



Unit 5: Study Guide
Multilevel models for macro and micro data
MIMAS
The University of Manchester

5.1 Introduction

5.2 Learning objectives

5.3 Single level models

5.4 Multilevel models

5.5 Theoretical background

5.5.1. Model 1: Single level model: logistic regression

5.5.2. Model 2: Multilevel model: null model

5.5.3. Model 3: Multilevel model: varying intercepts

5.5.4. Model 4: Multilevel model: varying intercepts and slopes

5.5.5. Model 5: Multilevel model: combining survey and aggregate data

5.5.6. Model 6: Multilevel model: interactions of survey and aggregate data

5.6 Using MLwiN and interpreting the results

5.6.1 MLwiN background

5.6.2 MLwiN data

5.6.3 MLwiN exercise conclusions

5.7 Information about the datasets

5.7.1 The variables used in the Lmmd6.ws dataset

5.8 References/further reading

5.1. Introduction

In this unit we see how the multilevel model provides a framework for combining individual level survey data with aggregate group level data. We illustrate this through an example where individual level data from the **European Social Survey** are combined with aggregate, country level data from the **Eurostat New Cronos** data that may be accessed via **ESDS international** (www.esds.ac.uk/international). The dependent variable in our example is whether or not the individual turned out to vote in the most recent election in their country of residence. We restrict the analysis to those people who were of voting age at the most recent election in their country of residence.

5.2 Learning objectives

By the end of this unit you will be able to:

- Comprehend the basic idea of multilevel modelling.
- Explain why multilevel modelling is useful when linking macro (group level aggregate) and micro (individual survey) data.
- Present the kinds of substantive research questions that can be asked when linking macro and micro data in a multilevel model.
- Outline software that permits multilevel models to be fitted.
- Explain how this software may be used to fit a multilevel model with a binary outcome.
- Give an example of multilevel modelling a binary outcome with micro data from the European Social Survey (ESS).
- Give an example of linking micro and macro data in the multilevel model framework by combining the ESS micro data with country-level macro data from Eurostat New Cronos, for long term unemployment.
- Outline the various multilevel models in this context – both substantively and theoretically.
- Explain how interactions between aggregate and individual level measures work in these models and why they might answer important substantive research questions.

5.3. Single level models

Before we discuss multilevel modelling it is worthwhile doing a quick review of traditional single-level analysis, including multiple linear regression and logistic regression. 'Single-level' means that the analysis is carried out at one analytical level – typically the individual level, although sometimes the single level is an aggregate construct, such as the "country". For example, a single level analysis at an aggregate level might be carried out to assess the relationship between the unemployment rate and the crime rate for a set of countries. In this example there would be one pair of values of each country: the unemployment rate and the crime rate. A positive relationship between these two rates would indicate that countries with high unemployment rates would also have high crime rates. However this analysis would not allow any inferences to be made about individual level relationships, such as the individual level relationship between crime and unemployment.

You would use multiple linear regression analysis to relate a set of explanatory variables (sometimes also called 'independent variables' or 'x' variables) to an outcome of interest (sometimes also called a 'dependent variable', or a 'y' variable) that has an interval (continuous) scale. The explanatory variables can be either interval scale (such as age in years), categorical (such as ethnic group), and typically the explanatory variables will be a mixture of these two types. When the response variable is an interval scale and can be assumed to have a normal distribution, we can use multiple linear regression models to assess the nature and strength of the associations of the explanatory variables with the dependent variable. An example would be using multiple linear regression models to investigate the relationship between blood pressure – the outcome variable; an interval scale dependent variable with a normal distribution – with several explanatory variables: age (interval scale), gender, and occupation (categorical). Often in social science, the dependent variable is categorical, and often has two categories or can be re-coded to have two categories. This outcome is binary (and is sometimes also referred to as a dichotomous or 0/1 variable). Examples of binary outcomes are: whether or not someone considers themselves to have limiting long term illness, whether or not someone is unemployed, or whether or not someone turns out to vote. In these situations, logistic regression models are used instead of multiple linear regression models. For example, you could do a logistic regression analysis to model the chance of someone turning out to vote

given information about their age, gender, highest educational qualification and employment status.

5.4. Multilevel models

Single level modelling approaches – multiple linear and logistic regression – are valuable methods to look at the nature and extent of associations of explanatory variables with an **outcome of interest**. However, many populations of interest in social science have a multi-level structure. If we ignore the structure and use a single level model, our analyses may be flawed because we have ignored the context in which processes may occur. Examples of multilevel populations include pupils (level 1) in schools (level 2), or people (level 1) in areas (level 2). Taking the second example, if we choose a single level modelling approach, we must decide whether to carry out the analysis at the individual level or at the area level. If we carry out the analysis at the individual level and ignore the context we may miss important group level effects – this problem is often referred to as the atomistic fallacy. This may occur, for example, when we consider unemployment as an outcome of interest and look at this with respect to individual characteristics such as gender, ethnic group and qualifications but do not take the local labour market conditions into account. If we carry out a single level analysis at the group level and assume the results also apply at the individual level our analyses may be flawed because there are problems of making individual level inferences from group level analyses. This phenomenon is known as the ecological fallacy. This would occur, for example, if the unemployment rate was the outcome of interest and this was related to an area level explanatory variable such as the proportion of people in rented accommodation in each area. This analysis would provide an estimate of the area level relationship between the proportion renting and the unemployment rate but it could not be immediately inferred that this relationship holds at the individual level for unemployed people and people who rent.

Multilevel models have been developed to allow analysis at several levels simultaneously, rather than having to choose at which level to carry out a single level analysis. Multilevel models can be fitted for dependent variables that are interval scale or with categorical outcomes. As well as allowing the relationship between the explanatory variables and dependent variables to be estimated, having taken into account the population structure, multilevel models enable the extent of variation in the outcome of interest to be measured at each level assumed in the model – both before and after the inclusion of explanatory variables in the model. For example, we may wish to assess the extent of variation in examination performance at 16 at the pupil level and at the school level, this would allow us to answer the following research questions:

What proportion of variation in examination performance occurs between schools and what proportion occurs between pupils?

How much of this pupil and school level variation is explained when explanatory variables such as prior examination performance and gender are included in the model?

Multilevel modelling techniques developed rapidly in the late 1980s, when the computing methods and resources for this modelling procedure improved dramatically. Much of the literature on multilevel modelling from this period focuses on educational data, and explores the hierarchy of pupils, classes, schools and sometimes also local education authorities. Measures of educational performance, such as exam scores are usually the dependent variables in this research.

The multilevel model also has other useful properties. Firstly, models can be specified to allow different relationships between the dependent variable and explanatory variables within different groups. For example, to allow a school-specific relationship between prior and current examination performance. Conceptually, this is similar to allowing a separate regression line for each school but statistically the multilevel model is a much more efficient way to proceed than via a separate regression analysis within each school. Multilevel models are also more statistically efficient (i.e. make better use of the available data) than an alternative fixed effects approach which would involve adding dummy variables and their interactions to the multiple linear or logistic regression models.

Secondly the multilevel model provides a natural and appropriate framework for combining data from different sources at one of the levels assumed in the model. For example if we specify a multilevel model with individual at level 1 and country at level 2 and we have sample survey data for a number of countries such as the **European Social Survey (ESS)**. We can use this dataset to assess the associations of age, gender, employment status etc with the chance that someone turns out to vote. If we have additional country level data, such as information from **Eurostat New Cronos** on social cohesion or long term unemployment, we can include this information in the model as a set of country level variables.

