Should We Avoid Random Effects Model When Covariates Correlate with Group Effects? - *A Simulation Study for Binary Outcomes and Covariates*

Ta Liu, Ph.D., Battelle Center for Analytics and Public Health

Introduction

For multilevel data such as patient data where patients are nested within hospitals, or panel data where there are both cross-sectional and time components, researchers often face a decision of whether to use fixed effects or random effects models to estimate the effect of the covariates of interest. When the covariates vary systematically by groups and correlate with group effects, the estimates from random effects models are no longer reliable as demonstrated by the influential work of Hausman and Taylor (1981).

Since consistency in estimation is a much desired property and covariates are often correlated with group effects, fixed effects models often end up as the default choice in practice. However, despite the fact that fixed effects model can yield unbiased estimates by sweeping away unobserved group effects through group dummy variables, they have serious limitations. To start, fixed effects models are not efficient by discarding all information at group level and can have larger variance than random effects models. They are not able to measure group effects that may also be of research interest and policy concerns. Additional efforts such as two or three-step regressions are proposed to estimate group level variations within the framework of fixed effects models (Plumper and Troeger. 2007). The fixed effects models are also limited in not being able to make inference outside existing groups.

Is there a method that allows us to continue to use random effects models while at the same time derive consistent estimates? The answer is yes. The solution is simply to bring group averages of correlated covariates into the model to control for group level heterogeneity. The idea appears in the original article of Hausman and Taylor and has been the subject of several recent studies (Bafumi and Gelman, 2006; Ebbes, Böckenholt and Wedel, 2004). These recent simulation-based studies show that random intercept models can yield unbiased estimates with added group means. These simulations, however, are all based on linear models with normally distributed data. It is not clear whether same results apply for models with binary outcomes.

Motivation

Binary outcomes and related models are common in many applied areas. It would be interesting and useful to find out whether adding group means to random effects models can correct the bias caused by the correlation between covariates and group effects in binary models. In addition, we want to know how sample size and strength of correlation affect the behaviors of the models.

Models for Estimation in Simulation

Let $\mathbf{y_{ij}}$ be the outcome for the $\mathbf{j}^{th}$ individual in the $\mathbf{i}^{th}$ group, being 1 if the individual experiences certain event and 0 otherwise. Let $\mathbf{x_{ij}}$ be the covariate of interest that correlates with groups, and $\mathbf{z_{ij}}$ be the variable to control for other characteristics of the individual. Let $\mathbf{z_{ij}}$ not correlate with either group effects or $\mathbf{x_{ij}}$. Both $\mathbf{x_{ij}}$ and $\mathbf{z_{ij}}$ are also binary.

$$\mathbf{Pr(y_{ij}=1) = logit^{-1}(\alpha_j + \beta x_{ij} + \gamma z_{ij}),\quad for\ i=1,\ldots,n\ and\ j=1,\ldots,m}$$

where n is the number of individuals within each group and m is the number of groups, and α_j is group dummies for the fixed effect model and is subject to a normal distribution **$\alpha_j \sim N(\mu, \sigma^2)$ for j=1,…,m** in random intercepts model.

