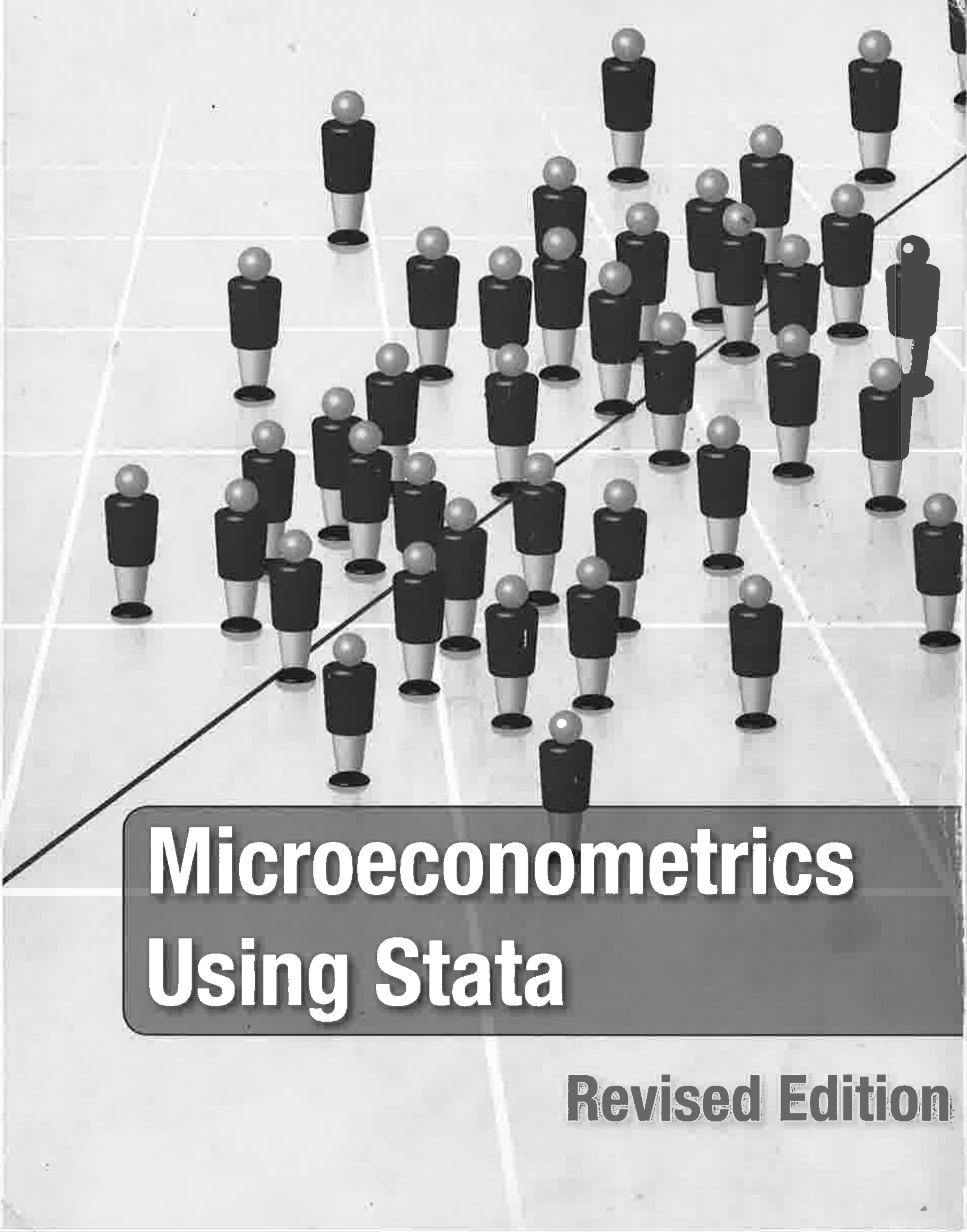

Microeconometrics Using Stata

Revised Edition

A. COLIN CAMERON
PRAVIN K. TRIVEDI

Stata Press

8 Linear panel-data models: Basics

8.1 Introduction

Panel data or longitudinal data are repeated measurements at different points in time on the same individual unit, such as person, firm, state, or country. Regressions can then capture both variation over units, similar to regression on cross-section data, and variation over time.

Panel-data methods are more complicated than cross-section–data methods. The standard errors of panel-data estimators need to be adjusted because each additional time period of data is not independent of previous periods. Panel data requires the use of much richer models and estimation methods. Also different areas of applied statistics use different methods for essentially the same data. The Stata `xt` commands, where `xt` is an acronym for cross-section time series, cover many of these methods.

We focus on methods for a short panel, meaning data on many individual units and few time periods. Examples include longitudinal surveys of many individuals and panel datasets on many firms. And we emphasize microeconometrics methods that attempt to estimate key marginal effects that can be given a causative interpretation.

The essential panel-data methods are given in this chapter, most notably, the important distinction between fixed-effects and random-effects models. Chapter 9 presents many other panel-data methods for the linear model, including those for instrumental-variables (IV) estimation, estimation when lagged dependent variables are regressors, estimation when panels are long rather than short, and estimation of mixed models with slope parameters that vary across individuals. Chapter 9 also shows how methods for short panels are applicable to other forms of clustered data or hierarchical data, such as cross-section individual data from a survey conducted at a number of villages, with clustering at the village level. Nonlinear models are presented in chapter 18.

8.2 Panel-data methods overview

There are many types of panel data and goals of panel-data analysis, leading to different models and estimators for panel data. We provide an overview in this section, with subsequent sections illustrating many of the various models and estimation methods.

8.2.1 Some basic considerations

First, panel data are usually observed at regular time intervals, as is the case for most time-series data. A common exception is growth curve analysis where, for example, children are observed at several irregularly spaced intervals in time, and a measure such as height or IQ is regressed on a polynomial in age.

Second, panel data can be balanced, meaning all individual units are observed in all time periods ($T_i = T$ for all i), or unbalanced ($T_i \neq T$ for some i). Most `xt` commands can be applied to both balanced and unbalanced data. In either case, however, estimator consistency requires that the sample-selection process does not lead to errors being correlated with regressors. Loosely speaking, the missingness is for random reasons rather than systematic reasons.

Third, the dataset may be a short panel (few time periods and many individuals), a long panel (many time periods and few individuals), or both (many time periods and many individuals). This distinction has consequences for both estimation and inference.

Fourth, model errors are very likely correlated. Microeconometrics methods emphasize correlation (or clustering) over time for a given individual, with independence over individual units. For some panel datasets, such as country panels, there additionally may be correlation across individuals. Regardless of the assumptions made, some correction to default ordinary least-squares (OLS) standard errors is usually necessary and efficiency gains using generalized least squares (GLS) may be possible.

Fifth, regression coefficient identification for some estimators can depend on regressor type. Some regressors, such as gender, may be time invariant with $x_{it} = x_i$ for all t. Some regressors, such as an overall time trend, may be individual invariant with $x_{it} = x_t$ for all i. And some may vary over both time and individuals.

Sixth, some or all model coefficients may vary across individuals or over time.

Seventh, the microeconometrics literature emphasizes the fixed-effects model. This model, explained in the next section, permits regressors to be endogenous provided that they are correlated only with a time-invariant component of the error. Most other branches of applied statistics instead emphasize the random-effects model that assumes that regressors are completely exogenous.

Finally, panel data permit estimation of dynamic models where lagged dependent variables may be regressors. Most panel-data analyses use models without this complication.

In this chapter, we focus on short panels (T fixed and $N \rightarrow \infty$) with model errors assumed to be independent over individuals. Long panels are treated separately in section 8.10. We consider linear models with and without fixed effects, and both static and dynamic models. The applications in this chapter use balanced panels. Most commands can also be applied to unbalanced panels, as demonstrated in some of the exercises, though one should also then check for panel-attrition bias.

8.2.2 Some basic panel models

There are several different linear models for panel data.

The fundamental distinction is that between fixed-effects and random-effects models. The term "fixed effects" is misleading because in both types of models individual-level effects are random. Fixed-effects models have the added complication that regressors may be correlated with the individual-level effects so that consistent estimation of regression parameters requires eliminating or controlling for the fixed effects.

Individual-effects model

The individual-specific–effects model for the scalar dependent variable y_{it} specifies that

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \tag{8.1}$$

where $\mathbf{x}_{it}$ are regressors, α_i are random individual-specific effects, and ε_{it} is an idiosyncratic error.

Two quite different models for the α_i are the fixed-effects and random-effects models.

Fixed-effects model

In the fixed-effects (FE) model, the α_i in (8.1) are permitted to be correlated with the regressors $\mathbf{x}_{it}$. This allows a limited form of endogeneity. We view the error in (8.1) as $u_{it} = \alpha_i + \varepsilon_{it}$ and permit $\mathbf{x}_{it}$ to be correlated with the time-invariant component of the error (α_i), while continuing to assume that $\mathbf{x}_{it}$ is uncorrelated with the idiosyncratic error ε_{it}. For example, we assume that if regressors in an earnings regression are correlated with unobserved ability, they are correlated only with the time-invariant component of ability, captured by α_i.

One possible estimation method is to jointly estimate $\alpha_1, \ldots, \alpha_N$ and $\boldsymbol{\beta}$. But for a short panel, asymptotic theory relies on $N \rightarrow \infty$, and here as $N \rightarrow \infty$ so too does the number of fixed effects to estimate. This problem is called the incidental-parameters problem. Interest lies in estimating $\boldsymbol{\beta}$, but first we need to control for the nuisance or incidental parameters, α_i.

Instead, it is still possible to consistently estimate $\boldsymbol{\beta}$, for time-varying regressors, by appropriate differencing transformations applied to (8.1) that eliminate α_i. These estimators are detailed in sections 8.5 and 8.9.

The FE model implies that $E(y_{it}|\alpha_i, \mathbf{x}_{it}) = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}$, assuming $E(\varepsilon_{it}|\alpha_i, \mathbf{x}_{it}) = 0$, so $\beta_j = \partial E(y_{it}|\alpha_i, \mathbf{x}_{it})/\partial x_{j,it}$. The attraction of the FE model is that we can obtain a consistent estimate of the marginal effect of the jth regressor on $E(y_{it}|\alpha_i, \mathbf{x}_{it})$, provided $x_{j,it}$ is time varying, even if the regressors are endogenous (albeit, a limited form of endogeneity).

At the same time, knowledge of $\boldsymbol{\beta}$ does not give complete information on the process generating y_{it}. In particular for prediction, we need an estimate of $E(y_{it}|\mathbf{x}_{it}) = E(\alpha_i|\mathbf{x}_{it}) + \mathbf{x}_{it}'\boldsymbol{\beta}$, and $E(\alpha_i|\mathbf{x}_{it})$ cannot be consistently estimated in short panels.

In nonlinear FE models, these results need to be tempered. It is not always possible to eliminate α_i, which is shown in chapter 18. And even if it is, consistent estimation of $\boldsymbol{\beta}$ may still not lead to a consistent estimate of the marginal effect $\partial E(y_{it}|\alpha_i, \mathbf{x}_{it})/\partial x_{j,it}$.

Random-effects model

In the random-effects (RE) model, it is assumed that α_i in (8.1) is purely random, a stronger assumption implying that α_i is uncorrelated with the regressors.

Estimation is then by a feasible generalized least-squares (FGLS) estimator, given in section 8.6. Advantages of the RE model are that it yields estimates of all coefficients and hence marginal effects, even those of time-invariant regressors, and that $E(y_{it}|\mathbf{x}_{it})$ can be estimated. The big disadvantage is that these estimates are inconsistent if the FE model is appropriate.

Pooled model or population-averaged model

Pooled models assume that regressors are exogenous and simply write the error as u_{it} rather than using the decomposition $\alpha_i + \varepsilon_{it}$. Then

$$y_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it} \qquad (8.2)$$

Note that $\mathbf{x}_{it}$ here does not include a constant, whereas in cross-section chapters, $\mathbf{x}_i$ additionally included a constant term.

OLS estimation of the parameters of this model is straightforward, but inference needs to control for likely correlation of the error u_{it} over time for a given individual (within correlation) and possible correlation over individuals (between correlation). FGLS estimation of (8.2) given an assumed model for the within correlation of u_{it} is presented in section 8.4. In the statistics literature, this is called a population-averaged model. Like RE estimators, consistency of the estimators requires that regressors be uncorrelated with u_{it}.

Two-way–effects model

A standard extension of the individual effects is a two-way–effects model that allows the intercept to vary over individuals and over time:

$$y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} \qquad (8.3)$$

For short panels, it is common to let the time effects γ_t be fixed effects. Then (8.3) reduces to (8.1), if the regressors in (8.1) include a set of time dummies (with one time dummy dropped to avoid the dummy-variable trap).

Mixed linear models

If the RE model is appropriate, richer models can permit slope parameters to also vary over individuals or time. The mixed linear model is a hierarchical linear model that is quite flexible and permits random parameter variation to depend on observable variables. The random-coefficients model is a special case that specifies

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

where $(\alpha_i\ \boldsymbol{\beta}'_i)' \sim (\boldsymbol{\beta}, \boldsymbol{\Sigma})$. For a long panel with few individuals, α_i and $\boldsymbol{\beta}_i$ can instead be parameters that can be estimated by running separate regressions for each individual.

8.2.3 Cluster–robust inference

Various estimators for the preceding models are given in subsequent sections. These estimators are usually based on the assumption that the idiosyncratic error $\varepsilon_{it} \sim (0, \sigma^2_\varepsilon)$. This assumption is often not satisfied in panel applications. Then many panel estimators still retain consistency, provided that ε_{it} are independent over i, but reported standard errors are incorrect.

For short panels, it is possible to obtain cluster–robust standard errors under the weaker assumptions that errors are independent across individuals and that $N \rightarrow \infty$. Specifically, $E(\varepsilon_{it}\varepsilon_{js}) = 0$ for $i \neq j$, $E(\varepsilon_{it}\varepsilon_{is})$ is unrestricted, and ε_{it} may be heteroskedastic. Where applicable, we use cluster–robust standard errors rather than the Stata defaults. For some, but not all, `xt` commands, the `vce(robust)` option is available. This leads to a cluster–robust estimate of the variance–covariance matrix of the estimator (VCE) for some commands and a robust estimate of the VCE for some commands. Otherwise, the `vce(bootstrap)` or `vce(jackknife)` options can be used because, for `xt` commands, these usually resample over clusters.

8.2.4 The xtreg command

The key command for estimation of the parameters of a linear panel-data model is the `xtreg` command. The command syntax is

```
xtreg depvar [indepvars] [if] [in] [weight] [, options]
```

The individual identifier must first be declared with the `xtset` command.

The key model options are population-averaged model (`pa`), FE model (`fe`), RE model (`re` and `mle`), and between-effects model (`be`). The individual models are discussed in detail in subsequent sections. The *weight* modifier is available only for `fe`, `mle`, and `pa`.

The `vce(robust)` option provides cluster–robust estimates of the standard errors, for all models but `be` and `mle`. Stata 10 labels the estimated VCE as simply "Robust" because the use of `xtreg` implies that we are in a clustered setting.

8.2.5 Stata linear panel-data commands

Table 8.1 summarizes `xt` commands for viewing panel data and estimating the parameters of linear panel-data models.

