# POLS 6481. Research Design and Quantitative Methods II

## Lecture 24. Time-Series Cross-Section Data and Methods

Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

Outline

1. Three Types of Data (other than Cross-Sections and Time-Series)
    - A. Pooled Independent Cross-Sections (a.k.a. Rolling Cross-Sections)
    - B. Panel Data
    - C. Longitudinal Data
2. Panel Data Example: How Unemployment Affects the Crime Rate
3. Explanation: How and *why* first differencing works
4. Panel Data Example: The Tradeoff between Work and Sleep
5. Least-Squares-Dummy-Variable Model
6. Fixed Effects
7. Random Effects
8. A Hausman Test
9. Longitudinal Data Example: Divorce Rates

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 1. Three Types of Data

### A. Pooled Independent Cross-Sections

- Cross-sectional samples are drawn randomly (independently) from a large population at different points in time
- Many public opinion polls, like the Gallup poll, fall into this category
- Sometimes referred to as “rolling cross sections”

Example 13.3 (p. 454-457) uses houses sold in North Andover, MA in 1978 and 1981

Example 13.1 (p. 449-451) uses the General Social Survey, biennial from 1972 to 1984

Exercise C15 (p. 481) uses the GSS questions on happiness, biennial from 1994 to 2006

## POLS 6481. Research Design and Quantitative Methods II
## Lecture 24. Time-Series Cross-Section Data and Methods

Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

### 1. Three Types of Data

### B. Panel Data

- Cross-sectional sample is drawn randomly once, then revisited after some treatment
- Typically small *T*, large *N*
- “balanced” panel has every case at each time period
  “unbalanced” panel has missing cases at some time periods

Example 13.5 (p. 463-464) uses survey on sleep in 1975 (n > 700) and 1981 (n = 239)

Example 13.7 (p. 467-468) uses 50 states + DC to examine traffic deaths in 1985 and 1990; numbers of open container laws and laws allowing pre-trial suspension of licenses increases

Exercise C7 (p. 478) uses 366 student-athletes’ grades in Fall and Spring semesters

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 1. Three Types of Data

### C. Longitudinal Data

- Cross-sectional sample is drawn randomly once, then revisited at regular intervals
- Typically large *T*, small *N*
- Autoregression and serial autocorrelation can be issues
- Sometimes these data are populations, not exactly samples (but… use your imagination)

Example 13.9 (p. 471-472) and Exercise 14.C2 (p. 504) use 90 North Carolina counties' crime rates from 1981 to 1987

Exercises 13.C11 (p. 479) and 14.C8 (p. 505) use Michigan school districts' test scores from 1992 to 1998

## 2. Panel Data Example: How Unemployment Affects the Crime Rate

Data: CRIME2.dta
Script: Lecture 24 crime.R

Sample is 46 cities in 1982 ($t = 1$) and in 1987 ($t = 2$)

response variable: crime rate ($crmrte_t$); explanatory variable: unemployment rate ($unem_t$)

Wooldridge Example 9.4 regressed $\log(crmrte_t)$ on:

- current unemployment, i.e., $unem_t$
- the natural log of law enforcement expenditures, i.e., $\log(lawexpc_t)$
- the lagged dependent variable, i.e. $\log(crmrte_{t-1})$

Only the lagged dependent variable was statistically significant and roughly = 1.

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

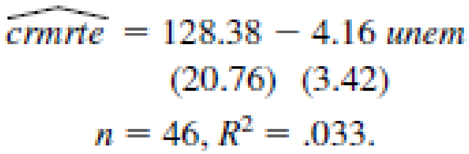
## 2. Panel Data Example: How Unemployment Affects the Crime Rate

Data: CRIME2.dta
Script: Lecture 24 crime.R

Sample is 46 cities in 1982 ($t$ = 1) and in 1987 ($t$ = 2)

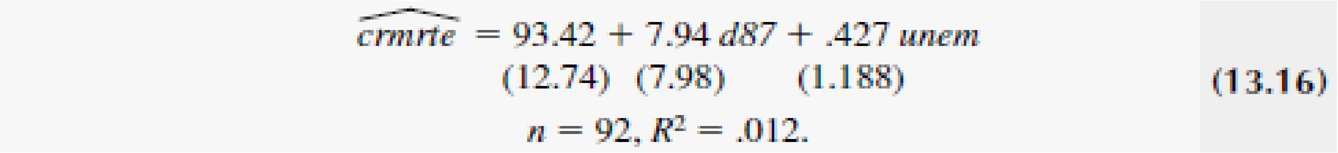
response variable: crime rate ($crmrte_t$); explanatory variable: unemployment rate ($unem_t$)

First, use 1987 cross-section only and regress *crmrte* on *unem* :



$$\begin{aligned}\widehat{crmrte} &= \underset{(20.76)}{128.38} - \underset{(3.42)}{4.16}\, unem \\ n &= 46,\ R^2 = .033.\end{aligned}$$

