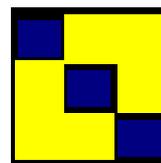


MLwiN Macros for advanced Multilevel modelling



by

Min Yang

Jon Rasbash

Harvey Goldstein

Maria Barbosa

**Multilevel Models Project
Institute of Education
University of London**

Version 2.0a: June 2001

***MLwiN* Macros for Advanced Multilevel Modelling**

Min Yang, Jon Rasbash, Harvey Goldstein, and Maria Barbosa

© 1999 M. Yang, J. Rasbash, H. Goldstein, M. Barbosa

All rights reserved

No part of this document may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, for any purpose other than the owner's personal use without the prior written permission of one of the copyright holders.

ISBN: 085473 549 6

Printed in the United Kingdom

MLwiN Macros for advanced Multilevel modelling

by

Min Yang

Jon Rasbash

Harvey Goldstein

Maria Barbosa

**Multilevel Models Project
Institute of Education
University of London**

Version 2.0a: June 2001

Web site: <http://multilevel.ioe.ac.uk>

email: m.yang@ioe.ac.uk

Acknowledgements

This work was partly funded by the Economic and Social Research Council (UK) under the programme for the Analysis of Large and Complex Datasets.

Maria Barbosa was sponsored by CNPq-Brazil.

Table of Contents

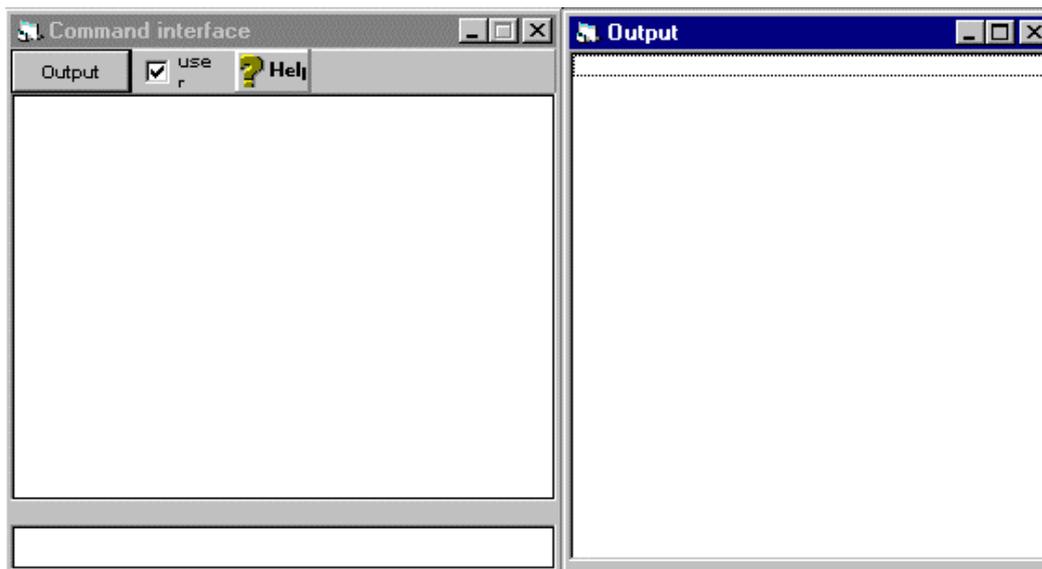
<i>Introduction</i>	4
Setting up the macros	5
Constraints	5
Problems	6
<i>Chapter 1. Multi-category response models</i>	5
Unordered categories	5
Ordered categories	6
Mixed continuous and discrete response data	6
Estimation	7
Running the macros	7
Examples	7
An unordered categorical response model	7
An ordered categorical response model	10
Mixed model I: continuous and unordered categorical responses	12
Mixed model II: continuous and ordered categorical responses	15
Some tips	18
<i>Chapter 2. Survival and Event duration models</i>	19
Log duration (accelerated failure time) models	19
Semiparametric Cox Models	19
Estimation	20
Running the macros	21
Examples	21
A log duration model	21
A semiparametric model	25
<i>Chapter 3. Time series and non-linear variance models</i>	30
Time series models	30
Non-linear level-1 variance functions	31
Estimation	31
Running the macros	31
Examples	32
Continuous time series models	32
The extended time series model	36
A heterogeneous level-1 variance model	38

Some tips	40
<i>Chapter 5. Time Series with discrete responses</i>	<i>41</i>
Description of the model	41
Estimation	42
Running the macros	42
Example	43
Time series with discrete response model	43
<i>References</i>	<i>48</i>

Introduction

This volume describes the use of *MLwiN* macros to fit multilevel models for a number of non-Normal non-linear models using commands issued through the **Command interface** window. It should be read in conjunction with the standard *MLwiN* user's guide (Goldstein *et al.*, 1998)

In the **command interface**, click on the **user** button and then the **output** button to obtain the following pair of windows (after resizing and moving)



The commands are typed into the lower box in the left hand window and the response to these commands appear in the right hand output window.

In version 1.0 of *MLwiN*, graphical user interface (GUI) tools have not been fully implemented for these models. The complete syntax for the commands is given in the *MLwiN* help system.

The reader is referred to Goldstein (1995) for the statistical background and references to specific chapters will be given. References will also be made to other publications, including working papers. *MLwiN* has a web site, <http://multilevel.ioe.ac.uk>, which contains upgrades and further documentation for these macros as well as general information about *MLwiN*.

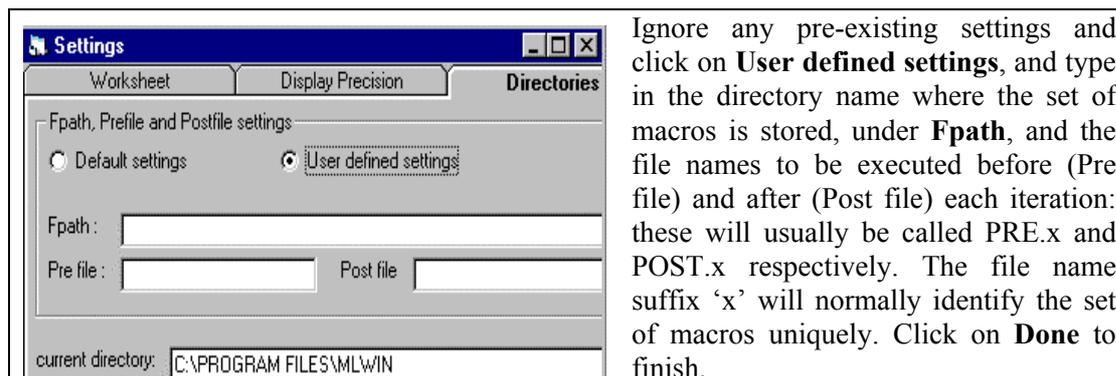
Each chapter of this volume contains a short description of the model and how to use it. This is followed by a worked example. The data for these examples are stored in worksheets distributed with the macros.

Generalised linear models with binomial, Poisson and negative binomial errors can be run directly from *MLwiN*'s graphical interface.

Setting up the macros

Four sets of macro are currently available in MLwiN V1.1, namely MULTICAT, SURVIVAL, TS and DTS for modelling categorical response, survival data, time series of Normal and Binary responses respectively.

Each set of macros should be stored in a subdirectory and the name of this subdirectory is communicated to *MLwiN* via the **Fpath** box in the **Settings** screen of the **Directories** on the **Options** menu.



In the examples, particular subdirectories have been assigned, but these can be changed by the user. *It is recommended that you do not run MLwiN from any of the macro subdirectories, i.e. do not set **current directory** as one of the macro subdirectories.* The macros described in this manual use file suffixes 'mc', 'su', 'ts' and 'dts' to refer to the multi-category, survival, time series (for Normal response) and time series (for binary response) macros respectively.

For each set of macros there is a file called OPTIONS.x. In the **Command interface** window, typing

OBEY OPTIONS.x

calls up in the output window (click on the **Output** button to display this) a 'setting screen' which will guide the user through the various choices available.

Once a model has been specified the estimation is controlled using the START, NEXT, RESiduals or LIKELihood commands. Note, however, that the LIKELihood command will only produce (possibly very) approximate values.

Constraints

Because the macros use specific locations on the worksheet you should be careful not to overwrite these unless the macros specifically require them to be set up by the user. They are:

Columns: C100-C101 C110-C112 C160-C190 C201-C210

Boxes: B1-B20 B200-B202 B300-B301

Groups: G1-G5 G9-G13 G17-G20

If you need extra columns or groups note that you may specify any number of these for your worksheet using the INIT command from the **Command interface** window or using the **worksheet** dialogue screen from the options menu.

The macros should allow you to interrupt them by pressing the 'escape' key but in some circumstances this may cause problems. It is often a good idea to run the first couple of iterations with 'BATCh' off to make sure the model set up is correct.

Problems

If you encounter any problems you cannot correct please contact us by sending email to m.yang@ioe.ac.uk.

Chapter 1. Multi-category response models

In this chapter we consider generalised linear models where the response is a vector of proportions. They are described in Chapter 7 of Goldstein (1995).

Unordered categories

Consider a simple 2-level model with just an intercept plus one explanatory variable. At level-2 we have a random intercept u_0 . The response is a vector of t proportions, where one is chosen as the base category.

$$\pi_{ij}^{(s)} = \exp(\beta_0 + \beta_1 x_{1ij}^{(s)} + u_{0j}^{(s)}) \left[1 + \sum_{h=1}^{t-1} \exp(\beta_0 + \beta_1 x_{1ij}^{(h)} + u_{0j}^{(h)}) \right]^{-1} \quad (1.1)$$

which assumes a multivariate logit link function, with $s = 1, \dots, t-1$.

Assuming a multinomial distribution for the observed proportions the covariance matrix for the set of observed proportions is given by

$$n_{ij}^{-1} \begin{pmatrix} \pi_{ij}^{(1)}(1 - \pi_{ij}^{(1)}) & & & & \\ -\pi_{ij}^{(1)}\pi_{ij}^{(2)} & \cdot & & & \\ \cdot & & \cdot & & \\ \cdot & & & \cdot & \\ -\pi_{ij}^{(1)}\pi_{ij}^{(t-1)} & \cdot & \cdot & \cdot & \pi_{ij}^{(t-1)}(1 - \pi_{ij}^{(t-1)}) \end{pmatrix} \quad (1.2)$$

By defining the following dummy variables which are functions of the predicted values we can fit this in *MLwiN* as a multivariate model

$$z_{1ij} = \sqrt{\pi_{ij} / n_{ij}}, \quad z_{2ij} = \pi_{ij} / \sqrt{2n_{ij}}$$

$$z_{3ij} = -\pi_{ij} / \sqrt{2n_{ij}}, \quad \pi_{ij} = \{\pi_{ij}^{(s)}\}$$

We specify Z_1 to have a random coefficient at level-1 with variance constrained to 1.0 and Z_2, Z_3 to have random coefficients at level-2 constraining their variances to zero and their covariance to 1.0. This produces the components of the structure (1.2) and extra-multinomial variation can be achieved by allowing the variance and covariance to be different from 1.0 but constraining them to be equal.

Ordered categories

The response now is a set of cumulative proportions with cumulative probabilities given by

$$\gamma_{ij}^{(s)} = \{1 + \exp[-(\beta_0^{(s)} + \beta_1 x_{1ij} + u_{0j}^{(s)})]\}^{-1} \quad (1.3)$$

$$\gamma_{ij}^{(s)} = \{1 - \exp[-\exp(\beta_0^{(s)} + \beta_1 x_{1ij} + u_{0j}^{(s)})]\} \quad (1.4)$$

for the logit and log-log links respectively. The coefficient of β_1 as well as those associated with other covariates will generally be estimated with a negative sign. This implies that increasing values of these covariates are associated with increasing probabilities as S increases. For two categories r, s the covariance of the observed cumulative proportions is

$$\gamma_{ij}^{(s)}(1 - \gamma_{ij}^{(r)}) / n_{ij} \quad s \leq r$$

As with the unordered models we can specify extra-multinomial variation. Note that in (1.3) ~ (1.4) we assume only that the intercept varies across the higher level units.

Mixed continuous and discrete response data

For a mixed response model, as in all the multivariate models, level-1 is used to define both the categorical response vector and the continuous response structure. Level-2 is the true level-1 and defines the covariance structure for the categorical and continuous responses, and coefficients associated with the continuous responses can be made random at this level to provide a heterogeneous ‘level-1’ variance or covariance structure. For more details, see Goldstein (1995). Level 3 is the real level-2, any of the coefficients can be made random at this level. We write the model in a general form as

$$F_{hij} = (1 - \delta_{hij}) F_{1ij} + \delta_{hij} F_{2ij} \quad (1.5)$$

where δ is a dummy variable coded as one for the continuous responses, zero for the categorical responses. F_1 expresses the models for the multinomial responses, and F_2 is for that of the continuous responses. Note that $F_{1ij} = \pi_{1ij}$ so that (1.6) contains a non-linear and a linear component. At the lowest level (2) any covariances between the discrete and continuous variables can be interpreted as biserial correlations (Goldstein, 1995) and at higher levels we assume multivariate Normality for the set of residuals from both parts of the model.

