

# Linear Regression Models for Panel Data Using SAS, Stata, LIMDEP, and SPSS\*

Hun Myoung Park (kucc625)

*This document summarizes linear regression models for panel data and illustrates how to estimate each model using SAS 9.1, Stata 10, LIMDEP 9, and SPSS 16. This document does not address nonlinear models (i.e., logit and probit models), but focuses on linear regression models.*

1. Introduction
2. Least Squares Dummy Variable Regression
3. Panel Data Models
4. The Fixed Group Effect Model
5. The Fixed Time Effect Model
6. The Fixed Group and Time Effect Model
7. Random Effect Models
8. The Poolability Test
9. Conclusion

## 1. Introduction

Panel (or longitudinal) data are cross-sectional and time series. U.S. Census Bureau's Census 2000 data at state or county level are cross-sectional, while the cumulative census data are considered panel data. Annual sales figures of Apple Computer Inc. for the past 30 years are time series, but they are not cross-sectional at all. The cumulative General Social Survey (GSS), American National Election Studies (ANES), and Current Population Survey (CPS) data are not panel data in the sense that individual respondents vary across survey years. Panel data may have group effects, time effects, or the both. These effects are analyzed by fixed effect and random effect models.

### 1.1 Data Arrangement

A panel data set contains a series of observations per each of  $n$  entities (e.g., firms and states). Each entity (or subject) includes  $T$  observations (1 through  $t$  time period). Thus, the total number of observations is  $nT$ .

Panel data have a cross-sectional (or group) variable and a time-series variable. In Stata, this arrangement is called the long form (as opposed to the wide form) with a group and a time

---

\* The citation of this document should read: "Park, Hun Myoung. 2008. *Linear Regression Models for Panel Data Using SAS, Stata, LIMDEP, and SPSS*. Technical Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University."

variables. Look at the following data set to see how panel data are arranged. There are 15 time periods. See Appendix for the details.

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

. list airline year load cost output fuel in 1/20, sep(15)
```

	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 different format like the wide form, you need to rearrange data first. Stata has the `.reshape` command to rearrange a data set back and forth between the long and wide form. The following changes from the current long form to wide one. Note the wide form has only six observations that have as many variables as the time period (4\*15 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

If you wish to rearrange the data set back to the long form, run the following command.

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

## 1.2 Fixed Effect versus Random Effect Models

Panel data models examine fixed and/or random effects of group of time. The core difference between fixed and random effect models lies in the role of dummies. If dummies are considered

as a part of the intercept, this is a fixed effect model. In a random effect model, the dummies act as an error term (see Table 1).

The fixed effect model examines group differences in intercepts, assuming the same slopes and constant variance across groups. Fixed effect models use least squares dummy variable (LSDV), within effect, and between effect estimation methods. Thus, ordinary least squares (OLS) regressions with dummies, in fact, are fixed effect models.

Table 1. Fixed Effect and Random Effect Models

	Fixed Effect Model	Random Effect Model
Functional form*	$y_{it} = (\alpha + \mu_i) + X'_{it}\beta + v_{it}$	$y_{it} = \alpha + X'_{it}\beta + (\mu_i + v_{it})$
Intercepts	Varying across groups and/or times	Constant
Error variances	Constant	Varying across groups and/or times
Slopes	Constant	Constant
Estimation	LSDV, within effect, between effect	GLS, FGLS
Hypothesis test	Incremental F test	Breusch-Pagan LM test

\*  $v_{it} \sim IID(0, \sigma_v^2)$

The random effect model, by contrast, estimates variance components for groups and error, assuming the same intercept and slopes. The difference among groups (or time periods) lies in the variance of the error term. This model is estimated by generalized least squares (GLS) when the  $\Omega$  matrix, a variance structure among groups, is known. The feasible generalized least squares (FGLS) method is used to estimate the variance structure when  $\Omega$  is not known. A typical example is the groupwise heteroscedastic regression model (Greene 2003). There are various estimation methods for FGLS including maximum likelihood methods and simulations (Baltagi and Cheng 1994).

Fixed effects are tested by the (incremental) F test, while random effects are examined by the Lagrange Multiplier (LM) test (Breusch and Pagan 1980). If the null hypothesis is not rejected, the pooled OLS regression is favored. The Hausman specification test (Hausman 1978) compares fixed effect and random effect models. Table 1 compares the fixed effect and random effect models.

If one grouping 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 group and/or time variables.

### 1.3 Estimation and Software Issues

The LSDV regression, within effect model, between effect model (group or time mean model), GLS, and FGLS are fundamentally based on OLS in terms of estimation. Thus, any procedure and command for OLS is good for linear panel data models.

The REG procedure of SAS/STAT, Stata `.regress (.cnsreg)`, LIMDEP `regress$`, and SPSS `regression` commands all fit LSDV1 by dropping one dummy and have options to suppress the intercept (LSDV2). SAS, Stata, and LIMDEP can estimate OLS with restrictions (LSDV3),

but SPSS cannot. Note that the Stata `.cnsreg` command requires the `.constraint` command that defines a restriction (Table 2).

SAS, Stata, and LIMDEP also provide the procedures (commands) that estimate panel data models in a convenient way. SAS/ETS has the TSCSREG and PANEL procedures to estimate one-way and two-way fixed and random effect models.<sup>1</sup> For the fixed effect model, these procedures estimate LSDV1, which drops one of dummy variables. For the random effects model, they by default use the Fuller-Battese method (1974) to estimate variance components for group, time, and error. These procedures also support other estimation methods such as Parks (1967) autoregressive model and Da Silva moving average method.

Table 2. Procedures and Commands in SAS, Stata, LIMDEP, and SPSS

	SAS	Stata	LIMDEP	SPSS
Regression (OLS)	PROC REG	<code>.regress</code>	Regress\$	Regression
LSDV1	w/o a dummy	w/o a dummy	w/o a dummy	w/o a dummy
LSDV2	/NOINT	Noconstant	w/o One in Rhs	/Origin
LSDV3	RESTRICT	<code>.cnsreg</code>	Cls:	N/A
Fixed effect (within effect)	TSCSREG /FIXONE PANEL /FIXONE	<code>.xtreg w/ fe</code> <code>.areg w/ abs</code>	Regress;Panel;St r=;Pds=;Fixed\$	N/A
Two-way fixed (within effect)	TSCSREG /FIXTWO PANEL /FIXTWO	N/A	Regress;Panel;St r=;Pds=;Period=\$	N/A
Between effect	PANEL /BTWNG PANEL /BTWNT	<code>.xtreg w/ be</code>	Regress;Panel;St r=;Pds=;Means\$	N/A
Random effect	TSCSREG /RANONE PANEL /RANONE	<code>.xtreg w/ re</code> <code>.xtmixed</code>	Regress;Panel;St r=;Pds=;Random\$	N/A
Two-way random	TSCSREG /RANTWO PANEL /RANTWO	N/A	Regress;Panel;St r=;Pds=;Period=\$	N/A

The TSCSREG procedure can handle balanced data only, whereas the PANEL procedure is able to deal with balanced and unbalanced data. The former provides one-way and two-way fixed and random effect models, while the latter supports the between effect model (`/BTWNT` and `/BTWNG`) and pooled OLS regression (`/POOLED`) as well. The PANEL procedure has BP and BP2 options to conduct the Breusch-Pagan LM test for random effects, while TSCSREG does not.<sup>2</sup> Despite advanced features of PANEL, the output of the two procedures is similar.

The Stata `.xtreg` command estimates within effect (fixed effect) models with the `fe` option, between effect models with the `be` option, and random effect models with the `re` option. This command, however, does not fit two-way fixed and random effect models.<sup>3</sup> The `.areg` command with `absorb` option, equivalent to the `.xtreg` with the `fe` option, fits the one-way within effect model that has a large dummy variable set. A random effect model can be also estimated using the `.xtmixed` command. Stata also has `.xtgls` that fits panel data models with heteroscedasticity across groups and/or autocorrelation within groups.

<sup>1</sup> SAS recently announced the PROC PANEL, an experimental procedure, for panel data models.

<sup>2</sup> However, BP and BP2 produce invalid Breusch-Pagan statistics in cases of unbalanced data.  
[http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/etsug\\_panel\\_sect041.htm](http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/etsug_panel_sect041.htm).

<sup>3</sup> However, you may fit the 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`.

The LIMDEP `regress$` command with the `panel` subcommand estimates panel data models. SPSS has limited ability to analyze panel data. Table 2 summarizes procedures and commands used for panel data analysis.

## 2. Least Squares Dummy Variable Regression

A dummy variable is a binary variable that is coded either 1 or zero. It is commonly used to examine group and time effects in regression. Consider a simple model of regressing R&D expenditure in 2002 on 2000 net income and firm type. The dummy variable  $d1$  is set to 1 for equipment and software firms and zero for telecommunication and electronics. The variable  $d2$  is coded in the opposite way. Take a look at the data structure (Figure 2).

Figure 2. Dummy Variable Coding for Firm Type

firm	rnd	income	type	d1	d2
Samsung	2,500	4,768	Electronics	0	1
AT&T	254	4,669	Telecom	0	1
IBM	4,750	8,093	IT Equipment	1	0
Siemens	5,490	6,528	Electronics	0	1
Verizon	.	11,797	Telecom	0	1
Microsoft	3,772	9,421	Service & S/W	1	0
...	...	...	...	...	...

### 2.1 Model 1 without a Dummy Variable

The ordinary least squares (OLS) regression without dummy variables, a pooled regression model, assumes a constant intercept and slope regardless of firm types. In the following regression equation,  $\beta_0$  is the intercept;  $\beta_1$  is the slope of net income in 2000; and  $\varepsilon_i$  is the error term.

$$\text{Model 1: } R \text{ \& } D_i = \beta_0 + \beta_1 \text{income}_i + \varepsilon_i$$

The pooled model has the intercept of 1,482.697 and slope of .223. For a \$ one million increase in net income, a firm is likely to increase R&D expenditure in 2002 by \$ .223 million.

. regress rnd income

Source	SS	df	MS	Number of obs = 39		
Model	15902406.5	1	15902406.5	F( 1, 37) =	7.07	
Residual	83261299.1	37	2250305.38	Prob > F =	0.0115	
				R-squared =	0.1604	
				Adj R-squared =	0.1377	
Total	99163705.6	38	2609571.2	Root MSE =	1500.1	

rnd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.2230523	.0839066	2.66	0.012	.0530414	.3930632
_cons	1482.697	314.7957	4.71	0.000	844.8599	2120.533

Pooled model:  $R\&D = 1,482.697 + .223 \cdot \text{income}$

Despite moderate goodness of fit statistics such as F and t, this is a naïve model. R&D investment tends to vary across industries.

## 2.2 Model 2 with a Dummy Variable

You may assume that equipment and software firms have more R&D expenditure than other types of companies. Let us take this group difference into account.<sup>4</sup> We have to drop one of the two dummy variables in order to avoid perfect multicollinearity. That is, OLS does not work with both dummies in a model. The  $\delta_1$  in model 2 is the coefficient that is valid in equipment and software companies only.

$$\text{Model 2: } R \& D_i = \beta_0 + \beta_1 \text{income}_i + \delta_1 d_{1i} + \varepsilon_i$$

Unlike Model 1, this model results in two different regression equations for two groups. The difference lies in the intercepts, but the slope remains unchanged.

. regress rnd income d1

Source	SS	df	MS			
Model	24987948.9	2	12493974.4	Number of obs =	39	
Residual	74175756.7	36	2060437.69	F( 2, 36) =	6.06	
				Prob > F =	0.0054	
				R-squared =	0.2520	
				Adj R-squared =	0.2104	
				Root MSE =	1435.4	
Total	99163705.6	38	2609571.2			

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
rnd						
income	.2180066	.0803248	2.71	0.010	.0551004	.3809128
d1	1006.626	479.3717	2.10	0.043	34.41498	1978.837
_cons	1133.579	344.0583	3.29	0.002	435.7962	1831.361

$$d1=1: R\&D = 2,140.205 + .218 * \text{income} = 1,113.579 + 1,006.626 * 1 + .218 * \text{income}$$

$$d1=0: R\&D = 1,133.579 + .218 * \text{income} = 1,113.579 + 1,006.626 * 0 + .218 * \text{income}$$

The slope .218 indicates a positive impact of two-year-lagged net income on a firm's R&D expenditure. Equipment and software firms on average spend \$1,007 million more for R&D than telecommunication and electronics companies.

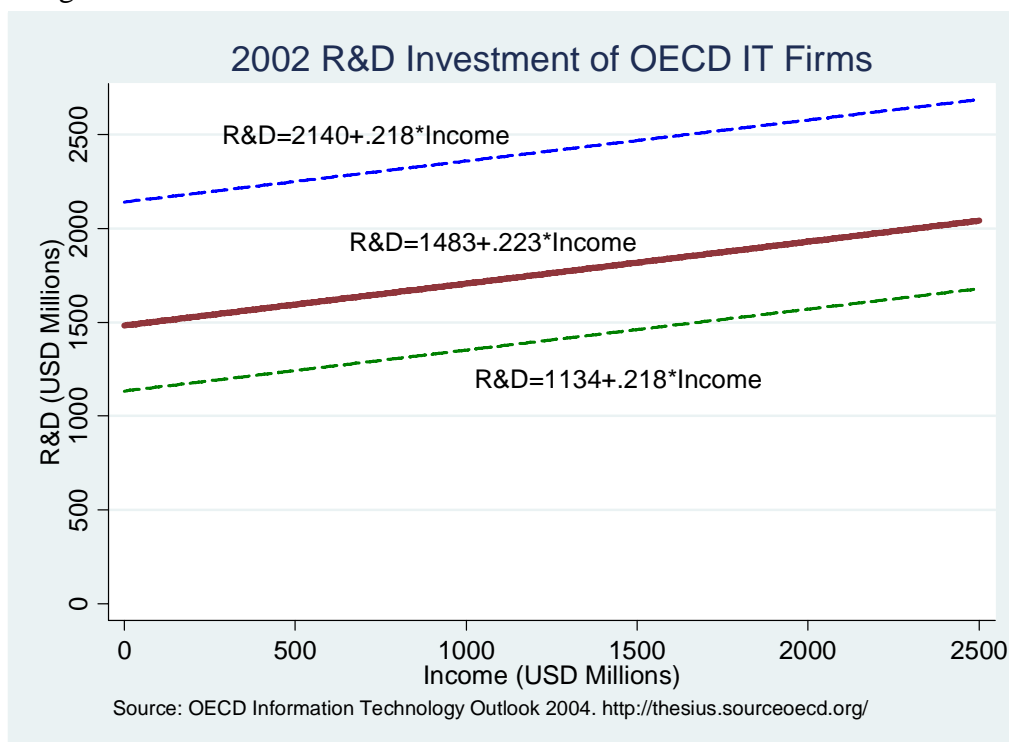
## 2.3 Visualization of Model 1 and 2

There is only a tiny difference in the slope (.223 versus .218) between Model 1 and Model 2. The intercept 1,483 of Model 1, however, is quite different from 1,134 for equipment and software companies and 2,140 for telecommunications and electronics in Model 2. This result appears to support Model 2.

Figure 3 highlights differences between Model 1 and 2 more clearly. The red line (pooled) in the middle is the regression line of Model 1; the blue line at the top is one for equipment and software companies ( $d1=1$ ) in Model 2; finally the green line at the bottom is for telecommunication and electronics firms ( $d2=1$  or  $d1=0$ ).

<sup>4</sup> The dummy variable (firm types) and regressors (net income) may or may not be correlated.

Figure 3. Regression Lines of Model 1 and Model 2



This plot shows that Model 1 ignores the group difference, and thus reports the misleading intercept. The difference in the intercept between two groups of firms looks substantial. Moreover, the two models have the similar slopes. Consequently, Model 2 considering fixed group effects seems better than the simple Model 1. Compare goodness of fit statistics (e.g.,  $F$ ,  $t$ ,  $R^2$ , and SSE) of the two models. See Section 3.2.2 and 4.7 for formal hypothesis testing.

## 2.4 Alternatives to LSDV1

The least squares dummy variable (LSDV) regression is ordinary least squares (OLS) with dummy variables. The critical issue in LSDV is how to avoid the perfect multicollinearity or the so called “dummy variable trap.” LSDV has three approaches to avoid getting caught in the trap. They produce different parameter estimates of dummies, but their results are equivalent.

The first approach, LSDV1, drops a dummy variable as in Model 2 above. The second approach includes all dummies and, in turn, suppresses the intercept (LSDV2). Finally, include the intercept and all dummies, and then impose a restriction that the sum of parameters of all dummies is zero (LSDV3). Take a look at the following functional forms to compare these three LSDVs.

$$\text{LSDV1: } R \& D_i = \beta_0 + \beta_1 \text{income}_i + \delta_1 d_{1i} + \varepsilon_i \text{ or } R \& D_i = \beta_0 + \beta_1 \text{income}_i + \delta_2 d_{2i} + \varepsilon_i$$

$$\text{LSDV2: } R \& D_i = \beta_1 \text{income}_i + \delta_1 d_{1i} + \delta_2 d_{2i} + \varepsilon_i$$

$$\text{LSDV3: } R \& D_i = \beta_0 + \beta_1 \text{income}_i + \delta_1 d_{1i} + \delta_2 d_{2i} + \varepsilon_i, \text{ subject to } \delta_1 + \delta_2 = 0$$

The main differences among these approaches exist in the meanings of the dummy variable parameters. Each approach defines the coefficients of dummy variables in different ways (Table 3). The parameter estimates in LSDV2 are actual intercepts of groups, making it easy to interpret substantively. LSDV1 reports differences from the reference point (dropped dummy variable). LSDV3 computes how far parameter estimates are away from the average group effect. Accordingly, null hypotheses of t-tests in the three approaches are different. Keep in mind that the  $R^2$  of LSDV2 is not correct. Table 3 contrasts the three LSDVs.

Table 3. Three Approaches of Least Squares Dummy Variable Models

	LSDV1: Drop one dummy	LSDV2: Suppress the intercept	LSDV3: Impose a restriction
Dummy included	$\alpha^a, d_2^a - d_d^a$	$d_1^* - d_d^*$	$\alpha^c, d_1^c - d_d^c$
Intercept?	Yes	No	Yes
All dummy?	No (d-1)	Yes (d)	Yes (d)
Restriction?	No	No	$\sum d_i^c = 0^*$
Dummy coefficients	$d_i^* = \alpha^a + d_i^a,$ $d_{dropped}^* = \alpha^a$	$d_1^*, d_2^*, \dots, d_d^*$	$d_i^* = \alpha^c + d_i^c,$ where $\alpha^c = \frac{1}{d} \sum d_i^*$
Meaning of a dummy coefficient	How far away from the reference point (dropped)?	Fixed group effect	How far away from the average group effect?
$H_0$ of T-test	$d_i^* - d_{dropped}^* = 0$	$d_i^* = 0$	$d_i^* - \frac{1}{d} \sum d_i^* = 0$

Source: Constructed from David Good's Lecture (2004)

\* This restriction reduces the number of parameters to be estimated, making the model identified.

## 2.5 Estimating Three LSDVs

The SAS REG procedure, Stata `.regress` command, LIMDEP `Regress$` command, and SPSS `Regression` command all fit OLS and LSDVs. Let us estimate three LSDVs using SAS and Stata.

### 2.5.1 LSDV 1 without a Dummy

LSDV 1 drops a dummy variable. The intercept is the actual parameter estimate of the dropped dummy variable. The coefficient of the dummy included means how far its parameter estimate is away from the reference point or baseline (i.e., the intercept).

Here we include d2 instead of d1 to see how a different reference point changes the result. Check the sign of the dummy coefficient included and the intercept. Dropping other dummies does not make any significant difference.

