36-Random effect

By Jaka Nugraha
RANDOM EFFECT MODEL AND GENERALIZED ESTIMATING EQUATIONS FOR
BINARY PANEL RESPONSE

Jaka Nugraha

Dept. of Statistics, Islamic University of Indonesia, Kampus Terpadu UII, Jl. Kalirang Km.14,
Yogyakarta, Indonesia

email: jk.nugraha@gmail.com

Abstract

Panel data models are widely used in empirical analysis because they allow researchers to control for unobserved individual time-invariant characteristics. However, these models pose important technical challenges. In particular, if individual heterogeneity is left completely unrestricted, and then estimates of model parameters in nonlinear and/or dynamic models suffer from the incidental parameters problem. This problem arises because the unobserved individual characteristics are replaced by inconsistent sample estimates, which, in turn, biases estimates of model parameters. Logit model or probit model on panel data with using univariate approximation (neglect correlation) result consistent estimator but not efficient. In many cases, data are multivariate or correlated (e.g., due to repeated observations on a study subject or for subjects within centers) and it is appealing to have a model that maintains a marginal logistic regression interpretation for the individual outcomes.

In this paper, we studied modeling binary panel response using Random Effects Model (REM). Using Monte Carlo Simulation, we research correlations effects to maximum likelihood estimator (MLE) of random effects model. We also compare MLE of REM to Generalized Estimating Equations (GEE) of logit model. Data were generated by using software R.2.8.1 as well as the estimation on the parameters. Based on the result, it can be concluded that (a) in some value of individual effect, random effects model is more better GEE. (b) REM can be accommodating individual effects and closer to parameter than the other. (c) REM is appropriate method to estimate covariances of utility at individual effect having value about one.

Keywords: Random Utility Models, Maximum Likelihood Estimator Generalized Estimating Equations, Logit Models, Probit Models

1. Introduction

Panel data models are widely used in empirical analysis because they allow researchers to control for unobserved individual time-invariant characteristics. However, these models pose important technical challenges. In particular, if individual heterogeneity is left completely unrestricted then estimates of model parameters in nonlinear and/or dynamic models suffer from the incidental parameters problem. This problem arises because the unobserved individual characteristics are replaced by inconsistent sample estimates, which, in turn, biases estimates of model parameters (Greene, 2003). Liang and Zeger (1986), shown that Logit model or Probit
model on panel data with using univariate approximation (neglect correlation) result consistent estimator but not efficient. In many cases, data are multivariate or correlated (e.g., due to repeated observations on a study subject or for subjects within centers) and it is appealing to have a model that maintains a marginal logistic regression interpretation for the individual outcomes. Commonly used logistic random effects models do not have this property, since the logistic structure is lost in integrating out the random effects. An alternative is to use a marginal analysis that avoids complete specification of the likelihood (Liang and Zeger, 1986; Prentice, 1988; Lipsitz et al.). Prentice (1988) proposed modeling strategy by GEE to obtain consistent estimator and normal asymptotic. GEE are hindered multiple integral by marginal distribution.

Nugraha et al. (2008) have tested Logit Model in multivariate binary response using Monte Carlo simulation. They concluded that GEE more proper on height correlation, although estimators of correlation was underestimated. MLE of Probit Model could not derive by analytic, because the likelihood functions formed a multiple integral. Simulation approximation to compute multiple integral caused bias. The others problem on Probit Model are the log function likelihood not global concave, so there are no one solutions. Simulation methods rely on approximating an integral (that does not have a closed form) through Monte Carlo integration. Draws are taken from the underlying distribution of the random variable of integration and used to calculate the numeric integral. Simulated maximum log likelihood estimation is a common estimator used for Probit Model and random effects model. Such estimators exhibit a non-negligible bias when too few draws are used in estimation, and prior research exists regarding the magnitude and properties of this bias with respect to quasi-random draws (Bhat, 2001). Train (1999) provide further evidence of the benefits of intelligent drawing techniques such as Halton and Shuffled Halton, which require fewer numbers of draws than pseudo-random in order to uncover identification issues.

In this paper, we studied modeling binary panel response using random effects model. Using Monte Carlo Simulation, we research correlations effects to maximum likelihood estimator (MLE) of Random Effects model. We also compare MLE on Random Effects model to GEE on Logit Model.

