DYNAMIC PANEL DATA ESTIMATION USING DPD98 FOR GAUSS: A GUIDE FOR USERS*

Manuel Arellano† and Stephen Bond‡

December 1998

INTRODUCTION

DPD98 is a program written in the Gauss matrix programming language to compute estimates for dynamic models from panel data. A number of estimators are available, including the generalised method of moments (GMM) techniques developed in Arellano and Bond (1991) and Arellano and Bover (1995), as well as more familiar OLS, within-groups and instrumental variables procedures. Standard errors and test statistics that are robust to the presence of heteroskedasticity are provided. Tests for serial correlation and instrument validity are automatically computed. Further tests of linear restrictions are available as options, and the parameter vectors and covariance matrices can be saved as Gauss matrices, so that tests of non-linear restrictions and Hausman specification tests can be implemented by users familiar with Gauss. Lagged and differenced series are easily constructed, with many other data transformations available. A particularly attractive feature of DPD98 is that it allows estimates to be computed from panels that are unbalanced in the sense of having a variable number of time-series observations per individual unit. In many contexts this allows a much larger sample to be exploited than would be the case if a balanced panel were required.

DPD98 was developed to use with data on a panel of companies, but is applicable to many other situations in which the number of time-series observations is small and the number of cross-section observations is large. We concentrate on estimators that do not require regressors to be strictly exogenous, and which require only the cross-section dimension of the data set to become large for consistency.

The main new features of DPD98 compared to earlier versions of DPD (see Arellano and Bond, 1988) are:

- system GMM estimators, combining moment conditions for equations in first differences with moment conditions for equations in levels, are easily computed
- batch operation is supported
- lag operator can be used to construct transformed series
- except where the model has been estimated in levels, tests for serial correlation are based on estimates of the residuals in first differences

*DPD was originally developed to use with the IFS company database. We have benefited from the input of many colleagues, but would particularly like to thank Richard Blundell for his encouragement and many helpful suggestions.

†CEMFI, Casado del Alisal 5, 28014 Madrid, Spain

‡Institute for Fiscal Studies, 7 Ridgmount St, London WC1E 7AE, UK and Nuffield College, Oxford OX1 1NF, UK

- parameter vectors and covariance matrices can be saved as Gauss matrices

Section 1 of this guide describes how DPD98 can be installed and how data should be organised for use with DPD98. Section 2 contains an account of the econometric methods employed by DPD98. Section 3 provides detailed instructions on how to use DPD98, and Section 4 contains an example.

1. INSTALLATION

DPD98 is contained in 3 files: DPD98.RUN, DPD98.FNS and DPD98.PRG. These were written using Version 3.2.13 of the Gauss language, but will run with all versions higher than Gauss386. We have used DPD98 successfully with Gauss for Windows, but this guide describes the use of DPD98 with the standard (DOS) versions of Gauss. These can of course be used under Windows95 and WindowsNT.

To install DPD98, simply copy the 3 DPD98 files onto any subdirectory (folder) of your hard disk. Provided Gauss is correctly installed on your computer (or network), there is no need for these files to be kept in the \Gauss subdirectory (folder). After entering Gauss you can change to any subdirectory (folder) where your DPD98 files are located. At the Gauss $\gg$ prompt, simply type:

dos cd \dname$\hookleftarrow$

where dname is the name of the subdirectory (folder) where your DPD98 files are located, and $\hookleftarrow$ represents the return or enter key.

1.1. DATA

The main requirement for running DPD98 is a suitably ordered Gauss data set. This can be created from a sorted ASCII data file, using the atog386 utility (see the Gauss manual for more details), or directly from other packages using a conversion utility like Stat/Transfer. DPD98 has strict requirements concerning how your data is ordered. Failure to observe these requirements normally causes the program to abort, and this is the most common source of problems experienced by users who are new to DPD98.

Each row of the Gauss data set should contain observations on a set of variables for some cross-sectional unit and some time period. Each column should contain observations on the same variable in every row. One column must contain the year to which the observation refers, and this must be in the format 19xx. Annual data is the only frequency which DPD98 supports, though other data frequencies can be used provided the observations are suitably labelled. For example, if your panel contains 4 monthly observations for April, May, June and July, or

four observations on decade-averages for the 1950s, 1960s, 1970s and 1980s, these panels can be used with DPD98 provided the observations are given labels such as 1991, 1992, 1993 and 1994 in the data set (and coefficients on 'year' dummies are interpreted accordingly). If any computers are still running after Jan 1 2000, the next release of DPD will allow years after 1999 to be used without relabelling.

All the time-series observations for each individual unit must be consecutive, and these observations must occur sequentially in the data set. Thus if you have 4 observations for 1991, 1992, 1993 and 1994, the first 4 rows of the data set should contain these observations for the first individual, the next 4 rows should contain these observations for the second individual, and so on.

Where the panel is unbalanced, in the sense of having more time-series observations for some individuals than for others, the individual units on which there is a common number of time-series observations should be grouped together in the data set. For example, if you have 4 observations for 100 individuals and 5 observations for 50 individuals, either the first 400 rows or the last 400 rows of the data set should contain the individuals with 4 time-series observations. Notice that it does not matter here whether an individual has the first observation or the last observation missing. Thus if the 5 observations are for 1991-95, the 4 observations may relate either to the period 1991-94 or to the period 1992-95.[1] It is sometimes useful, although not essential, for these groups of cross-sectional units to appear in ascending (or descending) order of the number of time-series observations per unit.

In addition to this main Gauss data set, DPD98 requires a secondary or auxiliary Gauss data set which describes the structure of the main data. This auxiliary data set must contain two columns: elements of the first column contain the number of time-series observations per individual unit in the appropriate section of the main data file; and elements of the second column contain the number of individual units which have this number of time-series observations. This structured description begins from the top of the main data set and moves down it. Continuing with our previous example, if we order the main data set so that the 100 individuals with 4 observations appear first, followed by the 50 individuals with 5 observations, then the auxiliary data set will take the form:

$$\begin{array}{cc} 4 & 100 \\ 5 & 50 \end{array}$$

Again this auxiliary Gauss data set can be created from an ASCII text file using the atog386 utility. It is used to facilitate the reading of unbalanced data sets.

[1] If the data available cover only the 5 years 1991-95, then these are the only two possibilities in this example, given that the annual observations must be consecutive. DPD98 also allows the use of rotating panels, where the data covers more years than are observed for any particular individual.

