Econometrics 1

Lecture 10: Regression with Panel Data

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼2613

E-mail: huangjp@szu.edu.cn

Tel: (0755) 2695 0548

Website: https://huangjp.com

Panel data

Panel data

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Notation for panel data

- Panel data consist of observations on the same n entities at two or more time periods T. If the data set contains observations on the variables X and Y, then the data are denoted

$$(X_{it}, Y_{it}), \quad i = 1, \ldots, n \text{ and } t = 1, \ldots, T.$$

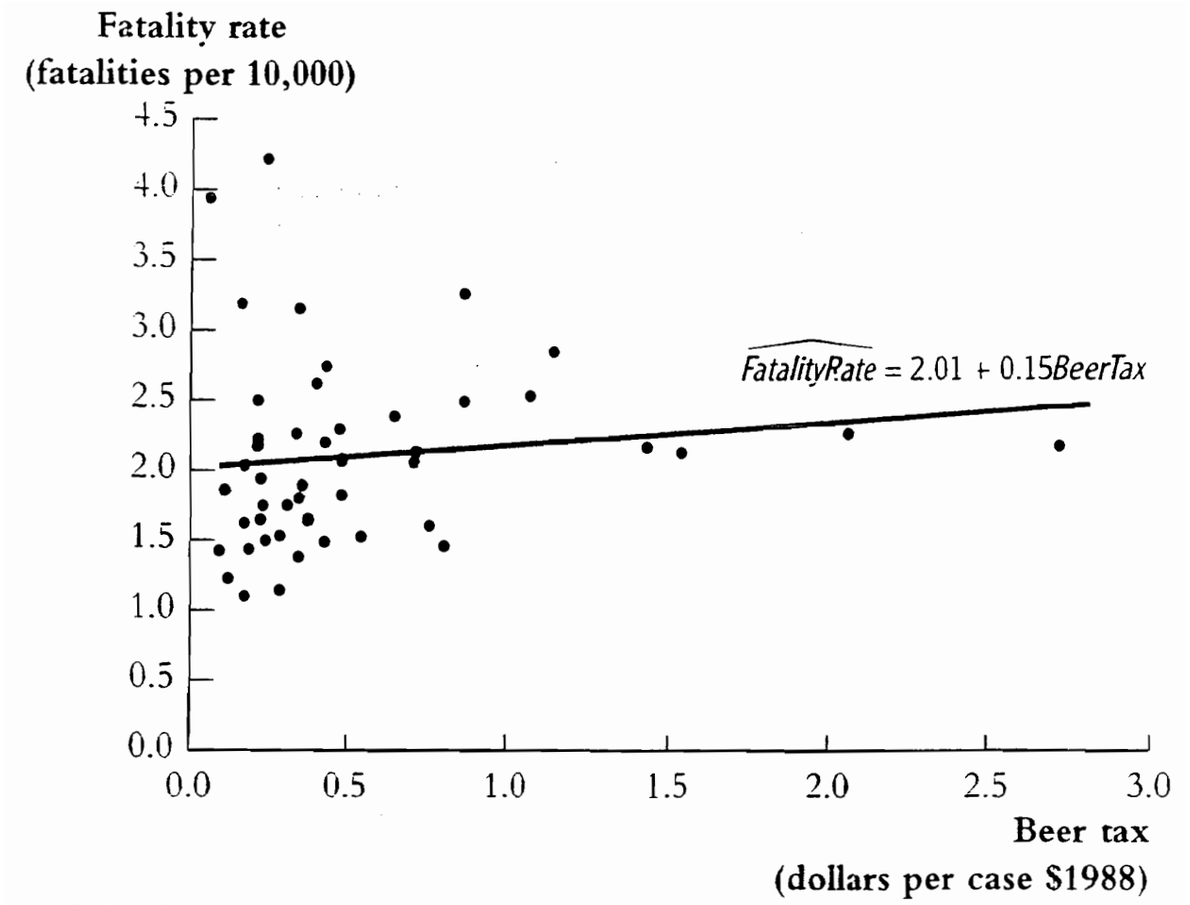
  where the first subscript, i, refers to the entity being observed and the second subscript, t, refers to the date at which it is observed.

