Workshop on multilevel modeling – Part II

Bangalore, October 17–October 21, 2011

Stata practicals for advanced MLM

Belkacem Abdous
`Belkacem.Abdous@fmed.ulaval.ca`

Thierry Duchesne
`Thierry.Duchesne@mat.ulaval.ca`

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Workshop on multilevel modeling – Part II
Bangalore, October 17–October 21, 2011

Review of two-level logistic regression: Practical

Belkacem Abdous
Belkacem.Abdous@fmed.ulaval.ca

Thierry Duchesne
Thierry.Duchesne@mat.ulaval.ca

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Exercise 1

Exercise 6.10 of MLMUS2 reconsiders the wine tasting data analyzed by Fahrmeir & Tutz (2001). In that study, nine judges were asked whether various white wines had to be classified as bitter. The goal of the study was to see if conditions that can be controlled while pressing the grapes can influence the bitterness of the wine.

Variables present in the dataset `wine.dta` are

- `bitter` and `dichot:` response variable, with value of bitterness contained in `bitter` and recoded in `dichot` as y_{ij} that takes on value 1 if the wine is classified as bitter and 0 otherwise;
- `judge:` judge identifier, j;
- `temp:` temperature while pressing the grapes, x_{2ij}, that takes on value 1 if the temperature is low, 0 if the temperature is high;
- `contact:` skin contact while pressing the grapes, x_{3ij}, that takes on value 1 if there was skin contact, 0 otherwise;
- `repl:` replication (two bottles at each combination of temperature and skin contact were randomly sampled for each judge).

(a) Start with a descriptive analysis of the dataset (e.g., means of the variables `dichot`, `temp`, `contact`, 2×2 tables of `dichot` crossed with each of `temp` and `contact`, same summary statistics but done separately for each judge). From this descriptive analysis, try to guess what the effect of each of `temp` and `contact` will be on `dichot` and whether a judge-level random effect will have a significant variance.

(b) Each judge is likely to feel the bitterness of the wines differently. We would therefore like our analyses to include judge-level random effects. Fit the following two random intercept logistic models: both models with `dichot` as response, but one with only `temp` as covariate and one with only `contact` as covariate. Is any of these factors significant at the 0.25 level?

[Note: For part (b) a Laplace approximation of the integral involved in the likelihood will suffice.]

(c) Build the best random intercept model possible that may depend on `temp`, `contact` and/or their interaction. Make sure that your inferences are based on estimates obtained with likelihood approximations of high quality.

(d) Do judge-level random intercepts appear to be really needed here?

(e) Get estimates of each judge's random intercept term. Produce a caterpillar plot of these estimates. Does the plot agree with your answer to part (d)?

(f) Give an estimate of the within-judge correlation for the latent response variable.

(g) Test this hypothesis at the 0.05 significance level: the effect of lowering the temperature on the log of the odds of obtaining a bitter wine is twice the effect of touching the grapes with the skin.

[Hint: Very easy if you use the `lincom` post-estimation command.]

(h) A colleague of yours wonders if each judge reacts differently to a change in temperature. Fit a model that can give you an idea of the answer to your colleague's question.

Workshop on multilevel modeling – Part II

Bangalore, October 17–October 21, 2011

Subject-specific and population-averaged inferences in two-level logistic regression: Practical

Belkacem Abdous
Belkacem.Abdous@fmed.ulaval.ca

Thierry Duchesne
Thierry.Duchesne@mat.ulaval.ca

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Exercise 1

Let us go back to the wine tasting example.

(a) Using your best model from the previous practical, get estimates the probability of classifying a wine as bitter when the temperature is low and there was no skin contact for each of the nine judges (you may use Stata and get the answers from the Data Editor).

(b) Get an estimate of the probability from part (a), but for a new "typical" judge.

(c) What is the subject-specific effect of lowering the temperature on the log of the odds of classifying a wine as bitter?

(d) Use an approximation to give an estimate of the same effect as in part (c), but for the population-averaged probabilities.

(e) Repeat part (a), but use `gllamm` and `gllapred` to get population-averaged probabilities.

(f) In the end, you conclude that what is really interesting is inferring about the effect of skin contact and temperature on the overall proportion of wines that are bitter, not the judge-specific effects. Repeat parts (b) and (c) of of the previous practical using a population-averaged approach to inference. What could be an appropriate correlation structure for these data?

(g) Redo the test from part (g) of the previous practical.

(h) Redo part (e) above with the new model. Use the Data Editor to compare the predicted probabilities from the two approaches.

Workshop on multilevel modeling – Part II
Bangalore, October 17–October 21, 2011

Using MLwiN from within Stata

Belkacem Abdous
Belkacem.Abdous@fmed.ulaval.ca

Thierry Duchesne
Thierry.Duchesne@mat.ulaval.ca

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Introduction

In today's practical we will analyze subsets of two databases of actual ANC data (pregnant women attending antenatal clinics and tested for HIV) that were collected from 2003 to 2008 in several districts. We would like to fit two-level logistic regression models to these databases to see (i) how district-level variables on the high risk groups predic HIV in ANC women and (ii) whether the rate of change in HIV prevalence over time is different between district that have received an certain type of intervention compared to those where no such intervention was applied.

Getting to the final model is quite challenging; not only do we have a very large number of observations (several 10,000), but we have a large number of explanatory variables and, in some models, many random coefficients. We did fit these models using the maximum likelihood method and `xtmelogit` and, hence, it is possible to do everything covered in this practical session with the Stata code learnt so far. Unfortunately, the execution time required to evaluate and maximize the likelihood function for each model was large (several dozens of minutes per model), which makes it inappropriate for a practical session. The solution that we have found is to install the `MLwiN`[1] software package and call it from within `Stata`. This allows us to fit the same models as with `xtmelogit`, but by using an approximate likelihood method instead of an exact method based on numerical integration. For the models considered here, the difference in the values of the estimates

[1] Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009). MLwiN Version 2.1. Centre for Multilevel Modelling, University of Bristol.

is negligible, but the difference in execution time is astronomical! By using the Stata command runmlwin[2], we do not really need to learn how to use MLwiN in detail.

Setting up runmlwin

Since we create our datasets and perform all post-estimation (e.g., tests, predictions) with Stata, we would like to keep using Stata as much as possible and only send the model fitting task out to MLwiN. This is exactly what we can do with the Stata command runmlwin produced by Leckie and Charlton. Though all the information can be found on the runmlwin website (http://www.bristol.ac.uk/cmm/software/runmlwin/), we review the basic operations here.