2. GEE Model

In the panel response within exponential family distribution, Liang dan Zeger (1986) proposed the GEE model. For the binary response \( Y_{it} \) with each \( Y_{it} \) binary value (dichotomous), both link logit and link probit can be utilized for GEE model. Conroyannis et al. (2001) has been constructed probit model on binary panel data by the model of:

\[
Y_{it} = \beta X_{it} + \xi_t + \epsilon_{it} \quad \text{for } i=1,...,n \text{ and } t=1,...,T
\]

\( \xi_t \) is individual effect within the normal distribution and mean value of null and the variance of \( \sigma_{\xi_t}^2 \). Whereas \( \epsilon_{it} \sim N(0, \sigma^2) \) and independent with \( \xi_t \). \( X_{it} \) are vector px1 respect to the independent variable for responden i on t period. \( \beta \) is a parameter vector in px1 size. Individual probability \( i \) for making a decision series was calculated by using conditional probability.

We assume that each of \( n \) individual observed \( T \) times. \( Y_{it} \) is the \( t^{th} \) response on \( i^{th} \) individual/subject and each response are binary. So, for response \( t \) individual can be

\( Y_i = (Y_{i1},...,Y_{iT}) \)

that is vector 1xT. \( Y_{it} = 1 \) if \( i^{th} \) subject and \( t^{th} \) response choose the first alternative and \( Y_{it} = 0 \) if choose the second alternative. Each subject have covariate \( X_{it} \) (individual characteristic) dan covariate \( Z_{ij} \) (characteristic of alternative \( j=0,1 \)). To simplify, we choose one of individual characteristic and one of characteristic of alternative. Utility of subject \( i \) choose alternative \( j \) on response \( t \) is
\[ U(i_t) = V(i_t, i_t) + e(i_t) \quad \text{for} \quad t=1,2,...,T; \quad i=1,2,...,n; \quad j=0,1 \] (2)

with \( V(i_t) = \beta_0 + \beta_1 x_i + \gamma z_{i_t} \).

By assumption that decision makers choosing alternative based on maximum utility, model can be represented in different of utility,
\[ U(i_t) = V(i_t, i_t) + e(i_t) \]

(3)

with \( V(i_t) = (V(i_t) - V(a)) \) and \( e(i_t) = (e(i_t) - e(a)) \).

Expectation of eq. (3) are
\[ E(e(i_t)) = E(e(i_t)) - E(e(a)) = 0.5772 - 0.5772 = 0 \]
\[ E(U(i_t)) = E(U(i_t)) - E(V(a)) \]

and theirs variances are
\[ \text{Var}(e(i_t)) = \text{Var}(e(i_t)) + \text{Var}(e(a)) = \frac{\pi^2}{6} + \frac{\pi^2}{6} = \frac{\pi^2}{3} \]
\[ \text{Var}(\alpha(i_t)) = \text{Var}(\alpha(a)) + \text{Var}(\alpha(i_t)) = 2\sigma^2 \]
\[ \text{Var}(U(i_t)) = \text{Var}(U(a)) + \text{Var}(U(i_t)) = 2\sigma^2 + \frac{\pi^2}{3} \]

Covariance and Correlation among utilities are
\[ \text{Cov}(U(i_t) U(a)) = \text{Cov}(\alpha(i_t), \alpha(a)) = 2\sigma^2 \text{ for all } t \neq a \]

\[ \text{Cor}(U(i_t) U(a)) = \frac{2\sigma^2}{\sqrt{2\sigma^2 + \frac{\pi^2}{3}}} \]

Probability of subject \( i \) choose \( (y_{i1},...,y_{iT}) = 1 \) is
\[ P(y_{i1} = 1,...,y_{iT} = 1) = \int \text{f}(-V(i_t) < e(i_t)) \cdot f(e(i_t)) \cdot de(i_t) \quad \forall t \] (4)

This probability value is multiple integral and depending on parameters \( \beta, \gamma \) and distribution of \( e(i_t) \) (Train, 2003).

The logit model can be derived by assumption that \( e(i_t) \) have Extreme Value Type I distribution (Gumbel) and independence each other (all \( i, j \) and \( t \)). Probability of subject \( i \) choose \( j=1 \) for response \( t \text{th} \) is
\[ P(y_{it} = 1) = \pi_{it} = \frac{\exp(V(i_t))}{\exp(V(i_a)) + \exp(V(i_t))} = \frac{\exp(V(i_t))}{[1 + \exp(V(i_t))]} \] (5)

with
\[ V(i_t) = \beta_0 + \beta_1 x_i + \gamma z_{i_t} \quad \text{for} \quad t=1,2,...,T; \quad i=1,2,...,n; \quad j=0,1. \]