- Balanced panel — the variables are observed for each entity and each time period.

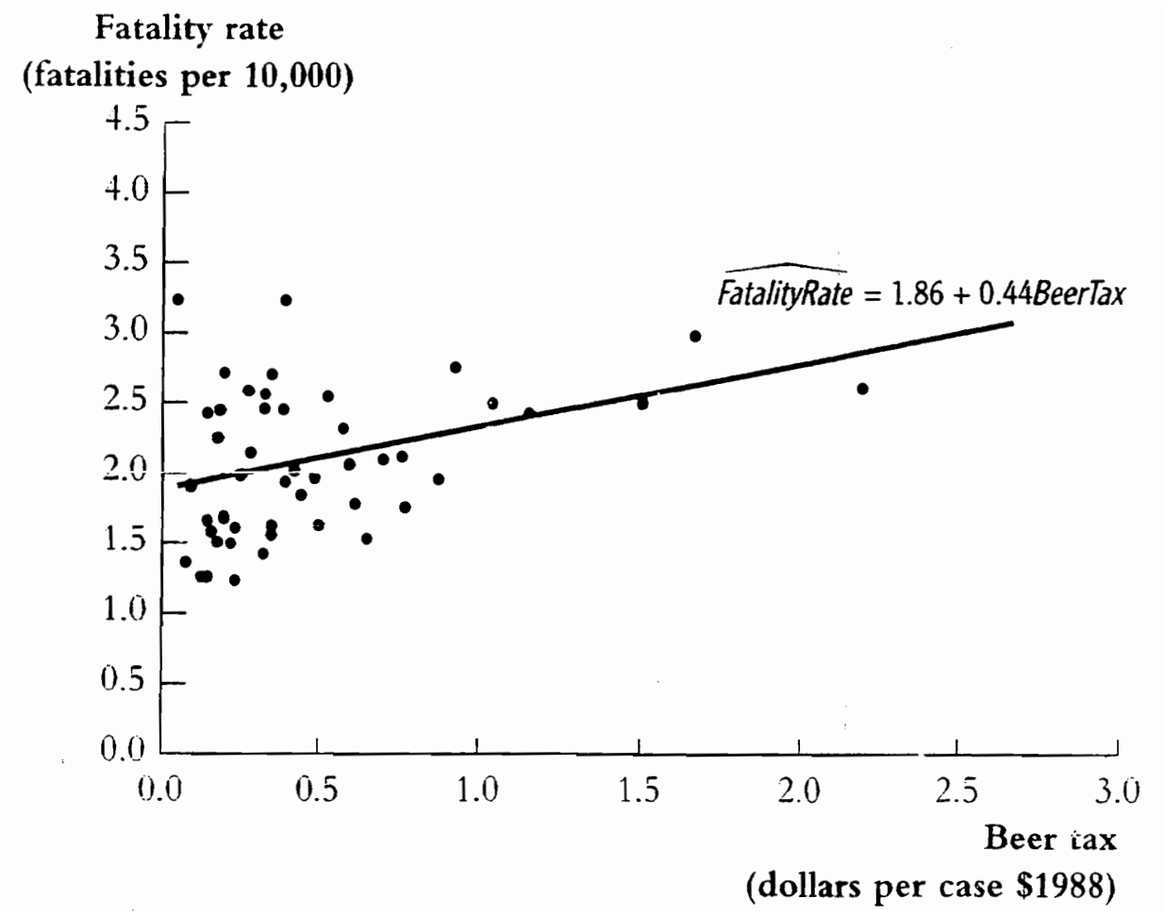
The U.S. state traffic fatality data set

- `fatality.xlsx` and `fatality.docx`
- Entities: 48 U.S. states (excluding Alaska and Hawaii).
- Time periods: 7 (1982–1988).
- The data is balanced.
- Variables:
  traffic fatality rate, beer tax, punishment for drunk driving, personal income, unemployment rate, population, etc.