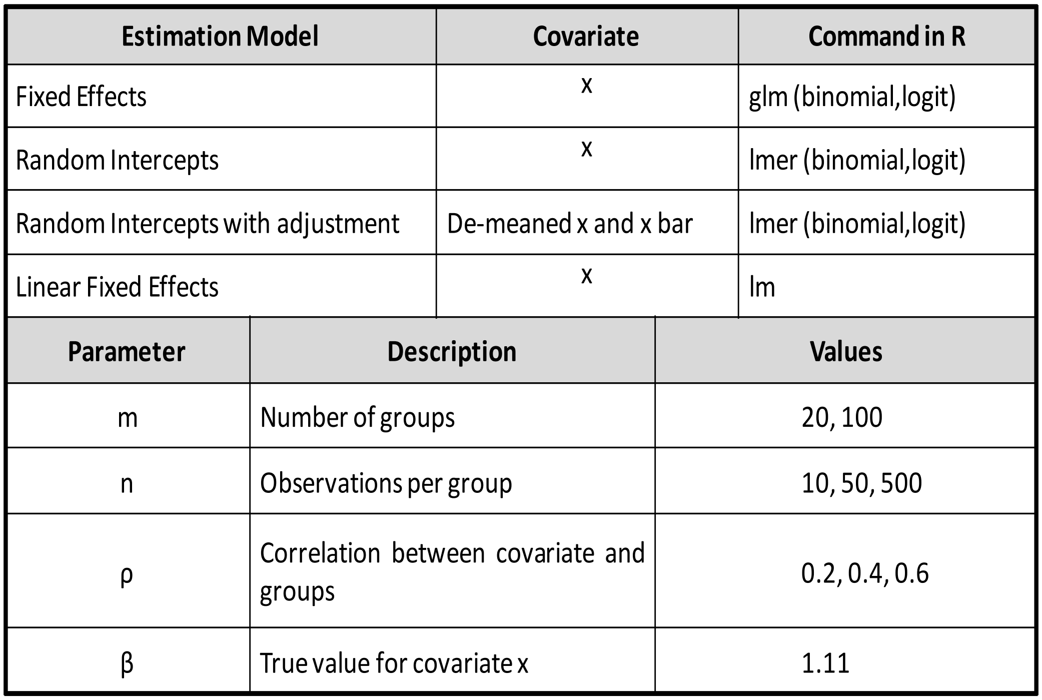
The proposed random effects model with adjustment is

$$\mathbf{Pr(y_{ij}=1) = logit^{-1}(\alpha_j + \beta\check{x}_{ij} + \gamma z_{ij} + \delta\bar{x}_j), \quad for\ i=1,\ldots,n\ and\ j=1,\ldots,m}$$

where $\bar{\mathbf{x}}_j$ is group averages of the original covariate $\mathbf{x}_{ij}$, and $\check{\mathbf{x}}_{ij}$ is the de-meaned $\mathbf{x}_{ij}$ as $\check{\mathbf{x}}_{ij} = \mathbf{x}_{ij} - \bar{\mathbf{x}}_j$. Basically, the original covariate is divided into two parts, i.e., within-group and between-group components. Since the within-group component is removed of group level variation, it is no longer correlated with group effects. It is also independent of the between-group component by design. It works as in fixed effects models by removing group level variations so its parameter estimate is consistent. In contrast to the fixed effects models, having the between-group component $\bar{\mathbf{x}}_j$ in the same model allows measurement of the covariate's contribution to group level variations and potential interactions among variables at different levels.

The fourth model is the linear probability model with group dummies. Even though linear models are typically considered inappropriate for binary outcomes, studies show that linear models offer results close to those from logistic models when the outcome is at the middle of the distribution. In addition, unlike estimates from logistic model in logits or odds ratios, estimates from linear models are invariant to omitted variables (Mood, 2010).

Table 1. Simulation Models and Parameters



Estimation Model	Covariate	Command in R
Fixed Effects	x	glm (binomial,logit)
Random Intercepts	x	lmer (binomial,logit)
Random Intercepts with adjustment	De-meaned x and x bar	lmer (binomial,logit)
Linear Fixed Effects	x	lm

Parameter	Description	Values
m	Number of groups	20, 100
n	Observations per group	10, 50, 500
ρ	Correlation between covariate and groups	0.2, 0.4, 0.6
β	True value for covariate x	1.11

Simulation Steps

- Choose number of groups and generate the same number of group effects by using random normal function with mean 0 and standard deviation of 1.
- Choose the number of individuals per group and assign group effects to each individual depending on the groups.
- Calculate the sample size = number of groups x number of individual in each group.
- Generate covariate x with a certain degree of correlation through the probability in random binomial function by multiplying group effects.
- Generate covariate z with a fixed probability (0.4) using random binomial function.
- Assign true values of parameters (1.11 to beta for x and 2.22 to gamma for z).
- Generate binary outcome by using random binomial function with a probability equal to the inverse logit of linear predictors of group effects and covariates x and z.
- Calculate group averages and de-means for the covariate of x.
- Run four models for a new randomly generated outcome data in each simulation using three regression functions in R, glm for fixed effects, lmer for random and adjusted random effects, and lm for linear probability model.
- Collect two statistics from each model, beta estimate and t or t-equivalent statistics for x or de-meaned x. Save instances of randomly generated data for Hausman test in Stata.

