Multiple Imputation for Multilevel Data

Multilevel Data

A unit of analysis is the what or whom being studied (e.g., observations, individuals, classrooms, groups, families, etc.)

Multilevel data structures have multiple units of analysis that are hierarchically nested

Lower-level units are nested within higher-level units

Some Terminology

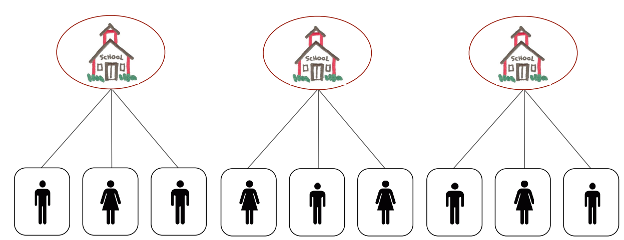
Observations at the lowest level of the data hierarchy are level-1 units

Observations at the second level of the hierarchy are level-2 units (or clusters)

e.g., Students at level-1 are nested within schools (clusters) at level-2, daily diary observations at level-1 are nested within individuals (clusters) at level-2

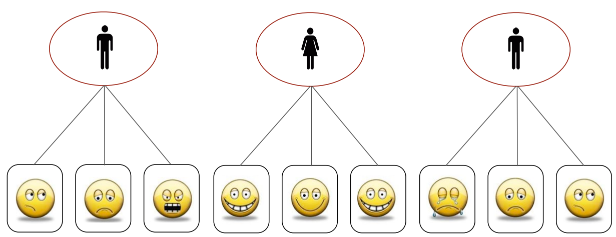
Multilevel Data Example 1

Sample comprised of multiple schools and several students in each school (i.e., students nested within schools)



Multilevel Data Example 2

Sample comprised of multiple individuals, each with several daily assessments of mood (i.e., observations nested within individuals)



Other Examples

Clients nested within therapists

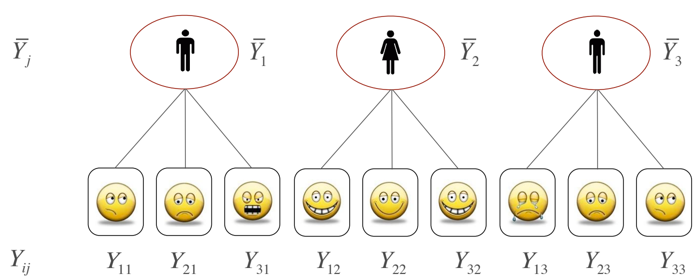
Twins nested within dyads

Individuals nested within families

Employees nested within workgroups

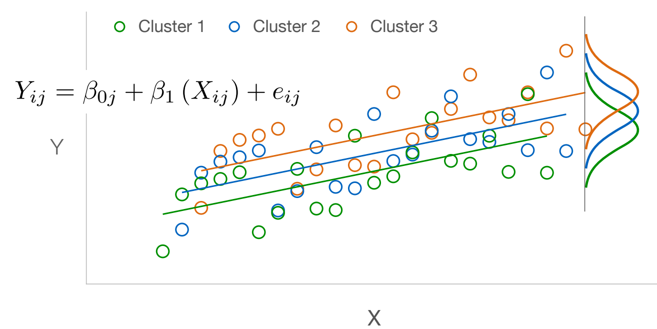
Mood Example

Mood scores vary across days within person, and average mood scores vary across persons



Random Intercept Model

A random intercept model is one where each cluster has unique intercepts (means) but the same slope



Level-2 Model and Reduced Form Model

Each cluster's intercept can be expressed as a mean intercept plus a residual

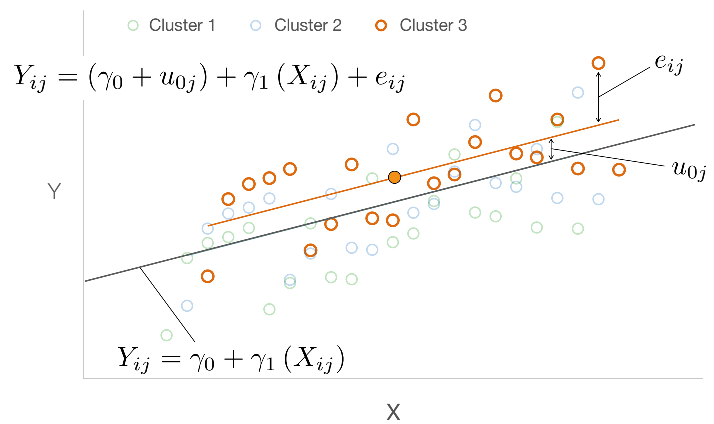
$$\beta_{0j} = \gamma_0 + u_{0j}$$

$$\beta_1 = \gamma_1$$

Substituting the right side of each equation into the level-1 model gives the reduced form model

