

Dynamic Panel of Count Data with Initial Conditions and Correlated Random Effects
: Application for Health Data

Sungjoo Yoon

Department of Economics, Indiana University

April, 2009

Key words: dynamic panel model, count data, initial conditions problem, correlated random effects.

JEL classification: C23, C51, I10.

INTRODUCTION

Unlike the linear cases, the numbers of papers on the dynamics of count data are not that much. It might be because of the properties of count data. Count data is composed of zero and positive integers, where the zeros make it difficult to make dynamic model. However, there are some papers that show several ways to bring dynamics into a count data model which is nonlinear.

(Literature review) Hausman et al. (1984), Cameron and Trivedi (1998), Bockenholt (1999) and so on.

One of the well known model is the ‘dynamic fixed effects panel model by Blundell, Griffiths, and Windmeijer (1995, 2002)’, which is called linear feedback model. This model uses first difference transformations and GMM for estimations. It assumes that the lagged dependent variable has linear relationship with current dependent variables, and the other covariates are related by nonlinear ways, exponential function. Therefore, we cannot use prepackaged computer program like STATA and SAS for the estimations. Also, the model employs fixed effects model, but if we use fixed effects model, we lose information of time-invariant variables like gender, race, and so on, which is regarded as important variables in microeconometrics. Finally, the initial values are important in the case of short panel. For example, if the number of doctor visits of this quarter is 5, then the number of doctor visits for next several quarters may highly related to the number of this quarter. However, the previous studies have not explicitly included the effects of initial values into their models. So, what I want to do in this paper is to make model which has low estimation cost, explicitly includes initial values as one of its covariates, and use adjusted random effects model where individual specific effects are related to covariates.

(I will add the summary of result here)

(The composition of this paper is written here. The big picture of this paper is that I first make model. Then do simulations with the model, and show that my model is better than others. Finally apply my model to real data).

MODEL

First, consider very simple panel model of count data which does not consider dynamics.

$$E[y_{i,t}|\alpha_i, x_{i,t}] = \exp(x'_{i,t}\beta + \alpha_i) = \exp(\alpha_i)\exp(x'_{i,t}\beta), \quad \text{where } \exp(\alpha_i) \sim \text{Gamma}(\mu_\alpha, \sigma_\alpha^2).$$

Here, α_i is individual specific effect, unobservable heterogeneity. There are several ways to bring lagged dependent variable into the model to consider dynamics. One of my objective in this paper is to find methods of estimation having low cost, so I just put the lagged value into the exponential function for using MLE which is in the prepackaged computer program in STATA or SAS. Then it becomes

$$E[y_{i,t}|y_{i,t-1}, \alpha_i, x_{i,t}] = \exp(\gamma y_{i,t-1} + x'_{i,t}\beta + \alpha_i) = \exp(\alpha_i)\exp(\gamma y_{i,t-1} + x'_{i,t}\beta) \\ , \text{where } \exp(\alpha_i) \sim \text{Gamma}(\mu_\alpha, \sigma_\alpha^2).$$

Panel model is divided by three parts on the basis of the relations between the individual specific effect and covariates, $x_{i,t}$. If the individual specific effect is regarded as constant in the model, then it is pooled model; if it is related to the covariates, then it is fixed effect model; if it is not related to the covariates, then it is random effect model. The fixed effect model is quite reasonable, but the assumption of random effect model is not reasonable. For example, consider that the individual specific effect includes gene, and one of covariates is health status. If this is the case, there should be some relation between those two. The fixed effect model catches this relationship, but the random effect model does not. However, if we use fixed effect model, we lose the information of time-invariant variables like gender, race, and so on. Also, it is known that in the case of nonlinear model, the fit of fixed effects model is not so good. So, we may need to use random effects model. Considering the unreasonable assumption of random effects model, I employ the correlated random effects model (Mundlak, 1978), *i.e.*, I add a variable, z_i , which is the average of $x_{i,t}$ over time, t . Finally, I add one more variable, $y_{i,0}$, which is the initial value of individual i . At the simple model, the two added variables are considered into the α_i , so now we can say

$$\alpha_i = \tau_0 y_{i,0} + \tau_1 z_i + a_i$$

, where a_i is purely random effect. Finally, my model becomes

$$E[y_{i,t}|a_i, y_{i,0}, y_{i,t-1}, x_{i,t}, z_i] = \exp(\gamma y_{i,t-1} + x'_{i,t}\beta + \tau_0 y_{i,0} + \tau_1 z_i + a_i) \\ = \exp(a_i)\exp(\gamma y_{i,t-1} + x'_{i,t}\beta + \tau_0 y_{i,0} + \tau_1 z_i) \\ , \text{where } z_i = \frac{1}{T+1} \sum_{t=0}^{t=T} x_{i,t} \text{ and } \exp(a_i) \sim \text{Gamma}(\mu_a, \sigma_a^2).$$

Now, we can use low-cost method of estimation, MLE; explicitly includes the effect of initial value into the model; and employ random effect model.

