Cluster Randomized (Multilevel) Designs

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Cluster Randomized (Multilevel) Designs

Introduction

- Consider a classic two-group completely randomized design in which one group gets the treatment and the other group is a control.
- We saw earlier in the course how such a design can be handled by a 2-sample, independent sample t-test.
- We also saw that, if we code treatment condition as a binary 0-1 variable, the design is also handled easily as a linear regression model.
- One advantage of the linear regression approach is that it is easy to expand the design to include continuous *covariates*.
- Another, more esoteric advantage, is that it leads us to think of the design, and its statistical test, in terms of an underlying *regression model*.

Revisiting the Voucher Experiment

- Murnane and Willett(Chapter 4, Table 4.1) analyzed an outcome of the NYSP evaluation study.
- In this study, students were divided into two groups. The experimental group was randomly selected (from a larger group of volunteer applicants) to receive a tuition voucher, while the control group did not receive a voucher.
- The key experimental question was whether receipt of a voucher affected achievement as measured by a `POST_ACH` variable.

Revisiting the Voucher Experiment

- If we "dummy-code" a VOUCHER variable as a binary 0-1 variable, we can process the data as a linear regression model

$$\mathtt{POST\_ACH}_i = \beta_0 + \beta_1 \mathtt{VOUCHER}_i + e_i \tag{1}$$

- We recall that the standard assumption is that the e_i are independent and identically distributed as normal variables with constant variance σ^2.
- We also recall that this model implies a conditional mean model

$$E(\mathtt{POST\_ACH}|\mathtt{VOUCHER}) = \beta_0 + \beta_1 \mathtt{VOUCHER} \tag{2}$$

  and so β_0 represents the mean of the control group ($\mathtt{VOUCHER} = 0$), and $\beta_0 + \beta_1$ is the mean of the experimental group ($\mathtt{VOUCHER} = 1$).
- The conditional variance model is

$$\mathrm{Var}(\mathtt{POST\_ACH}|\mathtt{VOUCHER}) = \sigma^2 \tag{3}$$

Revisiting the Voucher Experiment

- When the assumptions of the linear regression model hold, the estimate of the treatment effect ($\hat{\beta}_1$) has a number of good properties:
  1. It is unbiased, $E(\hat{\beta}_1) = \beta_1$.
  2. It is consistent. As n increases without bound, the probability that the estimate will be within any non-zero error bound of the parameter approaches 1.
  3. It is efficient, there is no other linear combination of the data that has a smaller variance.

Cluster-Randomized Designs

The Need for Sampling in Clusters

- In many if not most cases in education, a fully randomized design is difficult to implement.
- When studying educational interventions, it is often necessary, for a variety of practical reasons, to process students in intact classes. (See Murnane and Willett, Chapter 7, p. 107–110, for a brief discussion of this.)
- In this case all students within a given class are exposed to similar conditions.
- These within-class or within-cluster similarities will create violations of the assumptions of the classic regression model if the clustering is ignored.

Cluster-Randomized Designs

Dangers of Ignoring Clustering

- When observations are not, in effect, independent, we can be misled by a classical analysis.
- And when observations occur in clusters, they are in general not independent.
- How are we misled? One consequence is that we overrate the amount of information contained in the data.
- A classic example of this was in the audiology literature, in which some experimenters used the ear as the unit of analysis.
- Ears tend to be clustered within person. Most people have two ears. And because these two ears are attached to the same body, their properties tend to be correlated.
- Suppose a researcher gathered 20 individuals and measured their 40 ears on some audiological characteristic, then tried to compute a confidence interval for the mean value of that characteristic.

Cluster-Randomized Designs

Dangers of Ignoring Clustering

- We know that the confidence interval is roughly the sample mean plus or minus two (estimated) standard errors, or approximately

$$\bar{X} \pm 2\hat{\sigma}/\sqrt{n} \tag{4}$$

- The square root of n in the denominator of the above formula requires the assumption of independence of errors.
- Suppose ears were perfectly correlated within individual. Then the 40 ears would in effect represent only 20 observations, and the classic formula would have to be corrected as

$$\bar{X} \pm 2\frac{\hat{\sigma}}{\sqrt{n/2}} \tag{5}$$

- In Psychology 310, we derive the formula required to compensate for the within-person correlation between ears.

Multi-Level Modeling

- Suppose we sample subjects in intact classes, and randomly assign the intact classes to either an experimental or control condition.
- Let's introduce a model designed specifically to handle this situation.
- Consider observation Y_{ij} representing the outcome score for the ith person within the jth school.
- Within-school, we can model each student's performance as simply randomly varying around the school's mean.

$$Y_{ij} = \beta_{0j} + \epsilon_{ij} \tag{6}$$

  The errors ϵ_{ij} are assumed to be independently and identically distributed normally with variance σ_e^2.
- The within-school model is often referred to as a Level-1 model.