$$Y_{ij} = (\gamma_0 + u_{0j}) + \gamma_1 (X_{ij}) + e_{ij}$$

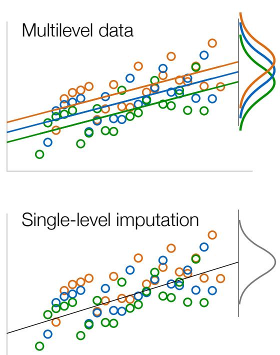
Random Intercept Model



Single-Level Imputation

Standard imputation routines assume a common distribution for all clusters (same means and variance-covariance matrix)

Single-level imputation will introduce substantial bias under any mechanism



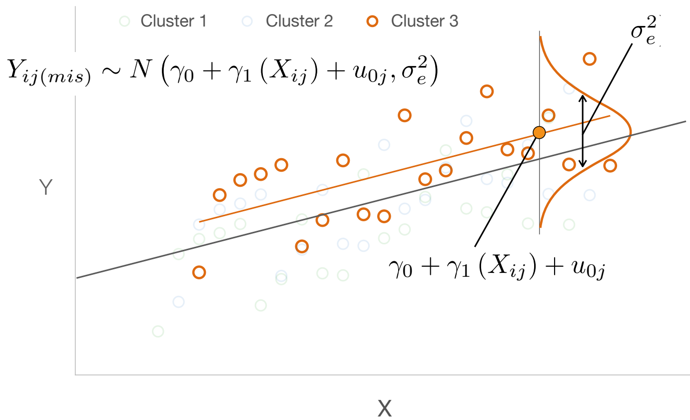
Multilevel Imputation

Multilevel imputation defines replacement values as a predicted score plus a normally distributed residual term

Each cluster has a unique regression line, so predicted scores account for between-cluster (level-2) differences

Within-cluster (level-1) variability defines the amount of random noise in the residuals

Random Intercept Imputation Model

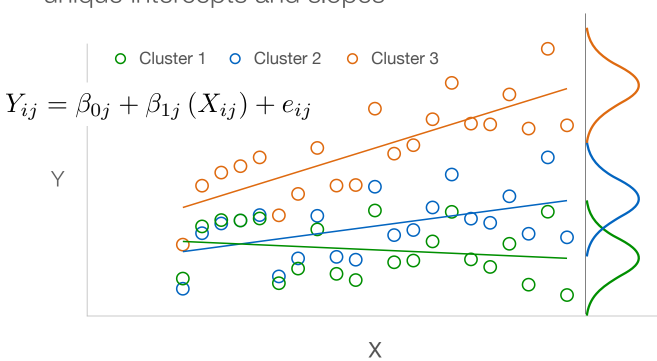


Blimp Imputation Script for Random Intercepts

```
DATA: ~/desktop/examples/data.csv;
VARNAMES: cluster y x;
MISSING: -99;
MODEL: cluster ~ y x;
SEED: 90291;
BURN: 1000;
THIN: 1000;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Random Slope Model

A random slope model is one where each cluster has unique intercepts and slopes



Level-2 Model and Reduced Form Model

Each cluster's intercept and slope can be expressed as a mean intercept and mean slope plus a residual

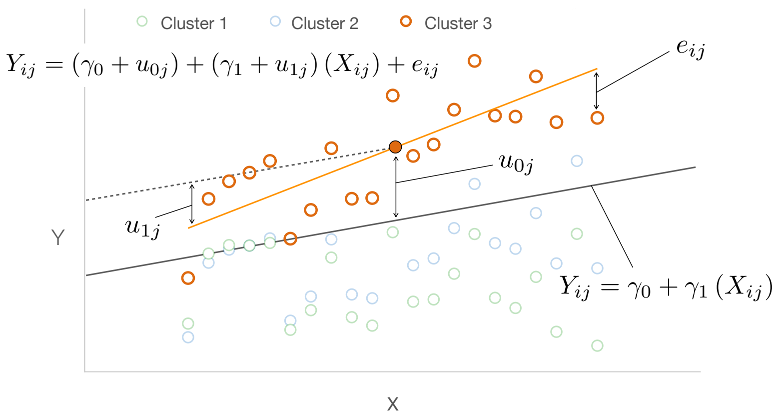
$$\beta_{0j} = \gamma_0 + u_{0j}$$

$$\beta_{1j} = \gamma_1 + u_{1j}$$

Substituting the right side of each equation into the level-1 model gives the reduced form model

$$Y_{ij} = (\gamma_0 + u_{0j}) + (\gamma_1 + u_{1j})(X_{ij}) + e_{ij}$$

Random Slope Model



Blimp Imputation Script for Random Slopes

