First version: 15/10/02
Latest revision25/05/04

A review of random effects modeling in Stata 8.0

Philippe Mourouga
Ligue nationale Contre Le Cancer, Paris
mourougap@ligue-cancer.net

1. Introduction

1.1 Background

Stata is an integrated statistical package for Windows and other platforms (UNIX, Macintosh, LINUX). This package is more and more used in the statistical community, and its many good points explain why users are coming to it:

- a broad range of statistics
- good data-management capabilities
- a point-and-click interface very useful for beginners
- excellent programmable capacities for more advanced users
- a large set of commands already available to analyze complex data
- an outstanding technical support
- and last but not least a very reasonable price

If we focus on random effects analysis Stata has a set of commands:

- under the generic *xt* prefix, commands were developed for panel data using random-effect estimation
- in addition, specific commands were developed by users such as:
    - *glamm*, Generalized linear latent and mixed models (GLLAMMs)
    - *metareg*, Random-effects meta-analysis
    - *rpoisson*, Poisson regression with a random effect
    - *reoprob*, Random-effects ordered probit

Our review of Stata for random effects modeling will:

- first consider the models available under the *xt* family procedures in release 8.0
- then the *gllamm* program will be presented

1.2 Software and hardware requirements

Release 8.0 of Stata is supported on Microsoft Windows 95, 98 and 2000. It can also run under Windows NT-Version 4.0, UNIX and Macintosh.

1.3 Data input/output functionality

For data input, Open tool in the File dropdown list brings Stata files (.dta) in the variables window. For ASCII data files or other text files specific commands are available (*insheet, infile, infix, edit, input*) by using the command window:

- *insheet* will read tab- or comma- separated data
- *infile* will read unformatted data
- *infix* will read formatted data
- *edit* and *input* will enter the data from the keyboard

For data output, the Save As tool can be used to output files with the file type as Stata file (.dta). Using the *outsheet* command in command window, one can output data either in standard data formats with space, or in comma and Tab delimited format.

1.4 Interface features: data manipulation, commands, menus

In Stata, a statistical *command* consists of a collection of *statements*, and each statement can be followed by many *options*. A statistical task such as model fitting can be conventionally carried out through syntax consisting of commands with options. This review will focus on the use of syntax for fitting a variety of random effect models.

Within the system, one mainly works with four windows named *Review, Variables, Stata Results and Stata Command. Review* allows users to check the last commands computed by Stata, *Variables* gives a description of the variables in use. The *Stata Command window* is used to input data and to write syntax for the statistical tasks to be submitted to run. Once a task has been run, the syntax executed will appeare in the *Stata Results* window. If mistakes in the syntax are detected by the system, they will be reported in this window too. Results of the analysis will be saved in a specific file if this file has been defined by the user.

Before carrying out statistical analysis or fitting models, data should be read and manipulated if needed in the system. In this review, we only use the necessary and basic statements in the step to prepare data for specific random effect models to be fitted, instead of exploring the step in depth.

2. Available tools for random effects modeling in Stata

2.1. In the current Stata version

The *xt family* commands. In Stata 8.0 the *xt* commands are documented in a specific manual named “cross-sectional time-series”. This set of commands provides tools for analyzing cross-sectional time-series datasets. These datasets are on a longitudinal form, two commands are used to specify to Stata that the data are of this form:

- *iis* is used to specify the unit index *i*
- *tis* is used to specify the time index *t*
-

After the data are set with *xtset* the *xt* command may be used. One other option is to set the data using the *i* option of each command.

Random-effects estimator is provided for the following models:

- *xtcloglog*, random-effects and population averaged clog (complementary log-log) models
- *xtintreg*, random-effects interval data regression models
- *xtlogit*, fixed-effects, random-effects and population averaged logit models
- *xtnbreg*, fixed-effects, random-effects and population-averaged negative binomial models
- *xtpoisson*, fixed-effects, random-effects and population-averaged Poisson models
- *xtprobit*, random-effects and population-averaged probit models
- *xtreg*, fixed-,between-, and random-effects, and population-averaged linear models
- *xtregar*, fixed-and random-effects linear models with an AR(1) error distribution
- *xttobit*, random-effects tobit models

Other models are available under the *xt* prefix and specific command (*xtdata, xt, xtdes,xtsum,xtttab)* are used to describe the dataset.

