Unbalanced Panel Data Estimation Using SAS/IML® Software
Clay Greene, The Consumers' Gas Company Limited, Toronto, CANADA
John Ah-Sen, The Consumers' Gas Company Limited, Toronto, CANADA

ABSTRACT
The most recent release of SAS/ETS® software introduced a procedure to estimate a model with panel data. This procedure requires the time series elements to be equal among all cross-sections. This restriction limits the estimation to a complete, or balanced panel. This paper examines the use of SAS/IML software, in a specific modelling application, to incorporate an unbalanced panel data set. The first section outlines the motivation for the model. The second section shows how SAS/IML software was used to incorporate logic to deal with unbalanced data. The final section states results.

INTRODUCTION
The analysis of panel data has definite empirical advantages. The panel data concept involves data which incorporates cross-section and time series elements. Under SAS/ETS software, Version 6, a procedure was introduced to estimate models using this type of data. The current routines underlyng PROC TSCSREG require the time series elements to be equal among all cross-sections. This is referred to as a 'balanced' panel data set. Circumstances may arise where a panel approach may be desired yet the time series elements for each cross-section are not equal. This case is referred to as an 'unbalanced' panel data set. Many researchers and forecasters would like to model panel data, but must use means other than PROC TSCSREG to complete the estimation in the presence of unbalanced data. In this paper SAS/IML software is used to deal directly with the issue of unbalanced data. This paper presents a given modelling strategy but the ideas used are applicable to a wider range of more general applications.

MODELLING METHODOLOGY
The methodology presented here was used in an energy industry modelling exercise. Consumers Gas is a natural gas distribution company who has undertaken to revise its forecasting approach. Within the Company, the focus of some elements of the forecasting exercise has changed from aggregate forecasts to appliance level forecasts. The problem facing forecasters was to create models to forecast natural gas consumption at the appliance level. The richness of company data allowed the use of two-step Bayesian estimation technique.

The data was organized in a panel framework where each customer represented a cross-sectional element and their data over time comprised the time series counterpart. A 'balanced panel' is hard to come by in company data. If we think about customers, some have had a relationship with the company for over fifty years while new homes are built each year. There may be an opportunity to group specific customers together but, one is still not guaranteed a balanced time series.

The two-step Bayesian approach involved a prior estimation over the appliance level data and a second step conditional demand over billing data. The reason for using a Bayesian approach was to reflect the richness of the appliance level estimates in the billing conditional demand. The prior estimation involved a simple OLS with correction for autocorrelation and heteroskedasticity. The panel for the prior estimation was completed by appliance type and the panel was balanced. The result of the prior estimation over the appliance level data yielded elements to be passed to the second stage.

These elements were;
\[ \hat{p}_i \] autocorrelation coefficient for appliance i
\[ \hat{\sigma}_i^2 \] variance of residuals for appliance i
\[ \hat{\beta} \] estimates of all priors coefficients for all appliances, i to l
\[ \text{cov}(\hat{p}) \] covariances of all priors coefficients

The company's ample Load Research data was key to these prior estimates and was an alternative to engineering estimates for appliance level estimates.

Having completed the prior estimation, the first step for the second stage was the construction of a matrix, \( \Omega_n^{-1} \). The omega matrix was constructed as follows;
\[ \Omega_n^{-1} = \left( \sum_{i=1}^{l} d_{i} \Omega_{i} + \Omega_{cd} \right)^{-1} \]

where \( \Omega_{i} = \frac{\Omega_{i}}{\sigma_{i}^{2}} \)

\[ i \text{ indexes appliances, } i=1 \text{ to } l \]

\[ n \text{ indexes cross-sectional elements, } n=1 \text{ to } N \]

NB: the structure of \( \Omega_{cd} \) is similar to \( \Omega_{i}^{n} \)