Estimation

Quasilikelihood estimation is used and is analogous to that for the generalised linear models for binary data. First and second order approximations are available with MQL and PQL procedures.

Running the macros

Having read in your data file type OBEY OPTIONS.MC which causes the following screen menu to appear

```

      MULTIPLE RESPONSE CATEGORY OPTIONS (RELEASE 1.1)
      =====
ERROR DISTN.           :B10=* - MULTINOMIAL(0), ORDERED MULTINOMIAL(1)
APPROXIMATION         :B11=* - 1ST ORDER(1), 2ND ORDER(2)
NONLINEAR PREDICTION :B12=* - FIXED:MQL(0), FIXED+RESIDUALS:PQL(1)
LINK FUNCTION         :B13=* - LOGIT(0), LOGLOG(1)
VARIANCE FUNCTION     :B14=* - DISTRIBUTIONAL(0), UNCONSTRAINED(1)
MIXED RESPONSE       :B16=* - NO(0), YES(1)
* = UNSET OR OUT OF ALLOWABLE RANGE

```

This is the default screen with no options set within the worksheet loaded. The option settings are clearly labelled and the model settings are constructed in the usual way. The extra requirements are

1. You must name a variable DENOM which contains the denominator for the response vector. The macros detect this and use it.
2. Declare the level-1 variance defining columns linked to G9 using the LINK C C... G9 command. The group G9 must be used to define the level-1 variation prior to starting the iterations. The columns in this group should not be used for anything else.
3. For models with both a proportion vector and continuous responses, we require dummy variables defining the lowest level variation for the vector of proportions and these are all linked to G9. We must also link *all* the explanatory variables for the continuous response to group G11.

The commands START, NEXT, RESID are used in the usual way.

Examples

An unordered categorical response model

The data are a sub-sample of the Junior School Project data (Mortimore et al, 1988). Open the worksheet 'JSPMIX.WS1', supplied with the macros in *MLwiN* and click on the **Names**

window to obtain the following screen. The variables ERY3 and TBY are two types of responses, continuous and categorical. There are missing values on ERY3 coded as -1 which we shall omit after the multivariate structure has been formed. We study first the categorical response TBY, and later we shall illustrate the mixed response model by including ERY3.

Name	n	missing	min	max
1 SCY3	1313	0	1	50
2 ID	1313	0	280	2286
3 SEX	1313	0	0	1
4 STAG	1313	0	1	3
5 RAV1	1313	0	4	36
6 ERY3	1313	0	-1	98
7 TBY	1313	0	1	3
8 C8	0	0	0	0
9 C9	0	0	0	0

SCY3 : school identification
 ID : pupil identification
 SEX : girl=0, boy=1
 STAG : stage of fluency in English at year 1.
 fully fluent=1, intermediate=2, beginner=3
 RAV1 : RAVENS test score year 1
 ERY3 : English reading test yr 3.
 TBY : Teacher's behaviour rating at yr 3:
 top 25%=1, middle 50%=2, bottom 25%=3.
 Missing data = -1

In the **Command interface** window type the following commands to form the multivariate structure before doing any model fitting. We may choose the 'top' category of TBY as the baseline, and remove this group entirely from the worksheet. We also leave out the continuous outcome ERY3 from this worksheet. The command VECTorise stacks the remaining two categorical responses of TBY of an individual into a single record in C8 to form a new response and puts an indicator in C9. The command REPEat simply repeats other variables two times to produce a further column the same length as the new response one. Explanatory variables associated with SEX are generated for each category of the response.

Two columns 'MVAR' and 'DENOM' are named after constructing the multivariate structure.

```

NAME c8 'tby1' c9 'tby2' c10 'tby3'
DUMM 'tby' c8 c9 c10
ERAS 'ery3' 'tby' 'tby1'
MOVE
VECT 2 'tby2' 'tby3' c8 c9
NAME c8 'resp' c9 'index'
DUMM 'index' 'tby2' 'tby3'
REPE 2 c1-c5 c1-c5
PUT 2626 1 c10
CALC c11=c10
NAME c10 'denom' c11 'mvar'
CALC c12='tby2' * 'sex'
CALC c13='tby3' * 'sex'
NAME c12 'tby2boy' c13 'tby3boy'
  
```

C10 contains all ones as denominator and C11 defines the level-1 variance. We must call the denominator 'DENOM' and link the variable 'mvar' to group 9. Here level-2 is the real level-1, and the macros will supply the necessary structure for categories at this level. Typing in the following commands will specify the structure and set up model A of table 1.1.

```
IDEN 3 'scy3' 2 'id' 1 'index'
RESP 'resp'
EXPL 1 'tby2' 'tby3' 'mvar'
FPAR 0 'mvar'
SETV 3 'tby2' 'tby3'
SETV 1 'mvar'
LINK 'mvar' g9
```

We now click on **Options** to bring up the **Directories** screen and enter the full file path, as described above. The default is `c:\program files\mlwin\multicat` and the pre and post file names are: `PRE.MC` and `POST.MC`. If *MLwiN* and the macros have been installed in another directory/folder this name should be used.

Alternatively we can create a macro file by clicking on File/New macro, and type in all commands listed above. Then click on Execute button to run the commands in batch mode. This macro can be saved, edited, and reused for models in later sections.

For the simplest multinomial logistic model with only intercepts we choose the 1st order MQL procedure and type in the following commands:

```
OBEY options.mc
SET b10 0
SET b11 1
SET b12 0
SET b13 0
SET b14 0
SET b16 0
BATCh 1
MAXI 100
START 1
```

The estimates after convergence are in Table 1.1 (model A). They can be shown by using `FIXE` and `RAND` commands.

To model gender effects, we type

```
EXPL 1 'tby2boy' 'tby3boy'
NEXT 1
```

This produces the estimates for model B in Table 1.1. To fit a model with a log-log link, we simply type `SET b13 1`, then type `START`. If you receive a 'numeric warning' message click 'yes' to proceed. The estimates are given by model C in the table. To fit extra-multinomial variation, we type `SET B14 1`, then type `NEXT 1` till convergence. For this example, there is little evidence for extra-multinomial variation, and we can rely on the standard assumption. Table 1.1 also presents the estimates using the second order approximation for both logistic and log-log links. For the logistic link, the model shows some variation of `TBY2` and `TBY3` on the log-odds scale between schools but little correlation between the two. The same results occur at level 3 for the log-log link.

The parameter 1 following commands `START` and `NEXT` ensures the model being fitted through the backend of the program with the number of iterations being controlled by users.

Table 1.1 Estimates from different model specifications for unordered TBY response. (S.E. in brackets)

	Model A b11=1, logit	Model B b11=1, logit	Model C b11=1, log-log	Model D b11=2, logit	Model E b11=2, log-log
Fixed					
TBY2	0.72(0.08)	0.51(0.09)	-0.32(0.07)	0.50(0.09)	-0.324(0.07)
TBY3	-0.08(0.09)	-0.59(0.12)	-1.66(0.11)	-0.64(0.12)	-1.72(0.11)
TBY2BOY		0.48(0.11)	-0.003(0.08)	0.57(0.11)	-0.003(0.08)
TBY3BOY		1.01(0.14)	0.60(0.12)	1.14(0.14)	0.61(0.12)
Random					
Level 3					
TBY2/TBY2	0.12(0.06)	0.12(0.05)	0.06(0.03)	0.12(0.05)	0.06(0.03)
TBY2/TBY3	0.02(0.05)	0.02(0.05)	0.01(0.03)	0.02(0.05)	0.01(0.03)
TBY3/TBY3	0.17(0.08)	0.18(0.08)	0.14(0.06)	0.19(0.08)	0.15(0.06)
Level-2					
-P/P	1	1	1	1	1
level-1					
MVAR/MVAR	1	1	1	1	1

The fixed effects of Model A suggest a larger proportion of pupils in the middle category and a smaller proportion of pupils in the lowest level category than in the highest category for behaviour. The gender effects of Model B suggest that more boys were assessed as being in the middle and bottom categories rather than the top category, compared to girls.

To show the findings more clearly, we have calculated the predicted proportions based on models B, and C using equations (1.1) and (1.2). The results in Table 1.2 suggest that these two links fit the data almost equally well. Note, however that these predictions are at the mean of the level-3 distribution on the logistic and log-log scales. Because of the non-linearity of the transformation these will not exactly coincide with the population mean on the probability scale. See Goldstein (1995 Section 5.3) for a further discussion of how to estimate functions of population parameters on the probability scale.

Table 1.2 Comparison between raw proportions and predicted ones, by gender

	Raw proportion		Predicted (logistic)		Predicted (log-log)	
	Girl	Boy	Girl	Boy	Girl	Boy
TBY1	30.4	19.6	31.0	19.1	30.9	20.0
TBY2	52.0	51.2	51.8	51.7	51.8	51.7
TBY3	17.6	29.2	17.2	29.3	17.3	29.2

An ordered categorical response model

In the above example the teachers' behaviour rating is in fact an ordered categorical variable. We shall fit an ordered multinomial model and compare the results with the unordered model.

Open the same worksheet 'JSPMIX.WS1', and type the following commands to get the data into the correct multivariate structure.

```

Eras 'ery3'
Dumm 'tby' c10-c12
Name c10 'tby1' c11 'tby2' c12 'tby3'
Vect 3 c10-c12 c8 c9
Repe 3 c1-c5 c1-c5
Name c8 'resp' c9 'index'
Dumm c9-c12
Eras 'tby'
Move
Mlcv 'id' 'resp' c11
Name c11 'cumu'
Omit 3 'index' c1-c11 'index' c1-c11

```

The command MLCU forms the cumulative response, OMIT removes the last category ('INDEX'=3).

The next step is to define the subdirectory and pre and post files for the macros from the **Options** window as before.

Finally we set up the simplest variance component model with a logit link which is similar to model A in Table 1.1 but with only the single term 'CONS' at level 3. The level-2 random part is put in by the macros after the first iteration. The macro settings are the same as in section 4 except that the box B10 is now set to 1. Then run the model.

```

PUT 2626 1 C12
CALC C13=C12
CALC C14=C13
NAME C12 'cons' C13 'denom' C14 'mvar'
IDEN 3 'scy3' 2 'id' 1 'index'
RESP 'cumu'
EXPL 1 'tby1' 'tby2' 'cons' 'mvar'
FPAR 0 'cons' 'mvar'
SETV 3 'cons'
SETV 1 'mvar'
LINK 'mvar' g9
OBEY options.mc
SET b10 1
SET b11 1
SET b12 0
SET b13 0
SET b14 0
SET b16 0
BATC 1
MAXI 50
START 1

```

After the first iteration 'MVAR' will be removed by the macros from level-1. In the ordered response model, setting 'MVAR' at level-1 to start with produces OLS estimates as starting values for the model at the first iteration. After the first iteration the macros will insert the term 'P' for the covariance structure at level-2.

This produces the results in Table 1.3 (model A) at convergence. To model the effects of any other explanatory variables we enter them in the model in the standard way, as given by (1.4) rather than as interactions with response category indicator variables. For example we use the following commands to fit gender:

```
EXPL 1 'sex'
NEXT 1
```

The results for this model are in Table 1.3 (model B), and show that boys are less likely than girls to be assessed by teachers at both the top and middle categories on their behaviour. The predictions in table 1.4 make this clearer.

To switch the link function from logit to log-log, type `SET B13 1`, then type `NEXT 1` till convergence.

Table 1.3 Estimates from different model specifications for ordered TBY response. (SE in brackets)

	Model A b11=1,logistic	Model B b11=1, logistic	Model C b11=1, log-log	Model D b11=1, log-log
Fixed				
TBY1	-1.087 (0.097)	-0.785 (0.111)	-1.237 (0.068)	-1.052 (0.074)
TBY2	1.2 (0.098)	1.551 (0.118)	0.382 (0.052)	0.586 (0.063)
SEX		-0.662 (0.109)		-0.394 (0.067)
Random				
Lev. 3 Var(cons)	0.240 (0.078)	0.266 (0.084)	0.067 (0.025)	0.074 (0.027)
Lev. 1 P*	1.0	1.0	1.0	1.0

Fitting the extra-multinomial variance at level-2 (`SET B14 1`) shows little departure from the value of 1.0 with an estimate of 0.964 for model B with a logit link and 0.975 for model D with the log-log link.

Applying equation (1.4) we can predict the cumulative proportions based on model B in Table 1.3 for boys and girls separately in the third column in Table 1.4. Applying equation (1.5) and based on model D in Table 1.3 produces predictions for the log-log link in Table 1.4. Compared with the raw proportions in Table 1.4, the logistic link function fits the data slightly better than the log-log link does.