Key Results

The adjusted random intercepts model is able to correct for the bias in the covariate correlated with group effects and does as well as the fixed effects models in all scenarios. The histograms below show the similarity in estimates and t statistics between the first row (fixed effects) and the third row (adjusted random effects) in two scenarios with high correlations (ρ=0.6). The mean (the red line) of simulated estimates is close to the true value of 1.11. When the sample is larger, even the estimate in the random effects model without adjustment is getting closer to the true value.

Figure 1. Histograms for β (True Value = 1.11) and t Statistics in Three Estimation Models for 20 Groups with 50 Observations and High Correlation (ρ=0.6) with 1000 Simulations

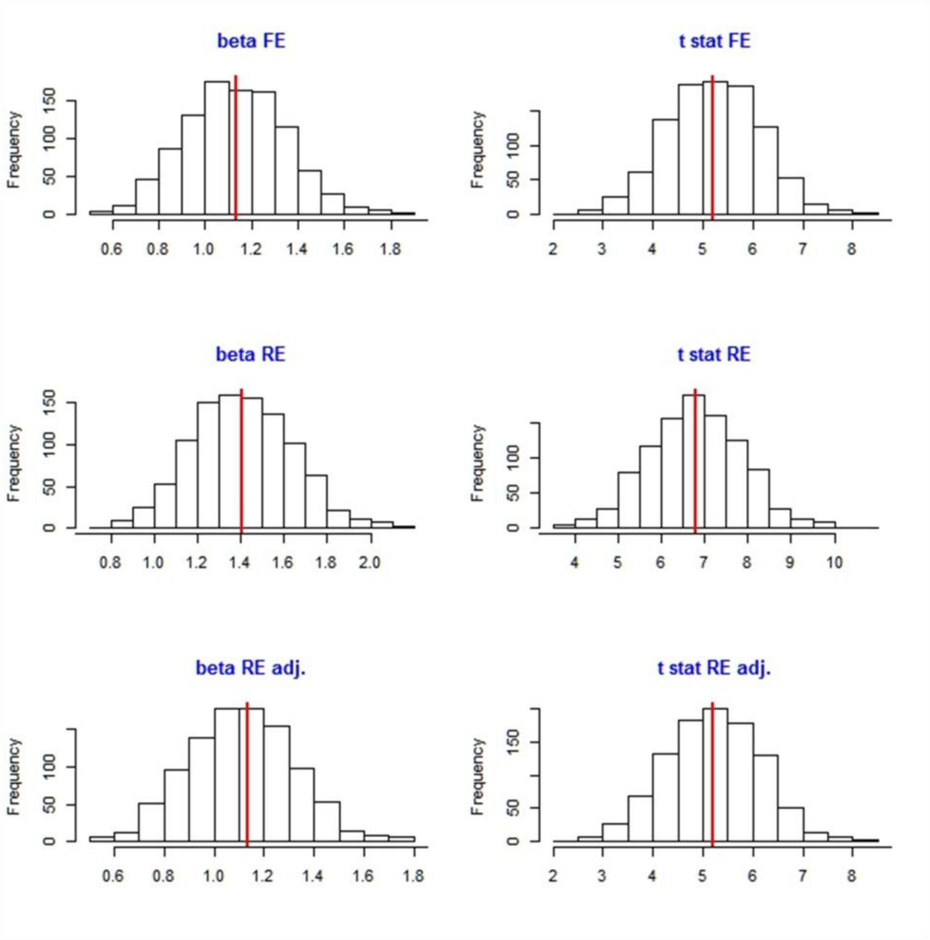
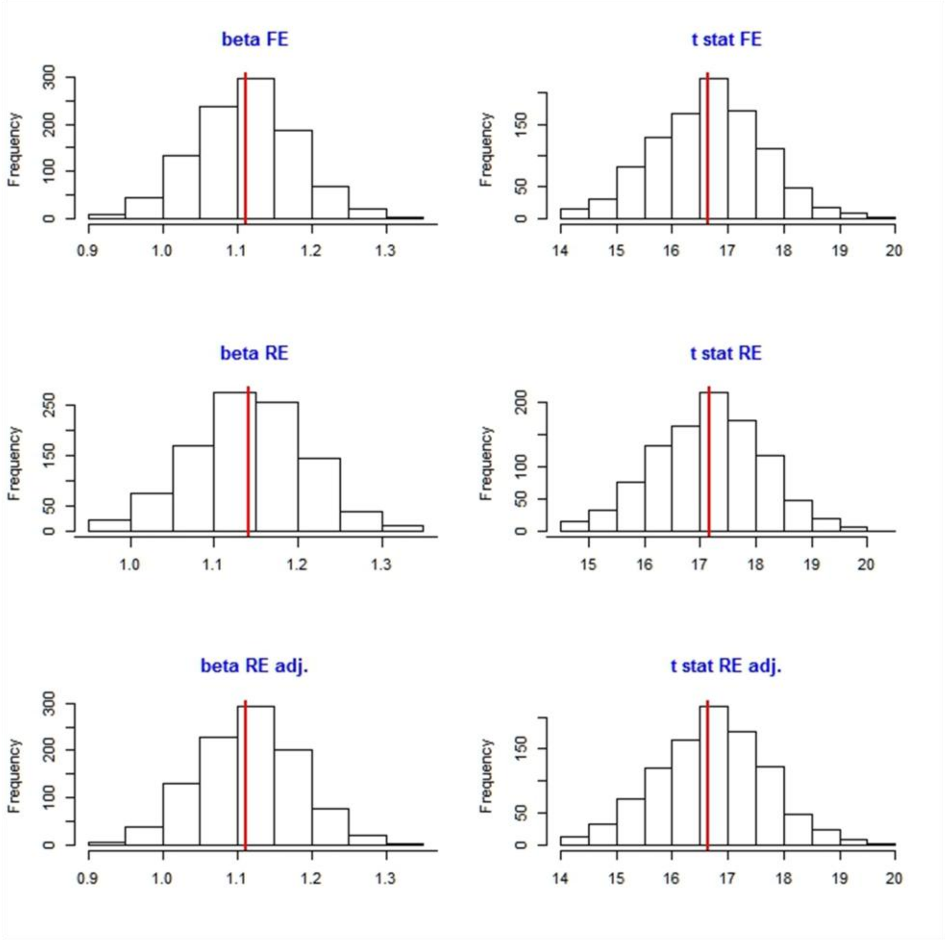