Some random-effects estimators in Stata used Gauss-Hermite quadrature to compute the log-likelihood and its derivative. The *quadchk* command is used to check the sensitivity of the quadrature approximation. The principle is to rerun the estimation using two different numbers of quadrature points and to compare the log-likelihood and the coefficients between the original model and the two refitted models. If the change is less than 0.01% then the choice of the quadrature point is thought not to significantly affect the results. For a change greater than 0.1% the quadrature approach must be questioned for the model under scrutiny.

2.2. Gllamm program

gllamm is a Stata program to fit Generalised Linear Latent and Mixed Models. These models are a class of multilevel latent variable models for multivariate responses of mixed type. The *gllamm* command can be downloaded at www.gllamm.org. A full description of the command with the manual and datasets are available on the website. The following models are described in the manual:

- multilevel generalized linear models
- multilevel factor models
- discrete random effects
- mixed response models
- continuous time to event survival data
- ordinal responses
- nominal or polytomous responses

2.3. Specific programs (STB)

Specific programs are available on the Stata net resources and were described in the Stata Technical Bulletin (STB) or in the new Stata Journal (which now replaces STB). The following list is a selection of these programs:

- STB-46 sg98: Poisson regression with a random effect - *rpoisson*

This command fits the negative binomial regression model for over dispersed count data. We will not describe this command in the present review due to the lack of appropriate data

- STB-59 sg158: Random-effects ordered probit - *reoprob, ghquadm*

This estimates a random-effects ordered probit model. This command will not be described in the present review.

- STB-37 sbe23: Meta-analysis regression – *metareg*

Estimate a random-effect meta-analysis. This estimates the extent to which one or more covariates, with values defined for each study in the analysis, explain heterogeneity in the study-specific effects.

By using the Search submenu in the Help menu one can easily check the available programs:

- from the Help menu go to the Search sub-menu
- tick for net resources
- type in the box: multilevel or random
- click on the program you want to install

2.4. Computer specifications

Analysis were performed on a Dell Inspiron 5150, Mobile Intel® Pentium®, 4 CPU 3.06 GHz, 1.59 GHz, 512 Mb of RAM.

3. Model specifications – standard xt commands

In this section, we explore some basic multilevel models that can be fitted by some standard procedures. The model includes normal models, logistic model for binary data, Poisson model for count data and repeated measures models for data with two -or three- level hierarchy. As the capacity of model fitting and efficiency of estimation procedures are concerned, we shall explore the syntax in specifying the models, information on model estimates and time for a model to converge.

3.1. Do file

The models in this review can be retrieved by running the analysis-review do file (see Annex)

3.2. Two-level Normal models

The data set to be used in the example appeared in the user's guide to *MLwiN* (exam.txt ascii file). It consists of 4,059 students (level 1 units) nested within 65 schools (level 2 units) with examination score as outcome (EXAM) at age 16 and the pre-school London Reading Test score (STANDLRT, x_1) taken at age 11. Both the outcome and the reading scores were standardised with zero mean and unit variance. The other student level variable is gender of the student (GENDER, x_2: code 1 for girl and 2 for boy). School level variables are school gender (SCHGEND, coded 1 for mixed schools, 2 for boy and 3 for girls schools respectively), school average of intake score, etc. For illustration purpose, we only consider the first three covariates STANDLRT, GENDER and SCHGEND in models. Dummy variables for boys' and girls' schools contrasting with mixed school are derived of the school gender.

Two models are fitted with one nested in the other, using the *xtreg* command:

- Model **A** is a variance component model with all three covariates fitted as fixed effects. Only the intercept is allowed to have random effects u_{0j} among schools. The model can be written as

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + \beta_4 x_{4j} + e_{0ij}$$

$$\beta_{0j} = \beta_0 + u_{0j},\ u_{0j} \sim N(0,\sigma_u^2),\ e_{0ij} \sim N(0,\sigma_e^2)$$

  The β s here are fixed effects, and var(u_{0j}) and var(e_{0ij}) are two variance parameters to be estimated.

- Model **B** has extended Model **A** by including an extra interaction term between the two student level variables, London Reading Score and student gender.

