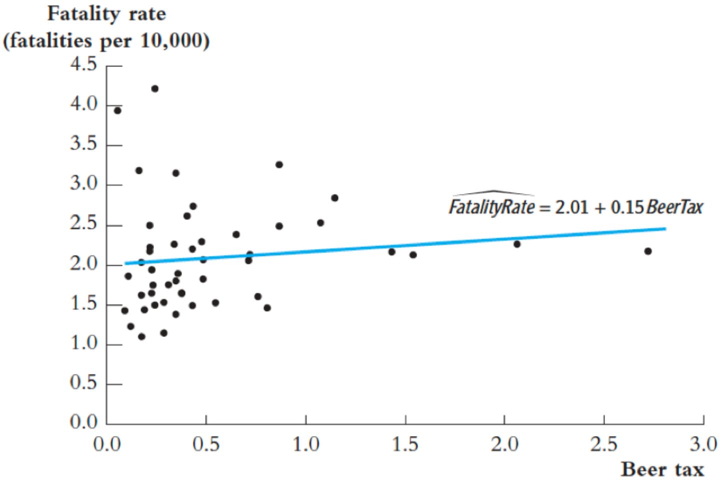
| Observation Number | State | Year | Cigarette Sales (packs per capita) | Average Price per Pack (including taxes) | Total Taxes (cigarette excise tax + sales tax) |
|---|---|---|---|---|---|
| 1 | Alabama | 1985 | 116.5 | $1.022 | $0.333 |
| 2 | Arkansas | 1985 | 128.5 | 1.015 | 0.370 |
| 3 | Arizona | 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

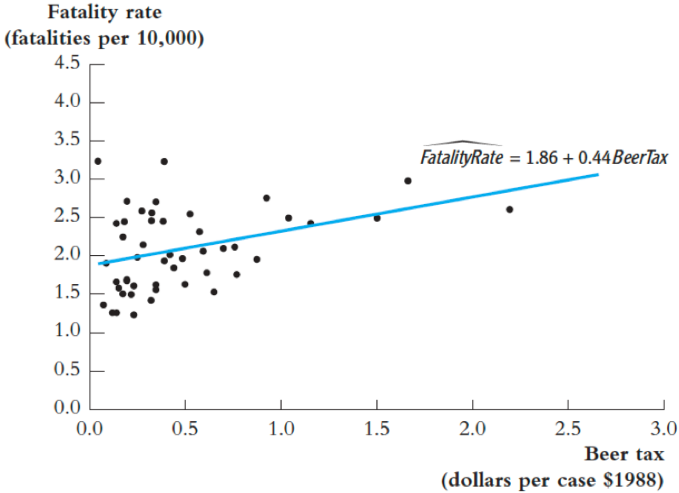
- Observational unit: a year in a U.S. state
- 48 U.S. states, so n = of entities = 48
- 7 years (1982,..., 1988),so T = # of time periods = 7
- Balanced panel, so total # observations $= 7»48 = 336$
- Variables:
- Dependent Variable: Traffic fatality rate (# traffic deaths in that state in that year, per 10,000 state residents)
- Independent Variable: Tax on a case of beer
- Other Controls (legal driving age, drunk driving laws, etc.)

# U.S. traffic death data for 1982

- Higher alcohol taxes, more traffic deaths



# U.S. traffic death data for 1988



**(b)** 1988 data

Figure 2:

# Simple Case: Panel Data with Two Time Periods

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}$$

- $Z_i$ is a factor that does not change over time; its omission might cause omitted variable bias; 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.

$$E[Y] = 1 \times Pr(Y = 1) + 0 \times Pr(Y = 0) = Pr(Y = 1)$$

## Panel Data with Two Time Periods

- The math: Consider the regressions for 1982 and 1988...