1. Install MLwiN. You are entitled to a free trial period of 30 days. After that time, you can either purchase MLwiN or run the models using xtmelogit or gllamm. To install the software, visit http://www.bristol.ac.uk/cmm/software/mlwin/download/. Write down the path of the directory where mlwin.exe is located on your computer once the installation is complete, as it will be useful later.

2. If you will be dealing with large datasets (as will be the case in this practical), change the default spreadsheet size in MLwiN by selecting the Worksheet ... option from the Options menu. Simply change the 10000 default value for, say, 30000 then click on the Use as defaults button. See figures 1 and 2 below.

3. Install runmlwin. You can do this within Stata by typing

   . net from http://www.bristol.ac.uk/cmm/media/runmlwin

4. You are ready to start running MLwiN from within Stata.

A simple example

Consider the wine tasting data from yesterday's practical. We ran the following Stata code:

```
use http://www.stata-press.com/data/mlmus2/wine
* Declare the dataset as longitudinal, with judge as level 2 units
xtset judge
* In case model from Part I not saved:
generate Con_Temp = contact*temp
xtmelogit dichot ///
   contact temp ///
```

[2] Leckie, G. and Charlton, C. (2011). runmlwin: Stata module for fitting multilevel models in the MLwiN software package. Centre for Multilevel Modelling, University of Bristol.

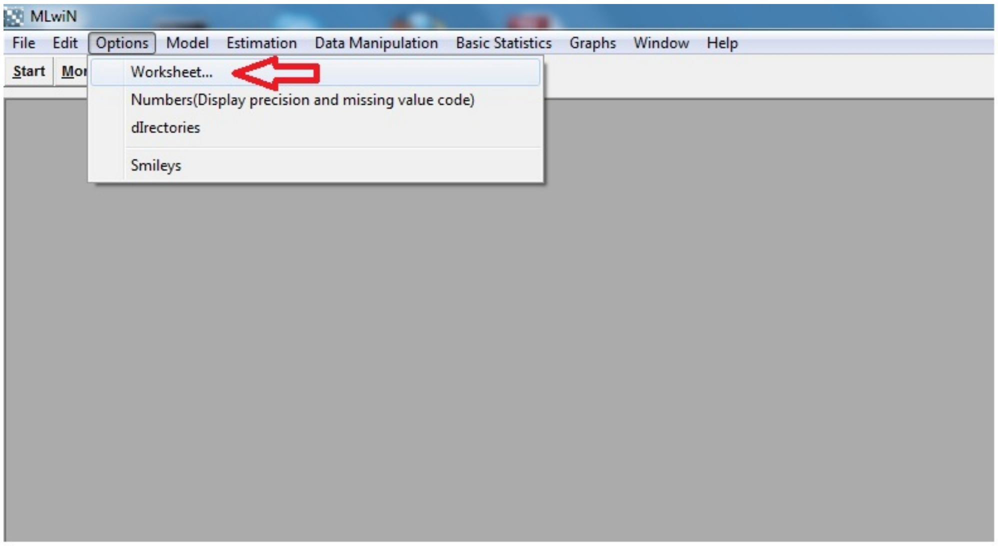
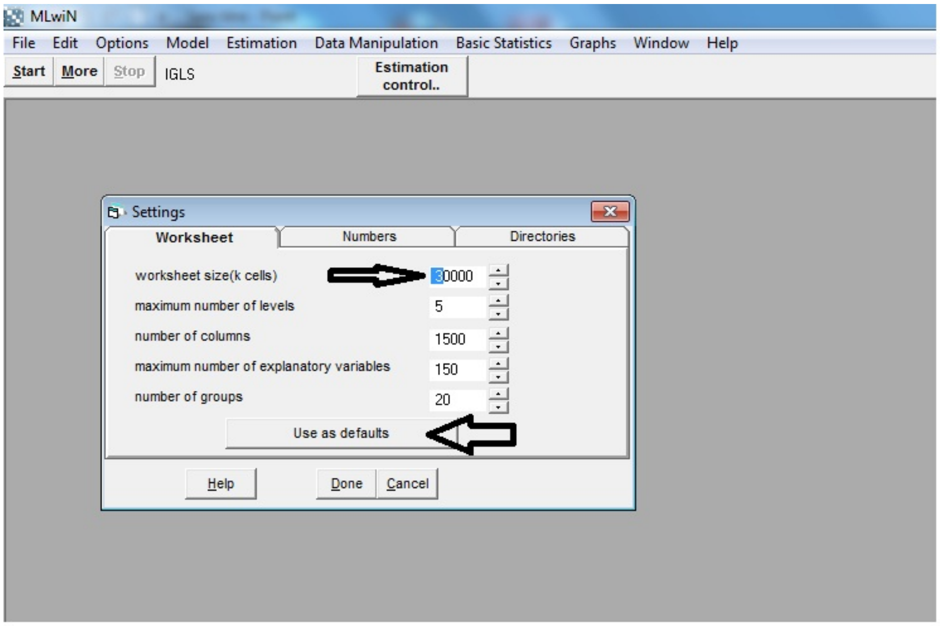

MLwiN
File Edit Options Model Estimation Data Manipulation Basic Statistics Graphs Window Help
Start Mor
Worksheet...
Numbers(Display precision and missing value code)
dIrectories
Smileys

Figure 1: Enabling MLwiN to deal with large datasets



worksheet size(k cells) 30000

maximum number of levels 5

number of columns 1500

maximum number of explanatory variables 150

number of groups 20

Figure 2: Enabling MLwiN to deal with large datasets

```
   || judge: ///
, variance intpoints(1)
estimates store contemp1
matrix ct1 = e(b)
xtmelogit dichot ///
   contact temp ///
   || judge: ///
, variance from(ct1,skip) refineopts(iterate(0)) intpoints(15)
estimates store contemp15
matrix ct15 = e(b)
```

How can we re-do this example, but by sending the model-fitting task out to MLwiN? Reading an preparing the data is done as previously:

```
use http://www.stata-press.com/data/mlmus2/wine
* Declare the dataset as longitudinal, with judge as level 2 units
xtset judge
generate Con_Temp = contact*temp
```

Now we must tell Stata where mlwin.exe can be found:

```
global MLwiN_path C:\Program Files (x86)\MLwiN trial\MLwiN.exe
```

We must also label the level 1 units (we have a variable judge that labels the level 2 units, but we do not have a label for level 1 units). The simplest way to do this is to number each line of the dataset 1, 2, etc. and put this number in a variable called, say, bottle:

```
generate bottle = _n
```

A variable taking on value 1 for all observations must be specified, as the intercept must be entered in the model as a covariate.

```
generate cons = 1
```

The dataset must then be sorted according to the level 2, level 1 hierarchy.

```
sort judge bottle
```

We are now ready to fit the model. You can get all the information about the syntax and options of runmlwin by typing help runmlwin in Stata.