The data for the second stage included all appliance level cross-sections, thus creating an unbalanced panel. The key to introducing the nature of unbalanced data is through the construction of the omega matrix. If the data is balanced; each cross-sectional element has exactly the same number of time series elements. The omega matrix would be a \( T \times T \) matrix. If \( T=5 \) then this matrix could be characterized, incorporating the prior estimate of \( P \), in the following form;
\[ \Omega_{i}^{n} = \begin{bmatrix}
1 & -\hat{p}_i & 0 & 0 & 0 \\
-\hat{p}_i & 1 + \hat{p}_i^2 & -\hat{p}_i & 0 & 0 \\
0 & -\hat{p}_i & (1 + \hat{p}_i^2) & -\hat{p}_i & 0 \\
0 & 0 & -\hat{p}_i & (1 + \hat{p}_i^2) & -\hat{p}_i \\
0 & 0 & 0 & -\hat{p}_i & 1
\end{bmatrix} \]

In the presence of unbalanced data this matrix shrinks to a \( (T-m) \times (T-m) \) matrix, where \( m \) is the number of missing observations in a given time series. Consider an example where there are ten possible time series points and there is data for the first three points and the eighth and ninth points. Hence there are only five points with data. Under this example the structure of omega would be a block diagonal matrix of two sub-matrices, one \( 3 \times 3 \) and the other \( 2 \times 2 \).

\[ \Omega_{i}^{n} = \begin{bmatrix}
1 & -\hat{p}_i & 0 & 0 & 0 \\
-\hat{p}_i & 1 + \hat{p}_i^2 & -\hat{p}_i & 0 & 0 \\
0 & -\hat{p}_i & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & -\hat{p}_i \\
0 & 0 & 0 & -\hat{p}_i & 1
\end{bmatrix} \]
Having explained a method for incorporating unbalanced data; the complete specification of the model is given with the following definitions.

$\mathcal{D}_1$ is a matrix of variables for which priors were produced. This matrix represents the conditional demand data representation of the prior appliance level models. $\mathcal{D}_2$ are those variables for which no priors exist. $Y_i$ represents the dependent variable, (billing load). $g$ is the total number of coefficients. $t$ is the index of time, $t$ to $T$. $n$ is the index of cross-sections, $n$ to $N$. $i$ is the index of appliances, $i \rightarrow I$.

\[ \mathcal{D}_1' \Omega^{-1} \mathcal{D}_2 = \sum_{n=1}^{N} \mathcal{D}_1'_{n(x)} \Omega_{n(x)}^{-1} \mathcal{D}_2_{n(x)} \]  \hspace{1cm} (2)

\[ \mathcal{D}_2' \Omega^{-1} \mathcal{D}_2 = \sum_{n=1}^{N} \mathcal{D}_2'_{n(x)} \Omega_{n(x)}^{-1} \mathcal{D}_2_{n(x)} \]  \hspace{1cm} (3)

\[ \mathcal{D}_1' \Omega^{-1} \mathcal{D}_1 = \sum_{n=1}^{N} \mathcal{D}_1'_{n(x)} \Omega_{n(x)}^{-1} \mathcal{D}_1_{n(x)} \]  \hspace{1cm} (6)

\[ \mathcal{D}_1' Y_{i} = \sum_{n=1}^{N} \mathcal{D}_1'_{n(x)} \Omega_{n(x)}^{-1} Y_{i} \]  \hspace{1cm} (4)

\[ \mathcal{D}_2' Y_{i} = \sum_{n=1}^{N} \mathcal{D}_2'_{n(x)} \Omega_{n(x)}^{-1} Y_{i} \]  \hspace{1cm} (5)

\[ \mathcal{D}_1' \mathcal{D}_1 = (g)(g) - (2)(g)(2)' \]  \hspace{1cm} (1)

The conditional demand estimate then becomes:

\[ \begin{bmatrix} \hat{\beta}_{(x)} \\ \hat{\gamma} \end{bmatrix}_{(g+1) \times 1} = \begin{bmatrix} (7)_{(g)} \\ (8)_{(g)} \end{bmatrix} \]  \hspace{1cm} (7)

\[ \begin{bmatrix} (6)_{(g)} \\ (2)_{(g)} \end{bmatrix}^{-1} \begin{bmatrix} (4)_{(g)} \\ (5)_{(g)} \end{bmatrix} \]  \hspace{1cm} (8)