On Logit Model, GEE are easier to implement than MLE. GEE use approximation by marginal distribution and can be represented by
\[ G(\theta) = \sum_{i=1}^{n} W(i) (Y(i_t) - \pi(i_t)) = 0 \] (6)

with \( W(i) = \text{diag}(\begin{bmatrix} 1 & 1 \\ X_j & ... & X_j \\ (Z_{i,t} - Z_{i,a}) & ... & (Z_{i,T} - Z_{i,a}) \end{bmatrix}) \)

\[ \Delta_j = \text{diag}(\begin{bmatrix} \pi(i_1)(1-\pi(i_1)) & ... & \pi(i_T)(1-\pi(i_T)) \end{bmatrix}) \]
S_i = A_i^{1/2} R_i A_i^{1/2} \text{ with } A_i^{1/2} = \begin{pmatrix} \sqrt{\text{Var}(Y_{i1})} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & \sqrt{\text{Var}(Y_{iT})} \end{pmatrix}

with \( Y_i = (Y_{i1}, \ldots, Y_{iT}) \); \( \pi_i = (\pi_{i1}, \ldots, \pi_{iT}) \). Estimators GEE are solving equations (6) on sample data \( W \) (Nugraha et al., 2008).

GEE on Probit Model are solving of estimating equation

\[ G(\theta) = \sum_{i=1}^{n} W_i \Delta_i S_i^{-1}(Y_i - \pi_i) = 0 \] (7)

with \( \pi_i = \Phi(\psi(Y_i)) \); \( \Delta_i = \text{diag}(\phi(Y_i)) \). Estimations of parameter correlations are underestimated. GEE on probit model are equivalent to GEE on Logit Model.

3. Random Effects Model

From the equation of utility difference (3), we added the individual effect \( \alpha_i \)

\[ U_i = V_i + \alpha_i + \epsilon_i \] (8)

\( \alpha_i \) is effect of individual \( i \) having normal distribution, \( \alpha_i \sim \text{NID}(0, \sigma^2) \) and independent to \( \epsilon_i \). \( \epsilon_i \) have Extreme Value Distribution.

Based on equation (8), we will estimate parameters \( (\sigma, \beta, \gamma) \) for \( t=1, \ldots, T \). By equation (5), we have conditional probability:

\[ g_i = p(y_{it} = 1 | \alpha_i) = p(\epsilon_i < V_i + \alpha_i) = \pi_i \alpha_i = \frac{\exp(V_i + \alpha_i)}{1 + \exp(V_i + \alpha_i)} \]

\[ = \frac{\exp(V_i + \alpha_i)}{1 + \exp(V_i + \alpha_i)} \] (9)

Marginal probabilities from equation (4) for Random Effect Model are

\[ P(y_{it} = 1) = \pi_i = \int \mathcal{N}(y_i \mid \alpha_i) f(\alpha_i) d\alpha_i \]

\[ = \int_{-\infty}^{\infty} \frac{\exp(V_i + \alpha_i)}{1 + \exp(V_i + \alpha_i)} \phi(\alpha_i) d\alpha_i \] (10)

\( \phi(\alpha_i) \) is standard normal density.

\[ P(y_{it} = 1, \ldots, y_{iT} = 1) = \prod_{t=1}^{T} P(y_{it} \mid \alpha_i) f(\alpha_i) d\alpha_i \]

\[ = \prod_{t=1}^{T} \frac{\exp(V_i + \alpha_i)}{1 + \exp(V_i + \alpha_i)} \phi(\alpha_i) d\alpha_i \] (11)

So,

\[ P(y_{it}, \ldots, y_{iT}) = \prod_{t=1}^{T} \left( \frac{\exp(V_i + \alpha_i)}{1 + \exp(V_i + \alpha_i)} \right)^{y_{it}} \left( 1 - \frac{\exp(V_i + \alpha_i)}{1 + \exp(V_i + \alpha_i)} \right)^{1-y_{it}} \phi(\alpha_i) d\alpha_i \]

MLE of parameters \( (\beta, \gamma, \sigma) \), \( t=1, \ldots, T \), can be obtained from likelihood function.
\[ L(\beta, \gamma, \sigma) = \prod_{i=1}^{n} \prod_{t=1}^{T} \left( g_{it} \right)^{\alpha_i} \left( 1 - g_{it} \right)^{-\alpha_i} \phi(\alpha_i) \hat{e_i} \]  

