XPReg: Software for the Analysis of Cross-Product Matrices

1. Introduction

XPReg is a GAUSS program which takes as its input a GAUSS matrix containing either variables organised into columns or in the form of a cross-product matrix. XPReg analyses the time-varying cross-section (TVCS) and time-varying fixed effect (TVFE) models:

$$TVCS \quad y_{it} = x_{it}\beta_t + \mu_t + \varepsilon_{it}$$

$$TVFE \quad y_{it} = x_{it}\beta_t + \lambda_t + \alpha_i + \eta_{it}$$

The two models are essentially the same expect for the different time dummies and error terms. The reason is that the TVFE takes account of any individual-specific effects which do not vary over time for the individual (including observed variables which are time-invariant).

XPReg is unique in two respects; first, it is the only cross-sectional or panel program to take the variation of the coefficients over time as its basic model. In the case of panel models, the panel relationship of the TVFE embodied in equation (1) is almost non-existent The restriction of time-invariance has been criticised at length in Bell and Ritchie (1996), and all the estimation methods in the programme assume time-varying parameters unless restricted.

The second difference is that XPReg was expressly designed to take a cross-product matrix (X'X) as its input. Basic OLS regression can be carried out by analysing the cross-product matrix. The advantage of this is where there is a large amount of data or where the data are confidential. As the size of the input matrix is determined by the number of variables, not the number of observations, this method is appropriate for very large datasets. Even if the dataset is small enough to fit into memory, the time savings from this method can be considerable. Because the data in a cross-product matrix are aggregated, then data subject to confidentiality restrictions can be safely made available to researchers by storing it in a cross-product matrix. Further information can be found in Ritchie (1995).

XPReg has recently been amended to allow "unmomented" matrices to be taken as input.

XPReg has four estimation methods. The TVCS is estimated by OLS or analysis of covariance. The TVFE is estimated by analysis of covariance or by differencing. Models with pre-differenced data can also be estimated. Minimum distance is feasible but not yet implemented.

2. Data

Input data is one or more GAUSS matrix: either a "raw" matrix of observations or a cross-product matrix. The latter form requires a separate "information matrix" and may also require subsidiary matrices. The panel need not be balanced; all observations (except those with missing values) will be used.

2.1 Raw data input

The raw matrix contains observations ordered by individual and by period, with all the observations for one individual gathered together. The individual and year identifiers must be in the first two columns of the matrix. The remaining columns contain the data. Missing data will cause that observation to be ignored. The top row contains character elements which are the variable names. It is important that these are character elements as these will be the variable names display later; numerical fields will not be displayed properly.

A typical matrix with four variables might look like:

Name	year	wage	sex	hours	age
Eric	75	105	0	37.5	23
Eric	76	112	0	37.5	24
Eric	79	127	0	37	27
:	:	:	:	:	:
Lucy	76	89	1	37.5	18
Lucy	77	100	1	37.5	19
:	:	:	:	:	:

The scale of the year column is irrelevant: XPReg calculates the period index as offsets from the lowest value it finds in this column. Thus the above records would refer to periods 1, 2, 5...2, 3...The year index must therefore be numeric. The individual index column can be character or numeric, but should be numeric as certain character strings can be confused with GAUSS special numeric codes.

It is not necessary to include a constant term or individual or time dummies. These will be added as necessary by the program.

2.2 Cross-product matrix input

This matrix is the cross-product of the raw matrix, but without the individual or period indexes or the column headings and including a constant term. Thus, if X is the above matrix without the first row and first two columns, including a constant and padded with zeroes where observations are missing:

1	105	0	37.5	23
1	112	0	37.5	24
0	0	0	0	0
0	0	0	0	0
1	127	0	37	27
:	:	:	:	:
0	0	0	0	0
1	89	1	37.5	18
1	100	1	37.5	19
:	:	:	:	:

then the cross product matrix is a 5x5 matrix X'X. *However*, to allow for TV parameters, XPReg treats each variable set for each period as separate regressors. In other words, the X matrix should actually be:

1	105	0	37.5	23	0	0	0	0	0	0	0	0	0	...
0	0	0	0	0	1	112	0	37.5	24	0	0	0	0	...
0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
:				:				:				:		

Thus the cross-product matrix is actually TKxTK, where T is the maximum number of observations for any individual and K is the number of variables, including a constant. The formula for any of the T^2 KxK submatrices in the cross-product matrix is

$$X'X[t,s] \equiv \sum_{i \in S(t,s)} x_{i't} x_i$$

where S(t, s) is the set of people observed at time t and time s. See Ritchie (1995).

The constant does not have to be in the first column.

For TVFE regression on a unbalanced panel using the cross-product matrices, a second "means" matrix must be supplied. This has the formula

$$X'X[t,s] \equiv \sum_{i \in S(t,s)} \frac{1}{T_i} x_{i't} x_{is}$$

where T_i is the number of observations for individual i. Note that the "main" matrix is independent of the period studied, whereas changes in the period under consideration might lead to T_i being different. Thus the means matrix is constructed for a *particular estimation period*; again, see Ritchie (1995).