A standard multilevel dataset comprises a set of individual level data with group level indicators. An example would be ESS data where data are available for individuals (level 1) and an indicator of country (level 2) is available for each individual. If additional country level such as the **Eurostat New Cronos** data are available, these can be combined with the ESS data at country level in the multilevel model, as explained theoretically in models 5 and 6 in Section 4 and from a practical perspective in Section 5.

5.5 Theoretical Background

In this section we specify several models to allow an assessment of the propensity to vote. We begin with a single level model (Model 1), based on an individual level analysis, and then specify several multilevel models. We explain the model specification in terms of the available survey data from the ESS and aggregate country level data from Eurostat New Cronos. Models 2-4 are multilevel models that can be fitted with ESS data alone. Models 5 and 6 combine country level aggregate data from the Eurostat New Cronos with the ESS data.

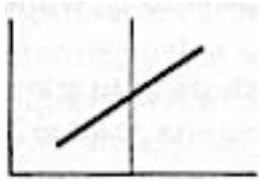
5.5.1 Model 1: Single level model

$$p_i = \Pr(y_i = 1 \mid x_i)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i$$

Where y_i is a 2 category dependent variable to indicate voter turnout. It takes the value 1 if the individual (subscript i) turned out to vote in the most recent election in their country and 0 if they did not. p_i is the probability that the person turns out to vote ($y_i = 1$) given some explanatory variable information we have about the individual, x_i . This could be their age, gender, highest level of education etc. the explanatory variables can be interval scale, categorical or a mixture of the two. In this theoretical discussion we will assume that x_i is an interval scale explanatory variable: age in years. The overall variation in voter turnout is denoted by $\text{Var}(y_i) = \sigma^2$.

Graphical interpretation: the graph below shows how this model works. One straight line is fitted to the data, relating the log of the odds of turning out to vote (vertical axis i.e. the y axis) to age (horizontal axis i.e. the x axis). In this model no country-level information is used; the assumption is that the same relationship applies for all 22 European countries.



Interpretation in words: we can use this model to relate the chance of someone voting to their age. If there is an increased chance of voting as people get older the line will have a positive slope as shown in the graph above.

Note: we could extend model on to allow a quadratic (curved) relationship with age by adding an age^2 term to the model.

5.5.2 Model 2: 'null' model

In the multilevel models specified in this section, the dependent variable, turnout to vote (0=no, 1=yes) now has two subscripts, i and j . There are two subscripts because the model has two levels. i is a subscript for individual (level 1) and j is a subscript for country (level 2).

$$p_{ij} = \Pr(y_i = 1)$$

$$\text{Logit} (P_{ij}) = \beta_0 + u_{0j}$$

$$\text{Var} (U_{0j}) = \sigma_{u_0}^2$$

This 'null' model is so-called because there are no explanatory variables, hence β_0 is the overall population log odds – in this example the overall log odds of turning out. u_{0j} is a country level residual term (also sometimes called an error term) with subscript j . there are 20 of these residuals, one for each European country in the ESS for which aggregate Eurostat New Cronos data is also available. If u_{0j} is positive, this indicates that the particular country it relates to has higher than average turnout. If u_{0j} is negative this indicates that the particular country it relates to has a lower than average turnout. If all countries

had the same turnout and there was no between country variation with respect to this variable, the values of the u_{0j} would be zero for every country.

We would fit model 2 as a starting point in a multilevel analysis, to answer the question:

Before we allow for any explanatory information, how much between country variation is there in the propensity to vote?

We would be able to assess this by looking at the estimated value of σ_u^2 , which is the variance of the u_{0j} terms.

Aside: we could also estimate the proportion of variation at the country level with a measure that has some parallels with the intra class correlation that can be used with interval scale dependent variables. We cannot use the intra class correlation here because our dependent variable is categorical and hence the 'mean' (chance of someone voting in this example) is directly related to the individual level variance. Hence, we need an alternative method appropriate to a categorical dependent variable. Several have been suggested, the simplest of which is usually referred to as a 'threshold model approach'. In this approach we use:

$$\text{Proportion of variance at group level} = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \frac{\pi^2}{3}}$$

Where σ_{u0}^2 is the estimate of the country level variance component, and $\pi = 3.14$ hence this leads to:

$$= \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + 3.29}$$

For a more detailed discussion of this issue see Snijders & Bosker (1999) Chapter 14, especially 14.3.3

5.5.3. Model 3: model with varying intercepts

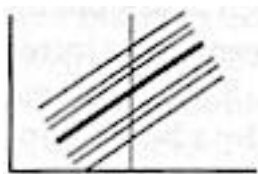
We can extend model 2 to include an explanatory variable, x_{ij} . In this example, let us assume that this variable is the age in years of person i in area j . p_{ij} is now the probability of person i in country j voting in the most recent election in their country, given that we know their age (denoted as $\Pr(y_i = 1 | x)$). Nb: the mathematical operator $|$ means 'given' or equivalently 'conditional on'. The log odds of person i in area j turning out to vote, $\text{Logit}(P_{ij})$, can now be expressed as a straight line, with intercept β_0 and slope (gradient) β_1 . These are the two coefficients of the 'overall relationship' between the chance of someone voting and their age. u_{0j} is a term which determines the change in the intercept for country j compared with the overall intercept. If u_{0j} is positive the intercept for the estimated linear relationship for country j is higher than the overall intercept. This would be the case for countries where there was a higher level of voting than generally in Europe, such as in Norway. If u_{0j} is negative the intercept for the estimated linear relationship for country j is lower than the overall intercept. This would be the case for countries where there was a lower level of voting than generally in Europe, such as in Poland. If u_{0j} is zero the intercept for the estimated linear relationship for country j is the same as the overall intercept. The estimated value of β_1 does not change from country to country; hence the lines are parallel as shown in the graph below. Because there is a different intercept for each country this model is sometimes referred to as the 'model with varying intercepts'. The estimated value of $\sigma_{u_0|x}^2$ shows the extent of variation in the intercepts, given that we know each person's age.

$$p_{ij} = \Pr(y_{ij} = 1 | x_{ij})$$

$$\text{Logit} (P_{ij}) = \beta_0 + \beta_1 x_{ij} + u_{0j}$$

$$\text{Var} (U_{0j} | x_{ij}) = \sigma_{u_0|x}^2$$

Graphical representation



5.5.4. Model 4: model with varying intercepts and slopes.

$$p_{ij} = \Pr(y_{ij} = 1 | x_{ij})$$

$$\text{Logit} (P_{ij}) = \beta_0 + \beta_{1j} x_{ij} + u_{0j}$$

Where the 'random slopes coefficient is:

$$\beta_{1j} = \beta_1 + u_{1j}$$

In this model an overall line relating the chance of someone voting with age is

fitted, with intercept and slope β_1 . The change in the intercept for country j is

u_{0j} and the change in the slope for country j is u_{1j} . If the overall relationship

between the chance of voting and age is positive and u_{1j} is positive then the line is steeper than the overall gradient for country j. If the overall relationship

between the chance of voting and age is positive and u_{1j} is negative then the line is less steep than the overall gradient for country j. For each country both the intercept and slope for the estimated relationship between the chance of voting

and age can vary from the overall line. Hence the relationship between u_{0j} and

u_{1j} is also of interest in Model 4, and this is summarised by the covariance

term $\sigma_{U_0U_1|x}$. If the overall relationship between chance of voting and age is

positive and $\sigma_{U_0U_1|x}$ is positive, this means that a line with a higher than overall intercept is also likely to have a steeper than overall slope. Hence the country-

