Paper 354-2011

Proc Panel vs. a Fast Algorithm for Estimation of Unbalanced Panel Data Models with Two Fixed Effects

Alan Ricardo da Silva, Universidade de Brasília, Dep. de Estatística, Brazil
Mariana Fernandes Teixeira, Universidade de Brasília, Dep. de Estatística, Brazil
Thaís Helena Fernandes Teixeira, Universidade de Brasília, Dep. de Estatística, Brazil

ABSTRACT

The panel data models are becoming more common in relation to cross-section and time-series models due to innumerable advantages, in addition to the computational advance that facilitates their utilization. This work aimed at presenting a faster general algorithm for estimation of panel data models with two fixed effects, when it is compared with PROC PANEL, working with balanced or unbalanced data. The results by simulation showed that until with a large data base (50,000 observations) only a few seconds were necessary for estimation, while in the PROC PANEL, the estimation wasn't possible due to out of memory.

1. INTRODUCTION

The panel data models are being widely used, amongst other reasons, due to the computational advancement. These models present advantages in relation to cross-section and time series models for, in addition to increasing the degrees of freedom, they manage to remove the influence of the individual and/or the time of the independent variables, thus making the estimates of model coefficients more realistic.

It is possible to estimate the model considering the "individual" and "time" effects as being fixed or random. Depending on the problem, one or both of the effects will be able to compose the cited model (The One-Way / Two-Way Fixed Effects or The One-Way / Two-Way Random Effects). In SAS 9.2, the PROC PANEL is responsible for this job. However, when the data are unbalanced (or incomplete), i.e, some observations are not present in the entire sample period, the estimation is too slow or even not possible in some cases.

This paper is aimed at presenting a more efficient computational algorithm developed using the IML module of the software SAS 9.2, to the estimation of the panel data models with two fixed effects, when the data are balanced or unbalanced. The paper is organized in the following manner: Section 2 introduces the unbalanced panel data models that will be used in the study. Section 3 presents the developed algorithm in SAS/IML and a case study to evaluate the potential of the program is presented in Section 4. Finally, the main conclusions about the study are presented.

2. UNBALANCED PANEL DATA MODELS

As described by Hsiao (1986), there are many benefits in using panel data, being the principal ones: control of the individual heterogeneity; panel data models have a greater variability, less colinearity between variables, more degrees of freedom and more efficiency; they are more capable to identify and measure effects that aren't detected in cross-section or time series data. However, these models also have some limitations, specially the requirement of more computational resources.

The general form of the balanced panel data model is shown in the Equation 1.

$$\begin{aligned} y_{it} &= \alpha + \beta X_{it} + \mu_{it} \qquad i = 1,\dots,N; \quad t = 1,\dots,T \\ \mu_{it} &= \mu_i + \lambda_t + v_{it} \end{aligned} \tag{1}$$

where *i* denotes the effect of the individual, the firm, etc (for simplicity, it will be referenced here as "individual"), and *t* denotes the time; y_{it} is the value of the dependent variable of the *i*-th individual in time *t*; α is the constant of the model; β is a vector (k x 1) of the coefficients; X_{it} is a matrix of the independent variables of the *i*-th individual in time *t*; and μ_{it} is the disturbance of the *i*-th individual in time *t*.

The disturbance term, in turn, is composed by some components called variance components: μ_i, λ_t and v_{it}. These represent, respectively, the non-observable effect specific to the individual *i*, the non-observable effect specific to time *t* and the remainder disturbance. To verify if such components are fixed effects (they influence directly the dependent variable) or random (they have correlation with the independent variables), specific tests are necessary. These tests will not be detailed here, but they can be found in Baltagi (2002), Greene (2002) and Hsiao (1986).

For the estimation of the parameters of model with two fixed effects in a balanced panel, *N-1* dummies referred to the “individuals” and *T-1* dummies referred to the “time”, or a matrix **Q*** (Within Transformation) can be used (Equation 2), which eliminates the necessity of dummies. The main difference between these two forms is that when the matrix **Q*** is used, in addition to the faster process, the intercept of the model considers the influence of all the individuals. Here,

$$\mathbf{Q}^* = \left(\mathbf{I}(T) - \frac{\mathbf{ee}'}{T} \right) \otimes \left(\mathbf{I}(N) - \frac{\mathbf{ll}'}{N} \right) \tag{2}$$

where $\mathbf{I}(T)$ is an identity matrix of dimension *T*, **e** is a vector of 1’s also of dimension *T*, $\mathbf{l}$ represents a vector of 1’s of dimension *N* and $\otimes$ represents the Kronecker product, i.e., the matrix **Q*** will have a final dimension of (*NT* x *NT)*. Thus, the vector of the parameters $\boldsymbol{\beta}$ referring to the independent variables is estimated by

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q}^* \mathbf{X}_i \right]^{-1} \left[\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q}^* \mathbf{y}_i \right] \tag{3}$$

where $\mathbf{X}$ is the matrix of the independent variables and $\mathbf{y}$ is the vector of the dependent variable. May be remarked that the matrix $\mathbf{X}$ may not contain the vector of 1’s. Therefore, when including the matrix **Q*** in the estimation of $\boldsymbol{\beta}$, the intercept is nulled. This way, after the estimation, the intercept of the model can be calculated by $\alpha = \bar{y} - \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}$, where $\bar{y}$ and $\bar{\mathbf{x}}$ are the averages of the dependent and independent variables, respectively. In this case, the parameters λ_t about “time” can be calculated by

$$\hat{\lambda}_t = (\bar{\mathbf{y}}_{.t} - \bar{y}_{..}) - (\bar{\mathbf{x}}_{.t} - \bar{\mathbf{x}}_{..})\hat{\boldsymbol{\beta}} \tag{4}$$