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + \beta_4 x_{4j} + \beta_5 (x_{1ij} \times x_{2ij}) + e_{0ij}$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

Two commands are available to fit these models:

- *xtreg*, estimates cross-sectional time-series linear regression model
- *xtgee*, estimates cross-sectional time-series analysis generalized linear model using GEE estimation of GLM as defined by Liang and Zeger (1986)

xtreg will estimate a simple variance component model and is not able to estimate models where both the slope and intercept differ between level 2 units. With the *re option* the GLS random-effects estimator will be used, the maximum-likelihood estimator being used with the *mle option*. The commands fitting Models A and B are:

```
use exam,clear
recode schgend 3=0
xi:xtreg normexam standlrt gender i.schgend,mle i(school)
dis -2*e(ll)
dis e(sigma_u)^2
dis e(sigma_e)^2

xi:xtreg normexam i.gender*standlrt i.schgend,mle i(school)
dis -2*e(ll)
dis e(sigma_u)^2
dis e(sigma_e)^2
```

xtgee with the *gaussian family, the identity link and the exchangeable correlation structure* will give similar results to *xtreg* but will not provide an estimate of the level 2 variance. Following are the format of the commands for fitting the two models.

```
xi:xtgee normexam i.gender standlrt i.schgend,fam(gauss) link(ident)
corr(exch) i(school)

xi:xtgee normexam i.gender*standlrt i.schgend,fam(gauss) link(ident)
corr(exch) i(school)
```

Results of Models A and B are presented in Tables 1 and 2 respectively.

Table 1 Estimates for Model A using the exam data

	Stata xtreg	Stata xtgee
Fixed part		
Intercept	-0.0103 (0.0763)	
Gender	0.1684 (0.0341)	0.1674 (0.0343)
Standlrt	0.5597 (0.0124)	0.5609 (0.0125)
Schgend1	-0.1583 (0.0873)	-0.1581 (0.0786)
Schgend2	0.0210 (0.1233)	0.0209 (0.1117)
Random Part		
Var(u_{0j})	0.0812	-
Var(e_{oij})	0.5621	-
Log likelihood	-4657.46	-
Computing time	<60s[f]	<60s[f]

[f] Dell workstation, Pentium III xeon, CPU:, 523MB RAM

Table 2 Estimates for Model B using the exam data

	Stata xtreg	Stata xtgee
Fixed part		
Intercept	-0.0103 (0.0764)	-0.0090 (0.0701)
Gender	0.1683 (0.0341)	0.1673 (0.0343)
Standlrt	0.5629 (0.0184)	0.5640 (0.0185)
Schgend1	-0.1581 (0.0873)	-0.1579 (0.0787)
Schgend2	0.0212 (0.1233)	0.0210 (0.1118)
Interaction Gender*standlrt	-0.0058 (0.0246)	-0.0057 (0.0248)
Random Part		
Var(u_{0j})	0.0812	-
Var(e_{oij})	0.5620	-
log likelihood	-4657.43	-
Computing time	<60s[f]	<60s[f]

[f] Dell workstation, Pentium III xeon, CPU , 523MB RAM

3.3. Two-level models for binary/binomial data

The commands *xtlogit* and *xtprobit* uses Gauss-Hermitte quadrature approximation. The quadrature formula requires that the integrated function be well-approximated by a polynomial and a specific command *quadchk* has been developed to investigate the applicability of the numeric technique.

xtlogit allows the user to fit the data with a logit link. *xtprobit* allows the user to fit the data with a probit link. For both commands, only variance component models are available.

For both commands the option *re* requests the random-effect estimator and is the default. An alternative will be to fit an equal-correlation logit (or probit) model with the *pa* option. To fit an equal-correlation model with the *xtgee* command the user must use the *corr(exchangeable)* option the results will be similar to those obtained with the *xtlogit,pa* command. [pa refers to apopulation average (GEE) model]

To evaluate overall goodness of fitting, the default output includes the log-likelihood and an overall Wald test. For each parameter estimate in the model a t-test value and a confidence interval based on the tail probability are given.

The panel-level variance component or level 2 variance component is labelled as `lnsig2u` in the output. The standard deviation is given under the label `sigma_u` with the proportion of the total variance contributed by the panel-level variance component labeled as `rho`. This measure is questionable in the context of GLM as the level 1 variance is fixed to be 1 and not estimated (see Goldstein et al 2002 for a discussion)

To illustrate how the procedure works, we fitted models to binary data from the 1989 Bangladesh Fertility Survey (bang.txt ascii file). The data are a sub-sample of 1934 women grouped into 60 districts. The outcome variable is use of contraception at the survey (y_{ij}) which equals 1 for using contraception and 0 otherwise. Three covariates are considered: age at survey centered at the sample mean (x_{1ij}); type of region of residence (x_{2ij}) which equals 1 for urban and 0 for rural; and number of living children (0=none, 1=one, 2=two, 3=three or more), represented by three dummy variables for the last three categories (x_{3ij}, x_{4ij} and x_{5ij} respectively). Model A is a variance component model with a logit link: in Model B the link is a probit.

