STATISTICAL HORIZONS

Multilevel and Mixed Models Using Stata

Stephen Vaisey, Ph.D.

Upcoming Seminar:
July 13-16, 2021, Remote Seminar

What are MLMs?

Most broadly, they are models that estimate parameters from data with:

- multiple observations from the same group (e.g., students in classes)
- repeated observations from the same person
  - sometimes called panel data or longitudinal data
  - not the primary focus here but *very similar*

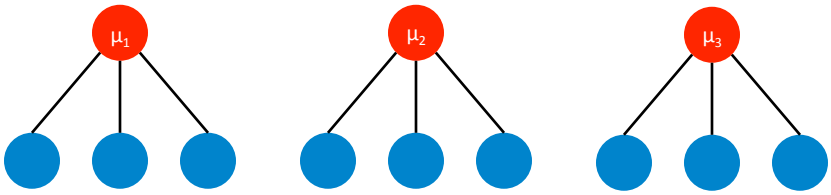
3

Why Use Multilevel Models?

- avoid problems with basic models (good)
- ask more interesting questions (better)

4

Hierarchical DGP



13

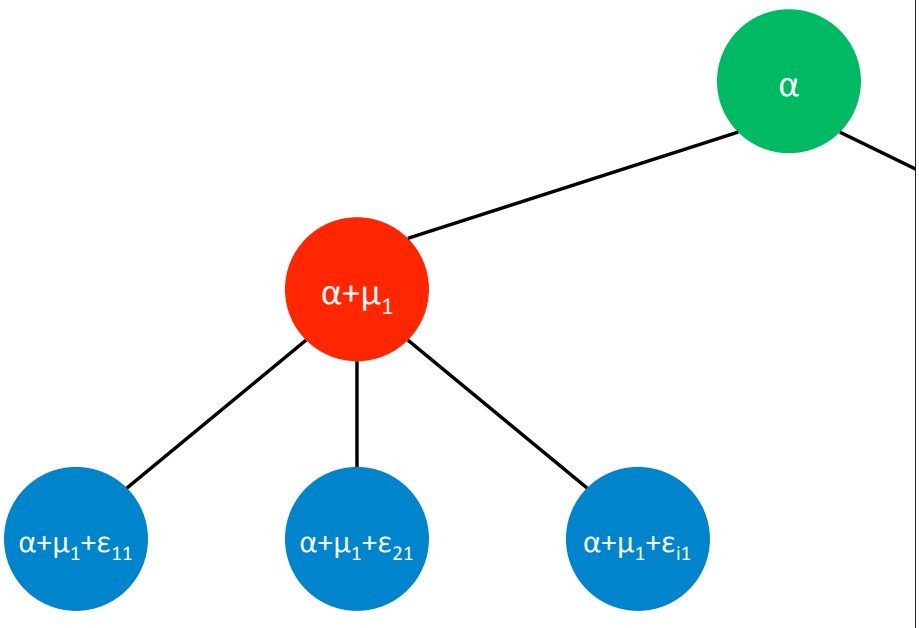
DGP Simulation Logic

```
set obs 5                    // # of clusters
gen u = rnormal(0,1)         // the different μ values
expand 10                    // how many "people" per cluster
gen y = u + rnormal(0,1)     // observed variable
```

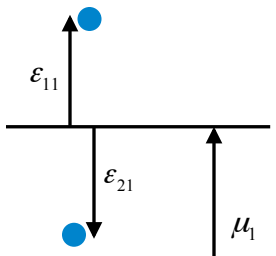
14

$$y_{ij} = \alpha + \mu_j + \varepsilon_{ij}$$

15



16



α

17

Different notations, same idea

$$y_{ij} = \alpha + \mu_j + \varepsilon_{ij}$$

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

$$y_{ij} = \beta + \zeta_j + \varepsilon_{ij}$$

$$y_{ij} = \beta_0 + u_j + e_{ij}$$

18

Main Point

Because the observations are not independent, using only a single error term (ε_i) is not justified. The errors of the individual observations may be correlated, violating a core assumption of OLS regression.