where $\bar{\mathbf{y}}_{.t}$ represents the average of $\mathbf{y}$ in time *t*, $\bar{y}_{..}$ represents the average of $\mathbf{y}$, $\bar{\mathbf{x}}_{.t}$ represents the vector (1 x k) of the averages of $\mathbf{X}$ in time *t* and $\bar{\mathbf{x}}_{..}$ represents the vector (1 x k) of the averages of $\mathbf{X}$. The parameters μ_i about “individual” can be calculated by

$$\hat{\mu}_i = (\bar{\mathbf{y}}_{i.} - \bar{y}_{..}) - (\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{..})\hat{\boldsymbol{\beta}} \tag{5}$$

where: $\bar{\mathbf{y}}_{i.}$ represents the average of $\mathbf{y}$ of the individual *i*, $\bar{y}_{..}$ represents the average of $\mathbf{y}$, $\bar{\mathbf{x}}_{i.}$ represents the vector (1 x k) of the averages of $\mathbf{X}$ of the individual *i* and $\bar{\mathbf{x}}_{..}$ represents the vector (1 x k) of the averages of $\mathbf{X}$.

Generally, incomplete panels are more likely to be the norm in typical economic empirical setting, because sometimes some observations can no longer be included in the database. For example, in collecting data on US airlines over time, a researcher may find that some firms have dropped out of the market while new entrants emerged over the sample period observed (Baltagi, 2002). Wansbeek and Kapteyn (1989) consider the regression model with unbalanced two-way error components disturbance

$$\begin{aligned} y_{it} &= \alpha + \beta X_{it} + \mu_{it} \qquad i = 1, \ldots, N_t; \qquad t = 1, \ldots, T \\ \mu_{it} &= \mu_i + \lambda_t + v_{it} \end{aligned} \tag{6}$$

where: *N*N_t ($N_t \le N$) denotes the number of individuals observed in the year *t*, with $n = \sum_t N_t$. Let $\mathbf{D}_t$ be the $(N_t \times N)$ matrix obtained from $\mathbf{I}_N$ by omitting the rows corresponding to individuals not observed in year *t*. Define

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2) = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_1\mathbf{1} & & \\ \vdots & & \ddots & \\ \mathbf{D}_\mathbf{T} & & & \mathbf{D}_\mathbf{T}\mathbf{1} \end{bmatrix} \tag{7}$$

where: $\mathbf{\Delta_1} = (\mathbf{D}_1^{'}, ..., \mathbf{D}_T^{'})'$ is (*n* x *N*) and $\mathbf{\Delta_2} = diag(\mathbf{D_T 1}) = diag(\mathbf{1}_{\mathbf{N_t}})$ is (*n* x *T*) and $\mathbf{1}_{\mathbf{N_t}}$ is $(N_t \times 1)$ vector of 1's. The matrix $\mathbf{\Delta}$ gives the dummy variable structure for the unbalanced panel model. For balanced panels, $\mathbf{\Delta_1} = \mathbf{e} \otimes \mathbf{I_N}$ and $\mathbf{\Delta_2} = \mathbf{I_T} \otimes \mathbf{1}$.

The Within Transformation for the unbalanced data case is a little complicated but nevertheless manageable (Wansbeek and Kapteyn, 1989). Note that $\mathbf{\Delta_N} = \mathbf{\Delta_1 \Delta_1}' = diag(\mathbf{T_i})$, where $\mathbf{T_i}$ is the number of years individual *i* is observed in the panel. Also, $\mathbf{\Delta_T} = \mathbf{\Delta}_2\mathbf{\Delta}_2' = diag(\mathbf{N_t})$ and $\mathbf{\Delta_{TN}} = \mathbf{\Delta}_2'\mathbf{\Delta}_1$ is the (*T* x *N*) matrix of zeros and ones indicating the absence or presence of an individual in a certain year. For balanced panels, $\mathbf{\Delta_N} = T\mathbf{I_N}$, $\mathbf{\Delta_T} = N\mathbf{I_T}$ and $\mathbf{\Delta_{TN}} = \mathbf{e1}'$. Define $\mathbf{P_{[\Delta]}} = \mathbf{\Delta(\Delta'\Delta)^{-1}\Delta'}$, then the Within Transformation is $\mathbf{Q^*_{[\Delta]}} = \mathbf{I_n} - \mathbf{P_{[\Delta]}}$. For the two fixed effects model in an unbalanced panel with $\mathbf{\Delta} = (\mathbf{\Delta_1}, \mathbf{\Delta_2})$, it can be shown that

$$\mathbf{P_{[\Delta]}} = \mathbf{P}_{[\Delta_1]} + \mathbf{P}_{[\mathbf{Q^*_{\Delta_1}\Delta_2}]} \tag{8}$$

and

$$Q^*_{[\mathbf{\Delta}]} = Q^*_{[\mathbf{\Delta_1}]} - Q^*_{[\Delta_1]}\Delta_2(\Delta_2^{'}Q^*_{[\Delta_1]}\Delta_2)^{-1}\Delta_2^{'}Q^*_{[\mathbf{\Delta_1}]} \tag{9}$$

However, $\mathbf{(\Delta'\Delta)^{-1}}$ is singular, i.e. $|\mathbf{\Delta'\Delta}| = 0$. Davis (2002) showed that:

Lemma 1. If A is (*n* x *m*) and B is (*m* x *n*), then $\mathbf{(I_n + AB)^{-1} = I_n - A(I_m + BA)^{-1}B}$, provided $|\mathbf{I_m + BA}| \neq 0$. Thus, $\mathbf{(I_n + XX')^{-1} = I_n - X(I_n + X'X)^{-1}X}$.