```
PROC REG DATA=masil.rnd2002;
    MODEL rnd = income d2;
RUN;
```

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: rnd

Number of Observations Read	50
Number of Observations Used	39
Number of Observations with Missing Values	11

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24987949	12493974	6.06	0.0054
Error	36	74175757	2060438		
Corrected Total	38	99163706			

Root MSE	1435.42248	R-Square	0.2520
Dependent Mean	2023.56410	Adj R-Sq	0.2104
Coeff Var	70.93536		

## Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	2140.20468	434.48460	4.93	<.0001
income	1	0.21801	0.08032	2.71	0.0101
d2	1	-1006.62593	479.37174	-2.10	0.0428

d2=0: R&D = 2,140.205 + .218\*income = 2,140.205 - 1,006.626\*0 + .218\*income

d2=1: R&D = 1,133.579 + .218\*income = 2,140.205 - 1,006.626\*1 + .218\*income

Alternatively you may use the GLM and MIXED procedures to get the same result.

```
PROC GLM DATA=masil.rnd2002;
  MODEL rnd = income d2 /SOLUTION;
RUN;
```

```
PROC MIXED DATA=masil.rnd2002;
  MODEL rnd = income d2 /SOLUTION;
RUN;
```

## 2.5.2 LSDV 2 without the Intercept

LSDV 2 includes all dummy variables and suppresses the intercept. The Stata `.regress` command has the `noconstant` option to fit LSDV2. The coefficients of dummies are actual parameter estimates; thus, you do not need to compute intercepts of groups. This LSDV, however, reports wrong  $R^2$  (.7135  $\neq$  .2520).

```
. regress rnd income d1 d2, noconstant
```

Source	SS	df	MS			
Model	184685604	3	61561868.1	Number of obs =	39	
Residual	74175756.7	36	2060437.69	F( 3, 36) =	29.88	
Total	258861361	39	6637470.79	Prob > F =	0.0000	
				R-squared =	0.7135	
				Adj R-squared =	0.6896	
				Root MSE =	1435.4	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.2180066	.0803248	2.71	0.010	.0551004	.3809128
d1	2140.205	434.4846	4.93	0.000	1259.029	3021.38
d2	1133.579	344.0583	3.29	0.002	435.7962	1831.361

d1=1: R&D = 2,140.205 + .218\*income

d2=1: R&D = 1,133.579 + .218\*income

### 2.5.3 LSDV 3 with a Restriction

LSDV 3 includes the intercept and all dummies and then imposes a restriction on the model. The restriction is that the sum of all dummy parameters is zero. The Stata `.constraint` command defines a constraint, while the `.cnsreg` command fits a constrained OLS using the `constraint()` option. The number in the parenthesis indicates the constraint number defined in the `.constraint` command.

```
. constraint 1 d1 + d2 = 0
. cnsreg rnd income d1 d2, constraint(1)
```

Constrained linear regression				Number of obs = 39		
				F( 2, 36) =	6.06	
				Prob > F =	0.0054	
				Root MSE =	1435.4	
( 1) d1 + d2 = 0						

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.2180066	.0803248	2.71	0.010	.0551004	.3809128
d1	503.313	239.6859	2.10	0.043	17.20749	989.4184
d2	-503.313	239.6859	-2.10	0.043	-989.4184	-17.20749
_cons	1636.892	310.0438	5.28	0.000	1008.094	2265.69

d1=1: R&D = 2,140.205 + .218\*income = 1,637 + 503 \*1 + (-503)\*0 + .218\*income

d2=1: R&D = 1,133.579 + .218\*income = 1,637 + 503 \*0 + (-503)\*1 + .218\*income

The intercept is the average of actual parameter estimates:  $1,636 = (2,140+1,133)/2$ . In the SAS output below, the coefficient of RESTRICT is virtually zero and, in theory, should be zero.

```
PROC REG DATA=masil.rnd2002;
  MODEL rnd = income d1 d2;
  RESTRICT d1 + d2 = 0;
RUN;
```

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: rnd

NOTE: Restrictions have been applied to parameter estimates.

Number of Observations Read				50	
Number of Observations Used				39	
Number of Observations with Missing Values				11	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24987949	12493974	6.06	0.0054
Error	36	74175757	2060438		
Corrected Total	38	99163706			
Root MSE		1435.42248	R-Square	0.2520	
Dependent Mean		2023.56410	Adj R-Sq	0.2104	
Coeff Var		70.93536			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	1636.89172	310.04381	5.28	<.0001
income	1	0.21801	0.08032	2.71	0.0101
d1	1	503.31297	239.68587	2.10	0.0428
d2	1	-503.31297	239.68587	-2.10	0.0428
RESTRICT	-1	1.81899E-12	0	.	.

\* Probability computed using beta distribution.

Table 4 compares how SAS, Stata, LIMDEP, and SPSS conducts LSDVs. SPSS is not able to fit the LSDV3. In LIMDEP, the  $b(2)$  of the `cls:` indicates the parameter estimate of the second independent variable. In SPSS, pay attention to the `/ORIGIN` option for LSDV2.

Table 4. Estimating Three LSDVs Using SAS, Stata, LIMDEP, and SPSS

	LSDV 1	LSDV 2	LSDV 3
<b>SAS</b>	PROC REG; MODEL rnd = income d2; RUN;	PROC REG; MODEL rnd = income d1 d2 /NOINT; RUN;	PROC REG; MODEL rnd = income d1 d2; RESTRICT d1 + d2 = 0; RUN;
<b>Stata</b>	. regress ind income d2	. regress rnd income d1 d2, noconstant	. constraint 1 d1+ d2 = 0 . cnsreg rnd income d1 d2 const(1)
<b>LIMDEP</b>	REGRESS; Lhs=rnd; Rhs=ONE,income, d2\$	REGRESS; Lhs=rnd; Rhs=income, d1, d2\$	REGRESS; Lhs=rnd; Rhs=ONE,income, d1, d2; Cls: b(2)+b(3)=0\$
<b>SPSS</b>	REGRESSION /MISSING LISTWISE /STATISTICS COEFF R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT rnd /METHOD=ENTER income d2.	REGRESSION /MISSING LISTWISE /STATISTICS COEFF R ANOVA /CRITERIA=PIN(.05) POUT(.10) /ORIGIN /DEPENDENT rnd /METHOD=ENTER income d1 d2.	N/A

### 3. Panel Data Models

Panel data models examine group effects, time effects, or both. These effects are either fixed effect or random effect. A *fixed effect model* examines if intercepts vary across groups or time periods, whereas a *random effect model* explores differences in error variances. A one-way model includes only one set of dummy variables (e.g., firm), while a two-way model considers two sets of dummy variables (e.g., firm and year). Model 2 in Chapter 2, in fact, is a one-way fixed group effect panel data model.

#### 3.1 Functional Forms and Notation

The parameter estimate of a dummy variable is a part of the intercept in a fixed effect model and a part of error in the random effect model. Slopes remain the same across groups or time periods. The functional forms of one-way panel data models are as follows.

Fixed group effect model:  $y_{it} = (\alpha + \mu_i) + X'_{it}\beta + v_{it}$ , where  $v_{it} \sim IID(0, \sigma_v^2)$

Random group effect model:  $y_{it} = \alpha + X'_{it}\beta + (\mu_i + v_{it})$ , where  $v_{it} \sim IID(0, \sigma_v^2)$

Note that errors are *independent identically distributed*,  $v_{it} \sim IID(0, \sigma_v^2)$ .

Notations used in this document include,

- $\bar{y}_i$ : dependent variable (DV) mean of group  $i$ .
- $\bar{x}_t$ : means of independent variables (IVs) at time  $t$ .
- $\bar{y}_{..}$  and  $\bar{x}_{..}$  for overall means of the DV and IVs, respectively.
- $n$ : the number of groups or firms
- $T$ : the number of time periods
- $N=nT$ : total number of observations
- $k$ : the number of regressors excluding dummy variables
- $K=k+1$  (including the intercept)

#### 3.2 Fixed Effect Models

There are several strategies for estimating fixed effect models. The *least squares dummy variable model (LSDV)* uses dummy variables, whereas the *within effect model* does not. These strategies, of course, produce the identical parameter estimates of non-dummy independent variables. The *between effect model* fits the model using group means of dependent and independent variables without dummies. Table 5 summarizes pros and cons of these models.

##### 3.2.1 Estimations: LSDV, Within Effect, and Between Effect Models

As discussed in Chapter 2, LSDV is widely used because it is relatively easy to estimate and interpret substantively. This LSDV, however, becomes problematic when there are many

groups or subjects in panel data. If  $T$  is fixed and  $N \rightarrow \infty$ , only coefficients of regressors are consistent. The coefficients of dummy variables,  $\alpha + \mu_i$ , are not consistent since the number of these parameters increases as  $N$  increases (Baltagi 2001). This is so called the *incidental parameter problem*. Under this circumstance, LSDV is useless, calling for another strategy, the within effect model.

The within effect model does not need dummy variables, but it uses deviations from group means. Thus, this model is the OLS of  $(y_{it} - \bar{y}_{i\bullet}) = \beta'(x_{it} - \bar{x}_{i\bullet}) + (\varepsilon_{it} - \bar{\varepsilon}_{i\bullet})$  without an intercept.<sup>5</sup> The incidental parameter problem is no longer an issue. The parameter estimates of the within effect model are identical to those of LSDV. The within effect model in turn has several disadvantages.

Since this model does not report dummy coefficients, you need to compute them using the formula  $d_g^* = \bar{y}_{g\bullet} - \beta' \bar{x}_{g\bullet}$ . Since no dummy is used, the within effect model has larger degrees of freedom for error, resulting in small MSE (mean square error) and incorrect (larger) standard errors of parameter estimates. Thus, you have to adjust the standard error using the formula

$$se_k^* = se_k \sqrt{\frac{df_{error}^{Within}}{df_{error}^{LSDV}}} = se_k \sqrt{\frac{nT - k}{nT - n - k}}. \text{ Finally, } R^2 \text{ of the within effect model is not correct}$$

because the intercept is suppressed.

Table 5. Comparison of Fixed Effect Models

	LSDV1	Within Effect	Between Effect
Functional form	$y_i = i\alpha_i + X_i\beta + \varepsilon_i$	$y_{it} - \bar{y}_{i\bullet} = x_{it} - \bar{x}_{i\bullet} + \varepsilon_{it} - \bar{\varepsilon}_{i\bullet}$	$\bar{y}_{i\bullet} = \alpha + \bar{x}_{i\bullet} + \varepsilon_i$
Dummy	Yes	No	No
Dummy coefficient	Presented	Need to be computed	N/A
Transformation	No	Deviation from the group means	Group means
Intercept (estimation)	Yes	No	No
$R^2$	Correct	Incorrect	
SSE	Correct	Correct	
MSE	Correct	Smaller	
Standard error of $\beta$	Correct	Incorrect (smaller)	
$DF_{error}$	$nT - n - k$	$nT - k$ (larger)	$n - K$
Observations	$nT$	$nT$	$n$

The between group effect model, so called the group mean regression, uses group means of the dependent and independent variables. This data aggregation reduces the number of observations down to  $n$ . Then, run OLS of  $\bar{y}_{i\bullet} = \alpha + \bar{x}_{i\bullet} + \varepsilon_i$ . Table 5 contrasts LSDV, the within effect model, and the between group models. In two-way fixed effect model, LSDV2 and the between effect model are not valid.

### 3.2.2 Testing Group Effects

<sup>5</sup> You need to follow three steps: 1) compute group means of the dependent and independent variables; 2) transform variables to get deviations from the group means; 3) run OLS with the transformed variables without the intercept.

The null hypothesis is that all dummy parameters except one are zero:  $H_0 : \mu_1 = \dots = \mu_{n-1} = 0$ . This hypothesis is tested by the F test, which is based on loss of goodness-of-fit. The robust model in the following formula is LSDV and the efficient model is the pooled regression.<sup>6</sup>

$$\frac{(e'e_{Efficient} - e'e_{Robust})/(n-1)}{(e'e_{Robust})/(nT-n-k)} = \frac{(R_{Robust}^2 - R_{Efficient}^2)/(n-1)}{(1-R_{Robust}^2)/(nT-n-k)} \sim F(n-1, nT-n-k)$$

If the null hypothesis is rejected, you may conclude that the fixed group effect model is better than the pooled OLS model.

### 3.2.3 Fixed Time Effect and Two-way Fixed Effect Models

For the fixed time effects model, you need to switch  $n$  and  $T$ , and  $i$  and  $t$  in the formulas.

- Model:  $y_{it} = \alpha + \tau_t + \beta' X_{it} + \varepsilon_{it}$
- Within effect model:  $(y_{it} - \bar{y}_{\bullet t}) = \beta'(x_{it} - \bar{x}_{\bullet t}) + (\varepsilon_{it} - \bar{\varepsilon}_{\bullet t})$
- Dummy coefficients:  $d_t^* = \bar{y}_{\bullet t} - \beta' \bar{x}_{\bullet t}$
- Correct standard errors:  $se_k^* = se_k \sqrt{\frac{df_{error}^{Within}}{df_{error}^{LSDV}}} = se_k \sqrt{\frac{Tn-k}{Tn-T-k}}$
- Between effect model:  $\bar{y}_{\bullet t} = \alpha + \bar{x}_{\bullet t} + \varepsilon_t$
- $H_0 : \tau_1 = \dots = \tau_{T-1} = 0$ .
- F-test:  $\frac{(e'e_{Efficient} - e'e_{Robust})/(T-1)}{(e'e_{Robust})/(Tn-T-k)} \sim F(T-1, Tn-T-k)$ .

The fixed group and time effect model uses slightly different formulas. The within effect model of this two-way fixed model is estimated by four approaches (see 6.1 for details).

- Model:  $y_{it} = \alpha + \mu_i + \tau_t + \beta' X_{it} + \varepsilon_{it}$ .
- Within effect Model:  $y_{it}^* = y_{it} - \bar{y}_{i\bullet} - \bar{y}_{\bullet t} + \bar{y}_{\bullet\bullet}$  and  $x_{it}^* = x_{it} - \bar{x}_{i\bullet} - \bar{x}_{\bullet t} + \bar{x}_{\bullet\bullet}$ .
- Dummy coefficients:  $d_g^* = (\bar{y}_{g\bullet} - \bar{y}_{\bullet\bullet}) - b'(\bar{x}_{g\bullet} - \bar{x}_{\bullet\bullet})$  and  $d_t^* = (\bar{y}_{\bullet t} - \bar{y}_{\bullet\bullet}) - b'(\bar{x}_{\bullet t} - \bar{x}_{\bullet\bullet})$
- Correct standard errors:  $se_k^* = se_k \sqrt{\frac{df_{error}^{Within}}{df_{error}^{LSDV}}} = se_k \sqrt{\frac{nT-k}{nT-n-T-k+1}}$
- $H_0 : \mu_1 = \dots = \mu_{n-1} = 0$  and  $\tau_1 = \dots = \tau_{T-1} = 0$ .
- F-test:  $\frac{(e'e_{Efficient} - e'e_{Robust})/(n+T-2)}{(e'e_{Robust})/(nT-n-T-k+1)} \sim F[(n+T-2), (nT-n-T-k+1)]$

<sup>6</sup> When comparing fixed effect and random effect models, the fixed effect estimates are considered as the robust estimates and random effect estimates as the efficient estimates.

### 3.3 Random Effect Models

The one-way random group effect model is formulated as  $y_{it} = \alpha + \beta' X_{it} + \mu_i + v_{it}$ ,  $w_{it} = \mu_i + v_{it}$  where  $\mu_i \sim IID(0, \sigma_\mu^2)$  and  $v_{it} \sim IID(0, \sigma_v^2)$ . The  $\mu_i$  are assumed independent of  $v_{it}$  and  $X_{it}$ , which are also independent of each other for all  $i$  and  $t$ . This assumption is not necessary in the fixed effect model. The components of  $Cov(w_{it}, w_{js}) = E(w_{it} w_{js})$  are  $\sigma_\mu^2 + \sigma_v^2$  if  $i=j$  and  $t=s$  and  $\sigma_\mu^2$  if  $i=j$  and  $t \neq s$ .<sup>7</sup>

A random effect model is estimated by generalized least squares (GLS) when the variance structure is known, and by feasible generalized least squares (FGLS) when the variance is unknown. Compared to fixed effect models, random effect models are relatively difficult to estimate. This document assumes panel data are balanced.

#### 3.3.1 Generalized Least Squares (GLS)

When  $\Omega$  is known (given), GLS based on the true variance components is BLUE and all the feasible GLS estimators considered are asymptotically efficient as either  $n$  or  $T$  approaches infinity (Baltagi 2001). The  $\Omega$  matrix looks like,

$$\Omega_{T \times T} = \begin{bmatrix} \sigma_\mu^2 + \sigma_v^2 & \sigma_\mu^2 & \dots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_v^2 & \dots & \sigma_\mu^2 \\ \dots & \dots & \dots & \dots \\ \sigma_\mu^2 & \sigma_\mu^2 & \dots & \sigma_\mu^2 + \sigma_v^2 \end{bmatrix}$$

In GLS, you just need to compute  $\theta$  using the  $\Omega$  matrix:  $\theta = 1 - \sqrt{\frac{\sigma_v^2}{T\sigma_\mu^2 + \sigma_v^2}}$ .<sup>8</sup> Then transform

variables as follows.

- $y_{it}^* = y_{it} - \theta \bar{y}_i$ .
- $x_{it}^* = x_{it} - \theta \bar{x}_i$  for all  $X_k$
- $\alpha^* = 1 - \theta$

Finally, run OLS on the transformed variables:  $y_{it}^* = \alpha^* + x_{it}^* \beta^* - \varepsilon_{it}^*$ . Since  $\Omega$  is often unknown, FGLS is more frequently used than GLS.

#### 3.3.2 Feasible Generalized Least Squares (FGLS)

<sup>7</sup> This implies that  $Corr(w_{it}, w_{js})$  is 1 if  $i=j$  and  $t=s$ , and  $\sigma_\mu^2 / (\sigma_\mu^2 + \sigma_v^2)$  if  $i=j$  and  $t \neq s$ .

<sup>8</sup> If  $\theta = 0$ , run pooled OLS. If  $\theta = 1$  and  $\sigma_v^2 = 0$ , then run the within effect model.

If  $\Omega$  is unknown, first you have to estimate  $\theta$  using  $\hat{\sigma}_\mu^2$  and  $\hat{\sigma}_v^2$ :

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

The  $\hat{\sigma}_v^2$  is derived from the SSE (sum of squares due to error) of the within effect model or from the deviations of residuals from group means of residuals:

$$\hat{\sigma}_v^2 = \frac{SSE_{within}}{nT - n - k} = \frac{e'e_{within}}{nT - n - k} = \frac{\sum_{i=1}^n \sum_{t=1}^T (v_{it} - \bar{v}_{i\cdot})^2}{nT - n - k}, \text{ where } v_{it} \text{ are the residuals of the LSDV1.}$$

The  $\hat{\sigma}_\mu^2$  comes from the between effect model (group mean regression):

$$\hat{\sigma}_\mu^2 = \hat{\sigma}_{between}^2 - \frac{\hat{\sigma}_v^2}{T}, \text{ where } \hat{\sigma}_{between}^2 = \frac{SSE_{between}}{n - K}.$$

Next, transform variables using  $\hat{\theta}$  and then run OLS:  $y_{it}^* = \alpha^* + x_{it}^* \beta^* - \varepsilon_{it}^*$ .

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

The estimation of the two-way random effect model is skipped here.

### 3.3.3 Testing Random Effects (LM test)

The null hypothesis is that cross-sectional variance components are zero,  $H_0: \sigma_u^2 = 0$ . Breusch and Pagan (1980) developed the Lagrange multiplier (LM) test (Greene 2003; Judge et al. 1988). In the following formula,  $\bar{e}$  is the  $n \times 1$  vector of the group specific means of pooled regression residuals, and  $e'e$  is the SSE of the pooled OLS regression. The LM follows chi-squared distribution with one degree of freedom.

$$LM_\mu = \frac{nT}{2(T-1)} \left[ \frac{e'DDe}{e'e} - 1 \right]^2 = \frac{nT}{2(T-1)} \left[ \frac{T^2 \bar{e}'\bar{e}}{e'e} - 1 \right]^2 \sim \chi^2(1).$$

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

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

The two way random effect model has the null hypothesis of  $H_0: \sigma_u^2 = 0$  and  $\sigma_v^2 = 0$ . The LM test combines two one-way random effect models for group and time,

$$LM_{\mu\nu} = LM_\mu + LM_\nu \sim \chi^2(2).$$

### 3.4 Hausman Test: Fixed Effects versus Random Effects

The Hausman specification test compares the fixed versus random effects under the null hypothesis that the individual effects are uncorrelated with the other regressors in the model (Hausman 1978). If correlated ( $H_0$  is rejected), a random effect model produces biased estimators, violating one of the Gauss-Markov assumptions; so a fixed effect model is preferred. Hausman's essential result is that the covariance of an efficient estimator with its difference from an inefficient estimator is zero (Greene 2003).

$$m = (b_{Robust} - b_{Efficient})' \hat{\Sigma}^{-1} (b_{Robust} - b_{Efficient}) \sim \chi^2(k),$$

$\hat{\Sigma} = Var[b_{Robust} - b_{Efficient}] = Var(b_{Robust}) - Var(b_{Efficient})$  is the difference between the estimated covariance matrix of the parameter estimates in the LSDV model (robust) and that of the random effects model (efficient). It is notable that an intercept and dummy variables SHOULD be excluded in computation.

### 3.5 Poolability Test

What is poolability? It asks if slopes are the same across groups or over time. Thus, the null hypothesis of the poolability test is  $H_0 : \beta_{ik} = \beta_k$ . Remember that slopes remain constant in fixed and random effect models; only intercepts and error variances matter.

The poolability test is undertaken under the assumption of  $\mu \sim N(0, s^2 I_{NT})$ . This test uses the F

statistic,  $F_{obs} = \frac{(e'e - \sum e_i' e_i) / (n-1)K}{\sum e_i' e_i / n(T-K)} \sim F[(n-1)K, n(T-K)]$ , where  $e'e$  is the SSE of the

pooled OLS and  $e_i' e_i$  is the SSE of the OLS regression for group  $i$ . If the null hypothesis is rejected, the panel data are not poolable. Under this circumstance, you may go to the random coefficient model or hierarchical regression model.

Similarly, the null hypothesis of the poolability test over time is  $H_0 : \beta_{tk} = \beta_k$ . The F-test is

$F_{obs} = \frac{(e'e - \sum e_t' e_t) / (T-1)K}{\sum e_t' e_t / T(n-K)} = F[(T-1)K, T(n-K)]$ , where  $e_t' e_t$  is SSE of the OLS

regression at time  $t$ .

## 4. The Fixed Group Effect Model

The one-way fixed group model examines group differences in the intercepts. The LSDV for this fixed model needs to create as many dummy variables as the number of groups or subjects. When many dummies are needed, the within effect model is useful since it transforms variables using group means to avoid dummies. The between effect model uses group means of variables.

### 4.1 The Pooled OLS Regression Model

Let us first consider the pooled model without dummy variables.