Table 8.1. Summary of `xt` commands

Data summary	`xtset`; `xtdescribe`; `xtsum`; `xtdata`; `xtline`; `xttab`; `xttrans`
Pooled OLS	`regress`
Pooled FGLS	`xtgee, family(gaussian)`; `xtgls`; `xtpcse`
Random effects	`xtreg, re`; `xtregar, re`
Fixed effects	`xtreg, fe`; `xtregar, fe`
Random slopes	`xtmixed`; `xtrc`
First-differences	`regress` (with differenced data)
Static IV	`xtivreg`; `xthtaylor`
Dynamic IV	`xtabond`; `xtdpdsys`; `xtdpd`

The core methods are presented in this chapter, with more specialized commands presented in chapter 9. Readers with long panels should look at section 8.10 (`xtgls`, `xtpcse`, `xtregar`) and data input may require first reading section 8.11.

8.3 Panel-data summary

In this section, we present various ways to summarize and view panel data and estimate a pooled OLS regression. The dataset used is a panel on log hourly wages and other variables for 595 people over the seven years 1976–1982.

8.3.1 Data description and summary statistics

The data, from Baltagi and Khanti-Akom (1990), were drawn from the Panel Study of Income Dynamics (PSID) and are a corrected version of data originally used by Cornwell and Rupert (1988).

The `mus08psidextract.dta` dataset has the following data:

```
. * Read in dataset and describe
. use mus08psidextract.dta, clear
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))
. describe
Contains data from mus08psidextract.dta
  obs:         4,165                          PSID wage data 1976-82 from Baltagi
                                                and Khanti-Akom (1990)
 vars:            22                          16 Aug 2007 16:29
 size:       295,715 (99.1% of memory free)   (_dta has notes)
-----------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-----------------------------------------------------------------------------------
exp             float  %9.0g                  years of full-time work
                                                experience
wks             float  %9.0g                  weeks worked
occ             float  %9.0g                  occupation; occ==1 if in a
                                                blue-collar occupation
ind             float  %9.0g                  industry; ind==1 if working in a
                                                manufacturing industry
south           float  %9.0g                  residence; south==1 if in the
                                                South area
smsa            float  %9.0g                  smsa==1 if in the Standard
                                                metropolitan statistical area
ms              float  %9.0g                  marital status
fem             float  %9.0g                  female or male
union           float  %9.0g                  if wage set be a union contract
ed              float  %9.0g                  years of education
blk             float  %9.0g                  black
lwage           float  %9.0g                  log wage
id              float  %9.0g
t               float  %9.0g
tdum1           byte   %8.0g                  t== 1.0000
tdum2           byte   %8.0g                  t== 2.0000
tdum3           byte   %8.0g                  t== 3.0000
tdum4           byte   %8.0g                  t== 4.0000
tdum5           byte   %8.0g                  t== 5.0000
tdum6           byte   %8.0g                  t== 6.0000
tdum7           byte   %8.0g                  t== 7.0000
exp2            float  %9.0g
-----------------------------------------------------------------------------------
Sorted by:  id  t
```

There are 4,165 individual–year pair observations. The variable labels describe the variables fairly clearly, though note that `lwage` is the log of hourly wage in cents, the indicator `fem` is 1 if female, `id` is the individual identifier, `t` is the year, and `exp2` is the square of `exp`.

(*Continued on next page*)

Descriptive statistics can be obtained by using the command summarize:

```
. * Summary of dataset
. summarize
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         exp |      4165    19.85378    10.96637          1         51
         wks |      4165    46.81152    5.129098          5         52
         occ |      4165    .5111645    .4999354          0          1
         ind |      4165    .3954382    .4890033          0          1
       south |      4165    .2902761    .4539442          0          1
-------------+--------------------------------------------------------
        smsa |      4165    .6537815     .475821          0          1
          ms |      4165    .8144058    .3888256          0          1
         fem |      4165     .112605    .3161473          0          1
       union |      4165    .3639856    .4812023          0          1
          ed |      4165    12.84538    2.787995          4         17
-------------+--------------------------------------------------------
         blk |      4165    .0722689    .2589637          0          1
       lwage |      4165    6.676346    .4615122    4.60517      8.537
          id |      4165         298    171.7821          1        595
           t |      4165           4     2.00024          1          7
       tdum1 |      4165    .1428571    .3499691          0          1
-------------+--------------------------------------------------------
       tdum2 |      4165    .1428571    .3499691          0          1
       tdum3 |      4165    .1428571    .3499691          0          1
       tdum4 |      4165    .1428571    .3499691          0          1
       tdum5 |      4165    .1428571    .3499691          0          1
       tdum6 |      4165    .1428571    .3499691          0          1
-------------+--------------------------------------------------------
       tdum7 |      4165    .1428571    .3499691          0          1
        exp2 |      4165     514.405    496.9962          1       2601
```

The variables take on values that are within the expected ranges, and there are no missing values. Both men and women are included, though from the mean of fem only 11% of the sample is female. Wages data are nonmissing in all years, and weeks worked are always positive, so the sample is restricted to individuals who work in all seven years.

8.3.2 Panel-data organization

The xt commands require that panel data be organized in so-called long form, with each observation a distinct individual–time pair, here an individual–year pair. Data may instead be organized in wide form, with a single observation combining data from all years for a given individual or combining data on all individuals for a given year. Then the data need to be converted from wide form to long form by using the reshape command presented in section 8.11.

Data organization can often be clear from listing the first few observations. For brevity, we list the first three observations for a few variables:

```
. * Organization of dataset
. list id t exp wks occ in 1/3, clean

       id   t   exp   wks   occ
  1.    1   1     3    32     0
  2.    1   2     4    43     0
  3.    1   3     5    40     0
```

The first observation is for individual 1 in year 1, the second observation is for individual 1 in year 2, and so on. These data are thus in long form. From summarize, the panel identifier id takes on the values 1–595, and the time variable t takes on the values 1–7. In general, the panel identifier need just be a unique identifier and the time variable could take on values of, for example, 76–82.

The panel-data xt commands require that, at a minimum, the panel identifier be declared. Many xt commands also require that the time identifier be declared. This is done by using the xtset command. Here we declare both identifiers:

```
. * Declare individual identifier and time identifier
. xtset id t
       panel variable:  id (strongly balanced)
        time variable:  t, 1 to 7
                delta:  1 unit
```

The panel identifier is given first, followed by the optional time identifier. The output indicates that data are available for all individuals in all time periods (strongly balanced) and the time variable increments uniformly by one.

When a Stata dataset is saved, the current settings, if any, from xtset are also saved. In this particular case, the original Stata dataset psidextract.dta already contained this information, so the preceding xtset command was actually unnecessary. The xtset command without any arguments reveals the current settings, if any.

8.3.3 Panel-data description

Once the panel data are xtset, the xtdescribe command provides information about the extent to which the panel is unbalanced.

```
. * Panel description of dataset
. xtdescribe
      id:  1, 2, ..., 595                                    n =        595
       t:  1, 2, ..., 7                                      T =          7
           Delta(t) = 1 unit
           Span(t)  = 7 periods
           (id*t uniquely identifies each observation)
Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         7       7       7         7         7       7       7
     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
      595    100.00  100.00 |  1111111
 ---------------------------+---------
      595    100.00         |  XXXXXXX
```

In this case, all 595 individuals have exactly 7 years of data. The data are therefore balanced because, additionally, the earlier `summarize` command showed that there are no missing values. Section 18.3 provides an example of `xtdescribe` with unbalanced data.

8.3.4 Within and between variation

Dependent variables and regressors can potentially vary over both time and individuals. Variation over time or a given individual is called within variation, and variation across individuals is called between variation. This distinction is important because estimators differ in their use of within and between variation. In particular, in the FE model the coefficient of a regressor with little within variation will be imprecisely estimated and will be not identified if there is no within variation at all.

The `xtsum`, `xttab`, and `xttrans` commands provide information on the relative importance of within variation and between variation of a variable.

We begin with `xtsum`. The total variation (around grand mean $\overline{x} = 1/NT \sum_i \sum_t x_{it}$) can be decomposed into within variation over time for each individual (around individual mean $\overline{x}_i = 1/T \sum_t x_{it}$) and between variation across individuals (for $\overline{x}$ around $\overline{x}_i$). The corresponding decomposition for the variance is

$$\text{Within variance:} \quad s^2_{\text{W}} = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \overline{x}_i)^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \overline{x}_i + \overline{x})^2$$

$$\text{Between variance:} \quad s^2_{\text{B}} = \frac{1}{N-1} \sum_i (\overline{x}_i - \overline{x})^2$$

$$\text{Overall variance:} \quad s^2_{\text{O}} = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \overline{x})^2$$

The second expression for s^2_{W} is equivalent to the first, because adding a constant does not change the variance, and is used at times because $x_{it} - \overline{x}_i + \overline{x}$ is centered on $\overline{x}$, providing a sense of scale, whereas $x_{it} - \overline{x}_i$ is centered on zero. For unbalanced data, replace NT in the formulas with $\sum_i T_i$. It can be shown that $s^2_{\text{O}} \simeq s^2_{\text{W}} + s^2_{\text{B}}$.

The `xtsum` command provides this variance decomposition. We do this for selected regressors and obtain

```
. * Panel summary statistics: within and between variation
. xtsum id t lwage ed exp exp2 wks south tdum1

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
id       overall |       298   171.7821          1        595 |     N =    4165
         between |              171.906          1        595 |     n =     595
         within  |                    0        298        298 |     T =       7
                 |                                            |
t        overall |         4    2.00024          1          7 |     N =    4165
         between |                    0          4          4 |     n =     595
         within  |              2.00024          1          7 |     T =       7
                 |                                            |
lwage    overall |  6.676346   .4615122    4.60517      8.537 |     N =    4165
         between |             .3942387     5.3364   7.813596 |     n =     595
         within  |             .2404023   4.781808   8.621092 |     T =       7
                 |                                            |
ed       overall |  12.84538   2.787995          4         17 |     N =    4165
         between |             2.790006          4         17 |     n =     595
         within  |                    0   12.84538   12.84538 |     T =       7
                 |                                            |
exp      overall |  19.85378   10.96637          1         51 |     N =    4165
         between |             10.79018          4         48 |     n =     595
         within  |              2.00024   16.85378   22.85378 |     T =       7
                 |                                            |
exp2     overall |   514.405   496.9962          1       2601 |     N =    4165
         between |             489.0495         20       2308 |     n =     595
         within  |             90.44581    231.405    807.405 |     T =       7
                 |                                            |
wks      overall |  46.81152   5.129098          5         52 |     N =    4165
         between |             3.284016   31.57143   51.57143 |     n =     595
         within  |             3.941881    12.2401   63.66867 |     T =       7
                 |                                            |
south    overall |  .2902761   .4539442          0          1 |     N =    4165
         between |             .4489462          0          1 |     n =     595
         within  |             .0693042  -.5668667   1.147419 |     T =       7
                 |                                            |
tdum1    overall |  .1428571   .3499691          0          1 |     N =    4165
         between |                    0   .1428571   .1428571 |     n =     595
         within  |             .3499691          0          1 |     T =       7
```

Time-invariant regressors have zero within variation, so the individual identifier `id` and the variable `ed` are time-invariant. Individual-invariant regressors have zero between variation, so the time identifier `t` and the time dummy `tdum1` are individual-invariant. For all other variables but `wks`, there is more variation across individuals (between variation) than over time (within variation), so within estimation may lead to considerable efficiency loss. What is not clear from the output from `xtsum` is that while variable `exp` has nonzero within variation, it evolves deterministically because for this sample `exp` increments by one with each additional period. The `min` and `max` columns give the minimums and maximums of x_{it} for `overall`, $\overline{x}_i$ for `between`, and $x_{it} - \overline{x}_i + \overline{x}$ for `within`.

In the `xtsum` output, Stata uses lowercase n to denote the number of individuals and uppercase N to denote the total number of individual–time observations. In our notation, these quantities are, respectively, N and $\sum_{i=1}^{N} T_i$.

The xttab command tabulates data in a way that provides additional details on the within and between variation of a variable. For example,

```
. * Panel tabulation for a variable
. xttab south
                  Overall              Between            Within
    south |    Freq.  Percent      Freq.  Percent        Percent
----------+-----------------------------------------------------
        0 |     2956    70.97        428    71.93          98.66
        1 |     1209    29.03        182    30.59          94.90
----------+-----------------------------------------------------
    Total |     4165   100.00        610   102.52          97.54
                                 (n = 595)
```

The overall summary shows that 71% of the 4,165 individual–year observations had south = 0, and 29% had south = 1. The between summary indicates that of the 595 people, 72% had south = 0 at least once and 31% had south = 1 at least once. The between total percentage is 102.52, because 2.52% of the sampled individuals (15 persons) lived some of the time in the south and some not in the south and hence are double counted. The within summary indicates that 95% of people who ever lived in the south always lived in the south during the time period covered by the panel, and 99% who lived outside the south always lived outside the south. The south variable is close to time-invariant.

The xttab command is most useful when the variable takes on few values, because then there are few values to tabulate and interpret.

The xttrans command provides transition probabilities from one period to the next. For example,

```
. * Transition probabilities for a variable
. xttrans south, freq
residence; |
  south==1 |  residence; south==1
  if in the|  if in the South area
South area |          0          1 |     Total
-----------+-----------------------+----------
         0 |      2,527          8 |     2,535
           |      99.68       0.32 |    100.00
-----------+-----------------------+----------
         1 |          8      1,027 |     1,035
           |       0.77      99.23 |    100.00
-----------+-----------------------+----------
     Total |      2,535      1,035 |     3,570
           |      71.01      28.99 |    100.00
```

One time period is lost in calculating transitions, so 3,570 observations are used. For time-invariant data, the diagonal entries will be 100% and the off-diagonal entries will be 0%. For south, 99.2% of the observations ever in the south for one period remain in the south for the next period. And for those who did not live in the south for one period, 99.7% remained outside the south for the next period. The south variable is close to time-invariant.

The xttrans command is most useful when the variable takes on few values.

8.3.5 Time-series plots for each individual

It can be useful to provide separate time-series plots for some or all individual units.

Separate time-series plots of a variable for one or more individuals can be obtained by using the xtline command. The overlay option overlays the plots for each individual on the same graph. For example,

```
. quietly xtline lwage if id<=20, overlay
```

produces overlaid time-series plots of lwage for the first 20 individuals in the sample.

We provide time-series plots for the first 20 individuals in the sample. The default is to provide a graph legend that identifies each individual that appears in the graph and takes up much of the graph if the graph uses data from many individuals. This legend can be suppressed by using the legend(off) option. Separate plots are obtained for lwage and for wks, and these are then combined by using the graph combine command. We have

```
. * Simple time-series plot for each of 20 individuals
. quietly xtline lwage if id<=20, overlay legend(off) saving(lwage, replace)
. quietly xtline wks if id<=20, overlay legend(off) saving(wks, replace)
. graph combine lwage.gph wks.gph, iscale(1)
```

Figure 8.1 shows that the wage rate increases roughly linearly over time, aside from two individuals with large increases from years 1 to 2, and that weeks worked show no discernible trend over time.

