Chapter 15
PANEL DATA

The analysis of panel data in economics has become increasingly important in recent years as the number of such datasets has grown along with econometric techniques to analyze them. The term "panel data" usually refers to data where the unit of observation varies in two or more dimensions. For example, you might have a sample of the same individuals observed at several points in time, or a set of time series, each for a different firm or country. Such data can be handled rather easily in TSP, although the inherent complexity of the data structure requires you to think a little harder about how to set things up.

This chapter provides some guidance on how to analyze panel data in TSP, and discusses several styles of research using such data. We begin with a few basic rules for setting up your data input, depending on the nature of the problem you are analyzing. Then we discuss the PANEL command, which produces total (pooled), between, within (fixed effects or conditional), and variance components (random effects) estimates for panel data. Finally we discuss how to estimate more complicated models in short panels using the minimum distance estimator in LSQ or GMM. This is a powerful methodology that can be used to estimate linear, nonlinear, and dynamic factor models with panel data, using the method of moments estimator to obtain asymptotically efficient estimates.

In the first part of this chapter we use the example of data on the patenting and R&D spending of a large number of firms, for three years each (see Hall, Griliches, and Hausman 1986; we use three years of data for simplicity, more would be needed to obtain real answers). Although the underlying patent data are application counts, we confine the analysis here to large firms so we can treat the patents as a continuous rather than discrete variable.

We are interested in the relationship between R&D spending (possibly lagged) and the resulting patent applications. The specification of the model that seems to have the most stable properties is to regress the log of patents on contemporaneous and lagged logs of R&D expenditures. However, as will become clear in the example, we expect a fixed difference in the propensity to patent across firms (because of differing technological characteristics of the industry and other reasons), and we expect that this propensity may be correlated with the level of R&D expenditure. This leads us to use many of the panel data techniques described in this chapter.

15.1. The basics of using panel data

15.1.1. Reading in panel data

The first decision to make when dealing with time series-cross section data is how to organize it. Usually, you are willing to assume conditional independence in one direction, but not in the other. For example, you may be willing to assume that observations on firms are conditionally, independently, and identically distributed, but not that there is no serial correlation within a set of observations on a single firm. If this is the case you should order the data so that the slowest varying index is the index of the dimension in which the data are independent. In the example of patents and R&D, use the order

FIRM	TIME PERIOD	VARIABLE
1	74	Patents, R&D for firm 1 in year 74
1	75	Patents, R&D for firm 1 in year 75
1	76	Patents, R&D for firm 1 in year 76
2	74	Patents, R&D for firm 2 in year 74
.	.	.
N	74	Patents, R&D for firm N in year 74
N	75	Patents, R&D for firm N in year 75
N	76	Patents, R&D for firm N in year 76

This order facilitates the construction of estimators that include serial correlation. It also allows you to regroup the datafile easily so that there is one cross section unit per observation with all the variables for all years (by reading in

a different format), for using the robust short panel methods of section 3.1. For example,

FIRM	VARIABLES
1	Patents, R&D for year 74; Patents, R&D for year 75;...
2	Patents, R&D for year 74; Patents, R&D for year 75;...
.	.
N	Patents, R&D for year 74; Patents, R&D for year 75;...

Here are two READ commands that read the identical dataset into TSP in the two formats shown above:

```
SET NOBS = 3*N ;
SMPL 1 NOBS ;
FREQ(PANEL,T=3, ID=@ID) ;
READ (FILE='PATDATA.DAT') @ID PATENTS LRND ;
```

Compare to the second method:

```
SMPL 1 N ;
READ (FILE='PATDATA.DAT') @ID PAT74 LRND74 @ID PAT75 LRND75
   @ID PAT76 LRND76 ;                    ? Only the value of @ID for 1976 will be stored.
```

There are situations where you want the data one observation per firm-year (the PANEL command and AR1 with the TSCS option) and situations where you want the data one observation per firm (when using complex lag structures or GMM methods); thus the data should be set up with this in mind. We refer to the first format as the *pooled* format and the second format as the *panel* format. The key factor that determines your choice between them is the statistical assumption of conditional independence: in general, when you assume that observations are independent (conditional on your model) across both dimensions, you will want the data in pooled format, and when you assume independence only in one direction (with the possible exception of simple first order serial correlation), the panel format.

15.1.2. Unbalanced panels

An unbalanced panel is one where there are a different number of observations for each cross section unit (or vice versa). These observations may be contiguous, or there may be holes in the data. That is, for the example dataset, we may have four years of data for one firm (1973 to 1976), three years for another (1973, 1975, 1976) and two years for a third (1975 and 1976). It is essential to use the FREQ(PANEL,ID=*ID_variable*); command with an unbalanced panel, to identify an *ID_variable* that indicates when one firm stops and the next one starts. You can also use the FREQ(PANEL) command with balanced panels, or to identify a time variable, the time series frequency of the data, etc.