Table 1.4 Comparison between raw cumulative proportions and predicted ones, by gender

	Raw proportion		Predicted (logit)		Predicted (log-log)	
	Girl	Boy	Girl	Boy	Girl	Boy
TBY1	30.4	19.6	31.3	19.0	29.5	21.0
TBY2	82.4	70.8	82.5	70.9	83.4	70.2
TBY3	100.0	100.0	100.0	100.0	100.0	100.0

We can also build up the model by adding more explanatory variables such as `RAV1`, `STAG` in the fixed part of the model, or allow the coefficient of `SEX` to vary randomly among schools.

Mixed model I: continuous and unordered categorical responses

In this section, using the same dataset, we add the continuous outcome `ERY3` to the response column. Missing values of `ERY3` will be omitted after the multivariate structure is formed.

Having opened the worksheet ' `jspmix.ws1` ', we now structure the data as follows:

```

DUMM 'tby' C10 C11 C12
NAME C10 'tby1' C11 'tby2' C12 'tby3'
VECT 3 'ery3' 'tby2' 'tby3' C8 C9
REPE 3 C1-C5 C1-C5
NAME C8 'resp' C9 'index'
DUMM 'index' 'ery3' 'tby2' 'tby3'
ERAS 'tby1' 'tby'
MOVE
OMIT -1 'resp' C1-C10 'resp' C1-C10
PUT 3745 1 C12
CALC C11='ery3'
CALC C13=1-'ery3'
NAME C11 'cons' C12 'denom' C13 'mvar'

```

There are 3745 records available for further modelling. This number is not just three times 1313 as one might have expected. The reason is that the 194 missing values of ERY3 have been omitted from the data. C12 contains ones, being the denominator column and C11 is one for the continuous ERY3 only and zero for the categorical responses. If there were another continuous response, another indicator column should be created. C13 contains ones for the categorical variables and zeroes for the continuous response.

Let us start from the simplest model, estimating only the mean score for ERY3 and mean proportions for the TBV categories.

The variance-covariance matrix of the categorical responses will be set by the macros at levels 1 and 2 through the terms P and $-P$.

Specify the models in the usual way and make MVAR define the level-1 variation:

```

IDEN 3 'scy3' 2 'id' 1 'index'
RESP 'resp'
EXPL 1 'cons' 'tby2' 'tby3' 'mvar'
FPAR 0 'mvar'
SETV 3 'cons' 'tby2' 'tby3'
SETV 2 'cons'
SETV 1 'mvar'
MAXI 20
BATCH 1

```

In addition to the model settings, we need to define the subdirectory for the macros as before. For the first model we choose the multinomial logit link with a 1st order MQL approximation.

```

SET B10 0
SET B11 1
SET B12 0
SET b13 0
SET b14 0
SET b16 1

```

The macros require us to link the explanatory variable associated with the level-1 variance of categorical responses to group G9 and to link explanatory variables associated with the continuous response to group G11.

```

LINK 'mvar' g9
LINK 'cons' g11

```

To check whether we have set macro options correctly we can OBEY OPTIONS.MC or NLSETT

To run the model, type START. At convergence, typing RAND and FIXE we obtain:

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR (U)	PREV. ESTIMATE	CORR
3	CONS /CONS	(9)	71.88	18.7	71.87	1
3	TBY2 /CONS	(7)	0.962	0.729	0.962	0.327
3	TBY2 /TBY2	(5)	0.121	0.0549	0.121	1
3	TBY3 /CONS	(4)	-1.383	0.885	-1.381	-0.401
3	TBY3 /TBY2	(1)	0.0213	0.0468	0.0215	0.151
3	TBY3 /TBY3	(2)	0.166	0.0763	0.165	1
2	CONS /CONS	(12)	396.4	16.95	396.4	1
2	-P /CONS	(7)	6.045	0.464	6.045	
2	P / -P	(11)	1	2.5e-009	1	
1	MVAR / MVAR	(11)	1	1.456e-009	1	
	PARAMETER		ESTIMATE	S. ERROR	PREV. ESTIMATE	
	CONS		40.58	1.39	40.58	
	TBY2		0.713	0.0759	0.714	
	TBY3		-0.0822	0.0896	-0.0828	

The macros define the level-1 random variable as $\sqrt{\pi_{ij}/n_{ij}}$, with coefficient variance constrained to be equal to 1. The variables, P and $-P$ at level-2, are defined as $\pi_{ij}/\sqrt{2n_{ij}}$ and $-\pi_{ij}/\sqrt{2n_{ij}}$ respectively with just a covariance term constrained to be equal to 1. This fits the covariance structure of the multinomial model with a logit link.

Both the fixed and random parameter estimates of TBY2 and TBY3 are similar to those of model A in Table 1.1. At school level the three correlation coefficients between v_{0j} and $v_j^{(s)}$ are $r_{v_{0,2}} = 0.33$, $r_{v_{0,3}} = -0.40$ and $r_{v_{2,3}} = 0.15$. Though the correlation coefficients are small, there is a tendency for schools with a high average reading score to have a higher probability of the behaviour assessment by teachers as category 2.

Suppose we were interested in looking at the main effects of gender, the initial English stage of pupils and Ravens (matrices) score on the final year's reading test as well as the behaviour assessment by teachers. This requires building up the basic model by adding more explanatory variables associated with the two kinds of response. For each category and continuous response a new explanatory variable should be formed by multiplying each indicator variable by the variable to be added. One explanatory variable will generate several new explanatory variables corresponding to each kind of response variable, in our case, three related to gender, three more related to Ravens and six more related to English stage.

Let us fit the main effects of gender first. Our assumptions on variances and covariances remain unchanged.

```

CALC C14='cons'*'sex'
CALC C15='tby2'*'sex'
CALC C16='tby3'*'sex'
NAME C14 'con_sex' C15 'tby2_sex' C16 'tby3_sex'

```

The explanatory variable associated with the continuous response is in C14. It should be linked to G11 together with the constant term. It needs to be declared also as an explanatory variable on the setting screen.

```

EXPL 1 'con_sex' 'tby2_sex' 'tby3_sex'
LINK 'cons' 'con_sex' g11
NEXT 1

```

At convergence we obtain the following estimates for the fixed part. Average reading score of boys is -5.75 lower than that of girls, and boys have greater odds-ratios of getting middle and bottom ratings on their behaviour than girls.

PARAMETER	ESTIMATE	S. ERROR	PREV. ESTIMATE
CONS	43.51	1.488	43.51
TBY2	0.501	0.0934	0.501
TBY3	-0.588	0.121	-0.587
CON_SEX	-5.752	1.192	-5.751
TBY2_SEX	0.485	0.112	0.486
TBY3_SEX	1.006	0.135	1.001

For complex variance models associated with the continuous response only, we can add appropriate terms at level-2, for example SETE 2 'cons' 'con_sex' to estimate variances for boys and girls separately.

Mixed model II: continuous and ordered categorical responses

We can use the same worksheet 'JSPMIX.WS1' but treating the TBY categories as ordered responses together with ERY3. Although the procedure for creating the multivariate structure is as the same as that in the previous section, there are two noteworthy differences. First, the TBY categories should be made cumulative, and second, the last category TBY3 is left out from the worksheet after the multivariate structure is formed.

```

DUMM 'tby' c10-c12
VECT 4 'ery3' c10-c12 c8 c9
REPE 4 c1-c5 c1-c5
NAME c10 'tby1' c11 'tby2' c12 'tby3'
NAME c8 'resp' c9 'index'
DUMM 'index' 'ery3' 'tby1' 'tby2' 'tby3'
OMIT -1 'resp' c1-c6 c9-c12 'resp' c1-c6 c9-c12
CALC 'ery3'='ery3'*'resp'
CHAN 2 100 'resp' 0 'resp'
MLCU 'id' 'resp' c13
CALC c13=c13 + 'ery3'
NAME c13 'mix-resp'
CHAN 2 100 'ery3' 1 'ery3'
ERAS 'tby3' 'resp' 'tby'
MOVE
OMIT 4 'index' c1-c10 'index' c1-c10
PUT 3745 1 c11
CALC c12=c11
NAME c11 'denom' c12 'mvar'

```

Above are the commands used to prepare the worksheet. The next step is to set up the simplest model with the following commands:

```

IDEN 3 'scy3' 2 'id' 1 'index'
RESP 'mix-resp'
EXPL 1 'ery3' 'tby1' 'tby2' 'mvar'
FPAR 0 'mvar'
SETV 3 'ery3' 'tby1' 'tby2'
SETV 2 'ery3'
SETV 1 'mvar'
LINK 'mvar' g9
LINK 'ery3' g11
MAXI 20
BATC 1

```

Click on **Options** to specify the path to the macros and the pre and post files. Now set these boxes:

```

SET b10 1
SET b11 1
SET b12 0
SET b13 0
SET b14 0
SET b16 1

```

Finally type `START 1` to fit the model.

During iterations, some 'reconstruction of data blocks' will occur, but convergence can still be achieved with the following estimates:

PARAMETER		ESTIMATE	S. ERROR(U)	PREV. ESTIMATE		
ERY3		40.09	1.39	40.09		
TBY1		-1.076	0.1135	-1.076		
TBY2		1.208	0.09029	1.208		
LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR. ---
---	---	---	---	---	---	---
ERY3	/ERY3	(3)	71.92	18.71	71.92	1 3
TBY1	/ERY3	(1)	0.04479	1.091	0.04479	0.00836 3
TBY1	/TBY1	(3)	0.3995	0.1248	0.3994	1 3
TBY2	/ERY3	(3)	1.077	0.8945	1.077	0.309 3
TBY2	/TBY1	(3)	0.182	0.07888	0.182	0.701 3
TBY2	/TBY2	(3)	0.1688	0.07766	0.1688	1 -----
---	---	---	---	---	---	---
/ERY3	(3)	395.3	16.83	395.3	1 2	ERY3 C397
/ERY3	(3)	12.32	0.9456	12.32	2	C398 /ERY3 (
3)	3.393	0.3018	3.393	2	P	* (5)
1	0	1				

In the fixed part of the model, the estimates for TBY1 and TBY2 are comparable with the earlier ones where only one common random parameter for the two categories was allowed at the school level. We see that the relationship between the school mean of the ERY3 reading score and the total grade of teachers' assessment on pupils' behaviour is weak (correlation coefficients as 0.008 and 0.309). The covariance terms between TBY2, TBY3 and ERY3 at the individual level may have no clear interpretation.

To model the gender effects associated with the two types of responses, we type the following:

```

CALC c13='sex'*'ery3'
CALC c14='sex'*(1-'ery3')
NAME c13 'ery3boy' c14 'tbyboy'
EXPL 1 c13 c14
LINK 'ery3' 'ery3boy' g11
NEXT 1

```

At convergence following results are obtained:

PARAMETER		ESTIMATE	S. ERROR(U)	PREV. ESTIMATE		
ERY3		42.83	1.482	42.83		
TBY1		-0.7694	0.1286	-0.7694		
TBY2		1.564	0.1102	1.564		
ERY3BOY		-5.454	1.19	-5.454		
TBYBOY		-0.6674	0.1086	-0.6674		
LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.
---	---	---	---	---	---	---
3	ERY3	/ERY3	(4)	68.47	17.91	68.47 1
3	TBY1	/ERY3	(1)	0.06196	1.124	0.06198 0.011
3	TBY1	/TBY1	(3)	0.4626	0.1386	0.4626 1
3	TBY2	/ERY3	(2)	0.9531	0.8832	0.9533 0.272
3	TBY2	/TBY1	(3)	0.2107	0.08519	0.2107 0.733
3	TBY2	/TBY2	(3)	0.1789	0.08033	0.1789 1
---	---	---	---	---	---	---
2	ERY3	/ERY3	(4)	387	16.51	387 1
2	C397	/ERY3	(4)	11.38	0.9097	11.38
2	C398	/ERY3	(3)	2.949	0.2841	2.949
2	P	*	(11)	1	0	1

Further elaboration of the model in the fixed part as well as random coefficients at level 3 can be carried out in the usual way. At level-2, complex variation associated with ERY3 may be modelled.

Some tips

The following may be helpful.

- Changing model settings one at time can avoid numerical problems in many cases. For example, if a model was fitted through 1st order + MQL with variance being constrained, the next model might be the 2nd order + MQL with extra multinomial variation and finally a 2nd order PQL model.
- Data which cannot be modelled satisfactorily by a multinomial logit model may be fitted better by a multinomial log-log model, and vice versa.
- Constructing the multivariate data structure from a large dataset with many categories can expand the worksheet enormously and create computational problems. Choosing the base category which has the largest proportion and combining some small sized categories (if justified) can reduce the worksheet size considerably.
- Always save the worksheet after major data manipulation or model changes.
- Residual calculations at and above level 3 and plotting can be done using the **Residuals** window.

Chapter 2. Survival and Event duration models

These macros will allow the fitting of log duration and semiparametric Cox models for survival or event duration (history) response data. See Goldstein (1995, Chapter 9) for a detailed discussion.

