International University of Japan
Public Management & Policy Analysis Program

Practical Guides To Panel Data Modeling: A Step by Step Analysis Using Stata*

Hun Myoung Park, Ph.D.
kucc625@iuj.ac.jp

Last modified on October 2011

Public Management and Policy Analysis Program
Graduate School of International Relations
International University of Japan
777 Kokusai-cho Minami Uonuma-shi, Niigata 949-7277, Japan
(025) 779-1424
http://www.iuj.ac.jp/faculty/kucc625

* The citation of this document should read: "Park, Hun Myoung. 2011. *Practical Guides To Panel Data Modeling: A Step-by-step Analysis Using Stata*. Tutorial Working Paper. Graduate School of International Relations, International University of Japan." This document is based on Park, Hun Myoung. 2005-2009. *Linear Regression Models for Panel Data Using SAS, Stata, LIMDEP, and SPSS*. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University

1. Introduction

Panel data are also called *longitudinal data* or *cross-sectional time-series data*. These longitudinal data have "observations on the same units in several different time periods" (Kennedy, 2008: 281); A panel data set has multiple entities, each of which has repeated measurements at different time periods. Panel data may have *individual (group) effect*, *time effect*, or both, which are analyzed by *fixed effect* and/or *random effect* models.

U.S. Census Bureau's Census 2000 data at the state or county level are cross-sectional but not time-series, while annual sales figures of Apple Computer Inc. for the past 20 years are time series but not cross-sectional. The cumulative Census data at the state level for the past 20 years are longitudinal. If annual sales data of Apple, IBM, LG, Siemens, Microsoft, Sony, and AT&T for the past 10 years are available, they are panel data. The National Longitudinal Survey of Labor Market Experience (NLS) and the Michigan Panel Study of Income Dynamics (PSID) data are cross sectional and time-series, while the cumulative General Social Survey (GSS) and American National Election Studies (ANES) data are not in the sense that individual respondents vary across survey year.

As more and more panel data are available, many scholars, practitioners, and students have been interested in panel data modeling because these longitudinal data have more variability and allow to explore more issues than do cross-sectional or time-series data alone (Kennedy, 2008: 282). Baltagi (2001) puts, "Panel data give more informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency" (p.6). Given well-organized panel data, panel data models are definitely attractive and appealing since they provide ways of dealing with heterogeneity and examine fixed and/or random effects in the longitudinal data.

However, panel data modeling is not as easy as it sounds. A common misunderstanding is that fixed and/or random effect models should always be employed whenever your data are arranged in the panel data format. The problems of panel data modeling, by and large, come from 1) panel data themselves, 2) modeling process, and 3) interpretation and presentation of the result. Some studies analyze poorly organized panel data (in fact, they are not longitudinal in a strong econometric sense) and some others mechanically apply fixed and/or random effect models in haste without consideration of relevance of such models. Careless researchers often fail to interpret the results correctly and to present them appropriately.

The motivation of this document is several IUJ master's theses that, I think, applied panel data models inappropriately and failed to interpret the results correctly. This document is intended to provide practical guides of panel data modeling, in particular, for writing a master's thesis. Students can learn how to 1) organize panel data, 2) recognize and handle ill-organized data, 3) choose a proper panel data model, 4) read and report Stata output correctly, 5) interpret the result substantively, and 6) present the result in a professional manner.

In order to avoid unnecessary complication, this document mainly focuses on linear regression models rather than nonlinear models (e.g., binary response and event count data models) and balanced data rather than unbalanced ones. Hopefully this document will be a good companion of those who want to analyze panel data for their master's theses at IUJ. Let us begin with preparing and evaluating panel data.

2. Preparing Panel Data

This section describes how to prepare panel data sets using Stata (release 11) and then discuss types and qualities of panel data.

2.1 Sample Panel Data Set

A sample panel data used here are total cost data for the U.S. airlines (1970-1984), which are available on http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm. The sample data set includes total cost, output index, fuel price, and loading factor of six U.S. airlines measured at 15 different time points. Let us type in the following command at the Stata's dot prompt.

```
. use http://www.indiana.edu/~statmath/stat/all/panel/airline.dta, clear
```

The `.use` command reads a data set `airline.dta` through Internet, and the `clear` option removes data in current memory and then loads new one in to the main memory. The `.keep` command below drops (deletes) all variables other than those listed in the command.

```
. keep airline year cost output fuel load
```

```
. describe airline year cost output fuel load
```

```
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------------------
airline         int    %8.0g                  Airline name
year            int    %8.0g                  Year
cost            float  %9.0g                  Total cost in $1,000
output          float  %9.0g                  Output in revenue passenger miles, index number
fuel            float  %9.0g                  Fuel price
load            float  %9.0g                  Load factor
```

The above `.describe` command displays basic information of variables listed after the command. The `.summary` command below provides descriptive statistics (e.g., mean, standard deviation, minimum, and maximum) of variables listed.[1] From the output below, we know that five airlines were coded from 1 to 6 and time periods were set from 1 through 15.

```
. sum airline year cost output fuel load
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     airline |        90         3.5    1.717393          1          6
        year |        90           8    4.344698          1         15
        cost |        90    13.36561    1.131971   11.14154    15.3733
      output |        90   -1.174309    1.150606  -3.278573   .6608616
        fuel |        90    12.77036    .8123749   11.55017     13.831
        load |        90    .5604602    .0527934    .432066    .676287
```

In order to use panel data commands in Stata, we need to declare cross-sectional (`airline`) and time-series (`year`) variables to tell Stata which variable is cross-sectional and which one is time-series. The `.tsset` command is followed by cross-sectional and time-series variables in order.

```
. tsset airline year
       panel variable:  airline (strongly balanced)
```

[1] You may use short versions of these commands; Stata knows that `.des` and `.sum` are equivalent to `.describe` and `.summary`, respectively.

```
        time variable:  year, 1 to 15
                delta:  1 unit
```

Let us first explore descriptive statistics of panel data. Run `.xtsum` to obtain summary statistics. The total number of observations is 90 because there are 6 units (entities) and 15 time periods. The overall mean (13.3656) and standard deviation (1.1320) of total cost below are the same as those in the `.sum` output above.

```
. xtsum cost output fuel load
```

```
Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
cost     overall |  13.36561   1.131971   11.14154    15.3733 |     N =      90
         between |             .9978636   12.27441   14.67563 |     n =       6
         within  |             .6650252   12.11545   14.91617 |     T =      15
                 |                                            |
output   overall | -1.174309   1.150606  -3.278573   .6608616 |     N =      90
         between |             1.166556   -2.49898   .3192696 |     n =       6
         within  |             .4208405  -1.987984   .1339861 |     T =      15
                 |                                            |
fuel     overall |  12.77036   .8123749   11.55017     13.831 |     N =      90
         between |             .0237151    12.7318    12.7921 |     n =       6
         within  |             .8120832   11.56883    13.8513 |     T =      15
                 |                                            |
load     overall |  .5604602   .0527934    .432066    .676287 |     N =      90
         between |             .0281511   .5197756   .5971917 |     n =       6
         within  |             .0460361   .4368492   .6581019 |     T =      15
```

Note that Stata lists three different types of statistics: overall, between, and within. Overall statistics are ordinary statistics that are based on 90 observations. “Between” statistics are calculated on the basis of summary statistics of six airlines (entities) regardless of time period, while “within” statistics by summary statistics of 15 time periods regardless of airline.

2.2 Type of Panel Data

A panel data set contains *n* entities or subjects, each of which includes *T* observations measured at 1 through *t* time period. Thus, the total number of observations in the panel data is *nT*. Ideally, panel data are measured at regular time intervals (e.g., year, quarter, and month). Otherwise, panel data should be analyzed with caution. A panel may be long or short, balanced or unbalanced, and fixed or rotating.

2.2.1 Long versus Short Panel Data

A *short panel* has many entities (large *n*) but few time periods (small *T*), while a *long panel* has many time periods (large T) but few entities (Cameron and Trivedi, 2009: 230). Accordingly, a short panel data set is *wide* in width (cross-sectional) and short in length (time-series), whereas a long panel is *narrow* in width. Both too small N (Type I error) and too large N (Type II error) problems matter. Researchers should be very careful especially when examining either short or long panel.

2.2.2 Balanced versus Unbalanced Panel Data

In a *balanced panel*, all entities have measurements in all time periods. In a contingency table (or cross-table) of cross-sectional and time-series variables, each cell should have only one frequency. Therefore, the total number of observations is *nT*. This tutorial document assumes that we have a well-organized balanced panel data set.

When each entity in a data set has different numbers of observations, the panel data are not balanced. Some cells in the contingency table have zero frequency. Accordingly, the total number of observations is not *nT* in an *unbalanced panel*. Unbalanced panel data entail some computation and estimation issues although most software packages are able to handle both balanced and unbalanced data.

2.2.3 Fixed versus Rotating Panel Data

If the same individuals (or entities) are observed for each period, the panel data set is called a *fixed panel* (Greene 2008: 184). If a set of individuals changes from one period to the next, the data set is a *rotating panel*. This document assumes a fixed panel.

2.3 Data Arrangement: Long versus Wide Form in Stata

A typical panel data set has a cross-section (entity or subject) variable and a time-series variable. In Stata, this arrangement is called the long form (as opposed to the wide form). While the long form has both individual (e.g., entity and group) and time variables, the wide form includes either individual or time variable. Most statistical software packages assume that panel data are arranged in the long form.

The following data set shows a typical panel data arrangement. Yes, this is a long form. There are 6 entities (`airline`) and 15 time periods (`year`).[2]

```
. list airline year load cost output fuel in 1/20, sep(20)
```

```
     +-------------------------------------------------------------+
     | airline   year      load       cost      output       fuel |
     |-------------------------------------------------------------|
  1. |       1      1   .534487    13.9471   -.0483954   11.57731 |
  2. |       1      2   .532328   14.01082   -.0133315   11.61102 |
  3. |       1      3   .547736   14.08521    .0879925   11.61344 |
  4. |       1      4   .540846   14.22863    .1619318   11.71156 |
  5. |       1      5   .591167   14.33236    .1485665   12.18896 |
  6. |       1      6   .575417    14.4164    .1602123   12.48978 |
  7. |       1      7   .594495   14.52004    .2550375   12.48162 |
  8. |       1      8   .597409   14.65482    .3297856    12.6648 |
  9. |       1      9   .638522   14.78597    .4779284   12.85868 |
 10. |       1     10   .676287   14.99343    .6018211   13.25208 |
 11. |       1     11   .605735   15.14728    .4356969   13.67813 |
 12. |       1     12    .61436   15.16818    .4238942   13.81275 |
 13. |       1     13   .633366   15.20081    .5069381   13.75151 |
 14. |       1     14   .650117   15.27014    .6001049   13.66419 |
 15. |       1     15   .625603    15.3733    .6608616   13.62121 |
 16. |       2      1   .490851   13.25215    -.652706   11.55017 |
 17. |       2      2   .473449   13.37018    -.626186   11.62157 |
 18. |       2      3   .503013   13.56404   -.4228269   11.68405 |
 19. |       2      4   .512501    13.8148   -.2337306   11.65092 |
 20. |       2      5   .566782   14.00113   -.1708536   12.27989 |
     +-------------------------------------------------------------+
```

If data are structured in a wide form, you need to rearrange data first. Stata has the `.reshape` command to rearrange a data set back and forth between long and short forms. The following `.reshape` with `wide` changes from the long form to wide one so that the resulting data set in a wide form has only six observations but in turn include an identification (entity)

[2] The `.list` command lists data items of individual observations. The `in 1/20` of this command displays data of the first 20 observations, and the `sep(20)` option inserts a horizontal separator line in every 20 observations rather than in the default every 5 lines.

variable `airline` and as many variables as the time periods (4×15), dropping a time variable `year`.

```
. reshape wide cost output fuel load, i(airline) j(year)
(note: j = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       90   ->       6
Number of variables                   6   ->      61
j variable (15 values)             year   ->   (dropped)
xij variables:
                                   cost   ->   cost1 cost2 ... cost15
                                 output   ->   output1 output2 ... output15
                                   fuel   ->   fuel1 fuel2 ... fuel15
                                   load   ->   load1 load2 ... load15
-----------------------------------------------------------------------------
```

The `i()` above specifies identification variables to be used as identification of observations.

If you wish to rearrange the data set back to the long counterpart, run the following `.reshape` command with `long`.

```
. reshape long cost output fuel load, i(airline) j(year)
```

2.4 Evaluating the Qualities of Your Panel Data

The first task that a research has to do after cleaning data is to check the quality of panel data in hand. When saying panel data, you are implicitly arguing that the data are well arranged by both cross-sectional and time-series variables and that you get a strong impression of presence of fixed and/or random effects. Otherwise, the data are simply (or physically) arranged in the panel data format but are no longer panel data in an econometric sense.

The most important issue is consistency in the unit of analysis (or measurement), which says that each observation in a data set deserves being treated and weighted equally. This requirement seems self-evident but is often overlooked by careless researchers. If each observation is not equivalent in many senses, any analysis based on such data may not be reliable. Here are some checkpoints that researchers should examine carefully.

- Make sure that your data are really longitudinal and that there are some fixed and/or random effects.
- Check if individuals (e.g., entities and subjects) are not consistent but changing. For instance, a company might be split or merged during the research period to become a completely new one.
- Similarly, check if time periods are not consistent but changing. A time period under some circumstances may not be fixed but almost random (e.g., second period is two days later the first period, third period is 100 days later the second period, forth period is one and a half years later the third period, etc.) In some data sets, time period is fixed but multiple time periods are used; both yearly and weekly data coexist in a data set.
- Check if an entity has more than one observation in a particular time period. For example, Apple has four observations for quarterly sales data in 2011, while each of other firms has one yearly sales observation in that year. In this simple case, you may aggregate quarterly data to obtain yearly figures.

- Check if measurement methods employed are not consistent. Measurements are not commensurable if 1) some entities were measured in method A and other entities in method B, 2) some time periods were measured in method C and other periods in method D, and/or 3) both 1) and 2) are mixed.[3]
- Be careful when you "darn" your data set by combining data sets measured and built by different institutions who employed different methods. This circumstance is quite understandable because a perfect data set is rarely ready for you; in many cases, you need to combine some sources of information to build a new data set for your research.

Another issue is if the number of entities and/or time-period is too small or too large. It is less valuable to contrast one group (or time period) with another in the panel data framework: n=2 or T=3). By contrast, comparing millions of individuals or time periods is almost useless because of high likelihood of Type II error. This task is almost similar to arguing that at least one company out of 1 million firms in the world has a different productivity. Is this argument interesting to you?; We already know that! In case of too large N (specifically n or T), you might try to reclassify individuals or time periods into several meaningful categories; for example, classify millions of individuals by their citizenships or ethnic groups (e.g., white, black, Asian, and Spanish).

Finally, many missing values are likely lower the quality of panel data. So called *listwise deletion* (an entire record is excluded from analysis if any single value of a variable is missing) tends to reduce the number of observations used in a model and thus weaken statistical power of a test. This issue is also related to discussion on balanced versus unbalanced panel data.

Once a well organized panel data is prepared, we are moving forward to discuss panel data models that are used to analyze fixed and/or random effects embedded in the longitudinal data.

[3] Assume that methods A and B, and methods C and D are not comparable each other in terms of scale and unit of measurements.

3. Basics of Panel Data Models

Panel data models examine group (individual-specific) effects, time effects, or both in order to deal with *heterogeneity* or *individual effect* that may or may not be observed.[4] These effects are either fixed or random effect. A *fixed effect model* examines if intercepts vary across group or time period, whereas a *random effect model* explores differences in error variance components across individual or time period. A *one-way model* includes only one set of dummy variables (e.g., firm1, firm2, ...), while a *two-way model* considers two sets of dummy variables (e.g., city1, city2, … and year1, year2, …).

This section follows Greene's (2008) notations with some modifications, such as lower-case k (the number of regressors excluding the intercept term; He uses K instead), w_{it} (the composite error term), and v_{it} (traditional error term; He uses ε_{it}).

3.1 Pooled OLS

If individual effect u_i (cross-sectional or time specific effect) does not exist (u_i =0), ordinary least squares (OLS) produces efficient and consistent parameter estimates.

$$y_{it} = \alpha + X_{it}'\beta + \varepsilon_{it} \ \ (u_i = 0)$$

OLS consists of five core assumptions (Greene, 2008: 11-19; Kennedy, 2008: 41-42).

1. **Linearity** says that the dependent variable is formulated as a linear function of a set of independent variable and the error (disturbance) term.
2. **Exogeneity** says that the expected value of disturbances is zero or disturbances are not correlated with any regressors.
3. Disturbances have the same variance (**3.a homoskedasticity**) and are not related with one another (**3.b nonautocorrelation**)
4. The observations on the independent variable are **not stochastic** but fixed in repeated samples without measurement errors.
5. **Full rank** assumption says that there is no exact linear relationship among independent variables (no multicollinearity).