Syntax of the `runmlwin` command

runmlwin *response_then_fixed_effects*
, *random_effects*
discrete(distribution(binomial) link(logit) denominator(*intercept_variable***)** *estimation_method***)**
[**constraints(***num***) initsmodel=***model_name* **or nopause**]

The "*response_then_fixed_effects*" works exactly as with `xtmelogit`: we simply write the name of the response variable and then the names of the covariates, the latter including the name of the intercept variable.

The syntax of the "*random_effects*" part is somewhat different from that of `xtmelogit`. Commands **level***i***(***label***:** [*random_coefs***,** *covar_struct* **residuals(***name***)**]**)** must be given to specify the label of the level 1 and level 2 units as well as the list of random coefficients at level 2. For instance for a random intercept model with the Ramesh data, if we label the level 1 units `fsw`, then the random part would be specified as `level2(dist:  cons) level1(fsw)`. If there are many random effects at the same level, the covariance matrix will be unstructured by default. The option **diagonal** must be specified to set the covariances between random effects to zero. The option `residuals`*name* creates variables *name*`0` *name*`1 ...` *name*`0se` *name*`1se ...` that contain the random effect estimates as well as their standard errors.

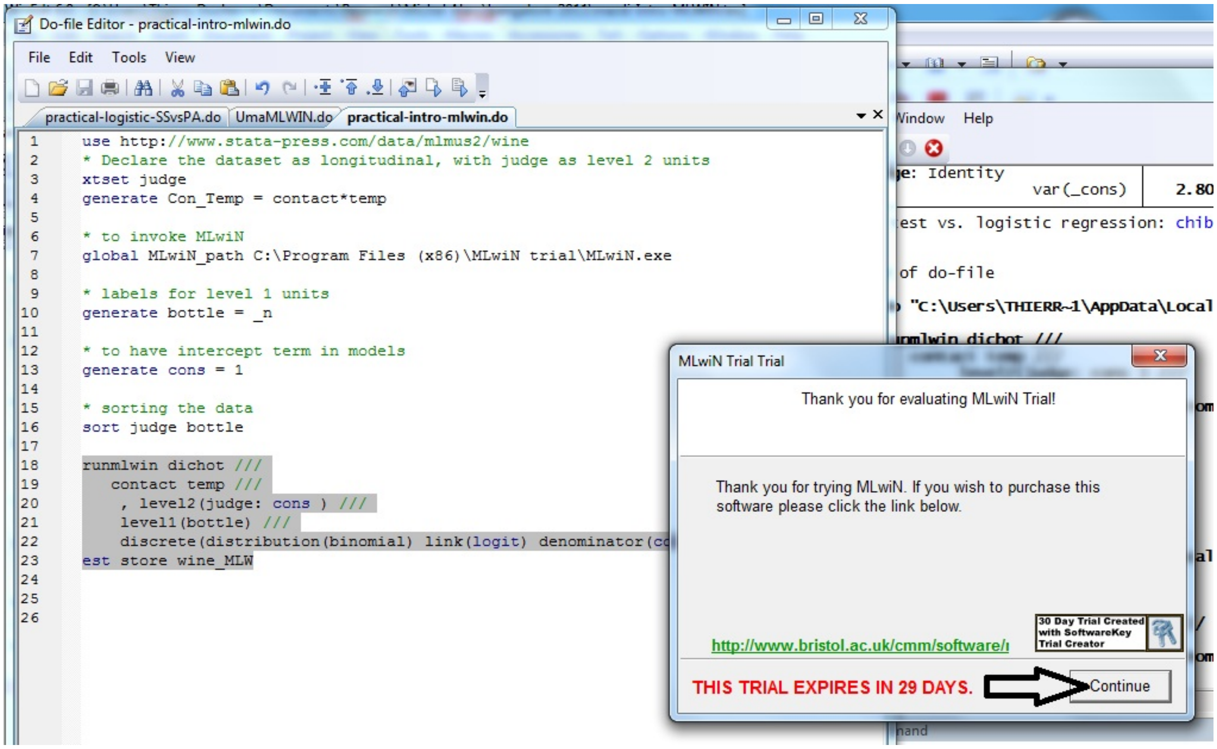
The "*estimation_method*" option tells `MLwiN` which method to use to approximate the likelihood function. They are `mql1, mql2, pql1` or `pql2`. Here "mql" stands for "marginal quasi-likelihood" while "pql" stands for "pseudo quasi-likelihood". The "1" and "2" indicate whether 1st or 2nd order approximations are used. In practice, it is recommended to use `mql1` to get a first fit then use `pql2` with the model obtained with `mql1` as starting point to get accurate results. In a way, this is similar to using `xtmelogit` with the Laplace approximation first, then with a larger number of integration points. To pass initial values to a model, we use the **initsmodel( )** option.

The **or** option tells `MLwiN` that we would like to see the final results in terms of odds ratios instead of in terms of the model coefficients. The **nopause** option tells `MLwiN` to return the results directly to `Stata` without pausing to show some model information to the user. Finally, the **constraints** option is only necessary when dealing with crossed-random effects models for 3-level models.

Back to the wine tasting example

Just for comparison's sake, we can start by fitting a random intercept model with 8 integration points with `xtmelogit`:

```
xtmelogit dichot ///
   contact temp ///
```


Do-file Editor - practical-intro-mlwin.do
File
Edit
Tools
View
practical-logistic-SSvsPA.do
UmaMLWIN.do
practical-intro-mlwin.do
Window
Help
e: Identity
var(_cons)
2.80
est vs. logistic regression: chib
of do-file
"C:\Users\THIERR~1\AppData\Local
nmlwin dichot ///
MLwiN Trial Trial
30 Day Trial Created with SoftwareKey Trial Creator
Continue
nand

```
use http://www.stata-press.com/data/mlmus2/wine
* Declare the dataset as longitudinal, with judge as level 2 units
xtset judge
generate Con_Temp = contact*temp

* to invoke MLwiN
global MLwiN_path C:\Program Files (x86)\MLwiN trial\MLwiN.exe

* labels for level 1 units
generate bottle = _n

* to have intercept term in models
generate cons = 1

* sorting the data
sort judge bottle

runmlwin dichot ///
   contact temp ///
    , level2(judge: cons ) ///
    level1(bottle) ///
    discrete(distribution(binomial) link(logit) denominator(cc
est store wine_MLW
```

Thank you for evaluating MLwiN Trial!