```
. regress cost output fuel load // pooled model
```

Source	SS	df	MS			
Model	112.705452	3	37.5684839	Number of obs =	90	
Residual	1.33544153	86	.01552839	F( 3, 86) =	2419.34	
				Prob > F	= 0.0000	
				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

cost = 9.517 + .883\*output +.454\*fuel -1.628\*load.

This model fits the data well ( $p < .0000$  and  $R^2 = .9883$ ). We may, however, suspect fixed group effects that produce different intercepts across groups. As discussed in Chapter 2, there are three equivalent approaches of LSDV. They report the identical parameter estimates of regressors excluding dummies. Let us begin with LSDV1.

### 4.2 LSDV1 without a Dummy

LSDV1 drops a dummy variable to identify the model. LSDV1 produces correct ANOVA information, goodness of fit, parameter estimates, and standard errors. As a consequence, this approach is commonly used in practice. LSDV produces six regression equations for six groups (airlines).

```
Group1: cost = 9.706 + .919*output +.417*fuel -1.070*load
Group2: cost = 9.665 + .919*output +.417*fuel -1.070*load
Group3: cost = 9.497 + .919*output +.417*fuel -1.070*load
Group4: cost = 9.891 + .919*output +.417*fuel -1.070*load
Group5: cost = 9.730 + .919*output +.417*fuel -1.070*load
Group6: cost = 9.793 + .919*output +.417*fuel -1.070*load
```

In SAS, the REG procedure fits the OLS regression model. Let us drop the last dummy g6, the reference point.

```
PROC REG DATA=masil.airline;
  MODEL cost = g1-g5 output fuel load;
```

**RUN;**

The REG Procedure  
Model: MODEL1  
Dependent Variable: cost

Number of Observations Read 90  
Number of Observations Used 90

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	113.74827	14.21853	3935.79	<.0001
Error	81	0.29262	0.00361		
Corrected Total	89	114.04089			

Root MSE 0.06011 R-Square 0.9974  
Dependent Mean 13.36561 Adj R-Sq 0.9972  
Coeff Var 0.44970

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	9.79300	0.26366	37.14	<.0001
g1	1	-0.08706	0.08420	-1.03	0.3042
g2	1	-0.12830	0.07573	-1.69	0.0941
g3	1	-0.29598	0.05002	-5.92	<.0001
g4	1	0.09749	0.03301	2.95	0.0041
g5	1	-0.06301	0.02389	-2.64	0.0100
output	1	0.91928	0.02989	30.76	<.0001
fuel	1	0.41749	0.01520	27.47	<.0001
load	1	-1.07040	0.20169	-5.31	<.0001

Note that the parameter estimate of g6 is presented in the intercept (9.793). Other dummy parameter estimates are computed with the reference point. The actual intercept of the group 1, for example, is computed as  $9.706 = 9.793 + (-.087)*1 + (-.1283)*0 + (-.2960)*0 + (.0975)*0 + (-.0630)*0$ , where 9.793 is the reference point.

Stata has the `.regress` command for OLS regression (LSDV).

`. regress cost g1-g5 output fuel load`

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

```

Now, run the `LIMDEP Regress$` command to fit the `LSDV1`. Do not forget to include `ONE` for the intercept in the `Rhs`;

```
--> REGRESS;Lhs=COST;Rhs=ONE,G1,G2,G3,G4,G5,OUTPUT,FUEL,LOAD$
```

```

+-----+
| Ordinary least squares regression   Weighting variable = none   |
| Dep. var. = COST   Mean= 13.36560933   , S.D.= 1.131971444   |
| Model size: Observations = 90, Parameters = 9, Deg.Fr.= 81   |
| Residuals: Sum of squares= .2926207777   , Std.Dev.= .06010   |
| Fit: R-squared= .997434, Adjusted R-squared = .99718   |
| Model test: F[ 8, 81] = 3935.82, Prob value = .00000   |
| Diagnostic: Log-L = 130.0865, Restricted(b=0) Log-L = -138.3581 |
| LogAmemiyaPrCrt.= -5.528, Akaike Info. Crt.= -2.691   |
| Autocorrel: Durbin-Watson Statistic = 1.02645, Rho = .48677   |
+-----+
+-----+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+-----+-----+-----+-----+-----+-----+
Constant  9.793021272   .26366104   37.142   .0000
G1        -.8707201949E-01   .84199161E-01   -1.034   .3042   .16666667
G2        -.1283060033   .75727781E-01   -1.694   .0940   .16666667
G3        -.2959885994   .50022855E-01   -5.917   .0000   .16666667
G4        .9749253376E-01   .33009146E-01   2.954   .0041   .16666667
G5        -.6300770422E-01   .23891796E-01   -2.637   .0100   .16666667
OUTPUT    .9192881432   .29889967E-01   30.756   .0000   -1.1743092
FUEL      .4174910457   .15199071E-01   27.468   .0000   12.770359
LOAD      -1.070395015   .20168924     -5.307   .0000   .56046016
(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

```

What if you drop a different dummy variable, say `g1`, instead of `g6`? Since the different reference point is applied, you will get different dummy coefficients. The other statistics such as goodness-of-fits, however, remain unchanged.

```
. regress cost g2-g6 output fuel load // LSDV1 dropping g1
```

```

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

When you have not created dummy variables, take advantage of the `.xi` prefix command.<sup>9</sup> Note that Stata by default drops the first dummy variable while the SAS TSCSREG and PANEL procedures in 4.5.2 drops the last dummy.

```
. 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
Model	113.74827	8	14.2185338	F( 8, 81) =	3935.79
Residual	.292622872	81	.003612628	Prob > F	= 0.0000
				R-squared	= 0.9974
				Adj R-squared	= 0.9972
				Root MSE	= .06011
Total	114.040893	89	1.28135835		

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

### 4.3 LSDV2 without the Intercept

LSDV2 reports actual parameter estimates of the dummies. Because LSDV2 suppresses the intercept, you will get incorrect F and  $R^2$  statistics.

In the SAS REG procedure, you need to use the `/NOINT` option to suppress the intercept. Note that the F value of 497,985 and  $R^2$  of 1 are not likely.

```
PROC REG DATA=masil.airline;
  MODEL cost = g1-g6 output fuel load /NOINT;
RUN;
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: cost
```

```
Number of Observations Read      90
Number of Observations Used      90
```

<sup>9</sup> The Stata `.xi` is used either as an ordinary command or a prefix command like `.bysort`. This command creates dummies from a categorical variable specified in the term `i.` and then run the command following the colon.

NOTE: No intercept in model. R-Square is redefined.

#### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	16191	1799.03381	497985	<.0001
Error	81	0.29262	0.00361		
Uncorrected Total	90	16192			

Root MSE	0.06011	R-Square	1.0000
Dependent Mean	13.36561	Adj R-Sq	1.0000
Coeff Var	0.44970		

#### Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
g1	1	9.70594	0.19312	50.26	<.0001
g2	1	9.66471	0.19898	48.57	<.0001
g3	1	9.49702	0.22496	42.22	<.0001
g4	1	9.89050	0.24176	40.91	<.0001
g5	1	9.73000	0.26094	37.29	<.0001
g6	1	9.79300	0.26366	37.14	<.0001
output	1	0.91928	0.02989	30.76	<.0001
fuel	1	0.41749	0.01520	27.47	<.0001
load	1	-1.07040	0.20169	-5.31	<.0001

Stata uses the `noconstant` option to suppress the intercept. Note that `noc` is its abbreviation.

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

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

In LIMDEP, you need to drop ONE out of the `Rhs`; to suppress the intercept. Unlike SAS and Stata, LIMDEP reports correct R2 and F even in LSDV2.

```
--> REGRESS;Lhs=COST;Rhs=G1,G2,G3,G4,G5,G6,OUTPUT,FUEL,LOAD$
```

```
+-----+
| Ordinary least squares regression Weighting variable = none |
| Dep. var. = COST Mean= 13.36560933 , S.D.= 1.131971444 |
| Model size: Observations = 90, Parameters = 9, Deg.Fr.= 81 |
| Residuals: Sum of squares= .2926207777 , Std.Dev.= .06010 |
| Fit: R-squared= .997434, Adjusted R-squared = .99718 |
| Model test: F[ 8, 81] = 3935.82, Prob value = .00000 |
| Diagnostic: Log-L = 130.0865, Restricted(b=0) Log-L = -138.3581 |
| LogAmemiyaPrCrt.= -5.528, Akaike Info. Crt.= -2.691 |
| Model does not contain ONE. R-squared and F can be negative! |
| Autocorrel: Durbin-Watson Statistic = 1.02645, Rho = .48677 |
+-----+
+-----+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+-----+-----+-----+-----+-----+-----+
G1          9.705949253      .19312325  50.258  .0000  .16666667
G2          9.664715269      .19898117  48.571  .0000  .16666667
G3          9.497032673      .22495746  42.217  .0000  .16666667
G4          9.890513806      .24176245  40.910  .0000  .16666667
G5          9.730013568      .26094094  37.288  .0000  .16666667
G6          9.793021272      .26366104  37.142  .0000  .16666667
OUTPUT      .9192881432      .29889967E-01  30.756  .0000  -1.1743092
FUEL        .4174910457      .15199071E-01  27.468  .0000  12.770359
LOAD       -1.070395015      .20168924   -5.307  .0000  .56046016
(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)
```

#### 4.4 LSDV3 with Restrictions

LSDV3 imposes a restriction that the sum of the dummy parameters is zero. The SAS REG procedure uses the RESTRICT statement to impose restrictions.

```
PROC REG DATA=masil.airline;
  MODEL cost = g1-g6 output fuel load;
  RESTRICT g1 + g2 + g3 + g4 + g5 + g6 = 0;
RUN;
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: cost
```

NOTE: Restrictions have been applied to parameter estimates.

```
Number of Observations Read      90
Number of Observations Used      90
```

#### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
--------	----	----------------	-------------	---------	--------

Model	8	113.74827	14.21853	3935.79	<.0001
Error	81	0.29262	0.00361		
Corrected Total	89	114.04089			

Root MSE	0.06011	R-Square	0.9974
Dependent Mean	13.36561	Adj R-Sq	0.9972
Coeff Var	0.44970		

## Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	9.71353	0.22964	42.30	<.0001
g1	1	-0.00759	0.04562	-0.17	0.8683
g2	1	-0.04882	0.03798	-1.29	0.2023
g3	1	-0.21651	0.01606	-13.48	<.0001
g4	1	0.17697	0.01942	9.11	<.0001
g5	1	0.01647	0.03669	0.45	0.6547
g6	1	0.07948	0.04050	1.96	0.0532
output	1	0.91928	0.02989	30.76	<.0001
fuel	1	0.41749	0.01520	27.47	<.0001
load	1	-1.07040	0.20169	-5.31	<.0001
RESTRICT	-1	3.01674E-15	1.51088E-10	0.00	1.0000*

\* Probability computed using beta distribution.

The dummy coefficients mean deviations from the averaged group effect (9.714). The actual intercept of group 2, for example, is  $9.665 = 9.714 + (-.049)$ . Note that the 3.01674E-15 of RESTRICT below is virtually zero.