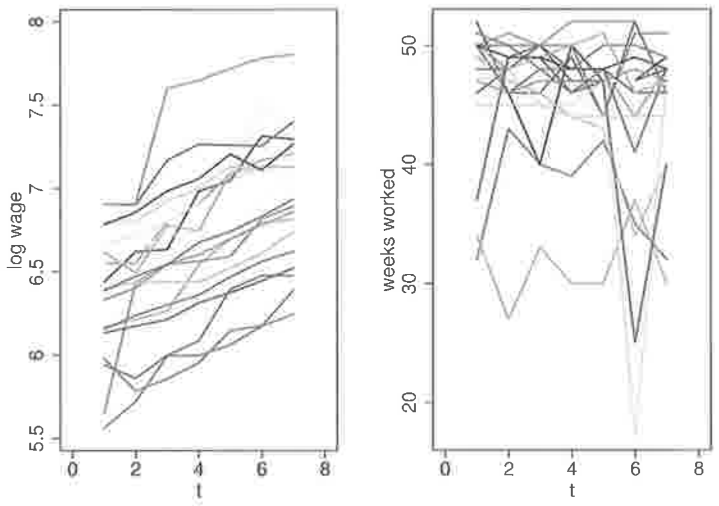

Figure 8.1. Time-series plots of log wage against year and weeks worked against year for each of the first 20 observations

8.3.6 Overall scatterplot

In cases where there is one key regressor, we can begin with a scatterplot of the dependent variable on the key regressor, using data from all panel observations.

The following command adds fitted quadratic regression and lowess regression curves to the scatterplot.

```
. graph twoway (scatter lwage exp) (qfit lwage exp) (lowess lwage exp)
```

This produces a graph that is difficult to read as the scatterplot points are very large, making it hard to then see the regression curves.

The following code presents a better-looking scatterplot of lnwage on exp, along with the fitted regression lines. It uses the same graph options as those explained in section 2.6.6. We have

```
. * Scatterplot, quadratic fit and nonparametric regression (lowess)
. graph twoway (scatter lwage exp, msize(small) msymbol(o))
> (qfit lwage exp, clstyle(p3) lwidth(medthick))
> (lowess lwage exp, bwidth(0.4) clstyle(p1) lwidth(medthick)),
> plotregion(style(none))
> title("Overall variation: Log wage versus experience")
> xtitle("Years of experience", size(medlarge)) xscale(titlegap(*5))
> ytitle("Log hourly wage", size(medlarge)) yscale(titlegap(*5))
> legend(pos(4) ring(0) col(1)) legend(size(small))
> legend(label(1 "Actual Data") label(2 "Quadratic fit") label(3 "Lowess"))
```

Each point on figure 8.2 represents an individual–year pair. The dashed smooth curve line is fitted by OLS of lwage on a quadratic in exp (using qfit), and the solid line is fitted by nonparametric regression (using lowess). Log wage increases until thirty or so years of experience and then declines.

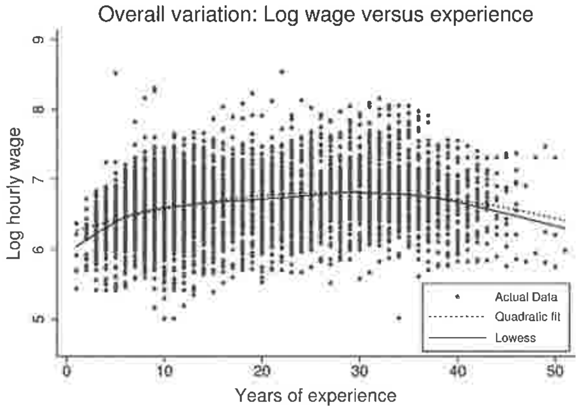

Figure 8.2. Overall scatterplot of log wage against experience using all observations

8.3.7 Within scatterplot

The xtdata command can be used to obtain similar plots for within variation, using option fe; between variation, using option be; and RE variation (the default), using option re. The xtdata command replaces the data in memory with the specified transform, so you should first preserve the data and then restore the data when you are finished with the transformed data.

For example, the fe option creates deviations from means, so that $(y_{it} - \overline{y}_i + \overline{y})$ is plotted against $(x_{it} - \overline{x}_i + \overline{x})$. For lwage plotted against exp, we obtain

```
. * Scatterplot for within variation
. preserve
. xtdata, fe
. graph twoway (scatter lwage exp) (qfit lwage exp) (lowess lwage exp),
> plotregion(style(none)) title("Within variation: Log wage versus experience")
. restore
```

The result is given in figure 8.3. At first glance, this figure is puzzling because only seven distinct values of exp appear. But the panel is balanced and exp (years of work experience) is increasing by exactly one each period for each individual in this sample of people who worked every year. So $(x_{it} - \overline{x}_i)$ increases by one each period, as does $(x_{it} - \overline{x}_i + \overline{x})$. The latter quantity is centered on $\overline{x} = 19.85$ (see section 8.3.1), which is the value in the middle year with $t = 4$. Clearly, it can be very useful to plot a figure such as this.

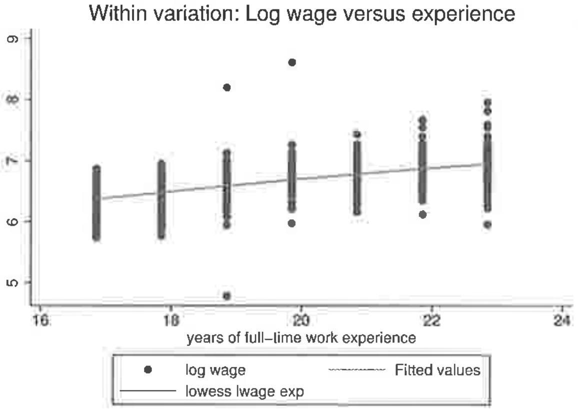

Figure 8.3. Within scatterplot of log-wage deviations from individual means against experience deviations from individual means

8.3.8 Pooled OLS regression with cluster–robust standard errors

A natural starting point is a pooled OLS regression for log wage using data for all individuals in all years.

We include as regressors education, weeks worked, and a quadratic in experience. Education is a time-invariant regressor, taking the same value each year for a given individual. Weeks worked is an example of a time-varying regressor. Experience is also time-varying, though it is so in a deterministic way as the sample comprises people who work full-time in all years, so experience increases by one year as t increments by one.

Regressing y_{it} on $\mathbf{x}_{it}$ yields consistent estimates of $\boldsymbol{\beta}$ if the composite error u_{it} in the pooled model of (8.2) is uncorrelated with $\mathbf{x}_{it}$. As explained in section 8.2, the error u_{it} is likely to be correlated over time for a given individual, so we use cluster–robust standard errors that cluster on the individual. We have

```
. * Pooled OLS with cluster-robust standard errors
. use mus08psidextract.dta, clear
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

. regress lwage exp exp2 wks ed, vce(cluster id)

Linear regression                                      Number of obs =     4165
                                                       F(  4,   594) =    72.58
                                                       Prob > F      =   0.0000
                                                       R-squared     =   0.2836
                                                       Root MSE      =   .39082

                                  (Std. Err. adjusted for 595 clusters in id)

                             Robust
       lwage       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

         exp     .044675   .0054385     8.21   0.000     .0339941     .055356
        exp2   -.0007156   .0001285    -5.57   0.000    -.0009679   -.0004633
         wks     .005827   .0019284     3.02   0.003     .0020396    .0096144
          ed    .0760407   .0052122    14.59   0.000     .0658042    .0862772
       _cons    4.907961   .1399887    35.06   0.000     4.633028    5.182894
```

The output shows that $R^2 = 0.28$, and the estimates imply that wages increase with experience until a peak at 31 years $[= 0.0447/(2 \times 0.00072)]$ and then decline. Wages increase by 0.6% with each additional week worked. And wages increase by 7.6% with each additional year of education.

For panel data, it is essential that OLS standard errors be corrected for clustering on the individual. In contrast, the default standard errors assume that the regression errors are independent and identically distributed (i.i.d.). Using the default standard errors, we obtain

```
. * Pooled OLS with incorrect default standard errors
. regress lwage exp exp2 wks ed

      Source        SS        df       MS              Number of obs =    4165
                                                       F(  4,  4160) =  411.62
       Model   251.491445      4  62.8728613           Prob > F      =  0.0000
    Residual   635.413457   4160  .152743619           R-squared     =  0.2836
                                                       Adj R-squared =  0.2829
       Total   886.904902   4164  .212993492           Root MSE      =  .39082

       lwage       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

         exp     .044675   .0023929    18.67   0.000     .0399838    .0493663
        exp2   -.0007156   .0000528   -13.56   0.000    -.0008191   -.0006121
         wks     .005827   .0011827     4.93   0.000     .0035084    .0081456
          ed    .0760407   .0022266    34.15   0.000     .0716754     .080406
       _cons    4.907961   .0673297    72.89   0.000     4.775959    5.039963
```

These standard errors are misleadingly small; the cluster–robust standard errors are, respectively, 0.0054, 0.0001, 0.0019, and 0.0052.

It is likely that if log wage is overpredicted in one year for a given person, then it is likely to be overpredicted in other years. Failure to control for this error correlation leads to underestimation of standard errors because, intuitively, each additional observation for a given person actually provides less than an independent piece of new information.

The difference between default and cluster–robust standard errors for pooled OLS can be very large. The difference increases with increasing T, increasing autocorrelation in model errors, and increasing autocorrelation of the regressor of interest. Specifically, the standard-error inflation factor $\tau \simeq \sqrt{1 + \rho_u \rho_x (T-1)}$, where ρ_u is the intraclass correlation of the error, defined below in (8.4), and ρ_x is the intraclass correlation of the regressor. Here $\rho_u \simeq 0.80$, shown below, and for time-invariant regressor `ed`, $\rho_x = 1$, so $\tau \simeq \sqrt{1 + 0.80 \times 1 \times 6} = \sqrt{5.8} \simeq 2.41$ for `ed`. Similarly the regressor `exp` has $\rho_x = 1$ because for this sample experience increases by one year as t increments by one.

Cluster–robust standard errors require that $N \rightarrow \infty$ and that errors are independent over i. The assumption of independence over i can be relaxed to independence at a more aggregated level, provided that the number of units is still large and the units nest the individual. For example, the PSID is a household survey and errors for individuals from the same household may be correlated. If, say, `houseid` is available as a household identifier, then we would use the `vce(cluster houseid)` option. As a second example, if the regressor of interest is aggregated at the state level, such as a state policy variable, and there are many states, then it may be better to use the `vce(cluster state)` option.

8.3.9 Time-series autocorrelations for panel data

The Stata time-series operators can be applied to panel data when both panel and time identifiers are set with the `xtset` command. Examples include `L.lwage` or `L1.lwage` for `lwage` lagged once, `L2.lwage` for `lwage` lagged twice, `D.lwage` for the difference in `lwage` (equals `lwage` – `L.lwage`), `LD.lwage` for this difference lagged once, and `L2D.lwage` for this difference lagged twice.

Use of these operators is the best way to create lagged variables because relevant missing values are automatically and correctly created. For example, `regress lwage L2.wage` will use (7 − 2) × 595 observations because forming `L2.wage` leads to a loss of the first two years of data for each of the 595 individuals.

The `corrgram` command for computing autocorrelations of time-series data does not work for panel data. Instead, autocorrelations can be obtained by using the `correlate` command. For example,

```
. * First-order autocorrelation in a variable
. sort id t

. correlate lwage L.lwage
(obs=3570)

             |                L.
             |    lwage    lwage
-------------+------------------
       lwage |
         --. |   1.0000
         L1. |   0.9189   1.0000
```

calculates the first-order autocorrelation coefficient for `lwage` to be 0.92.

We now calculate autocorrelations at all lags (here up to six periods). Rather than doing so for `lwage`, we do so for the residuals from the previous pooled OLS regression for `lwage`. We have

```
. * Autocorrelations of residual
. quietly regress lwage exp exp2 wks ed, vce(cluster id)

. predict uhat, residuals

. forvalues j = 1/6 {
  2.      quietly corr uhat L`j'.uhat
  3.      display "Autocorrelation at lag `j' = " %6.3f r(rho)
  4.      }
Autocorrelation at lag 1 =  0.884
Autocorrelation at lag 2 =  0.838
Autocorrelation at lag 3 =  0.811
Autocorrelation at lag 4 =  0.786
Autocorrelation at lag 5 =  0.750
Autocorrelation at lag 6 =  0.729
```

The `forvalues` loop leads to separate computation of each autocorrelation to maximize the number of observations used. If instead we gave a one-line command to compute the autocorrelations of `uhat` through `L6.uhat`, then only 595 observations would have been used. Here 6 × 595 observations are used to compute the autocorrelation at lag 1, 5 × 595 observations are used to compute the autocorrelation at lag 2, and so on. The average of the autocorrelations, 0.80, provides a rough estimate of the intraclass correlation coefficient of the residuals.

Clearly, the errors are serially correlated, and cluster–robust standard errors after pooled OLS are required. The individual-effects model provides an explanation for this correlation. If the error $u_{it} = \alpha_i + \varepsilon_{it}$, then even if ε_{it} is i.i.d. (0, σ_ε^2), we have $\mathrm{Cor}(u_{it}, u_{is}) \neq 0$ for $t \neq s$ if $\alpha_i \neq 0$. The individual effect α_i induces correlation over time for a given individual.

The preceding estimated autocorrelations are constant across years. For example, the correlation of `uhat` with `L.uhat` across years 1 and 2 is assumed to be the same as that across years 2 and 3, years 3 and 4, ..., years 6 and 7. This presumes that the errors are stationary.

In the nonstationary case, the autocorrelations will differ across pairs of years. For example, we consider the autocorrelations one year apart and allow these to differ across the year pairs. We have

```
. * First-order autocorrelation differs in different year pairs
. forvalues s = 2/7 {
  2.      quietly corr uhat L1.uhat if t == `s'
  3.      display "Autocorrelation at lag 1 in year `s' = " %6.3f r(rho)
  4.      }
Autocorrelation at lag 1 in year 2 =  0.915
Autocorrelation at lag 1 in year 3 =  0.799
Autocorrelation at lag 1 in year 4 =  0.855
Autocorrelation at lag 1 in year 5 =  0.867
Autocorrelation at lag 1 in year 6 =  0.894
Autocorrelation at lag 1 in year 7 =  0.893
```

The lag-1 autocorrelations for individual–year pairs range from 0.80 to 0.92, and their average is 0.87. From the earlier output, the lag-1 autocorrelation equals 0.88 when it is constrained to be equal across all year pairs. It is common to impose equality for simplicity.

8.3.10 Error correlation in the RE model

For the individual-effects model (8.1), the combined error $u_{it} = \alpha_i + \varepsilon_{it}$. The RE model assumes that α_i is i.i.d. with a variance of σ_α^2 and that u_{it} is i.i.d. with a variance of σ_ε^2.

Then u_{it} has a variance of $\mathrm{Var}(u_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$ and a covariance of $\mathrm{Cov}(u_{it}, u_{is}) = \sigma_\alpha^2$, $s \neq t$. It follows that in the RE model,

$$\rho_u = \mathrm{Cor}(u_{it}, u_{is}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2), \text{ for all } s \neq t \tag{8.4}$$

This constant correlation is called the intraclass correlation of the error.

The RE model therefore permits serial correlation in the model error. This correlation can approach 1 if the random effect is large relative to the idiosyncratic error, so that σ_α^2 is large relative to σ_ε^2.