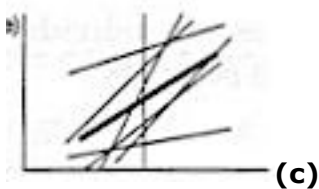
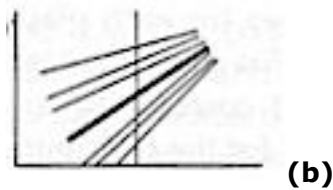
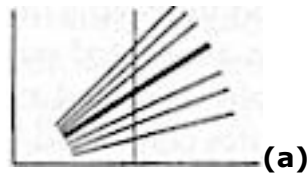
specific lines will diverge as shown in diagram (a) below. If $\sigma_{U_0U_1|x}$ is negative the country-specific lines will converge as shown in diagram (b) below. If there is no obvious pattern between intercept and slope, as shown in diagram (c), the estimated value of will be zero.

$$Var \left(\begin{matrix} U_{0j} \\ U_{1j} \end{matrix} \middle| x_{ij} \right) = \begin{pmatrix} \sigma_{U_0|x}^2 & \sigma_{U_0U_1|x} \\ \sigma_{U_0U_1|x} & \sigma_{U_1|x}^2 \end{pmatrix}$$

Alternatively, but equivalently, we can write the Model 5 as:

$$Logit(P_{ij}) = \beta_0 + \beta_1 x_{ij} + u_{1j} x_{ij} + u_{0j}$$

Graphical representation



5.5.5 Model 5: combining survey and aggregate data.

$$p_{ij} = \Pr(y_{ij} = 1 \mid x_{ij}, X_j)$$

$$Logit(P_{ij}) = \beta_0 + \beta_1 x_{ij} + \beta_2 X_j + u_{0j}$$

$$\text{Var}(U_{oj} \mid x_{ij}, X_j) = \sigma_{u_0|x, X}^2$$

Multilevel modelling allows us to combine variables the survey data with aggregate data from another source. Hence in the current example we could extend, for example, Model 3 to include aggregate (country level) information from another source. We illustrate this in Section 5 when we combine ESS survey data by adding % long term unemployment as an additional explanatory variable. This information is from the aggregate Eurostat New Cronos data. As this is country level information based on a census of all economically active people (i.e. it is a census not a survey) we denote it as uppercase X_j . Note that there is only a j (country level) subscript. There is no i subscript for this variable as all people in country i have the same value of long term unemployment. The substantive reason for adding long term unemployment here is that this may explain some of the country level variations in voting. Perhaps people living in countries with higher long term unemployment are more likely to vote. We will investigate this later, in Section 5.

5.5.6 Model 6 interactions between aggregate data and survey data variables.

$$p_{ij} = \Pr(y_{ij} = 1 \mid x_{ij}, X_j)$$

$$\text{Logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + \beta_2 X_j + \beta_3 x_{ij} X_j + u_{0j}$$

$$\text{Var}(U_{oj} \mid x_{ij}, X_j) = \sigma_{u_0|x, X}^2$$

Finally, we may

wish to look at interactions between individual and aggregate explanatory variables. In this example we can look at the interaction between a person's age and the amount of long term unemployment in the country in which they live:

$\beta_3 x_{ij} X_j$ - this enables us to ask the question 'is there any evidence that age relates to the change of voting differently in countries with high long-term unemployment compared with countries with low long-term unemployment?'. We could also look at other kinds of relationship with this model framework e.g. include an individual level explanatory variable indicating whether or not someone is unemployed and interact this with long term unemployment in the model to assess whether unemployed people in countries with high long-term

unemployment are more or less likely to vote than unemployed people countries with low long-term unemployment.

5.6 Using MLwiN and interpreting the results

5.6.1 MLwiN background

Various software packages are available for multilevel analysis. Some are specialist packages for multilevel modelling such as **MLwiN** or **HLM**. More general statistical packages such as **SPSS**, **SAS** and **STATA** also allow some multilevel modelling to be carried out but the scope for model specification is currently more limited than that of MLwiN and HLM.

We will make use of MLwiN which was developed by the Centre for Multilevel Modelling at the University of Bristol. The software can also be obtained via: www.cmm.bristol.ac.uk.

We will not explain in detail here how to get data from SPSS or excel into MLwiN but briefly a very useful way to get data from excel into MLwiN version 2 is to copy the entire excel spreadsheet and paste it into MLwiN having opened the MLwiN software by first choosing '**free columns**' in MLwiN. This method also enables the researcher to specify that the first row of the data to be pasted is the name of each variable. It also has the advantage that it preserves any gaps in the original dataset and treats these as missing cases in MLwiN.

It is easy to save an SPSS dataset as excel by using '**save as**' and also choosing the option to '**put variable names in first row**'.

5.6.2 MLwiN data

The data has been prepared for this exercise as Immd6.ws (the .ws suffix indicates an MLwiN worksheet which contains the data). N.B. If MLwiN has been used to fit some models, and the worksheet is then saved, these model results will also be contained in the worksheet – this is useful for saving results of previous analyses.

To merge individual and group level data in SPSS each dataset to be merged must have a group level id. In our case the ESS has a country code and there is then one row of aggregate country level data from the Eurostat New Cronos. In our example the ESS data (a 10% sub sample of the original dataset) contains 3362 cases and the Eurostat New Cronos contains 20 rows – one of each country that is common to both ESS and Eurostat New Cronos.

To merge files in SPSS:

1. **Open** the individual level data file and choose **data > merge cases > add variables**.
2. Select the aggregate data file as data to be merged.
3. Choose the key variable (the group level id).
4. Select '**external file is keyed table**'.

The resulting data file should then contain all the individual level data and the values of the aggregate data for each individual are then added as new columns in the data file. Every individual in a particular country has the same value of these aggregate variables.

**Activity1 - using MLwiN**

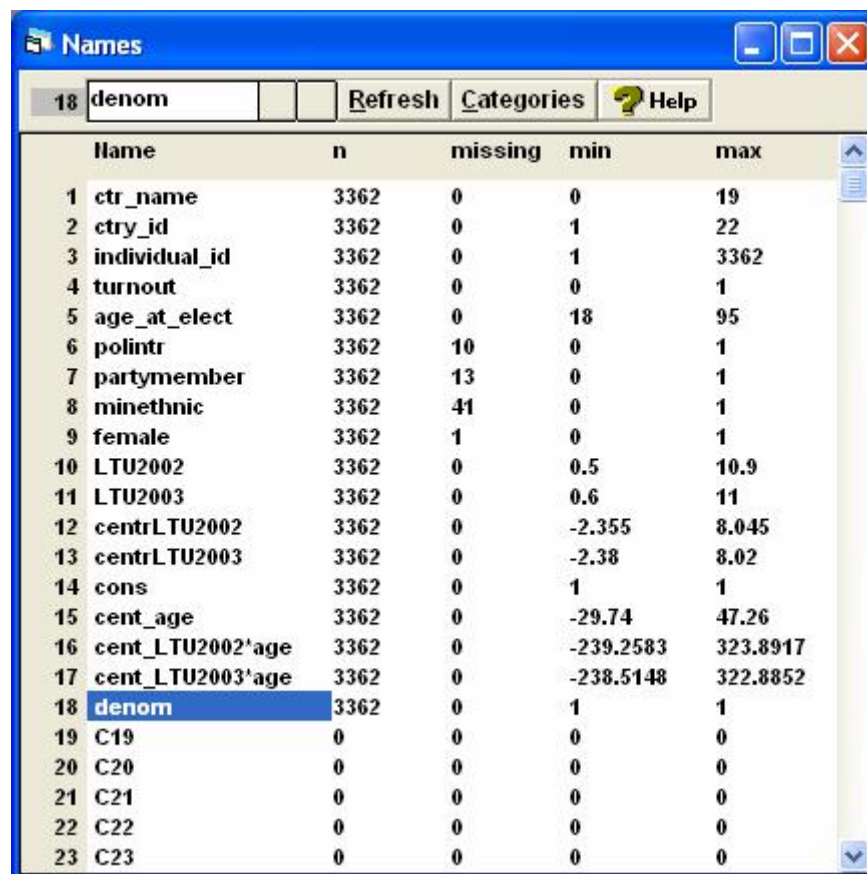
Open MLwiN by locating it in the programmes listed in the windows start menu or by clicking on the MLwiN icon on your desktop.