2. ECONOMETRIC METHODS

The general model that can be estimated with DPD98 is a single equation with individual effects of the form:

$$y_{it} = \sum_{k=1}^{p} \alpha_k y_{i(t-k)} + \beta'(L)x_{it} + \lambda_t + \eta_i + v_{it}$$

$$(t = q+1, ..., T_i;\ i = 1, ..., N)$$

where η_i and λ_t are respectively individual and time specific effects, x_{it} is a vector of explanatory variables, $\beta(L)$ is a vector of associated polynomials in the lag operator and q is the maximum lag length in the model. The number of time periods available on the ith individual, T_i, is small and the number of individuals, N, is large. Identification of the model requires restrictions on the serial correlation properties of the error term v_{it} and/or on the properties of the explanatory variables x_{it}. It is assumed that if the error term was originally autoregressive, the model has been transformed so that the coefficients α's and β's satisfy some set of common factor restrictions. Thus only serially uncorrelated or moving average errors are explicitly allowed. The v_{it} are assumed to be independently distributed across individuals with zero mean, but arbitrary forms of heteroskedasticity across units and time are possible. The x_{it} may or may not be correlated with the individual effects η_i, and for each of theses cases they may be strictly exogenous, predetermined or endogenous variables with respect to v_{it}. A case of particular interest is where the levels x_{it} are correlated with η_i but where Δx_{it} (and possibly Δy_{it}) are uncorrelated with η_i; this allows the use of (suitably lagged) Δx_{is} (and possibly Δy_{is}) as instruments for equations in levels.

The $(T_i - q)$ equations for individual i can be written conveniently in the form:

$$y_i = W_i\delta + \iota_i\eta_i + v_i$$

where δ is a parameter vector including the α_k's, the β's and the λ's, and W_i is a data matrix containing the time series of the lagged dependent variables, the x's and the time dummies. Lastly, ι_i is a $(T_i - q) \times 1$ vector of ones. DPD98 can be used to compute various linear GMM estimators of δ with the general form:

$$\widehat{\delta} = \left[\left(\sum_i W_i^{*\prime} Z_i\right) A_N \left(\sum_i Z_i' W_i^*\right)\right]^{-1} \left(\sum_i W_i^{*\prime} Z_i\right) A_N \left(\sum_i Z_i' y_i^*\right)$$

where

$$A_N = \left(\frac{1}{N}\sum_i Z_i' H_i Z_i\right)^{-1}$$

and W_i^* and y_i^* denote some transformation of W_i and y_i (e.g. levels, first differences, orthogonal deviations, combinations of first differences (or orthogonal deviations) and levels, deviations from individual means). Z_i is a matrix of instrumental variables which may or may not be entirely internal, and H_i is a possibly individual specific weighting matrix.

If the number of columns of Z_i equals that of W_i^*, A_N becomes irrelevant and $\widehat{\delta}$ reduces to

$$\delta = \left(\sum_i Z_i' W_i^*\right)^{-1} \left(\sum_i Z_i' y_i^*\right)$$

In particular, if $Z_i = W_i^*$ and the transformed W_i and y_i are deviations from individual means or orthogonal deviations[2], then $\widehat{\delta}$ is the within groups estimator. As another example, if the transformation denotes first differences, $Z_i = I_{T_i} \diagdown x_i'$ and $H_i = \widehat{v}_i^* \widehat{v}_i^{*\prime}$, where the $\widehat{v}_i^*$ are some consistent estimates of the first differenced residuals, then $\widehat{\delta}$ is the generalised three stage least squares estimator of Chamberlain (1984). These two estimators require the x_{it} to be strictly exogenous with respect to v_{it} for consistency. In addition, the within groups estimator can only be consistent as $N \to \infty$ for fixed T if W_i^* does not contain lagged dependent variables and all the explanatory variables are strictly exogenous.

When estimating dynamic models, we shall therefore typically be concerned with transformations that allow the use of lagged endogenous (and predetermined) variables as instruments in the transformed equations. Efficient GMM estimators will typically exploit a different number of instruments in each time period. Estimators of this type are discussed in Arellano (1988), Arellano and Bond (1991), Arellano and Bover (1995) and Blundell and Bond (1998). DPD98 can be used to compute a range of linear GMM estimators of this type.

Where there are no instruments available that are uncorrelated with the individual effects η_i, the transformation must eliminate this component of the error term. The first difference and orthogonal deviations transformations are two examples of transformations that eliminate η_i from the transformed error term, without at the same time introducing all lagged values of the distrubances v_{it} into the transformed error term.[3] Hence these transformations allow the use of

[2] Orthogonal deviations, as proposed by Arellano (1988) and Arellano and Bover (1995), express each observation as the deviation from the average of *future* observations in the sample for the same individual, and weight each deviation to standardise the variance (i.e.

$$x_{it}^* = \left(x_{it} - \frac{x_{i(t+1)} + ... + x_{iT}}{T-t}\right)\left(\frac{T-t}{T-t+1}\right)^{1/2} \text{ for } t = 1, ..., T-1)$$

If the original errors are serially uncorrelated and homoskedastic, the transformed errors will also be serially uncorrelated and homoskedastic.

[3] There are many other transformations which share these properties. See Arellano and Bover

suitably lagged endogenous (and predetermined) variables as instruments. For example, if the panel is balanced, $p = 1$, there are no explanatory variables nor time effects, the v_{it} are serially uncorrelated, and the initial conditions y_{i1} are uncorrelated with v_{it} for $t = 2, ..., T$, then using first differences we have:

Equations	Instruments available
$\Delta y_{i3} = \alpha \Delta y_{i2} + \Delta v_{i3}$	y_{i1}
$\Delta y_{i4} = \alpha \Delta y_{i3} + \Delta v_{i4}$	y_{i1}, y_{i2}
.	.
.	.
$\Delta y_{iT} = \alpha \Delta y_{i(T-1)} + \Delta v_{iT}$	$y_{i1}, y_{i2}, ..., y_{i(T-2)}$

In this case $y_i^* = (\Delta y_{i3}, ..., \Delta y_{iT})', W_i^* = (\Delta y_{i2}, ..., \Delta y_{i(T-1)})'$ and

$$Z_i = Z_i^D = \begin{pmatrix} y_{i1} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & y_{i1} & y_{i2} & \cdots & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & y_{i1} & y_{i2} & \cdots & y_{i(T-2)} \end{pmatrix}$$

Notice that precisely the same instrument set would be used to estimate the model in orthogonal deviations. Where the panel is unbalanced, for individuals with incomplete data the rows of Z_i corresponding to the missing equations are deleted, and missing values in the remaining rows are replaced by zeros.

