TRAMSS

Teaching Resources and Materials for Social Scientists

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

- Who is this site for?

Anyone with an interest in data discovery and statistical analysis.

- What do I need to get something out this site?

About an hour or so and an understanding of multiple regression. Then you may wish to return to print off material or download data and software.

- So where will it take me?

The site provides a taste of statistical software applications in event history analysis and multilevel modelling.

- How will I learn?

You can learn to search the Data Archive's catalogue and then download software and data to run analyses. Examples are presented in a substantive framework with specially prepared datasets.

- Am I about to get lost?

Use the left-hand menu to explore the site. Typically pages are structured so that there are layers of information if you want to pursue any aspect of the site

- Feedback your experience.

Please take the time to let us know how you get on. Use the electronic form available under feedback.

Top

TRAMSS

Teaching Resources and Materials for Social Scientists

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

- Who is this site for?

Anyone with an interest in data discovery and statistical analysis.

- What do I need to get something out this site?

About an hour or so and an understanding of multiple regression. Then you may wish to return to print off material or download data and software.

- So where will it take me?

The site provides a taste of statistical software applications in event history analysis and multilevel modelling.

- How will I learn?

You can learn to search the Data Archive's catalogue and then download software and data to run analyses. Examples are presented in a substantive framework with specially prepared datasets.

- Am I about to get lost?

Use the left-hand menu to explore the site. Typically pages are structured so that there are layers of information if you want to pursue any aspect of the site

- Feedback your experience.

Please take the time to let us know how you get on. Use the electronic form available under feedback.

Top

Project overview

Home page
Project overview
Aims and objectives
Background
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

ESRC-ALCD Training & Dissemination Project *by Dick Wiggins*

Our focus

The focus of our target audience is full-time and part-time Master's students in quantitative social science research together with research students in ESRC related Departments. It is also expected that the appeal of the project will be broad enough to include professional social science researchers and young academics keen to develop their methodological skills and knowledge of data resources. A central aspect of the project is also to put exemplar analyses in a substantive context. You will be introduced to data sources and method via a series of research questions. Once data have been extracted the training materials should take users through standard analyses and encourage them to ask questions that will lead them to more complex analyses, and possibly deepen their reading of text and journals. Obviously, the flexibility of the medium itself also allows those of you with prior experience either of accessing data or complex analysis to find your own route and use of the material.

Use the left hand sub-menu to find out more

Top

Research questions

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Migration

What factors determine an individual's propensity to migrate? Do people tend to move at certain ages, at particular life events, for employment opportunities or as a response to external factors such as the economic climate or the housing market? Are there people who are likely never to move? How can we disentangle the three temporal effects: age, calendar year and duration of stay at one address?

Modelling migration histories example ▶
Exemplar search ▶

Youth

Changes in social and in economic policy and the expansion of post-compulsory education in Britain in the 1980s have had a marked effect on the opportunities available to young people on reaching minimum school leaving age. How do structural factors such as social class, ethnicity, gender and parental education influence young people's routes beyond school?

Post-compulsory education routes example ▶
Exemplar search ▶

Mortality

What is happening to mortality rates in England and Wales over time? How much variation in mortality rates is there between districts? Is this variation just between districts, or are there also differences between the mortality rates of counties? Does mortality vary according to the type of area? What is happening to the variation in mortality rates over time? Multilevel modelling

Mortality example ▶
Exemplar search ▶

Education

Do schools differ in the effects they have on pupil attainment? How can we fairly compare schools accounting for differences in pupil intakes? Do low ability pupils fare better when they are educated alongside higher ability pupils or are they discouraged and fare worse?

Education example ▶
Exemplar search ▶

Top

Searching for data

Home page
Project overview
Research questions
Searching for data
About the UK Data Archive's catalogue
How to use the online catalogue
Online catalogue exemplar searches
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Discovering Data for Secondary Analysis *by Hilary Beedham*

The UK Data Archive (UKDA) is a unique source of data for secondary analysis with several thousands of datasets in our holdings. This includes data collected under the original TRAMSS project which were deposited with us to enable the research and teaching community to benefit from the original investment in these key datasets.

One aim of this part of the TRAMSS project is to encourage increased use of these and other complex datasets that are available from UKDA. So, the original datasets used in the exemplars, which form part of this project, can be obtained from UKDA.

The following pages are designed to help users to search for data appropriate to their needs by offering training material for our online catalogue.

The following modules are available:

Information about UKDA's online catalogue ▶
This module provides a general introduction to the online catalogue with brief, descriptive information about the information it provides.

How to use the online catalogue▶
This module shows different ways of searching the online catalogue, the different fields that are available to search and tips on searching.

Exemplar searches using the online catalogue▶
These are sample searches that are associated with the research question and data modules that form part of these web pages. They demonstrate how to use our catalogue to find data that will answer specific research questions.

Help on Searching▶
Go directly to detailed help on searching the online catalogue.

The online catalogue▶
Allows direct entry to our web-based catalogue.

Next section: Information about the UK Data Archive's catalogue ▶

Top

Statistical modelling

TRAMSS

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Context for Learning
More statistical modelling
Methodological framework
Selected readings
Bibliography
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Statistical Modelling *By Dick Wiggins*

What are these pages for? To help put the statistical applications in a wider framework. To find out more you'll see each of the keywords below marked up on your sub-menu along the left hand side. They are:

- What is the context of this project?
- Can I learn more about statistical modelling?
- Putting the data analysis into a methodological framework.
- Journal articles for further illustrations of modelling.
- Bibliography . All text referred to in these pages.

Next section: Context for learning ▶

Top

Exemplar datasets

TRAMSS

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Migration data

The data are derived from a large retrospective survey of life and work histories carried out in 1986 under the Social Change and Economic Life Initiative (SCELI). The data contain the migration histories of 348 males aged 20 to 60, starting from the completion of education up to 1985. The data set is longitudinal, with one observation for each individual per calendar year. There are a total of 6349 annual observations. The response variable is binary, indicating for each individual for each year whether there was a migration move. The explanatory variables include age, calendar year, duration of stay at each address and information on family and work histories.

Modelling migration histories example ▶

Youth data

The data set contains a random sample of 800 young people taken from the Youth Cohort Study of England and Wales, Cohort 3. The data were collected three times at yearly intervals in the late 1980s when the young people were 16 to 19. There are therefore a total of 2400 annual observations. The response variable is a four category hierarchical outcome, indicating for each year whether the young person was in education, unemployed, in employment or training, or out of the labour market. The explanatory variables are educational attainment, gender, ethnicity, and parental social class and education.

Post-compulsory education routes example ▶

Education data

This data set contains GCSE exam scores on 4059 pupils in 65 inner london schools. The data are for pupils sitting GCSE exams in 1990. There are also intake ability measures on the pupils at age 11(entry into secondary school). Pupil gender and school gender(boys school, girls school or mixed school) are also recorded.

Education example ▶

Mortality data

The data are taken from the local mortality datapack and detail deaths from all causes in England and Wales in the period 1979 to 1992. The data comprise the Standardised Mortality Ratio (SMR) for each of 403 districts in the 54 counties of England and Wales, with one observation for each year from 1979 to 1992. (The SMR is the ratio of the observed number of deaths in an area to the number of deaths that would be expected if national age- and sex- specific death rates were applied to each area.) Our model therefore has three levels: years nested within districts in counties. We also have information on the classification of the district into one of six types: rural areas, prospering areas, maturer areas, urban centres, mining and industrial

Mortality example ▶

Top ↑

Analysis software

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
MLwiN
SABRE
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

SOFTWARE *by Brian Francis*

- The training materials provided on this web site are designed to be used in conjunction with the software packages SABRE and MLwiN .
- SABRE (software for the statistical analysis of binary recurrent events) is freeware and can also be downloaded from the SABRE web site. The version provided here is smaller than the standard freeware version and will run on most PCs.
- MLwiN is a commercial, licensed, Windows-based software package for fitting multilevel models - the special version of MLwiN provided here is free and fully-functional but works only with the teaching datasets provided.
- Both packages were developed and enhanced under the ESRC Analysis of Large and Complex Datasets initiative. The statistical software, together with the teaching datasets can be downloaded from the DOWNLOAD page - see the left menu.
- The download and installation of the software is straightforward. Once they have been installed, SABRE or MLwiN can be run from the START menu as with all other software. The datasets and software manuals are stored in the same directory as the software.
- Both software packages will run under WINDOWS 95, WINDOWS 98, WINDOWS NT 3.0 or 4.0 and WINDOWS 2000. They each require at least a 486 PC with 32Mb of memory or higher.
- While reading the web based training material, you can run the relevant software package in another window, reading in the relevant teaching datasets, and comparing your results with the results on the screen. Alternatively, you may prefer to print out the tutorials and work through the examples using the downloaded software and the printed tutorials.
- You can also have the opportunity to also try out your own analyses and to challenge the analyses provided by the site developers!

  For more details on each of the packages, see the left-hand menu.

Top

Download Data and software

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Download Analysis software

Please read the following information before proceeding to download the software at the bottom of this page.

- Special editions of statistical modelling software (MlwiN and SABRE) can be downloaded from this site.
- The software is free and can be downloaded with tutorials and exemplar data sets derived from the Archive.
- Before downloading you will be required to endorse an undertaking agreement.
- Tutorials provide a step by step guide to the principles of each modelling application.
- Each tutorial is framed by a number of substantive research questions.
- For more information about the analysis software click on the left sub-menu. Alternatively, move straight on to download.
- Please let us know how you get on with downloading and using the material provided. There is an electronic feedback form available under the 'feedback' option.

ACCESS AGREEMENT FOR USE OF DATA

The depositors of the data used as exemplar material for this ALCD project have generously waived the usual requirement for users to sign a written access agreement before accessing the data.

Users are nevertheless required to agree the following conditions before accessing the data:

This access agreement concerns the conditions of use of data and explanatory documentation supplied to me by The Data Archive. These data and explanatory documentation are hereafter referred to as `the materials' which will also include any additional data or explanatory documentation which are not the subject of a separate agreement.

I hereby undertake:

(1) Purpose: To use the materials only for the purposes of learning or teaching via the TRAMSS web site.

(2) Confidentiality: To act at all times so as to preserve the confidentiality of individuals and institutions recorded in the materials. In particular I undertake not to use or attempt to use the materials to derive information relating neither specifically to an identified individual or institution nor to claim to have done so[1].

(3) Acknowledgement: To acknowledge in any publication, whether printed, electronic or broadcast, based wholly or in part on such materials, both the original depositors and the Archive. The wording of the citation for individual datasets is to be

found in the documentation distributed by the Archive. To declare in any such work that those who carried out the original collection and analysis of the data bear no responsibility for their further analysis or interpretation. To acknowledge Copyright where appropriate.

(4) Access to others: Only to give access to others via the TRAMSS web site.

(5) Errors: To notify the Archive of any errors discovered in the materials.

(6) Liability: To accept that the Archive and the depositor of the materials supplied bear no legal responsibility for their accuracy or comprehensiveness.

1 *This clause does not apply to certain historical data which are based on sources which are in the public domain. Please check with the Archive for exceptions.*

I have read and agreed these conditions ▶

Top

The project team

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

The TRAMSS team

From left to right: Dick Wiggins *(City)*, Fred Smith *(Southamption)*, Jon Rasbash *(IOE)* Hilary Beedham *(Data Archive)*, Martin Han avy *(Data Archive)*, Juliet Harman *(Lancaster)* Alistair Leyland *(Glasgow)*
Crouching: Brian Francis *(Lancaster)*

Professor Fred Smith, second from the left, is the Director of the Analysis of Large and Complex Datasets programme of which the teaching material project is a part. Dick Wiggins has co-ordinated the project. Jon Rasbash and Alastair Leyland are responsible for the multilevel modelling material and Brian Francis and Juliet Harman have designed the event history analysis modules. Martin Hanavy has responsibility for the web site and Hilary Beedham has co-ordinated work at Essex and is responsible for the data finding modules.

If you make use of material from this site, please acknowledge the authorship as follows:

Wiggins, R.D., Beedham, H., Francis, B., Goldstein, H., Hanavy, M., Harman, J., Leyland, A., Musgrave, S., Rasbash, J. and Smith, A.F. (2000) *Teaching Resources And Materials for Social Scientists http://tramss.data-archive.ac.uk*

Top

Feedback

TRAMSS

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

This project is in its early phases of development. Your comments are extremely important. Please take a couple of minutes to provide some feedback. The response boxes provided below enable you to enter as much text as you require. Your views will remain anonymous.

Date	20/07/2010
What are the best aspects of this site?	
What are the worst aspects of this site?	
How would you improve this site?	
In your opinion, who is going to benefit ost from this site?	
Why did you decide to visit this site?	
Thanks for taking the time - please add your contact details and brief description of what you do if you would like to keep up to date with in future developments.	
Name	
Email address	
Telephone	
How would you describe yourself? -	
- As a postgraduate student Are you studying full / part-time? What is your main discipline ?	
- As someone with responsibility for training postgraduates What area of application ?	
-As an academic researcher In what context ?	
-As something else Please specify	

Submit details Reset form

Top

Search/route map

TRAMSS

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Search facility

This option takes you to the TRAMSS search facility where you can perform boolean searches based on keywords. Instructions are given.

Search web site ▶

Routemap

The routemap will open a separate window - a routemap - which will guide you through the website.

Route map ▶

Top

Project overview

TRAMSS

Home page
Project overview
Aims and objectives
Background
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Aims

To develop a web-based learning and teaching resource for quantitative social science researchers, students and trainers.

Objectives

- to provide both downloadable data from the Archive and free executable software in advanced statistical modelling
- to provide illustrations of exemplar analyses that have substantive meaning for social science researchers
- to place data, substance and method in a context which can be easily adapted to the needs of trainers and their students
- to promote the effective use of existing data resources
- to enhance the researchers ability to make sense of complex data
- to create a learning environment which is both flexible and responsive to the user's training needs

Top

Project overview

TRAMSS

Home page
Project overview
Aims and objectives
Background
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Background to the project *by Dick Wiggins*

This exciting new project has been set-up under the auspices of the ESRC's Analysis of Large and Complex Data (ALCD) programme to disseminate the research results of the programme to social science researchers in the form of a teaching resource.

Professor Fred Smith (University of Southampton), as programme co-ordinator, has invited a team of ALCD researchers to join forces with representatives of the Data Archive to develop a Web Site to enable both trainers and students of advanced applications of statistical modelling to access both data and computing power in a training environment. Materials are intended to provide a training experience to harness both aspects of large and complex social science datasets and the use of appropriate analytical tools. The broad aims of the project are to increase the productivity of research analysis and to expand the size of the research community able to exploit the potential in archived datasets.

At the heart of the project are two major software projects, SABRE and MLwiN , developed under ALCD at Lancaster (Centre for Applied Statistics and the Longitudinal Data Analysis Research Unit directed by Professor Richard Davies) and The Institute of Education (The Multilevel Modelling project directed by Professor Harvey Goldstein). Broadly, SABRE is a powerful tool for analysing longitudinal event history data (Dale and Davies, 1994) and MLwiN handles a range of data analytic applications wherever data is structured hierarchically (Goldstein, 1995). Representatives of these projects, Brian Francis at Lancaster and Alastair Leyland (of Glasgow University also MlM project) are working closely with members of the Data Archive (Hilary Beedham, Martin Hanavy and Rowan Currie) to develop the learning material. Dick Wiggins, formerly Director of Graduate Studies at the Social Statistics Research Unit at City University and now a member of the Department of Sociology at City is assisting in the co-ordination of the project and carrying out evaluations with potential users.

Top

Modelling Migration Histories

Juliet Harman, Brian Francis and Richard Davies
Centre for Applied Statistics, Lancaster University

The main substantive questions

1. **Are some people more likely to move than others?**
What factors determine an individual's propensity to migrate? Are there people who are likely never to move?

2. **Does an individual's migration behaviour vary with time?**
Do people tend to move at certain ages, at particular life events (marriage, children, schooling), for employment opportunities, or as a response to external factors such as the economic climate or the housing market?

3. **How can we separate different temporal effects?**
Differing patterns of migration behaviour with age are likely for different birth cohorts, as individual life histories take place in different and changing economic conditions. Cumulative inertia effects (the increasing tendency to stay as length of residence in the same place increases) may complicate the variation of migration propensity with age. How can we disentangle the three temporal effects: age, calendar year and duration of stay?

What data set is analysed?

- To address these substantive questions, we need a data set on each of a large number on individuals, with information for each individual on their *migration* history, their *marital* history, their *employment* history and their *family* history.
- Such historical information is needed from the start of each individual's adult life until the date of data collection.
- We can use the UK Data Archive online catalogue to find a suitable data set. An example has been constructed on how to search for such a data set on migration.
- The data set chosen is a large retrospective survey of life and work histories carried out in 1986 under the Social Change and Economic Life Initiative (SCELI), funded by the ESRC.

Will I understand this module?

- We assume that you have a certain amount of statistical knowledge already. The most important requirement is to be able to understand the output of a multiple regression. A basic knowledge of logistic regression and Poisson regression (regression models for count data) would also be useful, but this is not essential. We provide an explanation of new technical terms, and explain results through the use of graphs.

Give me a quick overview of this module

- We first analyse a summary data set containing the total number of moves for each individual, and demonstrate the limitations of such cross-sectional analysis for drawing inference about the dynamics of migration.

- We then explore the longitudinal data set containing the life and work histories, and model the annual binary migration data using a conventional logistic model. We discuss the limitations of using conventional models for longitudinal data and demonstrate the importance of controlling for individual specific explanatory variables omitted from the analysis.

What software do I need?

- You will need to use SABRE, which is a statistical software package for the analysis of discrete longitudinal data. SABRE runs on all Windows machines and also on UNIX and Linux platforms. SABRE and the teaching data sets can be downloaded from here free of charge.

- SABRE is a specialist package, with a restricted range of commands; it has no facility for instance to plot graphs. However, the parameter estimates from model fitting can be copied into other packages. We use the statistical package GLIM to supplement SABRE.

How do I use this module?

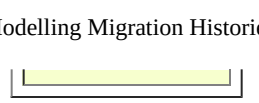
The best way is to follow the module page by page on the Web, loading the data set into SABRE in a new window, and following the instructions onscreen. Alternatively, it is possible to download the entire module as an ADOBE portable document file.

Acknowledgement

This example is based on research work carried out by R. B. Davies and R. Flowerdew (1992) and by Haghighi A. Borhani and R. B. Davies (1999a, 1999b), using data collected under the Social Change and Economic Life Initiative funded by the ESRC. The work by Haghighi Borhani and Davies was partially supported by ESRC research grant L315253007.

NEXT:Table of contents

Home page



Exemplar searches

Searching for data

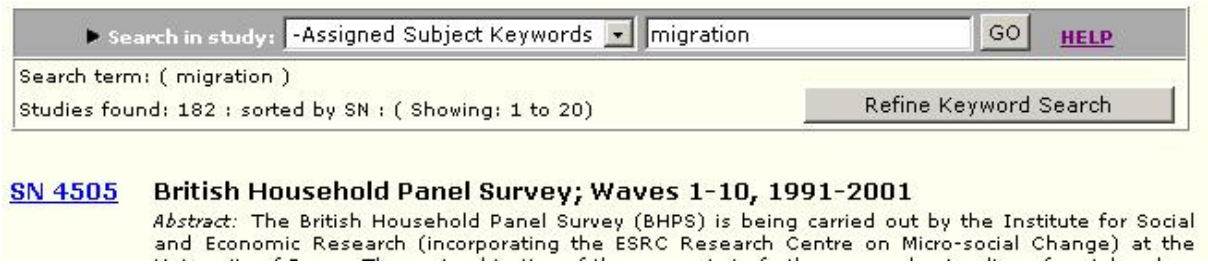
*If you would like to open a separate browser window to view the online catalogue whilst reading these notes **click here**.*

A sample search: UK Family Migration - Geography and Planning

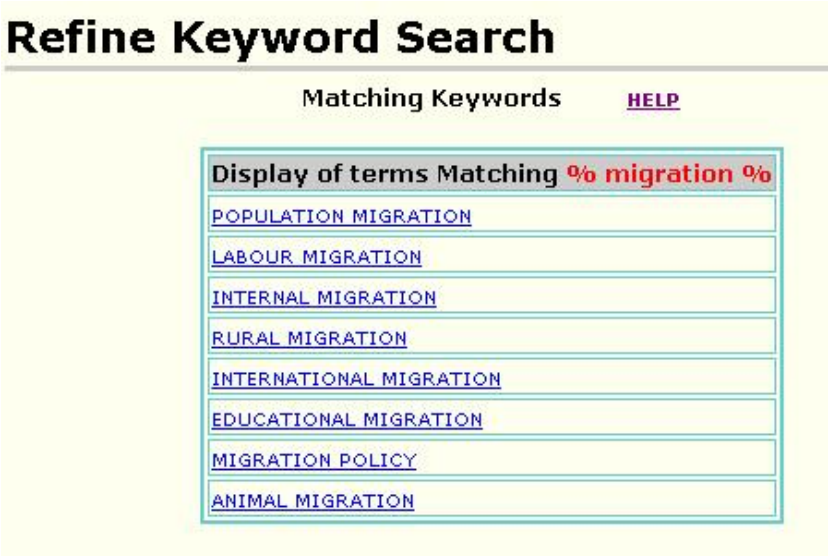
There follows a sample search for labour migration, using the assigned subject keyword search.

From the drop down list select 'Assigned Subject Keyword', type in 'Migration' and click on 'Go' (or use the Enter key).

This returns a list of over 180 studies.

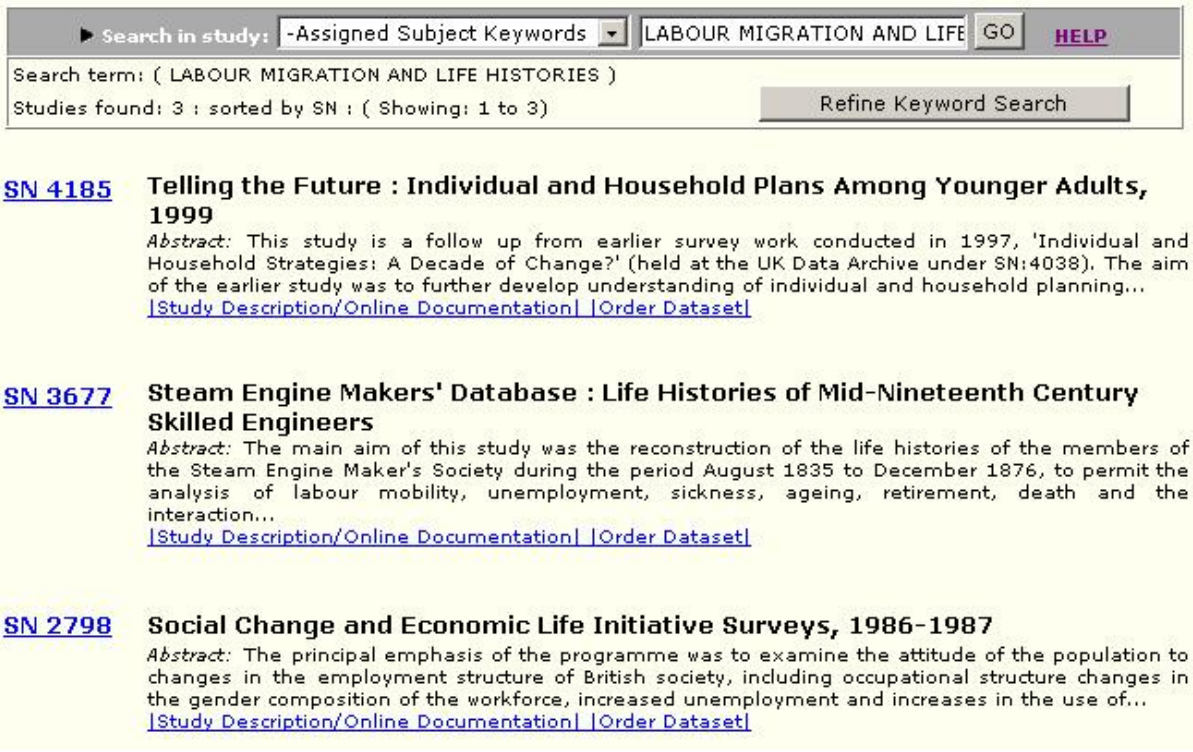


To give a more focused search, click on the 'Refine Keyword Search' button. This takes you to the UKDA thesaurus HASSET where you can browse terms to refine your search. In this instance, select 'Labour Migration' from the HASSET list and then click on 'Search on Keyword'.



This returns a list of over 70 studies which is still rather a lot of datasets to check for content. We can therefore narrow the search by combining our search with an additional keyword. From the list of search results, at the top of the screen add the keyword 'Life histories' to your search to give 'Labour migration AND Life histories'.

You now have 3 datasets to choose from:



Top

Modelling young people's post-compulsory education routes

Juliet Harman and Damon Berridge
Centre for Applied Statistics, Lancaster University

The sociological context

Changes in social and in economic policy and the expansion of post-compulsory education provision in Britain in the 1980s have had a marked effect on the opportunities available to young people on reaching the end of their period of compulsory education (Maguire and Maguire, 1997; MacDonald, 1999).

The collapse of the youth labour market, the introduction of youth training schemes, changes in state benefits together with the expansion of further education have encouraged 16 to 19 year olds to stay in education longer.

In 1973/74 33% of male and 37% of female 16 year olds remained in full- time education, compared to 70% of males and 76% of females in 1993/94 (Furlong and Cartmel 1997).

There is a debate amongst sociologists between those who hold that social divisions are of declining significance and that it is individual action that determines young people's pathways beyond school (Chisholm *et al.*, 1990; Beck, 1992), and a more orthodox sociological perspective which sees life experiences and aspirations shaped by social class and family background (Banks *et al.*, 1992; Jones and Wallace, 1992; Furlong and Cartmel, 1997).

In this analysis we aim to gain an understanding of the factors which influence young people's routes after they reach minimum school leaving age. Clearly educational attainment is a most important factor.

The main substantive question

* How do structural factors such as ethnicity, social class, gender and parental education affect a young person's experiences of the school to work transition in Britain near the end of the 20th century?

What data set is analysed?

To address this question, we need data on a large number of young people aged 16 and over, with information for each individual on their *educational* history, their *work* history and *family* and *demographic* information.

We can use the UK Data Archive online catalogue to find a suitable data set. An example has been constructed on how to search for such a data set on youth.

The data set chosen is derived from the Youth Cohort Study of England and Wales (YCS); a longitudinal study of young people's experiences as they complete their period of compulsory education and enter further education, training or employment.

Will I understand this module?

We assume that you have a certain amount of statistical knowledge already. The most important requirement is to be able to understand the output of a multiple regression. A basic knowledge of logistic regression would also be useful, but this is not essential. New technical terms are explained and the results of analyses are interpreted.

Give me a quick overview of this module

Young people's choices when they reach minimum school leaving age and either continue in education or enter the labour market may be seen as a decision tree with a number of branching points, with a hierarchical order.

Such hierarchical (or ordered) outcomes can be modelled using the **continuation ratio model** (Fienberg and Mason, 1979).

First we explain the continuation ratio model and demonstrate its application using a small cross-sectional data set containing ordered data on young people's educational attainment.

We then use the continuation ratio model to analyse the longitudinal YCS data, which contains hierarchical outcomes measured repeatedly over time for each individual. We examine which explanatory variables influence young people's activities during the teenage years after they leave school.

Conventional modelling approaches do not allow for the unmeasured and possibly unmeasurable factors which may account for the possible large variations between individuals (residual population heterogeneity). Fitting conventional models may therefore lead to biased results. We explain the importance of allowing for *population heterogeneity* and fit random effects (mixture) models to the longitudinal data.

What software do I need?

You will need to use SABRE, which is a statistical software package for the analysis of discrete longitudinal data. SABRE runs on all Windows machines and also on UNIX and Linux platforms. SABRE and the teaching data sets can be downloaded from here free of charge.

SABRE is a specialist package, with a restricted range of commands. We use the statistical package GLIM for the initial analysis of aggregate tabular data.

How do I use this module?

The best way is to follow the module page by page on the Web, loading the data set into SABRE in a separate window and following the instructions on screen. Alternatively, it is possible to download the entire module as an ADOBE portable document file.

Acknowledgement

This example is based on research work carried out by V. Gayle (1996, 1998) and by V. Gayle, D. M. Berridge and R. B. Davies (1999). Dr Vernon Gayle carried out the work for the most recent paper as part of an ESRC ALCD Phase II Visiting Fellowship (Award H519 44 5002 97).

NEXT:Contents of module

Home page

Exemplar searches

Searching for data

Home page
Project overview
Research questions
Searching for data
About the UK Data Archive's catalogue
How to use the online catalogue
Online catalogue exemplar searches
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

*If you would like to open a separate browser window to view the online catalogue whilst reading these notes **click here**.*

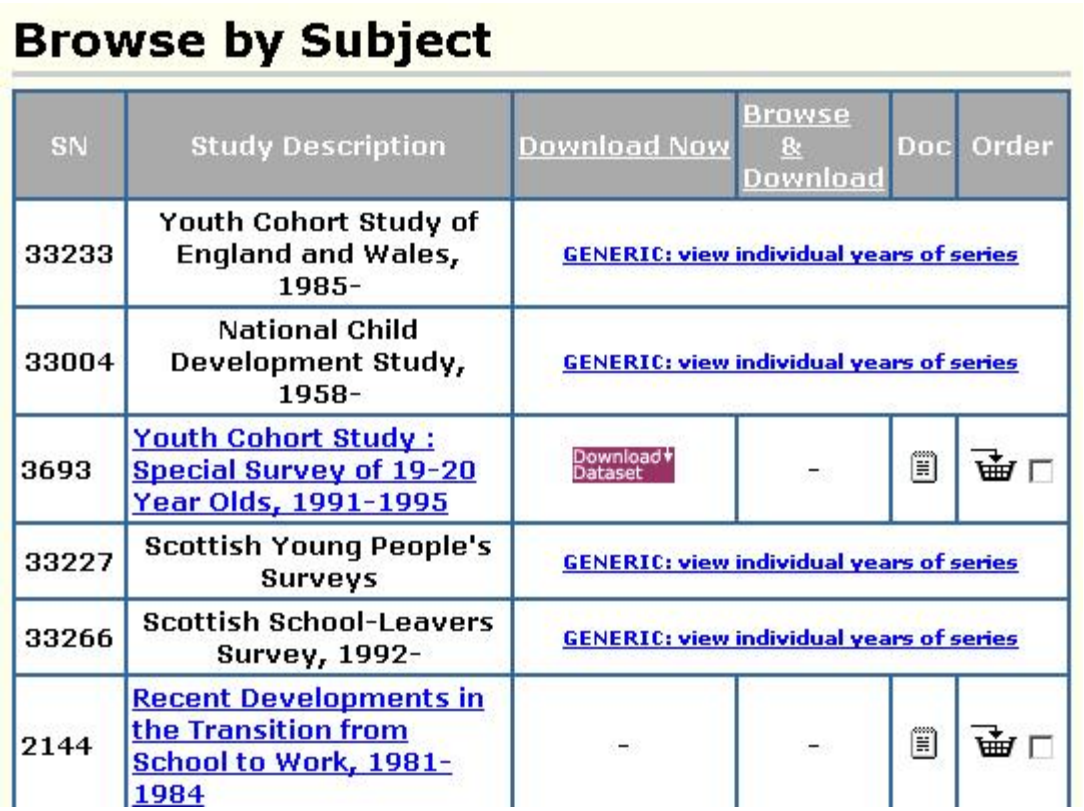
A sample search: Continuation in post-compulsory education.

There follows a sample search for data to help you answer the research question on continuation in post-compulsory education.

From the search catalogue web page, click on 'browsing by subject category'.

Click on 'Education' to show the expanded list of terms in this section and check the box next to 'School leaving'. If the box next to 'Education' were checked a search would be carried out on ALL subsections of 'Education'. Click on 'Go':

From the list of search results, take a look at some of the study descriptions to see which study best lends itself to answering the research question.



Browse by Subject

SN	Study Description	Download Now	Browse & Download	Doc	Order
33233	Youth Cohort Study of England and Wales, 1985-	GENERIC: view individual years of series			
33004	National Child Development Study, 1958-	GENERIC: view individual years of series			
3693	Youth Cohort Study : Special Survey of 19-20 Year Olds, 1991-1995	Download Dataset	-		
33227	Scottish Young People's Surveys	GENERIC: view individual years of series			
33266	Scottish School-Leavers Survey, 1992-	GENERIC: view individual years of series			
2144	Recent Developments in the Transition from School to Work, 1981-1984	-	-		

We can also explore the contents of the dataset further by clicking the 'Online Documentation' link at the top of the catalogue record or by scrolling down to the bottom of the record where we can view and download the User Guides in Adobe Acrobat PDF format.

Top

MLwiN - Mortality Overview

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Education Overview
Mortality Overview
Tutorials
Software
Back to main site

Mortality in England and Wales, 1979-1992

Alastair H Leyland and Alice McLeod

MRC Social and Public Health Sciences Unit, University of Glasgow

Introduction to the dataset

The data are taken from the local mortality datapack and detail deaths from all causes in England and Wales in the period 1979 to 1992. The full dataset is stored at the Data Archive at the University of Essex.

The data comprise the Standardised Mortality Ratio (SMR) for each of 403 districts in the 54 counties of England and Wales, with one observation for each year from 1979 to 1992. (The SMR is the ratio of the observed number of deaths in an area to the number of deaths that would be expected if national age- and sex- specific death rates were applied to each area.) Our model therefore has three levels: observations are made on ***years*** nested within ***districts*** in ***counties***. We also have information on the classification of the district into one of six types: rural areas, prospering areas, maturer areas, urban centres, mining and industrial areas, and inner London.

Research questions

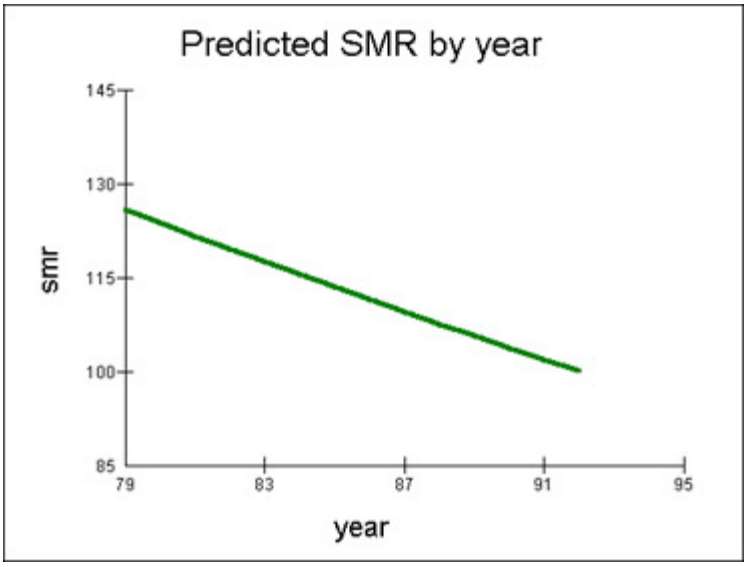
The full tutorial addresses the following research questions:

1. What is happening to mortality rates over time?
2. How much variation in mortality rates is there between districts of England and Wales?
3. Is this variation just between districts, or are there also differences between the mortality rates of counties?
4. Does mortality vary according to the type of area?
5. What is happening to the variation in mortality rates over time?

The full tutorial takes the user through the detailed analysis of the data set using

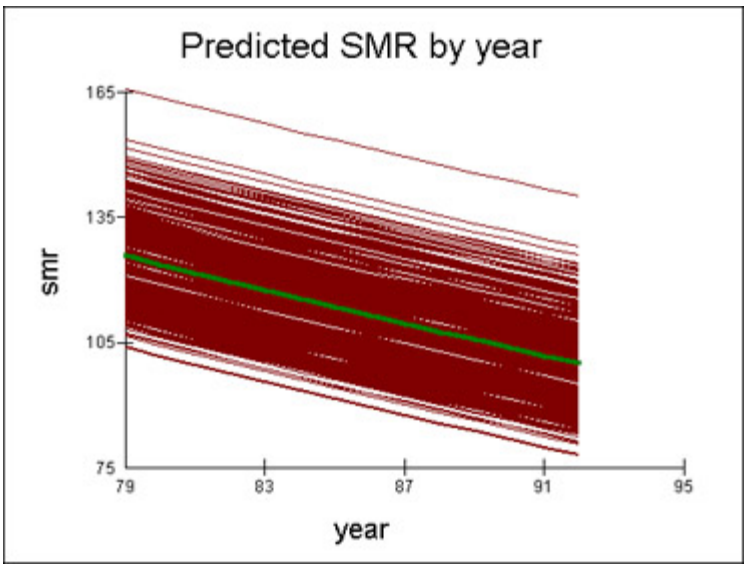
MLwiN ; for short answers to any of the above, illustrated using the output from the detailed analysis, see below.

What is happening to mortality rates over time?



The plot above shows the average predicted trend in the standardised mortality ratio (SMR) for all districts in England and Wales. SMRs have been decreasing over this time; the average decrease in SMR was approximately 2 points per year, from about 126 in 1979 to 100 in 1991.

How much variation in mortality rates is there between districts of England and Wales?



The above graph shows the predicted SMR for each district, under the assumption that the reduction in the SMR in each district is the same. That is, the predicted

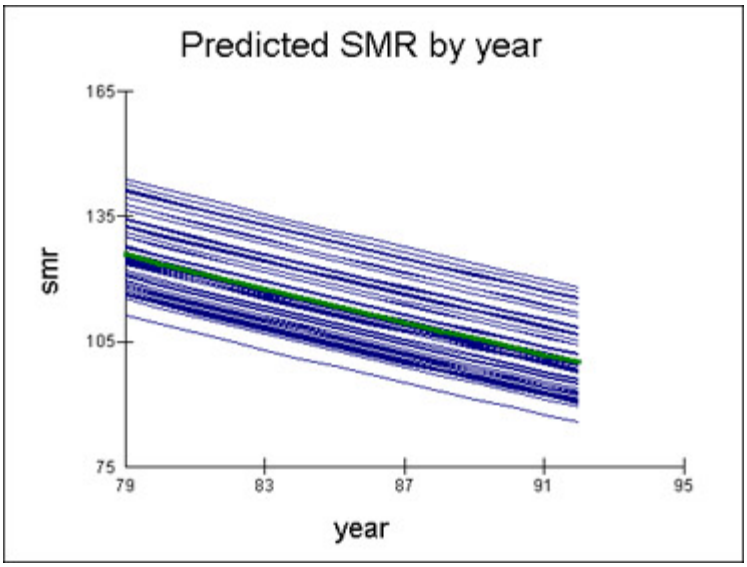
SMR for each district – show in red – is parallel to the average for England and Wales – shown in green. Thus, for example, the district with the highest SMR throughout this period (the uppermost line in the graph) always has an SMR about 40 points above the average – from 166 in 1979 to 140 in 1991. The assumption that all of these lines should be parallel may not be valid – we can test this later – but this means that the variation of districts around the mean is the same every year. We can quantify this variance; our estimate is 112.9.

However, this is only part of the picture. The graph above shows the *predicted* mortality for each district; the actual observations made in each year within each district will fluctuate around the district means. This gives rise to a second variance, quantifying the variation between years within districts. Our estimate of this variance is 24.5.

The total variance therefore has two components, one at each level of our multilevel analysis. There is year-on-year variation within districts in addition to variation between districts. The variation between districts accounts for about 82% of the total variation

$$\left(\frac{112.9}{112.9+24.5}=0.82\right).$$

Is this variation just between districts, or are there also differences between the mortality rates of counties?



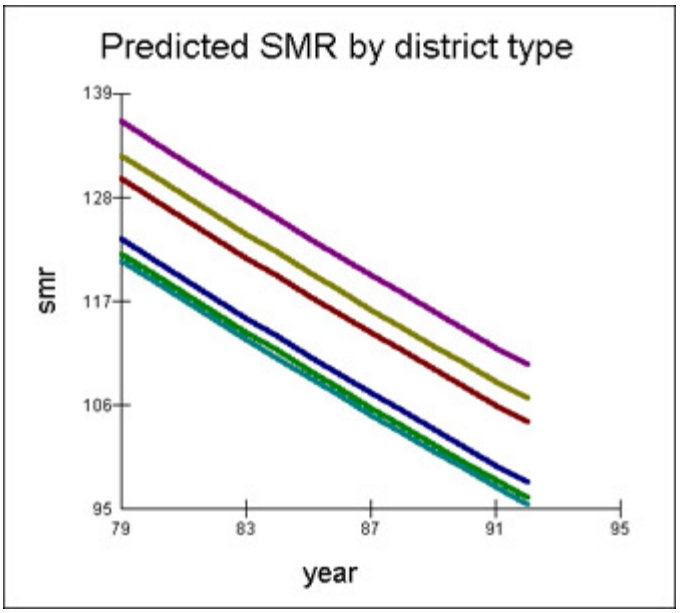
The above graph shows the predicted SMR for each county and indicates that there is indeed variation between counties as well as between districts and from

one year to another. Whilst the estimated variance between years within districts remains unchanged at 24.5, the higher level variance is partitioned; we estimate a variance of 75.8 between counties and of 42.9 between districts within counties. Leaving aside the apparently random fluctuations from one year to the next, we therefore find that 64%

$$\left(\frac{75.8}{75.8+42.9}=0.64\right)$$

of the total higher level variation between districts is in fact due to different mortality rates between the larger geographic units of counties.

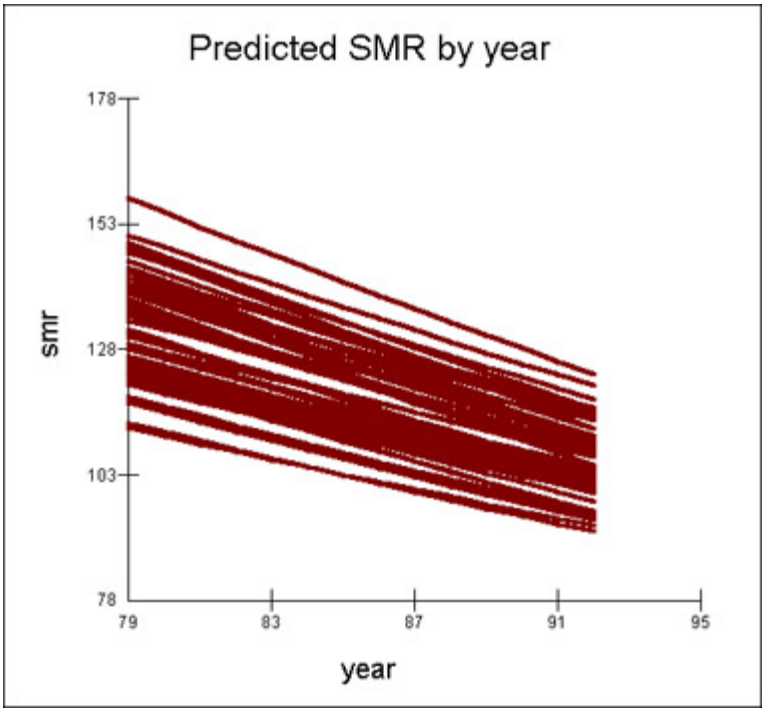
Does mortality vary according to the type of area?



The above graph shows the mean predicted SMR for six different types of district; starting from the top these are mining and industrial areas, inner London, urban centres, rural areas, prospering areas and maturer areas respectively. There are substantial differences between the types of area, with the SMR in the mining and industrial areas being on average nearly 15 points higher than in the maturer areas.

The introduction of the area classifications does nothing to alter the year-on-year variation in the SMRs. However, the between district and between county variances decrease by 29.4% and 52.5% respectively. The inclusion of this *district* level classification therefore has the greatest impact upon the variation between *counties*. This is not altogether surprising given the tendency for districts of the same type to cluster together within counties; for example, all districts of the classification "inner London" will lie within the same county (London).

What is happening to the variation in mortality rates over time?



The above graph shows how the predicted SMRs vary over time for urban districts. The slopes of counties and districts have been allowed to vary – in other words, the lines in the graph above are no longer parallel (and some of them cross). Different counties and districts have experienced differential rates of decrease in their SMRs between 1979 and 1992. You may note that the overall variation has been decreasing over time – put crudely, the predicted SMRs are more "spread out" in 1979 than they are in 1992. Leaving aside the year-on-year variation, the total geographical variation (i.e. the sum of the district and county level variances) has decreased by approximately 50% as mortality rates have fallen.

A more detailed step-by-step explanation of this example is available as a tutorial.

Top

Exemplar searches

Searching for data

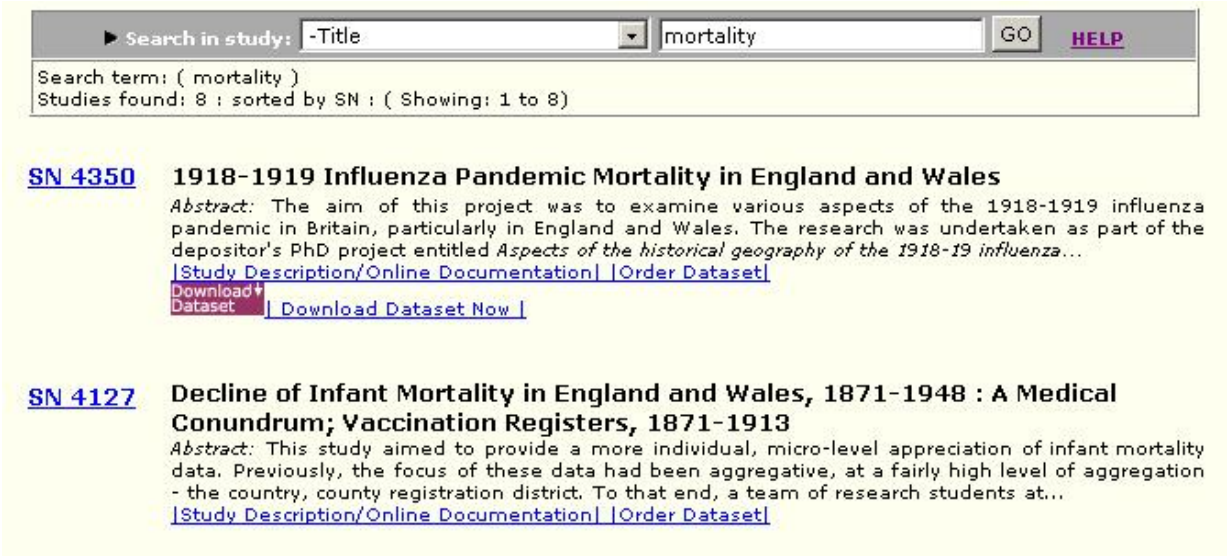
If you would like to open a separate browser window to view the online catalogue whilst reading these notes ***click here****.*

A sample search: Mortality data

There follows a sample title search for mortality data.

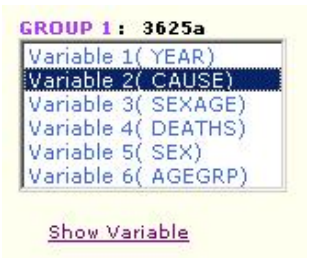
From the drop down list select 'Title', type in 'Mortality' and click on 'Go'.

Explore the contents of some of the datasets in the list of search results. To go into an individual catalogue record click on the blue 'SN****' (where **** is the study number) or 'Study Description...' hyperlinks.



If you scroll down the list of search results and look at the catalogue record for 'SN 3625 Local Mortality Datapack: Population and Deaths by Cause, 1979-1992', it should be clear that this dataset will enable us to answer our research questions; populations and counts of deaths are available by sex and for 5 year age groups for county districts in England and Wales over a 13 year period.

Now let's look at some variables: click on the 'Variable List' link at the top of the catalogue record. Each Group corresponds to a data file. In 'Group 1', highlight 'Variable 2 (CAUSE)' and click on 'Show Variable':



This shows you the value labels:

Value Labels	
Value 1	ALL CAUSES OF DEATH (001-999)
Value 2	INFECTIOUS AND PARASITIC DISEASES (001-139)
Value 3	INTESTINAL INFECTIOUS DISEASES (001-009)
Value 4	TUBERCULOSIS (ALL FORMS) (EXCLUDING LATE EFFECTS) (010-018)
Value 5	TUBERCULOSIS, RESPIRATORY (010-012)
Value 6	WHOOPING COUGH (033)
Value 7	STREPTOCOCCAL SORE THROAT, SCARLATINA & ERYSIPELAS (034,035)
Value 8	MENINGOCOCCAL INFECTION (036)
Value 9	SEPTICAEMIA (038)
Value 10	MEASLES (055)
Value 11	MALARIA (084)
Value 12	LATE EFFECTS OF TUBERCULOSIS (137)
Value 13	NEOPLASMS (140-239)
Value 14	MALIGNANT NEOPLASMS (140-208)
Value 15	LIP, ORAL CAVITY AND PHARYNX (140-149)
Value 16	DIGESTIVE ORGANS AND PERITONEUM (150-159)
Value 17	OESOPHAGUS (150)
Value 18	STOMACH (151)
Value 19	SMALL AND LARGE INTESTINE (EXCL. RECTUM) (152-153)
Value 20	COLON (153)
Value 21	RECTUM, RECTOSIGMOID JUNCTION AND ANUS (154)
Value 22	LIVER, GALL BLADDER AND BILE DUCTS (155-156)
Value 23	PANCREAS (157)
Value 24	LARYNX (161)

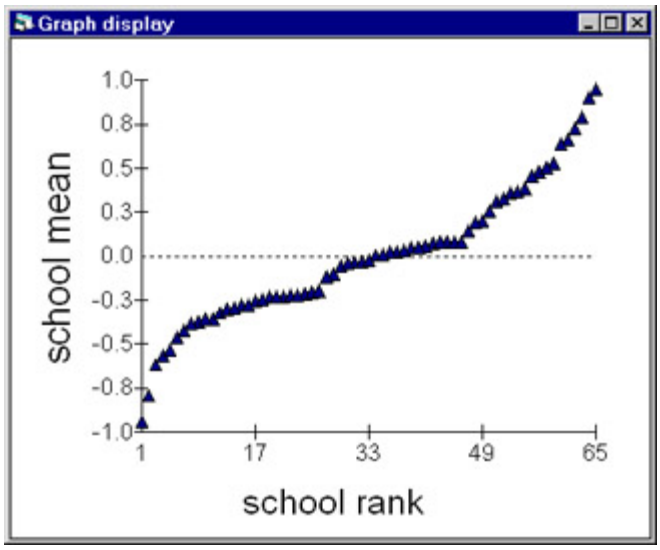
Top

MLwiN - Education overview

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Education Overview
Mortality Overview
Tutorials
Software
Back to main site

DO SCHOOLS DIFFER?

We can plot the school means out against their ranks - a graphical league table :

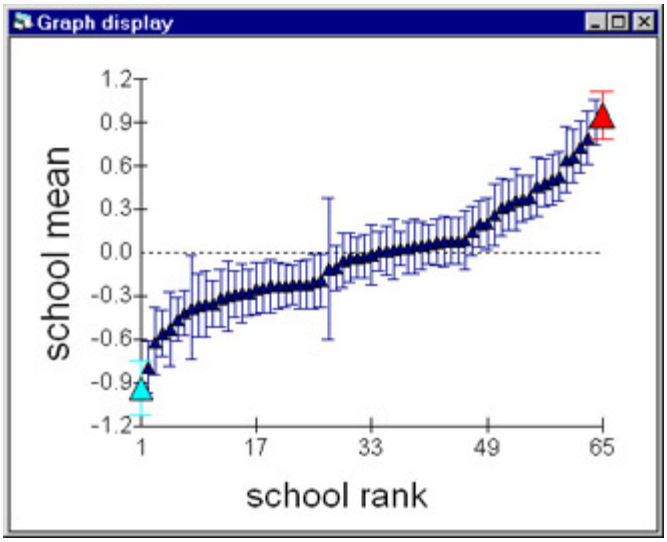


Our response variable has bean normalised. Therefore the difference between the highest school mean and the lowest school mean is 2 standard deviations.

From this graph it appears that different schools do have very different effects.

ADDING UNCERTAINTY INTERVALS AROUND THE ESTIMATES

We should put confidence intervals around our estimates of the means :



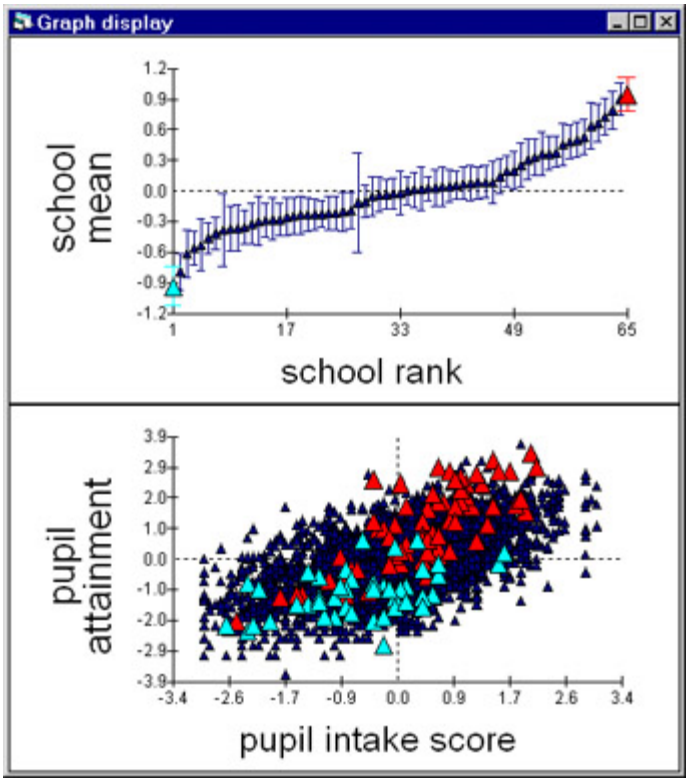
Even taking account of sampling error there are large statistically significant differences between the school means.

The shool with the lowest mean(school B) is highlighted in blue and the school

with the highest mean(school A) is higlighted in red.

WHAT ABOUT INTAKE ABILITY?

We also have data on tests children took when they entered secondary school at age 12. That is we have a measure of intake ability.



The lower graph plots pupil level attainment against pupil level intake score. There are 4000 pupils and therefore 4000 points on the graph. Pupils from the school A are picked out in red and pupils from the school B are picked out in blue.

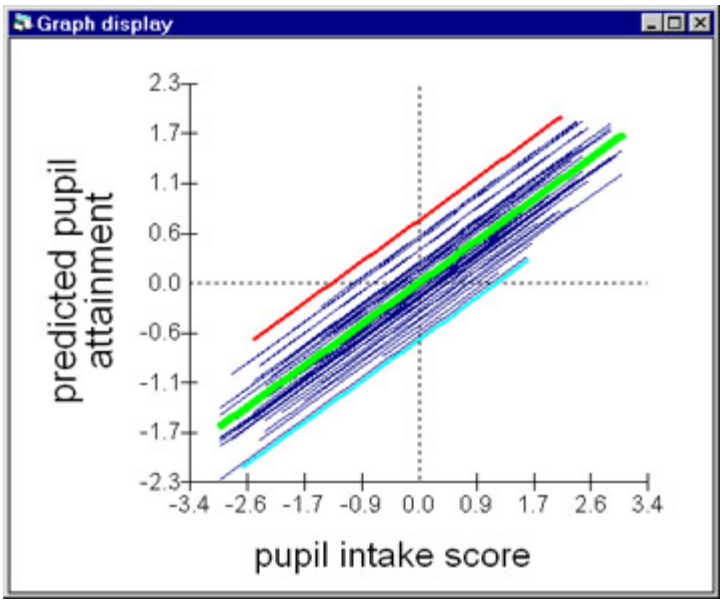
We see that there are more pupils in school A with high intake scores. School A attracted more able pupils than school B which must contribute to its higher outcome mean.

Rather than looking at raw(unajusted) school means we should be adjusting our model for the school's intake. We will then be looking at progress pupils make while attending a school. This is a more meaningful measure of school effectiveness.

ADJUSTING FOR INTAKE ABILITY

We can adjust for intake ability by regressing pupil attainment on pupil intake score. The model becomes multilevel because we allow each of our 65 schools to depart(be raised or lowered) from the overall regression line. These school level departures are known as school level residuals and can be thought of as a measure of the effect of the school.

The results of the model are illustrated in the graph below.



The central green line is the regression based on all pupils from all schools, from which the 65 school lines depart.

The equation of this line is

Predicted attainment = 0.092 + 0.566* *intake*

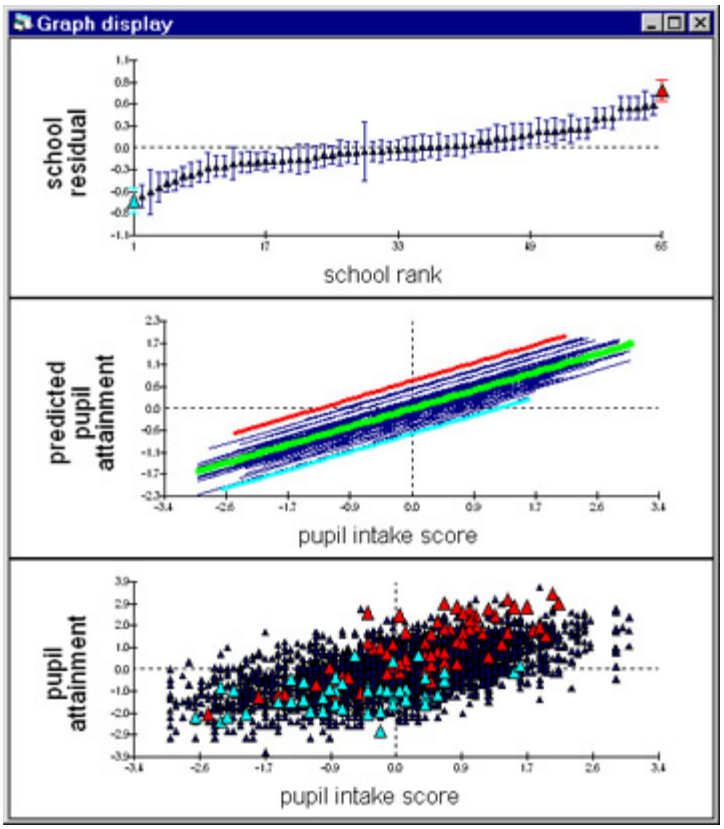
That is on average an increase of 1 unit of intake score results in an increase 0.566 units in outcome attainment.

We can see that even adjusting for intake score School A has the largest positive residual, its line is at the top, and school B has the largest negative residual, its line being at the bottom.

Can we now say that having adjusted for pupil intake ability School A is more effective than school B?

ADJUSTING FOR INTAKE ABILITY – A CLOSER INSPECTION

The graphs below reveal some interesting patterns:



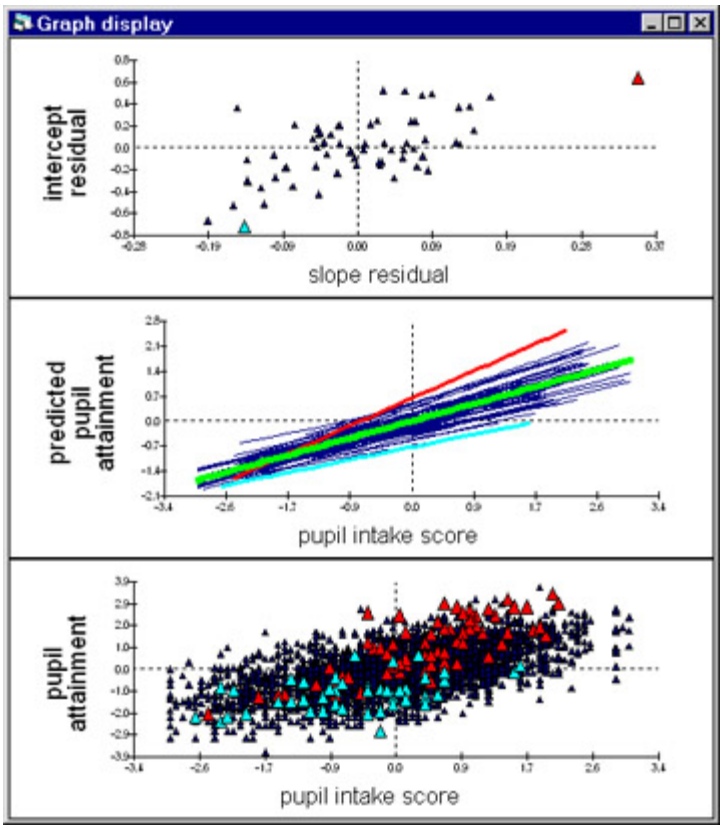
In the top panel we see that the school residual(effect) for school A is still statistically different from school B.

The set of school lines in the middle graph must be parallel because the model fitted contructs a schools line by adding that school's residual to the average(green) line's intercept. We could allow the schools lines to have different slopes.

The bottom graph suggests that lines with different slopes, certainly in the case of school A and school B, would be more realistic. Eyeballing the graph the points for school B suggest a line with a flatter slope than for school A.

ALLOWING DIFFERENT SLOPES FOR THE SCHOOLS LINES

If we allow every school to depart from the overall average line in terms of both its intercept and slope we get the folowing :



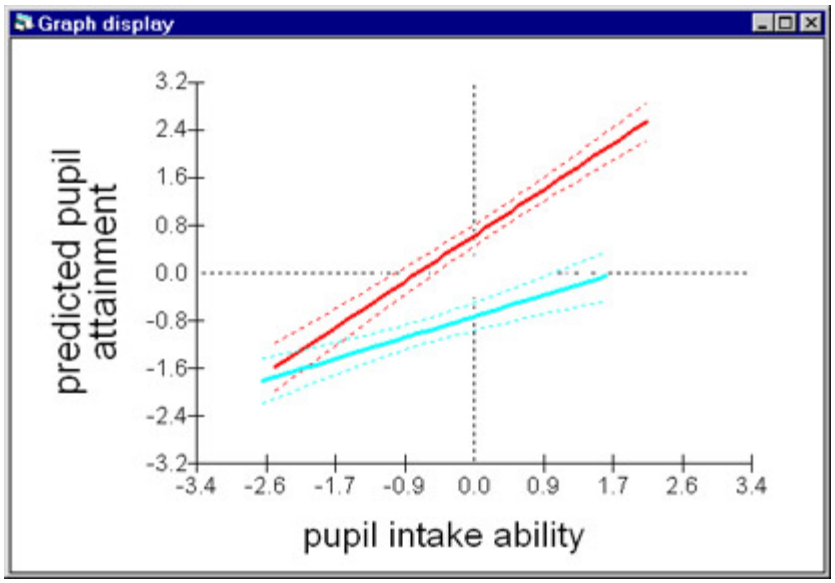
Every school now has an intercept residual and a slope residual. The corresponding 65 points are plotted in the top panel. We see that school A has the highest slope and the highest intercept. We therefore expect school A 's line to be the steepest line and to cross the y-axis at the point x=0 at a higher point than all the other schools' lines. If we look at the middle panel we can see this is the case.

Convesely, school B has the lowest intercept residual and a very low slope residual, which combine to create a flat line located at the bottom of the set of school's lines.

Consider again the question is school A more effective than school school B? The extent of the difference between the two schools depends on pupil's intake scores. For pupils with low intake scores the difference is small. For pupils with high intake scores the difference is large.

ONCE MORE WITH CONFIDENCE INTERVALS

Below is a graph with just the lines for schools A and B along with their associated confidence intervals.



Remember we are comparing in unadjusted terms the top and bottom schools. Once we correct for intake and allow schools to have their own intercepts and slopes we find that we can not definatively claim that school A is more effective than school B. For low inatke ability pupils statistically there is no difference between the two schools.

Schools are differentially effective for different types of pupils. Here we have only explored differential school effectiveness in terms of intake abilility. Schools can also be differentially effective with respect to other pupil characteristics. For example, gender, ethnicity and socio-economic status. Multilevel modelling provides a framework for describing and explaining these between school differences.

CONTEXTUAL EFFECTS

Another reason why multilevel modelling is attractive to social science researchers is that it is useful for exploring interactions between people and the social contexts they are situated in.

For example, do low ability pupils fare better when they are eduacted alongside higher ability pupils or are they discouraged and fare worse?

We can catgorise our 65 schools into 3 groups with respect to the intake scores of their pupils. We do the following.

- for each school calculate the mean intake score ability of all it pupils
- rank these 65 means
- assign schools in the bottom quartile to one group, the middle 50% to a second group and the top 25% to a high group.

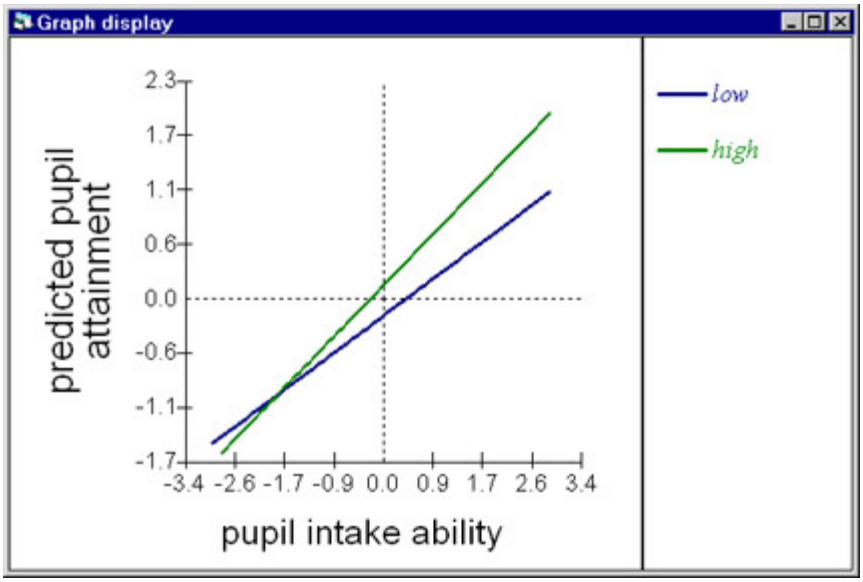
We now have three types of schools low, middle and high which correspond to low ability, middle ability and high ability schools.

We can include the school ability contextual variable in a multilevel model and allow it to interact with the pupil level intake ability. This tells us how pupils across the spectrum of pupil level intake ability are effected by being educated amoungst

low, middle or high ability peers.

The graph on the next page shows the results for low versus high ability schools.

THE CONTEXTUAL EFFECT OF PEER GROUP ABILITY



Consider first a high ability pupil. If we look at value of 2.6 on the x axis this corresponds to a pupil who on entry to secondary school hs a score of 2.6 standard deviations above the mean. If that pupil attends a school where her peers are on average low ability then the pupils predicted outcome attainment is 1.0 standard deviation above the average attainment; this is the height of the blue line at x value 2.6. However, if that same pupil attended a high ability school the model predicts that her outcome attainment would be 1.8 standard deviations above the mean outcome attainment; the height of the green line at x = 2.6.

For high ability pupils the model suggests there is a large positive effect of being eduacated amoungst high ability pupils. The difference between the green and blue lines represents the effect of being in a high ability group. As we move down the spectrum of pupil intake ability(leftwards along the the x axis) we see this effect lessening.

For values x less than –1.8 the blue line is higher. This means that very low ability pupils(x < -1.8) actually fare better when they are situated in a low ability school than in a high ability school.

Top

Exemplar searches

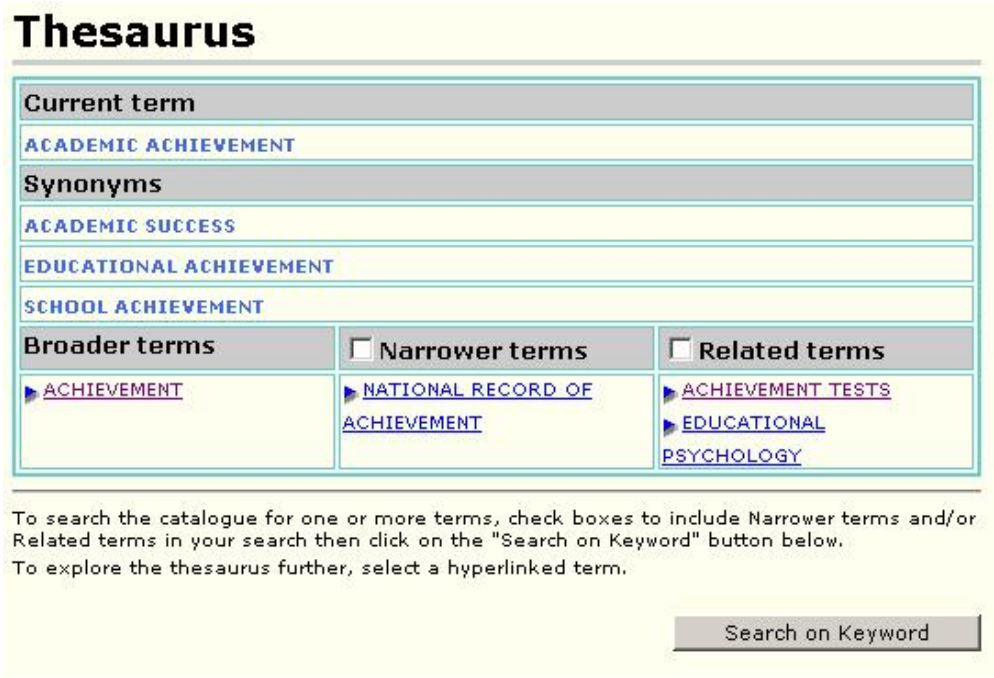
Searching for data

If you would like to open a separate browser window to view the online catalogue whilst reading these notes ***click here****.*

A sample search: Secondary education.

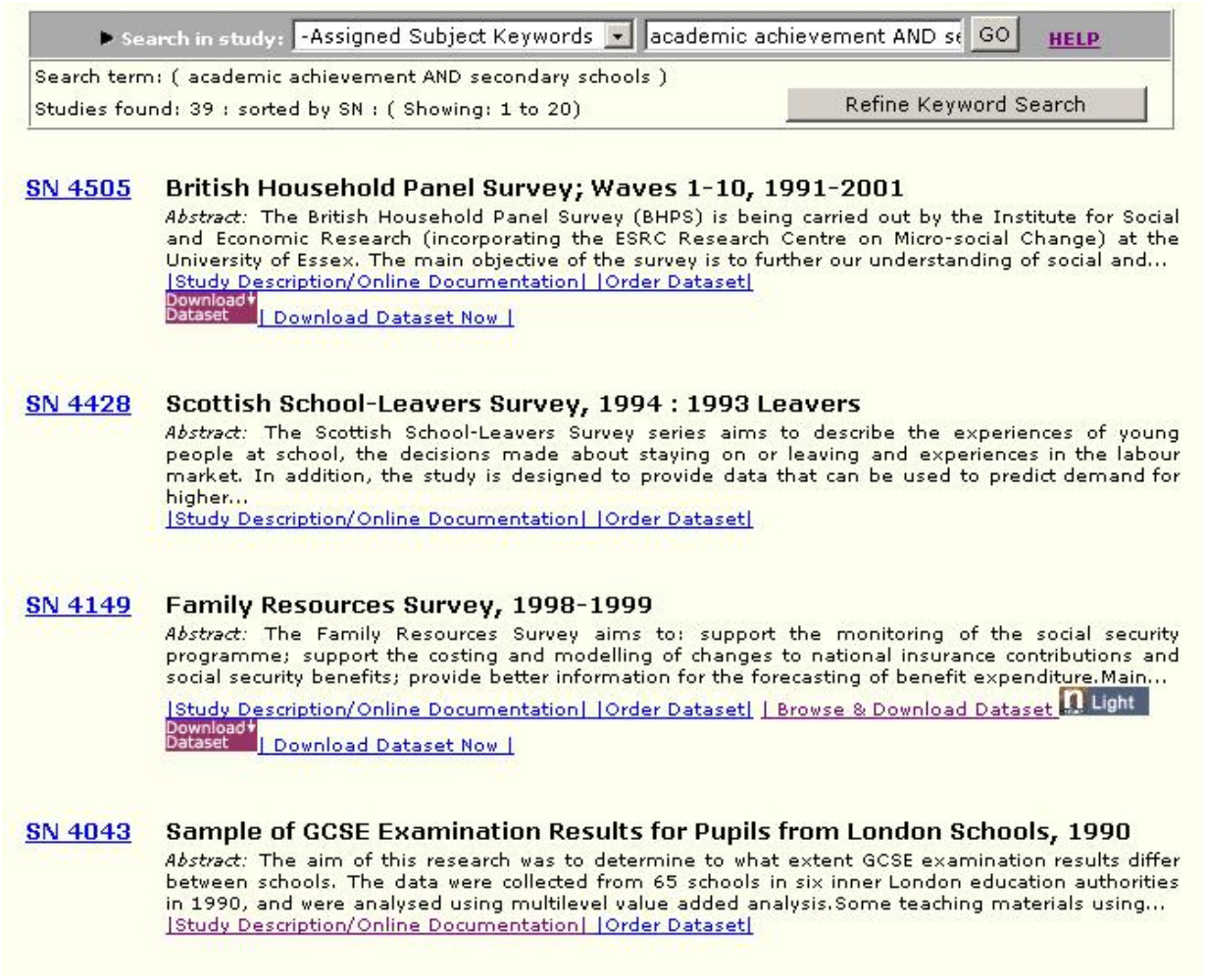
Given our research questions we might select 'Assigned Subject Keyword' from the drop down list and use 'school achievement' as our search term. However, this returns the following message: "No Documents matching your query school achievement were found".

This is where it is useful to search the HASSET thesaurus. From the search catalogue web page, follow the link to the HASSET Thesaurus. Next to 'Enter Keyword', type in 'School achievement' and click on 'Go'. This tells us that the preferred term is 'Academic achievement':



To search using the term 'Academic achievement' click on 'Search on Keyword'. However, this gives us over 160 studies. To narrow the results, add 'Secondary schools' to the search term, to give 'Academic achievement AND Secondary schools' and click on 'Go'.

This returns the following results:



If we look at the catalogue record for study number 4043 (Sample of GCSE Examination Results for Pupils from London Schools, 1990) we find the following information listed under the section 'Main Topics'.

The variables recorded are:
GCSE examination scores for English and mathematics;
a combined examination score from all subjects;
pupil and school identifiers (numerical only);
school gender (mixed, boys or girls school);
pupil gender;
continuous intake measure of reading ability;
categorical intake measure of verbal reasoning ability.
percentage of children in a school on free school meals, an indicator for social deprivation.

There is sufficient data here to explore our research questions.

An overview of a multilevel analysis of this data set is given in Education Overview

Top

Information about the Data Archive's catalogue

Searching for data

TRAMSS

Home page
Project overview
Research questions
Searching for data
About the UK Data Archive's catalogue
How to use the online catalogue
Online catalogue exemplar searches
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

The UK Data Archive's online catalogue

The online catalogue contains a record for each of the several thousand studies held in the UK Data Archive (UKDA) and is freely accessible by anyone with access to the Internet.

At its simplest, the online catalogue answers the following basic questions for each of the studies in the collection:

- What topics were investigated?
- Where was the research carried out?
- When was the data collected?
- Who was responsible fro the research/data collection?

A record will include most or all of the following information:

- Study number - this is allocated by UKDA and will be needed when ordering data.
- Study title - the title of the study agreed with the creators of the dataset.
- Subject Categories - each study is assigned one or more categories to reflect the overall subject of the data at study level.
- Assigned Keyword List - keywords, taken from the UKDA's thesaurus (HASSET) are assigned at the level of variables and offer much more detail than the subject categories.
- Name(s) of depositor(s), principal investigator(s), data collector(s), and sponsor(s).
- Abstract - a brief description of the study.
- Main topics covered by the dataset.
- Coverage - time period covered, geographical coverage, observation unit (e.g. individuals, households).
- Universe sampled - details of population included in the sample.
- Methodology - how the data were collected, number of units (cases).
- References to publications by principal investigators or resulting from secondary analysis. UKDA requests that all users of data inform us of any publications that result from their work with the data.
- Online documentation - the large majority of catalogue records provide access to freely downloadable User Guides supplied by the data depositors. These usually include detailed background information, such as methodology and sampling, the original questionnaires and the codebook.
- Variable list - a large number of catalogue records provide access to a list of variable names and variable and value labels.

Access to the data via the online catalogue

Users who have registered for a UKDA account, which involves signing an access agreement to agree to certain conditions of use, can:

- Download data via a 'Download Dataset' link in the online catalogue record. A large number of datasets are downloadable in this way.
- Browse, analyse, subset and download a number of major datasets using a service called NESSTAR via a 'Browse and Download' link in the online catalogue record.
- Add a dataset to an order from within a catalogue record (for example, to request the data on a CD).

Access to the data prepared for use in this project

The depositors of the data have agreed to waive the usual requirement for a signed access agreement. This means that users of this site may download data simply by agreeing to the conditions which appear on the screen.

The TRAMSS team would like to thank those depositors for their support.

Users of these pages should note that the exemplar datasets have been specially prepared to demonstrate the functions of the accompanying software. They will not necessarily reflect the complete range of variables or files available by ordering the full datasets from UKDA.

If you don't find what you want in the online catalogue email user support at UKDA.

Next section: How to use the online catalogue

Top

How to use the online catalogue
Searching for data

Home page
Project overview
Research questions
Searching for data
About the UK Data Archive's catalogue
How to use the online catalogue
Online catalogue exemplar searches
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

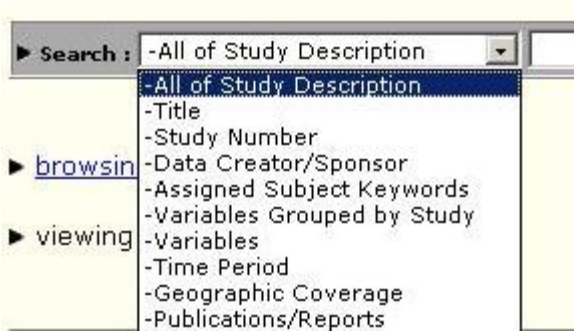
*If you would like to open a separate browser window to view the online catalogue whilst reading these notes **click here**.*

How do I use the online catalogue?

- Free-text or fielded search

Users can choose a field from a drop-down list that includes the following fields:

- All of Study Description
- Title
- Study Number
- Data Creator/Sponsor
- Assigned Subject Keywords
- Variables Grouped by Study
- Variables
- Time Period
- Geographic Coverage
- Publications/Reports



In any field, users can:

- search for terms with the same prefix. For example, to find scotland, scottish, scotsman and so on, users can enter the search term scot*.
- search for an exact phrase by putting quotation marks around the search term.
- combine search terms using AND, OR and AND NOT.

'All of Study Description' is the first and default search. This will produce a free-text search on all fields in the list (with the exception of variables), as well as the abstract and methodology e.g. sampling methods, method of data collection etc. Selecting one of the other fields will provide a more focused search.

'Assigned Subject Keywords' are contained in each catalogue record and cover all topics included in the data, including those at variable level. The keywords are taken from a controlled vocabulary list held in the UKDA thesaurus, HASSET, which is available to help decide on the appropriate search term. Having conducted an 'Assigned Subject Keywords' search a 'Refine Keyword Search' button is available at the top of the list of search results. This takes the user to the thesaurus where a more specific term may be selected to refine the search, or broader and related terms may be selected to widen the search.

'Variables' and 'Variables Grouped by Study' are searches that are only available for our most popular datasets or for those datasets deposited in a suitable format. These fields can be used to search for variables and then access the associated variable and value labels.

For more detailed Help on Searching click here

- Browse by subject category

Each dataset is assigned one or more categories to reflect the overall subject of the data at study level. Users can select the subject categories that they are interested in and browse the UKDA catalogue for studies with the chosen subject coverage.

- View a list of new data releases

Users can search for datasets or new editions released in the last 1, 2, 3, 6, 9 or 12 months.

- Conduct a thesaurus-aided search using the Humanities and Social Science Electronic Thesaurus (HASSET)

Instead of an 'Assigned Subject Keywords' search, users can access HASSET directly to search the online catalogue.

Next section: Exemplar searches using the online catalogue

Top

Exemplar searches

Searching for data

Home page
Project overview
Research questions
Searching for data
About the UK Data Archive's catalogue
How to use the online catalogue
Online catalogue exemplar searches
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

Exemplar searches using the UK Data Archive (UKDA) online catalogue

The following pages provide exemplar searches of the UKDA online catalogue to complement each of the exemplar research questions.

Please be aware that these exemplars are fixed. If you try and reproduce them in the live version the catalogue, the number of datasets resulting from the searches may not match those indicated in the exemplars. This is because we are constantly updating our catalogue by adding to the collection.

Click for sample catalogue searches:

- Migration search ▶
- Youth search ▶
- Education search ▶
- Mortality search ▶

Top

Statistical modelling

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Context for Learning
More statistical modelling
Methodological framework
Selected readings
Bibliography
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Context for learning

In their introduction to *Analyzing Social and Political Change* (1994) Professors Angela Dale and Richard Davies begin 'whilst qualitative researchers have long been concerned with process, limitations of data and method have confined many quantitative researchers to cross-sectional studies with inferences about process requiring bold assumptions or heavy reliance upon untested substantive theory'. They go on to argue that the social and political sciences are moving through a period of rapid methodological development. Statistical theory and application are developing an increasingly symbiotic relationship. Statistics cannot flourish without data whilst the complexities of the data collection process cannot be handled without appropriate conceptual frameworks and accompanying analytical tools. It is possible to attach new meaning to the ideas of methodological rigour and thoroughness in our approach to empirical research. They suggest that there is a 'growing recognition that analyses of social life based upon static, cross-sectional data are partial at best and misleading at worst. This changing emphasis has also brought about a corresponding increase in the longitudinal data available for secondary analysis'. What this project does is it unlock both the potential for a fuller understanding of the role of analytical tools in the search for knowledge based on large and complex data sources. It explicitly recognises the gap between 'everyday social science research' and the development work of a few statisticians fortunate to be funded under the ALCD programme. Our aim is to begin to bridge that gap by putting both the complexities of data together with appropriate software tools for analysis. Obviously, data and software cannot be joined up out of context. What stimulates the collection of data in substantive terms must be placed alongside the search for a methodological approach, which actually captures the research question and endeavours to reflect any assumptions, which shape its articulation. We present the user with such a context as well as providing an opportunity to taste the richness of secondary data available for analysis in the Archive.

- *Acknowledgement*. The intellectual stimulus for this overview has been largely drawn from Dale and Davies (1994).
- Dale, A. and Davies, R. (1994) Analyzing Social and Political Change, Sage, London.

Next section: More statistical modelling ▶

Top

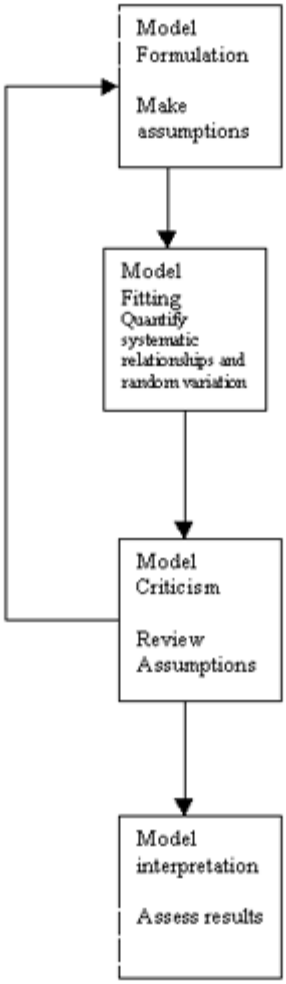
More statistical modelling

The project takes two particular methods of analysis, event history analysis (Tuma, 1994) and multilevel analysis (Goldstein, 1995) and their accompanying software products SABRE and MLwiN , both developed under the ALCD programme. Each method is illustrated in the context of a central core of substantive questions, which drive both the search for the data source and methodology. Obviously, the user who simply wishes to browse an aspect of either method or data source is free to do so. No one is required to be linear as analytical quests often involve an interaction between data, substance and method.

The general principles that underlie each method are provided so that the user is introduced to their analytical potential by means of an illustration. There is no attempt to provide either a complete description of method or software. What you have is a resource, which can alert the serious researcher to new possibilities and opportunities for acquiring knowledge. The methodologies illustrated in this project fall within the general inferential framework of statistical modelling. The process of modelling reveals three important benefits: the ability to distinguish systematic relationships in complex data, to make explicit the role of substantive theory for inference and the use of *fitted* models for prediction. The latter may have important consequences in policy research.

Dale and Davies (1994) provide an excellent description of the sequence of procedures that articulate statistical modelling. Namely, model formulation, model fitting, model criticism or assessment and model interpretation. The researcher recycles the first three stages until s/he is satisfied that the model is adequate. Model interpretation concludes the process.

The first stage entails a description of the process of interest. Put another way, specify or formulate a probability model (*model formulation*) which itself will imply a number of assumptions regarding the sampling scheme, level of measurement, error distribution and unobserved heterogeneity. The formulation and choice of variables entering the analysis will also be guided by substantive theoretical considerations. The second stage involves *model fitting*. The probability model is used to fit the observed data. We quantify the systematic relationships and random variation produced by the model. Parameters are estimated with accompanying measures of reliability (e.g. standard errors). *Model criticism* involves making an assessment of the adequacy of the model. Is it a parsimonious description of the observed process? Do the



assumptions hold ? This may involve using goodness-of-fit measures, analysis of residuals and comparison with alternative models involving greater or lesser complexity.

Finally, attention turns to the substantive significance of the findings. This is often the most interesting part of the process. To what extent has the analysis answered the research questions? What new insights have been revealed? Have any potential methodological pitfalls been addressed ? These considerations are rarely examined in formal textbooks on statistical methodology. Our approach is to allow you to try out a statistical investigation for yourself. For a sample of applications of multilevel modelling and event history analysis try ***selected journal articles***.

- Dale, A. and Davies, R. (1994) Analyzing Social and Political Change, Sage, London.
- Goldstein, H. (1995) Multilevel Statistical Models, 2nd edition, Edward Arnold, London.
- Tuma, N. (1994) Event History Analysis, Chapter 7, *Analyzing Social & Political Change*, Dale, A. and Davies, R. (eds), Sage, London.

Next section: Methodological framework▶

Top

Methodological framework

The first example uses data from a large *retrospective* survey of life and work histories carried out in the mid-1980's under the Social Change and Economic Life Initiative (SCELI), funded by the ESRC. The data are longitudinal and the illustration explains the limitations of cross-sectional data analysis for drawing inference about the dynamics of migration. Put another way, one of the key advantages of using longitudinal data is to overcome the problems of *inferring process from cross-sectional data*. Cross-sectional data that simply record levels of migration (both in- and out-migration) by age by geographical area may confound these effects and may give a misleading impression of *life course* changes. Data which record housing and labour market experience as well as information on particular life events for example, births and marriages allow analyses between different age cohorts and areas which relate decisions to move over the occurrence of life events and changing economic circumstances. By relating such outcomes to earlier circumstances for the same individuals it is possible to explain *how* the process unfolds for different age cohorts. This is only made possible by collecting data for individuals over successive time points.

A major drawback with retrospective information covering a wide range of variables is the problem of *recall* (Dex, 1991). Accurate data relating to the distant past may be difficult to collect even with careful prompting. One solution is to *repeat surveys*, with retrospective questions to cover any gaps in the intervening periods of data collection. Repeat surveys are commonly referred to as *panel studies*, for example The British Household Panel Survey (BHPS).

Rather than collect data retrospectively information can be collected as events occur or *prospectively*. Prospective birth cohort studies, such as The 1958 British Cohort Study (The National Child Development Survey (NCDS)), are used to study the developmental process. They largely, overcome the problems of recall but are expensive to administer and often undermined by *attrition* (Kaspryk et al., 1989). They are used to analyse development as they specifically control for age. Typically, they are concerned with the impact of the early life course on subsequent adult outcomes, for example the relationship between the accumulation of social and material disadvantage and ill health.