The default worksheet size for this exercise is 5000 cells which is too small to permit the analysis. However, it is easy to increase the worksheet size.

To do this go to options and make the worksheet 10000 cells (change from 5000). N.B. Do not save worksheet when prompted.

No go to the file menu in MLwiN and open Immd6.ws

Choose data **manipulation > names**.



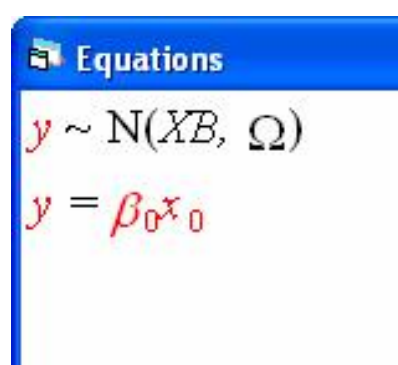
	Name	n	missing	min	max
1	ctr_name	3362	0	0	19
2	ctry_id	3362	0	1	22
3	individual_id	3362	0	1	3362
4	turnout	3362	0	0	1
5	age_at_elect	3362	0	18	95
6	polintr	3362	10	0	1
7	partymember	3362	13	0	1
8	minethnic	3362	41	0	1
9	female	3362	1	0	1
10	LTU2002	3362	0	0.5	10.9
11	LTU2003	3362	0	0.6	11
12	centrLTU2002	3362	0	-2.355	8.045
13	centrLTU2003	3362	0	-2.38	8.02
14	cons	3362	0	1	1
15	cent_age	3362	0	-29.74	47.26
16	cent_LTU2002'age	3362	0	-239.2583	323.8917
17	cent_LTU2003'age	3362	0	-238.5148	322.8852
18	denom	3362	0	1	1
19	C19	0	0	0	0
20	C20	0	0	0	0
21	C21	0	0	0	0
22	C22	0	0	0	0
23	C23	0	0	0	0

View the data and notice that the data have been sorted by country code (second column) – all the observations for Austria – the first country in the dataset appear together, then all the observations from the second country and so on.

N.B. Variables with uppercase names are from aggregate (macro) data. Variables with Lower case names are from the ESS survey (micro).

We have a binary outcome (turnout: 0=didn't vote, 1=voted) so we need to set up a multilevel logistic regression model to model the chance of someone voting. Do this as follows.

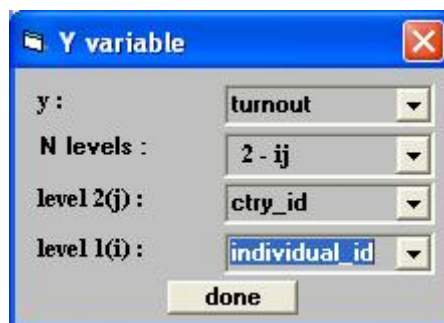
Go to model equations and you see this



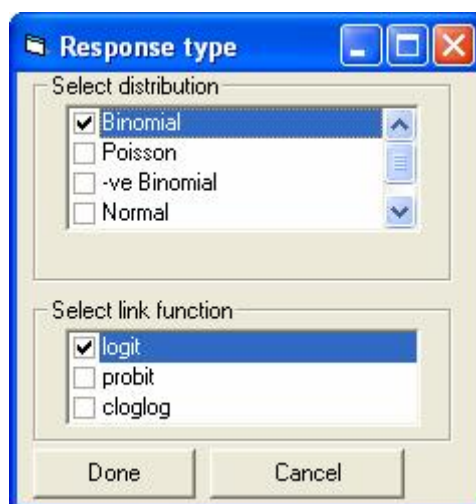
$$y \sim N(XB, \Omega)$$

$$y = \beta_0 x_0$$

Click on the red y variable and choose turnout. We have a 2 level structure with country at level 2 and individual at level 1 specify this structure like this:



We need to change the model specification from the basic assumption that y (the dependent variable) is a normally disturbed interval scale variable. Click on the N to change the distribution. Choose binomial **logit**.

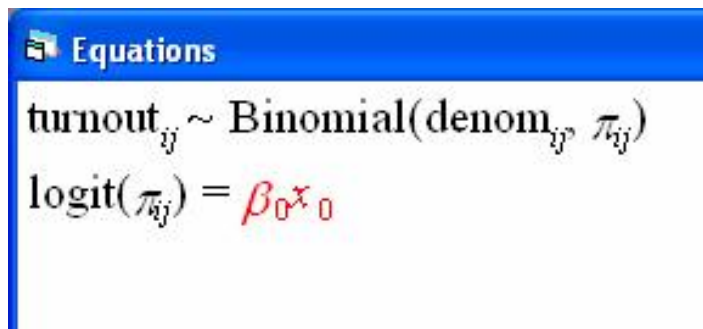


Now the equation looks like this:

$$\text{turnout}_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_0 x_0$$

Click on the red n and choose 'denom'.



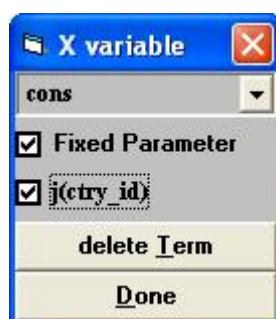
Equations

$$\text{turnout}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

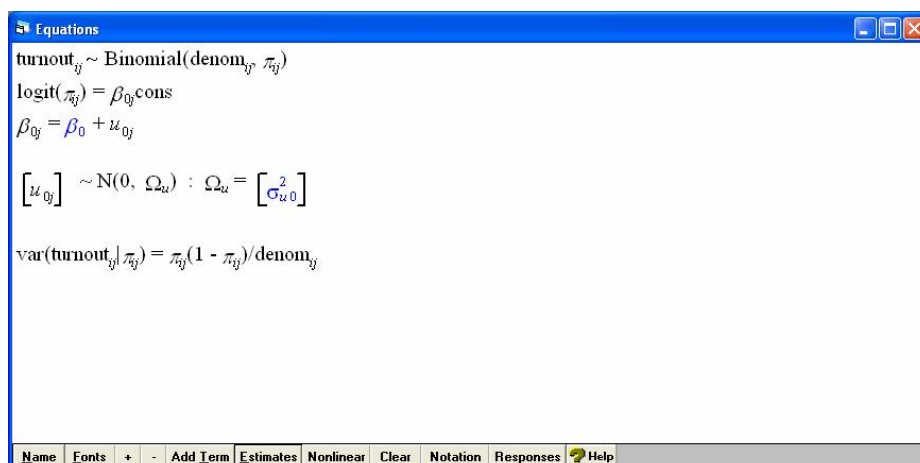
$$\text{logit}(\pi_{ij}) = \beta_0 x_0$$

Click on red x and choose 'cons' and allow this to vary from country to country by clicking the **ctry_id** box

N.B. 'cons' and 'denom' are two variables that are needed to allow MLwiN to fit a multilevel logistic model. In this example (which is typical of the situation for social science data) both 'cons' and 'denom' are just columns of 1s with the same number of observations as there are individuals in the dataset.