```
DATA: ~/desktop/examples/data.csv;
VARNAMES: cluster y x;
MISSING: -99;
MODEL: cluster ~ y:x;
OUTCOME: y;
SEED: 90291;
BURN: 1000;
THIN: 1000;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Analysis Example

Analysis Example

Data from a math problem solving intervention with students nested in classrooms

50 classrooms with 25 students each are randomly assigned to an intervention or a control curriculum

Researchers collect a number of student-level variables (e.g., pre and post problem-solving scores, math self-efficacy) and school-level variables (percentage of minority students, intervention dummy code)

Analysis Model

The analysis is a linear growth curve model with a random slope for months since the start of the school year and a condition-by-time interaction effect

$$Probsolve_{ij} = (\gamma_0 + u_{0j}) + (\gamma_1 + u_{1j})(Months_{ij}) + \gamma_2(Efficacy_{ij})$$

$$+ \gamma_3(Stanmath_j) + \gamma_4(Condition_j) + \gamma_5(Months_{ij})(Condition_j) + e_{ij}$$

Ex8.1.imp
Blimp Diagnostic Script

```
DATA: ~/desktop/examples/probsolve.csv;
VARIABLES: school condition esolpercent student abilitygrp
   female stanmath frlunch wave months probsolve efficacy;
ORDINAL: condition efficacy frlunch;
OUTCOME: probsolve;
MISSING: -99;
MODEL: student ~  probsolve:months efficacy stanmath
condition months*condition frlunch;
NIMPS: 2;
BURN: 10000;
THIN: 1;
SEED: 90291;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```

Diagnostic Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations 4351 to 8700 for 2 chains.
                         ------------------------------------------------------
                         |      Fix Eff| Ran Eff Var|      Err Var|    Threshold|
                         ------------------------------------------------------
                 Max PSR |        1.004|        1.003|        1.001|        1.059|
        Missing Variable |     efficacy|     efficacy|    probsolve|     efficacy|
                         ------------------------------------------------------

  Comparing iterations 4401 to 8800 for 2 chains.
                         ------------------------------------------------------
                         |      Fix Eff| Ran Eff Var|      Err Var|    Threshold|
                         ------------------------------------------------------
                 Max PSR |        1.005|        1.002|        1.000|        1.048|
        Missing Variable |     efficacy|     efficacy|    probsolve|     efficacy|
                         ------------------------------------------------------
```

Ex8.2.imp
Blimp Imputation Script (Mplus Format)

```
DATA: ~/desktop/examples/probsolve.csv;
VARIABLES: school condition esolpercent student abilitygrp
   female stanmath frlunch wave months probsolve efficacy;
ORDINAL: condition efficacy frlunch;
OUTCOME: probsolve;
MISSING: -99;
MODEL: student ~  probsolve:months efficacy stanmath
condition months*condition frlunch;
NIMPS: 20;
BURN: 5000;
THIN: 2500;
SEED: 90291;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

Blimp Output

```
---------------------------------------------------------------

VARIABLE ORDER IN SAVED DATA:

   school condition esolpercent student abilitygrp
   female stanmath frlunch wave months probsolve efficacy

---------------------------------------------------------------
```

Ex8.3.inp
Mplus Analysis Script

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names =  school condition esolpercent student abilitygrp
   female stanmath frlunch wave months probsolve
 efficacy;
usevariables = probsolve months efficacy stanmath
 condition;
cluster = student;
within = months efficacy;
between = stanmath condition;
ANALYSIS:
type = twolevel random;
```

Ex8.3.inp
Mplus Analysis Script

```
MODEL:
%within%
ranslope | probsolve on months;
probsolve on efficacy;
%between%
probsolve on stanmath condition;
ranslope on condition;
probsolve with ranslope;
```

Mplus Analysis Output

```
MODEL RESULTS

                                                    Two-Tailed   Rate of
                    Estimate       S.E.  Est./S.E.    P-Value    Missing

Within Level

 PROBSOLVE  ON
    EFFICACY           0.720      0.133      5.418      0.000      0.494

 Residual Variances
    PROBSOLVE         62.549      1.760     35.539      0.000      0.321
```