Multi-Level Modeling

- Different schools have different means. These means vary around a grand average, γ_{00}.
- We can view variation in these means as arising from two sources:
  1. A non-random change due to the treatment effect
  2. A random component (representing other random sources of variation between schools).
- We can write this Level-2 model as

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + u_{0j} \tag{7}$$

where the u_{0j} are normally distributed random variables with mean 0 and variance σ_u^2.

Multi-Level Modeling

- We can combine the Level-1 and Level-2 models to produce a composite model,

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \gamma_{01} T_j + (e_{ij} + u_{0j}) && (8) \\ &= \gamma_{00} + \gamma_{01} T_j + e^*_{ij} && (9) \end{aligned}$$

- The second version of the equation looks just like the classic simple linear regression model!
- However, there is an important difference.
- Under the stated assumptions, what is the covariance of the errors for any two individuals in the same group?

Multi-Level Modeling

- Since the errors have zero means, we may write, for the first two individuals in the jth group,

$$\begin{aligned}
\mathrm{Cov}(e^*_{1j}, e^*_{2j}) &= E(e^*_{1j} e^*_{2j}) \\
&= E((e_{1j} + u_{0j})(e_{2j} + u_{0j})) \\
&= E(e_{1j} e_{2j} + e_{1j} u_{0j} + u_{0j} e_{2j} + u^2_{0j}) \\
&= E(e_{1j} e_{2j}) + E(e_{1j} u_{0j}) + E(u_{0j} e_{2j}) + E(u^2_{0j}) \\
&= \mathrm{Cov}(e_{1j} e_{2j}) + \mathrm{Cov}(e_{1j} u_{0j}) + \mathrm{Cov}(u_{0j} e_{2j}) + \mathrm{Var}(u_{0j}) \\
&= 0 + 0 + 0 + \sigma^2_u
\end{aligned} \tag{10}$$

- Any two individuals in the same group would have correlated errors.

Multi-Level Modeling

- Note that we could also rewrite the combined model as

$$
\begin{aligned}
Y_{ij} &= (\gamma_0 0 + u_{0j}) + \gamma_{01} T_j + \epsilon_{ij} && (11) \\
&= \gamma^*_{00} + \gamma_{01} T_j + \epsilon_{ij} && (12)
\end{aligned}
$$

- In this form, the intercepts term is a random variable. Such a model is sometimes referred to as a random intercepts model as a result.
- Each group has its own intercept that is conceptualized as a random variable.
- Because some effects are constants and others are random variables, this model is also called a mixed model, or a Linear Mixed Effects(LME) model.

Fitting the Multi-Level Model in R

- To fit a LME model in R, we need to first write the model in composite form.
- Then, we use the `lmer` function from the `lme4` library, and we employ a modified syntax, in which the random effects are presented with parentheses and conditionalized on their grouping variables.
- For example, consider the model $Y_{ij} = \gamma_{00} + \gamma_{01} T_j + (e_{ij} + u_{0j})$. It is coded as

```
fit <- lmer(Y ~ 1 + T +(1|School))
```

- There is a random intercept component that varies conditionally with school, and note how this is indicated within parentheses.

Multilevel Modeling of an SFA Intervention

- The School For All (SFA) program is discussed on page 108 of Murnane and Willett.
- The SFA program was first introduced into public schools in Baltimore, Maryland, in 1987.
- During the next two decades, its use spread rapidly.
- Today more than 1,200 schools, most with economically disadvantaged student bodies, use this school-wide approach to developing students' reading skills.
- Sparking the rapid early expansion of SFA were the findings of several dozen non-experimental evaluations of the intervention conducted during the 1990s, which showed that the reading skills of children in schools that adopted SFA were better than those of children in "comparison" schools that implemented other reading curricula.
- However, these were not randomized experiments. Instead, the researchers who conducted the evaluations sought out and selected non-randomly several "comparison" schools that they believed served student populations that were demographically similar to those of the SFA schools and had a history of similarly low reading achievement.

Multilevel Modeling of an SFA Intervention

- A necessary condition for such evaluations to provide unbiased estimates of the causal impact of SFA is that treatment and comparison groups must be equal in expectation on all unobserved dimensions that are correlated with student reading outcomes, prior to treatment.

Multilevel Modeling of an SFA Intervention

- There are at least two reasons to question whether this condition was satisfied in the early non-experimental evaluations of SFA.
  1. Schools that adopted SFA were required to spend about $75,000 in the first year of program implementation, $35,000 in the second year, and $25,000 in the third year, to pay for the materials and training that the SFA Foundation provided. Schools that were able to obtain agreement from stakeholders to devote such substantial resources to a single program may have differed from other schools along other important dimensions, such as the quality of their leadership
  2. Before a school introduces SFA, the Success for All Foundation requires that four-fifths of the faculty members in the school vote to adopt the school-wide intervention. A result of this requirement may have been that schools that voted to adopt SFA possessed a greater sense of common purpose, on average, than those that adopted more conventional curricular approaches to teaching reading. This difference in commitment to improving children's reading skills could itself have influenced student outcomes positively even if the SFA approach itself was no better than the alternatives.