In DPD98 we call one-step estimates those which use some known matrix as the choice for H_i. For a first-difference procedure, the one-step estimator uses

$$H_i = H_i^D = \begin{pmatrix} 2 & -1 & \cdots & & 0 \\ -1 & 2 & \cdots & & 0 \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & -1 \\ 0 & 0 & \cdots & -1 & 2 \end{pmatrix}$$

while for a levels or orthogonal deviations procedure the one-step estimator sets H_i to an identity matrix. If the v_{it} are heteroskedastic, a two-step estimator which uses

$$H_i = \widehat{v}_i^* \widehat{v}_i^{*\prime}$$

(1995) for further discussion.

where $\widehat{v}_i^*$ are one-step residuals, is more efficient (cf. White (1982)). DPD98 produces both one-step and two-step GMM estimators, with asymptotic variance matrices that are heteroskedasticity-consistent in both cases. Users should note that, particularly when the v_{it} are heteroskedastic, simulations suggest that the asymptotic standard errors for the two-step estimators can be a poor guide for hypothesis testing in typical sample sizes. In these cases, inference based on asymptotic standard errors for the one-step estimators seems to be more reliable.[4]

In models with explanatory variables, Z_i may consist of sub-matrices with the block diagonal form illustrated above (exploiting all or part of the moment restrictions available), concatenated to straightforward one-column instruments. A judicious choice of the Z_i matrix should strike a compromise between prior knowledge (from economic theory and previous empirical work), the characteristics of the sample and computer limitations (see Arellano and Bond (1991) for an extended discussion and illustration). For example, if a predetermined regressor x_{it} correlated with the individual effect, is added to the model discussed above, i.e.

$$\begin{aligned} E(x_{it}v_{is}) &= 0 \text{ for } s \geq t \\ &\neq 0 \text{ otherwise} \\ E(x_{it}\eta_i) &\neq 0 \end{aligned}$$

then the corresponding optimal Z_i matrix is given by

$$Z_i = \begin{pmatrix} y_{i1} & x_{i1} & x_{i2} & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & y_{i1} & y_{i2} & x_{i1} & x_{i2} & x_{i3} & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & y_{i1} & \cdots & y_{i(T-2)} & x_{i1} & \cdots & x_{i(T-1)} \end{pmatrix}$$

Where the number of columns in Z_i is very large, computational considerations may require those columns containing the least informative instruments to be deleted. Even when computing speed is not an issue, it may be advisable not to use the whole history of the series as instruments in the later cross-sections. For a given cross-sectional sample size (N), the use of too many instruments may result in (small sample) overfitting biases. When overfitting results from the number of time periods (T) becoming large relative to the number of individuals (N), and there are no endogenous regressors present, these GMM estimators are biased towards within groups, which is not a serious concern since the within groups estimator is itself consistent for models with predetermined variables as T

[4]See Arellano and Bond (1991) and Blundell and Bond (1998) for further discussion.

becomes large.[5] However, in models with endogenous regressors, using too many instruments in the later cross-sections could result in seriously biased estimates. This possibility can be investigated in practice by comparing the GMM and within groups estimates.

The assumption of no serial correlation in the v_{it} is essential for the consistency of estimators such as those considered in the previous examples, which instrument the lagged dependent variable with further lags of the same variable. Thus DPD98 reports tests for the absence of first-order and second-order serial correlation in the first-differenced residuals. If the disturbances v_{it} are not serially correlated, there should be evidence of significant negative first order serial correlation in differenced residuals (i.e. $\widehat{v}_{it} - \widehat{v}_{i,t-1}$), and no evidence of second order serial correlation in the differenced residuals. These tests are based on the standardised average residual autocovariances which are asymptotically $N(0,1)$ variables under the null of no autocorrelation. The tests reported are based on estimates of the residuals in first differences, even when the estimator is obtained using orthogonal deviations.[6] More generally, Sargan tests of overidentifying restrictions are also reported. That is, if A_N has been chosen optimally for any given Z_i, the statistic

$$S = \left(\sum_i \widehat{v}_i^{*\prime} Z_i\right) A_N \left(\sum_i Z_i' \widehat{v}_i^*\right)$$

is asymptotically distributed as a chi-square with as many degrees of freedom as overidentifying restrictions, under the null hypothesis of the validity of the instruments. Note that only the Sargan test based on the two-step GMM estimator is heteroskedasticity-consistent. Again, Arellano and Bond (1991) provides a complete discussion of these procedures.

Where there are instruments available that are uncorrelated with the individual effects η_i, these variables can be used as instruments for the equations in levels. Typically this will imply a set of moment conditions relating to the equations in first differences (or orthogonal deviations) and a set of moment conditions relating to the equations in levels, which need to be combined to obtain the efficient GMM estimator.[7] For example, if the simple AR(1) model considered earlier is mean-stationary, then the first differences Δy_{it} will be uncorrelated with η_i, and this implies that $\Delta y_{i(t-1)}$ can be used as instruments in the levels equations.[8] In

[5] See Alvarez and Arellano (1998).

[6] Although the validity of orthogonality conditions is not affected, the transformation to orthogonal deviations can induce serial correlation in the transformed error term if the v_{it} are serially uncorrelated but heteroskedastic.

[7] In special cases it may be efficient to use only the equations in levels; for example, in a model with no lagged dependent variables and all regressors strictly exogenous and uncorrelated with individual effects.

[8] See Arellano and Bover (1995) and Blundell and Bond (1998) for further discussion.

addition to the instruments available for the first-differenced equations that were described earlier, we then have:

Equations	Instruments available
$y_{i3} = \alpha y_{i2} + \eta_i + v_{i3}$	Δy_{i2}
$y_{i4} = \alpha y_{i3} + \eta_i + v_{i4}$	Δy_{i3}
.	.
.	.
$y_{iT} = \alpha y_{i(T-1)} + \eta_i + v_{iT}$	$\Delta y_{i(T-1)}$

Notice that no instruments are available in this case for the first levels equation (i.e. $y_{i2} = \alpha y_{i1} + \eta_i + v_{i2}$), and that using further lags of Δy_{is} as instruments here would be redundant, given the instruments that are being used for the equations in first differences. In a balanced panel, we could use only the last levels equation (i.e. $y_{iT} = \alpha y_{i(T-1)} + \eta_i + v_{iT}$), where $(\Delta y_{i2}, \Delta y_{i3}, ..., \Delta y_{i(T-1)})$ would all be valid instruments; however this approach does not extend conveniently to unbalanced panels.

In this case, we use $y_i^* = (\Delta y_{i3}, ..., \Delta y_{iT}, y_{i3}, ..., y_{iT})', W_i^* = (\Delta y_{i2}, ..., \Delta y_{i(T-1)}, y_{i2}, ..., y_{i(T-1)})'$ and

$$Z_i = \begin{pmatrix} \mathbf{Z}_i^D & 0 & \cdots & 0 \\ 0 & \Delta y_{i2} & \cdots & 0 \\ . & . & & . \\ 0 & 0 & \cdots & \Delta y_{i(T-1)} \end{pmatrix}$$

where $\mathbf{Z}_i^D$ is the matrix of instruments for the equations in first differences, as described above. Again Z_i would be precisely the same if the transformed equations in y_i^* and W_i^* were in orthogonal deviations rather than first differences. In models with explanatory variables, it may be that the levels of some variables are uncorrelated with η_i, in which case suitably lagged levels of these variables can be used as instruments in the levels equations, and in this case there may be instruments available for the first levels equation.