Mplus Analysis Output

```
                                                    Two-Tailed   Rate of
                    Estimate       S.E.  Est./S.E.    P-Value    Missing
Between Level

 RANSLOPE   ON
    CONDITION          0.631      0.111      5.688      0.000      0.258

 PROBSOLVE  ON
    STANMATH           0.049      0.002     22.529      0.000      0.089
    CONDITION         -0.406      0.518     -0.784      0.433      0.157

 PROBSOLV WITH
    RANSLOPE           0.497      0.495      1.002      0.316      0.358

 Intercepts
    PROBSOLVE         73.424      1.150     63.822      0.000      0.116
    RANSLOPE           0.677      0.088      7.669      0.000      0.314

 Residual Variances
    PROBSOLVE         18.075      2.901      6.230      0.000      0.350
    RANSLOPE           0.347      0.130      2.673      0.008      0.410
```

Ex8.4.imp
Blimp Imputation Script (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/probsolve.csv;
VARIABLES: school condition esolpercent student abilitygrp
   female stanmath frlunch wave months probsolve efficacy;
ORDINAL: condition efficacy frlunch;
OUTCOME: probsolve;
MISSING: -99;
MODEL: student ~  probsolve:months efficacy stanmath
  condition months*condition frlunch;
NIMPS: 20;
BURN: 5000;
THIN: 2500;
SEED: 90291;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Blimp Output

```
--------------------------------------------------------------

VARIABLE ORDER IN SAVED DATA:

  imp# school condition esolpercent student abilitygrp
  female stanmath frlunch wave months probsolve efficacy

--------------------------------------------------------------
```

Ex8.5.r
R Analysis Script

```
# Required packages
library(mitml)
library(lme4)

# Read data
filepath <- "~/desktop/examples/imps.csv"
impdata <- read.csv(filepath, header = F)
names(impdata) <- c("imputation", "school", "condition",
  "esolpercent", "student", "abilitygrp", "female", "stanmath",
  "frlunch", "wave", "months", "probsolve", "efficacy")
impdata$conditionbytime <- impdata$months * impdata$condition

# Analyze data and pool estimates
implist <-  as.mitml.list(split(impdata, impdata$imputation))
analysis <- with(implist, lmer(probsolve ~ months + efficacy +
  stanmath + condition + conditionbytime + (months|student), REML = F))
estimates <- testEstimates(analysis, var.comp = T, df.com = NULL)
estimates
```

R Analysis Output

```
Final parameter estimates and inferences obtained from 20 imputed data sets.

               Estimate Std.Error   t.value        df   P(>|t|)       RIV       FMI
(Intercept)      73.423     1.120    65.564  1288.367     0.000     0.138     0.123
months            0.677     0.090     7.526   216.433     0.000     0.421     0.303
efficacy          0.720     0.132     5.459    79.432     0.000     0.957     0.501
stanmath          0.049     0.002    23.827  1961.783     0.000     0.109     0.099
condition        -0.406     0.520    -0.782   796.053     0.434     0.183     0.157
conditionbytime   0.631     0.112     5.651   304.107     0.000     0.333     0.255

                             Estimate
Intercept~~Intercept|student   18.075
Intercept~~months|student       0.496
months~~months|student          0.347
Residual~~Residual             62.549
ICC|student                     0.224

Unadjusted hypothesis test as appropriate in larger samples.
```

Ex8.6.sps
SPSS Analysis Script

```
data list free file = '/users/craig/desktop/examples/imps.csv'
 /imputation_ school condition esolpercent student abilitygrp
  female stanmath frlunch wave months probsolve efficacy.
exe.

* Initiate pooling routines.
sort cases by imputation_.
split file layed by imputation_.

* Analysis and pooling.
mixed probsolve with months efficacy stanmath condition
  /print = solution testcov
  /fixed = intercept months efficacy stanmath condition
   months*condition
  /random = intercept months | subject(student) covtype(un).
```

