Fixed/Random Effects Analysis of Repeated Measures Data

Flora Shi and Lindsay Xu

Github: https://github.com/cshi56/STA640-FinalProject

Video: https://github.com/cshi56/STA640-FinalProject

Abstract

Employing fixed or random effects models in longitudinal or panel data analysis is commonly considered to mitigate the impact of bias, particularly unmeasured confounding, under suitable conditions. Nevertheless, the underlying mechanisms through which fixed or random effects models address unmeasured confounders remain unclear. In this study, we use a data simulation design to elucidate how, under specific circumstances, fixed and random effects models can limit the influence of unmeasured confounders and improve causal estimation. Our findings suggest that the fixed effect model reduces the impact of unmeasured time-invariant confounders because the fixed effect model estimators rely on the variation within individuals. Therefore, the effects of unmeasured time-invariant confounders are controlled by the fixed effect model. In contrast, the random-effect model estimator uses variations from both within and between individuals. Meanwhile, in most cases, between-individual differences will introduce observed and unobserved confounders that the random-effect model cannot remove, causing bias. However, when the data collection points (T) and number of individuals (N) are large, the random model estimate will be largely determined by within-individual variation. Under this scenario, the difference between estimates from fixed effects and random-effects models is marginal. In addition, time-varying unmeasured confounders can hardly be controlled by simple fixed effects and random-effects models.

1. The Challenge with Panel Data

The main challenge of causal inference using panel data lies in the potential presence of unmeasured confounders that can affect the participants' exposure. In observational studies, participants are not randomly assigned to control and treatment groups. Rather, exposed and unexposed participants can differ significantly due to personal features or various social processes. In the real world, scientists try to include all important observed confounders (such as gender, race, education, etc.) in the model. However, there can always be unmeasured confounders (such as genetics, personalities, emotions, etc.) that affect the participants' exposure. Panel survey data (which involves repeated measures of the same individual/unit over time) give us a chance to limit the effects of unmeasured confounders in the model, while the estimation remains representative of the whole population (Gunasekara et la., 2014). Econometrics introduced the fixed effect model to estimate based on the change within the individual over time. The following discussions focus on how fixed effects and random effects models can control for unmeasured confounders under different assumptions.

2. Simulation Design

In this study, we mainly considered three data generating processes to investigate how fixed effects and random effects models perform under different data generating processes (DGPs) with respect to unmeasured confounding.

Borrowing the idea from Angrist and Pischke (chapter 5, 2009), we are interested in studying whether union membership affects a worker's wage. Specifically, the data generating process creates three panel datasets, each with 100 people observed over five periods. We first assign each individual some observed covariates that affect wages, including gender (male, female), race (black, white, others), and education (high school, college, above college). Then, we create an unobserved confounder (skill). For the first two panel datasets, we assume the unobserved confounder is time-invariant, while for the third panel dataset, we assume the confounder is time-varying. The effects of time-invariant confounder is also time-invariant (constant and additive). The treatment variable (whether the individual is a union member) is generated under two different scenarios: in one case, it has no correlation with the unmeasured confounders, while in the other case, it has correlation with both measured and unmeasured confounders. There are three scenarios:

1. Time-invariant confounding with no correlation between the confounder and treatment variable.

2. Time-invariant confounding with a correlation between the confounder and treatment variable.

3. Time-varying confounding with a correlation between the confounder and treatment variable.

Most importantly, we vary the number of individuals (N) and the number of time points (T). We have tried four scenarios:

1. small N and small T;

2. small N and large T;

3. large N and small T;

4. large N and large T.

3. Main Findings

3.1. Time-invariant unmeasured confounding with no correlation between the confounder and treatment variable

When the time-invariant unmeasured confounders have no correlation with the treatment variables, both the fixed effects model and the random effects model perform well (Figure 1). The fixed effects model effectively accounts for unmeasured confounders because the confounders do not change over time; therefore, estimators within an individual will remove their effects on the outcome (wages in this case). As the assumption that there is no correlation between the individual-specific effects and the treatment variable holds, the random effects model also performs well.

However, the random effects model is more reasonable in this case because, as Table 1 shows, the fixed effects model does not estimate the effects of time-invariant covariates; rather, it "controls" for them. In contrast, the random effects model estimates the effects of time-invariant variables.

Table 1:

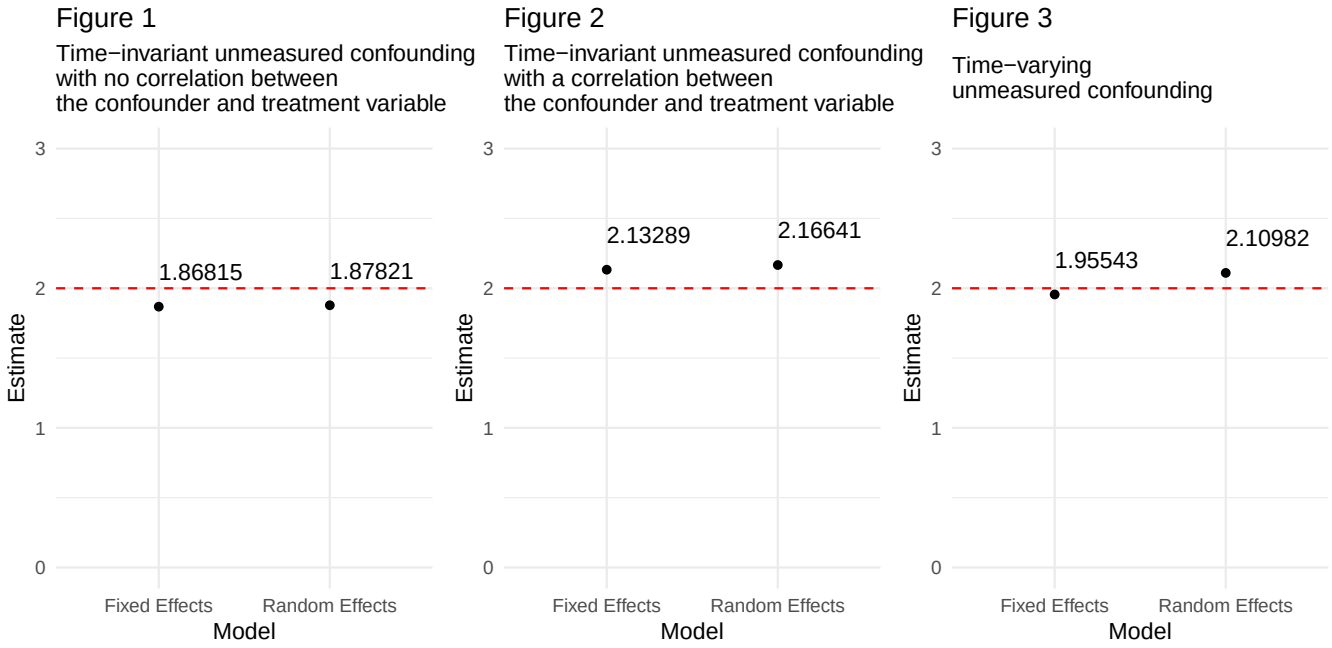
	Dependent variable:	
	wage	
	(1)	(2)
treatment	1.868***	1.878***
	(0.103)	(0.103)
gendermale		0.495
		(0.425)
raceothers		0.094
		(0.489)
racewhite		−0.532
		(0.523)
educationcollege_above		0.705
		(0.604)
educationhigh_school		0.400
		(0.496)
Constant		5.977***
		(0.510)
Observations	500	500
R^2	0.450	0.408
Adjusted R^2	0.312	0.401
F Statistic	326.121*** (df = 1; 399)	339.909***
Note:		*p<0.1; **p<0.05; ***p<0.01

3.2. Time-invariant unmeasured confounding with a correlation between the confounder and treatment variable

When the time-invariant unmeasured confounder is correlated with the treatment variable, the fixed effects model is much more effective because the within-individual transformation removes the effect of the unmeasured confounders (Figure 2). In this approach, each individual serves as their own control, and the effects of the time-invariant confounder are also time-invariant. In contrast, the results from the random effects model may be biased if the unmeasured confounder is correlated with the treatment. This is because the confounder is also correlated with the error term, and the random effects model may not fully control for the confounder.

3.3 Time-varying unmeasured confounding

When there is a time-varying unmeasured confounder, the fixed effects model cannot control for the unmeasured confounding since unmeasured confounders can vary over time within each individual. However, it can limit the bias related to time-varying confounders (Figure 3). Other techniques, such as instrumental variables, should be used to control for time-varying confounders. Meanwhile, the performance of the random effects model may be good if there is no correlation between the unmeasured confounder and explanatory variables in the model. This is because random effects models assume that unmeasured confounders are constant over time for each individual, which are captured by the random effects term. However, in this case, the confounder is correlated with the treatment variable, making the random effects model biased.



3.4 How the number of individual (N) and the number of data collection time points (T) affect the performance of fixed effects and random effects models

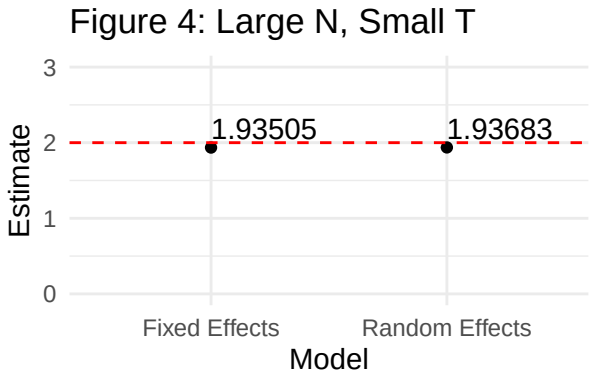
Again, we assume that the time-invariant unmeasured confounder is correlated with the treatment and generate the panel data under four different scenario: (1) small N and small T; (2) small N and large T; (3) large N and small T; (4) large N and large T.