If individual effect u_i is not zero in longitudinal data, heterogeneity (individual specific characteristics like intelligence and personality that are not captured in regressors) may influence assumption 2 and 3. In particular, disturbances may not have same variance but vary across individual (*heteroskedasticity*, violation of assumption 3.a) and/or are related with each other (*autocorrelation*, violation of assumption 3.b). This is an issue of *nonspherical* variance-covariance matrix of disturbances. The violation of assumption 2 renders random effect estimators biased. Hence, the OLS estimator is no longer best unbiased linear estimator. Then panel data models provide a way to deal with these problems.

3.2 Fixed versus Random Effects

Panel data models examine fixed and/or random effects of individual or time. The core difference between fixed and random effect models lies in the role of dummy variables

[4] Country, state, agency, firm, respondent, employee, and student are examples of a unit (individual or entity), whereas year, quarter, month, week, day, and hour can be examples of a time period.

(Table 3.1). A parameter estimate of a dummy variable is a part of the intercept in a fixed effect model and an error component in a random effect model. Slopes remain the same across group or time period in either fixed or random effect model. The functional forms of one-way fixed and random effect models are,[5]

Fixed effect model: $y_{it} = (\alpha + u_i) + X'_{it}\beta + v_{it}$

Random effect model: $y_{it} = \alpha + X'_{it}\beta + (u_i + v_{it})$,

where u_i is a fixed or random effect specific to individual (group) or time period that is not included in the regression, and errors are *independent identically distributed*, $v_{it} \sim IID(0, \sigma_v^2)$.

A fixed group effect model examines individual differences in intercepts, assuming the same slopes and constant variance across individual (group and entity). Since an individual specific effect is time invariant and considered a part of the intercept, u_i is allowed to be correlated with other regressors; That is, OLS assumption 2 is not violated. This fixed effect model is estimated by least squares dummy variable (LSDV) regression (OLS with a set of dummies) and within effect estimation methods.

Table 3.1 Fixed Effect and Random Effect Models

	Fixed Effect Model	Random Effect Model
Functional form	$y_{it} = (\alpha + u_i) + X'_{it}\beta + v_{it}$	$y_{it} = \alpha + X'_{it}\beta + (u_i + v_{it})$
Assumption	-	Individual effects are not correlated with regressors
Intercepts	Varying across group and/or time	Constant
Error variances	Constant	Randomly distributed across group and/or time
Slopes	Constant	Constant
Estimation	LSDV, within effect estimation	GLS, FGLS (EGLS)
Hypothesis test	F test	Breusch-Pagan LM test

A random effect model assumes that individual effect (heterogeneity) is not correlated with any regressor and then estimates error variance specific to groups (or times). Hence, u_i is an individual specific random heterogeneity or a component of the composite error term. This is why a random effect model is also called an error component model. The intercept and slopes of regressors are the same across individual. The difference among individuals (or time periods) lies in their individual specific errors, not in their intercepts.

A random effect model is estimated by generalized least squares (GLS) when a covariance structure of an individual *i*, Σ (sigma), is known. The feasible generalized least squares (FGLS) or estimated generalized least squares (EGLS) method is used to estimate the entire variance-covariance matrix V (Σ in all diagonal elements and 0 in all off-diagonal elements) when Σ is not known. There are various estimation methods for FGLS including the maximum likelihood method and simulation (Baltagi and Cheng, 1994).

A random effect model reduces the number of parameters to be estimated but will produce inconsistent estimates when individual specific random effect is correlated with regressors (Greene, 2008: 200-201).

Fixed effects are tested by the F test, while random effects are examined by the Lagrange multiplier (LM) test (Breusch and Pagan, 1980). If the null hypothesis is not rejected in either

[5] Let us focus here on cross-sectional (group) effects. For time effects, switch *i* with *t* in the formula.

test, the pooled OLS regression is favored. The Hausman specification test (Hausman, 1978) compares a random effect model to its fixed counterpart. If the null hypothesis that the individual effects are uncorrelated with the other regressors is not rejected, a random effect model is favored over its fixed counterpart.

If one cross-sectional or time-series variable is considered (e.g., country, firm, and race), this is called a one-way fixed or random effect model. Two-way effect models have two sets of dummy variables for individual and/or time variables (e.g., state and year) and thus entail some issues in estimation and interpretation.

3.3 Estimating Fixed Effect Models

There are several strategies for estimating a fixed effect model. The *least squares dummy variable* model (LSDV) uses dummy variables, whereas the "*within" estimation* does not. These strategies, of course, produce the identical parameter estimates of regressors (non-dummy independent variables). The "*between" estimation* fits a model using individual or time means of dependent and independent variables without dummies.

LSDV with a dummy dropped out of a set of dummies is widely used because it is relatively easy to estimate and interpret substantively. This LSDV, however, becomes problematic when there are many individuals (or groups) in panel data. If T is fixed and $n\rightarrow\infty$ (n is the number of groups or firms and T is the number of time periods), parameter estimates of regressors are consistent but the coefficients of individual effects, $\alpha + u_i$, are not (Baltagi, 2001: 14). In this short panel, LSDV includes a large number of dummy variables; the number of these parameters to be estimated increases as n increases (*incidental parameter problem*); therefore, LSDV loses n degrees of freedom but returns less efficient estimators (p.14). Under this circumstance, LSDV is useless and thus calls for another strategy, the within effect estimation.

Unlike LSDV, the "within" estimation does not need dummy variables, but it uses deviations from group (or time period) means. That is, "within" estimation uses variation within each individual or entity instead of a large number of dummies. The "within" estimation is,[6]

$$(y_{it} - \bar{y}_{i\bullet}) = (x_{it} - \bar{x}_{i\bullet})'\beta + (\varepsilon_{it} - \bar{\varepsilon}_{i\bullet}),$$

where $\bar{y}_{i\bullet}$ is the mean of dependent variable (DV) of individual (group) i, $\bar{x}_{i\bullet}$ represent the means of independent variables (IVs) of group i, and $\varepsilon_{i\bullet}$ is the mean of errors of group i.

In this "within" estimation, the incidental parameter problem is no longer an issue. The parameter estimates of regressors in the "within" estimation are identical to those of LSDV. The "within" estimation reports correct the *sum of squared errors* (SSE). The "within" estimation, however, has several disadvantages.

First, data transformation for "within" estimation wipes out all time-invariant variables (e.g., gender, citizenship, and ethnic group) that do not vary within an entity (Kennedy, 2008: 284). Since deviations of time-invariant variables from their average are all zero, it is not possible

[6] This "within" estimation needs three steps: 1) compute group means of the dependent and independent variables; 2) transform dependent and independent variables to get deviations from their group means; 3) run OLS on the transformed variables without the intercept term.

to estimate coefficients of such variables in "within" estimation. As a consequence, we have to fit LSDV when a model has time-invariant independent variables.

Second, "within" estimation produces incorrect statistics. Since no dummy is used, the within effect model has larger degrees of freedom for errors, accordingly reporting small *mean squared errors* (MSE), *standard errors of the estimates* (SEE) or *square root of mean squared errors* (SRMSE), and incorrect (smaller) standard errors of parameter estimates. Hence, we have to adjust incorrect standard errors using the following formula.[7]

$$se_k^* = se_k \sqrt{\frac{df_{error}^{within}}{df_{error}^{LSDV}}} = se_k \sqrt{\frac{nT-k}{nT-n-k}}$$

Third, R^2 of the "within" estimation is not correct because the intercept term is suppressed. Finally, the "within" estimation does not report dummy coefficients. We have to compute them, if really needed, using the formula $d_i^* = \bar{y}_{i\bullet} - \bar{x}_{i\bullet}'\beta$.

Table 3.2 Comparison of Three Estimation Methods

	LSDV	Within Estimation	Between Estimation
Functional form	$y_i = i\alpha_i + X_i\beta + \varepsilon_i$	$y_{it} - \bar{y}_{i\bullet} = x_{it} - \bar{x}_{i\bullet} + \varepsilon_{it} - \bar{\varepsilon}_{i\bullet}$	$\bar{y}_{i\bullet} = \alpha + \bar{x}_{i\bullet} + \varepsilon_i$
Time invariant variables	Yes	No	No
Dummy variables	Yes	No	No
Dummy coefficients	Presented	Need to be computed	N/A
Transformation	No	Deviation from the group means	Group means
Intercept estimated	Yes	No	Yes
R^2	Correct	Incorrect	
SSE	Correct	Correct	
MSE/SEE (SRMSE)	Correct	Incorrect (smaller)	
Standard errors	Correct	Incorrect (smaller)	
DF_{error}	*nT-n-k*[*]	*nT-k* (*n* larger)	*n-k-1*
Observations	*nT*	*nT*	*n*

[*] It means that the LSDV estimation loses *n* degrees of freedom because of dummy variables included.

The "between group" estimation, so called the group mean regression, uses variation between individual entities (groups). Specifically, this estimation calculates group means of the dependent and independent variables and thus reduces the number of observations down to *n*. Then, run OLS on these transformed, aggregated data: $\bar{y}_{i\bullet} = \alpha + \bar{x}_{i\bullet} + \varepsilon_i$. Table 3.2 contrasts LSDV, "within group" estimation, and "between group" estimation.

3.4 Estimating Random Effect Models

The one-way random effect model incorporates a composite error term, $w_{it} = u_i + v_{it}$. The u_i are assumed independent of traditional error term v_{it} and regressors X_{it}, which are also independent of each other for all *i* and *t*. Remember that this assumption is not necessary in a fixed effect model. This model is,

[7] Fortunately, Stata and other software packages report adjusted standard errors for us.

$y_{it} = \alpha + X_{it}'\beta + u_i + v_{it}$, where $u_i \sim IID(0, \sigma_u^2)$, and $v_{it} \sim IID(0, \sigma_v^2)$.

The covariance elements of $Cov(w_{it}, w_{js}) = E(w_{it}w'_{js})$ are $\sigma_u^2 + \sigma_v^2$ if *i*=*j* and *t*=*s* and σ_u^2 if *i*=*j* and $t \neq s$. Therefore, the covariance structure of composite errors $\Sigma = E(w_i w_i')$ for individual *i* and the variance-covariance matrix of entire disturbances (errors) V are,

$$\underset{T\times T}{\Sigma} = \begin{bmatrix} \sigma_u^2 + \sigma_v^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_v^2 & \dots & \sigma_u^2 \\ \dots & \dots & \dots & \dots \\ \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2 + \sigma_v^2 \end{bmatrix} \text{ and } \underset{nT\times nT}{V} = I_n \otimes \Sigma = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Sigma \end{bmatrix}$$

A random effect model is estimated by generalized least squares (GLS) when the covariance structure is known, and by feasible generalized least squares (FGLS) or estimated generalized least squares (EGLS) when the covariance structure of composite errors is unknown. Since Σ is often unknown, FGLS/EGLS is more frequently used than GLS. Compared to a fixed effect counterpart, a random effect model is relatively difficult to estimate.

In FGLS, you first have to estimate θ using $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$. The $\hat{\sigma}_u^2$ comes from the between effect estimation (group mean regression) and $\hat{\sigma}_v^2$ is derived from the SSE (sum of squared errors) of the within effect estimation or the deviations of residuals from group means of residuals.

$$\hat{\theta} = 1 - \sqrt{\frac{\hat{\sigma}_v^2}{T\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} = 1 - \sqrt{\frac{\hat{\sigma}_v^2}{T\hat{\sigma}_{between}^2}},$$

where $\hat{\sigma}_u^2 = \hat{\sigma}_{between}^2 - \frac{\hat{\sigma}_v^2}{T}$, where $\hat{\sigma}_{between}^2 = \frac{SSE_{between}}{n-k-1}$,

$\hat{\sigma}_v^2 = \frac{SSE_{within}}{nT-n-k} = \frac{e'e_{within}}{nT-n-k} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(v_{it} - \bar{v}_{i\bullet})^2}{nT-n-k}$, where v_{it} are the residuals of the LSDV.

Then, the dependent variable, independent variables, and the intercept term need to be transformed as follows,

$y_{it}^* = y_{it} - \hat{\theta}\,\bar{y}_{i\bullet}$
$x_{it}^* = x_{it} - \hat{\theta}\,\bar{x}_{i\bullet}$ for all x_k
$\alpha^* = 1 - \hat{\theta}$

Finally, run OLS on those transformed variables with the traditional intercept suppressed.

$y_{it}^* = \alpha^* + x_{it}^{*\prime}\beta^* + \varepsilon_{it}^*$.

3.5 Testing Fixed and Random Effects

How do we know if fixed and/or random effects exist in panel data in hand? A fixed effect is tested by F-test, while a random effect is examined by Breusch and Pagan's (1980) Lagrange multiplier (LM) test. The former compares a fixed effect model and OLS to see how much the fixed effect model can improve the goodness-of-fit, whereas the latter contrast a random effect model with OLS. The similarity between random and fixed effect estimators is tested by a Hausman test.

3.5.1 F-test for Fixed Effects

In a regression of $y_{it} = \alpha + \mu_i + X_{it}'\beta + \varepsilon_{it}$, the null hypothesis is that all dummy parameters except for one for the dropped are all zero, $H_0 : \mu_1 = ... = \mu_{n-1} = 0$. The alternative hypothesis is that at least one dummy parameter is not zero. This hypothesis is tested by an F test, which is based on loss of goodness-of-fit. This test contrasts LSDV (robust model) with the pooled OLS (efficient model) and examines the extent that the goodness-of-fit measures (SSE or R^2) changed.

$$F(n-1, nT-n-k) = \frac{(e'e_{pooled} - e'e_{LSDV})/(n-1)}{(e'e_{LSDV})/(nT-n-k)} = \frac{(R^2_{LSDV} - R^2_{pooled})/(n-1)}{(1-R^2_{LSDV})/(nT-n-k)}$$

If the null hypothesis is rejected (at least one group/time specific intercept u_i is not zero), you may conclude that there is a significant fixed effect or significant increase in goodness-of-fit in the fixed effect model; therefore, the fixed effect model is better than the pooled OLS.

3.5.2 Breusch-Pagan LM Test for Random Effects

Breusch and Pagan's (1980) Lagrange multiplier (LM) test examines if individual (or time) specific variance components are zero, $H_0 : \sigma_u^2 = 0$. The LM statistic follows the chi-squared distribution with one degree of freedom.

$$LM_u = \frac{nT}{2(T-1)}\left[\frac{T^2\bar{e}'\bar{e}}{e'e} - 1\right]^2 \sim \chi^2(1),$$

where $\bar{e}$ is the $n \times 1$ vector of the group means of pooled regression residuals, and $e'e$ is the SSE of the pooled OLS regression.

Baltagi (2001) presents the same LM test in a different way.

$$LM_u = \frac{nT}{2(T-1)}\left[\frac{\sum\left(\sum e_{it}\right)^2}{\sum\sum e_{it}^2} - 1\right]^2 = \frac{nT}{2(T-1)}\left[\frac{\sum(T\bar{e}_{i\bullet})^2}{\sum\sum e_{it}^2} - 1\right]^2 \sim \chi^2(1).$$

If the null hypothesis is rejected, you can conclude that there is a significant random effect in the panel data, and that the random effect model is able to deal with heterogeneity better than does the pooled OLS.

3.5.3 Hausman Test for Comparing Fixed and Random Effects

How do we know which effect (fixed effect or random effect) is more relevant and significant in the panel data? The Hausman specification test compares fixed and random effect models

under the null hypothesis that individual effects are uncorrelated with any regressor in the model (Hausman, 1978). If the null hypothesis of no correlation is not violated, LSDV and GLS are consistent, but LSDV is inefficient; otherwise, LSDV is consistent but GLS is inconsistent and biased (Greene, 2008: 208). The estimates of LSDV and GLS should not differ systematically under the null hypothesis. The Hausman test uses that "the covariance of an efficient estimator with its difference from an inefficient estimator is zero" (Greene, 2008: 208).

$$LM = \left(b_{LSDV} - b_{random}\right)' \hat{W}^{-1} \left(b_{LSDV} - b_{random}\right) \sim \chi^2(k),$$

where $\hat{W} = Var[b_{LSDV} - b_{random}] = Var(b_{LSDV}) - Var(b_{random})$ is the difference in the estimated covariance matrices of LSDV (robust model) and GLS (efficient model). Keep in mind that an intercept and dummy variables SHOULD be excluded in computation. This test statistic follows the chi-squared distribution with *k* degrees of freedom.

The formula says that a Hausman test examines if "the random effects estimate is insignificantly different from the unbiased fixed effect estimate" (Kennedy, 2008: 286). If the null hypothesis of no correlation is rejected, you may conclude that individual effects u_i are significantly correlated with at least one regressors in the model and thus the random effect model is problematic. Therefore, you need to go for a fixed effect model rather than the random effect counterpart. A drawback of this Hausman test is, however, that the difference of covariance matrices *W* may not be positive definite; Then, we may conclude that the null is not rejected assuming similarity of the covariance matrices renders such a problem (Greene, 2008: 209).