However, if that's your *only* concern, that can be dealt with using clustered SEs.

19

Example: Life Satisfaction in Europe

20

Data

200 person per country subsample of 2014 ESS (for faster estimation)

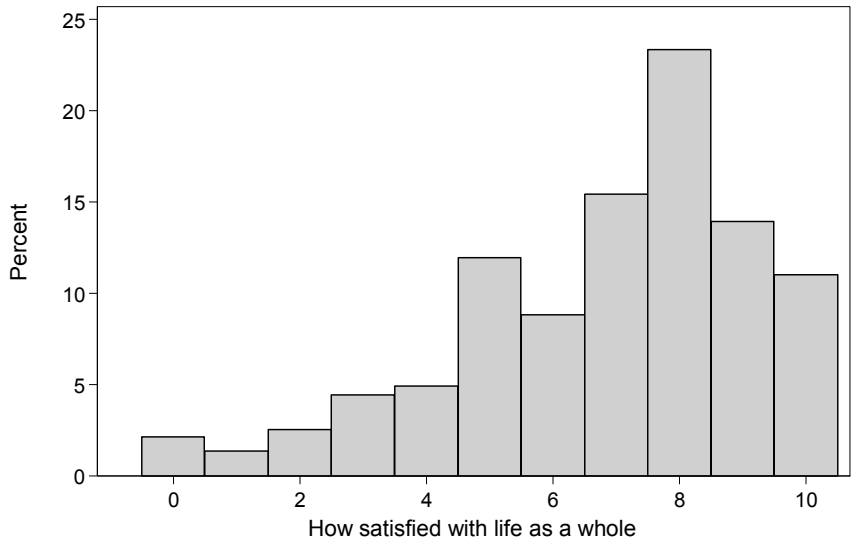
B20 **CARD 13** All things considered, how satisfied are you with your life as a whole nowadays? Please answer using this card, where 0 means extremely[18] dissatisfied and 10 means extremely satisfied.

Extremely dissatisfied										**Extremely satisfied**
00	01	02	03	04	05	06	07	08	09	10

21

How satisfied with life as a whole	Freq.	Percent	Cum.
Extremely dissatisfied	125	2.16	2.16
1	79	1.36	3.52
2	148	2.55	6.07
3	258	4.45	10.52
4	286	4.93	15.45
5	694	11.97	27.41
6	512	8.83	36.24
7	895	15.43	51.67
8	1,354	23.34	75.02
9	809	13.95	88.97
Extremely satisfied	640	11.03	100.00
Total	5,800	100.00	

22



23

```
. sum stflife, d

                How satisfied with life as a whole

      Percentiles      Smallest
 1%             0             0
 5%             2             0
10%             3             0       Obs                5,800
25%             5             0       Sum of Wgt.        5,800

50%             7                     Mean           6.783966
                        Largest       Std. Dev.      2.414616
75%             8            10
90%            10            10       Variance       5.830369
95%            10            10       Skewness      -.8581936
99%            10            10       Kurtosis       3.29755
```