We have now set up Model 2 – the null model. It looks like this (click on **Estimates** button at the bottom of the equations window to see this representation. As you can see the items in blue are the parameters to be estimated – on the log odds (logit) scale these are the overall mean beta 0 and the between country variance component sigma squared u 0.



Equations

$$\text{turnout}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j} \text{cons}$$

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

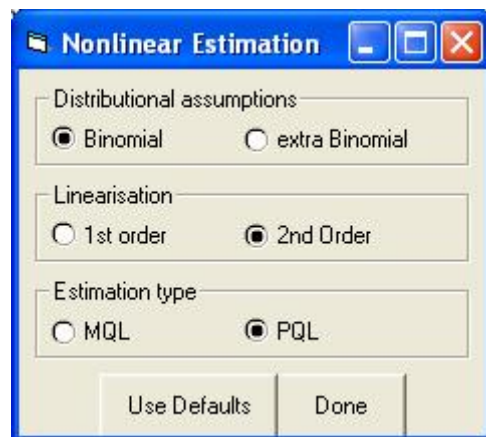
$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma_u^2 \end{bmatrix}$$

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

Name Fonts + - Add Term **Estimates** Nonlinear Clear Notation Responses ? Help

We now need to specify the estimation type. Click on nonlinear at the bottom of the equations window.

Choose 2nd order PQL – for technical reasons this gives better estimates of the variance components than the default.



An aside: Using MCMC estimation instead – some research shows that PQL variance estimates, whilst better than MQL estimates (the default in MLwiN) as still downwardly biased i.e. underestimate the extent of variation. Once we have estimated the parameters in MLwiN using PQL we can switch to Monte Carlo Markov Chain estimation by clicking on the 'estimation control' window and choosing 'MCMC'. Then re-estimate the model parameters using the PQL estimates as 'starting values' in the iterative process. We illustrate this below for this model. We could use this approach for any of the multilevel models. For more details see references on www.cmm.bristol.ac.uk.

Now click on '**start**' in the top left of the programme window. The parameters will turn from blue to green when the estimation process has converged. Click on the **Estimates** button at the bottom of the equations screen to see the estimated values:

$$\text{turnout}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j} \text{cons}$$

$$\beta_{0j} = 1.643(0.132) + u_{0j}$$

$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [0.299(0.110)]$$

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

The mean is 1.643 (on the logit scale). To convert back from logit to probability use

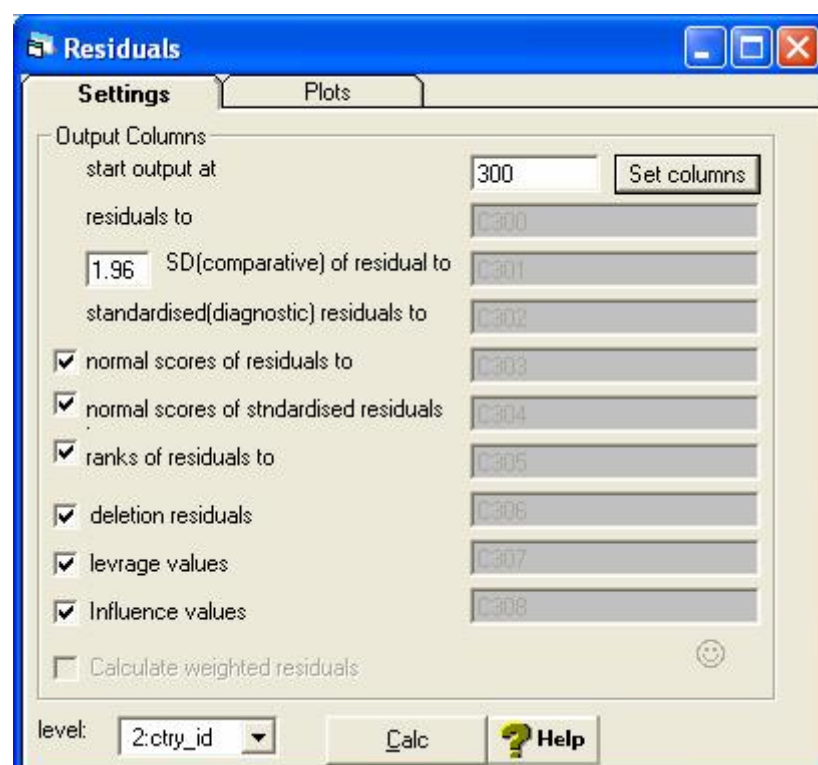
$$e^{1.643} / (1 + e^{1.643}) = 0.838$$

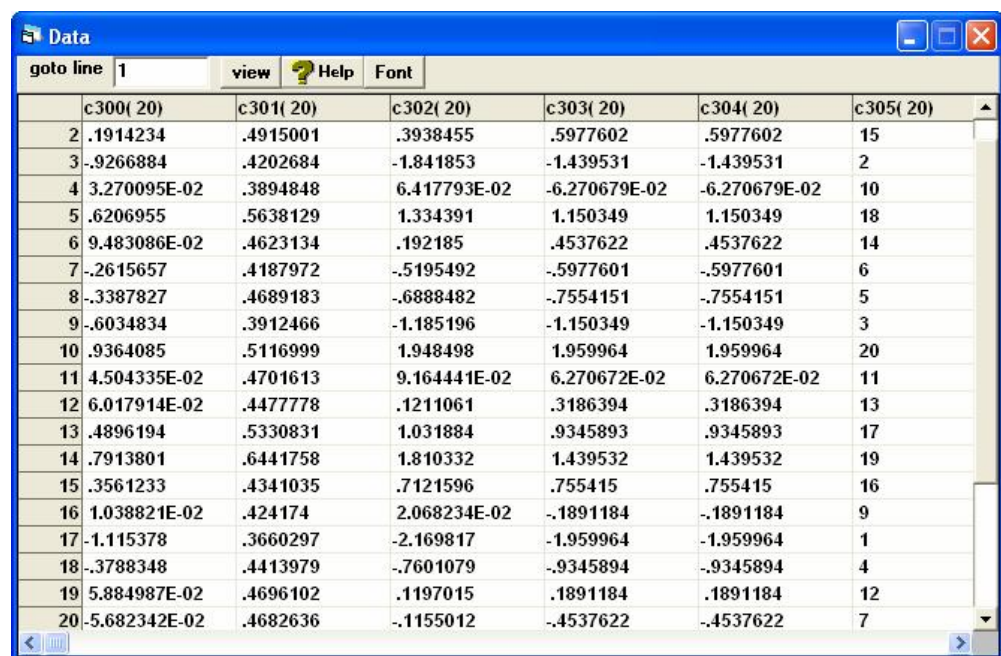
where 'e' is the exponential function.