But depending on the magnitude of the matrix **X**, adding one to the identity matrix $\mathbf{I_n}$ may be too much. For example, adding one to 150 is significantly different from adding one to 0.300. To avoid this problem, the term $\mathbf{I_n}/n^n$ instead $\mathbf{I_n}$ can be used just to eliminate the singularity, i. e., $\mathbf{(I_n + XX')^{-1}} = \mathbf{I_n - X(I_n}/n^n \mathbf{+ X'X)^{-1}X}$.

3. FAST ALGORITHM

The algorithm was developed using the same idea of Silva and Alves (2007). Briefly, the authors showed that working with matrices of dimension (*T* x *T*), computing N^2-times is more efficient from the computational point of view than working with matrices with dimension *NT* x *NT*, generating the same results. As we have seen before, Wansbeek and Kapteyn (1989) presented a method to estimate the parameters for the unbalanced case, which use the data ordered on the *N* individuals in *T* consecutive sets, so that *t* runs slowly and *i* runs fast. This is exactly the opposite ordering used for the balanced case and for Silva and Alves (2007). Because of that, the processing time for estimating the unbalanced two-way model is higher than the balanced two-way model.

So, the idea of the fast algorithm is to use the data ordered on the *T* individuals in *N* consecutive sets, so that the unbalanced two-way model is estimated by the unbalanced one-way model with dummy variables for time. The matrix **Q** used on the unbalanced one-way model is:

$$\mathbf{Q} = \mathbf{I}(T) - \frac{\mathbf{ee}'}{T} \tag{10}$$

The dummy variables are easily created as follows (where TT is the number of years):

```
dummyT=j(nrow(y),TT-1,0);
  do j=1 to TT-1;
    do i=1 to nrow(y);
           if &year[i]=_year_[j] then dummyT[i,j]=1;
    end;
  end;
x=x||dummyT;
```

The main difference between the unbalanced and the balanced cases is the number of years of each individual. Hence, it is necessary to compute a vector with those frequencies as follows:

```
_freq_=j(NN,1,0);
  _n_=j(nrow(&cross),1,0);
  _n_[1]=1;
    do k=2 to nrow(&cross);
    if &cross[k]=&cross[k-1] then _n_[k]=_n_[k-1];
                             else _n_[k]=_n_[k-1]+1;
    end;
    do f=1 to NN;
      do h=1 to nrow(&cross)-1;
        if f=_n_[h] then do;
          if &cross[h+1] = &cross[h] then _freq_[f]=_freq_[f]+1;
                                     else _freq_[f]=_freq_[f]+1;
          if h=nrow(&cross)-1 then _freq_[f]=_freq_[f]+1;
        end;
      end;
    end;
```

If the data are balanced then the vector is simply `_freq_=TT`. Finally, the parameters are estimated as in (3).

```
b1=j(nvar,nvar,0);
b2=j(nvar,1,0);
l2=0;
do i=1 to NN;
  if NT ^= _TOTAL_ then _dim_=_freq_[i];
                   else _dim_=_freq_[1];
  if _dim_ ^= 0 then do;
    e=j(_dim_,_dim_,1/_dim_);
    Qe=(I(_dim_)-e);
    l1=l2+1;
    l2=l1+_dim_-1;
    b1=b1+x[l1:l2,]`*Qe*x[l1:l2,];
    b2=b2+x[l1:l2,]`*Qe*y[l1:l2];
  end;
end;
```

In the next section, a case study will be presented to evaluate the potential of the algorithm, as well as to verify the differences with the PROC PANEL and the processing time.

4. CASE STUDY

To evaluate the potential of the algorithm, a database has been created, by simulation, from 100 to 10,000 individuals with 5 times, totalizing 50,000 observations. The computer used for processing was a PC CORE™ 2 DUO CPU E7300 2.66 Ghz with 2 GB of RAM memory and 150 GB of hard disk. The explicative variables have been created by random numbers of Binomial (6;0.5) and Normal (0,1) and the depended variable by a linear combination of the explicative variables. The unbalanced dataset was created by PROC SURVEYSELECT using Simple Random Sampling without replacement, with 75% of data in the first time period, 56% of data in the second time period, 90% of data in the third time period, 80% of data in the fourth time period and 95% of data in the fifth time period. The results of processing times using the fast algorithm are in the Figures 1 and 2.

It is interesting to note that with balanced data, the greater processing time was less than one second, while with unbalanced data the greater processing time was less than two minutes. The processing times of model estimated by PROC PANEL are given in the Figures 3 and 4. Note that for balanced case, from 26,000 observations (or 5,200 individuals in 5 times) the parameters were not estimated due to out of memory. In the unbalanced case, from 18,000 observations (or approximately 4,600 individuals in 5 times) the parameters were not estimated also due to out of memory.

