S-043/Stat-151
Analysis for Clustered and Longitudinal Data
(Multilevel & Longitudinal Models)

Lecture 6.2:
Cluster Robust Standard Errors

Note: A few slides taken from various decks found on Internet. These decks also posted in handouts folder.

Goals of this lecture

Building on Heteroskedastic Robust Standard Errors (the Sandwich Estimator) from last lecture…

Cluster Robust Standard Errors

- ★ Again we use classic OLS to get *population* trends.
- ★ Then, if we believe clusters are themselves an independent sample, we can use residuals to get very bad estimates of residual structure for each cluster.
- ★ We then average to get decent overall estimates.

This is a close sister to the sandwich, sharing the same intuition.

All of this is **only for the fixed effects**

Reminder: The residual variance matrix we assume for the sandwich

$$E(uu') = \begin{bmatrix} E(u_1u_1) & E(u_1u_2) & E(u_1u_n) \\ E(u_2u_1) & E(u_2u_2) & E(u_2u_n) \\ \\ E(u_nu_1) & E(u_nu_2) & E(u_nu_n) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ \\ 0 & 0 & \sigma_n^2 \end{bmatrix} = \Sigma$$

In this case we give each observation its own variance, but says that all the covariances between observations are zero.

To move to clustering, we change to a ***block diagonal*** matrix where some (but not all) of the off-diagonal terms can be non-zero.

The Huber-White Sandwich Estimator

$$Var(\hat{\beta}) = (X'X)^{-1} X'\hat{\Sigma}X (X'X)^{-1}$$

Bread

Meat

More Bread

The square root of the diagonal of this matrix gives

Heteroskedastic-Robust Standard Errors

The "meat" is made by plugging the residuals in to Sigma, our residual variance matrix structure.

Estimating the meat:
Estimating our residual covariance with a "plug in"

$$X'\hat{\Sigma}X = X'\begin{bmatrix} \hat{u}_1^2 & 0 & 0 \\ 0 & \hat{u}_2^2 & 0 \\ & & \\ 0 & 0 & \hat{u}_n^2 \end{bmatrix} X$$

We use our residuals for each observation to estimate the variance of each observation.

For any observation, this estimate is *terrible*.

Happily, when we average over them all (which the matrix multiply does) our average estimate is good!

Why this doesn't solve all our problems

With clustered data, the observations within a cluster are correlated.

I.e., the residuals for people in a cluster are similar.

This means our residual covariance matrix is **NOT** a diagonal matrix.

Example: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Consider students a, b in same school: the ϵ_a, ϵ_b are *correlated*. (Good school -> expect both residuals to be positive)

Consider students a, c in *different* schools: ϵ_a, ϵ_c are *NOT correlated*. (Knowing about student a tells us nothing about student b)

What is the covariance of a random intercept model?

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + r_j + \epsilon_{ij}$$

In the above random intercept model we have ***two*** error terms, one for the group, and one for the residual.

If we think of this as an *overall* error term, we have for the residuals of a group:

$$\Sigma = \begin{pmatrix} \sigma^2+\tau & \tau & 0 & 0 & 0 & 0 \\ \tau & \sigma^2+\tau & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2+\tau & \tau & \tau & 0 \\ 0 & 0 & \tau & \sigma^2+\tau & \tau & 0 \\ 0 & 0 & \tau & \tau & \sigma^2+\tau & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2+\tau \end{pmatrix}$$

Group 1

Group 2

Group 3

But we want to not say anything about what is going on *within group?*

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{ij}$$

If we just say residuals within group are correlated, but don't say *how*, we get the following

$$\Sigma = \begin{pmatrix} ?? & & & & & & \\ ?? & ?? & & & & & \\ 0 & 0 & ?? & & & & \\ 0 & 0 & ?? & ?? & & & \\ 0 & 0 & ?? & ?? & ?? & & \\ 0 & 0 & 0 & 0 & 0 & ?? & \\ 0 & 0 & 0 & 0 & 0 & ?? & ?? \end{pmatrix}$$

Group 1

Group 2

Group 3

This matrix is called "block diagonal"

Multilevel models put *structure* on our residuals. This is an assumption we might want to avoid.

I want more meat!