Log duration (accelerated failure time) models

Given length of time t with censoring indicator $z = 1$ (if censored) and one single explanatory variable x , a simple two level log duration model can be written in standard notation as

$$l_{ij} = \ln(t_{ij}) = (1 - z)(\beta_{0j} + \beta_1 x_{ij} + e_{1ij}) + z(\pi_{ij}(\beta_{0j}, \beta_1, e_{2ij})) \quad (2.1)$$

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

$$e_{1ij} \sim N(0, \sigma_{e_1}^2), \quad e_{2ij} \sim B(1, \pi_{ij}), \quad \text{cov}(e_{1ij}, e_{2ij}) = 0$$

Model (2.1) consists of two parts. The first part fits non-censored observations. The second part is for censoring with π_{ij} being the probability of censoring, which can be based upon different error distributions; Normal, Extreme Value, Gamma and Logistic. For data with no censored observations ($z = 0$), model (2.1) effectively is fitted just as a standard Normal two-level model. This set of macros allows only right censoring. They will also fit mixed models where there is a single duration response and several ordinary Normal response measurements.

Note that where a dataset contains more than 50% censored observations, applying these models may produce biased estimates. More work is being carried out on this topic.

Semiparametric Cox Models

Consider a simple 2-level model with, say, students within schools or occasions within subjects. We consider each time point in the data as defining a *block* indicated by l at which some observations come to the end of their duration due to either failure or censoring and some remain to the next time or block. At each block there is therefore a set of observations we denote as the total *risk set*. We may think of each observation within a block as a level-1 unit, above which, in the repeated measures case, there are occasions at level-2 and subjects at level 3. The ratio of the hazard for the unit which experiences a failure at a given occasion referred to by (j', k') to the sum of the hazards of the remaining risk set units (see McCullagh and Nelder, 1989) is

$$\frac{\exp(\beta_1 x_{1ij'k'} + u_{j'k'})}{\sum_{j,k} \exp(\beta_1 x_{1ij'k'} + u_{j'k'})} \quad (2.2)$$

where j and k refer to the real levels 2 and 3, for example occasion and subject or student and school.

In *MLwiN* we set up the model as follows. At each block denoted by l the response variable is defined for each member of the risk set as

$$y_{ijk(l)} = \begin{cases} 1 & \text{failed} \\ 0 & \text{not} \end{cases}$$

Because of equivalence between the likelihood for the multinomial and Poisson distributions, the latter is used to fit model (2.2). This can be written as

$$y_{ijk(l)} = \exp(\alpha_l + X_{jk} \beta_k) \quad (2.3)$$

Where there are ties within a block then more than one response will be non-zero. The terms α_l fit the 'blocking factor', and can be estimated from fitting either a set of parameters, one for each block, or a smoothed polynomial curve over the blocks numbered $1, \dots, p$. Thus if the h^{th} block is denoted by h , α_l is replaced by a low order polynomial, order m , $\sum_{t=0}^m \gamma_t h^t$, where the γ_t are (nuisance) parameters to be estimated.

The data are sorted into a 2-level structure, for example in the repeated measures case by failure times within subjects and occasions within the failure times. This retains proportional hazards within subjects. In this formulation the Poisson variation is defined at level-1 as described in Chapter 1, there is no variation at level-2 and the between-subject variation is at level 3. Alternatively we may wish to preserve *overall* proportionality, in which case the failure times define level 3 with no variation at that level. See Goldstein (1995, Chapter 9) for a discussion of this.

Estimation

As seen in equation (2.1), estimation for the log duration model involves a bivariate mixed response model where the natural logarithms of the known duration lengths are the continuous response and the censored observations are treated as binary responses with a logit link function.

Estimation for the semiparametric model is as for the Poisson model with offsets as described in the *MLwiN* help.

Running the macros

The default directory for these macros is c:\program files\mlwin\survival, and the pre and post file names are PRE.SU and POST.SU.

Having specified these from the **Options** window, typing OBEY OPTIONS.SU will cause the following screen to appear

```

SURVIVAL MODEL OPTIONS (RELEASE 1.0)
=====
PARAMETRIC      :B10=* - NORMAL(1), EXTREME VALUE(2), GAMMA(3), LOGISTIC(4)
SEMI-PARAMETRIC:B11=* - YES(1), LINK PVAR IN (G9)
MIXED RESPONSE :B12=* - YES(1), LINK SURVIVAL RELATED VARS. IN (G10)
  ADDITIONAL OPTIONS FOR SEMI-PARAMETRIC MODEL (IF B11=1)
  =====
APPROXIMATION   :B13=* - 1ST ORDER (1), 2ND ORDER (2)
VARIANCE FUNCTION :B14=* - POISSON (0), EXTRA-POISSON (1)
NONLINEAR PREDICTION:B15=* - FIXED: MQL(0), FIXED+RESIDUALS: PQL(1)
*=UNSPECIFIED

```

The options require the following to be specified

- A column named 'UNCENS' contains a zero if the observation is censored at the time coded in the column for time and one if not censored.
- A column named 'RIGHT' which is coded as one if the censoring is right censored and zero otherwise.
- For the semiparametric model the variable defining the level-1 (Poisson) variation must be linked to G9.
- For the mixed response model the explanatory variables associated with the survival time response must be linked to G10.

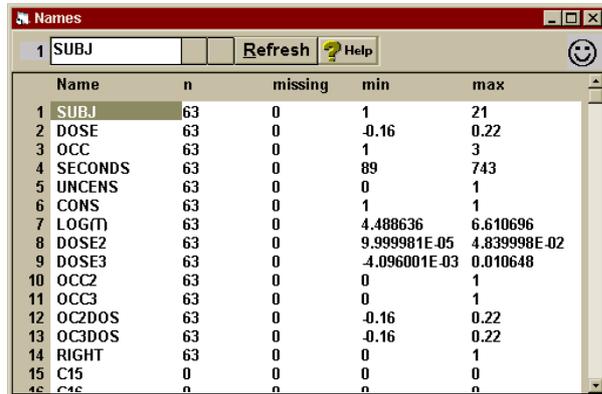
The commands START, NEXT, and RESID may be used in the usual way. The command LIKE provides the value of -2 loglikelihood and should be used only where a Normal error option is specified.

Examples

A log duration model

The data to be modelled are exercise times of subjects with angina pectoris in seconds after being given high dose oral isosorbide dinitrate (mg/kg) (Danahy et al, 1977). Twenty one subjects were involved, each with observations at three occasions. Right censoring was recorded when fatigue occurred.

This is a two-level structure with subjects at level-2 and repeated measures at level-1. The response variable is length of time in seconds with some censored observations. The worksheet is named 'ANGINA.WS1' and retrieving it and opening the **Names** window gives the following:



	Name	n	missing	min	max
1	SUBJ	63	0	1	21
2	DOSE	63	0	-0.16	0.22
3	OCC	63	0	1	3
4	SECONDS	63	0	89	743
5	UNCENS	63	0	0	1
6	CONS	63	0	1	1
7	LOG(T)	63	0	4.488636	6.610696
8	DOSE2	63	0	9.999981E-05	4.839998E-02
9	DOSE3	63	0	-4.096001E-03	0.010648
10	OCC2	63	0	0	1
11	OCC3	63	0	0	1
12	OCC2DOS	63	0	-0.16	0.22
13	OCC3DOS	63	0	-0.16	0.22
14	RIGHT	63	0	0	1
15	C15	0	0	0	0
16	C16	0	0	0	0

Column 'SUBJ' is the level-2 identifier or subject number; 'DOSE' is a level-2 explanatory variable centred about 0.36; 'OCC' is the occasion and used as the level-1 ID, 'SECONDS' is the exercise time to angina pectoris. 'UNCENS' is the flag for censored observations coded as 1 for non-censored and 0 otherwise. The remaining columns have been recoded from the first five. 'LOG(T)' is the logarithm of C4 and is the response variable. C8 and C9 are quadratic and cubic terms of 'DOSE' respectively. C10 and C11 are dummy variables for occasions 2 and 3 as contrasts with occasion 1. C12, C13 are interactions between 'OCC2', 'OCC3' and 'DOSE' respectively. Finally 'RIGHT' is coded as 1 for right censored observations and 0 otherwise.

We are interested in modelling the relation between dose and exercise time and the differences among occasions.

The model may be written as

$$l_{ij} = \beta_{0j}(\text{cons})_{ij} + \beta_1(\text{dose})_j + \beta_2(\text{occ2})_{ij} + \beta_3(\text{occ3})_{ij} \\ + \beta_4(\text{occ2})_{ij}(\text{dose})_j + \beta_5(\text{occ3})_{ij}(\text{dose})_j + e_{1ij} \\ \beta_{0j} = \beta_0 + u_{0j}, u_{0j} \sim N(0, \sigma_u^2), e_{1ij} \sim \text{Extreme}(0, \sigma_{e1}^2)$$

In this model, for the uncensored observations, the first two terms estimate a linear relation between dose and log duration for occasion 1, the third and fourth terms estimate differences of intercepts between occasions 2, 3 and 1 respectively. The two interaction terms estimate differences of linear slopes between occasions 2, 3, and 1 respectively. The last term e_{1ij} is for the level-1 residuals which, in this case, we assume follow an Extreme Value distribution. The difference of intercepts among subjects is described by the Normal random intercept term u_{0j} . For the censored observations, a binomial error is assumed with a logit link function (see Goldstein, 1995, Chapter 9 for details).

Note that if one uses the gamma option, then it is the actual duration rather than its logarithm which should be used as the response. The user should remember, however, that there is an assumption of Normality at level 2 and higher on this scale. Whether it is more reasonable to assume Normality for higher level random effects on the duration scale or the log(duration)

scale will depend on the data, and comparisons of higher level residual plots will be useful here. (An option to use the logarithm of a gamma distributed variable is unnecessary since the gamma distribution and the log-Normal are similar).

Remember to redefine the directories from the **Options** menu. To set up the model, go to the **Command Interface** window and type

```
IDEN 3 'subj' 2 'occ' 1 'cons'
EXPL 1 'cons' 'dose' 'occ2' 'occ3' 'oc2dos' 'oc3dos'
RESP 'log(t)'
SETV 3 'cons'
TOLE 2                                (set the convergence criteria to be less than 0.01)
```

Note that the data structure is shifted up one level. The level-1 variation is placed by the macros at level-2 with no variation at level-1. The user defines only variation at levels 3 and above.

Type the following in the command interface on the left below and the output on the right will be seen:



Command `fset` displays the current directory of the macros and the status of the `pre.su` and `post.su` files.

```
PREFile : pre.su enabled
POSTfile : post.su enabled
FPATH : c:\program files\mlwin\survival\
->obey options.su

SURVIVAL MODEL OPTIONS (Release 1.0)
=====
parametric      :b10=2 - normal(1), extreme value(2), gamma(3), logistic(4)
semi-parametric:b11=* - yes(1), link pvar in (g9)
mixed response :b12=* - yes(1), link survival related vars. in (g10)

ADDITIONAL OPTIONS for SEMI-PARAMETRIC MODEL (if B11=1)
=====
APPROXIMATION   :b13=* - 1st order (1), 2nd order (2)
VARIANCE FUNCTION :b14=* - Poisson (0), extra-Poisson (1)
NONLINEAR PREDICTION:b15=* - fixed: MQL(0), fixed+residuals: PQL(1)

*=unspecified
```

Now we can click on the **Start** button to fit the model. At convergence we have the following outputs, having typed commands `FIXE` and `RAND`, displayed in the **Output** window on the left.

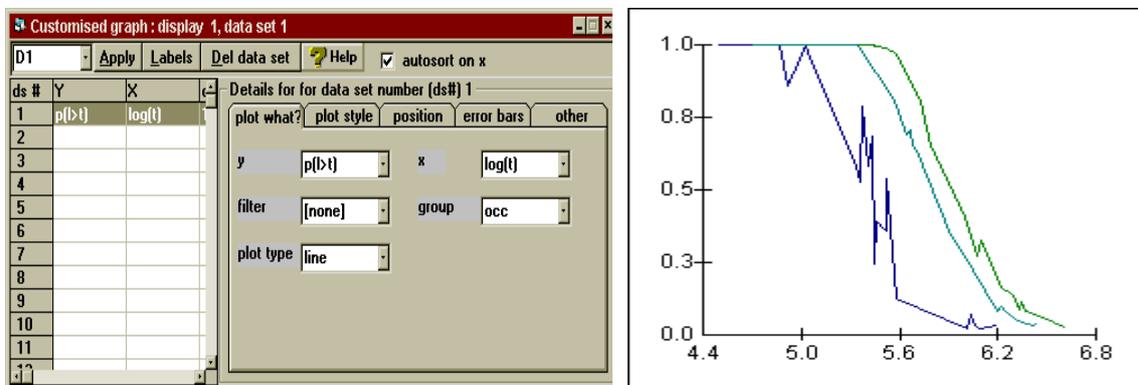
```
->fixe
PARAMETER      ESTIMATE  S. ERROR(U)  PREV. ESTIMATE
cons           5.445     0.1309      5.445
dose           -1.191     1.211       -1.191
occ2           0.5144    0.09629     0.5139
occ3           0.3746    0.0911      0.3742
oc2dos         1.468     0.9627      1.47
oc3dos         1.408     0.8653      1.409
->rand
LEV. PARAMETER (NCONV) ESTIMATE S. ERROR(U) PREV. ESTIM CORR.
-----
3   cons /cons (2)    0.2803  0.09662    0.28    1
-----
2   uncens /uncens (2) 0.07702 0.01945   0.07745 1
2   bin_cens /bin_cens (6) 1      0        1        1
```

At level-2 macros put in the term `UNCENS` for uncensored observations. The variance is the estimated $\text{var}(e_{1ij})$, and the variance term associated with `BIN_CENS` is constrained to be 1.0, for binomial variation. There is no covariance between `BIN_CENS` and any other variables modelled at this level.