In Stata, you have to use the `.cnsreg` command rather than `.regress`. The command, however, does not provide an ANOVA table and goodness-of-fit statistics.

```
. 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     = .06011
```

```
( 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

LIMDEP has the `CLS$` subcommand to impose restrictions. Again, do not forget to include `ONE` in the `Rhs`:

```
--> REGRESS;Lhs=COST;Rhs=ONE,G1,G2,G3,G4,G5,G6,OUTPUT,FUEL,LOAD;
      Cls:b(1)+b(2)+b(3)+b(4)+b(5)+b(6)=0$
```

```
+-----+
| Linearly restricted regression                                     |
| Ordinary least squares regression   Weighting variable = none |
| Dep. var. = COST      Mean= 13.36560933 , S.D.= 1.131971444   |
| Model size: Observations = 90, Parameters = 9, Deg.Fr.= 81    |
| Residuals: Sum of squares= .2926207777 , Std.Dev.= .06010    |
| Fit:      R-squared= .997434, Adjusted R-squared = .99718     |
|           (Note: Not using OLS. R-squared is not bounded in [0,1] |
| Model test: F[ 8, 81] = 3935.82, Prob value = .00000         |
| Diagnostic: Log-L = 130.0865, Restricted(b=0) Log-L = -138.3581 |
|           LogAmemiyaPrCrt.= -5.528, Akaike Info. Crt.= -2.691 |
| Note, when restrictions are imposed, R-squared can be less than zero. |
| F[ 1, 80] for the restrictions = .0000, Prob = 1.0000         |
| Autocorrel: Durbin-Watson Statistic = 1.02645, Rho = .48677 |
+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+-----+-----+-----+-----+-----+-----+-----+-----+
Constant  12.12205614  .27886962  43.469  .0000
G1        -2.416106889  .89836871E-01 -26.894  .0000  .16666667
G2        -2.457340873  .82929154E-01 -29.632  .0000  .16666667
G3        -2.625023469  .56175656E-01 -46.729  .0000  .16666667
G4        -2.231542336  .41557714E-01 -53.697  .0000  .16666667
G5        -2.392042574  .29995908E-01 -79.746  .0000  .16666667
G6        -2.329034870  .33569388E-01 -69.380  .0000  .16666667
OUTPUT    .9192881432  .29889967E-01  30.756  .0000  -1.1743092
FUEL      .4174910457  .15199071E-01  27.468  .0000  12.770359
LOAD      -1.070395015  .20168924  -5.307  .0000  .56046016
```

LSDV3 in LIMDEP reports different dummy coefficients. But you may draw actual intercepts of groups in a manner similar to what you would do in SAS and Stata. The actual intercept of group 3, for example, is  $9.497 = 12.122 + (-2.625)$ .

## 4.5 Within Group Effect Model

The within effect model does not use the dummies and thus has larger degrees of freedom, smaller MSE, and smaller standard errors of parameters than those of LSDV. As a consequence, you need to adjust standard errors. This model does not report individual dummy coefficients either. The SAS `TSCSREG` procedure and LIMDEP `Regress$` command report the adjusted (correct) MSE, SEE (Root MSE),  $R^2$ , and standard errors.

### 4.5.1 Estimating the Within Effect Model

First, let us manually estimate the within group effect model in Stata. You need to compute group means and transform dependent and independent variables using group means (log is skipped here).

```
. egen gm_cost=mean(cost), by(airline) // compute group means
. egen gm_output=mean(output), by(airline)
. egen gm_fuel=mean(fuel), by(airline)
. egen gm_load=mean(load), by(airline)
```

You will get the following group means of variables.

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

```
. gen gw_cost = cost - gm_cost // compute deviations from the group means
. gen gw_output = output - gm_output
. gen gw_fuel = fuel - gm_fuel
. gen gw_load = load - gm_load
```

Now, we are ready to run the within effect model. Keep in mind that you have to suppress the intercept. Carefully check MSE, SEE,  $R^2$ , and standard errors.

```
. regress gw_cost gw_output gw_fuel gw_load, noc // within effect
```

Source	SS	df	MS	Number of obs =	90
Model	39.0683861	3	13.0227954	F( 3, 87) =	3871.82
Residual	.292622861	87	.003363481	Prob > F	= 0.0000
				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

You may compute group intercepts using  $d_g^* = \bar{y}_g - \beta' \bar{x}_g$ . For example, the intercept of airline 5 is computed as  $9.730 = 12.363 - \{.919*(-2.286) + .417*12.792 + (-1.073)*.566\}$ . 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 the LSDV. For example, the standard error of the logged output is computed as  $.0299 = .0288 * \sqrt{87/81}$ .

#### 4.5.2 Using the SAS TSCSREG and PANEL Procedures

The TSCSREG and PANEL procedures of SAS/ETS allows users to fit the within effect model conveniently. The procedures, in fact, report LSDV1, but you do not need to create dummy

variables and compute deviations from the group means. This procedure reports correct MSE, SEE,  $R^2$ , and standard errors, and conducts the F test for the fixed group effect as well.

```
PROC SORT DATA=masil.airline;
  BY airline year;

PROC TSCSREG DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /FIXONE;
RUN;
```

## The TSCSREG Procedure

Dependent Variable: cost

## Model Description

Estimation Method	FixOne
Number of Cross Sections	6
Time Series Length	15

## Fit Statistics

SSE	0.2926	DFE	81
MSE	0.0036	Root MSE	0.0601
R-Square	0.9974		

## F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
5	81	57.73	<.0001

## Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr >  t	Label
CS1	1	-0.08706	0.0842	-1.03	0.3042	Cross Sectional Effect 1
CS2	1	-0.1283	0.0757	-1.69	0.0941	Cross Sectional Effect 2
CS3	1	-0.29598	0.0500	-5.92	<.0001	Cross Sectional Effect 3
CS4	1	0.097494	0.0330	2.95	0.0041	Cross Sectional Effect 4
CS5	1	-0.06301	0.0239	-2.64	0.0100	Cross Sectional Effect 5
Intercept	1	9.793004	0.2637	37.14	<.0001	Intercept
output	1	0.919285	0.0299	30.76	<.0001	
fuel	1	0.417492	0.0152	27.47	<.0001	
load	1	-1.0704	0.2017	-5.31	<.0001	

Note that a data set needs to be sorted in advance by variables to appear in the ID statement of the TSCSREG and PANEL procedures. The following PANEL procedure returns the same output.

```
PROC PANEL DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /FIXONE;
RUN;
```

### 4.5.3 Using Stata

The Stata `.xtreg` command fits the within group effect model without creating dummy variables. The command reports correct standard errors and the F test for fixed group effects. This command, however, does not provide an analysis of variance (ANOVA) table and correct  $R^2$  and F statistics. The `.xtreg` command should follow the `.tsset` command that specifies grouping and time variables.

```
. tsset airline year
   panel variable:  airline, 1 to 6
   time variable:  year, 1 to 15
```

The `fe` of `.xtreg` indicates the within effect model and `i(airline)` specifies airline as the independent unit. Note that this command reports adjusted (correct) standard errors.

```
. xtreg cost output fuel load, fe i(airline) // within group effect

Fixed-effects (within) regression              Number of obs   =          90
Group variable (i): 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

corr(u_i, Xb) = -0.3475                       F(3,81)         =    3604.80
                                                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
```

The last line of the output tests the null hypothesis that all dummy parameters in LSDV1 are zero (e.g.,  $g_1=0$ ,  $g_2=0$ ,  $g_3=0$ ,  $g_4=0$ , and  $g_5=0$ ). Note the intercept of 9.714 is that of LSDV3.

### 4.5.4 Using LIMDEP

In LIMDEP, you have to specify the panel data model and stratification or time variables. The `Panel$` and `Fixed$` subcommands mean a fixed effect panel data model. The `Str$` subcommand specifies a stratification variable.

```
--> REGRESS;Lhs=COST;Rhs=ONE,OUTPUT,FUEL,LOAD;Panel;Str=AIRLINE;Fixed$
```

```
+-----+
| OLS Without Group Dummy Variables |
| Ordinary least squares regression | Weighting variable = none |
| Dep. var. = COST Mean= 13.36560933 | , S.D.= 1.131971444 |
| Model size: Observations = 90, | Parameters = 4, Deg.Fr.= 86 |
| Residuals: Sum of squares= 1.335449522 | , Std.Dev.= .12461 |
| Fit: R-squared= .988290, Adjusted R-squared = | .98788 |
| Model test: F[ 3, 86] = 2419.33, Prob value = | .00000 |
| Diagnostic: Log-L = 61.7699, Restricted(b=0) Log-L = | -138.3581 |
| LogAmemiyaPrCrt.= -4.122, Akaike Info. Crt.= | -1.284 |
| Panel Data Analysis of COST [ONE way] |
| Unconditional ANOVA (No regressors) |
| Source Variation Deg. Free. Mean Square |
| Between 74.6799 5. 14.9360 |
| Residual 39.3611 84. .468584 |
| Total 114.041 89. 1.28136 |
+-----+
```

```
+-----+
|Variable | Coefficient | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+-----+
OUTPUT .8827386341 .13254552E-01 66.599 .0000 -1.1743092
FUEL .4539777119 .20304240E-01 22.359 .0000 12.770359
LOAD -1.627507797 .34530293 -4.713 .0000 .56046016
Constant 9.516912231 .22924522 41.514 .0000
(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)
```

```
+-----+
| Least Squares with Group Dummy Variables |
| Ordinary least squares regression | Weighting variable = none |
| Dep. var. = COST Mean= 13.36560933 | , S.D.= 1.131971444 |
| Model size: Observations = 90, | Parameters = 9, Deg.Fr.= 81 |
| Residuals: Sum of squares= .2926207777 | , Std.Dev.= .06010 |
| Fit: R-squared= .997434, Adjusted R-squared = | .99718 |
| Model test: F[ 8, 81] = 3935.82, Prob value = | .00000 |
| Diagnostic: Log-L = 130.0865, Restricted(b=0) Log-L = | -138.3581 |
| LogAmemiyaPrCrt.= -5.528, Akaike Info. Crt.= | -2.691 |
| Estd. Autocorrelation of e(i,t) .573531 |
+-----+
```

```
+-----+
|Variable | Coefficient | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+-----+
OUTPUT .9192881432 .29889967E-01 30.756 .0000 -1.1743092
FUEL .4174910457 .15199071E-01 27.468 .0000 12.770359
LOAD -1.070395015 .20168924 -5.307 .0000 .56046016
(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)
```

LIMDEP reports both the pooled OLS regression and the within effect model. Like the SAS TSCSREG procedure, LIMDEP provides correct MSE, SEE,  $R^2$ , and standard errors.

#### 4.6 Between Group Effect Model: Group Mean Regression

The between effect model uses aggregate information, group means of variables. In other words, the unit of analysis is not an individual observation, but groups or subjects. The number of observations jumps down to  $n$  from  $nT$ . This group mean regression produces different goodness-of-fits and parameter estimates from those of LSDV and the within effect model.

Let us compute group means and run the OLS regression with them. The `.collapse` command computes aggregate information and saves into a new data set. 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)
```

```
. regress gm_cost gm_output gm_fuel gm_load
```

Source	SS	df	MS	Number of obs =	6
Model	4.94698124	3	1.64899375	F( 3, 2) =	104.12
Residual	.031675926	2	.015837963	Prob > F	= 0.0095
				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

The SAS PANEL procedure has the `/BTWNG` and `/BTWNT` option to estimate the between effect model. The TSCSREG procedure does not have this option.

```
PROC PANEL DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /BTWNG;
RUN;
```

#### The PANEL Procedure Between Groups Estimates

Dependent Variable: cost

#### Model Description

Estimation Method	BtwGrps
Number of Cross Sections	6
Time Series Length	15

#### Fit Statistics

SSE	0.0317	DFE	2
MSE	0.0158	Root MSE	0.1258

R-Square            0.9936

## Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr >  t	Label
Intercept	1	85.80901	56.4830	1.52	0.2681	Intercept
output	1	0.782455	0.1088	7.19	0.0188	
fuel	1	-5.52398	4.4788	-1.23	0.3427	
load	1	-1.75102	2.7432	-0.64	0.5886	

The Stata `.xtreg` command has the `be` option to fit the between effect model. This command, however, does not report the ANOVA table.

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

```
Between regression (regression on group means) Number of obs = 90
Group variable (i): 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

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

	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

LIMDEP has the `Means;` subcommand to fit the between effect model.

```
--> REGRESS;Lhs=COST;Rhs=ONE,OUTPUT,FUEL,LOAD;Panel;Str=AIRLINE;Means$
```

```
+-----+
| Group Means Regression |
| Ordinary least squares regression Weighting variable = none |
| Dep. var. = YBAR(i.) Mean= 13.36560933 , S.D.= .9978636346 |
| Model size: Observations = 6, Parameters = 4, Deg.Fr.= 2 |
| Residuals: Sum of squares= .3167277206E-01, Std.Dev.= .12584 |
| Fit: R-squared= .993638, Adjusted R-squared = .98410 |
| Model test: F[ 3, 2] = 104.13, Prob value = .00953 |
| Diagnostic: Log-L = 7.2185, Restricted(b=0) Log-L = -7.9538 |
| LogAmemiyaPrCrt.= -3.635, Akaike Info. Crt.= -1.073 |
+-----+
+-----+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+-----+-----+-----+-----+-----+-----+
OUTPUT .7824472689 .10876126 7.194 .0000 .23025612E-11
FUEL -5.524437466 4.4786519 -1.234 .2174 .18642891
LOAD -1.750947653 2.7430470 -.638 .5233 .32541105
Constant 85.81483169 56.481148 1.519 .1287
```

#### 4.7 Testing Fixed Group Effects (F-test)

How do we know whether there are fixed group effects? The null hypothesis is that all dummy parameters except one are zero:  $H_0 : \mu_1 = \dots = \mu_{n-1} = 0$ .

In order to conduct a F-test, let us take the SSE ( $e'e$ ) of 1.3354 from the pooled OLS regression and .2926 from the LSDVs (LSDV1 through LSDV3) or the within effect model. Alternatively, you may draw  $R^2$  of .9974 from LSDV1 or LSDV3 and .9883 from the pooled OLS. Do not, however, use LSDV2 and the within effect model for  $R^2$ .

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

The large F statistic rejects the null hypothesis in favor of the fixed group effect model ( $p < .0000$ ).

The SAS TSCSREG and PANEL procedures and Stata `.xtreg` command by default conduct the F test. Alternatively, you may conduct the same test with LSDV1. In SAS, add the TEST statement in the REG procedure and run the procedure again (other outputs are skipped).

```
PROC REG DATA=masil.airline;
  MODEL cost = g1-g5 output fuel load;
  TEST g1 = g2 = g3 = g4 = g5 = 0;
RUN;
```

The REG Procedure  
Model: MODEL1

Test 1 Results for Dependent Variable cost

Source	DF	Mean Square	F Value	Pr > F
Numerator	5	0.20856	57.73	<.0001
Denominator	81	0.00361		

In Stata, run the `.test` command, a follow-up command for the Wald test, right after estimating the model.

```
. quietly regress cost g1-g5 output fuel load // LSDV1
. 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
```

## 4.8 Summary

Table 6 summarizes the estimation of panel data models in SAS, Stata, and LIMDEP. The SAS REG and TSCSREG procedures are generally preferred to Stata and LIMDEP commands.

Table 6 Comparison of the Fixed Effect Model in SAS, Stata, LIMDEP\*

	SAS 9	Stata 9	LIMDEP 8
OLS estimation	PROC REG;	. regress (cnsreg)	Regress\$
LSDV1	Correct	Correct	Correct (slightly different F)
LSDV2	Incorrect F, (adjusted) R <sup>2</sup>	Incorrect F, (adjusted) R <sup>2</sup>	Correct (slightly different F) Correct R <sup>2</sup>
LSDV3	Correct	. cnsreg command No R <sup>2</sup> , ANOVA table but F	Correct (slightly different F) Different dummy coefficients
Panel Estimation	PROC TSCSREG; PROC PANEL;	. xtreg	Regress; Panel\$
Estimation type	LSDV1	Within and between effect	Within effect
SSE (e'e)	Correct	No	Correct
MSE or SEE	Correct (adjusted)	No	Correct (adjusted) SEE
Model test (F)	No	Incorrect	Slightly different F
(adjusted) R <sup>2</sup>	Correct	Incorrect	Correct
Intercept	Correct	LSDV3 intercept	No
Coefficients	Correct	Correct	Correct
Standard errors	Correct (adjusted)	Correct (adjusted)	Correct (adjusted)
Effect test (F)	Yes	Yes	No
Between effect	Yes (PROC PANEL;)	Yes (the be option)	Yes (Means option)

\* "Yes/No" means whether the software reports the statistics. "Correct/incorrect" indicates whether the statistics are different from those of the least squares dummy variable (LSDV) 1 without a dummy variable.

## 5. The Fixed Time Effect Model

The fixed time effect model investigates how time affects the intercept using time dummy variables. The logic and method are the same as those of the fixed group effect model.

### 5.1 Least Squares Dummy Variable Models

The least squares dummy variable (LSDV) model produces fifteen regression equations. This section does not present all outputs, but one or two for each LSDV approach.

```
Time01: cost = 20.496 + .868*output - .484*fuel -1.954*load
Time02: cost = 20.578 + .868*output - .484*fuel -1.954*load
Time03: cost = 20.656 + .868*output - .484*fuel -1.954*load
Time04: cost = 20.741 + .868*output - .484*fuel -1.954*load
Time05: cost = 21.200 + .868*output - .484*fuel -1.954*load
Time06: cost = 21.412 + .868*output - .484*fuel -1.954*load
Time07: cost = 21.503 + .868*output - .484*fuel -1.954*load
Time08: cost = 21.654 + .868*output - .484*fuel -1.954*load
Time09: cost = 21.830 + .868*output - .484*fuel -1.954*load
Time10: cost = 22.114 + .868*output - .484*fuel -1.954*load
Time11: cost = 22.465 + .868*output - .484*fuel -1.954*load
Time12: cost = 22.651 + .868*output - .484*fuel -1.954*load
Time13: cost = 22.617 + .868*output - .484*fuel -1.954*load
Time14: cost = 22.552 + .868*output - .484*fuel -1.954*load
Time15: cost = 22.537 + .868*output - .484*fuel -1.954*load
```

#### 5.1.1 LSDV1 without a Dummy

Let us begin with the SAS REG procedure. The test statement examines fixed time effects.