Moving on to the cluster robust variance estimator (CRVE)

Making ties to MLM and residuals. Consider our school model...

$$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$$ **Level 1 Model**

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

$$\alpha_j = \gamma_0 + \gamma_1 s_j + r_j$$ **Level 2 Model**

$$r_j \sim N(0, \tau)$$

Implied Assumptions:

Level 1 and Level 2 error terms are not correlated with each other or themselves.

The expected value of an error given the covariates is 0.

The variance of the error terms does not change with the covariate.

...collapsed into a single regression

This is the gateway to the econometric view

$$
\begin{aligned}
y_{ij} &= \alpha_j + \beta x_{ij} + \epsilon_{ij} \\
&= (\gamma_0 + \gamma_1 s_j + r_j) + \beta x_{ij} + \epsilon_{ij} \\
&= \gamma_0 + \gamma_1 s_j + \beta x_{ij} + r_j + \epsilon_{ij}
\end{aligned}
$$

We have three "fixed effects" coefficients to estimate.

The r_j are the "school-level random effects"

The error term.
These are NOT independent
We say $u_j = r_j + \epsilon_{ij}$

Random error is still an error

Our collapsed error has mean zero.

$$E[r_j + \epsilon_{ij}] = E[r_j] + E[\epsilon_{ij}] = 0$$

But for two units in the same cluster, we don't have independence

$$cov(r_j + \epsilon_{ij}, r_j + \epsilon_{kj}) = \tau + \sigma^2 \neq 0$$

Econometricians don't like all the distributional assumptions involved with MLM. They work with the **overall** error term *without imposing this entire structure*.

A taste of economics

We have the model

$$y_{ij} = \gamma_0 + \gamma_1 s_j + \beta x_{ij} + u_{ij}$$

$$= \gamma_0 + \gamma_1 s_j + \beta x_{ij} + r_j + \epsilon_{ij}$$

We know that for two units in the same cluster, we don't have independence.

Cluster Robust Standard Errors will let the errors *within* a cluster be all tangled up (but they all have mean 0 in expectation).

Errors *between* clusters are assumed independent.

Quick peek as to where we are going: More meat is the punchline

Our sandwich formula is true for clustered data as well:

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

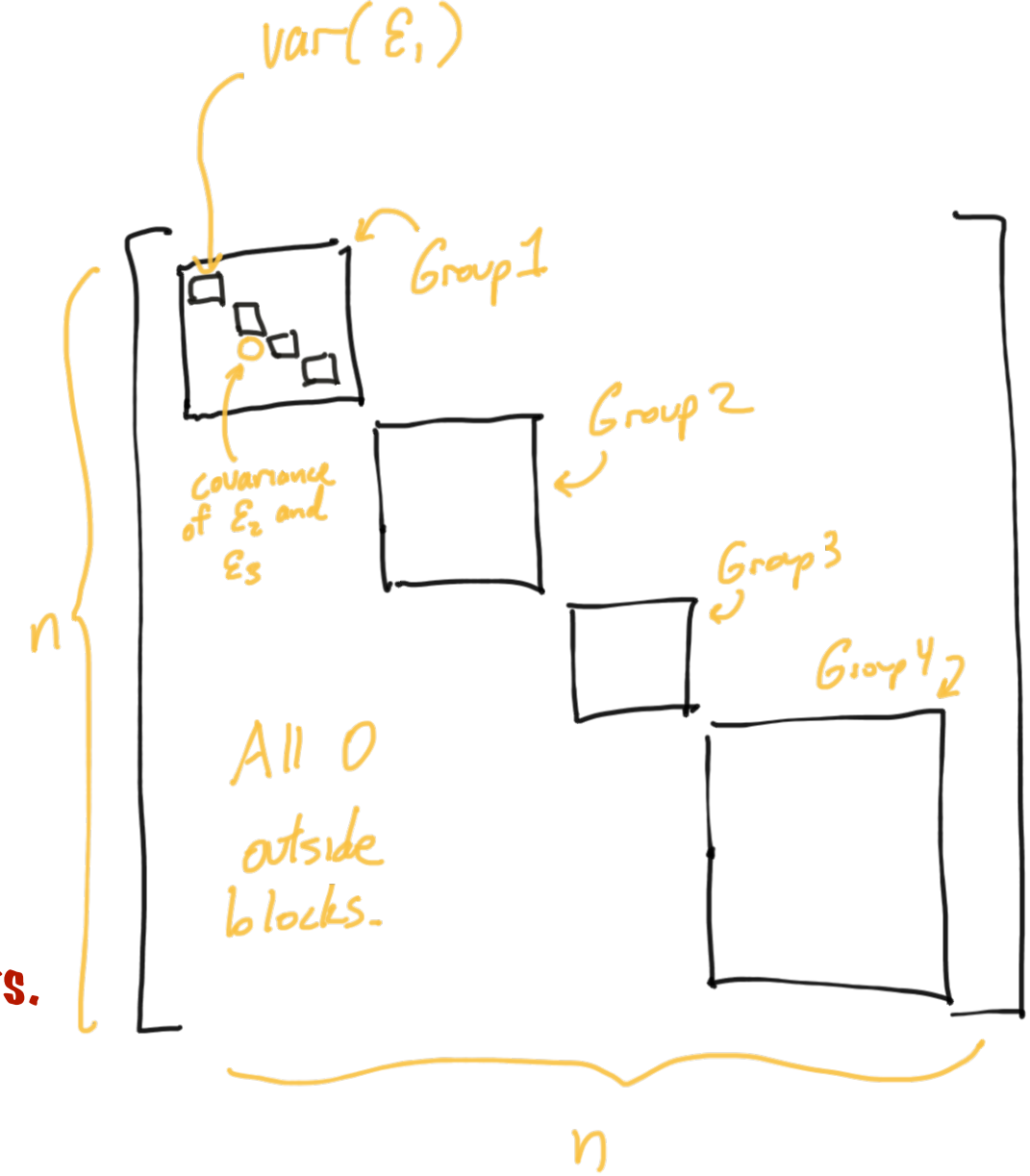
The key is that the residual covariance matrix Σ captures the clustering.