An inherent weakness of Birth Cohort Studies is that there is no opportunity to explore any differences between cohorts. A *cross-cohort* comparison allows the analyst to assess to what extent there have been changes across time. For example, comparing the labour market experiences of the 1958 British Cohort with those of the 1970 British Cohort facilities a comparison of changes in employment conditions for young people between the 1970's and the 1980's.

Whilst it is important to see the power of longitudinal data it is also important to appreciate that prospective studies are only as powerful as the questions allow the secondary analyst to pursue. As new research questions present themselves it is

often difficult to change the direction of longitudinal research.

Longitudinal data, whether prospective or retrospective, not only begins to unravel the nature of change at an individual level but also presents opportunities to explicitly recognise that a lot of behaviour is characterised by strong temporal tendencies. Dale and Davies (1994) draw the distinction between *duration dependence* (the interval since commencing a job and the decision to quit) and *state dependence* (how a person votes in relation to what they voted previously) and suggest that most of the factors creating such temporal dependencies generate *inertia effects* in behaviour. Longitudinal data then becomes essential if we are to understand these temporal tendencies in micro-level behaviour.

The authors go on to argue that with longitudinal data it is possible to achieve greater control over 'the myriad of variables that are inevitably *omitted* from any analysis'. They continue 'because of our limited ability to model human behaviour, there is considerable *heterogeneity* in the response variable, even among people with the same characteristics on all explanatory variables'. Using longitudinal data the effects of omitted variables can be explicitly accounted for in the model. Formally, this is referred to as *residual or unobserved heterogeneity*. See Heckman (1979) for an example.

Advanced statistical methodology has developed in response to a need both to model complex reality (the recognition of temporal dependency in behaviour) and incorporates ways of recognising the dangers of oversimplification (the explicit recognition of residual heterogeneity).

- Dale, A. and Davies, R. (1994) Analyzing Social and Political Change, Sage, London.
- Dex, S. (1991) The reliability of recall data; a literature review. Occasional Papers of the ESRC Research Centre on Micro-Social Change, Paper 6. Colchester, University of Essex.
- Heckman, J.J. (1979) New evidence on the dynamics of female labor supply, in C.B. Lloyd, E.Andrews and C.Gilroy (eds), *Handbook of Econometrics*, Vol. III, Amsterdam: Elseiver. pp. 1689-786.
- Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (eds.) Panel Surveys, New York, Wiley Interscience.

Next section: Selected readings ▶

Top

Selected readings

A selected number of readings which present the application of event history analysis and multilevel modelling are listed below to help you to find appropriate examples in your own area or discipline.

Journal of Health Services Research and Policy Vol. 1 No. 3, 1996: 154-164

Multilevel models: applications to health data

Nigel Rice, Alastair Leyland*

Centre or Health Economics, University of York;

*Public Health Research Unit, University of Glasgow, UK

Abstract

This paper presents an introductory account of multilevel models, highlighting the potential benefits that may be gained by the use of these methods. It draws on recent applications in health services research that have appeared in the literature. Methodological advances in these statistical techniques have taken place in the field of education, where empirical studies have mainly been concerned with comparing pupil achievement across different schools by exploring the relationship between individual and institutional factors. Although recent widespread availability of suitable software packages has enabled other disciplines to adopt these methods, to date they have received little attention in the health services research literature (the investigation of effects of geographical areas on health being a possible exception) despite their obvious application in many areas of current interest. Key areas that could benefit greatly from these techniques include the exploration of variations in clinical practice, comparisons of institutional performance and resource allocation.

Soc. Sci. Med. Vol. 37, No. 6, pp. 725-733, 1993

Do Places Matter? A Multi-Level Analysis of Regional Variations in Health-Related Behaviour in Britain

Craig Duncan,[1] Kelvyn Jones[1] and Graham Moon[2]

1Department of Geography, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth, PO1 3HE, England and [2]School of Social and Historical Studies, University of

Portsmouth, Milldam, Burnaby Road, Portsmouth PO1 3AS, England

Abstract

A number of commentators have argued that there is a distinctive geography of health-related behaviour. Behaviour has to be understood not only in terms of individual characteristics, but also in relation to local cultures. Places matter, and the context in which behaviour takes place is crucial for understanding and policy. Previous empirical research has been unable to operationalise these ideas and take simultaneous account of both individual compositional and aggregate contextual factors. The present paper addresses this shortcoming through a multi-level analysis of smoking and drinking behaviours recorded in a large-scale national survey. It suggests that place, expressed as regional differences, may be less important than previously implied.

Journal of Epidemiology and Community Health 1993; 47: 481-484

Effect of the remuneration system on the general practitioner's choice between surgery consultations and home visits

Ivar Sønbø Kristiansen, Knut Holtedahl

Abstract

Objective – To assess the influence of the remuneration system, municipality, doctor, and patient characteristics on general practitioners' choices between surgery and home visits.

Design – Prospective registration of patient contacts during one week for 116 general practitioners (GPs).

Setting – General practice in rural areas of northern Norway.

Main outcome measure – Type of GP visit (surgery v home visit).

Results – The estimated home visit rate was 0.14 per person per year. About 7% (range 0-39%) of consultations were home visits. Using multilevel analysis it was found that doctors paid on a "fee for service" basis tended to choose home visits more often than salaried doctors (adjusted odds ratio 1.90, 99% confidence interval 0.98, 3.69), but this was statistically significant for "scheduled" visits only (adjusted OR 4.50, 99% CI 1.67, 12-08). Patients who were older, male, and who were living .in areas well served by doctors were more likely to receive home visits.

Conclusion – In the choice between home visits and surgery consultations, doctors seem to be influenced by the nature of the remuneration when the patient's problem is not acute. Although home visiting is a function of tradition, culture, and organisational characteristics, the study indicates that financial incentives may be used to change behaviour and encourage home visiting.

Soc. Sc. Med. Vol. 33, No. 4, pp. 501-508, 1991

Ecological and Individual Effects in Childhood Immunisation Uptake: A Multi-Level Approach

Kelvyn Jones[1], Graham Moon[2] and Andrew Clegg[3]

1Department of Geography, Portsmouth Polytechnic, Buckingham Building, Lion Terrace,

Portsmouth PO1 3AS, [2]School of Social and Historical Studies, Portsmouth Polytechnic, Milldam, Burnaby Road, Portsmouth PO1 3AS and [3]Community and Small Hospitals Unit, Portsmouth and South East Hampshire Health Authority, St Marks House, Derby Road, North End, Portsmouth, UK

Abstract

Analyses of childhood immunisation uptake have traditionally been conducted at either the ecological or the individual scale. In this paper the problems stemming from these distinct strategies are explored and the potential of a multi-level modelling approach taking simultaneous account of processes at both levels is discussed. This discussion is set in the context of a case-study of pertussis immunisation uptake using data gathered from routine child health surveillance and immunisation uptake monitoring. The role of multi-level modelling in medical geographic research is briefly evaluated.

Health-related behaviour in context: a multilevel modelling approach

Craig Duncan[1], Kelvyn Jones[1], Graham Moon[2]

1Department of Geography, University of Portsmouth

2School of Social and Historical Studies, University of Portsmouth

Abstract

Recent attempts to place individual health-related behaviour in context have been judged largely unsuccessful. This paper examines how this situation might be improved and is especially concerned with the role of quantitative methodologies. It is argued that, whilst recent developments in social theory help provide important theoretical guidelines, they can only be implemented with difficulty in empirical health-related behaviour research if traditional quantitative methodologies are used. It is suggested that the best way to implement social theory within a quantitative framework is to apply the newly developed technique of multilevel modelling. This paper offers an overview of the multilevel approach and outlines its significance for health-related behaviour research. In addition, it details a number of ways in which the multilevel framework can be extended so as to achieve further improvements in the conceptualisation of health-related behaviour. To illustrate the value of the technique, the paper finishes by considering one of these extension is detail and applying it to data recording smoking behaviour in the United Kingdom.

Social Science and Medicine Vol. 46 No. 1, 1998: 97-117

Context, composition and heterogeneity: using multilevel models in health research

Craig Duncan[1], Kelvyn Jones[1], Graham Moon[2]

1Department of Geography, University of Portsmouth

2School of Social and Historical Studies, University of Portsmouth

Abstract

This paper considers the use of multilevel models in health research. Attention focuses on the structure and potential of such models and particular consideration is given to their use in elucidating the importance of contextual effects in relation to individual level social and demographic factors in understanding health outcomes, health-related behaviour and health service performance. Four graphical typologies are used to outline the questions that multilevel models can address and the paper illustrates their potential by drawing on published examples in a number of different research areas.

London; Edward Arnold 1995

Multilevel Statistical Models

Harvey Goldstein

Institute of Education, University of London

Preface

In the mid 1980s a number of researchers began to see how to introduce systematic approaches to the statistical modelling and analysis of hierarchically structured data. The early work of Aitkin *et al* (1981) on the teaching styles' data and Aitkin's subsequent classic work with Longford (1986) initiated a series of developments that, by the early 1990s, had resulted in a core set of established techniques, experience and software packages that could be applied routinely. These methods and further extensions of them are described in this book and are

coming to be applied widely in areas such as education, epidemiology, geography, child growth, household surveys and many others.

This second edition aims to integrate existing methodological developments within a consistent terminology and notation, provide examples and explain a number of new developments, especially in the area of discrete response data, time series models, random cross classifications, errors of measurement, missing data and nonlinear models. In many cases these developments are the subject of continuing research, so that we can expect further elaborations of the procedures described.

The main text seeks to avoid undue statistical complexity, with methodological derivations occurring in appendices. Examples and diagrams are used to illustrate the application of the technique and references are given to other work. The book is intended to be suitable for graduate level courses and as a general reference.

Read/download this text

Amsterdam; TT-Publikaties 1995

Applied Multilevel Analysis

JJ Hox

Faculty of Educational Sciences, University of Amsterdam

Preface

This book is meant as a basic and fairly nontechnical introduction to multilevel analysis, for applied researchers in the social sciences. The term ‘multilevel’ refers to a hierarchical or nested data structure, usually people within organizational groups, but the nesting may also consist of repeated measures within people, or respondents within clusters as in cluster sampling. The expression *Multilevel model* or *multilevel analysis* is used as a generic term for all models for nested data. This book presents two multilevel models: the multilevel regression model and a model for multivariate covariance structures.

Read/download this text

Oxford Bulletin of Economics and Statistics, 1998, Vol. 60, No,1, p.79.

Women’s employment transitions around child bearing

Dex, S., Joshi, H., Macran, S., McCullogh, A.

Abstract

The dynamics of women’s labour supply are examined at a crucial stage in their lifecycle. This paper uses the longitudinal employment history records for 3898 33-year-old mothers in the Fifth Sweep of the 1958 National Child Development Study cohort in the United Kingdom. Models of binary recurrent events are estimated, which correct for unobserved heterogeneity, using SABRE software. These focus on women’s first transition to employment after the first childbirth, and on the monthly transitions from first childbirth until censoring at the interview. Evidence of a polarisation is found between highly educated, high-wage mothers and lower-educated, low-wage mothers.

Oxford Bulletin of Economics and Statistics, 1998, Vol. 60, No,2, p.261-265.

The relationship between event history and discrete time duration

models: An application to the analysis of personnel absenteeism

Tim Barmby

The relationship between the parameter estimates obtained from an event history model for binary recurrent events and the parameters which would be obtained from a direct analysis of the durations in either of the two states described by the event history, is discussed in the context of worker absenteeism. The method shows how with suitable model parameterization and using existing software (SABRE software is used for the event history analysis), an important link between two types of analysis, commonly undertaken in econometrics, can be established.

Journal of Statistical Computation and Simulation, 1996, Vol. 55, No. 1-2, pp.73-86

Fitting a random effects model to ordinal recurrent events using existing software

Berridge, D. M., DosSantos, D. M.

Abstract

The continuation ratio model is a direct generalization of the familiar binary logistic model. In this paper, it is proposed to model ordinal recurrent events by generalizing the logistic-normal model for binary recurrent events in a similar manner. This new model is implemented in the statistical software package SABRE.

Statistical Methods in Medical Research,1994, Vol. 3, No.3, p.244-262

Some approaches to the analysis of recurrent event data

David Clayton

Abstract

Methodological research in biostatistics has been dominated over the last twenty years by further development of Cox's regression model for life tables and of Nelder and Wedderburn's formulation of generalized linear models. In both of these areas the need to address the problems introduced by subject level heterogeneity has provided a major motivation, and the analysis of data concerning recurrent events has been widely discussed within both frameworks. This paper reviews this work, drawing together the parallel development of 'marginal' and 'conditional' approaches in survival analysis and in generalized linear models. Frailty models are shown to be a special case of a random effects generalization of generalized linear models, whereas marginal models for multivariate failure time data are more closely related to the generalized estimating equation approach to longitudinal generalized linear models.

Computational methods for inference are discussed, including the Bayesian Markov chain Monte Carlo approach.

Statistical Methods in Medical Research,1994, Vol. 3, No.3, p.263-278.

Generalizations and applications of frailty models for survival and event data

Andrew Pickles and Robert Crouchley

Abstract

A variety of survival models with both discrete and continuously distributed frailty is considered within a framework that involves the specification of three sub-models. An intensity sub-model specifies how the intensity is related to values of covariates and frailty; a measurement sub-model specifies how fallible measures of frailty are related to it; and an exposure sub-model specifies how frailty is distributed within the population. The models include those in which frailty is due to omitted covariates and those where it represents a covariate that has been measured subject to error. Multivariate frailty is also considered, with particular emphasis on models suitable for application to genetically related individuals, notably twins. A numerical example illustrates the use of a model with multivariate frailty for data on repeated exercise times.

Oxford Bulletin of Economics and Statistics, 1992, Vol.54, No,2, p.145-171.

The relationship between a husband's unemployment and his wife's participation in the labour force

Richard B. Davies, Peter Elias and Roger Penn

A number of studies have found that the wives of unemployed men in Britain are less likely to be in paid work than the wives of employed men. Using life and work history data from six localities collected under the Social Change and Economic Life Initiative, this paper investigates how far the observed relationship is due to a true cross-couple state dependence (due to factors such as the benefit system, for instance) and the extent to which it is due to a heterogeneous population, where the differing personal and labour market characteristics of both partners influence their employment status. Mixture models are fitted using SABRE software, to control for omitted variables. The observed relationship between the employment status of husbands and wives is shown to be due not only to causal factors but also due to heterogeneity: men who tend to experience unemployment (due to low skills, for instance) are more likely than previously thought, to marry women who have difficulty in finding employment in the labour market.

Lindsey, J. K., (1999), Models for repeated measurements (Second edition), Oxford University Press, Oxford.

This book presents a wide range of methods and examples on the analysis of repeated measurements. It assumes familiarity with the basic methods of discrete data and survival analysis, and is a suitable text for research students. It is also an important reference book for research statisticians in fields such as agriculture, medicine, economics and psychology. The first part of the book introduces the

three basic types of response variable: continuous, categorical and count, and duration, with a discussion of the ways in which such repeated observations are interdependent. It develops a framework of suitable models, with the introduction of multivariate distributions and stochastic processes. The following three parts of the book present a large number of examples corresponding to the different types of response; the section on duration data includes frailty, heterogeneity and event histories. Some major revisions have taken place since the first edition, in line with new developments in the field.

Blossfield, H-P. and Rohwer, G., (1995), Techniques of Event History Modeling: New approaches to causal analysis, Lawrence Erlbaum Associates, Mahway, New Jersey

A comprehensive introductory account of event history modelling techniques using continuous-time models. Intended to be used as a student textbook and also as a reference work for researchers. Models allowing for unobserved heterogeneity are introduced in the final chapter. There is accompanying software called TDA, with practical examples from sociology and labour market studies.

Trussell, J., Hankinson, R, and Tilton, J. (eds.) (1992), Demographic Applications of Event History Analysis, Clarendon Press, Oxford.

A selection of papers on event history analysis with demographic applications, such as home ownership, fertility, non-marital union dissolution. Contributions include papers on the incorporation of unmeasured heterogeneity into the analysis of event histories and discussions of methodological issues such as the treatment of missing data .

Yamaguchi, K., (1991), Event History Analysis, Sage, Newbury Park.

An introduction to event history analysis, using both discrete-time logit and continuous-time models. The discrete-time section covers such issues as one- and two-way transitions, duration dependence, covariate duration effects and other temporal factors. The section on continuous-time survival models covers only the Cox model, but explores non-proportionality and stratified models, as well as time-dependent models. There is a final section which discusses practical problems in event history modelling.

Next section: Bibliography ▶

Top

Bibliography

All of the material referenced in these pages are listed here.

Dale, A. and Davies, R. (1994) Analyzing Social and Political Change, Sage, London.

Dex, S. (1991) The reliability of recall data; a literature review. Occasional Papers of the ESRC Research Centre on Micro-Social Change, Paper 6. Colchester, University of Essex.

Goldstein, H. (1995) Multilevel Statistical Models, 2nd edition, Edward Arnold, London.

Heckman, J.J. (1979) New evidence on the dynamics of female labor supply, in C.B. Lloyd, E.Andrews and C.Gilroy (eds), *Handbook of Econometrics*, Vol. III, Amsterdam: Elseiver. pp. 1689-786.

Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (eds.) Panel Surveys, New York, Wiley Interscience.

Tuma, N. (1994) Event History Analysis, Chapter 7, *Analyzing Social & Political Change*, Dale, A. and Davies, R. (eds), Sage, London.

Top

MLwiN
Analysis software

TRAMSS

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
MLwiN
SABRE
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

MLwiN - a Visual Interface for Multilevel Modelling

- Multilevel modelling is a powerful new statistical technique extending regression modelling to the analysis of data from hierarchical population structures. *More* ▶
- Examples of such structures include people within families, children within classes within schools, and repeated measurements on individuals. *More* ▶
- Many substantive research questions cannot be addressed unless these structures are modelled appropriately. *More* ▶
- This website provides overviews of the findings from two multilevel analyses guiding the user through the concepts and principles of multilevel modelling. *More* ▶
- Detailed step-by-step tutorials based on these analyses can be viewed on-line or downloaded. *More* ▶
- A special edition of the widely used multilevel modelling software package MLwiN can be downloaded from this site together with the necessary datasets to enable you to work through the tutorials. *More* ▶

Top

SABRE
Analysis software

TRAMSS

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
MLwiN
SABRE
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

SABRE - Software for the Analysis of Binary Recurrent Events

SABRE is a program for the statistical analysis of binary, ordinal and count recurrent events. Such data are common in many surveys either with recurrent information collected over time or with a clustered sampling scheme. It is particularly appropriate for the analysis of work and life histories, and has been used intensively on many longitudinal datasets. Its development has been funded by ESRC, ALCD and Lancaster University. In 1989, SABRE 2.0 was released, written by Jon Barry, Brian Francis and Richard Davies. SABRE 3.0, developed by Dave Stott, with substantially enchanced statistical facilities, was released as freeware on the WWW in 1996. The current release is version 3.1. As part of the ALCD initiative, SABRE was also incorporated as a function in the S-plus Lancaster University OSWALD package , which provides comprehensive facilities for the analysis of longitudinal data. The details of using and loading OSWALD are not given here but can be found on the OSWALD website.

Specification

- A command driven package, with over 35 commands. However a basic set of only a few commands is needed to fit models to data.
- Fits the mover-stayer model, conventional logistic, logistic-normal and logistic-normal with end-points models to binary data.
- Fits a continuation ratio-type generalisation of the above models to ordinal data.
- Fits conventional log-linear, log-linear normal and log-linear normal with end-point models to count data.
- Substantial control is available over the parameters of the algorithm for the sophisticated user
- Can deal with very long sequences of data
- Comprehensive user manual and on-line help system.

Typical Applications

- Studies of voting behaviour, trade union membership, economic activity and

migration.
- Demographic surveys
- Studies of infertility in humans
- Animal husbandry
- Absenteeism studies
- Clustered sampling schemes

For lots more information, see the SABRE web pages

Download Data and software

TRAMSS

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Advice for new users

If you are not already familiar with the SABRE or MLwiN software then please spend some time working with the tutorial material.

- SABRE - Software for the Analysis of Binary Recurrent Events

  Download teaching version of SABRE, manual and datasets ▶
  (for Windows 95, Windows 98 and Windows NT)

- MLwiN

  Download MLwiN and data ▶
  Get the MLwiN tutorials ▶

Notes on problems with downloading software

A caution: clicking on the links above should enable you to save the software to your local disk or network drive. *However you may need to check permissions with your systems administrator.* If you do not see a dialog box giving you this option to save to disk then use the right mouse button on the link and select 'save link as' to save to disk.

Top

Download Data and software

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Download Analysis software

Please read the following information before proceeding to download the software at the bottom of this page.

- Special editions of statistical modelling software (MlwiN and SABRE) can be downloaded from this site.
- The software is free and can be downloaded with tutorials and exemplar data sets derived from the Archive.
- Before downloading you will be required to endorse an undertaking agreement.
- Tutorials provide a step by step guide to the principles of each modelling application.
- Each tutorial is framed by a number of substantive research questions.
- For more information about the analysis software click on the left sub-menu. Alternatively, move straight on to download.
- Please let us know how you get on with downloading and using the material provided. There is an electronic feedback form available under the 'feedback' option.

ACCESS AGREEMENT FOR USE OF DATA

The depositors of the data used as exemplar material for this ALCD project have generously waived the usual requirement for users to sign a written access agreement before accessing the data.

Users are nevertheless required to agree the following conditions before accessing the data:

This access agreement concerns the conditions of use of data and explanatory documentation supplied to me by The Data Archive. These data and explanatory documentation are hereafter referred to as `the materials' which will also include any additional data or explanatory documentation which are not the subject of a separate agreement.

I hereby undertake:

(1) Purpose: To use the materials only for the purposes of learning or teaching via the TRAMSS web site.

(2) Confidentiality: To act at all times so as to preserve the confidentiality of individuals and institutions recorded in the materials. In particular I undertake not to use or attempt to use the materials to derive information relating neither specifically to an identified individual or institution nor to claim to have done so[1].

(3) Acknowledgement: To acknowledge in any publication, whether printed, electronic or broadcast, based wholly or in part on such materials, both the original depositors and the Archive. The wording of the citation for individual datasets is to be

found in the documentation distributed by the Archive. To declare in any such work that those who carried out the original collection and analysis of the data bear no responsibility for their further analysis or interpretation. To acknowledge Copyright where appropriate.

(4) Access to others: Only to give access to others via the TRAMSS web site.

(5) Errors: To notify the Archive of any errors discovered in the materials.

(6) Liability: To accept that the Archive and the depositor of the materials supplied bear no legal responsibility for their accuracy or comprehensiveness.

1 *This clause does not apply to certain historical data which are based on sources which are in the public domain. Please check with the Archive for exceptions.*

I have read and agreed these conditions ▶

Top

Search/route map

TRAMSS

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Search this site

Enter a term in the box below to search the TRAMSS web site

Search for: Submit Results per page: 10

Match: any search words all search words

Top

TRAMSS

Teaching Resources and Materials for Social Scientists

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

- Who is this site for?

Anyone with an interest in data discovery and statistical analysis.

- What do I need to get something out this site?

About an hour or so and an understanding of multiple regression. Then you may wish to return to print off material or download data and software.

- So where will it take me?

The site provides a taste of statistical software applications in event history analysis and multilevel modelling.

- How will I learn?

You can learn to search the Data Archive's catalogue and then download software and data to run analyses. Examples are presented in a substantive framework with specially prepared datasets.

- Am I about to get lost?

Use the left-hand menu to explore the site. Typically pages are structured so that there are layers of information if you want to pursue any aspect of the site

- Feedback your experience.

Please take the time to let us know how you get on. Use the electronic form available under feedback.

Top

Modelling Migration Histories

Juliet Harman, Brian Francis and Richard Davies
Centre for Applied Statistics, Lancaster University

The main substantive questions

1 **Are some people more likely to move than others?**
What factors determine an individual's propensity to migrate? Are there people who are likely never to move?

2 **Does an individual's migration behaviour vary with time?**
Do people tend to move at certain ages, at particular life events (marriage, children, schooling), for employment opportunities, or as a response to external factors such as the economic climate or the housing market?

3 **How can we separate different temporal effects?**
Differing patterns of migration behaviour with age are likely for different birth cohorts, as individual life histories take place in different and changing economic conditions. Cumulative inertia effects (the increasing tendency to stay as length of residence in the same place increases) may complicate the variation of migration propensity with age. How can we disentangle the three temporal effects: age, calendar year and duration of stay?

What data set is analysed?

- To address these substantive questions, we need a data set on each of a large number on individuals, with information for each individual on their *migration* history, their *marital* history, their *employment* history and their *family* history.
- Such historical information is needed from the start of each individual's adult life until the date of data collection.
- We can use BIRON to search the Data Archive catalogue to find a suitable data set. An example has been constructed on how to search for such a data set on migration.
- The data set chosen is a large retrospective survey of life and work histories carried out in 1986 under the Social Change and Economic Life Initiative (SCELI), funded by the ESRC.

Will I understand this module?

- We assume that you have a certain amount of statistical knowledge already. The most important requirement is to be able to understand the output of a multiple regression. A basic knowledge of logistic regression and Poisson regression (regression models for count data) would also be useful, but this is not essential. We provide an explanation of new technical terms, and explain results through the use of graphs.

Give me a quick overview of this module

- We first analyse a summary data set containing the total number of moves for each individual, and demonstrate the limitations of such cross-sectional analysis for drawing inference about the dynamics of migration.
- We then explore the longitudinal data set containing the life and work histories, and model the annual binary migration data using a conventional logistic model. We discuss the limitations of using conventional models for longitudinal data and demonstrate the importance of controlling for individual specific explanatory variables omitted from the analysis.

What software do I need?

- You will need to use SABRE, which is a statistical software package for the analysis of discrete longitudinal data. SABRE runs on all Windows machines and also on UNIX and Linux platforms. SABRE and the teaching data sets can be downloaded from

here free of charge.

- SABRE is a specialist package, with a restricted range of commands; it has no facility for instance to plot graphs. However, the parameter estimates from model fitting can be copied into other packages. We use the statistical package GLIM to supplement SABRE.

How do I use this module?

The best way is to follow the module page by page on the Web, loading the data set into SABRE in a new window, and following the instructions onscreen. Alternatively, it is possible to download the entire module as an ADOBE portable document file.

Acknowledgement

This example is based on research work carried out by R. B. Davies and R. Flowerdew (1992) and by Haghighi A. Borhani and R. B. Davies (1999a, 1999b), using data collected under the Social Change and Economic Life Initiative funded by the ESRC. The work by Haghighi Borhani and Davies was partially supported by ESRC research grant L315253007.

NEXT:Table of contents

Home page

MODELLING MIGRATION HISTORIES

List of contents:

Home page **Previous**

MODELLING MIGRATION HISTORIES

Introduction

- This example is concerned with individuals' migration histories within Great Britain, where migration is a residential move between two localities.
- Boundary choice is crucial in defining what is a migration move (White and Meuser, 1988).
- In this analysis migration is taken as an inter-county move. It is therefore concerned with moves which involve breaking away from social and community ties.
- For a recent text on migration see for instance Boyle, Halfacree and Vaughan (1998).

The data

- The data are derived from a large retrospective survey of life and work histories carried out in 1986 under the Social Change and Economic Life Initiative (SCELI), funded by the ESRC.
- The data were therefore not specifically collected for the study of migration, but were drawn from an existing data set which includes information on where individuals had lived all their working lives.
- The variables selected from the primary data set are those which are suggested in the research literature as important for explaining individual migration behaviour.
- Temporary moves of a few months duration do not imply commitment to a new area and are not regarded as migration. Migration data are therefore recorded on an annual basis.
- The respondents were aged 20 to 60 and lived in the travel-to-work area of Rochdale, just to the north of Manchester. (Rochdale was one of six localities chosen for the SCELI survey for their contrasting experience of recent economic change.)
- As the analysis is concerned with internal migration within Great Britain, individuals who had lived abroad during their working lives are excluded from the data set.
- The information for 1986 is incomplete and is therefore not included.
- The data set contains the migration histories of 348 males during their working, or potentially working lives, starting from the completion of education up to 1985.
- The data set is longitudinal, with one observation for each individual per calendar year. There are a total of 6349 annual observations.
- The start year for the collection of data for each individual is different, but the final year is the same.
- The response variable of interest is binary, indicating for each individual and for each calendar year, whether or not there was a migration move.
- The explanatory variables are age, calendar year, duration of stay at each address, education, and information on marriage, children, employment and occupational status for each year.

NEXT:The longitudinal data set

Home page | Contents | Previous

The longitudinal data set

Typical data matrix

The longitudinal data set is stored in the file *rochmig.dat*. The data matrix for a typical individual is of the form:

Case number	50016								
Move/No move	0	0	1	0	1	1	1	1	0
Age	17	18	19	20	21	22	23	24	25
Year	77	78	79	80	81	82	83	84	85
Duration of stay (dur)	1	2	3	1	2	1	1	1	1
Education (ed)	4	4	4	4	4	4	4	4	4
Children age 11-12 (ch1)	0	0	0	0	0	0	0	0	0
Children age 13-14 (ch2)	0	0	0	0	0	0	0	0	0
Children age 15-16 (ch3)	0	0	0	0	0	0	0	0	0
Children age 17-18 (ch4)	0	0	0	0	0	0	0	0	0
Marital status (msb)	1	1	1	1	1	1	1	1	1
Marital status (mse)	1	1	1	1	1	1	1	1	1
Employment status (esb)	7	7	7	7	7	7	7	0	0
Employment status (ese)	7	7	7	7	7	7	0	0	0
Occupational status (osb)	71	71	71	71	71	71	71	0	0
Occupational status (ose)	71	71	71	71	71	71	0	0	0
Marital break-up (mbu) ***	0	0	0	0	0	0	0	0	0
Remarriage (mrm) ***	0	0	0	0	0	0	0	0	0
First marriage (mfm) ***	0	0	0	0	0	0	0	0	0
Marital status (msb1) {msb collapsed} ***	1	1	1	1	1	1	1	1	1
Promotion to manager (epm) ***	0	0	0	0	0	0	0	0	0
Obtaining a job (eoj) ***	0	0	0	0	0	0	0	0	0
Employment status (esb1) {esb collapsed} ***	3	3	3	3	3	3	3	4	4
Promotion to service class (ops) ***	0	0	0	0	0	0	0	0	0
Occupation (osb1) {osb collapsed} ***	2	2	2	2	2	2	2	1	1
Marital status (msb2) {msb1 collapsed} ***	0	0	0	0	0	0	0	0	0
Employment (esb2) {esb1 collapsed} ***	2	2	2	2	2	2	2	3	3
Occupation (osb2) {osb1 collapsed***	1	1	1	1	1	1	1	1	1
Occupation (osb3) {osb2 collapsed} ***	0	0	0	0	0	0	0	0	0

The core variables are marked in bold; other variables have been derived from these and are marked with asterisks. Some are new variables which indicate a change in marital, occupational or

employment status during the year, - these are seen as important in explaining the dynamics of migration - , others are simplified versions of the core variables, formed by collapsing categories.

For a **detailed description of the variables** click here.

Limitations of the data set

- The data is restricted to those residing in the study area in 1986; it includes individuals who had moved to Rochdale before 1986, but not those who had moved away. Therefore those who had left cannot be compared with those remaining.
- The data contains the complete, or nearly complete histories for those aged sixty at the time of interview but only short histories for younger respondents.
- Therefore the data are comparatively sparse on migration behaviour during later career stages and during the more distant past. For earlier periods the maximum age is reduced.
- There is no information on retirement or post-retirement migration.
- As the data were not specifically collected for studying migration, some explanatory variables which may be important, such as family income for instance, were not available.
- The reliability of retrospective data may also be called into question (Dex 1995; Dex and McCulloch 1998).

Do we need such a large and complex longitudinal data set to answer the substantive questions?

We can sum up the number of migrations for each individual and produce a summary data set, with one line of information for each individual. This will give cross-sectional information for the years up to 1985.

What questions can be answered by cross-sectional analysis?

Next:Cross-sectional data

Home page | Contents | Previous

Cross-sectional data

Summarizing the data

For each individual, we can sum the number of migrations recorded in the survey, to produce one line of information containing:

- Case number
- Number of migrations since leaving school (n)
- Time (t), number of years since leaving school

Only time independent explanatory variables are included in these cross-sectional data.

- Educational qualification (ed), with 5 levels:
  1=Degree or equivalent; professional qualifications with a degree
  2=Education above A-level but below degree level; includes professional qualifications without a degree
  3=A-level or equivalent
  4=Other educational qualification
  5=None

The data matrix for the individual shown on the longitudinal data page can be summarized as follows:

case number	n	t	ed
50016	5	9	4

This person is one of the eight in the data set to have 5 migrations during the time in the survey. See Table 1.
The data sets can be downloaded from here. The cross-sectional data set is available in the file *rochmigx.dat*.

TABLE 1: Observed migration frequencies

Number of moves	0	1	2	3	4	5	>=6
Observed frequency	228	34	42	17	9	8	10

Table 1 summarizes the observed migration frequencies for the 348 respondents in the sample.

As the individuals ranged in age from 20 to 60, they had varying lengths of migration history.

If complete randomness in migration behaviour is assumed, then a Poisson model may be used to represent the aggregate count data.

NEXT:The Poisson model

Home page | Contents | Previous

Cross-sectional analysis:Poisson model for aggregate count data

The Poisson model

If complete randomness in migration behaviour is assumed, then a Poisson model may be used to represent the aggregate count data. Strictly, we should use a Binomial model as each individual is only allowed one migration per year so that the total number of migrations has an upper limit. However, for a large sample and and a low migration rate the Poisson model provides a good approximation.

For a homogeneous population, the probability of obtaining n_i outcomes in time t_i may be written as

$$Pr(n_i)=(m_i)^{n}_{i} \exp(-m_i) / n_i!$$

where m_i is the mean (or expected) number of migrations in time t_i.

For a constant annual migration rate r,

$$m_i=r*t_i$$

or

$$\log(m_i)=\log(r)+\log(t_i)$$

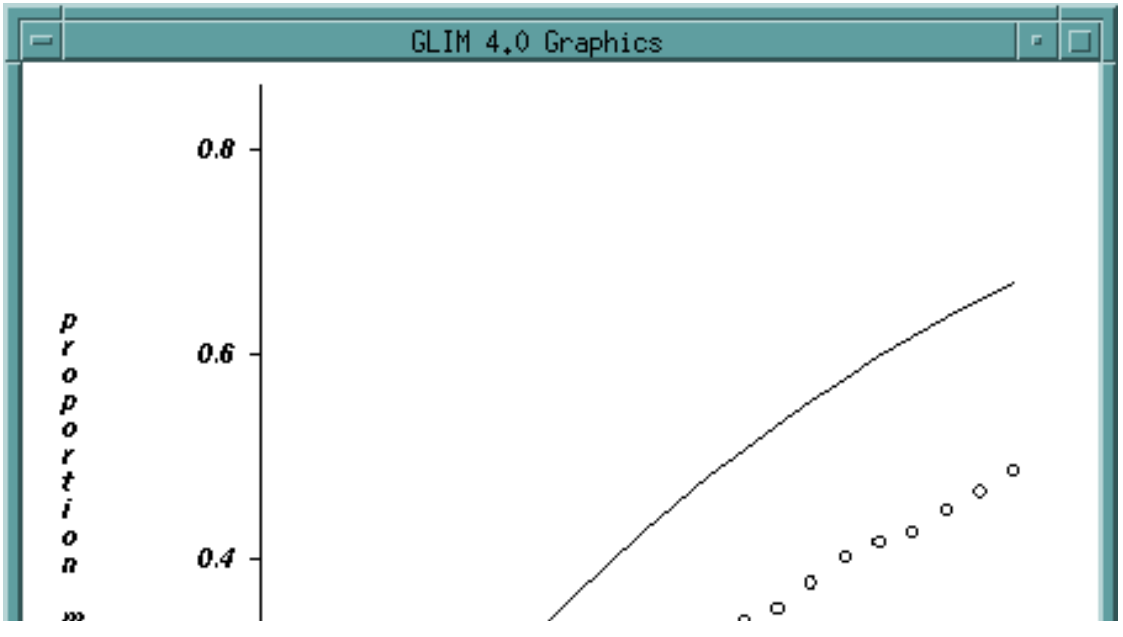
This model is an example of a generalised linear model. We will see how to fit such models a little later. When this model is fitted (using *log(r)* as an *OFFSET* in SABRE), the average annual migration rate comes out as 0.049 moves per individual per year.

For the time being, we note that this figure can also be calculated by simply dividing the total number of moves in the data set by the total time exposure to migration opportunities for the sample. Thus, there are 312 moves and 6349 annual observations, giving an average of 0.049 moves per individual per year.

This implies that each year a proportion of 0.049 of the population (or 4.9%) migrates, and that a proportion of 0.951 (or 95.1%) remains.

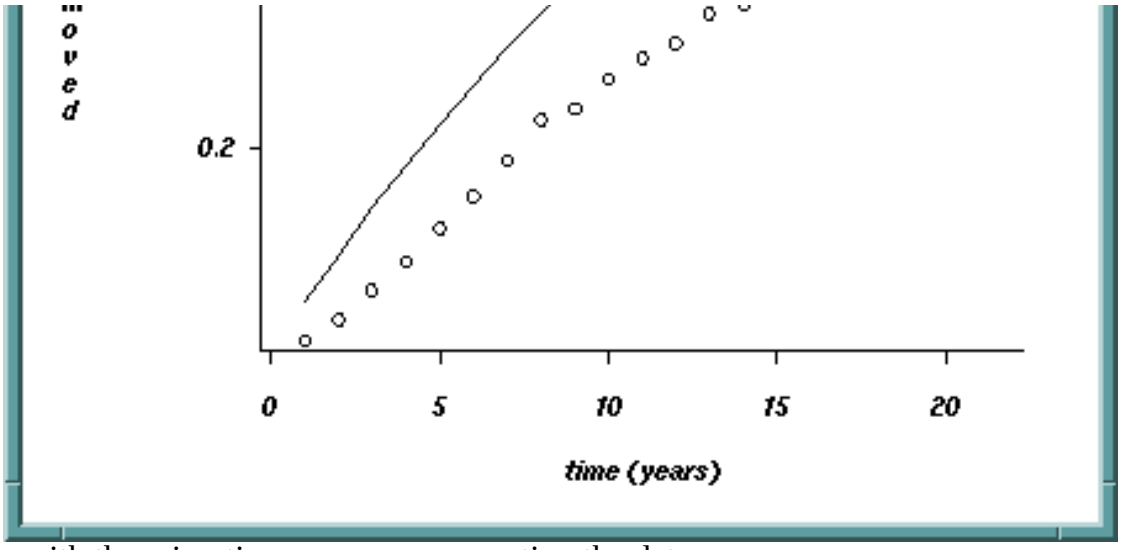
Using this model, the projected proportion moving at least once over a period of T years is equal to $[1-(0.951)^{T}]$.

The projected proportion migrating over different time periods is shown by the line on the graph. It is considerably higher than the observed proportion calculated from the data, which is indicated by circles.



It is evident that this model substantially and systematically overpredicts the proportion moving, and therefore underestimates population stability. This is a

consequence of assuming that migration behaviour over one time period can be used to predict migration behaviour over a longer time period, and is an example of a general problem, which Coleman (1973) calls the "deficient diagonal" effect.



The assumption that all individuals have the same propensity to migrate, which is not subject to change over time, does not seem compatible with the migration processes generating the data.

Allowing the migration rate to vary with time

The migration rate can be allowed to vary systematically with time in this simple model by replacing (t_i) in the above equation by (t_i) $^{b}{}_{1}$. Now the migration rate decreases through migration history if b_1 is less than 1 and increases if b_1 is greater than 1. One reason why we may expect b_1 to be less than 1 is due to *inertia* effects, with people increasingly less likely to move with duration in a specific locality.

It is convenient to write

$$r=\exp(b_0)$$

where b_0 is an unknown constant, and the exponentiation ensures that r is always non-negative.

The mean number of migrations may now be written as:

$$m_i=\exp(b_0)*(t_i)^{b_1}= \exp(b_0+b_1*\log(t_i))$$

or

$$\log(m_i)=b_0+ b_1*\log(t_i)$$

This model is typical of a **generalised linear model**, which contains:

1. a linear regression function or linear predictor in the explanatory variables, $[b_0+b_1*\log(t_i)]$,
2. a transformation, (logarithmic), which relates the linear predictor to the mean m_i,
3. a response variable n_i, which has a Poisson distribution with mean m_i.

The model may be fitted using SABRE software as follows. To run the example interactively, you will need to download the SABRE software and data sets.

SABRE SESSION:INPUT AND OUTPUT

```
C read in variables from data file
data case n t ed
read rochmigx.dat

        348 observations in dataset
```

```
C declare response variable
yvar n
C declare model
poisson yes
C calculate log(time)
transform ltime log t
C fit Poisson model with intercept
C and log(time) as explanatory variable
lfit int ltime
```

Iteration	Deviance	Reduction
1	1299.5140	
2	754.34418	545.2
3	658.72919	95.61
4	648.79228	9.937
5	648.49783	0.2945
6	648.49747	0.3547E-03
7	648.49747	0.5484E-09

```
C display parameter estimates
dis est
```

Parameter	Estimate	S. Error
int	-3.2884	0.35114
ltime	1.0887	0.11119

```
C display model fitted
dis m
```

X-vars	Y-var
int	n
ltime	

```
    Model type: standard Poisson log-linear

    Number of observations              =    348

    X-vars df=  2

    Deviance = 648.49747 on 346 residual degrees of freedom

stop
```

Results and conclusion

1 The estimated coefficient b_1 of *ltime* is 1.0887, with a standard error of 0.1112, and is therefore not significantly different from 1. The migration rate does not appear to decline or increase through migration history, but is constant.

Table 2:Observed and expected frequencies

Number of moves	0	1	2	3	4	5	=6
Observed frequency	228	34	42	17	9	8	10
Expected frequency	164.3	101.6	50.4	21.1	7.5	2.3	0.80

2 The observed migration frequencies are compared in Table 2 with the values predicted by the Poisson model. The model does not seem to fit the data, with the number of individuals making no moves or making four or more moves substantially underpredicted. There appears to be a systematic variation in migration frequency over and above the variation attributable by chance.

3 The fit of the model may be assessed by comparing the value of the sum of [(Expected frequency-Observed frequency) ^{2}/Expected frequency] with the χ^2 distribution on 5 degrees of freedom (7 cells - 2 estimated coefficients). The critical value at the 5% significance level is 11.07. The calculated value is in fact 192.5, an order of magnitude higher.

4 The degree of model misspecification may be measured by the dispersion parameter, which is the ratio of the scaled deviance and the residual degrees of freedom.(648.5/346)=1.87). If the model were well specified, this ratio would be approximately 1.

5 One explanation for the poor fit of the model is that the assumption of a *homogeneous* population is not valid. Individuals may vary in their likelihood of migration; the assumption of a migration rate which depends only on time may be incorrect. Thus, it may be possible to improve the model specification by including explanatory variables which distinguish between individuals.

Next:Poisson model with explanatory variable

Home page | Contents | Previous

Cross-sectional analysis:Poisson model with explanatory variables

Introduction

The Poisson model may be used for inference about explanatory variables even when the model is seriously misspecified, provided that:

1. The explanatory variables do not change over the migration histories.
2. Interest focuses on the relationship between the explanatory variables and the rate of migration.

Education is recognised as the single most important individual-level factor governing rates of internal migration, as it is related to the opportunity to progress in careers. (Sandefur and Scott, 1981; Goss, 1985; Liaw, 1990)

Five levels of educational attainment are available in the data, and may be included in the Poisson model.

The model

The previous equation for the mean number of migrations

$$\log(m_i)=b_0+ b_1*\log(t_i)$$

may be extended by writing:

$$\log(m_i)=b_0+ b_1*\log(t_i)+b_2*x_{i1} +b_3*x_{i2}+b_4*x_{i\,3}+b_5*x_{i4}+b_6*x_{i\,5}$$

where x_{ij}=1 if individual *i* has educational qualification *j* and 0 otherwise. These x_{ij} are known as dummy variables. SABRE constructs dummy variables internally for any variable defined as a factor.
Education has 5 levels: *j*=1 is the reference group, with no qualifications. The coefficient estimate for this level is absorbed into the intercept term and b_2 is set to zero by SABRE; the parameter estimates of the higher levels (b_3,b_4,b_5 and b_6) provide appropriate contrasts with this level.

We now add the 5-level factor **educational qualification** to the previous model.
For the lowest level to correspond to 'No qualifications', the educational levels in the data, which are coded 1 for 'Degree or equivalent' and 5 for 'No qualifications', are reversed. This is done by two *transform* commands.

SABRE SESSION:INPUT AND OUTPUT

```
data case n t ed
read rochmigx.dat

        348 observations in dataset

transform ltime log t
C reverse order of levels for ed in two stages
transform ned ed - 6
transform reved ned * -1
C check reversed levels
look ed reved

          ed          reved
      ------------------------
     1   4.000        2.000
     2   4.000        2.000
     3   5.000        1.000
     4   4.000        2.000
     5   3.000        3.000
     6   5.000        1.000
     7   2.000        4.000
```

```
      8   4.000       2.000
      9   5.000       1.000
     10   4.000       2.000
     11   3.000       3.000
     12   5.000       1.000
     13   2.000       4.000
     14   3.000       3.000
     15   4.000       2.000
     16   2.000       4.000
     17   5.000       1.000
     18   5.000       1.000
     19   3.000       3.000
     20   3.000       3.000

C convert variable reved to factor fed
C and fit previous model
fac reved fed
yvar n
poisson yes

lfit int ltime

    Iteration        Deviance        Reduction
    ---------------------------------------------
        1           1299.5140
        2           754.34418        545.2
        3           658.72919        95.61
        4           648.79228        9.937
        5           648.49783       0.2945
        6           648.49747       0.3547E-03
        7           648.49747       0.5484E-09

C now add in education
lfit +fed

    Iteration        Deviance        Reduction
    ---------------------------------------------
        1           1297.1251
        2           748.76297        548.4
        3           649.04377        99.72
        4           637.92142        11.12
        5           637.56670       0.3547
        6           637.56619       0.5089E-03
        7           637.56619       0.1140E-08

dis est

    Parameter              Estimate         S. Error
    ---------------------------------------------------
    int                    -3.7435          0.39195
    ltime                   1.1610          0.11553
    fed   ( 1)                 0.           ALIASED [I]
    fed   ( 2)             0.35868          0.13633
    fed   ( 3)            -0.15726E-01      0.24772
    fed   ( 4)             0.49562          0.22760
    fed   ( 5)             0.40762          0.20645

dis m

    X-vars      Y-var
    ----------------
    int         n
    ltime
    fed
```

```
    Model type: standard Poisson log-linear

    Number of observations             =    348

    X-vars df          =     6

    Deviance          =637.56619 on 342 residual degrees of freedom
    Deviance decrease =10.931280 on   4 residual degrees of freedom

stop
```

Results and conclusion

1. The addition of educational qualification to the model has reduced the deviance from 648.49 to 637.56 i.e. by 10.93 on 4 degrees of freedom. This is significant at the 5% level when compared with $\chi^2_{(4)}=9.49$.
   Thus, the addition of educational qualification appears to produce a modest improvement on the fit of the Poisson model.
2. The estimated coefficient of *ltime* is still close to 1; the migration rate again appears to be constant over time.
3. The coefficient estimate for the reference level of educational attainment shown as fed(1) has been absorbed into the **intercept** term.
   The coefficient estimates of other levels *j* give the difference between the reference level and level *j*. Due to the logarithmic link, the additive effect of b_j on the linear predictor, has a multiplicative effect of $\exp(b_j)$ on mean migration rates. For example fed(2), estimated as 0.35868, produces a multiplicative effect of exp(0.35868)=1.4 on the migration rate. Starting with the highest educational level, the multiplicative effects are as follows:

Education	Multiplicative factor
Degree or equivalent	1.5
Other higher education	1.6
A-level or equivalent	1.0
Other educational qualification	1.4
No qualification	1.0

4. These results do provide some evidence of migration propensity increasing with education, though the standard errors of the coefficient estimates are relatively large and the results are somewhat anomalous.
   This may be a particular feature of this data set, or it is possible that some explanation for the anomalies could be found if more precise categories of educational qualifications were available.
5. It must also be noted that there is no control for other variables which might influence migration behaviour and which may be correlated with the level of education.
6. The dispersion parameter, which is the ratio of the scaled deviance to the residual degrees of freedom = 637.566/342=1.86 has only slightly been reduced.
7. It is clear that adding educational qualification to the model, accounts only in a small way for the differences between individuals.

How can we control for other differences?

Next:A Mixture model for cross-sectional data

Home page **Contents** **Previous**

Allowing for unmeasured heterogeneity: a mixture model for cross-sectional data

Omitted explanatory variables

Educational qualification accounts only in a small way for the heterogeneity (ie. the variation in migration behaviour) of the population. Other important individual differences have not been measured, or indeed may be unmeasurable.

To model heterogeneity in migration propensity due to unmeasured and unmeasurable factors, we add an individual specific term, or nuisance parameter, e_i to the linear predictor, to represent the omitted explanatory variables. This term is assumed to be constant for each individual over time. The conventional assumption is that e_i is distributed independently of the included variables. The model equation, with 5 levels of educational attainment as before, becomes:

$$\log(m_i)=b_0+ b_1*\log(t_i)+b_2*x_{i1} +b_3*x_{i2}+b_4*x_{i\,3}+b_5*x_{i4}+b_6*x_{i\,5}+e_i$$

The mixture model

The term e_i which represents the effect of the omitted variables for each individual *i* is assumed to have some probability distribution over the population. This distribution has to be modelled in addition to the Poisson model for the count data. The model is now said to have a **mixing distribution** ; or alternatively the model is called a **random effects** or a **mixture model**.
Different methods may be used to fit mixture models, depending on the assumptions made about the probability distribution of the error terms. SABRE uses a standard approach (see for example Lancaster and Nickel 1980; Heckman and Singer 1984).

SABRE assumes a Normal distribution for e_i, with mean zero and variance σ^2, and uses a Gaussian quadrature method to fit the model. The tails of the Normal distribution cause a problem, as they assume zero probability at the extremes of the distribution. In fact, there is strong evidence that there are individuals for whom, in many situations, there will be a finite probability of never taking part in the process under investigation. These are the "stayers"; in the context of migration, these are the people who are likely never to move (over and above those who, by chance, do not move in the period covered by the study).

SABRE can allow for "stayers" by supplementing the quadrature mass points with endpoints at plus and minus infinity when this is appropriate. In this model, a nuisance parameter value of minus infinity implies zero probability of migration for that individual.

The standard SABRE mixture model is fitted using the *FIT* command, and includes endpoints by default. For the Poisson model, a single endpoint at minus infinity is included, which estimates the proportion of stayers. There is an option to omit the endpoints from the model and to allow the standard Poisson-Normal mixture model to be fitted, by using the *ENDPOINT* command. The parameterisation of the model is given in the SABRE reference guide.

We fit the log-linear Poisson-Normal mixture model for count data, first with endpoints and second without endpoints as follows:

Model with endpoints

SABRE SESSION:INPUT AND OUTPUT

```
data case n t ed
read rochmigx.dat
```

```
        348 observations in dataset
```

```
transform ltime log t
C reverse order of levels for ed
transform ned ed - 6
transform reved ned * -1
fac reved fed
poisson y
yvar n
C fit random effects model
C endpoints fitted by default
fit int ltime fed
```

```
    Initial Log-Linear Fit:

    Iteration        Deviance        Reduction
    ---------------------------------------------
        1           1297.1251
        2           748.76297         548.4
        3           649.04377         99.72
        4           637.92142         11.12
        5           637.56670        0.3547
        6           637.56619       0.5089E-03
        7           637.56619       0.1140E-08


    Iteration      Deviance          Step       End-point    Orthogonality
                                     length                    criterion
    ------------------------------------------------------------------------
        1         549.93673         1.0000        free          13.255
        2         531.94684         1.0000        free        0.28295E-01
        3         529.77935         0.0156        free          7.2279
        4         522.42322         0.5000        free          24.948
        5         495.13658         1.0000        free          16.832
        6         487.98913         1.0000        free          3.9855
        7         486.09574         1.0000        free          72.511
        8         486.07703         1.0000        free          15.212
        9         486.07703         1.0000        free
```

```
dis est
```

```
    Parameter               Estimate         S. Error
    ---------------------------------------------------
    int                     -2.6932          0.57967
    ltime                   0.97307          0.15646
    fed   ( 1)                   0.          ALIASED [I]
```

```
    fed   ( 2)             0.44283          0.18502
    fed   ( 3)            -0.34053E-01      0.32219
    fed   ( 4)             0.67497          0.32448
    fed   ( 5)             0.32705          0.27775
    scale                  0.45004          0.13086
                                                          PROBABILITY
                                                          ___________
    end-point 0            0.92752          0.19029       0.48120

dis m

    X-vars      Y-var       Case-var
    ____________________________________
    int         n           case
    ltime
    fed

    Model type: standard Poisson log-linear normal mixture with end-point

    Number of observations              =    348
    Number of cases                     =    348

    X-vars df          =     6
    Scale df           =     1
    End-point df       =     1

    Deviance      =   486.07703   on   340 residual degrees of freedom

fit -fed

    Iteration      Deviance       Step       End-point      Orthogonality
                                  length                      criterion
    ____________________________________________________________________
        1         619.14491      1.0000       free            91.345
        2         521.04026      1.0000       free            28.699
        3         497.87490      1.0000       free            23.435
        4         494.73843      1.0000       free            5.2864
        5         494.52771      1.0000       free            8.9229
        6         494.49902      1.0000       free            3.4139
        7         494.49442      1.0000       free            5.9225
        8         494.49442      1.0000       free

dis m

    X-vars      Y-var       Case-var
    ____________________________________
    int         n           case
    ltime

    Model type: standard Poisson log-linear normal mixture with end-point

    Number of observations              =    348
    Number of cases                     =    348

    X-vars df          =     2
```

```
Scale df           =     1
End-point df       =     1

Deviance          =  494.49442 on 344 residual degrees of freedom
Deviance increase =  8.4173895 on   4 residual degrees of freedom
```

Results and conclusion

1. The addition of the individual specific random term and left endpoint to the model has reduced the deviance from 637.56 to 486.08 ie. by 151.48 on 2 degrees of freedom. Although the χ^2 test is not strictly correct, as the standard Poisson model lies on the boundary of the parameter space of the Poisson mixture model, such a large reduction in deviance indicates a significant improvement in model fit. There appears to be considerable residual heterogeneity in the population.
2. The dispersion parameter has decreased to 486.08/340=1.43, confirming the improved fit.
3. The parameter estimates have changed little (by approximately one standard error); the standard errors of the parameter estimates have all increased. This result is typical when comparing models with and without unmeasured heterogeneity, provided all the explanatory variables are exogenous. We leave a discussion of the term *exogenous* until slightly later in this example.
4. Even though the standard Poisson model seems misspecified, the parameter estimates are *consistent*, ie. they *tend to the true values* when the sample size is increased. However, standard errors are underestimated and may lead us to conclude that an explanatory variable is significant, when in fact it is not. For instance, in the standard Poisson model, as the ratio of the parameter estimate to the standard error (t-ratio) for **fed(5)** is at about the 5% significance level of 2 , we might conclude that this factor is significant, whereas in the Poisson mixture model it is well below the 5% significance level, indicating that this factor is in fact not significant.
5. The small increase in deviance (8.42) compared to $\chi^2_{(4)}$=9.49 at the 5% level, when educational qualification is removed from the model confirms that education is not significant in the Poisson mixture model.
6. The scale parameter estimate is the standard deviation of the Normal distribution assumed for the individual specific terms e_i. It is significantly different from zero and indicates considerable residual heterogeneity.
7. Note the parameter estimate for the left endpoint. The parameter value of 0.9275 (standard error 0.1903) is significantly different from zero, and the associated probability of 0.48 suggests that the sample contains a significant number of "stayers".

Model without endpoints

We now continue the SABRE session, remove endpoints and refit the full model.

SABRE SESSION:CONTINUED

```
C put back fed
fit +fed
```

Iteration	Deviance	Step length	End-point	Orthogonality criterion
1	668.03445	1.0000	free	29.768
2	528.74742	1.0000	free	22.714
3	496.35404	1.0000	free	31.771

4	487.11433	1.0000	free	3.2170
5	486.70328	1.0000	free	12.330
6	486.41714	1.0000	free	8.5039
7	486.07846	1.0000	free	5.3708
8	486.07703	1.0000	free	6.6935
9	486.07703	1.0000	free	

```
dis m
```

```
    X-vars      Y-var       Case-var
    ____________________________________
    int         n           case
    ltime
    fed

    Model type: standard Poisson log-linear normal mixture with end-point

    Number of observations             =   348
    Number of cases                    =   348

    X-vars df          =     6
    Scale df           =     1
    End-point df       =     1

    Deviance           = 486.07703     on   340 residual degrees of freedom
    Deviance increase  =  8.4173895     on     4 residual degrees of freedom
```

```
C fit same model without endpoints
endpoint no
fit .
```

Iteration	Deviance	Step length	End-point	Orthogonality criterion
1	694.22855	1.0000	fixed	26.564
2	551.25040	1.0000	fixed	25.027
3	533.52981	1.0000	fixed	16.291
4	513.44427	1.0000	fixed	6.3297
5	511.82728	1.0000	fixed	8.5030
6	511.28156	1.0000	fixed	14.935
7	511.01450	1.0000	fixed	6.4983
8	511.00122	1.0000	fixed	4.4092
9	511.00114	1.0000	fixed	

```
dis est
```

Parameter	Estimate	S. Error
int	-4.5013	0.58650
ltime	1.1857	0.17733
fed (1)	0.	ALIASED [I]
fed (2)	0.26548	0.22422
fed (3)	0.16689	0.35579
fed (4)	0.51855	0.35699
fed (5)	0.61071	0.45804
scale	1.1940	0.99342E-01

```
dis m

    X-vars      Y-var       Case-var
    ____________________________________
    int         n           case
    ltime
    fed

    Model type: standard Poisson log-linear normal mixture

    Number of observations              =   348
    Number of cases                     =   348

    X-vars df           =     6
    Scale df            =     1

    Deviance            = 511.00114  on   341 residual degrees of freedom
    Deviance increase = 24.924106 on   1 residual degrees of freedom
```

Conclusion

When the same model is fitted without endpoints, the deviance *increases* by 24.9 on a change of 1 degree of freedom. Although the χ^2 test is again not strictly applicable, such a large change in deviance ($\chi^2_{(1)}$=3.84 at the 5% level) indicates that unobserved heterogeneity is in excess of that reflected by the Normal distribution. The model fits significantly better when allowance is made for "stayers".

What have we learnt from cross-sectional data analysis?

Next:Conclusions from cross-sectional analysis

Home page | Contents | Previous

Conclusions from cross-sectional data analysis

Summary

- Extrapolation of mean annual migration rates leads to an underprediction of population stability. This is because in a heterogeneous population, the individuals who are most likely to move, and who contribute to the mean annual migration rate will have moved away, leaving behind those who are less likely to move.
- Cross-sectional analysis of this data set does not indicate any systematic variation of the mean migration rate with time. Even for data sets which showed evidence of temporal variation, there would be no indication of whether this was due to age, cohort or inertial effects.
- Even though the standard Poisson model seems misspecified, because all the explanatory variables are exogenous the parameter estimates are *consistent*, ie. they tend to the true values when sample size is increased. However, standard errors are underestimated and may lead us to conclude that an explanatory variable is significant, when in fact it is not. For instance, results for the standard Poisson model suggest that educational qualifications do affect the likelihood of migration; the Poisson mixture model does not indicate significant educational qualification effects.
- There is evidence that the likelihood of migration varies markedly between individuals and that the sample contains a number of "stayers", individuals likely never to move.
- With a single count of outcomes for each individual, it is *impossible* to distinguish between a heterogeneous population, with some individuals having a consistently high and others a consistently low propensity to migrate, and a truly contagious process, in which an individual's experience of migration per se increases the probability of subsequent migration.

It is clear that the analysis of the cross-sectional data has answered only a few of the **substantive questions** of interest. No light has been shed on the *dynamics* of the migration process.
Longitudinal data analysis of individual event histories is necessary to explore the temporal variation in individual migration rates and to identify, for example, inertial effects.

Next:Introduction to longitudinal data analysis

Home page | **Contents** | **Previous**

Longitudinal data analysis: Introduction

The longitudinal data set

The response variable is now binary, indicating for each calendar year whether or not there was a migration move. As temporary moves of a few months duration do not imply commitment to an area, they are not considered as migration. Therefore migration events are recorded on an annual basis, with at most one move per year. We do not use annual count data.

We can now use time-varying explanatory variables. The variables age, calendar year, duration of stay and the presence of children of secondary age in the family are recorded each year, while marital status, employment status and occupational status are recorded at the beginning and end of each year. Other explanatory variables are derived from the raw data; some indicate a change in the status variables during the year, others have been created by collapsing categories of certain variables.

We look at the marital, employment and occupational status variables both at the beginning and at the end of each year, as it may be either the original status, the destination status or a change in status during the year which influences individual migration.

It is important to distinguish between two types of explanatory variable: an **endogenous** explanatory variable, which is in some way a function of an earlier outcome of the process under study, and an **exogenous** explanatory variable, in which there is no such relationship.

In this data set *duration of stay* is an endogenous explanatory variable, because the number of years of residence since the last migration move is related to the timing of that move.

Residual heterogeneity

Longitudinal data consist of repeated observations on each individual. The observations are independent between individuals, but correlated within individuals. The differences between individuals are measured by a range of explanatory variables which may differ over time. In practice not all the variables that characterize individuals are observable, and the omitted variables give rise to a residual heterogeneity.

In the cross-sectional analysis, as all explanatory variables were exogenous, the parameter estimates were consistent even though the standard Poisson model was misspecified. This is not the case for cross-sectional or longitudinal analyses if there are endogenous explanatory variables.

In the presence of endogenous explanatory variables, such as *duration of stay*, inference about temporal variation requires an explicit representation of residual heterogeneity, otherwise parameter estimates will be biased. This is only possible with longitudinal data; the problems posed by endogenous variables cannot be overcome using cross-sectional data.

The model

The response variable y_{it} is binary, defined as 1 if the individual *i* migrates in year *t*, and 0 otherwise. It has a Bernoulli probability distribution with

$$\Pr(y_{it}) = p_{it}^{y_{it}} (1-p_{it})^{1-y_{it}}$$

where p_{it} is the probability of a migration move by individual *i* in year *t*. The relation between p_{it} and the explanatory variables is made through a suitable linear predictor and the **logistic link** function. This transforms the linear predictor of explanatory variables, which may have any value between plus and minus infinity, to a probability which necessarily lies between zero and one.

Using the logistic link function $\log[p_{it}/(1-p_{it})]$, the **simple logistic** regression model is:

$$\log[p_{it}/(1-p_{it})] = \underline{\beta}' \underline{x}_{it}$$

where $\underline{\beta}' \underline{x}_{it} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \ldots$.

$\underline{\beta}' \underline{x}_{it}$ is a shorthand (vector) way of denoting the linear predictor, which may contain a large number of explanatory variables.

This can be rewritten as

$$p_{it} = \exp(\underline{\beta}' \underline{x}_{it}) / [1+\exp(\underline{\beta}' \underline{x}_{it})]$$

and the model including **residual heterogeneity** as

$$p_{it} = \exp(\underline{\beta}' \underline{x}_{it} + e_i) / [1+\exp(\underline{\beta}' \underline{x}_{it} + e_i)]$$

where $\underline{x}_{it}$ is a vector of explanatory variables, $\underline{\beta}'$ is a vector of unknown parameters and e_i is an individual specific term summarizing the effect of the omitted variables.

The large number of possible explanatory variables in the longitudinal data set require a pragmatic approach to model building. We first model the temporal variation.

Next:Longitudinal analysis: Temporal variation

Home page | Contents | Previous

Longitudinal data analysis: Temporal variation

As a first step we model the temporal variation, and fit models both with and without residual heterogeneity and compare them.

Temporal variation

The dynamic characteristics of the data are described by the three temporal explanatory variables: age, year, and duration of stay. Cohort effects are subsumed in the year and age components. Alternatively, it would be possible to reparameterise the model so that age and cohort rather than age and year effects are estimated. This would not affect the goodness of fit of the model.

Year effects are caused by external economic and social changes generating fluctuations in aggregate migration over time.

The variation of migration propensity with *age* is related to life cycle factors, such as marriage and children, and to career progression.

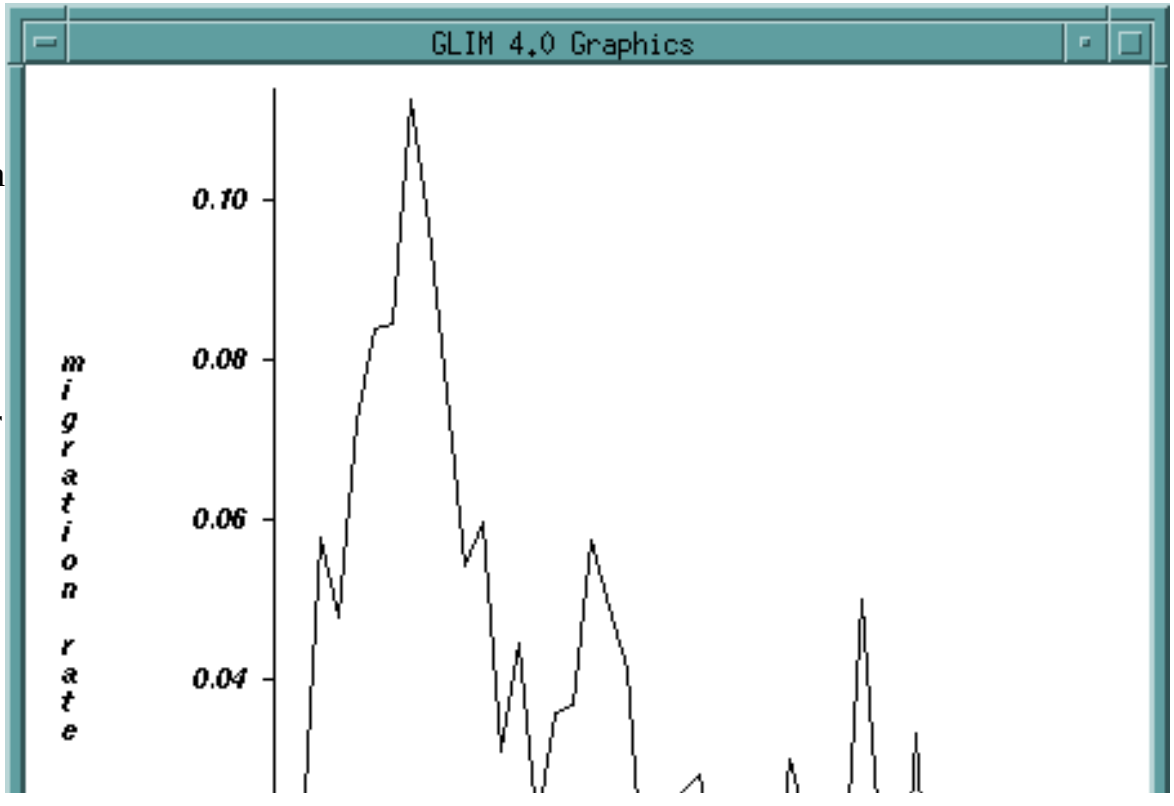
Duration of stay is a proxy variable for the many social, community and economic ties which strengthen with length of residence. It is a measure of *cumulative inertia*, which may compound the variation of migration propensity with age. (See Mc Ginnis, 1968; Huff and Clark, 1978.)

What functions of these explanatory variables are appropriate to use in the model?

We first explore the data to find a suitable starting point for model building.

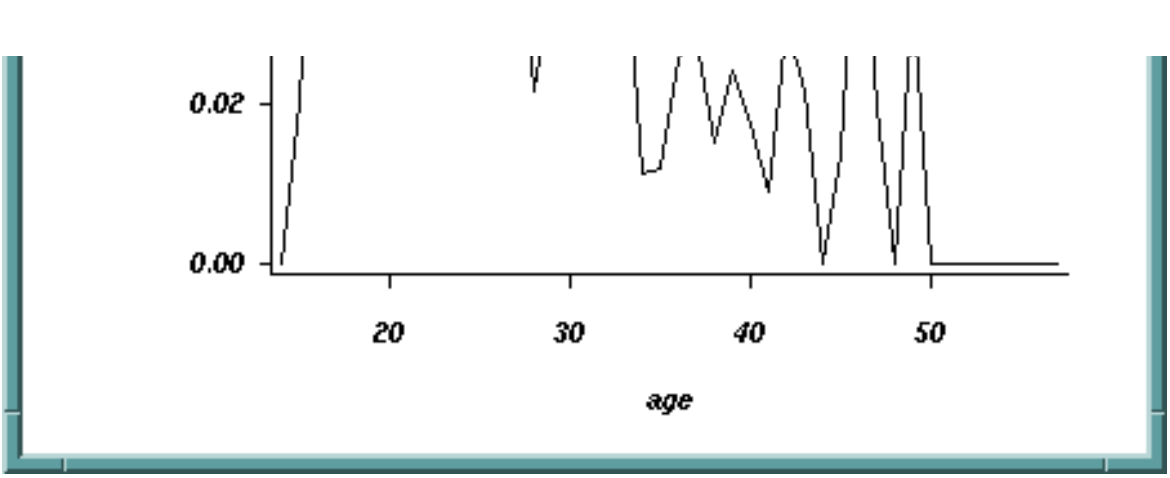
The age effect

As a first step, it is helpful to examine how the empirical mean migration rate varies with age. The mean migration rate is calculated by dividing the total number of moves by the total number of years of migration opportunity for each distinct age.



The results on the graph show a clear peak around age 20, some evidence of another peak at about 30 and at least two peaks close to each other just

under age 50. The latter peaks could be the result of fluctuations because the data are more sparse here.



It must be noted that there are no controls for other temporal variables in this graph. Nevertheless, there is evidence that the variation with age is multimodal (ie. has several peaks). This suggests using a polynomial representation of age in the models.

Modelling age, year and duration of stay as categorical variables

To explore how the migration rate varies with the three temporal variables, we split each variable into distinct categories, in such a way that we have a reasonable number of data points within each category. Thus the categories usually span five years, but are longer where the data are sparse near the edge of the data window. We fit the logistic model using these categories as levels of factors.

For age we choose cut-off points 20,25,30,35,40 and 45 years, so that the lowest category represents an age of less than 20 and the highest an age greater than 45. The cut-off points for year will be 55,60,65,70,75 and 80 and for duration of stay 5,10,15,20,25 and 30 years.
The model may be fitted using SABRE software as follows:

SABRE SESSION:INPUT AND OUTPUT

```
data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
osb ose mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
C convert variables to factors using the following
C cut-off points
factor age agegp 20 25 30 35 40 45
factor dur durgp 5 10 15 20 25 30
factor year yeargp 55 60 65 70 75 80
lfit int agegp  yeargp durgp

    Iteration        Deviance        Reduction
    ---------------------------------------------
        1           8801.5829
        2           2968.3684         5833.
        3           2335.8507         632.5
        4           2208.6718         127.2
        5           2187.8156         20.86
        6           2185.1153         2.700
        7           2184.8380        0.2772
```

```
        8           2184.8279       0.1014E-01
        9           2184.8278       0.2246E-04

dis est

    Parameter               Estimate         S. Error
    ------------------------------------------------------------
    int                      -2.1704          0.23184
    agegp ( 1)                    0.          ALIASED [I]
    agegp ( 2)               1.1042          0.16933
    agegp ( 3)              0.73531          0.21522
    agegp ( 4)               1.2723          0.23824
    agegp ( 5)               1.0235          0.32081
    agegp ( 6)               1.0312          0.42478
    agegp ( 7)               1.5378          0.51473
    yeargp( 1)                    0.          ALIASED [I]
    yeargp( 2)          -0.37839E-01          0.27795
    yeargp( 3)              -0.50404          0.28618
    yeargp( 4)              -0.74076          0.28944
    yeargp( 5)              -0.47078          0.27593
    yeargp( 6)              -0.86073          0.28758
    yeargp( 7)               -1.1719          0.28593
    durgp ( 1)                    0.          ALIASED [I]
    durgp ( 2)               -1.4236          0.15918
    durgp ( 3)               -1.9089          0.25098
    durgp ( 4)               -2.6716          0.38781
    durgp ( 5)               -4.1664           1.0210
    durgp ( 6)               -2.9408          0.77358
    durgp ( 7)               -3.0448           1.1063

stop
```

Results and conclusion

1 The parameter estimate of the intercept term refers to the lowest category of each categorical variable; the estimates for the higher levels give the contrasts between those categories and this reference level. The estimates for level 1 of each variable are therefore set to zero (and are said to be aliased).

2 Examination of the parameter estimates gives an indication of how the migration rate varies from category to category, when all three temporal variables are controlled for. For clarity the results are displayed on graphs.

3 The parameter estimates for **age** go up and down, rising three times as we go from category 1 to category 7 (Figure 1). This suggests including age in the model as a sixth order polynomial. We note that the age effect is likely to be better estimated at the lower ages than at the higher ages, because the data are sparse for the older age group.

4 For **year** there is a downward trend in parameter estimates, but with a small increase at category five (Figure 2). This may be a consequence of sparsity of data or it may show a real trend for these years. To allow for this rise and fall, we shall include year as a third order polynomial.

5 As **duration of stay** is increased, there is a general downward trend in parameter estimates, however the trend is not quite linear (Figure 3). The fluctuations at durations above 25 years may be due to sparsity of data. Plotting the parameter estimates against log duration (Figure 4) gives a more linear plot. This suggests trying this variable as either a linear or a logarithmic function.

6 From the parameter estimates we can calculate how the probability of migration varies with each of the explanatory variables for fixed values of the other two variables. Figure 5 illustrates the variation of the probability of migration with age in 1985 with duration of stay set to 10 years. Similar graphs may be plotted for the other variables.

Therefore the starting point for model building will be the following model:

age+age^2+age^3+age^4+age^5+age^6 +year+year^2+year^3+dur [or alternatively + log(dur)].

Next:Model development: A parsimonious main effects model for temporal data

Home page | Contents | Previous

Longitudinal data analysis: A parsimonious main effects model for temporal data

Model building strategy

In the first instance, we aim to find a **parsimonious** main effects model for the temporal variables. Using the results of our initial exploratory analysis we start by fitting the simple logistic model and comparing the fits of the following linear predictors:

$age+age^2+age^3+age^4+age^5+age^6 +year+year^2+year^3+dur$

and

$age+age^2+age^3+age^4+age^5+age^6 +year+year^2+year^3+\log(dur)$

We choose the better fitting model, and then fit a series of simple logistic models using a backward elimination technique. At each step we test if the removal of the least significant explanatory variable (lowest t-ratio) gives a significant deterioration in the model fit. If the removal of an explanatory variable results in an increase in deviance of less than 3.84 ie. $\chi^2_{(1)}$ at the 5% level, we exclude it from the model; otherwise it is retained.

Sabre analysis

SABRE SESSION:INPUT AND OUTPUT

```
data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
osb ose mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
transform age2 age * age
transform age3 age2 * age
transform age4 age3 * age
transform age5 age4 * age
transform age6 age5 * age
transform ldur log dur
transform year2 year * year
transform year3 year2 * year
lfit int dur year year2 year3 age age2 age3 age4 age5 age6

    Iteration        Deviance        Reduction
    _____________________________________________
        1           8801.5829
        2           2993.0684        5809.
        3           2373.6995        619.4
        4           2231.7859        141.9
        5           2195.4927        36.29
        6           2190.2053        5.287
        7           2190.0373       0.1680
        8           2190.0367       0.6502E-03
```

```
        9           2190.0367       0.1007E-06

dis est

    Parameter               Estimate         S. Error
    ____________________________________________________
    int                      -62.752          32.990
    dur                     -0.20904         0.17902E-01
    year                    -0.53834         0.94139
    year2                    0.71197E-02     0.14008E-01
    year3                   -0.33336E-04     0.68744E-04
    age                       11.740          4.8338
    age2                    -0.70751         0.33134
    age3                     0.20615E-01     0.10996E-01
    age4                    -0.29015E-03     0.17681E-03
    age5                     0.15811E-05     0.11036E-05
    age6                          0.         ALIASED [E]

C Extrinsic aliasing has occurred for age6.
C Fitting high order polynomials can often cause numerical problems.
C An option is to lower the tolerance for aliasing from the default value.
C As the parameter estimates for the higher order terms are very small
C We choose to transform 'age' to 'trage'=(age-30)/10, roughly
C standardising this variable.
C This is done in two stages.
transform tempage age - 30
transform trage tempage / 10
transform trage2 trage * trage
transform trage3 trage2 * trage
transform trage4 trage3 * trage
transform trage5 trage4 * trage
transform trage6 trage5 * trage
lfit int dur year year2 year3 trage trage2 trage3 trage4 trage5 trage6

    Iteration        Deviance        Reduction
    ____________________________________________
        1           8801.5829
        2           2992.7095        5809.
        3           2373.0803        619.6
        4           2230.9609        142.1
        5           2193.9097        37.05
        6           2187.7970        6.113
        7           2187.2527       0.5443
        8           2187.2013       0.5138E-01
        9           2187.2004       0.8804E-03
       10           2187.2004       0.3062E-06


dis est

    Parameter               Estimate         S. Error
    ____________________________________________________
    int                       12.878          20.980
    dur                     -0.20936         0.17929E-01
    year                    -0.53826         0.94677
    year2                    0.70876E-02     0.14085E-01
    year3                   -0.33068E-04     0.69111E-04
```

trage	0.36390	0.32000
trage2	-0.31495E-02	0.58966
trage3	-0.56019	0.51877
trage4	0.28100	0.54056
trage5	0.43264	0.20575
trage6	-0.22748	0.14640

```
C now try log(duration) instead of duration
lfit int ldur year year2 year3 trage trage2
trage3 trage4 trage5 trage6
```

Iteration	Deviance	Reduction
1	8801.5829	
2	2959.3492	5842.
3	2315.9106	643.4
4	2186.2580	129.7
5	2169.6448	16.61
6	2168.1606	1.484
7	2167.8240	0.3366
8	2167.7919	0.3208E-01
9	2167.7916	0.3470E-03
10	2167.7916	0.4665E-07

```
dis est
```

Parameter	Estimate	S. Error
int	12.117	21.298
ldur	-1.0483	0.72564E-01
year	-0.49783	0.96044
year2	0.65403E-02	0.14278E-01
year3	-0.30640E-04	0.70011E-04
trage	0.23216	0.32332
trage2	-0.11755	0.59711
trage3	-0.80204	0.52563
trage4	0.38544	0.55272
trage5	0.58007	0.20935
trage6	-0.29310	0.15118

```
C the model fits better with ldur
C start backward elimination using this model
C remove the highest polynomial term for year
lfit -year3
```

Iteration	Deviance	Reduction
1	8801.5829	
2	2959.3891	5842.
3	2315.9678	643.4
4	2186.3817	129.6
5	2169.8304	16.55
6	2168.3512	1.479
7	2168.0149	0.3363
8	2167.9828	0.3205E-01
9	2167.9825	0.3473E-03
10	2167.9825	0.4688E-07

```
dis est
```

Parameter	Estimate	S. Error
int	2.8950	3.3215
ldur	-1.0489	0.72558E-01
year	-0.79215E-01	0.96845E-01
year2	0.29616E-03	0.70291E-03
trage	0.24580	0.32189
trage2	-0.12526	0.59701
trage3	-0.80970	0.52543
trage4	0.38969	0.55254
trage5	0.57874	0.20935
trage6	-0.29289	0.15113

```
lfit -year2
```

Iteration	Deviance	Reduction
1	8801.5829	
2	2960.7613	5841.
3	2317.1450	643.6
4	2186.5548	130.6
5	2170.0008	16.55
6	2168.5289	1.472
7	2168.1916	0.3373
8	2168.1594	0.3224E-01
9	2168.1590	0.3511E-03
10	2168.1590	0.4787E-07

```
C the increase in deviance on removing year2 and year3
C is not significant at the 5% level
dis est
```

Parameter	Estimate	S. Error
int	1.5139	0.53900
ldur	-1.0488	0.72558E-01
year	-0.38518E-01	0.70233E-02
trage	0.24860	0.32199
trage2	-0.10853	0.59570
trage3	-0.81168	0.52582
trage4	0.38768	0.55271
trage5	0.57919	0.20955
trage6	-0.29282	0.15125

```
C remove the highest polynomial term for age
lfit -trage6
```

Iteration	Deviance	Reduction
1	8801.5829	
2	2961.4528	5840.
3	2318.8479	642.6
4	2189.1451	129.7
5	2173.5159	15.63
6	2172.9519	0.5640
7	2172.9473	0.4616E-02

8	2172.9473	0.7230E-05

```
dis est
```

Parameter	Estimate	S. Error
int	1.2943	0.53047
ldur	-1.0417	0.72482E-01
year	-0.37779E-01	0.70270E-02
trage	-0.46674E-01	0.26454
trage2	0.89932	0.31357
trage3	0.23829E-01	0.30000
trage4	-0.64032	0.15238
trage5	0.19928	0.90486E-01

```
C removing trage6 has produced an increase in deviance significant at
C the 5% level. Therefore keep all terms of sixth order polynomial


lfit +trage6
```

Iteration	Deviance	Reduction
1	8801.5829	
2	2960.7613	5841.
3	2317.1450	643.6
4	2186.5548	130.6
5	2170.0008	16.55
6	2168.5289	1.472
7	2168.1916	0.3373
8	2168.1594	0.3224E-01
9	2168.1590	0.3511E-03
10	2168.1590	0.4787E-07

```
C test year
lfit -year
```

Iteration	Deviance	Reduction
1	8801.5829	
2	2971.7962	5830.
3	2340.6810	631.1
4	2216.2755	124.4
5	2200.6027	15.67
6	2199.2849	1.318
7	2199.0021	0.2828
8	2198.9772	0.2493E-01
9	2198.9770	0.2284E-03
10	2198.9770	0.2177E-07

```
C significant change in deviance

lfit +year
```

Iteration	Deviance	Reduction
1	8801.5829	
2	2960.7613	5841.
3	2317.1450	643.6

```
         4           2186.5548         130.6
         5           2170.0008         16.55
         6           2168.5289         1.472
         7           2168.1916        0.3373
         8           2168.1594       0.3224E-01
         9           2168.1590       0.3511E-03
        10           2168.1590       0.4787E-07

C test log(duration)
lfit -ldur

    Iteration        Deviance         Reduction
    ------------------------------------------
         1           8801.5829
         2           3024.8900         5777.
         3           2455.8074         569.1
         4           2369.9867         85.82
         5           2362.7628         7.224
         6           2362.0790        0.6839
         7           2361.9175        0.1615
         8           2361.9060       0.1150E-01
         9           2361.9059       0.6667E-04
C significant change in deviance
lfit +ldur

    Iteration        Deviance         Reduction
    ------------------------------------------
         1           8801.5829
         2           2960.7613         5841.
         3           2317.1450         643.6
         4           2186.5548         130.6
         5           2170.0008         16.55
         6           2168.5289         1.472
         7           2168.1916        0.3373
         8           2168.1594       0.3224E-01
         9           2168.1590       0.3511E-03
        10           2168.1590       0.4787E-07

C final model
dis est

    Parameter              Estimate         S. Error
    ------------------------------------------------
    int                     1.5139          0.53900
    trage                  0.24860          0.32199
    trage2                -0.10853          0.59570
    trage3                -0.81168          0.52582
    trage4                 0.38768          0.55271
    trage5                 0.57919          0.20955
    trage6                -0.29282          0.15125
    year                 -0.38518E-01      0.70233E-02
    ldur                  -1.0488          0.72558E-01
```

stop

Results and conclusions

The first two models fitted compare the effects of duration and log(duration) in the full model. The model with log(duration) gives a much better fit with a reduction of deviance of almost 20; this function of duration is kept in the model.

During the process of backward elimination the second and third order terms of year have been removed from the model. The sixth order term of age is statistically significant at the 5% level; therefore this and all the lower order terms are retained in this hierarchical model. Both year and log(duration) are highly significant and are retained.

The parameters for this parsimonious model are as follows:

Variable	Estimate	Standard Error
constant	1.5139	0.53900
ldur	-1.0488	0.72557E-01
year	-0.38518E-01	0.70233E-02
trage	0.24860	0.32199
trage**2	-0.10853	0.59570
trage**3	-0.81168	0.52582
trage**4	0.38768	0.55271
trage**5	0.57919	0.20955
trage**6	-0.29282	0.15125

It is noted that the χ^2 test used to compare the deviance of nested models is not very powerful with highly correlated explanatory variables, such as powers of age. It may be possible to improve on the above parsimonious model with more powerful tests for individual effects, but that is beyond the scope of the present analysis.

The negative coefficient estimate for *ldur* indicates that the probability of migration decreases with duration of stay. This may be due to cumulative inertia effects due to a strengthening of community ties with increasing length of residence. Alternatively, it may be due to residual heterogeneity; with increasing duration, the individuals most likely to migrate will be more and more underrepresented.

The probability of migration predicted by this parsimonious model may be plotted on graphs. In plotting these figures the year is taken as 1985, the individual to be aged 40 and the duration of residence to be 10 years, as appropriate. This is necessary because the precise relationship between an explanatory variable and the response variable depends on the values of the other explanatory variables.
As there are no interaction terms in the model, the patterns shown on the graphs are generally valid.

The probability of migration plotted against age shows peaks just above age 20, around 35 and the largest near age 50. As the data are sparse for the older age group, the size and location of the third peak must be interpreted with caution,

The plot against duration of stay shows the expected decrease in the probability of migration with increasing length of residence. The plot against year also shows a decreasing probability of migration with time over the years 1965 to 1985.

Next:Model development: Random effects model for temporal data

Home page | Contents | Previous

Longitudinal data analysis: A random effects model for temporal data

Model fitting

We compare the fit of the parsimonious simple logistic regression model with the same model with random effects to allow for residual heterogeneity.
For binary data, SABRE fits endpoints at plus and minus infinity by default.

SABRE SESSION:INPUT AND OUTPUT

```
data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
osb ose mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
C transform age as before
transform tempage age - 30
transform trage tempage / 10
transform trage2 trage * trage
transform trage3 trage2 * trage
transform trage4 trage3 * trage
transform trage5 trage4 * trage
transform trage6 trage5 * trage
transform ldur log dur
C first fit the simple logistic model

lfit int ldur year trage trage2 trage3 trage4 trage5 trage6

    Iteration        Deviance        Reduction
    ____________________________________________
        1           8801.5829
        2           2960.7613         5841.
        3           2317.1450         643.6
        4           2186.5548         130.6
        5           2170.0008         16.55
        6           2168.5289         1.472
        7           2168.1916        0.3373
        8           2168.1594       0.3224E-01
        9           2168.1590       0.3511E-03
       10           2168.1590       0.4787E-07

dis est

    Parameter              Estimate         S. Error
    ____________________________________________________
    int                     1.5139          0.53900
    ldur                   -1.0488          0.72558E-01
    year                  -0.38518E-01      0.70233E-02
```

```
    trage                0.24860          0.32199
    trage2              -0.10853          0.59570
    trage3              -0.81168          0.52582
    trage4               0.38768          0.55271
    trage5               0.57919          0.20955
    trage6              -0.29282          0.15125

C fit the same model with random effects
C endpoints are fitted by default
fit .

    Iteration        Deviance         Step      End-points     Orthogonality
                                      length    0          1      criterion
    ------------------------------------------------------------------------
        1          2198.4881          1.0000    free    free        4.6471
        2          2198.2943          0.2500    free    free        3.1137
        3          2185.4266          0.3033    free    free        13.365
        4          2174.8955          0.1175    free    fixed       9.2150
        5          2142.0094          1.0000    free    free        10.360
        6          2135.1201          1.0000    free    free        3.7965
        7          2133.8038          1.0000    free    free        11.834
        8          2133.7948          1.0000    free    free        37.114
        9          2133.7948          1.0000    free    free

dis est

    Parameter              Estimate         S. Error
    ------------------------------------------------------------
    int                     0.83341          0.77050
    ldur                   -0.65918          0.10463
    year                   -0.36521E-01      0.10873E-01
    trage                  -0.69598E-01      0.34063
    trage2                  0.76814E-01      0.59487
    trage3                 -0.82208          0.53734
    trage4                  0.33146          0.54900
    trage5                  0.56760          0.21311
    trage6                 -0.27657          0.15032
    scale                   0.47710          0.17447
                                                        PROBABILITY
                                                        -----------
    end-point 0             0.56682          0.19724    0.36113
    end-point 1             0.27460E-02      0.46361E-02  0.17495E-02

stop
```

Results and conclusion

The deviance has decreased from 2168.16 to 2133.79. This is a reduction of over 34 on 3 degrees of freedom, on adding the individual specific random term to the model. The extra three degrees of freedom are given by the scale of the Normal mixing distribution and the two estimated probabilities of the endpoints. Although the χ^2 test is not strictly correct as the simple logistic model lies on the edge of the parameter space of the mixture model, such a large change in deviance ($\chi^2_{(3)}$=7.81) demonstrates that there is considerable unobserved heterogeneity in the population.