Model A: $\log it(p_{ij(y=1)}) = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij}$

$$\beta_{0j} = \beta_0 + u_{0j}$$

Model B: $probit(p_{ij(y=1)}) = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + \beta_4 x_{4j} + \beta_5 x_{5ij}$

$$\beta_{0j} = \beta_0 + u_{0j}$$

The syntax for Models A and B are listed in the Annex and model estimates are in Table 3.

Table 3 Estimates for Model A and B using the Bangladesh fertility data

	Stata xtlogit	Stata xtprobit
Fixed part		
Intercept	-1.6966 (0.1483)	-1.0322 (0.0874)
Age	-0.0265 (0.0079)	-0.0163 (0.0048)
Urban	0.7282 (0.1177)	0.4464 (0.0718)
Nblive2	1.1098 (0.1581)	0.6704 (0.0948)
Nblive3	1.3764 (0.1748)	0.8346 (0.1049)
Nbliv4	1.3462 (0.1797)	0.8149 (0.1074)
Random Part		
Var(u_{0j})	0.2253	0.0834
Log likelihood	-1206.58	-1206.29
Computing time	<60s[f]	<60s[f]

[f] Dell workstation, Pentium III xeon, CPU, 523MB RAM

Although there are small differences, the results are very similar between *MLwiN* (PQL2) and Stata.

3.4. Two-level model for count data

The commands *xtpois*, *xtnbreg* and the user-programmed command *rpoisson* are available to model count data.

For *xtpois* and *xtnbreg* as for other *xt commands*, the *re* option is the default and requests the random-effect estimator. With the *pa* option the user will get the population-averaged model and the command will be equivalent to: *xtgee…,family(poisson) link(log) corr(exchangeable) exposure(time).*

The maximum-likelihood random-effects model as for the *xtlogit, xtprobit* commands will use an *M*-point Gauss-Hermite quadrature. The *quadchk* command could be used to investigate the applicability of the numeric technique.

The *rpoisson* procedure may be used for fitting the following models:

- negative binomial regression model for over dispersed count data
- Poisson regression model in longitudinal studies of event counts within subject
- exponential survival time models with random frailty

More details are available in the Stata technical bulletin STB-46 sg98.

We try this procedure on the Malignant Melanoma Mortality data (mmmec.txt ascii file) from 354 counties within 78 regions within 9 European countries. The data consist of observed deaths and expected deaths due to malignant melanoma, which produce the standard mortality rate (SMR) in the form of Sum(*observed death) /Sum (expected* d*eath).* One important environmental variable is the county level UV radiation exposure that might be associated with the mortality rate.

We consider the following simple two-level model

Model A: $$\lambda_{ij} = \frac{y_{ij}}{E_{ij}} = \exp(\beta_0 + \beta_1 x_{1ij} + u_{0j})$$

E is the expected death number; x_1 is the measure of UVB dose reaching the earth's surface. A three-level structure exists in the data with only 9 level 3 units. In Model A, the level 3 structure is simply ignored. The logarithm *E* is treated as an offset term in the model.

By default in *xtpois* the distribution of the random effects u_{0j} is assumed gamma. The model estimates using the syntax listed in the Annex for this model is presented in Table 4. To assume Normal or Gaussian distribution for the random effects in the Stata term, the following syntax is used with model estimates in the following table.

xtpois obsdth uvb, off(inexp) i(region) normal re

Only $\sqrt{\sigma_u^2}$ and its s.e. are reported assuming a Gaussian distribution.