So the overall proportion reporting that they turned out to vote in this sample is 0.838 (this represents an average turnout of nearly 84%). We know that in the actual elections a lower proportion turned out. Hence some people are reporting in the ESS that they turned out to vote in the most recent election when in fact they did not (and/or the sampling process has lead to an oversampling of voters). We can account for this partially by using weights. See for example, the post-stratification approach used by Fieldhouse, Tranmer and Russell (2007). For now we will continue with the figures as they are.

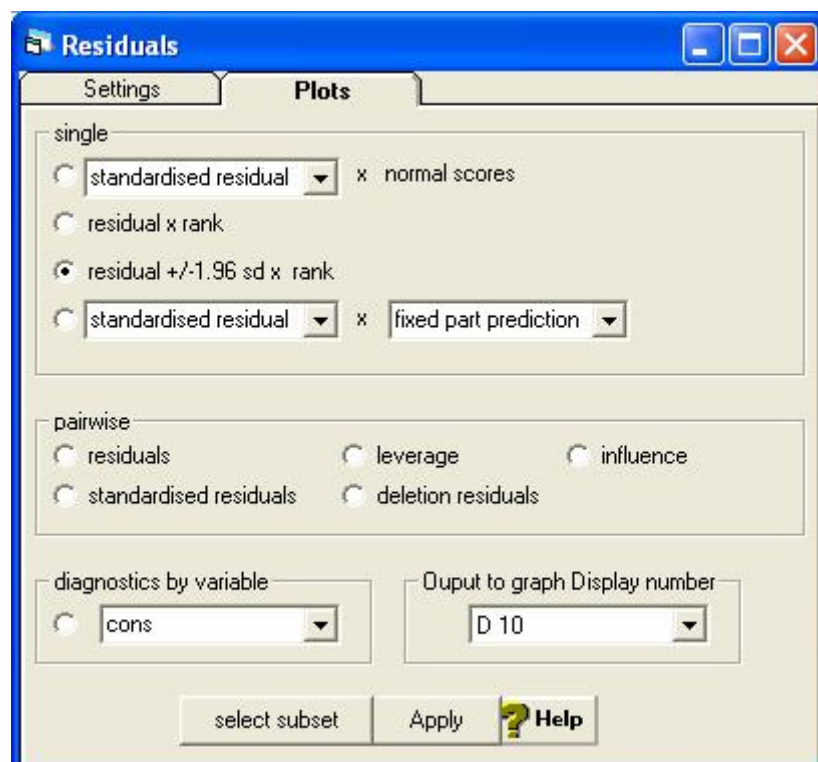
The country level variation is estimated as 0.299 on the logit scale, suggesting there is considerable variation between countries with respect to voter turnout.

We can save and plot the country level residuals from this model. Choose '**residuals**' from the model menu.





goto line	1	view	Help	Font		
	c300(20)	c301(20)	c302(20)	c303(20)	c304(20)	c305(20)
2	.1914234	.4915001	.3938455	.5977602	.5977602	15
3	-.9266884	.4202684	-1.841853	-1.439531	-1.439531	2
4	3.270095E-02	.3894848	6.417793E-02	-6.270679E-02	-6.270679E-02	10
5	.6206955	.5638129	1.334391	1.150349	1.150349	18
6	9.483086E-02	.4623134	.192185	.4537622	.4537622	14
7	-.2615657	.4187972	-.5195492	-.5977601	-.5977601	6
8	-.3387827	.4689183	-.6888482	-.7554151	-.7554151	5
9	-.6034834	.3912466	-1.185196	-1.150349	-1.150349	3
10	.9364085	.5116999	1.948498	1.959964	1.959964	20
11	4.504335E-02	.4701613	9.164441E-02	6.270672E-02	6.270672E-02	11
12	6.017914E-02	.4477778	.1211061	.3186394	.3186394	13
13	.4896194	.5330831	1.031884	.9345893	.9345893	17
14	.7913801	.6441758	1.810332	1.439532	1.439532	19
15	.3561233	.4341035	.7121596	.755415	.755415	16
16	1.038821E-02	.424174	2.068234E-02	-.1891184	-.1891184	9
17	-1.115378	.3660297	-2.169817	-1.959964	-1.959964	1
18	-.3788348	.4413979	-.7601079	-.9345894	-.9345894	4
19	5.884987E-02	.4696102	.1197015	.1891184	.1891184	12
20	-5.682342E-02	.4682636	-.1155012	-.4537622	-.4537622	7



Residuals

Settings | **Plots**

single

☐ standardised residual x normal scores

☐ residual x rank

☒ residual +/- 1.96 sd x rank

☐ standardised residual x fixed part prediction

pairwise

☐ residuals ☐ leverage ☐ influence

☐ standardised residuals ☐ deletion residuals

diagnostics by variable

☐ cons

Output to graph Display number

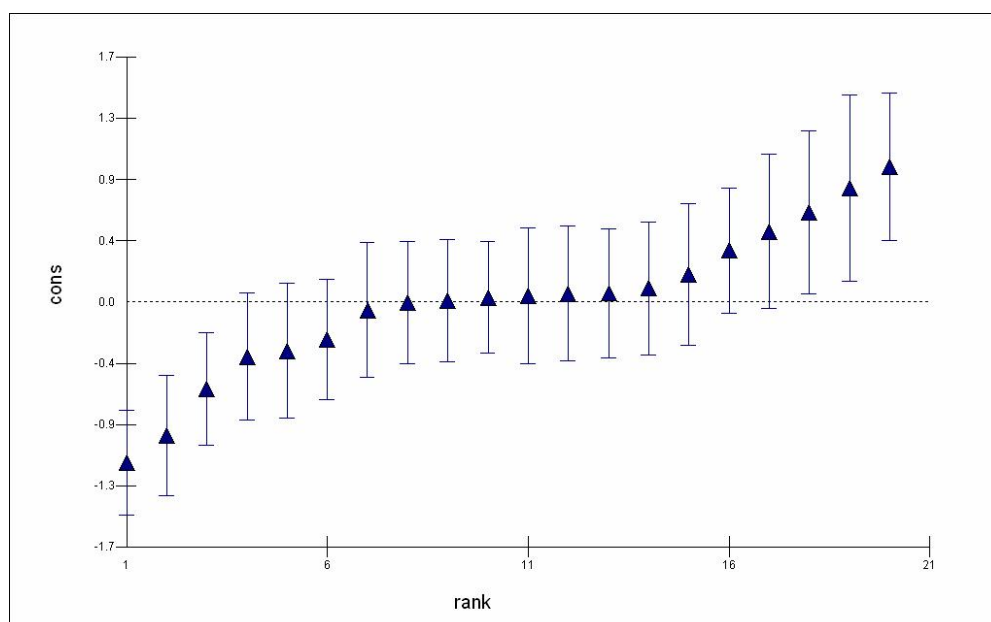
D 10

select subset Apply Help

And set the comparative s.d. as 1.96 and the level to be 2:ctry_id.

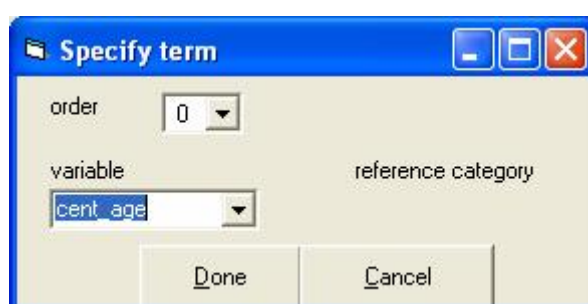
Also – click on **'set columns'**.