Core idea: Estimate Σ accounting for clustering and you have cluster-robust standard errors.

Variance-covariance of residuals with clustered data

Our variance-covariance of residuals

$$\Sigma =$$



If we can estimate this, we can plug it in to our sandwich formula to account for clustering!

We have 4 clusters. Cluster 1 has 4 units in it.

cluster-robust standard errors

```
> library( sandwich )
> M1 = lm( mathach ~ 1 + ses + sector, data = dat )
> vcov_clust = sandwich::vcovCL( M1, dat$id )
> vcov_clust
             (Intercept)         ses      sector
(Intercept)  0.04126811  0.00435265 -0.04263858
ses          0.00435265  0.01636795 -0.01173884
sector      -0.04263858 -0.01173884  0.10060102
> library( lmtest )
> coeftest( M1, vcov. = vcov_clust )

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 11.79325    0.20315 58.0532 < 2.2e-16 ***
ses          2.94856    0.12794 23.0469 < 2.2e-16 ***
sector       1.93501    0.31718  6.1007 1.111e-09 ***
```

Remember our High-School and Beyond data?

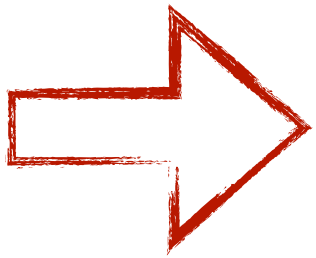
Compare with nominal SEs

```
> se.coef( M1 )
(Intercept)         ses      sector
 0.10610213  0.09783058  0.15249341
```

So what is happening? (Peeking into the math for cluster-robust SEs)

Stacking Many Worlds to Make One World

$$Y_1 = X_1\beta + u_1$$
$$Y_2 = X_2\beta + u_2$$
$$\dots$$
$$Y_G = X_G\beta + u_G$$



$$Y = X\beta + u$$

An n-vector of outcomes

An n X p matrix of all covariates

Each is a mini-regression with an n_g x p matrix of covariates, an n_g vector of outcomes, etc.

When we have grouping, we can write a single regression for each group.