Table 4 Estimates for Model A using the melanoma data

	Stata xtpois (gamma)	Stata xtpois (Gaussian)
Fixed part		
Intercept	-0.055 (0.049)	-0.138 (0.017)
uvb	-0.039 (0.010)	-0.056 (0.004)
Random Part		
Var(u_{0j})	0.164 (0.029)	0.102
Log likelihood		
	-1126.45	-1129.13
Computing time		
	<60s[f]	<60s[f]

[f] Dell workstation, Pentium III xeon, CPU:, 523MB RAM

3.5. Repeated measures data

In addition to general random intercepts models for repeated measures data, time series models for the structure of random effects can be fitted in Stata using three commands:

- *xtregar,* to fit random-effect linear models with an AR(1) disturbance
- *arch,* provide access to the Autoregressive Conditional Heteroskedasticity (ARCH) family of estimators
- *arima,* to fit autoregressive integrated moving average models

The *xtregar* command shares the characteristic of all the *xt* commands. The *re* option is the default and it requests the Baltagi-Wu GLS estimator of the random-effects model. This estimator can accommodate both unbalanced or unequally spaced datasets. Baltagi and Wu derive a transformation of the data that removes the AR(1) component, then a simple OLS is performed on the transformed data.

The height (cm) data from 26 boys between 11~13 years old in Oxford over 9 occasions from each boy with approximately 0.25 year apart are used here to illustrate the *xtregar* command.

Model A fits polynomial growth curve with random intercepts only at level 2. The variable t is the age centred at mean and the dependent variable is height (ht).

Model A:

$$ht_{ij} = \beta_{0j} + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_4 t_{ij}^4 + e_{ij}$$

$$\beta_{0ij} = \beta_0 + u_{0ij}\,,$$

$$u_{0ij} \sim N(0, \sigma_u^2),\ e_{ij} \sim N(0, \sigma_e^2)$$

The results in Table 5 are obtained using the syntax *xtregar* listed in the Annex.

Table 5 Estimates for Model A using the Oxford data

	Stata xtregar
Fixed part	
Intercept	148.9381 (1.5969)
Age	6.1814 (0.3241)
Age^2	1.3541 (0.5396)
Age^3	0.4188 (0.2910)
Age^4	-0.5921 (0.4407)
Random Part	
Var(u_{0j})	63.2108
-2 log likelihood	
Computing time	<60s[f]

[f] Dell workstation, Pentium III xeon, CPU, 523MB RAM

Further modeling the variance-covariance structure, the *arch* command estimates models of autoregressive conditional heteroskedasticity using conditional maximum likelihood. The basic model is of the following form:

$$y_t = x_t \beta + \varepsilon_t$$
$$Var\left(\varepsilon_t\right) = \sigma_t^2 = \gamma_0 + A(\sigma,\varepsilon) + B(\sigma,\varepsilon)^2$$

The y_t equation may optionally include ARCH-in-mean and/or ARMA terms. If no options are specified, $A() = B() = 0$ the model collapses to linear regression. For more details, please read the **arch** entry of the Help system. Specifications for the variance are also available with the *het option*. If specified the variable in the *het option* will have a variance specified as multiplicative heteroskedasticity.

The *arima* command estimates a model where the disturbances are allowed to follow a linear autoregressive moving-average (ARMA) specification. When independent variables are not specified this model reduce to autoregressive integrated moving average (ARIMA) model in the dependent variable. See the **Methods and Formulas** section of the **arima** command for more details.

4. Gllamm analysis

As illustrated in the previous sections the standard tools for random effects models in Stata is limited to two-level structure and random interects only models. The add-on program *gllamm* fits three level models with random slopes of Normal and discrete data, using adaptive quadrature and similar syntax to Stata's estimation command. A complete description of Generalised Linear Latent and Mixed Models is given in chapter 1 of the manual with a very clear description of the implementation and the syntax of the program in chapter 2. More information on the development and application as well as citations of Gllamm can be found in *http://www.gllamm.org/.*

Two data sets will be used in this analysis: the malignant melanoma mortality data set and the exam data set. For both examples we will give the number of quadrature points used for the analysis. The results are given in Table 6 and Table 7. It is noticed that both datasets are not a fortunate choice for Gllamm when compared to the PQL procedure used by, for example, SAS or MLwiN as the quadrature procedure tend to perform badly in the case for datasets with large counts (melanoma data) or introduces unnecessary approximation error and computation for purely continuous response (exam data). Nevertheless, the results are very similar to those obtain with MLwiN.

