Module 14: Missing Data
MLwiN/REALCOM Practical

Jonathan Bartlett & James Carpenter
London School of Hygiene & Tropical Medicine
www.missingdata.org.uk

Supported by ESRC grant RES 189-25-0103 and MRC grant G0900724

Pre-requisites

- You should be able to use MLwiN to fit single-level and multilevel regression models (see MLwiN practicals for Modules 3 and 5)
- Single-level logistic regression in MLwiN (Module 6)
- You need to have installed the MLwiN and REALCOM Impute packages, which are available from www.bristol.ac.uk/cmm.

Online resources:
www.missingdata.org.uk

Contents

Introduction to the Class Size Data

You will be analysing data derived from the class size study carried out by Peter Blatchford and colleagues at the Institute of Education, London, and kindly made available to us. Some study findings are described in Blatchford *et al* 2002. Please note that the dataset we use here has been extracted from the original project for the purposes of illustrating missing data concepts and methods, and so it is not representative of the original data. Furthermore, note that the dataset we use here contains different variables to the one referred to in the corresponding Concepts document.

To open the worksheet:

From within the LEMMA Learning Environment

- Go to **Module 14: Missing Data**, and scroll down to ***MLwiN*** **Datafiles**
- If you do not already have MLwiN to open the datafile with, click **(*get MLwiN*).**
- Click " 14.1.wsz"

The **Names** window will appear.

Table 14.1. Variables contained in the class size data

Variable name	Description and coding
pupil	Pupil identifier
school	School identifier
nlitpre	Pre-reception literacy score, with 1,741 missing values
nlitpost	Literacy score at the end of reception
fsmn	Eligible for free school meals (1=yes, 0=no)
gend	Gender (1=boys, 0=girls)
tentry	Term of school entry (1=Spring or Summer, 0=Autumn)
cons	Constant, set to 1
nlitpre_full	Pre-reception literacy score, with no missing values

The dataset contains literacy scores from pupils in the year before the reception year and scores from the end of the reception year. Note that the test scores have been normalised.

It also contains information on the child's gender and whether they were eligible to receive free school meals. The **nlitpre** variable contains 1,741 missing values. We have made these values missing artificially, in order to illustrate some of the concepts and methods we will use. We thus also have the ariable **nlitpre_full**, which is the **nlitpre** variable before any values were made missing. We emphasize that in practice the **nlitpre_full** variable would not be available.

The dataset is multilevel (clustered, hierarchical), because children are nested within schools. For the moment, we will completely ignore this multilevel structure in order to illustrate the concepts and methods in the more standard single-level setting. At the end of the practical, in 0, we consider a multilevel analysis of the data.

P14.1 The Model of Interest

Our model of interest throughout will be the linear regression of post-reception literacy scores on a child's pre-reception literacy score, eligibility for free school meals, and their gender.

We begin by fitting the model of interest to the full data, i.e. by using the **nlitpre_full** variable as a covariate, rather than **nlitpre**. We again emphasize that in practice this would not be possible, since only the partially observed variable **nlitpre** would be available.

- Click on the **Model** menu and then **Equations** to open the Equations window
- Check that the model matches the one shown in Figure 14.1
- Click the **Start** button to estimate the model parameters

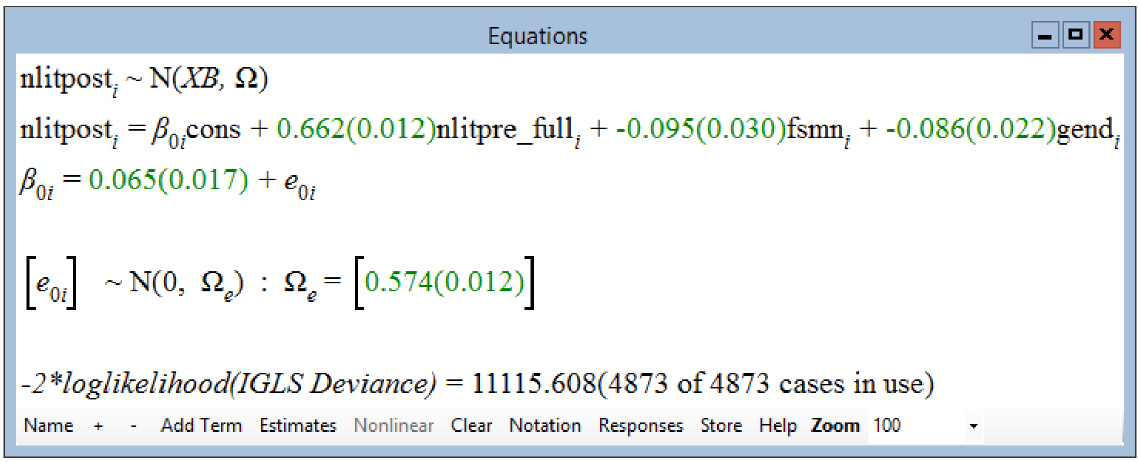


Figure 14.1. Model of interest fitted to full data

Figure 14.1 shows the fitted model, using the full data (note the message that 4873 of 4873 cases are in use). Comparing the coefficients to their standard errors, we see that all three covariates are statistically significantly associated with the post-reception literacy score. Children with high pre-reception scores tended to have higher post-reception literacy scores. Those who were eligible for free school meals had lower scores than those who were not, and boys had lower scores on average than girls.

We will take the coefficients shown in Figure 14.1 as the gold standard, to which we compare subsequent estimates which are based on the partially observed **nlitpre** variable (rather than **nlitpre_full**).