```
PROC REG DATA=masil.airline;
  MODEL cost = t1-t14 output fuel load;
RUN;
```

```

                                The REG Procedure
                                Model: MODEL1
                                Dependent Variable: cost

                                Number of Observations Read      90
                                Number of Observations Used       90

                                Analysis of Variance

                                Source                DF          Sum of
                                Squares                Mean
                                Square                F Value    Pr > F

                                Model                  17          112.95270
                                Error                  72           1.08819
                                Corrected Total        89          114.04089

                                Root MSE                0.12294    R-Square    0.9905
```



T1	20.49580478	4.2095283	4.869	.0000	.66666667E-01
T2	20.57803885	4.2215262	4.875	.0000	.66666667E-01
T3	20.65573100	4.2241771	4.890	.0000	.66666667E-01
T4	20.74075857	4.2457497	4.885	.0000	.66666667E-01
T5	21.19983202	4.4403312	4.774	.0000	.66666667E-01
T6	21.41162082	4.5386212	4.718	.0000	.66666667E-01
T7	21.50335085	4.5713968	4.704	.0000	.66666667E-01
T8	21.65402827	4.6228858	4.684	.0000	.66666667E-01
T9	21.82957108	4.6569062	4.688	.0000	.66666667E-01
T10	22.11380260	4.7926483	4.614	.0000	.66666667E-01
T11	22.46532734	4.9499089	4.539	.0000	.66666667E-01
T12	22.65133704	5.0085924	4.522	.0000	.66666667E-01
T13	22.61655508	4.9861391	4.536	.0000	.66666667E-01
T14	22.55222832	4.9559418	4.551	.0000	.66666667E-01
T15	22.53676562	4.9405321	4.562	.0000	.66666667E-01
OUTPUT	.8677267843	.15408184E-01	56.316	.0000	-1.1743092
FUEL	-.4844835367	.36410849	-1.331	.1875	12.770359
LOAD	-1.954404328	.44237771	-4.418	.0000	.56046015

(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

The following are the corresponding SAS REG procedure and Stata command for LSDV2 (outputs are skipped).

```
PROC REG DATA=masil.airline;
  MODEL cost = t1-t15 output fuel load /NOINT;
RUN;

. regress cost t1-t15 output fuel load, noc
```

### 5.1.3 LSDV3 with a Restriction

In SAS, you need to use the RESTRICT statement to impose a restriction.

```
PROC REG DATA=masil.airline;
  MODEL cost = t1-t15 output fuel load;
  RESTRICT t1+t2+t3+t4+t5+t6+t7+t8+t9+t10+t11+t12+t13+t14+t15=0;
RUN;
```

The REG Procedure  
Model: MODEL1  
Dependent Variable: cost

NOTE: Restrictions have been applied to parameter estimates.

Number of Observations Read	90
Number of Observations Used	90

#### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	112.95270	6.64428	439.62	<.0001
Error	72	1.08819	0.01511		
Corrected Total	89	114.04089			

Root MSE	0.12294	R-Square	0.9905
Dependent Mean	13.36561	Adj R-Sq	0.9882
Coeff Var	0.91981		

## Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	21.66698	4.62405	4.69	<.0001
t1	1	-1.17118	0.41783	-2.80	0.0065
t2	1	-1.08894	0.40586	-2.68	0.0090
t3	1	-1.01125	0.40323	-2.51	0.0144
t4	1	-0.92622	0.38177	-2.43	0.0178
t5	1	-0.46715	0.19076	-2.45	0.0168
t6	1	-0.25536	0.09856	-2.59	0.0116
t7	1	-0.16363	0.07190	-2.28	0.0258
t8	1	-0.01296	0.04862	-0.27	0.7907
t9	1	0.16259	0.06271	2.59	0.0115
t10	1	0.44682	0.17599	2.54	0.0133
t11	1	0.79834	0.32940	2.42	0.0179
t12	1	0.98435	0.38756	2.54	0.0132
t13	1	0.94957	0.36537	2.60	0.0113
t14	1	0.88524	0.33549	2.64	0.0102
t15	1	0.86978	0.32029	2.72	0.0083
output	1	0.86773	0.01541	56.32	<.0001
fuel	1	-0.48448	0.36411	-1.33	0.1875
load	1	-1.95440	0.44238	-4.42	<.0001
RESTRICT	-1	-3.946E-15	.	.	.

\* Probability computed using beta distribution.

In Stata, define the restriction with the `.constraint` command and specify the restriction using the `constraint()` option of the `.cnsreg` command.

```
. constraint define 3 t1+t2+t3+t4+t5+t6+t7+t8+t9+t10+t11+t12+t13+t14+t15=0
. cnsreg cost t1-t15 output fuel load, constraint(3)
```

```
Constrained linear regression                Number of obs =      90
                                             F( 17,    72) = 439.62
                                             Prob > F      = 0.0000
                                             Root MSE     = .12294

( 1)  t1 + t2 + t3 + t4 + t5 + t6 + t7 + t8 + t9 + t10 + t11 + t12 + t13 + t14 + t15 = 0
```

cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
t1	-1.171179	.4178338	-2.80	0.007	-2.004115 - .3382422
t2	-1.088945	.4058579	-2.68	0.009	-1.898008 - .2798816
t3	-1.011252	.4032308	-2.51	0.014	-1.815078 - .2074266
t4	-.9262249	.3817675	-2.43	0.018	-1.687265 - .1651852
t5	-.4671515	.1907596	-2.45	0.017	-.8474239 - .0868791
t6	-.2553627	.0985615	-2.59	0.012	-.4518415 - .0588839
t7	-.1636326	.0718969	-2.28	0.026	-.3069564 - .0203088
t8	-.0129552	.0486249	-0.27	0.791	-.1098872 .0839768
t9	.1625876	.0627099	2.59	0.012	.0375776 .2875976
t10	.4468191	.175994	2.54	0.013	.0959814 .7976568
t11	.7983439	.3294027	2.42	0.018	.1416916 1.454996
t12	.9843536	.3875583	2.54	0.013	.2117702 1.756937

t13		.9495716	.3653675	2.60	0.011	.2212248	1.677918
t14		.8852448	.3354912	2.64	0.010	.2164554	1.554034
t15		.8697821	.3202933	2.72	0.008	.2312891	1.508275
output		.8677268	.0154082	56.32	0.000	.8370111	.8984424
fuel		-.4844835	.3641085	-1.33	0.188	-1.210321	.2413535
load		-1.954404	.4423777	-4.42	0.000	-2.836268	-1.07254
_cons		21.66698	4.624053	4.69	0.000	12.4491	30.88486

The following are the corresponding LIMDEP command for LSDV3 (outputs are skipped).

```
REGRESS;Lhs=COST;Rhs=ONE,T1,T2,T3,T4,T5,T6,T7,T8,T9,T10,T11,T12,T13,T14,T15,OUTPUT,FUEL,LOAD;
CIs:b(1)+b(2)+b(3)+b(4)+b(5)+b(6)+b(7)+b(8)+b(9)+b(10)+b(11)+b(12)+b(13)+b(14)+b(15)=0$
```

## 5.2 Within Time Effect Model

The within effect mode for the fixed time effects needs to compute deviations from the time means. Keep in mind that the intercept should be suppressed.

### 5.2.1 Estimating the Time Effect Model

Let us manually estimate the fixed time effect model first.

```
. egen tm_cost = mean(cost), by(year) // compute time means
. egen tm_output = mean(output), by(year)
. egen tm_fuel = mean(fuel), by(year)
. egen tm_load = mean(load), by(year)
```

year	tm_cost	tm_output	tm_fuel	tm_load
1	12.36897	-1.790283	11.63606	.4788587
2	12.45963	-1.744389	11.66868	.4868322
3	12.60706	-1.577767	11.67494	.52358
4	12.77912	-1.443695	11.73193	.5244486
5	12.94143	-1.398122	12.26843	.5635266
6	13.0452	-1.393002	12.53826	.5541809
7	13.15965	-1.302416	12.62714	.5607425
8	13.29884	-1.222963	12.76768	.5670587
9	13.4651	-1.067003	12.86104	.6179098
10	13.70187	-.9023156	13.23183	.6233943
11	13.91324	-.9205539	13.66246	.5802577
12	14.05984	-.8641667	13.82315	.5856243
13	14.12841	-.7923916	13.75979	.5803183
14	14.23517	-.6428015	13.67403	.5804528
15	14.32062	-.5527684	13.62997	.5797168

```
. gen tw_cost = cost - tm_cost // transform variables
. gen tw_output = output - tm_output
. gen tw_fuel = fuel - tm_fuel
. gen tw_load = load - tm_load

. regress tw_cost tw_output tw_fuel tw_load, noc // within time effect
```

Source	SS	df	MS	Number of obs =	90
Model	75.6459391	3	25.215313	F( 3, 87) =	2015.95
Residual	1.08819023	87	.012507934	Prob > F	= 0.0000
				R-squared	= 0.9858
				Adj R-squared	= 0.9853
				Root MSE	= .11184
Total	76.7341294	90	.852601437		

tw_cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
tw_output	.8677268	.0140171	61.90	0.000	.8398663	.8955873
tw_fuel	-.4844836	.3312359	-1.46	0.147	-1.142851	.1738836
tw_load	-1.954404	.4024388	-4.86	0.000	-2.754295	-1.154514

If you want to get intercepts of years, use  $d_t^* = \bar{y}_{.t} - \beta' \bar{x}_{.t}$ . For example, the intercept of year 7 is  $21.503 = 13.1597 - \{.8677 * (-1.3024) + (-.4845) * 12.6271 + (-1.9544) * .5607\}$ . As discussed previously, the standard errors of the within effects model need to be adjusted. For instance, the correct standard error of fuel price is computed as  $.364 = .3312 * \text{sqrt}(87/72)$ .

## 5.2.2 Using the TSCSREG and PANEL procedures

You need to sort the data set by variables (i.e., `year` and `airline`) to appear in the `ID` statement of the TSCSREG and PANEL procedures.

```
PROC SORT DATA=masil.airline;
  BY year airline;

PROC PANEL DATA=masil.airline;
  ID year airline;
  MODEL cost = output fuel load /FIXONE;
RUN;
```

### The PANEL Procedure Fixed One Way Estimates

Dependent Variable: cost

#### Model Description

Estimation Method	FixOne
Number of Cross Sections	15
Time Series Length	6

#### Fit Statistics

SSE	1.0882	DFE	72
MSE	0.0151	Root MSE	0.1229
R-Square	0.9905		

#### F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
14	72	1.17	0.3178

#### Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr >  t	Label
----------	----	----------	----------------	---------	---------	-------

CS1	1	-2.04096	0.7347	-2.78	0.0070	Cross Sectional Effect 1
CS2	1	-1.95873	0.7228	-2.71	0.0084	Cross Sectional Effect 2
CS3	1	-1.88103	0.7204	-2.61	0.0110	Cross Sectional Effect 3
CS4	1	-1.79601	0.6988	-2.57	0.0122	Cross Sectional Effect 4
CS5	1	-1.33693	0.5060	-2.64	0.0101	Cross Sectional Effect 5
CS6	1	-1.12514	0.4086	-2.75	0.0075	Cross Sectional Effect 6
CS7	1	-1.03341	0.3764	-2.75	0.0076	Cross Sectional Effect 7
CS8	1	-0.88274	0.3260	-2.71	0.0085	Cross Sectional Effect 8
CS9	1	-0.70719	0.2947	-2.40	0.0190	Cross Sectional Effect 9
CS10	1	-0.42296	0.1668	-2.54	0.0134	Cross Sectional Effect 10
CS11	1	-0.07144	0.0718	-1.00	0.3228	Cross Sectional
CS12	1	0.114571	0.0984	1.16	0.2482	Cross Sectional Effect 12
CS13	1	0.079789	0.0844	0.95	0.3477	Cross Sectional Effect 13
CS14	1	0.015463	0.0726	0.21	0.8320	Cross Sectional Effect 14
Intercept	1	22.53677	4.9405	4.56	<.0001	Intercept
output	1	0.867727	0.0154	56.32	<.0001	
fuel	1	-0.48448	0.3641	-1.33	0.1875	
load	1	-1.9544	0.4424	-4.42	<.0001	

The following TSCSREG procedure gives the same outputs.

```
PROC TSCSREG DATA=masil.airline;
  ID year airline;
  MODEL cost = output fuel load /FIXONE;
RUN;
```

### 5.2.3 Using Stata

The Stata `.xtreg` command uses the `fe` option for the fixed effect model.

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

```
Fixed-effects (within) regression                Number of obs   =    90
Group variable (i): year                        Number of groups =    15

R-sq:  within = 0.9858                          Obs per group:  min =     6
          between = 0.4812                          avg =    6.0
          overall = 0.5265                          max =     6

corr(u_i, Xb) = -0.1503                          F(3,72)         = 1668.37
                                                Prob > F         = 0.0000
```

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

```

-----+-----
output | .8677268 .0154082 56.32 0.000 .8370111 .8984424
fuel   | -.4844835 .3641085 -1.33 0.188 -1.210321 .2413535
load   | -1.954404 .4423777 -4.42 0.000 -2.836268 -1.07254
_cons  | 21.66698 4.624053 4.69 0.000 12.4491 30.88486
-----+-----
sigma_u | .8027907
sigma_e | .12293801
rho     | .97708602 (fraction of variance due to u_i)
-----+-----
F test that all u_i=0:      F(14, 72) = 1.17          Prob > F = 0.3178

```

## 5.2.4 Using LIMDEP

You need to pay attention to the `str=;` subcommand for stratification.

--> **REGRESS;Lhs=COST;Rhs=ONE,OUTPUT,FUEL,LOAD;Panel;Str=YEAR;Fixed\$**

```

+-----+-----+
| OLS Without Group Dummy Variables |
| Ordinary least squares regression  Weighting variable = none |
| Dep. var. = COST      Mean= 13.36560933 , S.D.= 1.131971444 |
| Model size: Observations = 90, Parameters = 4, Deg.Fr.= 86 |
| Residuals: Sum of squares= 1.335449522 , Std.Dev.= .12461 |
| Fit: R-squared= .988290, Adjusted R-squared = .98788 |
| Model test: F[ 3, 86] = 2419.33, Prob value = .00000 |
| Diagnostic: Log-L = 61.7699, Restricted(b=0) Log-L = -138.3581 |
| LogAmemiyaPrCrt.= -4.122, Akaike Info. Crt.= -1.284 |
| Panel Data Analysis of COST [ONE way] |
| Unconditional ANOVA (No regressors) |
| Source Variation Deg. Free. Mean Square |
| Between 37.3068 14. 2.66477 |
| Residual 76.7341 75. 1.02312 |
| Total 114.041 89. 1.28136 |
+-----+-----+
+-----+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+-----+-----+-----+-----+-----+-----+
OUTPUT .8827386341 .13254552E-01 66.599 .0000 -1.1743092
FUEL .4539777119 .20304240E-01 22.359 .0000 12.770359
LOAD -1.627507797 .34530293 -4.713 .0000 .56046016
Constant 9.516912231 .22924522 41.514 .0000
(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

```

```

+-----+-----+
| Least Squares with Group Dummy Variables |
| Ordinary least squares regression  Weighting variable = none |
| Dep. var. = COST      Mean= 13.36560933 , S.D.= 1.131971444 |
| Model size: Observations = 90, Parameters = 18, Deg.Fr.= 72 |
| Residuals: Sum of squares= 1.088193393 , Std.Dev.= .12294 |
| Fit: R-squared= .990458, Adjusted R-squared = .98820 |
| Model test: F[ 17, 72] = 439.62, Prob value = .00000 |
| Diagnostic: Log-L = 70.9836, Restricted(b=0) Log-L = -138.3581 |
| LogAmemiyaPrCrt.= -4.010, Akaike Info. Crt.= -1.177 |
| Estd. Autocorrelation of e(i,t) .573531 |
+-----+-----+
+-----+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+-----+-----+-----+-----+-----+-----+

```

```

OUTPUT      .8677268093   .15408179E-01   56.316   .0000   -1.1743092
FUEL        -.4844946699    .36410984   -1.331   .1868   12.770359
LOAD        -1.954414378    .44237791   -4.418   .0000   .56046016
(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

```

```

+-----+
|                                     |
|               Test Statistics for the Classical Model               |
|                                     |
|      Model           Log-Likelihood      Sum of Squares      R-squared |
| (1) Constant term only      -138.35814      .1140409821D+03      .0000000 |
| (2) Group effects only      -120.52864      .7673414157D+02      .3271354 |
| (3) X - variables only       61.76991      .1335449522D+01      .9882897 |
| (4) X and group effects      70.98362      .1088193393D+01      .9904579 |
|                                     |
|                               Hypothesis Tests                       |
|                               Likelihood Ratio Test                 |
|                               Chi-squared   d.f.   Prob.           |
| (2) vs (1)      35.659      14      .00117      2.605      14      75      .00404 |
| (3) vs (1)      400.256      3      .00000      2419.329      3      86      .00000 |
| (4) vs (1)      418.684      17      .00000      439.617      17      72      .00000 |
| (4) vs (2)      383.025      3      .00000      1668.364      3      72      .00000 |
| (4) vs (3)      18.427      14      .18800      1.169      14      72      .31776 |
+-----+

```

### 5.3 Between Time Effect Model

The between effect model regresses time means of dependent variables on those of independent variables. See also 3.2 and 4.6.

```

. collapse (mean) tm_cost=cost (mean) tm_output=output (mean) tm_fuel=fuel ///
  (mean) tm_load=load, by(year)

. regress tm_cost tm_output tm_fuel tm_load // between time effect

```

Source	SS	df	MS	Number of obs = 15		
Model	6.21220479	3	2.07073493	F( 3, 11) =	4074.33	
Residual	.005590631	11	.000508239	Prob > F =	0.0000	
Total	6.21779542	14	.444128244	R-squared =	0.9991	
				Adj R-squared =	0.9989	
				Root MSE =	.02254	

tm_cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
tm_output	1.133337	.0512898	22.10	0.000	1.020449	1.246225
tm_fuel	.3342486	.0228284	14.64	0.000	.2840035	.3844937
tm_load	-1.350727	.2478264	-5.45	0.000	-1.896189	-.8052644
_cons	11.18505	.3660016	30.56	0.000	10.37949	11.99062

The SAS PANEL procedure has the /BTWNT option to estimate the between effect model.