This serial correlation is restricted to be the same at all lags, and the errors u_{it} are then called equicorrelated or exchangeable. From section 8.3.9, the error correlations were, respectively, 0.88, 0.84, 0.81, 0.79, 0.75, and 0.73, so a better model may be one that allows the error correlation to decrease with the lag length.

8.4 Pooled or population-averaged estimators

Pooled estimators simply regress y_{it} on an intercept and $\mathbf{x}_{it}$, using both between (cross-section) and within (time-series) variation in the data. Standard errors need to adjust for any error correlation and, given a model for error correlation, more-efficient FGLS estimation is possible. Pooled estimators, called population-averaged estimators in the statistics literature, are consistent if the RE model is appropriate and are inconsistent if the FE model is appropriate.

8.4.1 Pooled OLS estimator

The pooled OLS estimator can be motivated from the individual-effects model by rewriting (8.1) as the pooled model

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + (\alpha_i - \alpha + \varepsilon_{it}) \tag{8.5}$$

Any time-specific effects are assumed to be fixed and already included as time dummies in the regressors $\mathbf{x}_{it}$. The model (8.5) explicitly includes a common intercept, and the individual effects $\alpha_i - \alpha$ are now centered on zero.

Consistency of OLS requires that the error term $(\alpha_i - \alpha + \varepsilon_{it})$ be uncorrelated with $\mathbf{x}_{it}$. So pooled OLS is consistent in the RE model but is inconsistent in the FE model because then α_i is correlated with $\mathbf{x}_{it}$.

The pooled OLS estimator for our data example has already been presented in section 8.3.8. As emphasized there, cluster–robust standard errors are necessary in the common case of a short panel with independence across individuals.

8.4.2 Pooled FGLS estimator or population-averaged estimator

Pooled FGLS (PFGLS) estimation can lead to estimators of the parameters of the pooled model (8.5) that are more efficient than OLS estimation. Again we assume that any individual-level effects are uncorrelated with regressors, so PFGLS is consistent.

Different assumptions about the correlation structure for the errors u_{it} lead to different PFGLS estimators. In section 8.10, we present some estimators for long panels, using the `xtgls` and `xtregar` commands.

Here we consider only short panels with errors independent across individuals. We need to model the $T \times T$ matrix of error correlations. An assumed correlation structure, called a working matrix, is specified and the appropriate PFGLS estimator is obtained. To guard against the working matrix being a misspecified model of the error correlation, cluster–robust standard errors are computed. Better models for the error correlation lead to more-efficient estimators, but the use of robust standard errors means that the estimators are not presumed to be fully efficient.

In statistics literature, the pooled approach is called a population-averaged (PA) approach, because any individual effects are assumed to be random and are averaged out. The PFGLS estimator is then called the PA estimator.

8.4.3 The xtreg, pa command

The pooled estimator, or PA estimator, is obtained by using the `xtreg` command (see section 8.2.4) with the `pa` option. The two key additional options are `corr()`, to place different restrictions on the error correlations, and `vce(robust)`, to obtain cluster–robust standard errors that are valid even if `corr()` does not specify the correct correlation model, provided that observations are independent over i and $N \rightarrow \infty$.

Let $\rho_{ts} = \mathrm{Cor}(u_{it}u_{is})$, the error correlation over time for individual i, and note the restriction that ρ_{ts} does not vary with i. The `corr()` options all set $\rho_{tt} = 1$ but differ in the model for ρ_{ts} for $t \neq s$. With T time periods, the correlation matrix is $T \times T$, and there are potentially as many as $T(T-1)$ unique off-diagonal entries because it need not necessarily be the case that $\rho_{ts} = \rho_{st}$.

The `corr(independent)` option sets $\rho_{ts} = 0$ for $s \neq t$. Then the PA estimator equals the pooled OLS estimator.

The `corr(exchangeable)` option sets $\rho_{ts} = \rho$ for all $s \neq t$ so that errors are assumed to be equicorrelated. This assumption is imposed by the RE model (see section 8.3.10), and as a result, `xtreg, pa` with this option is asymptotically equivalent to `xtreg, re`.

For panel data, it is often the case that the error correlation ρ_{ts} declines as the time difference $|t-s|$ increases—the application in section 8.3.9 provided an example. The `corr(ar` *k*`)` option models this dampening by assuming an autoregressive process of order k, or AR(k) process, for u_{it}. For example, `corr(ar 1)` assumes that $u_{it} = \rho_1 u_{i,t-1} + \varepsilon_{it}$, which implies that $\rho_{ts} = \rho_1^{|t-s|}$. The `corr(stationary` *g*`)` option instead uses a moving-average process, or MA(g) process. This sets $\rho_{ts} = \rho_{|t-s|}$ if $|t-s| \leq g$, and $\rho_{ts} = 0$ if $|t-s| > g$.

The `corr(unstructured)` option places no restrictions on ρ_{ts}, aside from equality of $\rho_{i,ts}$ across individuals. Then $\mathrm{Cov}(u_{it}, u_{is}) = 1/N \sum_i (\widehat{u}_{it} - \overline{\widehat{u}_t})(\widehat{u}_{is} - \overline{\widehat{u}_s})$. For small T, this may be the best model, but for larger T, the method can fail numerically because there are $T(T-1)$ unique parameters ρ_{ts} to estimate. The `corr(nonstationary` *g*`)` option allows ρ_{ts} to be unrestricted if $|t-s| \leq g$ and sets $\rho_{ts} = 0$ if $|t-s| > g$ so there are fewer correlation parameters to estimate.

The PA estimator is also called the generalized estimating equations estimator in the statistics literature. The `xtreg, pa` command is the special case of `xtgee` with the `family(gaussian)` option. The more general `xtgee` command, presented in section 18.4.4, has other options that permit application to a wide range of nonlinear panel models.

8.4.4 Application of the xtreg, pa command

As an example, we specify an AR(2) error process. We have

```
. * Population-averaged or pooled FGLS estimator with AR(2) error
. xtreg lwage exp exp2 wks ed, pa corr(ar 2) vce(robust) nolog
GEE population-averaged model                   Number of obs      =      4165
Group and time vars:                 id t       Number of groups   =       595
Link:                            identity       Obs per group: min =         7
Family:                          Gaussian                      avg =       7.0
Correlation:                        AR(2)                      max =         7
                                                Wald chi2(4)       =    873.28
Scale parameter:                 .1966639       Prob > chi2        =    0.0000
                                    (Std. Err. adjusted for clustering on id)

                           Semirobust
       lwage      Coef.    Std. Err.      z    P>|z|     [95% Conf. Interval]

         exp    .0718915    .003999    17.98   0.000     .0640535    .0797294
        exp2   -.0008966   .0000933    -9.61   0.000    -.0010794   -.0007137
         wks    .0002964   .0010553     0.28   0.779     -.001772    .0023647
          ed    .0905069   .0060161    15.04   0.000     .0787156    .1022982
       _cons    4.526381   .1056897    42.83   0.000     4.319233    4.733529
```

The coefficients change considerably compared with those from pooled OLS. The cluster-robust standard errors are smaller than those from pooled OLS for all regressors except ed, illustrating the desired improved efficiency because of better modeling of the error correlations. Note that unlike the pure time-series case, controlling for autocorrelation does not lead to the loss of initial observations.

The estimated correlation matrix is stored in e(R). We have

```
. * Estimated error correlation matrix after xtreg, pa
. matrix list e(R)
symmetric e(R)[7,7]
            c1          c2          c3          c4          c5          c6          c7
r1           1
r2   .89722058           1
r3   .84308581   .89722058           1
r4   .78392846   .84308581   .89722058           1
r5   .73064474   .78392846   .84308581   .89722058           1
r6    .6806209   .73064474   .78392846   .84308581   .89722058           1
r7   .63409777    .6806209   .73064474   .78392846   .84308581   .89722058           1
```

By comparison, from section 8.3.9 the autocorrelations of the errors after pooled OLS estimation were 0.88, 0.84, 0.81, 0.79, 0.75, and 0.73.

In an end-of-chapter exercise, we compare estimates obtained using different error-correlation structures.

8.5 Within estimator

Estimators of the parameters $\boldsymbol{\beta}$ of the FE model (8.1) must remove the fixed effects α_i. The within transform discussed in the next section does so by mean-differencing. The within estimator performs OLS on the mean-differenced data. Because all the observations of the mean-difference of a time-invariant variable are zero, we cannot estimate the coefficient on a time-invariant variable.

Because the within estimator provides a consistent estimate of the FE model, it is often called the FE estimator, though the first-difference estimator given in section 8.9 also provides consistent estimates in the FE model. The within estimator is also consistent under the RE model, but alternative estimators are more efficient in the RE model.

8.5.1 Within estimator

The fixed effects α_i in the model (8.1) can be eliminated by subtraction of the corresponding model for individual means $\overline{y}_i = \overline{\mathbf{x}}_i'\boldsymbol{\beta} + \overline{\varepsilon}_i$, leading to the within model or mean-difference model

$$(y_{it} - \overline{y}_i) = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \overline{\varepsilon}_i) \tag{8.6}$$

where, for example, $\overline{\mathbf{x}}_i = T_i^{-1}\sum_{t=1}^{T_i}\mathbf{x}_{it}$. The within estimator is the OLS estimator of this model.

Because α_i has been eliminated, OLS leads to consistent estimates of $\boldsymbol{\beta}$ even if α_i is correlated with $\mathbf{x}_{it}$, as is the case in the FE model. This result is a great advantage of panel data. Consistent estimation is possible even with endogenous regressors $\mathbf{x}_{it}$, provided that $\mathbf{x}_{it}$ is correlated only with the time-invariant component of the error, α_i, and not with the time-varying component of the error, ε_{it}.

This desirable property of consistent parameter estimation in the FE model is tempered, however, by the inability to estimate the coefficients or a time-invariant regressor. Also the within estimator will be relatively imprecise for time-varying regressors that vary little over time.

Stata actually fits the model

$$(y_{it} - \overline{y}_i + \overline{\overline{y}}) = \alpha + (\mathbf{x}_{it} - \overline{\mathbf{x}}_i + \overline{\overline{\mathbf{x}}})'\boldsymbol{\beta} + (\varepsilon_{it} - \overline{\varepsilon}_i + \overline{\overline{\varepsilon}}) \tag{8.7}$$

where, for example, $\overline{\overline{y}} = (1/N)\overline{y}_i$ is the grand mean of y_{it}. This parameterization has the advantage of providing an intercept estimate, the average of the individual effects α_i, while yielding the same slope estimate $\boldsymbol{\beta}$ as that from the within model.

8.5.2 The xtreg, fe command

The within estimator is computed by using the **xtreg** command (see section 8.2.4) with the fe option. The default standard errors assume that after controlling for α_i, the error

ε_{it} is i.i.d. The `vce(robust)` option relaxes this assumption and provides cluster–robust standard errors, provided that observations are independent over i and $N \to \infty$.

8.5.3 Application of the xtreg, fe command

For our data, we obtain

```
. * Within or FE estimator with cluster-robust standard errors
. xtreg lwage exp exp2 wks ed, fe vce(cluster id)
note: ed omitted because of collinearity
Fixed-effects (within) regression               Number of obs      =      4165
Group variable: id                              Number of groups   =       595
R-sq:  within  = 0.6566                         Obs per group: min =         7
       between = 0.0276                                        avg =       7.0
       overall = 0.0476                                        max =         7
                                                F(3,594)           =   1059.72
corr(u_i, Xb)  = -0.9107                        Prob > F           =    0.0000
                                 (Std. Err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .1137879   .0040289    28.24   0.000     .1058753    .1217004
        exp2 |  -.0004244   .0000822    -5.16   0.000    -.0005858   -.0002629
         wks |   .0008359   .0008697     0.96   0.337    -.0008721    .0025439
          ed |  (omitted)
       _cons |   4.596396   .0600887    76.49   0.000     4.478384    4.714408
-------------+----------------------------------------------------------------
     sigma_u |  1.0362039
     sigma_e |  .15220316
         rho |  .97888036   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Compared with pooled OLS, the standard errors have roughly tripled because only within variation of the data is being used. The `sigma_u` and `sigma_e` entries are explained in section 8.8.1, and the R^2 measures are explained in section 8.8.2.

The most striking result is that the coefficient for education is not identified. This is because the data on education is time-invariant. In fact, given that we knew from the `xtsum` output in section 8.3.4 that `ed` had zero within standard deviation, we should not have included it as one of the regressors in the `xtreg, fe` command.

This is unfortunate because how wages depend on education is of great policy interest. It is certainly endogenous, because people with high ability are likely to have on average both high education and high wages. Alternative panel-data methods to control for endogeneity of the `ed` variable are presented in chapter 9. In other panel applications, endogenous regressors may be time-varying and the within estimator will suffice.

8.5.4 Least-squares dummy-variables regression

The within estimator of $\boldsymbol{\beta}$ is also called the FE estimator because it can be shown to equal the estimator obtained from direct OLS estimation of $\alpha_1, \ldots, \alpha_N$ and $\boldsymbol{\beta}$ in the original individual-effects model (8.1). The estimates of the fixed effects are then $\widehat{\alpha}_i = \overline{y}_i - \overline{\mathbf{x}}_i{}'\widehat{\boldsymbol{\beta}}$. In short panels, $\widehat{\alpha}_i$ is not consistently estimated, because it essentially relies on only T_i observations used to form $\overline{y}_i$ and $\overline{\mathbf{x}}_i$, but $\widehat{\boldsymbol{\beta}}$ is nonetheless consistently estimated.

Another name for the within estimator is the least-squares dummy-variable (LSDV) estimator, because it can be shown to equal the estimator obtained from OLS estimation of y_{it} on $\mathbf{x}_{it}$ and N individual-specific indicator variables $d_{j,it}$, $j = 1, \ldots, N$, where $d_{j,it} = 1$ for the ith observation if $j = 1$, and $d_{j,it} = 0$ otherwise. Thus we fit the model

$$y_{it} = \left(\sum\nolimits_{j=1}^{N} \alpha_i d_{j,it}\right) + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \tag{8.8}$$

This equivalence of LSDV and within estimators does not carry over to nonlinear models.

This parameterization provides an alternative way to estimate the parameters of the fixed-effects model, using cross-section OLS commands. The `areg` command, which fits the linear regression (8.8) with one set of mutually exclusive indicators, reports only the estimates of the parameters $\boldsymbol{\beta}$. We have

```
. * LSDV model fit using areg with cluster-robust standard errors
. areg lwage exp exp2 wks ed, absorb(id) vce(cluster id)
note: ed omitted because of collinearity
Linear regression, absorbing indicators                Number of obs =    4165
                                                       F(  3,   594) =  908.44
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9068
                                                       Adj R-squared =  0.8912
                                                       Root MSE      =   .1522
                                 (Std. Err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .1137879   .0043514    26.15   0.000     .1052418    .1223339
        exp2 |  -.0004244   .0000888    -4.78   0.000    -.0005988     -.00025
         wks |   .0008359   .0009393     0.89   0.374    -.0010089    .0026806
          ed |  (omitted)
       _cons |   4.596396   .0648993    70.82   0.000     4.468936    4.723856
-------------+----------------------------------------------------------------
          id |   absorbed                                     (595 categories)
```

The coefficient estimates are the same as those from `xtreg, fe`. The cluster–robust standard errors differ because of different small-sample correction, and those from `xtreg, fe` should be used. This difference arises because inference for `areg` is designed for the case where N is fixed and $T \to \infty$, whereas we are considering the short-panel case, where T is fixed and $N \to \infty$.

The model can also be fit using **regress**. One way to include a set of indicator variables for each individual is by inserting the **i.** operator before the categorical variable **id**. To do this, we need to increase the default setting of **matsize** to at least $N + K$, where K is the number of regressors in this model. The output from **regress** is very long because it includes coefficients for all the dummy variables. We instead suppress the output and use **estimates table** to list results for just the coefficients of interest.