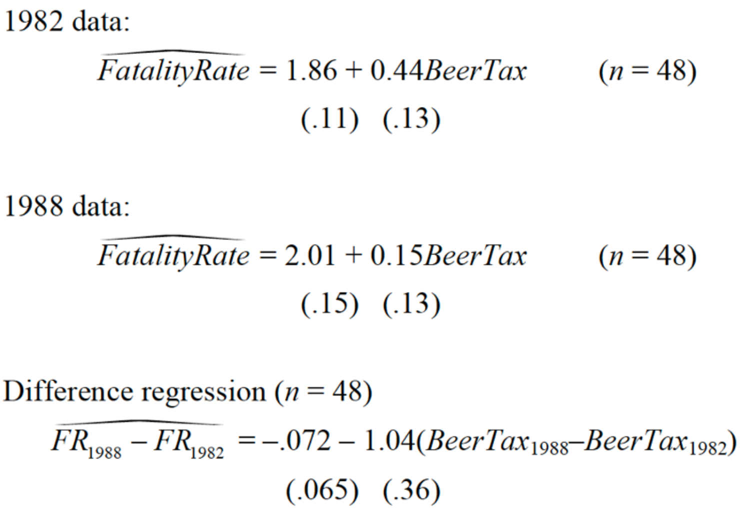
$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}$$
$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}$$

- So make a difference

$$FatalityRate_{i1988} - FatalityRate_{i1982} =$$
$$\beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i1982})$$

- Assumption: if $E(u_it)|BeerTax_{it}, Z_{it}) = 0$,then $(u_{i1988} - u_{i1982})$ is uncorrelated with either $BeerTax_{i1988}$ or $BeerTax_{i1982}$
- 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

# Traffic deaths and beer taxes



1982 data:

$$\widehat{FatalityRate} = \underset{(.11)}{1.86} + \underset{(.13)}{0.44}BeerTax \qquad (n = 48)$$

1988 data:

$$\widehat{FatalityRate} = \underset{(.15)}{2.01} + \underset{(.13)}{0.15}BeerTax \qquad (n = 48)$$

Difference regression $(n = 48)$

$$\widehat{FR_{1988} - FR_{1982}} = \underset{(.065)}{-.072} - \underset{(.36)}{1.04}(BeerTax_{1988} - BeerTax_{1982})$$

Figure 4:

# Traffic deaths and beer 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.

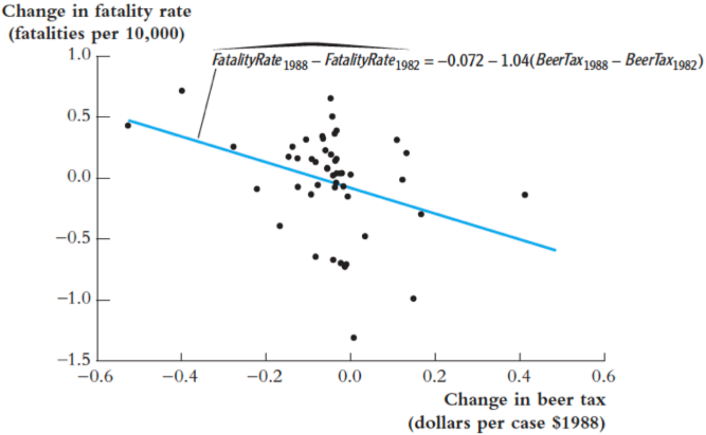


Figure 5:

# Wrap up

- In contrast to the cross-sectional regression results, the estimated effect of a change in the real beer tax is **negative**, as predicted by economic theory. The hypothesis that the population slope coefficient is zero is rejected at the 5% significance level.
- By examining changes in the fatality rate over time, the regression in Equation (10.8) 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 beer tax, then their omission will produce omitted variable bias.
- 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 the method of **fixed effects** regression

# Fixed Effect Model

# Introduction

- 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 of Section 10.2, fixed effects regression can be used when there are two or more time observations for each entity.