The β is shared across groups, giving our final equation at right

The cluster regression model

- Model for G clusters with N_g individuals per cluster:

$$\begin{aligned} y_{ig} &= \mathbf{x}'_{ig}\boldsymbol{\beta} + u_{ig}, \quad i = 1, ..., N_g,\ g = 1, ..., G, \\ \mathbf{y}_g &= \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}_g, \quad g = 1, ..., G, \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \end{aligned}$$

These three equations are three expressions of the same thing.

- OLS estimator

We are summing across our clusters, but this is really...

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{x}_{ig}\mathbf{x}'_{ig})^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{x}_{ig}y_{ig}) \\ &= (\sum_{g=1}^{G}\mathbf{X}'_g\mathbf{X}_g)^{-1}(\sum_{g=1}^{G}\mathbf{X}_g\mathbf{y}_g) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

...classic OLS

We also are assuming $E[u_{ig}|x_{ig}] = 0$

I.e., correct mean model

OLS with Clustered Errors (continued)

- As usual

Our estimate of the parameters is the truth plus an error term (that depends on the residuals and on X)

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g\mathbf{u}_g\right).\end{aligned}$$

Again, sum across groups

- Assume independence over g and correlation within g

$$\mathsf{E}[u_{ig}u_{jg'}|\mathbf{x}_{ig},\mathbf{x}_{jg'}] = 0, \text{ unless } g = g'.$$

This is a theorem

- Then $\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, \mathsf{V}[\widehat{\boldsymbol{\beta}}]]$ with asymptotic variance

$$\begin{aligned}\mathsf{Avar}[\widehat{\boldsymbol{\beta}}] &= \underbrace{(\mathsf{E}[\mathbf{X}'\mathbf{X}])^{-1}}_{\text{bread}}\underbrace{\left(\sum_{g=1}^{G}\mathsf{E}[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g']\right)}_{\text{meat}}\underbrace{(\mathsf{E}[\mathbf{X}'\mathbf{X}])^{-1}}_{\text{bread}} \\ &\neq \sigma_u^2(\mathsf{E}[\mathbf{X}'\mathbf{X}])^{-1}\end{aligned}$$

Asymptotic variance (i.e., lots of GROUPS).

I.e., we don't get our usual OLS uncertaintly.

The variance of our estimator

The general variance expression is

$$V[\hat{\beta}] = (X'X)^{-1}B(X'X)^{-1}$$

This is the bread and meat again!

with

$$B = X'V[u|X]X$$

meat

Clustered data has a *block diagonal structure* which gives

$$B_{clu} = \sum_{g=1}^{G} X_g' E[u_g u_g' | X_g] X_g = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} x_{ig} x_{jg}' \omega_{ig,jg}$$

with

$$\omega_{ig,jg} = E[u_{ig} u_{jg} | X_g]$$

being the error covariance between two observations

Concept Check

If we have a truly specified random-intercept model, then for observations *i* in group *g* (*g*=1,...,G) with

$$Y_{ig} = X_{ig}\beta + \alpha_g + \epsilon_{ig} \text{ with } \alpha_g \sim N(0, \tau)$$

$$u_{ig} = \alpha_g + \epsilon_{ig}$$

what will $\omega_{ig,jg} = E[u_{ig}u_{jg} \mid X_g]$ be for *i*, *j* with $i \neq j$, but same g?

A. 0

B. Some positive constant

C. Some negative constant

D. It will depend on the i and j

Our Cluster Robust Variance Estimator (CRVE)

We estimate our uncertainty in estimating our parameters as

$$\hat{V}_{clu}[\hat{\beta}] = (X'X)^{-1}\hat{B}_{clu}(X'X)^{-1}$$

with

$$\hat{B}_{clu} = \sum_{g=1}^{G} X_g'\hat{u}_g\hat{u}_g'X_g$$

This is just our variance expression with an estimate "plugged in." Same game as sandwich

using our G vectors of residuals calculated as

$$\hat{u}_g = y_g - X_g\hat{\beta}$$

Each GROUP gets a vector of residuals. We average the groups.


What do I need to know about these things?

When do we get big standard errors on our fixed effects?

The SEs are the diagonal of $V[\hat{\beta}]$, so we want to know when the diagonal of $V[\hat{\beta}]$ is large.