```

PROC PANEL DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /BTWNT;
RUN;

```

The PANEL Procedure

## Between Time Periods Estimates

Dependent Variable: cost

## Model Description

Estimation Method	BtwTime
Number of Cross Sections	6
Time Series Length	15

## Fit Statistics

SSE	0.0056	DFE	11
MSE	0.0005	Root MSE	0.0225
R-Square	0.9991		

## Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr >  t	Label
Intercept	1	11.18504	0.3660	30.56	<.0001	Intercept
output	1	1.133335	0.0513	22.10	<.0001	
fuel	1	0.334249	0.0228	14.64	<.0001	
load	1	-1.35073	0.2478	-5.45	0.0002	

You may use the `be` option in the Stata `.xtreg` command and the `Means;` subcommand in LIMDEP (outputs are skipped).

```
. xtreg cost output fuel load, be i(year) // between time effect model
--> REGRESS;Lhs=COST;Rhs=ONE,OUTPUT,FUEL,LOAD;Panel;Str=YEAR;Means$
```

#### 5.4 Testing Fixed Time Effects.

The null hypothesis is that all time dummy parameters except one are zero:

$H_0 : \tau_1 = \dots = \tau_{T-1} = 0$ . The F statistic is  $\frac{(1.3354 - 1.0882)/(15 - 1)}{(1.0882)/(6 * 15 - 15 - 3)} \sim 1.1683[14,72]$ . The p-

value of .3180 does not reject the null hypothesis.

The SAS TSCSREG and PANEL procedures and the Stata `.xtreg` command conduct the Wald test. You may get the same test using the TEST statement in LSDV1 and the Stata `.test` command (the output is skipped).

```
PROC REG DATA=masil.airline;
  MODEL cost = t1-t14 output fuel load;
  TEST t1=t2=t3=t4=t5=t6=t7=t8=t9=t10=t11=t12=t13=t14=0;
RUN;

. quietly regress cost t1-t14 output fuel load
. test t1 t2 t3 t4 t5 t6 t7 t8 t9 t10 t11 t12 t13 t14
```

## 6. The Fixed Group and Time Effect Model

The two-way fixed model considers both group and time effects. This model thus needs two sets of group and time dummy variables. LSDV2 and the between effect model are not valid in this model.

### 6.1 Least Squares Dummy Variable Models

There are four approaches to avoid the perfect multicollinearity or the dummy variable trap. You may not suppress the intercept under any circumstances.

- Drop one cross-section and one time-series dummy variables.
- Drop one cross-section dummy and impose a restriction on the time-series dummies of  $\sum \tau_t = 0$
- Drop one time-series dummy and impose a restriction on the cross-section dummies of  $\sum \mu_g = 0$
- Include all dummy variables and impose two restrictions on the cross-section and time-series dummies of  $\sum \mu_g = 0$  and  $\sum \tau_t = 0$

### 6.2 LSDV1 without Two Dummies

Let us first run LSDV1 using Stata.

```
. regress cost g1-g5 t1-t14 output fuel load
```

Source	SS	df	MS	Number of obs =	90
Model	113.864044	22	5.17563838	F( 22, 67) =	1960.82
Residual	.176848775	67	.002639534	Prob > F =	0.0000
				R-squared =	0.9984
				Adj R-squared =	0.9979
Total	114.040893	89	1.28135835	Root MSE =	.05138

cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
g1	.1742825	.0861201	2.02	0.047	.0023861 .346179
g2	.1114508	.0779551	1.43	0.157	-.0441482 .2670499
g3	-.143511	.0518934	-2.77	0.007	-.2470907 -.0399313
g4	.1802087	.0321443	5.61	0.000	.1160484 .2443691
g5	-.0466942	.0224688	-2.08	0.042	-.0915422 -.0018463
t1	-.6931382	.3378385	-2.05	0.044	-1.367467 -.0188098
t2	-.6384366	.3320802	-1.92	0.059	-1.301271 .0243983
t3	-.5958031	.3294473	-1.81	0.075	-1.253383 .0617764
t4	-.5421537	.3189139	-1.70	0.094	-1.178708 .0944011
t5	-.4730429	.2319459	-2.04	0.045	-.9360088 -.0100769
t6	-.4272042	.18844	-2.27	0.027	-.8033319 -.0510764
t7	-.3959783	.1732969	-2.28	0.025	-.7418804 -.0500762
t8	-.3398463	.1501062	-2.26	0.027	-.6394596 -.040233
t9	-.2718933	.1348175	-2.02	0.048	-.5409901 -.0027964
t10	-.2273857	.0763495	-2.98	0.004	-.37978 -.0749914
t11	-.1118032	.0319005	-3.50	0.001	-.175477 -.0481295
t12	-.033641	.0429008	-0.78	0.436	-.1192713 .0519893
t13	-.0177346	.0362554	-0.49	0.626	-.0901007 .0546315
t14	-.0186451	.030508	-0.61	0.543	-.0795393 .042249

output		.8172487	.031851	25.66	0.000	.7536739	.8808235
fuel		.16861	.163478	1.03	0.306	-.1576935	.4949135
load		-.8828142	.2617373	-3.37	0.001	-1.405244	-.3603843
_cons		12.94004	2.218231	5.83	0.000	8.512434	17.36765

---

The following is the corresponding SAS REG procedure (outputs are skipped).

```
PROC REG DATA=masil.airline;
  MODEL cost = g1-g5 t1-t14 output fuel load;
RUN;
```

The LIMDEP example is skipped here, since many dummy variables need to be listed in the Regress\$ command.

### 6.3 LSDV1 + LSDV3: Dropping a Dummy and Imposing a Restriction

In the second approach, you may drop either one group dummy or one time dummy. The following drops one time dummy, includes all group dummies, and imposes a restriction on group dummies.

```
PROC REG DATA=masil.airline;
  MODEL cost = g1-g6 t1-t14 output fuel load;
  RESTRICT g1 + g2 + g3 + g4 + g5 + g6 = 0;
RUN;
```

The REG Procedure  
Model: MODEL1  
Dependent Variable: cost

NOTE: Restrictions have been applied to parameter estimates.

Number of Observations Read	90
Number of Observations Used	90

#### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	113.86404	5.17564	1960.82	<.0001
Error	67	0.17685	0.00264		
Corrected Total	89	114.04089			

Root MSE	0.05138	R-Square	0.9984
Dependent Mean	13.36561	Adj R-Sq	0.9979
Coeff Var	0.38439		

#### Parameter Estimates

Parameter	Standard
-----------	----------

Variable	DF	Estimate	Error	t Value	Pr >  t
Intercept	1	12.98600	2.22540	5.84	<.0001
g1	1	0.12833	0.04601	2.79	0.0069
g2	1	0.06549	0.03897	1.68	0.0975
g3	1	-0.18947	0.01561	-12.14	<.0001
g4	1	0.13425	0.01832	7.33	<.0001
g5	1	-0.09265	0.03731	-2.48	0.0155
g6	1	-0.04596	0.04161	-1.10	0.2733
t1	1	-0.69314	0.33784	-2.05	0.0441
t2	1	-0.63844	0.33208	-1.92	0.0588
t3	1	-0.59580	0.32945	-1.81	0.0750
t4	1	-0.54215	0.31891	-1.70	0.0938
t5	1	-0.47304	0.23195	-2.04	0.0454
t6	1	-0.42720	0.18844	-2.27	0.0266
t7	1	-0.39598	0.17330	-2.28	0.0255
t8	1	-0.33985	0.15011	-2.26	0.0268
t9	1	-0.27189	0.13482	-2.02	0.0477
t10	1	-0.22739	0.07635	-2.98	0.0040
t11	1	-0.11180	0.03190	-3.50	0.0008
t12	1	-0.03364	0.04290	-0.78	0.4357
t13	1	-0.01773	0.03626	-0.49	0.6263
t14	1	-0.01865	0.03051	-0.61	0.5432
output	1	0.81725	0.03185	25.66	<.0001
fuel	1	0.16861	0.16348	1.03	0.3061
load	1	-0.88281	0.26174	-3.37	0.0012
RESTRICT	-1	-1.9387E-16	.	.	.

\* Probability computed using beta distribution.

Alternatively, you may run the Stata `.cnsreg` command with the second constraint (output is skipped).

```
. cnsreg cost g1-g6 t1-t14 output fuel load, constraint(2)
```

The following drops one group dummy and imposes a restriction on time dummies.

```
. cnsreg cost g1-g5 t1-t15 output fuel load, constraint(3)
```

```
Constrained linear regression                Number of obs =      90
                                             F( 22,   67) = 1960.82
                                             Prob > F      = 0.0000
                                             Root MSE     =  .05138

( 1)  t1 + t2 + t3 + t4 + t5 + t6 + t7 + t8 + t9 + t10 + t11 + t12 + t13 + t14 + t15 = 0
```

cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
g1	.1742825	.0861201	2.02	0.047	.0023861 .346179
g2	.1114508	.0779551	1.43	0.157	-.0441482 .2670499
g3	-.143511	.0518934	-2.77	0.007	-.2470907 -.0399313
g4	.1802087	.0321443	5.61	0.000	.1160484 .2443691
g5	-.0466942	.0224688	-2.08	0.042	-.0915422 -.0018463
t1	-.3740245	.191872	-1.95	0.055	-.7570026 .0089536
t2	-.3193228	.1860877	-1.72	0.091	-.6907554 .0521097
t3	-.2766893	.1833501	-1.51	0.136	-.6426576 .0892789
t4	-.2230399	.1729671	-1.29	0.202	-.5682837 .1222038
t5	-.1539291	.0864404	-1.78	0.079	-.3264649 .0186066
t6	-.1080904	.0448591	-2.41	0.019	-.1976296 -.0185513
t7	-.0768646	.0319336	-2.41	0.019	-.1406043 -.0131248
t8	-.0207326	.0204506	-1.01	0.314	-.061552 .0200869
t9	.0472205	.0290822	1.62	0.109	-.0108278 .1052688

t10	.0917281	.0811525	1.13	0.262	-.0702531	.2537092
t11	.2073105	.1491443	1.39	0.169	-.0903829	.5050039
t12	.2854727	.1756365	1.63	0.109	-.0650993	.6360447
t13	.3013791	.1660294	1.82	0.074	-.030017	.6327752
t14	.3004686	.1536212	1.96	0.055	-.0061606	.6070978
t15	.3191137	.1474883	2.16	0.034	.0247259	.6135015
output	.8172487	.031851	25.66	0.000	.7536739	.8808235
fuel	.16861	.163478	1.03	0.306	-.1576935	.4949135
load	-.8828142	.2617373	-3.37	0.001	-1.405244	-.3603843
_cons	12.62093	2.074302	6.08	0.000	8.480603	16.76125

You may run the following SAS REG procedure to get the same result (output is skipped).

```
PROC REG DATA=masil.airline; /* LSDV3 */
  MODEL cost = g1-g5 t1-t15 output fuel load;
  RESTRICT t1+t2+t3+t4+t5+t6+t7+t8+t9+t10+t11+t12+t13+t14+t15=0;
RUN;
```

## 6.4 LSDV3 with Two Restrictions

The third approach includes all group and time dummies and imposes two restrictions on group and time dummies.

```
. cnsreg cost g1-g6 t1-t15 output fuel load, constraint(2 3)
```

```
Constrained linear regression                               Number of obs =      90
                                                           F( 22,    67) = 1960.82
                                                           Prob > F      = 0.0000
                                                           Root MSE     = .05138

( 1)  g1 + g2 + g3 + g4 + g5 + g6 = 0
( 2)  t1 + t2 + t3 + t4 + t5 + t6 + t7 + t8 + t9 + t10 + t11 + t12 + t13 + t14 + t15 = 0
```

cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
g1	.1283264	.0460126	2.79	0.007	.0364849	.2201679
g2	.0654947	.0389685	1.68	0.097	-.0122867	.1432761
g3	-.1894671	.0156096	-12.14	0.000	-.220624	-.1583102
g4	.1342526	.0183163	7.33	0.000	.097693	.1708121
g5	-.0926504	.0373085	-2.48	0.016	-.1671184	-.0181824
g6	-.0459561	.0416069	-1.10	0.273	-.1290038	.0370916
t1	-.3740245	.191872	-1.95	0.055	-.7570026	.0089536
t2	-.3193228	.1860877	-1.72	0.091	-.6907554	.0521097
t3	-.2766893	.1833501	-1.51	0.136	-.6426576	.0892789
t4	-.2230399	.1729671	-1.29	0.202	-.5682837	.1222038
t5	-.1539291	.0864404	-1.78	0.079	-.3264649	.0186066
t6	-.1080904	.0448591	-2.41	0.019	-.1976296	-.0185513
t7	-.0768646	.0319336	-2.41	0.019	-.1406043	-.0131248
t8	-.0207326	.0204506	-1.01	0.314	-.061552	.0200869
t9	.0472205	.0290822	1.62	0.109	-.0108278	.1052688
t10	.0917281	.0811525	1.13	0.262	-.0702531	.2537092
t11	.2073105	.1491443	1.39	0.169	-.0903829	.5050039
t12	.2854727	.1756365	1.63	0.109	-.0650993	.6360447
t13	.3013791	.1660294	1.82	0.074	-.030017	.6327752
t14	.3004686	.1536212	1.96	0.055	-.0061606	.6070978
t15	.3191137	.1474883	2.16	0.034	.0247259	.6135015
output	.8172487	.031851	25.66	0.000	.7536739	.8808235
fuel	.16861	.163478	1.03	0.306	-.1576935	.4949135
load	-.8828142	.2617373	-3.37	0.001	-1.405244	-.3603843
_cons	12.66688	2.081068	6.09	0.000	8.513054	16.82071

The following SAS REG procedure gives you the same result (output is skipped).

```
PROC REG DATA=masil.airline;
```

<http://www.indiana.edu/~statmath>

```

MODEL cost = g1-g6 t1-t15 output fuel load;
RESTRICT g1 + g2 + g3 + g4 + g5 + g6 = 0;
RESTRICT t1+t2+t3+t4+t5+t6+t7+t8+t9+t10+t11+t12+t13+t14+t15=0;
RUN;

```

## 6.5 Two-way Within Effect Model

The two-way within group and time effect model requires a transformation of the data set as  $y_{it}^* = y_{it} - \bar{y}_{i\cdot} - \bar{y}_{\cdot t} + \bar{y}_{\cdot\cdot}$  and  $x_{it}^* = x_{it} - \bar{x}_{i\cdot} - \bar{x}_{\cdot t} + \bar{x}_{\cdot\cdot}$ . The following commands do this task.

```

. gen w_cost = cost - gm_cost - tm_cost + m_cost
. gen w_output = output - gm_output - tm_output + m_output
. gen w_fuel = fuel - gm_fuel - tm_fuel + m_fuel
. gen w_load = load - gm_load - tm_load + m_load

. tabstat cost output fuel load, stat(mean)

```

stats	cost	output	fuel	load
-----+-----				
mean	13.36561	-1.174309	12.77036	.5604602
-----+-----				

Now, run the OLS with the transformed variables. Do not forget to suppress the intercept.

```

. regress w_cost w_output w_fuel w_load, noc // within effect

```

Source	SS	df	MS	Number of obs =	90
-----+-----				F( 3, 87) =	307.86
Model	1.87739643	3	.625798811	Prob > F =	0.0000
Residual	.176848774	87	.002032745	R-squared =	0.9139
-----+-----				Adj R-squared =	0.9109
Total	2.05424521	90	.022824947	Root MSE =	.04509

w_cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
w_output	.8172487	.0279512	29.24	0.000	.7616927 .8728048
w_fuel	.16861	.1434621	1.18	0.243	-.1165364 .4537565
w_load	-.8828142	.2296907	-3.84	0.000	-1.339349 -.426279
-----+-----					

Note again that  $R^2$ , MSE, standard errors, and  $DF_{\text{error}}$  are not correct. The dummy variable coefficients are computed as  $d_g^* = (\bar{y}_{g\cdot} - \bar{y}_{\cdot\cdot}) - b'(\bar{x}_{g\cdot} - \bar{x}_{\cdot\cdot})$  and  $d_t^* = (\bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot}) - b'(\bar{x}_{\cdot t} - \bar{x}_{\cdot\cdot})$ . The standard errors also need to be adjusted; for instance, the standard error of the load factor is  $.2617 = .2297 * \sqrt{87/67}$ .

## 6.6 Using the TSCSREG and PANEL Procedures

The SAS TSCSREG and PANEL procedures have the /FIXTWO option to fit the two-way fixed effect model.