3.5.4 Chow Test for Poolability

What is poolability? Poolability asks if slopes are the same across group or over time (Baltagi 2001: 51-57). One simple version of poolability test is an extension of the Chow test (Chow, 1960). The null hypothesis of this Chow test is the slope of a regressor is the same regardless of individual for all *k* regrssors, $H_0 : \beta_{ik} = \beta_k$. Remember that slopes remain constant in fixed and random effect models; only intercepts and error variances matter.

$$F\left[(n-1)(k+1), n(T-k-1)\right] = \frac{(e'e - \sum e_i' e_i) \big/ (n-1)(k+1)}{\sum e_i' e_i \big/ n(T-k-1)},$$

where $e'e$ is the SSE of the pooled OLS and $e_i' e_i$ is the SSE of the pooled OLS for group *i*. If the null hypothesis is rejected, the panel data are not poolable; each individual has its own slopes for all regressors. Under this circumstance, you may try the random coefficient model or hierarchical regression model.

The Chow test assumes that individual error variance components follow the normal distribution, $\mu \sim N(0, s^2 I_{nT})$. If this assumption does not hold, the Chow test may not properly examine the null hypothesis (Baltagi, 2001: 53). Kennedy (2008) notes, "if there is reason to believe that errors in different equations have different variances, or that there is contemporaneous correlation between the equations' errors, such testing should be undertaken by using the SURE estimator, not OLS; … inference with OLS is unreliable if the variance-covariance matrix of the error is nonspherical" (p.292).

3.6 Model Selection: Fixed or Random Effect?

When combining fixed vs. random effects, group vs. time effects, and one-way vs. two-way effects, we get 12 possible panel data models as shown in Table 3.3. In general, one-way models are often used mainly due to their parsimony, and a fixed effect model is easier than a random counterpart to estimate the model and interpret its result. It is not, however, easy to sort out the best one out of the following 12 models.

Table 3.3 Classification of Panel Data Analysis

	Type	Fixed Effect	Random Effect
One-way	Group	One-way fixed group effect	One-way random group effect
	Time	One-way fixed time effect	One-way random time effect
Two-way	Two groups*	Two-way fixed group effect	Two-way random group effect
	Two times*	Two-way fixed time effect	Two-way random time effect
	Mixed	Two-way fixed group & time effect Two-way fixed time and random group effect Two-way fixed group and random time effect	Two-way random group & time effect

* These models need two group (or time) variables (e.g., country and airline).

3.6.1 Substantive Meanings of Fixed and Random Effects

The formal tests discussed in 3.5 examine presence of fixed and/or random effects. Specifically, the F-test compares a fixed effect model and (pooled) OLS, whereas the LM test contrasts a random effect model with OLS. The Hausman specification test compares fixed and random effect models. However, these tests do not provide substantive meanings of fixed and random effects. What does a fixed effect mean? How do we interpret a random effect substantively?

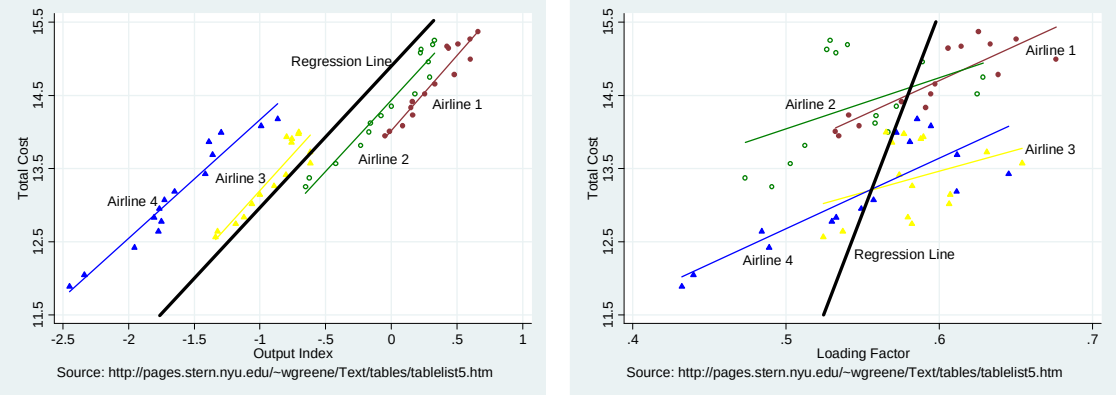
Here is a simple and rough answer. Suppose we are regressing the production of firms such as Apple, IBM, LG, and Sony on their R&D investment. A fixed effect might be interpreted as initial production capacities of these companies when no R&D investment is made; each firm has its own initial production capacity. A random effect might be viewed as a kind of consistency or stability of production. If the production of a company fluctuates up and down significantly, for example, its production is not stable (or its variance component is larger than those of other firms) even when its productivity (slope of R&D) remains the same across company.[8]

Kennedy (2008: 282-286) provides theoretical and insightful explanation of fixed and random effects. Either fixed or random effect is an issue of unmeasured variables or omitted relevance variables, which renders the pooled OLS biased. This heterogeneity is handled by either putting in dummy variables to estimate individual intercepts of groups (entities) or viewing "the different intercepts as having been drawn from a bowl of possible intercepts, so they may be interpreted as random … and treated as though they were a part of the error term" (p. 284); they are fixed effect model and random effect model, respectively. A random effect model has a "composite error term" that consists of the traditional random error and a "random intercept" measuring the extent to which individual's intercept differs from the

[8] Like dummy coefficients in a fixed effect model, parameter estimates of error components of individual companies can be calculated in a random effect model. The SAS MIXED procedure reports such error component estimators.

overall intercept (p. 284). He argues that the key difference between fixed and random effects is not whether unobserved heterogeneity is attributed to the intercept or variance components, but whether the individual specific error component is related to regressors.

Figure 3.1 Scatter Plots of Total Cost versus Output Index and Loading Factor



It will be a good practice to draw plots of the dependent and independent variables before modeling panel data. For instance, Figure 3.1 illustrates two scatter plots with linear regression lines of four airlines only. The left plot is of total cost versus output index, and the right one is of total cost versus loading factor (compare them with Kennedy's Figure 18.1 and 18.2). Assume that the thick black lines represent linear regression lines of entire observations. The key difference is that slopes of individual airlines are very similar to the overall regression line on the left plot, but different in the right plot.

As Kennedy (2008: 286) explains, OLS, fixed effect, and random effect estimators on the left plot are all unbiased, but random effect estimators are most efficient; a random effect is better. In the right plot, however, OLS and random effects estimators are biased because the composite error term seems to be correlated with a regressor, loading factor, but the fixed effects estimator is not biased; accordingly, a fixed effect model might be better.

3.6.2 Two Recommendations for Panel Data Modeling

The first recommendation, as in other data analysis processes, is to describe the data of interest carefully before analysis. Although often ignored in many data analyses, this data description is very important and useful for researchers to get ideas about data and analysis strategies. In panel data analysis, properties and quality of panel data influence model section significantly.

- Clean the data by examining if they were measured in reliable and consistent manners. If different time periods were used in a long panel, for example, try to rearrange (aggregate) data to improve consistency. If there are many missing values, decide whether you go for a balanced panel by throwing away some pieces of usable information or keep all usable observations in an unbalanced panel at the expense of methodological and computational complication.
- Examine the properties of the panel data including the number of entities (individuals), the number of time periods, balanced versus unbalanced panel, and fixed versus rotating panel. Then, try to find models appropriate for those properties.
- Be careful if you have "long" or "short" panel data. Imagine a long panel that has 10 thousand time periods but 3 individuals or a short panel of 2 (years) × 9,000 (firms).

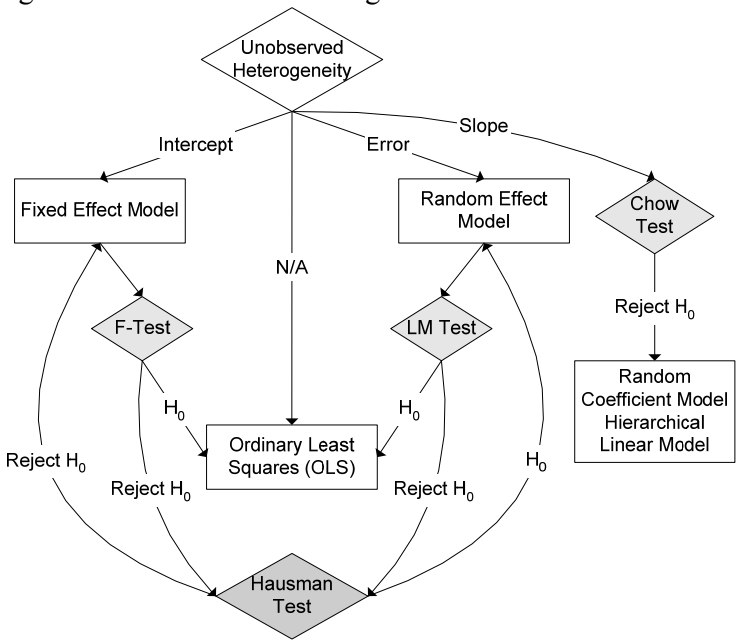
- If n and/or T are too large, try to reclassify individuals and/or time periods and get some manageable n' and T'. The null hypothesis of $u_1 = u_2 = \ldots = u_{999,999} = 0$ in a fixed effect model, for instance, is almost useless. This is just as you are seriously arguing that at least one citizen looks different from other 999,999 people! Didn't you know that before? Try to use yearly data rather than weekly data or monthly data rather than daily data.

Second recommendation is to begin with a simpler model. Try a pooled OLS rather than a fixed or random effect model; a one-way effect model rather than a two-way model; a fixed or random effect model rather than a hierarchical linear model; and so on. Do not try a fancy, of course, complicated, model that your panel data do not support enough (e.g., poorly organized panel and long/short panel).

3.6.3 Guidelines of Model Selection

On the modeling stage, let us begin with pooled OLS and then think critically about its potential problems if observed and unobserved heterogeneity (a set of missing relevant variables) is not taken into account. Also think about the source of heterogeneity (i.e., cross-sectional or time series variables) to determine individual (entity or group) effect or time effect.[9] Figure 3.2 provides a big picture of the panel data modeling process.

Figure 3.2 Panel Data Modeling Process



If you think that the individual heterogeneity is captured in the disturbance term and the individual (group or time) effect is not correlated with any regressors, try a random effect model. If the heterogeneity can be dealt with individual specific intercepts and the individual effect may possibly be correlated with any regressors, try a fixed effect model. If each individual (group) has its own initial capacity and shares the same disturbance variance with

[9] Kennedy (2008: 286) suggests that first examine if individual specific intercepts are equal; if yes, the panel data are poolable and OLS will do; if not, conduct the Hausman test; use random effect estimators if the group effect is not correlated with the error term; otherwise, use the fixed effect estimator.

other individuals, a fixed effect model is favored. If each individual has its own disturbance, a random effect will be better at figuring out heteroskedestic disturbances.

Next, conduct appropriate formal tests to examine individual group and/or time effects. If the null hypothesis of the LM test is rejected, a random effect model is better than the pooled OLS. If the null hypothesis of the F-test is rejected, a fixed effect model is favored over OLS. If both hypotheses are not rejected, fit the pooled OLS.

Conduct the Hausman test when both hypotheses of the F-test and LM test are all rejected. If the null hypothesis of uncorrelation between an individual effect and regressors is rejected, go for the robust fixed effect model; otherwise, stick to the efficient random effect model.

If you have a strong belief that the heterogeneity involves two cross-sectional, two time series, or one cross-section and one time series variables, try two-way effect models. Double-check if your panel data are well-organized, and *n* and *T* are large enough; do not try a two-way model for a poorly organized, badly unbalanced, and/or too long/short panel. Conduct appropriate F-test and LM test to examine the presence of two-way effects. Stata does not provide direct ways to fit two-way panel data models but it is not impossible. In Stata, two-way fixed effect models seem easier than two-way random effect models (see 3.7 below).

Finally, if you think that the heterogeneity entails slops (parameter estimates of regressors) varying across individual and/or time. Conduct a Chow test or equivalent to examine the poolability of the panel data. If the null hypothesis of poolable data is rejected, try a random coefficient model or hierarchical linear model.

3.7 Estimation Strategies in Stata

The least squares dummy variable (LSDV) regression, "within" estimation, "between" estimation (group or time mean model), GLS, and FGLS/EGLS are fundamentally based on ordinary least squares (OLS). Therefore, Stata `.regress` can fit all of these linear models.

Table 3.4 Stata Commands Used for Panel Data Analysis

	Commands	Options
Regression (OLS)	`.regress`	
LSDV1 without a dummy	`.regress`	
	`.xi: regress`	`i.`
LSDV2 without the intercept	`.regress`	`noconstant`
LSDV3 with a restriction	`.cnsreg` and `.constraint`	
One-way fixed effect ("within" estimation)	`.xtreg`	`fe`
	`.areg`	`abs`
Two-way fixed ("within" estimation)	`.xtreg` with a set of dummies	`fe`
"Between" estimation	`.xtreg`	`be`
One-way random effect	`.xtreg`	`re`
	`.xtgls`	
	`.xtmixed`	
Two-way random effect	`.xtmixed`	
Hieratical linear model	`.xtmixed`	
Random coefficient model	`.xtrc`	`betas`
Testing fixed effect (F-test)	`.test` (Included in `.xtreg`)	
Testing random effect (LM test)	`.xtest0`	
Comparing fixed and random effect	`.hausman`	

You can also use `.regress` with the `.xi` prefix command to fit LSDV1 without creating dummy variables (see 4.4.1). The `.cnsreg` command is used for LSDV3 with restrictions defined in `.constraint` (see 4.4.3). The `.areg` command with the `absorb` option, equivalent to the `.xtreg` with the `fe` option below, supports the one-way "within" estimation that involves a large number of individuals or time periods.

Stata has more convenient commands and options for panel data analysis. First, `.xtreg` estimates a fixed effect model with the `fe` option ("within" estimation), "between" estimators with `be`, and a random effect model with `re`. This command, however, does not directly fit two-way fixed and random effect models.[10] Table 3.4 summarizes related Stata commands.

A random effect model can be also estimated using `.xtmixed` and `.xtgls`. The `.xtgls` command fits panel data models with heteroscedasticity across group (time) and/or autocorrelation within a group (time). `.xtmixed` and `.xtrc` are used to fit hierarchical linear models and random coefficient models. In fact, a random effect model is a simple hierarchical linear model with a random intercept. `.logit` and `.probit` fit nonlinear regression models and examine fixed effects in logit and probit models.

`.xtmixed` with `fe` by default conducts the F-test for fixed effects. Of course, you can also use `.test` to conduct a classical Wald test to examine the fixed effects. Since `.xtmixed` does not report the Breusch-Pagan LM statistic for a random effect model, you need to conduct `.xtest0` after fitting a random effect model. Use `.hausman` to conduct Hausman test to compare fixed and random effect models.

[10] You may fit a two-way fixed effect model by including a set of dummies and using the `fe` option. For the two-way random effect model, you need to use the `.xtmixed` command instead of `.xtreg`.

4. Pooled OLS and LSDV

This section begins with classical least squares method called ordinary least squares (OLS) and explains how OLS can deal with unobserved heterogeneity using dummy variables. A dummy variable is a binary variable that is coded to either one or zero. OLS using dummy variables is called a least square dummy variable (LSDV) model. The sample model used here regresses total cost of airline companies on output in revenue passenger miles (output index), fuel price, and loading factor (the average capacity utilization of the fleet).[11]

4.1 Pooled OLS

The (pooled) OLS is a pooled linear regression without fixed and/or random effects. It assumes a constant intercept and slopes regardless of group and time period. In the sample panel data with five airlines and 15 time periods, the basic scheme is that total cost is determined by output, fuel price, and loading factor. The pooled OLS posits no difference in intercept and slopes across airline and time period.

OLS: $\cos t_i = \beta_0 + \beta_1 output_i + \beta_2 fuel_i + \beta_3 loading_i + \varepsilon_i$

Note that β_0 is the intercept; β_1 is the slope (coefficient or parameter estimate) of output; β_2 is the slope of fuel price; β_3 is the slope of loading factor; and ε_i is the error term.

Now, let us load the data and fit the pooled regression model.

```
. use http://www.indiana.edu/~statmath/stat/all/panel/airline.dta, clear
(Cost of U.S. Airlines (Greene 2003))

. regress cost output fuel load
      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  3,    86) = 2419.34
       Model |  112.705452     3  37.5684839           Prob > F      =  0.0000
    Residual |  1.33544153    86   .01552839           R-squared     =  0.9883
-------------+------------------------------           Adj R-squared =  0.9879
       Total |  114.040893    89  1.28135835           Root MSE      =  .12461

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .8827385   .0132545    66.60   0.000     .8563895    .9090876
        fuel |    .453977   .0203042    22.36   0.000     .4136136    .4943404
        load |   -1.62751    .345302    -4.71   0.000    -2.313948   -.9410727
       _cons |   9.516923   .2292445    41.51   0.000       9.0612    9.972645
------------------------------------------------------------------------------
```

This pooled OLS model fits the data well at the .05 significance level (F=2419.34 and p<.0000). R^2 of .9883 says that this model accounts for 99 percent of the total variance in the total cost of airline companies. The regression equation is,

cost = 9.5169 + .8827*output +.4540*fuel -1.6275*load

You may interpret these slopes in several ways. The *ceteris paribus* assumption, "holding all other variables constant," is important but often skipped in presentation. The p-values in parenthesis below are the results of t-tests for individual parameters.

[11] For details on the data, see http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm

Even in case of zero output index, zero fuel price, and zero loading factor, each airline company is expected to have 9.5169 units of total cost (p<.0000).