For the system of equations in first differences and levels, the one-step estimator computed in DPD98 uses the weighting matrix

$$H_i = \begin{pmatrix} H_i^D & 0 \\ 0 & I_i \end{pmatrix}$$

where H_i^D is the weighting matrix described above for the first differenced estimator, and I_i is an identity matrix with dimension equal to the number of levels equations observed for individual i. For the system of equations in orthogonal

deviations and levels, the one-step estimator computed in DPD98 sets H_i to an identity matrix with dimension equal to the total number of equations in the system for individual i. In both cases the corresponding two-step estimator uses $H_i = \widehat{v}_i^* \widehat{v}_i^{*\prime}$. We adopt these particular one-step weighting matrices because they are equivalent in the following sense: for a balanced panel where all the available linear moment restrictions are exploited (i.e. no columns of Z_i are omitted for computational or small sample reasons), the associated one-step GMM estimators are numerically identical, regardless of whether the first difference or orthogonal deviations transformation is used to construct the system. Notice though that the one-step estimator is asymptotically inefficient relative to the two-step estimator for both of these systems, even if the v_{it} are homoskedastic.[9] Again simulations have suggested that asymptotic inference based on the one-step versions may be more reliable than asymptotic inference based on the two-step versions, even in moderately large samples.[10]

The validity of these extra instruments in the levels equations can be tested using the Sargan statistic provided by DPD98. Since the set of instruments used for the equations in first differences (or orthogonal deviations) is a strict subset of that used in the system of first-differenced (or orthogonal deviations) and levels equations, a more specific test of these additional instruments is a Difference Sargan test which compares the Sargan statistic for the system estimator and the Sargan statistic for the corresponding first-differenced (or orthogonal deviations) estimator. Another possibility is to compare these estimates using a Hausman specification test, which can be computed here by including another set of regressors that take the value zero in the equations in first differences (or orthogonal deviations), and reproduce the levels of the right hand side variables for the equations in levels.[11] The test statistic is then a Wald test of the hypothesis that the coefficients on these additional regressors are jointly zero. Full details of these test procedures can be found in Arellano and Bond (1991) and Arellano (1993).

[9] With levels equations included in the system, the optimal weight matrix depends on unknown parameters (for example, the ratio of $\text{var}(\eta_i)$ to $\text{var}(v_{it})$) even in the homoskedastic case.

[10] See Blundell and Bond (1998).

[11] Thus in the AR(1) case described above we would have

$$W_i^* = \begin{pmatrix} 0 & \dots & 0 & y_{i2} & \dots & y_{i(T-1)} \\ \Delta y_{i2} & \dots & \Delta y_{i(T-1)} & y_{i2} & \dots & y_{i(T-1)} \end{pmatrix}'.$$

3. USING DPD98

All sample selection and model specification information are input to DPD98 by editing the DPD98.RUN file. This can be done using the Gauss editor or any other compatible editor, and only a very basic knowledge of Gauss is needed to operate DPD98. Options to include a constant and dummy variables, and selection from the menu of estimators and output options, can be specified using the DPD98.RUN file, or can be chosen interactively when running DPD98.

3.1. USER INPUT INFORMATION: THE DPD98.RUN FILE

The DPD98.RUN file is organized into several sections, each with a title and comments to provide 'on-line' assistance. These sections are each discussed below.

Data Set Selection

The data to be used in estimation are selected towards the top of DPD98.RUN. The main data set is specified by typing the name of the Gauss data set in the open statement, i.e.

open f1=mainname;

where mainname is the name of the main Gauss data set (without any extension). The auxiliary data set is selected in the same way at the following open statement, i.e.

open f2=auxname;

where auxname is the name of the auxiliary Gauss data set (without any extension). If these data sets are not on the same subdirectory (folder) as DPD98.RUN then their location must also be specified in these statements. It is recommended that data is read from the hard disk only.

Immediately beneath these open statements, the variables startf2 and stopf2 must be entered. These control where DPD98 begins and ends reading observations from the main data set, and should always take positive integer values. If all the data is to be used then startf2 should be set to 1 and stopf2 should be set to the number of rows in the auxiliary data set (which can be done automatically by specifying stopf2=rowsf(f2)). However, if the main data set is sorted by some characteristic of the individual units, this feature will allow estimation on a subsample. For example, if in an unbalanced panel the data are sorted in ascending order of time series observations per unit, then the balanced sub-panel with data for all time periods can be selected in this way. This would be obtained by setting startf2=rowsf(f2) and stopf2=rowsf(f2).

Other Data Information

The next section sets up several variables that are needed in reading the main data set and for creating dummy variables. All should take integer values. The first of these is ncomp, which controls how many cross-section units are read and processed at the same time. DPD98 operates by reading and processing the data in blocks of ncomp (or less) units. If an 'insufficient workspace' or 'read too large' error message is encountered, the solution is to reduce ncomp. The precise limit will depend on the number of instruments in the model and (to a lesser extent) on the maximum number of time-series observations per unit.

The next variable, yearcol, simply indicates the column of the main data set that contains the year to which each observation refers. One column must contain this information, and this must be in the format 19xx. The variable year1 should be set to the earliest year on which there is an observation in the (sub-sample of the) data being used. Again this must be in the format 19xx. The variable nyears should be set to the number of years covered by the (sub-sample of the) data being used. For example, if there are observations between 1984 and 1991, then year1 should be set to 1984 and nyears should be set to 8. Notice that, unlike in the earlier versions of DPD, it is not necessary for any single cross-section unit to have observations on all of these time periods.[12] However it is not advisable to include years for which there are very few observations, particularly when period-specific parameters are being estimated.

Cross-section units may also be associated by some time-invariant observed characteristic, and if such a group indicator is available in the data then DPD98 will create intercept dummies according to this characteristic. With company data this will often indicate an industry grouping to which the firm belongs. This is assumed in what follows, but other types of grouping (e.g. size, location) can be used in practice. The variable indcol indicates the column of the main data set that contains the industry code to which each observation is classified. The variable indmax indicates the number of groups or classes that have been used. Where this option is used, the industry codes in the data set must be integers running from one to indmax, with no gaps. Note that this is only an option. Where such a classification is either not desired or unavailable, indcol may be set to any arbitrary column number in the data set and indmax may be set to any arbitrary value. In this case the creation of 'industry' dummies (see below) should not be requested.

Model Information

The parameter lag controls how many equations are used in estimation, and

[12] For example, it may be that half the sample have observations from 1984-89 and half the sample have observations from 1986-91.

how many time-series observations on each unit are reserved to allow the creation of lagged series. This parameter should be set with reference to the model in levels, before considering any transformation needed to compute the estimator. When estimation is in first differences (or orthogonal deviations), DPD98 will automatically reserve one extra observation to allow the transformed series to be constructed. *Users familiar with earlier versions of DPD should note that this is a change from earlier versions of the program.*

The lag parameter will normally be set equal to the maximum lag length in the model. For example, in the AR(1) model we discussed in section 2,

$$y_{it} = \alpha y_{i(t-1)} + \eta_i + v_{it}$$

with v_{it} serially uncorrelated, the lag parameter will normally be set to 1. If the model was to be estimated using OLS in levels (perhaps for comparison to other estimators), the first levels equation used would be

$$y_{i2} = \alpha y_{i1} + \eta_i + v_{i2}$$

whilst if the same model was to be estimated using GMM in first differences, the first differenced equation to be used would be

$$\Delta y_{i3} = \alpha \Delta y_{i2} + \Delta v_{i3}$$

Similarly if the model to be estimated was

$$y_{it} = \alpha y_{i(t-1)} + \beta_0 x_{it} + \beta_1 x_{i(t-1)} + \beta_2 x_{i(t-2)} + \eta_i + v_{it}$$

the lag parameter would normally be set to 2.