# Fixed Effects Regression Model

- the dependent variable (FatalityRate) and observed regressor (BeerTax) denoted as $Y_{it}$ and $X_{it}$, respectively:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it} \tag{10.9}$$

- where $Z_i$ is an unobserved variable that varies from one state to the next but does not change over time (for example, $Z_i$ represents cultural attitudes toward drinking and driving). We want to estimate $\beta_1$, 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 $\alpha_i = \beta_0 + \beta_1 Z_i$,then Equation becomes

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{10.10}$$

- Equation (10.10) is the fixed effects regression model, in which $\alpha$ are treated as unknown intercepts to be estimated, one for each state. The interpretation of $\alpha_i$ as a state-specific intercept in Equation (10.10).
- Because the intercept $\alpha_i$ in Equation (10.10) can be thought of as the "effect" of being in entity $i$ (in the current application, entities are states), the terms $\alpha_i$,an are known as **entity fixed effects**.
- The variation in the entity fixed effects comes from omitted variables that, like $Z_i$ in Equation (10.9), vary across entities but not over time.

# Fixed Effects by using binary variables

- 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 (10.10) can be written equivalently as

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + ... + \gamma_n Dn_i + u_{it} \qquad (10.11)$$

- Thus there are two equivalent ways to write the fixed effects regression model, Equations (10.10) and (10.11).
- In Equation (10.10), it is written in terms of $n$ state specific intercepts.
- In Equation (10.11), the fixed effects regression model has a common intercept and $n-1$ binary regressors
- In both formulations, the slope coefficient on $X$ is the same from one

# Fixed Effects: Extension to multiple X's.

- The fixed effects regression model is

$$Y_{it} = \beta_1 X_{1,it} + ... + \beta_k X_{k,it} + \alpha_i + u_{it} \tag{10.12}$$

- Equivalently, the fixed effects regression can be expressed in terms of a common intercept

$$\begin{aligned} Y_{it} = &\beta_0 + \beta_1 X_{1,it} + ... + \beta_k X_{k,it} \\ &+ \gamma_2 D2_i + \gamma_3 D3_i + ... + \gamma_n Dn_i + u_{it} \end{aligned}$$

# Estimation and Inference

- In principle the binary variable specification of the fixed effects regression model [Equation (10.13)] 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.
- These special routines are equivalent to using OLS on the full binary variable regression, but are faster because they employ some mathematical simplifications that arise in the algebra of fixed effects regression.

# Estimation: The “entity-demeaned”

- take the average across times $t$ of both sides of Equation (10.10);

$$\bar{Y}_i = \beta_1 \bar{X}_i + \alpha_i + \bar{u}_t$$

- demeaned: let

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$$

- Accordingly,

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \tag{10.14}$$

- In fact, this estimator is identical to the OLS estimator of $\beta_1$ obtained by estimation of the fixed effects model in Equation (10.11)

# Extension: Regression with Time Fixed Effects

# Introduction

- 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.
- safety improvements in new cars as an omitted variable that changes over time but has the same value for all states.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it} \tag{10.16}$$

- 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

- suppose that the variables $Z_i$ are not present

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it} \tag{10.17}$$

- so the terms $\lambda$ are known as time fixed effects.The variation in the time fixed effects comes from omitted variables that, like $S_t$ in Equation (10.16), vary over time but not across entities.
- Just as the entity fixed effects regression model can be represented using $n-1$ binary indicators, so, too, can the time fixed effects regression model be represented using $T-1$ binary indicators:

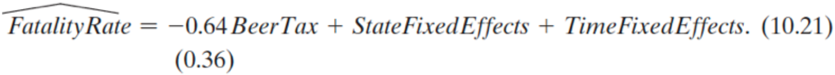
$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \delta_2 B2_t + ... + \delta_T BT_t + \alpha_i + u_{it} \tag{10.18}$$

# Both Entity and Time Fixed Effects

- If some omitted variables are constant over time but vary across states (such as cultural norms) while others are constant across states but vary over time (such as national safety standards),
- The combined entity and time fixed effects regression model is