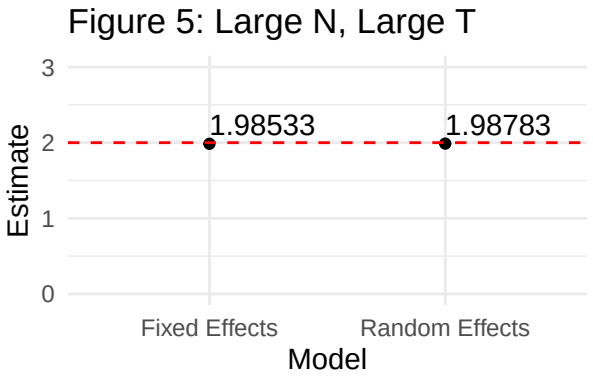
We found that the fixed effects model always performs better than the random effects model in scenario (1),(3), and (4), which is consistent with our findings in Section 3.2. Further, fixed effects model works best when the N is large. However, we also found that when the number of individuals (N) and the

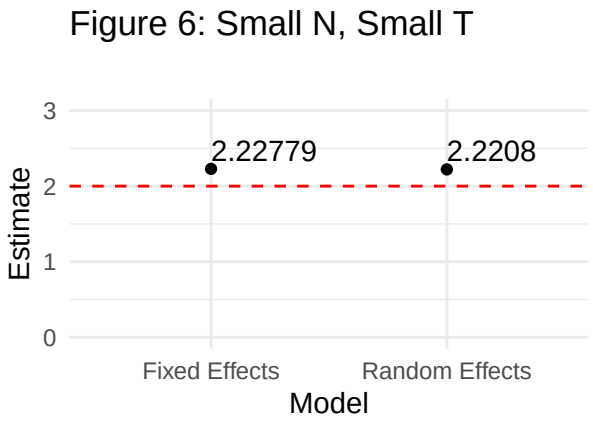
number of data collection time points (T) are both large, the performance of the fixed effects model and the random effects model are quite similar and close to the true treatment effect (see comparison between Figure 4 and 5). This is because when N and T are large, the estimate from the random effects model is dominated by within-individual variation, while the between-individual variation has a smaller impact. As a result, the difference between the estimates from the fixed effects and random models is eliminated. However, when N is small and T is large, random effects model performs better as they are more flexible to deal with persons across multiple data points.

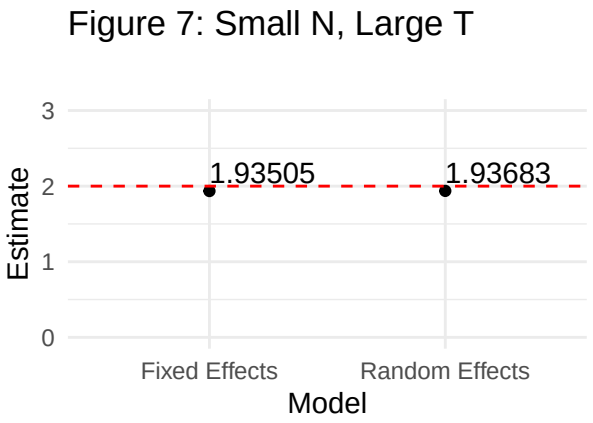
In particular, theoretically, there is more pooling when the group-level standard deviation σ_α is small, and more smoothing for groups with fewer observations. In general, the multilevel-modeling estimate of the group-level parameters α_j can be expressed as a weighted average of the no-pooling for its group $(\bar{y_j} - \beta\bar{x_j})$ and the mean μ_α (this expression is adopted from Gelman and Hill, 2007):

$$\alpha_j \approx \frac{n_j/\sigma_y^2}{n_j/\sigma_y^2 + 1/\sigma_\alpha^2}(\bar{y_j} - \beta\bar{x_j}) + \frac{1/\sigma_\alpha^2}{n_j/\sigma_y^2 + 1/\sigma_\alpha^2}\mu_\alpha$$









4. Conclusion

Overall, when analyzing panel data, fixed effects models are effective in assessing the relationship between the treatment variable and the outcome because they control for time-invariant unmeasured confounders and adjust for measured time-varying confounders (key covariates of interest). Moreover, we found that fixed effects models work best with large sample sizes (N) and low dropout rates over time. In contrast, for analyses involving individuals at different time points, where there may be a small number of individuals but a large number of time points, random effects models are often preferred to fixed effects models, as they can be more efficient and flexible in dealing with individual observations at multiple points. In summary, fixed effects models are useful tools to control for time-invariant confounding, which can bias

causal estimates. However, to address unmeasured time-varying confounding, more complex models such as instrumental variables or controlling for lagged dependent variables may be necessary (Angrist and Pischke, 2009).Ultimately, the model selection will depend on the research question, the available data, and the assumptions made about the underlying data-generating process.

5. References

1. Angrist, Joshua D., and J?rn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton university press, 2009.
2. Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2006.
3. Gunasekara, Fiona Imlach, et al. "Fixed effects analysis of repeated measures data." International journal of epidemiology 43.1 (2014): 264-269.