Many, but not all, of the estimators described in this chapter can be used with unbalanced panels. For example, the PANEL procedure, which assumes independence across time series and cross section units will work just the same whether the data are balanced or unbalanced. All command which use lags and leads recognize the *ID variable*, so that lags and leads will refer only within a single individual. One implication of this is that you can use a command like

```
SELECT MISS(IDV(-1));                    ? when IDV is the ID variable
```

to choose the first observations for all individuals. The AR1 command also recognizes the *ID variable*, so it will apply the special transformation to the first observation of each individual and avoid using any lags that would refer from one individual back to the previous individual.

In the panel format, unbalanced panels can be "balanced" by including missing data codes for the missing observations. Some of the methods described in section 15.3.1 may not work very well with unbalanced panels. This is in the nature of the data and the current state of econometric methodology; it is not necessarily a limitation of the program.

15.2. Random and Fixed Effects models -- the PANEL procedure

PANEL obtains estimates of linear regression models for panel data (several observations or time periods for each individual). The data may be unbalanced (different number of observations per individual). PANEL can also compute means by group and perform F tests between groups. To define the models estimated, assume we have observations on i=1,...,N individuals for each of t=1,...,T years. The dependent variable is denoted by y_{it} and the independent variables by X_{it}. The basic pooled or TOTAL regression model is

$$y_{it} = X_{it}\beta + \alpha + u_{it}$$

where α is the overall intercept and u_{it} is i.i.d. This model assumes a single set of slope coefficients for all the observations.

The fixed effect or WITHIN model assumes that there are common slopes, but that each cross section unit has its own intercept, which may or may not be correlated with the Xs:

$$y_{it} = X_{it}\beta + \alpha_i + u_{it}$$

The BYID model assumes that both the slopes and the intercepts vary across cross section units:

$$y_{it} = X_{it}\beta_i + \alpha_i + u_{it}$$

The BETWEEN model specifies the same relationship between the individual means:

$$y_{i.} = X_{i.}\beta + \alpha + u_{i.}$$

where

$$y_{i.} = \Sigma\, y_{it} /T$$

The random effects or VARCOMP model resembles the WITHIN model, but it assumes that the intercepts are drawn from a common distribution with mean α and variance σ_α^2. Unlike the WITHIN model, the estimates for this $\{\alpha_i\}$ model will not be consistent if the individual intercepts are correlated with the independent variables. Because of this, it is important to test for correlation. PANEL reports the Hausman test statistic for the difference between the fixed effects and random effects estimates, along with its p-value.

The VARCOMP estimator is computed by estimating the relative importance of between and within variation of the disturbance $\alpha_i + u_{it}$ and using this estimated ratio to combine the within and between estimators optimally. Under the null of uncorrelated intercepts, the VARCOMP estimator is asymptotically efficient, since it is a generalized least squares estimator. There are additional options for VARCOMP to control the actual variance components. Small or large sample formulas may be used, or you can supply the values directly. If negative variances are computed using the small sample formula, the program switches over to the large sample formulas, which always result in positive values.

All or some of these models can be estimated by a single PANEL statement. The basic PANEL statement is like the OLSQ statement: first list the dependent variable and then the independent variables. C is optional; an intercept term is central to these models and will be added if not present. Here is an example for the sample dataset:

```
PANEL LPAT C LRND ;
```

This command will produce estimates of the TOTAL, WITHIN, BETWEEN, and VARCOMP models, together with the value of an F-statistic for the hypothesis that all the intercepts are equal.

The observations over which the models are computed are determined by the current sample. Lags, leads, and missing values are handled properly.

Your data must be set up with all the time periods for each individual together (the pooled format). You must also specify when the data ends for one individual, and begins for the next. The best method is to provide an *ID variable*

series in the FREQ(PANEL) command that takes on different values for each individual, as we did in the sample dataset. If your data are balanced (the same number of time periods for every individual), the T= option can be used. Other options are also available (see the *Reference Manual*). If the data are not in this order, the SORT command can reorder them (you can also use SORT to reorder the data so that you can do variance components in the other (time) dimension). See Section 6.4 for an example.

