New Developments in Panel Data Estimation: Full-Factorial Panel Data Model

Patrick J. Howie
Vice President
TargetRx
220 Gibraltar Rd.
Horsham, Pa. 19044
215-444-8746
phowie@targetrx.com

Ewa J. Kleczyk
Sr. Econometrician
TargetRx
220 Gibraltar Rd.
Horsham, Pa. 19044
Ph.D. Candidate
Agricultural and Applied Economics Department
Virginia Tech
Blacksburg, Va. 24061
215-444-8806
ekleczyk@targetrx.com

Selected Paper prepared for presentation at the American Agricultural Economics Association, Portland, Oregon, July 29-August 1, 2007

New Developments in Panel Data Estimation: Full-Factorial Panel Data Model

April, 2007

Patrick J. Howie and Ewa J. Kleczyk[1]

Abstract

Panel data has been widely used in many social science studies. Pooling data across cross-sections and time-series improves quality of data analysis; however, the model is limited in its ability to actually accurately predict variables of interest due to severe practical data limitations and the ability of properly capturing varying market structures. In this article, a simple and innovative model of product share is introduced. The Full-Factorial Panel Data Model is based on the simple premises of re-conceptualization of any zero-sum group as a series of two-entity markets. This model solves the challenges associated with pooling data across disparate cross-sections and time-periods as well as the changing competitive market structure issues and therefore results in reliable variable of interest estimates.

Introduction

In recent years, panel data has become a widely utilized type of data set for econometric analysis in many social sciences. Panel data combines cross-sections and time-series data and therefore provides a more appealing structure of data analysis than either cross sectional or time-series alone. Although more costly to gather, the advantages of this data type include better and more precise parameter estimation due to a larger sample size as well as simplification of data modeling (Hsiao, 2005).

Panel data models are limited, however, in their ability to account for the structural differences resulting from pooling data across disparate cross-sections and time-series. This is particularly a problem when trying to estimate zero-sum dependent variables, such as market share, which is an important real world business application.

[1] Vice President of New Product Development, TargetRx; Ph.D. Candidate, Department of Agricultural and Applied Economics, Virginia Tech and Senior Econometrician, TargetRx. We would like to thank Dr. Christopher Parmeter for his valuable comments. All remaining errors are ours.

When estimating market share and other similar types of variables, panel data typically have varying degrees of competition across the markets or varying degrees of competition within a market across time. The limitation results in imprecise parameter estimates using current panel data analysis techniques. To tackle this problem, an innovative model for panel data estimation is introduced in order to deal with data reflecting zero-sum conditions. This "Full-Factorial Panel Data Model" is based on the re-conceptualization of any zero-sum group, such as a series of products battling for market share, as a series of two-entity groups. This model solves the challenges associated with varying group (i.e. market) structures when pooling data across cross-sections and time periods, while preserving the benefits of Panel Data Analysis, thus improving the reliability of the parameter estimates of the variables of interest.

Panel Data

Panel data analysis refers to data containing time-series for a cross-section or group of people who are surveyed periodically over a given period of time (Yaffee, 2003). The observations in panel data involve at least two dimensions: a cross-sectional dimension indicated by subscript *i* and a time-series dimension indicated by subscript *t*. Panel data analysis have become very popular in the social sciences, having been used in Economics to study behavior of firms and wages of people over time as well as in Marketing to study market share changes across different market structures (Hsiao, 2005; Yaffee, 2003).

Panel data analysis has many advantages over analysis using time-series and cross-sections alone. For example, the increased sample size due to the utilization of cross-sectional and time-series data improves the accuracy of model parameters estimates

due to a greater number of degrees of freedom and less multicollinearity compared to either cross-section or time-series data alone. In the case of non-stationary time-series data, the independence among cross-sections invokes the central limit theorem to ensure the estimators remain asymptotically normal. Additionally, since panel data contains information on both the inter-temporal dynamics and the individuality of entities, it controls for the effect of missing variables on the estimation results. Finally, panel data allows for identification of previously not identified model specification (Hsiao, 2005).

There are several types of panel data analytic models currently in use: constant coefficient models, fixed effects models, and random effects models. The most basic model is the constant coefficients model, where both the intercept and slope have constant coefficients. When there are no temporal or cross-sectional differences, the data can be pooled across cross-sections and time-series and ordinary least squares (OLS) regression can be performed to analyze the data. The constant coefficient model is specified as follows:

$$Y_{it} = \alpha + \beta X_{it} + v_{it}, \text{ where } i = 1 \ldots N \text{ and } t = 1 \ldots T \qquad (1)$$

where Y_{it} is the dependent variable, X_{it} is the independent variables, and v_{it} is the error term distributed normally ($v_{it} \sim NIID\ (0,\ \delta_v^2)$). The underlying assumptions of this model are: 1) the explanatory variables (X_{it}) in each time period are uncorrelated with the idiosyncratic error in each time period: $E(X_{it}' v_{it}) = 0$; and 2) the explanatory variables are uncorrelated with the unobserved effect in each time period: $E(X_{it}'\alpha_i) = 0$. The OLS regression estimation provides consistent estimators as long as the underlying assumptions are satisfied (Wooldridge, 2002).