Second, pool 1982 & 1987 cross-sections, and regress *crmrte* on *unem* and dummy for 1987:



$$\begin{aligned}\widehat{crmrte} &= \underset{(12.74)}{93.42} + \underset{(7.98)}{7.94}\, d87 + \underset{(1.188)}{.427}\, unem \\ n &= 92,\ R^2 = .012.\end{aligned} \qquad (13.16)$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

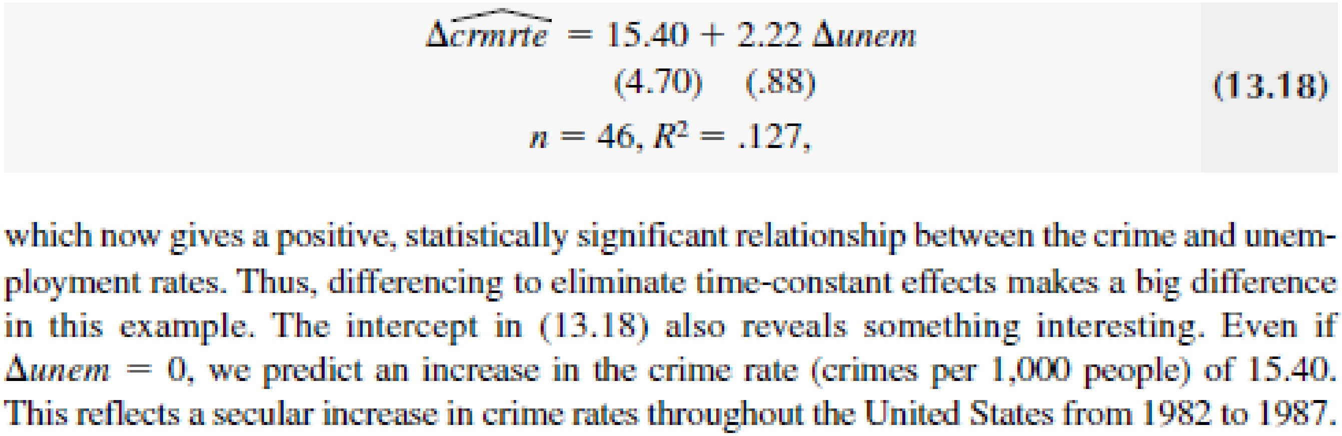
## 2. Panel Data Example: How Unemployment Affects the Crime Rate

Data: CRIME2.dta
Script: Lecture 24 crime.R

Sample is 46 cities in 1982 ($t = 1$) and in 1987 ($t = 2$)

response variable: crime rate ($crmrte_t$); explanatory variable: unemployment rate ($unem_t$)

Third, define changes from 1982 to 1987, and regress $\Delta crmrte$ on $\Delta\ unem$ :



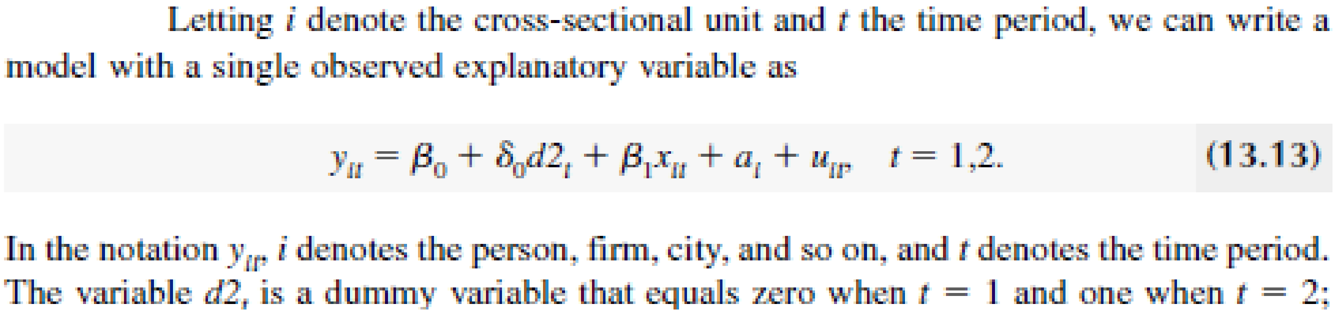
$$\widehat{\Delta crmrte} = \underset{(4.70)}{15.40} + \underset{(.88)}{2.22}\ \Delta unem \qquad (13.18)$$

$$n = 46,\ R^2 = .127,$$

which now gives a positive, statistically significant relationship between the crime and unemployment rates. Thus, differencing to eliminate time-constant effects makes a big difference in this example. The intercept in (13.18) also reveals something interesting. Even if $\Delta unem = 0$, we predict an increase in the crime rate (crimes per 1,000 people) of 15.40. This reflects a secular increase in crime rates throughout the United States from 1982 to 1987.