At present, PANEL does not automatically generate dummies for the time periods (although they can be included) or do variance components in the other dimension. To generate a set of time dummies for the sample dataset, use the TREND and DUMMY commands:

```
TREND(PER=3,START=74) YEAR ;          ? Makes a series = 74,75,76,74,75,76,... (for balanced data)
LIST YRDUM YEAR74-YEAR76 ;
DUMMY YEAR YRDUM ;
```

This creates three variables YEAR74, YEAR75, and YEAR76 with the following values:

OBS	YEAR74	YEAR75	YEAR76
1,1	1	0	0
1,2	0	1	0
1,3	0	0	1
2,1	1	0	0
2,2	0	1	0
	and so forth		

If you had loaded a variable YEAR (which is essential in the unbalanced case), you could have just used the DUMMY command directly, without using TREND.

This next example estimates all models including the individual firm regressions, and prints individual means:

```
PANEL(MEAN,BYID) LPAT C LRND YEAR75 YEAR76 ;
```

The output for this command will include F-statistics for the hypothesis that the slope coefficients are equal and for the joint hypothesis that both the slopes and intercepts are equal.

The following estimates VARCOMP only, using large sample formulas (note the use of year dummies with the intercept):

```
PANEL(NOTOT,NOBET,NOWITH,NOVSMALL) LPAT C LRND YEAR75 YEAR76;
```

15.3. Robust estimation with panel data

This section discusses how to obtain asymptotically efficient estimates of panel data models without imposing conditional homoskedasticity or independence over time on the disturbances of the model. The methods and estimators described here are due largely to Chamberlain (1982) and MaCurdy (1981a and 1982), although many others have contributed to their development. They are closely related to the GMM estimator proposed by Hansen and Singleton (1982) and the GMM command in TSP will compute some of them.

The estimators described here are minimum distance estimators that use an asymptotically optimal weighting matrix. In particular, they use the sample covariance of the distance measures (such as orthogonality conditions or residuals) as the weighting matrix. The key to understanding the computation of the estimators described here is to recognize that the SUR procedure (which computes multivariate regressions without imposing a diagonal covariance structure) can also compute a set of estimates of second moments or functions of second moments, *along with a robust estimate of their variance-covariance matrix*. This estimate is heteroskedastic-consistent and does not impose independence across the disturbances in each equation, where equation here refers to the moment equation.

This enables you to construct the optimal weighting matrix for many of these estimators easily, without special

programming. Using this matrix, which we call OMEGA, we can construct a minimum distance estimator for the second moments as functions of the parameters of interest using the SUR procedure again, but this time with only one observation (since the second moments and the estimated OMEGA are sufficient statistics for the problem).

This methodology can be applied to two different panel data estimation problems: the problem of describing the relationship between a set of endogenous variables (Y) and a set of exogenous variables (X), where the reduced form Π matrix is a sufficient statistic for the problem (Chamberlain's problem), and the problem of describing the relationship between a set of endogenous variables (Y) and a set of unobservable variables ("factors"), where the second moments of Y are a sufficient statistic. Obviously, the two types of models could be combined, but the presentation is simpler if they are treated separately.

15.3.1. The PI matrix method

Chamberlain (1982) showed that one way to estimate a whole range of panel data models was to summarize the data by regressing all the endogenous variables on all of the exogenous variables, obtaining an estimate of the reduced form matrix Π; and then to test various restrictions on this matrix implied by the models of interest (actually Chamberlain focused on the conditional expectation interpretation of regression so that the estimator in question was for the expectation of the Ys conditional on the Xs). If you use the minimum distance estimator

$$\text{argmin } (\pi - f(\delta))\, \Omega^{-1}\, (\pi - f(\delta))$$

together with an appropriate estimate of Ω to estimate the restricted parameter set δ, then the resulting estimates of δ are asymptotically efficient. The optimal estimate of Ω in this case is given by the sample covariance of w_i, where

$$w_i = (y_i - \Pi x_i) \otimes S_x^{-1}\, x_i$$

and S_x is the sample variance of the Xs. Note that this formula does not imply independence *within* each observational unit, nor does it impose homoskedasticity.

Using SUR, it is easy to estimate Π and its associated covariance Ω in TSP. For the sample dataset:

```
DOT 74-76 ;
  FRML PIEQ. LPAT. = PI.74*LRND74 + PI.75*LRND75 + PI.76*LRND76 ;
  PARAM PI.74-PI.76 ;
  MSD (NOPRINT) LPAT. LRND. ;  ? Removing all the year means.
  UNMAKE @MEAN PMEAN RMEAN ;
  LPAT. = LPAT.-PMEAN ; LRND. = LRND.-RMEAN ;
ENDDOT ;
SUR (HCOV=R) PIEQS ;              ? Note the robust option.
COPY @COEF PI ;                   ? Save the computed PI matrix and its
COPY @VCOV OMEGA ;                ? covariance estimate.
```

Removing the means of the data before forming the estimated Π simplifies things, because it implies that you do not have to carry around the X variable corresponding to the intercept. This two part strategy for estimation does not affect the asymptotics (MaCurdy 1982).

