Clustered Standard Errors

Erin Hartman

PS239 Sp2010

March 10, 2010

Clustered Standard Errors

"... analyses of group randomized trials that ignore clustering are an exercise in self-deception" (Cornfield 1978 pg. 101)

And, in the ever so enlightening words of Joshua Angrist:

"Making a data set larger by copying a smaller one n times generates no new information" (Angrist and Pischke 2009)

Clustering in the Modeling World

When we use parametric models, what do we usually assume about the errors?

Clustering in the Modeling World

When we use parametric models, what do we usually assume about the errors?

- Homoskedastic errors
- IID units

If we have clustering of our data, how are these assumptions violated?

Even when we run experiments in which, from Freedman, we know that the OLS estimates will be unbiased but the standard errors are wrong.

Malholtra and Kuo

TABLE 3 Rank-Ordered Logit Models Predicting Blame of Government Officials

Rep. Official	.61***	1.24***
	(.13)	(.22)
Rep. Official × Rep. Respondent	−.96***	−.76***
	(.20)	(.21)
Rep. Official × Party Cue	.56**	.57**
	(.20)	(.20)
Rep. Official × Rep. Respondent × Party Cue	−.96**	−.91**
	(.29)	(.30)
Rep. Official × Office Cue	−.37*	−.36*
	(.18)	(.18)
Rep. Official × Rep. Respondent × Office Cue	.16	.21
	(.28)	(.29)
Rep. Official × Both Cues	.06	.13
	(.19)	(.19)
Rep. Official × Rep. Respondent × Both Cues	−.41	−.55
	(.29)	(.29)
Rep. Official × Age	–	−.21
		(.22)
Rep. Official × Education	–	−.24
		(.16)
Rep. Official × White	–	−.07
		(.12)
Rep. Official × Male	–	.05
		(.11)
Rep. Official × South	–	−.28**
		(.11)
Rep. Official × Conservatism	–	−.75**
		(.24)
Log Likelihood	−2808.84	−2798.10
LR $\chi^2(8)$, $\chi^2(14)$	179.44***	200.91***

***$p<.001$; **$p<.01$; *$p<.05$; +$p<.10$ (two-tailed)
Note: N=2,380 for all regressions.

What is wrong with this picture?

- N of 340 respondents

- Each respondent is asked to rank 7 choices in a survey question.

Malholtra and Kuo

TABLE 3 Rank-Ordered Logit Models Predicting Blame of Government Officials

Rep. Official	.61***	1.24***
	(.13)	(.22)
Rep. Official × Rep. Respondent	−.96***	−.76***
	(.20)	(.21)
Rep. Official × Party Cue	.56**	.57**
	(.20)	(.20)
Rep. Official × Rep. Respondent × Party Cue	−.96**	−.91**
	(.29)	(.30)
Rep. Official × Office Cue	−.37*	−.36*
	(.18)	(.18)
Rep. Official × Rep. Respondent × Office Cue	.16	.21
	(.28)	(.29)
Rep. Official × Both Cues	.06	.13
	(.19)	(.19)
Rep. Official × Rep. Respondent × Both Cues	−.41	−.55
	(.29)	(.29)
Rep. Official × Age	–	−.21
		(.22)
Rep. Official × Education	–	−.24
		(.16)
Rep. Official × White	–	−.07
		(.12)
Rep. Official × Male	–	.05
		(.11)
Rep. Official × South	–	−.28**
		(.11)
Rep. Official × Conservatism	–	−.75**
		(.24)
Log Likelihood	−2808.84	−2798.10
LR $\chi^2(8)$, $\chi^2(14)$	179.44***	200.91***

***$p<.001$; **$p<.01$; *$p<.05$; +$p<.10$ (two-tailed)
Note: N=2,380 for all regressions.

What is wrong with this picture?

- N of 340 respondents

- Each respondent is asked to rank 7 choices in a survey question.

What is the random unit here?

Tennessee STAR experiment

The Tennessee STAR experiment randomly assigned 11,600 elementary school students and their teachers to a small class, regular-size class or regular-size class with a teacher-aide. The experiment began with the wave of students who entered kindergarten in 1985, and lasted for four years. After the third grade, all students returned to regular-size classes.

Say we have a dependent variable, test scores, Y_{ig} for each student i in group g, and the independent variable x_g which only varies at the group level. How would we model the response?

Tennessee STAR experiment

The Tennessee STAR experiment randomly assigned 11,600 elementary school students and their teachers to a small class, regular-size class or regular-size class with a teacher-aide. The experiment began with the wave of students who entered kindergarten in 1985, and lasted for four years. After the third grade, all students returned to regular-size classes.