- The coefficient estimate of *ldur* is still negative, but is considerably smaller in magnitude than in the simple logistic model. The estimate of this endogenous explanatory variable has changed by allowing for residual heterogeneity; the estimates of the other variables have changed little (by less than one standard error), and their standard errors are almost unchanged.

- The coefficient of *ldur* measures cumulative inertia effects, and its value confirms that there is an increasing disinclination to move with increasing length of residence. However the effect is smaller than suggested by the simple logistic model; that estimate was inflated because no account was taken of the fact that with increasing duration the individuals most likely to migrate are more and more underrepresented in the population. Inference about duration effects can be misleading unless there is control for omitted variables. (Lancaster 1979; Heckman and Singer 1985)

- The probability of 0.36 associated with the left endpoint gives a measure of the proportion of "stayers" in the population, i.e. those individuals never likely to migrate. Examination of the parameter estimate and standard error of the right endpoint (and corresponding probability of 0.0017) suggests that this parameter (which estimates the proportion of the population migrating every year) could be set to zero.

- The scale parameter estimate is the standard deviation of the Normal distribution assumed for the individual specific terms.

- The probability of migration predicted by this random effects model may be plotted on graphs to aid interpretation of the parameter estimates. As before, the year is taken as 1985, the individual to be aged 40, and the duration of residence to be 10 years, as appropriate. As no interaction terms have been considered, the trends shown on the graphs are generally valid.

- In calculating the probabilities, the individual specific term is given the estimated population median value, taking into account both the Normal distribution and the proportion of stayers.

- The plot against age now shows two clear peaks at just above age 20 and just below age 50. The relative size of the peaks has changed compared to the simple logistic model; the size and location of the peak near age 50 has again to be interpreted with caution as the data are sparse for this age group. The dominance of the first peak in the random effects model is more plausible substantively as this is the age at which geographical ties are at their minimum.

- The graph against duration of stay shows the decline in migration probability with duration for both the simple logistic and the random effects models. When unobserved heterogeneity is taken into account, the estimated decline is due to cumulative inertia effects; in the simple logistic model the estimate is inflated as discussed above.

- The shapes of the graphs of migration probability against year are the same for both models.

- The levels of probability estimated by the two models are not strictly comparable, as the simple logistic model gives the population average value for individuals with given values of the explanatory variables (age, year, duration of stay), whereas the random effects graphs show the probability values for individuals with the median value of the nuisance parameter.

Can we explain the pattern of migration with age by adding explanatory variables which measure life cycle factors, such as marriage, occupation and employment status and the presence of children in the family?

Next:Model development: Adding explanatory variables

Longitudinal data analysis: Adding explanatory variables

The variation of migration propensity with age has been linked to life cycle factors, such as marriage, employment, career moves, and the presence of children in the family. Similarly year effects can be linked to economic factors, and employment and career moves are seen to represent underlying economic health. Do explanatory variables which measure these effects explain the variation of migration behaviour with age and year?

The large number of possible explanatory variables require a pragmatic strategy to model building.

Model development

We start with the parsimonious main effects model for the temporal variables,

$$age+age^2+age^3+age^4+age^5+age^6+year+log(dur)$$

and add explanatory variables which measure individual life cycle effects.

We choose explanatory variables suggested by substantive considerations to include in our model. A number of such explanatory variables are present in the data set, giving information on education, occupation, marital status, employment, the presence of children of different ages, etc.

Although empirical evidence is mixed, **education** is often considered to increase the propensity to migrate, because it increases employment opportunities and gives access to better information about other areas. (Sandefur and Scott 1981, Goss 1985, Liaw 1990)

Marital status is an important feature of theories about migration behaviour, with evidence that married individuals are less likely to migrate. Getting married, marital break up and remarriage are expected to increase the probability of migration. (Devis 1983, Grundy 1989)

School age **children** create important ties to an area, and the fear of disrupting children's education may inhibit migration. (Long 1972, Davies and Flowerdew 1992)

Employment and **occupational** status variables also important in relation to migration (Warnes 1983, Greenwood 1985, Davies and Flowerdew 1992, Ellis et al. 1993, Herzog 1993).

Career progression is another important variable to affect migration (Salt 1990). We consider three variables measuring *changes* in employment or occupational status which, being "favourable to socio-economic achievement" (Cote 1997) might encourage migration: obtaining a job, promotion to manager and promotion to service class.

We fit a series of logistic models and use backward elimination to assess which explanatory variables to retain. As the parameter estimates, apart from that of the endogenous variable *ldur*, are very similar for the simple logistic and random effects models, and as the latter is much more

computer intensive, we use the simple logistic model for model development.

- We start with the model for the temporal variables, and add education (*ed*), occupational status (*osb3*), employment status (*esb2*) and marital status (*msb*), each measured at the beginning of the year, first marriage (*mfm*), marital break-up (*mbu*), remarriage (*mrm*), the presence of children of different ages (*ch1, ch2, ch3, ch4*), obtaining a job (*eoj*), promotion to manager (*epm*) and promotion to service class (*ops*).

- For education and marital status we use the original 5 level variables to include in the model; for employment and occupational status we have chosen for simplicity the collapsed variables *esb2* and *osb3* with 3 and 2 levels respectively, instead of the original 8 and 12 levels. The other variables are all 2-level factors.

- We note that some levels of the original employment and occupational status variables are likely to be highly correlated (eg. employment status: none, occupational status: none), and problems with aliasing are likely to occur in models which include such variables. Cross tabulation of the levels of these variables will help to identify possible problems, but that is beyond the scope of the present example.

- We use a cut-off significance level of 0.1 rather than the conventional 0.05. This is very conservative, as the simple logistic model tends to overestimate significance, as we noted earlier. However, as the model may be misspecified due to our pragmatic approach, conservatism is considered important to reduce the chance of rejecting a possibly relevant explanatory variable.

- At each step in the backward elimination we test if the removal of the explanatory variable with the lowest t-ratio (ratio of a parameter to its standard error) gives a significant deterioration in model fit by comparing the change in deviance with the appropriate value of χ^2.
At the 0.1 significance level the critical values of the chi-squared distribution for various degrees of freedom are $\chi^2_{(1)}=2.71$, $\chi^2_{(2)}=4.61$, $\chi^2_{(3)}=6.25$, $\chi^2_{(4)}=7.78$.

- When the preferred main effects model is found, the same model is refitted with random effects to allow for unobserved heterogeneity.

Next:The SABRE analysis

Home page | Contents | Previous

Adding explanatory variables: the SABRE analysis

We carry out the backward elimination as follows:

SABRE SESSION:INPUT AND OUTPUT

```
C input data and transform variables

data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
ocb oce mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
transform tempage age - 30
transform trage tempage / 10
transform trage2 trage * trage
transform trage3 trage2 * trage
transform trage4 trage3 * trage
transform trage5 trage4 * trage
transform trage6 trage5 * trage
transform ldur log dur
C convert explanatory variables to factors
factor ed fed
factor ch1 fch1
factor ch2 fch2
factor ch3 fch3
factor ch4 fch4
factor msb fmsb
factor msb1 fmsb1
factor msb2 fmsb2
factor mbu fmbu
factor mrm fmrm
factor mfm fmfm
factor eoj feoj
factor ops fops
factor epm fepm
factor esb2 fesb2
factor osb3 fosb3
C fit full model

lfit int ldur year trage trage2 trage3 trage4 trage5 trage6 &
fed fmbu fmfm fmrm fmsb fch1 fch2 fch3 fch4 fesb2 fosb3 fepm fops feoj

    Iteration        Deviance        Reduction
    _____________________________________________
```

```
      1           8801.5829
      2           2932.0559        5870.
      3           2260.2230        671.8
      4           2115.9063        144.3
      5           2095.5823        20.32
      6           2093.5142        2.068
      7           2093.1257       0.3885
      8           2093.0786       0.4706E-01
      9           2093.0765       0.2120E-02
     10           2093.0760       0.5102E-03
     11           2093.0758       0.1876E-03
```

```
dis est
```

Parameter	Estimate	S. Error
int	1.3741	0.74144
ldur	-0.97671	0.75658E-01
year	-0.42966E-01	0.77703E-02
trage	0.48422	0.34821
trage2	-0.81192E-01	0.63693
trage3	-0.58212	0.53301
trage4	0.30160	0.57210
trage5	0.42878	0.20849
trage6	-0.23204	0.15366
fed (1)	0.	ALIASED [I]
fed (2)	-0.29439E-01	0.29414
fed (3)	-0.42630	0.31085
fed (4)	0.19577E-01	0.21836
fed (5)	-0.25889	0.23502
fmbu (1)	0.	ALIASED [I]
fmbu (2)	1.2831	0.64008
fmfm (1)	0.	ALIASED [I]
fmfm (2)	0.46489	0.24075
fmrm (1)	0.	ALIASED [I]
fmrm (2)	0.97834	0.80128
fmsb (1)	0.	ALIASED [I]
fmsb (2)	-0.44557	0.19011
fmsb (3)	-0.26831	0.49968
fmsb (4)	0.78074	0.56836
fmsb (5)	-7.9406	82.102
fch1 (1)	0.	ALIASED [I]
fch1 (2)	-0.76060E-01	0.38951
fch2 (1)	0.	ALIASED [I]
fch2 (2)	-0.68220E-01	0.44099
fch3 (1)	0.	ALIASED [I]
fch3 (2)	-1.2554	0.75279
fch4 (1)	0.	ALIASED [I]
fch4 (2)	0.23823E-01	0.58680
fesb2 (1)	0.	ALIASED [I]
fesb2 (2)	0.52758	0.32558

```
    fesb2 ( 3)              0.90635          0.44986
    fosb3 ( 1)                   0.          ALIASED [I]
    fosb3 ( 2)              0.83994          0.16945
    fepm  ( 1)                   0.          ALIASED [I]
    fepm  ( 2)             -0.22312          0.50383
    fops  ( 1)                   0.          ALIASED [I]
    fops  ( 2)               1.1732          0.36420
    feoj  ( 1)                   0.          ALIASED [I]
    feoj  ( 2)              0.51723          0.43284

C note that the lowest level of each factor is set to zero
C fch1, fch2 and fch4 have very low t-ratios
C remove fch4 first, as this has lowest t-ratio

C To save space we use the MONITOR NO command to produce
C summary information only on the progress of the fitting algorithm

monitor no

lfit -fch4

    Deviance =     2093.0774     at iteration    11

lfit -fch1

    Deviance =     2093.1162     at iteration    11

lfit -fch2

    Deviance =     2093.1470     at iteration    11

lfit -fepm

    Deviance =     2093.3436     at iteration    11

lfit -feoj

    Deviance =     2094.8028     at iteration    11

C the changes in deviance above are not significant at the 10% level
C compared with 2.71, ie. chi-sq. for 1 degree of freedom
C for fed some levels appear more significant than others; test fed.

lfit -fed

    Deviance =     2100.8431     at iteration    11

C change in deviance of 6.04 is not significant at the 10% level
C compared with 7.78, ie. chi-sq. for 4 degrees of freedom
C fed can also be removed from the model
```

```
dis est
```

Parameter	Estimate	S. Error
int	1.0028	0.70183
ldur	-0.99577	0.74243E-01
year	-0.39207E-01	0.74412E-02
trage	0.43563	0.33628
trage2	-0.10207	0.62253
trage3	-0.54183	0.52659
trage4	0.29816	0.56929
trage5	0.41445	0.20703
trage6	-0.22861	0.15373
fmbu (1)	0.	ALIASED [I]
fmbu (2)	1.2363	0.64637
fmfm (1)	0.	ALIASED [I]
fmfm (2)	0.46619	0.24024
fmrm (1)	0.	ALIASED [I]
fmrm (2)	1.0371	0.79233
fmsb (1)	0.	ALIASED [I]
fmsb (2)	-0.44911	0.18890
fmsb (3)	-0.19336	0.49091
fmsb (4)	0.71328	0.55703
fmsb (5)	-7.8189	82.104
fch3 (1)	0.	ALIASED [I]
fch3 (2)	-1.2803	0.75074
fesb2 (1)	0.	ALIASED [I]
fesb2 (2)	0.55897	0.32382
fesb2 (3)	1.0902	0.39518
fosb3 (1)	0.	ALIASED [I]
fosb3 (2)	0.83672	0.16570
fops (1)	0.	ALIASED [I]
fops (2)	1.0891	0.28586

```
C level 2 of fmsb has a high t-ratio; the others are lower
C test fmsb

lfit -fmsb

    Deviance =     2110.0394     at iteration    10

C the change in deviance is significant at the 10% level
C compared with 7.78, ie. chi-sq. for 4 degree of freedom

C The factor fmsb is significant, but the effect of
C some levels is small. Therefore collapse some levels of fmsb
C and use the 3 level factor fmsb1 instead.
lfit +fmsb1

    Deviance =     2102.9664     at iteration    10
```

```
dis est
```

Parameter	Estimate	S. Error
int	0.95067	0.69857
ldur	-0.99776	0.74232E-01
year	-0.38286E-01	0.73967E-02
trage	0.42904	0.33667
trage2	-0.17240	0.61843
trage3	-0.57939	0.52601
trage4	0.34715	0.56440
trage5	0.43121	0.20700
trage6	-0.23906	0.15230
fmbu (1)	0.	ALIASED [I]
fmbu (2)	1.2313	0.64655
fmfm (1)	0.	ALIASED [I]
fmfm (2)	0.46823	0.24019
fmrm (1)	0.	ALIASED [I]
fmrm (2)	1.4241	0.75185
fch3 (1)	0.	ALIASED [I]
fch3 (2)	-1.2177	0.74682
fesb2 (1)	0.	ALIASED [I]
fesb2 (2)	0.55546	0.32356
fesb2 (3)	1.0911	0.39499
fosb3 (1)	0.	ALIASED [I]
fosb3 (2)	0.83211	0.16526
fops (1)	0.	ALIASED [I]
fops (2)	1.1032	0.28470
fmsb1 (1)	0.	ALIASED [I]
fmsb1 (2)	-0.44502	0.18885
fmsb1 (3)	0.11026	0.40049

```
C The change in deviance is significant at the
C 10% level compared with 4.6, ie. chi-sq. for 2 degree of freedom.
C Only level 2 seems significant.
C Collapse variable further; use 2 level factor msb2 instead.

lfit -fmsb1

    Deviance =     2110.0394     at iteration    10

lfit +fmsb2

    Deviance =     2103.0411     at iteration    10

dis est
```

Parameter	Estimate	S. Error
int	0.98308	0.68858
ldur	-0.99954	0.73925E-01

```
    year                  -0.38417E-01      0.73821E-02
    trage                  0.44814          0.32946
    trage2                -0.18073          0.61760
    trage3                -0.58213          0.52585
    trage4                 0.35071          0.56434
    trage5                 0.43121          0.20699
    trage6                -0.23961          0.15231
    fmbu  ( 1)                 0.           ALIASED [I]
    fmbu  ( 2)              1.2346          0.64645
    fmfm  ( 1)                 0.           ALIASED [I]
    fmfm  ( 2)             0.46040          0.23846
    fmrm  ( 1)                 0.           ALIASED [I]
    fmrm  ( 2)              1.5114          0.68339
    fch3  ( 1)                 0.           ALIASED [I]
    fch3  ( 2)             -1.2225          0.74642
    fesb2 ( 1)                 0.           ALIASED [I]
    fesb2 ( 2)             0.55741          0.32354
    fesb2 ( 3)              1.0932          0.39493
    fosb3 ( 1)                 0.           ALIASED [I]
    fosb3 ( 2)             0.83115          0.16524
    fops  ( 1)                 0.           ALIASED [I]
    fops  ( 2)              1.1069          0.28439
    fmsb2 ( 1)                 0.           ALIASED [I]
    fmsb2 ( 2)            -0.46453          0.17487

C The addition of fmsb2 to the model produces a change in
C deviance significant at the 10% level. The coefficient estimate is
C now significant. Keep fmsb2 in the model.

C Remove the remaining factors one by one and compare each
C change in deviance with 2.71 (chi-sq. at the 10% level,
C 1 degree of freedom).

lfit -fch3

    Deviance =     2106.7537     at iteration    10

lfit +fch3

    Deviance =     2103.0411     at iteration    10

lfit -fmbu

    Deviance =     2105.8325     at iteration    10

lfit +fmbu

    Deviance =     2103.0411     at iteration    10

lfit -fmrm
```

```
    Deviance =     2106.7500     at iteration    10

lfit +fmrm

    Deviance =     2103.0411     at iteration    10

lfit -fmfm

    Deviance =     2106.5408     at iteration    10

lfit +fmfm

    Deviance =     2103.0411     at iteration    10

lfit -fesb2

    Deviance =     2111.2846     at iteration    10

lfit +fesb2

    Deviance =     2103.0411     at iteration    10

lfit -fops

    Deviance =     2115.7878     at iteration    10

lfit +fops

    Deviance =     2103.0411     at iteration    10

lfit -fosb3

    Deviance =     2126.2913     at iteration    10

lfit +fosb3

    Deviance =     2103.0411     at iteration    10

C All the above factors are significant.

dis est

    Parameter               Estimate         S. Error
    ____________________________________________________________
    int                      0.98308          0.68858
    ldur                    -0.99954          0.73925E-01
    year                    -0.38417E-01      0.73821E-02
    trage                    0.44814          0.32946
    trage2                  -0.18073          0.61760
    trage3                  -0.58213          0.52585
    trage4                   0.35071          0.56434
```

```
    trage5                 0.43121          0.20699
    trage6                -0.23961          0.15231
    fch3  ( 1)                 0.           ALIASED [I]
    fch3  ( 2)              -1.2225          0.74642
    fmbu  ( 1)                 0.           ALIASED [I]
    fmbu  ( 2)               1.2346         0.64645
    fmrm  ( 1)                 0.           ALIASED [I]
    fmrm  ( 2)               1.5114         0.68339
    fmfm  ( 1)                 0.           ALIASED [I]
    fmfm  ( 2)              0.46040         0.23846
    fmsb2 ( 1)                 0.           ALIASED [I]
    fmsb2 ( 2)             -0.46453         0.17487
    fesb2 ( 1)                 0.           ALIASED [I]
    fesb2 ( 2)              0.55741         0.32354
    fesb2 ( 3)               1.0932         0.39493
    fops  ( 1)                 0.           ALIASED [I]
    fops  ( 2)               1.1069         0.28439
    fosb3 ( 1)                 0.           ALIASED [I]
    fosb3 ( 2)              0.83115         0.16524

C Is trage6 still significant?
lfit -trage6

    Deviance =     2106.0860     at iteration     8

C trage6 is significant at the 10% level
C The above model is therefore our final main effects model.
stop
```

Next:Random effects model with explanatory variables

Home page | Contents | Previous

The random effects model with explanatory variables

We now fit the random effects model with the explanatory variables we found significant in our previous analysis.

SABRE SESSION:INPUT AND OUTPUT

```
data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
osb ose mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
transform tempage age - 30
transform trage tempage / 10
transform trage2 trage * trage
transform trage3 trage2 * trage
transform trage4 trage3 * trage
transform trage5 trage4 * trage
transform trage6 trage5 * trage
transform ldur log dur
factor ch3 fch3
factor mbu fmbu
factor mrm fmrm
factor mfm fmfm
factor ops fops
factor esb2 fesb2
factor osb3 fosb3
factor msb2 fmsb2
C fit simple logistic main effects model

lfit int ldur year trage trage2 trage3 trage4 trage5 trage6 &
fch3 fesb2 fmbu fmrm fmfm fops fmsb2 fosb3

    Iteration        Deviance        Reduction
    ____________________________________________
        1           8801.5829
        2           2935.4172         5866.
        3           2266.9758         668.4
        4           2124.9723         142.0
        5           2105.4389         19.53
        6           2103.4563         1.983
        7           2103.0820        0.3743
        8           2103.0417       0.4028E-01
        9           2103.0411       0.5907E-03
       10           2103.0411       0.1442E-06

dis est

    Parameter              Estimate         S. Error
    ____________________________________________________
    int                    0.98308          0.68858
```

ldur	-0.99954	0.73925E-01
year	-0.38417E-01	0.73821E-02
trage	0.44814	0.32946
trage2	-0.18073	0.61760
trage3	-0.58213	0.52585
trage4	0.35071	0.56434
trage5	0.43121	0.20699
trage6	-0.23961	0.15231
fch3 (1)	0.	ALIASED [I]
fch3 (2)	-1.2225	0.74642
fesb2 (1)	0.	ALIASED [I]
fesb2 (2)	0.55741	0.32354
fesb2 (3)	1.0932	0.39493
fmbu (1)	0.	ALIASED [I]
fmbu (2)	1.2346	0.64645
fmrm (1)	0.	ALIASED [I]
fmrm (2)	1.5114	0.68339
fmfm (1)	0.	ALIASED [I]
fmfm (2)	0.46040	0.23846
fops (1)	0.	ALIASED [I]
fops (2)	1.1069	0.28439
fmsb2 (1)	0.	ALIASED [I]
fmsb2 (2)	-0.46453	0.17487
fosb3 (1)	0.	ALIASED [I]
fosb3 (2)	0.83115	0.16524

C fit the same model with random effects
fit .

Iteration	Deviance	Step length	End-points 0	1	Orthogonality criterion
1	2135.3134	1.0000	free	free	4.6620
2	2128.4000	0.2500	free	free	4.4184
3	2123.0042	0.4288	free	fixed	0.21340E-01
4	2122.5984	0.0078	free	fixed	7.1550
5	2088.6586	1.0000	free	free	4.0641
6	2079.1069	1.0000	free	free	4.5894
7	2075.6823	1.0000	free	free	24.604
8	2075.6458	1.0000	free	free	17.341
9	2075.6458	1.0000	free	free	

dis est

Parameter	Estimate	S. Error
int	0.73017	0.89590
ldur	-0.63527	0.10783
year	-0.37769E-01	0.10902E-01
trage	0.17360	0.34893
trage2	-0.16613E-01	0.61773
trage3	-0.54783	0.53830
trage4	0.28880	0.56026
trage5	0.40754	0.21096
trage6	-0.22024	0.15128
fch3 (1)	0.	ALIASED [I]

```
fch3  ( 2)              -1.3073          0.75078
fesb2 ( 1)                   0.          ALIASED [I]
fesb2 ( 2)              0.31615          0.37617
fesb2 ( 3)              0.77441          0.45042
fmbu  ( 1)                   0.          ALIASED [I]
fmbu  ( 2)               1.2513          0.66612
fmrm  ( 1)                   0.          ALIASED [I]
fmrm  ( 2)               1.5259          0.71835
fmfm  ( 1)                   0.          ALIASED [I]
fmfm  ( 2)              0.45266          0.25126
fops  ( 1)                   0.          ALIASED [I]
fops  ( 2)               1.2016          0.30209
fmsb2 ( 1)                   0.          ALIASED [I]
fmsb2 ( 2)             -0.56390          0.19405
fosb3 ( 1)                   0.          ALIASED [I]
fosb3 ( 2)              0.68677          0.18610
scale                   0.49269          0.18099

                                                      PROBABILITY
                                                      ___________
end-point 0             0.48867          0.19067      0.32760
end-point 1             0.29991E-02      0.43654E-02  0.20105E-02
```

stop

Next: Interpretation of results

Home page | Contents | Previous

Interpretation of results

The explanatory variables

Backward elimination using the simple logistic model has shown the following variables to be significant at the 10% level:

- **employment status**: esb2=1 (self employed), esb2=2 (employed), esb2=3 (not working)
- **occupational status**: osb3=1 (small proprietors, supervisors), osb3=0 (otherwise)
- **promotion to service class**: ops=0 (no), ops=1 (yes)
- **first marriage**: mfm=0 (no), mfm=1 (yes)
- **marital break-up**: mbu=0 (no), mbu=1 (yes)
- **remarriage**: mrm=0 (no), mrm=1 (yes)
- **presence of children age 15-16**: ch3=0 (no), ch3=1 (yes)
- **marital status**: msb2=0 (not married), msb2=1 (married)

Our preferred homogeneous main effects model is therefore:

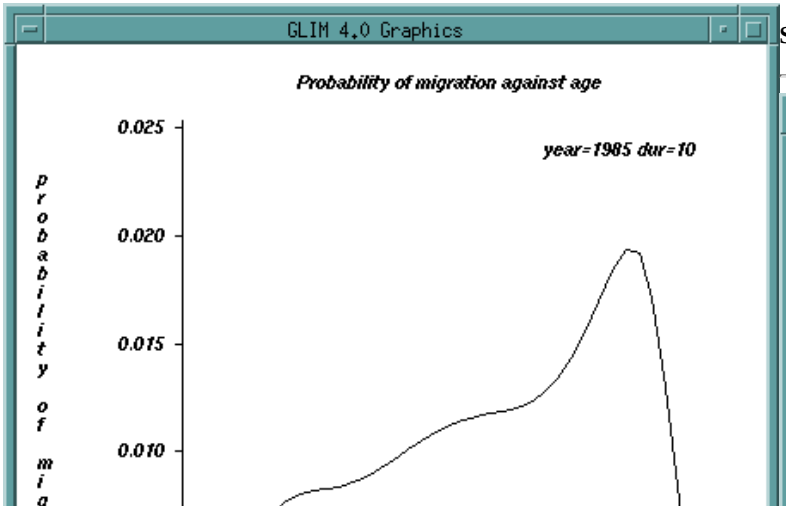
$age+age^2+age^3+age^4+age^5+age^6+year+log(dur)+esb2+osb3+ops+mfm+mbu+mrm+ch3+msb2$

Comparison of simple logistic and random effects models

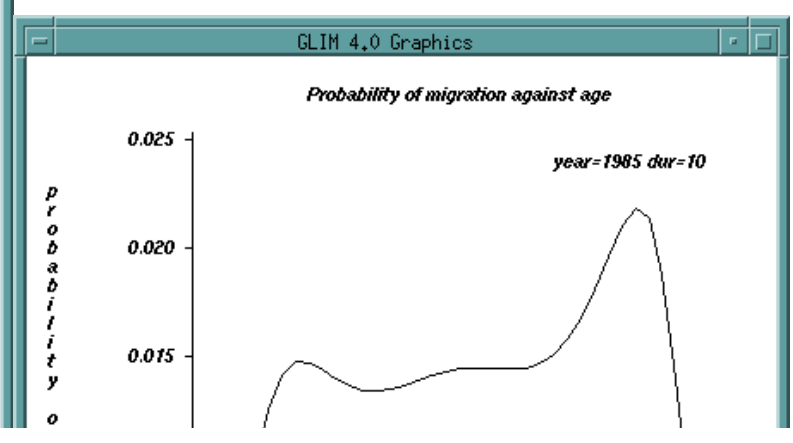
- When the same model is fitted with random effects, the deviance decreases by 27.4. Although it is not strictly correct to use the χ^2 test to compare the simple logistic and random effects models, such a substantial reduction in deviance for three extra parameters estimated (scale and two endpoints) provides evidence that in addition to the time varying explanatory variables included in the model, there remains unobserved heterogeneity.
- Comparison of the parameter estimates of the two models shows that, as before, only the estimate of the endogenous *log(dur)* has changed substantially (from -0.9995 to -0.6353): controlling for unobserved heterogeneity has decreased the observed negative duration of stay effect. (See Lancaster and Nickell 1980). The other parameter estimates for the two models are the same, within one standard error.
- The parameter estimates of *msb2* and *ch3* are both negative, providing evidence that being married significantly reduces the probability of migration, as does the presence of children in the age group 15-16, presumably for fear of disrupting schooling close to public examinations. There is no evidence that younger or older secondary school-age children increase ties to an area.
- The positive coefficient estimates for *mfm*, *mbu*, *mrm* and *ops* indicate that the events of first marriage, marital break-up, remarriage and promotion to service class all increase the probability of migration.
- The positive coefficients for levels 2 and 3 of *esb2* provides evidence that employed and unemployed individuals are more likely to migrate than the self-employed. Also the positive coefficient of *osb3* indicates that small proprietors and supervisors are more likely to migrate than others.
- The probability of 0.3276 estimated for the left hand endpoint again indicates a high proportion of stayers. The right endpoint is small and may be set to zero.

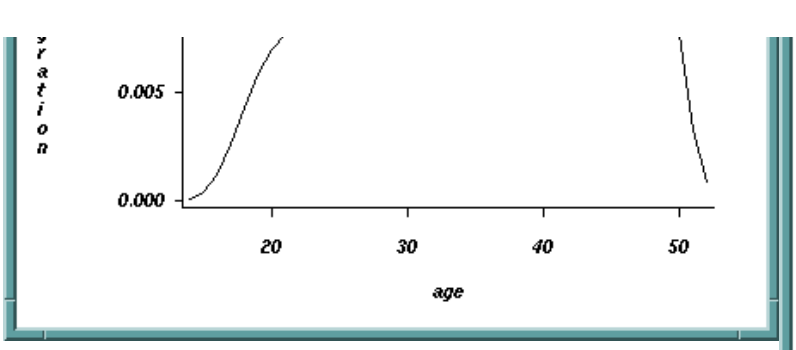
Variation with age

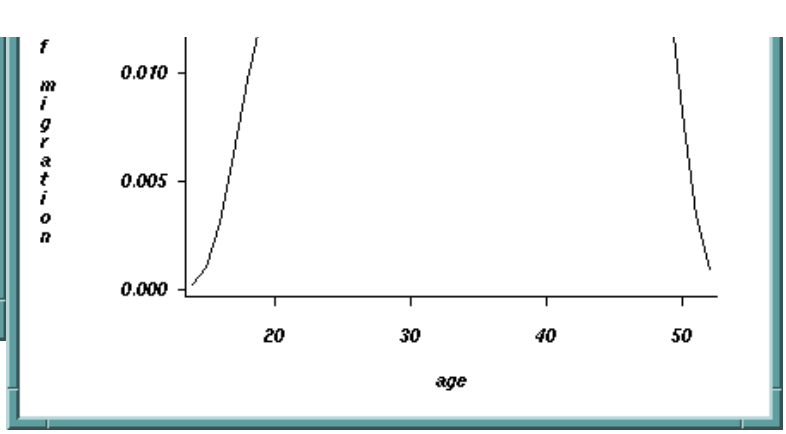
- To illustrate the difference between the homogeneous and random effects models, we plot the probability of migration against age, with the year taken as 1985, duration of stay 10 years and all other explanatory variables set to zero (ie. to their reference levels). As there are no interaction terms, the patterns shown on the graphs are generally valid.



Simple logistic model







Random effects model

Both graphs show a peak just below age 50, where the data are sparse; the random effects model, although flatter over the earlier years, has a more accentuated first peak just above age 20. The three peaks are less pronounced than in the original analysis without explanatory variables, but it is clear that controlling for life cycle effects provides only a partial explanation of the three peaks.

We shall examine the contribution of some of the explanatory variables to the peaks. Because of the excessive computing requirements of the random effects model, we shall use the simple logistic model in this analysis.

Next:Contribution of life cycle events to the peaks

Home page | Contents | Previous

Contribution of life cycle events to the peaks

To examine the contribution of an explanatory variable on the peaks, the variable is omitted from the preferred homogeneous main effects model and the simplified model is refitted. The probability of migration is plotted against age, with the year set to 1985, duration to 10 years and **all the explanatory variables set to zero**, as before.

The following graphs show the effects of removing in turn *msb2*, *mrm* and *ch3* from the full model. Similar graphs may be drawn for the other explanatory variables.

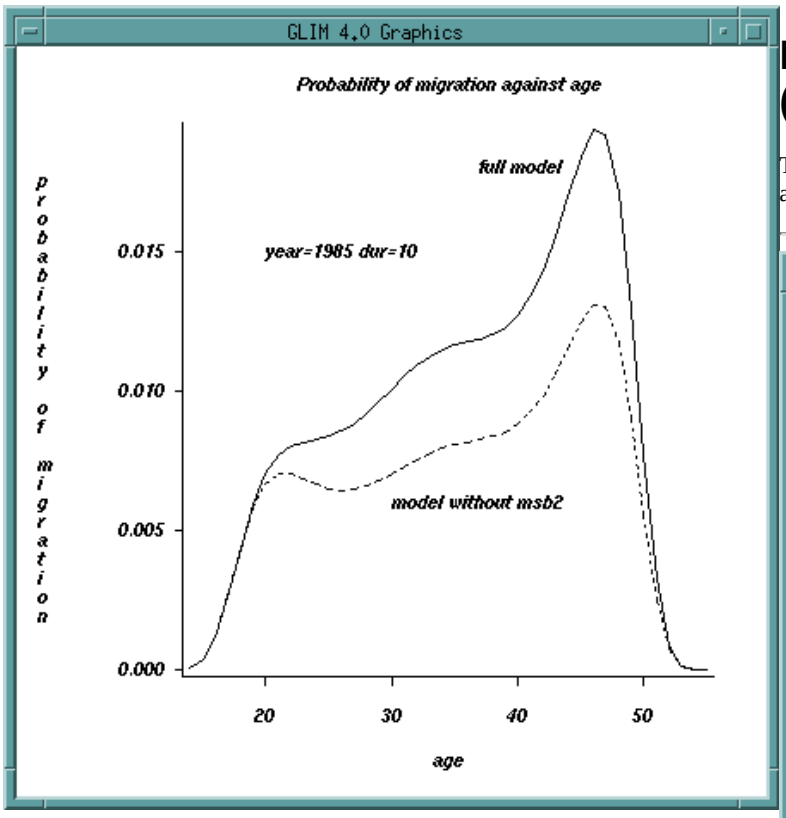


Figure 1: The effect of removing msb2 (marital status)

The basic shapes of the graphs are very similar, suggesting just a scaling effect, and no explanation of the peak.

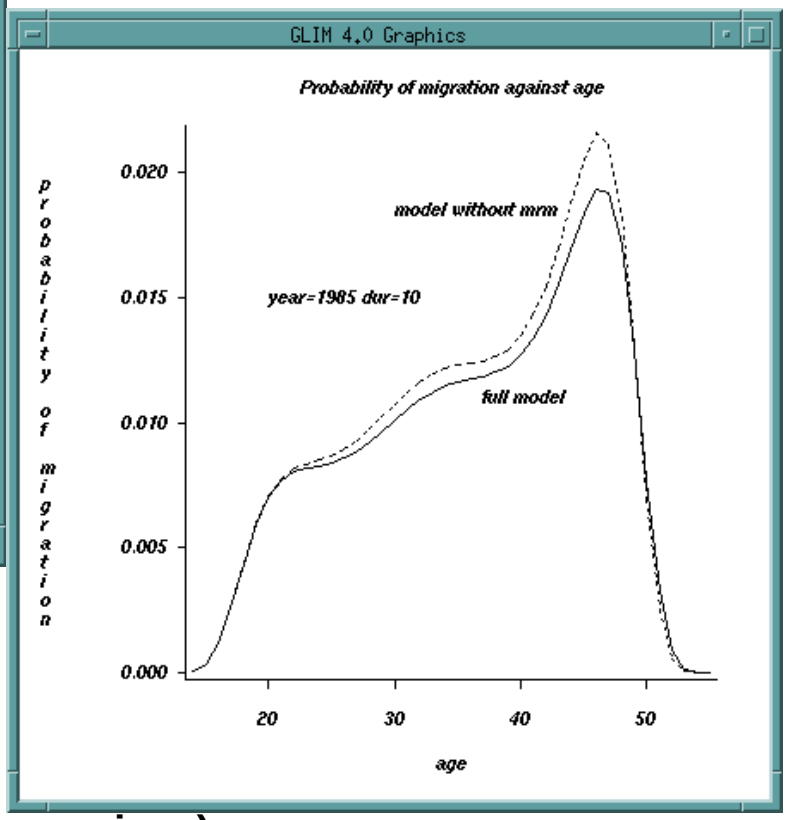


Figure 2: The effect of removing mrm (remarriage)

The peaks seem to be slightly attenuated in the full model with *mrm*=0 compared to the simplified model. It appears that the minor difference between the graphs is not just a scaling effect, but evidence that remarriage contributes to the third peak.

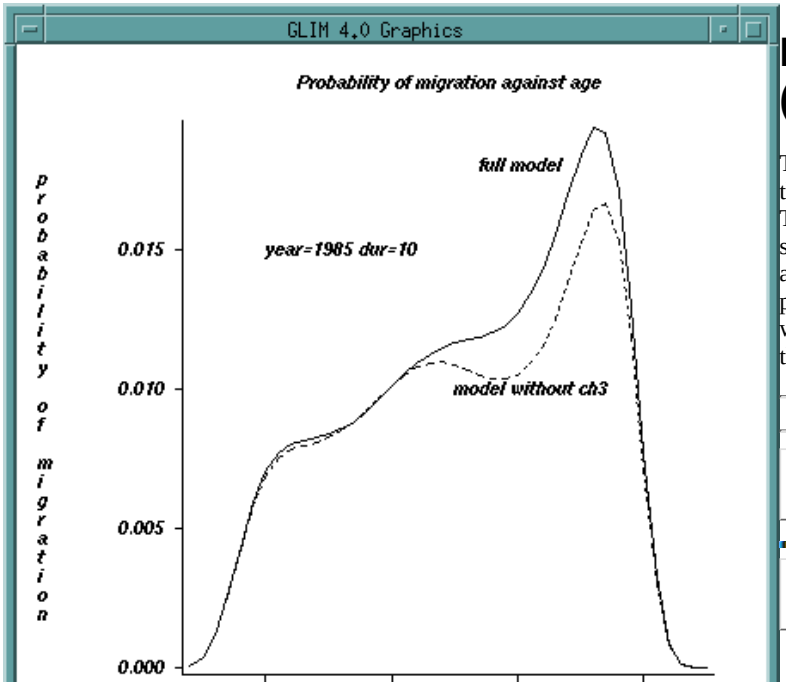


Figure 3: The effect of removing ch3 (children aged 15-16)

This variable appears to provide a partial explanation for migration behaviour in the age range 35 to 50 (the appropriate age for parents of children aged 15-16). The trough around age 40 with *ch3* excluded from the model is partially smoothed out in the full model with *ch3*=0. However, although having a child aged 15 to 16 does significantly reduce the probability of migration, the third peak is not attenuated in the full model, but is in fact increased, for those without children in this age range. This effect therefore does not explain the third peak.

Next:Conclusions and suggestions for further work

Home page | Contents | Previous


20
30
40
50
age

Conclusions and suggestions for further work

- We must be cautious about drawing general conclusions from this analysis as the sample was drawn from one locality. However, the extent to which migration behaviour with age can be explained by explanatory variables is likely to be informative about the process of migration.

- We have identified three statistically significant peaks in migration behaviour with age during individuals' working lives; at just above age 20, at around age 35 and just below age 50. The size and location of the third peak has to be interpreted with caution as the data are sparse here.

- We have shown that there is considerable heterogeneity in the population sampled, with a considerable proportion of individuals who are likely never to move.

- The negative coefficient estimate for *ldur* indicates that the probability of migration decreases with duration of stay in the locality, consistent with the concept of cumulative inertia.

- The simple logistic model takes no account of the fact that in a heterogeneous population, the individuals most likely to migrate are more and more underrepresented with increasing duration, and therefore inflates the duration of stay effect. To estimate the true effect of cumulative inertia, we must control for residual population heterogeneity.

- For the years studied the likelihood of migration decreased with calendar time for the population surveyed.

- The following time varying explanatory variables have been found to have a significant effect on migration (at the 10% level):

  Employment status
  Occupational status
  Promotion to service class
  First marriage
  Marital break-up
  Remarriage
  Presence of children age 15-16
  Marital status

- It is evident that the third peak in the pattern of migration with age persist even after controlling for the time-varying explanatory variables. Remarriage appears to make a small contribution to this peak, however controlling for the presence of children of age 15-16 actually increases the size of the peak for those without children of this age.

- The main effects model may be extended by the addition of interaction terms both between the time variables and between time and other explanatory variables. If these are confined to the linear term in age, there are 55 possible pairwise interactions. An interaction model has been fitted to this data by Borhani Haghighi and Davies (1999b). These throw light on questions such as:

1. Does the relative importance of the three peaks vary with calendar year?
2. Do patterns of migration behaviour for employed/self- employed/not working individuals relate to age?
3. Is the probability of migration after marriage break-up/remarriage age related?
We leave this for the student to explore.

 As we have analysed migration data from only one locality, it is not clear how far the results are generally characteristic of the process of inter-county migration and how far they are location specific. Analysing datasets from some of the other SCELI localities would throw light on this question. See Davies and Flowerdew (1992) for some early comparative work.

Next:References

Home page | Contents | Previous

References

1. Borhani Haghighi, A. and Davies, R. B. (1999a), Characterising temporal effects in social science data, *Computational Statistics and Data Analysis*, forthcoming
2. Borhani Haghighi, A. and Davies, R. B. (1999b), How migration propensity varies with age; the effects of life cycle and individual level characteristics, *Environment and Planning A*, forthcoming
3. Boyle, P., Halfacree, K. and Robinson V. (1998), *Exploring Contemporary Migration*, Longman
4. Coleman, J. S. (1973), *The Mathematics of Collective Action*, Heinemann
5. Cote, G. L. (1997), Socio-economic attainment, regional disparities and internal migration, *European Sociological Review* **13**, No.1, p. 55-77.
6. Davies, R. B. and Flowerdew R. (1992), Modeling migration careers, using data from a British survey, *Geographical Analysis*, **24**, No.1, p. 35-57
7. Devis, T. (1983), People changing address:1971 and 1981, *Population Trends*, **32**, p.15-20.
8. Dex, S. (1995), The reliability of recall data: A literature review, *Bulletin de Methodologie Sociologique*, **49**, p. 58-80.
9. Dex, S. and McCullough, A. (1998), The reliability of retrospective unemployment history data, *Work, Employment and Society*, **12**, no.3, p. 497-509.
10. Ellis, M., Barff, R., and Renard, B.(1993), Migration regions and interstate labor flows by occupation in the United States, *Growth and Change*, **24**, No. 2, p. 166-190.
11. Greenwood, M. J.(1985), Human migration: Theory, models and empirical studies, *Journal of Regional Science*, **25**, No. 4, p. 521-544.
12. Goss, E. P. (1985), General skills, specific skills and the migration decision, *Regional Science Perspective*, **15**, p. 17-26.
13. Grundy, E. M. C.(1989), *OPCS Longitudinal Study - Women's migration: Marriage, fertility and divorce*, HMSO
14. Heckman, J. J. and Singer, B. (1984), A method of minimising the impact of distributional assumptions in econometric models of duration, *Econometrica*, **52**, p. 271-320.
15. Heckman, J. J. and Singer, B. (1985), Social Science duration analysis, *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, p. 39-58.
16. Hertzog Jr., H. W., Schlottmann, A.M., and Boehm, T. P. (1993), Migration as a spatial job-search: A survey of empirical findings, *Regional Studies*, **27**, No. 4, p. 327-340.
17. Huff, J. O. and Clark, W. A. V. (1978), Cumulative stress and cumulative inertia: A behavioral model of decision to move, *Environment and Planning A*, **10**, p. 1101-1119.
18. Lancaster, T. (1979), Econometric methods for the duration of unemployment, *Econometrica*, **47**, No. 4.
19. Lancaster, T. and Nickell, S. (1980), The analysis of re-employment probabilities for the unemployed, *Journal of the Royal Statistical Society A* , **143**, Part 2, p. 141-165.
20. Liaw, K. L. (1990), Joint effects of personal factors and ecological variables on the interprovintial migration pattern of young adults in Canada: A nested logit analysis, *Geographical Analysis*, **22**, No. 3, p. 189-208.
21. Long, L. L. (1972), The influence of the number of children on residential mobility, *Demography* , **9**, No. 3.
22. Mc Ginnis, R. (1968), A stochastic model of social mobility, *American Sociological Review*, p. 712-722
23. Salt, J. (1990), Organisational labour migration: Theory and practice in the United Kingdom, in *Labour Migration*, ed. J. H. Johnson and J. Salt, p. 52-69 (David Fulton)
24. Sandefur, G. D. and Scott, W. J. (1981), A dynamic analysis of migration: an assessment of the effects of age, family and career variables, *Demography*, **18**, No. 3, p. 355-368.
25. Warnes, A. M. (1983), Migration in late working age and early retirement, *Socio-Economic Planning Sciences*, **17**, p.291- 302.

Home page | Contents | Previous

Response variable

The response variable (move/no move) is binary, indicating for each calendar year whether there was a migration move.

Explanatory variables

- Age in years, at the beginning of the year
- Year = year-1900.
- Duration of stay (dur), years since last inter-county move
- Educational qualification (ed), with 5 levels:

  1=Degree or equivalent; professional qualifications with a degree

  2=Education above A-level but below degree level; includes professional qualifications without a degree

  3=A-level or equivalent

  4=Other educational qualification

  5=None
- Presence of children in age group 11-12 (ch1)
- Presence of children in age group 13-14 (ch2)
- Presence of children in age group 15-16 (ch3)
- Presence of children in age group 17-19 (ch4)
- Marital status (msb), at the beginning of the year, with 5 levels:

  1=Single

  2=Married

  3=Separated

  4=Divorced

  5=Widowed
- Marital status (mse), at the end of the year, with 5 levels as for (msb)
- Employment status (esb), at the beginning of the year, with 8 levels:

  0=Not working

  1=Self employed with 25 or more employees

  2=Self employed with fewer than 25 employees

  3=Self employed without employees

  4=Manager in establishment with 25 or more employees

  5=Manager in establishment with fewer than 25 employees

  6=Foreman/Supervisor

  7=Employee, including family workers, apprentices and trainees
- Employment status (ese), at the end of the year, with 8 levels as for (esb)
- Occupational status (osb), at the beginning of the year, with 12 levels:

0=None
10=Service class, higher
20=Service class, lower
31=Routine non-manual, clerical and administrative
32=Routine non-manual, distributive
41=Small proprietors (1-25 employees)
42=Small proprietors (no employees)
43=Farmers and smallholders
50=Supervisors of manual workers, low grade technicians
60=Skilled manual
71=Semi and unskilled manual
72=Agricultural workers

- Occupational status (ose), at the end of the year, with 12 levels as for (osb).
- Marital break-up (mbu), change in marital status over the year defined as
  mbu=1 if msb=2 and mse=3 or 4
  mbu=0 otherwise
- Remarriage (mrm), change in marital status over the year defined as
  mrm=1 if msb=3 or 4 and mse=2
  mrm=0 otherwise
- First marriage (mfm), indicates transition from single to married state
  mfm=1 if msb=1 and mse=2
  mfm=0 otherwise
- Collapsed marital status variable (msb1) is defined as
  msb1=1 for single
  msb1=2 for married
  msb1=3 for separated/divorced/widowed
- Promotion to manager (epm) is defined as
  epm=1 if esb=6 or 7 and ese=4 or 5
  epm=0 otherwise
- Obtaining a job (eoj) is defined as
  eoj=1 if esb=0 and ese is >0
  eoj=0 otherwise
- Collapsed employment status variable (esb1) is defined as
  esb1=1 if esb=1,2 or 3 (self-employed)
  esb1=2 if esb=4 or 5 (manager)
  esb1=3 if esb=6 or 7 (employee, foreman, supervisor)
  esb1=4 if esb=0 (not working)
- Promotion to service class (ops) is defined as
  ops=1 if osb is >30 and ose=10 or 20

ops=0 otherwise

- Collapsed occupational status variable (osb1) is defined as
  osb1=1 if osb=0
  osb1=2 if osb=71, 72
  osb1=3 if osb=60
  osb1=4 if osb=50
  osb1=5 if osb=41,42,43
  osb1=6 if osb=31,32
  osb1=7 if osb=20
  osb1=8 if osb=10
- Collapsed marital status, msb1 variable recoded as (msb2):
  msb2=1 if msb1=2 (married)
  msb2=0 otherwise
- Collapsed employment status, fep variable recoded as (esb2):
  esb2=1 if esb1=1 (self-employed)
  esb2=2 if esb1=2 or 3 (manager, employee, foreman, supervisor)
  esb2=3 if esb1=4 (not working)
- Collapsed occupational status, osb1 variable recoded as (osb2):
  osb2=1 if osb1=1,2,3,6,7
  osb2=2 if osb1=4
  osb2=3 if osb1=5
  osb2=4 if osb1=8
- Collapsed occupational status, osb2 variable recoded as (osb3):
  osb3=1 if osb2=2 or 3
  osb3=0 otherwise

Poisson model: Calculation of expected frequency of migration

If we assume a constant migration rate r = 0.049 migrations per individual per year, then for a **ten year** period the mean migration rate is m = 0.49 moves per individual.

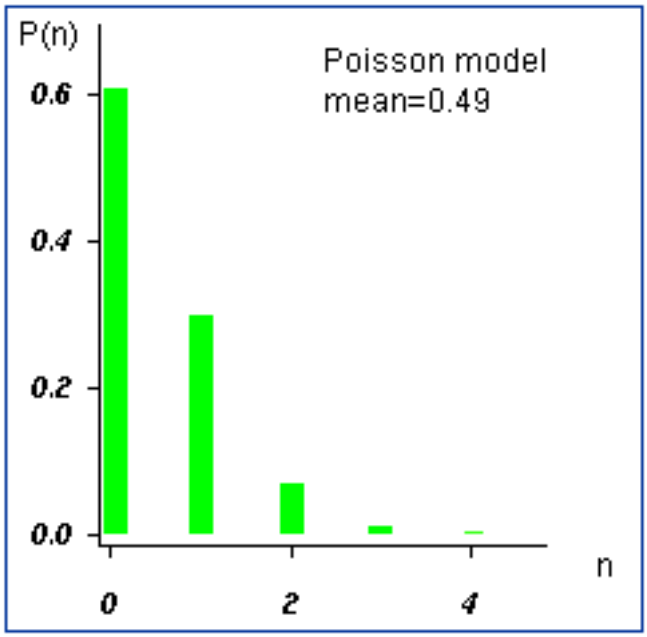
Using the Poisson model, the probability P(n) of n migrations per individual over **ten years** is given by:

$$P(n) = m^n \exp(-m)/n!$$

Substituting in this formula gives the following results:

n	P(n)
0	0.613
1	0.300
2	0.073
3	0.012
4	0.001
5	< 0.001
>=6	< 0.001

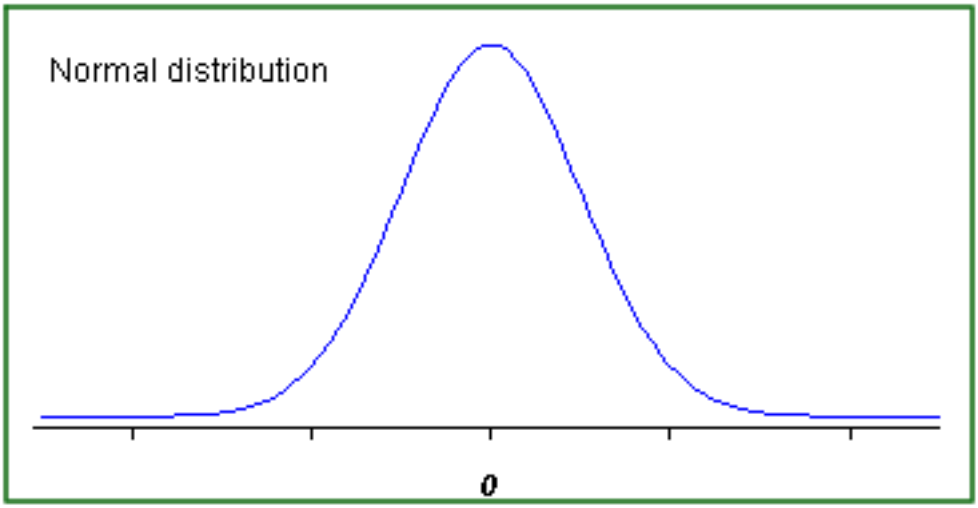
where P(n >=6) = 1-P(0)-P(1)-P(2)-P(3)-P(4)-P(5)



If there are *N* individuals in our sample, each with a **ten year** time exposure to migration opportunities, we can calculate the expected number to make n moves in this period by multiplying P(n) by *N*. Therefore out of a population of 100, for example, we expect 61.3 not to move, 30 to move once, 7.3 to move twice and so on.

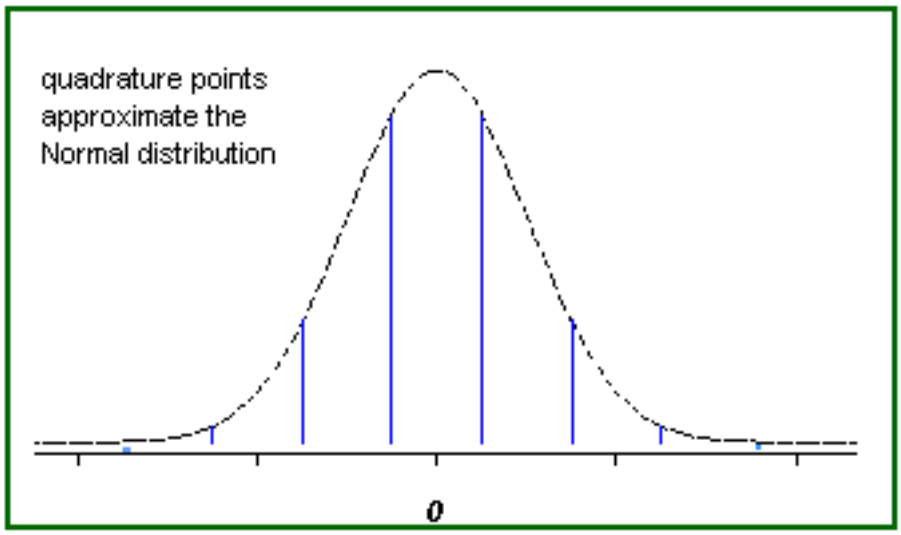
To work out the expected number of individuals making n moves for the whole data set, we repeat the above calculation for all lengths of migration history in the data and sum the results.

In practice, the calculation for Table 2 is most easily done by using an appropriate statistical computing package (such as GLIM) to fit the Poisson model and using the stored fitted mean migration rate corresponding to each individual's migration history to work out the expected probabilities of migration.

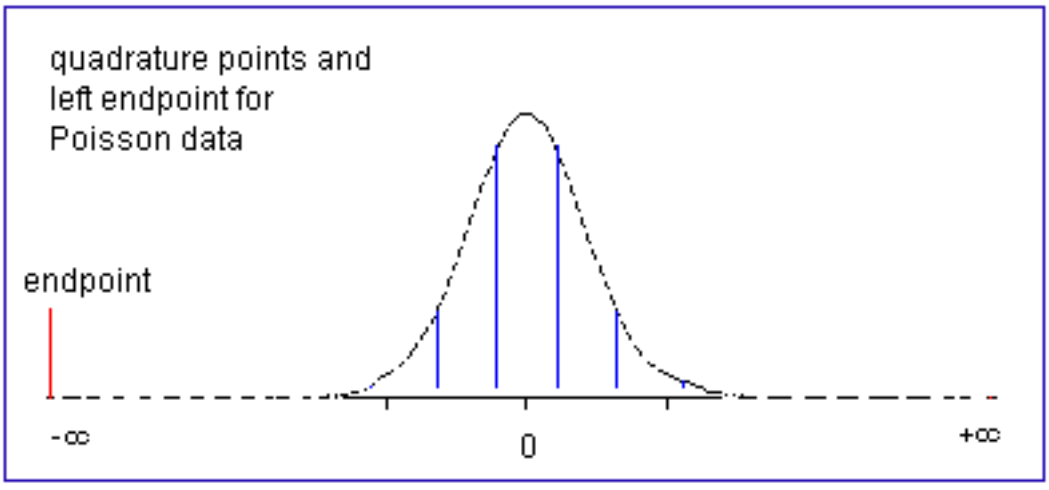


SABRE assumes a Normal probability distribution for the nuisance parameter with mean zero and standard deviation estimated from the data.

For ease of computation SABRE approximates the Normal distribution by a number of mass (or quadrature) points with specified probabilities at given locations, illustrated by the vertical lines. Increasing the number of quadrature points (see *MASS* command) increases the accuracy of the computation at the expense of computer time. The default number of quadrature points used by SABRE is eight.



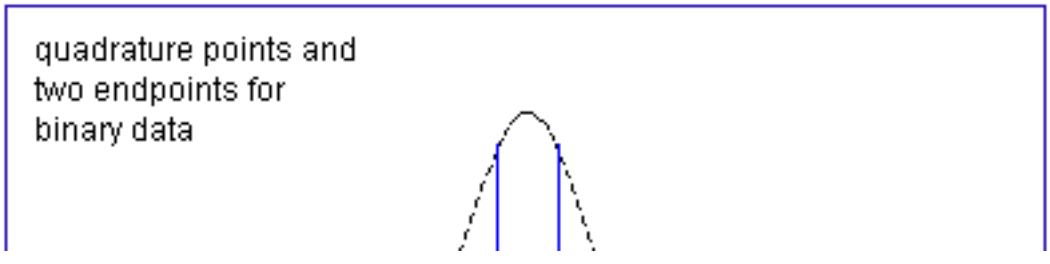
To compensate for the limitations of the Normal assumption for the distribution of the nuisance parameter (ie. tending to zero at the extremes), SABRE can supplement the quadrature points with endpoints (ie. delta functions at plus and/or minus infinity) with unknown probabilities which are estimated from the data.



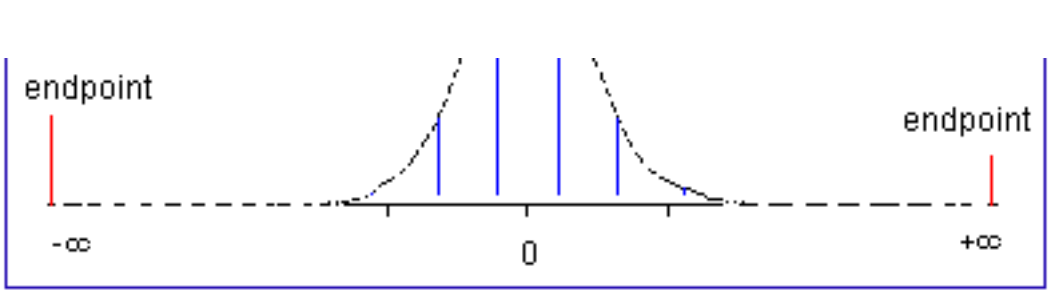
For the Poisson model a single left endpoint at minus infinity is included by default, implying zero probability of migration.

For binary data two endpoints are included by default, at plus and minus infinity.

The default settings can be changed by using the



ENDPOINT command.



Endogenous and exogenous variables

In the social sciences, interest often focuses on the *dynamics* of social or economic processes. Social science theory suggests that individual behaviour, choices or outcomes of a process are directly influenced by (or are a function of) previous behaviour, choices or outcomes. For instance, someone employed this week is more likely to be in employment next week than someone who is currently unemployed; someone who voted for a certain political party in the last elections is more likely to vote for that party in the next elections than someone who did not.

When analysing observational data in the social sciences, it is necessary to distinguish between two different types of explanatory variable; those which are **exogenous** (or external) to the process under study (for example age, sex, social class and education in studies of voting behaviour), and those which are **endogenous** . Endogenous variables have characteristics which in some way relate to previous decisions, choices or outcomes of a process. For example, in a study of voting behaviour *previous vote*, being a previous decision, is an endogenous variable; in the study of migration, *duration of stay* since the last residential move is endogenous as it relates to previous migration behaviour.

Endogenous variables may be seen as proxy variables for the many unmeasured and unmeasurable factors which affect individual choice or behaviour and which are therefore necessarily omitted from analyses. Thus *voting choice* may be seen as a proxy for individual social, economic and psychological characteristics, while *duration of stay* in a locality is a proxy for all the unknown social and economic factors which affect an individual's propensity to move.

Endogenous variables create problems in statistical analyses, because being related to the **outcomes** of the process of interest they will, by definition, be a function of the unobserved variables which govern the process. They will therefore be correlated with the random variation (or error structure) of the outcome. This leads to an infringement of the *basic* regression model assumption that the explanatory variables included in the model are independent of the error term. The consequence of this violation is risk of substantial and systematic bias.

In the presence of endogenous variables the **basic** statistical models are not robust against the infringement of assumptions. Expressed technically, parameter estimation is not *consistent*, ie. there is no guarantee that the parameter estimates will approach their true values as the sample size increases. *Consistency* is usually regarded as the minimum requirement of an acceptable estimation procedure.

To avoid spurious relationships and misleading results, with endogenous variables it is essential to use **longitudinal** data and models in which there is control for omitted variables. Longitudinal data, and in particular repeated measures on individuals are important because they provide scope for controlling for individual specific variables omitted from the analysis.

The conventional approach to representing the effect of omitted variables is to add an individual specific random term to the linear predictor, and to include an explicit distribution for this random term in the model.

There is no single agreed terminology for models which include this random tem. In econometrics the models are called random effect models; in epidemiology, frailty models; and statisticians also refer to them as multilevel models, mixture models or heterogeneous models. Models without random effects are sometimes called homogeneous models. An alternative terminology describes models without random effects as marginal models and models with random effects as conditional models. Marginal models correspond closely to the "population averaged" formulations

used in the General Estimating Equation literature.

- It is important to note that when interest focuses on the **causal relationship** in social processes inference can only be drawn by using **longitudinal data** and models in which there is control for unobserved (or **residual) heterogeneity**. Although this approach does not overcome all the problems of cross-sectional analysis with endogenous variables, there is ample evidence that it greatly improves inference.

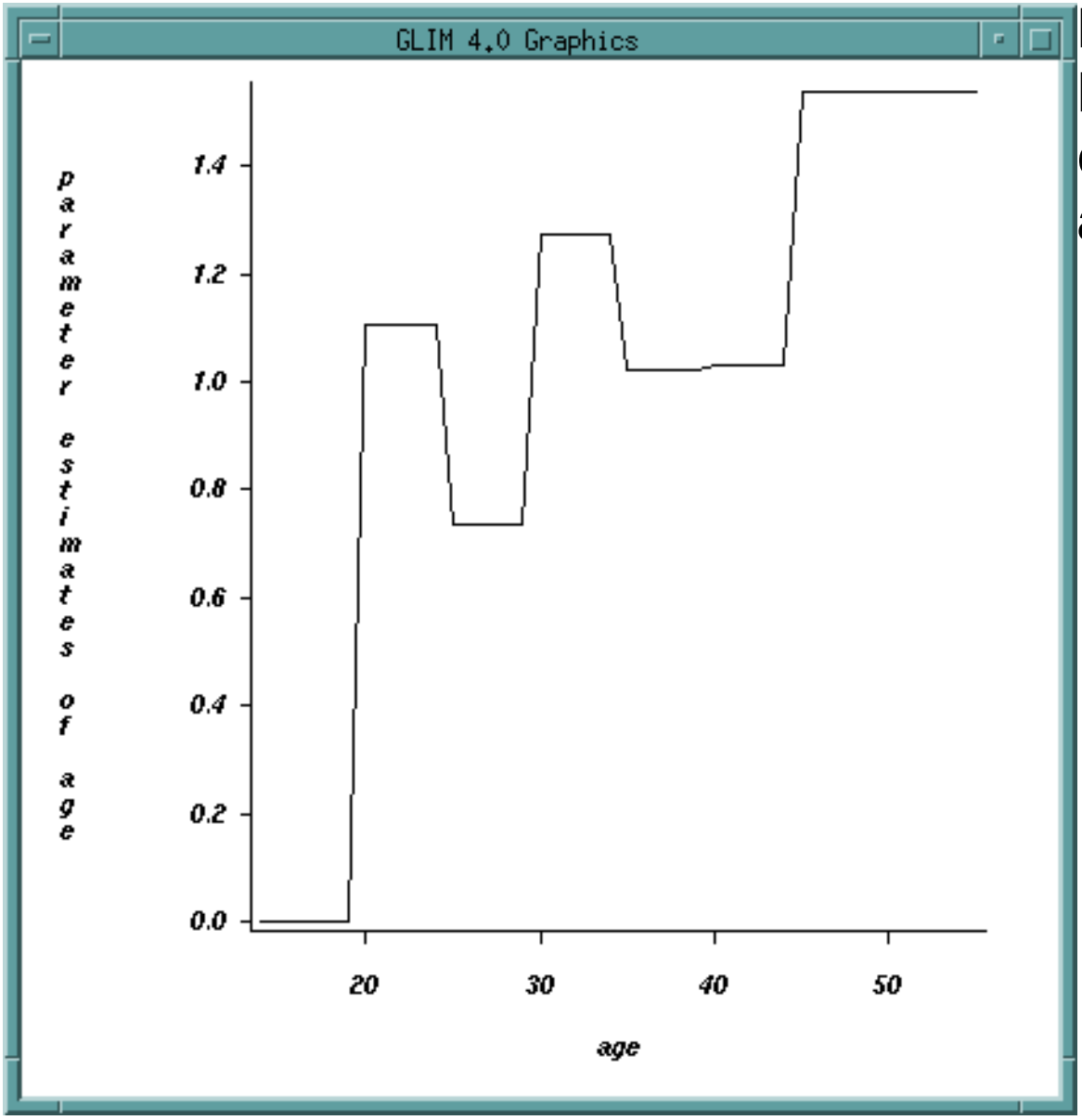


Figure 1: Parameter estimates of age

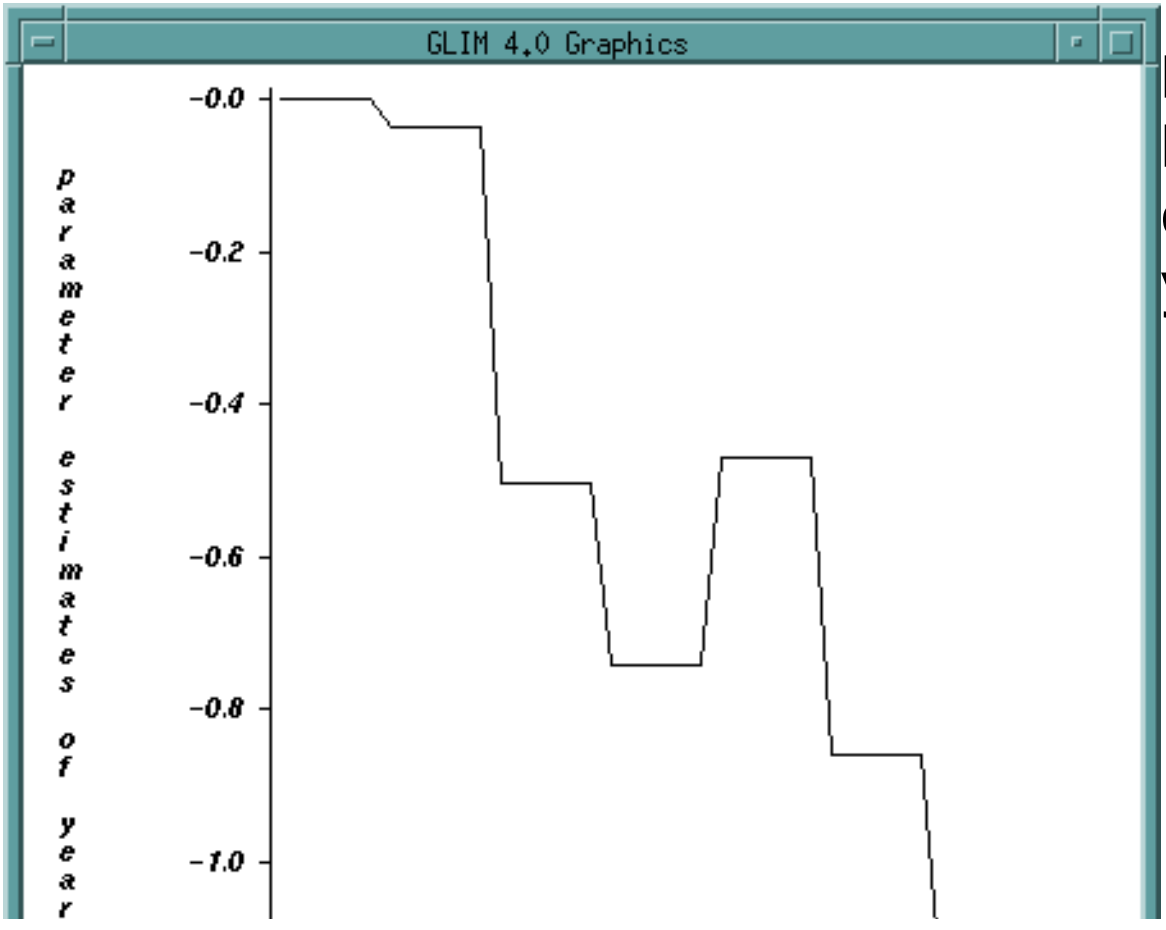
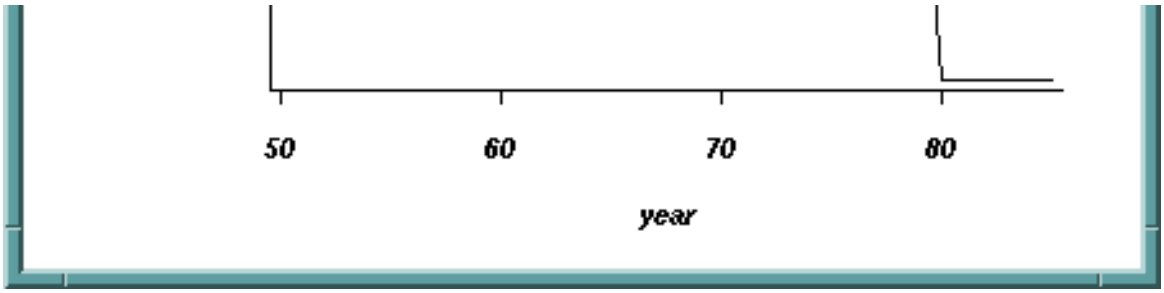


Figure 2: Parameter estimates of year



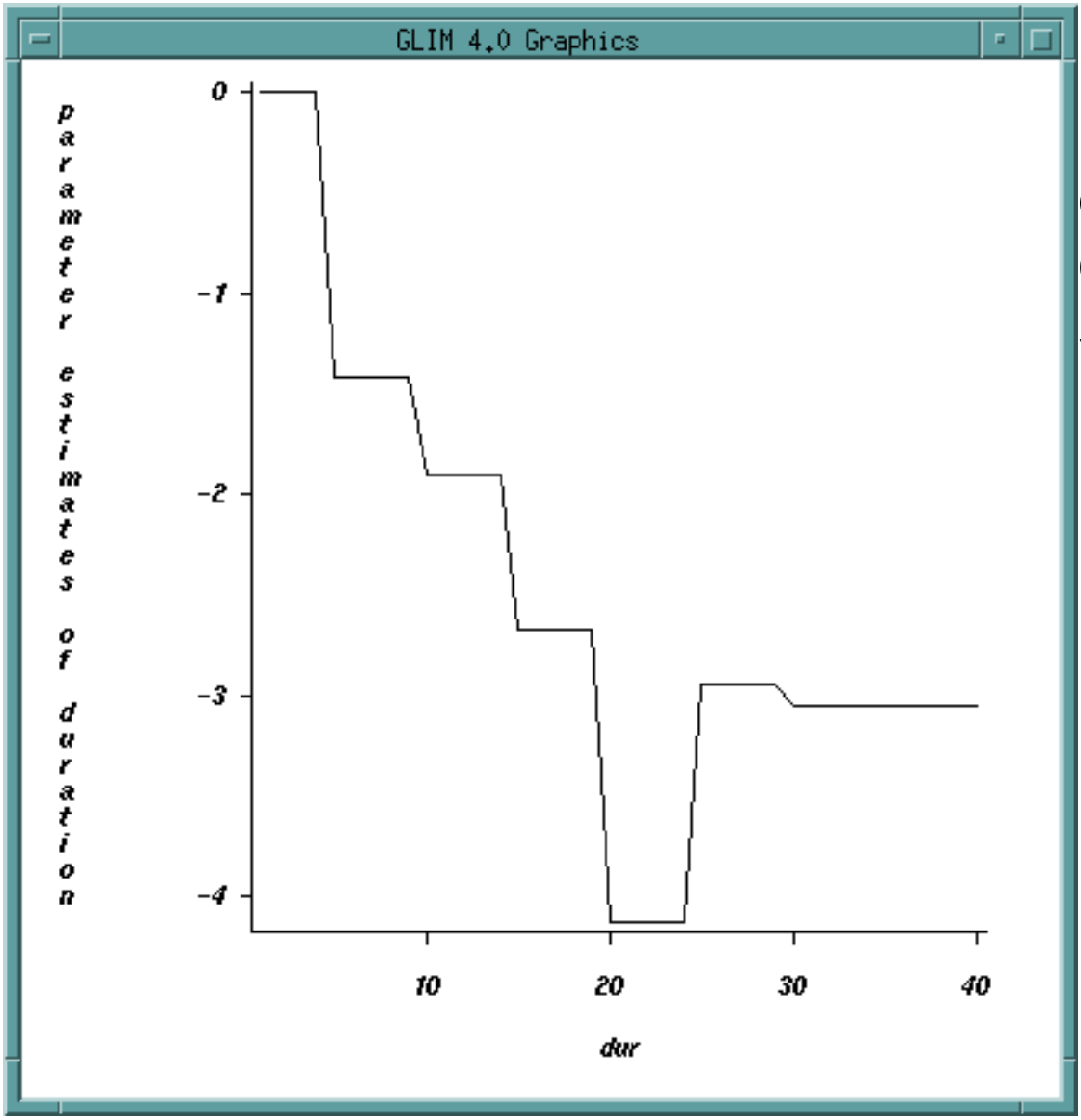


Figure 3: Parameter estimates of duration of stay

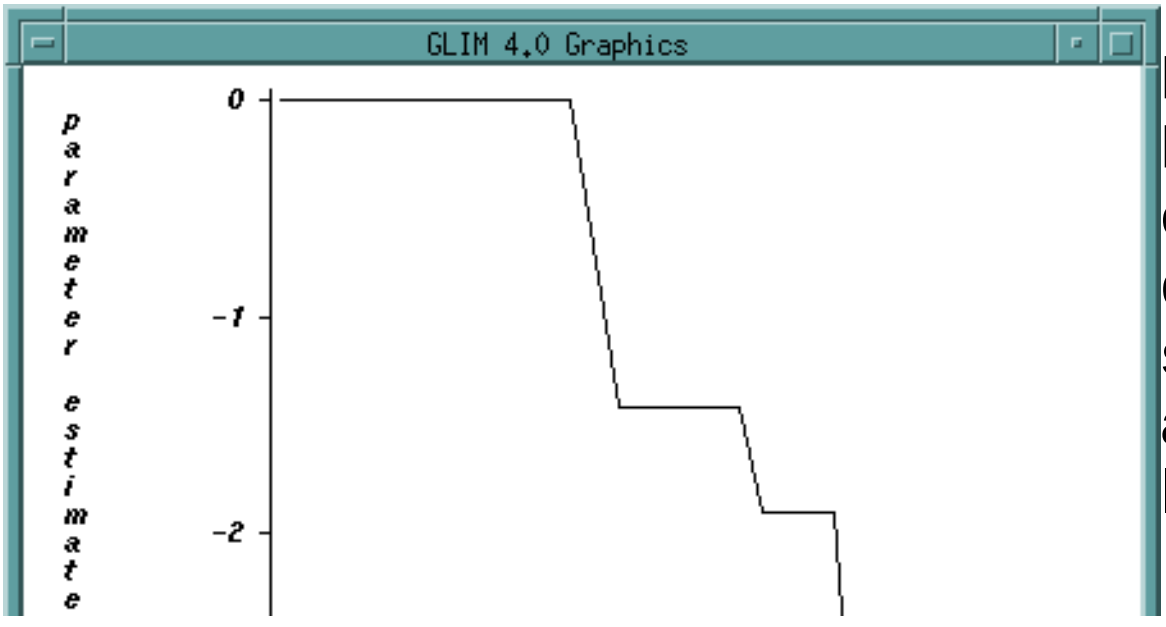
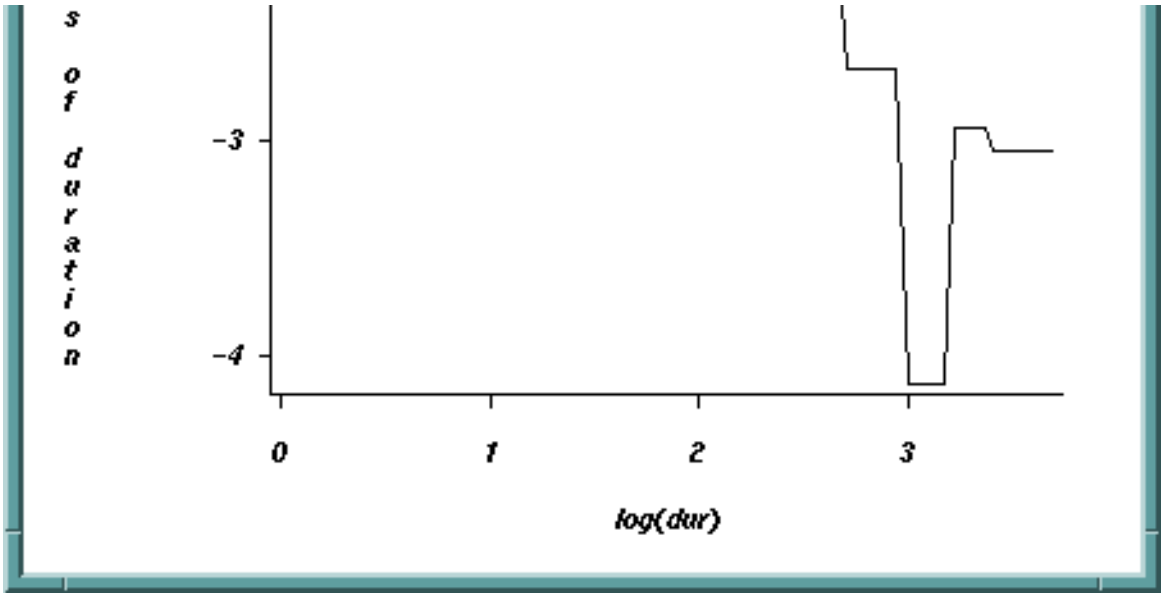


Figure 4: Parameter estimates of duration of stay plotted against log(duration)



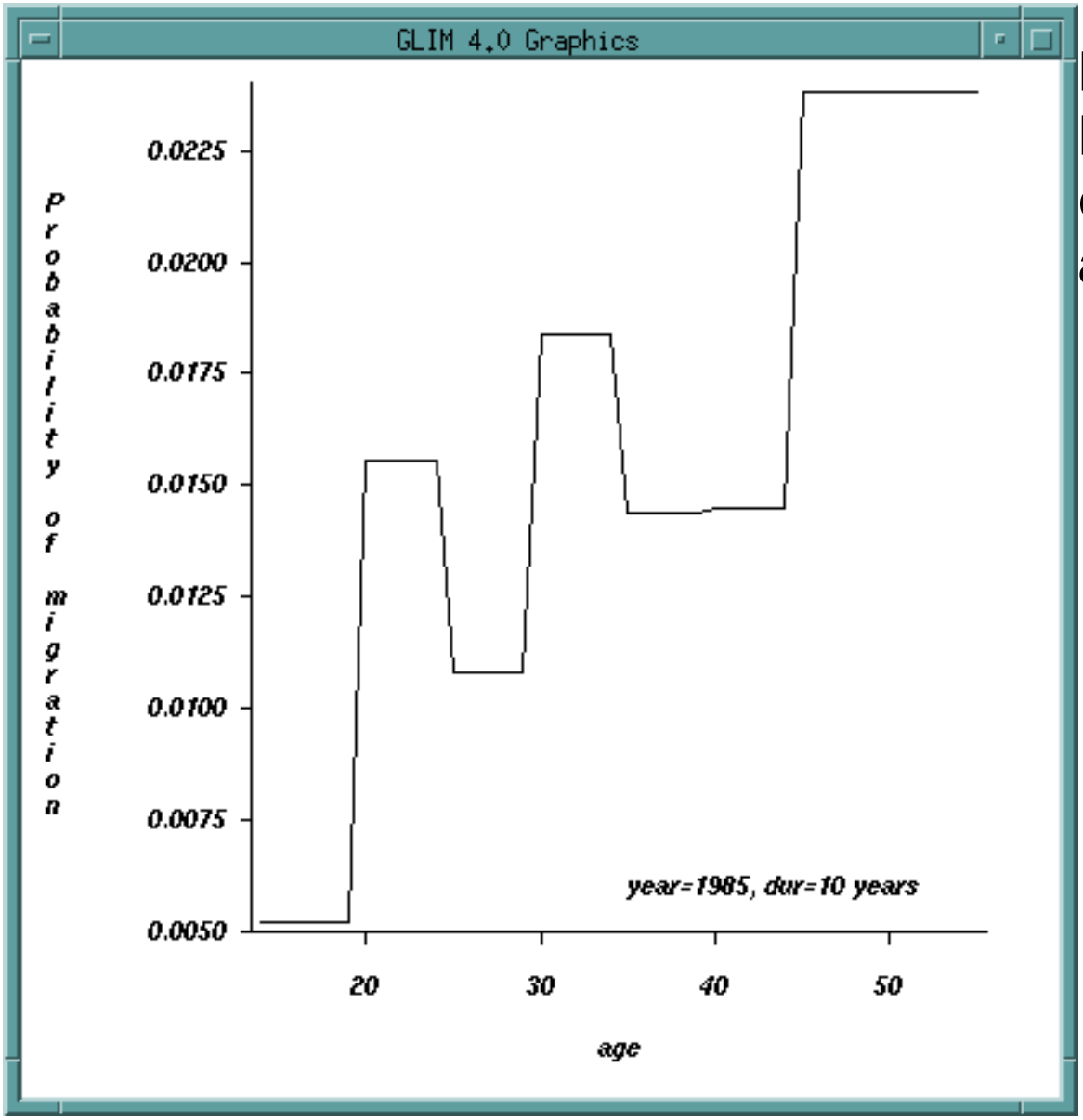
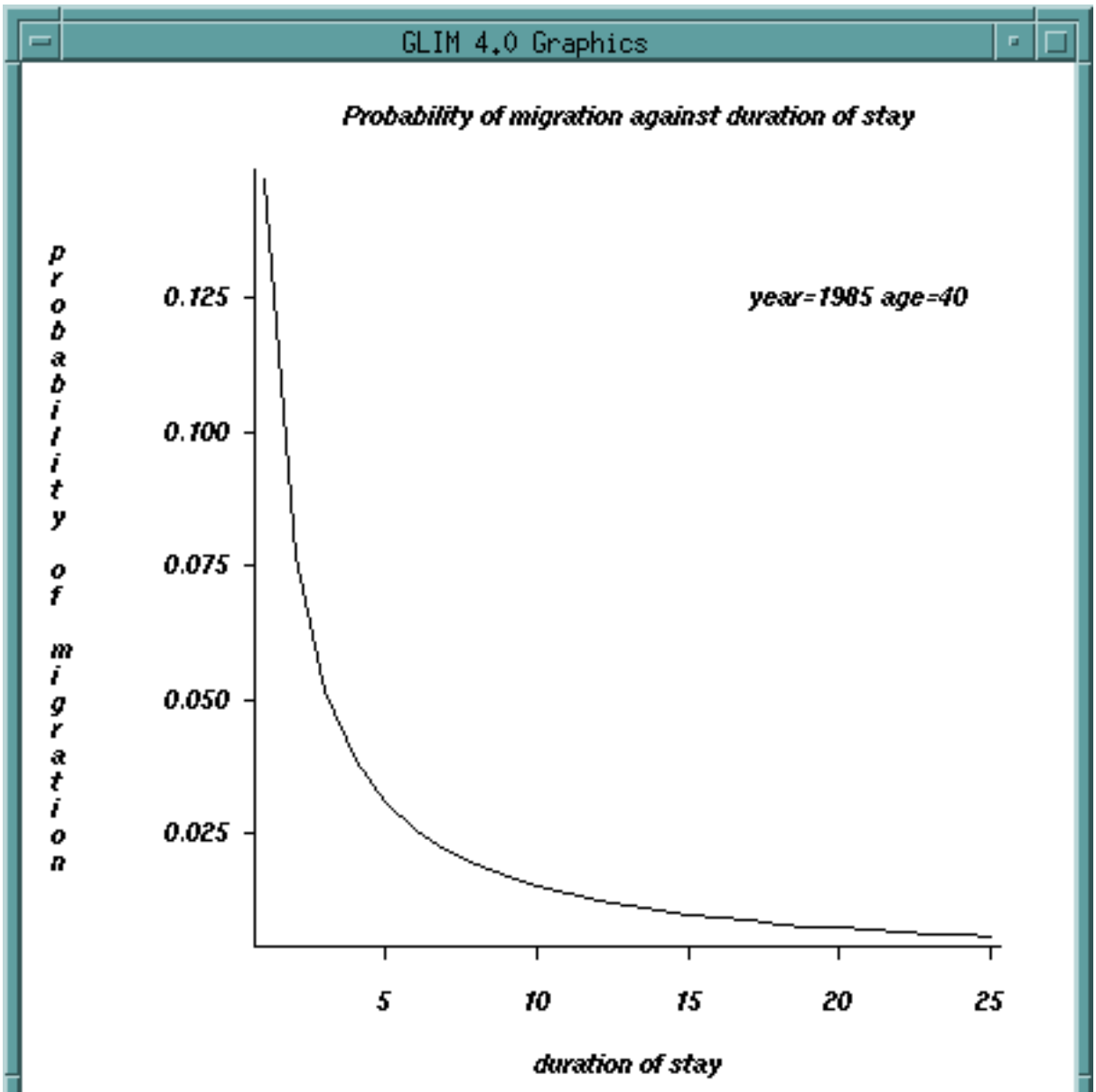
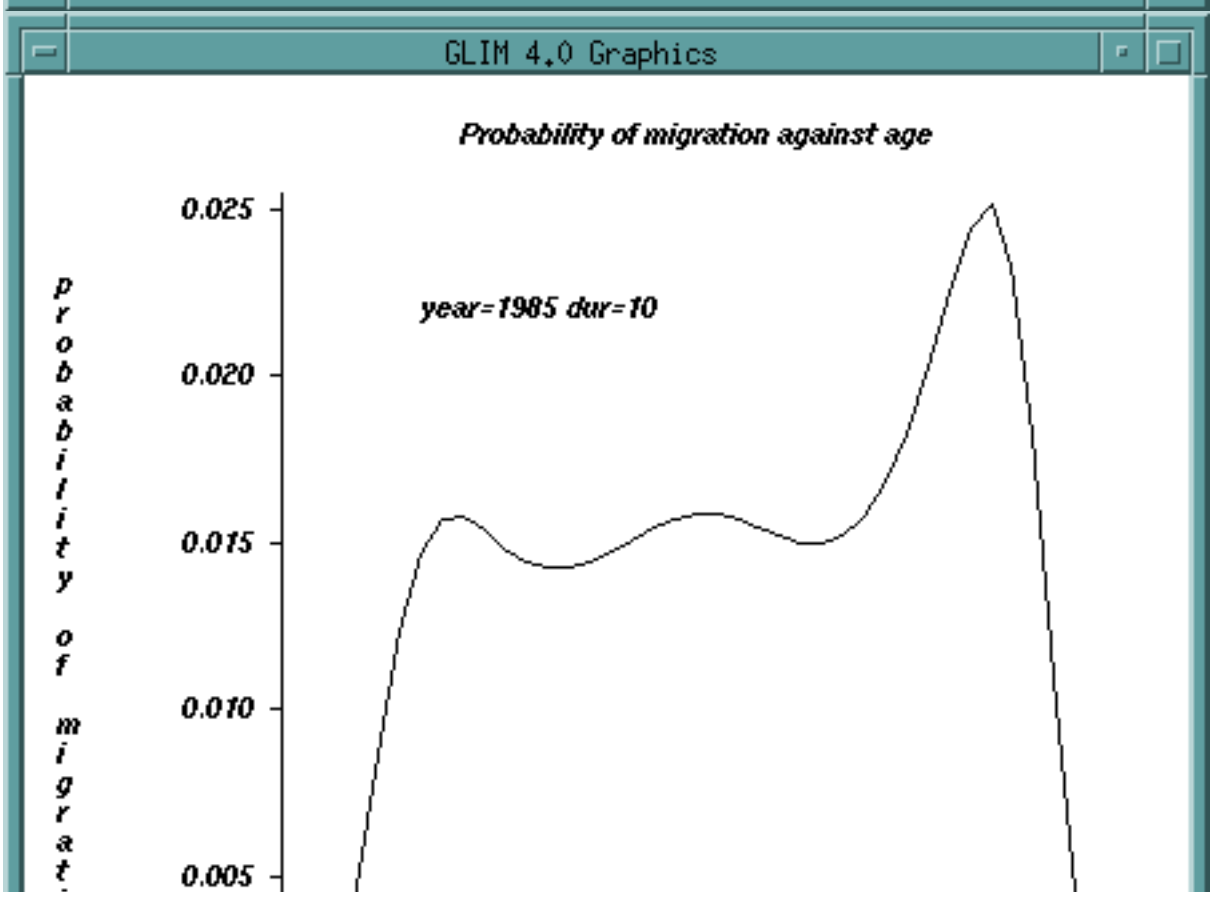
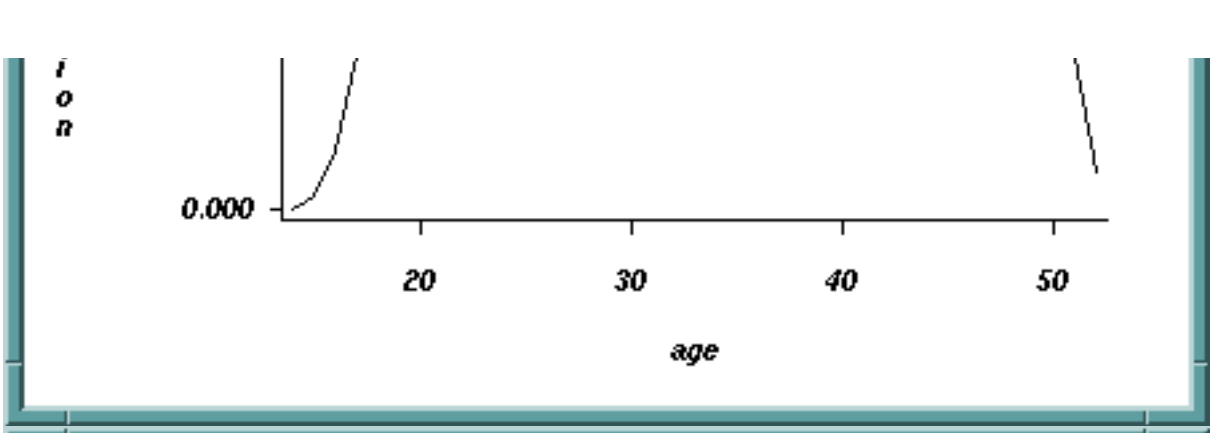
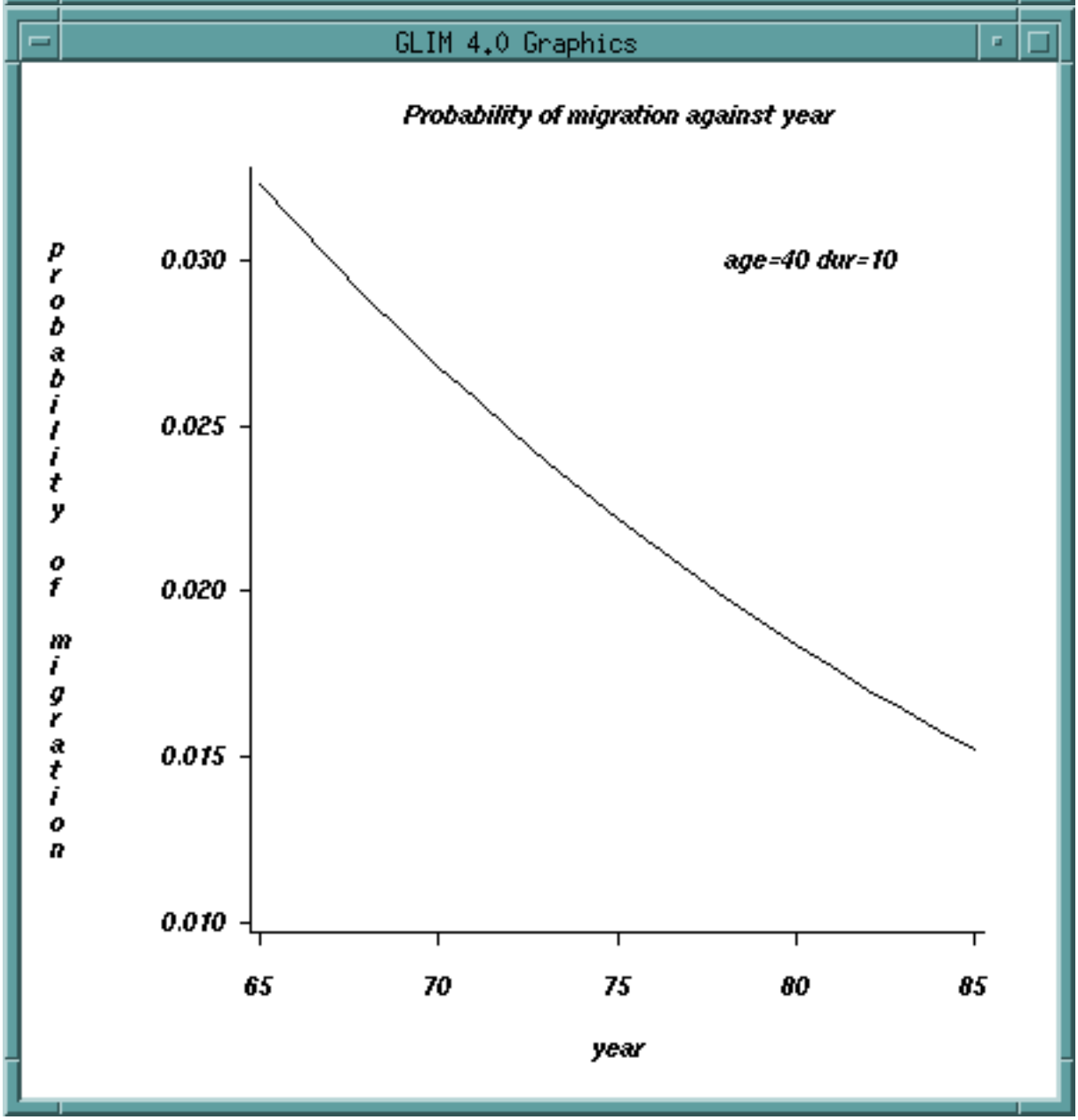