If only cross-section studies are wanted, then a TKxK matrix will suffice. These are the T diagonal KxK blocks from the full cross-product matrix:

$$X'X[t] \equiv \sum_{i \in S(t)} x_{i't} x_{it}$$

These can also be used for the TVFE regression but the means matrix must be full-size.

A program called MakeXX.GP constructs cross-product and means matrices from the raw X'X matrix. If your dataset is very large then use of the cross-product format can save a considerable amount of time. particularly for the TVCS.

The information file

The information file is a GAUSS matrix file. It is created automatically a program called MetaProg.GP, which is documented elsewhere; for users without access to the DE extraction software it must be created explicitly.

The information file holds a row of data for each variable "group":

groupName noOfItems item1 item2 ... itemN

groupName is the name given to a variable or group of variables. For most variables, this will just be the name of the variable. However, it may be convenient to refer to variables in groups; for example, it would usually make sense to treat occupation dummies as one group, area dummies as another, and so on. The program MetaProg automatically groups dummy and interactive variables together, leaving continuous variables ungrouped.

noOfItems refers to the number of elements in that group. This will obviously be one for ungrouped variables. This is followed by a list of individual names for each item in a group: *item1..itemN*. These names should be character elements, and not numbers; if numbers are used to identify group items, then they should be converted to their character equivalents before the information matrix is saved. MetaProg will generate item names for dummy variables from the breakpoints given; for interactive variables, the names are constructed from the names of individual elements.

For example, suppose the variables in the cross-product matrix consisted of the continuous variables

constant wage logwage hours age ageSq

the dummy variables

industry	with categories	0 1 2 3 4 5 6 7 8 9
newJob	"	1 2
basis	"	1 2 3 4

and the interactive variable

jobAge with elements age_1 age_2

Then the information matrix should contain:

"constant"	1	"constant"									
"wage"	1	"wage"									
"logwage"	1	"logwage"									
"hours"	1	"hours"									
"age"	1	"age"									
"ageSq"	1	"ageSq"									
"industry"	10	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
"newJob"	2	"1"	"2"								
"basis"	4	"1"	"2"	"3"							
"jobAge"	2	"age_1"	"age_2"								

The information in the empty cells is not used and so is irrelevant. Note that all cells bar those in the second column contain *character data*.

Suppose it was convenient to treat the age variables as one block. Then this could be rearranged as

"constant"	1	"constant"									
"wage"	1	"wage"									
"logwage"	1	"logwage"									
"hours"	1	"hours"									
"age"	2	"age"	"agesq"								
"industry"	10	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
"newJob"	2	"1"	"2"								
"basis"	4	"1"	"2"	"3"							
"jobAge"	2	"age_1"	"age_2"								

Clearly the order of the variables may not be altered.

3. Data/model/estimator combinations

The following combinations of data matrix, model and estimator are feasible:

Matrix type	TVCS OLS	TVCS covariance	TVFE unbalanced differenced	TVFE balanced differenced	TVFE covariance	MD
Raw X matrix	_	_			_	_
Small cross-product (TKxK)	_	_				_
Small cross-product (TKxK) with means matrix	_	_			_	
Full-size (TKxTK) unbalanced	_	_				_
Full-size (TKxTK) unbalanced with means matrix	_	_			_	
Full-size (TKxTK) balanced	_	_		_	_	
Small cross-product (TKxK) pre-differenced			_			

Unless the dataset is pre-differenced, the TVCS can be estimated from all datasets. Note that balanced differenced TVFE estimator is consistent but not unbiased, due to autocorrelation introduced by the differencing; this is *not* pointed out in Ritchie(1995)! The unbalanced TVFE on the pre-differenced dataset is unbiased and consistent as each period's parameters are estimated independently. However, when coefficients are pooled, then the bias reappears because of the joint estimation.

4. Running the program

These notes should be read whilst running the program. It saves me typing in all the questions you get asked...

Getting the data

The program first asks several questions to ascertain the type of input matrix (and hence the models available). The "standard" matrix is the small (TKxK) cross-product matrix. It then prints out the size and type of the matrix it is going to use, plus the number of "periods".

If a cross-product matrix and information file are given as inputs, the program checks for the internal consistency; that is, do the number of variables in the information file tie in with the number of rows/columns in the data matrix? The program will not continue unless the information file is consistent with the data.

It also asks for the name of the output file. All the results will be sent to a file with this name, with the extension ".RES" added to the end. All the coefficient tables will also be saved as GAUSS matrices, again using the output file name given. This matrix can then be analysed at leisure. A program called XPOutFmt.GP formats the coefficient tables into a comma-delimited format suitable for spreadsheet imports.

Choosing estimators and variables

The program first asks for the estimation method. This is one of the five above. Pre-differenced datasets are estimated by the TVCS estimators, as the fixed-effect has already been removed. The minimum distance estimator is not implemented at the moment. If anyone desperately wants it, I shall get round to it!