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 3. Explanation: How and *why* first differencing works



Letting $i$ denote the cross-sectional unit and $t$ the time period, we can write a model with a single observed explanatory variable as

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}, \quad t = 1,2. \tag{13.13}$$

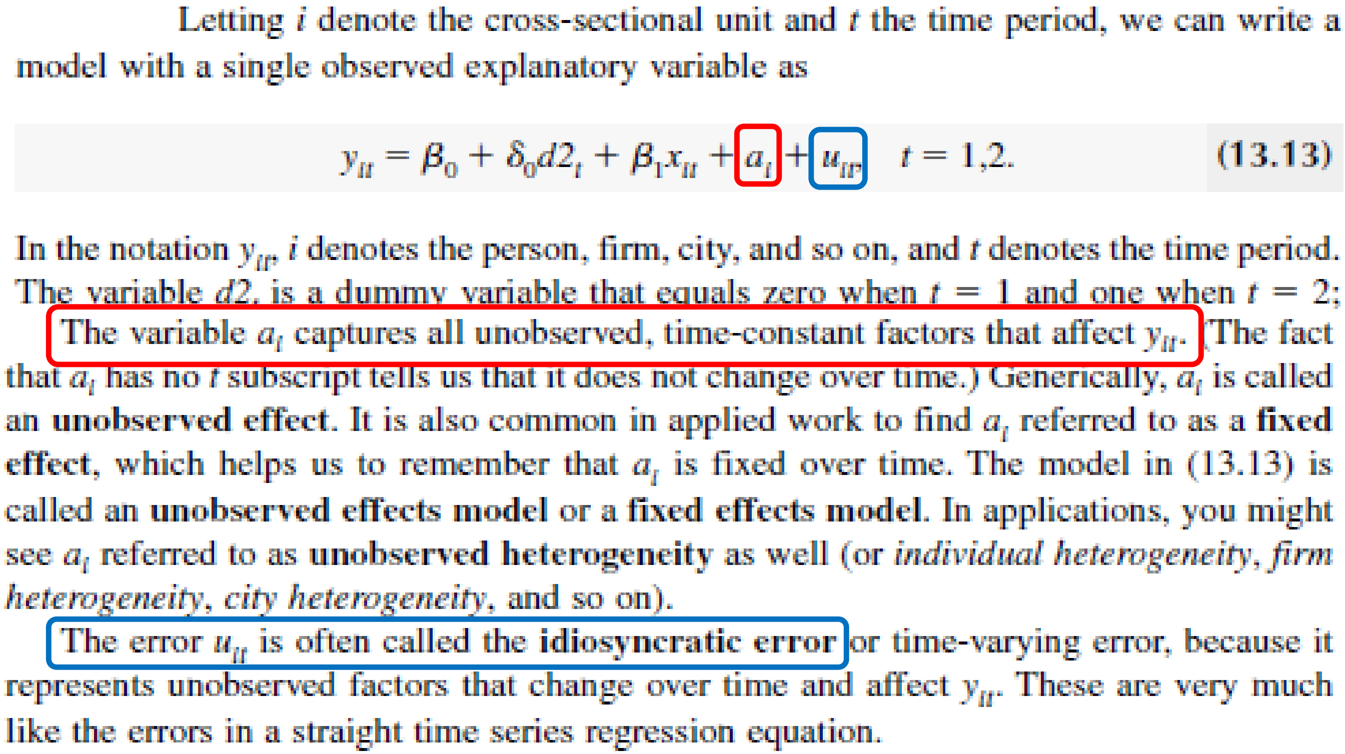
In the notation $y_{it}$, $i$ denotes the person, firm, city, and so on, and $t$ denotes the time period. The variable $d2_t$ is a dummy variable that equals zero when $t = 1$ and one when $t = 2$;

POLS 6481. Research Design and Quantitative Methods II

Lecture 24. Time-Series Cross-Section Data and Methods

Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 3. Explanation: How and *why* first differencing works



Letting $i$ denote the cross-sectional unit and $t$ the time period, we can write a model with a single observed explanatory variable as

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + \boxed{a_i} + \boxed{u_{it}}, \quad t = 1,2. \qquad \textbf{(13.13)}$$

In the notation $y_{it}$, $i$ denotes the person, firm, city, and so on, and $t$ denotes the time period. The variable $d2_t$ is a dummy variable that equals zero when $t = 1$ and one when $t = 2$;

The variable $a_i$ captures all unobserved, time-constant factors that affect $y_{it}$. (The fact that $a_i$ has no $t$ subscript tells us that it does not change over time.) Generically, $a_i$ is called an **unobserved effect**. It is also common in applied work to find $a_i$ referred to as a **fixed effect**, which helps us to remember that $a_i$ is fixed over time. The model in (13.13) is called an **unobserved effects model** or a **fixed effects model**. In applications, you might see $a_i$ referred to as **unobserved heterogeneity** as well (or *individual heterogeneity*, *firm heterogeneity*, *city heterogeneity*, and so on).