The next step is to prepare for the Bayesian weighting. We need to define two new matrices.

\[ (9)_{(g)} = \begin{bmatrix} \text{cov}(\hat{\beta}_1) \cdots 0 & \cdots & 0 \\ 0 & \text{cov}(\hat{\beta}_2) \cdots 0 & \cdots \\ 0 & 0 & \cdots & \text{cov}(\hat{\beta}_g) \end{bmatrix} \]  \hspace{1cm} (9)

\[ (10)_{(g)} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_i \end{bmatrix}_{(g)} \]  \hspace{1cm} (10)

Combination of the above matrices yield the posterior, given by,

\[ \begin{bmatrix} \beta^* \gamma^* \end{bmatrix}_{(g+1) \times 1} = \begin{bmatrix} (11)_{(g)} \\ (12)_{(x)} \end{bmatrix} \]  \hspace{1cm} (11)

where;

\[ (11)_{(g)} = [(1)_{(g)} + (9)_{(g)}]^{-1} \]

\[ * (1)_{(g)}(7)_{(g)} + (9)_{(g)}(10)_{(g)} \]

\[ (12)_{(x)} = [(8)_{(x)} + (3)_{(x)}(2)_{(x)}(1)_{(g)} + (9)_{(g)}]^{-1} \]

\[ *[(-9)_{(g)}(10)_{(g)} - (1)_{(g)}(7)_{(g)}] 

\[ + (3)_{(x)} (2)_{(x)} (7)_{(g)} \]

If one wishes to consider the variance-covariance matrix of the posterior mean,

\[ \mathcal{V}_{(g+1)|x} = \begin{bmatrix} (13)_{(g)} & (14)_{(x)} \\ (14)_{(g)} & (15)_{(x)} \end{bmatrix} \]  \hspace{1cm} (12)

where,

\[ (13)_{(g)} = ((1)_{(g)} + (9)_{(g)})^{-1} \]

\[ (14)_{(g)} = -((1)_{(g)} + (9)_{(g)})(3)_{(1x)}^{-1} \]

\[ (15)_{(x)} = (3)_{(x)}^{-1} + (3)_{(x)}^{-1}(2)_{(x)}(1)_{(g)} \]

\[ *[1(1)_{(g)} + (9)_{(g)}]^{-1}(2)_{(x)}(3)_{(1x)}^{-1} \]

Hence the standard errors of the posterior coefficients are given by the following,

\[ \sqrt{\mathcal{V}_{i,i}} \]  \hspace{1cm} (13)

and the standard deviation of forecasted values is,

\[ \sqrt{\mathcal{V}_{i,i}} = \sqrt{\mathcal{X}_i \mathcal{Y}_i} \]  \hspace{1cm} (14)

The evaluation of forecast errors has gained more prominence in recent times. Many forecasts are used in strategic planning and senior executives appreciate a measure of risk with each forecast. Researchers can evaluate these forecast errors at different points by varying the matrix of forecasted drivers. This point becomes more important if those drivers are themselves forecasts. Scenarios can be generated around supporting input forecasts.
USING SAS/IML SOFTWARE IN THE PRESENCE OF UNBALANCED DATA

The creation of the omega matrix is key to this estimation. The SAS/IML software code to do this is presented below. The original code was written in the macro language to accommodate differing N and T. The processing below is completed by cross-sectional element through a macro loop. The following macro references are required:

\&time = total potential time series points for cross-section n
\&typ = a code created to differentiate between appliance types, i
\&s = an index of cross-sectional elements, n = 1...N
\&val = non-missing time series elements for cross-section n
\&y = is a vector of the dependent variable
\&rcs = is a vector counting the number of continuous time series elements
\&rcb = is a vector containing locations of values in \&rcs

do j = 1 to \&time;
   /* do you have data */
   if \&y[j,1] ^= . then do;
      k=k+1;
   end;
   if (j<\&time) then do;
      if (\&y[j,1] ^= .) & (\&y[j+1,1] ^= .) then \&rcs[j,1] = k;
   end;
   else if (j=\&time) & (\&y[j,1] ^= .) then \&rcs[j,1] = k;
   if \&y[j,1] ^= . then k=0;
end;