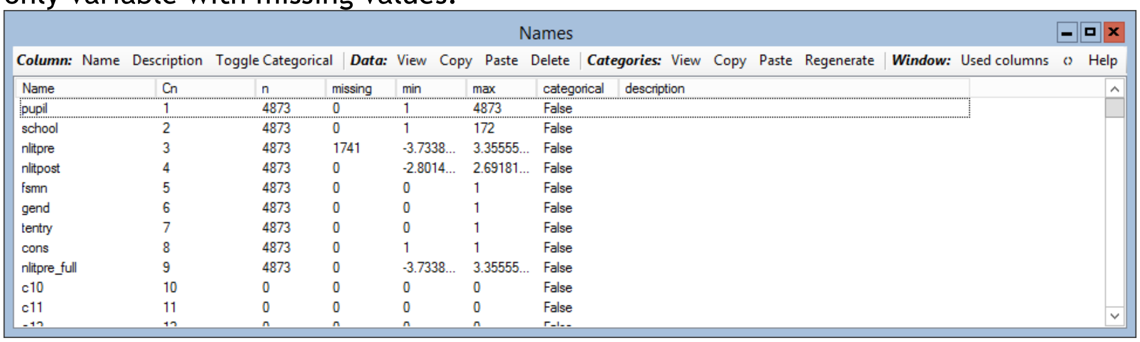
P14.2 Investigating Missingness

In this section we investigate the missing data in the class size dataset. We have already noted that the **nlitpre** variable contains missing values. First, open the worksheet:

From within the LEMMA Learning Environment

- Go to **Module 14: Missing Data**, and scroll down to ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).
- Click “ 14.2.wsz”

The number of missing values in each variable can be easily viewed in the Names window. The Names window should appear when you load the worksheet (otherwise click on **Data Manipulation**, then **Names**). This shows that **nlitpre** is the only variable with missing values.



Name	Cn	n	missing	min	max	categorical	description
pupil	1	4873	0	1	4873	False	
school	2	4873	0	1	172	False	
nlitpre	3	4873	1741	-3.7338...	3.35555...	False	
nlitpost	4	4873	0	-2.8014...	2.69181...	False	
fsmn	5	4873	0	0	1	False	
gend	6	4873	0	0	1	False	
tentry	7	4873	0	0	1	False	
cons	8	4873	0	1	1	False	
nlitpre_full	9	4873	0	-3.7338...	3.35555...	False	
c10	10	0	0	0	0	False	
c11	11	0	0	0	0	False	

Figure 14.2. The Names window for the class size dataset (14.2.wsz)

P14.2.1 Investigating the missingness mechanism

We now investigate the missingness mechanism for the **nlitpre** variable. This will help inform us as to what assumptions might be plausible for the mechanism, and consequently which approaches to handling the missing values will be appropriate. To do this, we first generate a variable which indicates whether the **nlitpre** variable is observed for a child or not.

- Click on **Data Manipulation** and then **Command interface**
- In the **Command interface** window which opens, type the following commands:

```
eqmi 0
calc c10 = 1*('nlitpre'!=MISSING)
name c10 'r'
eqmi 1
```

We use the `eqmi` command to temporarily change MLwiN’s treatment of missing values. In the `calc` command we then set column 10 to equal one when **nlitpre** is not missing and zero otherwise, and rename this column to **r**. Lastly we restore the `eqmi` setting back to its original value. To check that this worked:

- Using the **Names** window, highligh **nlitpre** and **r** and, under **Data**, click **View**.

Figure 14.3 shows these two columns for the first 16 rows in the dataset. This confirms that when **nlitpre** is missing, the new **r** variable is set to zero, and it is one when **nlitpre** is observed.

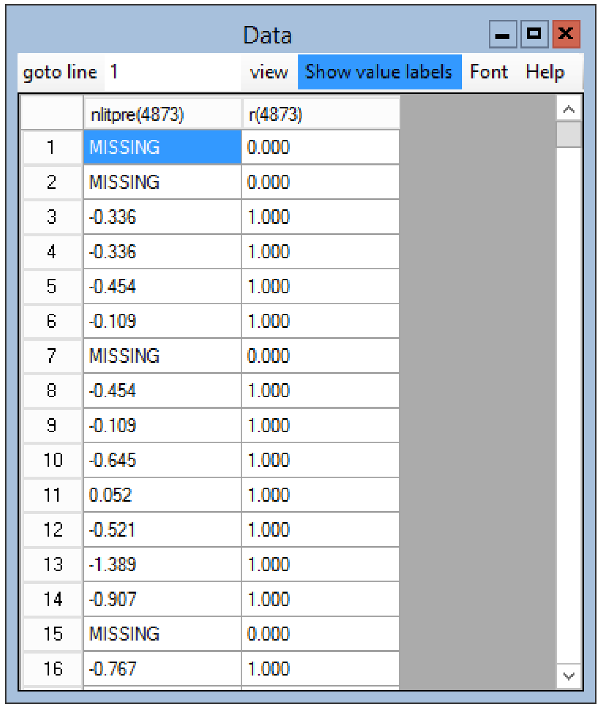


*Figure 14.3. Data view showing the contents of the **nlitpre** and **r** columns.*

We now investigate the missingness mechanism for **nlitpre** by fitting logistic regression models for the missingness indicator variable **r**.

- Open the **Equations** window (from the **Model** menu)
- Click the **Clear** button to clear the linear regression model for **nlitpost**
- Set up a logistic regression model. Select **r** as the dependent variable with **pupil** as the level 1 identifier. Click on the 'N' in r_i ~ N(XB, Ω) to specify the distribution of **r** and link function: check **Binomial** and **logit**. Click on the red **n_i** (the denominator) and select **cons**. Using **Add Term**, select **cons**, **gender**, **nlitpost** and **fsmn** as independent variables. (See the MLwiN practical of Module 6 for further details of single-level logistic regression in MLwiN.)
- Click **Start** to fit the model, and click Estimates twice to see the parameter estimates