We have relatively large variances (big standard errors) compared to conventional OLS when:

- ★ Covariates vary mostly across cluster rather than within cluster (note how a group level covariate only varies cross cluster!)
- ★ Errors within cluster are correlated so the omegas are non-zero (generally positive)
- ★ N_g is large (many units inside a cluster)
- ★ Within-cluster regressor and error correlations are the same sign (which is typical).

Take-away messages

1) There can be a great loss of efficiency in OLS estimation if errors are correlated within cluster

> "More observations in the cluster doesn't tell you much more that is new."

2) But sometimes these methods give you *smaller* errors than you would expect.

3) Failure to control for within-cluster error correlation generally gives overly small standard errors

4) Getting "cluster-robust" SEs is straightforward, *as long as you can rely on the assumption that the number of* ***clusters*** *is going to infinity (i.e., is larger than around 30).*

Why it works, and a simple improvement

For each cluster we estimate the variance-covariance matrix with just a single vector of residuals: this gives a bad estimate

The “magic” is from the averaging of G of these bad estimates (just like the Sandwich from before).

Aside: You can improve estimation by undoing overfitting by scaling the residuals as

$$\tilde{u}_g = c\hat{u}_g \text{ with } c = \frac{G}{G-1}\frac{N-1}{N-K} \approx \frac{G}{G-1}$$

Warning: Small Number of Clusters is Bad

The CRVE basically depends on the average of G poorly estimated covariance matrices.

This means the Central Limit Theorem doesn't kick in as it should.

There are a wide variety of fixes, patches, and hacks to repair this. They are all fairly technical.

One approach is to use the bootstrap (see appendix or readings) or the *clubSandwich package*.

Rule of thumb: J < 30 is small

A bonus feature

The Cluster-Robust estimator also allows for heteroskedasticity so it is technically both cluster- and heteroskedastic-robust!

Specifying the clusters: grain size

It is not always obvious how to specify the clusters.

Moulton (1986, 1990)

- ★ cluster at the level of an aggregated regressor.

Bertrand, Duáo and Mullainathan (2004)

- ★ with state-year data cluster on states (assumed to be independent) rather than state-year pairs.

Pepper (2002)

- ★ cluster at the highest level where there may be correlation
- ★ e.g. for individual in household in state may want to cluster at level of the state if state policy variable is a regressor.

! The issue here is we don't have multilevel modeling. E.g., classrooms are not independent within school. So we need to go big to get independence.

Fixed Effect Models
and
improved SEs for them

What is a "fixed effect" model?

If each group has its own intercept, we have:

$$y_{ij} = x'_{ig}\beta + \alpha_g + u_{ig}$$

$$= x'_{ig}\beta + \sum_{h=1}^{G} \alpha_g d_{hg} + u_{ig}$$

The d_{hg} are G dummy variables, one for each group. $d_{hg} = 1$ if group g is equal to h.

Compare to random effects where we just put a distribution on the α's

Sad thing: Note how we talk about parameters (e.g., population slopes) as fixed effects. This "fixed effects" means each group gets its own intercept. "Fixed effect" is not a consistently used term in the literature.

Two ways of estimating fixed effect models

Method 1: Least squares dummy variable (LSDV)

- ★ Just run OLS on your covariates including the cluster dummy variables

Method 2: The Fixed Effects Estimator

- ★ Also called the Within Estimator
- ★ Use OLS to estimate the "mean-differenced" model of

$$(y_{ig} - \bar{y}_g) = (x_{ig} - \bar{x}_g)'\beta + (u_{ig} - \bar{u}_g)$$

- ★ We have subtracted group means from everything.

Why do this (and why not)?

Fixed effect models control for a limited form of *endogeneity* of regressors, i.e., we assume

$$Cov[x_{ig}, \alpha_g] \neq 0 \text{ but } Cov[x_{ig}, u_{ig}] = 0$$

If we don't allow fixed effects, the *estimation itself* is **biased:** more data will not give us what we want.

Issues with this approach:

- ★ No cluster-level variables (they get wiped out)
- ★ Still need to use CRVE to deal with rest of within-cluster correlation of the error.
- ★ **Contrary to popular belief,** fixed effects DO NOT FIX EVERYTHING! (See Bell et al. reading.)

Why fixed effects might not control within-cluster correlation

Take a school with classrooms with students (three-level)

We have fixed effects for school.