For one unit increase in output index, the total cost of airlines is expected to increase by .8827 units, holding all other variables constant (p<.0000).

Whenever fuel price increases by ten units, the total cost will increase by 4.5398 units, holding all other variables constant (p<.0000).

If the loading factor increases by one unit, an airline company can save total cost on average by 1.6275 units (p<.0000).

Although this model fits the data well, you may suspect if each airline or year has different initial total cost. That is, each airline may have its own initial total cost, its Y-intercept, that is significantly different from those of other airline companies.What if you believe that error terms vary across airline and/or year? The former question suspect fixed effects, whereas the latter asks if there is any random effect.

4.2 LSDV with a Set of Dummy Variables

Let us here examine fixed group effects by introducing group (airline) dummy variables. The dummy variable `g1` is set to 1 for airline 1 and zero for other airline companies; similarly, the variable `g2` is coded as 1 for airline 2 and zero for other airline companies; and so on. See the following for the coding scheme of dummy variables.[12]

```
. generate g1=(airline==1)
. gen g2=(airline==2)
…

. list airline year g1-g6
```

	airline	year	g1	g2	g3	g4	g5	g6
1.	1	1	1	0	0	0	0	0
2.	1	2	1	0	0	0	0	0
3.	1	3	1	0	0	0	0	0
…	…	…						
14.	1	14	1	0	0	0	0	0
15.	1	15	1	0	0	0	0	0
16.	2	1	0	1	0	0	0	0
17.	2	2	0	1	0	0	0	0
…	…	…						
32.	3	2	0	0	1	0	0	0
33.	3	3	0	0	1	0	0	0
…	…	…						
46.	4	1	0	0	0	1	0	0
47.	4	2	0	0	0	1	0	0
…	…	…						
61.	5	1	0	0	0	0	1	0
62.	5	2	0	0	0	0	1	0
…	…	…						
88.	6	13	0	0	0	0	0	1
89.	6	14	0	0	0	0	0	1
90.	6	15	0	0	0	0	0	1

[12] The first `.generate` command creates a dummy variable and then assigns 1 if the condition (`airline==1`) provided is satisfied and 0 otherwise.

This LSDV model is,

$$\textbf{LSDV}: \ \text{cost}_i = \beta_0 + \beta_1 output_i + \beta_2 fuel_i + \beta_3 loading_i + u_1 g_1 + u_2 g_2 + u_3 g_3 + u_4 g_4 + u_5 g_5 + \varepsilon_i$$

You should find that five group dummies, g_1-g_5, are added to the pooled OLS equation. Notice that one of six dummies, g_6 in this case, was excluded from the regression equation in order to avoid perfect multicollinearity.[13] The dummy variables and regressors are allowed to be correlated in a fixed effect model. u_1-u_5 are respectively parameter estimates of group dummy variables g_1-g_5.

Let us fit this linear regression with dummies. In the following command, I intentionally added g_1-g_5 right after the dependent variable `cost` in order to emphasize their coefficients are part of intercepts (as opposed to error terms).

```
. regress cost g1-g5 output fuel load
```

```
      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  8,    81) = 3935.79
       Model |   113.74827     8  14.2185338           Prob > F      =  0.0000
    Residual |  .292622872    81  .003612628           R-squared     =  0.9974
-------------+------------------------------           Adj R-squared =  0.9972
       Total |  114.040893    89  1.28135835           Root MSE      =  .06011

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          g1 |  -.0870617   .0841995    -1.03   0.304    -.2545924     .080469
          g2 |  -.1282976   .0757281    -1.69   0.094    -.2789728    .0223776
          g3 |  -.2959828   .0500231    -5.92   0.000     -.395513   -.1964526
          g4 |    .097494   .0330093     2.95   0.004     .0318159    .1631721
          g5 |   -.063007   .0238919    -2.64   0.010    -.1105443   -.0154697
      output |   .9192846   .0298901    30.76   0.000     .8598126    .9787565
        fuel |   .4174918   .0151991    27.47   0.000     .3872503    .4477333
        load |  -1.070396     .20169    -5.31   0.000    -1.471696   -.6690963
       _cons |   9.793004   .2636622    37.14   0.000     9.268399    10.31761
------------------------------------------------------------------------------
```

This LSDV fits the data better than does the pooled OLS in 4.1. The F statistic increased from 2419.34 to 3935.79 (p<.0000); SSE (sum of squares due to error or residual) decreased from 1.3354 to .2926; and R^2 increased from .9883 to .9974. Due to the dummies included, this model loses five degrees of freedom (from 86 to 81). Parameter estimates of individual regressors are slightly different from those in the pooled OLS. For instance, the coefficient of fuel price decreased from .4540 to .4175 but its statistical significance remained almost unchanged (p<.0000).

This fixed effect model posits that each airline has its own intercept but shares the same slopes of regressors (i.e., output index, fuel price, and loading factor). Then, how do we get airline specific intercepts? How do we interpret the dummy coefficients u_1-u_5? How do we report regression equations in LSDV?

The parameter estimate of g_6 (dropped dummy) is presented in the LSDV intercept (9.7930), which is the baseline intercept (reference point). Each of u_1-u_5 represents the deviation of its group specific intercept from the baseline intercept 9.7930 (intercept of airline 6). For instance, u_1 = -.0871 means that the intercept of airline 1 is .0871 smaller than the reference

[13] The last dummy g_6 (airline 6) was dropped and used as the reference group. Of course, you may drop any other dummy to get the equivalent result.

point 9.7930. Accordingly, the intercept of airline 1 is 9.7059=9.7930+(-.0871).[14] More formal computation is 9.7059 = 9.7930 + (-.0871)*1 + (-.1283)*0 + (-.2960)*0 + (.0975)*0 + (-.0630)*0. Note that all group dummies other than g_1 are zero in case of airline 1. Similarly, we can compute other intercepts for airline 2-5 and eventually get the following six regression equations.

Airline 1: cost = 9.7059 + .9193*output +.4175*fuel -1.0704*load
Airline 2: cost = 9.6647 + .9193*output +.4175*fuel -1.0704*load
Airline 3: cost = 9.4970 + .9193*output +.4175*fuel -1.0704*load
Airline 4: cost = 9.8905 + .9193*output +.4175*fuel -1.0704*load
Airline 5: cost = 9.7300 + .9193*output +.4175*fuel -1.0704*load
Airline 6: cost = 9.7930 + .9193*output +.4175*fuel -1.0704*load

Notice that all parameter estimates of regressors are the same regardless of airline. The coefficients of g_1-g_5 are interpreted as,

> *The intercept of airline 2 is .1284 smaller than that of baseline intercept (airline 6) 9.7930, but this deviation is not statistically significant at the .05 significance level (p<.094).*
>
> *The intercept of airline 3 is .2960 smaller than that of baseline intercept 9.7930 and this deviation is statistically discernable from zero at the .05 level (p<.000).*
>
> *The intercept of airline 4 is 9.8905, .0975 larger than that of baseline intercept (p<.004).*

The question here is which model is better than the other? The pooled OLS or LSDV? And why? What are the costs and benefits of adding group dummies and get different group intercepts? Is addition of group dummies valuable?

4.3 Comparing Pooled OLS and LSDV (Fixed Effect Model)

There are some significant difference between the pooled OLS and LSDV (Table 4.1). LSDV improved all goodness-of-fit measures like F-test, SSE, root MSE, and (adjusted) R^2 significantly but lost 5 degrees of freedom by adding five group dummies. LSDV seems better than the pooled OLS.

Table 4.1 Comparing Pooled OLS and LSDV

	Pooled OLS	LSDV
Ouput index	.8827 (p<.000)	.9193 (p<.000)
Fuel price	.4540 (p<.000)	.4175 (p<.000)
Loading factor	-1.6275 (p<.000)	-1.0704 (p<.000)
Overall intercept (baseline intercept)	9.5169 (p<.000)	9.7930 (p<.000)
Airline 1 (deviation from the baseline)		-.0871 (p<.304)
Airline 2 (deviation from the baseline)		-.1283 (p<.094)
Airline 3 (deviation from the baseline)		-.2960 (p<.000)
Airline 4 (deviation from the baseline)		.0975 (p<.004)
Airline 5 (deviation from the baseline)		-.0630 (p<.010)
F-test	2419.34 (p<.0000)	3935.79 (p<.0000)
Degrees of freedom (error)	86	81

[14] However, the coefficient of g_1 is not statistically discernable from zero at the .05 level (t=-1.03, p<.304).

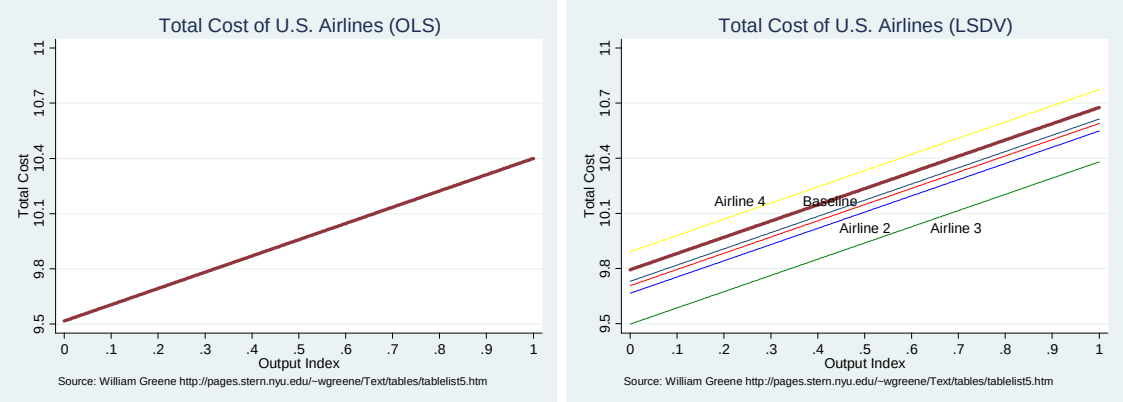
SSE (Sum of squares error)	1.3354	.2926
Root MSE	.1246	.0601
R^2	.9883	.9974
Adjusted R^2	.9879	.9972
N	90	90

Source: http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm

Parameter estimates of regressors show some differences between the pooled OLS and LSDV, but all of them are statistically significant at the .01 level. The pooled OLS reports the overall intercept, while LSDV presents the intercept of the dropped (baseline) and deviations of other five intercepts from the baseline. Large p-values of airline 1 and 2 suggest that the intercepts of airline 1 and 2 are not significantly deviated from the baseline intercept (intercept of airline 6).

Figure 4.1 highlights differences in intercepts between the pooled OLS (left) and LSDV (right). The red line on the left plot is the OLS regression line with the overall intercept of 9.5169. The red line on the right plot is the regression line of airline 6 whose dummy variable was excluded from the model. Other thin lines respectively represent regression lines of airline 1 through 5. For example, the top yellow line has the largest intercept of airline 4, while the bottom green line has the smallest intercept of airline 3.

Figure 4.1.Comparing Pooled OLS and LSDV



Note that the slopes of regression lines are similar in both plots because the coefficient of output index is similar in OLS and LSDV. If loading factor was used, the slopes of these lines would be different.

This eyeballing gives us subjective evidence of fixed group effect, but this evidence is not sufficient in a strong econometric sense. Section 5 will discus a formal test to examine the presence of the fixed effect.

4.4 Estimation Strategies: LSDV1, LSDV2, and LSDV3

The least squares dummy variable (LSDV) regression is ordinary least squares (OLS) with dummy variables. The key issue in LSDV is how to avoid the perfect multicollinearity or so called "dummy variable trap." Each approach has a constraint (restriction) that reduces the number of parameters to be estimated by one and thus makes the model identified. LSDV1 drops a dummy variable; LSDV2 suppresses the intercept; and LSDV3 imposes a restriction. These approaches are different from each other with respect to model estimation and

interpretation of dummy variable parameters (Suits 1984: 177). They produce different dummy parameter estimates, but their results are equivalent. You have to know the pros and cons of these three approaches.

4.4.1 Estimating LSDV1

The first approach, LSDV1, drops a dummy variable as shown in 4.2. That is, the parameter of the eliminated dummy variable is set to zero and is used as a baseline. You should be careful when selecting a variable to be dropped, $d_{dropped}^{LSDV1}$ (g_6 in 4.2), so that it can play a role of the reference group effectively. The functional form of LSDV1 is,

$$\cos t_i = \beta_0 + \beta_1 output_i + \beta_2 fuel_i + \beta_3 loading_i + u_1 g_1 + u_2 g_2 + u_3 g_3 + u_4 g_4 + u_5 g_5 + \varepsilon_i$$

Use the `.regress` command followed by a dependent variable and independent variables including a set of dummies (excluding one of dummies). The coefficient of a dummy included means how far its parameter estimate is away from the reference point or baseline (i.e., the overall intercept).

```
. regress cost g1-g5 output fuel load
```

What if we drop a different dummy variable, say g_1, instead of g_6? Since the different reference point is applied, we will get different dummy coefficients. But other statistics such as parameter estimates of regressors and goodness-of-fit measures remain unchanged. That is, choice of a dummy variable to be dropped does not change the model at all.

```
. regress cost g2-g6 output fuel load
```

```
      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  8,    81) = 3935.79
       Model |   113.74827     8  14.2185338           Prob > F      =  0.0000
    Residual |  .292622872    81  .003612628           R-squared     =  0.9974
-------------+------------------------------           Adj R-squared =  0.9972
       Total |  114.040893    89  1.28135835           Root MSE      =  .06011

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          g2 |  -.0412359   .0251839    -1.64   0.105    -.0913441    .0088722
          g3 |  -.2089211   .0427986    -4.88   0.000    -.2940769   -.1237652
          g4 |   .1845557   .0607527     3.04   0.003     .0636769    .3054345
          g5 |   .0240547   .0799041     0.30   0.764    -.1349293    .1830387
          g6 |   .0870617   .0841995     1.03   0.304     -.080469    .2545924
      output |   .9192846   .0298901    30.76   0.000     .8598126    .9787565
        fuel |   .4174918   .0151991    27.47   0.000     .3872503    .4477333
        load |  -1.070396     .20169    -5.31   0.000    -1.471696   -.6690963
       _cons |   9.705942    .193124    50.26   0.000     9.321686     10.0902
------------------------------------------------------------------------------
```

The intercept 9.7059 in this model is the parameter estimate (Y-intercept) of airline 1, whose dummy variable g_1 was excluded from the model. The coefficient -.0412 indicates the deviation of the intercept of airline 2 from the baseline 9.7059. That is, the intercept of airline 2 is .0412 smaller than the reference point of 9.7059. Therefore, the intercept of airline 2 is computed as 9.6647=9.7059-.0412. Similarly, the intercept of airline 3 is computed as 9.4970=9.7059-.2089.

When you have not created dummy variables, you may use the `.xi` prefix command (interaction expansion) to obtain the identical result.[15]

```
. xi: regress cost i.airline output fuel load

i.airline         _Iairline_1-6       (naturally coded; _Iairline_1 omitted)

      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  8,    81) = 3935.79
       Model |   113.74827     8  14.2185338           Prob > F      =  0.0000
    Residual |  .292622872    81  .003612628           R-squared     =  0.9974
-------------+------------------------------           Adj R-squared =  0.9972
       Total |  114.040893    89  1.28135835           Root MSE      =  .06011

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iairline_2 |  -.0412359   .0251839    -1.64   0.105    -.0913441    .0088722
 _Iairline_3 |  -.2089211   .0427986    -4.88   0.000    -.2940769   -.1237652
 _Iairline_4 |   .1845557   .0607527     3.04   0.003     .0636769    .3054345
 _Iairline_5 |   .0240547   .0799041     0.30   0.764    -.1349293    .1830387
 _Iairline_6 |   .0870617   .0841995     1.03   0.304     -.080469    .2545924
      output |   .9192846   .0298901    30.76   0.000     .8598126    .9787565
        fuel |   .4174918   .0151991    27.47   0.000     .3872503    .4477333
        load |  -1.070396     .20169    -5.31   0.000    -1.471696   -.6690963
       _cons |   9.705942   .193124    50.26   0.000     9.321686     10.0902
------------------------------------------------------------------------------
```

4.4.2 Estimating LSDV2

LSDV2 includes all dummies and, in turn, suppresses the intercept (i.e., set the intercept to zero). Its functional form is,

$$\cos t_i = \beta_1 output_i + \beta_2 fuel_i + \beta_3 loading_i + u_1 g_1 + u_2 g_2 + u_3 g_3 + u_4 g_4 + u_5 g_5 + u_6 g_6 + \varepsilon_i$$

You can fit LSDV2 using `.regress` with the `noconstant` option, which suppresses the intercept in the model. Notice that all group dummies g_1-g_6 are included in the model.

```
. regress cost g1-g6 output fuel load, noconstant

      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  9,    81) =       .
       Model |  16191.3043     9  1799.03381           Prob > F      =  0.0000
    Residual |  .292622872    81  .003612628           R-squared     =  1.0000
-------------+------------------------------           Adj R-squared =  1.0000
       Total |  16191.5969    90  179.906633           Root MSE      =  .06011

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          g1 |   9.705942   .193124    50.26   0.000     9.321686     10.0902
          g2 |   9.664706   .198982    48.57   0.000     9.268794    10.06062
          g3 |   9.497021   .2249584   42.22   0.000     9.049424    9.944618
          g4 |   9.890498   .2417635   40.91   0.000     9.409464    10.37153
          g5 |   9.729997   .2609421   37.29   0.000     9.210804    10.24919
          g6 |   9.793004   .2636622   37.14   0.000     9.268399    10.31761
      output |   .9192846   .0298901    30.76   0.000     .8598126    .9787565
        fuel |   .4174918   .0151991    27.47   0.000     .3872503    .4477333
        load |  -1.070396     .20169    -5.31   0.000    -1.471696   -.6690963
------------------------------------------------------------------------------
```

[15] The Stata `.xi` is used either as an ordinary command or a prefix command. `.xi` creates dummies from a categorical variable specified in the term `i.` and then run the command following the colon. Stata by default drops the first dummy variable.