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

# Application to traffic deaths

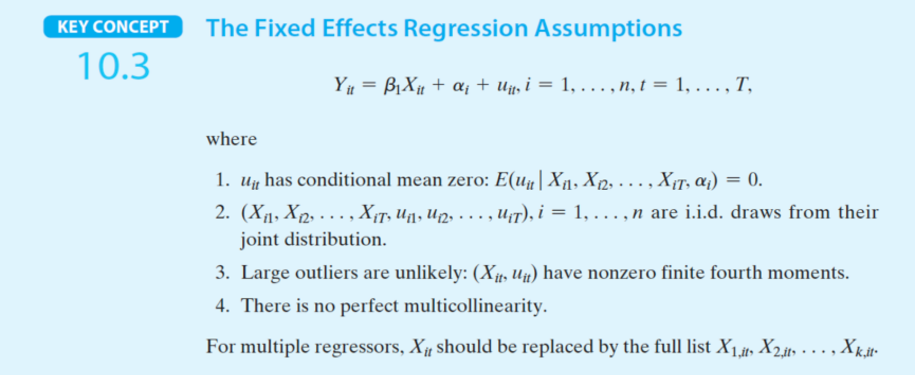


$$\underset{}{\widehat{FatalityRate}} = \underset{(0.36)}{-0.64}\,BeerTax + StateFixedEffects + TimeFixedEffects. \quad (10.21)$$

Figure 6:

# The Fixed Effects Regression Assumptions and Standard Errors

# The Fixed Effects Regression Assumptions



KEY CONCEPT 10.3

## The Fixed Effects Regression Assumptions

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

where

1. $u_{it}$ has conditional mean zero: $E(u_{it} \mid X_{i1}, X_{i2}, \dots, X_{iT}, \alpha_i) = 0$.
2. $(X_{i1}, X_{i2}, \dots, X_{iT}, u_{i1}, u_{i2}, \dots, u_{iT}), i = 1, \dots, n$ are i.i.d. draws from their joint distribution.
3. Large outliers are unlikely: $(X_{it}, u_{it})$ have nonzero finite fourth moments.
4. There is no perfect multicollinearity.

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

Figure 7:

# Autocorrelated in Panel Data

- An important difference between the panel data assumptions in Key Concept 10.3 and the assumptions for cross-sectional data in Key Concept 6.4 is Assumption 2.
    - *Cross-Section*: Assumption 2 holds: i.i.d sample.
    - *Panel data*: independent across entities but no such restricition **within** an entity.
- if $Cov(X_t, X_s)$ for some $t \neq s$, the $X_t$ is said to be **autocorrelated or serially correlated**.
- In the traffic fatality example, $X_{it}$, the beer tax in state i in year t, is autocorrelated:
    - Most of the time, the legislature does not change the beer tax, so if it is high one year relative to its mean value for state i,it will tend to be high the next year,too.

# Autocorrelated in Panel Data

- Similarly,$u_{it}$ would be also autocorrelated. It consists of time-varying factors that are determinants of $Y_{it}$ but are not included as regressors, and some of these omitted factors might be autocorrelated. It can formally be expressed as

$$Cov(u_{it}, u_{is}|X_{it}, X_{is}, \alpha_i) \neq 0 \; for \; t \neq s$$

  - eg. a downturn in the local economy and a road improvement project.
  - eg. severe winter driving conditions.
- The result: an analogy of heteroskedasticity.
- OLS panel data estimators of $\beta$ are unbiased and consistent but the standard errors will be wrong
  – usually the OLS standard errors understate the true uncertainty
- This problem can be solved by using **“heteroskedasticity and autocorrelation-consistent(HAC) standard errors”**

# Standard Errors for Fixed Effects Regression

- Standard errors that are valid if $u_{it}$ is potentially heteroskedastic and potentially correlated over time within an entity are referred to as heteroskedasticity and autocorrelation-consistent (HAC) standard errors.
- The standard errors used are one type of HAC standard errors, **clustered standard errors**.
- The term clustered arises because these standard errors allow the regression errors to have an arbitrary correlation within a cluster, or grouping, but assume that the regression errors are uncorrelated across clusters. In the context of panel data, each cluster consists of an entity.