The second type of panel data model is the fixed effects model, where the slopes are constant but the intercepts vary. In this type of model, there are significant differences among cross-sections and dummy variables are employed to represent each cross-section. Sometimes there might not be any significant differences across cross-sections, but an autoregressive time-series structure is present. Dummy variables are therefore utilized to represent temporal dependence between periods. The fixed effects model is specified as follows:

$$Y_{it} = \alpha_i + \beta X_{it} + v_{it}\,, \text{ where } i = 1\ldots N \text{ and } t = 1\ldots T \qquad (2)$$

$$\varepsilon_{it} = \alpha_i + v_{it} \qquad (3)$$

where $v_{it} \sim NIID\ (0,\ \delta_v^2)$; α_i denotes a cross-section-specific effect, and v_{it} is the idiosyncratic error term (Hsiao, 2002). In the fixed effects analysis, α_i is arbitrarily correlated with X_{it}, $E(X_{it}'\alpha_i) \neq 0$ (Wooldridge, 2002).

The third type of panel data model is the random effects model, where both the slopes and the intercepts vary. In a random effects model, the α_i is included in the error term and the model takes the following specification:

$$Y_{it} = \beta X_{it} + u_{it}, \text{ where } i = 1\ldots N \text{ and } t = 1\ldots T \qquad (4)$$

$$u_{it} = \alpha_i + v_{it} \qquad (5)$$

where $\alpha_i \sim NIID\ (0,\ \delta_\alpha^2)$; $v_{it} \sim NIID\ (0,\ \delta_v^2)$. In the random effects approach, α_i is in the composite error term that is orthogonal to the explanatory variables, (X_{it}), $E(X_{it}'\alpha_i) = 0$. Furthermore, the method accounts for the implied serial correlation in the composite error, $u_{it} = \alpha_i + v_{it}$, the same way as the generalized least squares (GLS) estimation technique (Wooldridge, 2002).

In order to identify whether a fixed or random effects model is appropriate for the data analysis, the Hausman test is usually performed to examine the appropriateness of the random effects estimator[2]. Based on the test result, either the fixed effects or random effects model is chosen (Wooldridge, 2002).

Panel Data Issues

Although panel data has greatly improved the ability of obtaining reliable parameter estimates thru eliminating important estimation issues such as omitted variable bias and non-stationary time-series, there are still severe practical data limitations impacting the precision and accuracy of panel data estimates. While pooling data across cross-sections and time-series increases the sample size of the data, it introduces asymmetries based on varying group structures across the cross-sections or different points in time that calls into question the soundness of the approach when applied to zero-sum datasets. A common example in the business literature is the pooling of data across various time periods or industries in order to determine the predictors of market share. In nearly every case, the number and strength of competitors varies considerably across the different time periods or the different industries. Models built on pooled data across such differing markets are therefore dominated by the differences across markets rather than by the differences across competitors within each market (Fok, 2003). For

[2] A Hausman test compares two estimators. Under the null, the fixed and random effects estimators are consistent, but one is more efficient; under the alternative, the more efficient of the two becomes inconsistent but the less efficient remains consistent. Thus if the null is not rejected, the two estimators should be similar; divergence indicates rejection of the null. This gives the test statistic:
$W = (\beta^{\wedge F} - \beta^{\wedge R})\Sigma^{\wedge -1}(\beta^{\wedge F} - \beta^{\wedge R}),\ W \sim X^2(k)$
where k is the number of estimated coefficients and $\Sigma^{\wedge -1}$ is the difference of the estimated covariance matrices from the two estimators. Rejection of the null implies the effects are correlated with the individual variances, and the fixed effects should be used (Stata 8, 2005).

example, a monopolistic competition market implies different parameter predictions compared to those found under a competitive market structure.

The varying group structures that occur when pooling data across cross-sections and time-series can severely limit one of the great advantages of using panel data, which is the increased sample size resulting in more reliable parameter estimates. Entry/exit of firms/products/etc. into or out of a group occurs rather frequently in both cross-sectional and time-series data. An adequately large sample is required for panel regression analysis to precisely capture the impact of each entering/existing firm/product on each group/market. If the sample size is not large enough, the parameter estimates of the independent variables will be inefficient (Fok, 2003). Market research studies analyzing the behavior of firms and products across different competitive market structures have struggled with the limitations imposed when pooling time-series and cross-sectional data to obtain large enough sample due to a high likelihood of underlying market structure changes, resulting in unreliable parameter estimates (Cooper and Nakanishi, 1996).