Find that all parameter estimates of regressors are the same as those in LSDV1. Also the coefficients of six dummies represent their group intercepts; that is, you do not need to compute individual group intercepts. This is the beauty of LSDV2.

LSDV2, however, reports incorrect (inflated) R^2 (1. > .9974) and F (very large > 3935.79). Obviously, the R^2 of 1 are not likely. This is because the X matrix does not, due to the suppressed intercept, have a column vector of 1 and produces incorrect sums of squares of model and total (Uyar and Erdem, 1990: 298). However, the sum of squares of errors (SSE) and their standard errors of parameter estimates are correct in any LSDV.

4.4.3 Estimating LSDV3

LSDV3 includes the intercept and all dummies, and then impose a restriction that the sum of parameters of all dummies is zero. The functional form of LSDV3 is,

$$\cos t_i = \beta_0 + \beta_1 output_i + \beta_2 fuel_i + \beta_3 loading_i + u_1 g_1 + u_2 g_2 + u_3 g_3 + u_4 g_4 + u_5 g_5 + u_6 g_6 + \varepsilon_i,$$
subject to $u_1 + u_2 + u_3 + u_4 + u_5 + u_6 = 0$

In Stata, you need to use both `.constraint` and `.cnsreg` commands to fit LSDV3. `.constraint` defines a constraint, while `.cnsreg` fits a constrained OLS using the `constraint()`option. The number in the parenthesis, 1 in the following example, indicates the constraint number defined in `.constraint`.

```
. constraint define 1 g1 + g2 + g3 + g4 + g5 + g6 = 0

. cnsreg cost g1-g6 output fuel load, constraint(1)

Constrained linear regression                   Number of obs   =         90
                                                F(   8,     81) =    3935.79
                                                Prob > F        =     0.0000
                                                Root MSE        =     0.0601

 ( 1)  g1 + g2 + g3 + g4 + g5 + g6 = 0
------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          g1 |  -.0075859   .0456178    -0.17   0.868    -.0983509    .0831792
          g2 |  -.0488218   .0379787    -1.29   0.202    -.1243875    .0267439
          g3 |  -.2165069   .0160624   -13.48   0.000    -.2484661   -.1845478
          g4 |   .1769698   .0194247     9.11   0.000     .1383208    .2156189
          g5 |   .0164689   .0366904     0.45   0.655    -.0565335    .0894712
          g6 |   .0794759   .0405008     1.96   0.053     -.001108    .1600597
      output |   .9192846   .0298901    30.76   0.000     .8598126    .9787565
        fuel |   .4174918   .0151991    27.47   0.000     .3872503    .4477333
        load |  -1.070396     .20169    -5.31   0.000    -1.471696   -.6690963
       _cons |   9.713528    .229641    42.30   0.000     9.256614    10.17044
------------------------------------------------------------------------------
```

LSDV3 returns the same parameter estimates of regressors and their standard errors as do LSDV1 and LSDV2. Stata `.cnsreg` command does not provide an ANOVA table and goodness-of-fit statistics other than F and square root of MSE.

Unlike LSDV1 and LSDV2, LSDV3 produces the intercept and six dummy coefficients but these coefficients have different meanings. The LSDV3 intercept is the average of individual group intercepts, while a dummy coefficient is the deviation of the group intercept from the averaged intercept. For example, 9.7135= (9.7059+9.6647+9.4970+9.8905+9.7300+9.7930)/6. The coefficient .0165 of airline 5 is the deviation from the averaged intercept 9.7135; that is, 0165=9.7300-9.7135.

4.4.3 Comparing LSDV1, LSDV2, and LSDV3

Three approaches end up fitting the same model and report the same parameter estimates of regressors and their standard errors (Table 4.2). LSDV1 and LSDV3 reports correct goodness-of-fit measures (Stata `.cnsreg` displays F-test and root MSE only), while LSDV2 reports correct SSE and root MSE but returns inflated (incorrect) F-test and R^2. Three LSDV approaches return different, but equivalent (representing the same group intercepts in different manners), dummy coefficients.

The key difference of three approaches lies in the meanings of the intercept and dummy coefficients (Table 4.3). A parameter estimate in LSDV2, δ_d^*, is the actual intercept (Y-intercept) of group *d*. It is easy to interpret substantively. The t-test examines if δ_d^* is zero.

Table 4.2 Comparing Results of LSDV1, LSDV2, and LSDV3

	LSDV1	LSDV2	LSDV3
Ouput index	.9193 (.0299)**	.9193 (.0299)**	.9193 (.0299)**
Fuel price	.4175 (.0152)**	.4175 (.0152)**	.4175 (.0152)**
Loading factor	-1.0704 (.2017)**	-1.0704 (.2017)**	-1.0704 (.2017)**
Intercept (baseline)	9.7930 (.2637)**	0.	9.7135 (.2296)**
Airline 1 (dummy)	-.0871 (.0842)	9.7059 (.1931)**	-.0076 (.0456)
Airline 2 (dummy)	-.1283 (.0757)	9.6647 (.1990)**	-.0488 (.0380)
Airline 3 (dummy)	-.2960 (.0500)**	9.4970 (.2250)**	-.2165 (.0161)**
Airline 4 (dummy)	.0975 (.0330)**	9.8905 (.2418)**	.1770 (.0194)**
Airline 5 (dummy)	-.0630 (.0239)**	9.7300 (.2609)**	.0165 (.0367)
Airline 6 (dummy)	0.	9.7930 (.2637)**	.0795 (.0405)
F-test	3935.79**	Large **	3935.79 **
Degrees of freedom	81	81	81
SSE	.2926	.2926	
Root MSE	.0601	.0601	.0601
R^2	.9974	1.	
Adjusted R^2	.9972	1.	
N	90	90	90

Source: http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm
* Standard errors in parenthesis; Statistical significance: * <.05, ** <.01

In LSDV1, a dummy coefficient shows the extent to which the actual intercept of group *d* deviates from the reference point (the parameter of the dropped dummy variable), which is the intercept of LSDV1, $\delta_{dropped}^* = \alpha^{LSDV1}$. The null hypothesis of t-test is that the deviation from the reference group is zero.
In LSDV3, a dummy coefficient means how far its actual parameter is away from the average group effect (Suits 1984: 178). The LSDV3 intercept is the averaged effect: $\alpha^{LSDV3} = \frac{1}{d}\sum \delta_i^*$.
Therefore, the null hypothesis is that the deviation of a group intercept from the averaged intercept is zero.

In short, each approach has a different baseline and restriction (u_5=0 in LSDV1; regression intercept=1 in LSDV2; and the sum group intercepts is 0) and thus tests a different hypothesis. But all approaches produce equivalent dummy coefficients and exactly the same parameter estimates of regressors. In other word, they all fit the same model; given one LSDV fitted, in other words, we can replicate the other two LSDVs.

Table 4.3. Summary of Dummy Coefficients in LSDV1, LSDV2, and LSDV3

	LSDV1	LSDV2	LSDV3
Dummies included	$d_1^{LSDV1} - d_d^{LSDV1}$ except for $d_{dropped}^{LSDV1}$	$d_1^* - d_d^*$	$d_1^{LSDV3} - d_d^{LSDV3}$
Intercept?	α^{LSDV1}	No	α^{LSDV3}
All dummies?	No (*d-1*)	Yes (*d*)	Yes (*d*)
Constraint (restriction)?	$\delta_{dropped}^{LSDV1} = 0$ (Drop one dummy)	$\alpha^{LSDV2} = 0$ (Suppress the intercept)	$\sum \delta_i^{LSDV3} = 0$ (Impose a restriction)
Actual dummy parameters	$\delta_i^* = \alpha^{LSDV1} + \delta_i^{LSDV1}$, $\delta_{dropped}^* = \alpha^{LSDV1}$	$\delta_1^*, \delta_2^*, \ldots \delta_d^*$	$\delta_i^* = \alpha^{LSDV3} + \delta_i^{LSDV3}$, $\alpha^{LSDV3} = \frac{1}{d}\sum \delta_i^*$
Meaning of a dummy coefficient	How far away from the reference group (dropped)?	Actual individual intercept	How far away from the averaged group effect?
H_0 of the t-test	$\delta_i^* - \delta_{dropped}^* = 0$	$\delta_i^* = 0$	$\delta_i^* - \frac{1}{d}\sum \delta_i^* = 0$

Source: Constructed from Suits (1984) and David Good's lecture (2004)

Which approach is better than the others? You need to consider both estimation and interpretation issues carefully. In general, LSDV1 is often preferred because of easy estimation in statistical software packages. Oftentimes researchers want to see how far dummy parameters deviate from the reference group rather than the actual group intercepts. If you have to report individual group intercepts, LSDV2 gives the answer directly. Finally, LSDV2 and LSDV3 involve some estimation problems; for example, LSDV2 reports an incorrect R^2.

5. Fixed Effect Model

A fixed group model examines group differences in intercepts. The LSDV for this fixed model needs to create as many dummy variables as the number of entities or subjects. When many dummies are needed, the within effect model is useful since it uses transformed variables without creating dummies.

The "within" estimation does not use dummy variables and thus has larger degrees of freedom, smaller MSE, and smaller standard errors of parameters than those of LSDV; therefore, we need to adjust these statistics. Because this estimation does not report individual dummy coefficients either, you need to compute them if really needed. Notice that R^2 reported in the within effect model is incorrect.

5.1 Estimating "Within Estimators" Manually

In order to estimate "within group" estimators manually, you need to compute group means of all dependent variables and regressors. The `quietly` below suppresses the terminal output of the command `.egen`, which produces group means in this case.

```
. quietly egen gm_cost=mean(cost), by(airline)
. quietly egen gm_output=mean(output), by(airline)
. quietly egen gm_fuel=mean(fuel), by(airline)
. quietly egen gm_load=mean(load), by(airline)
```

You will get the following group means of variables. For instance, 14.67563 is the mean of total costs of airline 1 from period 1 through 15.

```
+-------------------------------------------------------+
| airline    gm_cost   gm_output    gm_fuel    gm_load |
|-------------------------------------------------------|
|       1   14.67563    .3192696    12.7318   .5971917 |
|       2   14.37247    -.033027   12.75171   .5470946 |
|       3   13.37231   -.9122626   12.78972   .5845358 |
|       4    13.1358   -1.635174   12.77803   .5476773 |
|       5   12.36304   -2.285681    12.7921   .5664859 |
|       6   12.27441    -2.49898    12.7788   .5197756 |
+-------------------------------------------------------+
```

Then, transform dependent and independent variables to compute their deviations from group means.

```
. quietly gen gw_cost = cost - gm_cost
. quietly gen gw_output = output - gm_output
. quietly gen gw_fuel = fuel - gm_fuel
. quietly gen gw_load = load - gm_load
```

This transformation results in new variables as follows.

```
. list airline year gw_cost-gw_load
```

```
     +-------------------------------------------------------------------+
     | airline   year     gw_cost   gw_output     gw_fuel     gw_load |
     |-------------------------------------------------------------------|
  1. |       1      1   -.7285328    -.367665   -1.154494   -.0627047 |
  2. |       1      2   -.6648102   -.3326011   -1.120779   -.0648637 |
  3. |       1      3   -.5904226   -.2312771   -1.118361   -.0494557 |
  4. |       1      4   -.4469995   -.1573378   -1.020239   -.0563457 |
  5. |       1      5    -.343276   -.1707031   -.5428448   -.0060247 |
     |-------------------------------------------------------------------|
  6. |       1      6   -.2592325   -.1590573   -.2420216   -.0217747 |
  7. |       1      7   -.1555948   -.0642321   -.2501793   -.0026967 |
  8. |       1      8   -.0208158     .010516   -.0670052    .0002173 |
```

```
  9. |       1      9    .1103392    .1586588    .1268826    .0413303 |
 10. |       1     10    .3178005    .2825514    .5202751    .0790953 |
     |-------------------------------------------------------------------|
 11. |       1     11    .4716501    .1164272    .9463253    .0085433 |
 12. |       1     12    .4925508    .1046246    1.080944    .0171683 |
 13. |       1     13     .525177    .1876685    1.019706    .0361743 |
 14. |       1     14    .5945034    .2808353    .9323883    .0529253 |
 15. |       1     15     .697669    .3415919    .8894091    .0284113 |
     |-------------------------------------------------------------------|
 16. |       2      1   -1.120321   -.6196791   -1.201534   -.0562436 |
 17. |       2      2   -1.002288    -.593159   -1.130134   -.0736456 |
 18. |       2      3   -.8084316      -.3898   -1.067656   -.0440816 |
 19. |       2      4   -.5576658   -.2007037   -1.100786   -.0345936 |
 20. |       2      5    -.371335   -.1378266    -.471817    .0196874 |
```

Now, we are ready to run the within effect model with the intercept suppressed. The `noconstant` (or `noc`) option suppresses the intercept.

```
. regress gw_cost gw_output gw_fuel gw_load, noc
```

```
      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  3,    87) = 3871.82
       Model |  39.0683861     3  13.0227954           Prob > F      =  0.0000
    Residual |  .292622861    87  .003363481           R-squared     =  0.9926
-------------+------------------------------           Adj R-squared =  0.9923
       Total |   39.361009    90  .437344544           Root MSE      =    .058

------------------------------------------------------------------------------
     gw_cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   gw_output |   .9192846    .028841    31.87   0.000       .86196    .9766092
     gw_fuel |   .4174918   .0146657    28.47   0.000     .3883422    .4466414
     gw_load |  -1.070396   .1946109    -5.50   0.000    -1.457206   -.6835858
------------------------------------------------------------------------------
```

Compare this output with the LSDV output in 4.2. The within effect model reports correct SSE and parameter estimates of regressors but produces incorrect R^2 and standard errors of parameter estimates. Notice that the degrees of freedom increase from 81 (LSDV) to 87 since six dummy variables are not used.

You may compute group intercepts using $d_i^* = \bar{y}_{i\bullet} - \beta' \bar{x}_{i\bullet}$. For example, the intercept of airline 5 is computed as 9.730 = 12.3630 – {.9193*(-2.2857) + .4175*12.7921 + (-1.0704)*.5665}. In order to get the correct standard errors, you need to adjust them using the ratio of degrees of freedom of the within effect model and LSDV. For example, the standard error of output index is computed as .0299=.0288*sqrt(87/81).

5.2 "Within" Estimation Using `.xtreg`

The Stata `.xtreg` command estimates "within group" estimators without creating dummy variables. Let us first run the `.tsset` command and specifies cross-sectional and time-series variables. Note that both variables should be numeric in `.tsset`.

```
. quietly tsset airline year
```

The `.xtreg` command is followed by a dependent variable, regressors, and options. The `fe` option tells Stata to fit the within effect model.[16]

```
. xtreg cost output fuel load, fe i(airline)
```

[16] `i(airline)` specifies `airline` as the independent unit but this option is redundant because group and time variables are already defined in `.tsset`.

```
Fixed-effects (within) regression               Number of obs      =        90
Group variable: airline                         Number of groups   =         6

R-sq:  within  = 0.9926                         Obs per group: min =        15
       between = 0.9856                                        avg =      15.0
       overall = 0.9873                                        max =        15

                                                F(3,81)            =   3604.80
corr(u_i, Xb)  = -0.3475                        Prob > F           =    0.0000

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .9192846   .0298901    30.76   0.000     .8598126    .9787565
        fuel |   .4174918   .0151991    27.47   0.000     .3872503    .4477333
        load |  -1.070396     .20169    -5.31   0.000    -1.471696   -.6690963
       _cons |   9.713528    .229641    42.30   0.000     9.256614    10.17044
-------------+----------------------------------------------------------------
     sigma_u |   .1320775
     sigma_e |  .06010514
         rho |  .82843653   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(5, 81) =    57.73              Prob > F = 0.0000
```

Compare this within effect model with the LSDV output in 4.2. This command reports correct parameter estimates and their standard errors of regressors but returns incorrect F 3,604.80 and R^2 of .9926.

The F-test in the last line of the output examines the null hypothesis that five dummy parameters in LSDV1 are zero (e.g., $\mu_1=0$, $\mu_2=0$, $\mu_3=0$, $\mu_4=0$, and $\mu_5=0$). The large F statistic reject the null hypothesis in favor of the fixed group effect ($p<.0000$). Recall that the intercept of 9.7135 is the averaged intercept in LSDV3.