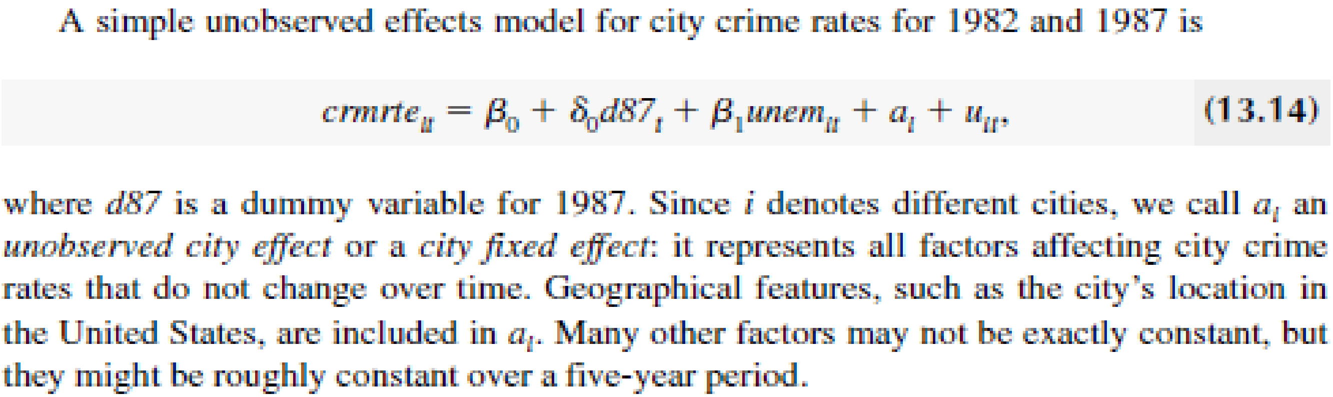
The error $u_{it}$ is often called the **idiosyncratic error** or time-varying error, because it represents unobserved factors that change over time and affect $y_{it}$. These are very much like the errors in a straight time series regression equation.

POLS 6481. Research Design and Quantitative Methods II

Lecture 24. Time-Series Cross-Section Data and Methods

Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 3. Explanation: How and *why* first differencing works



A simple unobserved effects model for city crime rates for 1982 and 1987 is

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}, \tag{13.14}$$

where *d87* is a dummy variable for 1987. Since *i* denotes different cities, we call $a_i$ an *unobserved city effect* or a *city fixed effect*: it represents all factors affecting city crime rates that do not change over time. Geographical features, such as the city's location in the United States, are included in $a_i$. Many other factors may not be exactly constant, but they might be roughly constant over a five-year period.

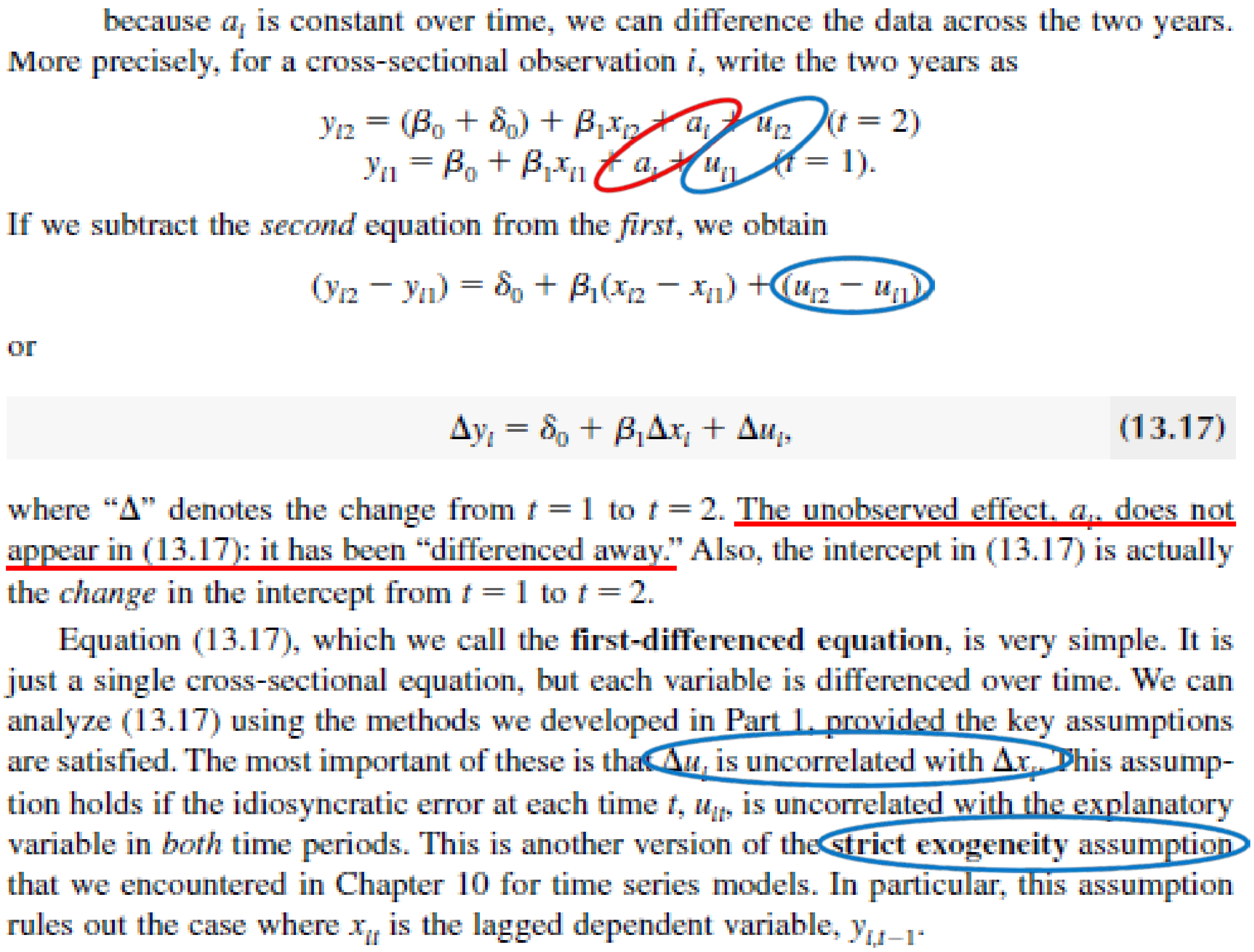
- Suppose we suspect that cities differ from each other in systematic ways, but...
- ... suppose no explanatory variables other than unemployment rate are available.
- Question: Is it possible to estimate *accurately* unemployment's effect on crime?
- Answer: Yes, if cities are observed for at least two periods and other factors affecting crime stay approximately constant over those periods