Traffic fatality rate and tax on beer



(a) 1982 data



(b) 1988 data

$$\widehat{FatalityRate} = 2.01 + 0.15BeerTax \quad \text{(1982 data)}.$$
$$(0.15) \quad (0.13)$$

$$\widehat{FatalityRate} = 1.86 + 0.44BeerTax \quad \text{(1988 data)}.$$
$$(0.11) \quad (0.13)$$

The positive coefficients may due to omitted variable bias.

OLS regression with fixed effects.

- Many factors can affect the fatality rate:
  quality of the automobiles, whether the highways are in good repair, whether most driving is rural or urban, the density of cars on road, whether it is socially acceptable to drink and drive.

- Some of these variable might be very hard or even impossible to measure.

- If these factors remain constant over time, with panel data, we can in effect **hold them constant** even though we cannot measure them.

Panel data regression

“Before and after” comparisons

- Data are obtained for $T = 2$.
- Let Z_i be a variable that determines the fatality rate in the ith state, but does not change over time.
- The regression model is

$$\text{FatalityRate}_{it} = \beta_0 + \beta_1 \text{BeerTax}_{it} + \beta_2 Z_i + u_{it}$$

“Before and after” comparisons

- Consider the model for the two years 1982 and 1988:

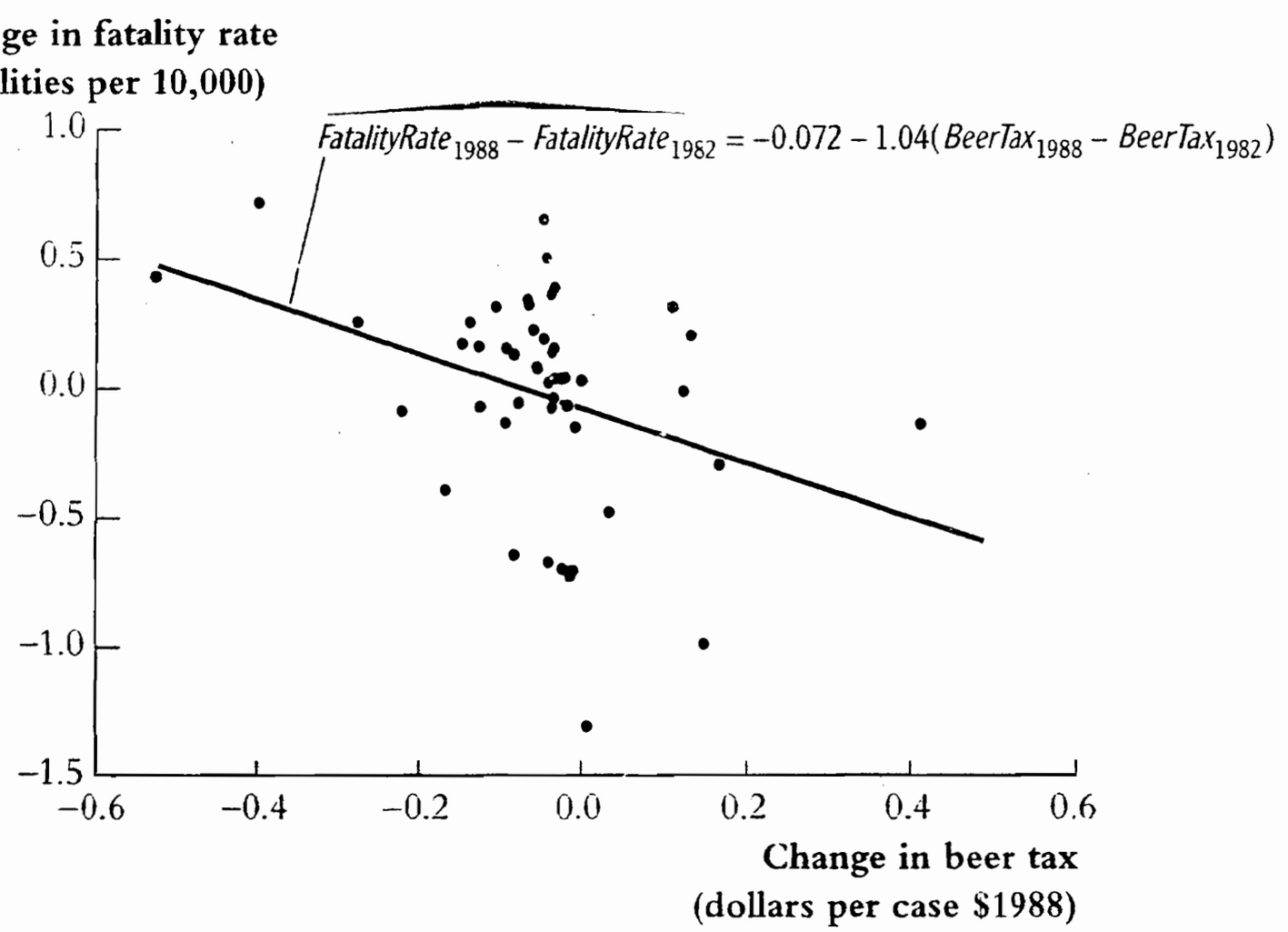
$$\text{FatalityRate}_{i1982} = \beta_0 + \beta_1 \text{BeerTax}_{i1982} + \beta_2 Z_i + u_{i1982}$$
$$\text{FatalityRate}_{i1988} = \beta_0 + \beta_1 \text{BeerTax}_{i1988} + \beta_2 Z_i + u_{i1988}$$