Now click on plots and choose **'residual'** +/- 1.96 s.d x rank and click **'apply'**.



We get this 'caterpillar' plot. The residuals U_{0j} are plotted in ascending order of magnitude with their confidence intervals. Where this confidence interval crosses the 0.0 line the turnout for that country is not significantly different from the overall turnout in Europe. If the confidence interval is entirely below the dotted line the turnout is significantly lower for that country and if the confidence interval is entirely above the dotted line the turnout is significantly higher for that country. The plot is interactive – we can click on a residual to find out the country id. For example the first residual on the plot is country id 19 (Poland) and the last one is country id 11 (Greece).

Now we extend the model to include an explanatory variable – age, which has been centred around its mean. This is Model 3.



We now re-estimate the model (press 'more' on top left of programme window).

Equations

$$\text{turnout}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j} \text{cons} + 0.014(0.003) \text{cent_age}_{ij}$$

$$\beta_{0j} = 1.660(0.134) + u_{0j}$$

$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [0.307(0.113)]$$

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

We now see that age has a positive coefficient (0.014) which is statistically significant (i.e. more than twice its standard error which is shown in brackets after the estimate and in this case is 0.003). A rule of thumb is to compare twice the standard error with the absolute (ignore sign) value of the coefficient. To do this exactly we would use 1.96 standard errors but as 1.96 is close to two it is a useful approximation to simply double it. As we can see $0.14 > 0.006$ so this coefficient is statistically significant. As people get older they are more likely to vote. There is still considerable variation between countries (0.307).

Conditional on knowing the age of each person in the model, so age does not explain all the country level variation in voting. We could produce a caterpillar plot of the residuals as before but we will now produce another kind of plot – one showing the predicted values. **Choose model > predictions**

predictions

$\text{logit}(y_{ij})^{\wedge} =$

variable	cons	cent_age _{ij}
fixed	β_0	β_1
level 2	u_{0j}	
level 1		

Fonts Name Calc ? Help output from prediction to

1.0 S.E.of output to

Click on β_0 , β_1 and u_{0j}

predictions

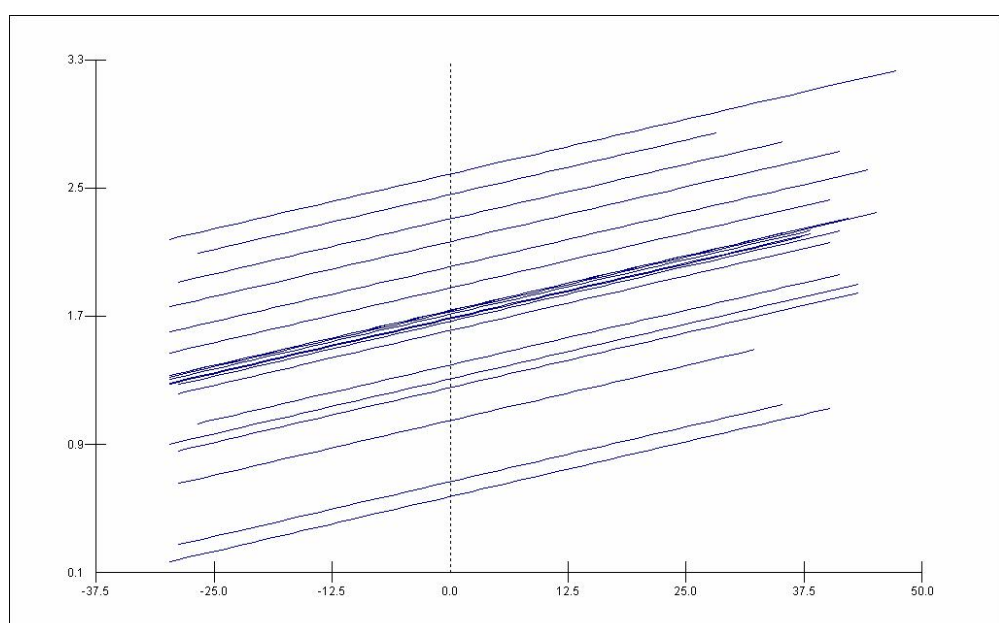
$$\text{logit}(y_{ij}) = \hat{\beta}_{0j} \text{cons} + \hat{\beta}_1 \text{cent_age}_{ij}$$

variable	cons	cent_age _{ij}
fixed	β_0	β_1
level 2	u_{0j}	
level 1		

Fonts Name Calc ? Help output from prediction to c20

1.0 S.E. of output to

Choose c20 as the output column and click 'calc'. No go to graphs, and choose customised graphs and set up the graph menu like this – a separate line for each country relating the predicted value of turnout on the log odds scale (c20) to (centred) age. Click on apply.



We now see a graph with 20 parallel lines – the gradient is positive. As people get older they are more likely to vote. On the centred age scale, 0

represents the average value of age – around 47. We can see that there is variation in terms of where the lines cross the vertical line at $x=0$; a linear effect of age does not explain all the country level variations in voting.

We can also allow the gradient of the line to be different in each country (Model 4). Click on the **cent_age** variable in the equations window. Tick the box that is marked **j(ctr_id)** we are now allowing each estimated line to have its own country-specific slope and intercept.



Our estimated model is:

$$\text{turnout}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

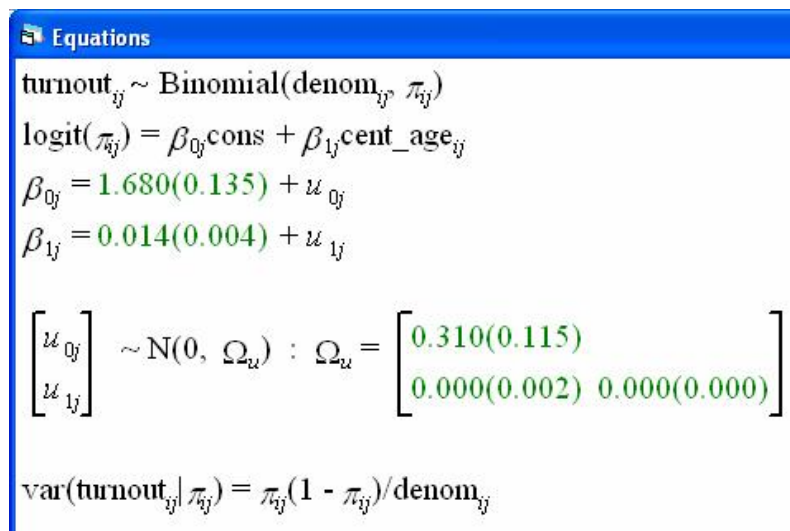
$$\text{logit}(\pi_{ij}) = \beta_{0j} \text{cons} + \beta_{1j} \text{cent_age}_{ij}$$

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

$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}$$

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



Equations

$$\text{turnout}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j} \text{cons} + \beta_{1j} \text{cent_age}_{ij}$$

$$\beta_{0j} = 1.680(0.135) + u_{0j}$$

$$\beta_{1j} = 0.014(0.004) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.310(0.115) & 0.000(0.000) \\ 0.000(0.002) & 0.000(0.000) \end{bmatrix}$$

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

We notice that both the variance of the slopes and the covariance of the slopes are estimated to be zero. There is no evidence that the gradient of the slope varies from country to country with respect to age in this sub sample of the ESS.