There are two reasons why the lag parameter may sometimes be set to higher values than the maximum lag length in the model. One reason is that the user may want to estimate the model using a later sub-period of the data. The other is that there may be no instruments for the earliest equations in the presence of moving average errors. For example, if in the AR(1) model it is known that the v_{it} disturbances are MA(1), then instruments dated $t-2$ are not valid in the differenced equations, but instruments dated $t-3$ and earlier remain valid. In this case the first differenced equation for which instruments are available is

$$\Delta y_{i4} = \alpha \Delta y_{i3} + \Delta v_{i4}$$

where y_{i1} is a valid instrument. By setting lag to 2, rather than 1, the redundant differenced equation for period 3 can be omitted.

Data Transformations

The next section of DPD98.RUN defines a Gauss subroutine in which data transformations and model selection are performed. Whenever this subroutine is called, all the columns of the main data set for the current block of units will have been read into a matrix called data. At this point, any operation that is available in Gauss may be performed on the columns of data in order to effect data transformations. Suppose for example that the data set contains 6 columns, and it is desired to use the ratio of the variables in columns 5 and 6 as a regressor in the model. This can be achieved by typing the following pair of statements in the subroutine (before the model is selected):

```
temp = data[.,5]./data[.,6];
data = data~temp;
```

The first statement here picks out column 5 of data and divides each element by the corresponding element of column 6. The result is assigned to a vector called temp. The second statement then attaches this vector to the right hand side of data ('horizontal concatenation'). Thus data now has 7 columns, and the new variable occupies column 7. In this way transformed variables can be used in the model without being permanently stored in the data set. Any Gauss operations (e.g. logarithms, powers) can be performed similarly, and the variable names temp$*$ are reserved for this purpose. Note that it is essential that, after performing data transformations, the modified data matrix continues to have the name data, since DPD98 will look for this matrix when selecting the model.

DPD98 also has 3 specific functions that can be used to construct transformed series at this stage. The function timdum(19xx) produces a column vector which takes the value of one for all observations in year 19xx and zero for observations in all other years. Similarly the function inddum(j) produces a column vector with ones for all observations for industry j, and zeros elsewhere. A dummy variable taking the value 1 in years 19xx and 19yy (or in industries j and k) can be obtained either by adding (i.e. timdum(19xx)+timdum(19yy)), or more simply by specifying timdum(19xx~19yy) (likewise inddum(j~k)). These functions can be used to obtain year-specific or industry-specific intercept dummies, in cases where the complete set of individual year or industry dummies is not desired.[13] By interacting these dummy variables with explanatory variables, these functions also allow the sub-period or sub-sample stability of some or all of the model coefficients to be investigated.

Finally the function back(c,l) allows lags of the basic series to be used in constructing transformations. Simple lags can be specified more easily at the model selection stage, but this function allows more complex transformations

[13] DPD98 will automatically include the complete set of year dummies and/or industry dummies as options (see below).

to be performed. The function works like a lag operator or backshift operator, producing the lth lag of the series in column c of data. For example, if instead of dividing observations in column 5 by the current value of the series in column 6, it is desired to divide by the first lag of the series in column 6, this can be achieved by the statement:

temp = data[.,5]./back(6,1);

As another example, the growth rate of the series in column 4 can be produced by the statement:

temp = (data[.,4]-back(4,1))./back(4,1);

Notice that in each of these examples there is no observation available on the transformed series for the first time period in which each individual unit is observed, and the transformed series will contain missing values for these periods. This should be taken into account when specifying the lag parameter which determines which observations are used in estimation.

Model Selection

The model to be estimated is selected using simple DPD98 functions. As in the econometric discussion above the dependent variable is y, the regressor matrix is x and the instrument matrix is z. The variables selected are also given names that will appear in the output.

a) Dependent variable

The dependent variable is selected with the functions lev(c,l), dif(c,l), dev(c,l), diflev(c,l) or devlev(c,l), which respectively return a series in levels, first-differences, orthogonal deviations, a stacked vector of first differences and levels, and a stacked vector of orthogonal deviations and levels. In all cases the first argument c indicates the column of data which contains the basic variable and the second argument l indicates the lag length to be produced. For example the statement:

y = lev(3,0);

selects the variable in column 3 of data to be the dependent variable, in levels form. Similarly:

y = dif(4,0);

selects the variable in column 4, and uses it in first-differenced form. Typically the lag length will be zero when selecting the dependent variable, although this is not essential.

After making y in this way, the selected variable must also be given a name. This is entered immediately below as the variable namey. The name has a maximum length of eight characters and must be enclosed between double inverted commas (”). Both upper and lower case may be used. For example, any of the following are acceptable:

namey = "v3";
or namey = "OUTPUT";
or namey = " log N";

b) Regressors

The same functions lev(c,l), dif(c,l), dev(c,l), diflev(c,l) or devlev(c,l) are used to select the matrix of regressors. Single columns are combined into a matrix using the horizontal concatenation operator (~) in Gauss. Any lag lengths may be used, up to the user-specified maximum lag (see above). For example, the statement:

x = dev(7,0)~dev(7,1);

selects a matrix of 2 regressors, both formed from the basic variable in column 7 of data and both in orthogonal deviations form. The first regressor is not lagged, and the second regressor is lagged one period. This requires lag to be set to 1 (or higher), and a minimum of at least 3 time-series observations to be available on each unit.

Each of the regressors chosen must again be given a name. Names are also combined using horizontal concatenation, and each name must be enclosed in double inverted commas. These are entered as the variable namex. For example, the statement:

namex = "Q"~"Q(-1)";

could correspond to the regressor matrix specified above.

In addition to these basic functions, DPD98 offers two additional functions which may be used when selecting the matrix of regressors (and instruments). To combine both levels of some series and first differences of other series in the same regressor matrix, the function lev1(c,l) should be used rather than lev(c,l). Since one observation is automatically lost when constructing first differences, the function lev1(c,l) returns the level of the series with the first available level omitted. The function zerolev(c,l) returns a stacked vector of the same length as diflev(c,l) (or devlev(c,l)), but with zeros in place of the observations in first differences (or orthogonal deviations). This can be used to construct tests of the validity of the instruments for the levels equations, as discussed in section 2.