Now suppose the class of restricted models of interest have the following form:

$$y_{it} = \beta_1 x_{it} + \beta_2 x_{i,t-1} + \ldots + \gamma_t \alpha_i + u_{it}$$

where α_i is the firm effect, which may be correlated with the x's:

$$\alpha_i = \lambda_1 x_{i1} + \ldots + \lambda_T x_{iT}$$

For the sample data, with three ys and three xs, the Π matrix has the following form:

$$\Pi = \begin{matrix} \beta_1 + \gamma_{74}\lambda_{74} & \gamma_{74}\lambda_{75} & \gamma_{74}\lambda_{76} \\ \beta_2 + \gamma_{75}\lambda_{74} & \beta_1 + \gamma_{75}\lambda_{75} & \gamma_{75}\lambda_{76} \\ \beta_3 + \gamma_{76}\lambda_{74} & \beta_2 + \gamma_{76}\lambda_{75} & \beta_1 + \gamma_{76}\lambda_{76} \end{matrix}$$

There are nine elements in Π and nine coefficients to be estimated, but there is one normalization restriction ($\gamma_{74}=1$), so there is one over-identifying restriction. If there are no correlated effects, the γs and λs will be zero and there will be six over-identifying restrictions.

With the estimated Π and Ω matrices obtained above, you can test for the two levels of restrictions implied by this model:

1) a stable lag structure and correlated firm effect.
2) a stable lag structure and uncorrelated firm effects.

Here is how to do it using the minimum distance procedure (LSQ or SUR):

```
?   Define the lists of PI coefficients and equations.
?
LIST PILIST PI7474-PI7476 PI7574-PI7576 PI7674-PI7676 ;
LIST PIEQLIST PIEQ7474-PIEQ7476 PIEQ7574-PIEQ7576 PIEQ7674-PIEQ7676 ;
LENGTH PILIST NPI ;                    ? Find out how many elements in PI matrix.
UNMAKE PI PILIST ;                     ? Unmake the estimated PI matrix into its individual elements.
CONST PILIST ;                         ? Treat the estimated PI coefficients as data below.
SUPRES COVU W COVT REGOUT;             ? Reduce the output for Minimum Distance estimation
?
?   Define the equations that express PI as a function of the underlying delta parameters (beta,  lambda, and gamma)
?  which are to be estimated.
?
FRML PIEQ7474 PI7474 = BETA1 + LAM74*GAM74 ;
FRML PIEQ7574 PI7574 = BETA2 + LAM74*GAM75 ;
FRML PIEQ7674 PI7674 = BETA3 + LAM74*GAM76 ;
FRML PIEQ7475 PI7475 =       LAM75*GAM74 ;
FRML PIEQ7575 PI7575 = BETA1 + LAM75*GAM75 ;
FRML PIEQ7675 PI7675 = BETA2 + LAM75*GAM76 ;
FRML PIEQ7476 PI7476 =       LAM76*GAM74 ;
FRML PIEQ7576 PI7576 =       LAM76*GAM75 ;
FRML PIEQ7676 PI7676 = BETA1 + LAM76*GAM76 ;

CONST LAM74-LAM76 GAM74-GAM76 ;    ? Starting values for model with
PARAM BETA1 1 BETA2 0 BETA3 0 ;    ? uncorrelated Xs (only betas to be estimated).
CONST LAM74 1 ;                    ? Free normalization.

SMPL 1,1 ;                         ? In effect we now have one observation on each element of PI.
SUR(WNAME=OMEGA) PIEQLIST ;
LENGTH @RNMS NOPAR ;
SET DF = NPI-NOPAR ;               ? Degrees of freedom for constrained model.
CDF(CHISQ,DF=DF) @TR ;              ? Test constraints.

PARAM LAM74-LAM76 GAM75 GAM76 ;     ? Starting values for model with correlated Xs.
SUR(WNAME=OMEGA) PIEQLIST ;         ? Estimation.
LENGTH @RNMS NOPAR ;
SET DF = NPI-NOPAR ;
CDF(CHISQ,DF=DF) @TR ;              ? Test the single constraint remaining.
```

The "TRACE OF MATRIX" criterion printed out by SUR after convergence is precisely the Chi-squared statistic for

the over-identifying constraints (with degrees of freedom equal to the number of elements of Π less the number of parameters being estimated). Note that changing the sample size to one is essential if you want standard error estimates to have the correct size (assuming OMEGA has been computed as shown).