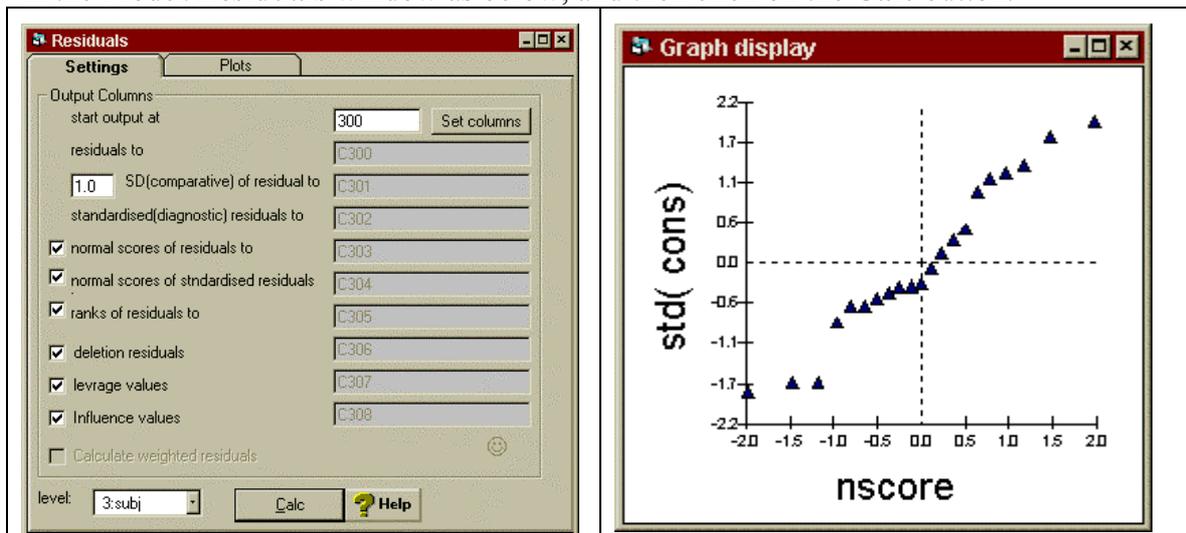
There is little evidence of a linear relation between $\log(t)$ and dose. The log exercise time increases significantly for occasions 2 and 3, suggesting the lowest survival probability at occasion 1 and the highest at occasion 2.

We can plot their survival fractions against $\log(t)$ as follows.

Click on the **Graph/Customised graph(s)** item to bring up the graph window. Then from the drop down list for the Y axis choose the variable 'P(L>T)', which is the survival fraction defined and used by the macros and updated after each iteration. Specify 'LOG(T)' for X, and 'OCC' as group indicator. Having specified the **plot type** as LINE and the **rotate colour** option (for the three occasions), click on the **Apply** button to bring up the following figure. The three lines, from left to right, are for occasions 1, 3, and 2 respectively, showing the same trend as noted with the fixed parameter estimates. We reset the scale for the Y axis in the graph to go from 0.0 to 1.0.



To check the Normality assumption for the residuals at level 3, we need to specify the settings in the **Model/Residuals** window as below, and then click on the **Calc** button.



Move to the **Plots** tab and choose the standardised residual against its normal score to obtain the above (approximately) Normal plot for subject level residuals.

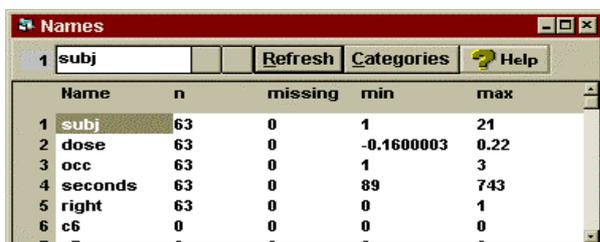
At level-2 we can also model complex variation and estimate any covariance between 'UNCENS' and other variables as usual.

To change the model from one distribution for level-1 residuals to another, for example from Extreme Value to Normal, simply type `SET B10 1`, then click on the **More** button.

A semiparametric model

We now open the same worksheet and convert the data into the required structure with the total 'risk set' and 'failures' corresponding to each time block so that the probability of 'failure' at time block l can be modelled.

To save space and time, we display only the first few columns of the worksheet as below



Name	n	missing	min	max
1 subj	63	0	1	21
2 dose	63	0	-0.1600003	0.22
3 occ	63	0	1	3
4 seconds	63	0	89	743
5 right	63	0	0	1
6 c6	0	0	0	0

The commands used to obtain this worksheet are,
ERAS c5-c13

MOVE

SORT 'seconds' c1-c3 c5 'seconds' c1-c3 c5

Note that the data must be sorted by the time variable before using the command SURVIVAL which converts the data.

Now we convert the data using the command SURV. For these data we have,

SURVIVAL times in <C4>, censors indicator in <C5>, input data in <C1-C3>, response to <C6>, number of failures in interval to <C7>, risk set indicators to <C8>, risk set time to <C9>, risk set size to <C10>, carried data to <C1-C3>.

Apply this and we obtain the expanded worksheet displayed using the **Names** command.

```
SURV 'seconds' 'right' c1-c3 c6-c10 c1-c3
ERAS 'seconds' 'right'
MOVE
NAME c4 'count' c5 'failure' c6 'rsi' c7 'rst' c8 'rss'
TOLE 2
```

Name	n	missing	min	max
1 SUBJ	1728	0	1.0000	21.000
2 DOSE	1728	0	-0.16000	0.22000
3 OCC	1728	0	1.0000	3.0000
4 COUNT	1728	0	0.0000	1.0000
5 FAILURE	1728	0	1.0000	3.0000
6 RSI	1728	0	1.0000	48.000
7 RST	1728	0	89.000	651.00
8 RSS	1728	0	2.0000	62.000

Column 'COUNT' contains the value one where a failure occurred at the particular time point and zero otherwise. It is the response variable for this model. C5 is the total number of failures at each time interval. Column 'RSI' works as an indicator of 'risk sets' or blocks corresponding to failure times. If we wish the 'blocking factor' can be fitted as 48 dummy variables from this column. Column 'RST' is simply the failure time and 'RSS' is the size of 'risk set' at the end of a time interval. Not all of these columns will be used in the estimation. We can VIEW columns C4-C8 thus:

	COUNT	FAILURE	RSI	RST	RSS
N =	1728	1728	1728	1728	1728
1	1.0000	1.0000	1.0000	89.000	62.000
2	0.0000	1.0000	1.0000	89.000	62.000
3	0.0000	1.0000	1.0000	89.000	62.000
⋮	⋮	⋮	⋮	⋮	⋮
63	0.0000	1.0000	1.0000	89.000	62.000
64	1.0000	2.0000	2.0000	110.00	60.000
65	1.0000	2.0000	2.0000	110.00	60.000
66	0.0000	2.0000	2.0000	110.00	60.000
⋮	⋮	⋮	⋮	⋮	⋮
125	0.0000	2.0000	2.0000	110.00	60.000
126	1.0000	1.0000	3.0000	121.00	59.000
127	0.0000	1.0000	3.0000	121.00	59.000
⋮	⋮	⋮	⋮	⋮	⋮
1726	1.0000	1.0000	48.000	651.00	2.0000
1727	0.0000	1.0000	48.000	651.00	2.0000
1728	0.0000	1.0000	48.000	651.00	2.0000

We now re-sort the data by subject and occasion. By doing this we treat subjects as level 3 units, occasions at level-2 and failure times as level-1.

To avoid fitting 48 nuisance parameters for the ‘blocking factor’ for each failure time held in ‘RSI’, we shall fit a cubic curve to the block code centred around 18, to graduate the blocking factor. Columns C13 and C14 are the linear and quadratic terms.

```

SORT 2 'subj' 'occ' c2 c4-c8 'subj' 'occ' c2 c4-c8
IDEN 3 'subj' 2 'occ' 1 'rsi'
PUT 1728 1 c9
CALC c10=c9
NAME c9 'cons' c10 'pvar'
DUMM 'occ' c11 c12
CALC c13='rsi'-18
CALC c14=c13^2
CALC c15=c13^3
CALC c16=c11*'dose'
CALC c17=c12*'dose'
NAME c11 'occ2' c12 'occ3' c13 'crsi' c14 'crsi^2' c15 'crsi^3'
NAME c16 'oc2dos' c17 'oc3dos'

```

To set up the model, type

```

EXPL 1 'cons' 'dose' 'occ2' 'occ3' c13-c14 'oc2dos' 'oc3dos' 'pvar'
FPAR 0 'pvar'
RESP 'count'
SETV 3 'cons'
SETV 2 'cons'
SETV 1 'pvar'

```

Now set the macros parameters (remember to redefine the directories from the **Options** menu).

```

SET b10 0
SET b11 1
SET b12 0
SET b13 1
SET b14 0
SET b15 0
LINK 'pvar' G9
OBEY
settings.su

```

The following screen shows the estimation procedure we have chosen

```

SEMI-PARAMETRIC COX MODEL SETTINGS:  V1.0
=====
APPROXIMATION(B13)      :  FIRST ORDER (1)
VARIANCE FUNCTION(B14)  :  DISTRIBUTIONAL(0)
NONLINEAR PREDICTION(B15):  FIXED PART ONLY:MQL(0)

```

Type `START` or click on start button now to fit the model. You can set batch mode on and change the maximum number of iterations at any stage. At convergence, the following results are obtained.

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE
CONS	-2.129	0.3035	-2.132
DOSE	7.653	2.758	7.623
OCC2	-1.856	0.2	-1.849
OCC3	-1.157	0.171	-1.153
CRSI	0.0962	0.0106	0.0959
CRSI^2	0.001789	0.000454	0.001772
OC2DOS	-5.233	1.51	-5.207
OC3DOS	-5.910	1.405	-5.887

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.	
3	CONS	/CONS	(1)	1.37	0.514	1.369	1
2	CONS	/CONS	(20)	0	0	0	1
1	PVAR	/PVAR	(18)	1	0	1	

As shown in the fixed part of the model, the risk of failure increases as the dose increases, implying a negative effect of dose on survival fraction. Note the zero estimate level-2. This is because level-2 (occasion) has no variation associated with it using this model. Therefore we can collapse the model down to two levels as follows

```

CLRV 3 'cons'
IDEN 2 'subj'
IDEN 3
NEXT

```

And the estimates are the same.

We can choose a higher order estimation procedure by setting `B13=2` (second order approximation) and `B15=1` (PQL) to get more accurate estimates, and we may also fit an unconstrained level-1 variance to check for extra Poisson variation. However, this combination presents numerical problems. We get the following estimates using a first order PQL approximation.

Type

```

SET b14 1
SET b15 1
NEXT 1
FIXE
RAND

```

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE
CONS	-2.933	0.508	-2.933
DOSE	6.160	4.832	6.160
OCC2	-2.033	0.4042	-2.033
OCC3	-1.254	0.3472	-1.254
CRSI	0.130	0.0201	0.130
CRSI^2	0.00150	0.00095	0.00150
OC2DOS	-5.829	3.577	-5.829
OC3DOS	-6.639	3.206	-6.639

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.
2	CONS /CONS	(1)	3.975	1.333	3.974	1
1	PVAR /PVAR	(1)	0.8091	0.0277	0.8088	

This shows that the effect of dose on exercise time is not significant and there is little evidence of an interaction between dose and occasion. There is also some evidence that the variation is less than Poisson. This result agrees with the log duration model.

For this model, the predicted survival fraction for the baseline mean where $X_{jk} = 0$ is given by $\hat{S}_h = \exp(-\sum_{l < h} \hat{\alpha}_l)$ and the estimated mean survival fraction with specific covariate X_{jk} is

$$\hat{S}_h^{\exp(X_{jk}\beta)}$$

For individual survival fractions the subject residuals at level-2 can be calculated and added to the equations above.