Then the log-likelihood function is

\[ LL = \log L(\beta, \gamma, \sigma) = \sum_{i=1}^{n} \log \left( \prod_{t=1}^{T} \left( g_{it} \right)^{\alpha_i} \left( 1 - g_{it} \right)^{-\alpha_i} \phi(\alpha_i) \right) \hat{e_i} \]  

4. Generating Data Simulation and Result

We will generate simulation data with \( T=3, n=1000 \). Then, the equations of utility are

\[ U_{it} = \alpha_{it} + \beta_{0i} + \beta_{1i}X_{it} + \gamma Z_{it} + \epsilon_{it} \quad \text{and} \]

\[ U_{it} = \alpha_{it} + \beta_{0i} + \beta_{1i}X_{it} + \gamma Z_{it} + \epsilon_{it} \]

for \( i=1,...,N; j=0,1 \) and \( t=1,...,3; \alpha_{it} \sim \text{Extreme Value Type I, } \alpha \sim N(0, \sigma^2) \).

Equation (14) can be presented in difference of utility \( U_i = U_{it} - U_{i0} \). On Logit Model, equations of utility difference are

\[ U_i = \beta_0 + \beta_1X_i + \gamma Z_i + \epsilon_i \]

or

\[ U_{it} = V_{it} + \alpha_i + \epsilon_i; U_{it} = V_{i1} + \alpha_i + \epsilon_i; U_{it} = V_{i3} + \alpha_i + \epsilon_i \]

with

\[ V_i = (V_{i0} - V_{it}) = \beta_0 + \beta_1X_i + \gamma Z_i \]

\[ Z_i = (Z_{i0} - Z_{it}); \beta_0 = \beta_{00} - \beta_{01}; \beta_1 = \beta_{10} - \beta_{11}; \alpha_i = (\alpha_{i0} - \alpha_{i1}) \]

We generate data on \( \beta_0 = 1; \beta_1 = 0.5, \gamma = 0.3 \) and some of variance \( \sigma^2 = 0; 0.5; 1; 2; 4, 6 \) using program R.2.8.1. From the data simulation, we built the Logit Model using three approximations. First, we assume independence each other for all \( i,j \) and \( t \) (independent logit model) and estimate parameters using MLE \( (\theta_{MLE}) \). Second, we estimate parameter using GEE \( (\theta_{GEE}) \). The last approximation is Random Effect Model using MLE \( (\theta_{MLE}) \).

Results of the simulations presented in Figure 1 to Figure 10.

![Figure 1. Estimator of \( \sigma^2 \)](image)

![Figure 2. Estimator of \( \beta_0 \)](image)

Based on simulation result representing at Figure 1., estimator of \( \sigma^2 \) using Random effect Model \( (\theta_{MLE}) \) more accurate than GEE \( (\theta_{GEE}) \), \( \theta_{MLE} \) have value closer to real value of parameter than \( \theta_{GEE} \).

Overall, estimator of \( \beta_0, \beta_1 \) and \( \gamma \) in MLE as same as GEE (see Figure 2 to Figure 10).
We can see that on $\sigma^2 = 0$ (no effect individual) estimators by three approximations are same. On general, estimator of independent Logit Model ($\hat{\theta}_{MLE}$) and estimator of GEE ($\hat{\theta}_{GEE}$) are not different but increasing of $\sigma^2$ impact to increasing bias of estimator. On data having individual effect (see Figure 2 to Figure 10), the random effect model ($\hat{\theta}_{MLRE}$) better than $\hat{\theta}_{MLE}$ and $\hat{\theta}_{GEE}$.

By GEE, we estimate coefficient regressions and correlation among alternative. Using effect random model we can estimate coefficient regressions, individual effect and covariates/correlations among alternative. On value of individual effect $\sigma^2$ 1, random effect model can estimate covariates of utility more appropriate than the others value of $\sigma^2$.

5. Conclusion

On modelling binary panel response, we can use Random Effects Model. This model use the Extreme Value distribution and the standart normal distribution. The model is

$$P(y_{it} = 1 | y) = \int \prod_{t=1}^{T} (g_{it})^{y_{it}} \left(1 - g_{it}\right)^{1-y_{it}} \phi(\alpha_i) d\alpha_i$$

with $g_{it} = \frac{\exp(V_{it} + \alpha_i)}{[1 + \exp(V_{it} + \alpha_i)]}$. The log-likelihood function is

$$\log L(\beta, \gamma, \sigma) = \sum_{i=1}^{n} \log \left( \int \prod_{t=1}^{T} (g_{it})^{y_{it}} \left(1 - g_{it}\right)^{1-y_{it}} \phi(\alpha_i) d\alpha_i \right)$$

Based on simulation, Random Effects Model more appropriate than GEE.

