A simulation study of estimating the treatment effect with hierarchical models

Jenny Liu & Kristen Wingrove
December 13th, 2019

Motivation

In most settings, there is a natural hierarchical structure of the collected data. Hierarchical structures exist across social, behavioral, health, biological, and physical sciences. Some examples of this are patients and doctors in hospitals, and the classic example of students in classrooms in schools. In these examples, individual observations are naturally clustered in meaningful ways and are not independent of one another (Hill, 2013). When hierarchical data are used to identify causal effects, a treatment effect estimate may be biased due to the fact that potential outcomes are correlated with one another and standard errors can be underestimated. Middleton (2008) shows that linear regression estimates are biased depending on the size of the groups. We will need to take into account the group structure when estimating our treatment effect through the use of either fixed effects or random effects.

There are many assumptions we need to make in order to make causal claims of any experiment. In multi-level settings, it is hard to fully believe we meet SUTVA considering the grouped structure of the data. For example, if we were to look at an intervention that took place at school when treatment is randomly assigned at the individual, it is difficult to believe that there is not interference between all individuals. For our final project, we took the opportunity to simulate data to meet the assumptions needed for causal inference and to see what would happen when we violate certain assumptions.

Assumptions

Since we approach our study in 'God mode', we have numerous assumptions that we are inferring in our base case simulation and models. Throughout the study, we will violate these assumptions in different ways to evaluate the effects on our estimand. Our assumptions for this simulation study include randomized experiment assumptions as outlined below:

- Independence between potential outcomes and treatment
- Ignorability
- Stable Unit Treatment Value Assumption (SUTVA)

In addition, we are assuming full compliance with the assigned treatment. In each scenario, we run three different models to assess the impact of violating certain assumptions. The models

include linear regression, fixed effects, and random effects models. As we have simulated our data, it is important to note that we know we are violating some of these assumptions already.

Linear Regression Assumptions:

- Validity: outcome measure reflects phenomenon of interest and model includes all relevant predictors
- Additivity and linearity: outcome is a linear function of the predictors
- Independence of errors
- Equal variance of errors
- Normality of errors

Since we are simulating multi-level data, the independence of errors assumption in the linear regression model is not met. We expect these models to perform worse than the fixed effects or random effects models outlined below which take into account the group structure.

Fixed Effects Effects:

- Standard parametric assumptions
- Fixed effect captures all of the group level differences

Random Effects:

- Standard parametric assumptions
- Random effect is normally distributed where $\alpha_i \sim N(0, \sigma_{\alpha}^2)$
- Random effects are uncorrelated with all of the predictors in the model, including the treatment variable

Describe the design & estimators

Our study design will look at classroom and student level data and simulates the effects of an after school program on test scores. We will simulate data for 10th grade classrooms and include both classroom-level and student-level variables. Treatment assignment corresponds with participation in an after-school program which provides participants additional curriculum support. Variables that we will simulate include the following:

- *Classroom-Level:* Teacher's years of experience teaching, education level, and average test scores from prior years

- *Student-Level:* Minority indicator, parent's education level, family income, qualification for free-lunch, distance to school in hours, gender, and 9th grade test scores

In this simulation study, we will have four scenarios where we evaluate three models and the estimated average treatment effect (ATE): base case scenario, violation of random effects assumptions, violation of ignorability, and assignment of treatment at the group level. In each scenario, we compare the estimated ATE ('researcher mode') to the sample average treatment effect (SATE) which is calculated in 'god mode.' We use simulations to generate randomization distributions (randomizing treatment only within a sample) to evaluate the performance of the ATE in regards to bias, root mean squared error, and confidence interval coverage.

In each scenario, we either alter the data generating process or the models (linear regression, fixed effects, or random effects). Our estimate for the ATE is the coefficient of the treatment variable from the models. A summary of the process and approach is provided in Table 1.

Table 1. Summary of the different settings and models we are exploring for our simulations

Scenario	**Simulation Alteration**	
	DGP	**Models**
Base		
Violate Ignorability	Treatment assignment based on 9th grade pretest scores	Confounder will not be included in the models
Violate Random Effects Assumption	Correlate classroom level variance with the classroom's family income average	
Group-level Treatment Assignment	Treatment assignment occurs at the classroom level. All students within a classroom receive treatment.	Treatment variable (Z) in the models is group-level (i.e. Z_j instead of Z_{ij}, where i is student and j is classroom).

The base case scenario will use the simulated hierarchical data to estimate ATE and compare to the true SATE. In this scenario, we do not have any violations and use student-level treatment assignment. However, as discussed in the assumptions section, we know that the linear regression model independence of errors assumption is not met when we have simulated group-level data. Therefore, we expect the linear regression model to estimate the ATE with a larger bias and higher variance around the SATE.