SPSS Analysis Output

Estimates of Fixed Effects

imputation_	Parameter	Estimate	Std. Error	df	t	Sig.
1.00	Intercept	73.060509	1.024685	1028.985	71.300	.000
	months	.702627	.076895	1017.335	9.138	.000
	efficacy	.796354	.094220	5076.361	8.452	.000
	stanmath	.049140	.001896	963.958	25.921	.000
...						
20.00	Intercept	73.068045	1.054519	1025.566	69.290	.000
	months	.656013	.073626	1035.066	8.910	.000
	efficacy	.747865	.095300	5183.331	7.847	.000
	stanmath	.048990	.001959	961.319	25.012	.000
	condition	-.183669	.477103	956.877	-.385	.700
	months * condition	.655700	.094142	966.332	6.965	.000
Pooled	Intercept	73.423899	1.121607		65.463	.000
	months	.677472	.090082		7.521	.000
	efficacy	.719974	.132043		5.453	.000
	stanmath	.048831	.002053		23.789	.000
	condition	-.406429	.520246		-.781	.435
	months * condition	.630883	.111729		5.647	.000

SPSS Analysis Output

Estimates of Covariance Parameters

imputation_	Parameter		Estimate	Std. Error	Wald Z	Sig.
1.00	Residual		62.236143	1.291955	48.172	.000
	Intercept + months [subject = student]	UN (1,1)	15.574233	2.313213	6.733	.000
		UN (2,1)	.661883	.401679	1.648	.099
		UN (2,2)	.434035	.105553	4.112	.000
...						
20.00	Residual		63.251133	1.312133	48.205	.000
	Intercept + months [subject = student]	UN (1,1)	17.577185	2.425691	7.246	.000
		UN (2,1)	.819120	.394488	2.076	.038
		UN (2,2)	.218377	.097110	2.249	.025
Pooled	Residual		62.556498	1.631058		.000
	Intercept + months [subject = student]	UN (1,1)	18.230427	2.977649		.000
		UN (2,1)	.484011	.502966		.337
		UN (2,2)	.351111	.131212		.008

Ex8.7.do
Stata Analysis Script

```
// Import and save original data
import delimited "~/desktop/examples/probsolve.csv"
rename (v1 - v12)(school condition esolpercent student abilitygrp female stanmath
  frlunch wave months probsolve efficacy)
generate imp = 0
generate id = student * 10000 + wave

// Recode missing values
foreach var of varlist school condition esolpercent student abilitygrp female stanmath
  frlunch wave months probsolve efficacy {
  replace `var' = . if `var'== -99
}
save original, replace

// Import and save imputed data
clear
import delimited "~/desktop/examples/imps.csv"
rename (v1 - v13)(imp school condition esolpercent student abilitygrp female stanmath
  frlunch wave months probsolve efficacy)
generate id = student * 10000 + wave
save imputed, replace
```

Ex8.7.do
Stata Analysis Script

```
// Append original and imputed data
use original, clear
append using imputed

// Convert to mi data
mi import flong, m(imp) id(id) imputed(school condition
  esolpercent student abilitygrp female stanmath frlunch wave
  months probsolve efficacy) clear

// Analyze data and pool results
mi estimate, cmdok: mixed probsolve months efficacy stanmath
  condition c.months#c.condition || student: months,
  covariance(un) var
```

Stata Analysis Output

```
Multiple-imputation estimates                  Imputations       =         20
Mixed-effects ML regression                    Number of obs     =      6,496

Group variable: student                        Number of groups  =        928
                                               Obs per group:
                                                            min =          7
                                                            avg =        7.0
                                                            max =          7
                                               Average RVI       =     0.3720
                                               Largest FMI       =     0.5015
DF adjustment:   Large sample                  DF:     min       =      79.43
                                                       avg       =     524.59
                                                       max       =   1,961.79
Model F test:       Equal FMI                  F(   5, 1453.9)   =     206.86
                                               Prob > F          =     0.0000

-------------------------------------------------------------------------------
          probsolve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------+----------------------------------------------------------
             months |   .6774599   .0900142     7.53   0.000     .5000432    .8548766
           efficacy |   .7202682   .1319383     5.46   0.000     .4576737    .9828626
           stanmath |   .0488297   .0020493    23.83   0.000     .0448106    .0528488
          condition |  -.4063836   .5195651    -0.78   0.434    -1.426263    .6134959
c.months#c.condition |   .6308712    .111636     5.65   0.000     .4111945    .8505479
              _cons |   73.42347   1.119879    65.56   0.000     71.22648    75.62045
-------------------------------------------------------------------------------
```

Stata Analysis Output

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
student: Unstructured        |
                  sd(months) |   .5799701   .1216618      .3827207    .8788794
                   sd(_cons) |    4.24296   .3508909      3.604052    4.995131
            corr(months,_cons) |   .2281315   .2984953     -.3708364    .6930637
-----------------------------+------------------------------------------------
               sd(Residual) |   7.908325   .1030426      7.707233    8.114663
------------------------------------------------------------------------------
```