```
. * LSDV model fit using factor variables with cluster-robust standard errors
. set matsize 800
. quietly regress lwage exp exp2 wks ed i.id, vce(cluster id)
. estimates table, keep(exp exp2 wks ed _cons) b se b(%12.7f)

    Variable    active

         exp    0.1137879
                0.0043514
        exp2   -0.0004244
                0.0000888
         wks    0.0008359
                0.0009393
          ed    0.1022134
                0.0046744
       _cons    4.3476807
                0.0443191

                           legend: b/se
```

The coefficient estimates and standard errors are exactly the same as those obtained from **areg**, aside from the constant. For **areg** (and **xtreg, fe**), the intercept is fit so that $\overline{y} - \overline{\overline{\mathbf{x}}}'\widehat{\boldsymbol{\beta}} = 0$, whereas this is not the case using **regress**. The standard errors are the same as those from **areg**, and as already noted, those from **xtreg, fe** should be used.

8.6 Between estimator

The between estimator uses only between or cross-section variation in the data and is the OLS estimator from the regression of $\overline{y}_i$ on $\overline{\mathbf{x}}_i$. Because only cross-section variation in the data is used, the coefficients of any individual-invariant regressors, such as time dummies, cannot be identified. We provide the estimator for completeness, even though it is seldom used because pooled estimators and the RE estimator are more efficient.

8.6.1 Between estimator

The between estimator is inconsistent in the FE model and is consistent in the RE model. To see this, average the individual-effects model (8.1) to obtain the between model

$$\overline{y}_i = \alpha + \overline{\mathbf{x}}_i'\boldsymbol{\beta} + (\alpha_i - \alpha + \overline{\varepsilon}_i)$$

The between estimator is the OLS estimator in this model. Consistency requires that the error term $(\alpha_i - \alpha + \overline{\varepsilon}_i)$ be uncorrelated with $\mathbf{x}_{it}$. This is the case if α_i is a random effect but not if α_i is a fixed effect.

8.6.2 Application of the xtreg, be command

The between estimator is obtained by specifying the **be** option of the **xtreg** command. There is no explicit option to obtain heteroskedasticity-robust standard errors, but these can be obtained by using the **vce(bootstrap)** option.

For our data, the bootstrap standard errors differ from the default by only 10%, because averages are used so that the complication is one of heteroskedastic errors rather than clustered errors. We report the default standard errors that are much more quickly computed. We have

```
. * Between estimator with default standard errors
. xtreg lwage exp exp2 wks ed, be
Between regression (regression on group means)  Number of obs      =      4165
Group variable: id                              Number of groups   =       595
R-sq:  within  = 0.1357                         Obs per group: min =         7
       between = 0.3264                                        avg =       7.0
       overall = 0.2723                                        max =         7
                                                F(4,590)           =     71.48
sd(u_i + avg(e_i.))=   .324656                  Prob > F           =    0.0000

       lwage       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

         exp     .038153   .0056967     6.70   0.000     .0269647    .0493412
        exp2   -.0006313   .0001257    -5.02   0.000    -.0008781   -.0003844
         wks    .0130903   .0040659     3.22   0.001     .0051048    .0210757
          ed    .0737838   .0048985    15.06   0.000     .0641632    .0834044
       _cons    4.683039   .2100989    22.29   0.000     4.270407    5.095672
```

The estimates and standard errors are closer to those obtained from pooled OLS than those obtained from within estimation.

8.7 RE estimator

The RE estimator is the FGLS estimator in the RE model (8.1) under the assumption that the random effect α_i is i.i.d. and the idiosyncratic error ε_{it} is i.i.d. The RE estimator is consistent if the RE model is appropriate and is inconsistent if the FE model is appropriate.

(Continued on next page)

8.7.1 RE estimator

The RE model is the individual-effects model (8.1)

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + (\alpha_i + \varepsilon_{it}) \tag{8.9}$$

with $\alpha_i \sim (\alpha, \sigma_\alpha^2)$ and $\varepsilon_{it} \sim (0, \sigma_u^2)$. Then from (8.4), the combined error $u_{it} = \alpha_i + \varepsilon_{it}$ is correlated over t for the given i with

$$\text{Cor}(u_{it}, u_{is}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2), \text{ for all } s \neq t \tag{8.10}$$

The RE estimator is the FGLS estimator of $\boldsymbol{\beta}$ in (8.9) given (8.10) for the error correlations.

In several different settings, such as heteroskedastic errors and AR(1) errors, the FGLS estimator can be calculated as the OLS estimator in a model transformed to have homoskedastic uncorrelated errors. This is also possible here. Some considerable algebra shows that the RE estimator can be obtained by OLS estimation in the transformed model

$$(y_{it} - \widehat{\theta}_i\overline{y}_i) = (1 - \widehat{\theta}_i)\alpha + (\mathbf{x}_{it} - \widehat{\theta}_i\overline{\mathbf{x}}_i)'\boldsymbol{\beta} + \{(1 - \widehat{\theta}_i)\alpha_i + (\varepsilon_{it} - \widehat{\theta}_i\overline{\varepsilon}_i)\} \tag{8.11}$$

where $\widehat{\theta}_i$ is a consistent estimate of

$$\theta_i = 1 - \sqrt{\sigma_\varepsilon^2/(T_i\sigma_\alpha^2 + \sigma_\varepsilon^2)}$$

The RE estimator is consistent and fully efficient if the RE model is appropriate. It is inconsistent if the FE model is appropriate, because then correlation between $\mathbf{x}_{it}$ and α_i implies correlation between the regressors and the error in (8.11). Also, if there are no fixed effects but the errors exhibit within-panel correlation, then the RE estimator is consistent but inefficient, and cluster–robust standard errors should be obtained.

The RE estimator uses both between and within variation in the data and has special cases of pooled OLS ($\widehat{\theta}_i = 0$) and within estimation ($\widehat{\theta}_i = 1$). The RE estimator approaches the within estimator as T gets large and as σ_α^2 gets large relative to σ_ε^2, because in those cases $\widehat{\theta}_i \rightarrow 1$.

8.7.2 The xtreg, re command

Three closely related and asymptotically equivalent RE estimators can be obtained by using the `xtreg` command (see section 8.2.4) with the `re`, `mle`, or `pa` option. These estimators use different estimates of the variance components σ_ε^2 and σ_α^2 and hence different estimates $\widehat{\theta}_i$ in the RE regression; see [XT] **xtreg** for the formulas.

The RE estimator uses unbiased estimates of the variance components and is obtained by using the `re` option. The maximum likelihood estimator, under the additional assumption of normally distributed α_i and ε_{it}, is computed by using the `mle` option. The RE model implies the errors are equicorrelated or exchangeable (see section 8.3.10), so `xtreg` with the `pa` and `corr(exchangeable)` options yields asymptotically equivalent results.

For panel data, the RE estimator assumption of equicorrelated errors is usually too strong. At the least, one should use the `vce(cluster id)` option to obtain cluster–robust standard errors. And more-efficient estimates can be obtained with `xtreg, pa` with a better error structure than those obtained with the `corr(exchangeable)` option.

8.7.3 Application of the xtreg, re command

For our data, `xtreg, re` yields

```
. * Random-effects estimator with cluster-robust standard errors
. xtreg lwage exp exp2 wks ed, re vce(cluster id) theta

Random-effects GLS regression                   Number of obs      =      4165
Group variable: id                              Number of groups   =       595

R-sq:  within  = 0.6340                         Obs per group: min =         7
       between = 0.1716                                        avg =       7.0
       overall = 0.1830                                        max =         7

Random effects u_i - Gaussian                   Wald chi2(4)       =   1598.50
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
theta              = .82280511

                                  (Std. Err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0888609   .0039992    22.22   0.000     .0810227    .0966992
        exp2 |  -.0007726   .0000896    -8.62   0.000    -.0009481   -.000597
         wks |   .0009658   .0009259     1.04   0.297    -.000849     .0027806
          ed |   .1117099   .0083954    13.31   0.000     .0952552    .1281647
       _cons |   3.829366   .1333931    28.71   0.000     3.567921    4.090812
-------------+----------------------------------------------------------------
     sigma_u |  .31951859
     sigma_e |  .15220316
         rho |  .81505521   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Unlike the within estimator, the coefficient of the time-invariant regressor ed is now estimated. The standard errors are somewhat smaller than those for the within estimator because some between variation is also used. The entries `sigma_u`, `sigma_e`, and `rho`, and the various R^2 measures, are explained in the next section.

The `re`, `mle`, and `pa corr(exchangeable)` options of `xtreg` yield asymptotically equivalent estimators that differ in typical sample sizes. Comparison for these data is left as an exercise.

(*Continued on next page*)

8.8 Comparison of estimators

Output from xtreg includes estimates of the standard deviation of the error components and R^2 measures that measure within, between, and overall fit. Prediction is possible using the postestimation predict command. We present these estimates before turning to comparison of OLS, between, RE, and within estimators.

8.8.1 Estimates of variance components

Output from the fe, re, and mle options of xtreg includes estimates of the standard deviations of the error components. The combined error in the individual-effects model that we label $\alpha_i + \varepsilon_{it}$ is referred to as $u_i + e_{it}$ in the Stata documentation and output. Thus Stata output sigma_u gives the standard deviation of the individual effect α_i, and sigma_e gives the standard deviation of the idiosyncratic error ε_{it}.

For the RE model estimates given in the previous section, the estimated standard deviation of α_i is twice that of ε_{it}. So the individual-specific component of the error (the random effect) is much more important than the idiosyncratic error.

The output labeled rho equals the intraclass correlation of the error ρ_u defined in (8.4). For the RE model, for example, the estimate of 0.815 is very high. This is expected because, from section 8.3.9, the average autocorrelation of the OLS residuals was computed to be around 0.80.

The theta option, available for the re option in the case of balanced data, reports the estimate $\widehat{\theta}_i = \widehat{\theta}$. Because $\widehat{\theta} = 0.823$, here the RE estimates will be much closer to the within estimates than to the OLS estimates. More generally, in the unbalanced case the matrix e(theta) saves the minimum, 5th percentile, median, 95th percentile, and maximum of $\widehat{\theta}_1, \ldots, \widehat{\theta}_N$.

8.8.2 Within and between R-squared

The table header from xtreg provides three R^2 measures, computed using the interpretation of R^2 as the squared correlation between the actual and fitted values of the dependent variable, where the fitted values ignore the contribution of $\widehat{\alpha}_i$.

Let $\widehat{\alpha}$ and $\widehat{\beta}$ be estimates obtained by one of the xtreg options (be, fe, or re). Let $\rho^2(x, y)$ denote the squared correlation between x and y. Then

$$\begin{array}{ll} \text{Within } R^2\text{:} & \rho^2\{(y_{it} - \overline{y}_i), (\mathbf{x}_{it}'\widehat{\beta} - \overline{\mathbf{x}}_i'\widehat{\beta})\} \\ \text{Between } R^2\text{:} & \rho^2(\overline{y}_i, \overline{\mathbf{x}}_i'\widehat{\beta}) \\ \text{Overall } R^2\text{:} & \rho^2(y_{it}, \mathbf{x}_{it}'\widehat{\beta}) \end{array}$$

The three R^2 measures are, respectively, 0.66, 0.03, and 0.05 for the within estimator; 0.14, 0.33, and 0.27 for the between estimator; and 0.63, 0.17, and 0.18 for the RE estimator. So the within estimator best explains the within variation ($R_{\mathrm{w}}^2 = 0.66$), and the between estimator best explains the between variation ($R_{\mathrm{b}}^2 = 0.33$). The within estimator has a low $R_{\mathrm{o}}^2 = 0.05$ and a much higher $R^2 = 0.91$ in section 8.5.4, because R_{o}^2 neglects $\widehat{\alpha}_i$.

8.8.3 Estimator comparison

We compare some of the panel estimators and associated standard errors, variance components estimates, and R^2. Pooled OLS is the same as the xtreg command with the corr(independent) and pa options. We have

```
. * Compare OLS, BE, FE, RE estimators, and methods to compute standard errors
. global xlist exp exp2 wks ed
. quietly regress lwage $xlist, vce(cluster id)
. estimates store OLS_rob
. quietly xtreg lwage $xlist, be
. estimates store BE
. quietly xtreg lwage $xlist, fe
. estimates store FE
. quietly xtreg lwage $xlist, fe vce(robust)
. estimates store FE_rob
. quietly xtreg lwage $xlist, re
. estimates store RE
. quietly xtreg lwage $xlist, re vce(robust)
. estimates store RE_rob
. estimates table OLS_rob BE FE FE_rob RE RE_rob,
> b se stats(N r2 r2_o r2_b r2_w sigma_u sigma_e rho) b(%7.4f)

------------------------------------------------------------------------------
    Variable |  OLS_rob      BE          FE        FE_rob        RE       RE_rob
-------------+----------------------------------------------------------------
         exp |   0.0447    0.0382      0.1138      0.1138      0.0889    0.0889
             |   0.0054    0.0057      0.0025      0.0040      0.0028    0.0040
        exp2 |  -0.0007   -0.0006     -0.0004     -0.0004     -0.0008   -0.0008
             |   0.0001    0.0001      0.0001      0.0001      0.0001    0.0001
         wks |   0.0058    0.0131      0.0008      0.0008      0.0010    0.0010
             |   0.0019    0.0041      0.0006      0.0009      0.0007    0.0009
          ed |   0.0760    0.0738   (omitted)   (omitted)      0.1117    0.1117
             |   0.0052    0.0049                              0.0061    0.0084
       _cons |   4.9080    4.6830      4.5964      4.5964      3.8294    3.8294
             |   0.1400    0.2101      0.0389      0.0601      0.0936    0.1334
-------------+----------------------------------------------------------------
           N |     4165      4165        4165        4165        4165      4165
          r2 |   0.2836    0.3264      0.6566      0.6566
        r2_o |             0.2723      0.0476      0.0476      0.1830    0.1830
        r2_b |             0.3264      0.0276      0.0276      0.1716    0.1716
        r2_w |             0.1357      0.6566      0.6566      0.6340    0.6340
     sigma_u |                         1.0362      1.0362      0.3195    0.3195
     sigma_e |                         0.1522      0.1522      0.1522    0.1522
         rho |                         0.9789      0.9789      0.8151    0.8151
------------------------------------------------------------------------------
                                                                 legend: b/se
```

Several features emerge. The estimated coefficients vary considerably across estimators, especially for the time-varying regressors. This reflects quite different results according to whether within variation or between variation is used. The within estimator did not provide a coefficient estimate for the time-invariant regressor **ed** (with the coefficient reported as 0.00). Cluster–robust standard errors for the FE and RE models exceed the default standard errors by one-third to one-half. The various R^2 measures and variance-components estimates also vary considerably across models.