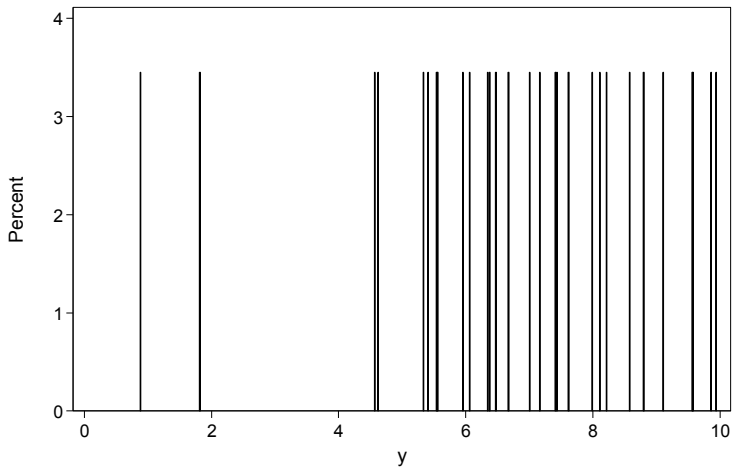
24

```
. tab country
```

Country	Freq.	Percent	Cum.
AL	200	3.45	3.45
BE	200	3.45	6.90
BG	200	3.45	10.34
CH	200	3.45	13.79
CY	200	3.45	17.24
CZ	200	3.45	20.69
DE	200	3.45	24.14
DK	200	3.45	27.59
EE	200	3.45	31.03
ES	200	3.45	34.48
FI	200	3.45	37.93
FR	200	3.45	41.38
GB	200	3.45	44.83
HU	200	3.45	48.28
IE	200	3.45	51.72
IL	200	3.45	55.17
IS	200	3.45	58.62
IT	200	3.45	62.07
LT	200	3.45	65.52
NL	200	3.45	68.97
NO	200	3.45	72.41
PL	200	3.45	75.86
PT	200	3.45	79.31
RU	200	3.45	82.76
SE	200	3.45	86.21
SI	200	3.45	89.66
SK	200	3.45	93.10
UA	200	3.45	96.55
XK	200	3.45	100.00
Total	5,800	100.00	