By default, `.xtreg` does not display an analysis of variance (ANOVA) table including SSE. Since many related statistics are stored in macros, you need to run `.display` (or `.di`) to get them.[17] The following commands return SSM, SSE, SEE or square root of MSE=**`e(rss)/e(df_r)`**, R^2, and adjusted R^2, respectively. Notice that SEE is reported under the label `sigma_e`.

```
. display e(mss) e(rss) sqrt(e(rss)/e(df_r))
39.068386 .29262287.06010514

. di e(r2) e(r2_a)
.99256567.99183141
```

Alternatively, you may use `.areg` to get the same result except for the correctR^2. Like `.xtreg`, the `.areg` command returns the same intercept, the averaged intercept in LSDV3.

```
. areg cost output fuel load, absorb(airline)

Linear regression, absorbing indicators                Number of obs =      90
                                                       F(  3,    81) = 3604.80
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9974
                                                       Adj R-squared =  0.9972
                                                       Root MSE      =  .06011

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .9192846   .0298901    30.76   0.000     .8598126    .9787565
        fuel |   .4174918   .0151991    27.47   0.000     .3872503    .4477333
```

[17] In order to view the list of macros available in `.xtreg`, run `.help xtreg`.

```
        load |  -1.070396     .20169    -5.31   0.000    -1.471696   -.6690963
       _cons |   9.713528    .229641    42.30   0.000     9.256614    10.17044
-------------+------------------------------------------------------------------
     airline |          F(5, 81) =     57.732   0.000           (6 categories)
```

Let us get SSE from the macro variable `e(rss)`.

```
. di e(rss) e(tss)-e(rss)
.29262287 113.74827
```

Table 5.1 Comparison of OLS, LSDV, and Within Effect Models

	OLS	LSDV	"Within"	.xtreg	.areg
Ouput index	.8827**	.9193**	.9193**	.9193**	.9193**
	(.0133)	(.0299)	(.0288)	(.0299)	(.0299)
Fuel price	.4540**	.4175**	.4175**	.4175**	.4175**
	(.0203)	(.0152)	(.0147)	(.0152)	(.0152)
Loading factor	-1.6275**	-1.0704**	-1.0704**	-1.0704**	-1.0704**
	(.3453)	(.2017)	(.1946)	(.2017)	(.2017)
Intercept (baseline)	9.5169**	9.7930**		9.7135**	9.7135**
	(.2292)	(.2637)		(.2296)	(.2296)
Airline 1 (dummy)		-.0871			
		(.0842)			
Airline 2 (dummy)		-.1283			
		(.0757)			
Airline 3 (dummy)		-.2960**			
		(.0500)			
Airline 4 (dummy)		.0975**			
		(.0330)			
Airline 5 (dummy)		-.0630**			
		(.0239)			
F-test (model)	2419.34**	3935.79**	3871.82**	3604.80**	3604.80**
Degrees of freedom	86	81	81	81	81
SSM (model)	112.7054	113.7483	39.0684	*39.0684*	*113.7483*
SSE (error/residual)	1.3354	.2926	.2926	*.2926*	*.2926*
Root MSE (SEE)	.1246	.0601	.0580	.0601	.0601
R^2	.9883	.9974	.9926	*.9926*	.9974
Adjusted R^2	.9879	.9972	.9923	*.9918*	.9972
F-test (fixed effect)				57.73**	57.732**
N	90	90	90	90	90

Source: http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm
* Standard errors in parenthesis; Statistics hidden in macros are italicized; Statistical significance: * <.05, ** <.01

Table 5.1 contrasts the output of the pooled OLS and four fixed effect estimations (i.e., LSDV, the within effect model, `.xtreg`, and `.areg`). All for estimations produce the same SSE and parameter estimates but reports a bit different standard errors and goodness-of-fit measures. The original within effect model reports incorrect standard errors, F statistics, SEE, and R^2 (See the numbers in red). The estimation using `.xtreg` and `.areg` return adjusted (corrected) standard errors and SEE; conduct F-test for fixed effect; and report the correct averaged intercept and its standard error. However, they report the same, wrong F statistic and do not, by default, display SSE. While `.areg` reports correct (adjusted) R^2, `.xtreg` holds wrong correct (adjusted) R^2 in macro variables.

So which estimation is best for you? LSDV is generally preferred because of correct estimation, goodness-of-fit, and group/time specific intercepts (in particular LSDV2). If the number of entities and/or time periods is large enough, say 100 time periods, `.xtreg` and `.areg` will provide less painful and more elegant solutions including F-test for fixed effects. However, you should keep in mind that they produce an incorrect F score for model test and (adjusted) R^2 (in `.xtreg`). Again DO NOT read F score and R^2 from the `.xtreg` output but get correct ones from LSDV.

If you want to try a random time effect model, add `i(year)` to `.xtreg`. Or switch cross-sectional and time series variables using `.tsset` and then run `.xtreg` again.

```
. quietly xtreg cost output fuel load, fe i(year)

. tsset year airline
       panel variable:  year (strongly balanced)
        time variable:  airline, 1 to 6
                delta:  1 unit

. quietly xtreg cost output fuel load, fe
```

5.3 Testing a Fixed Effect (F-test)

How do we know if there is a significant fixed group effect? The F-test based on loss of fit is the case. The null hypothesis of this F-test is that all dummy parameters except for one are zero: $H_0 : \mu_1 = ... = \mu_{n-1} = 0$.

In order for the F-test, let us obtain the SSE (e'e) of 1.3354 from the pooled OLS regression and .2926 from the LSDVs (LSDV1 through LSDV3). Alternatively, you may draw R^2 of .9974 from LSDV1 or LSDV3 and .9883 from the pooled OLS. DO NOT, however, read R^2 from LSDV2, the original within effect model, or the `.xtreg` output.

The F statistic is computed as,

$$\frac{(1.3354 - .2926)/(6-1)}{(.2926)/(90-6-3)} = \frac{(.9974 - .9883)/(6-1)}{(1-.9974)/(90-6-3)} \sim 57.7319[5,81].$$

53.7319 seems large enough to reject the null hypothesis. We already know that `.xtreg` and `.areg` by default conduct the F-test for fixed effects. Alternatively, we can run the `.test` command, a follow-up command for the Wald test, right after fitting LSDV.

```
. quietly regress cost g1-g5 output fuel load
. test g1 g2 g3 g4 g5

 ( 1)  g1 = 0
 ( 2)  g2 = 0
 ( 3)  g3 = 0
 ( 4)  g4 = 0
 ( 5)  g5 = 0

       F(  5,    81) =   57.73
            Prob > F =    0.0000
```

6. Random Effect Model

A random effect model examines how group and/or time influence error variances. This section discusses the feasible generalized least squares (FGLS) and various estimation methods available in Stata.[18] In order to get θ for FGLS, we need "between" estimation first.

6.1 "Between" Estimation: Group Mean Regression

In a between group effect model, the unit of analysis is not an individual observation, but entity. Accordingly, the number of observations jumps down from nT to n. Since this model uses aggregate group means of variables, it is often called as group mean regression.

Let us compute group means using the `.collapse` command. This command computes aggregate information, group means in this case, and stores into a new data set in memory. Note that /// links two command lines.

```
. collapse (mean) gm_cost=cost (mean) gm_output=output (mean) gm_fuel=fuel (mean)
///
  gm_load=load, by(airline)
```

Individual group means are listed below.

```
. list airline gm_cost-gm_load

     +-------------------------------------------------------+
     | airline    gm_cost   gm_output    gm_fuel    gm_load |
     |-------------------------------------------------------|
  1. |       1   14.67563    .3192696    12.7318   .5971917 |
  2. |       2   14.37247    -.033027   12.75171   .5470946 |
  3. |       3   13.37231   -.9122626   12.78972   .5845358 |
  4. |       4    13.1358   -1.635174   12.77803   .5476773 |
  5. |       5   12.36304   -2.285681    12.7921   .5664859 |
  6. |       6   12.27441    -2.49898    12.7788   .5197756 |
     +-------------------------------------------------------+
```

Now run OLS on these new variables in order to get SSE .0317.

```
. regress gm_cost gm_output gm_fuel gm_load

      Source |       SS       df       MS              Number of obs =       6
-------------+------------------------------           F(  3,     2) =  104.12
       Model |  4.94698124     3  1.64899375           Prob > F      =  0.0095
    Residual |  .031675926     2  .015837963           R-squared     =  0.9936
-------------+------------------------------           Adj R-squared =  0.9841
       Total |  4.97865717     5  .995731433           Root MSE      =  .12585

------------------------------------------------------------------------------
     gm_cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   gm_output |   .7824568   .1087646     7.19   0.019     .3144803    1.250433
     gm_fuel |  -5.523904   4.478718    -1.23   0.343    -24.79427    13.74647
     gm_load |  -1.751072   2.743167    -0.64   0.589    -13.55397    10.05182
       _cons |    85.8081   56.48199     1.52   0.268    -157.2143    328.8305
```

[18] Baltagi and Cheng (1994) introduce various ANOVA estimation methods, such as a modified Wallace and Hussain method, the Wansbeek and Kapteyn method, the Swamy and Arora method, and Henderson's method III. They also discuss maximum likelihood (ML) estimators, restricted ML estimators, minimum norm quadratic unbiased estimators (MINQUE), and minimum variance quadratic unbiased estimators (MIVQUE). Based on a Monte Carlo simulation, they argue that ANOVA estimators are Best Quadratic Unbiased estimators of the variance components for the balanced model, whereas ML, restricted ML, MINQUE, and MIVQUE are recommended for the unbalanced models.

```
------------------------------------------------------------------------------
```

You can also use `be` option in `.xtreg` to fit the same between effect model, but it does not report an ANOVA table. In the following output, R^2 .9936 is reported under the label `between =` and SEE .1258 under `sd(u_i + avg(e_i.))=` .

```
. xtreg cost output fuel load, be i(airline)

Between regression (regression on group means)  Number of obs      =        90
Group variable: airline                         Number of groups   =         6

R-sq:  within  = 0.8808                         Obs per group: min =        15
       between = 0.9936                                        avg =      15.0
       overall = 0.1371                                        max =        15

                                                F(3,2)             =    104.12
sd(u_i + avg(e_i.))=  .1258491                  Prob > F           =    0.0095

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .7824552   .1087663     7.19   0.019     .3144715    1.250439
        fuel |  -5.523978   4.478802    -1.23   0.343    -24.79471    13.74675
        load |  -1.751016    2.74319    -0.64   0.589    -13.55401    10.05198
       _cons |   85.80901   56.48302     1.52   0.268    -157.2178    328.8358
------------------------------------------------------------------------------
```

6.2 Estimating a Random Effect Model Manually

Since the covariance structure of individual *i*, Σ, is not known, we have to estimate θ using the SSEs of the between group effect model (.0317) and the fixed group effect model (.2926). See the formula in 3.4 and computation below.

The variance component of error $\hat{\sigma}_v^2$ is .00361263 = .292622872/(6*15-6-3)

The variance component of group $\hat{\sigma}_u^2$ is .01559712 =.031675926/(6-4) - .00361263/15

Thus, $\hat{\theta}$ is $.87668488 = 1 - \sqrt{\frac{\hat{\sigma}_v^2}{T\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} = 1 - \sqrt{\frac{\hat{\sigma}_v^2}{T\hat{\sigma}_{between}^2}} = 1 - \sqrt{\frac{.00361263}{15 * .031675926/(6-3-1)}}$,

where $\hat{\sigma}_{between}^2 = \frac{SSE_{between}}{n-k-1} = \frac{.031675926}{6-3-1} = .01583796$.

Next, transform the dependent and independent variables including the intercept using $\hat{\theta}$ .8767.

```
. gen rg_cost = cost - .87668488*gm_cost
. gen rg_output = output - .87668488*gm_output
. gen rg_fuel = fuel - .87668488*gm_fuel
. gen rg_load = load - .87668488*gm_load
. gen rg_int = 1 - .87668488 // for the intercept

. list airline year rg_cost-rg_int

     +-------------------------------------------------------------------------+
     | airline   year    rg_cost   rg_output    rg_fuel     rg_load     rg_int |
     |-------------------------------------------------------------------------|
  1. |       1      1   1.081195   -.3282942   .4155294    .0109381   .1233151 |
  2. |       1      2   1.144917   -.2932304   .4492446    .0087791   .1233151 |
  3. |       1      3   1.219305   -.1919063   .4516622    .0241871   .1233151 |
  4. |       1      4   1.362728    -.117967   .5497848    .0172971   .1233151 |
  5. |       1      5   1.466451   -.1313323   1.027179     .067618   .1233151 |
     |-------------------------------------------------------------------------|
  6. |       1      6   1.550495   -.1196865   1.328002    .0518681   .1233151 |
```

```
  7. |       1      7   1.654133   -.0248613   1.319844    .0709461   .1233151 |
  8. |       1      8   1.788912    .0498868   1.503018    .0738601   .1233151 |
  9. |       1      9   1.920067    .1980295   1.696906    .1149731   .1233151 |
 10. |       1     10   2.127528    .3219222   2.090299    .1527381   .1233151 |
     |-------------------------------------------------------------------------|
 11. |       1     11   2.281378     .155798   2.516349    .0821861   .1233151 |
 12. |       1     12   2.302278    .1439954   2.650968    .0908111   .1233151 |
 13. |       1     13   2.334904    .2270392    2.58973    .1098171   .1233151 |
 14. |       1     14   2.404231    .3202061   2.502412     .126568   .1233151 |
 15. |       1     15   2.507396    .3809627   2.459433    .1020541   .1233151 |
     |-------------------------------------------------------------------------|
 16. |       2      1   .6520216   -.6237518    .370944    .0112215   .1233151 |
 17. |       2      2    .770055   -.5972317   .4423447   -.0061806   .1233151 |
 18. |       2      3   .9639112   -.3938727   .5048227    .0233835   .1233151 |
 19. |       2      4   1.214677   -.2047764   .4716921    .0328715   .1233151 |
 20. |       2      5   1.401008   -.1418994   1.100661    .0871525   .1233151 |
```

Finally, run OLS with these transformed variables. Do not forget to add `noconstant` to suppress the OLS intercept.

```
. regress rg_cost rg_output rg_fuel rg_load rg_int, noc

      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  4,    86) =19642.72
       Model |  284.670313     4  71.1675783           Prob > F      =  0.0000
    Residual |  .311586777    86  .003623102           R-squared     =  0.9989
-------------+------------------------------           Adj R-squared =  0.9989
       Total |    284.9819    90  3.16646556           Root MSE      =  .06019

------------------------------------------------------------------------------
     rg_cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   rg_output |   .9066808   .0256249    35.38   0.000     .8557401    .9576215
     rg_fuel |   .4227784   .0140248    30.15   0.000      .394898    .4506587
     rg_load |    -1.0645   .2000703    -5.32   0.000    -1.462226   -.6667731
      rg_int |   9.627911   .2101638    45.81   0.000     9.210119     10.0457
------------------------------------------------------------------------------
```

Parameter estimates are similar to those in the fixed effect model in 4.2 and 5.2. SSE and SEE are .3116 and .0602, respectively. The (adjusted) R^2 reported is .9989 but is not correct because the intercept is suppressed; correct R^2 is .9923.

6.3 Random Effect Model Using `.xtreg`

We need to use the `re` option in `.xtreg` to produce FGLS estimates. The `theta` option reports an estimated theta (.8767). The parameter estimates and their standard errors are the same as those in 6.2. The R^2 of .9925 under the label `within =` is similar to correct .9923.

```
. xtreg cost output fuel load, re theta

Random-effects GLS regression                   Number of obs      =        90
Group variable: airline                         Number of groups   =         6

R-sq:  within  = 0.9925                         Obs per group: min =        15
       between = 0.9856                                        avg =      15.0
       overall = 0.9876                                        max =        15

Random effects u_i ~ Gaussian                   Wald chi2(3)       =  11091.33
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
theta              = .87668503

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .9066805    .025625    35.38   0.000     .8564565    .9569045
        fuel |   .4227784   .0140248    30.15   0.000     .3952904    .4502665
        load |  -1.064499   .2000703    -5.32   0.000    -1.456629    -.672368
       _cons |   9.627909    .210164    45.81   0.000     9.215995    10.03982
-------------+----------------------------------------------------------------
```

```
     sigma_u |  .12488859
     sigma_e |  .06010514
         rho |  .81193816   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The `sigma_u` (σ_u) and `sigma_e` (σ_v) are square roots of the variance components for groups and errors, respectively (.0156=.1249^2, .0036=.0601^2). Note that SEE is .0602 displayed under `sigma_e`.

The label `rho` represents the ratio of individual specific error variance to the composite (entire) error variance; that is, .8119=$.1249^2$ / ($.1249^2$ + $.0601^2$). A large ratio means that individual specific errors account for large proportion of the composite error variance; In this random effect model, for instance, the individual specific error can explain 81 percent of entire composite error variance. Accordingly, this ratio may be interpreted as a goodness-of-fit of random effect model.

The `.xtmixed` command also provides estimation methods for random effects. The `|| airline:,` option (the comma should not be omitted) tells Stata to use the subject variable `airline`. Parameter estimates and their standard errors are slightly different from those in 6.2. Variance components for groups and errors are reported under the labels `sd(_cons)` and `sd(Residual)`.