```

PROC TSCSREG DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /FIXTWO;
RUN;

```

## The TSCSREG Procedure

Dependent Variable: cost

## Model Description

Estimation Method	FixTwo
Number of Cross Sections	6
Time Series Length	15

## Fit Statistics

SSE	0.1768	DFE	67
MSE	0.0026	Root MSE	0.0514
R-Square	0.9984		

## F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
19	67	23.10	<.0001

## Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr >  t	Label
CS1	1	0.174283	0.0861	2.02	0.0470	Cross Sectional Effect 1
CS2	1	0.111451	0.0780	1.43	0.1575	Cross Sectional Effect 2
CS3	1	-0.14351	0.0519	-2.77	0.0073	Cross Sectional Effect 3
CS4	1	0.180209	0.0321	5.61	<.0001	Cross Sectional Effect 4
CS5	1	-0.04669	0.0225	-2.08	0.0415	Cross Sectional Effect 5
TS1	1	-0.69314	0.3378	-2.05	0.0441	Time Series Effect 1
TS2	1	-0.63844	0.3321	-1.92	0.0588	Time Series Effect 2
TS3	1	-0.5958	0.3294	-1.81	0.0750	Time Series Effect 3
TS4	1	-0.54215	0.3189	-1.70	0.0938	Time Series Effect 4
TS5	1	-0.47304	0.2319	-2.04	0.0454	Time Series Effect 5
TS6	1	-0.4272	0.1884	-2.27	0.0266	Time Series Effect 6
TS7	1	-0.39598	0.1733	-2.28	0.0255	Time Series Effect 7
TS8	1	-0.33985	0.1501	-2.26	0.0268	Time Series Effect 8

TS9	1	-0.27189	0.1348	-2.02	0.0477	Time Series Effect 9
TS10	1	-0.22739	0.0763	-2.98	0.0040	Time Series Effect 10
TS11	1	-0.1118	0.0319	-3.50	0.0008	Time Series Effect 11
TS12	1	-0.03364	0.0429	-0.78	0.4357	Time Series Effect 12
TS13	1	-0.01773	0.0363	-0.49	0.6263	Time Series Effect 13
TS14	1	-0.01865	0.0305	-0.61	0.5432	Time Series Effect 14
Intercept	1	12.94004	2.2182	5.83	<.0001	Intercept
output	1	0.817249	0.0319	25.66	<.0001	
fuel	1	0.16861	0.1635	1.03	0.3061	
load	1	-0.88281	0.2617	-3.37	0.0012	

## 6.7 Using Stata and LIMDEP

The Stata `.xtreg` command does not fit the two-way fixed or random effect model.

The Stata `.xtreg` command does not have an option for the two-way fixed and random effect model. However, the command fits the two-way fixed effect model by including a set of dummies for a group (LSDV1) and using the `fe` option.

```
. xtreg cost t1-t14 output fuel load, fe

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

R-sq:  within = 0.9955                       Obs per group:  min =       15
        between = 0.9859                      avg =      15.0
        overall = 0.9885                      max =       15

corr(u_i, Xb) = 0.3361                       F(17,67)       =     873.24
                                                Prob > F       =     0.0000
```

cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
t1	-.6931382	.3378385	-2.05	0.044	-1.367467 - .0188098
t2	-.6384366	.3320802	-1.92	0.059	-1.301271 .0243983
t3	-.5958031	.3294473	-1.81	0.075	-1.253383 .0617764
t4	-.5421537	.3189139	-1.70	0.094	-1.178708 .0944011
t5	-.4730429	.2319459	-2.04	0.045	-.9360088 -.0100769
t6	-.4272042	.18844	-2.27	0.027	-.8033319 -.0510764
t7	-.3959783	.1732969	-2.28	0.025	-.7418804 -.0500762
t8	-.3398463	.1501062	-2.26	0.027	-.6394596 -.040233
t9	-.2718933	.1348175	-2.02	0.048	-.5409901 -.0027964
t10	-.2273857	.0763495	-2.98	0.004	-.37978 -.0749914
t11	-.1118032	.0319005	-3.50	0.001	-.175477 -.0481295
t12	-.033641	.0429008	-0.78	0.436	-.1192713 .0519893
t13	-.0177346	.0362554	-0.49	0.626	-.0901007 .0546315
t14	-.0186451	.030508	-0.61	0.543	-.0795393 .042249
output	.8172487	.031851	25.66	0.000	.7536739 .8808235
fuel	.16861	.163478	1.03	0.306	-.1576935 .4949135
load	-.8828142	.2617373	-3.37	0.001	-1.405244 -.3603843
_cons	12.986	2.225402	5.84	0.000	8.544076 17.42792
sigma_u	.1306712				

```

sigma_e | .05137639
rho      | .86611203   (fraction of variance due to u_i)
-----
F test that all u_i=0:      F(5, 67) =      69.05          Prob > F = 0.0000

```

The F statistic of 69.05 tests only if parameters of g1 through g5 are all zero. When running the following command after the LSDV1 above, you should get the identical result.

```

. test g1=g2=g3=g4=g5=0

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

      F( 5, 67) =      69.05
      Prob > F =      0.0000

```

The following LIMDEP command fits the two-way fixed model. Note that this command has `Str$` and `Period$` specifications to specify stratification and time variables. This command presents the pooled model and one-way group effect model as well, but reports the incorrect intercept in the two-way fixed model, 12.667 (2.081).

```
REGRESS;Lhs=COST;Rhs=ONE,OUTPUT,FUEL,LOAD;Panel;Str=AIRLINE;Period=YEAR;Fixed$
```

## 6.8 Testing Fixed Group and Time Effects

The null hypothesis is that parameters of group and time dummies are zero:

$H_0 : \mu_1 = \dots = \mu_{n-1} = 0$  and  $\tau_1 = \dots = \tau_{T-1} = 0$ . The F test compares the pooled regression and two-way group and time effect model. The F statistic of 23.1085 rejects the null hypothesis at the .01 significance level ( $p < .0000$ ).

$$\frac{(1.3354 - .1768)/(6 + 15 - 2)}{(.1768)/(6 * 15 - 6 - 15 - 3 + 1)} \sim 23.1085[19,67]$$

The SAS TSCSREG and PANEL procedures conduct the F-test for the group and time effects. You may also run the following SAS REG procedure and `.regress` command to perform the same test.

```

PROC REG DATA=masil.airline;
  MODEL cost = g1-g5 t1-t14 output fuel load;
  TEST g1=g2=g3=g4=g5=t1=t2=t3=t4=t5=t6=t7=t8=t9=t10=t11=t12=t13=t14=0;
RUN;

. quietly regress cost g1-g5 t1-t14 output fuel load
. test g1 g2 g3 g4 g5 t1 t2 t3 t4 t5 t6 t7 t8 t9 t10 t11 t12 t13 t14

```

## 7. Random Effect Models

The random effects model examines how group and/or time affect error variances. This model is appropriate for  $n$  individuals who were drawn randomly from a large population. This chapter focuses on the feasible generalized least squares (FGLS) with variance component estimation methods from Baltagi and Chang (1994), Fuller and Battese (1974), and Wansbeek and Kapteyn (1989).<sup>10</sup>

### 7.1 The One-way Random Group Effect Model

When the omega matrix is not known, you have to estimate  $\theta$  using the SSEs of the pooled model (.0317) and the fixed effect model (.2926).

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

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

$$\text{Thus, } \hat{\theta} \text{ is } .87668488 = 1 - \sqrt{\frac{.00361263}{15 * .031675926 / (6 - 4)}}$$

Now, transform the dependent and independent variables including the intercept.

```
. gen rg_cost = cost - .87668488*gm_cost // transform variables
. 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
```

Finally, run the OLS with the transformed variables. Do not forget to suppress the intercept. This is the groupwise heteroscedastic regression model (Greene 2003).

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

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

rg_cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
rg_int	9.627911	.2101638	45.81	0.000	9.210119	10.0457

<sup>10</sup> 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.

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

## 7.2 Estimations in SAS, Stata, and LIMDEP

The SAS TSCSREG and PANEL procedures have the /RANONE option to fit the one-way random effect model. These procedures by default use the Fuller and Battese (1974) estimation method, which produces slightly different estimates from FGLS.

```
PROC TSCSREG DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /RANONE;
RUN;
```

### The TSCSREG Procedure

Dependent Variable: cost

#### Model Description

Estimation Method	RanOne
Number of Cross Sections	6
Time Series Length	15

#### Fit Statistics

SSE	0.3090	DFE	86
MSE	0.0036	Root MSE	0.0599
R-Square	0.9923		

#### Variance Component Estimates

Variance Component for Cross Sections	0.018198
Variance Component for Error	0.003613

#### Hausman Test for Random Effects

DF	m Value	Pr > m
3	0.92	0.8209

#### Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr >  t
Intercept	1	9.637	0.2132	45.21	<.0001
output	1	0.908024	0.0260	34.91	<.0001

fuel	1	0.422199	0.0141	29.95	<.0001
load	1	-1.06469	0.1995	-5.34	<.0001

The PANEL procedure has the /VCOMP=WK option for the Wansbeek and Kapteyn (1989) method, which is close to groupwise heteroscedastic regression. The BP option of the MODEL statement, not available in the TSCSREG procedure, conducts the Breusch-Pagan LM test for random effects. Note that two procedures estimate the same variance component for error (.0036) but a different variance component for groups (.0182 versus .0160),

```
PROC PANEL DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /RANONE BP VCOMP=WK;
RUN;
```

The PANEL Procedure  
Wansbeek and Kapteyn Variance Components (RanOne)

Dependent Variable: cost

Model Description

Estimation Method	RanOne
Number of Cross Sections	6
Time Series Length	15

Fit Statistics

SSE	0.3111	DFE	86
MSE	0.0036	Root MSE	0.0601
R-Square	0.9923		

Variance Component Estimates

Variance Component for Cross Sections	0.016015
Variance Component for Error	0.003613

Hausman Test for  
Random Effects

DF	m Value	Pr > m
2	1.63	0.4429

Breusch Pagan Test for Random  
Effects (One Way)

DF	m Value	Pr > m
1	334.85	<.0001

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr >  t
Intercept	1	9.629513	0.2107	45.71	<.0001
output	1	0.906918	0.0257	35.30	<.0001
fuel	1	0.422676	0.0140	30.11	<.0001
load	1	-1.06452	0.2000	-5.32	<.0001

The Stata `.xtreg` command has the `re` option to produce FGLS estimates. The `.iis` command specifies the panel identification variable, such as a grouping or cross-section variable that is used in the `i()` option.

```
. iis airline
```

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

```
Random-effects GLS regression                Number of obs   =       90
Group variable (i): 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
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
cost					
output	.9066805	.0256225	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 `theta` option reports the estimated `theta` (.8767). The `sigma_u` and `sigma_e` are square roots of the variance components for groups and errors (.0036=.0601<sup>2</sup>).

You may use maximum likelihood estimation to fit random effect (or random intercept) model. The `mle` option is used in `.xtreg` and `.xtmixed` commands.

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

```
Fitting constant-only model:
Iteration 0:  log likelihood = -4211.897
Iteration 1:  log likelihood = -2338.9839
Iteration 2:  log likelihood = -1296.6479
Iteration 3:  log likelihood = -721.39534
Iteration 4:  log likelihood = -408.67141
Iteration 5:  log likelihood = -243.2565
Iteration 6:  log likelihood = -160.06656
Iteration 7:  log likelihood = -122.03417
Iteration 8:  log likelihood = -107.60971
Iteration 9:  log likelihood = -103.87526
Iteration 10: log likelihood = -103.44067
Iteration 11: log likelihood = -103.4311
Iteration 12: log likelihood = -103.4311
```

Fitting full model:

```
Iteration 0: log likelihood = 108.79821
Iteration 1: log likelihood = 114.43768
Iteration 2: log likelihood = 114.72738
Iteration 3: log likelihood = 114.72896
Iteration 4: log likelihood = 114.72896
```

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

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

Log likelihood = 114.72896           LR chi2(3)      =    436.32
                                       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   .0138888   30.48  0.000   .3961557   .4505957
        load |  -1.064456   .1962311   -5.42  0.000  -1.449062  -.6798506
        _cons |   9.618648   .2066222   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

The `|| airline:` option tells Stata to fit the random-intercept model using the group variable `airline`.

```
. 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

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

```
-----+-----
      cost |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      output |   .9053099   .0246566   36.72  0.000   .8569851   .9536347
        fuel |   .4233757   .0136369   31.05  0.000   .3966488   .4501035
        load |  -1.064456   .1962309   -5.42  0.000  -1.449062  -.6798508
        _cons |   9.618648   .2026096   47.47  0.000   9.221541  10.01576
-----+-----
```

```
-----+-----
Random-effects Parameters |      Estimate   Std. Err.     [95% Conf. Interval]
-----+-----
airline: Identity        |
      sd(_cons) |   .1140843   .0345293     .0630373   .2064687
-----+-----
```

```
-----+-----
                sd(Residual) | .0591072 .0045701 .0507956 .0687787
-----+-----
LR test vs. linear regression: chibar2(01) = 105.92 Prob >= chibar2 = 0.0000
```

Alternatively you may use `.xtgls` that fits panel data models with heteroscedasticity and/or autocorrelation across and within groups. The estimators of `.xtreg` and `.xtgls` are slightly different.

```
. xtgls cost output fuel load, i(airline) panels(hetero)
```

Cross-sectional time-series FGLS regression

```
Coefficients: generalized least squares
Panels:       heteroskedastic
Correlation:  no autocorrelation
```

```
Estimated covariances      =          6      Number of obs      =          90
Estimated autocorrelations =          0      Number of groups     =          6
Estimated coefficients      =          4      Time periods         =          15
Log likelihood              = 90.34275      Wald chi2(3)         = 25228.72
                          Prob > chi2      =          0.0000
```

```
-----+-----
                cost |      Coef.   Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+-----
output |      .8921376   .0068822   129.63  0.000   .8786487   .9056266
fuel   |      .4296884   .0125486   34.24  0.000   .4050935   .4542833
load   |     -1.426202   .2064744   -6.91  0.000   -1.830885  -1.02152
_cons  |      9.730364   .1346436   72.27  0.000   9.466468   9.994261
-----+-----
```

In LIMDEP, you have to specify `Panel$` and `Het$` subcommands for the groupwise heteroscedastic model. Note that LIMDEP presents the pooled OLS regression and least squares dummy variable model as well.

```
--> REGRESS;Lhs=COST;Rhs=ONE,OUTPUT,FUEL,LOAD;Panel;Str=AIRLINE;Het=AIRLINE$
```

```
+-----+-----+-----+-----+-----+-----+
| OLS Without Group Dummy Variables |
| Ordinary least squares regression  Weighting variable = none |
| Dep. var. = COST Mean= 13.36560933 , S.D.= 1.131971444 |
| Model size: Observations = 90, Parameters = 4, Deg.Fr.= 86 |
| Residuals: Sum of squares= 1.335449522 , Std.Dev.= .12461 |
| Fit: R-squared= .988290, Adjusted R-squared = .98788 |
| Model test: F[ 3, 86] = 2419.33, Prob value = .00000 |
| Diagnostic: Log-L = 61.7699, Restricted(b=0) Log-L = -138.3581 |
| LogAmemiyaPrCrt.= -4.122, Akaike Info. Crt.= -1.284 |
| Panel Data Analysis of COST [ONE way] |
| Unconditional ANOVA (No regressors) |
| Source Variation Deg. Free. Mean Square |
| Between 74.6799 5. 14.9360 |
| Residual 39.3611 84. .468584 |
| Total 114.041 89. 1.28136 |
+-----+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+-----+-----+-----+-----+-----+-----+
OUTPUT .8827386341 .13254552E-01 66.599 .0000 -1.1743092
FUEL .4539777119 .20304240E-01 22.359 .0000 12.770359
```

```

LOAD          -1.627507797      .34530293   -4.713   .0000   .56046016
Constant      9.516912231      .22924522   41.514   .0000

```

(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

```

+-----+
| Least Squares with Group Dummy Variables |
| Ordinary least squares regression        |
| Weighting variable = none               |
| Dep. var. = COST      Mean= 13.36560933 |
|                               , S.D.= 1.131971444 |
| Model size: Observations = 90, Parameters = 9, Deg.Fr.= 81 |
| Residuals: Sum of squares= .2926207777 |
|                               , Std.Dev.= .06010 |
| Fit:      R-squared= .997434, Adjusted R-squared = .99718 |
| Model test: F[ 8, 81] = 3935.82, Prob value = .00000 |
| Diagnostic: Log-L = 130.0865, Restricted(b=0) Log-L = -138.3581 |
|                               LogAmemiyaPrCrt.= -5.528, Akaike Info. Crtr.= -2.691 |
| Estd. Autocorrelation of e(i,t) .573531 |
| White/Hetero. corrected covariance matrix used. |
+-----+

```

```

+-----+
| Variable | Coefficient | Standard Error | t-ratio | P[|T|>t] | Mean of X |
+-----+
OUTPUT    .9192881432 .19105357E-01  48.117 .0000 -1.1743092
FUEL      .4174910457 .13532534E-01  30.851 .0000 12.770359
LOAD      -1.070395015 .21662097 -4.941 .0000 .56046016

```

(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

```

+-----+
|                               Test Statistics for the Classical Model |
|                               |
| Model          Log-Likelihood   Sum of Squares   R-squared |
| (1) Constant term only      -138.35814      .1140409821D+03 .0000000 |
| (2) Group effects only      -90.48804       .3936109461D+02 .6548513 |
| (3) X - variables only       61.76991       .1335449522D+01 .9882897 |
| (4) X and group effects     130.08647       .2926207777D+00 .9974341 |
|                               |
|                               Hypothesis Tests |
|                               |
|                               Likelihood Ratio Test |
|                               |
| Chi-squared  d.f.  Prob.      F      num. denom.  Prob value |
| (2) vs (1)   95.740  5      .00000   31.875  5      84      .00000 |
| (3) vs (1)   400.256  3      .00000  2419.329  3      86      .00000 |
| (4) vs (1)   536.889  8      .00000  3935.818  8      81      .00000 |
| (4) vs (2)   441.149  3      .00000  3604.832  3      81      .00000 |
| (4) vs (3)   136.633  5      .00000   57.733  5      81      .00000 |
+-----+

```

Error: 425: REGR;PANEL. Could not invert VC matrix for Hausman test.