Hence we go back to the random intercepts only model (Model 3) by clicking on **cent_age** and choosing these options:



In the next model (Model 5) we add an aggregate country level variable: **centrLTU2002** – centred long term unemployment from the Eurostat New Cronos. We do this by first clicking '**add term**' in the equations window and choosing it. This model now has age as an explanatory variable from the micro data, long term unemployment from the macro data and the intercepts are allowed to vary from country to country.

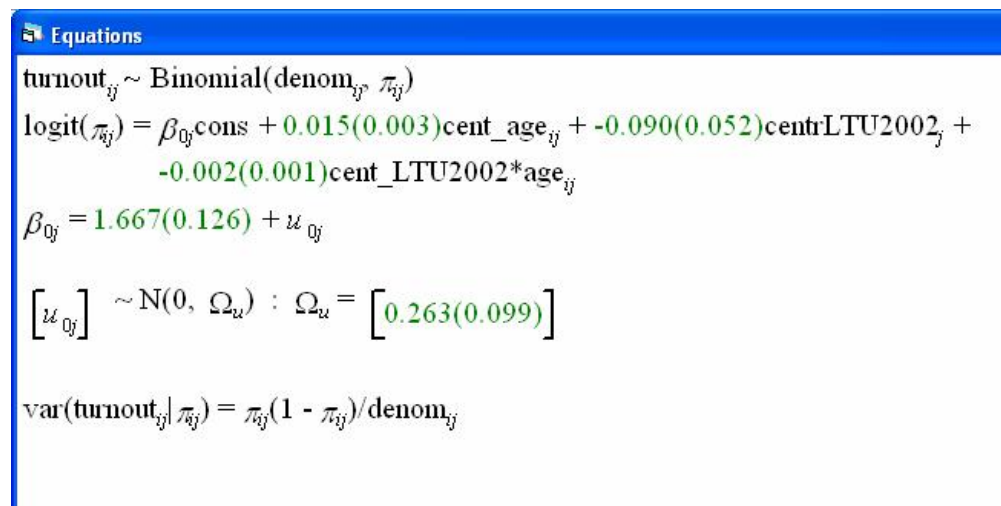
$$\begin{aligned} \text{turnout}_{ij} &\sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \beta_{0j} \text{cons} + 0.014(0.003) \text{cent_age}_{ij} + -0.082(0.052) \text{centrLTU2002}_j \\ \beta_{0j} &= 1.661(0.125) + u_{0j} \end{aligned}$$

$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [0.260(0.098)]$$

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

We notice that the coefficient of this term is negative: having controlled for age, the higher the level of long term unemployment the lower the voter turnout. We notice that this variable is not statistically significant at the usual 5% significance level, as twice its standard error is more than 0.082 but it has still lead to a 12% reduction in the estimated between country variance (0.260 compared with 0.302). When selected which variables for inclusion we take account of both of these factors, so a variable whose coefficient is not significant may still be included if it reduces the between groups variance. It is evident that many more variables may be needed at the country level to further reduce the variation. At present the relationship between age and chance of voting is assumed to be linear, so we might also want to explore the possibility of a quadratic (curved) relationship with age by adding $(\text{cent_age})^2$ to the model.

Finally we introduce the interaction term between age and long term unemployment (the product of the two variables) to the model and find that there is a significant coefficient for this term. It is negative (-0.002) and just significant – areas with higher long term unemployment tend to have a slightly shallower relationship with age with respect to voter turnout.



The screenshot shows the 'Equations' window in MLwiN. The equations are as follows:

$$\text{turnout}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j} \text{cons} + 0.015(0.003) \text{cent_age}_{ij} + -0.090(0.052) \text{centrLTU2002}_j + -0.002(0.001) \text{cent_LTU2002} * \text{age}_{ij}$$

$$\beta_{0j} = 1.667(0.126) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.263(0.099) \end{bmatrix}$$

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

5.6.3 MLwiN exercise conclusions

We have seen that the multilevel model is a useful framework for combining macro (aggregate) and micro (individual) data and applied it to an example based on voter turnout in 20 European countries using data from the European Social Survey and Eurostat New Cronos. We have seen that voter turnout increases with age and there is some evidence that voter turnout is lower in areas with high long term unemployment (Model 5). There is also some evidence that the rate of change in the chance of voter turnout is slightly less in areas of higher long-term unemployment than areas with lower long term unemployment.

5.7 Information about the datasets

Lmmd6.ws is an MLwiN dataset containing data from the ESS (variable names in lower case) and data from the Eurostat New Cronos in variable names in (UPPER CASE). The data have been pre-sorted by country id, to allow multilevel modelling to be carried out. Age and the long term unemployment variables have each been centred by subtracting the mean. This improves the substantive interpretation of the multilevel models because a value of 0 on a centred variable represents the mean of that variable.

The MLwiN information on '**cons**' and '**demon**' necessary for multilevel logistic regression analysis has also been added to this dataset. Some additional variables on political interest, member of a group and gender are also available on this dataset to allow further explanatory variables to be added to the models described here. An interaction between the long term unemployment for 2002 in each country (from the Eurostat New Cronos) and the age of each person (which has been centred) has already been created by multiplying these two variables together.

5.7.1 The variables used in the Lmmd6.ws dataset

Lmmd6.ws is an MLwiN worksheet containing the variables. No models have been previously specified or run on this dataset. The variables on this dataset are:

Micro data:

Ctry_name – name of country (string variable)

Ctry_id – country level id

Individual id – individual id

Turnout – voter turnout (dependent variable 0=didn't vote, 1=voted)

Age_at_elec – age of respondent at most recent election

Polintr – interest in politics

Partymember – member of political party?

Minethnic – in minority ethnic group in country of residence?

Female – 0=male, 1=female

Macro data:

LTU2002 - % long term unemployment 2002

LTU2003 - % long term unemployment 2003

centrLTU2002 = LTU2002 – mean (centred)

centrLTU2003 = LTU2003 – mean (centred)

Micro / Macro data interactions:

Cent_LTU2002*age = centred long term unemployment 2002 * centred age of respondent.

Cent_LTU2003*age = centred long term unemployment 2003 * centred age of respondent.

MLwiN variables:

Cons – a column of 1s

Denom – a column of 1s.

The Lmmd6 dataset has 3362 cases and is sorted by **ctry_id**. This is a 10% sub-sample of the original ESS dataset 20 of the original 22 countries in the ESS are common to both ESS and Eurostat New Cronos.

Lmmd6_example.sav is an SPSS .sav file containing all variables listed above except the MLwiN specific variables

Lmmd6_example.xls is an Excel spreadsheet containing all variables listed above, except the MLwiN specific variables.

5.8. References/further reading

Web:

European social survey: www.europeansocialsurvey.org

Eurostat New Cronos: www.esds.ac.uk/international - choose Eurostat New Cronos

Centre for Multilevel modelling: useful resources and links. MLwiN software and manuals and courses on basic and advanced multilevel modelling.
www.cmm.bristol.ac.uk/

Centre for Census and Survey Research: courses on advanced data analysis and multilevel modelling. Research is carried out here on methods for combining data and multi level modelling and the ESS. See: www.ccsr.ac.uk

Books:

Snijders T and Bosker R (1999) 'Multilevel modelling' Sage. – a good introduction to the topic.

Goldstein (2003) 'Multilevel statistical models' Edward Arnold – a more technical discussion.

Papers:

Fieldhouse E, Tranmer M, Russell A (2007) "Something about young people or something about elections? Electoral participation of young people in Europe: evidence from a multilevel analysis of the European Social Survey."
European Journal of Political Research