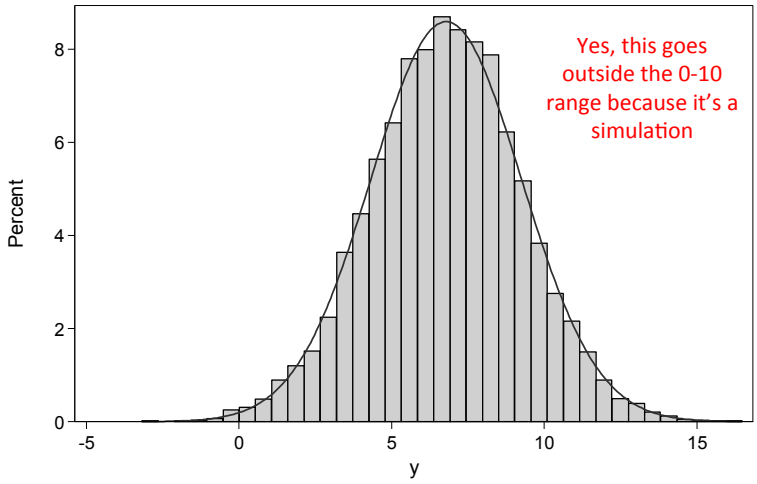
25

Two paths to the same variance: (1) all at the country level



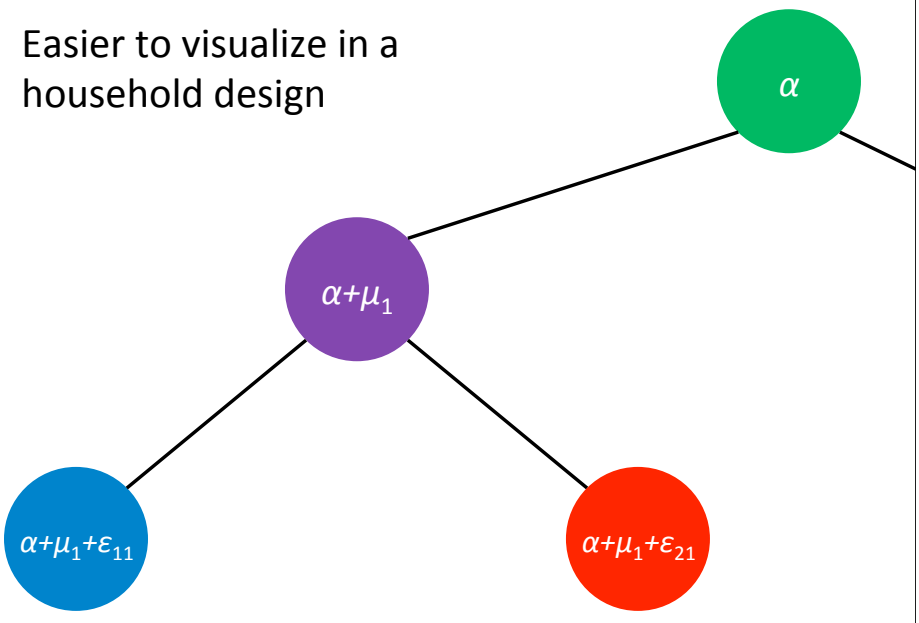
26

Two paths to the same variance: (2) all at the individual level



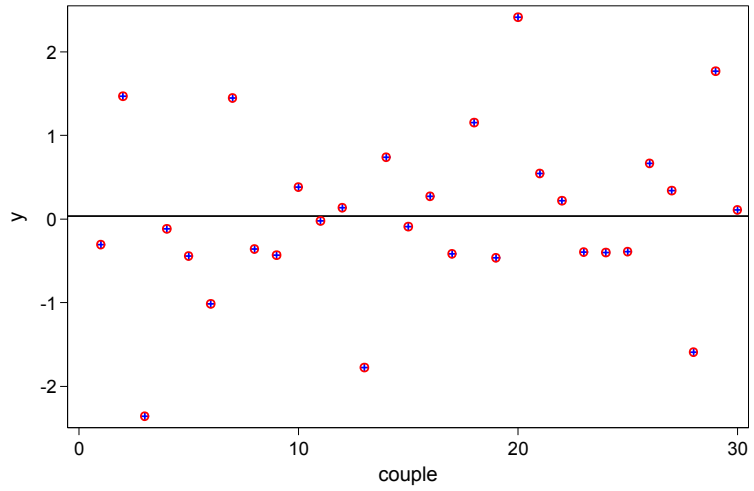
27

Easier to visualize in a household design



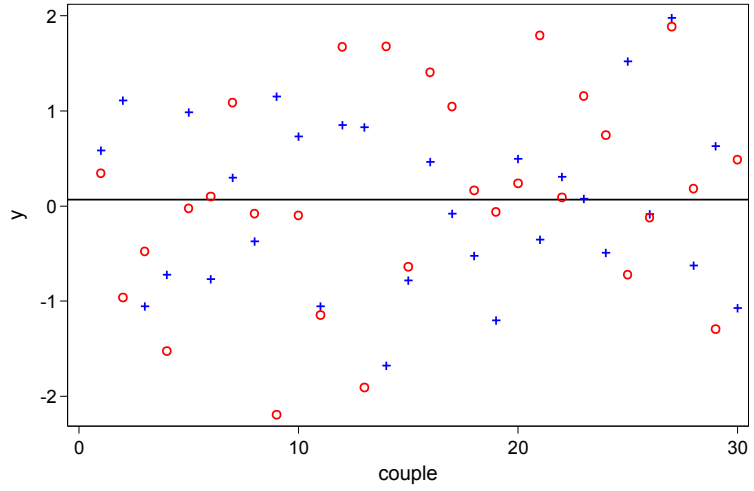
28

All variance at the household level



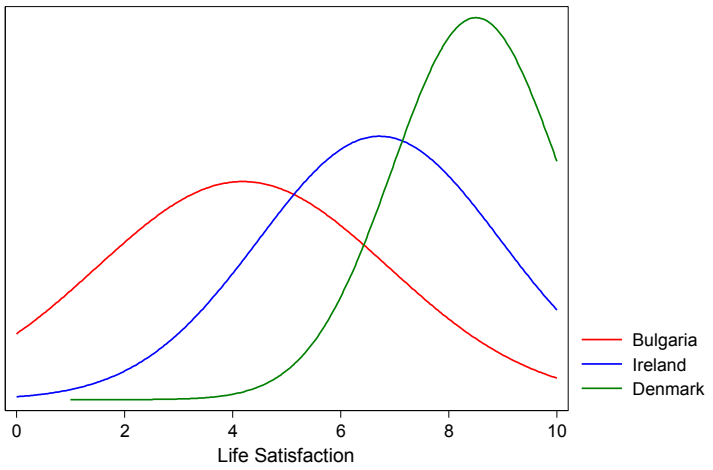
29

No variance at the household level



30

(As usual) a mix of both



31

Useful utilities

We use `xtset` to tell Stata that our dataset is clustered (although this is not really necessary for most analyses)