The only regressors that are not chosen in this way are the constant and intercept dummies. These can be added to the model automatically as described below, and if they are selected then names are automatically assigned by DPD98.

c) Instruments

We first describe how to specify the matrix of instruments when estimation uses the equations in levels, first differences or orthogonal deviations only. We then explain how this extends to the systems of transformed and levels equations.

A matrix of instruments may be formed using the basic lev(c,l) and dif(c,l) functions, in the same way as the matrix of regressors above. In addition DPD98 has a further function, gmm(c,l_1,l_2), which automatically returns all or part of the optimal instrument matrix required for the Generalised Method of Moments estimators discussed in section 2. Again c indicates the column position of the basic variable in data. Here l_1 refers to the lag length of the latest instrument to be used in each cross section, and l_2 refers to the lag length of the earliest instrument to be used. Thus if observations dated $t-2$, $t-3$ and $t-4$ are to be used as instruments in each of the cross-section equations, then l_1 would be set to 2 and l_2 would be set to 4.[14] *Users familiar with earlier versions of DPD should note that this is a change from earlier versions of the program.* If the number of cross-section units is large enough that all observations dated $t-2$ and earlier are to be used as instruments, this can be achieved by setting l_1 to 2 and l_2 to the default value of 99. Where future values of strictly exogenous variables are used as instruments, this is achieved by setting l_1 to a negative integer. For example, setting l_1 to -2 and l_2 to 3 will use the observations dated $t+2$, $t+1$, t, $t-1$, $t-2$ and $t-3$ as instruments. Similarly setting l_1 to -99 and l_2 to 99 will use all past, present and future observations on the series as instruments in each of the cross-section equations.[15]

Where series have been constructed using the backshift operator, or for other reasons the first m observations on each unit are missing, DPD98 has a corresponding function gmmb(c,l_1,l_2,m). The first 3 arguments work in the same way as for the basic gmm(c,l_1,l_2) function. The fourth argument allows columns of the instrument matrix corresponding to the missing observations to be deleted.[16] For example, if the variable is a growth rate constructed using back(c,1), the parameter m should be set to 1.

Matrices produced by gmm(c,l_1,l_2) or gmmb(c,l_1,l_2,m) may be combined with each other, or with vectors produced by lev(c,l) or dif(c,l), again using horizontal

[14] If any of these lagged values are not observed in the earlier cross-sections, DPD98 will automatically delete these columns from the instrument matrix. Unbalanced panel considerations are dealt with in the way discussed in section 2.

[15] Note that the same instrument set should be specified when estimating in orthogonal deviations as when estimating in first differences. The precise timing of the orthogonal deviations transformation used by DPD98 is

$$x^*_{it} = \left(x_{i(t-1)} - \frac{x_{it} + ... + x_{iT)}}{T-t+1} \right) \left(\frac{T-t+1}{T-t+2} \right)^{1/2} \text{ for } t = 2, ..., T$$

rather than that given in footnote 2. Thus if instruments dated $t-s$ are valid using first differences, they are also valid using orthogonal deviations.

[16] In unbalanced panels, missing values not removed by these column deletions are replaced by zeros.

concatenation. Where year dummies and/or industry dummies are requested, these are automatically included in the instrument matrix as well as the regressor matrix. If the total number of instruments (columns of z) specified is less than the total number of regressors (columns of x), so that the model is not identified, an error message is returned.

Note that the columns of x are not automatically included in the instrument matrix. Where some of the x variables are exogenous and are used simply to instrument themselves, these variables must be included explicitly in the instrument matrix as well as in the regressor matrix. This can be done using the lev(c,l), dif(c,l) or dev(c,l) functions, or by including a sub-matrix of x in the instrument matrix.[17] When OLS rather than an instrumental variables estimator is required, this can be achieved either by specifying:

```
    z = x;
or  z = ols;
```

When within groups is required, this can be achieved either by specifying y and x in orthogonal deviations and using OLS, or by specifying y and x in levels, specifying:

```
    z = wgroups;
```

and requesting within groups from the menu of estimation options (see below).[18]

In contrast to x, a name does not have to be specified for each column of the instrument matrix, z. A list of names which summarises the instrument set should however be entered as the variable namez. With OLS and within groups estimators, this selection is quite arbitrary, but some name must be entered.

We close this section with two examples that illustrate the syntax:

```
    z = gmm(7,2,99);
    namez = "Q(2,ALL)";
or  z = gmmb(6,2,4,1)~gmm(7,2,3)~dif(3,2);
    namez ="Y*(2,4)"~"Q(2,3)"~"DN(-2)";
```

System Estimators

When using systems of equations in first differences (or orthogonal deviations) and in levels, the instrument matrix is constructed in two parts. First, the instruments for the transformed equations in the system is set up in just the same way as the instrument matrix would be specified for estimation using first differenced (or orthogonal deviations) equations alone. Second, the additional instruments

[17] For example, the statement x[.,3 6 7] selects the sub-matrix containing columns 3, 6 and 7 of the x matrix. See the Gauss manual for further details.

[18] In either case, DPD98 calculates within groups estimates using the equivalence between the classical within groups estimator and OLS after transforming to orthogonal deviations. This equivalence is exact for balanced panels, and asymptotic for unbalanced panels. See Arellano and Bover (1995) for further discussion.

used in the levels equations must be specified. DPD98 provides two functions for this purpose: gmmlev(c,l_3) uses the level of the series in column c of data, lagged l_3 times, as an instrument in each of the levels equations of the system; gmmlevd(c,l_3) uses the first difference of the series in column c, lagged l_3 times, as an instrument in each of the levels equations.[19] Typically the lag length l_3 will be equal to the corresponding $l_1 - 1$, since if, for example, $x_{i(t-2)}$ is uncorrelated with the first-differenced error term Δv_{it}, then both $x_{i(t-1)}$ and $\Delta x_{i(t-1)}$ will be uncorrelated with the time-varying component v_{it} of the error term in levels.

After constructing each part of the instrument matrix separately, the two parts must be combined to give the block diagonal instrument matrix for the system described in section 2. This is achieved using the function combine(zd,zl), where zd is the instrument matrix set up for the first-differenced equations, and zl is the instrument matrix set up for the levels equations in the system. Note that the order of these two arguments is important. A list of names describing the full instrument matrix should then be specified as namez.

To illustrate this, suppose we are estimating a simple AR(1) model for the variable in column 3 of data, using a system of equations in orthogonal deviations and levels, under the maintained assumption of mean-stationarity. The complete model specification section of the DPD98.RUN file would then have the form:

```
y = devlev(3,0);
namey = "Y";
x = devlev(3,1);
namex = "Y(-1)";
zd = gmm(3,2,99);
zl = gmmlevd(3,1);
z = combine(zd,zl);
namez = "Y(2,ALL)"~"+DY(-1)";
```

Note that to test the validity of these instruments in the levels equations, we could replace this matrix of regressors by:

```
x = devlev(3,1)~zerolev(3,1);
namex = "Y(-1)"~"ZY(-1)";
```

whilst keeping the instrument matrix unchanged.