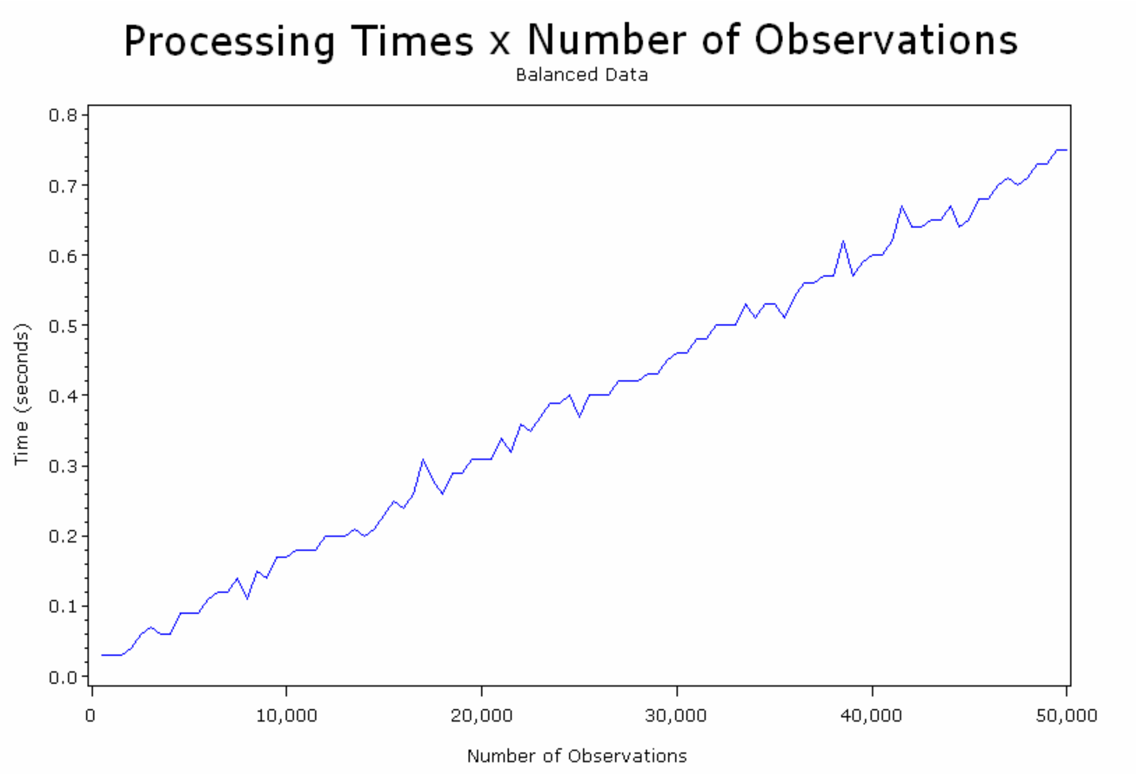

Figure 1. Number of observations vs. processing times in balanced two-way model estimated by FAST ALGORITHM.

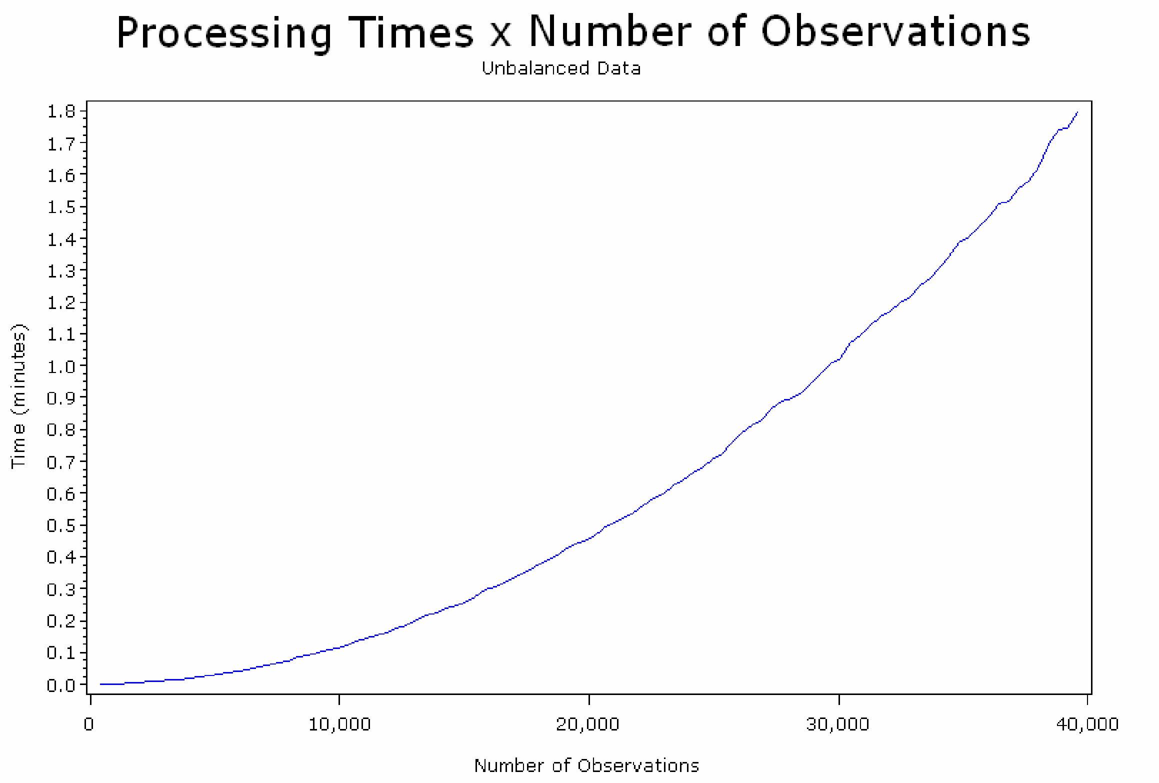

Figure 2. Number of observations vs. processing times in unbalanced two-way model estimated by FAST ALGORITHM.