Reference


36-Random effect

ORIGINALITY REPORT

0%

SIMILARITY INDEX

PRIMARY SOURCES

| EXCLUDE QUOTES | ON |
| EXCLUDE BIBLIOGRAPHY | ON |
| EXCLUDE MATCHES | < 15 WORDS |
Berita Acara Hasil Pengecekan Keaslian Karya Ilmiah
Atas Nama Jaka Nugraha, Dr., S.Si., M.Si. Untuk kenaikan Jabatan
Dari Lektor Kepala (400 AK) ke Guru Besar (850 AK)

Pada tanggal 23 Oktober 2020 telah dilakukan pengecekan Originality atau Similarity terhadap karya Ilmiah Dosen Tetap Universitas Islam Indonesia:

<table>
<thead>
<tr>
<th>NO</th>
<th>KARYA</th>
<th>REPORT ORIGINALITY</th>
<th>KETERANGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Menghasilkan karya ilmiah berupa jurnal internasional terindeks bereputasi dengan impact factor pada Sustainable Chemistry and Pharmacy, Scopus &amp; Scimagojr, SJR=0.62 pada bulan Juni 2018, dengan judul 'Preparation, characterization, and modelling activity of potassium fluoride modified hydrotalcite for microwave assisted biodiesel conversion'. ISSN: 2352-5541. Volume 8; issue 43252; pages 63-70; jumlah halaman 8</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Menghasilkan karya ilmiah berupa jurnal internasional terindeks pada The Turkish Online Journal of Design Art and Communicatio, Copernicus-DOAJ pada bulan September 2018, dengan judul 'Brown's Weighted Exponential Moving Average (B-Wema) with Levenberg-Marquardt Optimization to Forecasting Rate of Return'. ISSN: 2146-5193. Volume Special Edition; issue ; pages 1744-1749; jumlah halaman 6</td>
<td>19%</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Menghasilkan karya ilmiah berupa jurnal internasional terindeks pada International Journal of Engineering and Technology, ProQuest-DOAJ pada bulan September 2018, dengan judul 'Sentiment Analysis on Mobile Banking Application Using Naive Bayes Classifier and Association Methods'. ISSN: 2227-524X. Volume 7; issue 4.15 (Special Issue); pages 244-247; jumlah halaman 4</td>
<td>2%</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>Menghasilkan karya ilmiah berupa prosiding seminar internasional terindeks pada 2nd International Conference on</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>Judul Artikel</td>
<td>Detail Penerbitan</td>
<td>Halaman</td>
</tr>
<tr>
<td>-----</td>
<td>--------------</td>
<td>-------------------</td>
<td>---------</td>
</tr>
<tr>
<td>No.</td>
<td>Judul dan detail prosiding seminar internasional</td>
<td>Halaman</td>
<td>Persentase Validasi</td>
</tr>
<tr>
<td>-----</td>
<td>-------------------------------------------------</td>
<td>---------</td>
<td>-------------------</td>
</tr>
</tbody>
</table>

Pengecekan di atas menggunakan alat Thenticate dengan meniadakan (exclude) beberapa hal dengan ketentuan sebagai berikut:

1. Meniadakan (exclude) hasil cek kesamaan karya yang kurang dari 2 persen.
2. Meniadakan (exclude) hasil cek kesamaan karya yang disitas oleh pihak lain.
3. Meniadakan (exclude) hasil cek kesamaan karya yang terindikasi plagiasi kepada karya ilmiah yang bersangkutan.
4. Meniadakan (exclude) hasil cek kesamaan karya yang menunjukan url atau laman karya ilmiah yang bersangkutan.
5. Meniadakan (exclude) hasil cek kesamaan karya yang diupload dalam bentuk yang berbeda (online pribadi) yang terdeteksi merupakan karya sendiri bukan merupakan laman publikasi Jurnal resmi hanya untuk kepentingan sharing (seperti https://www.researchgate.net, facebook.com dll) sehingga bukan termasuk auto-plagisasi/self plagiarism.

Berdasarkan hasil pencekian di atas, maka karya ilmiah tersebut di atas dapat diteruskan usulannya ke Lembaga Layanan Dikti Wilayah V.

Yogyakarta, 23 Oktober 2020
Rektor

[Signature]

Prof. Fathul Wahid, S.T., M.Sc., Ph.D.