The program then presents a selection of variables. Those variables which have been designated groups in the information matrix may be selected as a group. The first time an estimate is made on a dataset, you will be asked to identify the constant. The select the dependent and explanatory variables (and the constant), enter the number of the variables you wish to use, not the name. Instruments may be selected in the same manner as variables; if no instruments are selected (or the instruments do not differ from the variables) then OLS estimation will be used. Typing "all" or "prev" will result in all variables or the previously selected list being used.

You need to select the periods to be used. These are numbered from 1 to the maximum. The maximum number of periods is determined automatically from the dataset.

Although the default is that all coefficients are TV, you may restrict subsets of the variables. Note that this does affect the TVCS estimates. With all coefficients varying, TVCS estimates are effectively a set of T separate regressions; with some pooled coefficients, all T periods need to be estimated jointly. This affects the calculation of the error term, as the error term is now assumed to be homoscedastic over all periods. This is also the assumption made by the TVFE, and in the pooled and restricted models to be discussed later. This does not affect the estimates; however, it does affect the T-statistics, and error terms. The reason is simply lack of interest on my part; again, if there is sufficient demand for homoscedastic errors, I shall investigate the feasiblity of heteroscedastic errors.

If you have selected the TVFE estimator on an unbalanced panel, you will also be asked for the name of the means matrix. If this matrix is not consistent with the estimation period you have selected, you will not be allowed to continue.

Finally, you will be presented with a list of variables and asked if you want to drop any. This is to allow individual variables to be removed where the variables are arranged in groups. Note that a variable deleted in one period will be deleted in all periods.

Singularity and multicollinearity

XPReg carries out tests on the selection of variables to ensure that estimation is possible. It tests for the singularity of the matrix; if the matrix is singular, you will be asked to delete variables. Estimation cannot proceed unless the matrix is non-singular. Variables which are zero in all years will be deleted

automatically.

The multicollinearity test comes from Greene (1990, p280). If the X'X matrix shows a high degree of collinearity, estimation can proceed although the error variance may be large.

Note that a variable deleted in one period will be deleted in all periods.

Other tests

Several tests are available, which depend on the estimator and dataset. The residuals test option is only available for raw X matrices, but does nothing at the moment anyway. The "variance analysis" lists the amount of the variance attributable to variables' own- and cross-correlations; this produces an enormous amount of output if the number of variable groups is large. The "pooling" and "restricted" options are available for the covariance estimators. The "pooled" model forces all parameters to be constant over time; the "restricted" model allows for time dummies. The "unrestricted" model is the default, TV, model. F-tests test the restrictions, along with the restriction on pooling subsets of the coefficients.

Output

Output should be fairly self-explanatory. Output is organised with blocks of coefficients printed for each period. These blocks are saved in the .FMT file mentioned earlier. Regressions statistics are printed after each regression; note that the TVFE and TVCS with pooled parameters are estimated jointly for all periods; for the normal TVCS T separate estimates are carried out.

If the option to save the covariance matrices was selected, you will be asked for the name of the save file (.fmt) to be used. *Do not* include the .fmt extension.

5. XPOutFmt

This program takes the .fmt saved by the program and writes it out as an ASCII text file, with quote marks around the variable names and commas separating the numbers. This can then be imported into spreadsheets using as a "comma/quote delimited" text file. When several blocks of coefficients have been saved with the same number of variables, the program will print out all the means, then all the coefficients, then all the errors, and so on. This is to save you reformatting, for example, TVFE output.

XpOutFmt.GP asks for the name of the output file (with no extension). From this file name it will create a new file with the extension ".PRN". Thus entering *extract1* as the file name would lead to a new file *extract1.PRN* being created. The .PRN extension is added by the program to make searching in Quattro/Lotus easier.

6. XPToQP

This program takes a cross-product matrix and rewrites it in a format suitable for QuattroPro or Lotus-123 to read (although there are severe limits on the number of column that can be read by Lotus v2.2 and non-Windows QuattroPro). XpToQp.GP reads in the cross-product matrix, adds row and column headings, and writes out the matrix, plus headings, in an ASCII file with strings surrounded by quote marks (") and numbers separated by commas. This can be read using the Quattro Tools/Import/Comma-delimited or Lotus File/Import/Numbers commands.

XpToQp.GP asks for the name of the data file and the associated information file. From the data file name it will create a new file with the extension ".PRN". Thus entering *extract1* as the file name would lead to a new file *extract1.PRN* being created. The .PRN extension is added by the program to make searching in Quattro/Lotus easier.

XpToQP attaches row and column headers to the cross-product matrix and then prints out the result.

The resulting ASCII file is usually around 25% larger than the original GAUSS matrix file. If the matrix contains several sub-matrices (ie different periods), then XpToQp will add an extra row and column header indicating that this is submatrix 1, 2, ... and so on.

7. MakeXX

This converts raw X matrices into cross-product matrices. It will also construct means matrices, asking which periods you want to create means for. The routines in this program are the same as those used in XPReg.

References

Bell, DNF and FJ Ritchie (1996) *Time-varying parameters in panel models*, University of Stirling Discussion Paper 96/11 March

Greene, WH (1990) *Econometric Analysis*, Philip Allan

Ritchie, FJ (1995) *Efficient access to large datasets for linear regression models*, University of Stirling Discussion Paper 95/12 December