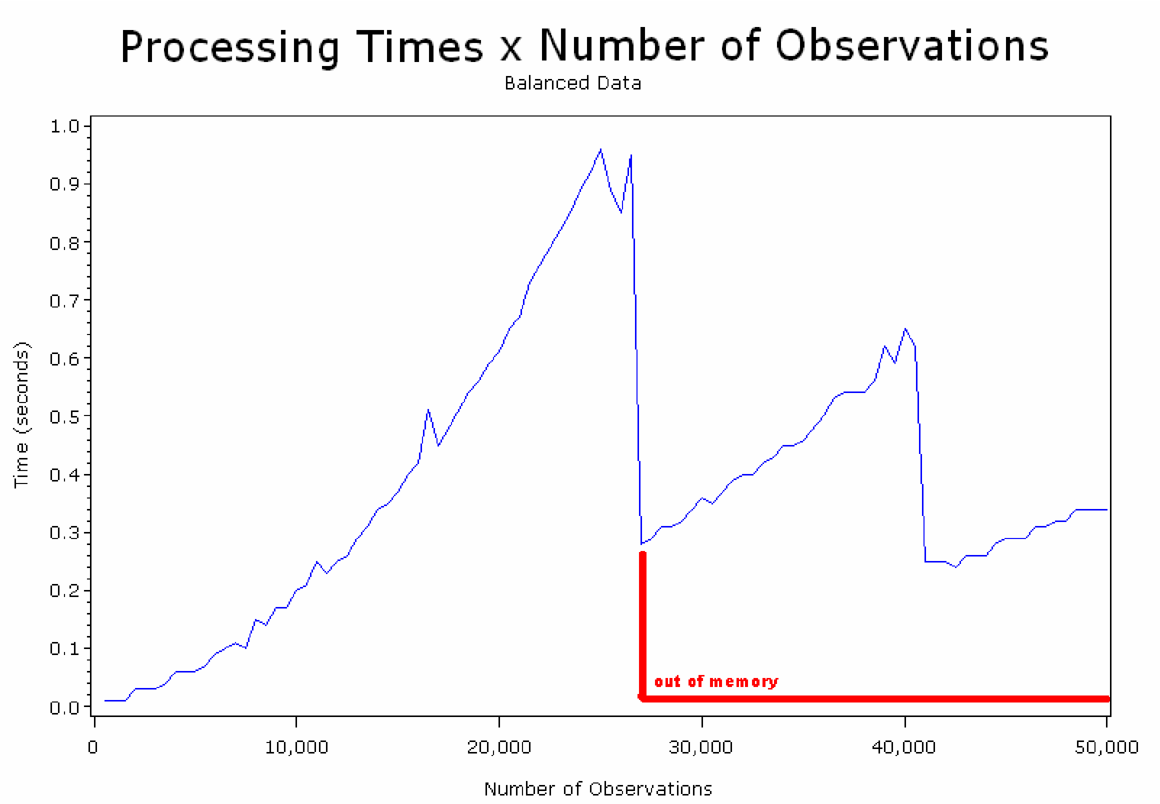

Figure 3. Number of observations vs. processing times in balanced two-way model estimated by PROC PANEL.

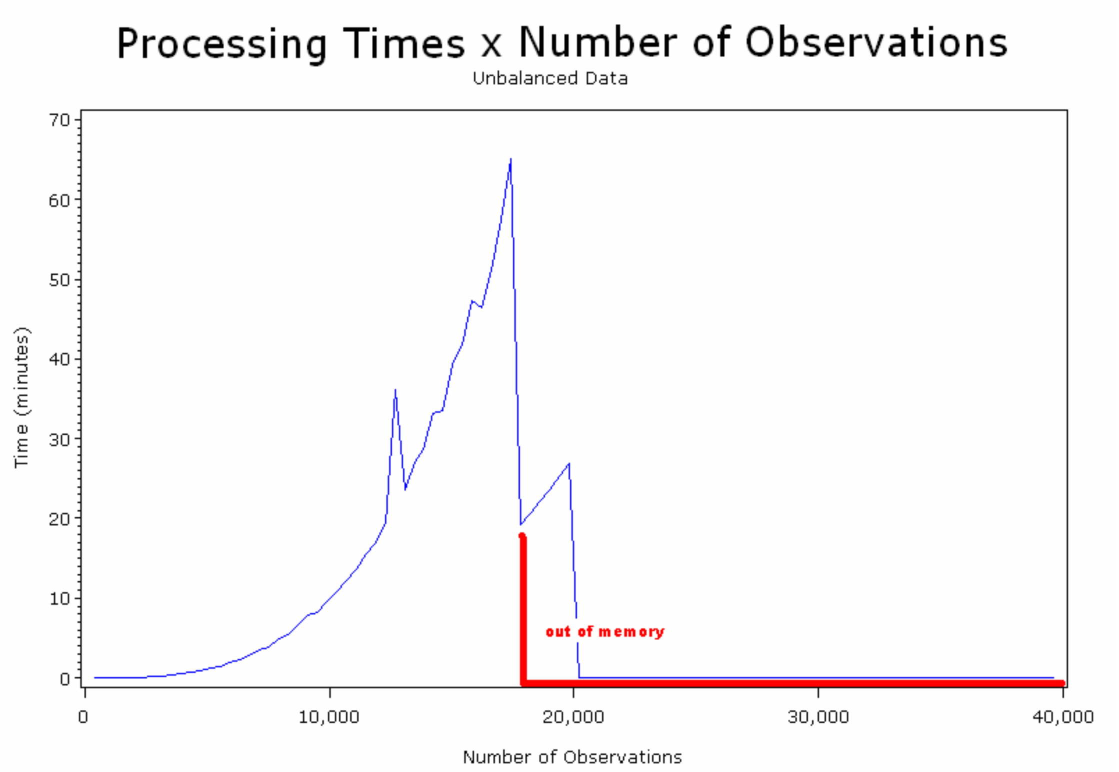

Figure 4. Number of observations vs. processing times in unbalanced two-way model estimated by PROC PANEL.