Say we have a dependent variable, test scores, Y_{ig} for each student i in group g, and the independent variable x_g which only varies at the group level. How would we model the response?

$$Y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}$$

Tennessee STAR experiment

The Tennessee STAR experiment randomly assigned 11,600 elementary school students and their teachers to a small class, regular-size class or regular-size class with a teacher-aide. The experiment began with the wave of students who entered kindergarten in 1985, and lasted for four years. After the third grade, all students returned to regular-size classes.

Say we have a dependent variable, test scores, Y_{ig} for each student i in group g, and the independent variable x_g which only varies at the group level. How would we model the response?

$$Y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}$$

What is the random unit here?

Decomposing the Error

One way to give structure to the error is to assume that e_{ig} has a group structure.

$$e_{ig} = \nu_g + \eta_{ig}$$

where ν_g is a random component specific to class g and η_{ig} is a mean-zero student-level error component that is left over. We will assume that each of these components is homoskedastic and that ν_g captures all the within-group correlation. This is referred to as a random effects model.

Intraclass Correlation Coefficient

The intraclass correlation coefficient refers to the correlation that we see for two observations within a group:

$$E[e_{ig}e_{jg}] = \rho_e \sigma_e^2 > 0$$

where ρ_e refers to the intraclass correlation.
Given our error structure, we can define the intraclass correlation as:

$$\rho_e = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$$

Note that a ρ_e of 1 indicates that all the errors within a group are the same. What does this imply about the Y_{ig}?

Intraclass Correlation Coefficient

The intraclass correlation coefficient refers to the correlation that we see for two observations within a group:

$$E[e_{ig}e_{jg}] = \rho_e \sigma_e^2 > 0$$

where ρ_e refers to the intraclass correlation.
Given our error structure, we can define the intraclass correlation as:

$$\rho_e = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$$

Note that a ρ_e of 1 indicates that all the errors within a group are the same. What does this imply about the Y_{ig}? They are all the same! Thus, Angrist's quote at the beginning of the lecture.

The Moulton Factor

Let $V_c(\hat{\beta}_1)$ refer to the conventional OLS variance formula (where homoskedastic errors are assumed) and let $V(\hat{\beta}_1)$ refer to the correct sampling variance given our error structure. We can derive the ratio of these as follows:

$$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = 1 + (n-1)\rho_e$$

This is referred to as the Moulton factor. It tells us how to scale our naive standard errors to compensate for our clustered standard errors. If $\rho_e = 1$, what does this say about our naive standard errors that ignore intraclass correlation?

Holding sample size constant, the Moulton factor increases with group size (i.e. there are fewer clusters). Why?

The Moulton Factor

Let $V_c(\hat{\beta}_1)$ refer to the conventional OLS variance formula (where homoskedastic errors are assumed) and let $V(\hat{\beta}_1)$ refer to the correct sampling variance given our error structure. We can derive the ratio of these as follows:

$$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = 1 + (n-1)\rho_e$$

This is referred to as the Moulton factor. It tells us how to scale our naive standard errors to compensate for our clustered standard errors. If $\rho_e = 1$, what does this say about our naive standard errors that ignore intraclass correlation?

Holding sample size constant, the Moulton factor increases with group size (i.e. there are fewer clusters). Why? Because as number of clusters decreases, the amount of independent information decreases. Information is independent across groups, but not within.

Small Numbers of Groups are Problematic

Number of groups is very important... too few groups will cause problems! At least in the modeling world.

Unfortunately, we often only have a few groups in our experiments.

So, in effect, we don't have a lot of information when we block because our randomization is done at a very high level. The lack of independent observations limits what we can say. It is important to see what we only get identification at the level at which we randomize!

We can use bootstrapping and randomization inference to get consistent standard errors. We will get back to this.

Robust Clustered Standard Errors

One solution to the Moulton problem is to used a robust-clustered standard errors approach. This is based on a "sandwich" estimator. Recall that our usual OLS variance estimate is:

$$\begin{aligned}\hat{\Sigma}_{ols} &= (X'X)^{-1}X'E(\epsilon\epsilon'|X)X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_{n\times n}X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\end{aligned}$$

where we estimate this using sample variance.

What part of this is no longer true with clustered data?

Robust Clustered Standard Errors

One solution to the Moulton problem is to used a robust-clustered standard errors approach. This is based on a "sandwich" estimator. Recall that our usual OLS variance estimate is:

$$\begin{aligned}\hat{\Sigma}_{ols} &= (X'X)^{-1}X'E(\epsilon\epsilon'|X)X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_{n\times n}X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\end{aligned}$$

where we estimate this using sample variance.