In the ignorability violation scenario, we are taking a two-pronged approach to evaluate the effect on the estimated ATE. First, we will alter the data generating process so that the treatment assignment (10th grade after-school program) is dependent on the student's 9th grade test scores. The student will have a higher probability of being assigned treatment if his or her 9th grade test scores were low. Additionally, the models used to estimate the ATE will not include a variable which is a known confounder, as defined in our data generating process.

The random effects violation scenario aims at breaking the assumption that the random effects are uncorrelated with all of the predictors in the model. To do this, we will correlate the simulated classroom level variance with one of the student-level confounders (family income). We will implement this as an alteration to the data generating process.

Lastly, the group-level scenario uses the same data generating process as the base case scenario up until treatment assignment and student outcome. This is because we are assuming SUTVA in all of our settings, which may not be true in classrooms. However, the problem with interference goes away when treatment assignment is assigned at group level. In this simulation setting, student outcomes are altered to account for a group-level treatment assignment. In our models, we will use a classroom level treatment assignment variable which affects the ATE interpretation. The ATE estimand will be interpreted as the average treatment effect at the classroom level, instead of at the student level.

Simulation set-up

For our simulation study, we simulated a randomized experiment taking place in suburban classrooms in New York state high schools. Where possible, characteristics and demographics from the New York population were used to inform the simulated data. The

estimand of interest is the average treatment effect (ATE). We are interested in evaluating the average effect of an after school program on 10th grade test scores through a randomized experiment. The ATE will measure the difference between those who attended our after school program on 10th grade test scores (treatment group) and those who did not (control group). We are able to do this with the use of a randomized experiment where each student or classroom is assigned to treatment with a known probabilistic rule with a non-zero probability of being assigned to treatment. We will estimate our estimand, the ATE, through three different modeling strategies: linear regression, fixed effect models, and random effect models. Linear regression ignores the group structure of our data and should give us a biased estimate of our ATE and an underestimate of our standard errors. Both fixed effect and random effect models take into account the group structure and should give us an unbiased estimate of our ATE.

Our data generating process contains three classroom-level covariates and seven student-level covariates. We simulated our data for 40 classrooms and 25 students in each classroom, in three different scenarios, for 1000 simulations:

I. Base case scenario

First we created a baseline model of all covariates with treatment randomized at the student-level. The classroom level predictors we were interested in were the teacher's work experience, the teacher's education, and an average test score for a given classroom (Table 2). We are using the number of years the teacher has been teaching as a proxy for their work experience. To do this, we simulated data with a normal distribution with mean 5 and standard deviation 5. We did not want any negative values for teachers experience, so we took the absolute values of our simulated yearstea covariate. For teachers education we wanted a majority of teachers to have a bachelor's degree, a quarter of teachers to have a master's degree, and 2.5% having a doctorate. This was calculated by using our settings of 40 classrooms to have only one classroom with a teacher with a Ph.D. Next, we simulated average test scores to be normally distributed with mean test score centered at 72 and a standard deviation of 8 and to be correlated with years teaching. Lastly we needed to create the group structure by adding variance to classrooms. This was centered at 0 with a standard deviation of 3.

Table 2. Classroom-level covariates

Variable Name	Description	Sampling
yearstea	-Number of years teaching	$N(5,\ 5^2)$
teach.edu	-Teacher's education level -Factor variable where 1 indicates Bachelor's degree, 2 indicates Master's degree, and 3 indicates PhD.	*Random sample:* *-1 at 72.5%* *-2 at 25%* *-3 at 2.5%*
avgtest	-Average test score of students in prior years -Correlated with teacher's work experience	$N(72,\ 8^2) +$ $.75 * yearstea$
classroom_score	-Variance to classrooms for group structure	$N(0,\ 3^2)$

The next step was to create our student-level covariates: minority, parent's education, family income, if the student is eligible for free lunch, their distance to school in hours, gender, and their 9th grade test scores. First, we simulated if the student was a minority with 35% of the students as a minority and 65% not a minority. Next, we wanted to create covariates the represent socioeconomic class through parent's education and family income. We did this in two steps, using the minority covariate. The parent's education covariate consists of highest level of education in five different categories: less than high school, high school degree, college degree, master's/professional degree, doctorate degree. If the student was a minority, there was a higher probability that parents had less than a high school degree and at least a high school degree, while non-minority students had a higher probability of having had at least a college degree or higher. Family income was then simulated correlated with both if the student was a minority and parent's education. First, family income was centered at income at around 50 (in thousands of dollars) with a standard deviation of 20 with all negative incomes made positive by

taking the absolute value. If parents had at least a high school degree, there was a boost in family income and a greater boost if they had at least a college degree. Additionally, if the student was a minority, there was a decrease in family income. The free lunch covariate was created to be an indicator variable and dependent on if the family income was less than $46,645. Next, distance to school was created with a normal distribution with mean 20 minutes and standard deviation of 4, and we took the absolute value of all negative values. There was an equal representation of gender in our data. Lastly, we generated our 9th grade test scores using both classroom and student-level covariates.