Figure 2. Histograms for β (True Value = 1.11) and t Statistics in Three Estimation Models for 20 Groups with 500 Observations and High Correlation (ρ=0.6) with 1000 Simulations

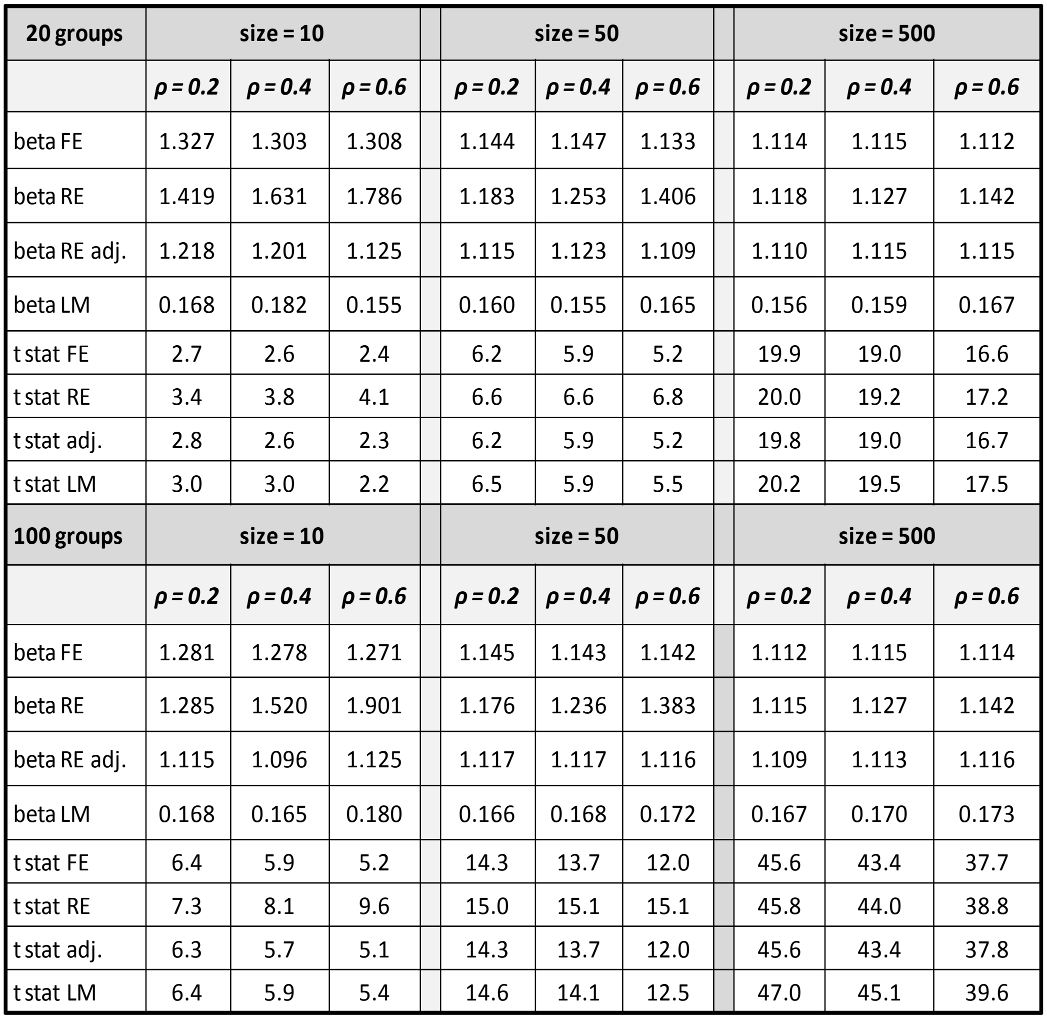


Results Details

- Random effects models without adjustment work only when sample size is large and correlation is moderate, as shown in the second row of beta RE with size of 500 and correlation ρ of 0.2 in Table 2 below.
- When sample size is small, even fixed effect estimates deviate substantially from the true value, possibly because of sampling errors.
- Adjusted random effects models with group averages and de-meaned covariates perform as good as fixed effects models, and better when sample size is small.
- Linear probability models yield the most reliable estimates throughout various combinations of number of groups, size of the group and correlation. It is worth further investigation to find out why there is such a property and how their effects compare with those of logit models for multilevel data.
- In terms of bias, the size of group is far more important than the number of groups as shown by the similarity between top and bottom panels with 20 and 100 groups. Statistical significance shown by t-statistics, however, is mostly determined by the total sample size.

- The strength of correlation between covariates and group effects only affects random effects models. All other three models are insensitive to the increase in correlation.
- The outcomes of the Hausman test depend on sample size. When sample size is small, the Hausman test often fails to reject the random effects model even when its estimate is very different from that of the fixed effects model. When sample size is large, the Hausman test tends to reject the random effects model even when the real difference is very small.