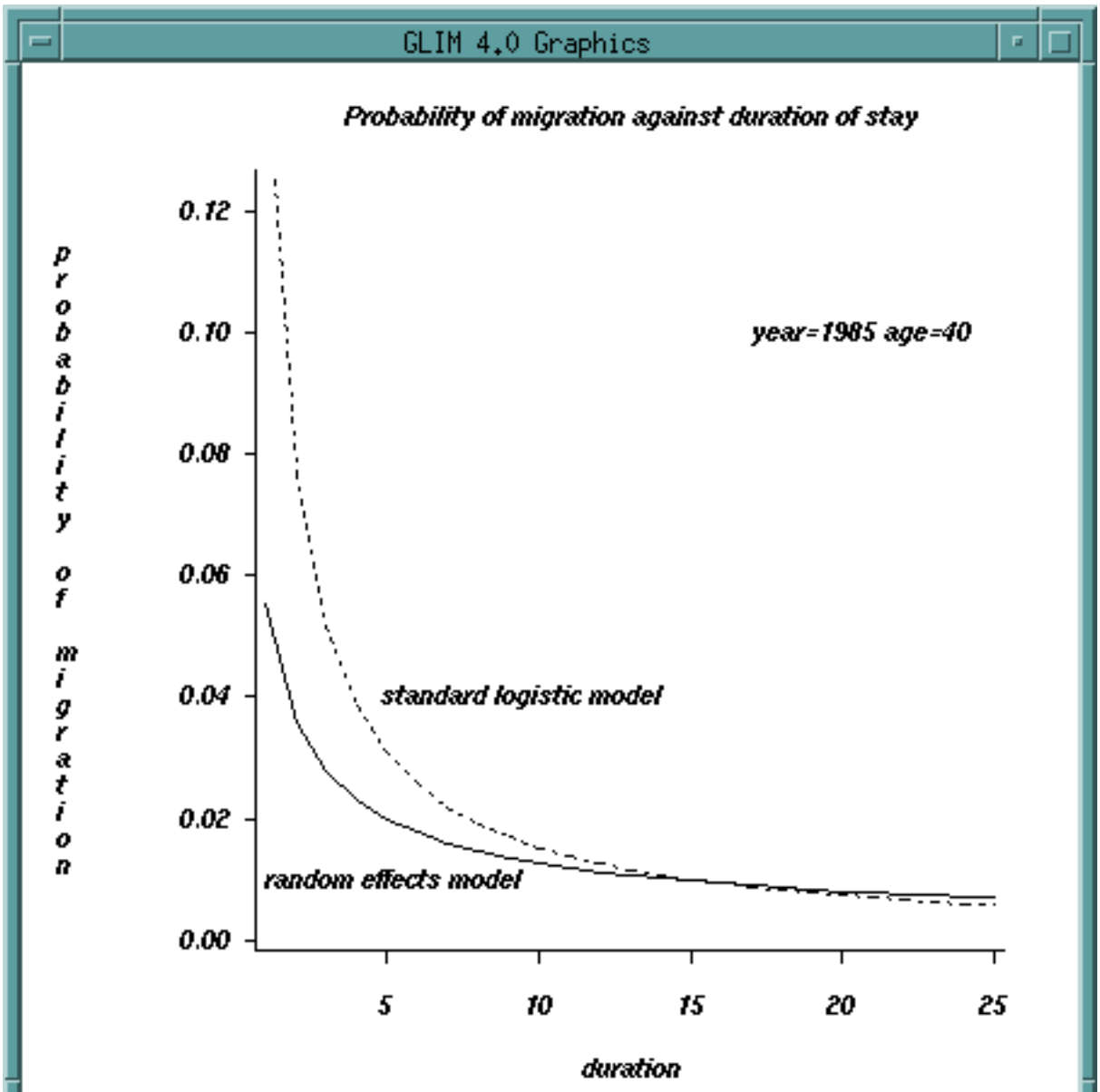
Figure 5: Probability of migration against age

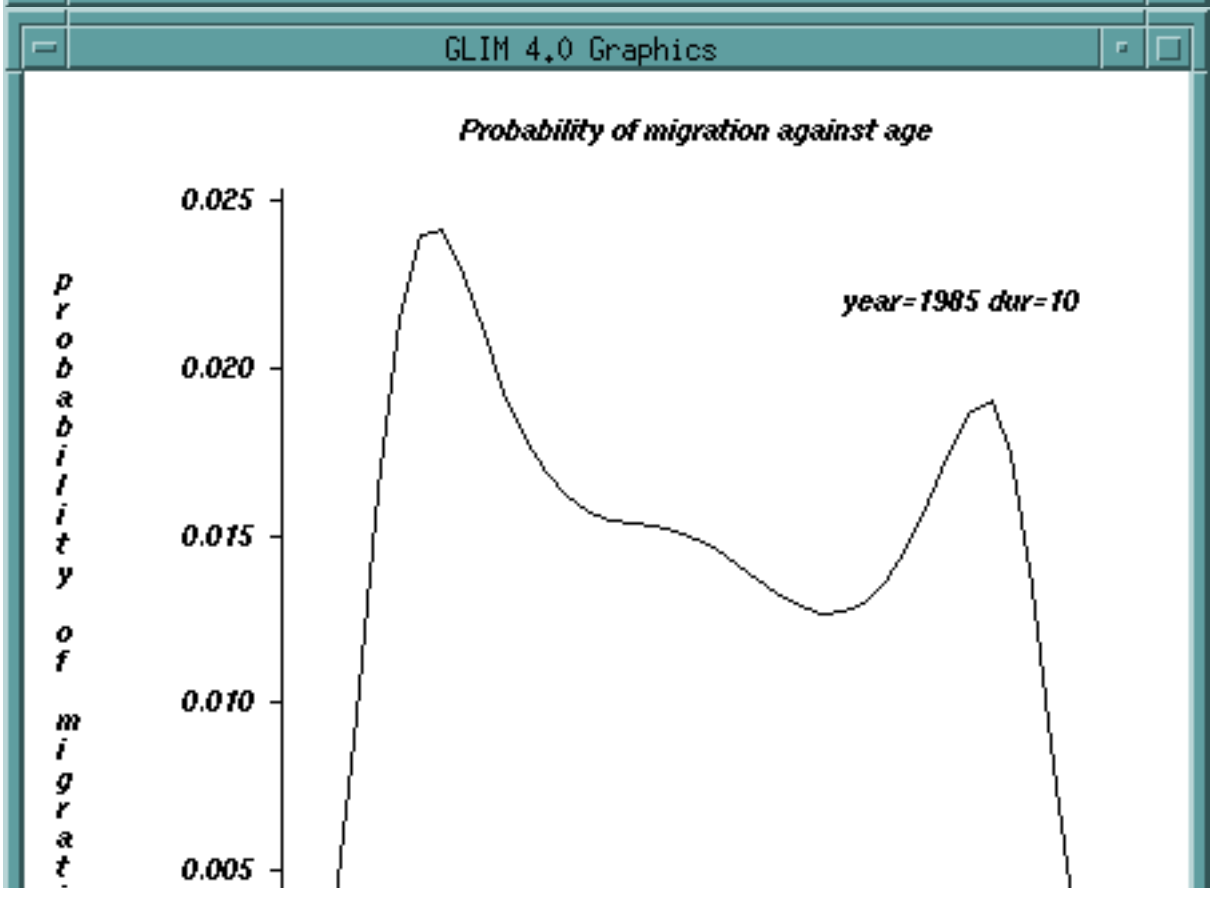

GLIM 4.0 Graphics
Probability of migration against duration of stay
probability of migration
0.125
0.100
0.075
0.050
0.025
year=1985 age=40
5
10
15
20
25
duration of stay

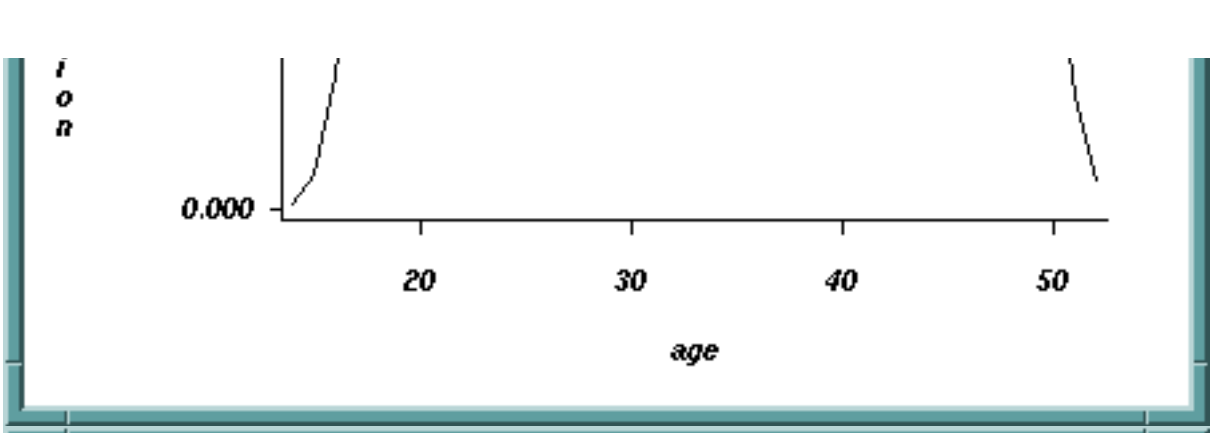

GLIM 4.0 Graphics
Probability of migration against age
probability of migrat
0.025
0.020
0.015
0.010
0.005
year=1985 dur=10


i
o
n
0.000
20
30
40
50
age

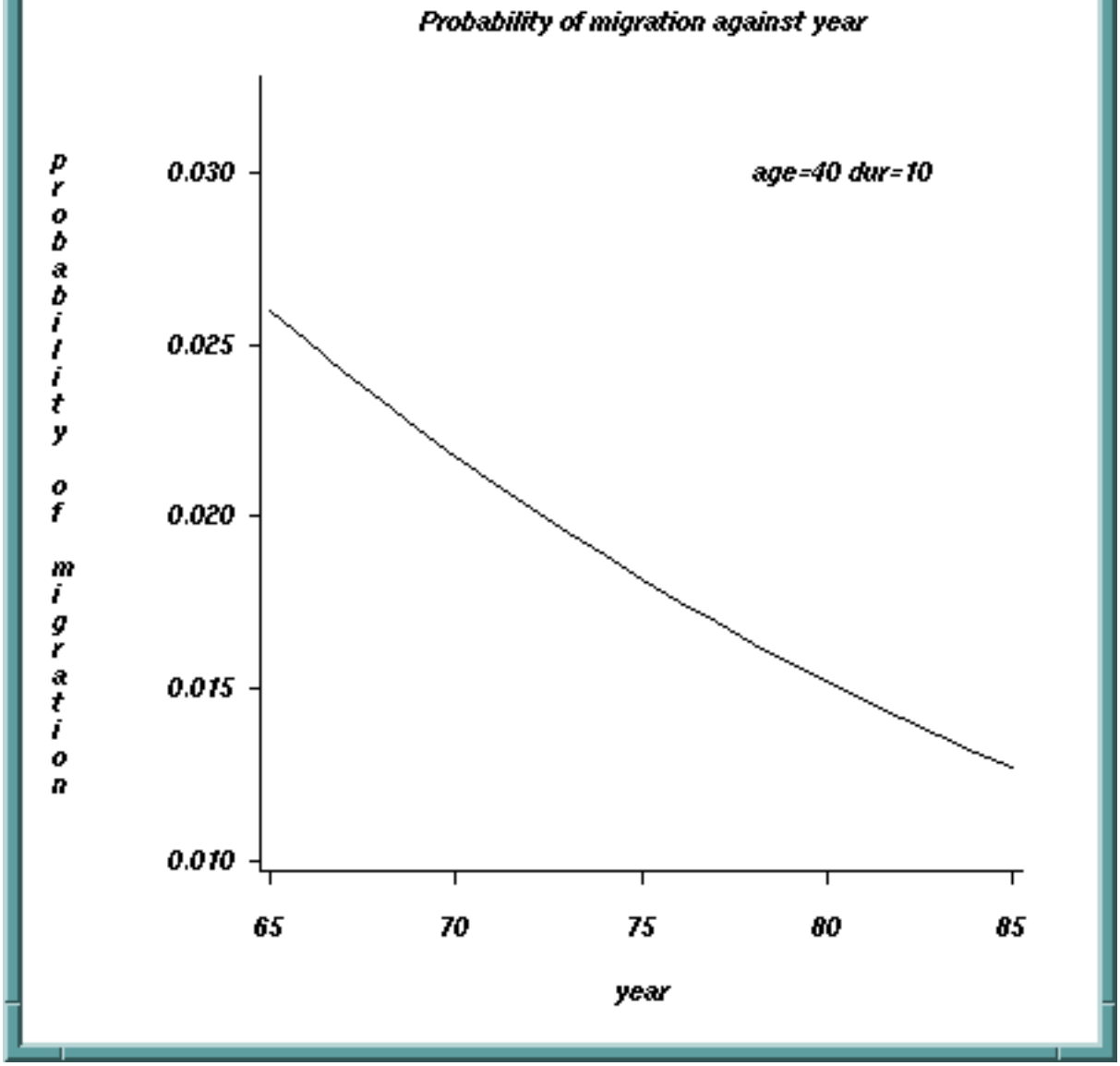

GLIM 4.0 Graphics
Probability of migration against year
p r o b a b i l i t y o f m i g r a t i o n
0.030
0.025
0.020
0.015
0.010
age=40 dur=10
65
70
75
80
85
year


GLIM 4.0 Graphics
Probability of migration against duration of stay
probability of migration
0.12
0.10
0.08
0.06
0.04
0.02
0.00
year=1985 age=40
standard logistic model
random effects model
5
10
15
20
25
duration


GLIM 4.0 Graphics
Probability of migration against age
probability of migrat
0.025
0.020
0.015
0.010
0.005
year=1985 dur=10

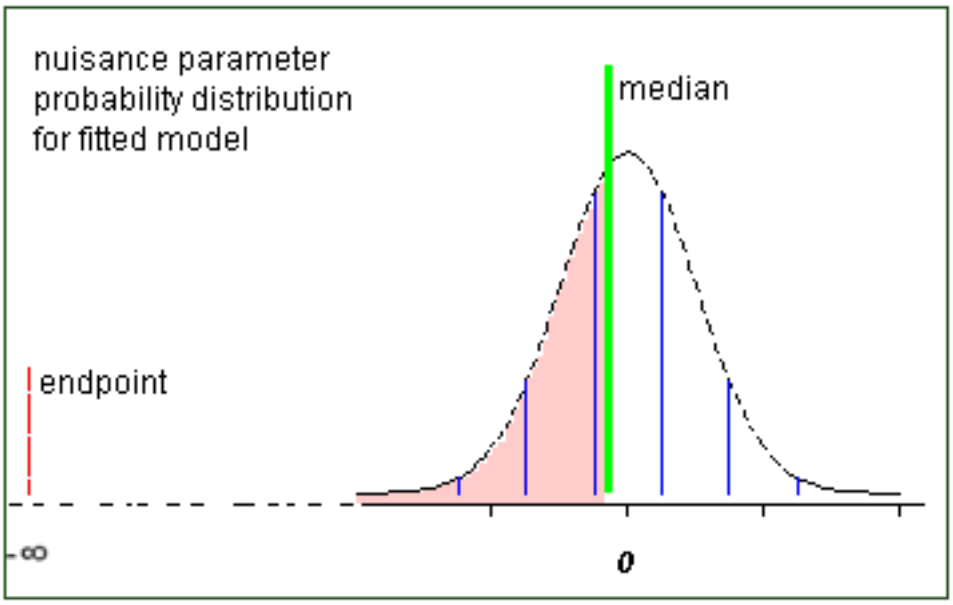


GLIM 4.0 Graphics



Calculation of the median value of the nuisance parameter

In our model the probability distribution of the individual specific random term (or nuisance parameter) e_i is represented by a Normal distribution (approximated by quadrature points) which is supplemented by endpoints at plus and minus infinity.
The fitted model shows only the left hand endpoint to be significant. Therefore the estimated probability distribution of the nuisance parameter for this data set may be represented as in the diagram:



The standard deviation (or scale parameter) σ for the Normal distribution and the probability P_L associated with the left hand endpoint are estimated from the data.
The total probability associated with the nuisance parameter (left hand endpoint plus area under the Normal curve) adds up to 1, so that the area under the Normal curve is $(1 - P_L)$.
The median splits the total probability distribution in half.

Therefore

Probability of left hand endpoint (P_L) + Probability of (e_i <= median) within Normal distribution = 0.5

Probability of (e_i <= median) within Normal distribution =
Area under the Normal curve from minus infinity to the median = $0.5 - P_L$

For the **standard Normal** distribution with mean=0, standard deviation=1 and cumulative distribution function Φ:

Φ(z) gives the area under the curve from minus infinity to z; the total area from minus infinity to plus infinity is 1.

For a **Normal** distribution with mean 0 and **standard deviation** σ, there is a scaling effect, so that

Area under the curve from minus infinity to median = Area under **standard Normal** curve from minus infinity to (median/σ) = Φ(median/σ).

In our model, there is additional scaling as the area under the Normal curve is reduced by a factor of $(1 - P_L)$.

Therefore, $(1 - P_L) * \Phi (\text{median} / \sigma) = 0.5 - P_L$

and

$\text{median} = \sigma * \Phi^{-1}[(0.5 - P_L)/ (1 - P_L)]$

From fitting the model we have found: $\sigma = 0.47710$ and $P_L = 0.36113$

Therefore median = $0.47710 * [\Phi^{-1} [(0.5 - 0.36113)/(1 - 0.36113)]$

$= 0.47710 * \Phi^{-1}(0.21737) = 0.47710 * (-0.78115) = - 0.372685$

$\Phi^{-1}( )$ may be looked up in statistical tables or obtained from statistical computing packages such as GLIM.

MODELLING MIGRATION HISTORIES

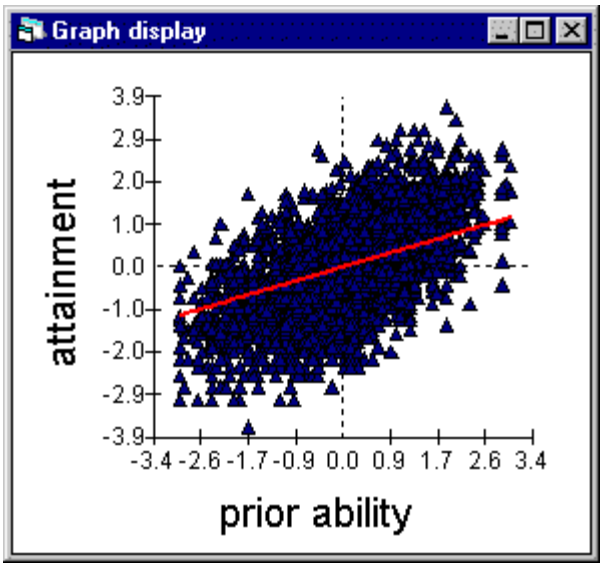
List of contents:

Home page	Previous

MLwiN - What is Multilevel Modelling?

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Tutorials
Software
Back to main site

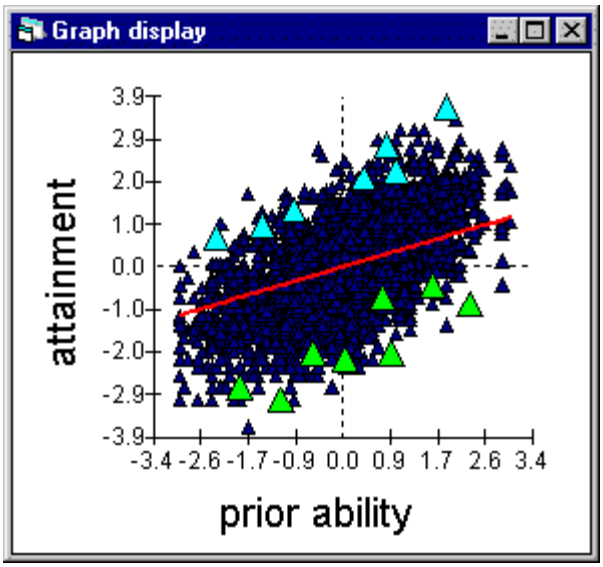
Multiple regression estimates average relationships between response (eg educational attainment) and predictor variables e.g. socio-economic status, gender, previous (baseline) ability.



The above graph illustrates a typical linear regression relationship, in this case between outcome attainment and prior-ability among a sample of students. The red line shows that on average an increase in prior ability is associated with an increase in outcome attainment.

A fundamental assumption of this regression model is that the residuals (the distance of the data points from the red regression line) are independent. However, data often have a multilevel structure which violates this assumption.

In this example students are grouped within schools. If we believe that the process of student selection by schools or the education given by schools may influence outcome attainment, then two students within a particular school will tend to be more similar than two students from different schools.



The pupils at two schools are highlighted in the above graph to illustrate this point. If we ignore the nesting of pupils within schools - that is, we analyse the data as though all pupils were independent - then we will tend to underestimate the standard errors of the regression coefficients. This problem, called "misestimated precision", means that we will tend to find too many relationships to be statistically significant.

Generally we are interested not only in the average relationship (the red line) but in how this relationship varies from school to school.

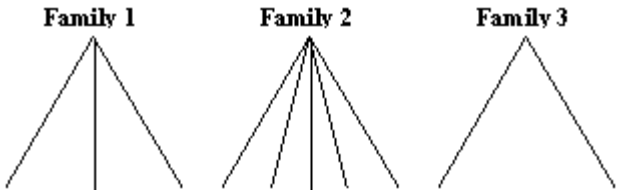
Multilevel modelling provides a powerful framework for exploring how average relationships vary across hierarchical structures.

Next Section: *Hierarchical Structures* ▶

Top

MLwiN - Hierarchical Structures

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Tutorials
Software
Back to main site



An example of a simple hierarchical structure is people nested in families. This structure has two levels: people are at level one (the lower level), and they are grouped within families at level two (the higher level).

The reason for taking the family structure into account in our analysis is because we believe that individuals from the same family are likely to be more similar than people from different families. This may be because of shared social characteristics (e.g. household income), environmental characteristics (e.g. housing conditions) or genetic predisposition.

Other examples of two level structures are students in schools or patients in hospitals. The lower level units (students or patients) are nested within higher level units (schools or hospitals).

A feature worth noting is that multilevel modelling does not require the data to be balanced i.e. the number of people in each family does not have to be the same.

Other Hierarchical Data Structures

Often social data are only available at a level above the individual – for example household or area. Such data may still have a hierarchical structure.

For example, we may have information collected on households, and the households are nested within areas. In other words, households are our level one units and areas are at level two.

In many studies we are interested in the growth or development of individuals over time. For example, we may measure the height of the same children at different ages or the voting behaviour of adults over time.

These data again form two-level hierarchies with measurement occasions at level one nested within individuals at level two.

In the above example, person 2 may have been measured on five occasions; person 1 three times and person 3 just twice. Recall that multilevel modelling does not require a balanced structure, so can accommodate a varying number of occasions for each individual.

In many situations we will find that the data structure requires more than two levels. For example, we may have pupils within classes within schools, people within households within areas, or repeated measures on pupils within classes within schools.

Next Section: *Research Questions* ▶

Top

MLwiN - Research Questions

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Tutorials
Software
Back to main site

What Kind of Research Question Can I Address with Multilevel Modelling?

How much of the variability in attainment is attributable to differences between schools and how much to differences between students within schools?

That is to what extent does the school attended influence the students' attainment?

Can we find factors at the student and school levels (for example gender, ethnicity, school size, school type) which account for the variability at either level?

Multilevel modelling allows us to determine the relative impact of each level of the hierarchy on the response and to identify the factors at each level that are associated with that level's impact.

For example, do low ability children fare better when educated alongside children of the same ability or children of higher ability? How would you begin to answer this question?

Although the above points relate to an educational example, multilevel modelling is a general technique and can address similar types of question for hierarchical structures from other disciplines.

Next Section: *Overviews* ▶

Top

MLwiN - Overviews

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Education Overview
Mortality Overview
Tutorials
Software
Back to main site

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Overviews of two analyses are available giving examples of the potential of multilevel modelling for social data.

The first example shows some findings from a multilevel analysis of educational attainment data from pupils attending a secondary school in London. The response variable is attainment in exams taken by pupils at age 16. There are data on 4000 pupils in 65 schools. The analysis is particularly concerned with the effect of schools. Are some schools more "effective" than others? *More* ▶

The second example is an exploration of variations in mortality rates in England and Wales. The data comprise repeated measures of the standardised mortality ratio (SMR) for 403 county districts which are nested in 54 counties over the period 1979 to 1991. The particular concerns are in how mortality changed over that time period and the nature and extent of regional variations. *More* ▶

The full tutorials and datasets may also be downloaded; these are designed to take you through the analyses on which the above overviews are based.

Next Section: *Tutorials* ▶

Top

MLwiN - Tutorials

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Tutorials
Software
Back to main site

This page contains the detailed tutorials. These can be opened directly or downloaded.

Educational example:

- Chapter 1 : Random intercept and random slope models
- Chapter 2: Residuals
- Chapter 3: Graphical procedures for exploring the model
- Chapter 4: Contextual effects
- Chapter 5: Variance Functions

Mortality example:

View/download tutorial

The tutorial files are in Acrobat *.pdf format. You can read Acrobat files either after copying or downloading them, or directly within a suitable web browser. If you wish to view acrobat files from within a web browser then you will need Internet Explorer 3 or later or Netscape 3.0 or later. Please consult your browser documentation for configuration information. In either case you will need to install the free Reader (version 3.0 or later) on your computer.

- If you wish to have more information about Acrobat go to Adobe's web site http://www.adobe.com/acrobat/ or go directly to http://www.adobe.com/prodindex/acrobat/readstep.html from where you will be able to download the reader.

If you wish to work through the tutorials on the example datasets with MLwiN , go to the software download page.

Next Section: *Software* ▶

Top

MLwiN - Software

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Tutorials
Software
Back to main site

- This version of MLwiN which you can download from this page is restricted to the analysis of the example datasets only. Apart from this limitation it is a fully functional version so there are no restrictions on the analyses that can be conducted on these data.
- The tutorials and the software provide an excellent introductory course into the theory and practice of multilevel modelling. They have been designed to be suitable for individual study or as a useful component of graduate or post-graduate courses in social statistics.
- To download and install the software and datasets click here.

To find out more about MLwiN , including details of how to order a full copy, visit the MLwiN web page

Top

Analysis software

Home page
Project overview
Research questions
Searching for data
Statistical modelling
Exemplar datasets
Software
MLwiN
SABRE
Download
The project team
Feedback
Search/Route map

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

SOFTWARE *by Brian Francis*

- The training materials provided on this web site are designed to be used in conjunction with the software packages SABRE and MLwiN .
- SABRE (software for the statistical analysis of binary recurrent events) is freeware and can also be downloaded from the SABRE web site. The version provided here is smaller than the standard freeware version and will run on most PCs.
- MLwiN is a commercial, licensed, Windows-based software package for fitting multilevel models - the special version of MLwiN provided here is free and fully-functional but works only with the teaching datasets provided.
- Both packages were developed and enhanced under the ESRC Analysis of Large and Complex Datasets initiative. The statistical software, together with the teaching datasets can be downloaded from the DOWNLOAD page - see the left menu.
- The download and installation of the software is straightforward. Once they have been installed, SABRE or MLwiN can be run from the START menu as with all other software. The datasets and software manuals are stored in the same directory as the software.
- Both software packages will run under WINDOWS 95, WINDOWS 98, WINDOWS NT 3.0 or 4.0 and WINDOWS 2000. They each require at least a 486 PC with 32Mb of memory or higher.
- While reading the web based training material, you can run the relevant software package in another window, reading in the relevant teaching datasets, and comparing your results with the results on the screen. Alternatively, you may prefer to print out the tutorials and work through the examples using the downloaded software and the printed tutorials.
- You can also have the opportunity to also try out your own analyses and to challenge the analyses provided by the site developers!

  For more details on each of the packages, see the left-hand menu.

Top

MODELLING MIGRATION HISTORIES

Introduction

- This example is concerned with individuals' migration histories within Great Britain, where migration is a residential move between two localities.
- Boundary choice is crucial in defining what is a migration move (White and Meuser, 1988).
- In this analysis migration is taken as an inter-county move. It is therefore concerned with moves which involve breaking away from social and community ties.
- For a recent text on migration see for instance Boyle, Halfacree and Vaughan (1998).

The data

- The data are derived from a large retrospective survey of life and work histories carried out in 1986 under the Social Change and Economic Life Initiative (SCELI), funded by the ESRC.
- The data were therefore not specifically collected for the study of migration, but were drawn from an existing data set which includes information on where individuals had lived all their working lives.
- The variables selected from the primary data set are those which are suggested in the research literature as important for explaining individual migration behaviour.
- Temporary moves of a few months duration do not imply commitment to a new area and are not regarded as migration. Migration data are therefore recorded on an annual basis.
- The respondents were aged 20 to 60 and lived in the travel-to-work area of Rochdale, just to the north of Manchester. (Rochdale was one of six localities chosen for the SCELI survey for their contrasting experience of recent economic change.)
- As the analysis is concerned with internal migration within Great Britain, individuals who had lived abroad during their working lives are excluded from the data set.
- The information for 1986 is incomplete and is therefore not included.
- The data set contains the migration histories of 348 males during their working, or potentially working lives, starting from the completion of education up to 1985.
- The data set is longitudinal, with one observation for each individual per calendar year. There are a total of 6349 annual observations.
- The start year for the collection of data for each individual is different, but the final year is the same.
- The response variable of interest is binary, indicating for each individual and for each calendar year, whether or

not there was a migration move.

- The explanatory variables are age, calendar year, duration of stay at each address, education, and information on marriage, children, employment and occupational status for each year.

NEXT:The longitudinal data set

Home page | **Contents** | **Previous**

The longitudinal data set

Typical data matrix

The longitudinal data set is stored in the file *rochmig.dat*. The data matrix for a typical individual is of the form:

Case number	50016								
Move/No move	0	0	1	0	1	1	1	1	0
Age	17	18	19	20	21	22	23	24	25
Year	77	78	79	80	81	82	83	84	85
Duration of stay (dur)	1	2	3	1	2	1	1	1	1
Education (ed)	4	4	4	4	4	4	4	4	4
Children age 11-12 (ch1)	0	0	0	0	0	0	0	0	0
Children age 13-14 (ch2)	0	0	0	0	0	0	0	0	0
Children age 15-16 (ch3)	0	0	0	0	0	0	0	0	0
Children age 17-18 (ch4)	0	0	0	0	0	0	0	0	0
Marital status (msb)	1	1	1	1	1	1	1	1	1
Marital status (mse)	1	1	1	1	1	1	1	1	1
Employment status (esb)	7	7	7	7	7	7	7	0	0
Employment status (ese)	7	7	7	7	7	7	0	0	0
Occupational status (osb)	71	71	71	71	71	71	71	0	0
Occupational status (ose)	71	71	71	71	71	71	0	0	0
Marital break-up (mbu) ***	0	0	0	0	0	0	0	0	0
Remarriage (mrm) ***	0	0	0	0	0	0	0	0	0
First marriage (mfm) ***	0	0	0	0	0	0	0	0	0
Marital status (msb1) {msb collapsed} ***	1	1	1	1	1	1	1	1	1
Promotion to manager (epm) ***	0	0	0	0	0	0	0	0	0
Obtaining a job (eoj) ***	0	0	0	0	0	0	0	0	0
Employment status (esb1) {esb collapsed} ***	3	3	3	3	3	3	3	4	4
Promotion to service class (ops) ***	0	0	0	0	0	0	0	0	0
Occupation (osb1) {osb collapsed} ***	2	2	2	2	2	2	2	1	1
Marital status (msb2) {msb1 collapsed} ***	0	0	0	0	0	0	0	0	0
Employment (esb2) {esb1 collapsed} ***	2	2	2	2	2	2	2	3	3
Occupation (osb2) {osb1 collapsed***	1	1	1	1	1	1	1	1	1
Occupation (osb3) {osb2 collapsed} ***	0	0	0	0	0	0	0	0	0

The core variables are marked in bold; other variables have been derived from these and are marked with asterisks.

Some are new variables which indicate a change in marital, occupational or employment status during the year, - these are seen as important in explaining the dynamics of migration - , others are simplified versions of the core variables, formed by collapsing categories.

For a **detailed description of the variables** click here.

Limitations of the data set

- The data is restricted to those residing in the study area in 1986; it includes individuals who had moved to Rochdale before 1986, but not those who had moved away. Therefore those who had left cannot be compared with those remaining.
- The data contains the complete, or nearly complete histories for those aged sixty at the time of interview but only short histories for younger respondents.
- Therefore the data are comparatively sparse on migration behaviour during later career stages and during the more distant past. For earlier periods the maximum age is reduced.
- There is no information on retirement or post-retirement migration.
- As the data were not specifically collected for studying migration, some explanatory variables which may be important, such as family income for instance, were not available.
- The reliability of retrospective data may also be called into question (Dex 1995; Dex and McCulloch 1998).

Do we need such a large and complex longitudinal data set to answer the substantive questions?

We can sum up the number of migrations for each individual and produce a summary data set, with one line of information for each individual. This will give cross-sectional information for the years up to 1985.

What questions can be answered by cross-sectional analysis?

Next:Cross-sectional data

Home page | **Contents** | **Previous**

Cross-sectional data

Summarizing the data

For each individual, we can sum the number of migrations recorded in the survey, to produce one line of information containing:

- Case number
- Number of migrations since leaving school (n)
- Time (t), number of years since leaving school

Only time independent explanatory variables are included in these cross-sectional data.

- Educational qualification (ed), with 5 levels:

  1=Degree or equivalent; professional qualifications with a degree

  2=Education above A-level but below degree level; includes professional qualifications without a degree

  3=A-level or equivalent

  4=Other educational qualification

  5=None

The data matrix for the individual shown on the longitudinal data page can be summarized as follows:

case number	n	t	ed
50016	5	9	4

This person is one of the eight in the data set to have 5 migrations during the time in the survey. See Table 1.
The data sets can be downloaded from here. The cross-sectional data set is available in the file *rochmigx.dat*.

TABLE 1: Observed migration frequencies

Number of moves	0	1	2	3	4	5	>=6
Observed frequency	228	34	42	17	9	8	10

Table 1 summarizes the observed migration frequencies for the 348 respondents in the sample.

As the individuals ranged in age from 20 to 60, they had varying lengths of migration history.

If complete randomness in migration behaviour is assumed, then a Poisson model may be used to represent the aggregate count data.

NEXT:The Poisson model

Cross-sectional analysis:Poisson model for aggregate count data

The Poisson model

If complete randomness in migration behaviour is assumed, then a Poisson model may be used to represent the aggregate count data. Strictly, we should use a Binomial model as each individual is only allowed one migration per year so that the total number of migrations has an upper limit. However, for a large sample and and a low migration rate the Poisson model provides a good approximation.

For a homogeneous population, the probability of obtaining n_i outcomes in time t_i may be written as

$$Pr(n_i)=(m_i)^{n_i} \exp(-m_i) / n_i!$$

where m_i is the mean (or expected) number of migrations in time t_i.

For a constant annual migration rate r,

$$m_i=r*t_i$$

or

$$\log(m_i)=\log(r)+\log(t_i)$$

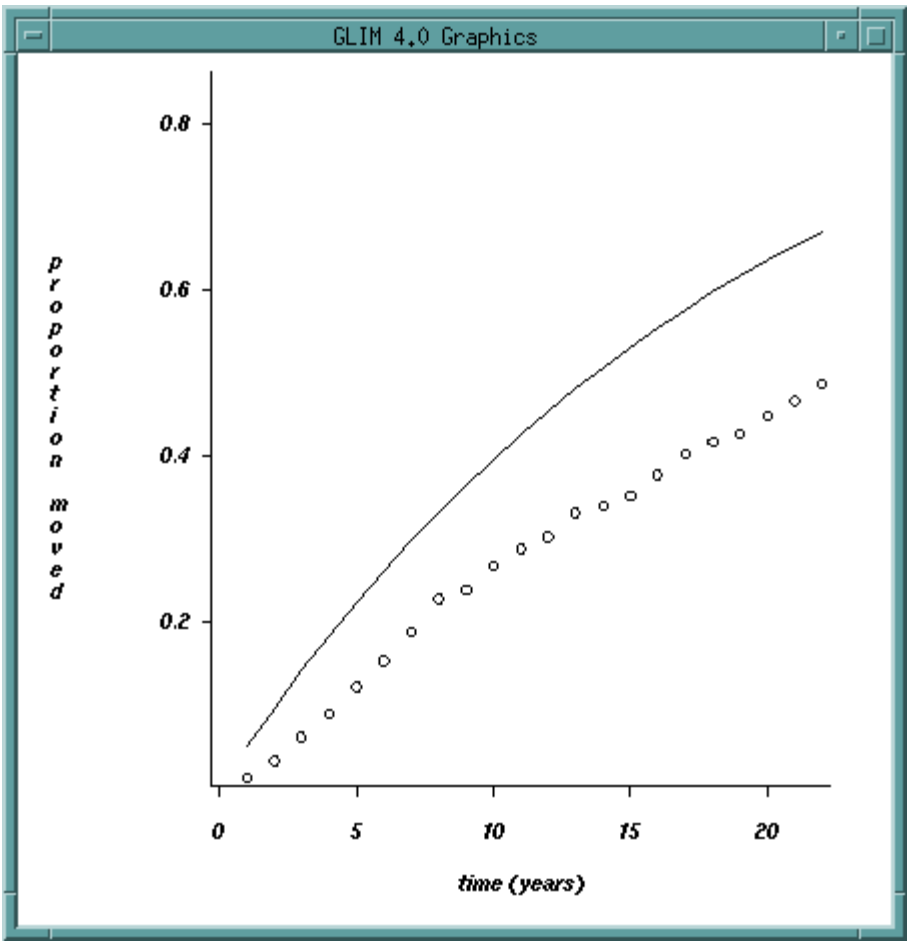
This model is an example of a generalised linear model. We will see how to fit such models a little later. When this model is fitted (using *log(r)* as an *OFFSET* in SABRE), the average annual migration rate comes out as 0.049 moves per individual per year.

For the time being, we note that this figure can also be calculated by simply dividing the total number of moves in the data set by the total time exposure to migration opportunities for the sample. Thus, there are 312 moves and 6349 annual observations, giving an average of 0.049 moves per individual per year.

This implies that each year a proportion of 0.049 of the population (or 4.9%) migrates, and that a proportion of 0.951 (or 95.1%) remains.

Using this model, the projected proportion moving at least once over a period of T years is equal to $[1-(0.951)^T]$.

The projected proportion migrating over different time periods is shown by the line on the graph. It is considerably higher than the observed proportion calculated from the data, which is indicated by circles.



It is evident that this model substantially and systematically overpredicts the proportion moving, and therefore underestimates population stability. This is a consequence of assuming that migration behaviour over one time period can be used to predict migration behaviour over a longer time period, and is an example of a general problem, which Coleman (1973) calls the "deficient diagonal" effect.

The assumption that all individuals have the same propensity to migrate, which is not subject to change over time, does not seem compatible with the migration processes generating the data.

Allowing the migration rate to vary with time

The migration rate can be allowed to vary systematically with time in this simple model by replacing (t_i) in the above equation by (t_i) ^{b_1}. Now the migration rate decreases through migration history if b_1 is less than 1 and increases if b_1 is greater than 1. One reason why we may expect b_1 to be less than 1 is due to *inertia* effects, with people increasingly less likely to move with duration in a specific locality.

It is convenient to write

$$r=\exp(b_0)$$

where b_0 is an unknown constant, and the exponentiation ensures that r is always non-negative.

The mean number of migrations may now be written as:

$$m_i=\exp(b_0)*(t_i)^{b_1}= \exp(b_0+b_1*\log(t_i))$$

or

$$\log(m_i)=b_0+ b_1*\log(t_i)$$

This model is typical of a **generalised linear model**, which contains:

1. a linear regression function or linear predictor in the explanatory variables, $[b_0+b_1*\log(t_i)]$,
2. a transformation, (logarithmic), which relates the linear predictor to the mean m_i,
3. a response variable n_i, which has a Poisson distribution with mean m_i.

The model may be fitted using SABRE software as follows. To run the example interactively, you will need to download the SABRE software and data sets.

SABRE SESSION:INPUT AND OUTPUT

```
C read in variables from data file
data case n t ed
read rochmigx.dat

        348 observations in dataset

C declare response variable
yvar n
C declare model
poisson yes
C calculate log(time)
transform ltime log t
C fit Poisson model with intercept
C and log(time) as explanatory variable
lfit int ltime

    Iteration        Deviance        Reduction
    ____________________________________________
        1           1299.5140
        2           754.34418        545.2
        3           658.72919        95.61
        4           648.79228        9.937
        5           648.49783       0.2945
        6           648.49747       0.3547E-03
        7           648.49747       0.5484E-09

C display parameter estimates
dis est

    Parameter              Estimate         S.
Error

_____________________________________________________
    int                    -3.2884
0.35114
    ltime                   1.0887
0.11119

C display model fitted
dis m

    X-vars      Y-var
    ____________
    int         n
    ltime

    Model type: standard Poisson log-linear

    Number of observations             =    348
```

```
    X-vars df=  2

    Deviance = 648.49747 on 346 residual degrees
of freedom

stop
```

Results and conclusion

1 The estimated coefficient b_1 of *ltime* is 1.0887, with a standard error of 0.1112, and is therefore not significantly different from 1. The migration rate does not appear to decline or increase through migration history, but is constant.

Table 2:Observed and expected frequencies

Number of moves	0	1	2	3	4	5	>=6
Observed frequency	228	34	42	17	9	8	10
Expected frequency	164.3	101.6	50.4	21.1	7.5	2.3	0.80

2 The observed migration frequencies are compared in Table 2 with the values predicted by the Poisson model. The model does not seem to fit the data, with the number of individuals making no moves or making four or more moves substantially underpredicted. There appears to be a systematic variation in migration frequency over and above the variation attributable by chance.

3 The fit of the model may be assessed by comparing the value of the sum of [(Expected frequency-Observed frequency) ^{2}/Expected frequency] with the c ^{2} distribution on 5 degrees of freedom (7 cells - 2 estimated coefficients). The critical value at the 5% significance level is 11.07. The calculated value is in fact 192.5, an order of magnitude higher.

4 The degree of model misspecification may be measured by the dispersion parameter, which is the ratio of the scaled deviance and the residual degrees of freedom.(648.5/346)=1.87). If the model were well specified, this ratio would be approximately 1.

5 One explanation for the poor fit of the model is that the assumption of a *homogeneous* population is not valid. Individuals may vary in their likelihood of migration; the assumption of a migration rate which depends only on time may be incorrect. Thus, it may be possible to improve the model specification by including explanatory variables which distinguish between individuals.

Next:Poisson model with explanatory variable

Home page | **Contents** | **Previous**

Cross-sectional analysis:Poisson model with explanatory variables

Introduction

The Poisson model may be used for inference about explanatory variables even when the model is seriously misspecified, provided that:

1. The explanatory variables do not change over the migration histories.
2. Interest focuses on the relationship between the explanatory variables and the rate of migration.

Education is recognised as the single most important individual-level factor governing rates of internal migration, as it is related to the opportunity to progress in careers. (Sandefur and Scott, 1981; Goss, 1985; Liaw, 1990)

Five levels of educational attainment are available in the data, and may be included in the Poisson model.

The model

The previous equation for the mean number of migrations

$$\log(m_i)=b_0+ b_1*\log(t_i)$$

may be extended by writing:

$$\log(m_i)=b_0+ b_1*\log(t_i)+b_2*x_{i1} +b_3*x_{i2}+b_4*x_{i3}+b_5*x_{i4}+b_6*x_{i5}$$

where x_{ij}=1 if individual *i* has educational qualification *j* and 0 otherwise. These x_{ij} are known as dummy variables. SABRE constructs dummy variables internally for any variable defined as a factor.
Education has 5 levels: *j*=1 is the reference group, with no qualifications. The coefficient estimate for this level is absorbed into the intercept term and b_2 is set to zero by SABRE; the parameter estimates of the higher levels (b_3,b_4,b_5 and b_6) provide appropriate contrasts with this level.

We now add the 5-level factor **educational qualification** to the previous model.
For the lowest level to correspond to 'No qualifications', the educational levels in the data, which are coded 1 for 'Degree or equivalent' and 5 for 'No qualifications', are reversed. This is done by two *transform* commands.

SABRE SESSION:INPUT AND OUTPUT

```
data case n t ed
read rochmigx.dat

        348 observations in dataset
```

```
transform ltime log t
C reverse order of levels for ed in two stages
transform ned ed - 6
transform reved ned * -1
C check reversed levels
look ed reved
```

	ed	reved
1	4.000	2.000
2	4.000	2.000
3	5.000	1.000
4	4.000	2.000
5	3.000	3.000
6	5.000	1.000
7	2.000	4.000
8	4.000	2.000
9	5.000	1.000
10	4.000	2.000
11	3.000	3.000
12	5.000	1.000
13	2.000	4.000
14	3.000	3.000
15	4.000	2.000
16	2.000	4.000
17	5.000	1.000
18	5.000	1.000
19	3.000	3.000
20	3.000	3.000

```
C convert variable reved to factor fed
C and fit previous model
fac reved fed
yvar n
poisson yes

lfit int ltime
```

Iteration	Deviance	Reduction
1	1299.5140	
2	754.34418	545.2
3	658.72919	95.61
4	648.79228	9.937
5	648.49783	0.2945
6	648.49747	0.3547E-03
7	648.49747	0.5484E-09

```
C now add in education
lfit +fed
```

Iteration	Deviance	Reduction
1	1297.1251	
2	748.76297	548.4
3	649.04377	99.72
4	637.92142	11.12
5	637.56670	0.3547
6	637.56619	0.5089E-03
7	637.56619	0.1140E-08

```
dis est
```

Parameter		Estimate	S. Error
int		-3.7435	0.39195
ltime		1.1610	0.11553
fed	(1)	0.	ALIASED [I]
fed	(2)	0.35868	0.13633
fed	(3)	-0.15726E-01	0.24772
fed	(4)	0.49562	0.22760
fed	(5)	0.40762	0.20645

```
dis m

    X-vars      Y-var
    ____________________
    int         n
    ltime
    fed

    Model type: standard Poisson log-linear

    Number of observations              =    348

    X-vars df           =     6

    Deviance           =637.56619 on 342 residual
degrees of freedom
    Deviance decrease =10.931280 on   4 residual
degrees of freedom

stop
```

Results and conclusion

1. The addition of educational qualification to the model has reduced the deviance from 648.49 to 637.56 i.e. by 10.93 on 4 degrees of freedom. This is significant at the 5% level when compared with $c^2_{(4)}$=9.49.
   Thus, the addition of educational qualification appears to produce a modest improvement on the fit of the Poisson model.
2. The estimated coefficient of *ltime* is still close to 1; the migration rate again appears to be constant over time.
3. The coefficient estimate for the reference level of educational attainment shown as fed(1) has been absorbed into the **intercept** term.
   The coefficient estimates of other levels *j* give the difference between the reference level and level *j*. Due to the logarithmic link, the additive effect of b_j on the linear predictor, has a multiplicative effect of $\exp(b_j)$ on mean migration rates. For example fed(2), estimated as 0.35868, produces a multiplicative effect of exp(0.35868)=1.4 on the migration rate. Starting with the highest educational level, the multiplicative effects are as follows:

Education	Multiplicative factor
Degree or equivalent	1.5
Other higher education	1.6
A-level or equivalent	1.0
Other educational qualification	1.4
No qualification	1.0

4. These results do provide some evidence of migration propensity increasing with education, though the standard errors of the coefficient estimates are relatively large and the results are somewhat anomalous.
   This may be a particular feature of this data set, or it is possible that some explanation for the anomalies could be found if more precise categories of educational qualifications were available.
5. It must also be noted that there is no control for other variables which might influence migration behaviour and which may be correlated with the level of education.
6. The dispersion parameter, which is the ratio of the scaled deviance to the residual degrees of freedom = 637.566/342=1.86 has only slightly been reduced.
7. It is clear that adding educational qualification to the model, accounts only in a small way for the differences between individuals.

How can we control for other differences?

Next:A Mixture model for cross-sectional data

Home page | Contents | Previous

Allowing for unmeasured heterogeneity: a mixture model for cross-sectional data

Omitted explanatory variables

Educational qualification accounts only in a small way for the heterogeneity (ie. the variation in migration behaviour) of the population. Other important individual differences have not been measured, or indeed may be unmeasurable.

To model heterogeneity in migration propensity due to unmeasured and unmeasurable factors, we add an individual specific term, or nuisance parameter, e_i to the linear predictor, to represent the omitted explanatory variables. This term is assumed to be constant for each individual over time. The conventional assumption is that e_i is distributed independently of the included variables. The model equation, with 5 levels of educational attainment as before, becomes:

$$\log(m_i)=b_0+ b_1*\log(t_i)+b_2*x_{i1} +b_3*x_{i2}+b_4*x_{i3}+b_5*x_{i4}+b_6*x_{i5}+e_i$$

The mixture model

The term e_i which represents the effect of the omitted variables for each individual *i* is assumed to have some probability distribution over the population. This distribution has to be modelled in addition to the Poisson model for the count data. The model is now said to have a **mixing distribution** ; or alternatively the model is called a **random effects** or a **mixture model**.
Different methods may be used to fit mixture models, depending on the assumptions made about the probability distribution of the error terms. SABRE uses a standard approach (see for example Lancaster and Nickel 1980; Heckman and Singer 1984).

SABRE assumes a Normal distribution for e_i, with mean zero and variance s^2, and uses a Gaussian quadrature method to fit the model. The tails of the Normal distribution cause a problem, as they assume zero probability at the extremes of the distribution. In fact, there is strong evidence that there are individuals for whom, in many situations, there will be a finite probability of never taking part in the process under investigation. These are the "stayers"; in the context of migration, these are the people who are likely never to move (over and above those who, by chance, do not move in the period covered by the study).

SABRE can allow for "stayers" by supplementing the quadrature mass points with endpoints at plus and minus infinity when this is appropriate. In this model, a nuisance parameter value of minus infinity implies zero probability of migration for that individual.

The standard SABRE mixture model is fitted using the *FIT* command, and includes endpoints by default. For the Poisson model, a single endpoint at minus infinity is included, which estimates the proportion of stayers. There is an option to omit the endpoints from the model and to allow the standard Poisson-Normal mixture model to be fitted, by using the *ENDPOINT* command. The parameterisation of the model is given in the SABRE reference guide.

We fit the log-linear Poisson-Normal mixture model for count data, first with endpoints and second without endpoints

as follows:

Model with endpoints

SABRE SESSION:INPUT AND OUTPUT

```
data case n t ed
read rochmigx.dat

        348 observations in dataset

transform ltime log t
C reverse order of levels for ed
transform ned ed - 6
transform reved ned * -1
fac reved fed
poisson y
yvar n
C fit random effects model
C endpoints fitted by default
fit int ltime fed

    Initial Log-Linear Fit:

    Iteration         Deviance         Reduction
    ____________________________________________
        1            1297.1251
        2            748.76297         548.4
        3            649.04377         99.72
        4            637.92142         11.12
        5            637.56670        0.3547
        6            637.56619       0.5089E-03
        7            637.56619       0.1140E-08


    Iteration      Deviance          Step      End-point
Orthogonality
                                    length
criterion

_____________________________________________________________________________
        1          549.93673         1.0000        free
13.255
        2          531.94684         1.0000        free
0.28295E-01
        3          529.77935         0.0156        free
7.2279
        4          522.42322         0.5000        free
24.948
        5          495.13658         1.0000        free
16.832
        6          487.98913         1.0000        free
3.9855
        7          486.09574         1.0000        free
72.511
        8          486.07703         1.0000        free
15.212
        9          486.07703         1.0000        free

dis est

    Parameter               Estimate          S. Error
    ____________________________________________________
    int                      -2.6932          0.57967
    ltime                    0.97307          0.15646
    fed    ( 1)                   0.          ALIASED [I]
    fed    ( 2)              0.44283          0.18502
    fed    ( 3)             -0.34053E-01      0.32219
    fed    ( 4)              0.67497          0.32448
    fed    ( 5)              0.32705          0.27775
    scale                    0.45004          0.13086

PROBABILITY

_____________
```

```
    end-point 0            0.92752          0.19029        0.48120

dis m

    X-vars      Y-var       Case-var
    ____________________________________
    int         n           case
    ltime
    fed

    Model type: standard Poisson log-linear normal mixture with
end-point

    Number of observations              =    348
    Number of cases                     =    348

    X-vars df          =     6
    Scale df           =     1
    End-point df       =     1

    Deviance      =   486.07703   on   340 residual degrees of
freedom

fit -fed

    Iteration      Deviance       Step      End-point
Orthogonality
                               length
criterion

_____________________________________________________________________________
        1          619.14491      1.0000       free
91.345
        2          521.04026      1.0000       free
28.699
        3          497.87490      1.0000       free
23.435
        4          494.73843      1.0000       free
5.2864
        5          494.52771      1.0000       free
8.9229
        6          494.49902      1.0000       free
3.4139
        7          494.49442      1.0000       free
5.9225
        8          494.49442      1.0000       free

dis m

    X-vars      Y-var       Case-var
    ____________________________________
    int         n           case
    ltime

    Model type: standard Poisson log-linear normal mixture with
end-point

    Number of observations              =    348
    Number of cases                     =    348

    X-vars df          =     2
    Scale df           =     1
    End-point df       =     1

    Deviance          =  494.49442 on 344 residual degrees of
freedom
    Deviance increase =  8.4173895 on   4 residual degrees of
freedom
```

Results and conclusion

1. The addition of the individual specific random term and left endpoint to the model has reduced the deviance

from 637.56 to 486.08 ie. by 151.48 on 2 degrees of freedom. Although the c test is not strictly correct, as the standard Poisson model lies on the boundary of the parameter space of the Poisson mixture model, such a large reduction in deviance indicates a significant improvement in model fit. There appears to be considerable residual heterogeneity in the population.

2. The dispersion parameter has decreased to 486.08/340=1.43, confirming the improved fit.
3. The parameter estimates have changed little (by approximately one standard error); the standard errors of the parameter estimates have all increased. This result is typical when comparing models with and without unmeasured heterogeneity, provided all the explanatory variables are exogenous. We leave a discussion of the term *exogenous* until slightly later in this example.
4. Even though the standard Poisson model seems misspecified, the parameter estimates are *consistent*, ie. they *tend to the true values* when the sample size is increased. However, standard errors are underestimated and may lead us to conclude that an explanatory variable is significant, when in fact it is not. For instance, in the standard Poisson model, as the ratio of the parameter estimate to the standard error (t-ratio) for **fed(5)** is at about the 5% significance level of 2 , we might conclude that this factor is significant, whereas in the Poisson mixture model it is well below the 5% significance level, indicating that this factor is in fact not significant.
5. The small increase in deviance (8.42) compared to $c^2_{(4)}$=9.49 at the 5% level, when educational qualification is removed from the model confirms that education is not significant in the Poisson mixture model.
6. The scale parameter estimate is the standard deviation of the Normal distribution assumed for the individual specific terms e_i. It is significantly different from zero and indicates considerable residual heterogeneity.
7. Note the parameter estimate for the left endpoint. The parameter value of 0.9275 (standard error 0.1903) is significantly different from zero, and the associated probability of 0.48 suggests that the sample contains a significant number of "stayers".

Model without endpoints

We now continue the SABRE session, remove endpoints and refit the full model.

SABRE SESSION:CONTINUED

```
C put back fed
fit +fed

    Iteration         Deviance          Step       End-point
Orthogonality
                                      length
criterion

_____________________________________________________________________
         1            668.03445         1.0000        free
29.768
         2            528.74742         1.0000        free
22.714
         3            496.35404         1.0000        free
31.771
         4            487.11433         1.0000        free
3.2170
         5            486.70328         1.0000        free
12.330
         6            486.41714         1.0000        free
8.5039
         7            486.07846         1.0000        free
5.3708
         8            486.07703         1.0000        free
6.6935
         9            486.07703         1.0000        free

dis m

    X-vars       Y-var        Case-var
    ____________________________________
    int          n            case
    ltime
    fed
```

```
    Model type: standard Poisson log-linear normal mixture with end-
point

    Number of observations             =   348
    Number of cases                    =   348

    X-vars df          =     6
    Scale df           =     1
    End-point df       =     1

    Deviance           = 486.07703     on   340 residual degrees of
freedom
    Deviance increase  =  8.4173895     on    4 residual degrees of
freedom
C fit same model without endpoints
endpoint no
fit .

    Iteration       Deviance         Step      End-point
Orthogonality
                                    length                   criterion

_____________________________________________________________________
        1          694.22855        1.0000       fixed         26.564
        2          551.25040        1.0000       fixed         25.027
        3          533.52981        1.0000       fixed         16.291
        4          513.44427        1.0000       fixed         6.3297
        5          511.82728        1.0000       fixed         8.5030
        6          511.28156        1.0000       fixed         14.935
        7          511.01450        1.0000       fixed         6.4983
        8          511.00122        1.0000       fixed         4.4092
        9          511.00114        1.0000       fixed

dis est

    Parameter               Estimate         S. Error
    ____________________________________________________
    int                     -4.5013          0.58650
    ltime                    1.1857          0.17733
    fed    ( 1)                 0.           ALIASED [I]
    fed    ( 2)             0.26548          0.22422
    fed    ( 3)             0.16689          0.35579
    fed    ( 4)             0.51855          0.35699
    fed    ( 5)             0.61071          0.45804
    scale                    1.1940          0.99342E-01

dis m

    X-vars      Y-var       Case-var
    ____________________________________
    int         n           case
    ltime
    fed

    Model type: standard Poisson log-linear normal mixture

    Number of observations             =   348
    Number of cases                    =   348

    X-vars df          =     6
    Scale df           =     1

    Deviance           = 511.00114   on   341 residual degrees of
freedom
    Deviance increase = 24.924106 on    1 residual degrees of freedom
```

Conclusion

When the same model is fitted without endpoints, the deviance *increases* by 24.9 on a change of 1 degree of freedom. Although the c^2 test is again not strictly applicable, such a large change in deviance ($c^2_{(1)}$=3.84 at the 5% level) indicates that unobserved heterogeneity is in excess of that reflected by the Normal distribution. The model fits significantly better when allowance is made for "stayers".

What have we learnt from cross-sectional data analysis?

Next:Conclusions from cross-sectional analysis

Home page | **Contents** | **Previous**

Conclusions from cross-sectional data analysis

Summary

- Extrapolation of mean annual migration rates leads to an underprediction of population stability. This is because in a heterogeneous population, the individuals who are most likely to move, and who contribute to the mean annual migration rate will have moved away, leaving behind those who are less likely to move.

- Cross-sectional analysis of this data set does not indicate any systematic variation of the mean migration rate with time. Even for data sets which showed evidence of temporal variation, there would be no indication of whether this was due to age, cohort or inertial effects.

- Even though the standard Poisson model seems misspecified, because all the explanatory variables are exogenous the parameter estimates are *consistent*, ie. they tend to the true values when sample size is increased. However, standard errors are underestimated and may lead us to conclude that an explanatory variable is significant, when in fact it is not. For instance, results for the standard Poisson model suggest that educational qualifications do affect the likelihood of migration; the Poisson mixture model does not indicate significant educational qualification effects.

- There is evidence that the likelihood of migration varies markedly between individuals and that the sample contains a number of "stayers", individuals likely never to move.

- With a single count of outcomes for each individual, it is *impossible* to distinguish between a heterogeneous population, with some individuals having a consistently high and others a consistently low propensity to migrate, and a truly contagious process, in which an individual's experience of migration per se increases the probability of subsequent migration.

It is clear that the analysis of the cross-sectional data has answered only a few of the **substantive questions** of interest. No light has been shed on the *dynamics* of the migration process.
Longitudinal data analysis of individual event histories is necessary to explore the temporal variation in individual migration rates and to identify, for example, inertial effects.

Next:Introduction to longitudinal data analysis

Home page | **Contents** | **Previous**

Longitudinal data analysis: Introduction

The longitudinal data set

- The response variable is now binary, indicating for each calendar year whether or not there was a migration move. As temporary moves of a few months duration do not imply commitment to an area, they are not considered as migration. Therefore migration events are recorded on an annual basis, with at most one move per year. We do not use annual count data.

- We can now use time-varying explanatory variables. The variables age, calendar year, duration of stay and the presence of children of secondary age in the family are recorded each year, while marital status, employment status and occupational status are recorded at the beginning and end of each year. Other explanatory variables are derived from the raw data; some indicate a change in the status variables during the year, others have been created by collapsing categories of certain variables.

- We look at the marital, employment and occupational status variables both at the beginning and at the end of each year, as it may be either the original status, the destination status or a change in status during the year which influences individual migration.

- It is important to distinguish between two types of explanatory variable: an **endogenous** explanatory variable, which is in some way a function of an earlier outcome of the process under study, and an **exogenous** explanatory variable, in which there is no such relationship.

- In this data set *duration of stay* is an endogenous explanatory variable, because the number of years of residence since the last migration move is related to the timing of that move.

Residual heterogeneity

Longitudinal data consist of repeated observations on each individual. The observations are independent between individuals, but correlated within individuals. The differences between individuals are measured by a range of explanatory variables which may differ over time. In practice not all the variables that characterize individuals are observable, and the omitted variables give rise to a residual heterogeneity.

In the cross-sectional analysis, as all explanatory variables were exogenous, the parameter estimates were consistent even though the standard Poisson model was misspecified. This is not the case for cross-sectional or longitudinal analyses if there are endogenous explanatory variables.

In the presence of endogenous explanatory variables, such as *duration of stay*, inference about temporal variation requires an explicit representation of residual heterogeneity, otherwise parameter estimates will be biased. This is only possible with longitudinal data; the problems posed by endogenous variables cannot be overcome using cross-sectional data.

The model

The response variable y_{it} is binary, defined as 1 if the individual *i* migrates in year *t*, and 0 otherwise. It has a Bernoulli probability distribution with

$$Pr(y_{it}) = p_{it}^{y_{it}} (1-p_{it})^{1-y_{it}}$$

where p_{it} is the probability of a migration move by individual *i* in year *t*. The relation between p_{it} and the explanatory variables is made through a suitable linear predictor and the **logistic link** function. This transforms the linear predictor of explanatory variables, which may have any value between plus and minus infinity, to a probability which necessarily lies between zero and one.

Using the logistic link function $\log[p_{it}/(1-p_{it})]$, the **simple logistic** regression model is:

$$\log[p_{it}/(1-p_{it})] = \underline{b}' \underline{x}_{it}$$

where $\underline{b}' \underline{x}_{it} = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4}+\dots$.
$\underline{b}' \underline{x}_{it}$ is a shorthand (vector) way of denoting the linear predictor, which may contain a large number of explanatory variables.

This can be rewritten as

$$p_{it}=\exp(\underline{b}' \underline{x}_{it})/[1+\exp(\underline{b}' \underline{x}_{it})]$$

and the model including **residual heterogeneity** as

$$p_{it}=\exp(\underline{b}' \underline{x}_{it}+e_i) /[1+\exp(\underline{b}' \underline{x}_{it}+e_i)]$$

where $\underline{x}_{it}$ is a vector of explanatory variables, $\underline{b}'$ is a vector of unknown parameters and e_i is an individual specific term summarizing the effect of the omitted variables.

The large number of possible explanatory variables in the longitudinal data set require a pragmatic approach to model building. We first model the temporal variation.

Next:Longitudinal analysis: Temporal variation

Home page | **Contents** | **Previous**

Longitudinal data analysis: Temporal variation

As a first step we model the temporal variation, and fit models both with and without residual heterogeneity and compare them.

Temporal variation

The dynamic characteristics of the data are described by the three temporal explanatory variables: age, year, and duration of stay. Cohort effects are subsumed in the year and age components. Alternatively, it would be possible to reparameterise the model so that age and cohort rather than age and year effects are estimated. This would not affect the goodness of fit of the model.

- *Year* effects are caused by external economic and social changes generating fluctuations in aggregate migration over time.
- The variation of migration propensity with *age* is related to life cycle factors, such as marriage and children, and to career progression.
- *Duration of stay* is a proxy variable for the many social, community and economic ties which strengthen with length of residence. It is a measure of *cumulative inertia*, which may compound the variation of migration propensity with age. (See Mc Ginnis, 1968; Huff and Clark, 1978.)

What functions of these explanatory variables are appropriate to use in the model?

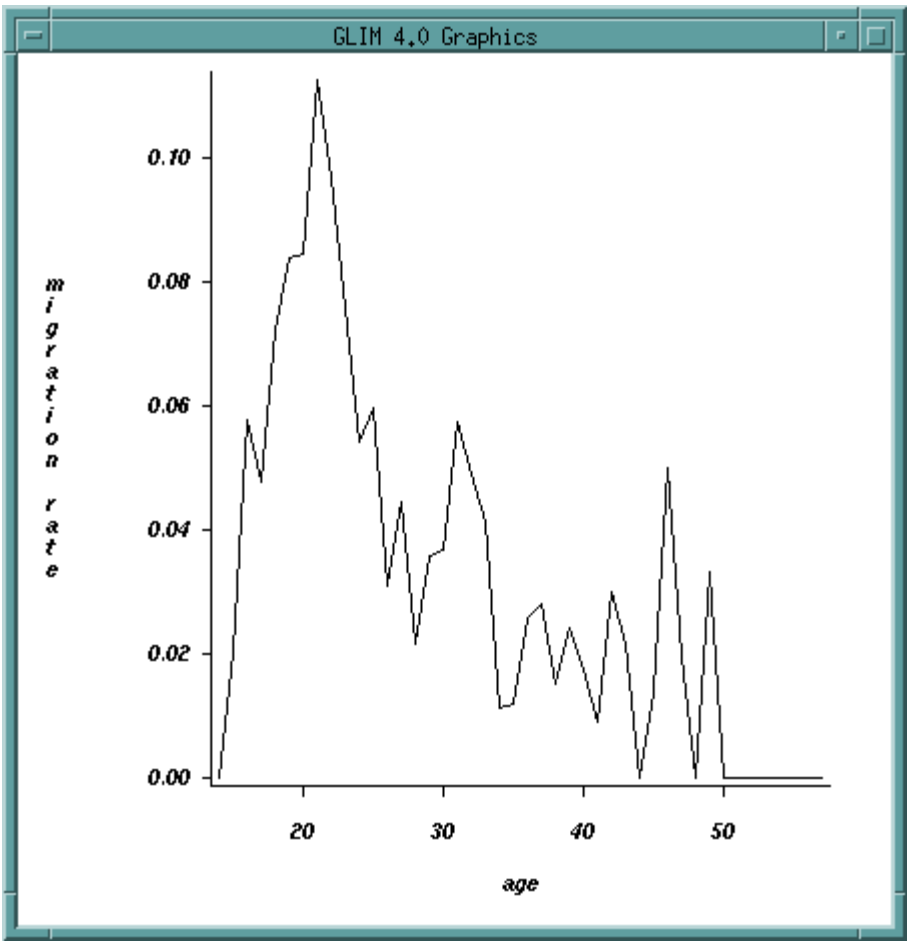
We first explore the data to find a suitable starting point for model building.

The age effect

As a first step, it is helpful to examine how the empirical mean migration rate varies with age. The mean migration rate is calculated by dividing the total number of moves by the total number of years of migration opportunity for each distinct age.

The results on the graph show a clear peak around age 20, some evidence of another peak at about 30 and at least two peaks close to each other just under age 50. The latter peaks could be the result of fluctuations because the data are more sparse here.

It must be noted that there are no controls for other temporal variables in this graph. Nevertheless, there is evidence that the variation with age is multimodal (ie. has several peaks). This suggests using a polynomial representation of age in the models.



Modelling age, year and duration of stay as categorical variables

To explore how the migration rate varies with the three temporal variables, we split each variable into distinct categories, in such a way that we have a reasonable number of data points within each category. Thus the categories usually span five years, but are longer where the data are sparse near the edge of the data window. We fit the logistic model using these categories as levels of factors.

For age we choose cut-off points 20,25,30,35,40 and 45 years, so that the lowest category represents an age of less than 20 and the highest an age greater than 45. The cut-off points for year will be 55,60,65,70,75 and 80 and for duration of stay 5,10,15,20,25 and 30 years.
The model may be fitted using SABRE software as follows:

SABRE SESSION:INPUT AND OUTPUT

```
data case move age year dur ed ch1 ch2 ch3 ch4
msb mse esb ese &
osb ose mbu mrm mfm msb1 epm eoj esb1 ops osb1
msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
C convert variables to factors using the following
C cut-off points
factor age agegp 20 25 30 35 40 45
factor dur durgp 5 10 15 20 25 30
factor year yeargp 55 60 65 70 75 80
lfit int agegp  yeargp durgp

    Iteration        Deviance        Reduction
```

```
         1           8801.5829
         2           2968.3684        5833.
         3           2335.8507        632.5
         4           2208.6718        127.2
         5           2187.8156        20.86
         6           2185.1153        2.700
         7           2184.8380       0.2772
         8           2184.8279       0.1014E-01
         9           2184.8278       0.2246E-04

dis est

    Parameter               Estimate         S.
Error

------------------------------------------------------------
    int                     -2.1704
0.23184
    agegp ( 1)                   0.
ALIASED [I]
    agegp ( 2)               1.1042
0.16933
    agegp ( 3)              0.73531
0.21522
    agegp ( 4)               1.2723
0.23824
    agegp ( 5)               1.0235
0.32081
    agegp ( 6)               1.0312
0.42478
    agegp ( 7)               1.5378
0.51473
    yeargp( 1)                   0.
ALIASED [I]
    yeargp( 2)           -0.37839E-01
0.27795
    yeargp( 3)              -0.50404
0.28618
    yeargp( 4)              -0.74076
0.28944
    yeargp( 5)              -0.47078
0.27593
    yeargp( 6)              -0.86073
0.28758
    yeargp( 7)               -1.1719
0.28593
    durgp ( 1)                   0.
ALIASED [I]
    durgp ( 2)               -1.4236
0.15918
    durgp ( 3)               -1.9089
0.25098
    durgp ( 4)               -2.6716
0.38781
    durgp ( 5)               -4.1664
1.0210
    durgp ( 6)               -2.9408
0.77358
    durgp ( 7)               -3.0448
1.1063

stop
```

Results and conclusion

1 The parameter estimate of the intercept term refers to the lowest category of each categorical variable; the estimates for the higher levels give the contrasts between those categories and this reference level. The estimates for level 1 of each variable are therefore set to zero (and are said to be aliased).

2 Examination of the parameter estimates gives an indication of how the migration rate varies from category to category, when all three temporal variables are controlled for. For clarity the results are displayed on graphs.

3 The parameter estimates for **age** go up and down, rising three times as we go from category 1 to category 7 (Figure 1). This suggests including age in the model as a sixth order polynomial. We note that the age effect is likely to be better estimated at the lower ages than at the higher ages, because the data are sparse for the older age group.

4 For **year** there is a downward trend in parameter estimates, but with a small increase at category five (Figure 2). This may be a consequence of sparsity of data or it may show a real trend for these years. To allow for this rise and fall, we shall include year as a third order polynomial.

5 As **duration of stay** is increased, there is a general downward trend in parameter estimates, however the trend is not quite linear (Figure 3). The fluctuations at durations above 25 years may be due to sparsity of data. Plotting the parameter estimates against log duration (Figure 4) gives a more linear plot. This suggests trying this variable as either a linear or a logarithmic function.

6 From the parameter estimates we can calculate how the probability of migration varies with each of the explanatory variables for fixed values of the other two variables. Figure 5 illustrates the variation of the probability of migration with age in 1985 with duration of stay set to 10 years. Similar graphs may be plotted for the other variables.

Therefore the starting point for model building will be the following model:

$age+age^2+age^3+age^4+age^5+age^6 +year+year^2+year^3+dur$ [or alternatively + log(dur)].

Next:Model development: A parsimonious main effects model for temporal data

Home page | **Contents** | **Previous**

Longitudinal data analysis: A parsimonious main effects model for temporal data

Model building strategy

In the first instance, we aim to find a **parsimonious** main effects model for the temporal variables. Using the results of our initial exploratory analysis we start by fitting the simple logistic model and comparing the fits of the following linear predictors:

$$age+age^2+age^3+age^4+age^5+age^6+year+year^2+year^3+dur$$

and

$$age+age^2+age^3+age^4+age^5+age^6+year+year^2+year^3+\log(dur)$$

We choose the better fitting model, and then fit a series of simple logistic models using a backward elimination technique. At each step we test if the removal of the least significant explanatory variable (lowest t-ratio) gives a significant deterioration in the model fit. If the removal of an explanatory variable results in an increase in deviance of less than 3.84 ie. c $^2_{(1)}$ at the 5% level, we exclude it from the model; otherwise it is retained.

Sabre analysis

SABRE SESSION:INPUT AND OUTPUT

```
data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
osb ose mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
transform age2 age * age
transform age3 age2 * age
transform age4 age3 * age
transform age5 age4 * age
transform age6 age5 * age
transform ldur log dur
transform year2 year * year
transform year3 year2 * year
lfit int dur year year2 year3 age age2 age3 age4 age5 age6

    Iteration         Deviance         Reduction
    ____________________________________________
        1            8801.5829
        2            2993.0684         5809.
        3            2373.6995         619.4
        4            2231.7859         141.9
        5            2195.4927         36.29
        6            2190.2053         5.287
        7            2190.0373        0.1680
        8            2190.0367       0.6502E-03
        9            2190.0367       0.1007E-06
```

```
dis est

    Parameter               Estimate         S. Error
    ____________________________________________________
    int                      -62.752           32.990
    dur                     -0.20904         0.17902E-01
    year                    -0.53834         0.94139
    year2                    0.71197E-02     0.14008E-01
    year3                   -0.33336E-04     0.68744E-04
    age                       11.740          4.8338
    age2                    -0.70751         0.33134
    age3                     0.20615E-01     0.10996E-01
    age4                    -0.29015E-03     0.17681E-03
    age5                     0.15811E-05     0.11036E-05
    age6                          0.         ALIASED [E]

C Extrinsic aliasing has occurred for age6.
C Fitting high order polynomials can often cause numerical problems.
C An option is to lower the tolerance for aliasing from the default
value.
C As the parameter estimates for the higher order terms are very small
C We choose to transform 'age' to 'trage'=(age-30)/10, roughly
C standardising this variable.
C This is done in two stages.
transform tempage age - 30
transform trage tempage / 10
transform trage2 trage * trage
transform trage3 trage2 * trage
transform trage4 trage3 * trage
transform trage5 trage4 * trage
transform trage6 trage5 * trage
lfit int dur year year2 year3 trage trage2 trage3 trage4 trage5 trage6

    Iteration        Deviance        Reduction
    ____________________________________________
        1           8801.5829
        2           2992.7095         5809.
        3           2373.0803         619.6
        4           2230.9609         142.1
        5           2193.9097         37.05
        6           2187.7970         6.113
        7           2187.2527        0.5443
        8           2187.2013       0.5138E-01
        9           2187.2004       0.8804E-03
       10           2187.2004       0.3062E-06


dis est

    Parameter               Estimate         S. Error
    ____________________________________________________
    int                       12.878           20.980
    dur                     -0.20936         0.17929E-01
    year                    -0.53826         0.94677
    year2                    0.70876E-02     0.14085E-01
    year3                   -0.33068E-04     0.69111E-04
    trage                    0.36390         0.32000
    trage2                  -0.31495E-02     0.58966
    trage3                  -0.56019         0.51877
    trage4                   0.28100         0.54056
    trage5                   0.43264         0.20575
    trage6                  -0.22748         0.14640

C now try log(duration) instead of duration
lfit int ldur year year2 year3 trage trage2
trage3 trage4 trage5 trage6

    Iteration        Deviance        Reduction
    ____________________________________________
        1           8801.5829
        2           2959.3492         5842.
        3           2315.9106         643.4
        4           2186.2580         129.7
        5           2169.6448         16.61
        6           2168.1606         1.484
        7           2167.8240        0.3366
        8           2167.7919       0.3208E-01
        9           2167.7916       0.3470E-03
       10           2167.7916       0.4665E-07

dis est
```

```
    Parameter                Estimate          S. Error
    ____________________________________________________
    int                        12.117            21.298
    ldur                      -1.0483          0.72564E-01
    year                     -0.49783           0.96044
    year2                     0.65403E-02      0.14278E-01
    year3                    -0.30640E-04      0.70011E-04
    trage                      0.23216           0.32332
    trage2                    -0.11755           0.59711
    trage3                    -0.80204           0.52563
    trage4                     0.38544           0.55272
    trage5                     0.58007           0.20935
    trage6                    -0.29310           0.15118

C the model fits better with ldur
C start backward elimination using this model
C remove the highest polynomial term for year
lfit -year3

    Iteration         Deviance         Reduction
    ___________________________________________
        1            8801.5829
        2            2959.3891         5842.
        3            2315.9678         643.4
        4            2186.3817         129.6
        5            2169.8304         16.55
        6            2168.3512         1.479
        7            2168.0149        0.3363
        8            2167.9828       0.3205E-01
        9            2167.9825       0.3473E-03
       10            2167.9825       0.4688E-07

dis est

    Parameter                Estimate          S. Error
    ____________________________________________________
    int                        2.8950            3.3215
    ldur                      -1.0489          0.72558E-01
    year                     -0.79215E-01      0.96845E-01
    year2                     0.29616E-03      0.70291E-03
    trage                      0.24580           0.32189
    trage2                    -0.12526           0.59701
    trage3                    -0.80970           0.52543
    trage4                     0.38969           0.55254
    trage5                     0.57874           0.20935
    trage6                    -0.29289           0.15113
lfit -year2

    Iteration         Deviance         Reduction
    ___________________________________________
        1            8801.5829
        2            2960.7613         5841.
        3            2317.1450         643.6
        4            2186.5548         130.6
        5            2170.0008         16.55
        6            2168.5289         1.472
        7            2168.1916        0.3373
        8            2168.1594       0.3224E-01
        9            2168.1590       0.3511E-03
       10            2168.1590       0.4787E-07
C the increase in deviance on removing year2 and year3
C is not significant at the 5% level
dis est

    Parameter                Estimate          S. Error
    ____________________________________________________
    int                        1.5139           0.53900
    ldur                      -1.0488          0.72558E-01
    year                     -0.38518E-01      0.70233E-02
    trage                      0.24860           0.32199
    trage2                    -0.10853           0.59570
    trage3                    -0.81168           0.52582
    trage4                     0.38768           0.55271
    trage5                     0.57919           0.20955
    trage6                    -0.29282           0.15125

C remove the highest polynomial term for age
lfit -trage6

    Iteration         Deviance         Reduction
```

```
        1            8801.5829
        2            2961.4528         5840.
        3            2318.8479         642.6
        4            2189.1451         129.7
        5            2173.5159         15.63
        6            2172.9519        0.5640
        7            2172.9473       0.4616E-02
        8            2172.9473       0.7230E-05

dis est

    Parameter              Estimate          S. Error
    ------------------------------------------------------------
    int                     1.2943           0.53047
    ldur                   -1.0417           0.72482E-01
    year                  -0.37779E-01       0.70270E-02
    trage                 -0.46674E-01       0.26454
    trage2                 0.89932           0.31357
    trage3                 0.23829E-01       0.30000
    trage4                -0.64032           0.15238
    trage5                 0.19928           0.90486E-01

C removing trage6 has produced an increase in deviance significant at
C the 5% level. Therefore keep all terms of sixth order polynomial


lfit +trage6

    Iteration         Deviance         Reduction
    ------------------------------------------------------------
        1            8801.5829
        2            2960.7613         5841.
        3            2317.1450         643.6
        4            2186.5548         130.6
        5            2170.0008         16.55
        6            2168.5289         1.472
        7            2168.1916        0.3373
        8            2168.1594       0.3224E-01
        9            2168.1590       0.3511E-03
       10            2168.1590       0.4787E-07

C test year
lfit -year

    Iteration         Deviance         Reduction
    ------------------------------------------------------------
        1            8801.5829
        2            2971.7962         5830.
        3            2340.6810         631.1
        4            2216.2755         124.4
        5            2200.6027         15.67
        6            2199.2849         1.318
        7            2199.0021        0.2828
        8            2198.9772       0.2493E-01
        9            2198.9770       0.2284E-03
       10            2198.9770       0.2177E-07
C significant change in deviance

lfit +year
    Iteration         Deviance         Reduction
    ------------------------------------------------------------
        1            8801.5829
        2            2960.7613         5841.
        3            2317.1450         643.6
        4            2186.5548         130.6
        5            2170.0008         16.55
        6            2168.5289         1.472
        7            2168.1916        0.3373
        8            2168.1594       0.3224E-01
        9            2168.1590       0.3511E-03
       10            2168.1590       0.4787E-07

C test log(duration)
lfit -ldur

    Iteration         Deviance         Reduction
    ------------------------------------------------------------
        1            8801.5829
        2            3024.8900         5777.
        3            2455.8074         569.1
```

```
         4            2369.9867         85.82
         5            2362.7628         7.224
         6            2362.0790        0.6839
         7            2361.9175        0.1615
         8            2361.9060        0.1150E-01
         9            2361.9059        0.6667E-04
C significant change in deviance
lfit +ldur

    Iteration         Deviance         Reduction
    ____________________________________________
        1            8801.5829
        2            2960.7613         5841.
        3            2317.1450         643.6
        4            2186.5548         130.6
        5            2170.0008         16.55
        6            2168.5289         1.472
        7            2168.1916        0.3373
        8            2168.1594        0.3224E-01
        9            2168.1590        0.3511E-03
       10            2168.1590        0.4787E-07

C final model
dis est

    Parameter                Estimate          S. Error
    ____________________________________________________
    int                       1.5139           0.53900
    trage                    0.24860           0.32199
    trage2                  -0.10853           0.59570
    trage3                  -0.81168           0.52582
    trage4                   0.38768           0.55271
    trage5                   0.57919           0.20955
    trage6                  -0.29282           0.15125
    year                 -0.38518E-01       0.70233E-02
    ldur                     -1.0488        0.72558E-01
stop
```

Results and conclusions

- The first two models fitted compare the effects of duration and log(duration) in the full model. The model with log(duration) gives a much better fit with a reduction of deviance of almost 20; this function of duration is kept in the model.