%do n = 1 to \&val;
   /* loop to create your omega */
   z=sum(\&rcs[\&rcb[\&n,1]]);
   /* create a scalar for loop */
   if z ^= 0 then do;
      \&ran=i(\&rcs[\&rcb[\&n,1]]);
   end;
   /* create sub-matrix */
   do m = 1 to z;
      /* loop to fill elements of sub-matrix */
      k=m-1;
      if (m=1) then do;
         \&ran[k,m]=-1*\&yho;
         \&ran[k,m]=-1*\&yho**2;
      end;
      if (m>1) then \&ran[m,m]=1;
   end;
end;

With each sub-matrix created for a given cross-section, these were diagonally blocked with the SAS/IML software 'block function'. Having created each specific cross-sectional omega, the formulas presented above are calculated.

RESULTS

The platform used for estimation was The SAS System for Windows, Version 6.08, operating under Windows 3.1. The hardware used was an Intel 486SX/2.66MHz with 16MB of RAM. The computer operates on a Novell network through 16Mb Token Ring network card. Run time for estimation varied. The appliance level information was over 70 to 120 cross-sections and varied by appliance. The prior estimation was structured to be a balanced time series for 14 months giving (14 x 70-120) observations. These appliance level equations took about 10 to 15 minutes to converge. Estimation of the fully specified prior conditional demand included 620 cross-sectional elements for 67 months giving (67 x 620) observations. This estimation took about 15 hours to converge. The results of this estimation lead to the creation of \( \Omega^{2d} \) mentioned above.

The largest part of computer time in the posterior estimation, the second stage, was occupied with calculating equations (2) through (8). The exercise among the 620 cross-sectional elements, with a minimum potential of 47 months of monthly data took about 12 hours to complete. The resulting data was used in the final conditional demand estimation, equations (11) and (12), took about 10 minutes for estimation convergence. The estimates that resulted from this exercise proved to be reasonable through validation checking. We believe the Bayesian weighting of the prior estimates provided better estimates than only performing a simple billing load conditional demand.

CONCLUSION

Panel data can be very effective in producing better forecasts. The SAS/ETS software, Version 6, procedure PROC TSCSREG has given users a powerful tool but fails to incorporate unbalanced data. Real time series data used by forecasters is rarely balanced. Understanding that unbalanced data is a truism, we have shown a framework where forecasters can incorporate the richness of a panel series in the presence of unbalanced data. Even though the approach is taken with respect to a given modelling strategy, the ideas used are widely applicable. The key to making this operational is dependent on your computer system. We had to make many changes, by trial and error, in order to handle the large amounts of data. The usage of panel data has helped our forecasting efforts. A call for further research in the code behind SAS/ETS software's PROC TSCSREG to incorporate an unbalanced time series would aid many researchers in offering them an alternate modelling specification.

BIBLIOGRAPHY


ACKNOWLEDGEMENTS

The authors wish to thank the following people for helping out with this project in various ways; Paula McNeil, SAS Institute Canada; Annette Sanders, SAS Institute, Cary NC; Ian Gebhardt, Consumers Gas Company Ltd.; and our families who were with us on the journey of this project; from waking up in the middle of the night quoting SAS code to working weekends. Special thanks to the author of the 'Bayesian twist', Dean Mountain, PhD, Professor, McMaster University, email 'mountain@mcmaster.ca'.

SAS, SAS/ETS and SAS/IML are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates a USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

Clay Greene & John Ah-Sen can be reached at, Consumers Gas, PO Box 650, Scarborough, ON, M1K 5E3, CANADA, e-mail 'cgreene@consumersgas.com' and 'jahsen@consumersgas.com'.

Statistics, Data Analysis, and Modeling