Our outcome variable was generated using number of years teaching, teacher's education, parent's education, family income, if the student was minority, distance to school, a classroom effect, and random error. Treatment was randomly assigned at the student-level and at the group level.

II. Violating ignorability

One of the key assumptions we need to make in order to estimate the ATE is ignorability. We will violate ignorability in our simulations by not measuring all of our covariates by removing removing distance to school from our models. We will also make 9th grade test scores a confounder in our models and correlate treatment and explore what happens to the estimate of ATE. We expect we will get a biased estimate of the ATE for all three of our models.

III. Violating random effects assumption

A key assumption of using random effects to estimate the treatment effect is that the random effect is not correlated with any confounders, We will also violate the random effects assumption by correlating the classroom effect with the family income. We expect to get a biased estimate of our ATE using a random effects model and basic linear regression as opposed to using a fixed effect model.

Simulation results

First, we looked at a table for each scenario where we calculated bias by subtracting the SATE from the estimated ATE or treatment variable coefficient. In Table 3, we see that the bias is low in all of the scenarios and models. However, there are some instances where certain models or scenarios have larger bias comparatively. In all scenarios, the fixed effects and random effects models perform better than the linear regression. This affirms our data generating process where we implemented a group-level structure. Across scenarios, we see that the fixed effects and random effects model bias is similar in the base case and random effects violation but is impacted more in the ignorability violation. Of the four scenarios, group level treatment has the highest bias.

Table 3. Bias comparisons of the sample average treatment effect for each setting and model from randomization distributions.

Bias.Comparison	Base.Case	Ig.Violation	RE.Violation	Group.Treatment
Linear Regression	-0.0178	0.0197	-0.0107	-0.0732
Fixed Effects	0.0016	0.0035	0.0015	N/A
Random Effects	0.0015	0.0036	0.0014	-0.0728

The root mean squared error (RMSE) table follows the same pattern as the bias comparison table above (Table 4). The group level treatment scenario has the highest RMSE, and the RMSE for fixed or random effects is similar across all scenarios. In all scenarios, the linear regression models have the highest RMSE of the three different models run.

Table 4. Root mean squared error comparisons of bias for the sample average treatment effect for each setting and model from randomization distributions.

RMSE.Comparison	Base.Case	Ig.Violation	RE.Violation	Group.Treatment
Linear Regression	0.23	0.19	0.21	1.24
Fixed Effects	0.04	0.05	0.06	N/A
Random Effects	0.04	0.05	0.06	1.23

Additionally, we plotted the randomization distributions from the 1,000 simulations run between the scenarios and models (Figure 1 and Figure 2). Linear regression histograms are on the left with fixed effects models in the middle and random effects on the right-hand side. Blue lines indicate the mean of the randomization distribution, and a red dashed line indicates the SATE. Due to the low bias, as seen in the bias comparison table, there are some instances where the differences in the red and blue lines is hard to distinguish. These histograms visually enforce the conclusions of the bias table, where we can see that the group level histograms have a slightly larger difference between the randomization distribution mean and the SATE. Additionally, we see that the efficiency of the ATE increases with the fixed effects and random effects models as the variance around the mean is smaller than the linear regression models. This reflects our lower bias and RMSE in these models.

