

**10<sup>th</sup> National Convention on Statistics (NCS)**  
EDSA Shangri-La Hotel  
October 1-2, 2007

**Generalized Method of Moments Estimation on a  
Linear Panel Data Model of a Clinical Trial**

by

Aristotle B. Magallanes

For additional information, please contact:

Author's name	:	Aristotle B. Magallanes
Designation	:	Assistant Professor
Affiliation	:	College of Public Health, University of the Philippines-Manila
Address	:	Taft Avenue, Manila
Telefax. no.	:	(0632) (5242703)
E-mail	:	<a href="mailto:abmagallanes@yahoo.com">abmagallanes@yahoo.com</a>

# Generalized Method of Moments Estimation on a Linear Panel Data Model of a Clinical Trial

by

Aristotle B. Magallanes<sup>1</sup>

## ABSTRACT

The Generalized Method of Moments (GMM) is a statistical tool used for estimating model of financial and economic panel data. The estimation technique is an improvement method over Ordinary Least-Squares because of the smaller asymptotic variances, no distributional assumption involving the error and even consistent under weak distributional assumption. With its growing impact in applications, the method on estimation is considered in some clinical trials analysis.

The GMM is used in the estimation of a model that characterizes the linear panel data of a randomized clinical trial on acute non-bloody diarrhea of 324-month old infants and children. The balanced panel data set was observed repeatedly for the presence and absence of diarrhea to every subject medicating to a treatment drug in a ten-day follow-up considering mean age of the subjects of each drug as a covariate. This study aims to estimate the linear panel data model of the drug efficacy and safety. The GMM estimates showed significant effect of the previous day  $t-1$  of the log(odds) of  $i$ th drug and followed a dynamic linear panel data model.

**Keywords:** Generalized method of moments, Linear panel data model, Randomized clinical trial on diarrhea

## I. Introduction

The Generalized Method of Moments (GMM) is a statistical method for obtaining estimates of parameters of economic models. A number of literatures on GMM-based inference techniques in econometrics are applied to various areas such as agriculture, business cycles, commodity markets, health care and many others (Hall, 2005). These methods and applications are aids to the consideration of the GMM estimation procedure to some clinical trial analysis.

The GMM is an improvement estimation technique over Ordinary Least-Squares (OLS) because of the smaller asymptotic variances. It has no distributional assumption involving the error and it is even consistent under weak distributional assumption. It has robust estimators to the failure of model assumptions on existence of heteroskedasticity. It also allows the parameters to be overidentified and suited for obtaining efficient estimators that account for the serial correlation. On the other hand, GMM can suffer from finite-sample problems especially adding many moment conditions that do not add much information, and

---

<sup>1</sup> Assistant Professor, Department of Epidemiology and Biostatistics, College of Public Health, University of the Philippines Manila. abmagallanes@yahoo.com

the finite-sample bias in GMM estimators becomes an issue with small sample sizes (Wooldridge, 2001).

In this study, GMM is used in the estimation of a model that characterizes the linear panel data of a randomized clinical trial (RCT) on acute non-bloody diarrhea of 3-24-month old infants and children. The balanced panel data set is contained observations of the presence and absence of diarrhea to every subject medicating to one of the treatment drugs in a ten-day clinical trial. It is a single binary response variable which is observed repeatedly for each subject with ten-day follow-ups. A set of covariates for each of the subjects is also considered and recorded like age, weight, height and other relevant variables. This aims to estimate the linear panel data model of the drug efficacy and safety of the RCT on diarrhea using GMM.

The paper is organized as follows. Section 2 discusses the panel data models. Section 3 presents the estimation procedure of the model using GMM. Section 4 reports the results of the model estimation of the applied data set of RCT on diarrhea. Finally, section 5 concludes.

## II. Panel Data Models

A data set containing observations on a single phenomenon observed over multiple time periods is called time series. In time series data, both the values and the ordering of the data points have meaning. On the other hand, a data set containing observations on multiple phenomena observed at a single point in time is called cross-sectional. In cross-sectional data sets, the values of the data points have meaning, but the ordering of the data points does not. Thus, a data set containing observations on multiple phenomena observed over multiple time periods is called panel data. Time series and cross-sectional data are both one-dimensional while panel data sets are two-dimensional.