With a larger number of Ys, Xs, or observations in the time dimension, the number of models that might be nested in this way becomes very large and the TSP program correspondingly larger. Using DOT loops and other shortcuts can make programming easier and streamline your program so that it is easier to read and debug. See the examples earlier in this chapter for ideas.

15.3.2. Dynamic factor models with panel data

An example of how to estimate a fairly complex dynamic factor model using SUR is available in the examples on the TSP web site. The example is drawn from Hall and Hayashi (1989).

15.3.3. GMM Estimation of panel data models.

A series of recent papers (Keane and Runkle 1992, Arellano and Bond 1991, Ahn and Schmidt 1992) have advocated the use of the GMM methodology for the estimation of dynamic panel models or panel data models with predetermined rather than exogenous right-hand side variables. These estimators are straightforward to implement in TSP using the GMM estimation command together with a MASK option that chooses the instruments you wish to use for each equation.

As an example, consider the one-variable model of y on x, with 3 years of data for each y, but 6 years, including 3 lags for each x. With this much data, it is possible to test not only for strong exogeneity of the x's, but also for weak exogeneity of lag order 0, 1, or 2, either unconditionally or conditional on the presence of individual effects. That is, you can test for whether there is zero correlation between the first-differenced disturbances and future x's and also for whether there is zero correlation between the first-differenced disturbances and current x's, or x's lagged once. As the number of lags of x variables that are not assumed to be exogenous increases, the number of moment restrictions imposed decreases. A simple TSP run that performs these tests conditional on individual effects is shown below:

```
?   Equations of the model.
FRML UEQ86 Y86 - BETA1*X86-BETA2*X85-BETA3*X84 ;
FRML UEQ87 Y87 - BETA1*X87-BETA2*X86-BETA3*X85 ;
FRML UEQ88 Y88 - BETA1*X88-BETA2*X87-BETA3*X86 ;
?   First-differenced versions.
FRML DUEQ87 UEQ87-UEQ86 ;
FRML DUEQ88 UEQ88-UEQ87 ;
DOT 87 88 ;
 EQSUB DUEQ. UEQ86-UEQ88 ;
ENDDOT ;
LIST XLIST X83-X88 ;                    ? List of potential instruments.

READ (NROW=6,NCOL=2) M1 ;               ? Mask for lag 1 and greater instruments.
1 1  1 1  1 1  0 1  0 0  0 0
;
GMM (HETERO,INST=XLIST,MASK=M1) DUEQ87 DUEQ88 ;
COPY @GMMOVID TRACE1 ;                  ? Chi-squared statistic for estimated model.
COPY @NOVID DF1 ;                       ? actual number of moment restrictions imposed.
READ (NROW=6,NCOL=2) M0 ;               ? Mask for lag 0 instruments (weak exog).
1 1  1 1  1 1  1 1  0 1  0 0
;
GMM (HETERO,INST=XLIST,MASK=M0) DUEQ87 DUEQ88 ;
COPY @GMMOVID TRACE0 ;
COPY @NOVID DF0 ;
```

```
? Strong exogeneity: no MASK means all instruments are used for all equations.
GMM (HETERO,INST=XLIST) DUEQ87 DUEQ88 ;
COPY @GMMOVID TRACES ;
COPY @NOVID DFS;

? Compute Test Statistics.
SET TEST0 = TRACE0-TRACE1 ; SET DFR = DF0-DF1 ;
CDF(DF=DFR,CHISQ) TEST0 ;                    ? Test lag 0 instruments, maintaining lag 1.
SET TESTS = TRACES-TRACE0 ; SET DFR = DFS-DF0 ;
CDF(DF=DFR,CHISQ) TESTS ;                    ? Test strong exogeneity, maintaining weak.
```

Note that the first model estimated in this example is the least constrained model; all the others will be tested relative to this one.

Modifying this example to perform the test without allowing for individual effects is straightforward: simply use the level equations UEQ86-UEQ8 and modify the M1 and M0 matrices accordingly. For example, if the order of the equations is UEQ86 UEQ87 UEQ88, the mask M1 should be the following matrix:

```
1 1 1
1 1 1
1 1 1
0 1 1
0 0 1
0 0 0
```

The TSP web page *www.tspintl.com* contains more examples of estimating a panel data model using both the GMM and PI matrix techniques. There are also examples of performing LM tests for serial correlation as in Arellano and Bond (1991), and computing the one-step covariance matrix estimate recommended by Blundell and Bond (1995).