We still have correlation *within the school* for the classrooms.

One approach:

- ★ Use the fixed school effects to deal with bias,
- ★ and then use CRVE to deal with this classroom correlation.

Messy technicalities, and Stata gets a point

- ★ The finite sample correction with fixed effects is delicate (we need to account for the G additional covariates even if we do the mean centering).
- ★ The asymptotic theory is not a clean story.
- ★ Stata has a bunch of built-in options and methods to handle this.
- ★ See posted (very technical) Cameron and Miller paper (Section IIIB) and mine it for these Stata commands.

Comparing SE estimators across methods

CHICKEN

PIGEON

Simple OLS (WRONG) on NYS Data

```
> M1 = lm( ATTIT ~ 1 + AGE + FEMALE, data = nys1 )
> arm::display( M1 )
            coef.est coef.se
(Intercept) -0.47     0.07
AGE          0.06     0.01
FEMALE      -0.07     0.02
---
n = 1079, k = 3
residual sd = 0.25, R-Squared = 0.13

> SE.wrong = se.coef( M1 )
> SE.wrong
(Intercept)         AGE      FEMALE
    0.07311     0.00554     0.01550
```

These SEs are very wrong since they ignore repeat observations

We have heteroskedasticity AND correlated residuals within group

Cluster-robust SEs

```
> vcov_clust = sandwich::vcovCL( M1, nys1$ID )
> vcov_clust
             (Intercept)         AGE     FEMALE
(Intercept)     0.004234 -3.24e-04 -6.52e-04
AGE            -0.000324  2.66e-05  2.67e-05
FEMALE         -0.000652  2.67e-05  6.81e-04
> SE.clust = sqrt( diag( vcov_clust ) )
> SE.clust
(Intercept)         AGE      FEMALE
    0.06507     0.00516     0.02610


> SE.clust / SE.wrong
(Intercept)         AGE      FEMALE
       0.89        0.93        1.68
```

Here is how much the SEs changed. Above 1 means cluster robust are larger.

Random Intercept Model

```
> M2 = lmer( ATTIT ~ 1 + AGE + FEMALE + (1|ID), data=nys1 )
> display( M2 )
            coef.est coef.se
(Intercept) -0.47     0.06
AGE          0.06     0.00
FEMALE      -0.07     0.03

Error terms:
 Groups   Name        Std.Dev.
 ID       (Intercept) 0.18
 Residual             0.18
---
number of obs: 1079, groups: ID, 239
AIC = -204.9, DIC = -258
deviance = -236.4
> SE.inter = se.fixef( M2 )
> SE.inter / SE.wrong
(Intercept)         AGE      FEMALE
      0.754       0.719       1.683
> SE.inter / SE.clust
(Intercept)         AGE      FEMALE
      0.847       0.773       1.000
```

Random intercept models also are supposed to handle clustering (but not heteroskedasticity)

Quite similar to cluster SEs for level 2 covariate.

Random Slope Model

```
> # And random slope model
> M3 = lmer( ATTIT ~ 1 + AGE + FEMALE + (1 + AGE|ID), data=nys1 )
> display( M3 )
            coef.est coef.se
(Intercept) -0.49     0.06
AGE          0.06     0.00
FEMALE      -0.06     0.02

Error terms:
 Groups   Name        Std.Dev. Corr
 ID       (Intercept) 0.52
          AGE         0.05     -0.98
 Residual             0.16
---
number of obs: 1079, groups: ID, 239
AIC = -308.7, DIC = -366
deviance = -344.4
> SE.slope = se.fixef( M3 )
> SE.slope / SE.wrong
(Intercept)         AGE      FEMALE
      0.814       0.891       1.427
```

Random slope could help with the heteroskedasticity

Compare them all!

```
> bind_rows( SE.wrong = SE.wrong,
+            SE.clust = SE.clust,
+            SE.inter = SE.inter,
+            SE.slope = SE.slope, .id = "Model")
# A tibble: 4 x 4
  Model      `(Intercept)`     AGE FEMALE
  <chr>              <dbl>   <dbl>  <dbl>
1 SE.wrong          0.0731 0.00554 0.0155
2 SE.clust          0.0651 0.00516 0.0261
3 SE.inter          0.0551 0.00399 0.0261
4 SE.slope          0.0595 0.00494 0.0221
```

Some observations:
Cluster SE: bigger on intercept and AGE
Random Slope smaller SE for FEMALE

Different data -> different trends entirely possible.
But larger SE for level 2 covariates generally true

Recap of robust standard errors

So what does all of this mean?