Multilevel Modeling of an SFA Intervention

- To help eliminate some of these issues, a cluster-randomized design was developed.
- 41 schools were randomly assigned, 21 to the experimental group and 20 to the control group.
- Murnane and Willett present a sample analysis of performance on a "Word- Attack test (WATTACK).
- The treatment variable is SFA.
- There was also a covariate tested in the third, most complex model. This covariate is SCHL_PPVT, i.e., the school-level average on the Peabody Picture Vocabulary Test.

Multilevel Modeling of an SFA Intervention

- Here is the code for the 3 models.

```
> data <- read.csv("ch07.csv")
> attach(data)
> library(lme4)
> ## Base Model -- No Treatment Effect, Random Intercepts
> fit.0 <- lmer(wattack ~ 1 + (1|schid))
> ## Adds Treatment Effect for sfa
> fit.1 <- lmer(wattack ~ 1 + sfa +(1|schid))
> ## Adds sch_ppvt as a covariate
> fit.2 <- lmer(wattack ~ 1 + sfa + sch_ppvt +(1|schid))
```

Multilevel Modeling of an SFA Intervention

- Let's examine some output. In the base model, we find an intercept of 477.54, agreeing precisely with the output in Murnane and Willett, page 114.
- There is a minor discrepancy in the estimate of σ_u^2, the random effects intercept. The textbook has it at 78.69, while R gives 79.13.

Multilevel Modeling of an SFA Intervention

```
> summary(fit.0)

Linear mixed model fit by REML ['lmerMod']
Formula: wattack ~ 1 + (1 | schid)

REML criterion at convergence: 20147.73

Random effects:
 Groups   Name        Variance Std.Dev.
 schid    (Intercept)  79.13    8.896
 Residual             314.19   17.725
Number of obs: 2334, groups: schid, 41

Fixed effects:
            Estimate Std. Error t value
(Intercept)  477.535      1.451   329.1
```

Multilevel Modeling of an SFA Intervention

- We add sfa, and the coefficient is not significant in the modified model.

```
> summary(fit.1)

Linear mixed model fit by REML ['lmerMod']
Formula: wattack ~ 1 + sfa + (1 | schid)

REML criterion at convergence: 20141.48

Random effects:
 Groups   Name        Variance Std.Dev.
 schid    (Intercept)  75.68    8.70
 Residual             314.23   17.73
Number of obs: 2334, groups: schid, 41

Fixed effects:
            Estimate Std. Error t value
(Intercept)  475.302      2.035  233.62
sfa            4.366      2.844    1.54

Correlation of Fixed Effects:
    (Intr)
sfa -0.715
```

Multilevel Modeling of an SFA Intervention

- Moreover, use of the anova command for model comparison shows no significant difference between the two models.

```
> anova(fit.0,fit.1)

Data:
Models:
fit.0: wattack ~ 1 + (1 | schid)
fit.1: wattack ~ 1 + sfa + (1 | schid)
      Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
fit.0  3 20156 20174 -10075    20150
fit.1  4 20156 20179 -10074    20148 2.4029      1     0.1211
```

Multilevel Modeling of an SFA Intervention

- The covariate sch_ppvt produces a significant improvement in model fit. The model summary shows a significant coefficient:

```
> summary(fit.2)
Linear mixed model fit by REML ['lmerMod']
Formula: wattack ~ 1 + sfa + sch_ppvt + (1 | schid)

REML criterion at convergence: 20127.22

Random effects:
 Groups   Name        Variance Std.Dev.
 schid    (Intercept)  49.33    7.024
 Residual             314.18   17.725
Number of obs: 2334, groups: schid, 41

Fixed effects:
            Estimate Std. Error t value
(Intercept) 419.8075    12.6427   33.21
sfa           3.5691     2.3555    1.52
sch_ppvt      0.6229     0.1408    4.43

Correlation of Fixed Effects:
         (Intr) sfa
sfa      -0.006
sch_ppvt -0.991 -0.089
```

Multilevel Modeling of an SFA Intervention

- The model comparison shows a significant change:

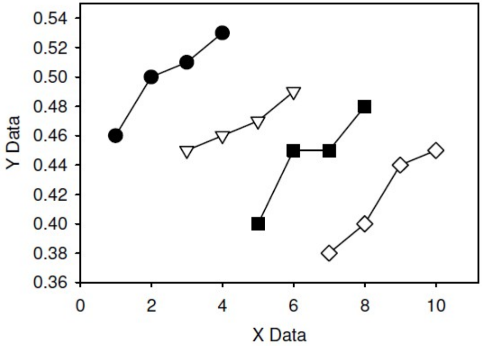
```
> anova(fit.1,fit.2)
Data:
Models:
fit.1: wattack ~ 1 + sfa + (1 | schid)
fit.2: wattack ~ 1 + sfa + sch_ppvt + (1 | schid)
      Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
fit.1  4 20156 20179 -10074    20148
fit.2  5 20141 20170 -10065    20131 17.173      1  3.412e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Intraclass Correlation

- Was multilevel modeling worth the trouble?
- A course in multilevel modeling will go into significant detail about how multilevel modeling improves on other techniques sometimes used in its stead.
- For example, one approach often used in the past is to compute within-group means on a predictor X and a criterion Y for each of the k classes, then perform linear regression on those means.
- This can lead to the *ecological fallacy*, in which the results on the aggregated data are assumed to apply to the data within each group.
- As the graph on the next slide shows, this can be false.

The Intraclass Correlation

- As the plot (generated by Kris Preacher, whose excellent course on Multilevel Modeling is offered here at Vanderbilt on a consistent basis and is highly recommended) shows, within the 4 groups, there is a positive regression slope.
- However, if one were to plot the group means on the two variables, one would have 4 points with a negative slope.



The Intraclass Correlation

- One way of characterizing the need for a multilevel approach to modeling is to compute an *intraclass correlation coefficient*.
- Consider again the model of Equation 6. Since both terms on the right are random variables, and they are uncorrelated, it shows that the total variance of Y is the sum of two components.
- That is,

$$\begin{aligned} \sigma_y^2 &= \sigma_\beta^2 + \sigma_e^2 \\ &= \sigma_u^2 + \sigma_e^2 \end{aligned} \tag{13}$$

The Intraclass Correlation

- The proportion of variance accounted for by groups is

$$\rho_{IC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{14}$$

- This coefficient is zero if there is no between groups variation.
- Let's calculate an estimate of it from our R output.

The Intraclass Correlation

- We simply substitute estimates of the quantities in the equation on the previous slide.

$$\hat{\rho}_{IC} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} \quad (15)$$

$$= \frac{75.68}{75.68 + 314.19} \quad (16)$$

$$= \frac{75.68}{389.87} \quad (17)$$

$$= 0.194 \quad (18)$$

The Intraclass Correlation

- Recall that the intraclass correlation actually is in the form of a squared correlation coefficient.
- We would not be too surprised, then to discover that methodologists tend to classify an intraclass correlation of 0.01 as "small," 0.09 as "medium", and 0.25 as "large."
- Note that, when the intraclass correlation approaches 1, it means that variation between schools dwarfs variation within schools. In that case, in effect, each school can be reduced to a single data point, and nothing is lost by doing regression on the group means.
- When the intraclass correlation is low, then multilevel modeling has much to offer.

The Intraclass Correlation

- Power to detect effects is much greater when the intraclass correlation is low.
- This is reflected in a couple of power charts from Murnane and Willett (p. 124).

The Intraclass Correlation

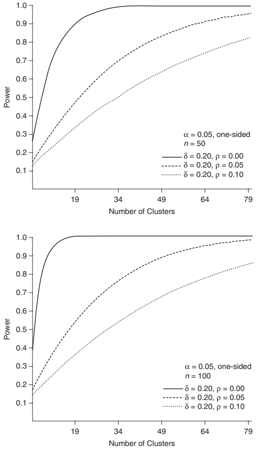


Figure 7.1 Anticipated relationship between statistical power of a cluster-randomized research design versus number of clusters (schools), for a small effect size (0.2), at three values of intraclass correlation (0, 0.05, and 0.1), α-level of 0.05, on a one-sided test. *Top panel*: 50 children/school. *Bottom panel*: 100 children/school.

The Intraclass Correlation

- A key lesson from the charts is that the number of clusters has a much greater influence on power than the number of observed subjects per cluster.
- A revealing equation (rearranged substantially from the form given in Murnane and Willett), shows the sampling variance of $\hat{\gamma}$ as

$$\text{Var}(\hat{\gamma}_{10}) = 4\sigma_e^2 \left(\frac{1}{nJ} + \frac{1}{J} \left[\frac{\rho}{1-\rho} \right] \right) \tag{19}$$

- When ρ is close to 0, the right term vanishes. When ρ is close to 1, the right term can become very large.
- With moderate values of ρ around 0.10, the right term can dominate the left, which explains how the number of clusters can assume greater importance than the number of subjects per cluster.
- And, of course, σ_e^2 dominates everything. Finding useful covariates can diminish σ_e^2 and greatly improve precision of estimation of your treatment effects.