The sample size issue is closely related to the estimation of effects of new entrant/exit from the market. According to Fok (2003), a combined model of pre- and post- entry/exit should be employed to capture the changing market structure. Methods such as standardizing/normalizing (log-centering transformation) the data are employed to facilitate the varying market structure. These methods, however, still yields biased estimates. An addition or removal of a single firm/product leads to a significant change in the standardized values and therefore biased estimates (Cooper and Nakanishi, 1996).

A substantive number of articles such as those by Shankar (1999) and Gatignon *et al.* (1990) deal with the changing market structures by simply estimating pre- and post-

entry/exit models; however, the pre- and post-entry models are interdependent and do not capture the dynamic market structure. Additionally, the approach might not guarantee an adequate sample for post-entry/exit model estimation and therefore providing unreliable share estimates (Fok, 2003). Finally, the changing market structure could be tested within models that exclude new entrants from the analysis. This approach is, however, not appropriate, because the new entrant behavior has a direct impact on the incumbents. So even with the assumption of constant competitive market structure, the parameter estimates are affected by the new market entrant (Fok, 2003).

Full Factorial Panel Data Model Introduction

In this article, a new and innovative method to panel data organization prior to regression estimation is proposed when dealing with data reflecting zero-sum conditions, such as market share. The Full Factorial Panel Data Model improves parameter estimates when modeling a single group (ie. market) and solves the challenges associated with pooling data across groups and time. The approach is based on a re-conceptualization of any group as a series of two-entity groups. In this "full factorial" model, the data is restructured to reflect every two-entity group combination. For example, for market *K (k* represents number of groups/markets in the study) with *N* number of cross-sections, the two-entity group/market is described as follows:

$$y_{kij} = Y_{ki}/(Y_{ki} + Y_{kj}), \text{ where } k = 1\ldots K,\ i = 1\ldots I \text{ and } j = 1\ldots I \text{ and } i \neq j \quad (6)$$

$$y_{kji} = Y_{kj}/(Y_{ki} + Y_{kj}), \text{ where } k = 1\ldots K,\ i = 1\ldots I \text{ and } j = 1\ldots I \text{ and } i \neq j \quad (7)$$

$$y_{kij} + y_{kji} = 1, \text{ where } k = 1\ldots K,\ i = 1\ldots I \text{ and } j = 1\ldots I \text{ and } i \neq j \quad (8)$$

for *N!/2!* (every combination of two-cross-section groups/markets), where *i*, *j* are entities/products in market *K*.

There are many desirable properties of this approach. First, every group has the same market structure. For market share, every firm/product has an expected market share of 50%. The property reduces the impact of market dynamics across cross-sections on estimation process. As a result, no distinction between market structures for each entity/product in necessary. Second, the total number of observations increases to:

$$Sample\ size = N!/2! \qquad (9)$$

where *N* is the number of cross-sections in the group/market. The significant increase in the sample further improves the efficiency of the estimates. Third, through the use of dummy variables and their interactions with continuous independent variables, this approach can be parameterized to capture estimates for every pair, such as the estimates of cross-elasticities. As the number of pair-wise parameters increases, however, the sample size advantages decrease. Finally, the changing market dynamic thru participants' entry/exit into/from the market does not affect the model since the group/market structure is always comprised of two entities/products.

As mentioned above, the marketing data can be pooled across firms/products and time periods without affecting the underlying structure of the data, which allows for employment of panel data estimation techniques including the constant, fixed and random effects models.

The independent variables can also be transformed to conform to the new way of data estimation. For example, additive models can be estimated by taking the gap

between the two cross-sections' independent variables while the multiplicative models can be estimated using the ratio of the two:

$$xgap_{ijt} = X_{it} - X_{jt}, \text{ where } i = 1...I,\ j = 1...I \text{ and } i \neq j;\ t = 1\ldots T \qquad (10)$$

$$xratio_{ijt} = X_{it}/X_{jt}, \text{ where } i = 1...I, j = 1...I \text{ and } i \neq j;\ t = 1\ldots T \qquad (11).$$

In order to control for the group/market size, an independent variable (continues or dummy variable) can be included in the estimation process.

A caveat of the pair-wise application is the problem surrounding the estimation of standard error. Although the regression parameter estimates are unbiased, the coefficients' standard errors are biased down due to the increased sample size, thus leading to an overestimation of the significance level of the independent variables. In order to adjust the standard errors for the overstatement of the degrees of freedom, the standard errors obtained from the double entry regression are multiplied by the factor of squared root of 2 or by the covariance matrix obtained in the initial regression when multiplied by the factor of 2 (Kohler and Rodgers, 2001).