Table 2. Simulation Results in Main Parameter β (True Value = 1.11) and t Statistics from Four Models under Combinations of Groups, Number of Observations Per Group and Correlation ρ between Main Covariate and Group Effects



20 groups	size = 10			size = 50			size = 500		
	ρ = 0.2	*ρ = 0.4*	*ρ = 0.6*	*ρ = 0.2*	*ρ = 0.4*	*ρ = 0.6*	*ρ = 0.2*	*ρ = 0.4*	*ρ = 0.6*
beta FE	1.327	1.303	1.308	1.144	1.147	1.133	1.114	1.115	1.112
beta RE	1.419	1.631	1.786	1.183	1.253	1.406	1.118	1.127	1.142
beta RE adj.	1.218	1.201	1.125	1.115	1.123	1.109	1.110	1.115	1.115
beta LM	0.168	0.182	0.155	0.160	0.155	0.165	0.156	0.159	0.167
t stat FE	2.7	2.6	2.4	6.2	5.9	5.2	19.9	19.0	16.6
t stat RE	3.4	3.8	4.1	6.6	6.6	6.8	20.0	19.2	17.2
t stat adj.	2.8	2.6	2.3	6.2	5.9	5.2	19.8	19.0	16.7
t stat LM	3.0	3.0	2.2	6.5	5.9	5.5	20.2	19.5	17.5
100 groups	**size = 10**			**size = 50**			**size = 500**		
	ρ = 0.2	*ρ = 0.4*	*ρ = 0.6*	*ρ = 0.2*	*ρ = 0.4*	*ρ = 0.6*	*ρ = 0.2*	*ρ = 0.4*	*ρ = 0.6*
beta FE	1.281	1.278	1.271	1.145	1.143	1.142	1.112	1.115	1.114
beta RE	1.285	1.520	1.901	1.176	1.236	1.383	1.115	1.127	1.142
beta RE adj.	1.115	1.096	1.125	1.117	1.117	1.116	1.109	1.113	1.116
beta LM	0.168	0.165	0.180	0.166	0.168	0.172	0.167	0.170	0.173
t stat FE	6.4	5.9	5.2	14.3	13.7	12.0	45.6	43.4	37.7
t stat RE	7.3	8.1	9.6	15.0	15.1	15.1	45.8	44.0	38.8
t stat adj.	6.3	5.7	5.1	14.3	13.7	12.0	45.6	43.4	37.8
t stat LM	6.4	5.9	5.4	14.6	14.1	12.5	47.0	45.1	39.6

Limitations

- In this simulation study, the covariate and group effects are created in two steps. As a result, the correlation between the two is not precisely controlled. It is possible to do just that if we use latent variable approach and create both from a single multivariate normal distribution with a defined covariate structure. That should make the simulation more rigorous.
- The simulation study has a generic control variable z and is in all four models. It would be interesting to leave it out and watch how an omitted variable that is not correlated with the main covariate affects model fitting in multilevel logistic models.
- There are only three correlation settings of 0.2, 0.4 and 0.6 to cover the common area of concern. A wider and finer scale may be helpful to examine the effect in various scenarios.

Conclusions

- Results here clearly show that random effects models with the adjustment of group averages are effective in correcting the bias in the unadjusted random effects models across all scenarios without exception, similar to what has been found for linear models.
- The adjustment is easy to implement. First, one runs a fixed effects model without the covariate to get intercepts as group effects. Then one can correlate the covariate with the group effects to see whether they are correlated. If they are correlated, one just needs to create a de-meaned version and the group averages of the covariate and include them in the random effect model.
- Given the ease of the adjustment and the advantages of random effects models over fixed effects models minus the concern of bias, random effects models should become the default choice in most situations. Major statistical software packages such as R, Stata, SAS and SPSS all have capabilities for random effects models, also called mixed effects or hierarchical models.

References

Bafumi J and Gelman A. 2006. “Fitting Multilevel Models When Predictors and Group Effects Correlate”. Paper presented at 2006 Annual Meeting of the Midwest Political Science Association, Chicago, Il.

Ebbes P, Böckenholt U, and Wedel M. 2004. “Regressor and Random-Effects Dependencies in Multilevel Models”. Statistica Neerlandica 58(2): 161-178.

Hausman J and Taylor W. 1981. “Panel Data and Unobserverable Individual Effects”. *Econometrica* 49: 1377-1398.

Mood C. 2010. “Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It”. *European Sociological Review* 26(1): 67-82.

Plumper T and Troeger V. 2007. “Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects”. *Political Analysis* 15:124-139.