Take-aways

- ★ All these methods are targeted towards getting a good estimate of the variance-covariance matrix for your fixed effect estimator.
- ★ The reasoning is fairly technical, but the purpose is always this point.
- ★ These classes have provided a taste of what is out there, but the expectation of this course is not deep knowledge in this area.

The core comparison:

- ★ Multilevel modeling explicitly models error structure.
- ★ Cluster-robust methods avoid doing this to avoid making assumptions.

An admission, and some resources

- ★ Unfortunately R does not have clear support for dealing with the many, many weird and special cases of this approach.
- ★ Stata, however, does. Usually by tacking on a “robust” option to the end of your regression command.
- ★ The Cameron and Miller paper is dense and hard to decipher, but it has posted code and examples that you can explore, and it does sometimes simply give concrete advice.

Appendix:
Some miscellaneous notes and observations

History of the CRVE

- ★ proposed by Liang and Zeger (1986) for grouped data
- ★ proposed by Arellano (1987) for the fixed effects estimator for short panels (where the grouping is on the individual)
- ★ popularized by incorporation in Stata as the cluster option (Rogers, 1993)

Equicorrelated Errors

This simple model allows us to see when clustering matters a lot.

- Suppose equicorrelation within cluster g

$$\text{Cor}[u_{ig}, u_{jg} | \mathbf{x}_{ig}, \mathbf{x}_{jg}] = \begin{cases} 1 & i = j \\ \rho_u & i \neq j \end{cases}$$

  - this arises in a random effects model with $u_{ig} = \alpha_g + \varepsilon_{ig}$, where α_g and ε_{ig} are i.i.d. errors.
  - an example is individual i in village g or student i in school g.

- The incorrect default OLS variance estimate should be inflated by

$$\tau_j \simeq 1 + \rho_{x_j} \rho_u (\bar{N}_g - 1),$$

  - ρ_{x_j} is the within cluster correlation of x_j
  - ρ_u is the within cluster error correlation
  - $\bar{N}_g$ is the average cluster size.
  - Kloek (1981), Scott and Holt (1982).

This is a "rule of thumb" formula made from our simple model.

(Prior page model gives us this moral)

Moral: Worry about correlation

- Moulton (1986, 1990) showed that the inflation can be large even if ρ_u is small
  - especially with a grouped regressor (same for all individuals in group) so that $\rho_x = 1$.
  - CPS data example: $N_g = 81$, $\rho_x = 1$ and $\rho_u = 0.1$ then $\tau_j \simeq 1 + \rho_{x_j}\rho_u(\bar{N}_g - 1) = 1 + 1 \times 0.1 \times 80 = 9$.
    - ★ true standard errors are three times the default!
- So should correct for clustering even in settings where not obviously a problem.

"Grouped regressor": A level-2 covariate that varies at the cluster level. These are hard to estimate because we need a covariate to vary to see how it is connected to outcome, and thus we effectively only have G observations, not n.

Aside: Function to compute Cluster Robust Errors in R by hand

```
> cl <- function(dat, fm, cluster){
    attach(dat, warn.conflicts = F)
    require(sandwich)
    require(lmtest)
    M <- length(unique(cluster))
    N <- length(cluster)
    K <- fm$rank
    dfc <- (M/(M-1))*((N-1)/(N-K))
    uj  <- apply(estfun(fm), 2, function(x)
                                tapply(x, cluster, sum));
    vcovCL <- dfc*sandwich(fm, meat=crossprod(uj)/N)
    coeftest(fm, vcovCL)
}

> cl(dat, mod, dat$school)

t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1180     0.0531    2.22    0.026 *
psafe         0.8533     0.0184   46.47   <2e-16 ***
```

Cluster bootstrap with asymptotic refinement

The bootstrap basically works, but you need to implement it correctly. See this paper, or others, or professor, for more information.