Thank you for trying MLwiN. If you wish to purchase this software please click the link below.

http://www.bristol.ac.uk/cmm/software/

THIS TRIAL EXPIRES IN 29 DAYS.

Figure 3: Invoking MLwiN from Stata

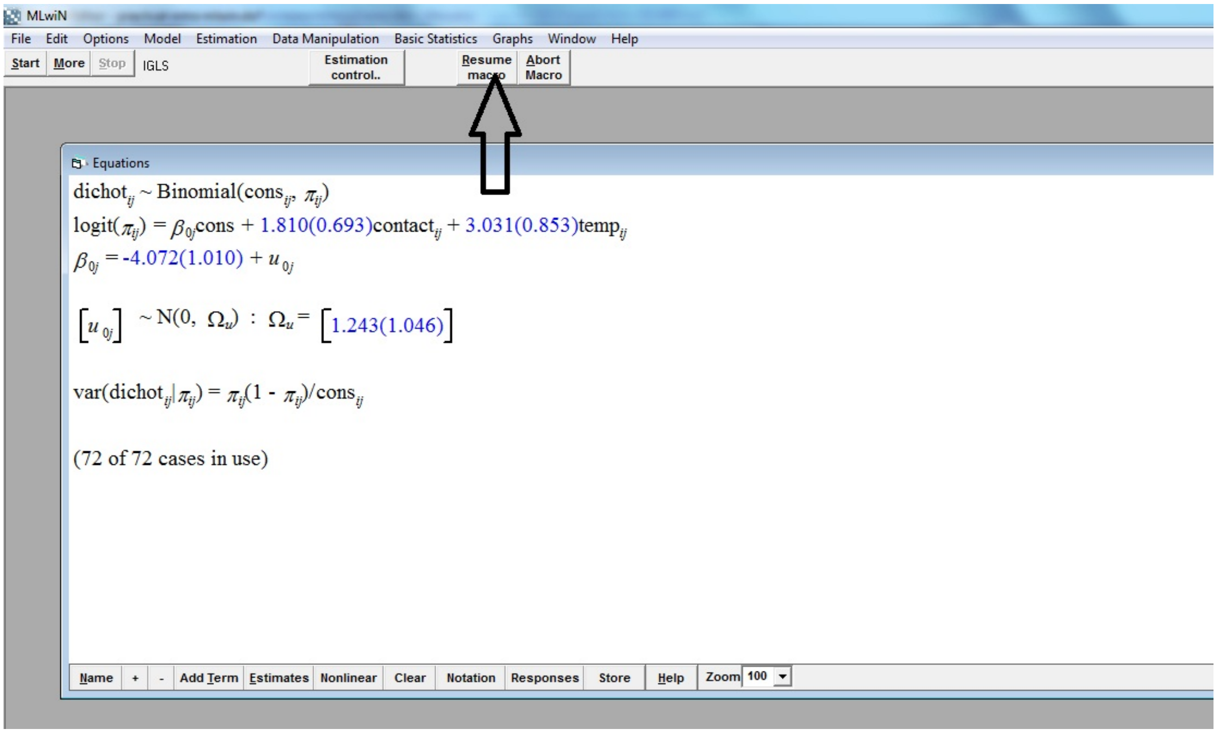
```
   || judge: ///
  , variance intpoints(1)

xtmelogit dichot ///
   contact temp ///
   || judge: ///
, variance intpoints(15)
```

To let MLwiN perform the estimation instead, we can try

```
runmlwin dichot cons contact temp ///
, level2(judge: cons ) ///
level1(bottle) ///
discrete(distribution(binomial) link(logit) denominator(cons) mql1) ///
    nopause
    estimates store wine_m1
```

When this code is run in Stata, the startup window of MLwiN pops up and we must press the Continue button (see Figure 3).

Then we can continue on:



$\text{dichot}_{ij} \sim \text{Binomial}(\text{cons}_{ij}, \pi_{ij})$

$\text{logit}(\pi_{ij}) = \beta_{0j}\text{cons} + 1.810(0.693)\text{contact}_{ij} + 3.031(0.853)\text{temp}_{ij}$

$\beta_{0j} = \text{-}4.072(1.010) + u_{0j}$

$\left[u_{0j}\right] \sim \text{N}(0,\ \Omega_u) \ : \ \Omega_u = \left[1.243(1.046)\right]$

$\text{var}(\text{dichot}_{ij}|\pi_{ij}) = \pi_{ij}(1 - \pi_{ij})/\text{cons}_{ij}$

(72 of 72 cases in use)

Figure 4: Must click on "Resume macro" twice to close MLwiN and return to Stata when the option nopause is not specified in the call to runmlwin

```
runmlwin dichot cons contact temp ///
, level2(judge: cons , residuals(judge) ///
level1(bottle) ///
discrete(distribution(binomial) link(logit) denominator(cons) pql2) ///
    initsmodel(wine_m1)
    estimates store wine_p2
```

Because the **nopause** option is not used, then the user will have to click the "Resume macro" button twice when the MLwiN window opens. See figure 4.

We can see in the Stata output that the results with the Laplace, pql2 and 15 point integration methods yield results that are close, but that the mql1 approximation misses the mark. This is why it is recommended to use it only to obtain starting values for the other estimation methods. Note that to this end, the command `matrix wm1=e(b)` could also have been run after runmlwin to save the parameter estimates in matrix form. The residuals(judge) option created two new variables, judge0 and judge0se that contain the estimates and standard errors of the random intercept estimates. To get the equivalent with xtmelogit, we must use the predict postestimation command with the reffects option.

Workshop on multilevel modeling – Part II
Bangalore, October 17–October 21, 2011

Determinants of HIV prevalence among pregnant women in four southern Indian states

Belkacem Abdous
Belkacem.Abdous@fmed.ulaval.ca

Thierry Duchesne
Thierry.Duchesne@mat.ulaval.ca

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Introduction

Usha Thamattoor and co-authors ran an analysis to identify district level high risk population parameters that influenced the HIV prevalence among pregnant women attending antenatal care clinics (ANC population). To this end, they used data obtained from integrated biological and behavioural assessments (IBBA) carried out between 2004 and 2007 among female sex workers (FSWs), their clients and men who have sex with men (MSM) as well as data from sentinel surveillance in the ANC population. In both cases, data were collected in 24 districts from 4 southern states from 2004 to 2007. Other district level variables were also available.