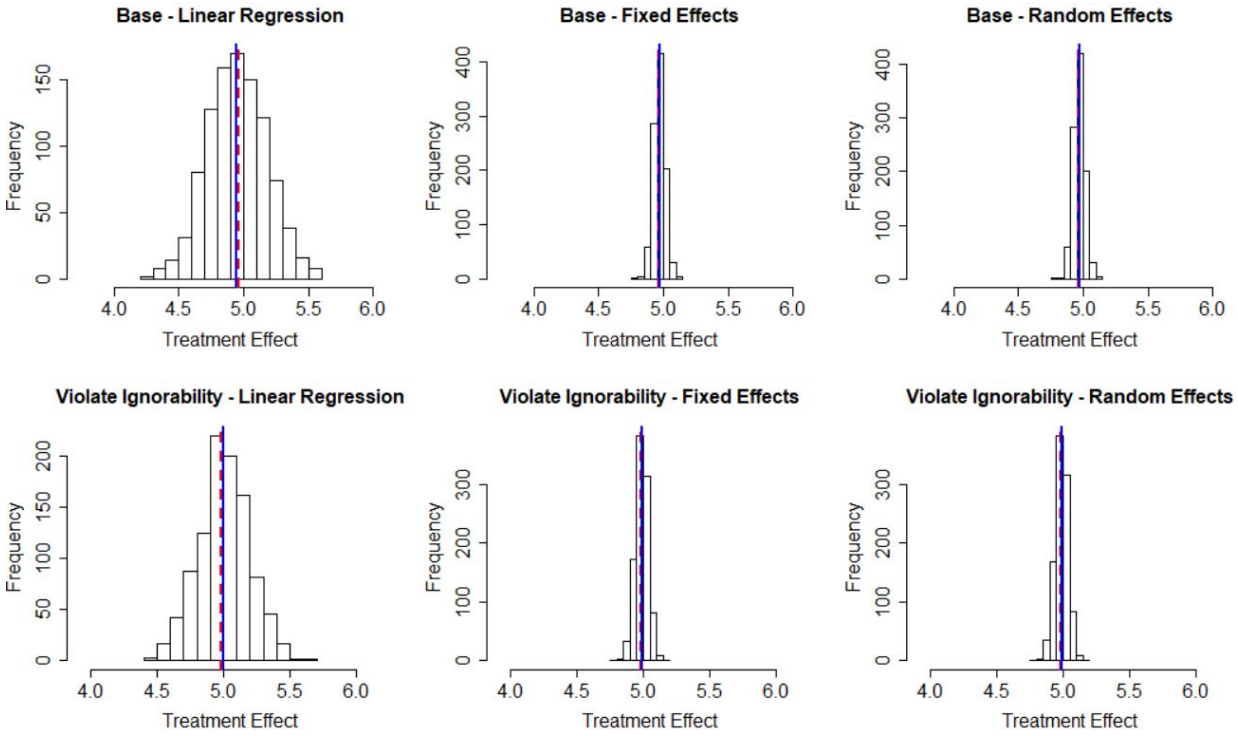

Figure 1. Randomization distribution of the average treatment effect for base case setting and violating ignorability setting.