CONCLUSION

The algorithm described here is really very fast than one used by PROC PANEL. The main problem of the algorithm used by PROC PANEL is the estimation by LSDV (Least Square Dummy Variables) method, i. e., a dummy variable is created for each individual and for each time. So, it is necessary to invert a matrix of dimension $(k+NT)\times(k+NT)$, while the fast algorithm described here invert a matrix of dimension $(k+T)\times(k+T)$, generating the same results. It is interesting to point out that the method described by Wansbeek and Kapteyn (1989) was not used because its main objective is to estimate the parameters of the model without dummy variables. However, Silva and Alves (2007) showed the saving of computational time when the algorithm with the matrix **Q** and dummy variables for the time were used. So, for the estimation of the unbalanced panel data model with two fixed effects it is necessary only to compute the vector with the frequencies of the number of years of each individual and to use the algorithm proposed by Silva and Alves (2007).

REFERENCES

BALTAGI, B. H., *Econometric Analysis of Panel Data*. Wiley, 3rd ed, 2002.

DAVIS, P., *Estimating Multi-way Error Components Models with Unbalanced data Structures Using Instrumental Variables.* Journal of Econometrics (106), p. 67-95, 2002.

GREENE, W. H., *Econometric Analysis*. Pearson Education, 5th ed, 2002.

HSIAO, C., *Analysis of Panel Data*. Econometric Society Monographs. New York: Cambridge University Press, 1986.

SAS, *SAS OnlineDoc*. http://support.sas.com/onlinedoc/913/docMainpage.jsp, 2008.

SILVA. A. R. and ALVES. P. F., *Computational Algorithm for Estimation of Panel Data Models with Two Fixed Effects.* Rev. Mat. Estat., São Paulo, v.25, n.2, p.19-32, Brazil, 2007.

WANSBEEK, T. and KAPTEYN, A., *Estimation of the Error Components Model with Incomplete Panels*. Journal of *Econometrics, 41, p. 341-361, 1989.*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Alan Ricardo da Silva
Enterprise: Universidade de Brasília
Address: Campus Universitário Darcy Ribeiro, Departamento de Estatística, ICC, CSS 361/31
City, State ZIP: Brasília, DF, Brazil, 70910-900
Work Phone: +5561 3107 6756
Fax: +5561 3107 6768
E-mail: alansilva@unb.br
Web: www.unb.br/ie/est

Name: Mariana Fernandes Teixeira and Thaís Helena Fernandes Teixeira
Enterprise: Universidade de Brasília
Address: Campus Universitário Darcy Ribeiro, Departamento de Estatística
City, State ZIP: Brasília, DF, Brazil, 70910-900
Work Phone: +5561 9292 3380 and +5561 9293 0921
E-mail: mari.f.t@hotmail.com and thaishft@hotmail.com.
Web: www.unb.br/ie/est