POLS 6481. Research Design and Quantitative Methods II

Lecture 24. Time-Series Cross-Section Data and Methods

Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 3. Explanation: How and *why* first differencing works



because $a_i$ is constant over time, we can difference the data across the two years. More precisely, for a cross-sectional observation $i$, write the two years as

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \quad (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \quad (t = 1).$$

If we subtract the *second* equation from the *first*, we obtain

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

or

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i, \qquad \textbf{(13.17)}$$

where "Δ" denotes the change from $t = 1$ to $t = 2$. The unobserved effect, $a_i$, does not appear in (13.17): it has been "differenced away." Also, the intercept in (13.17) is actually the *change* in the intercept from $t = 1$ to $t = 2$.

Equation (13.17), which we call the **first-differenced equation**, is very simple. It is just a single cross-sectional equation, but each variable is differenced over time. We can analyze (13.17) using the methods we developed in Part 1, provided the key assumptions are satisfied. The most important of these is that $\Delta u_i$ is uncorrelated with $\Delta x_i$. This assumption holds if the idiosyncratic error at each time $t$, $u_{it}$, is uncorrelated with the explanatory variable in *both* time periods. This is another version of the strict exogeneity assumption that we encountered in Chapter 10 for time series models. In particular, this assumption rules out the case where $x_{it}$ is the lagged dependent variable, $y_{i,t-1}$.

## 3½. *Whether* first differencing works

- **Discussion of first-differenced panel estimator**
  - Further explanatory variables may be included in original equation
  - If there is arbitrary correlation between the effects of unobserved, time-invariant characteristics of cross-sectional units ($a_i$) and the explanatory variables ($x_{itj}$), then OLS estimation of the static equation would be inconsistent (omitted variable bias)
  - The first-differenced panel estimator allows consistent estimation of causal effects in the presence of time-invariant endogeneity
  - Consistency requires strict exogeneity in the original equation (i.e., no correlation between the idiosyncratic errors ($u_{it}$) and the explanatory variables ($x_{itj}$))
  - First-differenced estimates will be imprecise if explanatory variables vary only little over time (and will be impossible if explanatory variables are time-invariant)

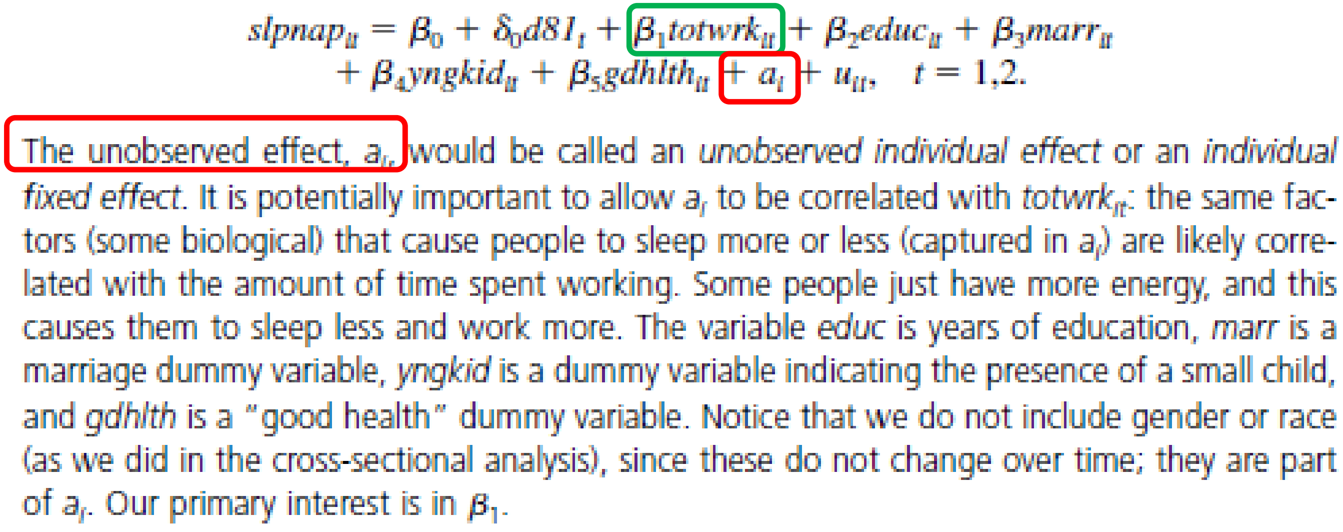
POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 4. Panel Data Example: The Tradeoff Between Work and Sleep