For the example above executing the following commands will produce the mean surviving fraction conditional on dose for each occasion stored in columns C18, C21, and C22 for occasions 1 to 3 respectively.

```

SORT 'rst' 'crsi' 'crsi^2' c26-c28
TAKE c26-c28 c26-c28
CALC c18=expo(-2.933+0.13*c27+0.001502*c28)
CUMU c18 c18
CALC c18=expo(-c18)
CALC c21=c18^expo(-2.033)
CALC c22=c18^expo(-1.254)
CALC c19=log(c26)
NAME c19 'log(t)' c18 's-occ1' c21 's-occ2' c22 's-occ3'

```

In the **Graph** window define the survival fraction curves for occasions 1, 3, and 2 and we see a similar pattern to that from fitting the log-duration model.

Customised graph : display 2, data set 3

D2 Apply Labels Del data set Help autosort on x

ds #	Y	X
1	s-occ1	log(t)
2	s-occ2	log(t)
3	s-occ3	log(t)
4		
5		
6		
7		
8		
9		
10		
11		
12		

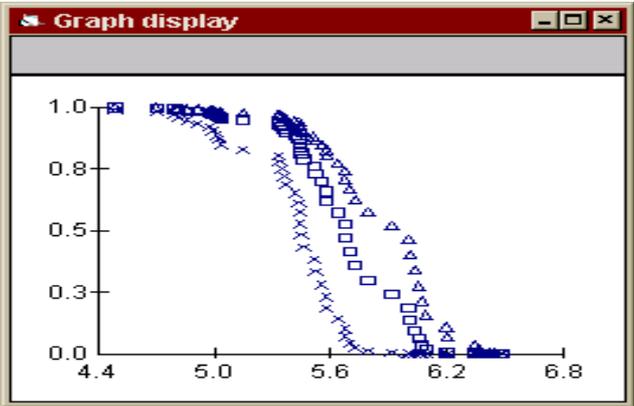
Details for for data set number (ds#) 3

plot what? plot style position error bars other

y s-occ3 x log(t)

filter [none] group [none]

plot type point



Chapter 3. Time series and non-linear variance models

We discuss two kinds of models which involve modelling covariance matrices or variances as non-linear functions of explanatory variables (See Goldstein, 1995 Chapters 5, 6).

Time series models

We consider a 2-level repeated measures model of the form

$$y_{ij} = \sum_p \beta_{pij} t_{ij}^p + e_{ij} \quad (3.1)$$

In these models we assume that the level-1 residuals are no longer independent but have an autocorrelation structure of the following form

$$\text{cov}(e_t e_{t-s}) = \sigma_e^2 \exp(-g(\alpha, s)) \quad (3.2)$$

where σ_e^2 is the level-1 variance term (assumed constant), s is the time difference, and α is a vector of parameters. The function g takes the general form $g = \sum_h \alpha_h s^{l_h}$, which is a polynomial function if the l_h are constrained to be positive integers (e.g. $l_h = h$), or possibly a fractional polynomial. As l_h changes, different models can be fitted. Thus, if an individual is measured on three occasions at times, say 1, 3, and 7, model (3.2) generates the following covariance structure for this individual:

Occ (Time)	1 (1)	2 (3)	3 (7)
1 (1)	σ_e^2		
2 (3)	$\sigma_e^2 \exp(-g(\alpha, 2))$	σ_e^2	
3 (7)	$\sigma_e^2 \exp(-g(\alpha, 6))$	$\sigma_e^2 \exp(-g(\alpha, 4))$	σ_e^2

We can extend this model to the case where the level-1 variance term is a function of time in addition to the time series component. The function g can be written as

$$g = \alpha s + \sum_h \beta_h^* (t_1^h + t_2^h) \quad (3.3)$$

where the second term structures the level-1 variance. It forms another polynomial with integer power terms $h=(1,2,\dots,H)$ only. As h changes, slightly different models are obtained. For example, for $h = 2$, the above table for model (3.3) becomes

Occ (Time)	1 (1)	2 (3)	3 (7)
1 (1)	$\sigma_e^2 e^{-(\beta_1^* t + \beta_2^* t^2)}$		
2 (3)	$\sigma_e^2 e^{-(\alpha 2 + \beta_1^* t + \beta_2^* t^2)}$	$\sigma_e^2 e^{-(\beta_1^* t + \beta_2^* t^2)}$	
3 (7)	$\sigma_e^2 e^{-(\alpha 6 + \beta_1^* t + \beta_2^* t^2)}$	$\sigma_e^2 e^{-(\alpha 4 + \beta_1^* t + \beta_2^* t^2)}$	$\sigma_e^2 e^{-(\beta_1^* t + \beta_2^* t^2)}$

As $s \rightarrow 0$ this implies that the variance is a quadratic function of time, i.e., $\text{var}(y_t) = \sigma_e^2 \exp(-2(\beta_1^* t + \beta_2^* t^2))$. As the correlation between t and $t+s$ is still $\text{corr}(y_t, y_{t+s}) = e^{-\alpha s}$ from this model, we have the same interpretation for α as before. Models in continuous time with any number of occasions for individuals can be fitted.

Another extension to the function g is to allow the covariance to depend on some individual level characteristic (e.g. gender) with the form

$$g = (\alpha_0 + \alpha_1 z_{1j} + \alpha_2 z_{2j})^s \quad (3.4)$$

where z_1 and z_2 are dichotomous variables at the individual level. The term α_0 estimates the covariance for the baseline individuals, and the other two coefficients estimate the differences of the covariance between the baseline individuals and those with the characteristics coded as 1. Only two such variables can be fitted in (3.4) currently.

Non-linear level-1 variance functions

We assume again that the level-1 variance is an exponential function of other explanatory variables such as age. Let z refer to a set of such explanatory variables. We can write a general model as

$$\text{var}(e_{ij}) = \exp(\sum_k \beta_k^* z_{kij}) \quad (3.5)$$

Estimation

The estimates produced are maximum likelihood under the assumption of multivariate Normality.

Running the macros

In the **Options** menu define the subdirectory and pre and post files; by default these are respectively `c:\program files\mlwin\ts`, `PRE.TS` and `POST.TS`. Typing the command `OBEY OPTIONS.TS` from the **Command interface** window brings up the following screen

```

NONLINEAR LEVEL-1 COVARIANCE OPTIONS(RELEASE 1.2)
=====
S (POLY.):B10=* - YES(1), POWERS (C185), START VALS. (C201)
T (POLY.):B11=* - YES(1), POWERS (C186), START VALS. (C202)
Z :B12=* - YES(1), LINK FUNCTION VARIABLE COLUMNS IN (G9)
RELATIONS:B13=* - NONE(0), ADDITIVE(1), PRODUCT(2)
*=UNSPECIFIED

```

The requirements are as follows

- For any time series model with the term associated with S B10 should be set to 1 and for a heterogeneous (non-linear) level-1 variance model to 0.
- B11 specifies whether the second term in (3.3) is fitted.
- B12 should be set to 1 for the heterogeneous model.
- B13 defines how S and T or S and Z are to be combined. For the additive function (3.3) to set B13 to 1; and for the product function (3.4) to set B13 to 2.
- For a heterogeneous level-1 variance model with no time series structure, B10, B11 and B13 should be set to zero.
- In time series models (3.2)-(3.4) the time variable must be named 'T'.

The commands `START`, `NEXT`, `LIKE` and `RESI` can be used in the usual way.

Examples

Continuous time series models

The data are in a worksheet supplied with *MLwiN*.

Open the worksheet 'OXBOYS.WS2' and view the data from the **Names** window.

	Name	n	missing	min	max
1	ID	234	0	1	26
2	AGE	234	0	-1	1.0055
3	HEIGHT	234	0	126.2	174.8
4	AGE2	234	0	0	1.011
5	AGE3	234	0	-1	1.0166
6	AGE4	234	0	0	1.0222
7	CONS	234	0	1	1
8	OCC	234	0	1	9
9	SEAS	234	0	0.84	9.36
10	C10	0	0	0	0
11	C11	0	0	0	0

The data consist of height measurements on a sample of 26 boys each measured on nine occasions between the ages of 11 and 14 years. C1 contains the individual ID, C2 the centred age, C3 height in cm, C4-C6 the quadratic, cubic and fourth degree terms of age. Variable 'OCC' indicates the measurement occasions corresponding to June, September, January, March, over three years. The measurements are approximately 0.25 years apart. Variable 'SEAS' is the month in decimals when measures were taken.

A simple 2-level model has already been fitted to the data; displayed by typing the commands `FIXE` and `RAND`.

PARAMETER	ESTIMATE	S. ERROR (U)	PREV. ESTIMATE
CONS	149	1.539	149
AGE	6.173	0.35	6.174
AGE2	1.128	0.348	1.128
AGE3	0.4542	0.1614	0.4542
AGE4	-0.3768	0.2983	-0.3768

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR (U)	PREV. ESTIM	CORR.	
2	CONS	/CONS	(13)	61.54	17.1	61.63	1
2	AGE	/CONS	(13)	7.992	3.022	8.005	0.614
2	AGE	/AGE	(13)	2.753	0.7799	2.755	1
2	AGE2	/CONS	(13)	1.36	1.416	1.367	0.216
2	AGE2	/AGE	(12)	0.8787	0.3437	0.8796	0.66
2	AGE2	/AGE2	(12)	0.6437	0.2278	0.644	1

We now model level-1 residuals by a continuous time model with $g = \alpha s$ only.

To setup the model, we type

`NAMES C2 'T'` (the macros require this name 'T' for the time variable)

We now need to set B10, C185 and C201 for the model. It is recommended to put a somehow large starting value in C201.

Type in the following commands in the **Command Interface** window.

<code>SET B10 1</code>	(turn on time series model's switch)
<code>JOIN C185 1 C185</code>	(put in value of the power term of s)
<code>JOIN C201 20 C201</code>	(put in starting value of α)
<code>SET B11 0</code>	(close switches for other models)
<code>SET B12 0</code>	
<code>SET B13 0</code>	
<code>MAXI 50</code>	(define upper limit for number of iterations)

You may like to confirm the model settings by obeying `OPTIONS.TS` again, then printing B10-B13, C185 and C201, then run the macros by click on **Start** button.

This iteration procedure stops after several iterations. One may find the parameter S1 in the random part is still not converged. Simply click on **More** twice to carry on two more iterations. Then type `FIXE` and `RAND` commands to display estimates as in the following **Output** window.

Note: it is worth pointing out that the starting value in C201 is important. Choose a value that is larger than the estimate you expect. You may find the warning message given by the program in the output window, 'ssp matrix has..... reconstruction used', and sometimes it may fail to converge. In this case go back to set up a new, larger starting value.

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE				
CONS	149	1.539	149				
T	6.185	0.3525	6.185				
AGE2	1.253	0.3773	1.253				
AGE3	0.433	0.174	0.433				
AGE4	-0.4954	0.3191	-0.4952				
LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.	
2	CONS	/CONS	(6)	61.48	17.07	61.48	1
2	T	/CONS	(6)	7.94	2.995	7.94	0.618
2	T	/T	(4)	2.681	0.7669	2.682	1
2	AGE2	/CONS	(3)	1.483	1.404	1.482	0.25
2	AGE2	/T	(4)	0.8547	0.3365	0.8548	0.695
2	AGE2	/AGE2	(3)	0.5641	0.2296	0.5643	1
2	T(0)	*	(3)	0.2496	0.04488	0.2495	
2	S1	*	(2)	7.189	2.248	7.202	
1	CONS	/CONS	(3)	0.2496	0.04488	0.2495	

Here both $T(0)$ at level-2 and the level-1 random part are for the same estimate of σ_e^2 , but constrained to be equal. We can ignore the $T(0)$ part and use the level-1 estimate to calculate the level-1 residuals and to check the model assumptions in the usual way. The estimate associated with S1 is that of α . Both estimates for $T(0)$ and S1 are not strictly level-2 random parameters, although they do appear in this section of the output for convenience.

We see that the estimated α is large compared with its standard error. We can work out autocorrelations at precise intervals of 0.25, 0.5, 0.75 and 1 by typing

```
JOIN C50 0.25 0.5 0.75 1 C50
CALC C51=expo(-7.189*C50)
PRINT C51
```

The estimated autocorrelations are 0.17, 0.027, 0.005 and 0.0007 in C51. We now fit a seasonal component with a year period in the fixed part of the model by typing

```
CALC C10=3.1416*'SEAS'/6
COS C10 C11
SIN C10 C12
NAME C11 'COS' C12 'SIN'
EXPL 1 'COS' 'SIN'
```

Click on the **More** button to resume iterations. At convergence we get following results.

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE				
CONS	148.9	1.539	148.9				
T	6.187	0.3511	6.187				
AGE2	2.138	0.4573	2.138				
AGE3	0.3842	0.1691	0.3842				
AGE4	-1.517	0.4425	-1.517				
COS	-0.2335	0.06784	-0.2335				
SIN	0.01503	0.05114	0.01502				
LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.	
2	CONS	/CONS	(10)	61.48	17.07	61.47	1
2	T	/CONS	(10)	7.93	2.988	7.923	0.618
2	T	/T	(8)	2.68	0.7638	2.674	
2	AGE2	/CONS	(1)	1.48	1.401	1.489	0.249
2	AGE2	/T	(8)	0.8523	0.3351	0.8501	0.687
2	AGE2	/AGE2	(2)	0.5746	0.2282	0.573	1
2	T(0)	*	(2)	0.2345	0.04347	0.2349	
2	S1	*	(1)	6.895	2.066	6.882	
1	CONS	/CONS	(2)	0.2345	0.04347	0.2349	

The $(-2 \times \log\text{-likelihood})$ value for this model is 611.435 using the `LIKE` command.