- During the process of backward elimination the second and third order terms of year have been removed from the model. The sixth order term of age is statistically significant at the 5% level; therefore this and all the lower order terms are retained in this hierarchical model. Both year and log(duration) are highly significant and are retained.

- The parameters for this parsimonious model are as follows:

Variable	Estimate	Standard Error
constant	1.5139	0.53900
ldur	-1.0488	0.72557E-01
year	-0.38518E-01	0.70233E-02
trage	0.24860	0.32199
trage**2	-0.10853	0.59570
trage**3	-0.81168	0.52582
trage**4	0.38768	0.55271
trage**5	0.57919	0.20955
trage**6	-0.29282	0.15125

- It is noted that the c^2 test used to compare the deviance of nested models is not very powerful with highly correlated explanatory variables, such as powers of age. It may be possible to improve on the above parsimonious model with more powerful tests for individual effects, but that is beyond the scope of the present analysis.

- The negative coefficient estimate for *ldur* indicates that the probability of migration decreases with duration of stay. This may be due to cumulative inertia effects due to a strengthening of community ties with increasing length of residence. Alternatively, it may be due to residual heterogeneity; with increasing duration, the individuals most likely to migrate will be more and more underrepresented.

- The probability of migration predicted by this parsimonious model may be plotted on graphs. In plotting these figures the year is taken as 1985, the individual to be aged 40 and the duration of residence to be 10 years, as appropriate. This is necessary because the precise relationship between an explanatory variable and the response variable depends on the values of the other explanatory variables.
As there are no interaction terms in the model, the patterns shown on the graphs are generally valid.

- The probability of migration plotted against age shows peaks just above age 20, around 35 and the largest near age 50. As the data are sparse for the older age group, the size and location of the third peak must be interpreted with caution,

- The plot against duration of stay shows the expected decrease in the probability of migration with increasing length of residence. The plot against year also shows a decreasing probability of migration with time over the years 1965 to 1985.

Next:Model development: Random effects model for temporal data

Home page | **Contents** | **Previous**

Longitudinal data analysis: A random effects model for temporal data

Model fitting

We compare the fit of the parsimonious simple logistic regression model with the same model with random effects to allow for residual heterogeneity.
For binary data, SABRE fits endpoints at plus and minus infinity by default.

SABRE SESSION:INPUT AND OUTPUT

```
data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
osb ose mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
C transform age as before
transform tempage age - 30
transform trage tempage / 10
transform trage2 trage * trage
transform trage3 trage2 * trage
transform trage4 trage3 * trage
transform trage5 trage4 * trage
transform trage6 trage5 * trage
transform ldur log dur
C first fit the simple logistic model

lfit int ldur year trage trage2 trage3 trage4 trage5 trage6

    Iteration         Deviance         Reduction
    ___________________________________________
        1            8801.5829
        2            2960.7613         5841.
        3            2317.1450         643.6
        4            2186.5548         130.6
        5            2170.0008         16.55
        6            2168.5289         1.472
        7            2168.1916        0.3373
        8            2168.1594       0.3224E-01
        9            2168.1590       0.3511E-03
       10            2168.1590       0.4787E-07

dis est

    Parameter               Estimate          S. Error
    ____________________________________________________
    int                       1.5139           0.53900
    ldur                     -1.0488           0.72558E-01
    year                    -0.38518E-01       0.70233E-02
    trage                     0.24860          0.32199
    trage2                   -0.10853          0.59570
    trage3                   -0.81168          0.52582
    trage4                    0.38768          0.55271
    trage5                    0.57919          0.20955
    trage6                   -0.29282          0.15125

C fit the same model with random effects
C endpoints are fitted by default
fit .

    Iteration         Deviance          Step       End-points
Orthogonality
```

```
                                        length    0          1
criterion

_____________________________________________________________________________
         1             2198.4881         1.0000    free    free
4.6471
         2             2198.2943         0.2500    free    free
3.1137
         3             2185.4266         0.3033    free    free
13.365
         4             2174.8955         0.1175    free    fixed
9.2150
         5             2142.0094         1.0000    free    free
10.360
         6             2135.1201         1.0000    free    free
3.7965
         7             2133.8038         1.0000    free    free
11.834
         8             2133.7948         1.0000    free    free
37.114
         9             2133.7948         1.0000    free    free

dis est

     Parameter               Estimate          S. Error
     _____________________________________________________
     int                      0.83341          0.77050
     ldur                    -0.65918          0.10463
     year                    -0.36521E-01      0.10873E-01
     trage                   -0.69598E-01      0.34063
     trage2                   0.76814E-01      0.59487
     trage3                  -0.82208          0.53734
     trage4                   0.33146          0.54900
     trage5                   0.56760          0.21311
     trage6                  -0.27657          0.15032
     scale                    0.47710          0.17447

PROBABILITY

_____________
     end-point 0              0.56682          0.19724          0.36113
     end-point 1              0.27460E-02      0.46361E-02      0.17495E-
02

stop
```

Results and conclusion

- The deviance has decreased from 2168.16 to 2133.79. This is a reduction of over 34 on 3 degrees of freedom, on adding the individual specific random term to the model. The extra three degrees of freedom are given by the scale of the Normal mixing distribution and the two estimated probabilities of the endpoints. Although the c^2 test is not strictly correct as the simple logistic model lies on the edge of the parameter space of the mixture model, such a large change in deviance ($c^2_{(3)}$=7.81) demonstrates that there is considerable unobserved heterogeneity in the population.

- The coefficient estimate of *ldur* is still negative, but is considerably smaller in magnitude than in the simple logistic model. The estimate of this endogenous explanatory variable has changed by allowing for residual heterogeneity; the estimates of the other variables have changed little (by less than one standard error), and their standard errors are almost unchanged.

- The coefficient of *ldur* measures cumulative inertia effects, and its value confirms that there is an increasing disinclination to move with increasing length of residence. However the effect is smaller than suggested by the simple logistic model; that estimate was inflated because no account was taken of the fact that with increasing duration the individuals most likely to migrate are more and more underrepresented in the population. Inference about duration effects can be misleading unless there is control for omitted variables. (Lancaster 1979; Heckman and Singer 1985)

- The probability of 0.36 associated with the left endpoint gives a measure of the proportion of "stayers" in the

population, i.e. those individuals never likely to migrate. Examination of the parameter estimate and standard error of the right endpoint (and corresponding probability of 0.0017) suggests that this parameter (which estimates the proportion of the population migrating every year) could be set to zero.

- The scale parameter estimate is the standard deviation of the Normal distribution assumed for the individual specific terms.

- The probability of migration predicted by this random effects model may be plotted on graphs to aid interpretation of the parameter estimates. As before, the year is taken as 1985, the individual to be aged 40, and the duration of residence to be 10 years, as appropriate. As no interaction terms have been considered, the trends shown on the graphs are generally valid.

- In calculating the probabilities, the individual specific term is given the estimated population median value, taking into account both the Normal distribution and the proportion of stayers.

- The plot against age now shows two clear peaks at just above age 20 and just below age 50. The relative size of the peaks has changed compared to the simple logistic model; the size and location of the peak near age 50 has again to be interpreted with caution as the data are sparse for this age group. The dominance of the first peak in the random effects model is more plausible substantively as this is the age at which geographical ties are at their minimum.

- The graph against duration of stay shows the decline in migration probability with duration for both the simple logistic and the random effects models. When unobserved heterogeneity is taken into account, the estimated decline is due to cumulative inertia effects; in the simple logistic model the estimate is inflated as discussed above.

- The shapes of the graphs of migration probability against year are the same for both models.

- The levels of probability estimated by the two models are not strictly comparable, as the simple logistic model gives the population average value for individuals with given values of the explanatory variables (age, year, duration of stay), whereas the random effects graphs show the probability values for individuals with the median value of the nuisance parameter.

Can we explain the pattern of migration with age by adding explanatory variables which measure life cycle factors, such as marriage, occupation and employment status and the presence of children in the family?

Next:Model development: Adding explanatory variables

Home page | **Contents** | **Previous**

Longitudinal data analysis: Adding explanatory variables

The variation of migration propensity with age has been linked to life cycle factors, such as marriage, employment, career moves, and the presence of children in the family. Similarly year effects can be linked to economic factors, and employment and career moves are seen to represent underlying economic health. Do explanatory variables which measure these effects explain the variation of migration behaviour with age and year?

The large number of possible explanatory variables require a pragmatic strategy to model building.

Model development

- We start with the parsimonious main effects model for the temporal variables,

$$age+age^2+age^3+age^4+age^5+age^6+year+log(dur)$$

and add explanatory variables which measure individual life cycle effects.

- We choose explanatory variables suggested by substantive considerations to include in our model. A number of such explanatory variables are present in the data set, giving information on education, occupation, marital status, employment, the presence of children of different ages, etc.

- Although empirical evidence is mixed, **education** is often considered to increase the propensity to migrate, because it increases employment opportunities and gives access to better information about other areas. (Sandefur and Scott 1981, Goss 1985, Liaw 1990)

- **Marital status** is an important feature of theories about migration behaviour, with evidence that married individuals are less likely to migrate. Getting married, marital break up and remarriage are expected to increase the probability of migration. (Devis 1983, Grundy 1989)

- School age **children** create important ties to an area, and the fear of disrupting children's education may inhibit migration. (Long 1972, Davies and Flowerdew 1992)

- **Employment** and **occupational** status variables also important in relation to migration (Warnes 1983, Greenwood 1985, Davies and Flowerdew 1992, Ellis et al. 1993, Herzog 1993).

- Career progression is another important variable to affect migration (Salt 1990). We consider three variables measuring *changes* in employment or occupational status which, being "favourable to socio-economic achievement" (Cote 1997) might encourage migration: obtaining a job, promotion to manager and promotion to service class.

- We fit a series of logistic models and use backward elimination to assess which explanatory variables to retain. As the parameter estimates, apart from that of the endogenous variable *ldur*, are very similar for the simple logistic and random effects models, and as the latter is much more computer intensive, we use the simple logistic model for model development.

- We start with the model for the temporal variables, and add education (*ed*), occupational status (*osb3*),

employment status (*esb2*) and marital status (*msb*), each measured at the beginning of the year, first marriage (*mfm*), marital break-up (*mbu*), remarriage (*mrm*), the presence of children of different ages (*ch1, ch2, ch3, ch4*), obtaining a job (*eoj*), promotion to manager (*epm*) and promotion to service class (*ops*).

- For education and marital status we use the original 5 level variables to include in the model; for employment and occupational status we have chosen for simplicity the collapsed variables *esb2* and *osb3* with 3 and 2 levels respectively, instead of the original 8 and 12 levels. The other variables are all 2-level factors.

- We note that some levels of the original employment and occupational status variables are likely to be highly correlated (eg. employment status: none, occupational status: none), and problems with aliasing are likely to occur in models which include such variables. Cross tabulation of the levels of these variables will help to identify possible problems, but that is beyond the scope of the present example.

- We use a cut-off significance level of 0.1 rather than the conventional 0.05. This is very conservative, as the simple logistic model tends to overestimate significance, as we noted earlier. However, as the model may be misspecified due to our pragmatic approach, conservatism is considered important to reduce the chance of rejecting a possibly relevant explanatory variable.

- At each step in the backward elimination we test if the removal of the explanatory variable with the lowest t-ratio (ratio of a parameter to its standard error) gives a significant deterioration in model fit by comparing the change in deviance with the appropriate value of c^2.

At the 0.1 significance level the critical values of the chi-squared distribution for various degrees of freedom are $c^2_{(1)}=2.71$, $c^2_{(2)}=4.61$, $c^2_{(3)}=6.25$, $c^2_{(4)}=7.78$.

- When the preferred main effects model is found, the same model is refitted with random effects to allow for unobserved heterogeneity.

Next:The SABRE analysis

Home page | **Contents** | **Previous**

Adding explanatory variables: the SABRE analysis

We carry out the backward elimination as follows:

SABRE SESSION:INPUT AND OUTPUT

```
C input data and transform variables

data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
ocb oce mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
transform tempage age - 30
transform trage tempage / 10
transform trage2 trage * trage
transform trage3 trage2 * trage
transform trage4 trage3 * trage
transform trage5 trage4 * trage
transform trage6 trage5 * trage
transform ldur log dur
C convert explanatory variables to factors
factor ed fed
factor ch1 fch1
factor ch2 fch2
factor ch3 fch3
factor ch4 fch4
factor msb fmsb
factor msb1 fmsb1
factor msb2 fmsb2
factor mbu fmbu
factor mrm fmrm
factor mfm fmfm
factor eoj feoj
factor ops fops
factor epm fepm
factor esb2 fesb2
factor osb3 fosb3
C fit full model

lfit int ldur year trage trage2 trage3 trage4 trage5 trage6 &
fed fmbu fmfm fmrm fmsb fch1 fch2 fch3 fch4 fesb2 fosb3 fepm fops feoj

    Iteration         Deviance         Reduction
    ---------------------------------------------
        1            8801.5829
        2            2932.0559         5870.
        3            2260.2230         671.8
        4            2115.9063         144.3
        5            2095.5823         20.32
        6            2093.5142         2.068
        7            2093.1257        0.3885
        8            2093.0786       0.4706E-01
        9            2093.0765       0.2120E-02
       10            2093.0760       0.5102E-03
       11            2093.0758       0.1876E-03

dis est

    Parameter               Estimate          S. Error
    ---------------------------------------------------
    int                       1.3741           0.74144
    ldur                     -0.97671         0.75658E-01
    year                    -0.42966E-01      0.77703E-02
    trage                    0.48422          0.34821
    trage2                  -0.81192E-01      0.63693
    trage3                  -0.58212          0.53301
    trage4                   0.30160          0.57210
    trage5                   0.42878          0.20849
```

```
    trage6                -0.23204          0.15366
    fed    ( 1)                0.           ALIASED [I]
    fed    ( 2)           -0.29439E-01      0.29414
    fed    ( 3)           -0.42630          0.31085
    fed    ( 4)            0.19577E-01      0.21836
    fed    ( 5)           -0.25889          0.23502
    fmbu   ( 1)                0.           ALIASED [I]
    fmbu   ( 2)             1.2831          0.64008
    fmfm   ( 1)                0.           ALIASED [I]
    fmfm   ( 2)            0.46489          0.24075
    fmrm   ( 1)                0.           ALIASED [I]
    fmrm   ( 2)            0.97834          0.80128
    fmsb   ( 1)                0.           ALIASED [I]
    fmsb   ( 2)           -0.44557          0.19011
    fmsb   ( 3)           -0.26831          0.49968
    fmsb   ( 4)            0.78074          0.56836
    fmsb   ( 5)            -7.9406           82.102
    fch1   ( 1)                0.           ALIASED [I]
    fch1   ( 2)           -0.76060E-01      0.38951
    fch2   ( 1)                0.           ALIASED [I]
    fch2   ( 2)           -0.68220E-01      0.44099
    fch3   ( 1)                0.           ALIASED [I]
    fch3   ( 2)            -1.2554          0.75279
    fch4   ( 1)                0.           ALIASED [I]
    fch4   ( 2)            0.23823E-01      0.58680
    fesb2  ( 1)                0.           ALIASED [I]
    fesb2  ( 2)            0.52758          0.32558
    fesb2  ( 3)            0.90635          0.44986
    fosb3  ( 1)                0.           ALIASED [I]
    fosb3  ( 2)            0.83994          0.16945
    fepm   ( 1)                0.           ALIASED [I]
    fepm   ( 2)           -0.22312          0.50383
    fops   ( 1)                0.           ALIASED [I]
    fops   ( 2)             1.1732          0.36420
    feoj   ( 1)                0.           ALIASED [I]
    feoj   ( 2)            0.51723          0.43284

C note that the lowest level of each factor is set to zero
C fch1, fch2 and fch4 have very low t-ratios
C remove fch4 first, as this has lowest t-ratio

C To save space we use the MONITOR NO command to produce
C summary information only on the progress of the fitting algorithm

monitor no

lfit -fch4

    Deviance =     2093.0774     at iteration    11

lfit -fch1

    Deviance =     2093.1162     at iteration    11

lfit -fch2

    Deviance =     2093.1470     at iteration    11

lfit -fepm

    Deviance =     2093.3436     at iteration    11

lfit -feoj

    Deviance =     2094.8028     at iteration    11

C the changes in deviance above are not significant at the 10% level
C compared with 2.71, ie. chi-sq. for 1 degree of freedom
C for fed some levels appear more significant than others; test fed.

lfit -fed

    Deviance =     2100.8431     at iteration    11

C change in deviance of 6.04 is not significant at the 10% level
C compared with 7.78, ie. chi-sq. for 4 degrees of freedom
C fed can also be removed from the model

dis est

    Parameter               Estimate         S. Error
```

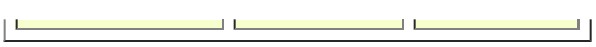
```
    ____________________________________________________________
    int                        1.0028          0.70183
    ldur                      -0.99577         0.74243E-01
    year                      -0.39207E-01     0.74412E-02
    trage                      0.43563         0.33628
    trage2                    -0.10207         0.62253
    trage3                    -0.54183         0.52659
    trage4                     0.29816         0.56929
    trage5                     0.41445         0.20703
    trage6                    -0.22861         0.15373
    fmbu  ( 1)                       0.        ALIASED [I]
    fmbu  ( 2)                 1.2363          0.64637
    fmfm  ( 1)                       0.        ALIASED [I]
    fmfm  ( 2)                 0.46619         0.24024
    fmrm  ( 1)                       0.        ALIASED [I]
    fmrm  ( 2)                 1.0371          0.79233
    fmsb  ( 1)                       0.        ALIASED [I]
    fmsb  ( 2)                -0.44911         0.18890
    fmsb  ( 3)                -0.19336         0.49091
    fmsb  ( 4)                 0.71328         0.55703
    fmsb  ( 5)                 -7.8189          82.104
    fch3  ( 1)                       0.        ALIASED [I]
    fch3  ( 2)                 -1.2803         0.75074
    fesb2 ( 1)                       0.        ALIASED [I]
    fesb2 ( 2)                 0.55897         0.32382
    fesb2 ( 3)                 1.0902          0.39518
    fosb3 ( 1)                       0.        ALIASED [I]
    fosb3 ( 2)                 0.83672         0.16570
    fops  ( 1)                       0.        ALIASED [I]
    fops  ( 2)                 1.0891          0.28586

C level 2 of fmsb has a high t-ratio; the others are lower
C test fmsb

lfit -fmsb

    Deviance =     2110.0394     at iteration    10

C the change in deviance is significant at the 10% level
C compared with 7.78, ie. chi-sq. for 4 degree of freedom

C The factor fmsb is significant, but the effect of
C some levels is small. Therefore collapse some levels of fmsb
C and use the 3 level factor fmsb1 instead.
lfit +fmsb1

    Deviance =     2102.9664     at iteration    10

dis est

    Parameter              Estimate         S. Error
    ____________________________________________________________
    int                        0.95067         0.69857
    ldur                      -0.99776         0.74232E-01
    year                      -0.38286E-01     0.73967E-02
    trage                      0.42904         0.33667
    trage2                    -0.17240         0.61843
    trage3                    -0.57939         0.52601
    trage4                     0.34715         0.56440
    trage5                     0.43121         0.20700
    trage6                    -0.23906         0.15230
    fmbu  ( 1)                       0.        ALIASED [I]
    fmbu  ( 2)                 1.2313          0.64655
    fmfm  ( 1)                       0.        ALIASED [I]
    fmfm  ( 2)                 0.46823         0.24019
    fmrm  ( 1)                       0.        ALIASED [I]
    fmrm  ( 2)                 1.4241          0.75185
    fch3  ( 1)                       0.        ALIASED [I]
    fch3  ( 2)                 -1.2177         0.74682
    fesb2 ( 1)                       0.        ALIASED [I]
    fesb2 ( 2)                 0.55546         0.32356
    fesb2 ( 3)                 1.0911          0.39499
    fosb3 ( 1)                       0.        ALIASED [I]
    fosb3 ( 2)                 0.83211         0.16526
    fops  ( 1)                       0.        ALIASED [I]
    fops  ( 2)                 1.1032          0.28470
    fmsb1 ( 1)                       0.        ALIASED [I]
    fmsb1 ( 2)                -0.44502         0.18885
    fmsb1 ( 3)                 0.11026         0.40049

C The change in deviance is significant at the
```

```
C 10% level compared with 4.6, ie. chi-sq. for 2 degree of freedom.
C Only level 2 seems significant.
C Collapse variable further; use 2 level factor msb2 instead.

lfit -fmsb1

    Deviance =     2110.0394     at iteration    10

lfit +fmsb2

    Deviance =     2103.0411     at iteration    10

dis est

    Parameter               Estimate         S. Error
    ____________________________________________________
    int                      0.98308          0.68858
    ldur                    -0.99954          0.73925E-01
    year                    -0.38417E-01      0.73821E-02
    trage                    0.44814          0.32946
    trage2                  -0.18073          0.61760
    trage3                  -0.58213          0.52585
    trage4                   0.35071          0.56434
    trage5                   0.43121          0.20699
    trage6                  -0.23961          0.15231
    fmbu   ( 1)                   0.          ALIASED [I]
    fmbu   ( 2)               1.2346          0.64645
    fmfm   ( 1)                   0.          ALIASED [I]
    fmfm   ( 2)              0.46040          0.23846
    fmrm   ( 1)                   0.          ALIASED [I]
    fmrm   ( 2)               1.5114          0.68339
    fch3   ( 1)                   0.          ALIASED [I]
    fch3   ( 2)              -1.2225          0.74642
    fesb2  ( 1)                   0.          ALIASED [I]
    fesb2  ( 2)              0.55741          0.32354
    fesb2  ( 3)               1.0932          0.39493
    fosb3  ( 1)                   0.          ALIASED [I]
    fosb3  ( 2)              0.83115          0.16524
    fops   ( 1)                   0.          ALIASED [I]
    fops   ( 2)               1.1069          0.28439
    fmsb2  ( 1)                   0.          ALIASED [I]
    fmsb2  ( 2)             -0.46453          0.17487

C The addition of fmsb2 to the model produces a change in
C deviance significant at the 10% level. The coefficient estimate is
C now significant. Keep fmsb2 in the model.

C Remove the remaining factors one by one and compare each
C change in deviance with 2.71 (chi-sq. at the 10% level,
C 1 degree of freedom).

lfit -fch3

    Deviance =     2106.7537     at iteration    10

lfit +fch3

    Deviance =     2103.0411     at iteration    10

lfit -fmbu

    Deviance =     2105.8325     at iteration    10

lfit +fmbu

    Deviance =     2103.0411     at iteration    10

lfit -fmrm

    Deviance =     2106.7500     at iteration    10

lfit +fmrm

    Deviance =     2103.0411     at iteration    10

lfit -fmfm

    Deviance =     2106.5408     at iteration    10

lfit +fmfm
```

```
    Deviance =     2103.0411     at iteration    10
lfit -fesb2
    Deviance =     2111.2846     at iteration    10
lfit +fesb2
    Deviance =     2103.0411     at iteration    10
lfit -fops
    Deviance =     2115.7878     at iteration    10
lfit +fops
    Deviance =     2103.0411     at iteration    10
lfit -fosb3
    Deviance =     2126.2913     at iteration    10
lfit +fosb3
    Deviance =     2103.0411     at iteration    10
C All the above factors are significant.
dis est
    Parameter               Estimate         S. Error
    ____________________________________________________
    int                      0.98308          0.68858
    ldur                    -0.99954          0.73925E-01
    year                    -0.38417E-01      0.73821E-02
    trage                    0.44814          0.32946
    trage2                  -0.18073          0.61760
    trage3                  -0.58213          0.52585
    trage4                   0.35071          0.56434
    trage5                   0.43121          0.20699
    trage6                  -0.23961          0.15231
    fch3   ( 1)                0.             ALIASED [I]
    fch3   ( 2)              -1.2225           0.74642
    fmbu   ( 1)                0.             ALIASED [I]
    fmbu   ( 2)               1.2346           0.64645
    fmrm   ( 1)                0.             ALIASED [I]
    fmrm   ( 2)               1.5114           0.68339
    fmfm   ( 1)                0.             ALIASED [I]
    fmfm   ( 2)              0.46040           0.23846
    fmsb2  ( 1)                0.             ALIASED [I]
    fmsb2  ( 2)             -0.46453           0.17487
    fesb2  ( 1)                0.             ALIASED [I]
    fesb2  ( 2)              0.55741           0.32354
    fesb2  ( 3)               1.0932           0.39493
    fops   ( 1)                0.             ALIASED [I]
    fops   ( 2)               1.1069           0.28439
    fosb3  ( 1)                0.             ALIASED [I]
    fosb3  ( 2)              0.83115           0.16524
C Is trage6 still significant?
lfit -trage6
    Deviance =     2106.0860     at iteration     8
C trage6 is significant at the 10% level
C The above model is therefore our final main effects model.
stop
```

Next:Random effects model with explanatory variables

Home page | **Contents** | **Previous**



The random effects model with explanatory variables

We now fit the random effects model with the explanatory variables we found significant in our previous analysis.

SABRE SESSION:INPUT AND OUTPUT

```
data case move age year dur ed ch1 ch2 ch3 ch4 msb mse esb ese &
osb ose mbu mrm mfm msb1 epm eoj esb1 ops osb1 msb2 esb2 osb2 osb3
read rochmig.dat

       6349 observations in dataset

yvar move
transform tempage age - 30
transform trage tempage / 10
transform trage2 trage * trage
transform trage3 trage2 * trage
transform trage4 trage3 * trage
transform trage5 trage4 * trage
transform trage6 trage5 * trage
transform ldur log dur
factor ch3 fch3
factor mbu fmbu
factor mrm fmrm
factor mfm fmfm
factor ops fops
factor esb2 fesb2
factor osb3 fosb3
factor msb2 fmsb2
C fit simple logistic main effects model

lfit int ldur year trage trage2 trage3 trage4 trage5 trage6 &
fch3 fesb2 fmbu fmrm fmfm fops fmsb2 fosb3

    Iteration         Deviance         Reduction
    ___________________________________________
        1            8801.5829
        2            2935.4172          5866.
        3            2266.9758          668.4
        4            2124.9723          142.0
        5            2105.4389          19.53
        6            2103.4563          1.983
        7            2103.0820         0.3743
        8            2103.0417        0.4028E-01
        9            2103.0411        0.5907E-03
       10            2103.0411        0.1442E-06

dis est

    Parameter                Estimate          S. Error
    ___________________________________________________
    int                       0.98308          0.68858
    ldur                     -0.99954          0.73925E-01
    year                     -0.38417E-01      0.73821E-02
    trage                     0.44814          0.32946
    trage2                   -0.18073          0.61760
    trage3                   -0.58213          0.52585
    trage4                    0.35071          0.56434
    trage5                    0.43121          0.20699
    trage6                   -0.23961          0.15231
    fch3   ( 1)                    0.          ALIASED [I]
    fch3   ( 2)                -1.2225          0.74642
    fesb2  ( 1)                    0.          ALIASED [I]
    fesb2  ( 2)               0.55741          0.32354
    fesb2  ( 3)                1.0932          0.39493
    fmbu   ( 1)                    0.          ALIASED [I]
    fmbu   ( 2)                1.2346          0.64645
    fmrm   ( 1)                    0.          ALIASED [I]
    fmrm   ( 2)                1.5114          0.68339
    fmfm   ( 1)                    0.          ALIASED [I]
```

```
    fmfm  ( 2)              0.46040          0.23846
    fops  ( 1)                   0.          ALIASED [I]
    fops  ( 2)               1.1069          0.28439
    fmsb2 ( 1)                   0.          ALIASED [I]
    fmsb2 ( 2)             -0.46453          0.17487
    fosb3 ( 1)                   0.          ALIASED [I]
    fosb3 ( 2)              0.83115          0.16524

C fit the same model with random effects
fit .

    Iteration        Deviance         Step      End-points
Orthogonality
                                     length    0          1
criterion

_____________________________________________________________________________
         1            2135.3134        1.0000    free   free
4.6620
         2            2128.4000        0.2500    free   free
4.4184
         3            2123.0042        0.4288    free   fixed
0.21340E-01
         4            2122.5984        0.0078    free   fixed
7.1550
         5            2088.6586        1.0000    free   free
4.0641
         6            2079.1069        1.0000    free   free
4.5894
         7            2075.6823        1.0000    free   free
24.604
         8            2075.6458        1.0000    free   free
17.341
         9            2075.6458        1.0000    free   free

dis est

    Parameter               Estimate         S. Error
    ____________________________________________________
    int                      0.73017          0.89590
    ldur                    -0.63527          0.10783
    year                    -0.37769E-01      0.10902E-01
    trage                    0.17360          0.34893
    trage2                  -0.16613E-01      0.61773
    trage3                  -0.54783          0.53830
    trage4                   0.28880          0.56026
    trage5                   0.40754          0.21096
    trage6                  -0.22024          0.15128
    fch3  ( 1)                    0.          ALIASED [I]
    fch3  ( 2)               -1.3073          0.75078
    fesb2 ( 1)                    0.          ALIASED [I]
    fesb2 ( 2)               0.31615          0.37617
    fesb2 ( 3)               0.77441          0.45042
    fmbu  ( 1)                    0.          ALIASED [I]
    fmbu  ( 2)                1.2513          0.66612
    fmrm  ( 1)                    0.          ALIASED [I]
    fmrm  ( 2)                1.5259          0.71835
    fmfm  ( 1)                    0.          ALIASED [I]
    fmfm  ( 2)               0.45266          0.25126
    fops  ( 1)                    0.          ALIASED [I]
    fops  ( 2)                1.2016          0.30209
    fmsb2 ( 1)                    0.          ALIASED [I]
    fmsb2 ( 2)              -0.56390          0.19405
    fosb3 ( 1)                    0.          ALIASED [I]
    fosb3 ( 2)               0.68677          0.18610
    scale                    0.49269          0.18099

PROBABILITY


____________
    end-point 0              0.48867          0.19067          0.32760
    end-point 1              0.29991E-02      0.43654E-02      0.20105E-
02

stop
```

Next: Interpretation of results

Home page | Contents | Previous

Interpretation of results

The explanatory variables

Backward elimination using the simple logistic model has shown the following variables to be significant at the 10% level:

- **employment status**: esb2=1 (self employed), esb2=2 (employed), esb2=3 (not working)
- **occupational status**: osb3=1 (small proprietors, supervisors), osb3=0 (otherwise)
- **promotion to service class**: ops=0 (no), ops=1 (yes)
- **first marriage**: mfm=0 (no), mfm=1 (yes)
- **marital break-up**: mbu=0 (no), mbu=1 (yes)
- **remarriage**: mrm=0 (no), mrm=1 (yes)
- **presence of children age 15-16**: ch3=0 (no), ch3=1 (yes)
- **marital status**: msb2=0 (not married), msb2=1 (married)

Our preferred homogeneous main effects model is therefore:

$$\text{age}+\text{age}^2+\text{age}^3+\text{age}^4+\text{age}^5+\text{age}^6+\text{year}+\log(\text{dur})+\text{esb2}+\text{osb3}+\text{ops}+\text{mfm}+\text{mbu}+\text{mrm}+\text{ch3}+\text{msb2}$$

Comparison of simple logistic and random effects models

- When the same model is fitted with random effects, the deviance decreases by 27.4. Although it is not strictly correct to use the c^2 test to compare the simple logistic and random effects models, such a substantial reduction in deviance for three extra parameters estimated (scale and two endpoints) provides evidence that in addition to the time varying explanatory variables included in the model, there remains unobserved heterogeneity.

- Comparison of the parameter estimates of the two models shows that, as before, only the estimate of the endogenous *log(dur)* has changed substantially (from -0.9995 to -0.6353): controlling for unobserved heterogeneity has decreased the observed negative duration of stay effect. (See Lancaster and Nickell 1980). The other parameter estimates for the two models are the same, within one standard error.

- The parameter estimates of *msb2* and *ch3* are both negative, providing evidence that being married significantly reduces the probability of migration, as does the presence of children in the age group 15-16, presumably for fear of disrupting schooling close to public examinations. There is no evidence that younger or older secondary school-age children increase ties to an area.

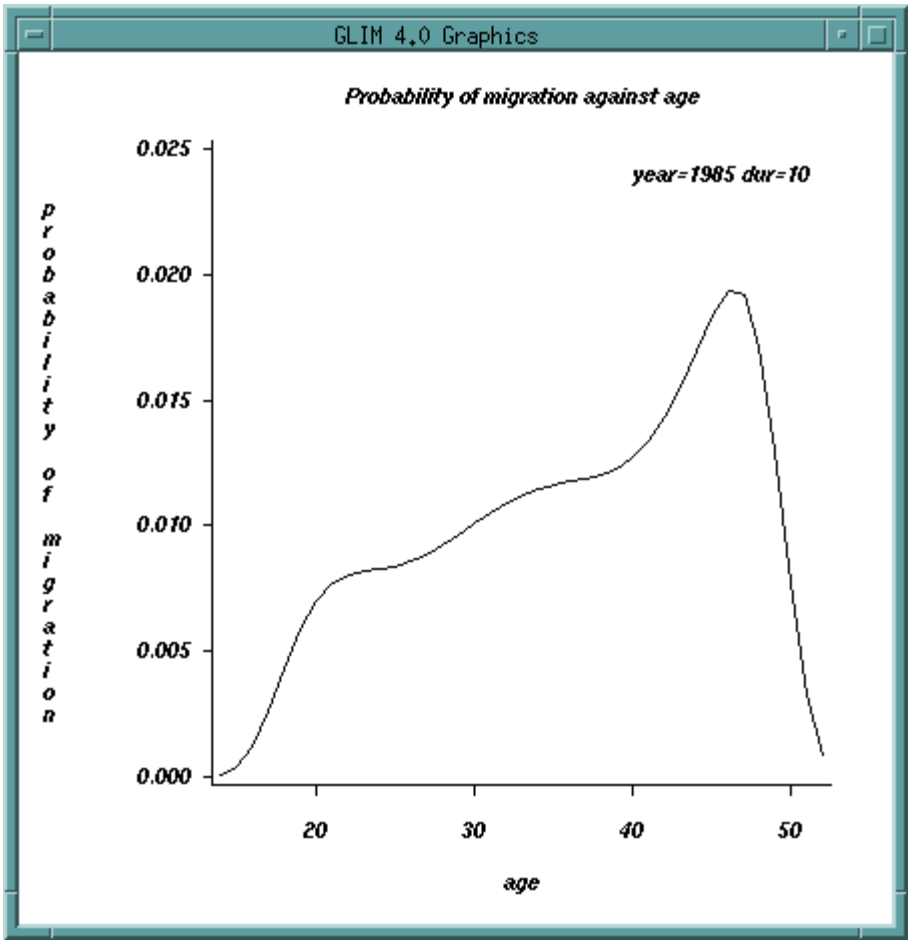
- The positive coefficient estimates for *mfm, mbu, mrm* and *ops* indicate that the events of first marriage, marital break-up, remarriage and promotion to service class all increase the probability of migration.

- The positive coefficients for levels 2 and 3 of *esb2* provides evidence that employed and unemployed individuals are more likely to migrate than the self-employed. Also the positive coefficient of *osb3* indicates that small proprietors and supervisors are more likely to migrate than others.

- The probability of 0.3276 estimated for the left hand endpoint again indicates a high proportion of stayers. The right endpoint is small and may be set to zero.

Variation with age

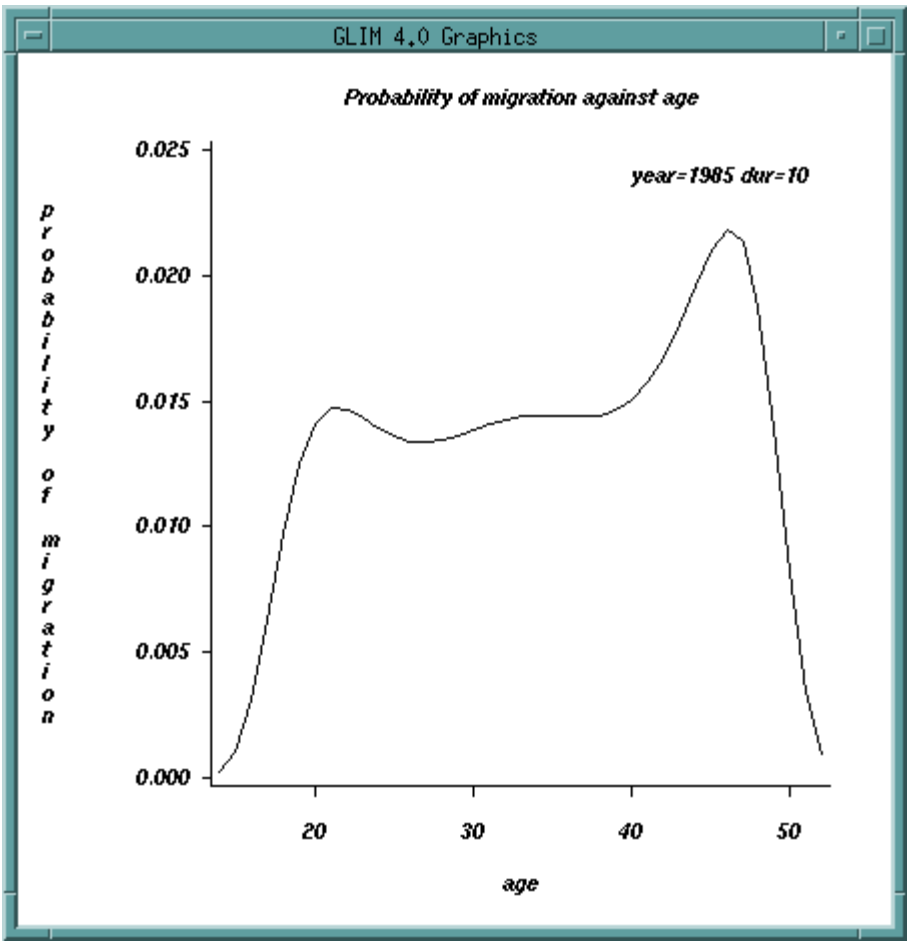
- To illustrate the difference between the homogeneous and random effects models, we plot the probability of migration against age, with the year taken as 1985, duration of stay 10 years and all other explanatory variables set to zero (ie. to their reference levels). As there are no interaction terms, the patterns shown on the graphs are generally valid.



Simple logistic model

Random effects model

Both graphs show a peak just below age 50, where the data are sparse; the random effects model, although flatter over the earlier years, has a more accentuated first peak just above age 20. The three peaks are less pronounced than in the original analysis without explanatory variables, but it is clear that controlling for life cycle effects provides only a partial explanation of the



three peaks.

We shall examine the contribution of some of the explanatory variables to the peaks. Because of the excessive computing requirements of the random effects model, we shall use the simple logistic model in this analysis.

Next:Contribution of life cycle events to the peaks

Home page | **Contents** | **Previous**

Contribution of life cycle events to the peaks

To examine the contribution of an explanatory variable on the peaks, the variable is omitted from the preferred homogeneous main effects model and the simplified model is refitted. The probability of migration is plotted against age, with the year set to 1985, duration to 10 years and **all the explanatory variables set to zero**, as before.

The following graphs show the effects of removing in turn *msb2*, *mrm* and *ch3* from the full model. Similar graphs may be drawn for the other explanatory variables.

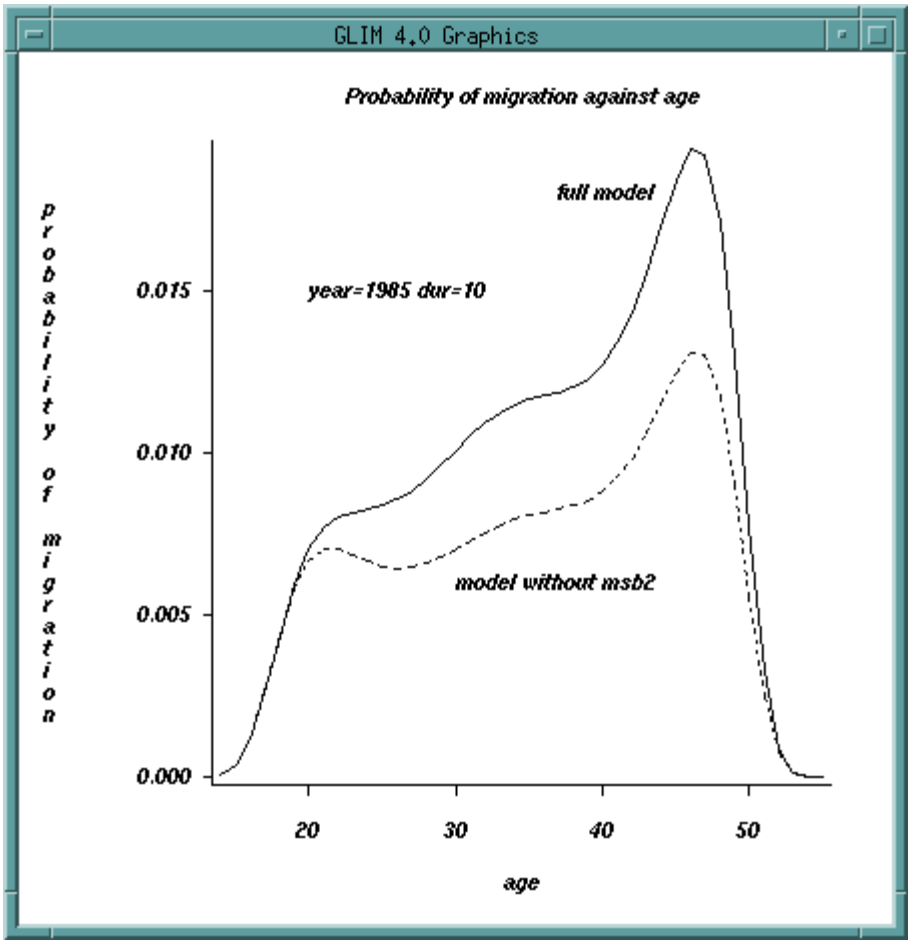
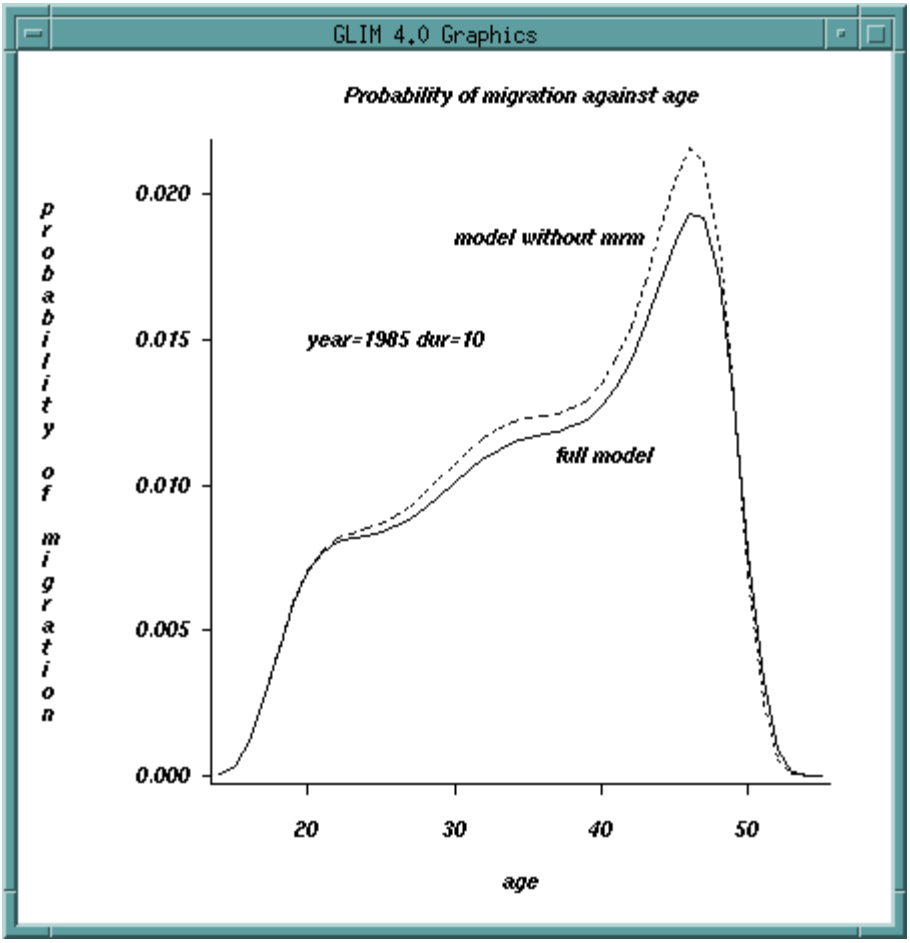


Figure 1: The effect of removing msb2 (marital status)

The basic shapes of the graphs are very similar, suggesting just a scaling effect, and no explanation of the peak.

Figure 2: The effect of removing mrm (remarriage)

The peaks seem to be slightly attenuated in the full model with *mrm*=0 compared to the simplified model. It appears that the minor difference between the graphs is not just a scaling effect, but evidence that remarriage contributes to the third peak.



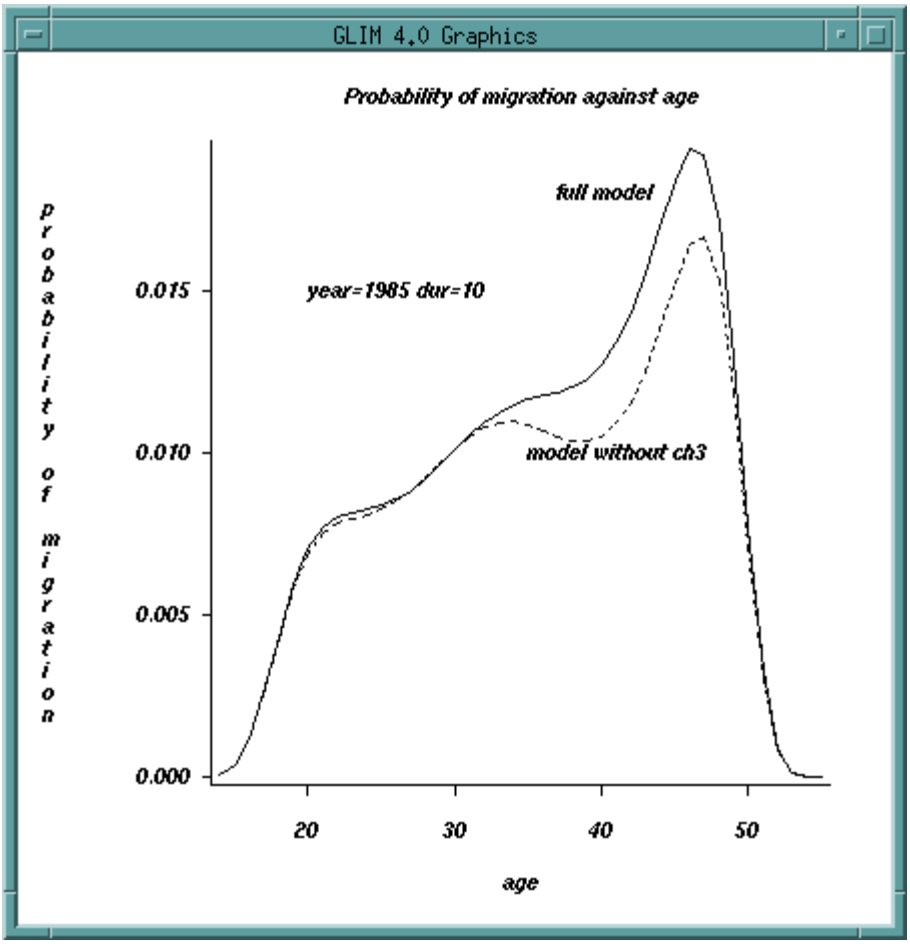


Figure 3: The effect of removing ch3 (children aged 15-16)

This variable appears to provide a partial explanation for migration behaviour in the age range 35 to 50 (the appropriate age for parents of children aged 15-16). The trough around age 40 with *ch3* excluded from the model is partially smoothed out in the full model with *ch3*=0. However, although having a child aged 15 to 16 does significantly reduce the probability of migration, the third peak is not attenuated in the full model, but is in fact increased, for those without children in this age range. This effect therefore does not explain the third peak.

Next:Conclusions and suggestions for further work

Home page | **Contents** | **Previous**

Conclusions and suggestions for further work

- We must be cautious about drawing general conclusions from this analysis as the sample was drawn from one locality. However, the extent to which migration behaviour with age can be explained by explanatory variables is likely to be informative about the process of migration.

- We have identified three statistically significant peaks in migration behaviour with age during individuals' working lives; at just above age 20, at around age 35 and just below age 50. The size and location of the third peak has to be interpreted with caution as the data are sparse here.

- We have shown that there is considerable heterogeneity in the population sampled, with a considerable proportion of individuals who are likely never to move.

- The negative coefficient estimate for *ldur* indicates that the probability of migration decreases with duration of stay in the locality, consistent with the concept of cumulative inertia.

- The simple logistic model takes no account of the fact that in a heterogeneous population, the individuals most likely to migrate are more and more underrepresented with increasing duration, and therefore inflates the duration of stay effect. To estimate the true effect of cumulative inertia, we must control for residual population heterogeneity.

- For the years studied the likelihood of migration decreased with calendar time for the population surveyed.

- The following time varying explanatory variables have been found to have a significant effect on migration (at the 10% level):

Employment status
Occupational status
Promotion to service class
First marriage
Marital break-up
Remarriage
Presence of children age 15-16
Marital status

- It is evident that the third peak in the pattern of migration with age persist even after controlling for the time-varying explanatory variables. Remarriage appears to make a small contribution to this peak, however controlling for the presence of children of age 15-16 actually increases the size of the peak for those without children of this age.

- The main effects model may be extended by the addition of interaction terms both between the time variables and between time and other explanatory variables. If these are confined to the linear term in age, there are 55 possible pairwise interactions. An interaction model has been fitted to this data by Borhani Haghighi and Davies (1999b). These throw light on questions such as:

1. Does the relative importance of the three peaks vary with calendar year?
2. Do patterns of migration behaviour for employed/self- employed/not working individuals relate to age?
3. Is the probability of migration after marriage break-up/remarriage age related?

We leave this for the student to explore.

As we have analysed migration data from only one locality, it is not clear how far the results are generally

characteristic of the process of inter-county migration and how far they are location specific. Analysing datasets from some of the other SCELI localities would throw light on this question. See Davies and Flowerdew (1992) for some early comparative work.

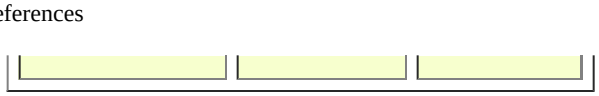
Next:References

Home page | **Contents** | **Previous**

References

1. Borhani Haghighi, A. and Davies, R. B. (1999a), Characterising temporal effects in social science data, *Computational Statistics and Data Analysis*, forthcoming
2. Borhani Haghighi, A. and Davies, R. B. (1999b), How migration propensity varies with age; the effects of life cycle and individual level characteristics, *Environment and Planning A*, forthcoming
3. Boyle, P., Halfacree, K. and Robinson V. (1998), *Exploring Contemporary Migration*, Longman
4. Coleman, J. S. (1973), *The Mathematics of Collective Action*, Heinemann
5. Cote, G. L. (1997), Socio-economic attainment, regional disparities and internal migration, *European Sociological Review* **13**, No.1, p. 55-77.
6. Davies, R. B. and Flowerdew R. (1992), Modeling migration careers, using data from a British survey, *Geographical Analysis*, **24**, No.1, p. 35-57
7. Devis, T. (1983), People changing address:1971 and 1981, *Population Trends*, **32**, p.15-20.
8. Dex, S. (1995), The reliability of recall data: A literature review, *Bulletin de Methodologie Sociologique*, **49**, p. 58-80.
9. Dex, S. and McCullough, A. (1998), The reliability of retrospective unemployment history data, *Work, Employment and Society*, **12**, no.3, p. 497-509.
10. Ellis, M., Barff, R., and Renard, B.(1993), Migration regions and interstate labor flows by occupation in the United States, *Growth and Change*, **24**, No. 2, p. 166-190.
11. Greenwood, M. J.(1985), Human migration: Theory, models and empirical studies, *Journal of Regional Science*, **25**, No. 4, p. 521-544.
12. Goss, E. P. (1985), General skills, specific skills and the migration decision, *Regional Science Perspective*, **15**, p. 17-26.
13. Grundy, E. M. C.(1989), *OPCS Longitudinal Study - Women's migration: Marriage, fertility and divorce*, HMSO
14. Heckman, J. J. and Singer, B. (1984), A method of minimising the impact of distributional assumptions in econometric models of duration, *Econometrica*, **52**, p. 271-320.
15. Heckman, J. J. and Singer, B. (1985), Social Science duration analysis, *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, p. 39-58.
16. Hertzog Jr., H. W., Schlottmann, A.M., and Boehm, T. P. (1993), Migration as a spatial job-search: A survey of empirical findings, *Regional Studies*, **27**, No. 4, p. 327-340.
17. Huff, J. O. and Clark, W. A. V. (1978), Cumulative stress and cumulative inertia: A behavioral model of decision to move, *Environment and Planning A*, **10**, p. 1101-1119.
18. Lancaster, T. (1979), Econometric methods for the duration of unemployment, *Econometrica*, **47**, No. 4.
19. Lancaster, T. and Nickell, S. (1980), The analysis of re-employment probabilities for the unemployed, *Journal of the Royal Statistical Society A* , **143**, Part 2, p. 141-165.
20. Liaw, K. L. (1990), Joint effects of personal factors and ecological variables on the interprovintial migration pattern of young adults in Canada: A nested logit analysis, *Geographical Analysis*, **22**, No. 3, p. 189-208.
21. Long, L. L. (1972), The influence of the number of children on residential mobility, *Demography* , **9**, No. 3.
22. Mc Ginnis, R. (1968), A stochastic model of social mobility, *American Sociological Review*, p. 712-722
23. Salt, J. (1990), Organisational labour migration: Theory and practice in the United Kingdom, in *Labour Migration*, ed. J. H. Johnson and J. Salt, p. 52-69 (David Fulton)
24. Sandefur, G. D. and Scott, W. J. (1981), A dynamic analysis of migration: an assessment of the effects of age, family and career variables, *Demography*, **18**, No. 3, p. 355-368.
25. Warnes, A. M. (1983), Migration in late working age and early retirement, *Socio-Economic Planning Sciences*, **17**, p.291- 302.

Home page | **Contents** | **Previous**



MLwiN - Overviews

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Overviews of two analyses are available giving examples of the potential of multilevel modelling for social data.

The first example shows some findings from a multilevel analysis of educational attainment data from pupils attending a secondary school in London. The response variable is attainment in exams taken by pupils at age 16. There are data on 4000 pupils in 65 schools. The analysis is particularly concerned with the effect of schools. Are some schools more "effective" than others? *More* ▶

The second example is an exploration of variations in mortality rates in England and Wales. The data comprise repeated measures of the standardised mortality ratio (SMR) for 403 county districts which are nested in 54 counties over the period 1979 to 1991. The particular concerns are in how mortality changed over that time period and the nature and extent of regional variations. *More* ▶

The full tutorials and datasets may also be downloaded; these are designed to take you through the analyses on which the above overviews are based.

Next Section: *Tutorials* ▶

Top

Chapter 1: Random Intercept and Random Slope Models

This chapter is a tutorial which will take you through the basic procedures for specifying a multilevel model in *MLwiN*, estimating parameters, making inferences, and plotting results. It provides both an introduction to the software and a practical introduction to multilevel modelling.

As we have seen, multilevel models are useful in a very wide range of applications. For illustration here, we use an educational data set for which an *MLwiN* worksheet has already been prepared. Usually, at the beginning of an analysis, you will have to create such a worksheet yourself either by entering the data directly or by reading a file or files prepared elsewhere. Facilities for doing this are described at the end of this chapter. The data in the worksheet we use have been selected from a very much larger data set, of examination results from six inner London Education Authorities (school boards). A key aim of the original analysis was to establish whether some schools were more 'effective' than others in promoting students' learning and development, taking account of variations in the characteristics of students when they started Secondary school. The analysis then looked for factors associated with any school differences found. Thus the focus was on an analysis of factors associated with examination performance after adjusting for student intake achievements. As you explore *MLwiN* using the simplified data set you will also be imitating, in a simplified way, the procedures of the original analysis. For a full account of that analysis see Goldstein *et al.* (1993).

Opening the worksheet and looking at the data

When you start *MLwiN* the main window appears. Immediately below the *MLwiN* title bar are the *menu bar* and below it the *tool bar* as shown:

These menus are fully described in the online Help system.This may be accessed either by clicking the **Help** button on the menu bar shown above or (for context-sensitive

Help) by clicking the **Help** button displayed in the window you are currently working with. You should use this system freely.

The buttons on the tool bar relate to model estimation and control, and we shall describe these in detail later. Below the tool bar is a blank workspace into which you will open windows using the **Window** menu These windows form the rest of the 'graphical user interface' which you use to specify tasks to *MLwiN*. Below the workspace is the *status bar*, which monitors the progress of the iterative estimation procedure. Open the tutorial worksheet as follows:

Select **File** menu

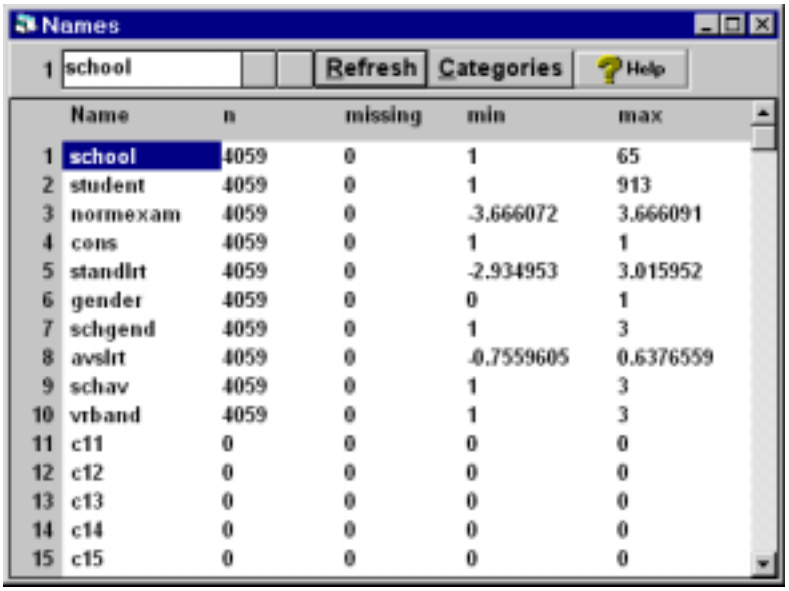
Select **Open worksheet**

Select **tutorial.ws**

Click **Open**

When this operation is complete the filename will appear in the title bar of the main window and the status bar will be initialised.

The *MLwiN* worksheet holds the data and other information in a series of *columns*. These are initially named c1, c2, …,but the columns can (and should) be given meaningful names to show what their contents relate to. This has already been done in the **Tutorial** worksheet that you have loaded. When a worksheetis loaded a summary of the variables, shown below, automatically appears.



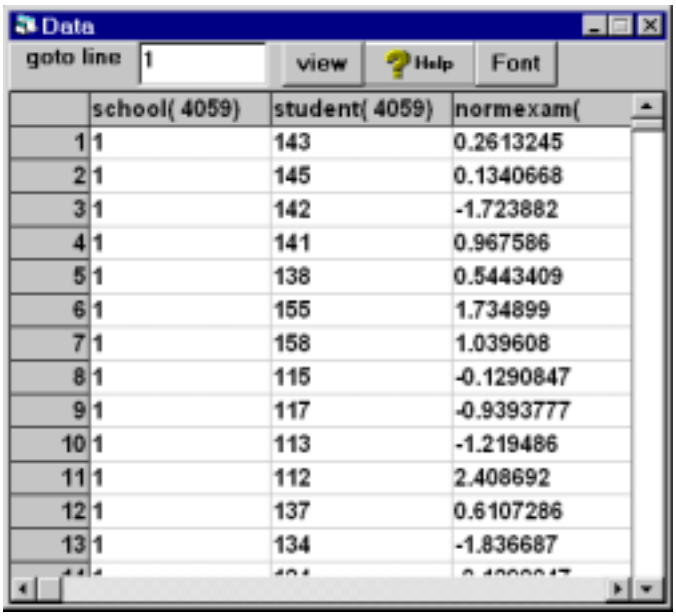
	Name	n	missing	min	max
1	school	4059	0	1	65
2	student	4059	0	1	913
3	normexam	4059	0	-3.666072	3.666091
4	cons	4059	0	1	1
5	standlrt	4059	0	-2.934953	3.015952
6	gender	4059	0	0	1
7	schgend	4059	0	1	3
8	avslrt	4059	0	-0.7559605	0.6376559
9	schav	4059	0	1	3
10	vrband	4059	0	1	3
11	c11	0	0	0	0
12	c12	0	0	0	0
13	c13	0	0	0	0
14	c14	0	0	0	0
15	c15	0	0	0	0

Each line in the body of the window summarises a column of data. In the present case only the first 10 of the 400 columns of the worksheet contain data. Each column contains 4059 items, one item for each student represented in the data set. There are no missing values, and the minimum and maximum value in each column are shown. Note the Help button on the tool bar. The remaining items on the tool bar of this window are for attaching a name to a column. We shall use these later.

You can view individual items in the data using the Data window as follows:

Select **Data manipulation** menu

Select **View or edit data**



	school(4059)	student(4059)	normexam(
1	1	143	0.2613245
2	1	145	0.1340668
3	1	142	-1.723882
4	1	141	0.967586
5	1	138	0.5443409
6	1	155	1.734899
7	1	158	1.039608
8	1	115	-0.1290847
9	1	117	-0.9393777
10	1	113	-1.219486
11	1	112	2.408692
12	1	137	0.6107286
13	1	134	-1.836687

When this window is first opened it always shows the first three columns in the worksheet. The exact number of items shown depends on the space available on your screen.

You can view any selection of columns, spreadsheet fashion, as follows:

Click the **View** button

Select columns to view

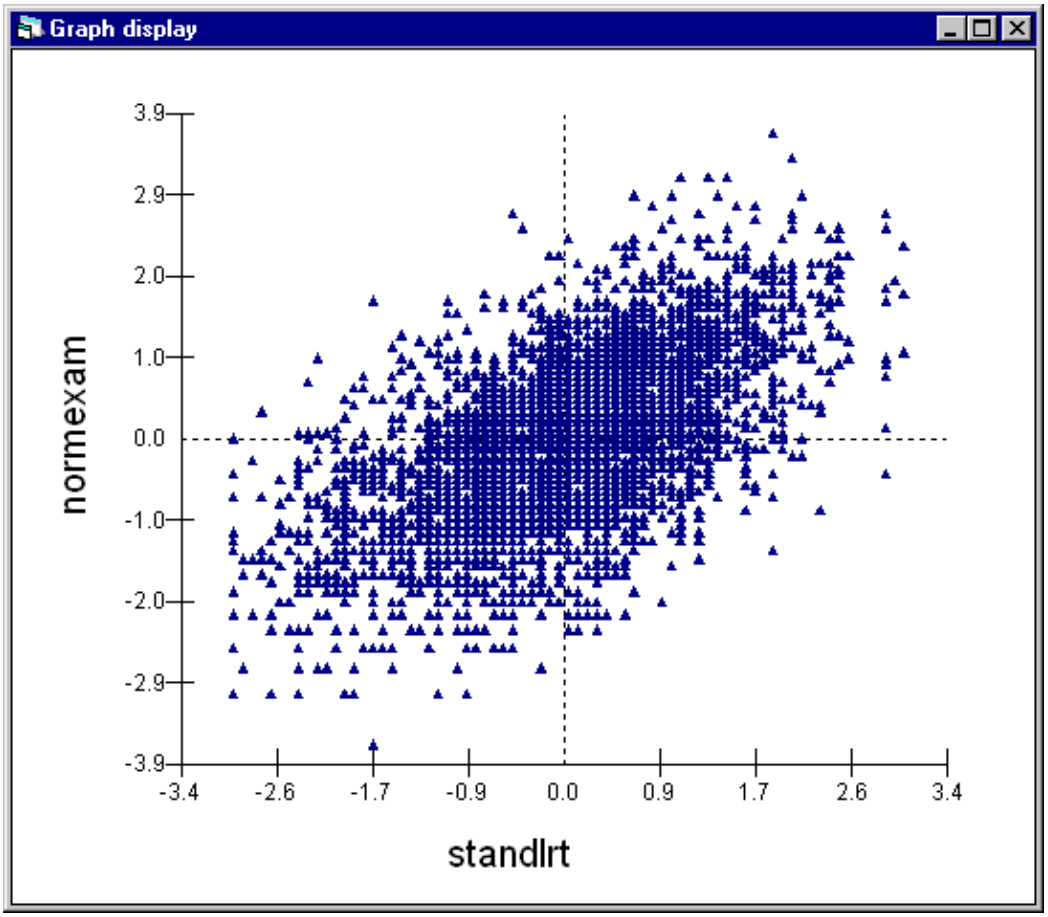
Click **OK**

You can select a block of adjacent columns either by pointing and dragging or by selecting the column at one end of the block and holding down 'Shift' while you select the column at the other end. You can add to an existing selection by holding down 'Ctrl' while you select new columns or blocks.

The **Font** button, which is present in several of the *MLwiN* windows, can be used to make the characters *in that window* larger or smaller. This can be useful when the space available for the windows is not too large.

The **school** and **student** columns contain identifiers; **normexam** is the exam score obtained by each student at age 16, Normalised to have approximately a standard Normal distribution, **cons** is a column of 1's, and **standlrt** is the score for each student at age 11 on the London Reading Test, standardised using *z*-scores. **Normexam** is going to be the y-variable and **cons** and **standlrt** the x-variables in our initial analysis. The other data columns will be used in later sections of the manual. Use the scroll bars of the **Data** window to move horizontally and vertically through the data, and move or resize the window if you wish. You can go straight to line 1035, for example, by typing 1035 in the **goto line** box, and you can highlight a particular cell by pointing and clicking. This provides a means to edit data: see the Help system for more details.

Having viewed your data you will typically wish to tabulate and plot selected variables, and derive other summary statistics, before proceeding to multilevel modelling. Tabulation and other basic statistical operations are available on the **basic statistics** menu. These operations are described in the help system. In our first model we shall be looking at the relationship between the outcome attainment measure **normexam** and the intake ability measure **standlrt** and at how this relationship varies across schools. The scatter plot of **normexam** against **standlrt** for the whole of the data looks like this:



The plot shows, as might be expected, a positive correlation with pupils with higher intake scores tending to have higher outcome scores. Our modelling will attempt to partition the overall variability shown here into a part which is attributable to schools and a part which is attributable to students. We will demonstrate later in the chapter how to produce such graphs in *MLwiN* but first we focus on setting up a basic model.

You can now proceed straight away to the next section of this chapter, or stop at this point and close *MLwiN*. No data have been changed and you can continue with the next section after re-opening the worksheet **Tutorial.ws**. Each of the remaining sections in this chapter is self-contained [but they must be read in the right order!], and you are invited to save the current worksheet (using a different name) where necessary to preserve continuity.

Setting up a variance components multilevel model

We now go through the process of specifying a two-level variance components model for the examination data. First, close any open windows in the workspace. Then:

Select **Model** menu

Select **Equations**

The following window appears:

This window shows the nucleus of a model, which you elaborate in stages to specify the one you want. The tool bar for this window is at the bottom, and we shall describe these buttons shortly.

The first line in the main body of the window specifies the default distributional assumption: the response vector has a mean specified in matrix notation by the *fixed part* XB, and a random part consisting of a set of random variables described by the *covariance matrix* Ω. This covariance matrix Ω incorporates the separate covariance matrices of the random coefficientss at each level. We shall see below how it is specified. Note that y and x_0 are shown in red. This indicates that they have not yet been defined.

To define the response variable we have to specify its name and also that there are two levels. The lowest level, level 1, represents the variability between students at the same school; the next higher level, level 2, represents the variability between different schools. To do all this

Click *y* (either of the *y* symbols shown will do)

The **Y variable** dialogue box appears, with two drop-down lists: one labelled **y**, the other labelled **N levels**.

In the **y** list, select **normexam**

In the **N levels** list, select **2-ij**

By convention, the suffix *i* is used by *MLwiN* for level 1 and *j* for level 2, but suffixes can be changed as we will show later.

This reveals two further drop-down lists, *level 2 (j)* and *level 1 (i).*

In the *level 2 (j)* list, select **school** In the *level 1 (i)* list, select **student** Click **done**

In the **Equations** window the red *y* has changed to y_{ij} in black indicating that the response and the number of levels have been defined.

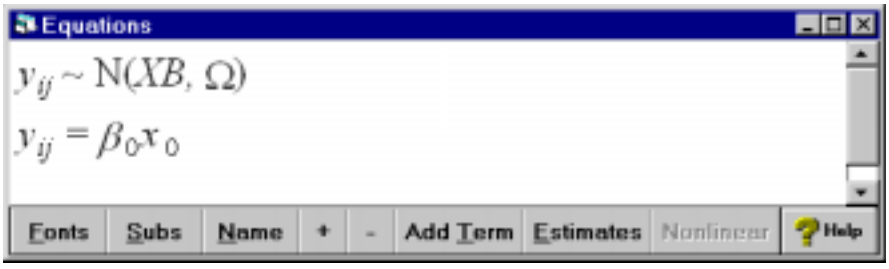
Now we must define the explanatory variables.

Click x_0 In the drop-down list, select **cons**

Note that the **fixed parameter** box is checked: by default, each explanatory variable is assumed to have a fixed parameter. We have just identified the explanatory variable x_0 with a column of 1's. This vector of 1's explicitly models the intercept. Other software packages may do this for you automatically, however, in the interests of greater flexibility, *MLwiN* does not.

Click **Done**

The **Equations** window now looks like this:



We are gradually building equation (1.8) which assumes the simple level 2 variation shown in figure 1.2. We have specified the fixed parameter associated with the intercept, and now require another explanatory variable.

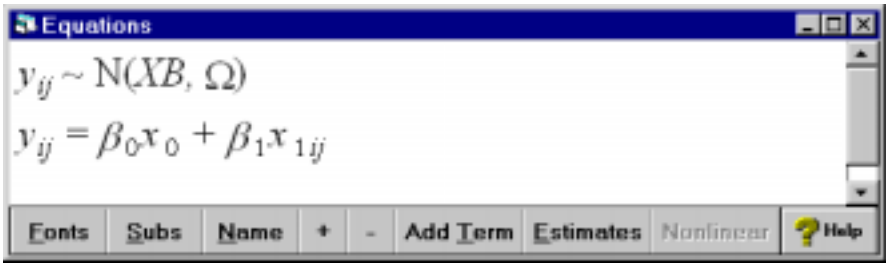
Click the **AddTerm** button on the tool bar

Click x_1

Select **standlrt**

Click **Done**

The **Equations** window looks like this –



This completes the specification of the fixed part of the model. Note that x_0 has no other subscript but that x_1 has collected subscripts *ij*. *MLwiN* detects that **cons** is constant over the whole data set, whereas the values of **standlrt** change at both level 1 and level 2.

To define the random part.

Click β_0 (or x_0)

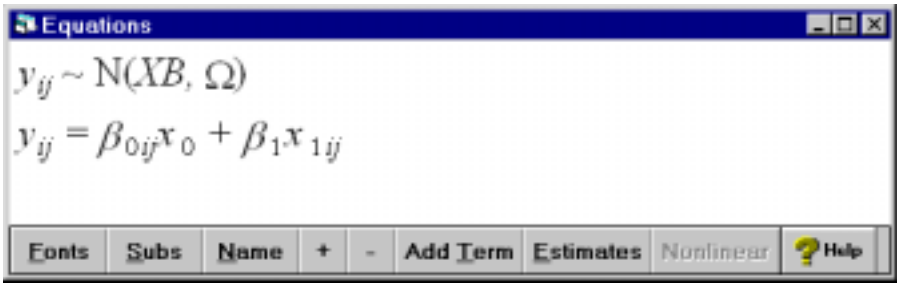
This redisplays the dialogue box for x_0, seen earlier. We wish to specify that the coefficient of x_0 is random at both school and student levels.

Check the box labelled **j(SCHOOL)**

Check the box labelled **i(STUDENT)**

Click **Done**

This produces

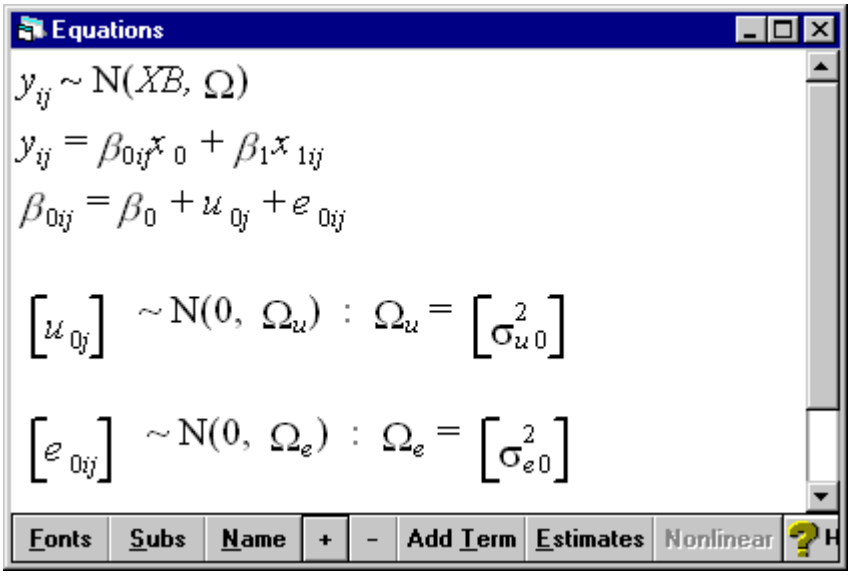


We have now defined the model. To see the composition of β_{0ij},

Click the + button on the tool bar

You should now see the model as defined in equation (1.8).

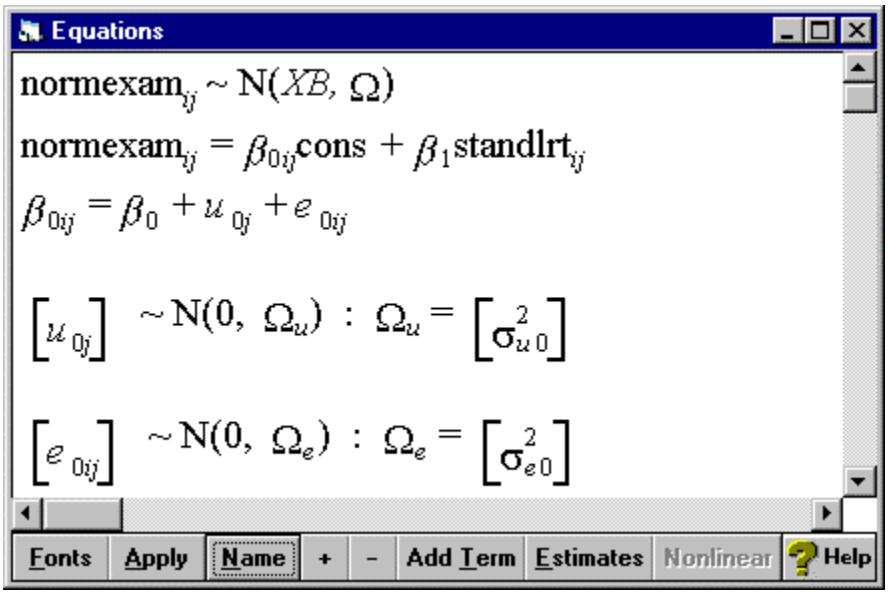
The + and – buttons control how much detail of the model is displayed. Click + a second time to reveal:



You may need to resize the window by dragging the lower border in order to see all the details, or alternatively change the font size.

To replace *y*, x_0 and x_1 by their variable names,

Click the **Name** button



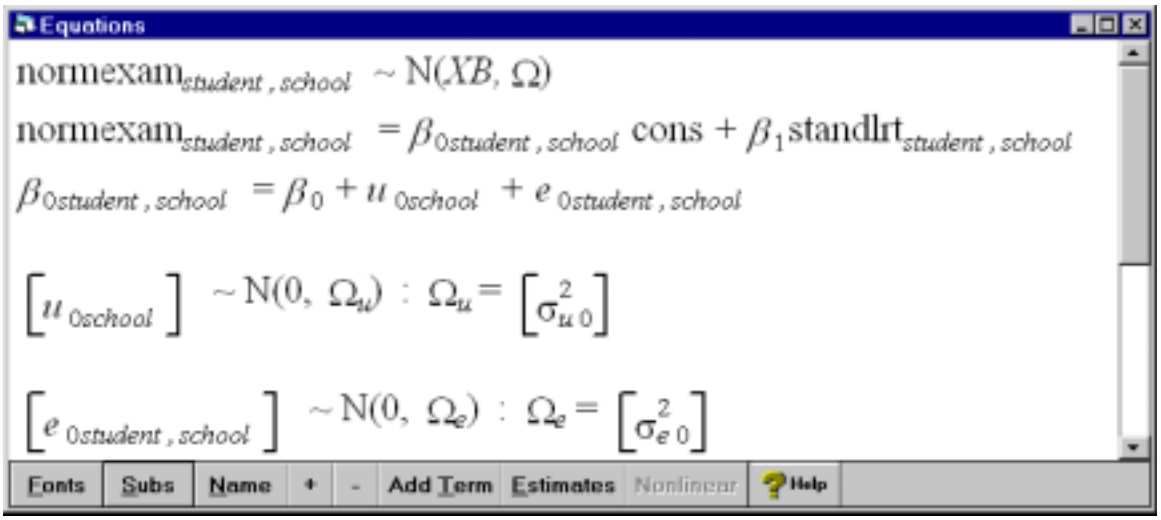
The **Name** button is a 'toggle': clicking again brings back the *x*'s and *y*'s.

In summary, the model that we have specified relates **normexam** to **standlrt.** The regression coefficients for the intercept and the slope of **standlrt** are (β_0, β_1). These coefficients define the average line across all students in all schools. The model is made multilevel by allowing each school's summary line to depart (be raised or lowered) from the average line by an amount u_{0j}. The *i*'th student in the *j*'th school departs from its school's summary line by an amount e_{0ij}. The information conveyed on the last two lines of the display is that the school level random departures u_{0j} are distributed Normally with mean 0 and variance σ^2_{u0} and the student level random departures e_{0ij} are distributed Normally with mean 0 and variance σ^2_{e0} (the Ω's can be ignored for the time being). The u_{0j} (one for each school) are called the level 2 or school level *residuals*; the e_{0ij} (one for each student) are the level 1 or student level residuals.

Just as we can toggle between *x*'s and actual variable names, so we can show actual variable names as subscripts. To do this

Click the **Subscripts** button

Which produces :

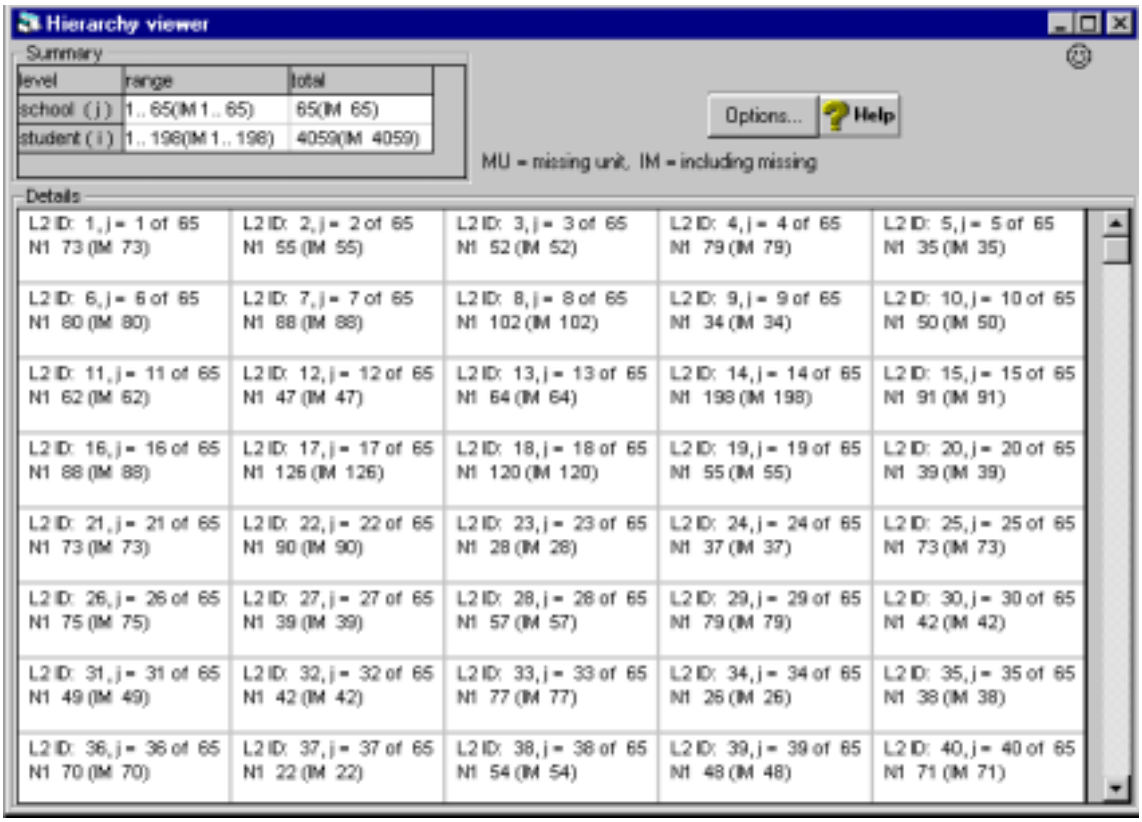


This display is somewhat verbose but a little more readable than the default subscript display. You can switch between the subscript formats by pressing the **subscripts** button. The screen shots in this chapter use the default subscript format. You can gain more control over how subscripts are displayed by clicking on **subscripts** from the model menu.

Before running a model it is always a good idea to get *MLwiN* to display a summary of the hierarchical structure to make sure that the structure *MLwiN* is using is correct. To do this

Select the **Model** menu Select **Hierarchy Viewer**

Which produces :



The top **summary** grid shows, in the **total** column, that there are 4059 pupils in 65 schools. The range column shows that there are maximum of 198 pupils in any school. The **details** grid shows information on each school. 'L2 ID' means 'level 2 identifier value', so that the first cell under **details** relates to school no 1. If when you come to analyse your own data the hierarchy that is reported does not conform to what you expect, then the most likely reason is that your data are not sorted in the manner required by *MLwiN*. In an *n* level model *MLwiN* requires your data to be sorted by level 1, within level 2, within level 3...level *n*. There is a sort function available from the **Data Manipulation** menu.

We have now completed the specification phase for this simple model. It is a good idea to save the worksheet which contains the specification of the model so far, giving it a different name so that you can return to this point in the manual at a later time.

Estimation

We shall now get *MLwiN* to estimate the parameters of the model specified in the previous section.

We going to estimate the two parameters β_0 and β_1 which in a single level model are the regression coefficients. In multilevel modelling regression coefficients are referred

to as constituting the **fixed** part of the model**.** We also estimate the variance of the school level random effects σ^2_{u0} and the variance of the pupil level random effects σ^2_{e0}. The random effects and their variances are referred to as the **random** part of the model**.**

Click the **Estimates** button on the **Equations** window tool bar

You should see highlighted in blue the parameters that are to be estimated. Initially, we will not estimate the 4059 individual pupil level random effects and 65 school level random effects, we will return to these later.

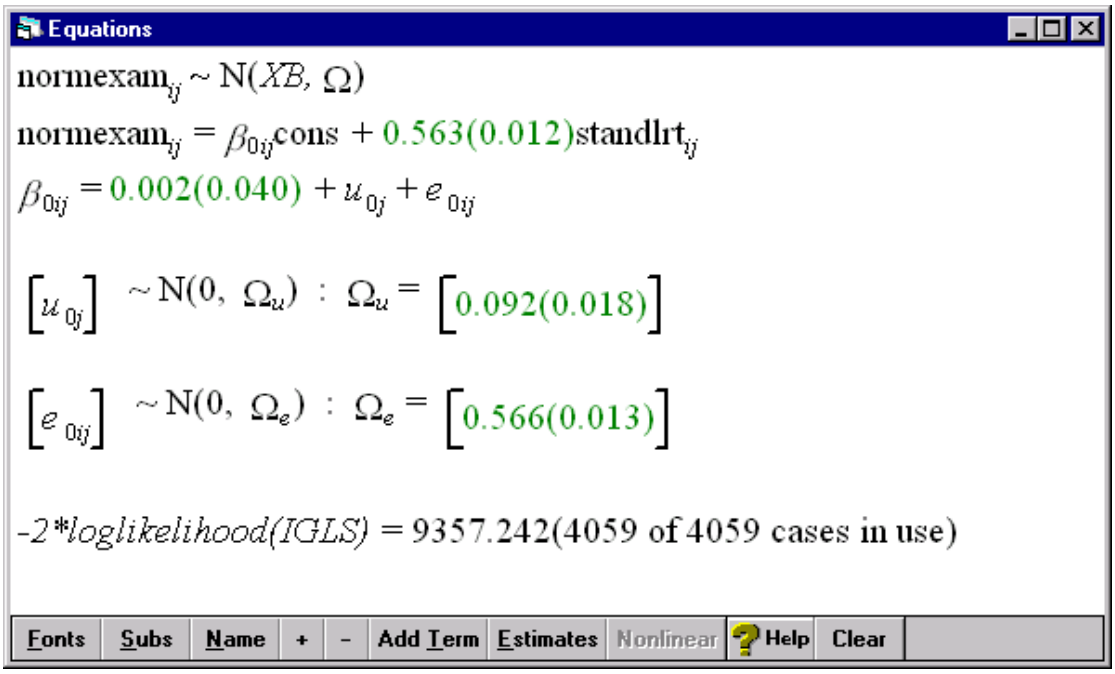
The estimation process is iterative. To begin the estimation we use the tool bar of the main *MLwiN* window. The **Start** button starts estimation, the **Stop** button stops it, and the **More** button resumes estimation after a stop. The default method of estimation is iterative generalised least squares (IGLS). This is noted on the right of the **Stop** button, and it is the method we shall use. The **Estimation control** button is used to vary the method, to specify convergence criteria, and so on. See the Help system for further details.