which implies

$$\begin{aligned}\text{FatalityRate}_{i1988} - \text{FatalityRate}_{i1982} \\ = \beta_1(\text{BeerTax}_{i1988} - \text{BeerTax}_{i1982}) + u_{i1988} - u_{i1982}\end{aligned}$$

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.



$$\overline{FatalityRate_{1988} - FatalityRate_{1982}} = -0.072 - 1.04(BeerTax_{1988} - BeerTax_{1982}).$$
$$(0.065)\ (0.36)$$

The coefficient of BeerTax is negative

Try to reproduce equation (10.8) in gretl

- Define fatality rate

```
series fatality = allmort / pop * 10000
```

- Useful commands

smpl — resets the sample range
store — save data to a file (by default a .gdt file)
open — opens a data file
append — opens a data file and appends to the current dataset

Fixed effects regression

- Consider the regression model with the dependent variable (FatalityRate) and observed regressor (BeerTax) denoted as Y_{it} and X_{it}

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

  Z_i is unobserved. We want to estimate β_1.

- Let $\alpha_i = \beta_0 + \beta_2 Z_i$, the above equation becomes

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

  This is the **fixed effects regression model**.
  $\alpha_1, \ldots, \alpha_n$ are known as *entity fixed effects*.

Binary variable specification

- The fixed effect regression model $Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$ can be written equivalently as

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it}$$

- The entity fixed effects are

$$\alpha_1 = \beta_0,$$
$$\alpha_i = \beta_0 + \gamma_i \quad \text{for } i \geq 2$$

General fixed effects regression model

The fixed effects regression model is

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

where $i = 1, \ldots, n; t = 1, \ldots, T$; $X_{k,it}$ is the the value of the kth regressor in time period t; and $\alpha_1, \ldots, \alpha_n$ are entity-specific intercepts.

Equivalently, the fixed effects regression model can be written in terms of a common intercept, the X's, and n - 1 binary variables representing all but one entity:

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

where $D2_i = 1$ if $i = 2$ and $D2_i = 0$ otherwise, and so forth.

Estimate and inference

- The binary variable specification of the fixed effects regression model

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

  has $k + n$ regressors.

- For the U.S. state traffic fatality data, $n = 48$.

The entity-demeaned OLS algorithm

- Step 1: subtract the entity-specific average from each variable.