```
. xtmixed cost output fuel load || airline:,

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood =  105.20458
Iteration 1:   log restricted-likelihood =  105.20458

Computing standard errors:

Mixed-effects REML regression                   Number of obs      =        90
Group variable: airline                         Number of groups   =         6

                                                Obs per group: min =        15
                                                               avg =      15.0
                                                               max =        15


                                                Wald chi2(3)       =  11114.85
Log restricted-likelihood =  105.20458          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .9073166    .025809    35.16   0.000      .856732    .9579013
        fuel |   .4225032   .0140598    30.05   0.000     .3949465      .45006
        load |  -1.064572   .1997763    -5.33   0.000    -1.456126   -.6730179
       _cons |   9.632212    .211559    45.53   0.000     9.217564    10.04686
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
airline: Identity            |
                   sd(_cons) |   .1293723   .0429029      .0675403    .2478107
-----------------------------+------------------------------------------------
                sd(Residual) |   .0600715   .0047138       .051508    .0700588
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   107.49 Prob >= chibar2 = 0.0000
```

Both `.xtreg` and `.xtmixed` commands with the `mle` option support maximum likelihood estimation. The following two commands produce the same result. Notice that error variance components are computed as .0130=1141^2 and .0035 = .0591^2.

```
. xtreg cost output fuel load, re mle

Random-effects ML regression                    Number of obs      =        90
Group variable: airline                         Number of groups   =         6

Random effects u_i ~ Gaussian                   Obs per group: min =        15
                                                               avg =      15.0
                                                               max =        15

                                                LR chi2(3)         =    436.32
Log likelihood  =  114.72896                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .9053099   .0253759    35.68   0.000     .8555741    .9550458
        fuel |   .4233757    .013888    30.48   0.000     .3961557    .4505957
        load |  -1.064456    .196231    -5.42   0.000    -1.449062   -.6798506
       _cons |   9.618648    .206622    46.55   0.000     9.213677    10.02362
-------------+----------------------------------------------------------------
    /sigma_u |   .1140843   .0345293                      .0630373    .2064687
    /sigma_e |   .0591072   .0045701                      .0507956    .0687787
         rho |   .7883772   .1047419                      .5365302    .9344669
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)=  105.92 Prob>=chibar2 = 0.000

. xtmixed cost output fuel load || airline:, mle

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood =  114.72896
Iteration 1:   log likelihood =  114.72896

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =        90
Group variable: airline                         Number of groups   =         6

                                                Obs per group: min =        15
                                                               avg =      15.0
                                                               max =        15


                                                Wald chi2(3)       =  11552.23
Log likelihood =  114.72896                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .9053099    .024656    36.72   0.000     .8569851    .9536348
        fuel |   .4233757   .0136369    31.05   0.000      .396648    .4501035
        load |  -1.064456   .1962309    -5.42   0.000    -1.449062   -.6798508
       _cons |   9.618648   .2026097    47.47   0.000     9.221541    10.01576
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
airline: Identity            |
                  sd(_cons) |   .1140844   .0345293      .0630373    .2064689
-----------------------------+------------------------------------------------
               sd(Residual) |   .0591071   .0045701      .0507956    .0687787
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   105.92 Prob >= chibar2 = 0.0000
```

If you want to try a random time effect model, add `i(year)` to `.xtreg` as shown in 5.3

```
. xtreg cost output fuel load, re i(year) theta
```

Table 6.1 Comparison of OLS and Various Random Effect Estimations

	OLS	Random Effect	.xtreg	.xtmixed	.xtreg mle
Ouput index	.8827**	.9067**	.9067**	.9073**	.9053**
	(.0133)	(.0256)	(.0256)	(.0258)	(.0254)
Fuel price	.4540**	.4228**	.4228**	.4225**	.4234**
	(.0203)	(.0140)	(.0140)	(.0141)	(.0139)
Loading factor	-1.6275**	-1.0645**	-1.0645**	-1.0646**	-1.0646**
	(.3453)	(.2001)	(.2001)	(.1998)	(.1962)
Intercept	9.5169**	9.6279**	9.6279**	9.6322**	9.6186**
	(.2292)	(.2102)	(.2102)	(.2116)	(.2066)
F, Wald, LR test	2419.34**	19642.72**	11091.33**	11114.85**	436.32**
SSM (model)	112.7054	284.6703			
SSE	1.3354	.3116			
SEE or $\hat{\sigma}_v$	.1246	.0602	.0601	.0601	.0591
$\hat{\sigma}_u$		.1249	.1249	.1294	.1141
θ		.8767	.8767		
R^2	.9883	.9989	.9925		
Adjusted R^2	.9879	.9989			
LR Test				107.49**	105.92**
N	90	90	90	90	90

Source: http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm
* Standard errors in parenthesis; Statistical significance: * <.05, ** <.01

The `.xtreg` without `mle` produces correct parameter estimates and their standard errors of the random effect model but incorrect R^2. The `.xtmixed` command without `mle` employes restricted maximum likelihood (REML) estimation and report slightly different parameter estimates and their standard errors. The `.xtreg` and `.xtmixed` commands with `mle` return the same full information maximum likelihood (FIML) estimates, which are slightly different from the first two methods. The maximum likelihood estimation conducts the likelihood ratio (LR) test to examine random effects.

6.4 Testing a Random Effect: LM test

The Breusch-Pagan Lagrange multiplier (LM) test examines if any random effect exists. The null hypothesis is that individual-specific or time-specific error variance components are zero: $H_0 : \sigma_u^2 = 0$. If the null hypothesis is not rejected, the pooled OLS is preferred; otherwise, the random effect model is better. See the formula in 3.5.2.

In order for the LM test, we need to know $e'e$, SSE of the pooled OLS, and $\bar{e}'\bar{e}$, the sum of squared group specific residuals. The $e'e$ of the pooled OLS is 1.33544153 and $\bar{e}'\bar{e}$ .0665147 is computed as follows.

```
. quietly regress cost output fuel load // run pooled OLS
. quietly predict r, resid              // calculate residuals and save into r
. collapse (mean) gm_r=r, by(airline)   // calculate group means of r
. quietly gen gm_r2=gm_r^2              // calculate squared group means of r

. list

     +--------------------------------+
     | airline        gm_r      gm_r2 |
     |--------------------------------|
  1. |       1    .0688689   .0047429 |
  2. |       2   -.0138781   .0001926 |
  3. |       3   -.1942235   .0377228 |
```

```
  4. |       4    .1527256   .0233251 |
  5. |       5   -.0215835   .0004658 |
  6. |       6    .0080906   .0000655 |
     +--------------------------------+
```

```
. tabstat gm_r2, stat(sum)       // obtain the sum of squared group means of r

    variable |       sum
-------------+----------
       gm_r2 |  .0665147
-----------------------
```

Finally the LM test is,

$$334.8496 = \frac{6*15}{2(15-1)}\left[\frac{15^2 * .0665}{1.3354} - 1\right]^2 \sim \chi^2(1).$$

With the large chi-squared of 334.8496, we reject the null hypothesis in favor of the random group effect model (p <.0000). In Stata, run the `.xttest0` command right after estimating the one-way random effect model in order to get the same result.

```
. quietly xtreg cost output fuel load, re i(airline)

. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        cost[airline,t] = Xb + u[airline] + e[airline,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                 ---------+-----------------------------
                     cost |   1.281358       1.131971
                        e |   .0036126       .0601051
                        u |   .0155972       .1248886

        Test:   Var(u) = 0
                             chi2(1) =   334.85
                          Prob > chi2 =     0.0000
```

7. Hausman Test and Chow Test

If you find both significant fixed and random effects, which effect is more significant and which model is better than the other? The Hausman specification test can answer this question by comparing fixed and random effects. What if you come to think that individual slopes of regressors are not constant but vary across airline or time? A poolability test will give you an answer.

Table 6.1 summarizes the results of pooled OLS, fixed effect, and random effect model. We may ask, "Which model is better than the others?"

7.1 Hausman Test To Choose Fixed or Random Effect

How do we compare a fixed effect model and its random counterpart? The Hausman specification test examines if the individual effects are uncorrelated with other regressors in the model. If individual effects are correlated with any other regressor, the random effect model violates a Gauss-Markov assumption and is no longer Best Linear Unbiased Estimate (BLUE). It is because individual effects are parts of the error term in a random effect model.

Therefore, if the null hypothesis is rejected, a fixed effect model is favored over the random counterpart. In a fixed effect model, individual effects are parts of the intercept and the correlation between the intercept and regressors does not violate any Gauss-Markov assumption; a fixed effect model is still BLUE.

Let us first check cross-sectional and time-series variables to make sure we are doing fine.

```
. tsset airline year
       panel variable:  airline (strongly balanced)
        time variable:  year, 1 to 15
                delta:  1 unit
```

The Hausman test requires that both fixed and random effect models are fitted. `.estimate store` saves the result of the random effect model into `random_group`.

```
. quietly xtreg cost output fuel load, re
. quietly estimates store random_group

. quietly xtreg cost output fuel load, fe
. quietly estimates store fixed_group

. hausman random_group fixed_group
```

Run the `.hausman` command followed by random and fixed effect results in order.

```
. hausman random_group fixed_group

                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |  random_group fixed_group     Difference          S.E.
-------------+------------------------------------------------------------
      output |    .9066805     .9192846       -.0126041               .
        fuel |    .4227784     .4174918        .0052867               .
        load |   -1.064499    -1.070396        .0058974               .
------------------------------------------------------------------------------
                           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)
                        =    -2.12    chi2<0 ==> model fitted on these
                                      data fails to meet the asymptotic
                                      assumptions of the Hausman test;
                                      see suest for a generalized test
```

Alternatively, you may replace `fixed_group` with a period (.) indicating last fitted model, the fixed effect model in this case.

```
. hausman random_group .
```

The Hausman test returns -2.12 and but warns that data fails to meet the asymptotic assumptions. Although the chi-squares score is small enough not to reject the null hypothesis, we may not conclude that the random effect model is better than its fixed counterpart; the test is not conclusive.

7.2 Chow Test for Poolability

The poolability test, here Chow test, examines if panel data are poolable so that the slopes of regressors are the same across individual entities or time periods (Bantagi, 2001: 51-55). If the null hypothesis of poolability is rejected, individual airlines may have their own slopes of regressors and then fixed and/or random effects are no longer appealing. Instead, you may try random coefficient model or hierarchical regression model that is skipped in the working paper.

In order for poolability test, we need to run group by group (or time by time) OLS regressions. In Stata, the `forvalues` loop and `if` qualifier make it easy to run group by group regressions. Open the Stata do editor by running `.doedit` at the dot prompt, type in the following commands, and then execute them.

```
forvalues i= 1(1)6 {
    display "OLS regression for group " `i'
    regress cost output fuel load if airline==`i'
}
```

```
OLS regression for group 1

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) = 1843.46
       Model |  3.41824348     3  1.13941449           Prob > F      =  0.0000
    Residual |  .006798918    11  .000618083           R-squared     =  0.9980
-------------+------------------------------           Adj R-squared =  0.9975
       Total |   3.4250424    14  .244645886           Root MSE      =  .02486

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |    1.18318   .0968946    12.21   0.000     .9699164    1.396444
        fuel |   .3865867   .0181946    21.25   0.000     .3465406    .4266329
        load |  -2.461629   .4013571    -6.13   0.000     -3.34501   -1.578248
       _cons |     10.846   .2972551    36.49   0.000     10.19174    11.50025
------------------------------------------------------------------------------

OLS regression for group 2

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) = 3129.50
       Model |  6.47622084     3  2.15874028           Prob > F      =  0.0000
    Residual |  .007587838    11  .000689803           R-squared     =  0.9988
-------------+------------------------------           Adj R-squared =  0.9985
       Total |  6.48380868    14  .463129191           Root MSE      =  .02626

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
      output |   1.459104   .0792856    18.40   0.000     1.284597     1.63361
        fuel |   .3088958   .0272443    11.34   0.000     .2489315      .36886
        load |  -2.724785   .2376522   -11.47   0.000    -3.247854   -2.201716
       _cons |   11.97243   .4320951    27.71   0.000     11.02139    12.92346
------------------------------------------------------------------------------

OLS regression for group 3

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) =  608.10
       Model |  3.79286673     3  1.26428891           Prob > F      =  0.0000
    Residual |  .022869767    11   .00207907           R-squared     =  0.9940
-------------+------------------------------           Adj R-squared =  0.9924
       Total |   3.8157365    14  .272552607           Root MSE      =   .0456

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .7268305   .1554418     4.68   0.001     .3847054    1.068956
        fuel |   .4515127   .0381103    11.85   0.000     .3676324    .5353929
        load |  -.7513069   .6105989    -1.23   0.244    -2.095226    .5926122
       _cons |   8.699815   .8985786     9.68   0.000     6.722057    10.67757
------------------------------------------------------------------------------

OLS regression for group 4

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) =  777.86
       Model |  7.37252558     3  2.45750853           Prob > F      =  0.0000
    Residual |  .034752343    11  .003159304           R-squared     =  0.9953
-------------+------------------------------           Adj R-squared =  0.9940
       Total |  7.40727792    14   .52909128           Root MSE      =  .05621

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .9353749   .0759266    12.32   0.000     .7682616    1.102488
        fuel |   .4637263    .044347    10.46   0.000     .3661192    .5613333
        load |  -.7756708   .4707826    -1.65   0.128    -1.811856    .2605148
       _cons |   9.164608   .6023241    15.22   0.000     7.838902    10.49031
------------------------------------------------------------------------------

OLS regression for group 5

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) = 1999.89
       Model |  7.08313716     3  2.36104572           Prob > F      =  0.0000
    Residual |  .012986435    11  .001180585           R-squared     =  0.9982
-------------+------------------------------           Adj R-squared =  0.9977
       Total |  7.09612359    14  .506865971           Root MSE      =  .03436

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   1.076299   .0771255    13.96   0.000     .9065471    1.246051
        fuel |   .2920542   .0434213     6.73   0.000     .1964845    .3876239
        load |  -1.206847   .3336308    -3.62   0.004    -1.941163   -.4725305
       _cons |   11.77079   .7430078    15.84   0.000     10.13544    13.40614
------------------------------------------------------------------------------

OLS regression for group 6

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) = 2602.49
       Model |  11.1173565     3  3.70578551           Prob > F      =  0.0000
    Residual |  .015663323    11  .001423938           R-squared     =  0.9986
-------------+------------------------------           Adj R-squared =  0.9982
       Total |  11.1330199    14  .795215705           Root MSE      =  .03774

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   .9673393   .0321728    30.07   0.000     .8965275    1.038151
        fuel |   .3023258   .0308235     9.81   0.000     .2344839    .3701678
        load |   .1050328   .4767508     0.22   0.830    -.9442886    1.154354
       _cons |   10.77381   .4095921    26.30   0.000     9.872309    11.67532
```

```
------------------------------------------------------------------------------
```

The null hypothesis of the poolability test across group is that all slopes of regressors are the same across group: $H_0 : \beta_{ik} = \beta_k$ for 1… *i*th group and 1… *k*th regressor.

The SSE of the pooled OLS regression, which represented by $e'e$, is 1.3354. And the sum of SSEs of group by group regression, $e_i'e_i$, is .1007 = .0068 + .0076 + .0229 + .0348 + .0130 + .0157. The F statistic is,

$$\frac{(1.3354 - .1007/(6-1)(3+1)}{.1007/6(15-3-1)} \sim 40.4812[20,66]$$

The large 40.4812 rejects the null hypothesis of poolability (p< .0000). We conclude that the panel data are not poolable with respect to airline. Both fixed and random effect models may be misleading and we need to try random coefficient model or hierarchical linear regression model.[19]

The following `.xtrc` estimates Swamy's (1970) random-coefficients linear regression model and `betas` presents group specific slopes. Theoretical discussion and interpretation of the result are skipped.

```
. xtrc cost output fuel load, betas

Random-coefficients regression                  Number of obs      =        90
Group variable: airline                         Number of groups   =         6

                                                Obs per group: min =        15
                                                               avg =      15.0
                                                               max =        15

                                                Wald chi2(3)       =   1377.99
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        cost |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      output |   1.062236   .1073229     9.90   0.000     .8518869    1.272585
        fuel |    .368607    .034347    10.73   0.000     .3012881    .4359258
        load |  -1.347659   .4734776    -2.85   0.004    -2.275658   -.4196602
       _cons |   10.54412   .5911727    17.84   0.000     9.385443     11.7028
------------------------------------------------------------------------------
Test of parameter constancy:    chi2(20) =   661.80       Prob > chi2 = 0.0000

                        Group-specific coefficients
------------------------------------------------------------------------------
                     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
     Group 1 |
-------------+----------------------------------------------------------------
      output |   1.179249   .0923552    12.77   0.000     .9982361    1.360262
        fuel |    .381204   .0164192    23.22   0.000     .3490229    .4133851
        load |  -2.315307   .4146611    -5.58   0.000    -3.128028   -1.502586
       _cons |   10.82836    .297623    36.38   0.000     10.24503    11.41169
-------------+----------------------------------------------------------------
     Group 2 |
-------------+----------------------------------------------------------------
      output |   1.411403   .0783685    18.01   0.000     1.257803    1.565002
        fuel |   .3247732   .0268443    12.10   0.000     .2721594     .377387
        load |  -2.661805   .2609131   -10.20   0.000    -3.173185   -2.150424
       _cons |   11.73411   .4343784    27.01   0.000     10.88274    12.58547
-------------+----------------------------------------------------------------
```

[19] However, this Chow test may be problematic when errors do not follow a normal distribution: $\varepsilon \sim N(0,\Omega)$ instead of $\varepsilon \sim N(0,\sigma^2 I)$. See Bantagi (2001: 52-55) for extensive discussion on this issue.