APPENDIX I – SAS® ALGORITHM

```
%macro tscsreg(tab=,y=,x=,cross=,year=);
/******************************************************/
/* Written By Alan Silva, Mariana Teixeira and Thais  */
/* Teixeira (09/2010)                                 */
/* E-mail: alansilva@unb.br; mari.f.t@hotmail.com;    */
/*          thaishft@hotmail.com                     */
/* References:                                       */
/* [1]HSIAO,C.(1986). "ANALYSIS OF PANEL DATA".      */
/*    NEW YORK: CAMBRIDGE UNIVERSITY PRESS           */
/*    ECONOMETRIC SOCIETY MONOGRAPHS                 */
/* [2]GREENE,W.H.(2002). "ECONOMETRIC ANALYSIS".     */
/*    PEARSON EDUCATION, 5th ed.                     */
/* [3]BALTAGI,C.H.(2002). "ECONOMETRIC ANALYSIS      */
/*    OF PANEL DATA". WILEY, 3rd ed.                 */
/******************************************************/
proc sort data=&tab;by &cross &year;run;
proc iml;
use &tab;
read all var {&y} into y;
read all var {&x} into x;
read all var {&cross &year};
nomes={Intercept &x}`;
NN=ncol(unique(&cross));
TT=ncol(unique(&year));
NT=NN*TT;
_TOTAL_=nrow(y);
if NT ^= _TOTAL_ then print "Unbalanced Panel";
                  else print "Balanced Panel";
nvar2=ncol(x);
_year_=unique(&year);
dummyT=j(nrow(y),TT-1,0);
do j=1 to TT-1;
  do i=1 to nrow(y);
    if &year[i]=_year_[j] then dummyT[i,j]=1;
  end;
end;
x=x||dummyT;
_year_=_year_`;
nvar=ncol(x);

/*********** computing the frequency by individual *******/
if NT ^= _TOTAL_ then do;
  _freq_=j(NN,1,0);
  _n_=j(nrow(&cross),1,0);
  _n_[1]=1;
    do k=2 to nrow(&cross);
      if &cross[k]=&cross[k-1] then _n_[k]=_n_[k-1];
                                else _n_[k]=_n_[k-1]+1;
    end;
    do f=1 to NN;
      do h=1 to nrow(&cross)-1;
        if f=_n_[h] then do;
          if &cross[h+1] = &cross[h] then _freq_[f]=_freq_[f]+1;
                                     else _freq_[f]=_freq_[f]+1;
          if h=nrow(&cross)-1 then _freq_[f]=_freq_[f]+1;
        end;
      end;
    end;
end;
                 else do;
                   _freq_=TT;
                 end;
/****************************************************/
```

```
b1=j(nvar,nvar,0);
b2=j(nvar,1,0);
l2=0;
do i=1 to NN;
  if NT ^= _TOTAL_ then _dim_=_freq_[i];
                     else _dim_=_freq_[1];
  if _dim_ ^= 0 then do;
    e=j(_dim_,_dim_,1/_dim_);
    Qe=(I(_dim_)-e);
    l1=l2+1;
    l2=l1+_dim_-1;
    b1=b1+x[l1:l2,]`*Qe*x[l1:l2,];
    b2=b2+x[l1:l2,]`*Qe*y[l1:l2];
  end;
end;

b=inv(b1)*b2;
yb=sum(y)/_TOTAl_;
xb=j(1,nvar,0);
do i=1 to nvar;
  xb[,i]=sum(x[,i])/_TOTAL_;
end;
b0=yb-xb*b;

rsqr1=0;
rsqr2=0;
l2=0;
do i=1 to NN;
  if NT ^= _TOTAL_ then _dim_=_freq_[i];
                     else _dim_=_freq_[1];
  if _dim_ ^= 0 then do;
    e=j(_dim_,1,1);
    Qe=I(_dim_)-(e*e`)/_dim_;
    l1=l2+1;
    l2=l1+_dim_-1;
    rsqr1=rsqr1+((y[l1:l2,]-e*b0-x[l1:l2,]*b)`*Qe*(y[l1:l2,]-e*b0-x[l1:l2,]*b));
    rsqr2=rsqr2+y[l1:l2]`*Qe*y[l1:l2];
  end;
end;
sigma2u=rsqr1/(_TOTAL_-NN-nvar);
varb=sigma2u*inv(b1);
stdb=sqrt(vecdiag(abs(varb)));
b=b[1:nvar2];
stdb=stdb[1:nvar2];
x=x[,1:nvar2];
nvar=ncol(x);
yb=sum(y)/_TOTAl_;
xb=j(1,nvar,0);
do i=1 to nvar;
  xb[,i]=sum(x[,i])/_TOTAL_;
end;
b0=yb-xb*b;
stdb0=sqrt(abs(sigma2u/(_TOTAL_)+xb*varb[1:nvar,1:nvar]*xb`));
b=b0//b;
tmp=stdb0//stdb;
tstat=b/tmp;
probt=2*(1-probt(abs(tstat),_TOTAL_-2));
param=b||tmp;
rsqr=1-rsqr1/rsqr2;

d1={"Estimation Method" "Number of Cross Sections "
"Number of Time Series" "Number of Observations"}`;
d2={"FixTwo"}//char(NN,6)//char(TT,6)//char(_TOTAL_,6);
print "Dependent Variable: &y";
print "Model Description",, d1[label=''] d2[label=''];
est={'Estimate' 'Std. Error'};
f1={"SSE" "MSE" "R-Square   "}`;
```

```
f2=char(sigma2u*(_TOTAL_-NN-TT+1-nvar),10)//char(sigma2u,10)//char(rsqr,10);
f11={"DFE" "Root MSE  "}`;
f22=char(_TOTAL_-NN-TT+1-nvar,6)//char(sqrt(sigma2u),6);
print "Fit Statistics",, f1[label=''] f2[label=''] {'       '}f11[label='']
f22[label=''];
print,"Parameter Estimates",, nomes[label='Variable']param[format=comma10.5
colname=est label='']
tstat[format=comma10.2 colname='   t Value' label='']
probt[format=pvalue6.4 colname='P > |t|' label=''];

/* F Test*/
ybi=j(NN,1,0);
l2=0;
do i=1 to NN;
  if NT ^= _TOTAL_ then _dim_=_freq_[i];
                     else _dim_=_freq_[1];
  l1=l2+1;
  l2=l1+_dim_-1;
  ybi[i]=(sum(y[l1:l2])/_dim_);