### 2.1. Linear Panel Data Model

A linear panel data model is expressed in the form

$$y_{it} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_{it}, \quad i = 1, 2, \dots, n, \quad t = 1, \dots, T_i \quad (2.1)$$

This can also be rewritten as

$$y_{it} = \sum_{j=0}^k \beta_j x_{ij} + \varepsilon_{it}, \quad x_{i0} = 1$$

where  $y_{it}$  be the dependent variable of  $i$ th subject at time  $t$ ,  $x_{jt}$  be the observed value of  $i$ th subject in the  $j$ th covariate,  $\beta_j$  be the parametric coefficient of the  $j$ th covariate, and  $\varepsilon_{it}$  be the  $i$ th unobserved error at time  $t$ . The  $T_i \geq 2$  and  $N \geq 2$  are considered. Let  $e_{it}$  have finite moments and in particular  $E(e_{it}) = E(e_{it} e_{is}) = 0$ , for  $t \neq s$  (Arellano and Bond, 1991). Moreover,  $e_{it} = v_i + v_t$  where  $E[v_i] = E[v_t] = 0$ ;  $E[v_i v_t] = 0$  (Anderson and Hsiao, 1981).

One such example is the case when the response variable  $y_{it}$  could be an indicator of diabetes where

$$y_{it} = \begin{cases} 0 & \text{if } i\text{th subject has no diabetes at time } t \\ 1 & \text{if } i\text{th subject has diabetes at time } t \end{cases}$$

and  $x_1$  could be an indicator for sex of the subject. Equation (2.1) can also be expressed in matrix form:

$$y_{it} = \mathbf{X}_{it}' \boldsymbol{\beta} + e_{it}, \quad (2.2)$$

where  $\mathbf{X}_{it}$  be a  $(k+1) \times 1$  vector of covariates for the  $i$ th subject at time  $t$  and  $\boldsymbol{\beta}$  be a  $(k+1) \times 1$  vector of parametric coefficients of covariates.

When the dependent variable  $y_{it}$  is a binary response variable, with output  $(0,1)$ ,  $E(y_{it}|x_{it}=x)$  is called a binary regression. Note that  $E(y_{it}|x_{it}=x) = \Pr(y_{it} = 1|x_{it};\boldsymbol{\beta})$ , which equals the proportion of population members who have  $y_{it}=1$  among those who have  $x_{it}=x$ . In addition,  $E(y_{it}|x_{it};\boldsymbol{\beta}) = \mathbf{X}_{it}' \boldsymbol{\beta}$ . This implies that  $\Pr(y_{it} = 1|x_{it};\boldsymbol{\beta}) = \mathbf{X}_{it}' \boldsymbol{\beta}$ . A model of the form  $g[E(y_{it}|x_{it};\boldsymbol{\beta})] = \mathbf{X}_{it}' \boldsymbol{\beta}$  is called a generalized linear model (Rothman and Greenland, 1998). The function  $g$  is called the link function for the model; thus, the link function is logit for the log-linear model. Hence, the generalized-linear form of the logistic-odds model is the logit-linear odds model. Therefore, analogous expression of (2.2) is

$$\begin{aligned} \log\left[\frac{\Pr(y_{it} = 1|x_{it};\boldsymbol{\beta})}{\Pr(y_{it} = 0|x_{it};\boldsymbol{\beta})}\right] &= \mathbf{X}_{it}' \boldsymbol{\beta}, \\ \text{logit}(\Pr(y_{it} = 1|x_{it};\boldsymbol{\beta})) &= \mathbf{X}_{it}' \boldsymbol{\beta}, \end{aligned} \quad (2.3)$$

that is, the logistic model corresponds to the logit function, because  $\text{logit}(\expit(u)) = u$ .

Equation (2.1) is static since the equation does not depend on  $y_{i,t-1}$  (Honore and Tamer, 2006). It is also an extension of logistic regression to the case where the binary outcome variable is observed repeatedly for each subject (Zeger, Liang and Self, 1985).

## 2.2. Dynamic Linear Panel Data Model

A linear panel data model can be considered also in the form

$$y_{it} = \sum_{j=1}^p \alpha_j y_{i,t-j} + \sum_{l=0}^k \beta_l x_{il} + \varepsilon_{it}, \quad i = 1, 2, \dots, N, \quad t = 1, 2, \dots, T_i \quad (2.4)$$

where  $\alpha_j$  be the parametric coefficients of the lagged dependent variables  $y_{i,t-j}$  and at least one  $\alpha_j \neq 0$ .

Equation (2.4) is called a dynamic linear panel data model since the equation depends on lagged dependent variable  $y_{i,t-1}$  (Honore and Tamer, 2006). The dynamic behavior of the model follows that at least one lagged dependent variable has a significant effect to the dependent variable.