Data: SLP75_81.dta
Script: Lecture 24 sleep.R

J.E. Biddle and D.S. Hamermesh (1990) "Sleep and the Allocation of Time" *Journal of Political Economy* 98: 922–943

More than 700 people were interviewed in 1975; 239 were re-interviewed in 1981.
The dependent variable is the number of minutes of sleep per week.



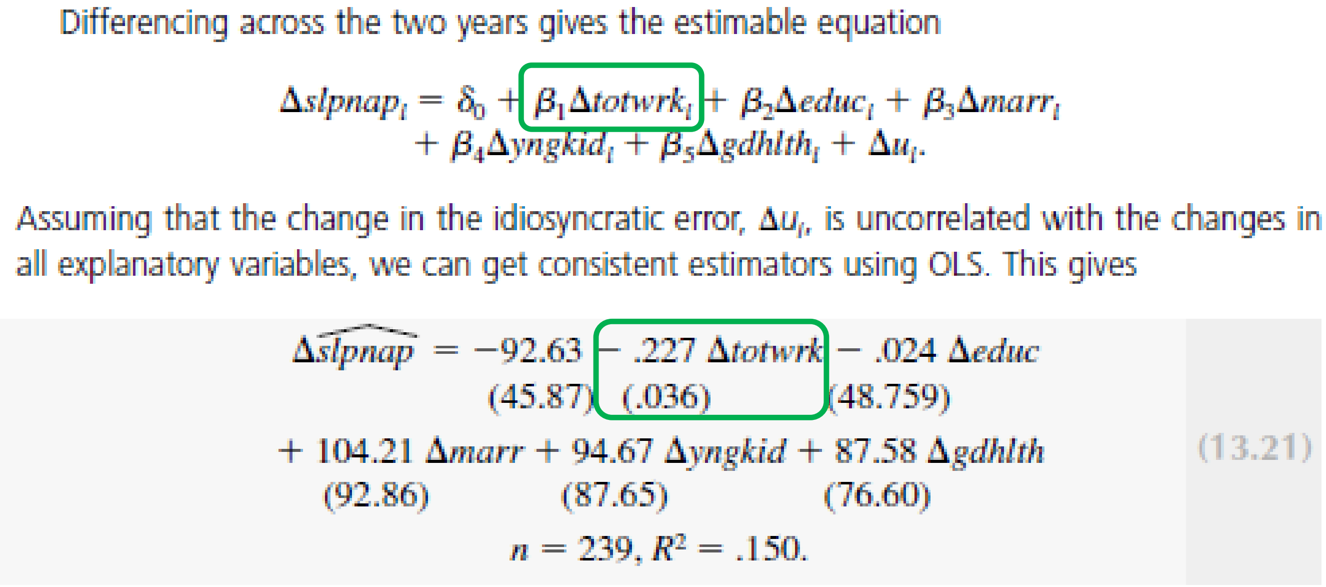
$$slpnap_{it} = \beta_0 + \delta_0 d81_t + \boxed{\beta_1 totwrk_{it}} + \beta_2 educ_{it} + \beta_3 marr_{it} + \beta_4 yngkid_{it} + \beta_5 gdhlth_{it} \boxed{+ a_i} + u_{it}, \quad t = 1,2.$$

The unobserved effect, $a_i$, would be called an *unobserved individual effect* or an *individual fixed effect*. It is potentially important to allow $a_i$ to be correlated with $totwrk_{it}$: the same factors (some biological) that cause people to sleep more or less (captured in $a_i$) are likely correlated with the amount of time spent working. Some people just have more energy, and this causes them to sleep less and work more. The variable *educ* is years of education, *marr* is a marriage dummy variable, *yngkid* is a dummy variable indicating the presence of a small child, and *gdhlth* is a "good health" dummy variable. Notice that we do not include gender or race (as we did in the cross-sectional analysis), since these do not change over time; they are part of $a_i$. Our primary interest is in $\beta_1$.