```

+-----+
| Random Effects Model: v(i,t) = e(i,t) + u(i) |
| Estimates: Var[e] = .361260D-02 |
|              Var[u] = .119159D-01 |
|              Corr[v(i,t),v(i,s)] = .767356 |
| Lagrange Multiplier Test vs. Model (3) = 334.85 |
| ( 1 df, prob value = .000000) |
| (High values of LM favor FEM/REM over CR model.) |
| Fixed vs. Random Effects (Hausman) = .00 |
| ( 3 df, prob value = 1.000000) |
| (High (low) values of H favor FEM (REM).) |
| Reestimated using GLS coefficients: |
+-----+

```

```

| Estimates: Var[e]          = .362491D-02 |
|           Var[u]          = .392309D-01 |
| Var[e] above is an average. Groupwise |
| heteroscedasticity model was estimated. |
|           Sum of Squares   .147779D+01 |
+-----+

```

```

+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+-----+-----+-----+-----+-----+
OUTPUT   .9041238041  .24615477E-01  36.730  .0000  -1.1743092
FUEL     .4238986905  .13746498E-01  30.837  .0000  12.770359
LOAD    -1.064558659  .19933132  -5.341  .0000  .56046016
Constant 9.610634379  .20277404  47.396  .0000
(Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

```

Like SAS TSCSREG and PANEL procedures, LIMDEP estimates a slightly different variance component for groups (.0119), thus producing different parameter estimates. In addition, the Hausman test is not successful in this example.

### 7.3 The One-way Random Time Effect Model

Let us compute  $\hat{\theta}$  using the SSEs of the between effect model (.0056) and the fixed effect model (1.0882).

The variance component for error  $\hat{\sigma}_\varepsilon^2$  is  $.01511375 = 1.08819022/(15*6-15-3)$

The variance component for time  $\hat{\sigma}_v^2$  is  $-.00201072 = .005590631/(15-4) - .01511375/6$

The  $\hat{\theta}$  is  $-1.226263 = 1 - \sqrt{\frac{.01511375}{6*.005590631/(15-4)}}$

```

. gen rt_cost = cost - (-1.226263)*tm_cost // transform variables
. gen rt_output = output - (-1.226263)*tm_output
. gen rt_fuel = fuel - (-1.226263)*tm_fuel
. gen rt_load = load - (-1.226263)*tm_load
. gen rt_int = 1 - (-1.226263) // for the intercept

. regress rt_cost rt_int rt_output rt_fuel rt_load, noc

```

Source	SS	df	MS	Number of obs =	90
Model	79944.1804	4	19986.0451	F( 4, 86) =	.
Residual	1.79271995	86	.020845581	Prob > F =	0.0000
Total	79945.9732	90	888.288591	R-squared =	1.0000
				Adj R-squared =	1.0000
				Root MSE =	.14438

rt_cost	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
rt_int	9.516098	.1489281	63.90	0.000	9.220038 9.812157
rt_output	.8883838	.0143338	61.98	0.000	.8598891 .9168785
rt_fuel	.4392731	.0129051	34.04	0.000	.4136186 .4649277
rt_load	-1.279176	.2482869	-5.15	0.000	-1.772754 -.7855982

However, the negative value of the variance component for time is not likely. This section presents examples of procedures and commands for the one-way time random effect model without outputs.

In SAS, use the TSCSREG or PANEL procedure with the /RANONE option.

```
PROC SORT DATA=masil.airline;
  BY year airline;

PROC TSCSREG DATA=masil.airline;
  ID year airline;
  MODEL cost = output fuel load /RANONE;
RUN;

PROC PANEL DATA=masil.airline;
  ID year airline;
  MODEL cost = output fuel load /RANONE BP;
RUN;
```

In Stata, you have to switch the grouping and time variables using the `.tsset` command.

```
. tsset year airline
      panel variable:  year, 1 to 15
      time variable:  airline, 1 to 6

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

In LIMDEP, you need to use the `Period$` and `Random$` subcommands.

```
REGRESS;Lhs=COST;Rhs=ONE,OUTPUT,FUEL,LOAD;Panel;Pds=15;Het=YEAR$
```

## 7.4 The Two-way Random Effect Model in SAS

The random group and time effect model is formulated as  $y_{it} = \alpha + \beta' X_{it} + u_i + \gamma_t + \varepsilon_{it}$ . Let us first estimate the two way FGLS using the SAS PANEL procedure with the /RANTWO option. The BP2 option conducts the Breusch-Pagan LM test for the two-way random effect model.

```
PROC PANEL DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /RANTWO BP2;
RUN;
```

The PANEL Procedure  
Fuller and Battese Variance Components (RanTwo)

Dependent Variable: cost

### Model Description

Estimation Method	RanTwo
Number of Cross Sections	6

Time Series Length 15

## Fit Statistics

SSE	0.2322	DFE	86
MSE	0.0027	Root MSE	0.0520
R-Square	0.9829		

## Variance Component Estimates

Variance Component for Cross Sections	0.017439
Variance Component for Time Series	0.001081
Variance Component for Error	0.00264

Hausman Test for  
Random Effects

DF	m Value	Pr > m
3	6.93	0.0741

Breusch Pagan Test for Random  
Effects (Two Way)

DF	m Value	Pr > m
2	336.40	<.0001

## Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr >  t
Intercept	1	9.362677	0.2440	38.38	<.0001
output	1	0.866448	0.0255	33.98	<.0001
fuel	1	0.436163	0.0172	25.41	<.0001
load	1	-0.98053	0.2235	-4.39	<.0001

Similarly, you may run the TSCSREG procedure with the /RANTWO option.

```
PROC TSCSREG DATA=masil.airline;
  ID airline year;
  MODEL cost = output fuel load /RANTWO;
RUN;
```

## 7.5 Testing Random Effect Models

The Breusch-Pagan Lagrange multiplier (LM) test is designed to test random effects. The null hypothesis of the one-way random group effect model is that variances of groups are zero:

$H_0 : \sigma_u^2 = 0$ . If the null hypothesis is not rejected, the pooled regression model is appropriate. The  $e'e$  of the pooled OLS is 1.33544153 and  $\bar{e}'\bar{e}$  is .0665147.

$$\text{LM is } 334.8496 = \frac{6 * 15}{2(15-1)} \left[ \frac{15^2 * .0665}{1.3354} - 1 \right]^2 \sim \chi^2(1) \text{ with } p < .0000.$$

With the large chi-squared, we reject the null hypothesis in favor of the random group effect model. The SAS PANEL procedure with the /BP option and the LIMDEP Panel\$ and Het\$ subcommands report the LM statistic. In Stata, run the .xttest0 command right after estimating the one-way random effect model.

```
. 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

The null hypothesis of the one-way random time effect is that variance components for time are zero,  $H_0 : \sigma_v^2 = 0$ . The following LM test uses Baltagi's formula. The small chi-squared of 1.5472 does not reject the null hypothesis at the .01 level.

$$\text{LM is } 1.5472 = \frac{Tn}{2(n-1)} \left[ \frac{\sum (n\bar{e}_{.t})^2}{\sum \sum e_{it}^2} - 1 \right]^2 = \frac{15 * 6}{2(6-1)} \left[ \frac{.7817}{1.3354} - 1 \right]^2 \sim \chi^2(1) \text{ with } p < .2135$$

```
. quietly xtreg cost output fuel load, re i(year)
. xttest0
```

Breusch and Pagan Lagrangian multiplier test for random effects:

```
cost[year,t] = Xb + u[year] + e[year,t]
```

Estimated results:

	Var	sd = sqrt(Var)
cost	1.281358	1.131971
e	.0151138	.122938
u	0	0

Test: Var(u) = 0

chi2(1) = 1.55  
Prob > chi2 = 0.2135

The two way random effects model has the null hypothesis that variance components for groups and time are all zero. The LM statistic with two degrees of freedom is  $336.3968 = 334.8496 + 1.5472$  ( $p < .0001$ ).

## 7.6 Fixed Effects versus Random Effects

How do we compare a fixed effect model and its counterpart random effect model? The Hausman specification test examines if the individual effects are uncorrelated with the other regressors in the model. Since computation is complicated, let us conduct the test in Stata.

```
. tsset airline year
      panel variable:  airline, 1 to 6
      time variable:  year, 1 to 15

. quietly xtreg cost output fuel load, fe

. estimates store fixed_group

. quietly xtreg cost output fuel load, re

. hausman fixed_group .
```

```

      ---- Coefficients ----
      |          (b)          (B)          (b-B)          sqrt(diag(V_b-V_B))
      |          fix_group          .          Difference          S.E.
-----+-----
output |          .9192846          .9066805          .0126041          .0153877
fuel   |          .4174918          .4227784          -.0052867          .0058583
load   |          -1.070396          -1.064499          -.0058974          .0255088
-----+-----
      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
      Prob>chi2 =          0.5469
      (V_b-V_B is not positive definite)
```

The Hausman statistic 2.12 is different from the PANEL procedure's 1.63 and Greene (2003)'s 4.16. It is because SAS, Stata, and LIMDEP use different estimation methods to produce slightly different parameter estimates. These tests, however, do not reject the null hypothesis in favor of the random effect model.

## 7.7 Summary

Table 7 summarizes random effect estimations in SAS, Stata, and LIMDEP. The SAS PANEL procedure is highly recommended.

Table 7 Comparison of the Random Effect Model in SAS, Stata, LIMDEP\*

	SAS 9		Stata 9	LIMDEP 8
Procedure/Command	PROC TSCSREG	PROC PANEL	. xtreg	Regress; Panel\$
One-way	/RANONE	/RANONE WK	re	Str=;Pds=;Het;Random\$
Two-way	/RANTWO	/RANTWO	No	Problematic

SSE ( $e'e$ )	Slightly different	Correct	No	No
MSE or SEE	Slightly different	Correct	No	No
Model test (F)	No	No	Wald test	No
(adjusted) $R^2$	Slightly different	Slightly different	Incorrect	No
Intercept	Slightly different	Correct	Correct	Slightly different
Coefficients	Slightly different	Correct	Correct	Slightly different
Standard errors	Slightly different	Correct	Correct	Slightly different
Variance for group	Slightly different	Correct	Correct (sigma)	Slightly different
Variance for error	Correct	Correct	Correct (sigma)	Correct
Theta	No	No	theta	No
Breusch-Pagan (LM)	No	BP option	. xttest0	Yes
Hausman Test (H)	Incorrect	Yes	. hausman	Yes (unstable)

\* “Yes/No” means whether the software reports the statistics. “Correct/incorrect” indicates whether the statistics are different from those of the groupwise heteroscedastic regression.

## 8. The Poolability Test

In order to conduct the poolability test, you need to run group by group OLS regressions and/or time by time OLS regressions. If the null hypothesis is rejected, the panel data are not poolable. In this case, you may consider the random coefficient model and hierarchical regression model.

### 8.1 Group by Group OLS Regression

In SAS, use the BY statement in the REG procedure. Do not forget to sort the data set in advance.

```
PROC SORT DATA=masil.airline;
  BY airline;

PROC REG DATA=masil.airline;
  MODEL cost = output fuel load;
  BY airline;
RUN;
```

In Stata, the `if` qualifier makes it easy to run group by group regressions.

```
. forvalues i= 1(1)6 { // run group by group regression
  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		
Model	3.41824348	3	1.13941449	F( 3, 11)	=	1843.46
Residual	.006798918	11	.000618083	Prob > F	=	0.0000
Total	3.4250424	14	.244645886	R-squared	=	0.9980
				Adj R-squared	=	0.9975
				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		
Model	6.47622084	3	2.15874028	F( 3, 11)	=	3129.50
Residual	.007587838	11	.000689803	Prob > F	=	0.0000
Total	6.48380868	14	.463129191	R-squared	=	0.9988
				Adj R-squared	=	0.9985
				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
Model	3.79286673	3	1.26428891	F( 3, 11) =	608.10
Residual	.022869767	11	.00207907	Prob > F	= 0.0000
				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
Model	7.37252558	3	2.45750853	F( 3, 11) =	777.86
Residual	.034752343	11	.003159304	Prob > F	= 0.0000
				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
Model	7.08313716	3	2.36104572	F( 3, 11) =	1999.89
Residual	.012986435	11	.001180585	Prob > F	= 0.0000
				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
Model	11.1173565	3	3.70578551	F( 3, 11) =	2602.49
Residual	.015663323	11	.001423938	Prob > F	= 0.0000
				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
-----+-----

```

## 8.2 Poolability Test across Groups

The null hypothesis of the poolability test across groups is  $H_0 : \beta_{ik} = \beta_k$ . The  $e'e$  is 1.3354, the SSE of the pooled OLS regression. The  $e_i'e_i$  is  $.1007 = .0068 + .0076 + .0229 + .0348 + .0130 + .0157$ .

Thus, the F statistic is  $\frac{(1.3354 - .1007)/(6-1)4}{.1007/6(15-4)} \sim 40.4812[20,66]$

The large 40.4812 rejects the null hypothesis of poolability ( $p < .0000$ ). We conclude that the panel data are not poolable with respect to group.

## 8.3 Poolability Test over Time

The null hypothesis of the poolability test over time is  $H_0 : \beta_{ik} = \beta_k$ . The sum of  $e_t'e_t$  is computed from the 15 time by time regression.

```

. di .044807673 + .023093978 + .016506613 + .012170358 + .014104542 + ///
      .000469826 + .063648817 + .085430285 + .049329439 + .077112957 + ///
      .029913538 + .087240016 + .143348297 + .066075346 + .037256216

.7505079

```

The F statistic is  $.4175[84,30] = \frac{(1.3354 - .7505)/(15-1)4}{.7505/15(6-4)}$

The small F statistic does not reject the null hypothesis in favor of poolable panel data with respect to time ( $p < .9991$ ).

## 9. Conclusion

Panel data models investigate group and time effects using fixed effect and random effect models. The fixed effect model asks how group and/or time affect the intercept, while the random effect model analyzes error variance structures affected by group and/or time. Slopes are assumed unchanged in both fixed effect and random effect models.

A panel data set needs to be arranged in the long format as shown in 1.1. If the number of groups (subjects) or time periods is extremely large, panel data models may be less useful because the null hypothesis of F test is too strong. Then, you may consider categorizing subjects to reduce the number of groups. If data are severely unbalanced, read output with caution and consider dropping subjects with many missing data points. This document assumes that data are balanced without missing values.

Fixed effect models are estimated by the least squares dummy variable (LSDV) regression, within effect model, and between effect model. LSDV has three approaches to avoid perfect multicollinearity. LSDV1 drops a dummy, LSDV2 suppresses the intercept, and LSDV3 includes all dummies and imposes restrictions instead. LSDV1 is commonly used since it produces correct statistics. LSDV2 provides actual parameter estimates of group intercepts, but reports incorrect  $R^2$  and F statistic. Note that the dummy parameters of three LSDV approaches have different meanings and thus conduct different t-tests.

The within effect model does not use dummy variables but deviations from group means. Thus, this model is useful when there are many groups and/or time periods in the panel data set (no incidental parameter problem at all). The dummy parameter estimates need to be computed afterward. Because of its larger degrees of freedom, the within effect model produces incorrect MSE and standard errors of parameters. As a result, you need to adjust the standard errors to conduct correct t-tests.

Random effect models are estimated by the generalized least squares (GLS) and the feasible generalization least squares (FGLS). When the variance structure is known, GLS is used. If unknown, FGLS estimates theta. Parameter estimates may vary depending on estimation methods.

Fixed effects are tested by the F-test and random effects by the Breusch-Pagan Lagrange multiplier test. The Hausman specification test compares a fixed effect model and a random effect model. If the null hypothesis of uncorrelation is rejected, the fixed effect model is preferred. Poolability is tested by running group by group or time by time regressions.

Among the four statistical packages addressed in this document, I would recommend SAS and Stata. In particular, the SAS PANEL procedure, although experimental now, provides various ways of analyzing panel data and report correct (adjusted) statistics (see Table 6 and 7). Stata is very handy to manipulate panel data, but it does not fit two-way effect models and reports incorrect F-test and  $R^2$ . LIMDEP is able to estimate various panel data models, but it appears to be less stable yet. SPSS is least recommended for panel data models.

## APPENDIX: Data sets

**Data set 1:** Data of the top 50 information technology firms presented in *OECD Information Technology Outlook 2004* (<http://thesius.sourceoecd.org/>).

*firm* = IT company name

*type* = type of IT firm

*rnd* = 2002 R&D investment in current USD millions

*income* = 2000 net income in current USD millions

*d1* = 1 for equipment and software firms and 0 for telecommunication and electronics

```
. tab type d1
```

Type of Firm	d1		Total
	0	1	
Telecom	18	0	18
Electronics	17	0	17
IT Equipment	0	6	6
Comm. Equipment	0	5	5
Service & S/W	0	4	4
Total	35	15	50

```
. sum rnd income
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rnd	39	2023.564	1615.417	0	5490
income	50	2509.78	3104.585	-732	11797

**Data set 2:** Cost data for U.S. airlines (1970-1984) presented in Greene (2003).

URL: <http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm>

*airline* = airline (six airlines)

*year* = year (fifteen years)

*output0* = output in revenue passenger miles, index number

*cost0* = total cost in \$1,000

*fuel0* = fuel price

*load* = load factor, the average capacity utilization of the fleet

```
. tsset
```

```
panel variable:  airline, 1 to 6
time variable:  year, 1 to 15
```

```
. sum output0 cost0 fuel0 load
```

Variable	Obs	Mean	Std. Dev.	Min	Max
output0	90	.5449946	.5335865	.037682	1.93646
cost0	90	1122524	1192075	68978	4748320
fuel0	90	471683	329502.9	103795	1015610
load	90	.5604602	.0527934	.432066	.676287

## 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.
- Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Newbury Park, CA: Sage.
- Freund, Rudolf J., and Ramon C. Littell. 2000. *SAS System for Regression*, 3<sup>rd</sup> 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. 2003. *Econometric Analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall.
- Greene, William H. 2007. *LIMDEP Version 9.0 Econometric Modeling Guide*. Plainview, New York: Econometric Software.
- Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica*, 46(6):1251-1271.
- 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.
- SPSS Inc. 2007. *SPSS 16.0 Command Syntax Reference*. Chicago, IL: SPSS Inc.
- Stata Press. 2007. *Stata Base Reference Manual, Release 10*. College Station, TX: Stata Press.
- Stata Press. 2007. *Stata Longitudinal/Panel Data Reference Manual, Release 10*. College Station, TX: Stata Press.
- Stata Press. 2007. *Stata Time-Series Reference Manual, Release 10*. College Station, TX: Stata Press.
- Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

## ACKNOWLEDGEMENTS

I have to thank Dr. Heejoon Kang of the Kelley School of Business and Dr. David H. Good of the School of Public and Environmental Affairs, Indiana University at Bloomington. I am also grateful to Jeremy Albright and Kevin Wilhite at the UITS Center for Statistical and Mathematical Computing for comments and suggestions. A special thanks to many readers around the world who have eagerly provided constructive feedback and encouraged me to keep improving this document.

## REVISION HISTORY

- 2005.11 First draft
- 2008.04 Corrected some errors and added Stata examples
- 2008.11 Second draft