Let  $\mathbf{X}_{it}' = (y_{i,t-1}, y_{i,t-2}, \dots, y_{i,t-p}, 1, x_{i1}, \dots, x_{ik})$  be the  $m \times 1$  vector of covariates for subject  $i$  at time  $t$ ,  $m = p + k + 1$ . Equation (2.4) can be represented in matrix form as

$$\mathbf{y}_{it} = \mathbf{X}_{it}' \mathbf{b} + e_{it} \quad (2.5)$$

where  $\mathbf{b}$  be the  $m \times 1$  vector of parametric coefficient of  $m \times 1$  vector of covariate  $\mathbf{X}_{it}$ . Analogous expression of Equation (2.5) is

$$\text{logit}(p_{it}) = \mathbf{X}_{it}' \mathbf{\beta},$$

which is an extended version of Equation (2.3).

### III. Generalized Method of Moments Estimation Procedure

Suppose a model of linear panel data in Equation (2.5) is expressed as

$$f(y_{it} | \mathbf{X}_{it}; \mathbf{\beta}) \quad (3.1)$$

The fitting of model (3.1) is done via Arellano-Bond estimation technique. The Arellano-Bond estimators are derived using the GMM estimation.

Let  $\mathbf{\beta}$  be a vector of unknown parameters which are to be estimated,  $\mathbf{X}_{it}$  be a vector of random variables and  $\mathbf{f}(\cdot)$  be a vector of functions then a population moment condition takes the form

$$E[f(y_{it} | \mathbf{X}_{it}; \mathbf{\beta})] = 0 \quad (3.2)$$

for all  $t$ . The Generalized Method of Moments estimator based on (3.2) is the value of  $\mathbf{\beta}$  which minimizes:

$$\mathbf{Q}_T(\mathbf{b}) = \mathbf{T}^{-1} \sum_{t=1}^T \dot{\mathbf{a}}(\mathbf{y}_{it} | \mathbf{X}_{it}; \mathbf{b})' \mathbf{W}_t \mathbf{T}^{-1} \sum_{t=1}^T \dot{\mathbf{a}}(\mathbf{y}_{it} | \mathbf{X}_{it}; \mathbf{b})$$

where  $W_T$  is a positive semi-definite matrix which may depend on the data but converges in probability to a positive definite matrix of constants (Hall, 2005).

GMM involves choosing parameter estimators to minimize a quadratic form in a weighting matrix,  $W_T$ , and the sample moment  $\frac{1}{T} \sum_{t=1}^T f(y_{it} | X_{it}; b)$ . The restrictions on the weighting matrix are required to ensure that  $Q_T(\beta)$  is a meaningful measure of distance. The estimation procedure was discussed in Wooldridge (2001) and in Arellano and Bond (1991)

#### IV. Application in a Panel Data of a Clinical Trial on Acute Non-Bloody Diarrhea

The GMM is applied for the estimation and testing to a model of drug efficacy and safety using a panel data of a randomized clinical trial (RCT) of acute non-bloody diarrhea of 3-24-month old infants and children. Randomization was aimed to control the effect of other intervening variables and other sources of biases that may affect the results of the trials. All subjects were given equal opportunity to be included in either of the two treatment drugs. There were 35 patients in each treatment group. Group 1 received drug A, which was a combination of OMX plus ORS, with a dosage of one capsule twice a day for a five-day treatment drug. Group 2 received drug B, which was ORS only. The treatments followed CDD protocol for both fluid therapy and nutrition. The detail of RCT on diarrhea was discussed in Magallanes (2007).

The data set was taken from the ten-day follow-up of RCT on diarrhea with respect to the occurrence of diarrhea per patient using a treatment drug. Suppose a dynamic drug efficacy and safety equation is of the form

$$\text{logodds}_t = \alpha * \text{logodds}_{i,t-1} + \beta_0 + \beta_1 * \text{age}_{it} + e_t$$

where  $\text{logodds}_t$  is the  $\log(\text{odds})$  of  $i$ th drug on day  $t$ ,  $i = 1, 2$   $t = 1, 2, \dots, 10$ , and the covariate  $\text{age}_{it}$  be the mean age of the 35 subjects of  $i$ th drug on day  $t$ .

The GMM estimates of the coefficients of the drug efficacy and safety model is given in Table 1. The estimated coefficients are derived via Arellano-Bond estimation procedure. The coefficient of  $\text{logodds}_{i,t-1}$  is statistically significant in the fitted model while the  $\text{age}_t$  is not significant. A 95% confidence interval of the parametric coefficient of  $\text{logodds}_{i,t-1}$  is [0.3876, 1.1862], which is significantly different from zero. Hence, the  $\log(\text{odds})$  of the  $i$ th drug on day  $t-1$  has a significant effect in the model of drug efficacy and safety on day  $t$ . This implies that the drug efficacy and safety equation is a dynamic linear panel data model. The

lagged logodds reported is up to day t-1 because it is the only lagged dependent variable has a significant effect in the fitted model.