8.8.4 Fixed effects versus random effects

The essential distinction in microeconometrics analysis of panel data is that between FE and RE models. If effects are fixed, then the pooled OLS and RE estimators are inconsistent, and instead the within (or FE) estimator needs to be used. The within estimator is otherwise less desirable, because using only within variation leads to less-efficient estimation and inability to estimate coefficients of time-invariant regressors.

To understand this distinction, consider the scalar regression of y_{it} on x_{it}. Consistency of the pooled OLS estimator requires that $E(u_{it}|x_{it}) = 0$ in the model $y_{it} = \alpha + \beta x_{it} + u_{it}$. If this assumption fails so that x_{it} is endogenous, IV estimation can yield consistent estimates. It can be difficult to find an instrument z_{it} for x_{it} that satisfies $E(u_{it}|z_{it}) = 0$.

Panel data provide an alternative way to obtain consistent estimates. Introduce the individual-effects model $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$. Consistency in this model requires the weaker assumption that $E(\varepsilon_{it}|\alpha_i, x_{it}) = 0$. Essentially, the error has two components: the time-invariant component α_i correlated with regressors that we can eliminate through differencing, and a time-varying component that, given α_i, is uncorrelated with regressors.

The RE model adds an additional assumption to the individual-effects model: α_i is distributed independently of x_{it}. This is a much stronger assumption because it implies that $E(\varepsilon_{it}|\alpha_i, x_{it}) = E(\varepsilon_{it}|x_{it})$, so consistency requires that $E(\varepsilon_{it}|x_{it}) = 0$, as assumed by the pooled OLS model.

For individual-effects models, the fundamental issue is whether the individual effect is correlated with regressors.

8.8.5 Hausman test for fixed effects

Under the null hypothesis that individual effects are random, these estimators should be similar because both are consistent. Under the alternative, these estimators diverge. This juxtaposition is a natural setting for a Hausman test (see section 12.7), comparing FE and RE estimators. The test compares the estimable coefficients of time-varying regressors or can be applied to a key subset of these (often one key regressor).

The hausman command

The **hausman** command implements the standard form of the Hausman test. We have already stored the within estimates as **FE** and the RE estimates as **RE**, so we can immediately implement the test.

For these data, the default version of the **hausman FE RE** command leads to a variance estimate $\{\widehat{V}(\widehat{\beta}_{\rm FE}) - \widehat{V}(\widehat{\beta}_{\rm RE})\}$ that is negative definite, so estimated standard errors of $(\widehat{\beta}_{j,{\rm FE}} - \widehat{\beta}_{j,{\rm RE}})$ cannot be obtained. This problem can arise because different estimates of the error variance are used in forming $\widehat{V}(\widehat{\beta}_{\rm FE})$ and $\widehat{V}(\widehat{\beta}_{\rm RE})$. Similar issues arise for a Hausman test comparing OLS and two-stage least-squares estimates.

It is better to use the **sigmamore** option, which specifies that both covariance matrices are based on the (same) estimated disturbance variance from the efficient estimator. We obtain

```
. * Hausman test assuming RE estimator is fully efficient under null hypothesis
. hausman FE RE, sigmamore
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
         exp |   .1137879     .0888609        .0249269        .0012778
        exp2 |  -.0004244    -.0007726        .0003482        .0000285
         wks |   .0008359     .0009658       -.0001299        .0001108
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =     1513.02
                Prob>chi2 =      0.0000
```

The output from **hausman** provides a nice side-by-side comparison. For the coefficient of regressor **exp**, a test of **RE** against **FE** yields $t = 0.0249/0.00128 = 19.5$, a highly statistically significant difference. And the overall statistic, here $\chi^2(3)$, has $p = 0.000$. This leads to strong rejection of the null hypothesis that RE provides consistent estimates.

Robust Hausman test

A serious shortcoming of the standard Hausman test is that it requires the RE estimator to be efficient. This in turn requires that the α_i and ε_{it} are i.i.d., an invalid assumption if cluster–robust standard errors for the RE estimator differ substantially from default standard errors. For our data example, and in many applications, a robust version of the Hausman test is needed. There is no Stata command for this. A panel bootstrap Hausman test can be conducted, using an adaptation of the bootstrap Hausman test example in section 13.4.6.

Simpler is to test $H_0: \gamma = 0$ in the auxiliary OLS regression

$$(y_{it} - \widehat{\theta}\overline{y}_i) = (1 - \widehat{\theta})\alpha + (\mathbf{x}_{it} - \widehat{\theta}\overline{\mathbf{x}}_i)'\boldsymbol{\beta} + (\mathbf{x}_{1it} - \overline{\mathbf{x}}_{1i})'\boldsymbol{\gamma} + v_{it}$$

where $\mathbf{x}_1$ denotes only time-varying regressors. A Wald test of $\boldsymbol{\gamma} = \mathbf{0}$ can be shown to be asymptotically equivalent to the standard test when the RE estimator is fully efficient under H_0 and is numerically equivalent to **hausman** with the **sigmaless** option. A summary of related tests for fixed versus random effects is given in Baltagi (2008, 72–78).

In the more likely case that the RE estimator is not fully efficient, Wooldridge (2002) proposes performing the Wald test using cluster–robust standard errors. To implement this test in Stata, we need to generate the RE differences $y_{it} - \widehat{\theta}\overline{y}_i$ and $\mathbf{x}_{it} - \widehat{\theta}\overline{\mathbf{x}}_i$, and the mean-differences $\mathbf{x}_{1it} - \overline{\mathbf{x}}_{1i}$.

```
. * Robust Hausman test using method of Wooldridge (2002)
. quietly xtreg lwage $xlist, re
. scalar theta = e(theta)
. global yandxforhausman lwage exp exp2 wks ed
. sort id
. foreach x of varlist $yandxforhausman {
  2.    by id: egen mean`x´ = mean(`x´)
  3.    generate md`x´ = `x´ - mean`x´
  4.    generate red`x´ = `x´ - theta*mean`x´
  5.    }
. quietly regress redlwage redexp redexp2 redwks reded mdexp mdexp2 mdwks,
> vce(cluster id)
. test mdexp mdexp2 mdwks
 ( 1)  mdexp = 0
 ( 2)  mdexp2 = 0
 ( 3)  mdwks = 0
       F(  3,   594) =  597.47
            Prob > F =    0.0000
```

The test strongly rejects the null hypothesis, and we conclude that the RE model is not appropriate. The code will become more complex in the unbalanced case, because we then need to compute $\widehat{\theta}_i$ for each observation. The user-written command `xtoverid` following command `xtreg, re vce(cluster id)` implements the preceding test in both balanced and unbalanced settings.

8.8.6 Prediction

The postestimation `predict` command after `xtreg` provides estimated residuals and fitted values following estimation of the individual-effects model $y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$.

The estimated individual-specific error $\widehat{\alpha}_i = \overline{y}_i - \overline{\mathbf{x}}_{it}'\widehat{\boldsymbol{\beta}}$ is obtained by using the `u` option; the estimated idiosyncratic error $\widehat{\varepsilon}_{it} = y_{it} - \widehat{\alpha}_i - \mathbf{x}_{it}'\widehat{\boldsymbol{\beta}}$ is obtained by using the `e` option; and the `ue` option gives $\widehat{\alpha}_i + \widehat{\varepsilon}_{it}$.

Fitted values of the dependent variable differ according to whether the estimated individual-specific error is used. The fitted value $y_{it} = \widehat{\alpha} + \mathbf{x}_{it}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\alpha} = N^{-1}\sum_i \widehat{\alpha}_i$, is obtained by using the `xb` option. The fitted value $y_{it} = \widehat{\alpha}_i + \mathbf{x}_{it}'\widehat{\boldsymbol{\beta}}$ is obtained by using the `xbu` option.

As an example, we contrast OLS and RE in-sample fitted values.

```
. * Prediction after OLS and RE estimation
. quietly regress lwage exp exp2 wks ed, vce(cluster id)
. predict xbols, xb
. quietly xtreg lwage exp exp2 wks ed, re
. predict xbre, xb
. predict xbure, xbu
. summarize lwage xbols xbre xbure
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       lwage |      4165    6.676346    .4615122    4.60517      8.537
       xbols |      4165    6.676346    .2457572   5.850037   7.200861
        xbre |      4165    6.676346    .6205324   5.028067    8.22958
       xbure |      4165    6.676346    .4082951    5.29993   7.968179
. correlate lwage xbols xbre xbure
(obs=4165)
             |    lwage    xbols     xbre    xbure
-------------+------------------------------------
       lwage |   1.0000
       xbols |   0.5325   1.0000
        xbre |   0.4278   0.8034   1.0000
       xbure |   0.9375   0.6019   0.4836   1.0000
```

The RE prediction $\widehat{\alpha} + \mathbf{x}_{it}'\widehat{\boldsymbol{\beta}}$ is not as highly correlated with `lwage` as is the OLS prediction (0.43 versus 0.53), which was expected because the OLS estimator maximizes this correlation.

When instead we use $\widehat{\alpha}_i + \mathbf{x}_{it}'\widehat{\boldsymbol{\beta}}$ so the fitted individual effect is included, the correlation of the prediction with `lwage` increases greatly to 0.94. In a short panel, however, these predictions are not consistent because each individual prediction $\widehat{\alpha}_i = \overline{y}_i - \overline{\mathbf{x}}_{it}'\widehat{\boldsymbol{\beta}}$ is based on only T observations and $T \nrightarrow \infty$.

8.9 First-difference estimator

Consistent estimation of $\boldsymbol{\beta}$ in the FE model requires eliminating the α_i. One way to do so is to mean-difference, yielding the within estimator. An alternative way is to first-difference, leading to the first-difference estimator. This alternative has the advantage of relying on weaker exogeneity assumptions, explained below, that become important in dynamic models presented in the next chapter. In the current chapter, the within estimator is traditionally favored as it is the more efficient estimator if the ε_{it} are i.i.d.

8.9.1 First-difference estimator

The first-difference (FD) estimator is obtained by performing OLS on the first-differenced variables

$$(y_{it} - y_{i,t-1}) = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}) \tag{8.12}$$

First-differencing has eliminated α_i, so OLS estimation of this model leads to consistent estimates of $\boldsymbol{\beta}$ in the FE model. The coefficients of time-invariant regressors are not identified, because then $x_{it} - x_{i,t-1} = 0$, as was the case for the within estimator.

The FD estimator is not provided as an option to xtreg. Instead, the estimator can be computed by using regress and Stata time-series operators to compute the first-differences. We have

```
. sort id t
. * First-differences estimator with cluster-robust standard errors
. regress D.(lwage exp exp2 wks ed), vce(cluster id) noconstant
note: _delete omitted because of collinearity
Linear regression                                      Number of obs =     3570
                                                       F(  3,    594) = 1035.19
                                                       Prob > F      =   0.0000
                                                       R-squared     =   0.2209
                                                       Root MSE      =   .18156
                                        (Std. Err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
     D.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |
         D1. |   .1170654   .0040974    28.57   0.000     .1090182    .1251126
             |
        exp2 |
         D1. |  -.0005321   .0000808    -6.58   0.000    -.0006908   -.0003734
             |
         wks |
         D1. |  -.0002683   .0011783    -0.23   0.820    -.0025824    .0020459
             |
          ed |
         D1. |  (omitted)
------------------------------------------------------------------------------
```

Note that the noconstant option is used. If instead an intercept is included in (8.12), say, δ, this would imply that the original model had a time trend because $\delta t - \delta(t-1) = \delta$.

As expected, the coefficient for education is not identified because ed here is time-invariant. The coefficient for wks actually changes sign compared with the other estimators, though it is highly statistically insignificant.

The FD estimator, like the within estimator, provides consistent estimators when the individual effects are fixed. For panels with $T = 2$, the FD and within estimators are equivalent; otherwise, the two differ. For static models, the FE model is used because it is the efficient estimator if the idiosyncratic error ε_{it} is i.i.d.

The FD estimator seemingly uses one less year of data compared with the within estimator, because the FD output lists 3,570 observations rather than 4,165. This, however, is misleading. Using the LSDV interpretation of the within estimator, the within estimator essentially loses 595 observations by estimating the T fixed effects $\alpha_1, \ldots, \alpha_T$.

8.9.2 Strict and weak exogeneity

From (8.6), the within estimator requires that $\varepsilon_{it} - \overline{\varepsilon}_i$ be uncorrelated with $\mathbf{x}_{it} - \overline{\mathbf{x}}_i$. This is the case under the assumption of strict exogeneity or strong exogeneity that

$$E(\varepsilon_{it}|\alpha_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \ldots, \mathbf{x}_{iT}) = 0$$

From (8.12), the FD estimator requires that $\varepsilon_{it} - \varepsilon_{i,t-1}$ be uncorrelated with $\mathbf{x}_{it} - \mathbf{x}_{i,t-1}$. This is the case under the assumption of weak exogeneity that

$$E(\varepsilon_{it}|\alpha_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}) = 0$$

This is a considerably weaker assumption because it permits future values of the regressors to be correlated with the error, as will be the case if the regressor is a lagged dependent variable.

As long as there is no feedback from the idiosyncratic shock today to a covariate tomorrow, this distinction is unnecessary when estimating static models. It becomes important for dynamic models (see section 9.4), because then strict exogeneity no longer holds and we turn to the FD estimator.

8.10 Long panels

The methods up to this point have focused on short panels. Now we consider long panels with many time periods for relatively few individuals (N is small and $T \rightarrow \infty$). Examples are data on a few regions, firms, or industries followed for many time periods.

Then individual fixed effects, if desired, can be easily handled by including dummy variables for each individual as regressors. Instead, the focus is on more-efficient GLS estimation under richer models of the error process than those specified in the short-panel case. Here we consider only methods for stationary errors, and we only briefly cover the growing area of panel data with unit roots and cointegration.

8.10.1 Long-panel dataset

The dataset used is a U.S. state–year panel from Baltagi, Griffin, and Xiong (2000) on annual cigarette consumption and price for U.S. states over 30 years. The ultimate goal is to measure the responsiveness of per capita cigarette consumption to real cigarette prices. Price varies across states, due in large part to different levels of taxation, as well as over time.

The original data were for $N = 46$ states and $T = 30$, and it is not clear whether we should treat $N \to \infty$, as we have done to date, or $T \to \infty$, or both. This situation is not unusual for a panel that uses aggregated regional data over time. To make explicit that we are considering $T \to \infty$, we use data from only $N = 10$ states, similar to many countries where there may be around 10 major regions (states or provinces).

The `mus08cigar.dta` dataset has the following data:

```
. * Description of cigarette dataset
. use mus08cigar.dta, clear
. describe
Contains data from mus08cigar.dta
  obs:           300
 vars:             6                          13 Mar 2008 20:45
 size:         8,400 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
state           float  %9.0g                  U.S. state
year            float  %9.0g                  Year 1963 to 1992
lnp             float  %9.0g                  Log state real price of pack of
                                                cigarettes
lnpmin          float  %9.0g                  Log of min real price in
                                                adjoining states
lnc             float  %9.0g                  Log state cigarette sales in
                                                packs per capita
lny             float  %9.0g                  Log state per capita disposable
                                                income
-------------------------------------------------------------------------------
Sorted by:
```

There are 300 observations, so each state–year pair is a separate observation because 10 × 30 = 300. The quantity demanded (`lnc`) will depend on price (`lnp`), price of a substitute (`lnpmin`), and income (`lny`).