(I will add mixture model for explanation later; covariate follows poisson distribution, and the unobserved heterogeneity follows gamma distributions, so the model is explained by mixture distribution.

Next, I will add the reason why we use gamma distribution with shape parameter equal to one, which is related to zero-inflation problems.

In addition, I plan to add some explanation of solving initial value problems; Heckman's and Wooldridge's methods, and the reason why I choose Wooldridge's one.

Finally, I will add some explanation of Mundlak's correlated random effect model).

SIMULATION

For checking the validity of my model, I do the simulation before applying the model to real data. Here, I generate very general dataset which includes the effects of initial values and so on. Then, I do the simulation with (1) full model which includes both $y_{i,0}$, and z_i , (2) initial-only model, (2) correlation-only model, and (4) without model which does not includes any of $y_{i,0}$, and z_i . In this simulation, the number of time periods is 9, from $t=0$ to $t=8$, so the correlated random effect comes to be $z_i = \frac{1}{9} \sum_{t=0}^{t=8} x_{i,t}$. The covariate $x_{i,t}$ is simply composed of $\beta_0 + \beta_1 x_{i,t}$, and $x_{i,t}$ s are randomly drawn from Normal (0,1). Note that the value of $x_{i,t}$ are fixed over replications to reduce variations from the random covariates, and focus on the lagged dependent variables. The initial values of each individual, $y_{i,0}$, are randomly drawn from mixture distribution and the parameter of the poisson distribution is $\mu_{i,0}$, which is equal to $\exp(\beta_0 + \beta_1 x_{i,0} + a_i)$. Finally the pure random effect, a_i , follows normal distribution with mean zero and variance σ_a^2 (Here, I employ normal distribution rather than gamma distribution because it is easier to control the variance of normal distribution rather than that of gamma distribution. Textbook mentions that the distribution of a_i is alternative to the gamma distribution of $\exp(a_i)$. However, I think that now I may be able to manage the variance in gamma distribution, so I plan to the following simulations again with gamma distribution. Then, I may get better results. I hope so).

[Table 1] shows the results of simulation where the number of simulation is 100. In each simulation, we get the estimates and calculate the average of estimates over 100 simulations. As you see, the averages of estimates overall approach to the true value, which I give in the model, as the sample size, N, increases from 300 to 1000 (It is not clear, so I may need to adjust the sample size, and do simulations again like N=50, 100, 200, 1000. Actually, I will try the simulation with gamma distribution rather than normal distribution for the unobserved heterogeneity. In this case, I expect a little bit more clear results).

[Table 1] Fit of estimation by sample size, N.

Variable	True value	N=300	N=600	N=1000
ave(gamma_hat)	0.1	0.0990046	0.1009095	0.1019942
ave(beta_hat)	0.1	0.1062831	0.0985971	0.1015778

ave(tau0_hat)	0.05	0.0566454	0.0558687	0.0590085
ave(tau1_hat)	0.05	0.021696	0.0651966	0.0363104
ave(cons)	0.5	0.5008193	0.4854546	0.488273

[Table 2] shows the comparison of four models when we use the very general dataset I generate. Here, the number of simulation is 100 and the sample size of each model is 1000. As you see, full model which consider both initial effect and use correlated random effect is better than other models.

[Table 2] Comparison of models.

N=1000	True value	Full	Initial only	Correlation only	Without
ave(gamma_hat)	0.1	0.1019942	0.1029056	0.1091489	0.1112947
ave(beta_hat)	0.1	0.1015778	0.1051418	0.0937829	0.105696
ave(tau0_hat)	0.05	0.0590085	0.0575164		
ave(tau1_hat)	0.05	0.0363104		0.023658	
ave(cons)	0.5	0.488273	0.5033911	0.5872398	0.5912744

APPLICATION: HEALTH DATA

I have not started this analysis because I have not finished the simulation part yet. Let me just explain the dataset I have. The data has eight-time period, two years, quarterly data. The dependent variable is number of doctor visits for each quarter, and covariates are standard socio-demographic variables; health insurance status variables; and health status variables. And I roughly checked that the coefficient of lagged dependent variable is about 4%. I simulated with 10%, and checked that it does not explode with the coefficient value. So, I may use my model for analyzing this health data.

ADDITIONAL WORKS

If time is permitted, then I would like to compare my model and previous linear feedback model using my dataset which is very generous.

CONCLUSION

REFERENCES

Wiji Arulampalam, Mark B. Stewart. 2009. Simplified implementation of the heckman estimator of the dynamic probit model and a comparison with alternative estimators.

Cameron, Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge University Press.

Cameron, Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

Cameron, Trivedi. 2009. *Microeconometrics: Using STATA*. STATA Press.

Wooldridge JM. 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* 20: 39-54.