Once data modeling is completed, estimates for the dependent variable of interest of the original "full" group/market ($\boldsymbol{Y}p_{k1t}...\boldsymbol{Y}p_{klt}$) can be directly recovered by "adding up" the dependent variables of each pair-wise market combinations times the relative shares in the original dependent variable of each pair. The following outlines the necessary steps for this conversion using market share example:

1. Put all entities/products from k groups/markets predicted variables of interest in terms of $Yp_{k1t...}\ Yp_{klt}$. $Y_{k1t}...\ Y_{klt}$ are the actual values of the dependent variable from the last month with observed data; if one of the firms/products has been just launched then the dependent variable for that entity/product is 0.

$yp_{kijt} = Yp_{kit}/(Y_{kit} + Y_{kjt})$, where $k = 1…K$; $i = 1…I$, $j = 1…I$ and $i \neq j$; $t = 1…T$

(12)

$yp_{kijt} * (Y_{kit} + Y_{kjt}) = Yp_{kit}$, where $k = 1…K$; $i = 1…I$, $j = 1…I$ and $i \neq j$; $t = 1…T$

(13)

2. Take the average value of dependent variable for each entity/product ($Yp_{k1t}…Yp_{kIt}$) at time t:

$Yp_{k1t} = \Sigma\, Yp_{k1t}/(N-1)$, where $k = 1…K$; $t = 1…T$ and N = number of entities/products in group/market (14)

$\mathbf{Y}p_{kIt} = \Sigma\, Yp_{knt}/(N-1)$, where $k = 1…K$; $t = 1…T$ and N = number of entities/products in group/market (15)

3. Rebalance $\mathbf{Y}p_{k1t}…\mathbf{Y}p_{kIt}$ in order for the firm's/product's dependent variable to sum up to unity:

$$\mathbf{Y}p_{k1t} +… + \mathbf{Y}p_{kIt} = 1, \text{ where } t = 1…T \qquad (16).$$

4. If an entity/product entered or left the group/market just add/remove the appropriate equation to each step and follow the same method.

For forecasting purposes, the same process applies in an iterative fashion. When a new entrant enters the group/market, the dependent variable for incumbents can be precisely forecasted without worrying about the group/market structure and just considering the current number of participants in the group/market. The parameter coefficients of the independent variables are based on the two-cross-section group with the universal group/market structure.

Conclusion

The object of this article was to introduce a new and innovative panel data model for more accurate and precise parameter estimation when dealing with data representing zero-sum conditions such as when predicting market share. The Full Factorial Panel Data Model re-conceptualizes any groups as a series of two entity groups. The model can be applied in all market related studies, including attraction models for market share estimation, competitive firm behavior studies, financial models, as well as agricultural market models. The model allows researchers to resolve the problem of pooling data across market and time-series as well as deal with the dynamic market structures by creating two-firm/product markets for all study participants. The innovative model framework results in better and more precise historical market evaluation and therefore more accurate forecasts of variables of interest.

Acknowledgments

We would like to thank Dr. Christopher Parmeter from Virginia Tech and Jim Colizzo from TargetRx for their valuable comments. Additionally, Ewa J. Kleczyk would like to thank TargetRx located in Horsham, Pa. for financially supporting her doctorate education at Virginia Tech.

References

1. Cooper, L.G. and Nakanishi, M. (1996). *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness.* Kluwer Academic Publishers: Boston.

2. Fok, D. (2003). *Advanced Econometric Marketing Models.* Erasmus Research Institute of Management and University of Rotterdam. Available online at: http://www.erim.eur.nl.
3. Gatignon, H., Weitz, B. and Bansal, P. (1990). "Brand Introduction Strategies and Competitive Environments". *Journal of Marketing Research*. p. 390–401.
4. Hsiao, C. (2002). *Analysis of Panel Data (2nd ed.)*, Cambridge University Press.
5. Kohler, H. P. and Rodgers, J. L. (2001). "DF-Analysis of Heritability with Double-Entry Twin Data: Asymptotic Standard Errors and Efficient Estimation." Available online at: http://www.user.demogr.mpg.de/kohler.
6. Shankar, V. (1999). "New Product Introduction and Incumbent Response Strategies: Their Interrelationship and the Role of Multimarket Contact". *Journal of Marketing Research*. p. 327–344.
7. *Stata 8 Manual*. (2005). Available online at: http://www.stata.com/sipport/faqxs/stat/panel.html.
8. Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.
9. Yaffee, R. (2005). "Primer for Panel Data Analysis." *Connect: Information Technology at NYU.* Available online at: http://www.nyu.edu/its/pubs/connect/fall03/yaffee_primer.html.