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

$$\overline{Y}_i = \beta_1 \overline{X}_i + \alpha_i + \overline{u}_i \quad \text{where } \overline{B}_i = \sum_{t=1}^{T} B_{it}/T,\ B \in \{Y, X, u\}$$

$$\widetilde{Y}_{it} = Y_{it} - \overline{Y}_i,\ \widetilde{X}_{it} = X_{it} - \overline{X}_i,\ \widetilde{u}_{it} = u_{it} - \overline{u}_i,$$

- Step 2: run the regression using the entity-demeaned variables.

$$\widetilde{Y}_{it} = \beta_1 \widetilde{X}_{it} + \widetilde{u}_{it}$$

Then, β_1 can be estimated by the OLS regression.

Practice using `fatality.xlsx`

- Set data as panel

```
setobs state year --panel-vars
```

entity var. time var.

- The dummy variable approach

```
genr unitdum  # create 48 unit(entity) dummies
ols fatality beertax du_* --robust
    # const is not needed
    # du_* indicates all unit dummies
```

- The entity-demeaned approach

```
panel fatality const beertax --robust  # const is needed
```

Reporting regression results

- The regression results for all years can be reported in the form

$$\widehat{\text{FatalityRate}} = \underset{(0.29)}{-0.66}\,\text{BeerTax} + \text{StateFixedEffects}$$

- The estimated state fixed intercepts are not of primary interest.
- The negative coefficient coincides with economic theory.
- No intercept.

Time fixed effects

- Time fixed effects can control for variables that are constant across entity but evolve over time.
  E.g, mobile safety.

- Let S_t denote the unobserved effect that changes over time but constant across states. The regression model becomes

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

Time effects only model

- The time effects regression model with a single regressor is

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

where $\lambda_1, \ldots, \lambda_T$ are known as **time fixed effects**.

- Binary variable specification

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_t + \cdots + \delta_T BT_t + u_{it}$$

where $\delta_t, \ldots, \delta_T$ are unknown coefficients and where $B2_t = 1$ if $t = 2$ and $B2_t = 0$ otherwise, and so forth.

Both entity and time fixed effects

- Entity and time fixed effects regression model

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

- Binary variable specification

$$\begin{aligned} Y_{it} =& \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \cdots + \gamma_n Dn_i \\ &+ \delta_2 B2_t + \cdots + \delta_T BT_t + u_{it} \end{aligned}$$

Estimation

- GretI uses entity demeaned OLS with time dummies

```
panel fatality const beertax --time-dummies --robust
```

- Reporting results

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

- How many variables on the right-hand side?

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

Fixed effects regression assumptions

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

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

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

Heteroskedasticity- and autocorrelation-consistent (HAC) standard errors

- Assumption 2 holds that variables are independent across entity, but makes no restriction within an entity.

- If a variable (or error term) is correlated over time for a given entity, it is said to be **autocorrelated** or **serially correlated**.

- The estimators are not biased, but the standard errors must be calculated using **HAC standard errors**.

- In the textbook the **clustered standard errors** are used, which is the default setting of `panel` command with `--robust` option (the Arellano HAC estimator).

Practice

- Read Section 10.6 and reproduce Table 10.1.

- Minimum legal drinking age is given in `mlda`. You need to take the integer part of the data, e.g., 18.5 should be changed to 18.

- The “Mandatory jail or community service?” variable should be defined from `jaild` and `comserd`.

- Average vehicle miles per driver is given in `vmiles`, but should be divided by 1000.

- Unemployment rate is given in `unrate`, not `unus`.

Further readings on panel data econometrics

1. Arellano, M. (2003) *Panel Data Econometrics*, Oxford University Press.

2. Baltagi, B. H. (1995) *Econometric Analysis of Panel Data*, Wiley.

3. Cameron, A. & Trivedi, P. (2005) *Microeconometrics: Methods and Applications*, Cambridge University Press.

4. Wooldridge, J. M. (2002) *Econometric Analysis of Cross Section and Panel Data*, MIT Press.