Click **Start**

You will now see the progress gauges at the bottom of the screen (R for random parameters and F for fixed parameters) fill up with green as the estimation proceeds alternately for the random and fixed parts of the model. In the present case this is completed at iteration 3 at which point the blue highlighted parameters in the **Equations** window change to green to indicate convergence. Convergence is judged to have occurred when all the parameters between two iterations have changed by less than a given tolerance, which is 10^{-2} by default but can be changed from the **Options** menu.

Click **Estimates**

once more and you will see the parameter estimates displayed together with their standard errors as in the following screen (the last line of the screen can be ignored for the time being).



The first two lines of this display reproduce equations 1.8, with the actual names of the different variables filled in. Recall that our model amounts to fitting a set of parallel straight lines to the results from the different schools. The slopes of the lines are all the same, and the fitted value of the common slope is 0.563 with a standard error of 0.012 (clearly, this is highly significant). However, the intercepts of the lines vary. Their mean is 0.002 and this has a standard error (in brackets) of 0.040. Not surprisingly with Normalized data, this is close to zero. The intercepts for the different schools are the level 2 residuals u_{0j} and these are distributed around their mean with a variance shown on line 4 of the display as 0.092 (standard error 0.018). The variance appears to be significantly different from zero. Judging significance for variances however, (and assignng confidence intervals) is not as straightforward as for the fixed part parameters. The simple comparison with the standard error and also the use of the **interval and tests** procedures (see help system) provides approximations that can act as rough guides. We shall deal with this further when discussing the likelihood ratio statistic and also in the part of this guide which deals with simulation based techniques. Of course, the actual data points do not lie exactly on the straight lines; they vary about them with amounts given by the level 1 residuals e_{0ij} and these have a variance estimated as 0.566, standard error 0.013. We shall see in the next chapter how *MLwiN* enables us to estimate and plot the residuals in order to obtain a better understanding of the model.

If we were to take children at random from the whole population, their variance would be the sum of the level 2 and level 1 variances, 0.092 + 0.566 = 0.658. The between-school variance makes up a proportion 0.140 of this total variance. This quantity is known as the *intra-school correlation*. It measures the extent to which the scores of children in the same school resemble each other as compared with those from children at different schools.

The last line of the display contains a quantity known as twice the *log likelihood*. This will prove to be useful in comparing alternative models for the data and carrying out significance tests. It can be ignored for the time being.

This is another place where you would do well to save the worksheet.

Graphing Predictions : Variance components

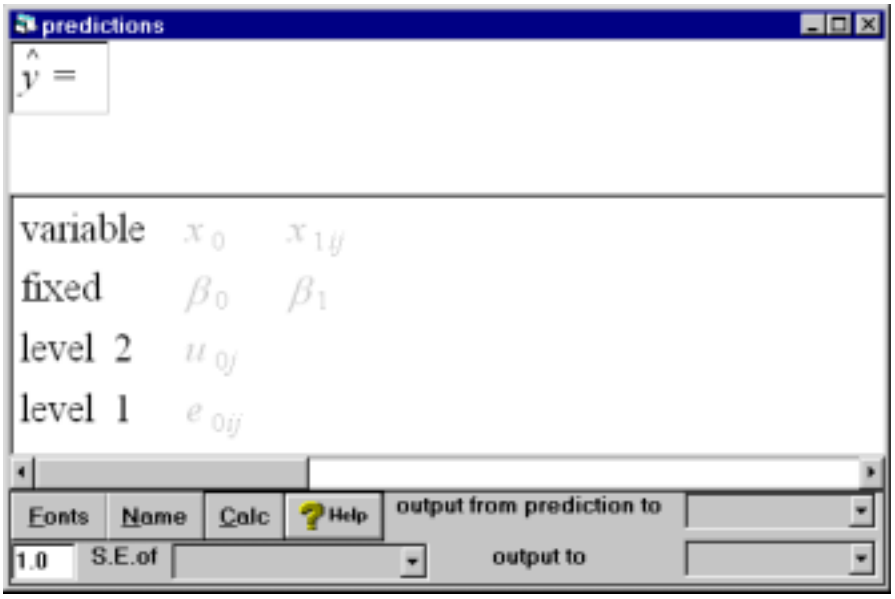
We have now constructed and fitted a *variance components* model in which schools vary only in their intercepts. It is a model of *simple* variation at level 2, which gives rise to the parallel lines illustrated in figure 1.2.

To demonstrate how the model parameters we have just estimated combine to produce the parallel lines of figure 1.2 we now introduce two new windows the **Predictions** window that can be used to calculate predictions from the model and the **Customised graphs** window which is a general purpose window for building graphs that can be used to graph our predicted values.

Lets start by calculating the average predicted line produced from the fixed part intercept and slope coefficients(β_0, β_1).

Select the **Model** menu Select **Predictions**

Which produces :

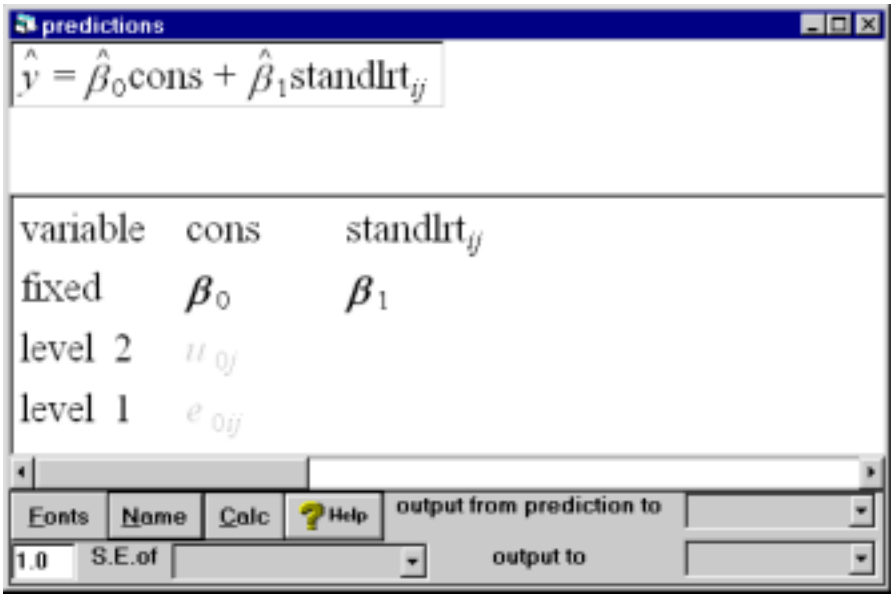


The elements of the model are arranged in two columns, one for each explanatory variable. Initially these columns are 'greyed out'. You build up a prediction equation in the top section of the window by selecting the elements you want from the lower section. Clicking on the variable name at the head of a column selects all the elements in that column. Clicking on an already-selected element deselects it.

Select suitable elements to produce the desired equation :

Click on β_0
Click on β_1
Click on **Names**

The **prediction** window should now look like this :



The only estimates used in this equation are $\hat{\beta}_0$ and $\hat{\beta}_1$, the fixed parameters – no random quantities have been included.

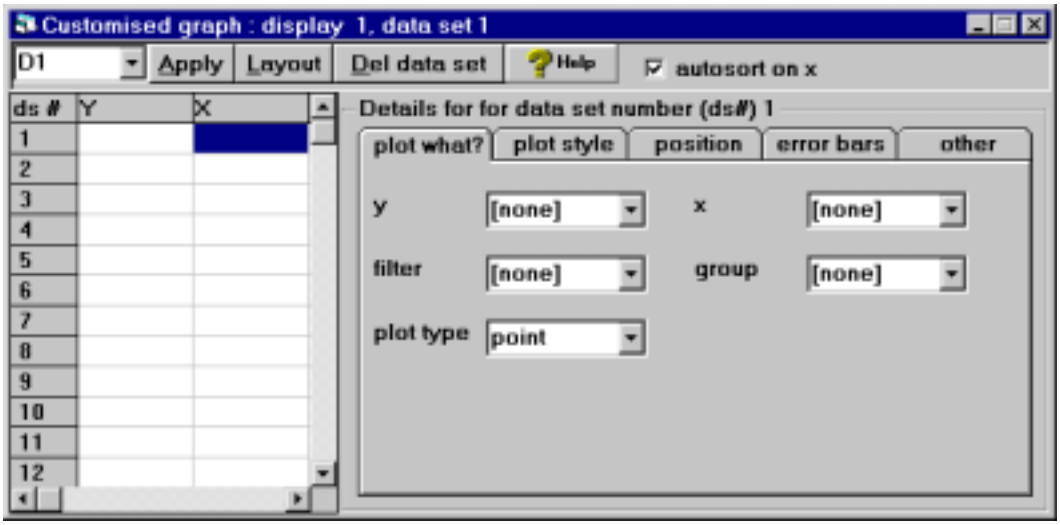
We need to specify where the output from the prediction is to go and then execute the prediction

In the **output from prediction to** drop-down list, select **C11** Click **Calc**

We now want to graph the predictions in column 11 against our predictor variable **standlrt**. We can do this using the **customised graph** window.

Select the **Graphs** menu Select **customised graph(s)**

This produces the following window :



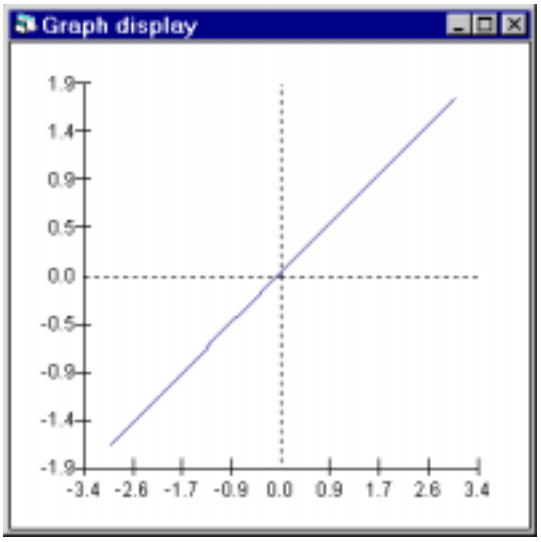
This general purpose graphing window has a great deal of functionality, which is described in more detail both in the help system and in the next chapter of this guide. For the moment we will confine ourselves to its more basic functions. To plot out the data set of predicted values :

In the drop down list labeled **y** in the **plot what ?** tab select **c11**
In the neighbouring drop down list labeled **x** select **standlrt**
In the drop down list labeled **plot type** select **line**
In the drop down list labeled **group** select **school**

This last action specifies that the plot will produce one line for each school. For the present graph all the lines will coincide, but we shall need this facility when we update our predictions to produce the school level summary lines. To see the graph :

Click the **Apply** button

The following graph will appear :



We are now going to focus on the **predictions** window and the **graph display** window.

Close the **Equations** Close the **Customised graph** window

Arrange the **predictions** and **graph display** windows so that they are both visible. If by mistake you click on the interior area of the **graph display** window, a window offering advanced options will appear; if this happens just close the advanced options window; we will be dealing with this feature in the next chapter.

The line for the *j*'th school departs from the above average prediction line by an amount u_{0j}. The school level residual u_{0j} modifies the intercept term, but the slope coefficient β_1 is fixed. Thus all the predicted lines for all 65 schools must be parallel. To include the estimated school level intercept residuals in the prediction function :

Select the **predictions** window click on the term u_{0j}

The prediction equation in the top part of the **predictions** window changes from

$$\hat{y} = \hat{\beta}_0 \text{cons} + \hat{\beta}_1 \text{standlrt}_{ij}$$

to

$$\hat{y} = \hat{\beta}_{0j}\text{cons} + \hat{\beta}_1\text{standlrt}_{ij}$$

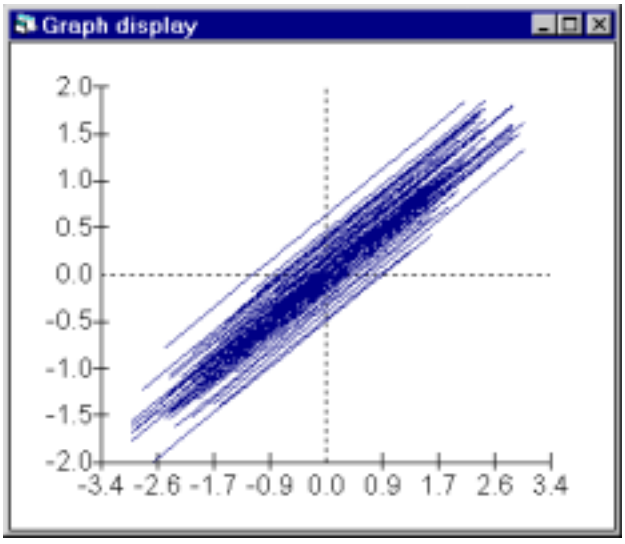
The crucial difference is that the estimate of the intercept $\hat{\beta}_0$ now has a *j* subscript. This subscript indicates that instead of having a single intercept, we have an intercept for each school, which is formed by taking the fixed estimate and adding the estimated residual for school *j*

$$\hat{\beta}_{0j} = \hat{\beta}_0 + \hat{u}_{0j}$$

We therefore have a regression equation for each school which when applied to the data produce 65 parallel lines. To overwrite the previous prediction in column 11 with the parallel lines

Press the **Calc** button in the prediction window

The graph display window is automatically updated with the new values in column 11 to show the 65 parallel lines.

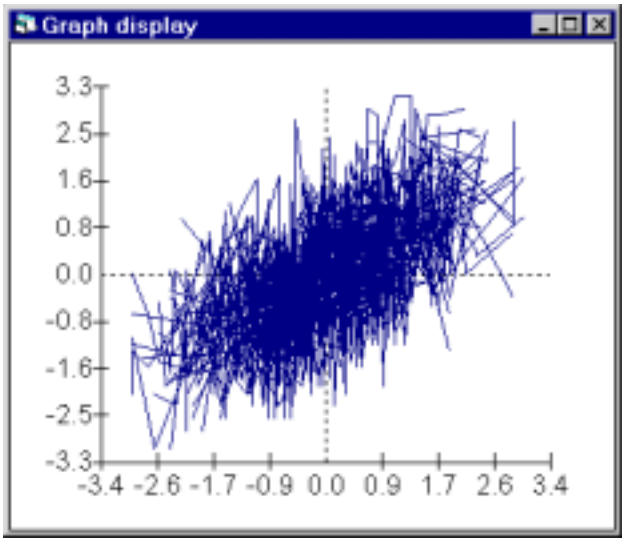


In this plot we have used the school level residuals(u_{0j}). Residuals and their estimation are dealt with in more detail in the next chapter.

Student *i* in school *j* departs from the school *j* summary line by an amount e_{0ij}. Recalculate the predictions to include e_{0ij} as well as u_{0j} as follows

Click on e_{0ij}

Press the **Calc** button



Which is a line plot through the original values of y_{ij}, i.e. we have predicted back onto the original data. Experiment including different combinations of $(\beta_0, \beta_1, u_{0j}, e_{0ij})$ in the prediction equation. Before pressing the **calc** button try and work out what pattern you expect to see in the graph window.

A Random slopes model

The *variance components* model which we have just specified and estimated assumes that the only variation between schools is in their intercepts. We should allow for the possibility that the school lines have different slopes as in Figure 1.3. This implies that the coefficient of standlrt will vary from school to school.

Still regarding the sample schools as a random sample from a population of schools, we wish to specify a coefficient of standlrt which is random at level 2. To do this we need to inform *MLwiN* that the coefficient of x_{1ij}, or $\mathbf{standlrt}_{ij}$, should have the subscript *j* attached.

To do this

Select the model menu

Select the Equations window

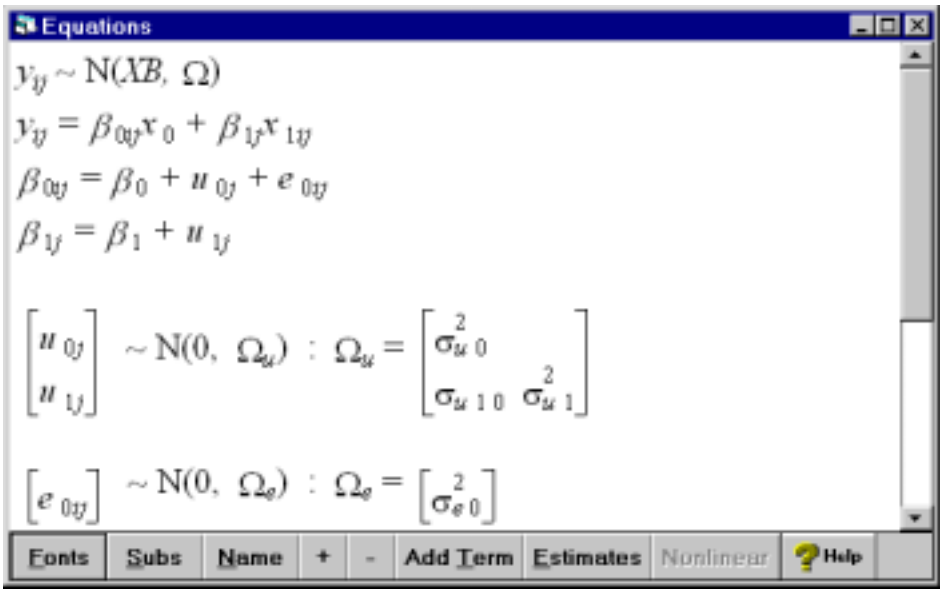
Click Estimates until β_0 etc. are displayed in black

Click β_1

Check the box labelled j(school)

Click **Done**

This produces the following result:



Now that the model is becoming more complex we can begin to explain the general notation. We have two explanatory variables x_0 and x_{1ij} (cons and standlrt). Anything containing a 0 subscript is associated with x_0 and anything containing a 1 subscript is associated with x_{1ij}. The letter *u* is used for random departures at level 2(in this case school). The letter *e* is used for random departures at level 1(in this case student).

The parameters β_0 and β_1 are the fixed part (regression coefficients) associated with x_0 and x_{1ij}. They combine to give the average line across all students in all schools.

The terms u_{0j} and u_{1j} are random departures or 'residuals' at the school level from β_0 and β_1. They allow the *j*'th school's summary line to differ from the average line in both its slope and its intercept.

The terms u_{0j} and u_{1j} follow a multivariate (in this case bivariate) Normal distribution with mean 0 and covariance matrix Ω_u. In this model we have two random variables at level 2 so Ω_u is a 2 by 2 covariance matrix. The elements of Ω_u are :

var(u_{0j}) = σ^2_{u0} (the variation across the schools' summary lines in their intercepts)

var(u_{1j}) = σ^2_{u1} (the variation across the schools' summary lines in their slopes)

cov(u_{0j}, u_{1j}) = σ_{u01} (the school level intercept/slope covariance).

Students' scores depart from their school's summary line by an amount e_{0ij}. (We associate the level 1 variation with x_0 because this corresponds to modelling constant or homogeneous variation of the student level departures. This requirement can be relaxed as we shall see later).

To fit this new model we could click Start as before, but it will probably be quicker to use the estimates we have already obtained as initial values for the iterative calculations. Therefore

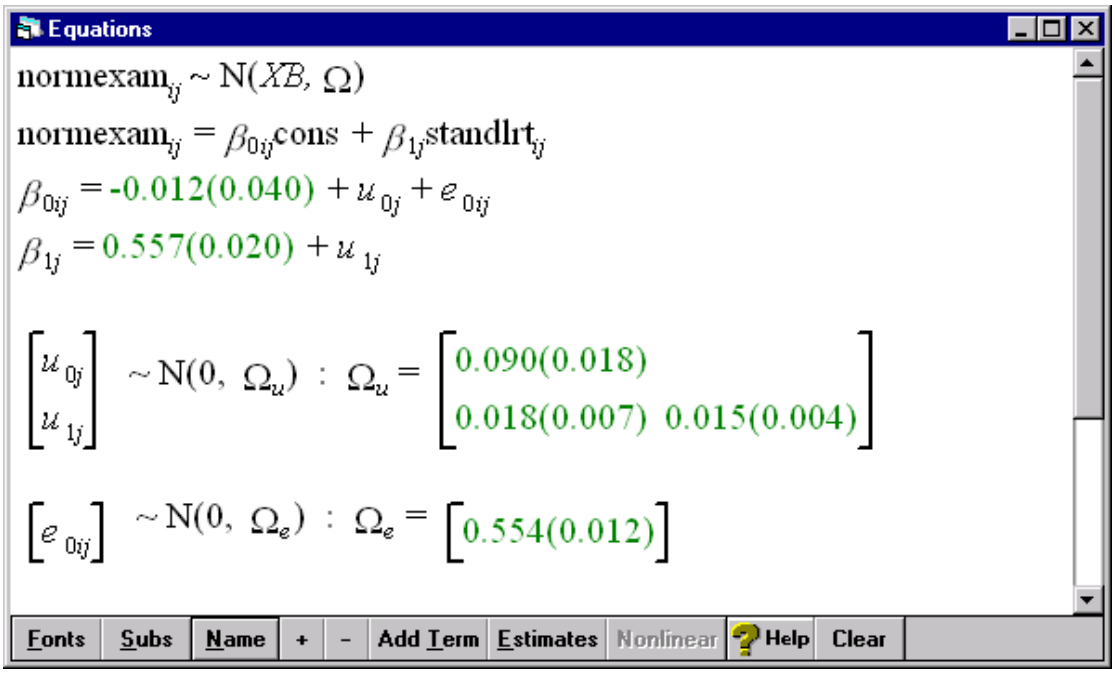
Click More

Convergence is achieved at iteration 7.

In order to see the estimates,

Click Estimates (twice if necessary) Click Names

To give



You should compare this display with that for the model where we did not fit a random slope. In line 2 of the display the coefficient of standlrt has acquired a suffix *j* indicating that it varies from school to school. In fact, its mean from line 4 is 0.557 (standard error 0.020), not far different from the model with a single slope. However, the individual school slopes vary about this mean with a variance estimated as 0.015 (standard error 0.004). The intercepts of the individual school lines also differ. Their mean is –0.012 (standard error 0.040) and their variance is 0.090 (standard error 0.018). In addition there is a positive covariance between intercepts and slopes estimated as +0.018 (standard error 0.007), suggesting that schools with higher intercepts tend to some extent to have steeper slopes and this corresponds to a correlation between the intercept and slope (across schools) of $0.018/\sqrt{0.015*0.090} = 0.49$. This will lead to a fanning out pattern when we plot the schools predicted lines.

As in the previous model the pupils' individual scores vary around their schools' lines by quantities e_{0ij}, the level 1 residuals, whose variance is estimated as 0.554 (standard error 0.012).

The quantity on the last line of the display, -2*log-likelihood can be used to make an overall comparison of this more complicated model with the previous one. You will see that it has decreased from 9357.2 to 9316.9, a difference of 40.3. The new model involves two extra parameters, the variance of the slope residuals u_{1j} and their covariance with the intercept residuals u_{0j} and the change (which is also the change in deviance, where the deviance for Normal models differs by a constant term for a fixed sample size) can be regarded as a χ^2 value with 2 degrees of freedom under the null hypothesis that the extra parameters have population values of zero. As such it is very highly significant, confirming the better fit of the more elaborate model to the data.

Graphing predictions : random slopes

We can look at pattern of the schools summary lines by updating the predictions in the graph display window. We need to form the prediction equation

$$\hat{y} = \hat{\beta}_{0j} x_0 + \hat{\beta}_{1j} x_{1ij}$$

One way to do this is

Select the **Model** menu

Select **Predictions**

In the predictions window click on the words Explanatory variables

From the menu that appears choose **Include all explanatory variables**

Click on e_{oij} to remove it from the prediction equation

In the **output from prediction to** drop-down list, select **c11**

Click **Calc**

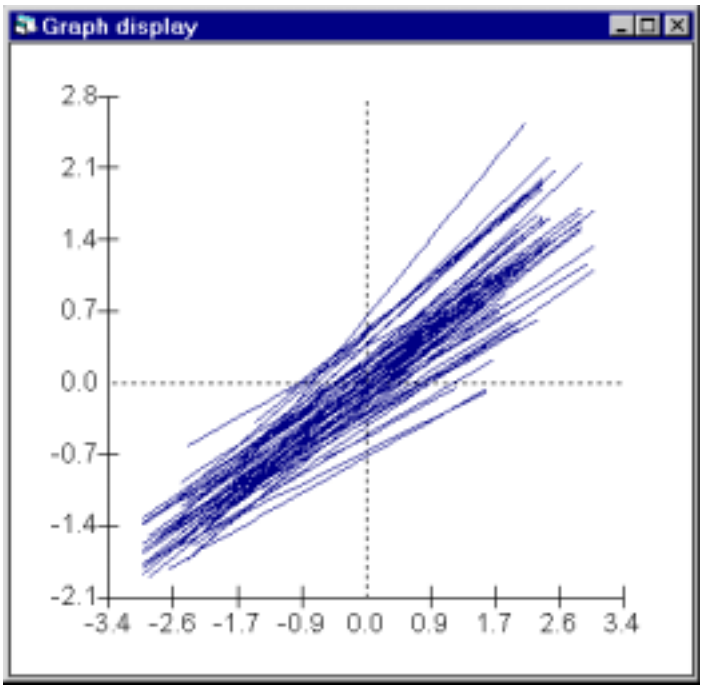
This will overwrite the previous predictions from the random intercepts model with the predictions from the random slopes model. The graph window will be automatically updated. If you do not have the graph window displayed, then

Select the **Graphs** menu

Select **customised graphs**

Click **Apply**

The graph display window should look like this :



The graph shows the fanning out pattern for the school prediction lines that is implied by the positive intercept/slope covariance at the school level.

To test your understanding try building different prediction equations in the **predictions** window; before you press the **calc** button try and work out how the graph in the **graph display** window will change.

That concludes the second chapter. It is a good idea to save your worksheet using the **save** option on the **File** menu.

What you should have learnt from this chapter

You should understand :

- What a random intercept model is
- What a random slope model is
- The equations used to describe these models
- How to construct, estimate and interpret these models using the equations window in *MLwiN*

- How to carry out simple tests of significance

- How to use the predictions window to calculate predictions from the model estimates

Chapter 2: Residuals

In this chapter we will work through the random slope model again. This time we shall explore the school and student random departures known as *residuals*.

Before we begin let's close any open windows :

Select the **Window** menu

Select **close all windows**

What are multilevel residuals?

In order to answer that question let's return to the random intercepts model. You can retrieve one of the earlier saved worksheets, or you can modify the random slopes model -

Select the **model** menu

Select **Equations**

Click on β_1

Uncheck the box labeled **j(school)**

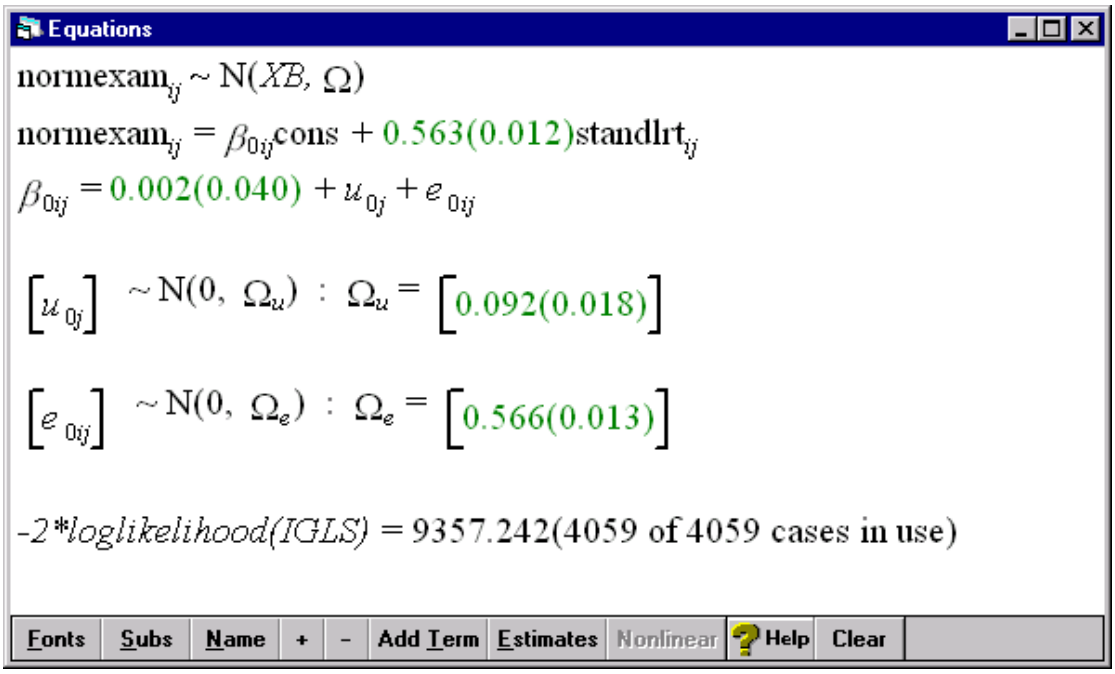
Click **Done**

The slope coefficient is now fixed with no random component. Now run the model and view the estimates:

Press **Start** on the main toolbar

Press **Name** then **Estimates** twice in the **Equations** window

Which produces :



This should be familiar from the previous chapter. The current model is a 2-level linear regression relationship of **normexam** on **standlrt**, with an average line defined by the two fixed coefficients β_0 and β_1. The model is made two-level by allowing the line for the *j*th school to be raised or lowered from the average line by an amount u_{0j}. These departures from the average line are known as the level 2 residuals. Their mean is zero and their estimated variance of 0.092 is shown in the **Equations** window. With educational data of the kind we are analyzing, they might be called the school effects. In other datasets, the level 2 residuals might be hospital, household or area effects.

The true values of the level 2 residuals are unknown, but we will often require to obtain estimates of them. We might reasonably ask for the effect on student attainment of one particular school. We can in fact predict the values of the residuals given the observed data and the estimated parameters of the model (see Goldstein, 1995, Appendix 2.2). In ordinary multiple regression, we can estimate the residuals simply by subtracting the predictions for each individual from the observed values. In multilevel models with residuals at each of several levels, a more complex procedure is needed.

Suppose that y_{ij} is the observed value for the *i*th student in the *j*th school and that $\hat{y}_{ij}$ is the predicted value from the average regression line. Then the *raw residual* for this subject is $r_{ij} = y_{ij} - \hat{y}_{ij}$. The raw residual for the *j*th school is the mean of these over the students in the school. Write this as r_{+j}. Then the predicted level 2 residual for this school is obtained by multiplying r_{+j}. by a factor as follows –

$$\hat{u}_{0j} = \frac{\sigma^2{}_{u0}}{\sigma^2{}_{uo} + \sigma^2{}_{e0} / n_j} r_{+j}$$

where n_j is the number of students in this school.

The multiplier in the above formula is always less than or equal to 1 so that the estimated residual is usually less in magnitude than the raw residual. We say that the raw residual has been multiplied by a *shrinkage factor* and the estimated residual is sometimes called a shrunken residual. The shrinkage factor will be noticeably less than 1 when σ^2_{e0} is large compared to σ^2_{u0} or when n_j is small (or both). In either case we have relatively little information about the school (its students are very variable or few in number) and the raw residual is pulled in towards zero. In future 'residual' will mean shrunken residual. Note that we can now estimate the level 1 residuals simply by the formula

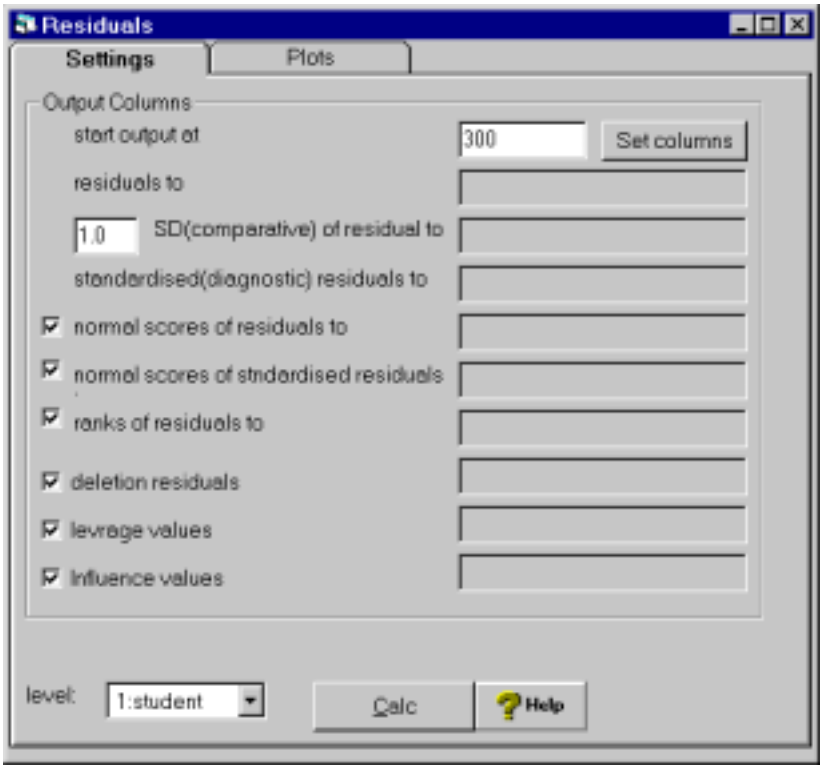
$$\hat{e}_{0ij} = r_{ij} - \hat{u}_{0j}$$

MLwiN is capable of calculating residuals at any level and of providing standard errors for them. These can be used for comparing higher level units (such as schools) and for model checking and diagnosis.

Calculating residuals in *MLwiN*

We can use the **Residuals** window in *MLwiN* to calculate residuals. Let's take a look at the level 2 residuals in our model.

Select **Model** menu
Select **Residuals**
Select **Settings** tab



The comparative standard deviation (SD) of the residual is defined as the standard deviation of $u_{0j} - \hat{u}_{0j}$ and is used for making inferences about the unknown underlying value u_{0j}, given the estimate $\hat{u}_{0j}$. The standardised residual is defined as $\hat{u}_{0j} / SD(\hat{u}_{0j})$ and is used for diagnostic plotting to ascertain Normality etc.

As you will see, this window permits the calculation of the residuals and of several functions of them. We need level 2 residuals, so at the bottom of the window

From the **level:** list **s**elect **2:school**

You also need to specify the columns into which the computed values of the functions will be placed.

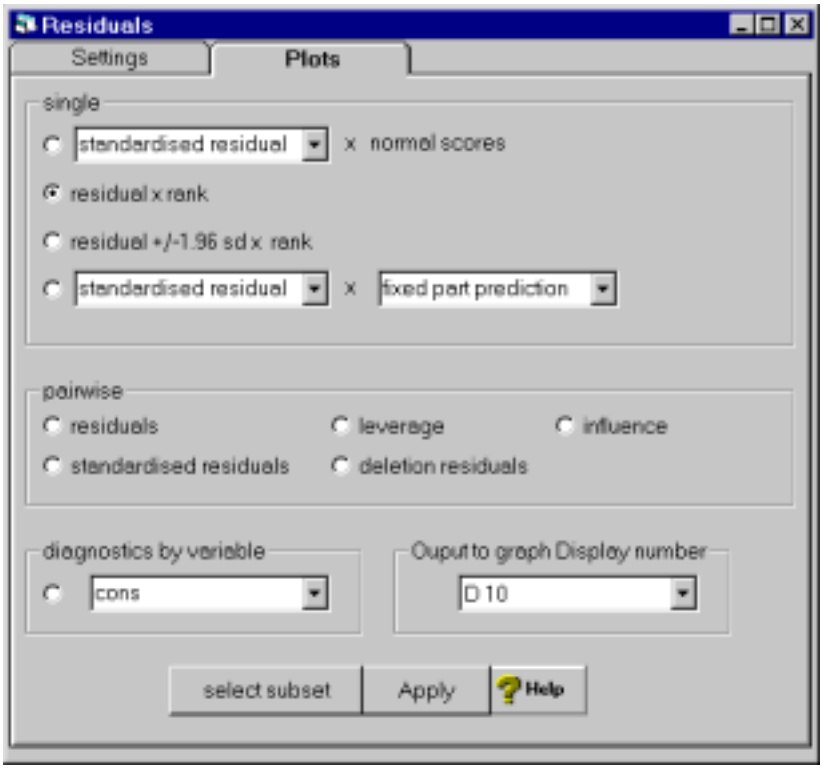
Click the **Set columns** button

The nine boxes beneath this button are now filled in grey with column numbers running sequentially from C300. These columns are suitable for our purposes, but you can change the starting column by editing the **start output at** box. You can also change the multiplier to be applied to the standard deviations, which by default will be stored in C301.

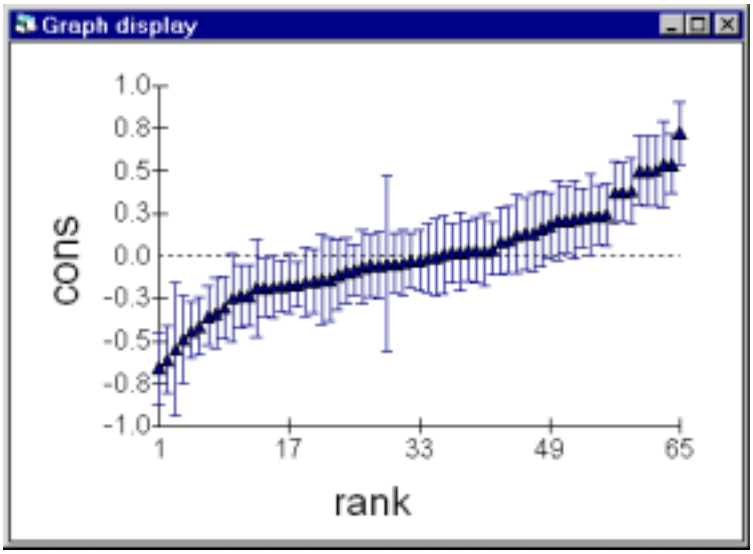
Edit the SD multiplier to 1.96

Click **Calc(**to calculate columns C300 to C308.

Having calculated the school residuals, we need to inspect them and *MLwiN* provides a variety of graphical displays for this purpose. The most useful of these are available from the **Residuals** window by clicking on the **Plots** tab. This brings up the following window –



One useful display plots the residuals in ascending order with their 95% confidence limit. To obtain this, click on the third option in the **single** frame (residual +/- 1.96 SD x rank) then click **Apply**. The following graph appears



This is sometimes known (for obvious reasons) as a caterpillar plot. We have 65 level 2 residuals plotted, one for each school in the data set. Looking at the confidence intervals around them, we can see a group of 10 or 15 schools at each end of the plot where the confidence intervals for their residuals do not overlap zero. Remembering that these residuals represent school departures from the overall average line predicted by the fixed parameters, this means that the majority of the schools do not differ significantly from the average line at the 5% level.

See Goldstein and Healy (1995) for further discussion on how to interpret and modify such plots when multiple comparisons among level 2 units are to be made. Comparisons such as these, especially of schools or hospitals, raise difficult issues: in many applications, such as here, there are large standard errors attached to the estimates. Goldstein and Spiegelhalter (1996) discuss this and related issues in detail.

Note: You may find that you sometimes need to resize graphs in *MLwiN* to obtain a clear labeling of axes.

What you should have learnt from this chapter

- Multilevel residuals are shrunken towards zero and shrinkage increases as n_j decreases

- How to calculate residuals in *MLwiN*

Chapter 3. Graphical procedures for exploring the model

Displaying graphs

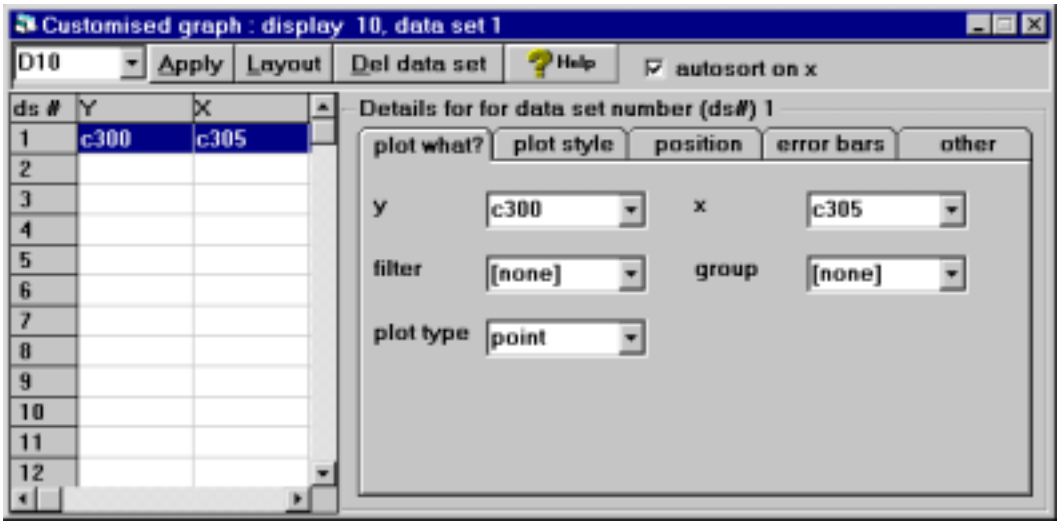
We have already produced a graphical display of the school level residuals in our random intercept model, using the **Residuals** window to specify what we wanted. *MLwiN* has very powerful graphical facilities, and in this chapter we shall see how to obtain more sophisticated graphs using the **Customised graphs** window. We will also use some of these graphical features to explore the random intercepts and random slopes models.

Graphical output in *MLwiN* can be described (very appropriately) at three levels. At the highest level, a *display* is essentially what can be displayed on the computer screen at one time. You can specify up to 10 different displays and switch between them as you require. A display can consist of several *graphs*. A graph is a frame with *x* and *y* axes showing lines, points or bars, and each display can show an array of up to 5x5 graphs. A single graph can plot one or more *datasets*, each one consisting of a set of *x* and *y* coordinates held in worksheet columns.

To see how this works,

Select the **graphs** menu Select **customised graphs**

The following window appears :



This screen is currently showing the construction details for display D10 – you may have noticed that the **plot** tab of the **Residuals** window in the previous chapter specified this in its bottom right hand corner. The display so far contains a single graph, and this in turn contains a single dataset, ds1 for which the *y* and *x* coordinates are in columns c300 and c305 respectively. As you can check from the **Residuals** window, these contain the level 2 residuals and their ranks.

Let us add a second graph to this display containing a scatterplot of *normexam* against *standlrt* for the whole of the data. First we need to specify this as a second dataset.

Select data set number 2(**ds #2**) by clicking on the row labeled **2** in the grid on the left hand side of the window

Now use the *y* and *x* dropdown lists on the **plot what?** tab to specify **normexam** and **standlrt** as the *y* and *x* variables in ds2.

Next we need to specify that this graph is to separate from that containing the caterpillar plot. To do this,

Click the **position** tab on the right hand side of the **customised graph** window

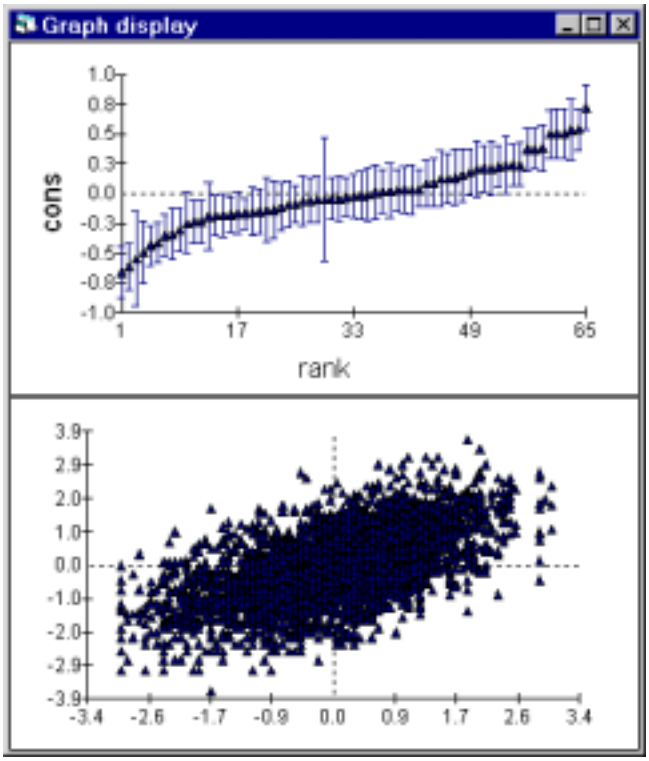
The display can contain a 5x5 grid or trellis of different graphs. The cross in the position grid indicates where the current data set, in this case (**normexam**, **standlrt**), will be plotted. The default position is row 1, column 1. We want the scatterplot to appear vertically below the caterpillar plot in row 2, column 1 of the trellis, so

Click the row 2 column 1 cell in the above grid

Now to see what we have got,

Press the **Apply** button at the top of the **Cutomised graph** window

and the following display will appear on the screen:



As a further illustration of the graphical facilities of *MLwiN*, let us create a third graph to show the 65 individual regression lines of the different schools and the average line from which they depart in a random manner. We can insert this between the two graphs that we already have. First we need to calculate the points for plotting in the new graph. For the individual lines

Select the **Model** window

Select **Predictions**

Click on **Variable**

Select **Include all explanatory variables**

Click on e_{0ij} to remove it

In the **output from prediction** list select c11

Press **calc**

This will form the predictions using the level 2 (school)residuals but not the level 1 (student) residuals. For the overall average line we need to eliminate the level 2 residuals, leaving only the fixed part of the model:

In the **Predictions** window click on u_{0j} to remove it

In the **output from prediction** list select c12

Press **calc**

Close the **Predictions** window

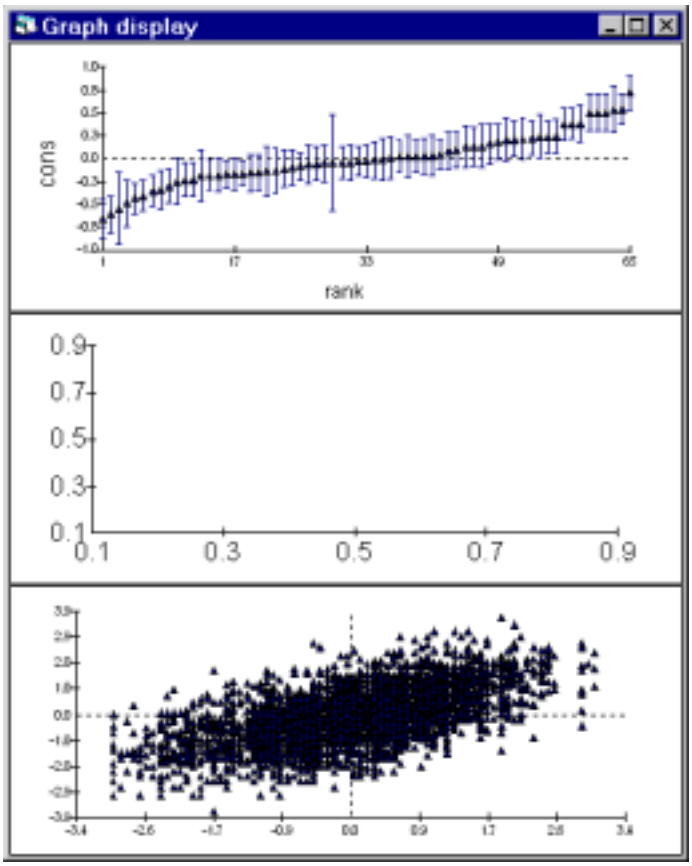
The **Customised graph** window is currently showing the details of dataset ds2, the scatterplot. With this dataset selected

Click on the **position** tab

In the grid click the cell in row 3, column 1

Press **Apply**

The display now appears as follows:



We have not yet specified any datasets for the middle graph so it is blank for the time being. Here and elsewhere you may need to resize and re-position the graph display window by pulling on its borders in the usual way.

Now let us plot the lines that we have calculated. We need to plot c11 and c12 against *standlrt*. For the individual school lines we shall need to specify the group, meaning that the 65 lines should be plotted separately. In the Customised graphs window

Select data set ds3 at the left of the window

In the **y** dropdown list specify c11

In the **x** dropdown list specify *standlrt*

In the **group** dropdown list select *school*

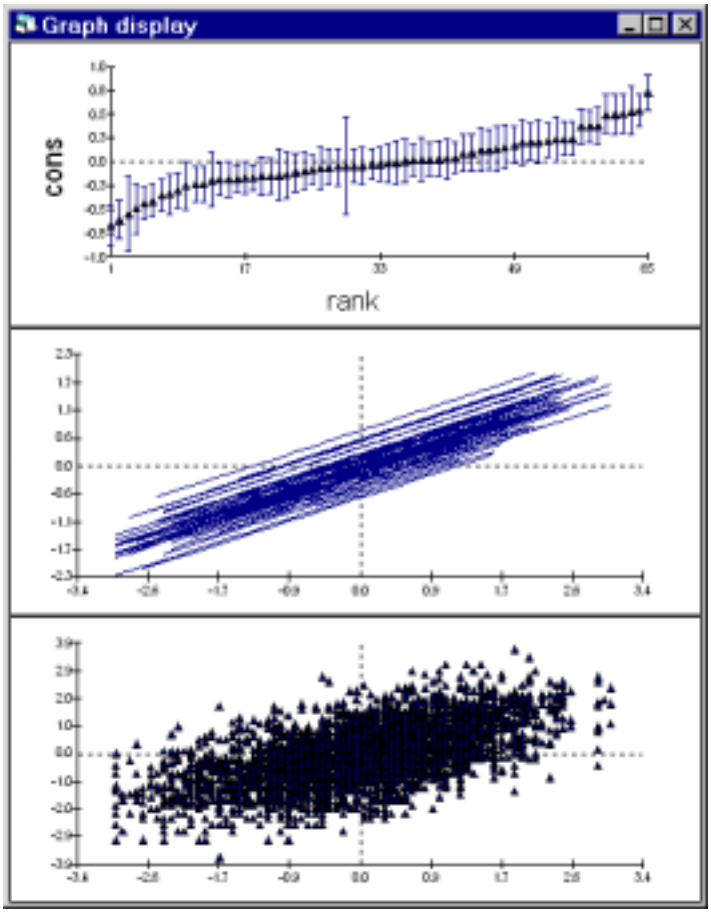
In the **plot type** dropdown list select *line*

Select the **position** tab

In the grid click the cell in row 2 column 1

Click **Apply**

This produces the following display:



Now we can superimpose the overall average line by specifying a second dataset for the middle graph. So that it will show up, we can plot it in red and make it thicker than the other lines:

Select dataset ds4 at the left hand side of the **Customised graphs** window

In the **y** dropdown list select c12

In the **x** dropdown list select *standlrt*

In the **plot type** dropdown list select *line*

Select the **plot styles** tab

In the **colour** dropdown list select *red*

In the **line thickness** dropdown list select 2

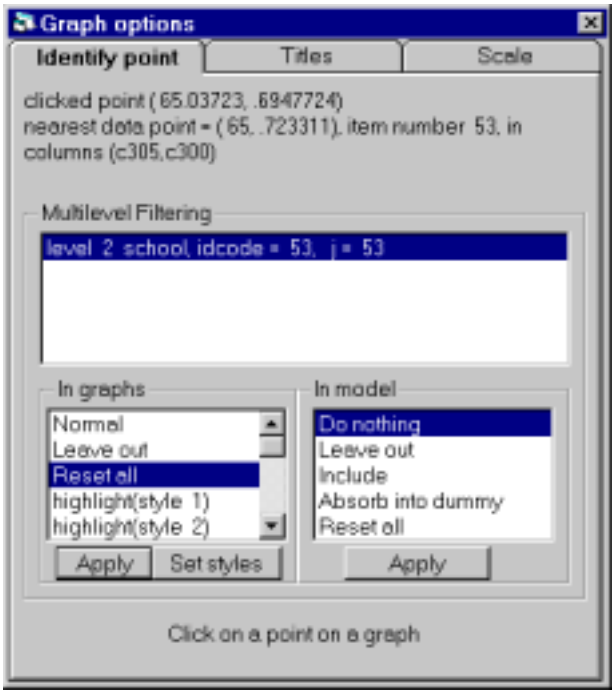
Select the **position** tab

In the grid click the cell in row 2, column 1

[I think the position should be OK following the previous manoeuvre]

Click **Apply**

There is a lot more that *MLwiN* makes it possible to do with the graphs that we have produced. To investigate some of this, click in the top graph on the point corresponding to the largest of the level 2 residuals, the one with rank 65. This brings up the following **Graph options** screen:



The box in the centre shows that we have selected the 53rd school out of the 65, whose identifier happens to be 53. We can *highlight* all the points in the display that belong to this school by selecting **highlight (style 1)** and clicking **Apply**. If you do this you will see that the appropriate point in the top graph, two lines in the middle graph and a set of points in the scatterplot have all become coloured red.

The individual school line is the the thinner of the two highlighted lines in the middle graph. As would be expected from the fact that it has the highest intercept residual, the school's line is at the top of the collection of school lines.

It is not necessary to highlight all references to school 53. To de-highlight the school's contribution to the overall average line which is contained in dataset ds4, in the **Customised graphs** window:

Select dataset 3

Click on the **other** tab

Click the Exclude from highlight box

Click **Apply**

In the caterpillar plot there is a residual around rank 30 which has very wide error bars. Let us try to see why. If you click on the point representing this school in the caterpillar plot, the **graph options** window will identify it as school 48. Highlight the points belonging to this school in a different colour:

Using the graph options window, in the in graphs box select highlight (style 2)

Click **Apply**

The points in the scatterplot belonging to this school will be highlighted in cyan, and inspection of the plot shows that there are only two of them. This means that there is very little information regarding this school. As a result, the confidence limits for its residual are very wide, and the residual itself will have been shrunk towards zero by an appreciable amount.

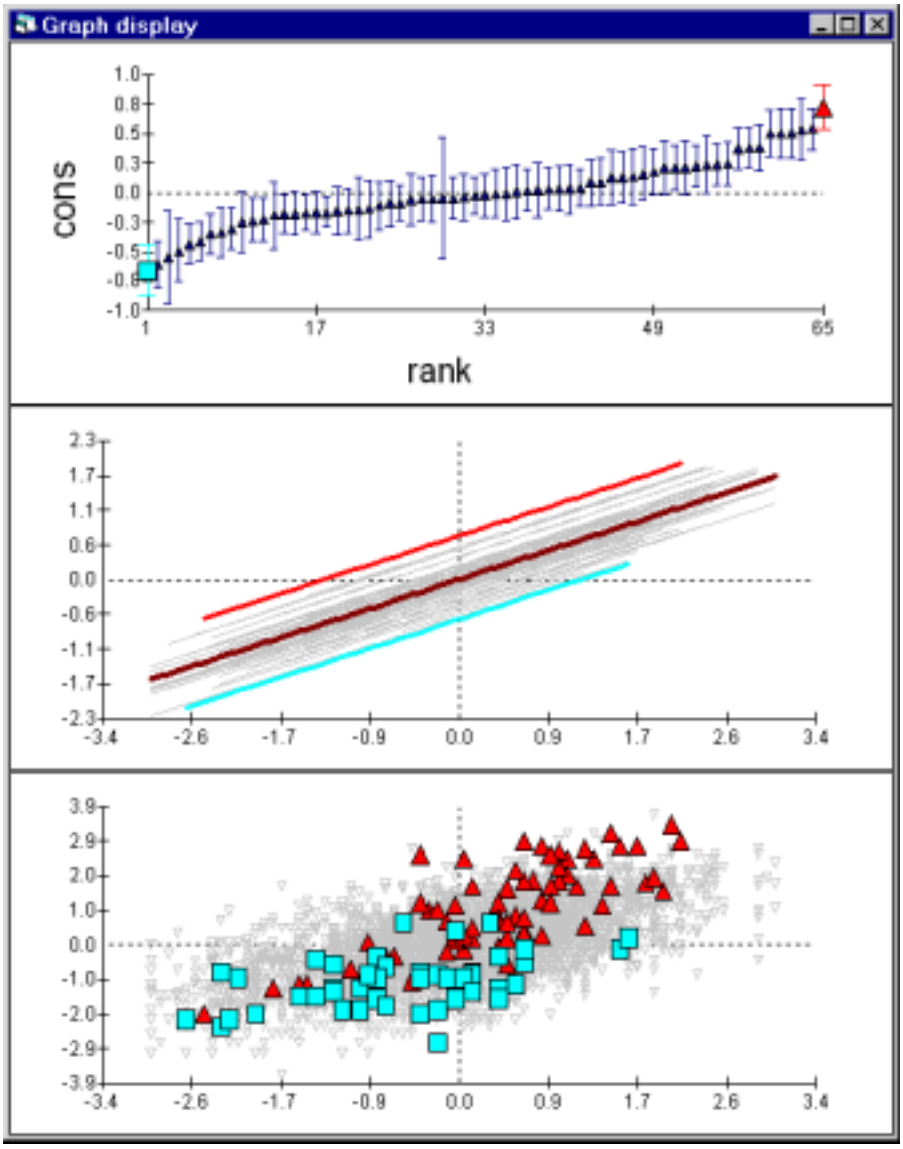
Next let us remove all the highlights from school 48. In the **graph options** window

In the **in graphs** box select **normal**

Click **Apply**

Now let us look at the school at the other end of the caterpillar, that with the lowest school level residual. Click on its point in the caterpillar (it turns out to be school 59) and in the **Graph options** window select **highlight (style 3)** and click **Apply**. The

highlighting will remain and the graphical display will look something like this, having regard to the limitations of monochrome reproduction:



The caterpillar tells us simply that school 49 and 53 have different intercepts – one is significantly below the average line, the other significantly above it. But the bottom graph suggests a more complicated situation. At higher levels of *standlrt*, the points for school 53 certainly appear to be consistently above those for school 49. But at the other end of the scale, at the left of the graph, there does not seem to be much difference between the schools. The graph indeed suggests that the two schools have different slopes, with school 53 the steeper.

To follow up this suggestion, let us keep the graphical display while we extend our model to contain random slopes. To do this:

From the **Model** menu select **Equation**

Click on β_1 and check the box labelled *j (school)* to make it random at level 2

Click **Done**

Click **More** on the main toolbar and watch for convergence

Close the **Equations** window

Now we need to update the predictions in column c11 to take account of the new model:

From the **Model** menu select **Predictions**

Click on u_{0j} and u_{1j} to include them in the predictions

In the **Output from predictions** dropdown list select c11

Click **Calc**

Notice that the graphical display is automatically updated with the new contents of column c11.

The caterpillar plot at the top of the display however is now out of date, having been calculated from the previous model. (Recall we used the residuals window to create the caterpillar plot). We now have two sets of level 2 residuals, one giving the intercepts for the different schools and one the slopes. To calculate and store these:

Select **Residuals** from the **Model** menu

Select **2:School** from the **level** dropdown list

Edit the **Start output at** box to 310

Click **Calc**

The intercept and slope residuals will be put into columns c310 and c311. To plot them

against each other:

In the **Customised graphs** window select dataset ds#1 and click **Delete dataset**

From the *y* dropdown list select c310

From the *x* dropdown list select c311

Click **Apply**

The axis titles in the top graph also need changing. Note that if you use the customised graph window to create graphs no titles are automatically put on the graphs. This is because a graph may contain many data sets so in general there is no obvious text for the titles. The existing titles appear because the graph was originally constructed by using the **plots** tab on the **residuals** window. You can specify or alter titles by clicking on a graph. In our case:

Click somewhere in the top graph to bring up the **Graph options** window

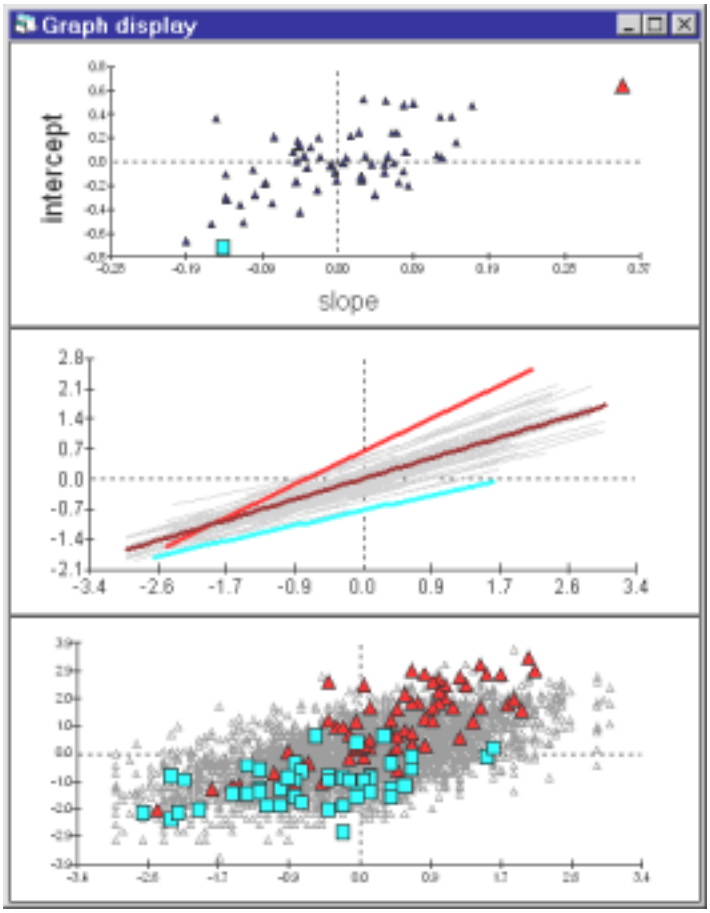
Select the **titles** tab

Edit the *y title* to be *Intercept*

Edit the *x title* to be *Slope*

Click **Apply**

You can add titles to the other graphs in the same way if you wish. Now the graphical display will look like this:


Graph display
intercept
slope
2.8
2.1
1.4
0.7
0.0
-0.7
-1.4
-2.1
-3.4
-2.6
-1.7
-0.9
0.0
0.9
1.7
2.6
3.4

The two schools at the opposite ends of the scale are still highlighted, and the middle graph confirms that there is very little difference between them at the lower levels of **standlrt**. School 53 stands out as exceptional in the top graph, with a high intercept and much higher slope than the other schools.

For a more detailed comparison between schools 53 and 49, we can put 95% confidence bands around their regression lines. To calculate the widths of the bands:

Select **Predictions** from the **Model** menu

Edit the multiplier of **S.E.** to 1.96

From the **S.E. of** dropdown list select *level 2 resid function*

From the **output to** dropdown list select column c13

Click **Apply**

Now plot the bands

In the **Customised graphs** window select dataset ds#2

Select the **errors** tab

From the **y error +** list se;ect c13

From the **y error –** list select c13

From the **y error type** list select *lines*

Click **Apply**

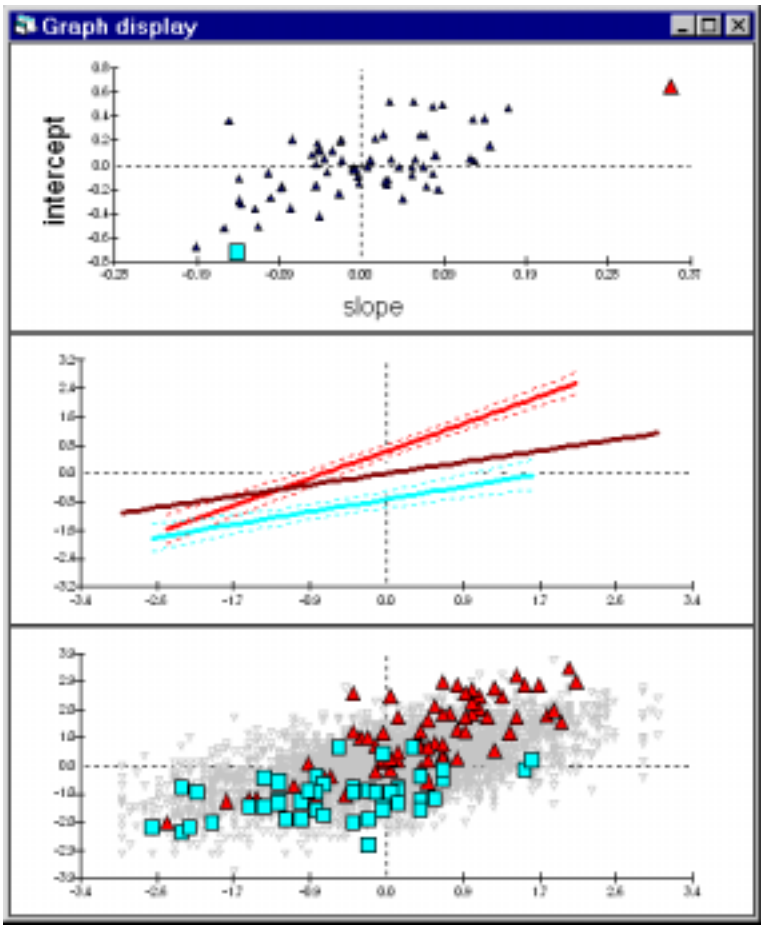
This draws 65 confidence bands around 65 school lines, which is not a particularly readable graph. However, we can focus in on the two highlighted schools by drawing the rest in white.

Select the **customised graphs** window

Select data set number 2 (**ds # 2**)

From the **colour** list select **white**

Click Apply



The confidence bands confirm that what appeared to be the top and bottom schools cannot be reliably separated at the lower end of the intake scale.

Looking at the intercepts and slopes may be able to shed light on interesting educational questions. For example, schools with high intercepts and low slopes, plotting in the top left quadrant of the top graph, are 'levelling up' – they are doing well by their students at all levels of initial ability. Schools with high slopes are differentiating between levels of intake ability. The highlighting and othere graphical features of *MLwiN* can be useful for exploring such features of complicated data. See Yang et al., (1999) for a further discussion of this educational issue.

What you should have learnt from this chapter

- How to make different graphical representations of complex data.
- How to explore aspects of multilevel data using graphical facilities such as highlighting.
- With random slopes models differences between higher level units(e.g. schools) can not be expressed by a single number.

Chapter 4: Contextual effects

Many interesting questions in social science are of the form how are individuals effected by their social contexts? For example,

- Do girls learn more effectively in a girls' school or a mixed sex school?

- Do low ability pupils fare better when they are educated alongside higher ability pupils or worse?

In this section we will develop models to investigate these two questions.

Before we go on let's close all open windows by

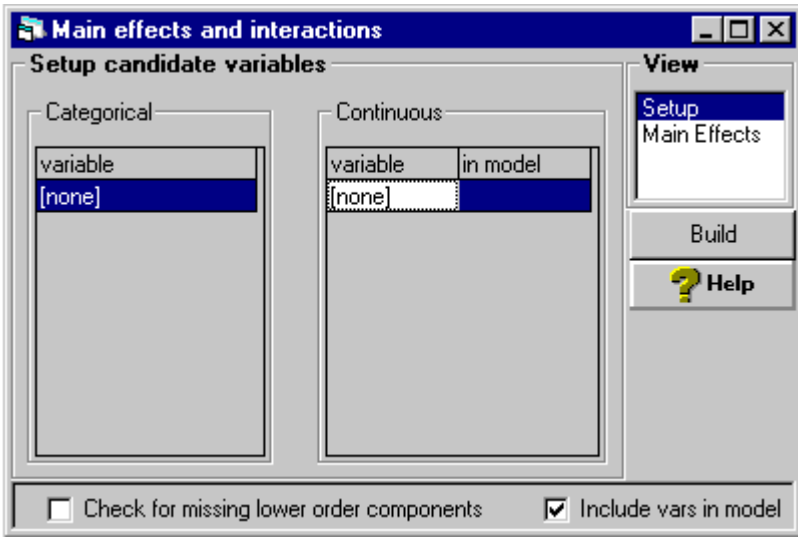
Select the **Window** menu Select **close all windows**

Pupil gender and school gender effects

We are now going use a new window which is useful for building models with categorical explanatory variables.

Select the **Model** menu Select **Main Effects and Interactions**

The following window appears



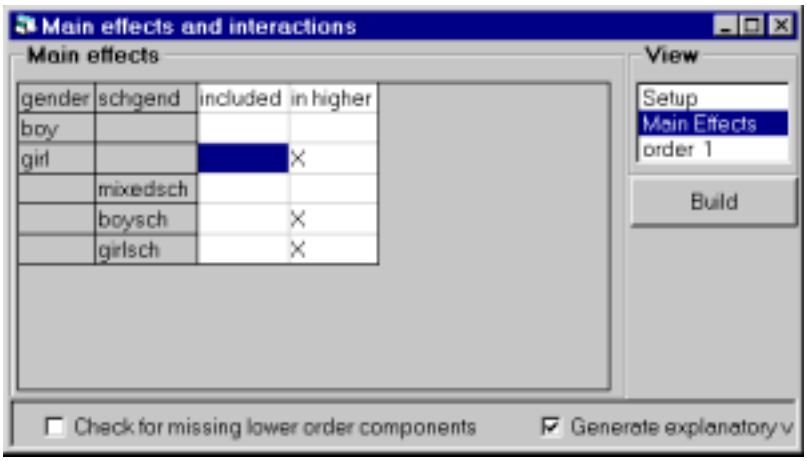
This screen automates the process of creating sets of dummy variables (and interactions between sets of dummy variables) that are required for modelling categorical predictors. To enter main effects for individual gender and school gender

In the panel marked **categorical** click on **[none]** From the list that appears select **gender** Click on **[none]** again and select **schgend**

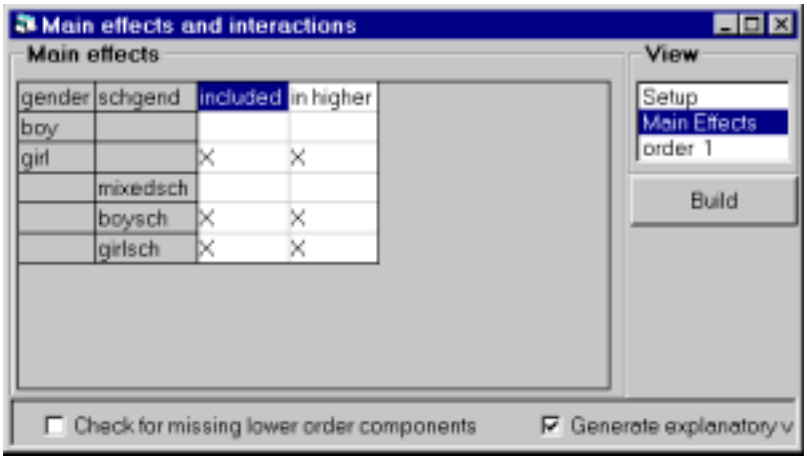
Note that now that we have defined two cataegorical variables the possiblility for a 1st order interaction exists and the **view** panel (on the right of the window) has been updated to include the option **order 1.** We just want to include main effects so

In the **View** panel click on **Main Effects**

The main effects and interactions window now displays a list of potential main effects:



At the moment no main effects are included, to fit gender and school gender with **boy** and **mixedsch** as the reference categories, click on the corresponding entries in the **included** column to produce the pattern :



The **in higher** column defines what categories are made available for higher order interactions. This is useful when you have large numbers of categorical variables and the number of possible combinations for higher order interactions is very large.

To add the main effects to the model :

Click **Build**

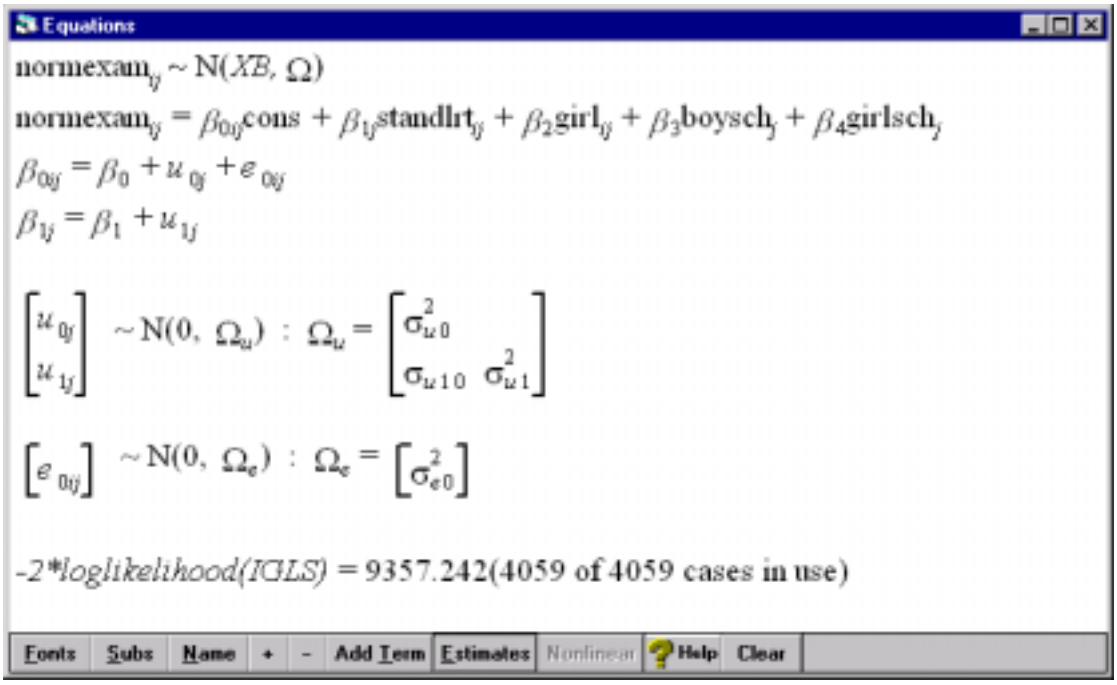
To view the model

Select the **Model** menu

Select **Equations**

Click **Names**

Which produces the following model

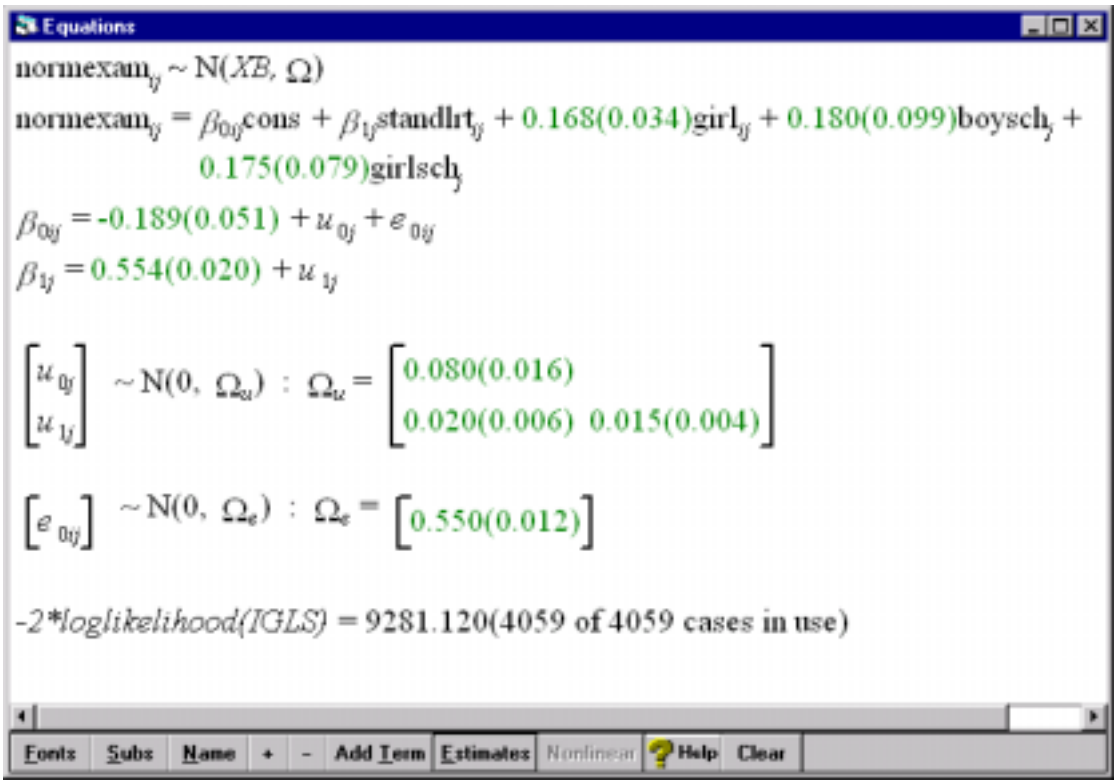


You can see that main effects for **girl, boysch** and **girlsch** have been added.**Girl** has subscript *ij* because it is a pupil level variable, whereas the two school level variables have subscript *j*. We can run the model and view the results by

Click **estimates** until numbers appear in the equations window

Press **More** on the main toolbar

The model converges to the results below:



The reference category is boys in a mixed school. Girls in a mixed school do 0.168 of a standard deviation better than boys in a mixed school. Girls in a girls school do 0.175 points better than girls in a mixed school and (0.175+0.168) points better than boys in a mixed school. Boys in a boys school do 0.18 points better than boys in a mixed school.

Adding these three parameters produced a reduction in the deviance of 35, which, under the null hypothesis of no effects, follows a chi-squred distribution with three degrees of freedom. You can look this probability up using the **Tail Areas** option on the **Basic Statistics** menu. The value is highly significant.

In the 2 by 3 table of gender by school gender there are two empty cells, there are no boys in a girls school and no girls in a boys school. We are currently using a reference group and three parameters to model a four entry table, therefore because of the empty cells the model is saturated and no higher order intercations can be added.

The pupil gender and school gender effects modify the intercept (**standlrt**=0). An interesting question is do these effects change across the intake spectrum. To address this we need to extend the model to include the interaction of the continuous variable **standlrt** with our categorical variables. To do this

Select the **Main Effects and Interactions** window

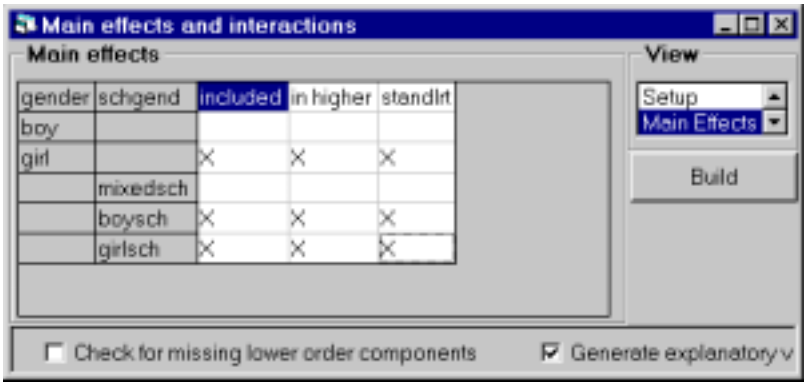
In the **View** panel select **setup**

In the **continous** panel click on **none**

From the list that appears select **standlrt**

In the **View** panel select **main effects**

The main effects screen now has a column for **standlrt** added. Click on entries in the column to produce the following pattern



Click on **Build**

The equations window will be automatically modified to include the thre new interaction terms. Run the model :

press **More** on the main toolbar

The deviance reduces by less than one unit. From this we conclude there is no evidence of an interaction between the gender variables and intake score. We can remove from the model by

Select the **main effects and interactions** window

Ensure **main effects** are selected in the **view** panel

Deselect all entries marked with a **X** in the standlrt column by clicking

Press **Build**

Note that we could have clicked on individual terms in **Equations** window and selected the **delete term** option. However, this would not have removed the terms from the main effects and interactions tables and every subsequent **build** would put them back into the model.

Contextual effects

The variable **schav** is contructed by taking the average intake ability(**standlrt**) for each school, based on these averages the bottom 25% of schools are coded 1(**low**), the middle 50 % coded 2(**mid**) and the top 25% coded 3(**high**). Let's include this categorical school level contextual variable in the model.

Select the **Main Effects and Interactions** window

Select S**etup** from the **View** panel

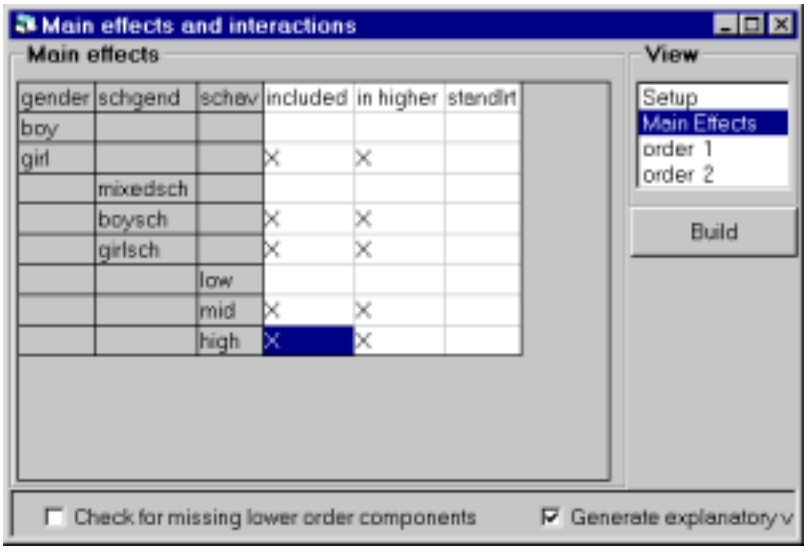
In the **Categorical** panel click on **[none]**

Select **schav** from the list that appears

Select **Main Effects** from the **View panel**

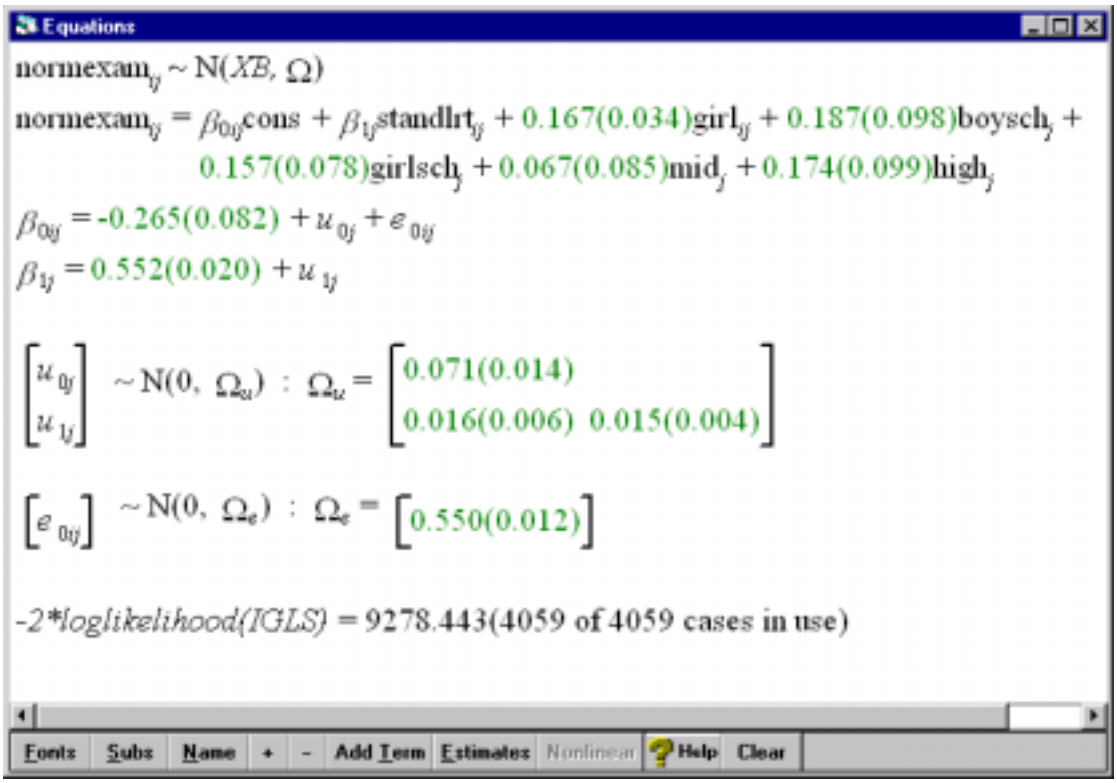
Click the **included** column for the **mid** and **high** entries

The main effects and interactions window should now look like this :



Click **Build**

Run the model by pressing **more** on the main toolbar



Children attending **mid** and **high** ability schools score 0.067 and 0.174 points more than children attending **low** ability schools. The effects are of borderline statistical significance. This model assumes the contextual effects of school ability are the same across the intake ability spectrum because these contextual effects are modifying the intercept term. That is the effect of being in a **high** ability school is the same for low ability and high ability pupils. To relax this assumption we need to include the interaction between **standlrt** and the school ability contextual variables. To do this :

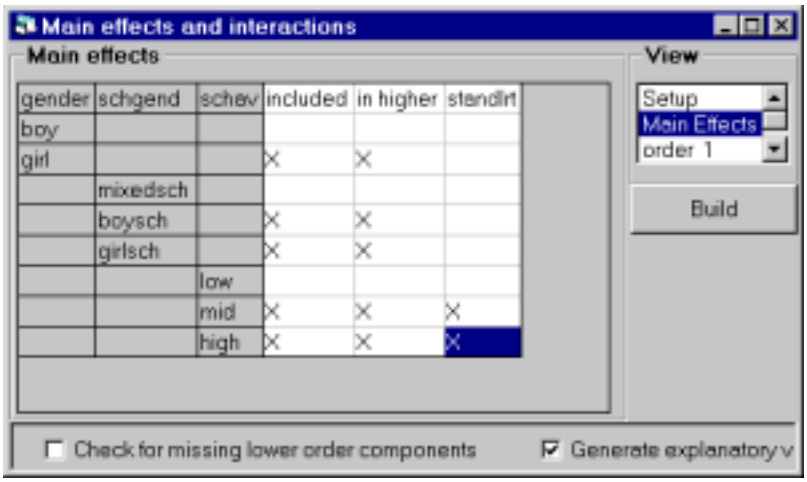
Select the **Main Effects and Interactions** window

Select **Main Effects** from the **View** panel

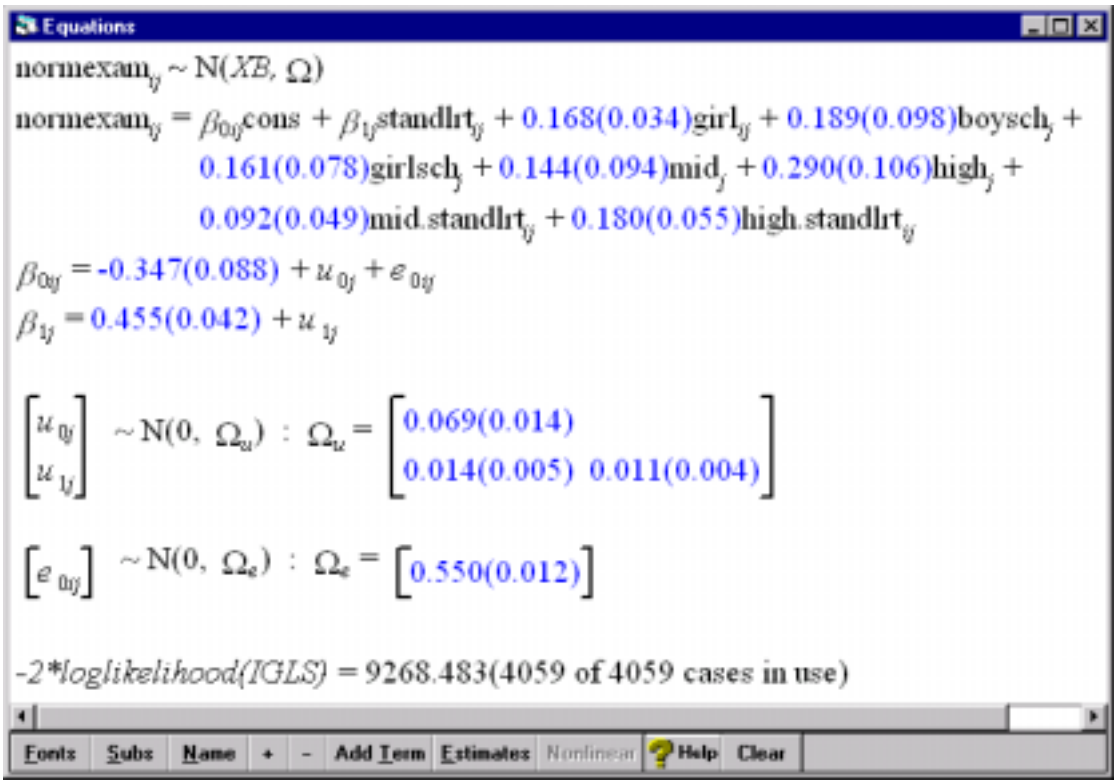
Click the **standlrt** column for the **mid** and **high** entries

Click **Build**

The **Main Effects and Interactions** window should look like this:



The model converges to :



The slope coefficient for **standlrt** for pupils from **low** intake ability schools is 0.455. For pupils from **mid** ability schools the slope is steeper 0.455+0.092 and for pupils from **high** ability schools the slope is steeper still 0.455+0.18. These two interaction terms have explained variabilty in the slope of **standlrt** in terms of a school level variable therefore the between school variability of the **standlrt** slope has been substantially reduced (from 0.015 to 0.011). Note that the previous contextual effects **boysch, girlsch, mid** and **high** all modified the intercept and therefore fitting these school level variables reduced the between school variabilty of the intercept.