end;

xbi=j(NN,nvar,0);
do j=1 to nvar;
  l2=0;
  do i=1 to NN;
    if NT ^= _TOTAL_ then _dim_=_freq_[i];
                       else _dim_=_freq_[1];
    l1=l2+1;
    l2=l1+_dim_-1;
    xbi[i,j]=sum(x[l1:l2,j])/(_dim_);
  end;
end;
close &tab;
sort &tab out=_&tab by &year &cross;
use _&tab;
read all var {&y} into y_;
read all var {&x} into x_;
read all var {&year} into t_;
nvar=ncol(x_);

/*********** computing the frequency by time *******/
_t_=j(nrow(t_),1,0);
_t_[1]=1;
do k=2 to nrow(t_);
  if t_[k]=t_[k-1] then _t_[k]=_t_[k-1];
                   else _t_[k]=_t_[k-1]+1;
end;
if NT ^= _TOTAL_ then do;
  _freqt_=j(TT,1,0);
  l2=0;
  do f=1 to TT;
    do h=1 to nrow(t_)-1;
      if f=_t_[h] then do;
        if t_[h+1] = t_[h] then _freqt_[f]=_freqt_[f]+1;
                           else _freqt_[f]=_freqt_[f]+1;
        if h=nrow(t_)-1 then _freqt_[f]=_freqt_[f]+1;
      end;
    end;
  end;
end;
                 else do;
                    _freqt_=NN;
                 end;
/**************************************/

ybt=j(TT,1,0);
xbt=j(TT,nvar,0);
```

```
l2=0;
do j=1 to TT;
  if NT ^= _TOTAL_ then _dimt_=_freqt_[j];
                     else _dimt_=_freqt_[1];
  l1=l2+1;
  l2=l1+_dimt_-1;
  ybt[j]=(sum(y_[l1:l2])/_dimt_);
  do k=1 to nvar;
    xbt[j,k]=sum(x_[l1:l2,k])/_dimt_;
  end;
end;
ym=j(nrow(y),1,0);
xm=j(nrow(x),nvar,0);
ymi=j(nrow(y),1,0);
xmi=j(nrow(x),nvar,0);
ymt=j(nrow(y),1,0);
xmt=j(nrow(x),nvar,0);
l2=0;
do i=1 to NN;
  if NT ^= _TOTAL_ then _dim_=_freq_[i];
                     else _dim_=_freq_[1];
  l1=l2+1;
  l2=l1+_dim_-1;
  t_1=_t_[l1:l2];
  do l=1 to _dim_;
    if l=1 then do;
      ybt_=ybt[t_1[l]];
      xbt_=xbt[t_1[l],];
    end;
           else do;
             ybt_=ybt_//ybt[t_1[l]];
             xbt_=xbt_//xbt[t_1[l],];
           end;
  end;
  ym[l1:l2]=y[l1:l2]-ybi[i]-ybt_+yb;
  ymi[l1:l2]=y[l1:l2]-ybi[i];
  ymt[l1:l2]=y[l1:l2]-ybt_;
  do j=1 to nvar;
    xm[l1:l2,j]=x[l1:l2,j]-xbi[i,j]-xbt_[,j]+xb[j];
    xmi[l1:l2,j]=x[l1:l2,j]-xbi[i,j];
    xmt[l1:l2,j]=x[l1:l2,j]-xbt_[,j];
  end;
end;
Wxxi=j(NN*nvar,nvar,0);
Wxyi=j(NN*nvar,1,0);
Wyyi=j(NN,1,0);
l2=0;
do i=1 to NN;
  if NT ^= _TOTAL_ then _dim_=_freq_[i];
                     else _dim_=_freq_[1];
  l1=l2+1;
  l2=l1+_dim_-1;
  k1=(i-1)*nvar+1;
  k2=i*nvar;
  Wxxi[k1:k2,]=Wxxi[k1:k2,]+xm[l1:l2,]`*xm[l1:l2,];
  Wxyi[k1:k2]=Wxyi[k1:k2]+xm[l1:l2,]`*ym[l1:l2];
  Wyyi[i]=Wyyi[i]+ym[l1:l2]`*ym[l1:l2];
end;
RSS=j(NN,1,0);
do i=1 to NN;
  k1=(i-1)*nvar+1;
  k2=i*nvar;
  if det(Wxxi[k1:k2,])=0 then RSS[i]=0;
                         else RSS[i]=Wyyi[i]-
Wxyi[k1:k2]`*inv(Wxxi[k1:k2,])*Wxyi[k1:k2];
end;
S1=sum(RSS);
```

```
Wxx=j(nvar,nvar,0);
Wxy=j(nvar,1,0);
do j=1 to nvar;
  do i=1 to NN;
    k=j+(nvar*(i-1));
    Wxx[j,]=Wxx[j,]+Wxxi[k,];
    Wxy[j]=Wxy[j]+Wxyi[k];
  end;
end;
Wyy=sum(Wyyi);
S2=Wyy-Wxy`*inv(Wxx)*Wxy;
xmb=j(nrow(x),nvar,0);
do i=1 to nvar;
  xmb[,i]=x[,i]-xb[i];
end;
ymb=y-yb;
Txx=xmb`*xmb;
Txy=xmb`*ymb;
Tyy=ymb`*ymb;
Txxi=xmi`*xmi;
Txyi=xmi`*ymi;
Tyyi=ymi`*ymi;
Txxt=xmt`*xmt;
Txyt=xmt`*ymt;
Tyyt=ymt`*ymt;
S3=Tyy-Txy`*inv(Txx)*Txy;
S3i=Tyyi-Txyi`*inv(Txxi)*Txyi;
S3t=Tyyt-Txyt`*inv(Txxt)*Txyt;
F1=((S2-S1)/((NN+TT-2)*nvar))/(S1/((_TOTAL_-NN-TT+1)*(nvar+1)));
gl11=(NN+TT-2)*nvar;
gl12=((_TOTAL_-NN-TT+1)*(nvar+1));
probf1=1-probf(F1,gl11,gl12);
F4=((S3-S2)/(NN+TT-2))/(S2/((_TOTAL_-NN-TT+1)-nvar));
gl41=(NN+TT-2);
gl42=((_TOTAL_-NN-TT+1)-nvar);
probf4=1-probf(F4,gl41,gl42);
F4i=((S3t-S2)/(NN-1))/(S2/((_TOTAL_-NN-TT+1)-nvar));
gl41i=(NN-1);
probf4i=1-probf(F4i,gl41i,gl42);
F4t=((S3i-S2)/(TT-1))/(S2/((_TOTAL_-NN-TT+1)-nvar));
gl41t=(TT-1);
probf4t=1-probf(F4t,gl41t,gl42);
*Print;
print,"F Test for Homogeneous Bi",,
gl11[label='Num DF'] gl12[label='Den DF'] F1[format=10.2 label='F Value']
probf1[format=pvalue6. label='Prob > F'];
print,"F Test for No Fixed Effect (ai=0 and lt=0)",,
gl41[label='Num DF'] gl42[label='Den DF'] F4[format=10.2 label='F Value']
probf4[format=pvalue6. label='Prob > F'];
print,"F Test for No Fixed Effect (ai=0 and lt ne 0)",,
gl41i[label='Num DF'] gl42[label='Den DF'] F4i[format=10.2 label='F Value']
probf4i[format=pvalue6. label='Prob > F'];
print,"F Test for No Fixed Effect (lt=0 and ai ne 0)",,
gl41t[label='Num DF'] gl42[label='Den DF'] F4t[format=10.2 label='F Value']
probf4t[format=pvalue6. label='Prob > F'];
quit;
%mend tscsreg;
```