Description and exploration of the dataset

Because of time considerations, we will only analyze a dataset comprised of a subset of the individual and district variables originally studied by Thamattoor et al. For each of the 46,255 pregnant women observed, we have the following information:

- `state:` the state of the woman's ANC clinic
- `district:` the district of the woman's ANC clinic
- `fsw_hiv:` HIV prevalence in FSW in the district

- hiv_msm: HIV prevalence in MSM in the district
- client_hiv: HIV prevalence among clients of FSW in the district
- per_women_marrying_under_18: proportion of women marrying under 18 in the district
- male_literate_rate_t: total male literacy rate in the district
- year: year at which the woman attended the clinic
- hiv_prev: 1 if woman HIV positive, 0 otherwise
- employment_1 to employment_5: the variable employment_j is 1 if the woman has employment of type j and is 0 otherwise
- age: the woman's age
- migrant_y: 1 if the woman is migrant, 0 otherwise
- Literate_0, Literate_5, Literate_12, Literate_ggreater: indicators that the woman has 0, 5, 12 or more years of education

First, read the dataset into Stata and define it as panel data with districtnum, a variable that contains a different numerical value for each district, as the group variable. Explore the dataset with the Data Editor. Try to get tables of the HIV prevalence as a function of year. First do it for the entire dataset, then for each state, then for each district. Does HIV prevalence seem to vary by district? Getting tables of prevalence as a function of some of the other categorical variables crossed with districts could be interesting as well. At first glance, are there some variables that appear to be associated with HIV prevalence? Might the effect of some variables vary across districts?

Building the model

Initial variable selection

Test whether each covariate listed above is significant at the 25% level using random intercept univariate logistic regression models. Also define an indicator variable age25 that is 1 if the woman is 25 years or older and that is 0 otherwise, and also fit a univariate random intercept logistic regression model with this new variable. For literacy, treat Literate_ggreater as baseline level and use employment_1 as reference level for employment. To be in line with Thamattoor et al, do not consider year in your analyses. Were your intuitions from the exploratory analysis correct? [Note: You can use MLwiN to speed things up.]

Going up towards the final model

Put all the variables that were significant at the 25% level into a multivariate random intercept model and perform a backward selection analysis with a 5% significance level to obtain a final model. Does adding some random coefficients in front of some of the individual-level covariates seem to be required? Draw caterpillar plots of the estimates of the random effects that you have decided to keep in your model. Do all these random effects appear to be needed?

Answering the questions

As discussed in the Introduction, the main purpose of the study of Thamattoor et al. was to find the predictors of the prevalence of HIV in ANC women. Give an answer to this question using your final model. Interpret the effect of all the variables and state whether this effect varies from district to district in your model.

Write the Stata code required to get population-averaged estimates of the effect of the type of employment on the odds of ANC women to test positive for HIV and to test whether this effect is significant. You do not have to run this code, as it is likely to be extremely slow!

Workshop on multilevel modeling – Part II
Bangalore, October 17–October 21, 2011

Tuesday practical

Determinants of time trends in HIV prevalence among pregnant women in Karnataka

Belkacem Abdous
`Belkacem.Abdous@fmed.ulaval.ca`

Thierry Duchesne
`Thierry.Duchesne@mat.ulaval.ca`

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Introduction

Uma Mahajan and co-authors had data on young women (aged less than 25) attending antenatal care clinics from 2003 to 2008 in 27 districts in Karnataka. They also had information about whether the district was an intensive prevention intervention (KHPT) district or a non-intensive intervention (non KPHT) district. The main objective of their study was to see whether HIV prevalence changed at the same rate over time in KHPT and non-KPHT districts, adjusting for other variables that are potentially important confounders.

Description and exploration of the dataset

Because of time considerations, we will not analyze the entire dataset considered. We will instead focus on a subset of the variables. For each of the 88,003 women observed, we have the following information:

- `year:` the year at which the observation was taken
- `age:` the age of the woman
- `locality:` whether the woman lives in a rural or urban area

- `typol`: the type of clinic visited (urban DH or rural FRU)

- `lit`: whether the woman is literate or illiterate

- `hiv`: whether the woman tested positive or negative for HIV

- `dist`: the district

- `khpt`: whether the district is an intensive or non-intensive prevention intervention district

- `hivprv2003`: HIV prevalence in the district at the start of the intervention program

- `sexratio`: sex ratio in the district

- `pfswest`: estimated proportion of FSW in district population

First, read the dataset into `Stata` and define it as panel data with `dist` as the group variable. Explore the dataset with the Data Editor. Try to get tables or to draw a plot of the HIV prevalence as a function of year. First do it for the entire dataset, then separately for the two types of districts, i.e., intensive (`khpt=1`) and non-intensive (`khpt=2`) intervention districts. At first glance, does the rate of change in HIV prevalence over the years seem to be the same in intensive and non-intensive districts? Is the difference between intensive and non-intensive intervention districts linear as a function of time?

Building the model

Defining new variables

An important conclusion that can be reached from the exploratory analyses is that entering time (i.e., `year`) in the models in a linear fashion appears to be inappropriate. `year` should therefore be included in the models as a categorical covariate. To this end, define 5 indicator variables `y1`, `y2`, ..., `y5` with `y1` equal to 1 if `year` is 2004, `y2` equal to 1 when `year` is 2005, ..., `y5` equal to 1 when `year` is 2008 (all indicators will be 0 when `year` is 2003, and therefore 2003 will serve as the reference/baseline year).

To ensure comparability with analyses of this dataset that have already been published, define the following variables:

- `khp`: 1 when `khpt` is 1, 0 when `khpt` is 2

- `local`: 1 when `locality` is 2, 0 when `locality` is 1

- `typ`: 1 when `typol` is 2, 0 when `typol` is 1

- `lite`: 1 when `lit` is 2, 0 when `lit` is 1

Initial variable selection

Because the interaction between time and type of intervention has to be in the final model if we want to estimate it, start by fitting four random intercept logistic models with a single covariate (one model with khp, one model with local, one model with typ and one model with lite). Are there covariates that are significant at the 25% level in these analyses? [Note: You can use MLwiN to speed things up ...]

The same univariate analyses can also be done for the district level variables hivprv03, sexratio and pfswest.

Going up towards the final model

As univariate analyses suggested, the baseline HIV prevalence in the district (hivprv03) is an important covariate. As a matter of fact, the interaction between time and the district baseline HIV prevalence is also an important covariate and any effect of the intervention in time must be measured after having corrected for the time×baseline prevalence interaction. Fit a random intercept model that contains year (as quantified by y1 to y5), the baseline prevalence, the year (as quantified by y1 to y5) by baseline prevalence interaction, khp, the year (as quantified by y1 to y5) by khp interaction, as well as any other variable that proved to be significant in the univariate analyses above.