We now have three different linear relationships between the output score(**normexam**) and the intake score(**standlrt**) for pupils from **low, mid** and **high** ability schools. The prediction line for **low** ability schools is

$$\hat{\beta}_0\text{cons} + \hat{\beta}_1\text{standlrt}_{ij}$$

The prediction line for the **high** ability schools is

$$\hat{\beta}_0\text{cons} + \hat{\beta}_1\text{standlrt}_{ij} + \hat{\beta}_6\text{high}_j + \hat{\beta}_8\text{high.standlrt}_{ij}$$

The difference between these two lines, that is the effect of being in a **high** ability school is

$\hat{\beta}_6 \text{high}_j + \hat{\beta}_8 \text{high.standlrt}_{ij}$

We can create this prediction function by

select the **Model** window

Select **Predictions**

Clear any existing prediction by clicking on **variable**

Select **Remove all explanatory variables** from the menu that appears

Click in turn on β_6, β_8

In the **output from prediction to** list select **c30**

Press Ctl-N and rename **C30** to **predab**

Click **Calc**

We can plot this function as follows :

Select the **Customised Graph** window

Select display number 5 **D5**

In the **y** list select **predtab**

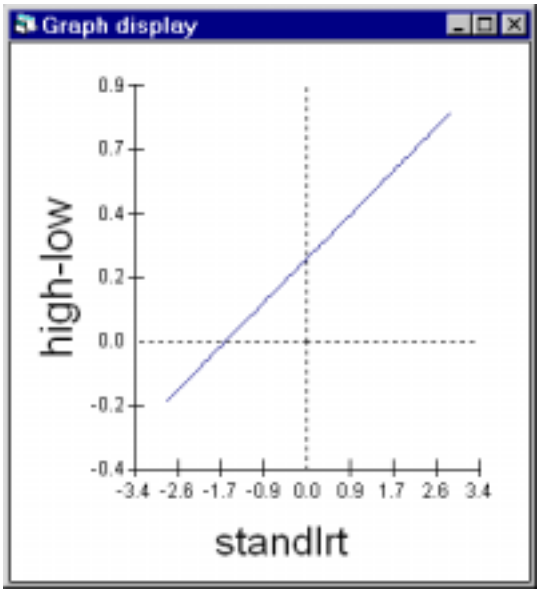
In the **x** list select **standlrt**

In the **plot type** list select **line**

In the **filter** list select **high**

Click **Apply**

Which produces



This graph shows how the effect of pupils being in a **high** ability school changes across the intake spectrum. On average very able pupils being educated in a **high** ability school score 0.9 of a standard deviation higher in their outcome score than they would if they were educated in a **low** ability school. Once a pupils' intake score drops below –1.7 then they fare progressively better in a **low** ability school. This finding has some educational interest but we do not pursue that here. We can put a 95% confidence band around this line by

Select the **Predictions** window

Edit the multiplier **S.E. of** to 1.96

In the **S.E. of** list select **Fixed**

In the corresponding **output to** list select **c31**

Click **Calc**

Select the **customised graph** window

Select **error bars** tab

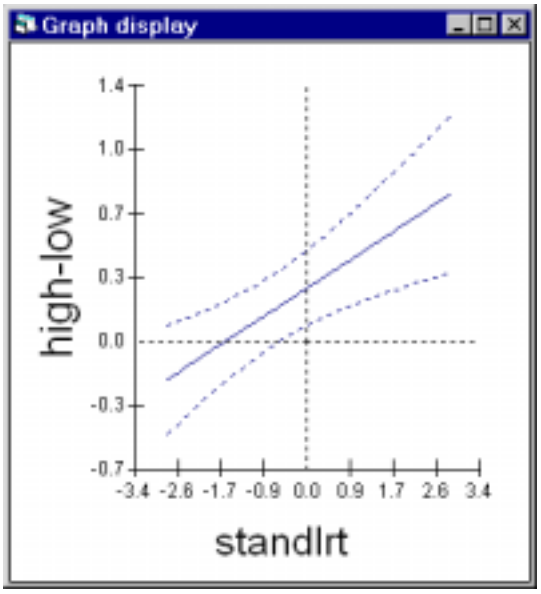
In the **y** errors + list select **c31**

In the **y** errors – list select c31

In the **y** error type list select **lines**

Click **Apply**

Which produces



Save your worksheet. We will be using it in the next chapter.

What you should have learnt from this chapter

- What is meant by contextual effects
- How to set up multilevel models with interaction terms

Chapter 5: Modelling the variance as a function of explanatory variables

From the fanning out pattern of the school summary lines seen in chapters 3 and 4 we know that schools are more variable for students with higher levels of **standlrt**. Another way of saying this is that the between school variance is a function of **standlrt**.

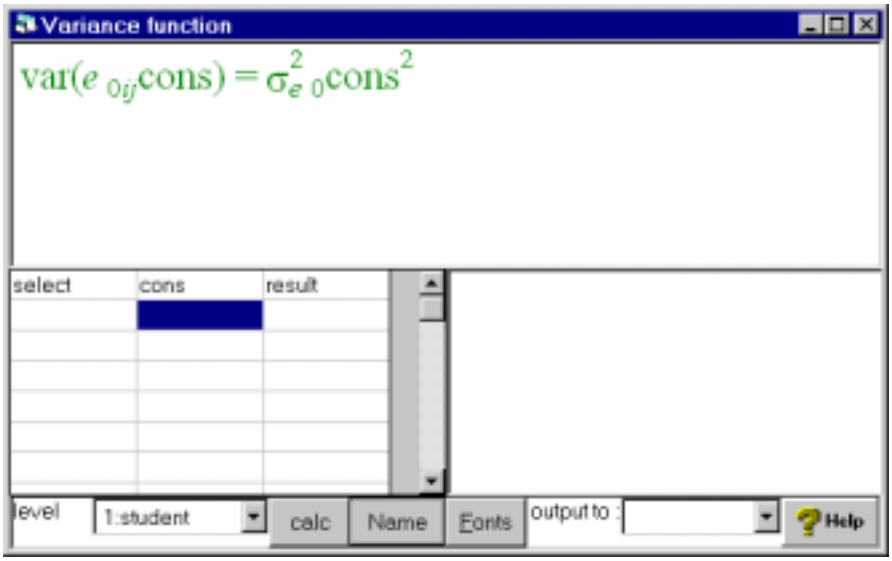
In *MLwiN* we always specify the random variation in terms of coefficients of explanatory variables, the total variance at each level is thus a function of these explanatory variables. These functions are displayed in the **Variance function** window.

Retreive the worksheet from the end of chapter 5

Select the **Model** menu

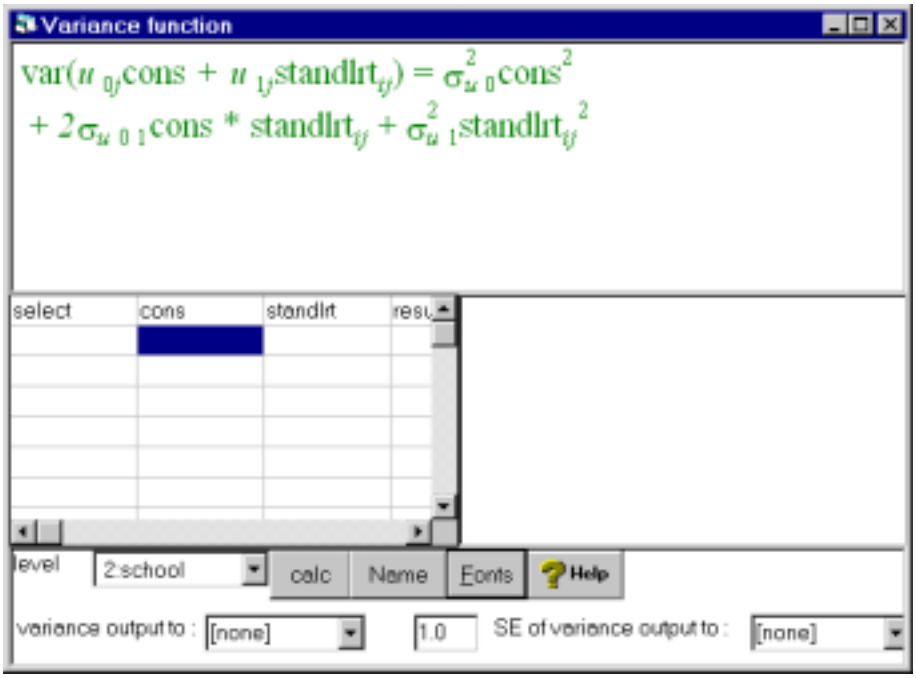
Select **Variance Function**

Click **Name** button in the V**ariance function** window



The initial display in this window is of the level 1 variance. In the present model we have simple (constant) variation at level 1, as the above equation shows. Now

In the **level** drop-down list, select **2:school**



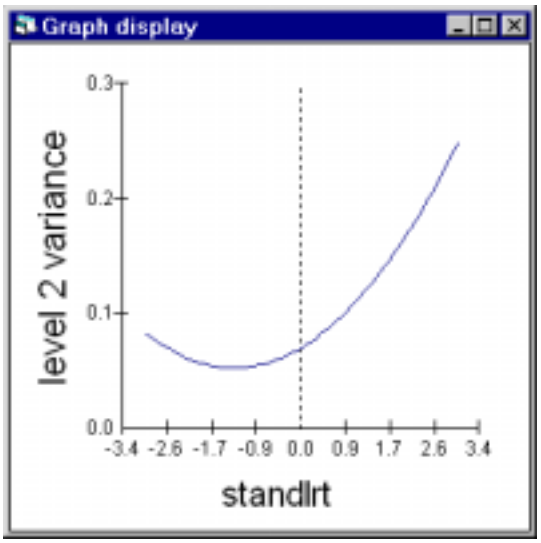
The function shown is simply the variance of the sum of two random coefficients times their respective explanatory variables, u_{0j}**cons** and u_{0j}**standlrt**$_{ij}$, written out explicitly. Given that **cons** is a vector of ones we see that the between school variance is a quadratic function of **standlrt** with coefficients formed by the set of level 2 random parameters. The intercept in the quadratic function is σ^2_{u0}, the linear term is $2\sigma_{u01}$ and the quadratic term is σ^2_{u1}. We can compute this function and the Variance function window provides us with a simple means of doing this.

The column in the window headed **select, cons, standlrt and result** are for computing individual values of the variance function. Since **standlrt** is a continuous variable it will be useful to calculate the level 2 variance for every value of **standlrt** that occurs**.**

In the **variance output to** list on the tool bar, select c30

Click **Calc**

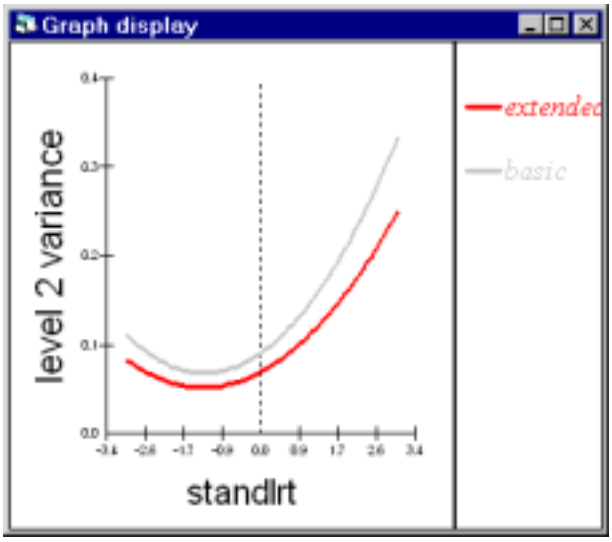
Now you can use the **customised graph** window to plot c30 against **standlrt:**



The above graph has had the y-axis rescaled to run between 0 and 0.3. The apparent pattern of greater variation between schools for students with extreme **standlrt** scores, especially high ones, is consistent with the plot of prediction lines for the schools we viewed earlier.

We need to be careful about over the interpretation of such plots. Polynomial functions are often unreliable at extremes of the data to which they are fitted. Another difficulty with using polynomals to model variances is that they may, for some values of the explanatory variables, predict a negative overall variance. To overcome this we can use nonlinear(negative exponential) functions to model variance. This is an advanced topic and for details see the *Advanced Modelling* Guide (Yang et al., 1999).

We can construct a similar level 2 variance plot for the basic random slope model, before extending the model by adding **gender, schgend** and **schav** explanatory variables. This can be illuminating because it shows us to what extent these variables are explaining between-school differences across the range of **standlrt**. This is left as an exercise for the reader but the graph comparing the between school variance for the two models is shown below.



We see in both models schools are more variable for students with high **standlrt** scores. The explanatory variables we added in the extended model explain about 25% of the between school variation across the spectrum of **standlrt.**

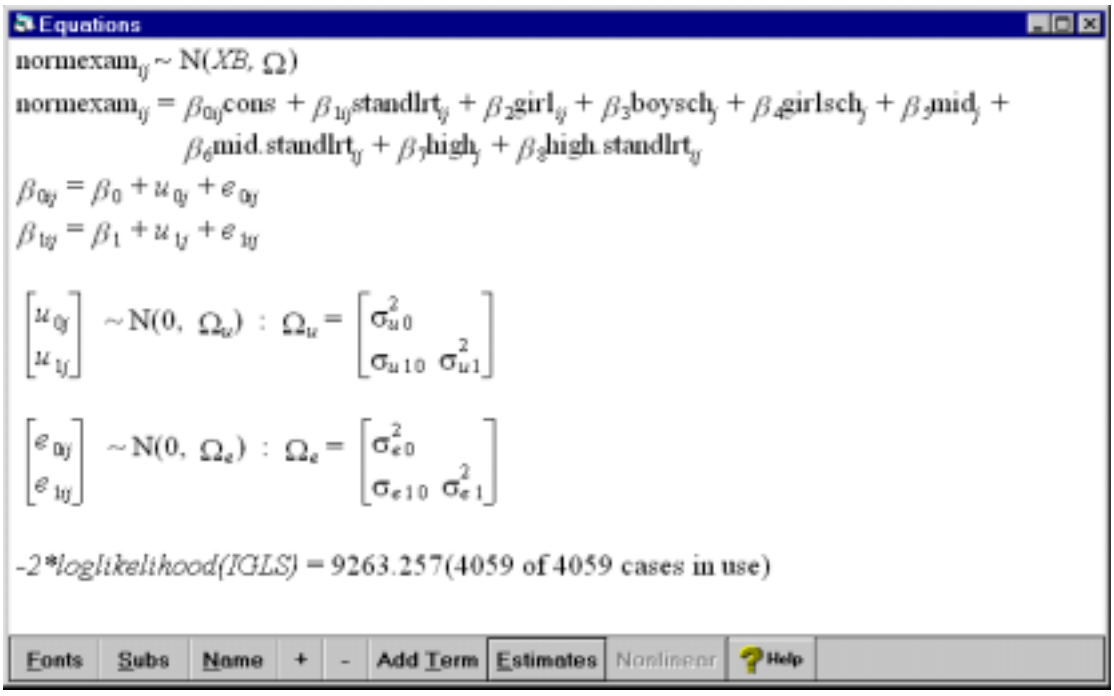
Complex variation at level 1

Until now we have assumed a constant variance at level 1. It may be that the student level departures around their school summary lines are not constant. They may change in magnitude at different levels of **standlrt** or be larger for boys than girls. In other words the student level variance may also be a function of explanatory variables.

Let's look and see if the pupil level variance changes as a function of **standlrt**. To do this we need to make the coefficient of **standlrt** random at the student level. To do this

In the equations window click on β_1 Check the box labeled **i(student)**

Which produces



Now β_1 the coefficient of **standlrt** has a school level random term u_{1j} and a student level random term e_{1ij} attached to it. As we have seen, at the school level we can think of the variance of the u_{1j} terms, that is σ^2_{u1} in two ways. Firstly, we can think of it as the between school variation in the slopes. Secondly we can think of it as a coefficient in a quadratic function that describes how the between school variation changes with respect to **standlrt**. Both conceptualisations are useful.

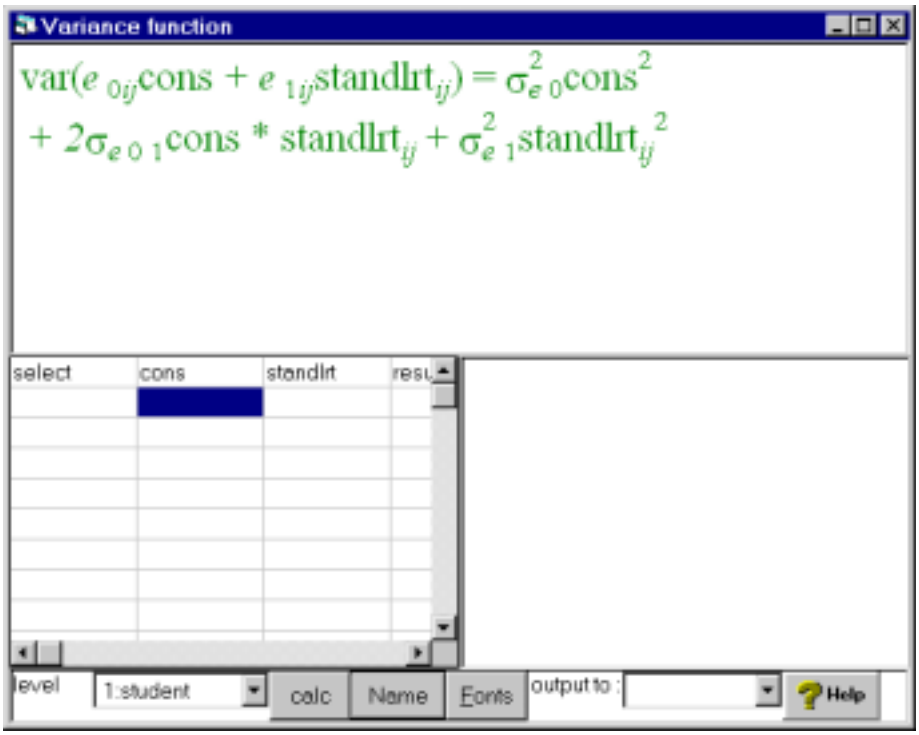
The situation at the student level is different. It does not make sense to think of the variance of the e_{1ij}'s, that is σ^2_{e1} as the between student variation in the slopes. This is because a student corresponds to only one data point and it is not possible to have a slope through one data point. However, the second conceptualisation where σ^2_{e1} is a coefficient in a function that describes how between student variation changes with respect to **standlrt** is both valid and useful**.** This means that in models with complex level 1 variation we do not think of the estimated random parameters as separate variances and covariances but rather as elements in a function that describes how the level 1 variation changes with respect to explanatory variables. The **variance function** window can be used to display the form of the function.

Run the model

Select the **variance function** menu

From the **level** drop down list select **1:student**

Which produces



As with level 2, we have a quadratic form for the level 1 variation. Let us evaluate the function for plotting

In the **output to** drop down list select **c31**

Click **calc**

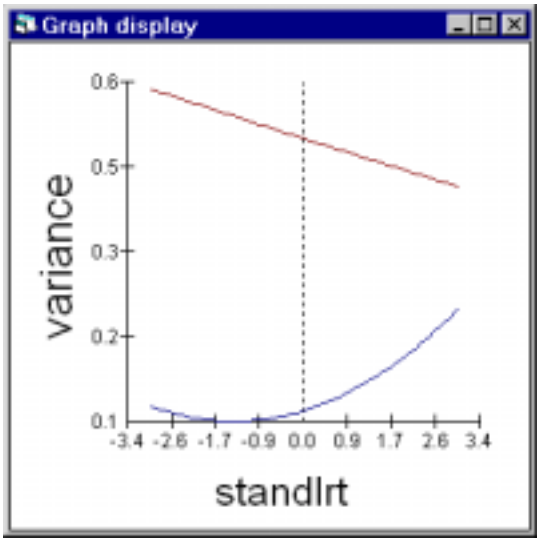
Now let's add the level 1 variance function to the graph containing the level 2 variance function.

Select the **customised graphs** window

Select the **display** used to plot the level 2 variance function

Add another data set with **y** as c31, **x** as **standlrt,** plotted as a red line

Which produces



The lower curved line is the between school variation. The higher straight line is the between student variation. If we look at the equations screen we can see that σ^2_{e1} is zero to 3 decimal places. The variance σ^2_{e1} acts as the quadratic coefficient in the level 1 variance function hence we have a straight line as the function is dominated by the other two terms. The general picture is that the between school variation increases as **standlrt** increases, whereas between student variation decreases with **standlrt.** This means the intra-school correlation(school variance / [school variance + student variance]) increases with **standlrt.** Therefore the effect of school is relatively greater for students with higher intake achievements.

Notice, as we pointed out earlier, that for high enough levels of **standlrt** the level 1 variance will be negative. In fact in the present data set such values of **standlrt** do not exist and the straight line is a reasonable approximation over the range of the data.

The student level variance functions are calculated from 4059 points, that is the 4059 students in the data set. The school level variance functions are calculated from only 65 points. This means that there is sufficient data at the student level to support estimation of more complex variance functions than at the school level.

Lets experiment by allowing the student level variance to be a function of gender as well as **standlrt**. We can also remove the σ^2_{e1} term which we have seen is negligible.

In the equations window click on β_2

Check the box labeled **i(student)**

The level 1 matrix Ω_e is now a 3 by 3 matrix.

Click on the σ^2_{e1} term.

You will be asked if you want to remove the term from the model. Click **yes**

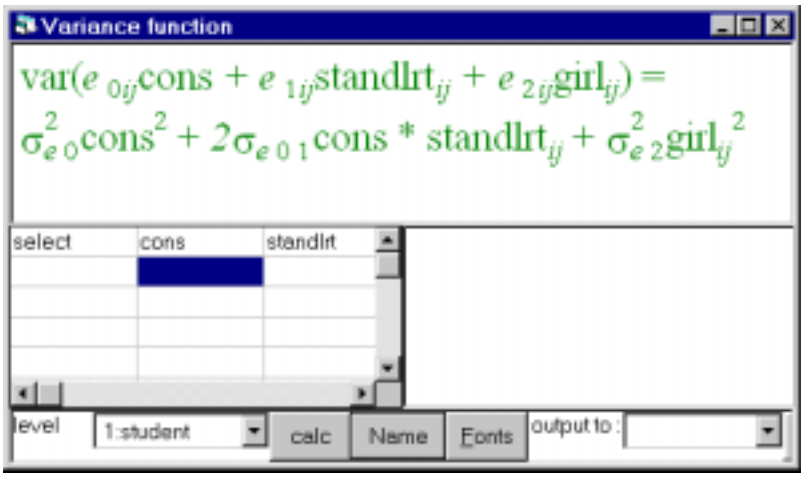
Do the same for σ_{e21} and σ_{e20}

When you remove terms from a covariance matrix in the **equations** window they are replaced with zeros. You can put back removed terms by clicking on the zeros.

Notice that the new level 1 parameter σ^2_{e2} is estimated as –0.054. You might be surprised at seeing a negative variance. However, remember at level 1 that the random parameters cannot be interpreted separately; instead they are elements in a function for the variance. What is important is that the function does not go negative within the range of the data.

[Note – *MLwiN* by default will allow negative values for individual variance parameters at level 1. However, at higher levels the default behaviour is to reset any negative variances and all associated covariances to zero. These defaults can be over-ridden in the **Estmation Control** window available by pressing the **Estimation Control** button on the main toolbar.]

Now use the variance function window to display what function is being fitted to the student level variance.



From the **equations** window we can see that $\{\sigma^2_{e0}, \sigma_{e01}, \sigma^2_{e2}\}$={0.583, -0.012, -.054}. Substituting these values into the function shown in the **variance function** window we get the student level variance for the boys is :

0.583 - 0.024 * **standlrt**

and for the girls is:

0.583 - 0.054 - 0.024***standlrt**

Note that we can get the mathematically equivilent result fitting the model with the following terms at level 1 : $\sigma^2_{e0}, \sigma_{e01}, \sigma_{e02}$. This is left as an exercise for the reader.

The line describing the between student variation for girls is lower than the boys line by 0.054. It could be that the lines have different slopes. We can see if this is the case by fitting a more complex model to the level 1 variance. In the **equations** window:

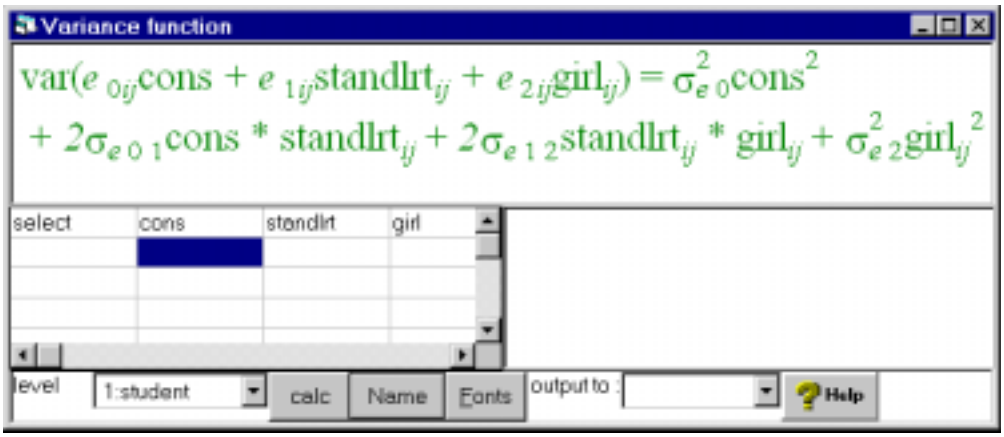
In the level 1 covariance matrix click on the right hand 0 on the bottom line.

You will be asked if you want to add term **girl/standlrt.** Click **Yes.**

Run the model

We obtain estimates for the level 1 parameters $\{\sigma^2_{e0}, \sigma_{e01}, \sigma_{e12}, \sigma^2_{e2}\}$={0.584, -0.032, 0.031,-0.058}

The updated variance function window now looks like this :



The level 1 variance for boys is now :

0.584 + 2 *(-0.032)**standlrt**=0.584 –0.064* **standlrt**

and for girls is:

0.584 +(2*(-0.032) +2*(0.031))* **standlrt**-0.058=0.526-0.02* **standlrt**

We can see the level 1 variance for girls is fairly constant across **standlrt.** For boys the level 1 variance function has a negative slope, indicating the boys who have high levels of **standlrt** are much less variable in their attainment**.** We can graph these functions :

In the **variance function** window set **output to:** list to c31

Press **calc**

Select the **customised graphs** window

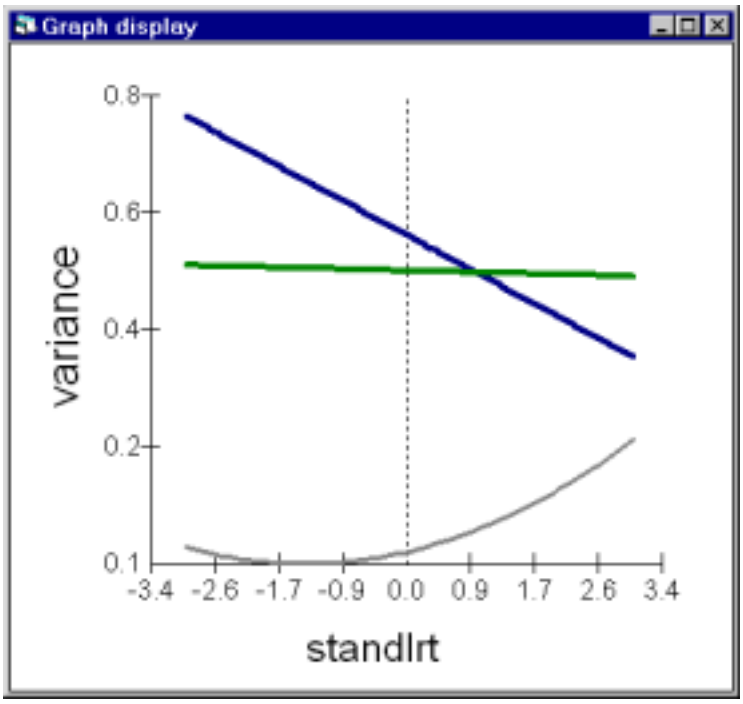
Select the **display** used to plot the level 2 variance function

Select the data set **y**=c31, **x=standlrt**

In the **group** list select **gender**

Click **Apply**

Which produces :



We see that the student level variance for boys drops from 0.8 to 0.4 across the spectrum of **standlrt**, whereas the student level variance for girls remains fairly constant at around 0.53.

We are now forming a general picture of the nature of the variability in our model at both the student and school levels of the hierarchy. The variability in schools' contributions to students progress is greater at extreme values of **standlrt,** particularly positive values. The variability in girls progress is fairly constant. However, the progress of low intake ability boys is very variable but this variability drops markedly as we move across the intake achievement range.

These complex patterns of variation give rise to intra-school correlations that change as a function of **standlrt** and **gender**. Modelling such intra-unit correlations that change as a function of explanatory variables provides a useful framework when addressing interesting substantive questions.

Fitting models which allow complex patterns of variation at level 1 can produce interesting substantive insights. Another advantage is that where there is very strong heterogeneity at level 1 failing to model it can lead to a serious model specification. In some cases the mis-specification can be so severe that the simpler model fails to converge but when the model is extended to allow for a complex level 1 variance structure convergence occurs. Usually the effects of the mis-specification are more subtle, you can find that failure to model complex level 1 variation can lead to inflated

estimates of higher level variances (that is between-student heterogeneity becomes incorporated in between-school variance parameters).

What you should have learnt from this chapter

That variance functions are a useful interpretation for viewing variability at the different levels in our model.

How to construct and graph variance functions in *MLwiN*

A more complex interpretation of intra unit correlation

Mortality in England and Wales, 1979-1992
An Introduction to Multilevel Modelling using MLwiN

Alastair H Leyland and Alice McLeod

MRC Social and Public Health Sciences Unit
University of Glasgow
4 Lilybank Gardens
Glasgow G12 8RZ

MRC Social and Public Health Sciences Unit occasional paper no. 1, 2000.

The Social and Public Health Sciences Unit is a Research Unit supported by the Medical Research Council and the Chief Scientist Office of the Scottish Executive Health Department at the University of Glasgow. The production of these training materials has been supported by the Economic and Social Research Council as part of the Analysis of Large and Complex Datasets programme.

Mortality in England and Wales, 1979-1992
An Introduction to Multilevel Modelling using MLwiN

Alastair H Leyland and Alice McLeod

INTRODUCTION TO THE DATASET

The data are taken from the local mortality datapack and detail deaths from all causes in England and Wales in the period 1979 to 1992. The data are stored at the Data Archives at the University of Essex, maintained by the Economic and Social Research Council. The raw data comprise two files: one containing information on deaths over this time period and the other detailing the populations of the relevant areas (counties in England and Wales) in each year. For further information on this and other available datasets the user should visit the Data Archives website: `http://dawww.sx.ac.uk/`

RESEARCH QUESTIONS

In the following tutorial we will attempt to answer the following research questions:

1. What is happening to mortality rates over time?
2. How much variation in mortality rates is there between districts of England and Wales?
3. Is this variation just between districts, or are there also differences between the mortality rates of counties?
4. Does mortality vary according to the type of area?
5. What is happening to the variation in mortality rates over time?

INTRODUCTION TO MLWIN

Opening a worksheet

MLwiN files are known as worksheets which include all the data and model settings from the last saved version.

Go to the **File** menu Select **Open worksheet** Open the worksheet called **lmdp.ws**

The name of the current file appears in the bar at the top of the MLwiN window.

Names window

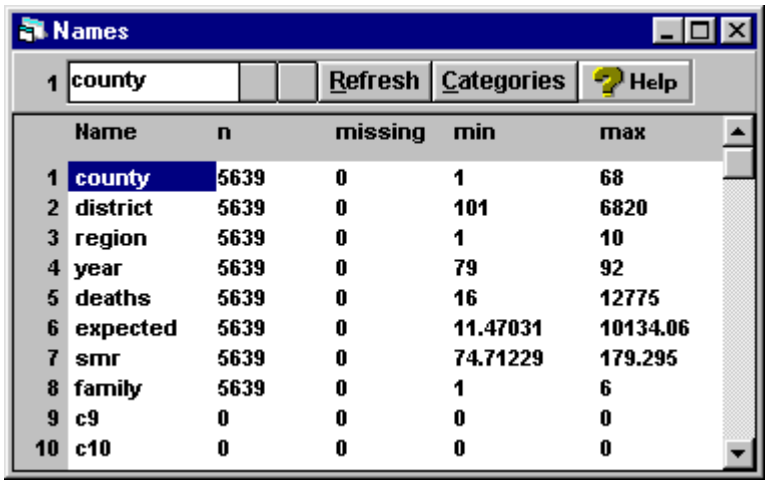
To view a summary of this worksheet

Go to the **Data manipulation** menu Select **Names**

This brings up a list of all the variables stored in the worksheet together with some summary information. The worksheet contains 8 variables, each with 5639 data points and no missing values. Each data point corresponds to the annual number of deaths for all districts in England and Wales in the period 1979-92. COUNTY, DISTRICT and REGION are area identifiers; there are 403 county DISTRICTs (coded from 101 to 6820) which are nested within 54 COUNTYs (coded from 1 to 68), and these in turn lie within one of 10 REGIONs. The data cover 14 YEARs from 1979 to 1992. Note that there are only 5639 data points rather than the 5642 that might be expected (403 county districts with an observation for each of 14 years); three data points have been removed because extreme outlying values made them implausible. The next two columns show the number of DEATHS observed in each district at each time point – ranging from 16 to 12,775 – and the number that would be EXPECTED. The expected number of deaths has been calculated on the basis of the age and sex structure of that area's population in each year had the 1992 national age- and sex-specific mortality rates applied throughout. This worksheet has been constructed using the two raw data files contained in the local mortality datapack – the number of deaths and the populations. The OBSERVED and EXPECTED deaths are combined to form the standardised mortality ratio (SMR) for each year in each district. This is calculated as

$$smr = \frac{\text{observed deaths}}{\text{expected deaths}} \times 100$$

and reflects the excess deaths in an area, standardised for age and sex, over the national average mortality rate in 1992 (average = 100). The range from 75 to 179 implies a minimum mortality rate for one area in one year 25% below the 1992 average and a maximum 79% above the average. Finally, the variable FAMILY is a classification of districts into 6 groups devised by the Office for National Statistics:- 1 – Inner London, 2 – Rural areas, 3 – Prospering areas, 4 – Maturer areas, 5 – Urban centres, 6 – Mining and industrial areas. All of the remaining columns are empty; the default name for such columns is 'C' followed by the column number.



	Name	n	missing	min	max
1	county	5639	0	1	68
2	district	5639	0	101	6820
3	region	5639	0	1	10
4	year	5639	0	79	92
5	deaths	5639	0	16	12775
6	expected	5639	0	11.47031	10134.06
7	smr	5639	0	74.71229	179.295
8	family	5639	0	1	6
9	c9	0	0	0	0
10	c10	0	0	0	0

Data window

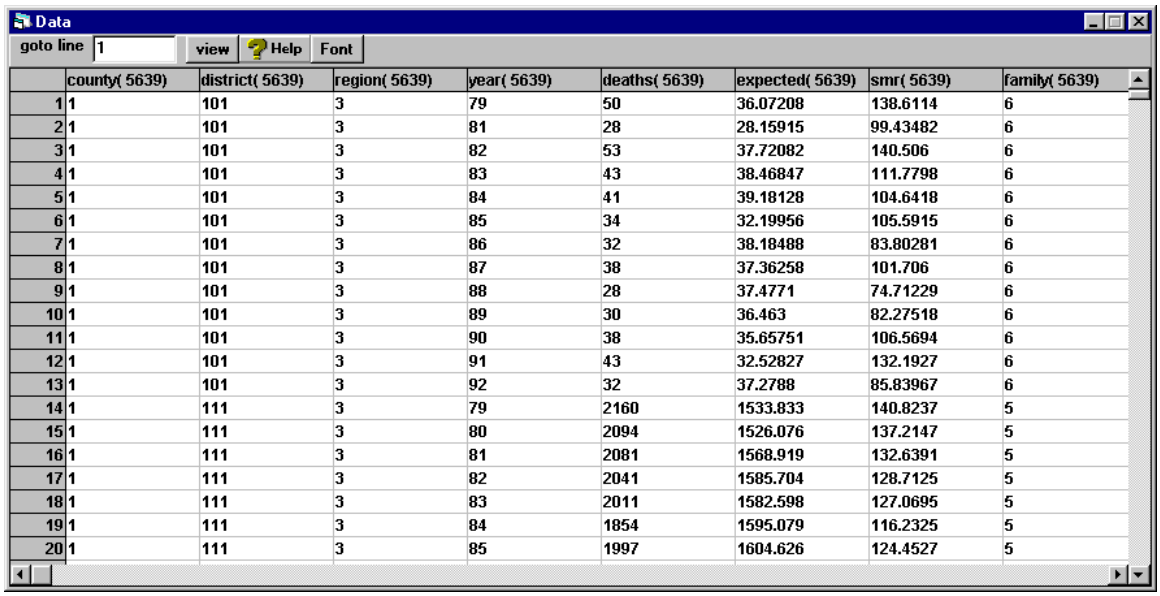
The data may be viewed and edited in a spreadsheet format

Go to the **Data manipulation** menu Select **View or edit data**

By default only the first 3 columns are shown. The **view** button within this window can be used to change or extend the selection of variables shown.

Click the **view** button To select all variables, click on COUNTY, hold down the left mouse button and drag to FAMILY Click the **OK** button

All windows can be re-sized by clicking on the borders and dragging; also the scroll bars at the bottom and on the right hand side can be used to view more of the selected data.



	county(5639)	district(5639)	region(5639)	year(5639)	deaths(5639)	expected(5639)	smr(5639)	family(5639)
1	1	101	3	79	50	36.07208	138.6114	6
2	1	101	3	81	28	28.15915	99.43482	6
3	1	101	3	82	53	37.72082	140.506	6
4	1	101	3	83	43	38.46847	111.7798	6
5	1	101	3	84	41	39.18128	104.6418	6
6	1	101	3	85	34	32.19956	105.5915	6
7	1	101	3	86	32	38.18488	83.80281	6
8	1	101	3	87	38	37.36258	101.706	6
9	1	101	3	88	28	37.4771	74.71229	6
10	1	101	3	89	30	36.463	82.27518	6
11	1	101	3	90	38	35.65751	106.5694	6
12	1	101	3	91	43	32.52827	132.1927	6
13	1	101	3	92	32	37.2788	85.83967	6
14	1	111	3	79	2160	1533.833	140.8237	5
15	1	111	3	80	2094	1526.076	137.2147	5
16	1	111	3	81	2081	1568.919	132.6391	5
17	1	111	3	82	2041	1585.704	128.7125	5
18	1	111	3	83	2011	1582.598	127.0695	5
19	1	111	3	84	1854	1595.079	116.2325	5
20	1	111	3	85	1997	1604.626	124.4527	5

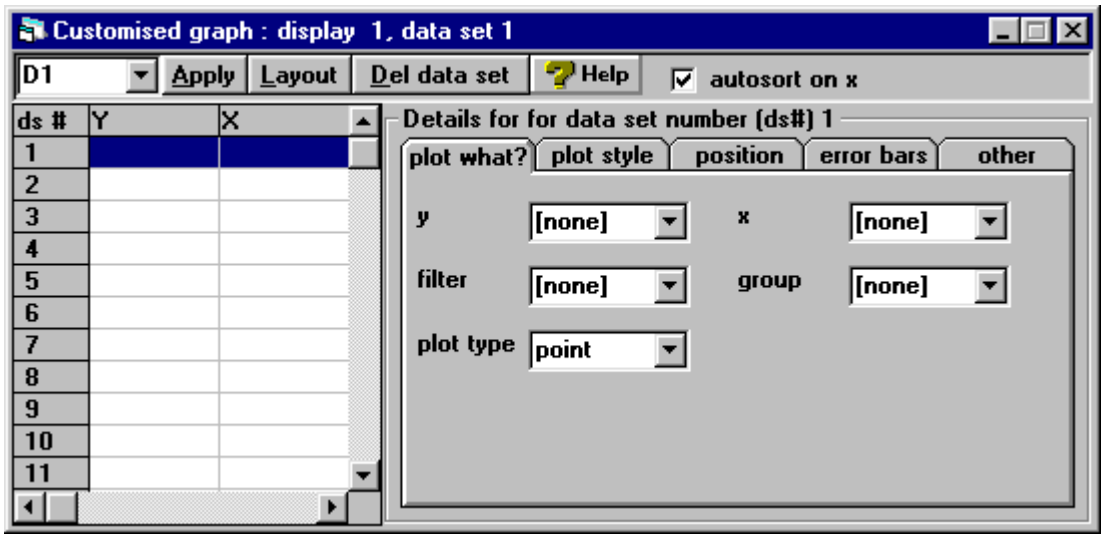
The first 13 observations are made on DISTRICT 101, COUNTY 1, REGION 3. The 13 observations on this DISTRICT can be seen to correspond to 13 YEARs of data; there is no observation for 1980. The estimated SMR in this district varies from 75 in 1988 to 141 in 1982. The district classification was group 1 – Inner London.

Graph window

Before starting to model we may wish to examine the data in a graph.

Go to the **Graphs** menu Select **Customised Graphs**

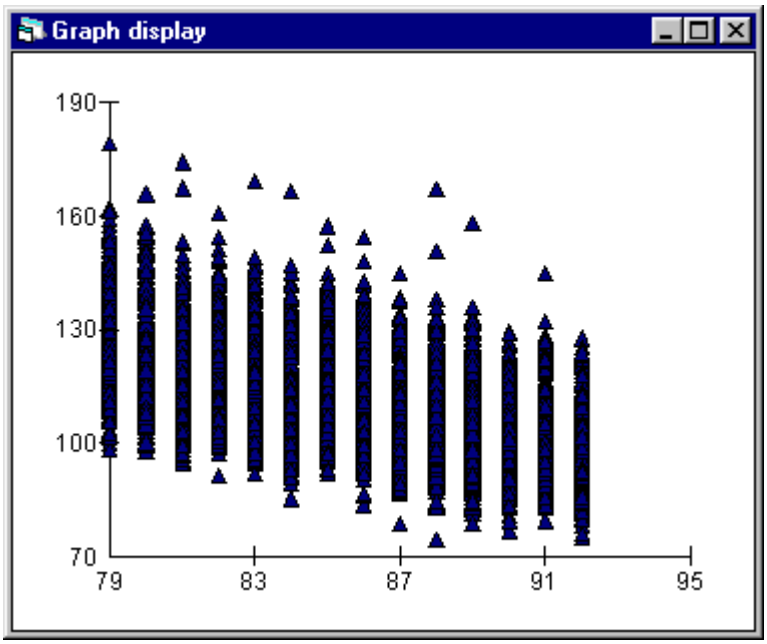
The graphical output in MLwiN is separated into three components. A *display* is what can be displayed on the computer screen at any one time and up to ten different displays may be specified. The pull-down menu at the top-left hand corner of the customised graph window corresponds to the display function – this currently shows **D1** denoting display 1. Each display can contain a number of *graphs*. A graph is a frame with x and y axes showing lines, points or bars, and each display can show an array of up to 5x5 graphs. The **Layout** button at the top of the customised graph window is used to specify the layout of the display. Finally, each graph can plot one or more *datasets*, each one consisting of a set of x and y coordinates held in the worksheet columns. Different datasets may be specified by clicking on different rows under the **ds#** heading shown at the right hand side of the customised graph display.



To obtain a scatter plot of SMRs by year, ensure that the **plot what?** tab is selected and

Select the **y** variable to be SMR Select the **x** variable to be YEAR

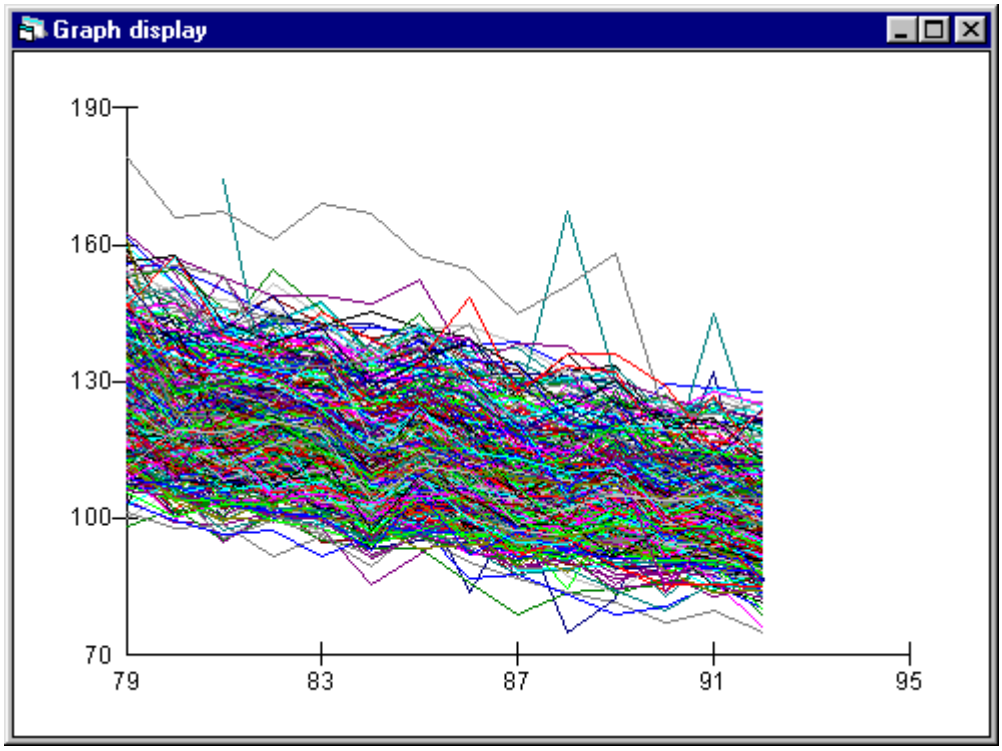
Click the **Apply** button



It is clear that there have been considerable reductions in SMR over these 14 years; nearly every district had an SMR greater than 100 in 1979. (The fact that standardisation was to 1992 means that the average SMR was 100 for that year.)

To change this graph to a line plot with a line for each district

In the **Customised Graph** window, select **group** to be DISTRICT
Change **plot type** to **line**
Select the **plot style** tab
Change **colour** to **16 rotate**
Click the **Apply** button



It is possible to identify points on the graph: point and click anywhere on the graph and the **Identify point** window will appear with details of the closest data point. Also included in the **Graph options** window are facilities for adding titles to the graph and axes, and for making other changes to the display including the scales.

Closing windows

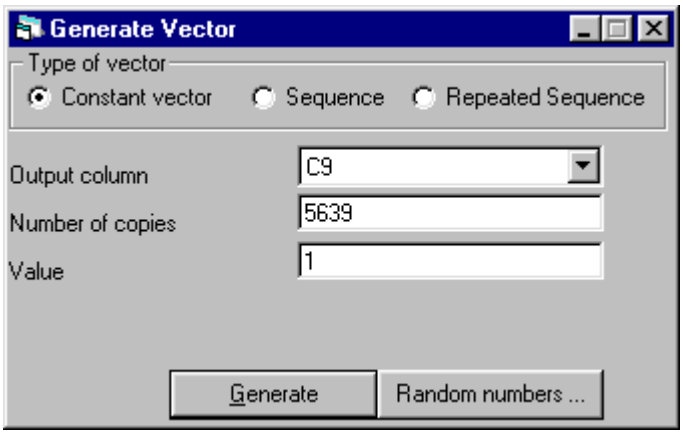
At any time you may wish to close or minimize windows to prevent your screen from becoming too cluttered. You may do this, as with any other Windows package, by clicking on the X or _ buttons respectively in the top right corner of each window. Alternatively you may go to the **Window** menu and select **close all windows**.

MODEL SPECIFICATION

Creating new variables

A number of functions are available in MLwiN that allow the creation of new variables or amendments to existing variables. In order to include a constant or intercept term in a model, we need to create a column of 1's that spans the entire data set.

Go to **Data manipulation** menu
Select **Generate vector**
Select **Type of vector** to be **Constant vector**
Select C9 to be the **Output column**
Enter 5639 (the number of data points) beside **Number of copies**
Enter 1 beside **Value**
Click the **Generate** button



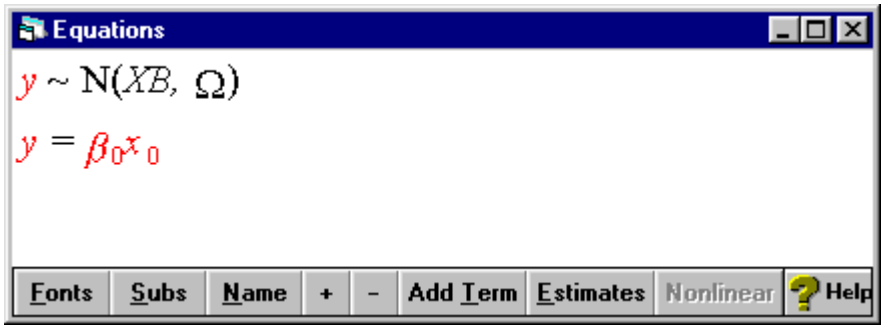
Returning to the **Names** window, column C9 now contains 5639 data points each with the value of 1. To name this new variable

Click on C9 in the **Names** window
Highlight C9 in the box at the top of the window
Type CONS and press <return>

Equations window

Specifying models in MLwiN is done mainly via the **Equations** window.

Go to **Models** menu Select **Equations**



The items in red must be defined before a model can be fitted to the data.

Click on either of the **y** terms Select SMR as the dependent variable

The structure of the hierarchical model is also specified at this stage; first by stating the number of levels the model will have and then by specifying what the levels of the hierarchy are using the appropriate identifier variables. For a 2-level model of YEARs nested within DISTRICTs

Select 2 – ij for **N levels** Select DISTRICT for **level 2(j)** Select YEAR for **level 1(i)** Click on **Done**

The red response variable **y** has changed colour indicating that this term has been defined; moreover, the addition of the subscripts **i** and **j** indicates that this is a 2-level model. The ordering tells us that YEAR **i** is nested within DISTRICT **j**. In a similar manner we can define CONS to be an independent variable.

Click on the $\beta_0 x_0$ term Select CONS from the drop-down list

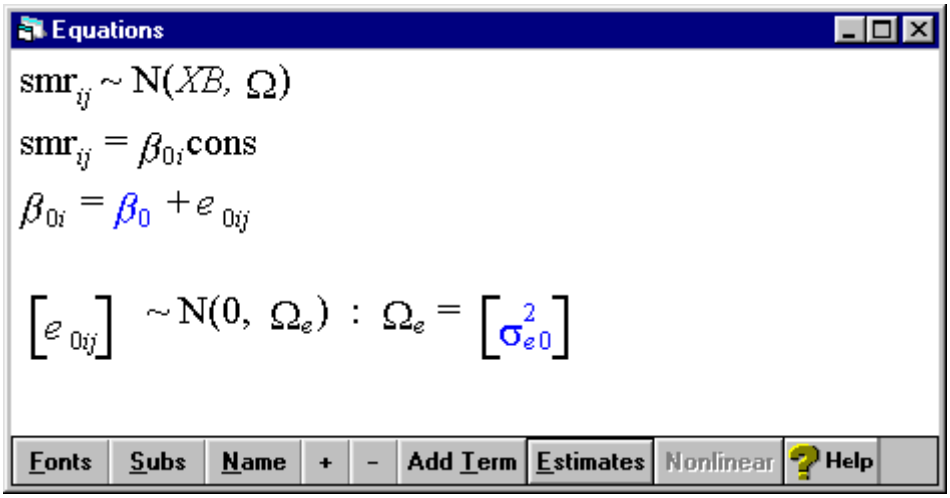
The check boxes indicate what part of the model each variable is in; by default, CONS has been added to the fixed part of the model and its coefficient will provide an estimate of the intercept. The other options in this window relate to the random part of the model. We will start by fitting a model with the intercept random at the level of YEAR (level 1).

Click on the check box by **i(YEAR)** Click on **Done**

Note that the term β_0 changes to β_{0i}, denoting the fact that it is random at level 1.

At any time we can toggle between a purely algebraic representation and one that gives a little more detail of the models that are being fitted.

In the **Equations** window, click on the **Name** button Click the **Estimates** button



Note that **y** and **x** have been replaced by the names of the specified dependent and independent variables, SMR and CONS. Clicking the **Estimates** button expands the model to include the distributional assumptions that are applied to the random terms, in this case to the error term $\mathbf{e_{0ij}}$. Two terms in the model are blue: the grand intercept β_0 and the level 1 variance σ_{e0}^2. This indicates that these terms are to be estimated; when the model converges, the blue will change to green. Clicking on the **Estimates** button again will replace these two terms with their current estimates (both the default value of 1.00 because no model has yet been estimated).

In addition to the mean we will add year as an independent variable in the fixed part of the model in order to answer our first research question.

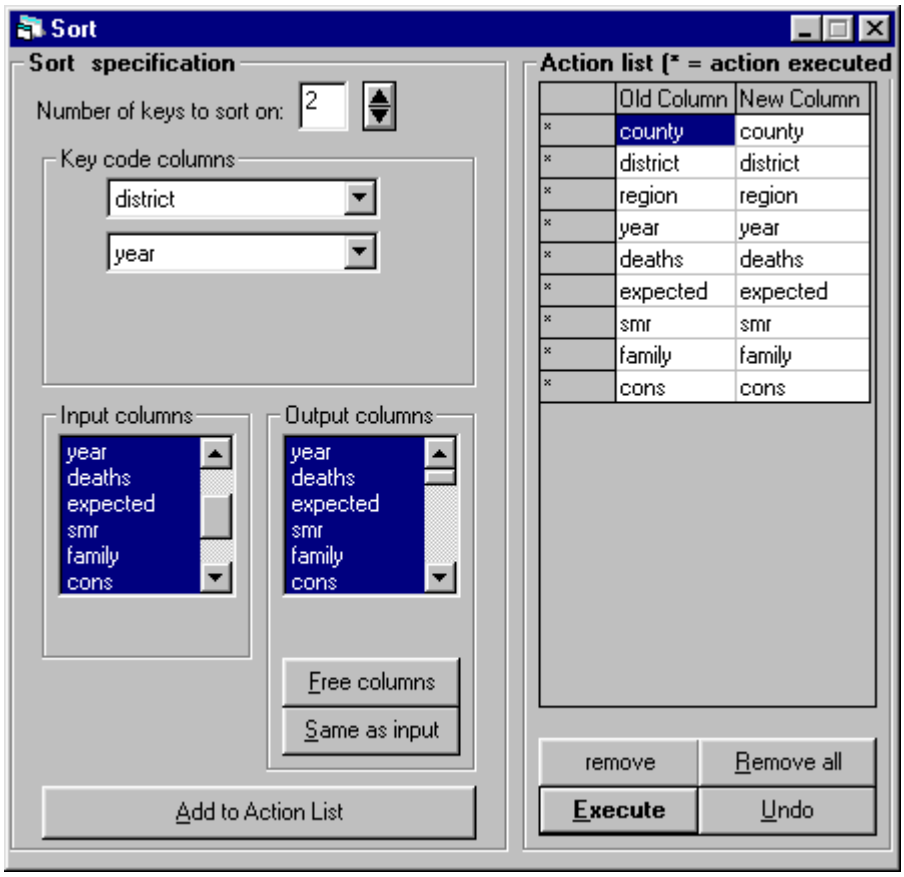
Click on the **Add Term** button
Click on the red $\mathbf{x_1}$ which appears and select YEAR from the drop-down list
Click **Done**

The third term to be estimated, $\boldsymbol{\beta_1}$, is the coefficient associated with year and this will estimate the trend in SMRs during the study period.

Sorting the data

Before fitting any model, the data need to be sorted within their hierarchy (in this example by YEAR within DISTRICTs).

Go to **Data manipulation** menu
Select **Sort**
Increase the **Number of keys to sort on** to 2
Select DISTRICT as the first **Key code column** and YEAR as the second
Select all named variables under the heading **Input columns**
Press **Same as input** button to overwrite current columns with sorted data
Press **Add to action list** and then **Execute**

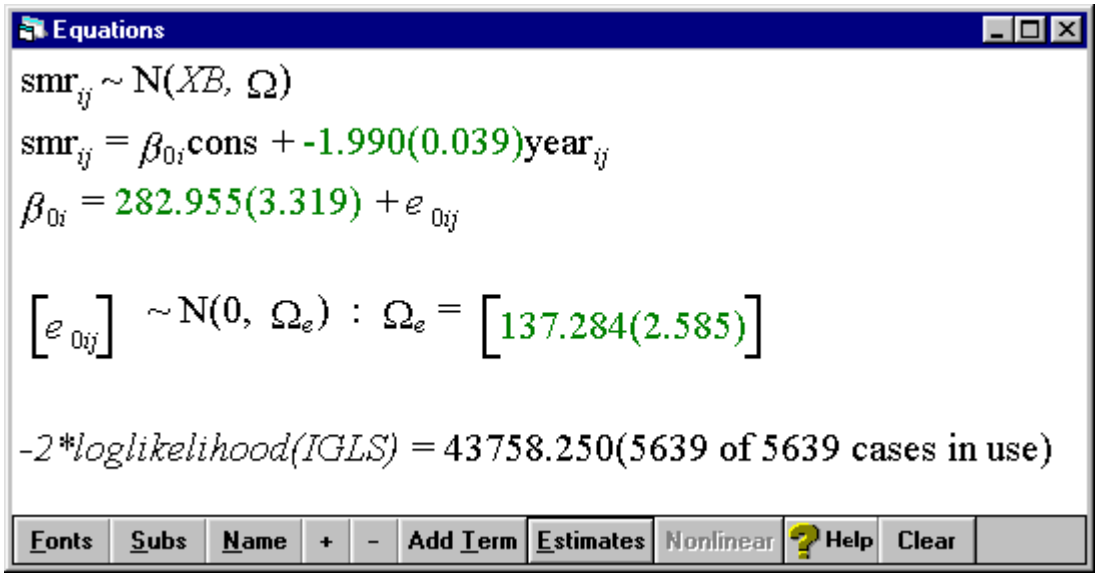


Fitting the model

The model is now ready to be estimated.

Click the **Start** button on the tool bar at the top of the MLwiN screen

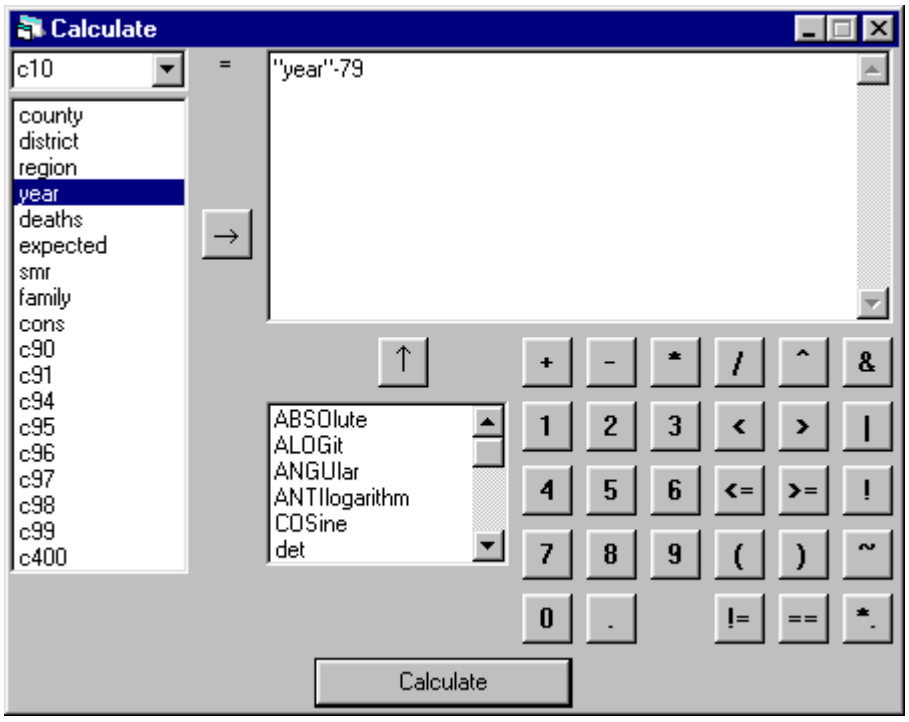
After one iteration the model converges and the blue estimates in the equation window turn green.



The estimated standard errors of each of the parameter estimates are shown in brackets. Our intercept is about 283, and the SMR has been decreasing at 1.990 per year. This decrease is highly significant in comparison with its standard error. The variance of all of the observations around this fitted trend is 137. The current model has a single term to describe the variation around the mean and is therefore just an ordinary least squares (OLS) regression model, but it is a starting point for our multilevel analysis. The value –2*log(likelihood) is provided as an aid to model selection.

Before continuing, consider the interpretation of the intercept term. This is the predicted value of the SMR in all districts when the variable YEAR takes the value 0 – in other words, in 1900. Since the data do not cover this period it is not sensible to make any inference about the SMR at this time, and we can change the origin to something more meaningful. Explanatory variables are frequently centred around an average value; in this case, however, we will set the origin at the first year for which we have data (1979).

Go to **Data manipulation** menu
Select **Calculate**
Select the empty column C10 from the list of variables and press the right arrow button
Click on the = button on the window's keypad
Select YEAR from the list of variables and press the right arrow button
Use the window's keypad to enter **–79**
Press **Calculate**

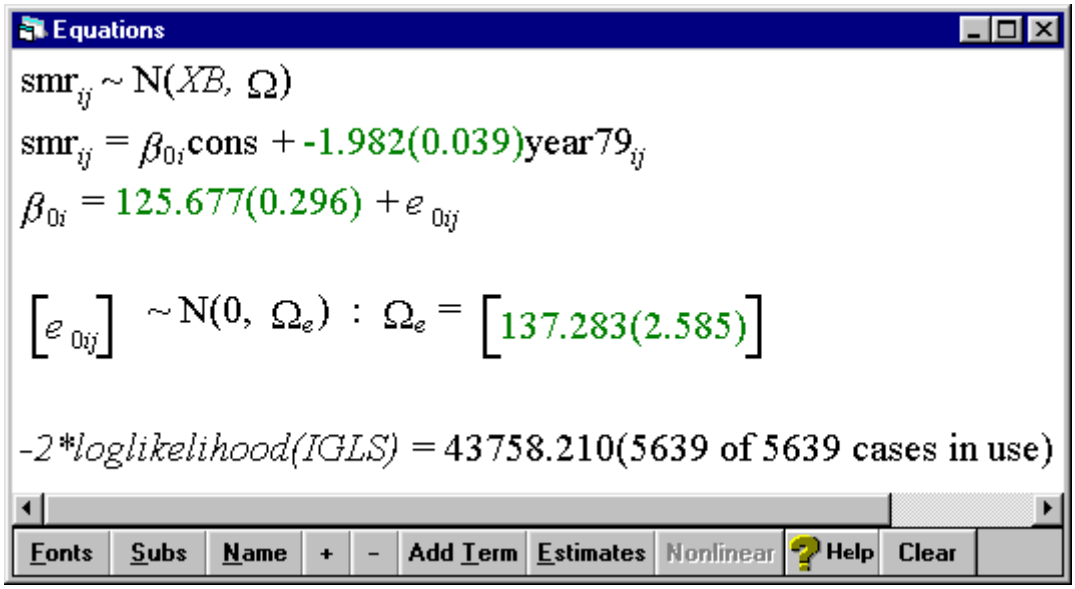


This has created a new year variable with an origin of 1979; name this new variable YEAR79 using the **Names** window.

To change the trend variable in the current model return to the **Equations** window

Click on $\mathbf{year}_{ij}$
Select YEAR79 from the drop-down list
Click **Done**

The parameter estimates turn blue indicating that the model has changed and that it must be re-estimated. Rather than click on the **Start** button again, click on the **More** button to continue estimation from the current values.



The estimated slope has changed very slightly, but the big change is in the intercept. In 1979 the average SMR was therefore about 126.

VARIANCE COMPONENTS

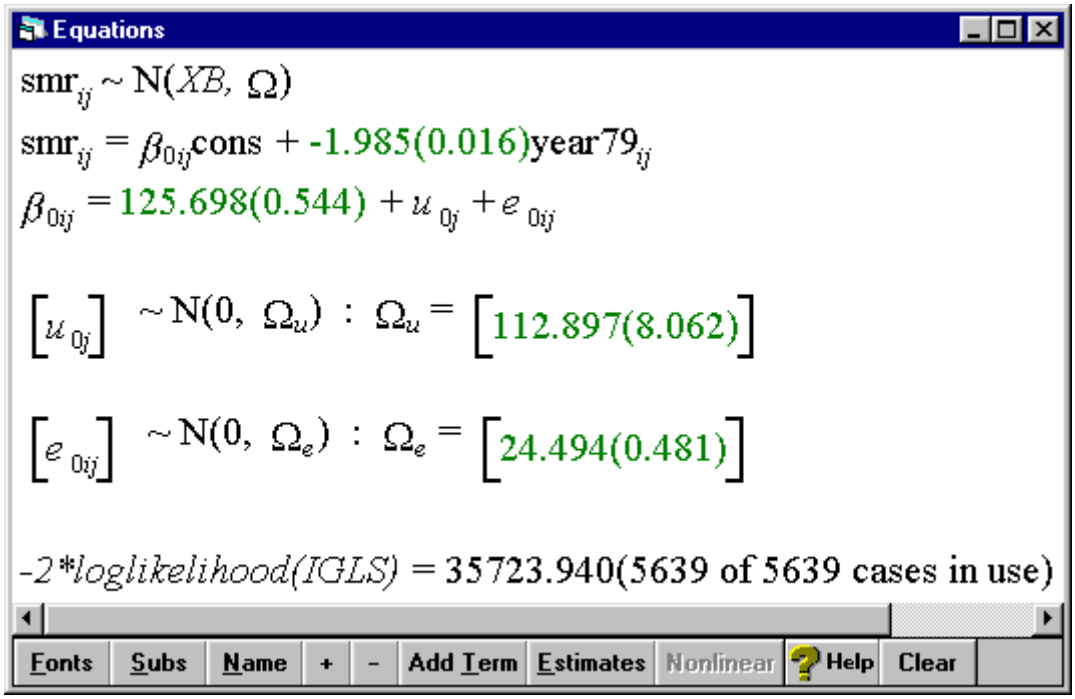
All of the variance in the current model is at the lowest level of observation; this is just an ordinary least squares (OLS) regression equation. This model may be expanded by partitioning the variance into that which is attributable to random variation between DISTRICTs and that which arises due to fluctuations between YEARs within DISTRICTs.

A 2-level variance components model

In the **Equations** window

Click on β_{0i}
Check the box by **j(DISTRICT)**
Click **Done**

The intercept term now has an additional subscript, indicating that it varies across DISTRICTs as well as across YEARs. The intercept is partitioned into three parts: the overall fixed part intercept for 1979, the error term $\mathbf{e_{0ij}}$ and a term $\mathbf{u_{0j}}$ which is specific to DISTRICT j. The $\mathbf{u_{0j}}$ are random effects at level-2 and are assumed to arise from a normal distribution. The intercept for the j^{th} district in 1979 will be given by $\beta_0 + \mathbf{u_{0j}}$. This model may be fitted by clicking **More**.



There is little change in the estimates of the intercept and slope in the fixed part of the model. However, most of the variation is now at level 2 – between DISTRICTs rather than YEARs. The total variation ($\sigma_{u0}^2 + \sigma_{e0}^2$) is 137.391, very close to the estimate of σ_{e0}^2 obtained in the single level model. The proportion of the total variance which arises due to differences between DISTRICTs is approximately 113/137 or 82.2%. This figure is known as the *intra-unit* or *intra-class correlation*, and indicates that the correlation between two observations made in different YEARs on the same DISTRICT is 0.822. The level 1 variance may be interpreted as the variation between years within districts. So, in answer to the second research question, it would appear that the majority of the variation in mortality is due to between-district differences rather than year-on-year fluctuations. Note that the addition of a single variance term has produced a substantial reduction in the value of –2*log(likelihood).

Adding a further level

We can add COUNTY as a third level to the model and examine the relative importance of these large areas compared to the smaller DISTRICTs. First resort the data according to this new hierarchy.

Go to **Data manipulation** menu
Select **Sort**
Increase the **Number of keys to sort on** to 3
Select COUNTY as the first **Key code column,** DISTRICT as the second and YEAR as the third

Select all named variables under the heading **Input columns**

Press **Same as input** button to overwrite current columns with sorted data

Press **Add to action list** and then **Execute**

Now return to the **Equations** window

Click on $\mathbf{y_{ij}}$ or $\mathbf{smr_{ij}}$

Change **N levels** to **3 – ijk**

Select COUNTY from the drop-down list by **level 3(k)**

Click **Done**

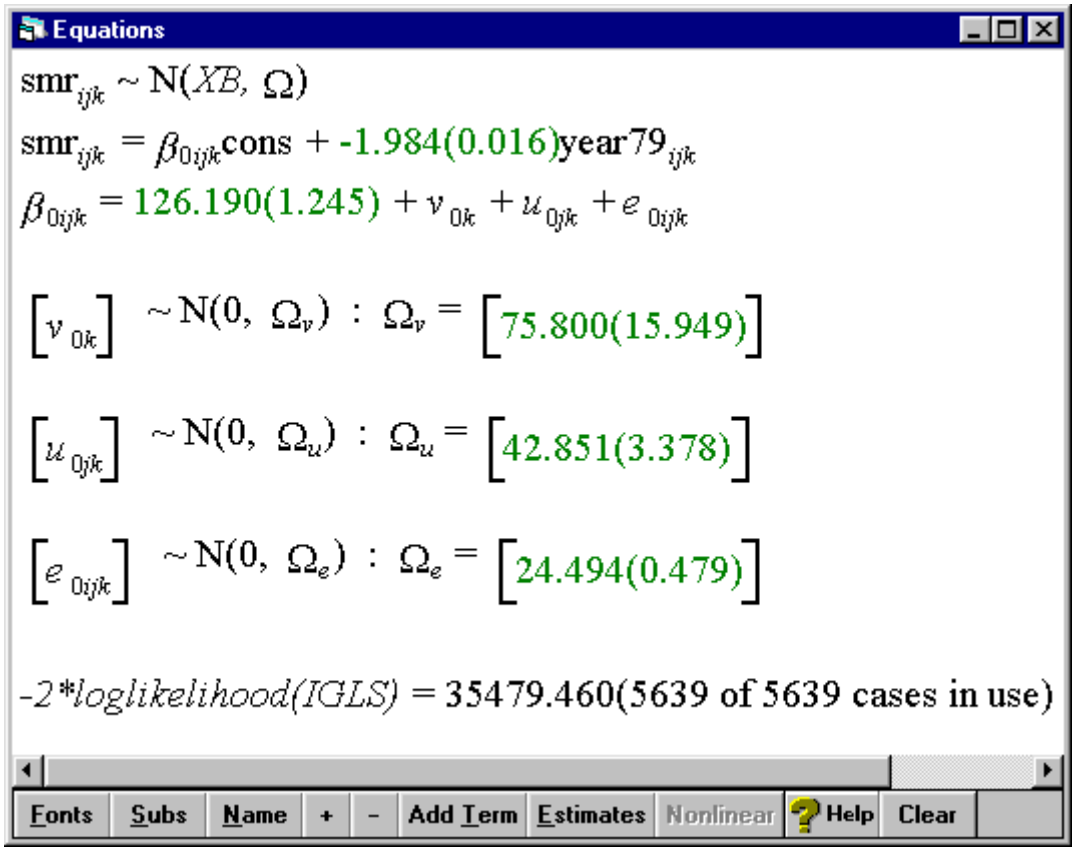
Click on $\boldsymbol{\beta_{0ij}}$

Check the box by **k(COUNTY)**

Click **Done**

Click **More** to estimate the new model

The intercept term now has an additional subscript to indicate that it varies across COUNTY as well as across DISTRICTs and YEARs. The terms $\mathbf{v_{0k}}$ are level-3 random effects and are again assumed to arise from a normal distribution.



The fixed part is unchanged as is the level 1 (between years within DISTRICTs) variance. However, the higher level variance has been partitioned further into that attributable to COUNTYs and that due to differences between DISTRICTs within COUNTYs. About 53% of the total variation can be seen to be between COUNTYs with 30% between DISTRICTs and just 17% between YEARs.

INTERPRETING THE MODEL

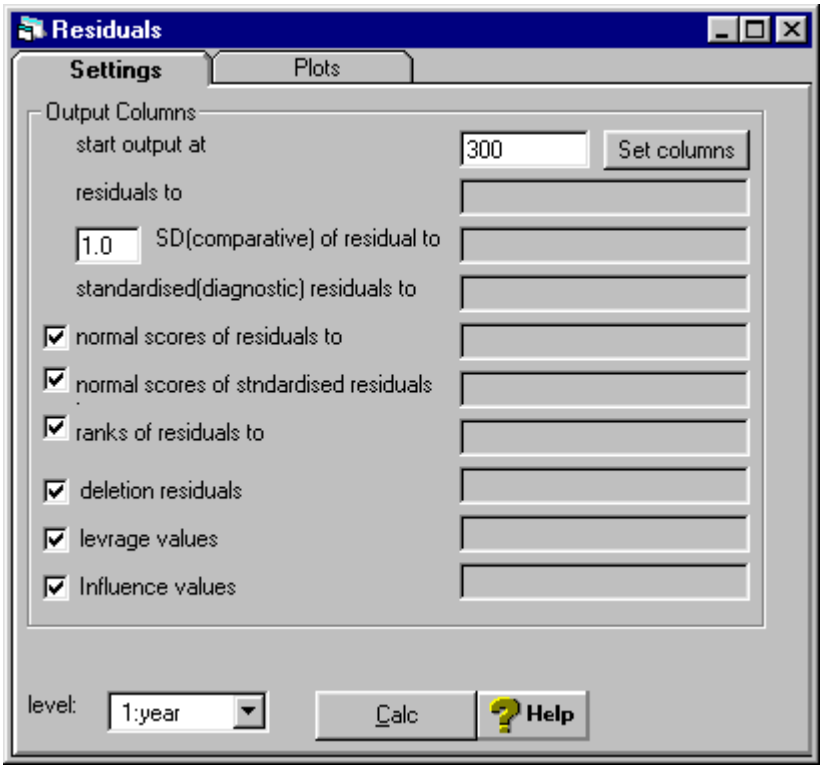
Residuals

In an ordinary least squares (OLS) regression equation, the residual or error term is the difference between the observed and fitted values. In the above model, the equation may be written as

$$y_{ijk} = \left(\beta_0 x_0 + \beta_1 x_{1ijk}\right) + \left(v_{0k} x_0 + u_{0jk} x_0 + e_{0ijk} x_0\right)$$

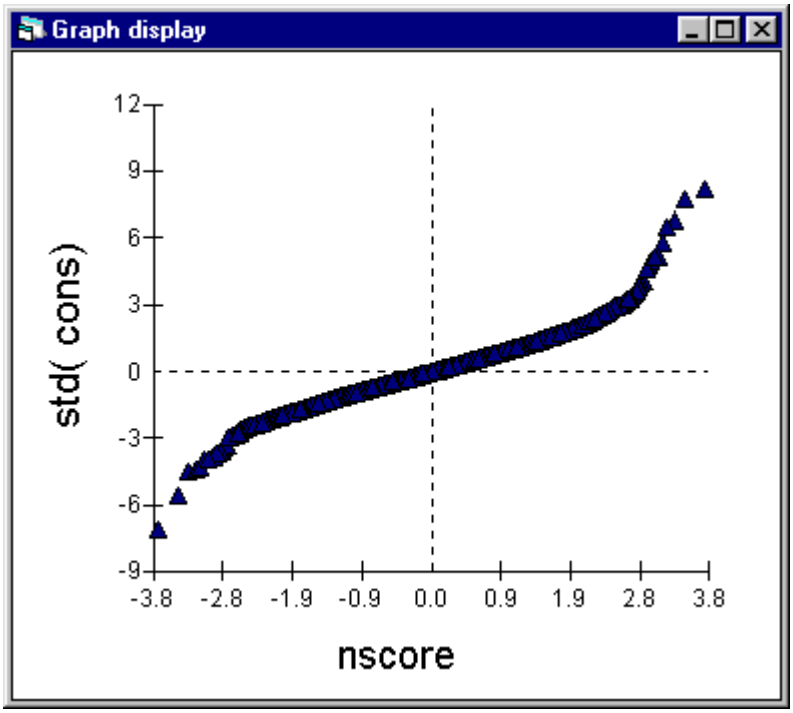
The terms inside the first set of brackets comprise the fixed part of the model, i.e. the fitted values for all data points. The terms inside the second set of brackets comprise the random part of the model and describe the variation around the fitted values at each level of the hierarchy.. Thus, the difference between the observed and fitted values is comprised of residuals at three levels – the $\mathbf{v_{0k}}$, $\mathbf{u_{0jk}}$ and $\mathbf{e_{0ijk}}$ in the regression equation. (Remember that $\mathbf{x_0}$ is the variable CONS i.e. it takes the value 1 for every observation.) Each set of residuals is assumed to follow a Normal distribution and this assumption may be checked in the same way as residual diagnostics in OLS. First consider the residuals at level 1.

Go to the **Models** menu Select **Residuals**



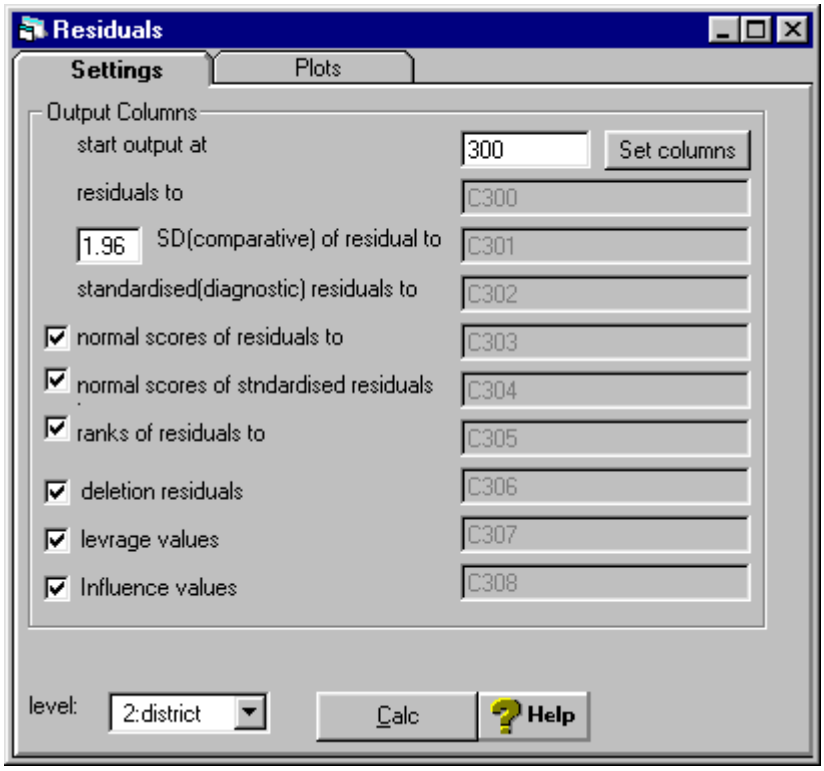
There are a variety of options which allow a range of standard diagnostic checks to be carried out – for example, to check the normality of the data or to look for outliers. By default all 9 functions are calculated and the results are stored in columns c300-c308. The box in the bottom left corner specifies the level at which the residuals are calculated; the default is level 1.

Click on the **Set columns** button
Click **Calc**
Select the **Plots** tab at the top of the **Residuals** window
Select the first option **standardised residual x normal scores**
Click **Apply**.

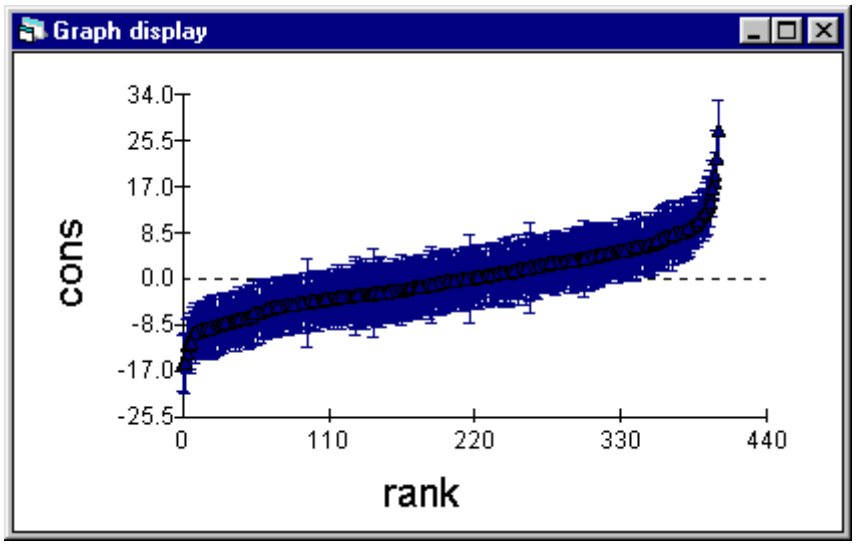


The points in the resulting graph should lie on a straight line; the fact that they don't suggests that there may be some departure from Normality. For the moment we will ignore this and look at the residuals at level 2 (DISTRICT).

Click on the **Settings** tab in the **Residuals** window
Select **2:DISTRICT** to be the **level** at which the residuals are calculated
Change the multiplier in the box by **SD(comparative) of residual** to 1.96
Click on **Set columns**
Click **Calc**



Select the **Plots** tab Choose a plot of **residual +/-1.96 sd x rank** Click on **Apply**

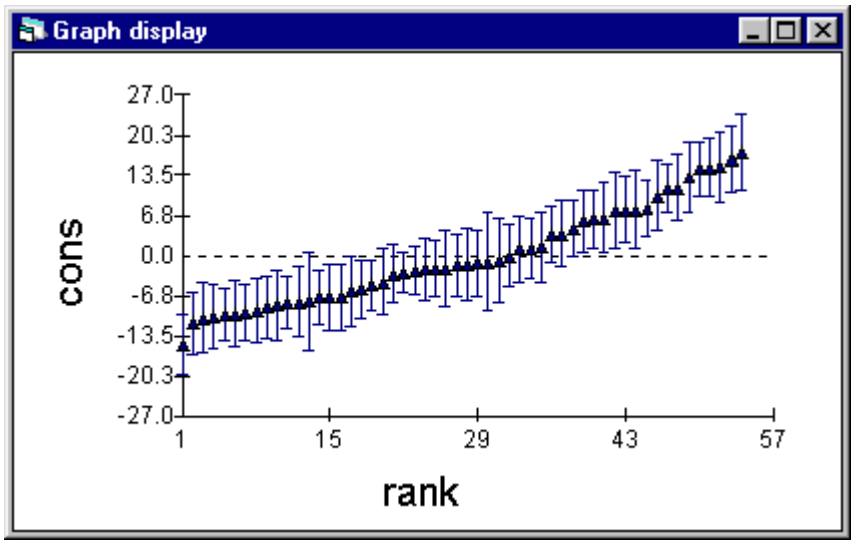


This plot shows the residuals or random effects for each DISTRICT, ordered from those with the smallest residuals on the left to those DISTRICTs with the largest residuals on the right. The range of values is from a reduction in the SMR of 16 points to an increase of 28 points. Since there is another level above DISTRICT, that of COUNTY, the residuals do not represent differences from the national

average but from the COUNTY average. The residuals are accompanied by error bars of half-width 1.96 S.D.; a DISTRICT whose error bar does not cross the horizontal line through zero has an SMR which is significantly different from the COUNTY average.

Finally, consider the residuals at level 3 (COUNTY).

Click on the **Settings** tab in the **Residuals** window
Select **3:COUNTY** to be the **level** at which the residuals are calculated
Ensure the multiplier in the box by **SD(comparative) of residual** is set to 1.96
Click on **Set columns**
Click **Calc**
Select the **Plots** tab
Choose a plot of **residual +/-1.96 sd x rank**
Click on **Apply**



The range of values of the COUNTY residuals is from a reduction in SMR of 15 points to an increase of 17 points. Although this is not as great as the range that is apparent among the DISTRICTs, bear in mind that there are considerably fewer COUNTYs than DISTRICTs (54 as opposed to 403). Thirty-three of the COUNTYs have residuals which are significantly different from zero. Note that not all DISTRICTs within these COUNTYs need have SMRs which are significantly different from 100; a REGION with a positive residual may contain DISTRICTs with negative residuals because the components of the composite random part – $\mathbf{u}_{0jk}$ and $\mathbf{v}_{0k}$ – are assumed to be independent.

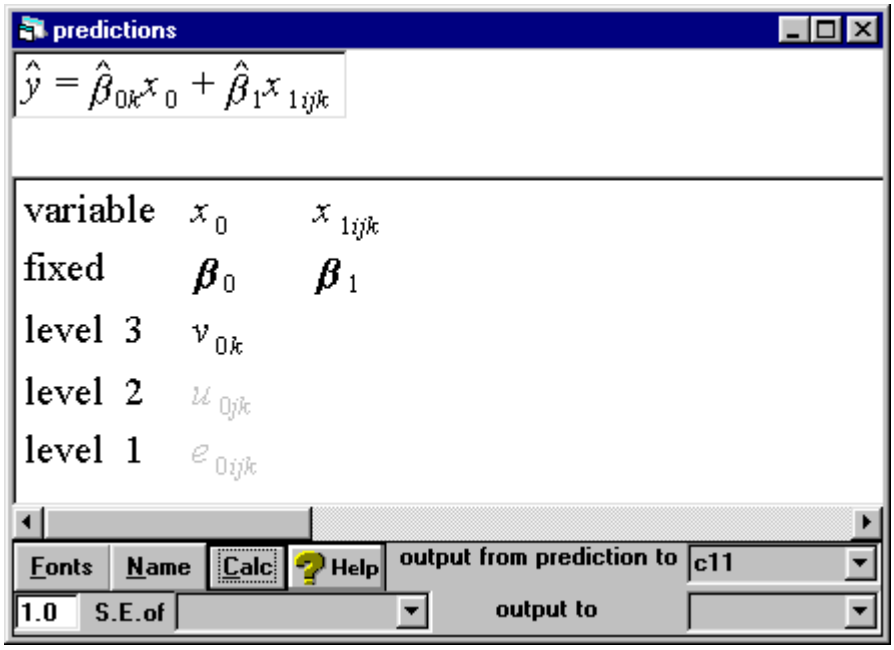
Predictions window

A number of different predictions may be made from a multilevel model depending on whether one includes fixed effects only or a combination of fixed and random effects. For example, prediction lines for COUNTYs are derived from the fixed part of the model together with the residuals from the appropriate level – the $\mathbf{v_{0k}}$ in this case.