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e,* 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 4. Panel Data Example: The Tradeoff Between Work and Sleep

| | | |
|---|---|---|
| Data: | SLP75_81.dta | J.E. Biddle and D.S. Hamermesh (1990) "Sleep and the |
| Script: | Lecture 24 sleep.R | Allocation of Time" *Journal of Political Economy* 98: 922–943 |



Differencing across the two years gives the estimable equation

$$\Delta slpnap_i = \delta_0 + \boxed{\beta_1 \Delta totwrk_i} + \beta_2 \Delta educ_i + \beta_3 \Delta marr_i + \beta_4 \Delta yngkid_i + \beta_5 \Delta gdhlth_i + \Delta u_i.$$

Assuming that the change in the idiosyncratic error, $\Delta u_i$, is uncorrelated with the changes in all explanatory variables, we can get consistent estimators using OLS. This gives

$$\begin{aligned}\Delta \widehat{slpnap} &= \underset{(45.87)}{-92.63} \boxed{\underset{(.036)}{-\ .227}\ \Delta totwrk} \underset{(48.759)}{-\ .024}\ \Delta educ \\ &+ \underset{(92.86)}{104.21}\ \Delta marr + \underset{(87.65)}{94.67}\ \Delta yngkid + \underset{(76.60)}{87.58}\ \Delta gdhlth \\ n &= 239,\ R^2 = .150.\end{aligned} \quad (13.21)$$

Only $\Delta$*totwrk* has a statistically significant coefficient ($t$ = –6.31). One additional hour of work per week is associated with 13.6 fewer minutes of sleep per week (13.6 ≈ –.227 * 60). **Few respondents experienced changes in education, marriage, young kids, or health during six-year interval, so std. err.'s are large, and these variables are insignificant.**

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 5. Least-Squares-Dummy-Variable Model

$a_i$ denotes effects of unobserved factors that are cross-sectional-unit-specific and time-invariant (reasons why one unit (firm, person, state, country) has different value of response variable from other units); potentially correlated to explanatory variables

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \; i = 1, \ldots, N, t = 1, \ldots, T$$

introduce a dummy for each individual in the original regression and use pooled OLS:

$$\begin{aligned} y_{it} = {} & a_1 ind1_{it} + a_2 ind2_{it} + \ldots + a_N indN_{it} \\ & + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + u_{it} \end{aligned}$$

Unit-specific intercepts follow a familiar pattern:

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - \cdots - \hat{\beta}_k \bar{x}_{ik}, \; i = 1, \ldots, N$$

## 6. Fixed Effects Model

$a_i$ denotes effects of unobserved factors that are cross-sectional-unit-specific and time-invariant (reasons why one unit (firm, person, state, country) has different value of response variable from other units); potentially correlated to explanatory variables

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \; i = 1, \ldots, N, t = 1, \ldots, T$$

Generate the mean values for each *i* , averaging across all *t* :

$$\bar{y}_i = \beta_1 \bar{x}_{i1} + \cdots + \beta_k \bar{x}_{ik} + \bar{a}_i + \bar{u}_i$$

Subtract the mean values at each *t* :

$$\Rightarrow \; [y_{it} - \bar{y}_i] = \beta_1 [x_{it1} - \bar{x}_{i1}] + \cdots + \beta_k [x_{itk} - \bar{x}_{ik}] + [u_{it} - \bar{u}_i]$$

Because $a_i$ is time-invariant, $a_i - \bar{a}_i = 0$ (i.e., the fixed effect is removed)

“Within” estimator uses across-time variation within cross-sectional units.

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 6. Fixed Effects Model

- **Discussion of fixed-effects panel estimator**
  - Consistency requires strict exogeneity in the original equation (i.e., no correlation between the idiosyncratic errors ($u_{it}$) and the explanatory variables ($x_{itj}$))
  - Degrees of freedom have to be adjusted because the N time averages are estimated in addition to $k$ coefficients (resulting degrees of freedom = $NT–N–k$)
  - The R-squared of the de-meaned equation is inappropriate
  - The effect of time-invariant variables cannot be estimated
  - But the effects of *interactions* with time-invariant variables *can* be estimated
  - If a full set of time dummies are included, the effect of variables whose change over time is constant cannot be estimated (e.g., age, experience)

## 6. Fixed Effects Model

- **Discussion of fixed-effects panel estimator versus first-differencing**
  - Fixed effects and first differencing are identical for $T = 2$
  - Fixed effects is more efficient than first differencing *if the classical assumptions hold* for $T > 2$
  - First differencing may be better in the case of severe serial correlation in the errors
  - First differencing may be better if $T$ is very large (and $N$ not so large) since the panel will have a pronounced time series character and problems like strong dependence arise
  - Otherwise, it is a good idea to compute both and check robustness

## 7. Random Effects Model

Suppose we treat the unit-specific, time-invariant effect, $a_i$, as completely random, i.e. unrelated to explanatory variables:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \ i = 1, \ldots, N, t = 1, \ldots, T$$

The random effects assumption is: $Cov(x_{itj}, a_i) = 0$

If this assumption holds, then the composite error, $a_i + u_{it}$, is uncorrelated with the explanatory variables but is serially correlated for observations coming from the same i:

$$Cov(a_i + u_{it}, a_i + u_{is}) = Cov(a_i, a_i) = \sigma_a^2$$

because both contain $a_i$, even under the assumption that idiosyncratic errors ($u_{it}$) are serially uncorrelated!