- Cameron, Gelbach and Miller (2007)
  - Test $H_0 : \beta_1 = \beta_1^0$ against $H_a : \beta_1 \neq \beta_1^0$ using $w = (\widehat{\beta}_1 - \beta_1^0)/s_{\widehat{\beta}_1}$
  - perform a cluster bootstrap with asymptotic refinement
  - then true test size is $\alpha + O(G^{-3/2})$ rather than usual $\alpha + O(G^{-1})$
  - hopefully improvement when G is small
  - wild cluster bootstrap is best.

This is one patch that is easy to implement and worth knowing about.
It seems to be superior to the mathematical adjustments

Wild cluster bootstrap

This is a specialized bootstrap where you keep your cluster residuals together as a block

1. Obtain the OLS estimator $\widehat{\boldsymbol{\beta}}$ and OLS residuals $\widehat{\mathbf{u}}_g$, $g = 1, ..., G$.
   - Best to use residuals that impose H_0.
2. Do B iterations of this step. On the b^{th} iteration:
   1. For each cluster $g = 1, ..., G$, form
      $\widehat{\mathbf{u}}_g^* = \widehat{\mathbf{u}}_g$ or $\widehat{\mathbf{u}}_g^* = -\widehat{\mathbf{u}}_g$ each with probability 0.5
      and hence form $\widehat{\mathbf{y}}_g^* = \mathbf{X}_g'\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}_g^*$.
      This yields wild cluster bootstrap resample $\{(\widehat{\mathbf{y}}_1^*, \mathbf{X}_1), ..., (\widehat{\mathbf{y}}_G^*, \mathbf{X}_G)\}$.
   2. Calculate the OLS estimate $\widehat{\beta}_{1,b}^*$ and its standard error $s_{\widehat{\beta}_{1,b}^*}$ and given these form the Wald test statistic $w_b^* = (\widehat{\beta}_{1,b}^* - \widehat{\beta}_1)/s_{\widehat{\beta}_{1,b}^*}$.
3. Reject H_0 at level α if and only if

$$w < w_{[\alpha/2]}^* \text{ or } w > w_{[1-\alpha/2]}^*,$$

where $w_{[q]}^*$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.

Appendix: Robust Errors for Random Effects Models

Generalized Least Squares

$$Y = X\beta + \epsilon$$
$$E[\epsilon|X] = 0$$
$$Var[\epsilon|X] = \Sigma$$

We assume we know all of this, including what Sigma looks like

In this framework, Sigma is known and we have

$$\hat{\beta} = \arg\min_b (Y - Xb)'\Sigma^{-1}(Y - Xb)$$

giving

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

This is OLS but we take the correlation structure of our errors into account *in the estimation process*.

One problem: we don't typically know Sigma.

Feasible Generalized Least Squares

In FGLS, we proceed in two stages:

(1) the model is estimated by OLS or another consistent (but inefficient) estimator, and the residuals are used to build a consistent estimator of the errors covariance matrix

(2) using the consistent estimator of the covariance matrix of the errors, we can implement GLS ideas.

In particular, we can get more plausible estimates of uncertainty.

Example: Within-cluster Equicorrelation

Each group has a specified covariance form and then

$$\Omega = Diag[\Omega_g]$$

If we have a consistent estimator of Sigma, we have

$$\hat{\beta}_{FGLS} = \left(\sum_{g=1}^{G} X_g' \Omega_g^{-1} X_g \right)^{-1} \sum_{g=1}^{G} X_g' \widehat{\Omega}_g^{-1} y_g$$

This estimator is efficient *only if the error variance is correctly specified.*

We can improve it with robust techniques...

Cluster-Robust FGLS

We have the following

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}}] = \left(\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\widehat{\Omega}_g^{-1}\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\widehat{\Omega}_g^{-1}\mathbf{X}_g\right)\left(\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X}\right)^{-1}$$

- where $\widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$ and $\widehat{\Omega} = \mathrm{Diag}[\widehat{\Omega}_g]$
- assumes $\mathbf{u}_g$ and $\mathbf{u}_h$ are uncorrelated, for $g \neq h$, and $G \longrightarrow \infty$.

And where do you get your candidate models for your Ω_g?

Random Effect Models!