What part of this is no longer true with clustered data?
With clustered standard errors, $E(\epsilon\epsilon'|X) \neq \sigma^2 I_{n\times n}$

Robust Clustered Standard Errors

Now we use a new covariance matrix structure in which we allow for variance within clusters.

$$\hat{\Sigma}_{cl} = (X'X)^{-1}\Big(\sum_g X_g \hat{\psi}_g X_g\Big)(X'X)^{-1}$$

where

$$\begin{aligned}\hat{\psi}_g &= a\hat{e}_g\hat{e}_g' \\ &= a\begin{bmatrix} \hat{e}_{1g}^2 & \hat{e}_{1g}\hat{e}_{2g} & \cdots & \hat{e}_{1g}\hat{e}_{n_g g} \\ \hat{e}_{1g}\hat{e}_{2g} & \hat{e}_{2g}^2 & \cdots & \hat{e}_{2g}\hat{e}_{n_g g} \\ \vdots & \vdots & & \hat{e}_{n_g-1,g}\hat{e}_{n_g g} \\ \hat{e}_{1g}\hat{e}_{n_g g} & \cdots & \hat{e}_{n_g-1,g}\hat{e}_{n_g g} & \hat{e}_{n_g g}^2 \end{bmatrix}\end{aligned}$$

Robust Clustered Standard Errors

X_g is a set of group regressors.

- Note that this is consistent for any within-group variance structure.

- We could draw out the whole covariance matrix as a blocked diagonal matrix.

- Notice that the sums are done over groups. LLN will kick in only when there are a large number of groups. This is why we get worse estimates for small numbers of groups.

Aggregating Data

Given that we think that the effects are identified at the group level, we can overcome our standard error problems by simply aggregating the data.

$$\bar{Y}_g = \beta_0 + \beta_1 x_g + \bar{e}_g$$

weighted by the group size. Because the group means are approximately normal for moderately sized groups, we might get slightly more reliable results than RCSE when we have fewer groups.

DIff-in-Diff Designs

Say we have a panel of wage earnings by various demographic groups over time by state. What is the random unit here?

DIff-in-Diff Designs

Say we have a panel of wage earnings by various demographic groups over time by state. What is the random unit here?

The state-year unit.
If we write a model for state-year effects:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \nu_{st} + \eta_{ist}$$

where D_{st} indicates treatment states in post-treatment years.

Then we could analyze it using the methods described above.

DIff-in-Diff Designs

Say we have a panel of wage earnings by various demographic groups over time by state. What is the random unit here?

The state-year unit.
If we write a model for state-year effects:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \nu_{st} + \eta_{ist}$$

where D_{st} indicates treatment states in post-treatment years.

Then we could analyze it using the methods described above.

However, if we only have two states and two time periods, how would we disentangle state-year shocks from our treatment effect?

DIff-in-Diff Designs

Say we have a panel of wage earnings by various demographic groups over time by state. What is the random unit here?

The state-year unit.
If we write a model for state-year effects:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \nu_{st} + \eta_{ist}$$

where D_{st} indicates treatment states in post-treatment years.

Then we could analyze it using the methods described above.

However, if we only have two states and two time periods, how would we disentangle state-year shocks from our treatment effect?

Even with more data, now we must deal with serial correlation in clustered data!

The Bootstrap

Clustered standard errors boil down to determining what is the unit level at which we have randomness. One way to determine standard errors when we know the unit of randomization is with the bootstrap.

The Block Bootstrap

Given we know the unit of randomization is the group, when we resample for our bootstrap, what should we be resampling?

The Bootstrap

Clustered standard errors boil down to determining what is the unit level at which we have randomness. One way to determine standard errors when we know the unit of randomization is with the bootstrap.

The Block Bootstrap

Given we know the unit of randomization is the group, when we resample for our bootstrap, what should we be resampling?

We should resample at the group level.

In the Tennessee STAR experiment, how should we resample?

The Bootstrap

Clustered standard errors boil down to determining what is the unit level at which we have randomness. One way to determine standard errors when we know the unit of randomization is with the bootstrap.

The Block Bootstrap

Given we know the unit of randomization is the group, when we resample for our bootstrap, what should we be resampling?

We should resample at the group level.

In the Tennessee STAR experiment, how should we resample?

- At the classroom level

The Bootstrap

If we use the block bootstrap, then we compute our standard error by:

1) bootstrap $\hat{\beta}_1$ B times.
2) calculate the standard error as:

$$s_{\hat{\beta}_1,B} = \left(\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}^*_{1b} - \overline{\hat{\beta}^*_1})^2 \right)^{1/2}$$

with:

$$\overline{\hat{\beta}^*_1} = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}^*_{1b}$$

References

- Green and Vavreck "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches"
- Angrist and Pischke Mostly Harmless Econometrics
- Bertrand et. al "How much should we trust differences-in-differences estimates?"
- Cameron et al "Bootstrap Based Improvements for Inference with Clustered Errors"
- David Freedman On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors"
  - This is a very important paper: it basically says that if we don't believe our model, then what good is it to have a more precise estimate of an already biased estimate.