It would be desirable to allow the effect of time to vary from one district to the other. Using the last model fitted as starting point, add random coefficients for each of y1, ..., y5. Once you have this model, use a backward selection approach to remove the unimportant covariates that are not significant at the 5% level.

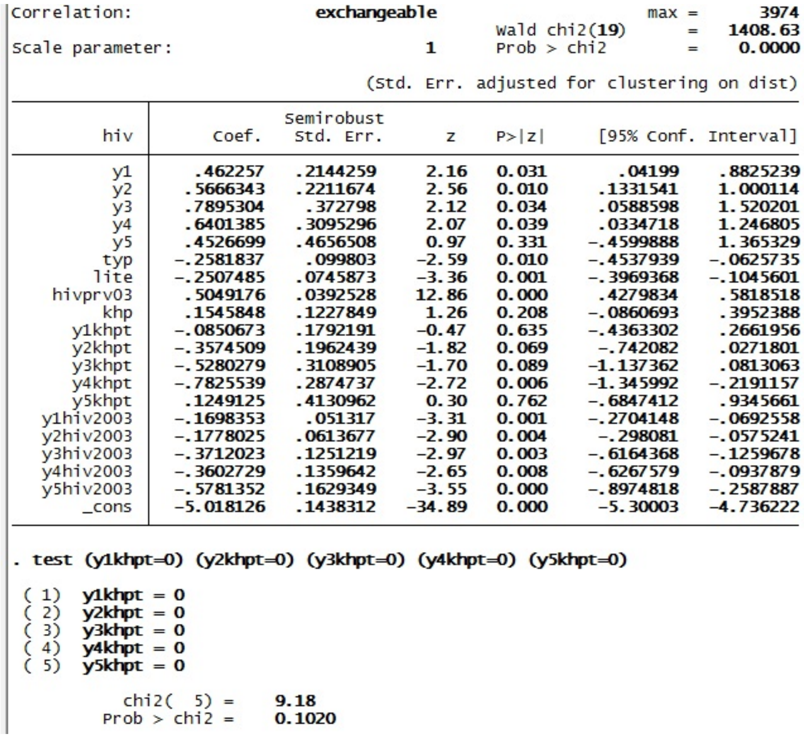
Answering the questions

Effect of intervention intensity over time

The primary objective is to assess whether the rate of change of the HIV prevalence in time is different between intensive and non-intensive intervention districts. This amounts to testing the significance of the year by khp interaction. Is this interaction significant? What can we conclude about the effect of the intensity of the intervention? Redo this test, but with a model that does not include pfswest.

Random effect estimation and prediction

To estimate the number of cases averted in the population, estimates of the HIV prevalence for each district with khp set to 0 and other covariates set to their district-level average value are useful.



```
Correlation:                exchangeable               max =       3974
                                        Wald chi2(19)      =    1408.63
Scale parameter:                       1   Prob > chi2     =     0.0000

                                  (Std. Err. adjusted for clustering on dist)
------------------------------------------------------------------------------
             |             Semirobust
         hiv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          y1 |    .462257   .2144259     2.16   0.031       .04199    .8825239
          y2 |   .5666343   .2211674     2.56   0.010     .1331541    1.000114
          y3 |   .7895304    .372798     2.12   0.034     .0588598    1.520201
          y4 |   .6401385   .3095296     2.07   0.039     .0334718    1.246805
          y5 |   .4526699   .4656508     0.97   0.331    -.4599888    1.365329
         typ |  -.2581837    .099803    -2.59   0.010    -.4537939   -.0625735
        lite |  -.2507485   .0745873    -3.36   0.001    -.3969368   -.1045601
    hivprv03 |   .5049176   .0392528    12.86   0.000     .4279834    .5818518
         khp |   .1545848   .1227849     1.26   0.208    -.0860693    .3952388
      y1khpt |  -.0850673   .1792191    -0.47   0.635    -.4363302    .2661956
      y2khpt |  -.3574509   .1962439    -1.82   0.069     -.742082    .0271801
      y3khpt |  -.5280279   .3108905    -1.70   0.089    -1.137362    .0813063
      y4khpt |  -.7825539   .2874737    -2.72   0.006    -1.345992   -.2191157
      y5khpt |   .1249125   .4130962     0.30   0.762    -.6847412    .9345661
   y1hiv2003 |  -.1698353    .051317    -3.31   0.001    -.2704148   -.0692558
   y2hiv2003 |  -.1778025   .0613677    -2.90   0.004     -.298081   -.0575241
   y3hiv2003 |  -.3712023   .1251219    -2.97   0.003    -.6164368   -.1259678
   y4hiv2003 |  -.3602729   .1359642    -2.65   0.008    -.6267579   -.0937879
   y5hiv2003 |  -.5781352   .1629349    -3.55   0.000    -.8974818   -.2587887
       _cons |  -5.018126   .1438312   -34.89   0.000     -5.30003   -4.736222
------------------------------------------------------------------------------

. test (y1khpt=0) (y2khpt=0) (y3khpt=0) (y4khpt=0) (y5khpt=0)

 ( 1)  y1khpt = 0
 ( 2)  y2khpt = 0
 ( 3)  y3khpt = 0
 ( 4)  y4khpt = 0
 ( 5)  y5khpt = 0

           chi2(  5) =    9.18
         Prob > chi2 =    0.1020
```

Figure 5: Output of `xtgee` for the ANC data

Along with estimates of the district-level random effects, this allows you to compute an estimate of the HIV prevalence in the district, had the intensive intervention not occurred. Compute the district-level average of all covariates, then compute an estimate of the prevalence in each district when `khp=0` in 2004 (year 1).

Also produce caterpillar plots of the estimates of the random coefficients in front of `y1` and `y5`. Comment.

Population-averaged effects

The GEE method can be used to obtain estimates of population-averaged effect. For instance if one were interested in comparing the difference in overall HIV prevalence between literate and illiterate women in the population, one would need a population-averaged estimate of the corresponding coefficient. This can be obtained with the GEE method. Unfortunately, with a dataset of this size, running `xtgee` on `Stata` takes a very long time. If you let it run long enough (a few hours), you get the output shown in Figure 5.