Table 6 Estimates for two and three-level models for melanoma data using 8-point adaptive quadrature

	Two-level model	Three-level model
Fixed part		
Intercept	-0.0345 (0.0143)	-0.0617 (0.0258)
uvb	-0.0389 (0.0034)	-0.0378 (0.0063)
Random Part		
Var(v_{0j})		0.1680 (0.0181)
Var(u_{0jk})	0.1556 (0.0108)	0.0160 (0.0077)
Log likelihood	-1143.24	-1124.70

Computing time	30 seconds[f]	90 seconds[f]

[f] Dell workstation, Pentium III xeon, CPU, 523MB RAM

Table 7: Estimates for random coefficient model for exam data

	8-point adaptive quadrature	10-point adaptive quadrature
Fixed part		
Intercept	-0.0617 (0.0447)	-0.0548 (0.0459)
Gender	0.1808 (0.0331)	0.1763 (0.0330)
Standlrt	0.5542 (0.0180)	0.5542 (0.0183)
Schgend1	-0.1041 (0.0424)	-0.1033 (0.0458)
Schgend2	0.0047 (0.0594)	0.0025 (0.0627)
Random Part		
Var(u_{0j})	0.0776 (0.0119)	0.0853 (0.0145)
Cov(u_{0j},u_{1j})	0.0207 (0.0056)	0.0229 (0.0071)
Var(u_{1j})	0.01527 (0.0045)	0.0159 (0.0049)
Var(e_{oij})	0.5503 (0.01239)	0.5499 (0.0124)
Log likelihood	-4632.88	-4.632.99
Computing time	29 minutes[f]	45 minutes[f]

[f] Dell workstation, Pentium III xeon, CPU, 523MB RAM

5. Documentation and user support

Stata provides users with an excellent set of manuals describing the commands with technical annex and examples as well as website. The Help system is also very useful but less detailed than in the manuals.

Within the package under the Help menu, submenus allow the user:

- to access Stata web site
- to download user-written programs
- to get official updates
- to search within the package or on the net specific statistical topics
- to look at Stata commands

6. Conclusion

Stata corporation has developed a set of commands for longitudinal data (time-series data) under the *xt* prefix. These commands when compared to other software give very similar results (see table of comparative timings and estimates). As the commands were not developed to handle hierarchical data, only variance component models were available in the Stata 7.0 core package.

One user-defined command, *gllamm*, extends Stata capacity to fit hierarchical models. This procedure is being continuously improved and can be downloaded from the gllamm website (*http://www.gllamm.org/)*. It is important to notice that the number of quadrature points is important and must be checked carefully. The choice of datasets is important in the comparison as some data sets perform better when using quadrature approximation and others worse. The Gllamm manual does give some comparisons with *MLinM* through the choice of other datasets.

References

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* **2**, 1-21.

Goldstein, H., Browne, W. and Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics* **1**: 223-232.

Skrondal, A. and Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika* **68** (2), 267-287.

Rabe-Hesketh, S. and Everitt, B. (2004). *A Handbook of Statistical Analyses Using Stata, 3rd edition*. Chapman & Hall/CRC.

Annex: Stata do-file to run the analysis

Please copy the do-file in the repertory “`c:\data\mlwin\datasets`” with the datasets before running the analysis.

```
/* preambule */
set more 1
cd c:\data\mlwin\datasets
set matsize 300
capture log close
log using analysis_review,replace text

/* first part of the review: xt command and stata command
we use the i option of the command to set level 2 identifier */
/* normal */
use exam,clear
recode schgend 3=0
xi:xtreg normexam standlrt gender i.schgend,mle i(school)
dis -2*e(ll)
dis e(sigma_u)^2
dis e(sigma_e)^2
xi:xtgee normexam i.gender standlrt i.schgend,fam(gauss) link(ident) corr(exch)
i(school)
xi:xtreg normexam i.gender*standlrt i.schgend,mle i(school)
dis -2*e(ll)
dis e(sigma_u)^2
dis e(sigma_e)^2
xi:xtgee normexam i.gender*standlrt i.schgend,fam(gauss) link(ident) corr(exch)
i(school)

/* binary */
use bang,clear
xi:xtlogit use age urban i.nbliv, i(district)
dis e(sigma_u)^2
xi:xtprobit use age urban i.nbliv, i(district)
dis e(sigma_u)^2

/* poisson */
use mmmec,clear
gen lnexp=ln(expdth)
xi:xtpoisson obsdth uvb,off(lnexp) i(region)

/* repeated measures */
use oxboys,clear
tsset id occasion
gen age2=age^2
gen age3=age^3
gen age4=age^4
xtregar height age age2 age3 age4
dis e(sigma_u)^2
dis e(sigma_e)^2
```