For example, in a wage equation, for a given individual, the same effects of unobserved ability would appear in the error term of each period.

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 7. Random Effects Model

- Under the random effects assumption, the explanatory variables are exogenous, and so pooled OLS provides consistent estimates
- If pooled OLS is used, then standard errors have to be adjusted for serial correlation in errors for given *i* (one method is clustered standard errors)
- Because of the serial correlation, OLS is not efficient
- However, one can transform the model by quasi-demeaning the data:

$$[y_{it} - \boxed{\lambda}\bar{y}_i] = \beta_1 [x_{it1} - \boxed{\lambda}\bar{x}_{i1}] + \cdots + \beta_k [x_{itk} - \boxed{\lambda}\bar{x}_{ik}] + \underbrace{[a_i - \boxed{\lambda}\bar{a}_i + u_{it} - \boxed{\lambda}\bar{u}_i]}_{\text{Error satisfies Gauss-Markov assumptions}}$$

$$\boxed{\lambda = 1 - \left[\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)\right]^{1/2}}, \quad 0 \leq \lambda \leq 1$$

## 7. Random Effects Model

- The quasi-demeaning parameter is unknown, but it can be estimated
- FGLS using the estimated λ is called random effects estimation
- If the random effect is relatively unimportant compared to the idosyncratic error, FGLS will be close to pooled OLS (because λ → 0)
- If the random effect is relatively important compared to the idiosyncratic term, FGLS will be similar to fixed effects (because λ → 1)
- Random effects estimation allows you to include time-invariant variables!

$$[y_{it} - \boxed{\lambda}\bar{y}_i] = \beta_1 [x_{it1} - \boxed{\lambda}\bar{x}_{i1}] + \cdots + \beta_k [x_{itk} - \boxed{\lambda}\bar{x}_{ik}] + [a_i - \boxed{\lambda}\bar{a}_i + u_{it} - \boxed{\lambda}\bar{u}_i]$$

$$\boxed{\lambda = 1 - \left[\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)\right]^{1/2}},\ \ 0 \leq \lambda \leq 1$$

$$\lambda = 1 - \sqrt{\frac{1}{1 + \frac{T \cdot \sigma_a^2}{\sigma_u^2}}}$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 8. A Hausman Test

Fixed effects is consistent – unit-specific effects automatically are uncorrelated to regressors; however it is also inefficient – consumes many degrees of freedom, some unnecessarily

Random effects is efficient – it treats unit-specific effects as part of the error to be minimized; however it may be inconsistent – unit-specific errors might be correlated to regressors, which would lead to omitted variable bias.

The Hausman test compares the vectors of FE and RE coefficients;
if the difference is small (high *p* value), then we conclude that RE is consistent; choose RE;
if the difference is large (small *p* value), then we conclude that RE is inconsistent; choose FE.

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 8. A Hausman Test

Earlier we examined unemployment and crime rates in 1982 and 1987. The R script also contains fixed- and random-effects estimators, which yield the expected, positive relationship. The Hausman test yields an insignificant *p* value, so we retain the null hypothesis (that the random effects estimator is consistent), telling us that we ought to prefer the random effects model (which is efficient).

In lab 14, you use data on beer taxes and vehicular accidents. The fixed effects estimator yields a negative relationship (higher beer taxes leads to fewer vehicular accidents), but the random effects estimator yields a positive but insignificant relationship; the Hausman test yields a significant *p* value, so we reject the null hypothesis in favor of the alternative hypothesis (that the random effects estimator is inconsistent), telling you that you ought to prefer the fixed effects model.

Stop and think: What was the random effects assumption? What do the results of these two analyses tell you about the random effects assumption with each dataset?

POLS 6481. Research Design and Quantitative Methods II
Lecture 24. Time-Series Cross-Section Data and Methods
Readings: Wooldridge, *Introductory Econometrics 5e*, 13.3 + 13.5 + 14.1 + 14.2 + 14.4

## 9. Longitudinal Data Example: Divorce Rates

Data: divorce_Frees.dta
Script: Lecture 24 divorce no panel.R
Source: *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*
Edward W. Frees (2004, Cambridge University Press)

Sample is 51 states (including DC) in 1965, 1975, 1985, and 1995
$time = \{ 1, 2, 3, 4 \}$; $N \times T = 204$ *

Response variable: number of divorces & annulments per 1000 population ($\mathit{divorce}_t$)
Explanatory variables: average monthly welfare payments per family ($\mathit{afdc}_t$)
total unemployment rate, as percent of workforce ($\mathit{unemploy}_t$)

Three models: pooled model
fixed effects model
random effects model (followed by Hausman test)

* Note there are some missing data, especially Louisiana and Nevada for most years…