Workshop on multilevel modeling – Part II

Bangalore, October 17-21, 2011

Higher level models with nested random effects

Belkacem Abdous
`Belkacem.Abdous@fmed.ulaval.ca`

Thierry Duchesne
`Thierry.Duchesne@mat.ulaval.ca`

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Exercise 1 : continuous response *see exercise 10.1 MLMUS2 and http://www.biostat.jhsph.edu/ fdominic/teaching/bio656/labs/lab.html*

The math-achievement dataset `achievement.dta` contains information from the U.S. Sustaining Effects Study, which is a longitudinal study of children's academic progress during the six years of elementary school (kindergarten and 1st through 5th grade).

The data have a three-level structure with repeated observations on 1,721 students from 60 public elementary schools in urban areas. Thus, we have repeated observation within child within school.

- Level 1 (repeated observations within a child)
  - `math`: math-test score derived from an item response model
  - `year`: year of study minus 3.5 (1 through 6 minus 3.5, values -2.5, -1.5, -0.5, 0.5, 1.5, 2.5) (a_{1ijk})
  - `grade`: grade level of child at time of observation - sometimes repeats
  - `retained`: indicator for child being retained in grade (1 = retained, 0 = not retained)
- Level 2 (child)
  - `child`: child identifier
  - `female`: dummy variable for gender (1 = female, 0 = male)
  - `black`: dummy variable for being African American (X_{1jk})
  - `hispanic`: dummy variable for being Hispanic (X_{2jk})
- Level 3 (school)

- `school`: school identifier
- `size`: number of students enrolled in the school
- `lowinc`: percentage of students from low income families (W_{1k})
- `mobility`: percentage of students moving during the course of a school year

Goals:

(1) Describe and explore data structure with three levels.

(2) Fit 3-level models with a Normal outcome using xtmixed.

(3) Interpret model parameters (effect coefficients and variance components).

I. Exploratory Data Analysis

1. Use the `xtdes` command to examine the different patterns of observations taken on children in the dataset, (`xtdes` *only accepts integer time variables*)
2. Use the `xtsum` command to give estimates of the mean math score, and its variability among schools and among children

II. Two-level variance component with a random intercept for `school`

1. Use the notations of MLMUS2 (see slides) and write the two-level variance component with a random intercept for `school`.
2. Fit the model using `xtreg`, `xtmixed` and `gllamm`. Compare and interpret the estimates.

III. Two-level variance component with a random intercept for `child`

1. Write the two-level variance component with a random intercept for `child`.
2. Fit this model using `xtreg`, `xtmixed` and `gllamm`. Compare and interpret the estimates.

IV. Three-level variance component, accounting for clustering of children within schools, including a random intercept for `child` and a random intercept for `school`

1. Write and fit this model.
2. Do we need to include a random intercept for `child` ? for `school`?
3. Compute ICC between measurements from same `child` but different `school`
4. Compute ICC between measurements from same `child` and same `school`

V. Incorporating covariates as fixed effects

1. First, add child-level covariates to the previous model. Interpret these results

2. Now, add some school-level covariates as fixed effects. Interpret these results

3. Do we need to include any of the covariates that control for SES?

VI. Add in a random slope on year at the child level

1. Write the corresponding equation. (one big model and three steps model)

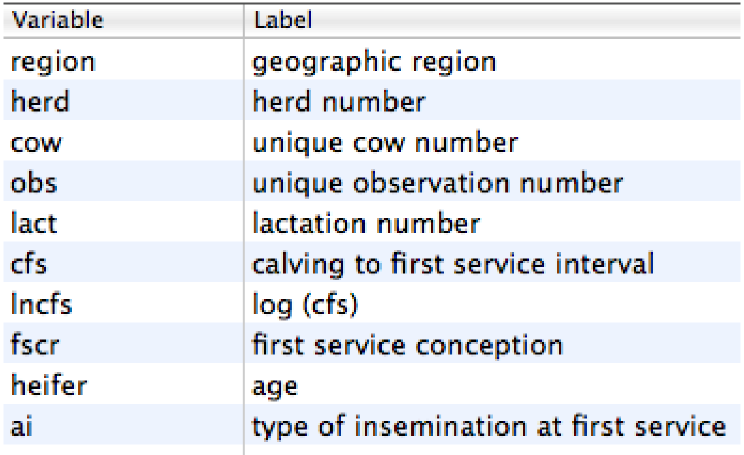
2. Assess the goodness of fit of this model.

3. Second, allow for correlation between random effects at the child level.

4. Which model should we select?

Exercise 2 : continuous response *see exercises 6.6, 10.4 and 10.5 MLMUS2*

Dohoo et al. (2001) and Dohoo et al. (2003) analyzed data on dairy cows from Reunion Island. One outcome considered was the ”risk” of conception at the first insemination attempt (first service) since the previous calving. This outcome was available for several lactations (calving: giving birth to a calf) per cow. The variables in the dataset `dairy.dta` used here are:



Variable	Label
region	geographic region
herd	herd number
cow	unique cow number
obs	unique observation number
lact	lactation number
cfs	calving to first service interval
lncfs	log (cfs)
fscr	first service conception
heifer	age
ai	type of insemination at first service

Four-level regression model: one outcome considered was the time interval between calving and the first service (attempt to inseminate the cow again). This outcome was available for several lactations per cow.

1. Fit a four-level random intercept model with `lncfs` as the response variable and with random intercepts for cows, herds, and geographic regions. Do not include any covariates. Use restricted maximum likelihood (REML) estimation. There are only five geographic regions so that it is arguable that region should be treated as fixed.

2. Obtain the estimated residual intraclass correlations of the latent responses for

   - two observations from the same cow.
   - observations for two different cows from the same herd
   - observations for two different cows from different herds in the same region

3. Fit a three-level model for lactations nested in cows nested in herds, including dummy variables for the five geographic regions using REML and omitting

the constant. Compare the estimates for this model with the estimates using a four-level model.

Exercise 3: binary response *see exercises 6.6, 10.4 and 10.5 MLMUS2*

1. Consider again the dataset `dairy.dat` an fit a two-level random-intercept logistic regression model for the response variable `fscr`, an indicator for conception at the first insemination attempt (first service). Include a random intercept for cow and the covariates `lncfs, ai, heifer`.

2. Obtain estimated odds ratios with 95% confidence intervals for the covariates and interpret them.

3. Obtain the estimated residual intraclass correlation between latent responses for two observations from the same cow. Is there much variability in the cow's fertility?

4. Obtain the estimated median odds ratio for the two randomly chosen cows with the same covariates, comparing the cow with the larger random intercept to the cow with the smaller random intercept.

5. Extend the model above by including a random intercept for herds as well. Use `xtmelogit` or `gllamm` with 5 integration points to speed up estimation. Is there any evidence for unobserved heterogeneity in fertility between herds?