Note finally that when some regressors are used simply to instrument themselves,[20] these should be concatenated to z only after z has been constructed using the combine(zd,zl) function.

[19] In each case there is a corresponding function gmmlevb(c,l_3,m) and gmmlevdb(c,l_3,m) for series with m missing observations.

[20] Here this requires that they are uncorrelated with the individual effects η_i, as well as being uncorrelated with the errors in the transformed equations.

Pseud's Corner

In cases where the total number of instruments is large relative to the cross-section dimension of the panel, there may be difficulty in inverting the matrix $\left(\frac{1}{N}\sum_i Z_i'\widehat{v}_i^*\widehat{v}_i^{*\prime}Z_i\right)$ which is required to compute the two-step GMM estimator. This will typically cause the program to abort and return a 'matrix not invertible' error message at the end of the second read through the data.[21] When this happens, the estimator can still be computed by using a Moore-Penrose pseudo-inverse to evaluate the weight matrix. DPD98 allows this as an option, by setting the parameter pseud to one in the Gauss options section of the DPD98.RUN file. This will occur automatically in cases where the total number of instruments exceeds the number of cross-section units.

For normal use, we recommend that the parameter pseud is set to zero, so that pseudo-inverses are not routinely used. Non-singularity of the matrix $\left(\frac{1}{N}\sum_i Z_i'\widehat{v}_i^*\widehat{v}_i^{*\prime}Z_i\right)$ can be taken as a signal that the number of instruments is becoming large for the given sample size, and this may be useful information to have when re-specifying the model.

The User-Defined Wald Test

When DPD98 is run it will automatically compute a Wald test of joint significance for all the variables entered in x (i.e. a test of the null hypothesis that their estimated coefficients are all zero). When intercept dummies are selected, similar tests of their joint significance are computed. In addition the user may select a subset of the regressors in x to be separately tested. This can be useful in testing for sub-sample stability, as well as more general linear restrictions.

This option is turned on by setting the variable waldtest to 1. Otherwise waldtest should be set to 0. When the option is selected, the columns of x that are to be tested are specified by entering the column numbers as the variable testcols. Note that these refer to column positions in x rather than in data, and they are combined using horizontal concatenation. For example, to test the joint significance of the variables in the first two columns of x, the statement is:

testcols =1~2;

As usual, an arbitrary value should be assigned to testcols when this option is not being used.

Saving the Output

Output from DPD98 will appear on the screen but should also be directed to an output file for subsequent inspection and printing. This is accomplished by typing a filename in the output file statement at the bottom of DPD98.RUN. Any

[21] If the program aborts with a 'matrix not invertible' error message at the end of the first read of the data, this is normally because some of the columns of the instrument matrix are perfectly collinear. In this case, the model should be re-specified.

valid DOS filename may be used here, and a location other than the default drive may be specified. The output file produced by DPD98 is an ASCII text file that can be edited in all standard text editors, or read directly into word processors. After the filename, one of the words on or reset must be included. For example

output file = c:\myoutput\results.txt on;

If on is used, the output from this run will be appended to the bottom of the output file. If reset is used, the output file will be overwritten. Care should be exercised when using the latter option!

3.2. RUNNING DPD98

Once DPD98.RUN has been edited the program is ready to run. From within Gauss, DPD98 can be run from command mode with the command:

run DPD98.run

Alternatively DPD98 can be run from the edit mode, which is often convenient. Enter the Gauss editor with the command:

edit DPD98.run

and use the F2 key from the editor to execute the program.

3.2.1. INTERACTIVE MODE: THE MENU OF OPTIONS

To run DPD98 interactively, the parameters sys and bat in the Gauss options section at the top of the DPD98.RUN file should both be set to zero. On running DPD98, the user is then presented with a series of options which are controlled by typing answers to prompts on the screen.

The first question asks for the form of the model to be entered. Type 0 if the model is specified in levels (unless within groups is required), 1 if the model is specified in first-differences, 2 if the model is specified in orthogonal deviations, 3 if the model is specified using a system of first-differenced and levels equations, 4 if the model is specified using a system of orthogonal deviations and levels equations, and 5 if the model is specified in levels and the z = wgroups command has been used. In each case the number entered should be followed by the return key. This information determines the form of the H_i matrix used to compute the one-step estimator as discussed in section 2, the error structure assumed for computing non-robust covariance matrices, and the form of the serial correlation tests reported.

The second question asks for a choice to be made from the options to include a constant and various intercept dummies. Type 0 for no constant, 1 for a single set of year dummies, 2 for indmax sets of year dummies (i.e. separate year dummy

coefficients for firms in different industries), 3 for a single constant, 4 for a set of industry dummies only (with no year dummies), and 5 for additive year dummies and industry dummies. Where year dummies are specified, DPD98 includes a constant term and excludes the year dummy for the first year available. The coefficient on the constant term estimates the intercept in the first period available, whilst the coefficients on the remaining year dummies estimate the difference between the intercept in these years and the intercept in the first period. This is equivalent to including a full set of year dummies, and when estimation is in levels this allows the hypothesis that the set of dummies can be replaced by a single intercept to be easily tested. This is the Wald test produced automatically when estimation is in levels. When estimation is in first differences (or orthogonal deviations), the same hypothesis implies that there should be no intercept at all in the first differenced (or orthogonal deviations) equations, and this is again the Wald test that is automatically computed by DPD98.

Where more than one set of year dummies is specified, DPD98 similarly includes a constant term, a set of year dummies for all industries (with the first year omitted), and then a full set of separate year dummies for each of the industries 2 to indmax. This is equivalent to including a full set of industry-year interactions, and facilitates testing of the hypothesis that the multiple sets of year dummies can be replaced by a single set of year dummies (common to all industries). This is the Wald test automatically produced by DPD98 in this case.

Where industry dummies alone are specified, DPD98 includes a constant term and excludes the industry dummy for the first industry. The Wald test produced here tests the hypothesis that the set of industry dummies can be replaced by a single intercept.

Where additive year and industry dummies are specified, DPD98 includes a constant term and excludes both the first year dummy and the first industry dummy available. Separate Wald tests test the hypotheses that the year dummies can be omitted, the industry dummies can be omitted, and that the additive year and industry dummies can be replaced by a single intercept (i.e. that both sets of intercept dummies can be omitted).

The third question asks whether standard errors and test statistics that are (large N asymptotically) consistent in the presence of general heteroskedasticity are to be computed. In this and the following questions, type 1 if this option is desired and 0 otherwise. Where appropriate the two-step instrumental variables estimator (see above) is also produced when this option is requested. This option requires the data set to be read a second time and so increases the execution time of the program. However in our experience heteroskedasticity is often present in panel data models. In this case the non-robust test statistics may be seriously misleading, so that this option is strongly recommended. For the system estima-

tors, DPD98 does not produce non-robust standard errors, so that this option is automatic and the question is omitted.