This model suggests a seasonal pattern of boys' height growth (see *Goldstein, Healy and Rasbash, 1994*, for details).

We can make more coefficients random at level-2 for this model. As the macros work, we have to clear the parameters $t(0)$ and `S1` at level two, then set these coefficients random at level-2, and remove the constraint between the estimates of $T(0)$ and the variance at level-1, and then restart. For example to allow the parameter of 'age3' to vary across boys, we do the following

```
CLRD 2 'T(0)'          (clear design variable at level-2)
CLRD 2 'S1'
RCON
SETV 2 'AGE3'
START
```

At convergence we obtain following random parameter estimates

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.
2	CONS/CONS	(5)	61.54	17.09	61.55	1
2	T/CONS	(4)	9.254	3.711	9.249	0.589
2	T/T	(2)	4.012	1.23	4.011	1
2	AGE2/CONS	(1)	1.379	1.403	1.373	0.222
2	AGE2/T	(4)	0.9587	0.4143	0.959	0.604
2	AGE2/AGE2	(2)	0.6288	0.2278	0.6298	1
2	AGE3/CONS	(1)	-1.701	1.5	-1.694	-0.408
2	AGE3/T	(2)	-0.9543	0.4905	-0.9533	-0.896
2	AGE3/AGE2	(1)	-0.1354	0.1689	-0.1344	-0.321
2	AGE3/AGE3	(2)	0.2825	0.2758	0.2825	1
2	T(0) *	(2)	0.196	0.03593	0.196	
2	S1 *	(1)	10.61	5.467	10.51	
1	CONS/CONS	(2)	0.196	0.03593	0.196	

The $-2 \times \log\text{-likelihood}$ for this model is 603.083. The difference between the likelihood values for the models is 8.355 with 4 degrees of freedom.

CPRO 8.355 4
0.079406

(probability for Chi-squared value and d.f.)
(display the probability)

This indicates little evidence for the variability of coefficient of 'age3' among boys, although the estimates of S1 and T(0) have changed.

To extend the model above for a higher order time series model, i.e. to change the g function, we need to retrieve the worksheet 'OXBOYS.WS2' and run 'OPTIONS.TS' to set-up a new model.

The extended time series model

To fit the model (3.3), we need to go back to retrieve the worksheet, and run 'OPTIONS.TS' to set-up the model. From the menu we must set B10=1, B11=1, B12=0 and B13=1, meanwhile entering corresponding values in C185, C201, C186 and C202. After that we then type START to run the model till convergence is achieved. The estimates do not converge for this dataset.

The model (3.4) is fitted to a simulated dataset based on the length (cm) growth of 0-2 years infant girls from two cities A and B. The dataset contains 120 individuals with 60 from each city measured two-monthly. A cubic polynomial growth curve with variable growth rates among girls has been assumed. The level-1 variance σ_e^2 is assumed to be 1 for convenience. Only a small difference between the growth curves and a rather weak difference of the autocorrelations are assumed between the two groups of infants. The model with population parameters is as follows.

$$y_{ij} = \sum_{p=0}^3 \beta_p t_{ij}^p + \sum_{p=0}^3 u_{pj} t_{ij}^p + e_{ij}$$

For city A $\beta = (75.0, 11.0, 1.0, 6)$ and for city B $\beta = (74.0, 12, -2, 5)$. The variance-covariance at level-2 are assumed the same for the two groups of girls, taking the following form,

$$\begin{pmatrix} \sigma_{u_0}^2 & & & & \\ \sigma_{u_{01}} & \sigma_{u_1}^2 & & & \\ \sigma_{u_{02}} & \sigma_{u_{12}} & \sigma_{u_2}^2 & & \\ \sigma_{u_{03}} & \sigma_{u_{13}} & \sigma_{u_{23}} & \sigma_{u_3}^2 & \end{pmatrix} = \begin{pmatrix} 5.5 & & & & \\ 0.7 & 2.0 & & & \\ -1.5 & 0.1 & 0.9 & & \\ 1.5 & -1.0 & -0.8 & 1.5 & \end{pmatrix}$$

The residuals term at level-1 is assumed to have autocorrelation coefficients 0.3 and 0.5 (or the covariances $\alpha_0=1.20$ and $\alpha_1=0.69$) for cities A and B separately for the equal time interval to be 1. For a two-weekly interval data of our example, we expect the covariances α s as around 7.2 and 4.2 respectively.

A single simulation carried out produced the example worksheet 'simu_lg.ws'. Opening the worksheet, we find that for a two-level model fitted to the data with no autocorrelation we have.

Name	n	missing	min	max
1 ID2	1560	0	1	160
2 ID1	1560	0	1	780
3 T	1560	0	-.1	1
4 AGE2	1560	0	0	1
5 AGE3	1560	0	-.1	1
6 HT	1560	0	55.05	102.4518
7 CONS	1560	0	1	1
8 CT-B	1560	0	0	1
9 B-T	1560	0	-.1	1
10 B-T2	1560	0	0	1
11 B-T3	1560	0	-.1	1
12 C12	0	0	0	0
13 C13	0	0	0	0

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE
CONS	75.63	0.3128	75.62
T	10.98	0.2404	10.98
AGE2	0.5647	0.1784	0.565
AGE3	6.283	0.2501	6.283
CT-B	-0.6611	0.4423	-0.6598
B-T	0.1024	0.34	0.1028
B-T2	0.571	0.2522	0.5707
B-T3	-0.256	0.3537	-0.2558

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.
2	CONS	/CONS (1)	5.746	0.7578	5.746	1
2	T	/CONS (1)	0.8663	0.4193	0.8665	0.226
2	T	/T (1)	2.555	0.4493	2.555	1
2	AGE2	/CONS (1)	-1.797	0.3549	-1.797	-0.622
2	AGE2	/T (1)	-0.3448	0.237	-0.3449	-0.179
2	AGE2	/AGE2 (1)	1.451	0.2473	1.451	1
2	AGE3	/CONS (1)	1.359	0.4461	1.359	0.386
2	AGE3	/T (1)	-1.571	0.4132	-1.572	-0.67
2	AGE3	/AGE2 (1)	-0.434	0.2476	-0.434	-0.246
2	AGE3	/AGE3 (1)	2.151	0.4894	2.151	1

1	CONS	/CONS (1)	0.707	0.03042	0.707
---	------	-----------	-------	---------	-------

The term 'T' in the worksheet is age in years and centred at year 1. The first four parameters, in the fixed part of the model, estimate the growth curve for individuals in city A. The next four parameters estimate the difference of the growth curves of the individuals between cities B and A. As expected there is little evidence for the difference of the growth curves between the two groups, and all growth coefficients are variable across individuals. However the level-1 variance is underestimated.

The following commands will fit the model (3.4) to this dataset.

OBEY OPTIONS.TS to bring up the following screen

```

NONLINEAR LEVEL-1 COVARIANCE OPTIONS (RELEASE 1.2)
=====
S (POLY.):B10=* - YES(1), POWERS (C185), START VALS. (C201)
T (POLY.):B11=* - YES(1), POWERS (C186), START VALS. (C202)
Z          :B12=* - YES(1), LINK FUNCTION VARIABLE COLUMNS IN (G9)
RELATIONS:B13=* - NONE(0), ADDITIVE(1), PRODUCT(2)
*=UNSPECIFIED

```

Now type

```

SET B10 1                (to switch on the model)
SET B11 0                (to switch off other models)
SET B12 1
SET B13 2
JOIN C185 1 C185
JOIN C201 15 C201
LINK 'CT-B' G9          (to link variable to group 9)

```

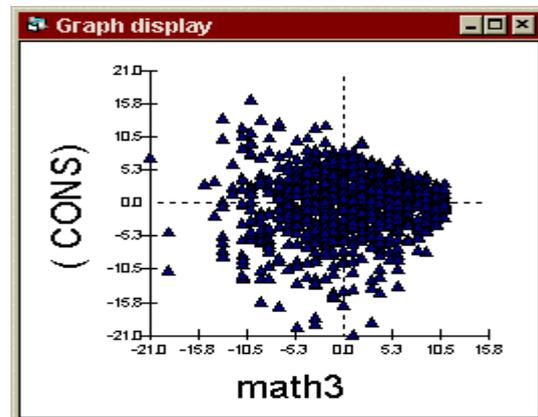
Now type **Start** 1 to estimate the model. At convergence the following model estimates are obtained. One may find the (NCONV) of random parameter estimate Z1 is 0 when the program stops iteration. In this case simply type **Next** 1 once or several times till (NCONV) of Z1 reaches 1.

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE				
CONS	75.63	0.3126	75.63				
T	10.96	0.2374	10.96				
AGE2	0.5546	0.1757	0.5546				
AGE3	6.312	0.244	6.312				
CT-B	-0.6617	0.442	-0.6615				
B-T	0.1337	0.3357	0.1336				
B-T2	0.5774	0.2484	0.5774				
B-T3	-0.3079	0.3451	-0.3077				
LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.	
2	CONS	/CONS	(7)	5.615	0.7575	5.615	1
2	T	/CONS	(8)	0.8815	0.4142	0.8814	0.276
2	T	/T	(5)	1.814	0.4722	1.815	1
2	AGE2	/CONS	(7)	-1.633	0.3518	-1.633	-0.659
2	AGE2	/T	(7)	-0.3562	0.2306	-0.3562	-0.253
2	AGE2	/AGE2	(5)	1.095	0.2526	1.095	1
2	AGE3	/CONS	(8)	1.33	0.435	1.33	0.515
2	AGE3	/T	(3)	-0.7897	0.4316	-0.7899	-0.538
2	AGE3	/AGE2	(5)	-0.4039	0.2376	-0.4039	-0.354
2	AGE3	/AGE3	(3)	1.186	0.5067	1.186	1
2	T(0)	*	(7)	0.8836	0.0652	0.8835	
2	S1	*	(5)	7.874	0.913	7.876	
2	Z1	*	(1)	0.05567	1.23	0.05582	
1	CONS	/CONS	(7)	0.8836	0.0652	0.8835	

With the time series structure in the model the level-1 variance is closer to 1 than before, although the estimated difference of covariances ($\hat{\alpha}_1 = 0.05564$) for girls from city B is not significantly different from those from city A ($\hat{\alpha}_0 = 7.87$). Other estimates are hardly changed.

A heterogeneous level-1 variance model

Open the worksheet 'JSP.WS2' (Woodhouse, 1996) for this exercise. A two-level model has been fitted to the data already. The level-1 residuals plotted against the initial score MATH3 (right) show a decreasing variance with increasing MATH3 score. We shall model the variance of the level-1 residuals as a negative exponential function of MATH3 score.



In the **Command Interface** window, type

```
OBEY OPTIONS.TS
```

The following screen is shown in the **Output** window

```

NONLINEAR LEVEL-1 COVARIANCE OPTIONS (RELEASE 1.2)
=====
S (POLY.):B10=* - YES(1), POWERS (C185), START VALS. (C201)
T (POLY.):B11=* - YES(1), POWERS (C186), START VALS. (C202)
Z          :B12=* - YES(1), LINK FUNCTION VARIABLE COLUMNS IN (G9)
RELATIONS:B13=* - NONE(0), ADDITIVE(1), PRODUCT(2)

*=UNSPECIFIED

```

Now type

```

SET B12 1          (to switch on the model)
SET B10 0          (to switch off other models)
SET B11 0
SET B13 0
LINK 'CONS' 'MATH3' g9      (to link variables in group 9)
MAXI 10
BATCh 1
START 1

```

The message 'CONVERGENCE NOT ACHIEVED YET' appears after two iterations. After 40 iterations the program has not converged due to a very small negative estimate for the variance of 'math3' at level-2 which in fact oscillates between iterations. To resolve this we leave out the math3 coefficient at level-2 from the model, then type NEXT 1 till convergence with the following results. You should not type START with the terms T(0) and Z1 in the model at this stage. This is because the command START begins the estimation from the first iteration when the macros will seek a single level-1 variance, then make a function of it to obtain a starting value of T(0) for later iterations. At this stage there is no such single variance at level-1, and the results will be unreliable. At convergence you will see a message 'macro procedure converged for this model' we now obtain

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE		
CONS	30.33	0.3545	30.33		
MATH3	0.5922	0.03278	0.5922		
NON_MAN	0.9464	0.3493	0.947		
SEX	-0.3824	0.3047	-0.3808		
LEV.	PARAMETER (NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.
2	CONS/CONS (0)	2.703	1.5	2.668	1
1	T(0)/T(0) (0)	3.249	0.2105	3.247	
1	Z1/T(0) (0)	-0.128	0.03563	-0.127	

The estimates of β_0^* and β_1^* are 3.249 and -0.1281 respectively. The NCONV column should be ignored because the convergence here is monitored by the macros rather than the program itself.