Workshop on multilevel modeling – Part II
Bangalore, October 17-21, 2011
Higher level models with crossed random effects

Belkacem Abdous
Belkacem.Abdous@fmed.ulaval.ca

Thierry Duchesne
Thierry.Duchesne@mat.ulaval.ca

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Exercise 1 : continuous response *see exercise 11.5 MLMUS2*

The dataset `neighborhood.dta` concerns neighborhood effects on educational attainment for young people in one education authority in Scotland who left school between 1984 and 1986.

- Student level
  - `attain`: a measure of educational attainment
  - `p7vrq`: primary 7 verbal reasoning quotient
  - `p7read`: primary 7 reading test scores
  - `dadocc`: father's occupation
  - `dadunemp`: father is unemployed (dummy: 1=unemployed, 0= not unemployed)
  - `daded`: dummy variable for father's schooling being past age of 15
  - `momed`: dummy variable for mother's schooling being past age of 15
  - `male`: student is male (dummy)
- Neighborhood level

- `neighed`: neighborhood identifier
- `deprive`: social-deprivation score

- School level
  - `schid`: school identifier

Goals:

(1) Describe and explore data structure with three levels.

(2) Fit crossed-level models using xtmixed.

(3) Interpret model parameters.

I. Exploratory Data Analysis

1. Use the table command to present the data as a two-way cross-tabulation of neighbourhoods (`neighid`) by schools (`schid`). For presentation purposes, restrict this cross-tabulation to the subset of neighbourhoods in the sample with identifier values in the range 1 to 38, 251 to 263 or 793 to 803.

2. Produce a table of descriptive statistics for the student variables used in the analyses.

II. Variance components models (unconditional models)

Model 1 : Fit a two-level (students within schools) variance components model which simply partitions the variation in attainment into between-school and within-school components.

Model 2 : Fit a two-level (students within neighborhoods) variance components model which simply partitions the variation in attainment into between-neighborhood and within-neighborhood components.

Model 3 : Model 1 accounted for school effects but ignored neighborhoods and Model 2 accounted for neighborhood effects but ignored schools, fit a model that will simultaneously account for both sources of attainment variation (neighborhoods are nested with schools)

III. Cross-classified model Clearly our data are not a pure hierarchy, we have to use cross-classified models

1. Fit a model for student educational attainment without covariates but with random intercepts for neighborhood and school using MLE.

2. Include a random interaction between school and neighborhood, and use a likelihood ratio test to compare this model with the previous model.

3. Include the neighborhood-level covariate `deprive`, and discuss both the estimated coefficient of `deprive` and the changes in the standard deviation estimates for the random effects due to including this covariate.

4. Remove the neighborhood-by-school random interaction and include all student-level covariates. Interpret the obtained results.

5. For the final model, estimate residual intraclass correlations due to being in the same neighborhood but not the same school, the same school but not the same neighborhood, and both in the same school and the same neighborhood.

Workshop on multilevel modeling – Part II
Bangalore, October 17–October 21, 2011

Determinants of HIV prevalence among pregnant women in four southern Indian states: Analysis with three nested levels

Belkacem Abdous
Belkacem.Abdous@fmed.ulaval.ca

Thierry Duchesne
Thierry.Duchesne@mat.ulaval.ca

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Introduction

Let us return to the analysis of the ANC data discussed by Thamattoor et al. The objective was to identify district level high risk population parameters that influenced the HIV prevalence among pregnant women attending antenatal care clinics (ANC population). When we analyzed those data, we did not explicitly take into account the fact that districts are nested within states in the inferences. Since unmodelled state-level decisions may impact the prevalence of HIV in a similar manner in all the districts of a same state, perhaps analyzing the data with a three level model where ANC are nested within districts nested within states could provide inferences based on a more realistic and accurate correlation structure.

Building the model

Data exploration

We have already looked at the data and the variables on Tuesday. Perhaps one additional data exploration task that could be performed is to look at the average HIV prevalence among ANC in each district and see whether districts within a same state tend to be more similar than districts from different states.

Initial variable selection

Repeat the initial variable selection step from the analysis done on Tuesday morning, but this time with a three-level nested random intercept logistic model. In other words, for each explanatory variable in the dataset, fit a univariate random intercept logistic regression model with ANC nested within districts nested within states. Are the results any different from Tuesday's two-level analysis at this stage?

Towards the final model

Starting with all variables that were significant at the 25% level in the univariate analysis, use a backward selection approach to get to a final model where all variables are significant at the 5% level.

A three-level model allows us to put random coefficients in front of level-2 variables. Try adding a state-level random coefficient in front of the district-level variables in your model. Does their effect appear to vary from state to state? Draw a caterpillar plot of the estimates of the state-level effects that you have decided to keep in your model, if there are any such effects.

Workshop on multilevel modeling – Part II
Bangalore, October 17–October 21, 2011

Determinants of time trends in HIV prevalence among pregnant women in Karnataka: Analysis with crossed random effects

Belkacem Abdous
`Belkacem.Abdous@fmed.ulaval.ca`

Thierry Duchesne
`Thierry.Duchesne@mat.ulaval.ca`

Université Laval and
Santé des populations: URESP, Centre de recherche FRSQ du
Centre hospitalier *affilié* universitaire de Québec

Introduction

Let us return to the analysis of the ANC data discussed by Mahajan et al. Back on Tuesday, we wanted to infer about the interaction between time (year as a categorical variable) and intervention (KHPT–intensive or non KHPT–non intensive). In this practical, we will consider a different question. Rather than infer about the effect of year or district, we will focus our attention on the effects of the other explanatory variables and we will treat year and district as a random effects. More precisely, we will build multi-level models with year crossed with district as the random effect structure.

Building the model

Because we have already explored the dataset, we will start with model building right away. We will follow a procedure inspired from the analysis outlined on p. 505 of MLMUS2.

1. Fit a model for HIV prevalence without any covariates but with random intercepts of year and district.

2. Include a random interaction between year and district. Use a likelihood ratio test to compare this model to that of the previous step.

3. Include the district-level covariates one by one (i.e., fit univariate multi-level logistic models). For each of these covariates, discuss whether it is significant and its effect on the estimates of the standard deviation of the random effects.

4. Include all covariates that were significant at the 25% level in a multivariate model. Is the random interaction between year and district still required?

5. Perform a backward selection procedure using a 5% significance level.

6. Give estimates of the residual intraclass correlations for the latent variable for women in a same year but different districts, the same district but different years, and both the same year and same district.