We use `xtsum` to ask for descriptive statistics about our variables

32

```
. xtset country
       panel variable:  country (balanced)

. xtsum stflife
```

Variable		Mean	Std. Dev.	Min	Max	Observations	
stflife	overall	6.783966	2.414616	0	10	N =	5800
	between		1.034484	4.185	8.5	n =	29
	within		2.190191	-1.351034	12.59897	T =	200

NOTE: the grand mean (here 6.784) is added back to the within estimates to make the numbers comparable.

33

Intraclass Correlation (ρ)

$$\rho = \frac{\sigma^2_{country}}{\sigma^2_{country} + \sigma^2_{individual}}$$

$$\rho = \frac{\psi}{\psi + \theta}$$

34

Estimating in Stata

We have a couple of options:

- `xtreg` is easy but can't do a lot
- `mixed` is more complicated and more flexible
  - called `xtmixed` in earlier versions of Stata
  - a few user-written postestimation commands work only after `xtmixed` (which still works)

35

```
. xtreg stflife, i(country) mle
Iteration 0:   log likelihood = -12845.508
Iteration 1:   log likelihood = -12845.508

Random-effects ML regression                    Number of obs      =      5,800
Group variable: country                         Number of groups   =         29

Random effects u_i ~ Gaussian                   Obs per group:
                                                              min =        200
                                                              avg =      200.0
                                                              max =        200

                                                Wald chi2(0)       =       0.00
Log likelihood  = -12845.508                    Prob > chi2        =          .
```

stflife	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	6.783966	.1887578	35.94	0.000	6.414007	7.153924
/sigma_u	1.004567	.1350565			.7718647	1.307424
/sigma_e	2.195497	.0204358			2.155807	2.235918
rho	.1731158	.0385865			.1079808	.2589746

LR test of sigma_u=0: chibar2(01) = 993.54 Prob >= chibar2 = 0.000

36

```
. mixed stflife || country:

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -12845.508
Iteration 1:   log likelihood = -12845.508  (backed up)

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =      5,800
Group variable: country                         Number of groups  =         29

                                                Obs per group:
                                                              min =        200
                                                              avg =      200.0
                                                              max =        200

                                                Wald chi2(0)      =          .
Log likelihood = -12845.508                     Prob > chi2       =          .
```

stflife	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	6.783966	.1887578	35.94	0.000	6.414007	7.153924

Random-effects Parameters		Estimate	Std. Err.	[95% Conf.	Interval]
country: Identity	var(_cons)	1.009155	.271346	.5957758	1.709356
	var(Residual)	4.820209	.0897336	4.647504	4.999331

LR test vs. linear model: chibar2(01) = 993.54 Prob >= chibar2 = 0.0000

37

```
. estat icc

Intraclass correlation
```

Level	ICC	Std. Err.	[95% Conf.	Interval]
country	.1731158	.0385865	.1098739	.2620427

We can say something like: “17% of the variance in life satisfaction is at the country level.”

This is why these are called
variance components models.

38

Getting the individual estimates

```
predict u, reffects                    // get the values
egen one = tag(country)                // pick one case per country
sort u
list country u if one, clean noobs
```

39

country	u
BG	-2.49272
UA	-1.536227
HU	-1.047103
RU	-.9942981
AL	-.9289942
FR	-.8954188
LT	-.8819007
EE	-.6304559
XK	-.6163967
IT	-.5671842
PT	-.5326819
CZ	-.4402117
SK	-.390595
IE	-.1248178
PL	.2240138
ES	.3247775
SI	.4002492
CY	.411091
GB	.4114189
IL	.5766028
BE	.6491744
DE	.6868622
NL	.8796129
SE	1.007696
IS	1.014848
FI	1.180357
NO	1.316033
CH	1.348022
DK	1.648247

40