The fourth question asks whether basic descriptive statistics should be included in the output file. The descriptive statistics available show the mean, standard deviation and extreme values of each series in y and x, together with a matrix of simple correlation coefficients. Except where the model is estimated in first-differences, these descriptive statistics are provided for the levels of the series.

The fifth question asks whether the full covariance matrices for the estimated parameters should be included in the output file. When they have been computed, the heteroskedasticity-consistent covariance matrices for the one-step and two-step estimators are reported when this option is requested.

The last question asks whether the vectors of coefficient estimates and covariance matrices should be saved as Gauss matrices. When this is requested, the one-step and two-step coefficient matrices are saved as beta1.fmt and beta2.fmt, their heteroskedasticity-consistent covariances matrices are saved as var1.fmt and var2.fms, and the non-robust covariance matrix for the one-step estimator is saved as var.fmt. For users familiar with Gauss, these saved matrices can be loaded into other Gauss programs, for example to compute Wald tests of non-linear restrictions and Hausman specification tests.

3.2.2. BATCH MODE

To run DPD98 in batch mode, the parameter bat should be set to one. In this case, the six questions described in the previous section will be suppressed, and the desired options must be declared in the DPD98.RUN file, as the parameters imod, icon, irob, ides, icov and isav in the section labelled Set Up For Batch Operation. If the parameter sys is set to zero, the program will return to the Gauss $\gg$ prompt when it finishes executing. If the parameter sys is set to one, the program will return to DOS when it finishes executing.

This latter option allows a series of DPD98.RUN files to be executed sequentially, using a DOS batch file. The precise details will depend on how Gauss is installed on your computer (or network). In all cases, first prepare a series of DPD98.RUN files, each with bat set to one and sys set to one, and saved with distinct names (for example, DPD98A.RUN, DPD98B.RUN, etc.). Then prepare a DOS batch file that will execute a series of programs written in Gauss.

For example, if Gauss is installed locally on your computer and you can enter Gauss (from DOS, or from a DOS window) simply by typing gaussi$\hookleftarrow$. Then all that is required is a batch (.bat) file which contains a series of statements of the form:

gaussi/b run DPD98.RUN

The /b option is recommended but not essential; this allows the batch file to move on to the next program should one of your jobs abort.

In this case, for example, a simple batch file rundpd.bat containing the two lines

gaussi/b run DPD98A.RUN

gaussi/b run DPD98B.RUN

can be used to run these two DPD98.RUN files sequentially. This is executed from the DOS prompt, simply by typing rundpd↩. Note that this can be run from a DOS window under Windows95 or WindowsNT.

Where Gauss is run over a network, the form of the batch file needed may be more complicated but the principle is the same. Often Gauss will be executed over the network using a local batch file (in Windows95 or WindowsNT, the shortcut to Gauss should reveal the name of this batch file, which should also be executable from a DOS window). Within this batch file, you should find the gaussi statement (probably preceded by a path, e.g. r:\apps\gauss\gaussi), which is where the batch file starts Gauss. Multiple Gauss programs can then be executed sequentially by replacing this statement by a series of statements of the form gaussi/b run DPD98.RUN, each preceded by the same path.

3.3. OUTPUT FROM DPD98

The output file produced by DPD98 is largely self-explanatory. The last column in the main tables, labelled P-Value, reports the probability of rejecting the null hypothesis that the coefficient is zero, using a two-tailed test. With the basic one-step estimates the residual sum of squares (RSS) and total sum of squares (TSS) are reported, along with the estimated variance of the error term. When the model is estimated in levels, this provides an estimate of the variance of $(\eta_i + v_{it})$. When the model is estimated in first-differences or orthogonal deviations, this provides an estimate of the variance of the time-varying component v_{it} only.

The Wald tests reported are asymptotically distributed as χ^2 variables, with the degrees of freedom (df) reported. The Sargan tests of overidentifying restrictions are also asymptotically distributed as χ^2. The tests for first-order and second-order serial correlation relate to the estimated residuals in first-differences, unless the model has been estimated using only the levels equations. Note that first-differencing will induce MA(1) serial correlation if the time-varying component of the error term in levels is a serially uncorrelated disturbance. These tests are discussed in Arellano and Bond (1991), and are asymptotically distributed as standard normal variables. For all these test statistics, p-values report the probability of rejecting the null hypothesis. Where robust test statistics are computed

and the data is read a second time, the complete serial correlation matrix (based on the one-step residuals) is also reported.

4. AN EXAMPLE

In this section we describe an example DPD98.RUN file together with the output file that it produces. The example data sets XDATA and AUXDATA are supplied with DPD98. XDATA has six columns which contain data for the sample of 140 UK quoted companies over the period 1976-1984 used in Arellano and Bond (1991). The variables in these columns are an industry code, the accounting year, employment, real wages, gross capital stock and an index of industry output respectively. The panel is unbalanced, with observations varying between 7 and 9 records per company.

The example DPD98.RUN file specifies a log-linear labour demand equation including 2 lags of the dependent variable, current and lagged real wages, current capital and current and lagged industry output. Notice that the log series are constructed internally at the data transformations stage. The model is estimated in first differences. The instrument set exploits all available linear moment restrictions involving the dependent variable (assuming no serial correlation in the time-varying component of the errors in levels), in combination with the remaining regressors in stacked form. A Wald test of the joint significance of the two real wage variables is computed.

On running DPD98 time dummies were requested, but not industry dummies. Robust test statistics and two-step estimates were selected. Descriptive statistics and covariance matrices of the estimates were omitted. This results are contained in the output file DPD98.OUT.

References

[1] Alvarez, J. and Arellano, M. (1998), "The time series and cross-section asymptotics of dynamic panel data estimators", mimeo, CEMFI, Madrid.

[2] Arellano, M. (1988), "An alternative transformation for fixed effects models with predetermined variables", Institute of Economics and Statistics, Oxford, Applied Economics Discussion Paper no.__.,

[3] Arellano, M. (1993), "__", *Journal of Econometrics*.

[4] Arellano, M. and Bond, S.R. (1988), "Dynamic panel data estimation using DPD - a guide for users", Institute for Fiscal Studies Working Paper no. 88/15.

[5] Arellano, M. and Bond, S.R. (1991), "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations", *Review of Economic Studies*, 58, 277-297.

[6] Arellano, M. and Bover, O. (1995), "Another look at the instrumental-variable estimation of error-components models", *Journal of Econometrics*, 68, 29-52.

[7] Blundell, R.W. and Bond, S.R. (1998), "Initial conditions and moment restrictions in dynamic panel data models", *Journal of Econometrics*, 87, 115-143.

[8] Chamberlain, G. (1984), "Panel Data", in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics*, Volume II, Elsevier Science Publications.

[9] White, H. (1982), "Instrumental Variables Regression with Independent Observations", *Econometrica*, 50, 483-499.