Descriptive statistics can be obtained by using `summarize`:

```
. * Summary of cigarette dataset
. summarize, separator(6)
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       state |       300         5.5     2.87708          1         10
        year |       300        77.5    8.669903         63         92
         lnp |       300    4.518424    .1406979   4.176332    4.96916
      lnpmin |       300      4.4308    .1379243     4.0428   4.831303
         lnc |       300    4.792591    .2071792   4.212128   5.690022
         lny |       300    8.731014    .6942426   7.300023    10.0385
```

The variables `state` and `year` have the expected ranges. The variability in per capita cigarette sales (`lnc`) is actually greater than the variability in price (`lnp`), with respective standard deviations of 0.21 and 0.14. All variables are observed for all 300 observations, so the panel is indeed balanced.

8.10.2 Pooled OLS and PFGLS

A natural starting point is the two-way–effects model $y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}$. When the panel has few individuals relative to the number of periods, the individual effects α_i (here state effects) can be incorporated into $\mathbf{x}_{it}$ as dummy-variable regressors. Then there are too many time effects γ_t (here year effects). Rather than trying to control for these in ways analogous to the use of `xtreg` in the short-panel case, it is usually sufficient to take advantage of the natural ordering of time (as opposed to individuals) and simply include a linear or quadratic trend in time.

We therefore focus on the pooled model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it}, \quad i = 1, \ldots, N, \ t = 1, \ldots, T \tag{8.13}$$

where the regressors $\mathbf{x}_{it}$ include an intercept, often time and possibly time-squared, and possibly a set of individual indicator variables. We assume that N is quite small relative to T.

We consider pooled OLS and PFGLS of this model under a variety of assumptions about the error u_{it}. In the short-panel case, it was possible to obtain standard errors that control for serial correlation in the error without explicitly stating a model for serial correlation. Instead, we could use cluster–robust standard errors, given a small T and $N \to \infty$. Now, however, T is large relative to N, and it is necessary to specify a model for serial correlation in the error. Also given that N is small, it is possible to relax the assumption that u_{it} is independent over i.

8.10.3 The xtpcse and xtgls commands

The `xtpcse` and `xtgls` commands are more suited than `xtgee` for pooled OLS and GLS when data are from a long panel. They allow the error u_{it} in the model to be correlated over i, allow the use of an AR(1) model for u_{it} over t, and allow u_{it} to be heteroskedastic. At the greatest level of generality,

$$u_{it} = \rho_i u_{i,t-1} + \varepsilon_{it} \tag{8.14}$$

where ε_{it} are serially uncorrelated but are correlated over i with $\mathrm{Cor}(\varepsilon_{it}, \varepsilon_{is}) = \sigma_{ts}$.

The `xtpcse` command yields (long) panel-corrected standard errors for the pooled OLS estimator, as well as for a pooled least-squares estimator with an AR(1) model for u_{it}. The syntax is

`xtpcse` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

The `correlation()` option determines the type of pooled estimator. Pooled OLS is obtained by using `correlation(independent)`. The pooled AR(1) estimator with general ρ_i is obtained by using `correlation(psar1)`. With a balanced panel, $y_{it} - \widehat{\rho}_i y_{it,t-1}$ is regressed on $\mathbf{x}^*_{it} = \mathbf{x}_{it} - \widehat{\rho}\mathbf{x}_{it,t-1}$ for $t > 1$, whereas $\sqrt{(1-\widehat{\rho}_i)^2} y_{i1}$ is regressed on $\sqrt{(1-\widehat{\rho}_i)^2}\mathbf{x}_{i1}$ for $t = 1$. The pooled estimator with AR(1) error and $\rho_i = \rho$ is obtained by using `correlation(ar1)`. Then $\widehat{\rho}$, calculated as the average of the $\widehat{\rho}_i$, is used.

In all cases, panel-corrected standard errors that allow heteroskedasticity and correlation over i are reported, unless the `hetonly` option is used, in which case independence over i is assumed, or the `independent` option is used, in which case ε_{it} is i.i.d.

The `xtgls` command goes further and obtains PFGLS estimates and associated standard errors assuming the model for the errors is the correct model. The estimators are more efficient asymptotically than those from `xtpcse`, if the model is correctly specified. The command has the usual syntax:

`xtgls` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

The `panels()` option specifies the error correlation across individuals, where for our data an individual is a state. The `panels(iid)` option specifies u_{it} to be i.i.d., in which case the pooled OLS estimator is obtained. The `panels(heteroskedastic)` option specifies u_{it} to be independent with a variance of $E(u_{it}^2) = \sigma_i^2$ that can be different for each individual. Because there are many observations for each individual, σ_i^2 can be consistently estimated. The `panels(correlated)` option additionally allows correlation across individuals, with independence over time for a given individual, so that $E(u_{it}u_{jt}) = \sigma_{ij}$. This option requires that $T > N$.

The `corr()` option specifies the serial correlation of errors for each individual state. The `corr(independent)` option specifies u_{it} to be serially uncorrelated. The `corr(ar1)` option permits AR(1) autocorrelation of the error with $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, where ε_{it} is i.i.d. The `corr(psar1)` option relaxes the assumption of a common AR(1) parameter to allow $u_{it} = \rho_i u_{i,t-1} + \varepsilon_{it}$. The `rhotype()` option provides various methods to compute this AR(1) parameter(s). The default estimator is two-step FGLS, whereas the `igls` option uses iterated FGLS. The `force` option enables estimation even if observations are unequally spaced over time.

Additionally, we illustrate the user-written `xtscc` command (Hoechle 2007). This generalizes `xtpcse` by applying the method of Driscoll and Kraay (1998) to obtain Newey–West-type standard errors that allow autocorrelated errors of general form, rather than restricting errors to be AR(1). Error correlation across panels, often called spatial correlation, is assumed. The error is allowed to be serially correlated for m lags. The default is for the program to determine m. Alternatively, m can be specified using the `lags(`*m*`)` option.

8.10.4 Application of the xtgls, xtpcse, and xtscc commands

As an example, we begin with a PFGLS estimator that uses the most flexible model for the error u_{it}, with flexible correlation across states and a distinct AR(1) process for the error in each state. In principle, this is the best estimator to use, but in practice when T is not much larger than N, there can be finite-sample bias in the estimators and standard errors; see Beck and Katz (1995). Then it is best, at the least, to use the more restrictive `corr(ar1)` rather than `corr(psar1)`.

We obtain

```
. * Pooled GLS with error correlated across states and state-specific AR(1)
. xtset state year
       panel variable:  state (strongly balanced)
        time variable:  year, 63 to 92
                delta:  1 unit

. xtgls lnc lnp lny lnpmin year, panels(correlated) corr(psar1)

Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic with cross-sectional correlation
Correlation:   panel-specific AR(1)

Estimated covariances      =        55          Number of obs      =       300
Estimated autocorrelations =        10          Number of groups   =        10
Estimated coefficients     =         5          Time periods       =        30
                                                Wald chi2(4)       =    342.15
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         lnc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |  -.3260683   .0218214   -14.94   0.000    -.3688375   -.2832991
         lny |   .4646236   .0645149     7.20   0.000     .3381768    .5910704
      lnpmin |   .0174759   .0274963     0.64   0.525    -.0364159    .0713677
        year |  -.0397666   .0052431    -7.58   0.000    -.0500429   -.0294902
       _cons |   5.157994   .2753002    18.74   0.000     4.618416    5.697573
------------------------------------------------------------------------------
```

All regressors have the expected effects. The estimated price elasticity of demand for cigarettes is −0.326, the income elasticity is 0.465, demand declines by 4% per year (the coefficient of `year` is a semielasticity because the dependent variable is in logs), and a higher minimum price in adjoining states increases demand in the current state. There are 10 states, so there are $10 \times 11/2 = 55$ unique entries in the 10×10 contemporaneous error covariance matrix, and 10 autocorrelation parameters ρ_i are estimated.

We now use `xtpcse`, `xtgls`, and user-written `xtscc` to obtain the following pooled estimators and associated standard errors: 1) pooled OLS with i.i.d. errors; 2) pooled OLS with standard errors assuming correlation over states; 3) pooled OLS with standard errors assuming general serial correlation in the error (to four lags) and correlation over states; 4) pooled OLS that assumes an AR(1) error and then gets standard errors that additionally permit correlation over states; 5) PFGLS with standard errors assuming an AR(1) error; and 6) PFGLS assuming an AR(1) error and correlation across states. In all cases of AR(1) error, we specialize to $\rho_i = \rho$.

```
. * Comparison of various pooled OLS and GLS estimators
. quietly xtpcse lnc lnp lny lnpmin year, corr(ind) independent nmk
. estimates store OLS_iid
. quietly xtpcse lnc lnp lny lnpmin year, corr(ind)
. estimates store OLS_cor
. quietly xtscc lnc lnp lny lnpmin year, lag(4)
. estimates store OLS_DK
. quietly xtpcse lnc lnp lny lnpmin year, corr(ar1)
. estimates store AR1_cor
. quietly xtgls lnc lnp lny lnpmin year, corr(ar1) panels(iid)
. estimates store FGLSAR1
. quietly xtgls lnc lnp lny lnpmin year, corr(ar1) panels(correlated)
. estimates store FGLSCAR
. estimates table OLS_iid OLS_cor OLS_DK AR1_cor FGLSAR1 FGLSCAR, b(%7.3f) se

--------------------------------------------------------------------------
    Variable |  OLS_iid    OLS_cor    OLS_DK     AR1_cor    FGLSAR1   FGLSCAR
-------------+------------------------------------------------------------
         lnp |   -0.583     -0.583     -0.583     -0.266     -0.264    -0.330
             |    0.129      0.169      0.273      0.049      0.049     0.026
         lny |    0.365      0.365      0.365      0.398      0.397     0.407
             |    0.049      0.080      0.163      0.125      0.094     0.080
      lnpmin |   -0.027     -0.027     -0.027      0.069      0.070     0.036
             |    0.128      0.166      0.252      0.064      0.059     0.034
        year |   -0.033     -0.033     -0.033     -0.038     -0.038    -0.037
             |    0.004      0.006      0.012      0.010      0.007     0.006
       _cons |    6.930      6.930      6.930      5.115      5.100     5.393
             |    0.353      0.330      0.515      0.544      0.414     0.361
--------------------------------------------------------------------------
                                                               legend: b/se
```

For pooled OLS with i.i.d. errors, the **nmk** option normalizes the VCE by $N-k$ rather than N, so that output is exactly the same as that from **regress** with default standard errors. The same results could be obtained by using **xtgls** with the **corr(ind) panel(iid) nmk** options. Allowing correlation across states increases OLS standard errors by 30–50%. Additionally, allowing for serial correlation (OLS_DK) leads to another 50–100% increase in the standard errors. The fourth and fifth estimators control for at least an AR(1) error and yield roughly similar coefficients and standard errors. The final column results are similar to those given at the start of this section, where we used the more flexible **corr(psar1)** rather than **corr(ar1)**.

8.10.5 Separate regressions

The pooled regression specifies the same regression model for all individuals in all years. Instead, we could have a separate regression model for each individual unit:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + u_{it}$$

This model has NK parameters, so inference is easiest for a long panel with a small N.

For example, suppose for the cigarette example we want to fit separate regressions for each state. Separate OLS regressions for each state can be obtained by using the **statsby** prefix with the **by(state)** option. We have

```
. * Run separate regressions for each state
. statsby, by(state) clear: regress lnc lnp lny lnpmin year
(running regress on estimation sample)

      command:  regress lnc lnp lny lnpmin year
           by:  state

Statsby groups
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........
```

This leads to a dataset with 10 observations on **state** and the five regression coefficients. We have

```
. * Report regression coefficients for each state
. format _b* %9.2f
. list, clean

       state     _b_lnp     _b_lny   _b_lnp~n    _b_year    _b_cons
  1.       1      -0.36       1.10       0.24      -0.08       2.10
  2.       2       0.12       0.60      -0.45      -0.05       5.14
  3.       3      -0.20       0.76       0.12      -0.05       2.72
  4.       4      -0.52      -0.14      -0.21      -0.00       9.56
  5.       5      -0.55       0.71       0.30      -0.07       4.76
  6.       6      -0.11       0.21      -0.14      -0.02       6.20
  7.       7      -0.43      -0.07       0.18      -0.03       9.14
  8.       8      -0.26       0.89       0.08      -0.07       3.67
  9.       9      -0.03       0.55      -0.36      -0.04       4.69
 10.      10      -1.41       1.12       1.14      -0.08       2.70
```

In all states except one, sales decline as price rises, and in most states, sales increase with income.

One can also test for poolability, meaning to test whether the parameters are the same across states. In this example, there are $5 \times 10 = 50$ parameters in the unrestricted model and 5 in the restricted pooled model, so there are 45 parameters to test.

8.10.6 FE and RE models

As noted earlier, if there are few individuals and many time periods, individual-specific FE models can be fit with the LSDV approach of including a set of dummy variables, here for each time period (rather than for each individual as in the short-panel case).

Alternatively, one can use the **xtregar** command. This model is the individual-effects model $y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}$, with AR(1) error $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$. This is a better model of the error than the i.i.d. error model $u_{it} = \varepsilon_{it}$ assumed in **xtreg**, so **xtregar** potentially will lead to more-efficient parameter estimates.

The syntax of **xtregar** is similar to that for **xtreg**. The two key options are **fe** and **re**. The **fe** option treats α_i as a fixed effect. Given an estimate of $\widehat{\rho}$, we first transform to eliminate the effect of the AR(1) error, as described after (8.14), and then transform again (mean-difference) to eliminate the individual effect. The **re** option treats α_i as a random effect.

We compare pooled OLS estimates, RE estimates using `xtreg` and `xtregar`, and within estimates using `xtreg`, `xtregar`, and `xtscc`. Recall that `xtscc` calculates either the OLS or regular within estimator but then estimates the VCE assuming quite general error correlation over time and across states. We have

```
. * Comparison of various RE and FE estimators
. use mus08cigar.dta, clear
. quietly xtscc lnc lnp lny lnpmin, lag(4)
. estimates store OLS_DK
. quietly xtreg lnc lnp lny lnpmin, fe
. estimates store FE_REG
. quietly xtreg lnc lnp lny lnpmin, re
. estimates store RE_REG
. quietly xtregar lnc lnp lny lnpmin, fe
. estimates store FE_REGAR
. quietly xtregar lnc lnp lny lnpmin, re
. estimates store RE_REGAR
. quietly xtscc lnc lnp lny lnpmin, fe lag(4)
. estimates store FE_DK
. estimates table OLS_DK FE_REG RE_REG FE_REGAR RE_REGAR FE_DK, b(%7.3f) se

------------------------------------------------------------------------------
    Variable |  OLS_DK     FE_REG     RE_REG    FE_RE~R    RE_RE~R     FE_DK
-------------+----------------------------------------------------------------
         lnp |  -0.611     -1.136     -1.110     -0.260     -0.282     -1.136
             |   0.428      0.101      0.102      0.049      0.052      0.158
         lny |  -0.027     -0.046     -0.045     -0.066     -0.074     -0.046
             |   0.026      0.011      0.011      0.064      0.026      0.020
      lnpmin |  -0.129      0.421      0.394     -0.010     -0.004      0.421
             |   0.338      0.101      0.102      0.057      0.060      0.168
       _cons |   8.357      8.462      8.459      6.537      6.708      8.462
             |   0.633      0.241      0.247      0.036      0.289      0.464
------------------------------------------------------------------------------
                                                                legend: b/se
```

There are three distinctly different sets of coefficient estimates: those using pooled OLS, those using `xtreg` to obtain FE and RE estimators, and those using `xtregar` to obtain FE and RE estimators. The final set of estimates uses the `fe` option of the user-written `xtscc` command. This produces the standard within estimator but then finds standard errors that are robust to both spatial (across panels) and serial autocorrelation of the error.