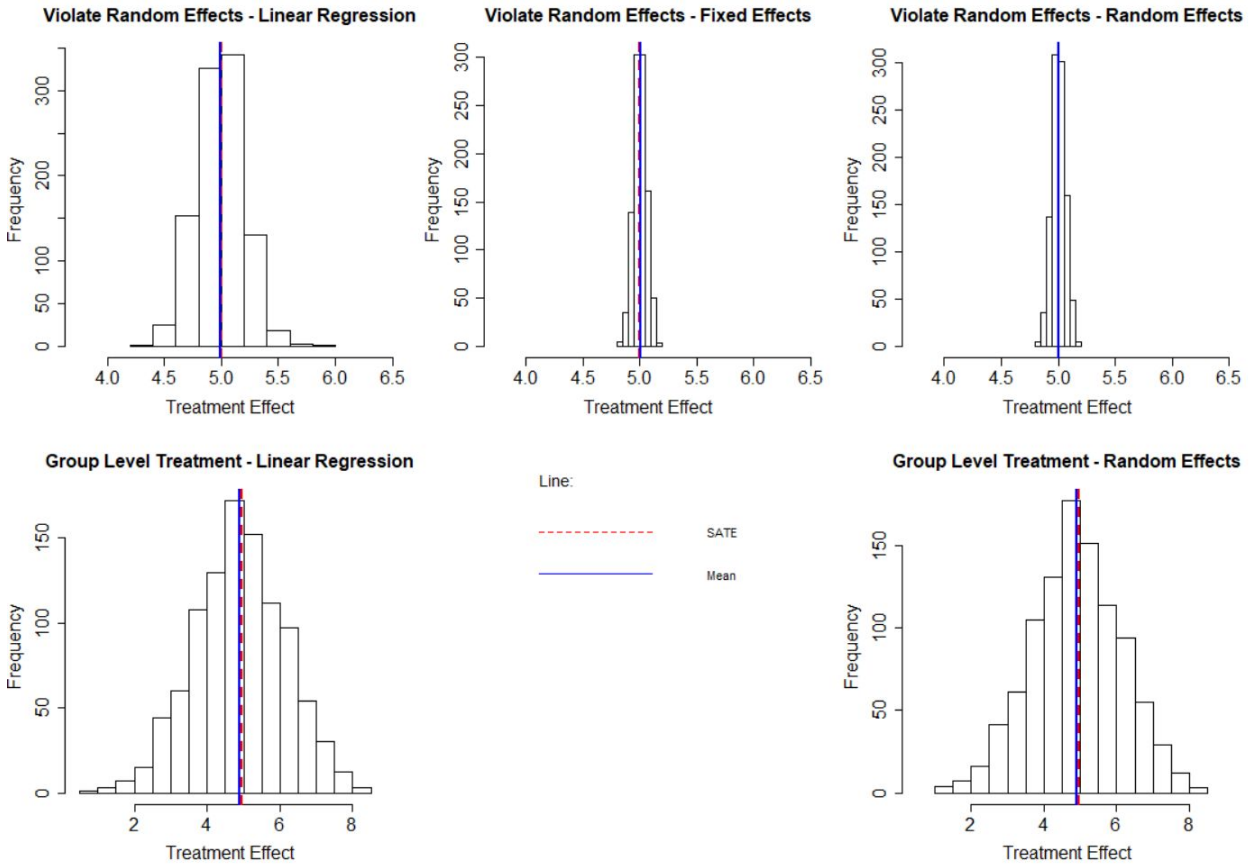

Figure 2. Randomization distribution of the average treatment effect for violating random effects assumptions and group-level randomization treatment.

Additionally, when we look at the confidence intervals for each of the three settings with randomization at the student-level, the SATE is captured within the confidence intervals a larger percent of the time as compared to the linear regression model with all models capturing the SATE at least 95% of the time. However, with treatment at the group-level, the SATE is only captured 29% of the time in the confidence intervals produced by the linear regression model and 93% of the time using random effects .

Table 5. Comparison table of percentage of times confidence intervals capture SATE for each setting and model from randomization distributions.

CI.Comparison	Base.Case	RE.Violation	Ig.Violation	Group.Treatment
Linear Regression	0.95	0.96	0.95	0.29
Fixed Effects	0.99	0.98	0.99	N/A
Random Effects	0.99	0.98	0.99	0.93

Discussion

From our simulation settings where randomization occurred at the student-level, our models including the group structure produced randomization distributions with less variance, less bias, and smaller root mean square errors. Looking only at the base case scenario when we ignore the group structure of our data, our treatment effect estimates are more biased compared to models that account for the group structure (either fixed effects and random effects). When we violate ignorability, by removing a confounder from our model, the linear regression model that ignores the group structure also gives us a more biased estimate of our treatment effect compared to using either fixed effect or random effect models. In both of these settings, base and violating ignorability, both fixed and random effects perform similarly. We hoped to see a difference in performance in our next setting where we violate the random effects assumption by correlating our random effect with family income. In this setting, we see again the linear regression model is more biased than both the fixed effect and random effect models, but both fixed and random effect models perform similarly even when we do not meet the random effect assumption. In all three settings, we should take into account the group structure of the data.

In our previous settings, we are assuming SUTVA, but in hierarchical structures that might not necessarily be the case. Our last simulation setting, randomization at classroom-level, looks at a scenario when the problem with interference goes away. In this setting, linear regression and random effects models perform similarly when estimating the treatment effect, however when we look at the confidence intervals produced these estimate the results vary. Both these models produce biased estimates of the treatment effect. If we run our simulations with a higher number of classrooms, this bias decreases and is line with the other methods. However, when we look at the confidence intervals produced by both models, the linear regression estimates only cover the SATE 29% of the time as compared to the random effects model covering the SATE 93% of the time. In this exploratory exercise, the two models (linear regression and random effects) still perform similarly to each other in regards to bias, but when we take into account the standard errors linear regression performs considerably worse.We are unable to and should not used fixed effects to estimate the treatment effect. This is because the fixed effects absorb any differences between the groups, including the treatment. When we

randomized at the group-level, we should still consider the hierarchical structure of the data by using random effects models, even if the linear regression model estimates are similar as it does not capture the standard errors correctly.

Our biggest takeaway is that causal inference is hard.

References

1. Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1) 1-48. doi: 10.18637/jss.v067.i01

2. Gill, J., Womack, A. (2013) The Multilevel Model Framework. *The SAGE Handbook of multilevel modeling* (pp. 201-220). London: SAGE Publications Ltd doi: 10.4135/9781446247600

3. Hill, J. (2013). Multilevel models and causal inference. *The SAGE Handbook of multilevel modeling* (pp. 201-220). London: SAGE Publications Ltd doi: 10.4135/9781446247600

4. Middleton, J. (2008). 'Bias of the regression estimator for experiments using clustered random assignment' *Statistics and Probability Letters*. 78(pp. 2654-2659).