Table 1. Dynamic Drug Efficacy and Safety Equation.

Variable	Coefficient	Z	Pr >  Z	95% Confidence Interval
logodds <sub>i,t-1</sub>	0.7869	3.86	0.000	[0.3876,1.1862]
age <sub>it</sub>	-0.0068	-0.16	0.871	[-0.0890,0.0754]

Table 2 presents the tests on over-identifying restriction, on no autocorrelation of orders 1 and 2, and on zero parametric coefficients except the constant. Sargan test fails to reject the over-identifying restriction which indicates the homoskedasticity of the data set. The non-rejection of the over-identifying restriction began with the homoskedastic version of this model and then the robust case for the coefficients is considered. Using the Arellano-Bond test that average autocovariance in residuals of order 1 is zero, the null hypothesis of no first-order autocorrelation is rejected. In addition, the null hypothesis of no second-order autocorrelation is also rejected.

Table 2. Test on Over-identifying Restriction, on No Autocorrelation of Orders 1 and 2, and on Zero Parametric Coefficients Except the Constant

Test	Test Statistic Value	P-value
Over-identifying Restriction	Chi2(35) = 9.43 <sup>a</sup>	1.0000
No autocorrelation (of order 1)	Z = -1.24 <sup>b</sup>	0.2137
No autocorrelation (of order 2)	Z = 1.32 <sup>b</sup>	0.1865
All Parametric Coefficients, except constant, are 0	Chi2(1) = 14.92	0.0001

<sup>a</sup> Sargan Test

<sup>b</sup> Arellano-Bond Test

The parametric coefficients of lagged dependent variable logodds<sub>i,t-1</sub> and the covariate age<sub>it</sub>, except the constant, given in Table 2 are significantly different from zero. This appears that the fitted dynamic model of drug efficacy and safety is described by the characteristics of lagged logodds<sub>i,t-1</sub> and covariate age<sub>it</sub>. Therefore, the drug efficacy and safety equation is expressed as

$$\text{logodds}_{it} = 0.7869 \cdot \text{logodds}_{i,t-1} - 0.0068 \cdot \text{age}_{it}$$

## V. Conclusion

The estimation of linear panel data model is discussed in this paper using the GMM.

The RCT on diarrhea was used in the study to perform the estimation techniques of GMM. The GMM is used because it is an improvement estimation technique over Ordinary Least-Squares (OLS) due to the smaller asymptotic variances, no distributional assumption involving the error and even consistent under weak distributional assumption (Wooldridge, 2001).

The method was applied to estimate the dynamic drug efficacy and safety equation using a balanced panel data of RCT on acute non-bloody diarrhea of 3-24-month old infants and children for a ten-day follow-up. The GMM estimators showed significant effect of the previous day of the log(odds) of ith drug.

## **Acknowledgement**

The author would like to express his gratitude to Dr. Erniel B. Barrios for inspiring him to write the paper, for providing reading materials and direction of the paper, and for extending his expertise.

## References

- Anderson, T.W. and Hsiao, C. (1981). "Estimation of dynamic models with error components". *Journal of the American Statistical Association*. 76(375); 598- 606.
- Arellano, M. And Bond, S. (1991). "Some tests of specification for panel data: Monte Carlo evidence ad an application to employment equations". *Review of Economic Studies*. 58, 277-297.
- Hall, A.R. (2005). *Generalized Method of Moments*. New York: Oxford University Press
- Honore, B.E. And E. Tamer. (2006). "Bounds on Parameters in Panel Dynamic Discrete Choice Models". *Econometrica*. 74(3), 611-629.
- Magallanes, A.B. (2007). A Stopping rule in the clinical trial on acute non-bloody diarrhea using a Bayesian approach. Unpublished Paper.
- Rothman, K.J. And Greenland, S. (1998). *Modern Epidemiology*. Second Edition. Philadelphia: Lippincott-Raven Publishers.
- STATA Corp. (2003). *Cross-Sectional Time Series. Reference Manual. Release 8*. College Station: Stata Press Publication.
- Wooldridge, J.M. (2001). "Applications of Generalized Method of Moments Estimation". *Journal of Econometric Perspective*. 15(4), 87-100
- Zeger, S.L., Yee, L.K., And Self, S.G. (1985). "The analysis of binary longitudinal data with time-independent covariates". *Biometrika*. 72(1); 31 - 38.