Some tips

- You can resume modelling by typing `NEXT 1` if only making changes to the fixed part or at level-2 or higher or leaving out some level-1 parameter terms. Otherwise it is recommended that you restart.
- If you do not get convergence with the time series models, restart with a larger starting value in C201.

Chapter 5. Time Series with discrete responses

Description of the model

In certain cases where there is a 2-level repeated measures structure with a response which is a proportion, the standard modelling procedure which assumes binomial variation at level 1, is inappropriate. For example in studies of voting behaviour many individuals will always vote for a particular political party or candidate on successive occasions. Their particular random effect will therefore lie at $\pm\infty$ so that the assumption of a Normal (or other finite distribution) is invalid. Barbosa and Goldstein (1999) give a discussion of such data and the following modelling procedure is based upon their approach.

We consider a two level repeated measures model where the response is discrete (see Goldstein, 1995 Chapter 7) as follows:

$$\pi_{ij} = f(X_{ij}\beta_j)$$

where π_{ij} is the expected value of the response for ij -th level 1 unit and f is a nonlinear function of the ‘linear predictor’ $X_{ij}\beta_j$. Here level 1 is occasion and level 2 is individual subject. The model is completed by specifying a distribution for the observed response $y_{ij}|\pi_{ij}$. At present the following macros will only deal with proportions or binary responses, but in principle the response can also be, say, a count with basic Poisson variation. The macros described here will work with responses which use a logit, log-log or probit link function. We write, for the logit link function

$$\pi_{ij} = \{1 + \exp(-[\beta_0 + \beta_1 x_{1ij} + u_{0j}])\}^{-1}$$

The observed responses y_{ij} are proportions with the standard assumption that they are binomially distributed, namely

$$y_{ij} \sim \text{Bin}(\pi_{ij}, n_{ij})$$

where n_{ij} is the denominator for the proportion. In most applications the response will be binary so that $n_{ij} = 1$.

Following Goldstein (1995) we can fit this into a standard multilevel framework by writing

$$y_{ij} = \pi_{ij} + e_{ij}z_{ij}, \quad z_{ij} = \sqrt{\pi_{ij}(1-\pi_{ij})/n_{ij}}, \quad \sigma_e^2 = 1$$

and we write the covariance between observations as

$$\text{cov}(y_{ij}, y_{(i+s)j} | \pi_{ij}, \pi_{(i+s)j}) = \sqrt{\pi_{ij}(1-\pi_{ij})\pi_{(i+s)j}(1-\pi_{(i+s)j})} f(s)$$

$$f(s) = \alpha + \exp(-h(\beta, z, s))$$

where s is the time between occasions (the lag) and $f(s)$ the correlation function between measurements s time units apart.

For further details see Barbosa & Goldstein (1999). This paper can be downloaded from the following web address (requires Adobe Acrobat reader)

http://www.ioe.ac.uk/hgoldstn/discrete_response_multilevel_models.pdf

Estimation

The estimation procedure in the macro DTS produces quasilielihood estimates of the model parameters (McCullagh and Nelder, 1989). MQL and PQL procedures are available.

Running the macros

In the **Options.dts** menu define the user directory in which you have installed the macros (see the introduction to the manual). The pre file and post file are respectively called PRE.DTS and POST.DTS. Typing the command `obey nlsett.dts` from the **Command Interface** window you will see the following screen:

```

NON-LINEAR SETTINGS (v. April 1999)
=====
ERROR DISTRIBUTION (B10)   : BINOMIAL(0)
APPROXIMATION(B11)       : FIRST ORDER(1)
NONLINEAR PREDICTION(B12) : FIXED PART ONLY:MQL(0)
LINK FUNCTION(B13)       : LOGIT (0)
VARIANCE FUNCTION(B14)   : DISTRIBUTIONAL(0)
DISCRETE TIME SERIES(B17) : YES(1)

```

The requirements are as follows

- B10 specifies the level 1 error assumption and must be set to zero (binomial);
- B11 specifies a first or second order approximation;
- B12 specifies MQL or PQL procedure;
- B13 specifies the link function;
- B14 specifies that the error distribution assumption for binomial variation is exactly satisfied (distributional), otherwise an under/over dispersion parameter will be estimated (unconstrained);
- B17 must be set to 1.

Example

Time series with discrete response model

The example data are in the worksheet 'BES.WS' supplied with MLwiN.

The data set consists of voting intentions for a sample taken from the British Election Study. Full details of the data set are given in Yang et al. (2000). In the present case we have measurements of voting intention together with a set of voter's attitudes taken in 1983, 1986 and 1987. In the present analysis we shall fit only average effects for the three years, without covariates. The data are at three levels; occasion, voter and voting area (constituency).

Open the worksheet 'BES.WS' and view the data from **Names** window.

	Name	n	missing	min	max
1	occ	3434	0	83	87
2	voter	3434	0	25	5999
3	area	3434	0	23	650
4	votecons	3434	0	0	1
5	cons	3434	0	1	1
6	83	3434	0	0	1
7	86	3434	0	0	1
8	87	3434	0	0	1
9	var(83)	3434	0	0	1
10	var(86)	3434	0	0	1
11	var(87)	3434	0	0	1
12	b_var	3434	0	1	1
13	t	3434	0	83	87
14	denom	3434	0	1	1
15	c15	0	0	0	0
16	c16	0	0	0	0
17	c17	0	0	0	0
18	c18	0	0	0	0
19	c19	0	0	0	0

'OCC' denotes the measurement occasion;

'VOTER' is the identification of the voter;

'AREA', is the area (constituency) code;

'VOTECONS' is the response variable and has value 1 if the respondent voted conservative and 0 otherwise;

'83', '86', '87' are dummy variables for each occasion;

'B_VAR', is the explanatory variable defining the level 1 variation. It is normally set to 1 and the macros require this named variable to be present;

'T' is the time variable, measured in years. The macros require this name for time.

'DENOM' is the denominator for the response proportion. In the present data this is 1.

To set up a 3-level model, where occasions are nested within voter and voters nested within area, type the following commands in the Command Interface window:

```
IDEN 1 'OCC'
IDEN 2 'VOTER'
IDEN 3 'AREA'
RESP 'VOTECONS'
EXPL 1 'CONS' '83' '86' '87' 'B_VAR'
FPAR 0 'CONS' 'B_VAR'
LINK 'B_VAR' G9
SETV 1 'B_VAR'
SETV 3 'CONS'
```

Make sure the **output** window is showing and then type SETT to display the following

```

EXPLAnatory variables in   CONS      83      86      87      B_VAR
FPARAmeters                83      86      87
RESPonse variable in      VOTECONS
FSDErrors : uncorrected          RSDErrors : uncorrected
MAXIterations 20  TOLerance   2    METHod is IGLS  BATCH is OFF
IDENTifying codes : 1-OCC, 2-VOTER, 3-AREA
LEVEL 3 RPM
      CONS
CONS      1
LEVEL 1 RPM
      B_VAR
B_VAR    1

```

We now need to specify the model to be fitted:

```

SET B10 0      (binomial error distribution)
SET B11 1      (first order approximation)
SET B12 0      (MQL prediction)
SET B13 0      (logit link function)
SET B14 0      (level 1 variance constrained to 1, i.e. assumes binomial variation)
SET B17 1      (mandatory)

```

Save your file now by typing, for example,

SAVE model0.ws

The macros require you to specify the autocorrelation function parameters to be specified in column c185 and starting values in C201. In the following table the currently available choices are listed, together with a sample command and an example of setting starting values for each parameter of 5.

Function $h(\beta, z, s)$	Command	Implementation
$\beta * s$	JOIN C185 1 C185 JOIN C201 5 C201	C185 has the power of s set to 1; C201 has a single starting value.
$\beta_0 + \beta_1 * s$	JOIN C185 0 1 C185 JOIN C201 5 5 C201	C185 has powers of s set to 0, 1; C201 has two starting values.
$\beta_0 * s + \beta_1 * s^{-1}$	JOIN C185 1 -1 C185 JOIN C201 5 5 C201	C185 has powers of s set to 1, -1; C201 has two starting values.

The last function is the most flexible and we are going to use it to model the data presented above. Type

```
JOIN C185 1 -1 C185
```

```
JOIN C201 5 5 C201
```

To run the estimation procedure type the command **start 1**. *Note that you can use the usual form of the start command without the argument 1, but this does not allow you to monitor the iterations one by one.* Do not run the model by clicking on the start button, if you do then MLwiN will fit the response as a Normal variable automatically.

If you type 'FIXE' when the single iteration is complete you will see the following;

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE
83	0.4056	0.01795	0
86	0.3212	0.01985	0
87	0.443	0.02108	0

Now type 'RAND' to obtain;

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.
3	CONS /CONS	(0)	0.01995	0.004227	0	1
1	B_VAR /B_VAR	(0)	0.2164	0.0245	0	

You may now type **next 1** for a single further iteration or **next** to iterate to convergence..

At about the 8th iteration convergence is achieved and the following estimates will be obtained (type 'FIXE' and 'RAND'):

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE
83	-0.3908	0.07143	-0.3909
86	-0.7818	0.07941	-0.7818
87	-0.3024	0.08036	-0.3024

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.
3	CONS /CONS	(2)	0.2672	0.06894	0.2682	1
2	T(0) *	(7)	1	1.427e-008	1	
2	S1 *	(1)	0.1092	0.006905	0.1084	
2	S2 *	(1)	0.3941	0.03177	0.3948	
1	B_VAR /B_VAR	(7)	1	1.816e-009	1	

'S1' is the estimate for the β_0 parameter and 'S2' is the estimate for the β_1 parameter.

Sometimes during iterations a 'numeric warning' window will appear. Click to continue and in the procedure will normally proceed to convergence.

To run the second order PQL procedure, with the binomial variance unconstrained (set B11=2, B12=1, B14=1), and type 'START 1', then 'NEXT 1' or 'NEXT'.

*Note that you must retrieve your worksheet (model0.ws in the present case) whenever you wish to change the model specification and then restart. Once this worksheet is open, go to the **Options/Directories** window to confirm the correct path by clicking on the **Done** button. This action will effectively activate the PRE.DTS and POST.DTS files.*

On convergence you will see the following estimates

PARAMETER	ESTIMATE	S. ERROR(U)	PREV. ESTIMATE		
83	-0.4219	0.07597	-0.4216		
86	-0.85	0.08401	-0.8495		
87	-0.3226	0.08479	-0.3224		

LEV.	PARAMETER	(NCONV)	ESTIMATE	S. ERROR(U)	PREV. ESTIM	CORR.
3	CONS /CONS	(1)	0.3335	0.07859	0.3329	1
2	T(0) *	(3)	0.971	0.02924	0.9704	
2	S1 *	(1)	0.1161	0.007575	0.1168	
2	S2 *	(2)	0.424	0.03496	0.4265	
1	B_VAR /B_VAR	(3)	0.971	0.02924	0.9704	

We see that the extra binomial parameter is close to 1 with no real evidence of underdispersion. We can compute the autocorrelations at lags $s = 1, 3, 4$ apart, using the function $(\beta_0 * s + \beta_1 * s^{-1})$ and the fitted values by typing

```
JOIN c60 1 3 4 c60
CALC c61=expo(-(0.1161*c60+0.424*c60^(-1)))
PRINT c61
```

The estimated autocorrelations are 0.58, 0.61 and 0.57.

References

- Barbosa, M. and Goldstein, H. (1999). Discrete Response Multilevel Models for Repeated Measures; an application to voting intentions data. *Submitted for publication*.
- Danahy, D.J., Buwell, D.T., Aranow, W.S. and Prakash, R (1977). Surfained hemodynamic and anti-anginal effect of high dose oral isosorbide dinitrate, *Circulation* **55**: 381-387.
- Goldstein, H. (1995) *Multilevel Statistical Models*, London: Edward Arnold; New York: Halsted Press.
- Goldstein, H., M. J. R. Healy, et al. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, **13**: 1643-55.
- Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne., Yang, M., Woodhouse, G. and Healy, M. (1998). *A user's guide to MLwiN*. London, Institute of Education.
- McCullagh P. & Nelder J.A. (1992). *Generalised Linear Models*, London: Chapman & Hall.
- Rasbash, J. and Woodhouse, G. (1996). *MLn Command Reference*, V1.0a, Multilevel Models Project, Institute of Education, University of London.
- Woodhouse, G. (1996). *Multilevel Modelling Applications, a Guide to users of MLn*, Multilevel Models Project, Institute of Education, University of London.
- Yang, M., Heath, A. and Goldstein, H. (2000). Multilevel models for repeated binary outcomes: attitudes and vote over the electoral cycle. *JRSS, A*:163, Part 1, 1-14.