8.10.7 Unit roots and cointegration

Panel methods for unit roots and cointegration are based on methods developed for a single time series and assume that $T \to \infty$. We now consider their application to panel data, a currently active area of research.

If N is small, say $N < 10$, then seemingly unrelated equations methods can be used. When N is large, the panel aspect becomes more important. Complications include the need to control for cross-section unobserved heterogeneity when N is large, asymptotic theory that can vary with exactly how N and T both go to infinity, and the possibility of cross-section dependence. At the same time, statistics that have nonnormal distributions for a single time series can be averaged over cross sections to obtain statistics with a normal distribution.

Unit-root tests can have low power. Panel data may increase the power because of now having time series for several cross sections. The unit-root tests can also be of interest per se, such as testing purchasing power parity, as well as being relevant to consequent considerations of cointegration. A dynamic model with cross-section heterogeneity is

$$y_{it} = \rho_i y_{i,t-1} + \phi_{i1}\Delta y_{i,t-1} + \cdots + \phi_{ip_i}\Delta y_{i,t-p_i} + \mathbf{z}_{it}'\boldsymbol{\gamma}_i + u_{it}$$

where lagged changes are introduced so that u_{it} is i.i.d. Examples of $\mathbf{z}_{it}$ include individual effects $[\mathbf{z}_{it} = (1)]$, individual effects and individual time trends $[\mathbf{z}_{it} = (1\ t)']$, and $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}$ in the case of homogeneity. A unit-root test is a test of $H_0: \rho_1 = \cdots \rho_N = 1$. Levin, C.-F. Lin, and C.-S. J. Chu (2002) proposed a test against the alternative of homogeneity, $H_a: \rho_1 = \cdots = \rho_N = \rho < 1$, that is based on pooled OLS estimation using specific first-step pooled residuals, where in both steps homogeneity ($\rho_i = \rho$ and $\phi_{ik} = \phi_k$) is imposed. The user-written `levinlin` command (Bornhorst and Baum 2006) performs this test. Im, Pesaran, and Shin (2003) instead test against an alternative of heterogeneity, $H_a: \rho_1 < 1, \ldots, \rho_{N_0} < 1$, for a fraction N_0/N of the ρ_i by averaging separate augmented Dickey–Fuller tests for each cross section. The user-written `ipshin` command (Bornhorst and Baum 2007) implements this test. Both test statistics are asymptotically normal and both assume $N/T \to 0$ so that the time-series dimension dominates the cross-section dimension.

As in the case of a single time series, cointegration tests are used to ensure that statistical relationships between trending variables are not spurious. A quite general cointegrated panel model is

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta}_i + \mathbf{z}_{it}'\boldsymbol{\gamma}_i + u_{it} \\ \mathbf{x}_{it} &= \mathbf{x}_{i,t-1} + \boldsymbol{\varepsilon}_{it} \end{aligned}$$

where $\mathbf{z}_{it}$ is deterministic and can include individual effects and time trends, and $\mathbf{x}_{it}$ are (co)integrated regressors. Most tests of cointegration are based on the OLS residuals $\widehat{u}_{it}$, but the unit-root tests cannot be directly applied if $\text{Cov}(u_{it}, \boldsymbol{\varepsilon}_{it}) \neq 0$, as is likely. Single-equation estimators have been proposed that generalize to panels fully modified OLS and dynamic OLS, and Johanssen's system approach has also been generalized to panels. The user-written `xtpmg` command (Blackburne and Frank 2007) implements the estimators of Pesaran and Smith (1995) and Pesaran, Shin, and Smith (1999) for nonstationary heterogeneous panels with a large N and T. For references, see Baltagi (2008) and Breitung and Pesaran (2005).

8.11 Panel-data management

Stata xt commands require panel data to be in long form, meaning that each individual–time pair is a separate observation. Some datasets instead store panel data in wide form, which has the advantage of using less space. Sometimes the observational unit is the individual, and a single observation has all time periods for that individual. And sometimes the observational unit is a time period, and a single observation has all individuals for that time period.

We illustrate how to move from wide form to long form and vice versa by using the reshape command. Our example is for panel data, but reshape can also be used in other contexts where data are grouped, such as clustered data grouped by village rather than panel data grouped by time.

8.11.1 Wide-form data

We consider a dataset that is originally in wide form, with each observation containing all years of data for an individual. The dataset is a subset of the data from the previous section. Each observation is a state and has all years of data for that state. We have

```
. * Wide form data (observation is a state)
. use mus08cigarwide.dta, clear

. list, clean
       state   lnp63   lnc63   lnp64   lnc64   lnp65   lnc65
  1.       1     4.5     4.5     4.6     4.6     4.5     4.6
  2.       2     4.4     4.8     4.3     4.8     4.3     4.8
  3.       3     4.5     4.6     4.5     4.6     4.5     4.6
  4.       4     4.4     5.0     4.4     4.9     4.4     4.9
  5.       5     4.5     5.1     4.5     5.0     4.5     5.0
  6.       6     4.5     5.1     4.5     5.1     4.5     5.1
  7.       7     4.3     5.5     4.3     5.5     4.3     5.5
  8.       8     4.5     4.9     4.6     4.8     4.5     4.9
  9.       9     4.5     4.7     4.5     4.7     4.6     4.6
 10.      10     4.5     4.6     4.6     4.5     4.5     4.6
```

The data contain a state identifier, state; three years of data on log price, lnp63–lnp65; and three years of data on log sales, lnc63–lnc65. The data are for 10 states.

8.11.2 Convert wide form to long form

The data can be converted from wide form to long form by using reshape long. The desired dataset will have an observation as a state–year pair. The variables should be a state identifier, a year identifier, and the current state–year observations on lnp and lnc.

The simple command reshape long actually does this automatically, because it interprets the suffixes 63–65 as denoting the grouping that needs to be expanded to long form. We use a more detailed version of the command that spells out exactly what we want to do and leads to exactly the same result as reshape long without arguments. We have

```
. * Convert from wide form to long form (observation is a state-year pair)
. reshape long lnp lnc, i(state) j(year)
(note: j = 63 64 65)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       10   ->      30
Number of variables                   7   ->       4
j variable (3 values)                     ->   year
xij variables:
                      lnp63 lnp64 lnp65   ->   lnp
                      lnc63 lnc64 lnc65   ->   lnc
-----------------------------------------------------------------------------
```

The output indicates that we have expanded the dataset from 10 observations (10 states) to 30 observations (30 state–year pairs). A year-identifier variable, year, has been created. The wide-form data lnp63–lnp65 have been collapsed to lnp in long form, and lnc63–lnc65 have been collapse to lnc.

We now list the first six observations of the new long-form data.

```
. * Long-form data (observation is a state)
. list in 1/6, sepby(state)

     +----------------------------+
     | state   year   lnp   lnc   |
     |----------------------------|
  1. |     1     63   4.5   4.5   |
  2. |     1     64   4.6   4.6   |
  3. |     1     65   4.5   4.6   |
     |----------------------------|
  4. |     2     63   4.4   4.8   |
  5. |     2     64   4.3   4.8   |
  6. |     2     65   4.3   4.8   |
     +----------------------------+
```

Any year-invariant variables will also be included in the long-form data. Here the state-identifier variable, state, is the only such variable.

8.11.3 Convert long form to wide form

Going the other way, data can be converted from long form to wide form by using reshape wide. The desired dataset will have an observation as a state. The constructed variables should be a state identifier and observations on lnp and lnc for each of the three years 63–65.

The reshape wide command without arguments actually does this automatically, because it interprets year as the relevant time-identifier and adds suffixes 63–65 to the variables lnp and lnc that are varying with year. We use a more detailed version of the command that spells out exactly what we want to do and leads to exactly the same result. We have

(Continued on next page)

```
. * Reconvert from long form to wide form (observation is a state)
. reshape wide lnp lnc, i(state) j(year)
(note: j = 63 64 65)
Data                                   long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                           30   ->     10
Number of variables                       4   ->      7
j variable (3 values)                  year   ->   (dropped)
xij variables:
                                        lnp   ->   lnp63 lnp64 lnp65
                                        lnc   ->   lnc63 lnc64 lnc65
-----------------------------------------------------------------------------
```

The output indicates that we have collapsed the dataset from 30 observations (30 state–year pairs) to 10 observations (10 states). The `year` variable has been dropped. The long-form data `lnp` has been expanded to `lnp63`–`lnp65` in wide form, and `lnc` has been expanded to `lnc63`–`lnc65`.

A complete listing of the wide form dataset is

```
. list, clean

       state   lnp63   lnc63   lnp64   lnc64   lnp65   lnc65
  1.       1     4.5     4.5     4.6     4.6     4.5     4.6
  2.       2     4.4     4.8     4.3     4.8     4.3     4.8
  3.       3     4.5     4.6     4.5     4.6     4.5     4.6
  4.       4     4.4     5.0     4.4     4.9     4.4     4.9
  5.       5     4.5     5.1     4.5     5.0     4.5     5.0
  6.       6     4.5     5.1     4.5     5.1     4.5     5.1
  7.       7     4.3     5.5     4.3     5.5     4.3     5.5
  8.       8     4.5     4.9     4.6     4.8     4.5     4.9
  9.       9     4.5     4.7     4.5     4.7     4.6     4.6
 10.      10     4.5     4.6     4.6     4.5     4.5     4.6
```

This is exactly the same as the original `mus08cigarwide.dta` dataset, listed in section 8.11.1.

8.11.4 An alternative wide-form data

The wide form we considered had each state as the unit of observation. An alternative is that each year is the observation. Then the preceding commands are reversed so that we have `i(year) j(state)` rather than `i(state) j(year)`.

To demonstrate this case, we first need to create the data in wide form with `year` as the observational unit. We do so by converting the current data, in wide form with `state` as the observational unit, to long form with 30 observations as presented above, and then use `reshape wide` to create wide-form data with `year` as the observational unit.

```
. * Create alternative wide-form data (observation is a year)
. quietly reshape long lnp lnc, i(state) j(year)

. reshape wide lnp lnc, i(year) j(state)
(note: j = 1 2 3 4 5 6 7 8 9 10)
Data                                   long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                           30   ->      3
Number of variables                       4   ->     21
j variable (10 values)                state   ->   (dropped)
xij variables:
                                        lnp   ->   lnp1 lnp2 ... lnp10
                                        lnc   ->   lnc1 lnc2 ... lnc10
-----------------------------------------------------------------------------

. list year lnp1 lnp2 lnc1 lnc2, clean

       year   lnp1   lnp2   lnc1   lnc2
  1.     63    4.5    4.4    4.5    4.8
  2.     64    4.6    4.3    4.6    4.8
  3.     65    4.5    4.3    4.6    4.8
```

The wide form has 3 observations (one per year) and 21 variables (`lnp` and `lnc` for each of 10 states plus `year`).

We now have data in wide form with `year` as the observational unit. To use `xt` commands, we use `reshape long` to convert to long-form data with an observation for each state–year pair. We have

```
. * Convert from wide form (observation is year) to long form (year-state)
. reshape long lnp lnc, i(year) j(state)
(note: j = 1 2 3 4 5 6 7 8 9 10)
Data                                   wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                            3   ->     30
Number of variables                      21   ->      4
j variable (10 values)                        ->   state
xij variables:
                         lnp1 lnp2 ... lnp10  ->   lnp
                         lnc1 lnc2 ... lnc10  ->   lnc
-----------------------------------------------------------------------------

. list in 1/6, clean

       year   state   lnp   lnc
  1.     63       1   4.5   4.5
  2.     63       2   4.4   4.8
  3.     63       3   4.5   4.6
  4.     63       4   4.4   5.0
  5.     63       5   4.5   5.1
  6.     63       6   4.5   5.1
```

The data are now in long form, as in section 8.11.2.

8.12 Stata resources

FE and RE estimators appear in many econometrics texts. Panel texts with complete coverage of the basic material are Baltagi (2008) and Hsiao (2003). The key Stata reference is [XT] *Longitudinal/Panel-Data Reference Manual*, especially [XT] `xt` and [XT] `xtreg`. Useful online `help` categories include `xt` and `xtreg`. For estimation with long panels, a useful Stata user-written command is `xtscc`, as well as several others mentioned in section 8.10.

8.13 Exercises

1. For the data of section 8.3, use `xtsum` to describe the variation in `occ`, `smsa`, `ind`, `ms`, `union`, `fem`, and `blk`. Which of these variables are time-invariant? Use `xttab` and `xttrans` to provide interpretations of how `occ` changes for individuals over the seven years. Provide a time-series plot of `exp` for the first ten observations and provide interpretation. Provide a scatterplot of `lwage` against `ed`. Is this plot showing within variation, between variation, or both?

2. For the data of section 8.3, manually obtain the three standard deviations of `lwage` given by the `xtsum` command. For the overall standard deviation, use `summarize`. For the between standard deviation, compute `by id: egen meanwage = mean(lwage)` and apply `summarize` to (`meanwage-`*grandmean*) for `t==1`, where *grandmean* is the grand mean over all observations. For the within standard deviation, apply `summarize` to `(lwage-meanwage)`. Compare your standard deviations with those from `xtsum`. Does $s_{\mathrm{O}}^2 \simeq s_{\mathrm{W}}^2 + s_{\mathrm{B}}^2$?

3. For the model and data of section 8.4, compare PFGLS estimators under the following assumptions about the error process: independent, exchangeable, AR(2), and MA(6). Also compare the associated standard-error estimates obtained by using default standard errors and by using cluster–robust standard errors. You will find it easiest if you combine results using `estimates table`. What happens if you try to fit the model with no structure placed on the error correlations?

4. For the model and data of section 8.5, obtain the within estimator by applying `regress` to (8.7). Hint: For example, for variable x, type `by id: egen avex = mean(`x`)` followed by `summarize` x and then `generate mdx = `x` - avex + r(mean)`. Verify that you get the same estimated coefficients as you would with `xtreg, fe`.

5. For the model and data of section 8.6, compare the RE estimators obtained by using `xtreg` with the `re`, `mle`, and `pa` options, and `xtgee` with the `corr(exchangeable)` option. Also compare the associated standard-error estimates obtained by using default standard errors and by using cluster–robust standard errors. You will find it easiest if you combine results using `estimates table`.

6. Consider the RE model output given in section 8.7. Verify that, given the estimated values of `e_sigma` and `u_sigma`, application of the formulas in that section leads to the estimated values of `rho` and `theta`.

7. Make an unbalanced panel dataset by using the data of section 8.4 but then typing `set seed 10101` and `drop if runiform() < 0.2`. This will randomly drop 20% of the individual–year observations. Type `xtdescribe`. Do you obtain the expected patterns of missing data? Use `xtsum` to describe the variation in `id`, `t`, `wage`, `ed`, and `south`. How do the results compare with those from the full panel? Use `xttab` and `xttrans` to provide interpretations of how `south` changes for individuals over time. Compare the within estimator with that in section 8.5 using the balanced panel.