```
     Group 3 |
-------------+----------------------------------------------------------------
      output |   .8686628   .1183818     7.34   0.000     .6366388    1.100687
        fuel |   .4207733   .0325501    12.93   0.000     .3569763    .4845703
        load |  -1.128314   .4015202    -2.81   0.005    -1.915279   -.3413491
       _cons |   9.443223   .6731321    14.03   0.000     8.123908    10.76254
-------------+----------------------------------------------------------------
     Group 4 |
-------------+----------------------------------------------------------------
      output |   .9344674     .07332    12.75   0.000     .7907628    1.078172
        fuel |   .4466961   .0379257    11.78   0.000     .3723631    .5210291
        load |  -.6759248   .3282028    -2.06   0.039     -1.31919   -.0326592
       _cons |   9.325361   .5744807    16.23   0.000     8.199399    10.45132
-------------+----------------------------------------------------------------
     Group 5 |
-------------+----------------------------------------------------------------
      output |   1.019989   .0623505    16.36   0.000     .8977842    1.142194
        fuel |   .3173035   .0378956     8.37   0.000     .2430295    .3915774
        load |  -1.036449    .276928    -3.74   0.000    -1.579218   -.4936802
       _cons |   11.22227   .6038026    18.59   0.000     10.03884     12.4057
-------------+----------------------------------------------------------------
     Group 6 |
-------------+----------------------------------------------------------------
      output |   .9596449   .0345695    27.76   0.000     .8918899      1.0274
        fuel |   .3208917   .0300269    10.69   0.000     .2620401    .3797434
        load |  -.2681565   .3158239    -0.85   0.396    -.8871599     .350847
       _cons |    10.7114   .4216672    25.40   0.000     9.884945    11.53785
------------------------------------------------------------------------------
```

8. Presenting Panel Data Models

The key question now is, "Which information do we have to report? And how?" Some studies report parameter estimates and their statistical significances only; Others include standard errors but exclude goodness-of-fit measures; And oftentimes researchers fail to interpret the results substantively for readers. This section discusses general guidelines for presenting panel data models. However, specific pieces of information to be presented and their styles depend on research questions and purpose of studies.

8.1 Presenting All Possible Models? No!

Some studies present all possible models including the pooled OLS, fixed effect model, random effect models, and two-way effect model. Is this practice reasonable? No. Strictly speaking, if one model is "right," the other models are "wrong." It must be absurd to present wrong models together unless comparison of models is the goal of the study. If a fixed effect turns out insignificant, why are you trying to present the "wrong" model? In short, you just need to report a "right" model or your final model only.

8.2 Which Information Should Be Reported?

You MUST report goodness-of-fit measures, parameter estimates with their standard errors, and test results (See Table 8.1).

8.2.1 Goodness-of-fit Measures

Goodness-of-fit examines the extant that the model fits data. In case of poor goodness-of-fit, you need to try other model. The essential goodness-of-fit measures that you need to report are,

- F-test (or likelihood rato test) to test the model and its significance (p-value).
- Sum of squared errors (residual), degrees of freedom for errors, and N (nT).
- R^2 in OLS and fixed effect models.
- Theta θ and variance components $\hat{\sigma}_u$ estimated in a random effect model.

Keep in mind that some estimation methods report incorrect statistics and standard errors. For example, `.xtreg` returns incorrect R^2 in a fixed effect model because the command fits the "within" estimator (running OLS on transformed data with the intercept suppressed). Both between R^2 and overall R^2 displayed in Stata output are almost meaningless. In order to get correct R^2 for a fixed effect model, use LSDV1 or `.areg`. Use macro variables, if needed, to obtain various goodness-of-fit measures that are not displayed in Stata output.

8.2.2 Parameter Estimates of Regressors

Obviously, you must report parameter estimates and their standard errors.
Fortunately, `.regress` and `.xtreg` produce correct parameter estimates and their adjusted standard errors. But the "within" estimation itself produces incorrect standard errors due to incorrect (larger) degrees of freedom (see Table 3.2).

8.2.3 Parameter Estimates of Dummy Variables

In a fixed effect model, a question is if individual intercepts need to be reported. In general, parameter estimates of regressors are of primary interest in most cases and accordingly individual intercepts are not needed. However, you have to report them if audience wants to know or individual effects are of main research interest. The combination of LSDV1 or LSDV2 will give you easy solutions for this case (see 4.2).[20] Do not forget that LSDV1, LSDV2, and LSDV3 have different meanings of dummy parameters and that null hypotheses of t-test differ from one another (see Table 4.3).

8.2.4 Test Results

Finally, you should report if fixed and/or random effect exists because panel data modeling is to examine fixed and/or random effects. Report and interpret the results of F-test for a fixed effect model and/or Breusch-Pagan LM test for a random effect model. When both fixed and random effects are statistically significant, you need to conduct a Hausman test and report its result.[21] If you doubt constant slopes across group and/or time, conduct a Chow test to examine the poolability of data.

Table 8.1 Examples of Presenting Analysis Results

	Pooled OLS	Fixed Effect Model	Random Effect Model
Ouput index	.8827**	.9193**	.9067**
	(.0133)	(.0299)	(.0256)
Fuel price	.4540**	.4175**	.4228**
	(.0203)	(.0152)	(.0140)
Loading factor	-1.6275**	-1.0704**	-1.0645**
	(.3453)	(.2017)	(.2001)
Intercept (baseline)	9.5169**	9.7930**	9.6279**
	(.2292)	(.2637)	(.2102)
Airline 1 (dummy)		-.0871	
		(.0842)	
Airline 2 (dummy)		-.1283	
		(.0757)	
Airline 3 (dummy)		-.2960**	
		(.0500)	
Airline 4 (dummy)		.0975**	
		(.0330)	
Airline 5 (dummy)		-.0630*	
		(.0239)	
F-test (model)	2419.34**	3935.79**	19642.72**
DF	86	81	81
R^2	.9883	.9974	
SSE (SRMSE)	1.3354	.2926	.3116
SEE or $\hat{\sigma}_v$	.1246	.0601	.0602
$\hat{\sigma}_u$			.1249
θ			.8767
Effect Test		57.7319**	334.8496**
N	90	90	90

Source: http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm
* Standard errors in parenthesis; ** Statistical significance: * <.05, **<.01

[20] In some uncommon cases, you need to report variance components estimated in a random effect model. Unlike the SAS MIXED procedure, `.xtreg` does not report these statistics.

[21] The null hypothesis is that group/time specific effects are not correlated with any regressors. Either "A random effect model is better than the fixed effect model" or "A random effect model is consistent" is NOT a correct null hypothesis. If the null hypothesis is rejected, a random effect model violates a key OLS assumption 2 and ends up with biased and inconsistent estimates; however, a fixed effect model still remains unbiased and consistent.

8.3 Interpreting Results Substantively

If your model fits the data well and individual regressors turn out statistically significant, you have to interpret parameter estimators in a "sensical" way. You may not simply report signs and magnitude of coefficients. Do not simply say, for example, an independent variable is "significant," "negatively (or positively) related to…", or "insignificantly related..."

A standard form of interpretation is, "For one unit increase in an IV, DV is expected to increase by OO units, holding all other variables constant." You may omit the *ceteris paribus* assumption (holding all other variables constant). However, try to make interpretation more sense to audience who does not know much about econometrics. See 4.1 and 4.2 for examples of substantive interpretation.

Provide statistical significance in a table and the p-value in parenthesis at the end of the interpretation sentence.

8.4 Presenting Results Professionally

Many studies often present results in tables but some of them fail to construct professional tables. Common bad table examples include 1) large and various fonts, 2) too small and/or too large numbers, 3) colorful and stylish border lines, 4) badly aligned numbers, and 5) non-systematic order. The following is the list of checkpoints to be considered when constructing a professional table (see Table 8.1 for an example).

- Title should describe the contents of a table appropriately. Provide unit of measurement (e.g., Million Dollars) and period (e.g., Year 2010) if needed.
- Organize a table systemically and compactly.
- Provide parameter estimates and their standard errors in parenthesis.
- Do not use variable names used in computer software as labels. Use loading factor instead of `load`.
- Use 10 point *Times New Roman* for labels and 10 point *Courier New* for numbers. Do not use stylish fonts (e.g., *Cooperplate*) and too big or too small size.
- Rescale numbers appropriately in order to avoid such numbers as "0.00004455" or "75,845,341,697,785."
- Report up to three or four digits below the decimal point. Do not round numbers arbitrarily.
- Do not use stylish border lines (e.g., their colors, thickness, and type of lines).
- Minimize use of vertical and horizontal lines. Use no vertical line in general.
- Align numbers to the right and consider the location of decimal point carefully.
- Use "Standardized coefficients," if needed, rather than "Beta," "β," or "beta coefficients." Nobody knows the true value of β.
- Provide the source of data, if applicable, at the bottom of the table.
- Indicate statistical significance as $^{*}<.05$, $^{**}<.01$.

8.5 Common Mistakes and Awkward Expressions

It is not difficult to find awkward expressions even in academic papers. Consider following suggestions for common mistakes in presentation.

8.5.1 Statistical Significance

Do not say, "significant level," "at 5% level," or "at the level of significance α=5%," and the like. These expressions should be "significance level," "at the .05 level," and "at the .05 (significance) level," respectively. Use a specific significance level (e.g., "at the .01 significance level") rather than "at the conventional level."

8.5.2 Hypothesis

A hypothesis is a conjecture about the unknown (e.g, α, β, δ, and σ). Therefore, "$b_1 = 0$" is not a valid hypothesis, but "$\beta_1 = 0$" is. Because the b_1 is already known (estimated from the sample), you do not need to test $b_1 = 0$.

8.5.3 Parameter Estimates

Say, "parameter estimates of β_1" or "the coefficient of an independent variable 1" instead of "The coefficient of β_1." Also say, "standardized coefficients" instead of "Beta," β, or "beta coefficient."

8.5.4 P-values

Do not say, "The p-value is significant." A p-value itself is neither significant nor insignificant. You may say, "The p-value is small enough to reject H_0" or "A small p-value suggests rejection of H_0."

8.5.5 Reject or Do Not Reject the Null Hypothesis

Say, "reject" or "do not reject" the null hypothesis rather than "accept (or confirm)" the null hypothesis. Also say "reject the H_0 at the .01 level" instead of "I do not believe that the H_0 is true" or "The test provides decisive evidence that the H_0 is wrong" (no one knows if a H_0 is really true or wrong). Always be simple and clear.

9. Conclusion

Panel data are analyzed to investigate individual (group) and/or time effects using fixed effect and random effect models. A fixed effect model asks how heterogeneity from group and/or time affects individual intercepts, while a random effect model hypothesizes error variance structures affected by group and/or time. Disturbances in a random effect model are assumed to be randomly distributed across group or time. But the key difference between fixed and random effect models is that individual effect u_i in a random effect model should not correlated with any regressor. Slopes are assumed unchanged in both fixed effect and random effect models.

A panel data set needs to be arranged in the long form as shown in 2.3. Longitudinal data are balanced or unbalanced, fixed or rotating, and long or short. If data are severely unbalanced, too long, or too short, read output with caution and, in case of an unbalanced panel, consider dropping subjects with many missing data points. If the number of groups (subjects) or time periods is extremely large, you may consider categorizing subjects to reduce the number of groups or time periods.

A fixed effect model is estimated by the least squares dummy variable (LSDV) regression and "within" estimation. LSDV has three approaches to avoid perfect multicollinearity. LSDV1 drops a dummy; LSDV2 suppresses the intercept; and LSDV3 includes all dummies and imposes a restriction instead. LSDV1 is commonly used since it produces correct statistics. LSDV2 provides actual individual intercepts, but reports incorrect R^2 and F score. Remember that the dummy parameters of three LSDV approaches have different meanings and thus conduct different t-tests.

The "within" estimation does not use dummy variables but deviations from group means. Thus, this estimation is useful when there are many groups and/or time periods in the panel data set since it is able to avoid the incidental parameter problem. In turn, time-invariant independent variables are wiped out in the data transformation process and the dummy parameter estimates need to be computed afterward. Because of its larger degrees of freedom, the "within" estimation produces incorrect R^2 and standard errors of parameters although Stata reports adjusted standard errors.

Fixed effect (F test)	Random effect (B-P LM test)	Your Selection
H_0 is not rejected (No fixed effect)	H_0 is not rejected (No random effect)	Pooled OLS
H_0 is rejected (fixed effect)	H_0 is not rejected (No random effect)	Fixed effect model
H_0 is not rejected (No fixed effect)	H_0 is rejected (random effect)	Random effect model
H_0 is rejected (fixed effect)	H_0 is rejected (random effect)	Choose a fixed effect model if the null hypothesis of a Hausman test is rejected; otherwise, fit a random effect model.

In order to determine an appropriate model for a panel, first describe data carefully by producing summary statistics and drawing plots. Then begin with a simple model like the pooled OLS.

Imagine four possible outcomes of hypothesis testing shown in the table above. If both null hypotheses of F-test and LM test are not rejected, your best model is the pooled OLS. If the null hypothesis of an F-test in a fixed effect model is rejected and the null of a Breusch-Pagan LM test in a random effect model is not, a fixed effect model is the case. If you find both significant fixed and random effects in your panel data, conduct a Hausman specification test and compares a fixed effect model and a random effect model. If the null hypothesis of uncorrelation between individual effects and regressors is rejected, fit a random effect model; otherwise, a fixed effect model is preferred.

If you think that your data are not poolable and each entity has different slopes of regressors, conduct a Chow test and then, if its null hypothesis is rejected, try to fit a random coefficient model or hierarchical linear model. For details about model selection, see 3.6.

It is important to present the result correctly. The essential information includes goodness-of-fit measures (e.g., F score and likelihood ratio, SSE, and R^2), parameter estimates with their standard errors, and test results (i.e., F-test, LM test, Hausman test, and Chow test). These pieces of information should be presented in a professional table. Researchers should interpret the results substantively so that audience without sophisticated econometric knowledge can understand.

References

Baltagi, Badi H. 2001. *Econometric Analysis of Panel Data*. Wiley, John & Sons.

Baltagi, Badi H., and Young-Jae Chang. 1994. "Incomplete Panels: A Comparative Study of Alternative Estimators for the Unbalanced One-way Error Component Regression Model." *Journal of Econometrics*, 62(2): 67-89.

Breusch, T. S., and A. R. Pagan. 1980. "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics." *Review of Economic Studies*, 47(1):239-253.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cameron, A. Colin, and Pravin K. Trivedi. 2009. *Microeconometrics Using Stata*. TX: Stata Press.

Chow, Gregory C. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica*, 28 (3): 591–605.

Freund, Rudolf J., and Ramon C. Littell. 2000. *SAS System for Regression*, 3rd ed. Cary, NC: SAS Institute.

Fuller, Wayne A. and George E. Battese. 1973. "Transformations for Estimation of Linear Models with Nested-Error Structure." *Journal of the American Statistical Association*, 68(343) (September): 626-632.

Fuller, Wayne A. and George E. Battese. 1974. "Estimation of Linear Models with Crossed-Error Structure." *Journal of Econometrics*, 2: 67-78.

Greene, William H. 2007. *LIMDEP Version 9.0 Econometric Modeling Guide 1*. Plainview, New York: Econometric Software.

Greene, William H. 2008. *Econometric Analysis*, 6th ed. Upper Saddle River, NJ: Prentice Hall.

Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica*, 46(6):1251-1271.

Kennedy, Peter. 2008. *A Guide to Econometrics*, 6th ed. Malden, MA: Blackwell Publishing

SAS Institute. 2004. *SAS/ETS 9.1 User's Guide*. Cary, NC: SAS Institute.

SAS Institute. 2004. *SAS/STAT 9.1 User's Guide*. Cary, NC: SAS Institute.

Stata Press. 2010. *Stata Base Reference Manual, Release 11*. College Station, TX: Stata Press.

Stata Press. 2010. *Stata Longitudinal/Panel Data Reference Manual, Release 11*. College Station, TX: Stata Press.

Suits, Daniel B. 1984. "Dummy Variables: Mechanics V. Interpretation." *Review of Economics & Statistics*, 66 (1):177-180.

Swamy, P. A. V. B. 1970. "Efficient Inference in a Random Coefficient Regression Model." *Econometrica*, 38: 311-323.

Uyar, Bulent, and Orhan Erdem. 1990. "Regression Procedures in SAS: Problems?" *American Statistician*, 44(4): 296-301.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.