Go to the **Model** menu Select **Predictions**

The elements of the model are arranged in two columns, one for each explanatory variable. Initially, all the terms are in grey indicating that they have not been selected and are not included in the prediction equation at the top of the **Predictions** window. The prediction equation is built by selecting the appropriate terms; clicking on the variable name at the head of the column ($\mathbf{x_0}$ or $\mathbf{x_{1ijk}}$) selects all the terms in that column, whilst clicking on individual terms (such as β_0 or $\mathbf{v_{0k}}$) toggles them in or out of the prediction equation.

Click on $\mathbf{x_0}$, $\mathbf{x_{1ijk}}$, then $\mathbf{u_{0jk}}$ and $\mathbf{e_{0ijk}}$ to remove them from the prediction In the drop-down list by **output from prediction to** select C11 Click on **Calc**



The results from this prediction are now in C11; using the **Names** window, name this variable PRED3 to indicate that it is a prediction including the level 3 (COUNTY) random effects. (Note: if the **Names** window was already open then column C11 may still appear to be empty; click on the **Refresh** button at the top of the **Names** window to overcome this problem.) Now plot the predicted values for each COUNTY against YEAR

Go to the **Graph** menu
Select **Customised Graph**

Note that details of earlier graphs are still held. D1 contains plots of the crude data while D10 contains the plot of residuals carried out in the previous section. To create a new graph

Select **D2** from the list
Select the **y** variable to be PRED3
Select the **x** variable to be YEAR
Select **group** to be COUNTY
Select **plot type** to be **line**
Click the **Apply** button

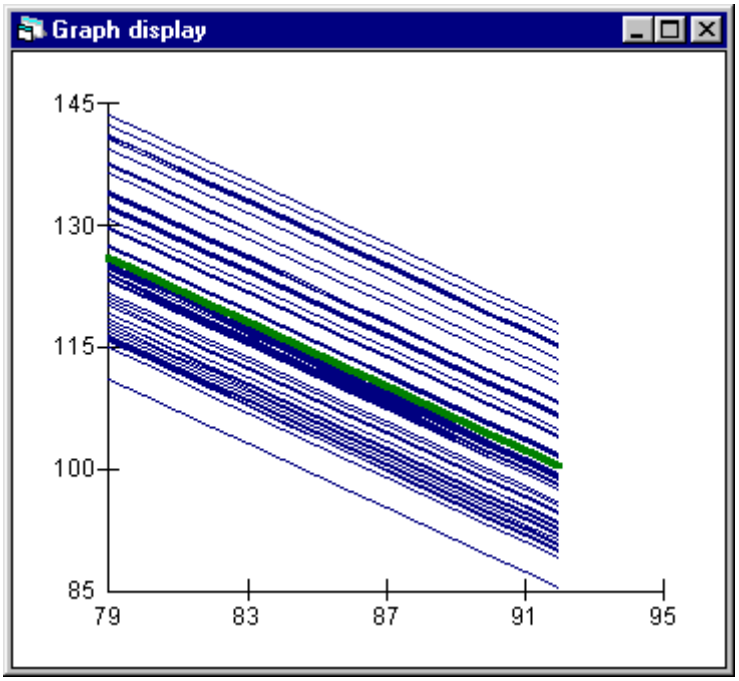
This produces a plot of 54 parallel lines, one for each COUNTY. We will superimpose on this graph the prediction of the fixed part of the model, the mean line given by

$$\hat{y}_{ijk} = \hat{\beta}_0 x_0 + \hat{\beta}_1 x_{1ijk}$$

Return to the **Predictions** window
Click on $\mathbf{v_{0k}}$ to remove it from the prediction equation
In the drop-down list by **output from prediction to** select C12
Click on **Calc**

In the **Names** window, change the name of C12 to PREDFP to indicate that it is a prediction from the fixed part only.

Return to the **Customised Graph** window
Ensure **D2** is selected
Under **ds #** (dataset number) click on number **2**
Select the **y** variable to be PREDFP
Select the **x** variable to be YEAR
Select **plot type** to be **line**
Click the **plot style** tab
Change the **colour** to **2 green**
Change the **line thickness** to **3**
Click the **Apply** button



The national mean SMR is highlighted in green with the predicted mean for each COUNTY shown around it. The lines are all parallel since the effect of each COUNTY, $\mathbf{v_{0k}}$, is assumed to be the same throughout the study period. This residual is the horizontal distance between the national mean and the COUNTY mean; a positive value of $\mathbf{v_{0k}}$ indicates the COUNTY mean SMR is greater than the national mean.

Now look at the predicted means for DISTRICTs within a specific COUNTY. First we need to generate the predicted values for each DISTRICT:

$$\hat{y}_{ijk} = \hat{\beta}_{0jk} x_0 + \hat{\beta}_1 x_{1ijk}$$

Return to the **Predictions** window
Click on $\mathbf{v_{0k}}$ and $\mathbf{u_{0jk}}$ to add them to the prediction equation
In the drop-down list by **output from prediction to** select C13
Click on **Calc**

In the **Names** window change the name C13 to PRED2 to indicate that these predicted values include the level 2 (DISTRICT) random effects.

To illustrate the different prediction lines in a single chart, select a single COUNTY, e.g. COUNTY number 1. To create an indicator for COUNTY number 1:

Go to **Data manipulation** menu
Select **Calculate**
Select the empty column C14 from the list of variables and press the right arrow button
Click on the = button on the window's keypad
Select COUNTY from the list of variables and press the right arrow button
Use the window's keypad to enter **==1**
Press **Calculate**

Note the logical command "==" (two equals signs) means "is equal to". This will create a dummy variable with the value 1 if the data are from COUNTY number 1, 0 otherwise.

Go to the **Names** window and change the name of C14 to COUNTY1.

Return to the **Customised Graph** window

Ensure **D2** is selected

Highlight data set number **1** under **ds #**

Select the **filter** to be COUNTY1

Click on the **plot style** tab

Change the **line thickness** to **3**

Click the **Apply** button

The resulting graph now has just two lines – one for the national mean and one for the selected COUNTY. To plot the predicted lines for the DISTRICTS in COUNTY number 1

Return to the **Customised Graph** window

Select **ds # 2**

Select the **filter** to be COUNTY1

Click the **Apply** button

Under **ds #** click on number **3**

Select the **y** variable to be PRED2

Select the **x** variable to be YEAR

Select the **filter** to be COUNTY1

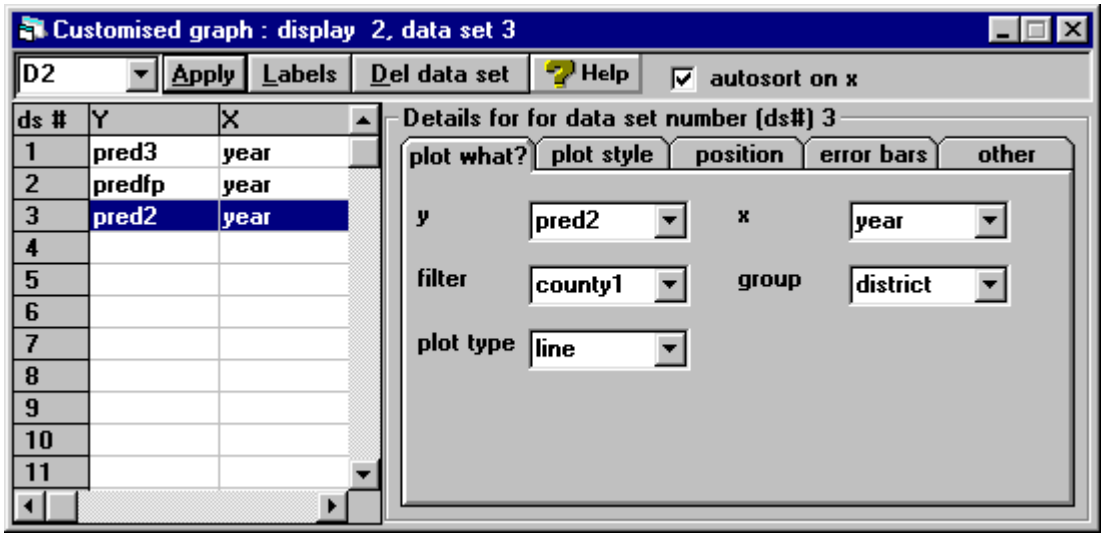
Select **group** to be DISTRICT

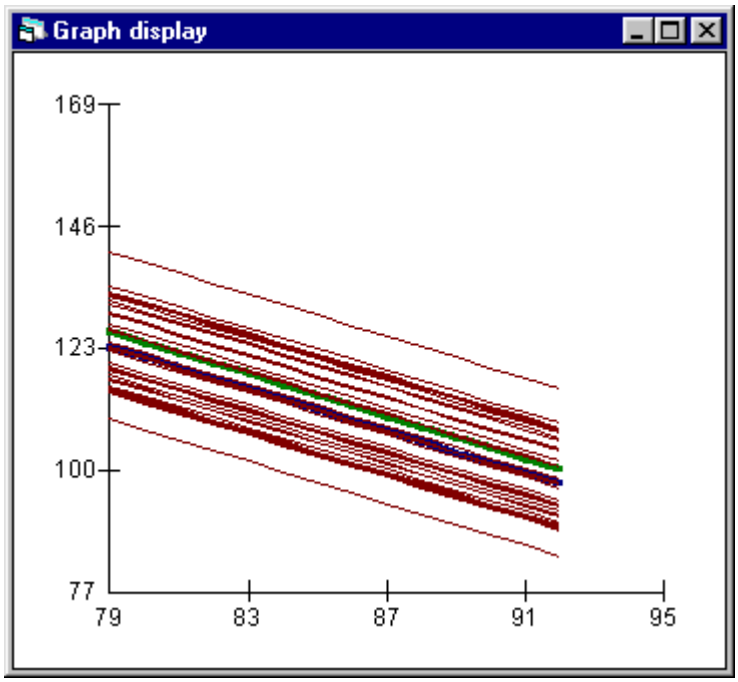
Select the **plot type** to be **line**

Click the **plot style** tab

Change the **colour** to **4 red**

Click the **Apply** button





In addition to the national mean (green) and COUNTY mean (blue) the graph now displays the DISTRICT means for the selected COUNTY. The vertical distance between the green and blue lines is the level 3 (COUNTY) residual $\mathbf{v_{01}}$ (the subscript k is replaced by the number of the COUNTY). The fact that the COUNTY mean is below the national mean indicates that this residual is negative. The vertical distance between each DISTRICT mean and the COUNTY mean is the level 2 (DISTRICT) residual $\mathbf{u_{0j1}}$. The vertical distance between each DISTRICT mean and the national mean is then the composite residual $\mathbf{v_{01}} + \mathbf{u_{0j1}}$. You may note that, despite the average for this COUNTY being below the national average, some of the DISTRICT means still lie above the national average (the green line) because the composite residual $\mathbf{v_{01}} + \mathbf{u_{0j1}}$ is greater than zero.

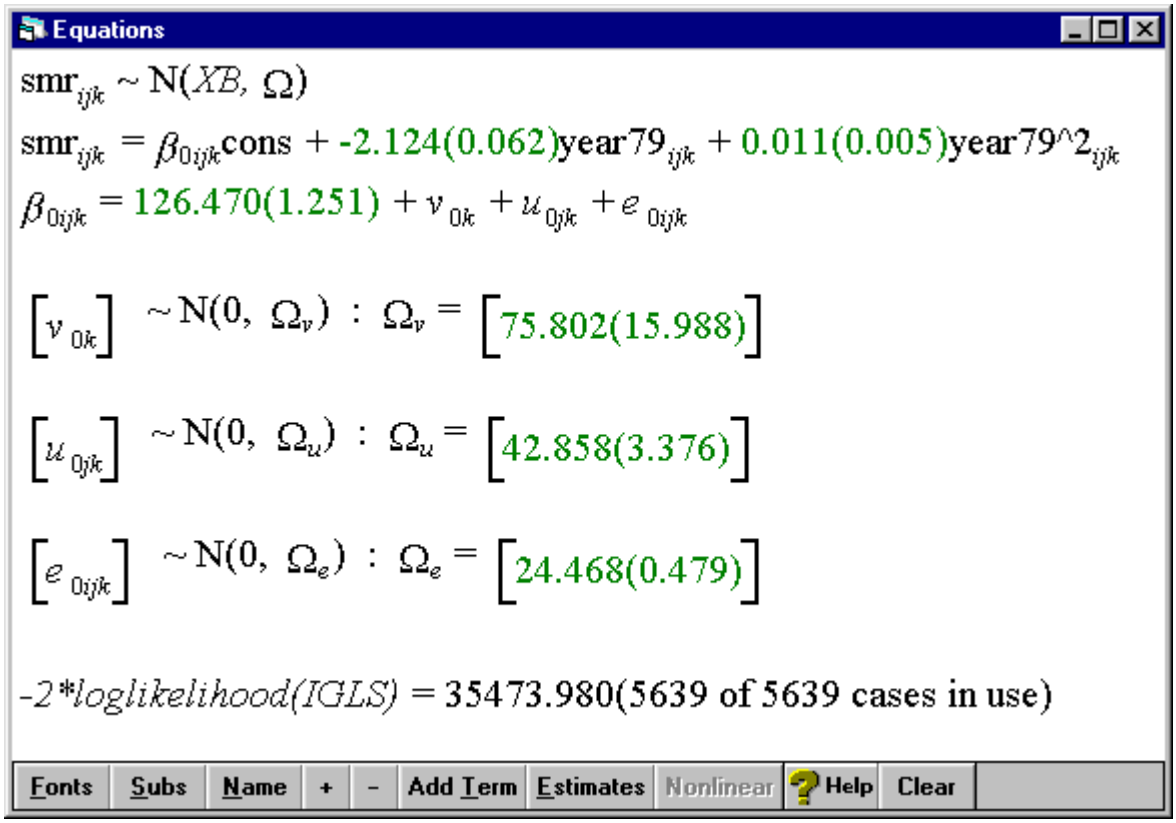
Adding more fixed effects

The models fitted so far include only an intercept term (CONS) and a trend coefficient (YEAR) in the fixed part. Now consider the addition of further variables. Firstly, add a quadratic term in year since the assumption of a linear trend may be too simplistic.

Go to **Data manipulation** menu
Select **Calculate**
Select the empty column C15 from the list of variables and press the right arrow button
Click on the = button on the window's keypad
Select YEAR79 from the list of variables and press the right arrow button
Use the window's keypad to enter **^2**
Press **Calculate**

In the **Names** window change the name of C15 to YEAR79^2.

Return to the **Equations window**
Click on **Add Term**
Click on $\mathbf{x_2}$ and select YEAR79^2, then click on **Done**
Click on the **More** button to re-estimate the model



The reduction in (-2*loglikelihood) is 5.48 from 1 degree of freedom so this term has significantly improved the fit of the model. The addition of this term has, however, done nothing to reduce the variance at any of the three levels in the model.

The next covariate to consider adding to the fixed part of the model is the variable FAMILY, a classification of the DISTRICTs into different types. It is a categorical variable with 6 categories and in order to model the effect of this covariate on the average SMR, it is necessary to create six dummy variables, one for each DISTRICT classification. The easiest way of entering categorical variables such as this is to use the **Main Effects and Interactions** window.

Under the **Model** menu select **Main Effects and Interactions**
Under **Categorical variable**, click on **[none]** and select FAMILY from the pull-down list
Click **OK** on the dialog box to open the **Set category names** window
Click on **0** under **name** and type LONDON
Click on **1** under **name** and type RURAL
Click on **2** under **name** and type PROSPER
Click on **3** under **name** and type MATURE
Click on **4** under **name** and type URBAN
Click on **5** under **name** and type MINING

Click on **Apply** and then close the **Set category names** window by clicking on **Quit**

Under **Categorical variable**, click on **[none]** and select FAMILY from the pull-down list

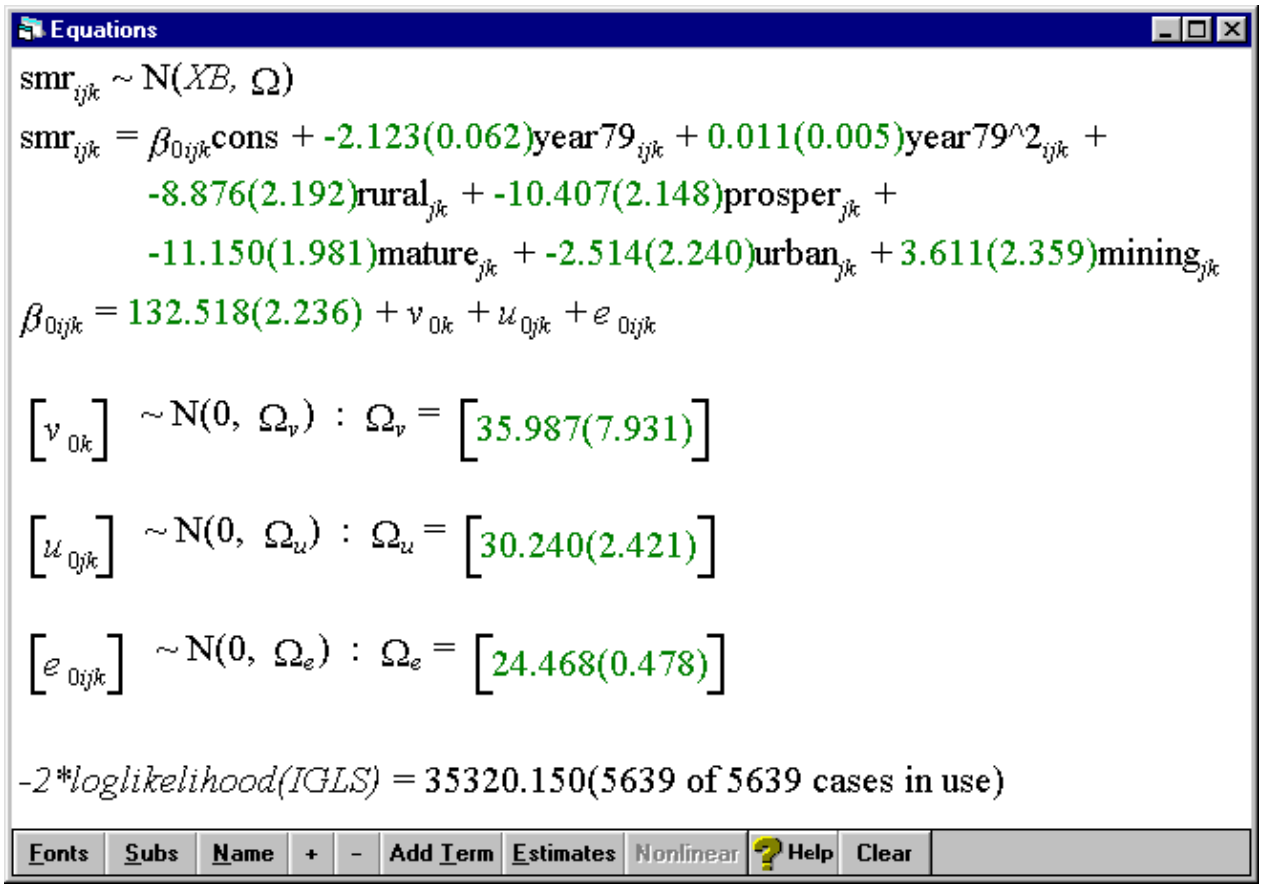
Under **View**, click on **Main effects**

Under **Main effects included**, click on each variable apart from LONDON

Click **Build**

Click on the **More** button to re-estimate the model

We have created six dummy variables named LONDON, RURAL etc., and each observation takes the value 1 if the DISTRICT is of that type, 0 otherwise. (The **Set category names** window may also be opened from the **Names** window by highlighting the relevant variable – in this case FAMILY – and clicking on the **Categories** button.) As with OLS regression, when a covariate has *n* categories, only *n-1* dummies are fitted in the model, the remaining category being used as a baseline against which comparisons are drawn. In this example, category 0, Inner London, will be used as the baseline.



The intercept or CONS term has changed as this is now the estimated mean in 1979 for areas in Inner London. There has been a significant reduction in –2*loglikelihood with the loss of just 5 degrees of freedom. The total variance has been reduced by 36.6% from 143 to 91; whilst the year-on-year (level 1) variation has changed little, the between DISTRICT (level 2) variance has been reduced by 29%

and the between COUNTY (level 3) variance by 53%. The addition of a level 2 variable has then had the greatest effect on the apparent variation between level 3 units, indicating that to a large extent there is homogeneity of the type of DISTRICT found within each COUNTY. (This is not surprising; as an example, consider the fact that all of the DISTRICTs classified as being Inner London must lie within the same COUNTY i.e. London.).

Intervals and tests window

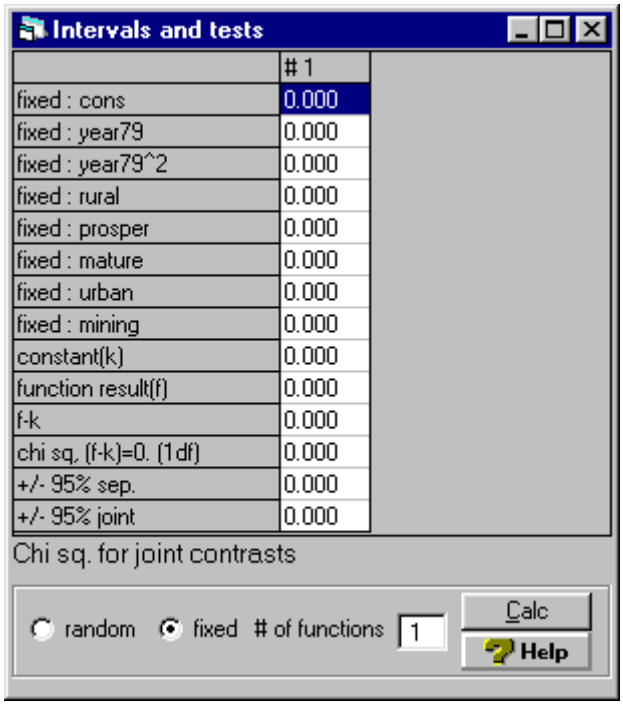
So far the change in likelihood has been used to assess improvement in the fit of the model to the data. It is also possible to carry out hypotheses tests for either fixed or random parameters using the **Intervals and tests** window. To illustrate how tests are formulated, consider the following two hypotheses. Firstly, if we are interested in testing whether SMRs in urban DISTRICTs are the same as those in Inner London then, since Inner London is the baseline category, this is equivalent to testing whether the coefficient for URBAN is significantly different from 0, i.e.

Hypothesis 1: $\beta_6=0$

We are not limited to single parameter tests but can also formulate significance tests involving a function of two or more parameters, as well as joint significant tests involving two or more functions of the model parameters. For example, consider a test of the hypothesis that SMRs in rural, prospering and mature DISTRICTs are the same, i.e.

Hypothesis 2: $\beta_3=\beta_4=\beta_5$ *or*
$(\beta_3-\beta_4=0)$ *and* $(\beta_4-\beta_5=0)$ *and* $(\beta_3-\beta_5=0)$

Go to **Model** menu
Select **Intervals and tests**
Select **fixed** at the bottom of the window



The **# of functions** relates to the number of functions or contrasts of the parameter estimates being tested under a single hypothesis; for hypothesis 1 only one function is necessary while three functions are required for hypothesis 2. The boxes beside each fixed parameter are used to enter the function of the parameters to be tested, while the constant (k) contains the value to which the function is compared which, in both of the following cases, is the default value zero. So for hypothesis 1:

Select the box beside **fixed : urban**
Type 1
Press **Calc**

Note that the function **f** is a single multiple of the URBAN parameter and so equals β_6, and because k=0, (f–k) also equals the parameter β_6. The test statistic, based on Wald's Test, appears in the bottom half of the window, **joint chi sq test(1df)=1.260**, and this may be compared to a chi-squared distribution to either accept or reject the hypothesis that $\beta_6=0$. We can obtain the p-value for the chi-squared or other distributions using the **Tail Areas** window:

Go to the **Basic statistics** menu
Select **Tail areas**
Select **Chi squared** under **Operation**
Enter the test statistic value **1.260** in the box beside **Value**
Enter the value **1** in the box beside **Degrees of freedom**

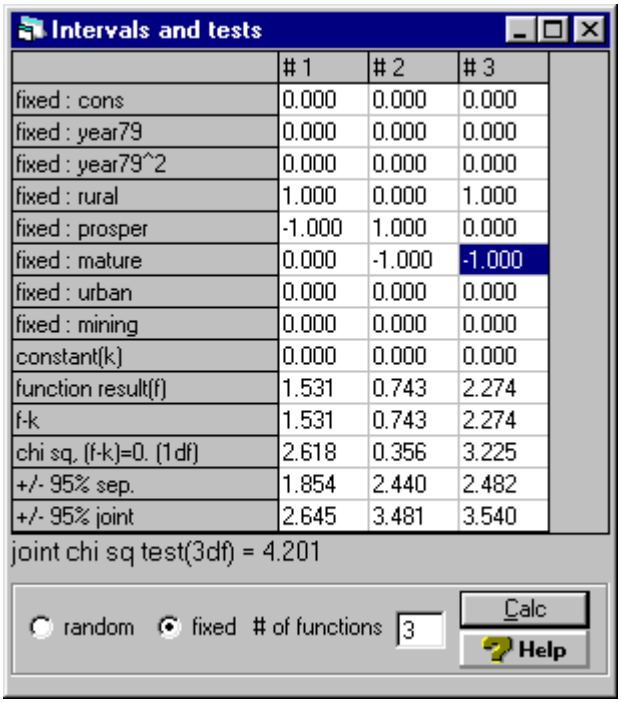
Press **Calculate**

The probability value of 0.26165 appears in the output window which should open automatically. In this case, $1.260<\chi^2(1, 0.95)=3.84$, so we do not reject the hypothesis that the mean SMR is the same in Inner London and urban DISTRICTs.

Now to formulate a test for *Hypothesis 2* (if the **Intervals and tests** window is still open, close it down and open it again to erase details of the previous test).

Ensure **fixed** is selected at the bottom of the **Intervals and tests** window
Change the **#of functions** to **3**
In the first column, enter a **1** beside **fixed:rural** and a **–1** beside **fixed:prosper**
In the second column, enter a **1** beside **fixed:prosper** and a **–1** beside **fixed:mature**
In the third column, enter a **1** beside **fixed:rural** and a **–1** beside **fixed:mature**
Press **Calc**

Each column specifies a function of the parameters which is compared to **constant (k)** equal to zero; for example, in column 1, the function is $(1 \times \beta_4)–(1 \times \beta_5)= 0$ (i.e. $\beta_4=\beta_5$).



We return to the **Tail areas** window

Go to the **Basic statistics** menu

Select **Tail areas**

Select **Chi squared** under **Operation**

Enter the test statistic value **4.201** in the box beside **Value**

Enter the value **3** in the box beside **Degrees of freedom**

Press **Calculate**

This time we are jointly testing 3 functions and therefore require 3 degrees of freedom. The resulting p-value of 0.24056 indicates that we cannot reject the hypothesis that the mean SMRs of RURAL, PROSPER and MATURE are the same.

In practice at this stage we might want to collapse the variable FAMILY into just 3 categories: a baseline category comprising Inner London and Urban areas, achieved by deleting the variable URBAN from the current model, and a combination of Rural, Prospering and Maturer areas, which would involve creating a new variable using the **Calculate** window and replacing the variables RURAL, PROSPER and MATURE in the model with this new variable. However, we will continue for the moment with all six categories.

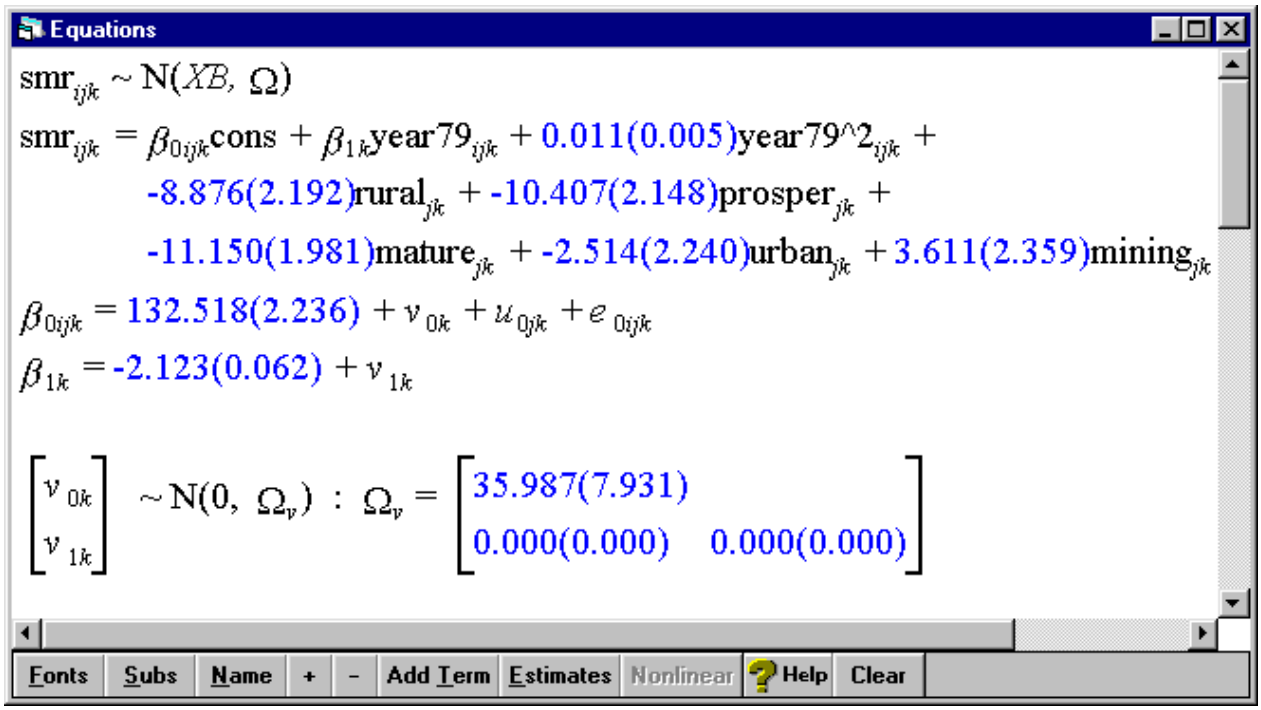
RANDOM COEFFICIENTS

We now consider another important class of multilevel model; random coefficients. In variance components models only the intercept is considered random; however, in the following model we will also allow the slope to vary across higher levels.

Random slopes

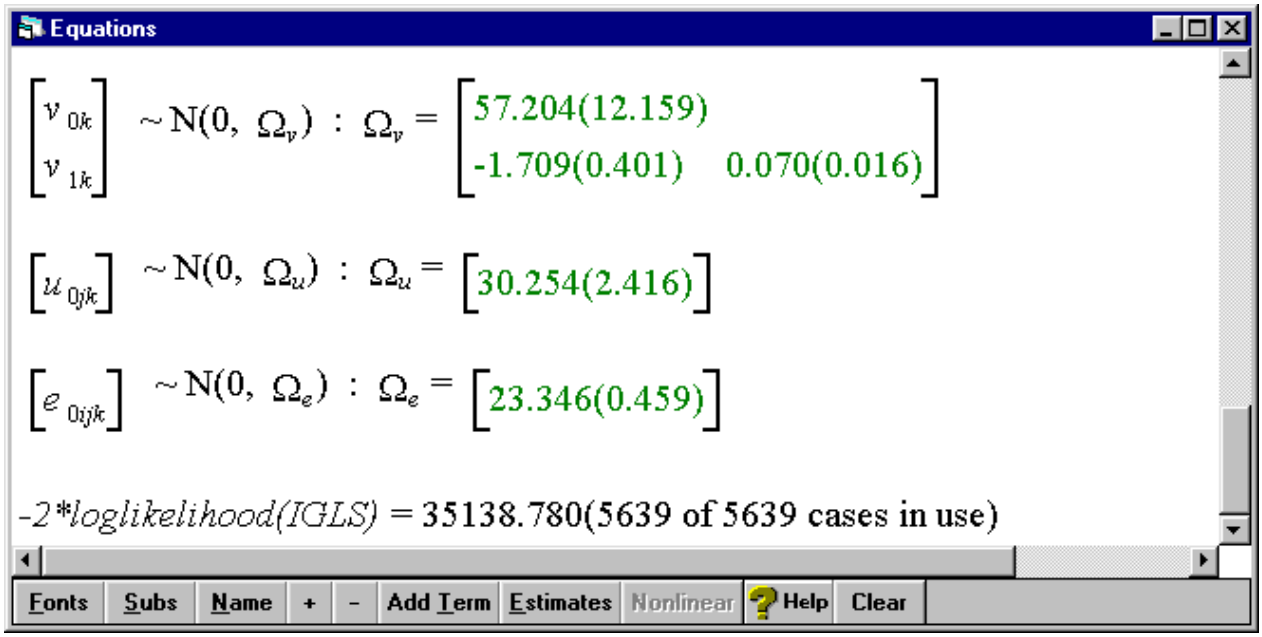
The following section considers the possibility that the rate at which the SMRs have been decreasing may vary from one COUNTY to another. The models fitted so far have contained random intercepts for both COUNTY and DISTRICT; however, the following model will also consider random slopes across the level 3 units (COUNTYs).

Return to the **Equations** window Click on **year79$_{ijk}$** and check the box by **k(COUNTY)** Then click **Done**



The coefficient of **year79$_{ijk}$** has changed from β_1 to β_{1k} indicating that this parameter now varies randomly across COUNTYs. The estimate of β_{1k} is now given as a mean β_1, common to all

COUNTYs, plus a level 3 residual $\mathbf{v_{1k}}$, unique to the k[th] COUNTY. The level 3 residuals $\mathbf{v_{0k}}$ and $\mathbf{v_{1k}}$ now have a joint multivariate Normal distribution with variances σ_{v0}^2 and σ_{v1}^2 respectively and covariance σ_{v10}. Click on **More** to estimate this model.



There is little change in the fixed part of the model, nor in the level 1 or level 2 variances. There has, however, been a large reduction in the value of –2*log(likelihood). Therefore, the addition of random slopes has improved the overall fit of the model. The three random terms at level 3 now refer to the variance of the intercept (CONS) for COUNTYs – σ_{v0}^2, the variance of the slope (YEAR79) for COUNTYs – σ_{v1}^2, and the covariance between the two, σ_{v10}. Whilst the two additional random terms appear large compared to their standard error, it is possible to test this formally using the **Intervals and tests** window.

Go to **Model** menu
Select **Intervals and tests**
Select **random** at the bottom of the window
In the box beside **# of functions** type 2

There are two functions to test; our hypothesis is

Hypothesis 3: $\sigma_{v1}^2=\sigma_{v10}=0$

or

$\sigma_{v1}^2 =0$ and $\sigma_{v10}=0$

In the first column, enter a **1** beside **county:year79/cons**

In the second column, enter a **1** beside **county:year79/year79**

Press **Calc**

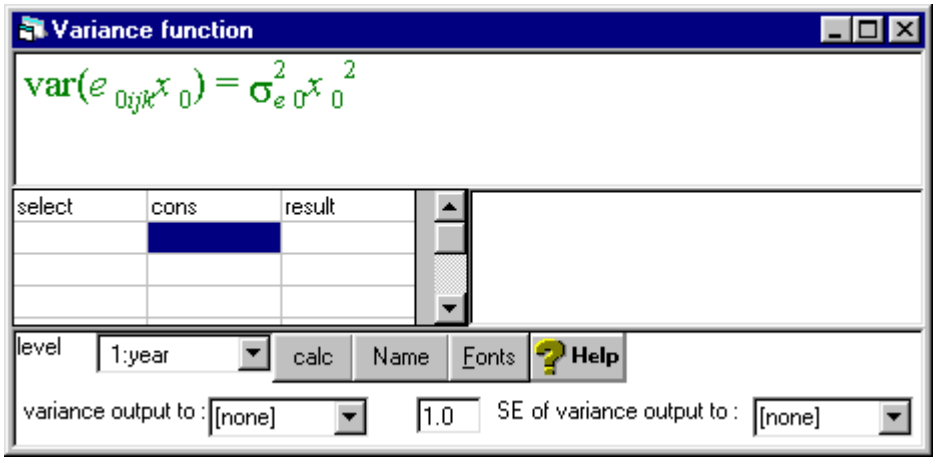
The value of 19.332 is extremely significant when compared with a chi-squared distribution with 2 degrees of freedom; we therefore reject the hypothesis that the two random terms are not significantly different from 0.

The level 3 variance is now more complex and more difficult to interpret; however, the **Variance function** window can be used as an aid.

Variance function window

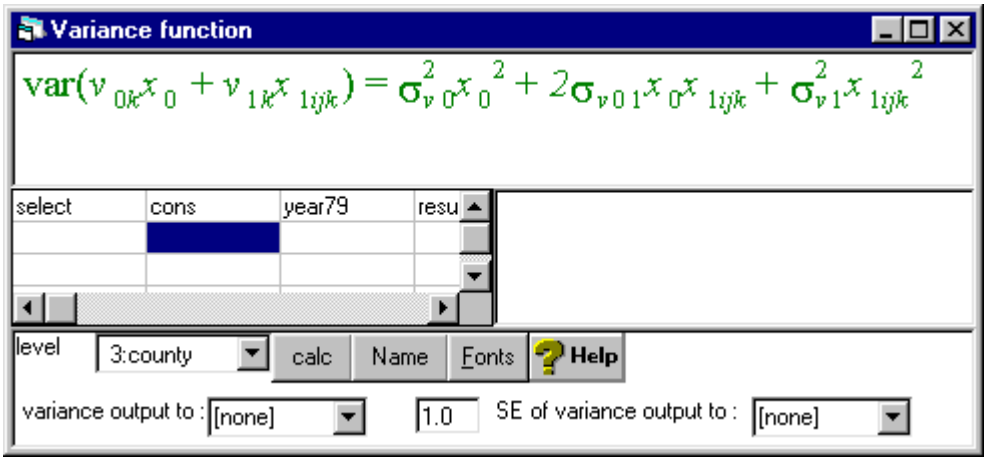
Go to **Model** menu

Select **Variance function**



The purpose of this window is to display and calculate the variance function at each level of the current model. The variance function for level 1 is currently shown; this only involves one term because the current model assumes that the level 1 variance is constant for all observations. To view the level 3 variance function:

In the drop-down list by **level** in the bottom left-hand corner, select **3:COUNTY**



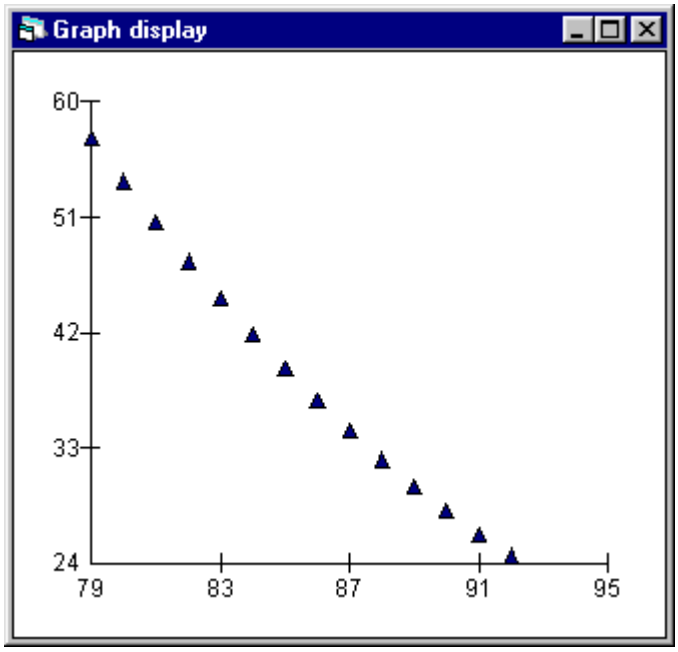
The current model has two terms random at level 3, the intercept and the slope, so the level 3 variance is a function of two random variables. The function shown is the variance of the sum of the two random terms $\mathbf{v_{0k}x_0}$ and $\mathbf{v_{1k}x_{1ijk}}$. Since $\mathbf{x_0}$ is just the CONStant term, taking the value 1, the level 3 variance is a quadratic in $\mathbf{x_{1ijk}}$ (YEAR79). We can use the **Variance function** window to calculate this function and use the **Graph** window to plot it.

Note that the columns in the table in the **Variance function** window named **select**, **cons**, **year79** and **result** allow us to estimate the variance function at specific values of YEAR79. However, rather than enter the values from 0 to 13 it is simpler to estimate the function for all data points.

In the drop-down menu by **variance output to**, select C21 **Click** calc

In the **Names window** name C23 VARF3. To plot the level-3 variance across the observed values of YEAR79:

Go to the **Customised Graph** window Select a new display **D3** Select **y** to be VARF3 Select **x** to be YEAR Click **Apply**



The level 3 (between COUNTY) variance has steadily decreased from a high of 57.2 in 1979 to a low of 24.6 in 1992. It therefore appears that absolute differentials between COUNTYs have been decreasing over time. Another way of examining this change is by looking at the prediction graphs. First calculate the predicted values using the random intercepts and slopes at COUNTY level:

Choose the **Predictions** window form the **Model** menu
Click on $\mathbf{x_0}$, $\mathbf{x_{1ijk}}$ and $\mathbf{x_{2ijk}}$ to ensure that they are included
Click on $\mathbf{u_{0jk}}$ and $\mathbf{e_{0ijk}}$ to remove them from the prediction
Select PRED3 for **output from prediction to**
Click **Calc**

Next re-calculate the predicted values using the fixed part of the model only:

Click on $\mathbf{v_{0k}}$ and $\mathbf{v_{1k}}$ to remove these terms from the prediction
Select PREDFP for **output from prediction to**
Click **Calc**

To plot these new predictions:

Return to the **Customised Graph** window
Select a new display **D4**
Under **ds # 1**

Select **y** to be PRED3

Select **x** to be YEAR

Select COUNTY as the **group**

Change **plot type** to **line**

Click **Apply**

Choose **ds # 2**

Select **y** to be PREDFP

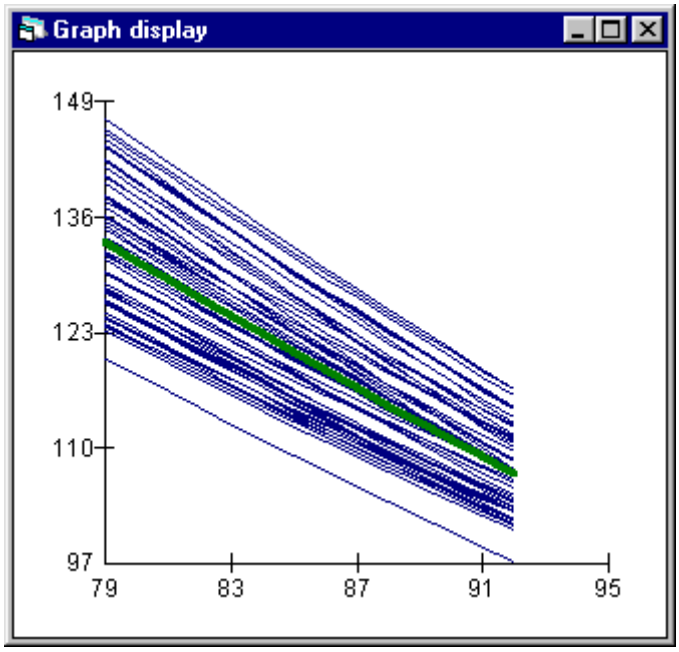
Select **x** to be YEAR

Change **plot type** to **line**

Under the **plot style** tab set **colour** to **2 green**

Set **line thickness** to **3**

Click **Apply**



The plot shows the individual predicted trends for each COUNTY plotted around the mean trend line shown in green. The fact that the COUNTY lines are converging toward the mean line over time demonstrates the decrease in level-3 variation over time.

Higher-level residuals

There are now two sets of residuals at COUNTY level.

Under the **Model** menu, open the **Residuals** window

Click on the **Settings** tab

Select the **level** to be **3:COUNTY**

Choose a multiplier of 1.96 for the **SD (comparative) of residual to**

Click on **Set columns**

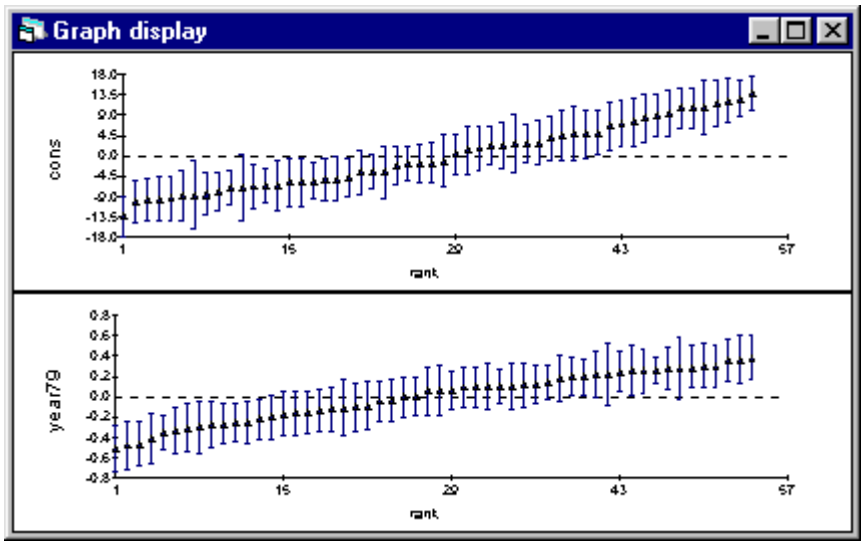
Each item in the list now has two columns assigned to it: the first column for the intercept residual, the second for the slope residual.

Click **Calc**

Select the **Plots** tab

Select **residual +/- 1.96 sd x rank**

Click **Apply**

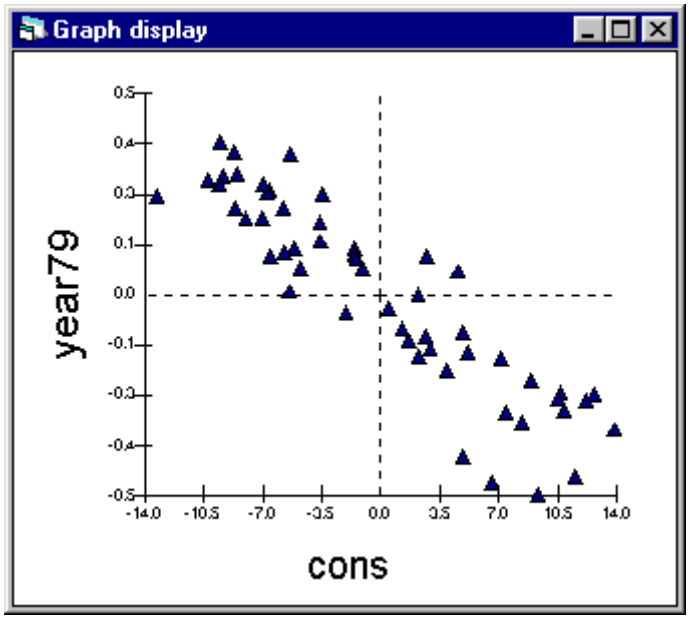


These plots can be used to examine how many COUNTYs have slopes which differ from the average as well as how many have intercepts which differ from the average. Note that a COUNTY's rank for the intercept residual will not necessarily be the same as its rank for the slope residual. To see how the intercept and slope residuals are correlated between COUNTYs:

Return to the **Plots** tab in the **Residuals** window

Under the **pairwise** heading, select a **residuals** plot

Click **Apply**



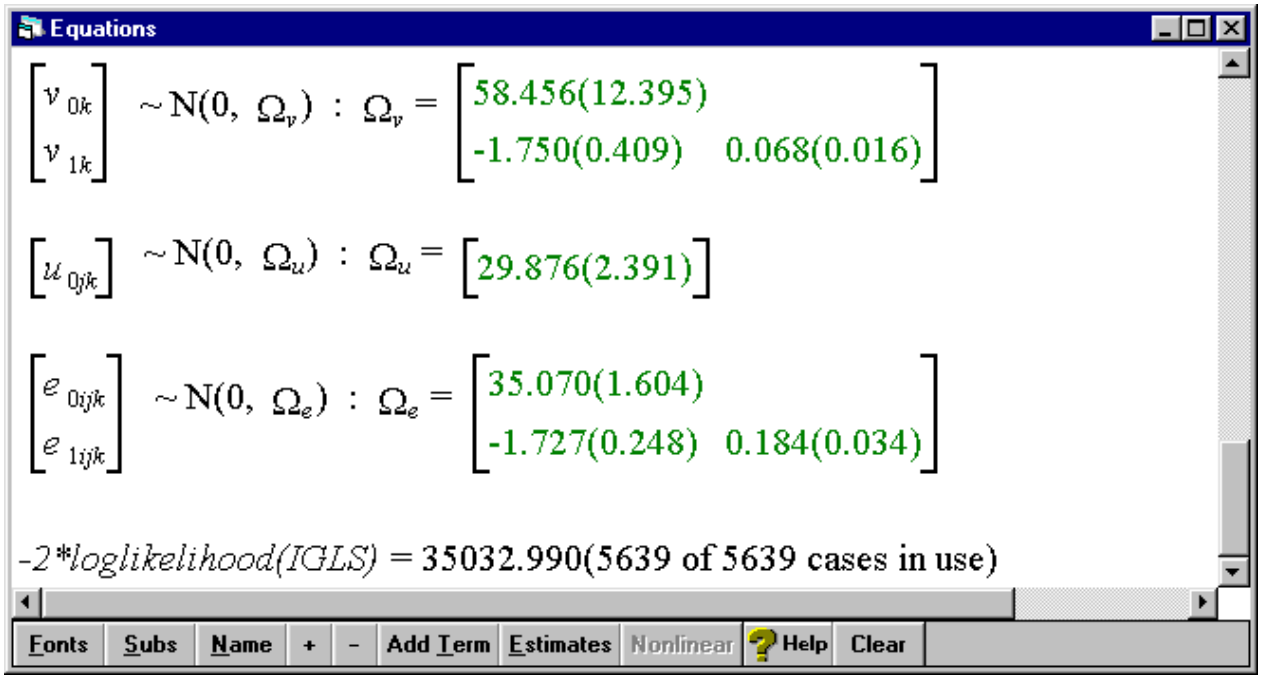
This shows the strong negative correlation between the two sets of residuals. Those in the top left quadrant refer to those COUNTYs with negative intercept (CONS) residuals and positive slope (YEAR79) residuals. This suggests that those COUNTYs which had the lower than average SMRs in 1979 experienced a more gradual decrease in SMR over the 14 YEARs. Similarly, the COUNTYs featured in the bottom right quadrant are those which had above average SMRs in 1979 (positive CONS residual) but which experienced mortality decreasing at a faster than average rate (negative YEAR79 residual).

Complex level 1 variation

The multilevel framework allows variables to be random at any level so, for example, we may wish to extend the previous model such that trends in SMR not only vary across COUNTYs but also vary across DISTRICTs at level-2. However, random variables at level 1 have a slightly different interpretation; this concerns the effects of heterogeneity (i.e. non-constant variance). In this example, we may consider whether the variation between observations is constant throughout the 14 years or whether it changes.

Return to the equations window click on **year79**$_{ijk}$ check the box at **i(year)** click done

Now estimate this model by clicking on the **More** button.



There is evidence of heterogeneity with a substantial reduction in –2*loglikelihood. This mean that the degree of scatter of individual observations about the predicted DISTRICT (level 2) means is not constant over time; it appears to have been decreasing. We can use the **Variance function** window to estimate the variance at each level, creating two new variables VARF2 and VARF1 and plotting these three variables against YEAR in the **Graph** window.

Open the **Variance function** window under the **Model** menu
Ensure that **1:YEAR** is selected to be the **level**
In the drop-down menu by **variance output to**, select C22
Click calc
Select **2:DISTRICT** to be the **level**
In the drop-down menu by **variance output to**, select C23
Click calc
Select **3:COUNTY** to be the **level**
In the drop-down menu by **variance output to**, select VARF3
Click calc

In the **Names window** name C22 VARF1 and C23 VARF2. To plot all of these variance functions across the observed values of YEAR79:

Go to the **Customised Graph** window

Select display **D3**

Select **y** to be VARF3

Select **x** to be YEAR

Select the **plot type** to be **line**

Under the **plot style** tab, select the **line thickness** to be **3**

Click **Apply**

Select **ds#2** with VARF2 as the **y** variable, YEAR as the **x** variable, and the **plot type** to be **line**

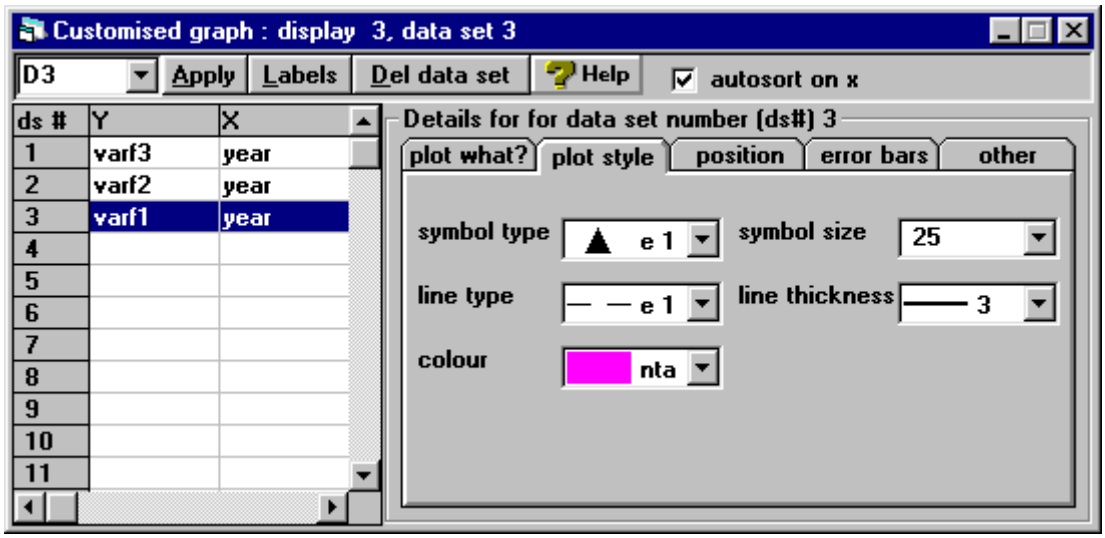
Under the **plot style** tab, select the **colour** to be **4 red** and the **line thickness** to be **3**

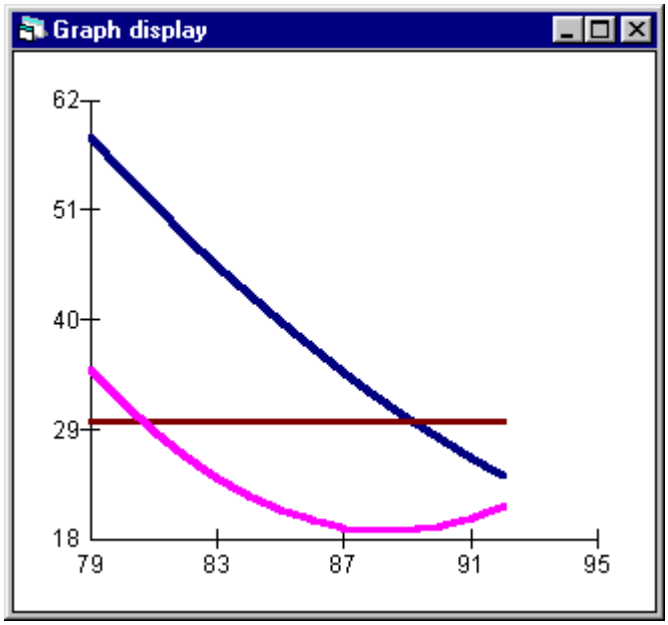
Click **Apply**

Select **ds#3** with VARF1 as the **y** variable, YEAR as the **x** variable, and the **plot type** to be **line**

Under the **plot style** tab, select the **colour** to be **13 Light magenta** and the **line thickness** to be **3**

Click **Apply**





We have not fitted any random effects at level 2, so the variation between DISTRICTs within COUNTYs is assumed to be constant. The variation between COUNTYs decreased steadily between 1979 and 1992; however, the level 1 variance decreased from 1979 to 1988 but it appears to have increased slightly since then. The total variation has decreased from 123 in 1979 to just 76 in 1992. In a similar manner it is possible to explore the extent to which the level 2 variation (between DISTRICTs) has also been varying over time.

By this stage the user has become familiar with the basics of model fitting for continuous (normally distributed) responses. The fixed part of the model can be built up as with an ordinary least squares (OLS) regression model, including any combination of continuous and categorical variables and interactions between them. The significance and effect of variables can be examined through changes in the likelihood or through comparisons of the parameter estimates with their estimated standard errors.

The difference between such models and OLS regression is the ability to separate the variance into the different levels in the model – COUNTY, DISTRICT and YEAR in this example – and then to model this variance by considering other variables to be random at any of the levels. At higher levels this has the interpretation of fitting random slopes; at the lowest level this is modelling heterogeneity (non-constant variance) within the data. We are again able to test for the significance of any of these random terms.

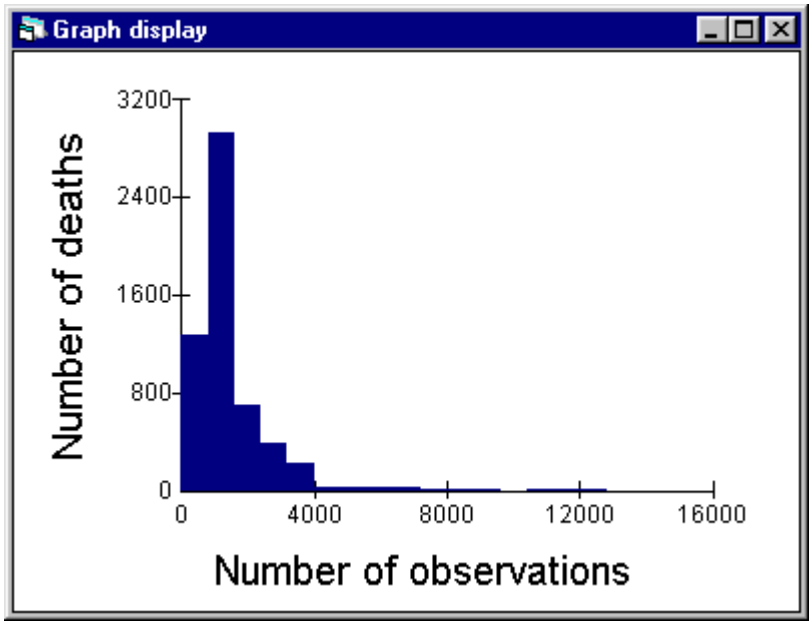
The example used has been illustrative of the methods employed when fitting a multilevel model; it is not, however, the way in which we would normally model such data. The next section goes on to

consider a generalised linear model for these data; however, before proceeding to the more complex modelling it is important to have a good understanding of the basics covered up to this point.

A POISSON MODEL – INTRODUCTION

The model that we have fitted assumes that the standardised mortality ratio follows a normal distribution. We found that the variance decreased over the period 1979-1992; over this time the standardised mortality ratio also fell. This suggests that there may be a link between the variance in a particular year and the average mortality rate in that year. We have also attached equal importance to every area and in every year; this is probably not sensible since the size of areas in terms of their populations and the number of deaths observed varies considerably both across areas and over time. One possibility would be to weight each observation according to the population of the district in that year; this requires weighting at each level of analysis and would ensure that areas from which we have the most information – the largest areas in terms of their populations – are afforded the most weight. In this section we adopt an alternative approach.

The local mortality datapack is based on **counts** of deaths. Instead of modelling a transformation of this response – the SMR – we can consider modelling the actual counts of deaths. Such data tend to be discrete rather than continuous – you can't observe fractions of deaths – and they also tend to be extremely skewed (see histogram below). Therefore, the assumption of a normal distribution is usually not sensible.



Instead we can fit a generalised linear model and approximate a Poisson distribution for the data.

First open the original worksheet LMDP.ws again.

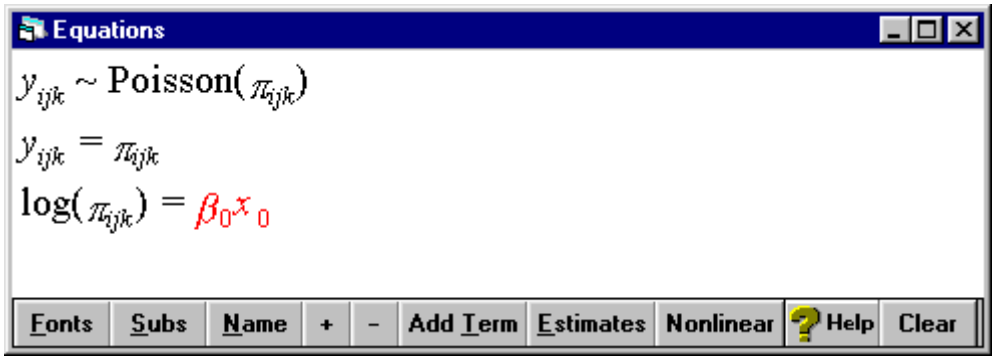
Go to the **File** menu
Select **Open worksheet**
Open the worksheet called **lmdp.ws**

Then go to the **Equations** window.

Click on either of the **y** terms
Select DEATHS as the dependent variable
Select 3 – ijk for **N levels**
Select COUNTY for **level 3(k)**
Select DISTRICT for **level 2(j)**
Select YEAR for **level 1(i)**
Click on **Done**

So far we have simply repeated the steps for the 3-level model in the introductory tutorial with the response variable being DEATHS rather than SMR. We now have to amend the default distribution for the response. In the **Equations** window

Click on the **N** that defines the Normal distribution
Select **Poisson** from the list
Change the **logit** link function to **log**



These steps have specified the response to be a Poisson random variable, which defines the lowest level variance function, and the linearising function of the response to be the natural logarithm. As in the normal multilevel model, we need to define a CONStant again.

Go to **Data manipulation** menu
Select **Generate vector**
Select **Type of vector** to be **Constant vector**
Select C9 to be the **Output column**
Enter 5639 (the number of data points) beside **Number of copies**
Enter 1 beside **Value**
Click the **Generate** button

The linearising function used when estimating generalised linear models means that the level one variance is on a different scale to the variance at higher levels. This means that we require two copies of the CONS column; one to model the level 1 variation and one to model the higher level variation.

Go to **Data manipulation** menu
Select **Calculate**
Select C10 and press the right arrow button
Click the '=' button on the keypad
Select C9 and press the right arrow button
Click the **Calculate** button

Open the **Names** window and name C9 and C10 CONS and PCONS respectively. Now return to the **Equations** window.

Click on $\beta_0 x_0$
Select CONS from the drop-down list
Click on the check box by **j(DISTRICT)**
Click on the check box by **k(COUNTY)**
Click on **Done**

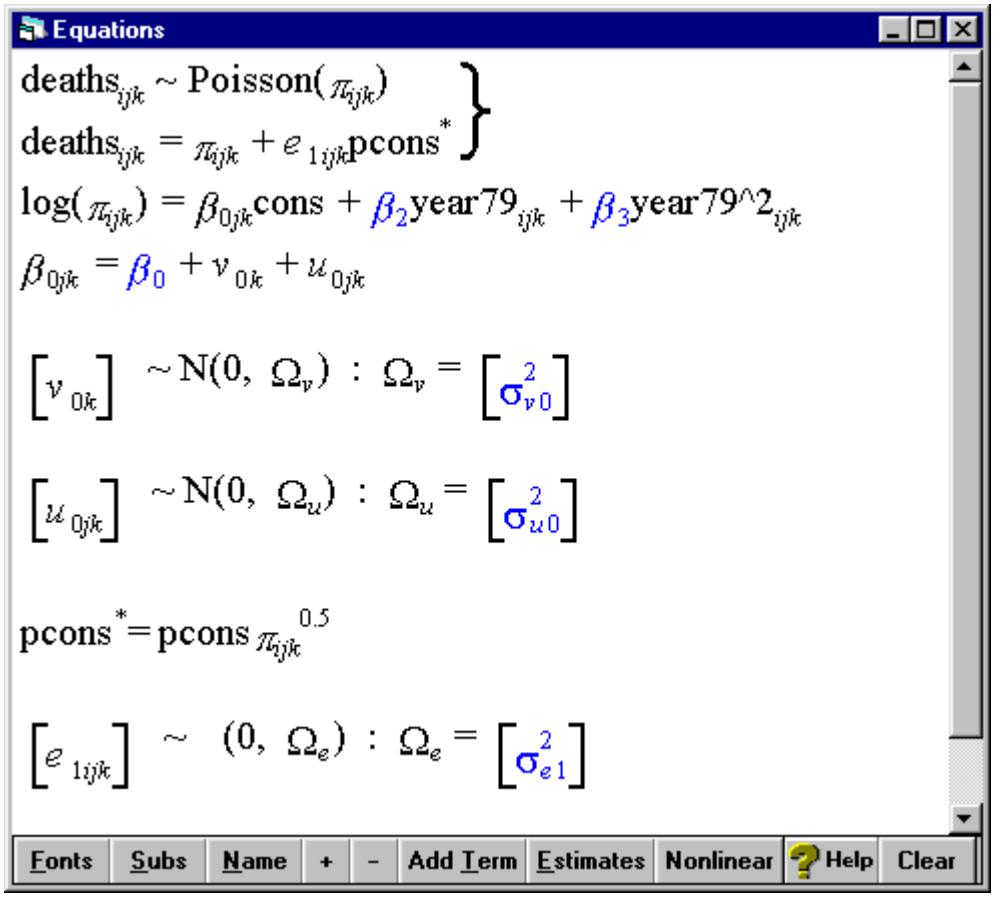
With these commands, we have added the constant to the fixed part of the model to estimate the intercept and have also allowed the intercept to vary across DISTRICTs and COUNTYs. Still in the **Equations** window:

Click on the **Add term** button
Click on the new explanatory variable and choose PCONS from the drop-down list
Remove this from the fixed part by clicking on the check-box by **Fixed parameter**
Click on the check box by **i(YEAR)**
Click on **Done**

These commands have added PCONS to the model in order to model the level-1 Poisson variation only. As in the introductory tutorial, we will fit a quadratic in YEAR.

Go to **Data manipulation** menu
Select **Calculate**
Select the empty column C11 and press the right arrow button
Click the '=' button on the keypad
Select YEAR from the list of variables and press the right arrow button
Use the window's keypad to enter **–79**
Press **Calculate**
Clear this calculation using the backspace or delete buttons on your keyboard
Next, select the empty column C12 and press the right arrow button
Click the '=' button on the keypad
Select C11 from the list of variables and press the right arrow button
Use the window's keypad to enter **^2**
Press **Calculate**

Use the **Names** window to name C11 and C12 YEAR79 and YEAR79^2 respectively. Next, use the **Add Terms** button to add both terms to the fixed part of the model only. The **Equations** window should now look like this (remember you can use the **Name** and **+** buttons to display more information about the current model in the **Equations** window):



The response, DEATHs, is set equal to the predicted number of deaths, $\boldsymbol{\pi_{ijk}}$, to be estimated, plus an error term $\mathbf{e_{1ijk}}$ (multiplied by a factor PCONS* which may be ignored temporarily). The e_{1ijk}'s are the level 1 residuals and the variance of these residuals is constrained by the assumption of a Poisson distribution. The predicted number of deaths is then estimated by taking the log of π_{ijk} (i.e. linearising the response) and setting this equal to the linear predictor on the right hand side. This linear predictor is estimated as a quadratic function of time and the intercept in the predictor, β_{0jk}, varies across both COUNTYs and DISTRICTs via the random effects v_{0k} and u_{0jk}, respectively. As in the first part of this tutorial, these random effects are assumed to be normally distributed with zero means. The current model will provide estimates of how the average number of deaths has changed over time (the fixed part) and how the average number of deaths varies between districts and counties (the random part).

The offset

The model described above will fit the observed number of DEATHS in an area using just a mean and a linear and quadratic term in YEAR. However, unlike the SMR this response variable has not been scaled. That is, the SMR of an average DISTRICT in 1992 should be 100; the number of DEATHS in

that DISTRICT may be 10 or 10000 depending on the size of the population. All that an SMR of 100 tells us is that the observed number of DEATHS is the same as the EXPECTED number; we are now trying to fit that observed number and so need to account for the EXPECTED number in our model. We will do this by including it as an *offset* term. We can think of this as modelling the log of the ratio of the predicted deaths $\boldsymbol{\pi_{ijk}}$ to the EXPECTED deaths $\mathbf{E_{ijk}}$ as

$$\log\left(\pi_{ijk} \Big/ E_{ijk}\right) = \beta_{0jk}x_0 + \beta_2 x_{2ijk} + \beta_3 x_{3ijk}$$

In terms of the predicted number of deaths this can be rewritten as

$$\log\left(\pi_{ijk}\right) = \log\left(E_{ijk}\right) + \beta_{0jk}x_0 + \beta_2 x_{2ijk} + \beta_3 x_{3ijk}$$

In other words, the logarithm of the EXPECTED number of deaths in each area, based on population size and age-sex composition, is entered into the regression equation but its coefficient is fixed at 1 rather than being estimated freely, as is the case with the covariate coefficients for CONS, YEAR79 and YEAR79^2. MLwiN provides a facility to do this; the variable to be offset must be named OFFS.

Go to **Data manipulation** menu
Select **Calculate**
Select the empty column C13 and press the right arrow button
Click the '=' button on the keypad
Select LOGE from the list of functions and press the up arrow button
Click the '(' button on the keypad
Select EXPECTED from the list of variables and press the right arrow button
Click the ')' button on the keypad
Click the **Calculate** button

In the **Names** window, name C13 OFFS. This variable is now included in all subsequent Poisson models unless it is renamed or removed using the OFFSET command.

Nonlinear estimation

As mentioned above, generalised linear models are approximated in MLwiN by using a linearising function based upon an expansion of the Taylor series. Specialist knowledge of this approximation is not necessary, however, users should be aware of the following options which are available when using nonlinear estimation.

Click the **Nonlinear** button at the bottom of the **Equations** window

A window appears and provides details of the options for three settings:

- **Distributional assumptions** gives us the options of **Poisson** or **extra Poisson** variation at level 1. A **Poisson** distribution has an equal mean and variance such that $E(y_{ijk}) = Var(y_{ijk}) = \pi_{ijk}$. However, it may be that such a distribution does not fit the data well; the most common situation is one in which the tail of the observed distribution is too heavy. We can sometimes obtain a better approximation to the data by allowing **extra Poisson** variation; the mean remains unchanged but we fit the variance as $Var(y_{ijk}) = \pi_{ijk}\sigma_e^2$. **Poisson** (distributional) variation can then be seen to be a special case of this in which $\sigma_e^2 = 1$.
- **Linearisation** gives us the choice of using a **1st order** or **2nd order** approximation to the Taylor series.

- **Estimation type** gives us the option of using marginal quasi-likelihood (**MQL**) or penalised quasi-likelihood (**PQL**).

The latter two options affect the way in which coefficients are estimated. Bias in parameter estimates tends to be lower when using **2nd order** approximations and **PQL** estimation; however, there is an associated cost in as much as estimation may take longer. The **PQL** estimation procedure is also somewhat less robust and you may experience problems with convergence. A guideline is often to use **1st order, MQL** when exploring the data and to use **2nd order, PQL** to test the model and obtain final estimates.

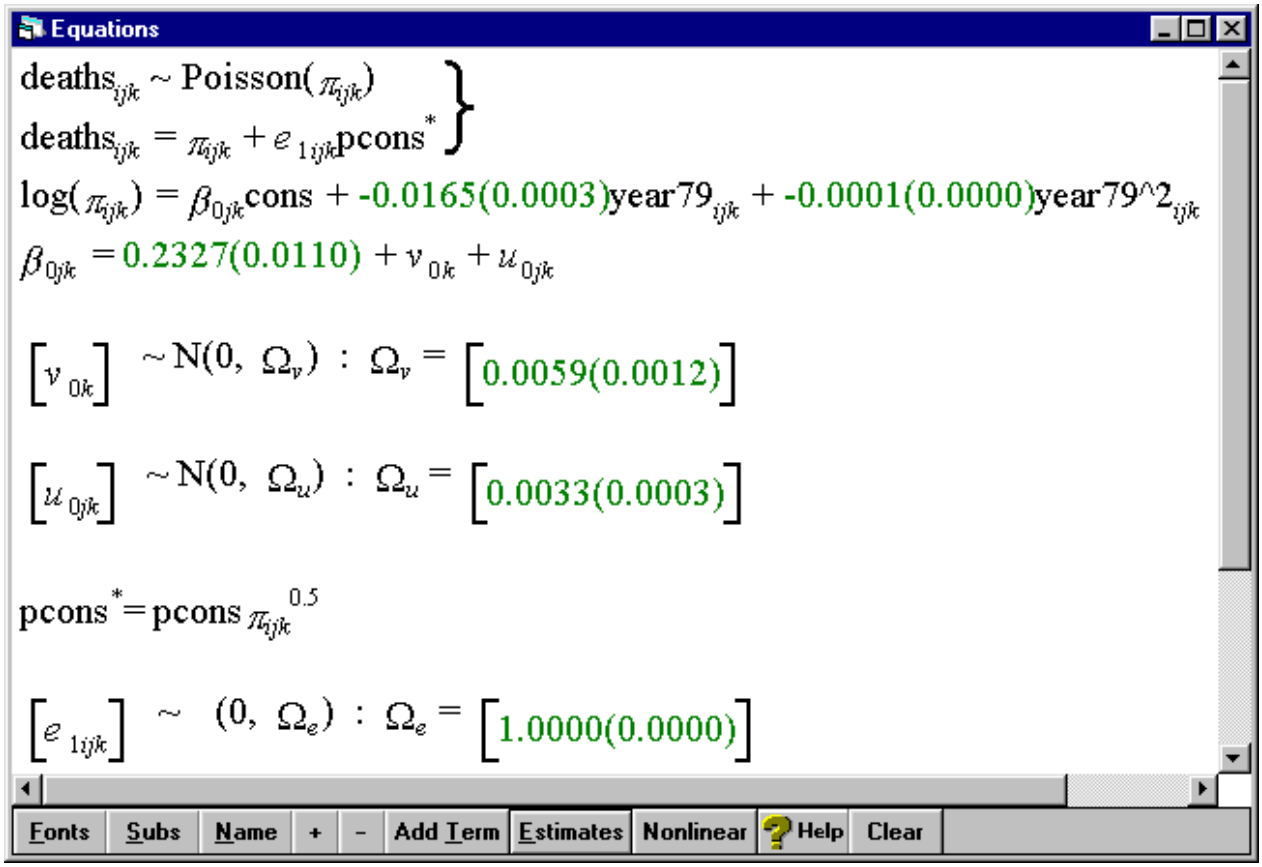
We will begin by using the default settings, assuming **Poisson** variation and a **1st order, MQL** estimation procedure. These options may be set by clicking the **Use Defaults** button in the **Nonlinear Estimation** window and then clicking **Done**.

Press the **Start** button to estimate the model

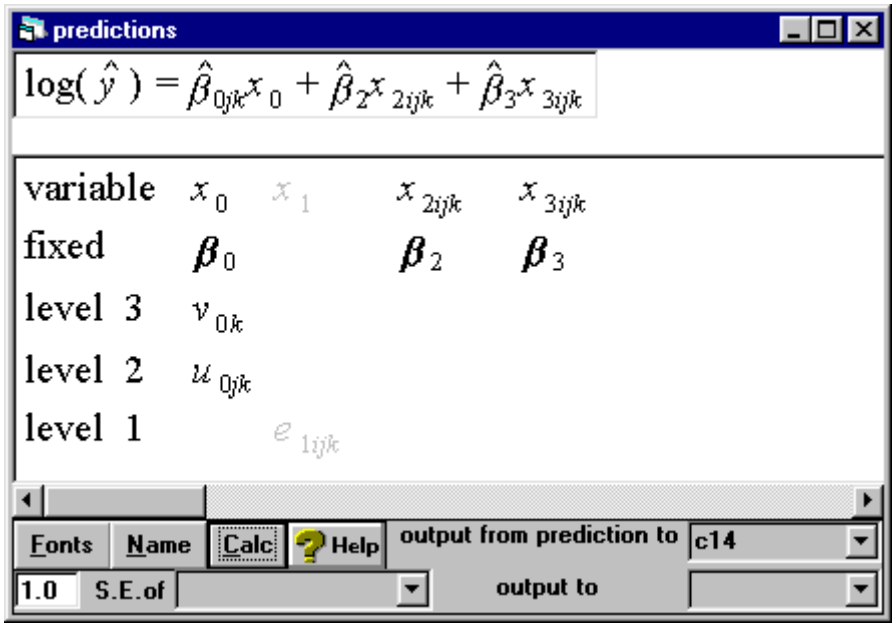
To view the estimates, it will be helpful to change the precision of the display.

Go to the **Options** menu
Select **Numbers**
increase the **# digits after decimal point** to 4
Click **Apply** and then **Done**

By clicking on the **Estimates** button in the **Equations** window, the following should appear:



The parameter estimates are now on the log scale and should be treated as such with the OFFSet term included; for example, the average number of deaths in 1979 has been fitted as 1.262 ($e^{0.2327}$) times the expected number. We can see what is going on more clearly using the **Graph** window. First of all we will get Predictions by DISTRICT, output these to c14 and name this column PRED2.



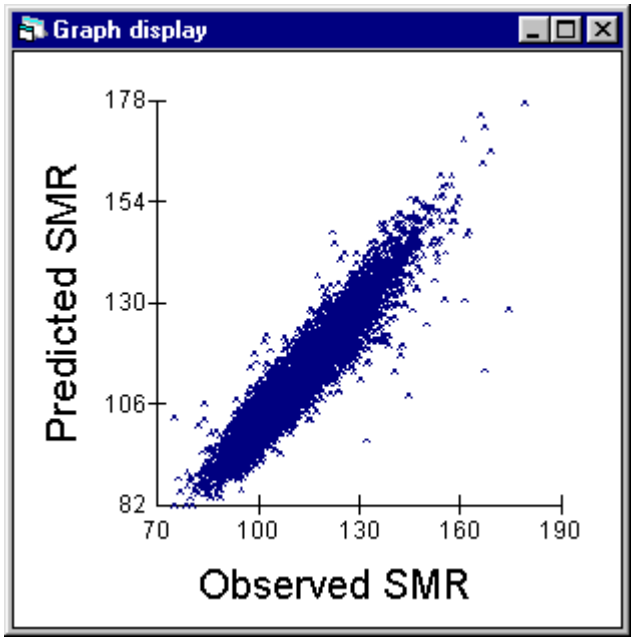
In a similar manner put the predicted values for the fixed part in c15 and name this column PREDFP, and the level 3 predictions in c16 and name this column PRED3. You will note from the summary statistics in the **Names** window that these prediction equations are on the log scale; they also do not include our OFFSet term. As such, we really have the predicted values

$$\log\left(\hat{y} \big/ E_{ijk}\right)$$

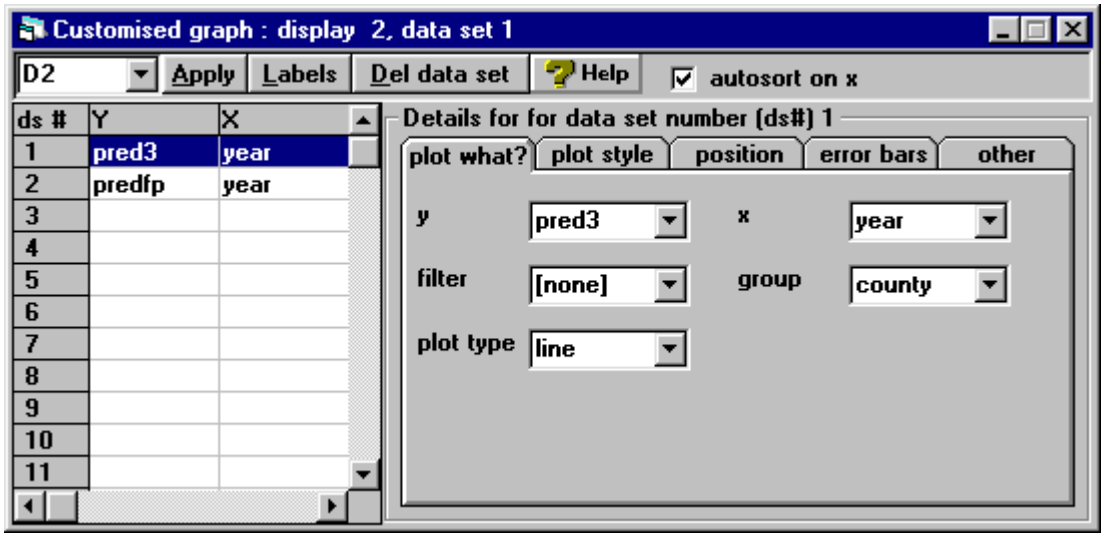
We can very easily convert these to predicted SMRs by taking the EXPOnents in the **Calculate** window:

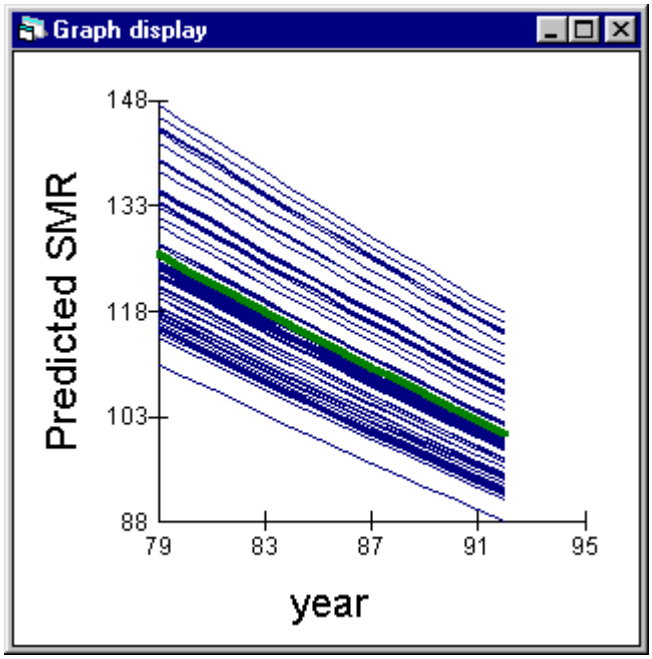
Go to **Data manipulation** menu
Select **Calculate**
Select the PRED2 and press the right arrow button
Type =100* using the keypad
Select the function EXPOnential from the list and press the up arrow button
Click the '(' button on the keypad
Select PRED2 from the list of variables and press the right arrow button
Click the ')' button on the keypad
Click the **Calculate** button

Repeat this process for the variables PREDFP and PRED3. We can now plot the predicted SMR against the observed values; PRED2 includes DISTRICT and COUNTY effects but assumes that the year-on-year fluctuations are part of a Poisson process.



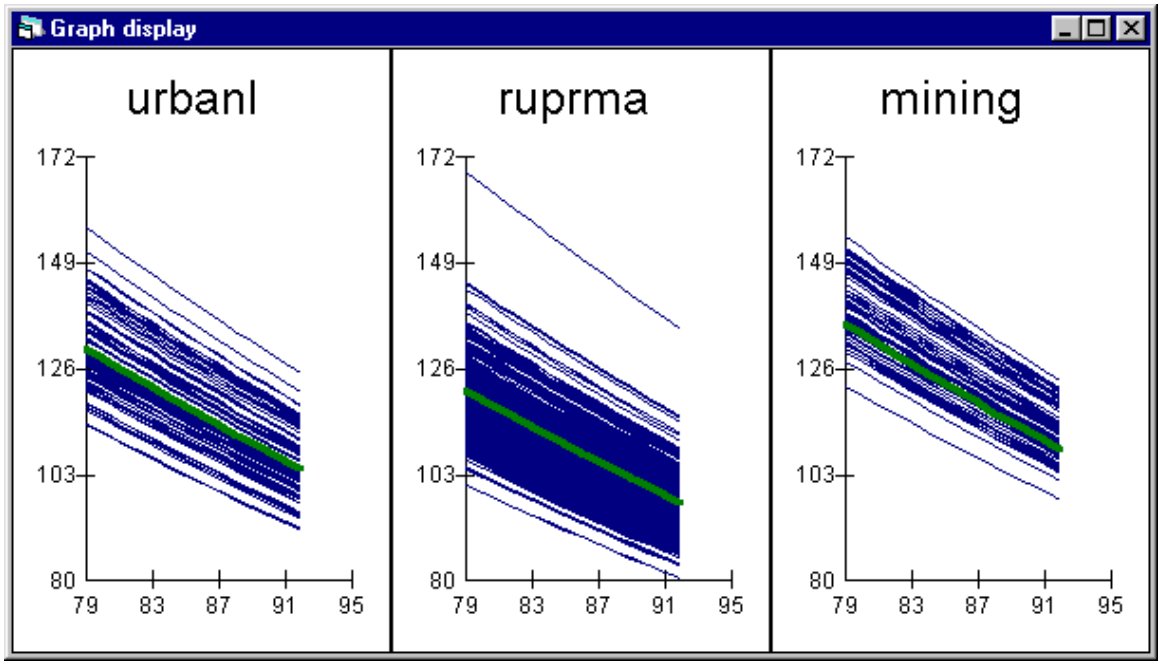
We can also plot the predicted values at national and COUNTY level by YEAR:





This graph illustrates the convergence of SMRs that we noted in the previous analysis; this is to be expected since the assumption of Poisson variation means that we can expect the variation to decrease as the number of DEATHS decreases.

You can continue to build up the model as before, entering random effects where appropriate. The plots of predicted SMRs can be broken down into the three area groupings – urban areas and inner London (URBANL), rural, prospering and maturer areas (RUPRMA) and MINING using the layout option of the Graph window. These might look as follows.



These graphs indicate that there are clear differences between the three types of area in terms of their mean SMR, with MINING areas tending to have the highest SMRs. One of the RURAL districts – DISTRICT 4820 – appears to be outlying with the highest predicted SMR over the period.

The user should now be familiar with the ideas behind modelling mortality data using Poisson regression models. There are many unanswered questions within this data set which stem from the above introductory analysis; for example, are the trends over time the same for the three area groupings or should different slopes be fitted for each? And are the slopes random across DISTRICTs and COUNTYs? The above plot also suggests that the variance of the intercept for DISTRICTs might be different for the three area types, possibly being higher for the RURAL, PROSPERing and MATUREr areas and lower for the MINING areas. The user is encouraged to explore the data further with these questions in mind.

MLwiN - Tutorials

What is Multilevel Modelling?
Hierarchical Structures
Research Questions
Overviews
Tutorials
Software
Back to main site

This page contains the detailed tutorials. These can be opened directly or downloaded.

Educational example:

- Chapter 1 : Random intercept and random slope models
- Chapter 2: Residuals
- Chapter 3: Graphical procedures for exploring the model
- Chapter 4: Contextual effects
- Chapter 5: Variance Functions

Mortality example:

View/download tutorial

The tutorial files are in Acrobat *.pdf format. You can read Acrobat files either after copying or downloading them, or directly within a suitable web browser. If you wish to view acrobat files from within a web browser then you will need Internet Explorer 3 or later or Netscape 3.0 or later. Please consult your browser documentation for configuration information. In either case you will need to install the free Reader (version 3.0 or later) on your computer.

- If you wish to have more information about Acrobat go to Adobe's web site http://www.adobe.com/acrobat/ or go directly to http://www.adobe.com/prodindex/acrobat/readstep.html from where you will be able to download the reader.

If you wish to work through the tutorials on the example datasets with MLwiN , go to the software download page.

Next Section: *Software* ▶

Top