Creating a list of missingness patterns in MLwiN

Rebecca Pillinger

17th April 2013

Centre for Multilevel Modelling
University of Bristol

Creating a list of missingness patterns in MLwiN

Rebecca Pillinger

17th April 2013

Frequently the data we analyse will be incomplete, with some cases lacking observations for some variables. It is generally useful to investigate this phenomenon, and one part of this investigation may be to examine the patterns of missingness among the variables of interest. For example, it may be that there are a few variables which are either all observed or all missing for most cases.

It is very easy in MLwiN to produce a list of the patterns of missingness which appear in the data, with a record of how many cases have each pattern.

A note on the example data

The data which appear in the screenshots in this document demonstrating the process are taken from the US National Collaborative Perinatal Project (NCPP)[1]. They are in the public domain and can be found at ftp://sph-ftp.jhsph.edu/cpp/[2], together with extensive documentation. Here we use a subset consisting of all twins who survived at least to age 8. The variables used have been renamed by the present author for ease of use. They consist of the zygosity of the twin pair (**zygosity**), a sex variable created by the present author using all the variables recording sex available in the data (so as to reduce the missingness of sex as far as possible; **sex**), the age when the WISC and WRAT were administered (**age**), the sex variable from the same dataset as the WISC and WRAT (**sex7**), scores for the three WRAT tests **spelling**, **reading** and **arithmetic**, raw full scale WISC score (**FS**), WISC full scale IQ (**FSIQ**), raw WISC verbal score (**verbal**), WISC verbal IQ (**verbalIQ**), raw WISC performance score (**perf**), WISC performance IQ (**perfIQ**), parental education (**e7**), occupation (**o7**) and income (**i7**) when the twins were age 7, and two different codings of parental education (**e01** and **e02**), occupation (**o01** and **o02**) and income (**i01** and **i02**) at registration into the study (before the birth of the twins).

Checking missingness is recorded correctly

First we make sure that after importation to MLwiN, all the values that represent missingness have been recoded to MLwiN's missing value. Depending on how the data have been imported and what values were used for missingess in the original dataset, this may have

[1] The LaTeX example table at the end of the document is not based on the real data.

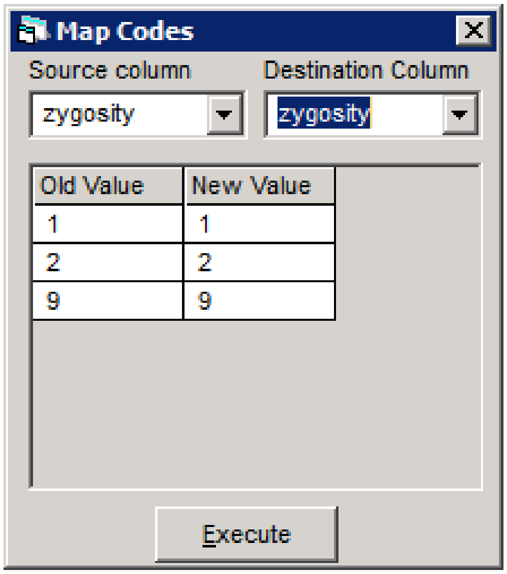
[2] Lawlor et al. (2005)

been done automatically or we may need to additionally do some recoding manually. If the data have been imported from Stata, SPSS, SAS or Minitab then all values that were designated as coding missingness in the original worksheet should have been automatically set to MLwiN's missing value. We can check by seeing what values appear in the data for each variable. We can do this by examining the **Names** window to see what the range of values appearing in the data is for each variable. A value that has been recoded to missing will not be shown as the minimum or maximum value - but be careful not to miss values that denote missingness that are between the minimum and maximum legitimate values for the variable!. Or we can use **Tabulate** under **Basic Statistics** menu. Or we can select **recode → by value** from the **Data Manipulation** menu.

If we do need to recode any value to MLwiN's MISSING value, we can do this by:

- From the **Data Manipulation** menu select **recode → by value**
- From both the drop-down boxes at the top of the window, select the variable which you need to recode
- In the right hand column, next to the value that needs to be recoded to missing, type MISSING
- Press **Execute**

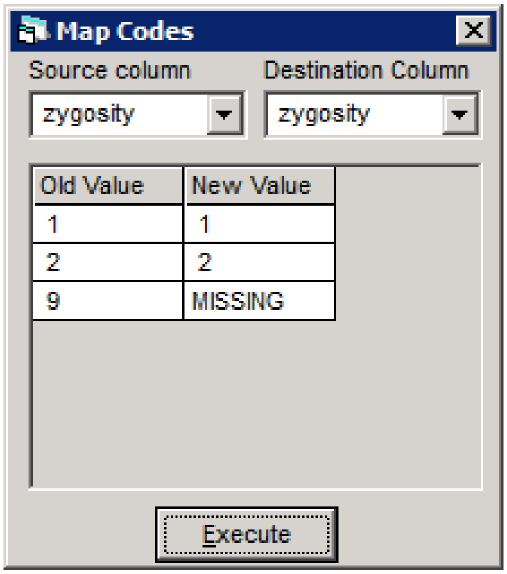
For example, recoding the value 9 to MISSING for the variable **zygosity**:



Source column: zygosity

Destination Column: zygosity

Old Value	New Value
1	1
2	2
9	9



Source column: zygosity

Destination Column: zygosity

Old Value	New Value
1	1
2	2
9	MISSING

Creating missingness indicator variables

Next we need to create a dummy variable for each variable of interest that takes the value 1 when that variable is missing and 0 when it is observed[3]. We can do this using a macro, like the one that appears below[4]. (Change c3 to c24 to whichever columns contain your variables of interest, and c103 to c124 to columns that are free in your worksheet). The second section, where the newly created columns are assigned names is optional; the later steps in the process do not use the names, but it can be useful to name the columns anyway to keep track of things.

```
▶ calc c103 = c3 == MISSING

▶ calc c104 = c4 == MISSING
```

[3] Creating a variable that takes the value 0 when that variable is missing and 1 when it is observed would be just as valid and work just as well in the following steps, and has perhaps a more sensible interpretation, but we do not do this since the code in this step would be slightly more complicated.

[4] Note that the macro could be more compactly coded:

```
▶ LOOP b1 3 24
▶ CALC b2 = b1 + 100
▶ CALC cb2 = cb1 == MISSING
▶ ENDLoop
▶
▶ NAME c103 '\zygmiss'c104 '\sexmiss'c105 '\agemiss'c106 '\sex7miss'c107
  '\spellmiss'c108 '\readmiss'c109 '\arithmiss'c110 '\FSmiss'c111 '\FSIQmiss'c112
  '\verbalmiss'c113 '\verbalIQmiss'c114 '\perfmiss'c115 '\perfIQmiss'c116
  '\e7miss'c117 '\o7miss'c118 '\i7miss'c119 '\e01miss'c120 '\o01miss'c121
  '\i01miss'c122 '\e02miss'c123 '\o02miss'c124 '\i02miss'
```

```
► calc c105 = c5 == MISSING
► calc c106 = c6 == MISSING
► calc c107 = c7 == MISSING
► calc c108 = c8 == MISSING
► calc c109 = c9 == MISSING
► calc c110 = c10 == MISSING
► calc c111 = c11 == MISSING
► calc c112 = c12 == MISSING
► calc c113 = c13 == MISSING
► calc c114 = c14 == MISSING
► calc c115 = c15 == MISSING
► calc c116 = c16 == MISSING
► calc c117 = c17 == MISSING
► calc c118 = c18 == MISSING
► calc c119 = c19 == MISSING
► calc c120 = c20 == MISSING
► calc c121 = c21 == MISSING
► calc c122 = c22 == MISSING
► calc c123 = c23 == MISSING
► calc c124 = c24 == MISSING
►
► name c103 '\zygmiss'
► name c104 '\sexmiss'
► name c105 '\agemiss'
► name c106 '\sex7miss'
► name c107 '\spellmiss'
► name c108 '\readmiss'
► name c109 '\arithmiss'
► name c110 '\FSmiss'
► name c111 '\FSIQmiss'
► name c112 '\verbalmiss'
► name c113 '\verbalIQmiss'
► name c114 '\perfmiss'
► name c115 '\perfIQmiss'
► name c116 '\e7miss'
► name c117 '\o7miss'
► name c118 '\i7miss'
► name c119 '\e01miss'
► name c120 '\o01miss'
► name c121 '\i01miss'
► name c122 '\e02miss'
► name c123 '\o02miss'
► name c124 '\i02miss'
```

Calculating the pattern of missingness for each case

Now we create a variable which records for each case the pattern of missingness among these variables:

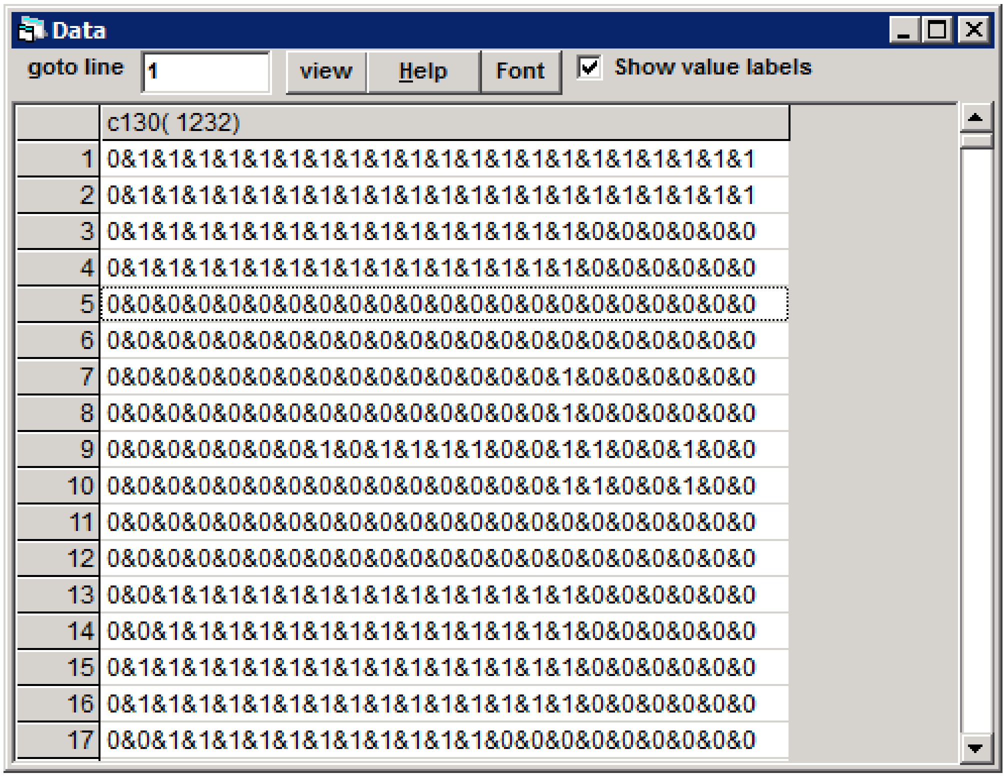
- In the **Command interface**, type

```
► comb c103 c104 c105 c106 c107 c108 c109 c110 c111 c112 c113 c114
  c115 c116 c117 c118 c119 c120 c121 c122 c123 c124 c130
```

Here the last column in the arguments (c130) is a free column where the missingness patterns will be stored, and the others are the missingness indicators we just created[5].

This creates a categorical variable whose category labels carry information about the missingness pattern. If we highlight **c130** in the **Names** window and click on **View** under **Data**, we see (after resizing the column to make it wider):

[5] Of course, we could add this command to the end of the macro that we use to create the missingness indicators, instead of typing it in the Command interface.



goto line 1

☑ Show value labels

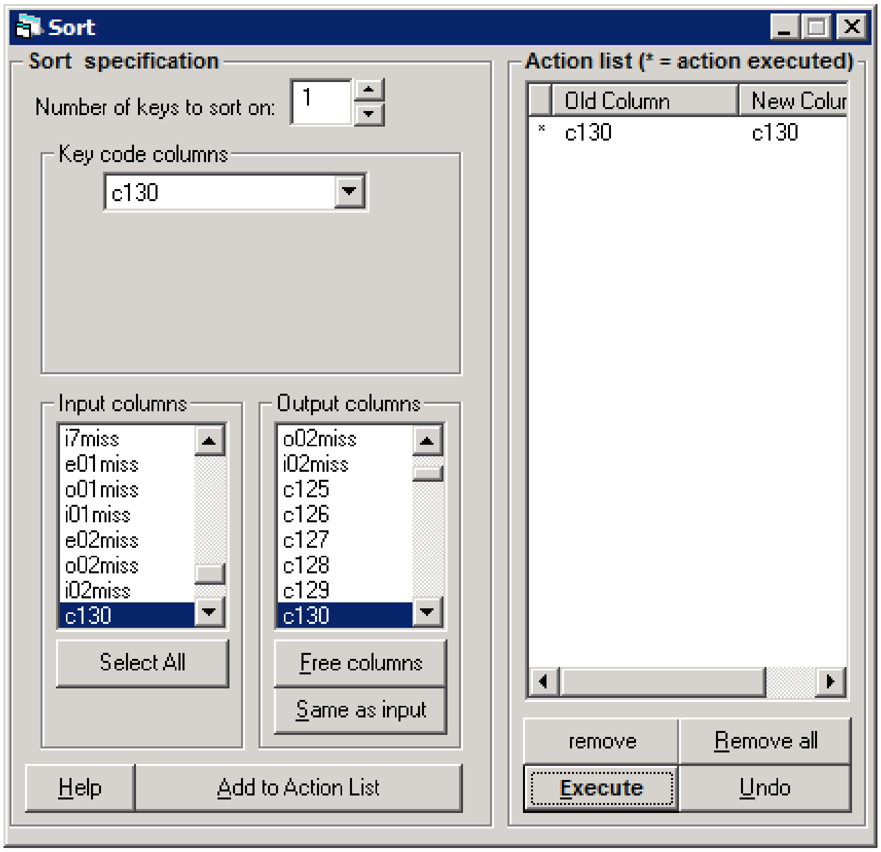
	c130(1232)
1	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1
2	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1
3	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
4	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
5	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
6	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
7	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0
8	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0
9	0&0&0&0&0&0&0&1&0&1&1&1&1&0&0&1&1&0&0&1&0&0
10	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&0&0
11	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
12	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
13	0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
14	0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
15	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
16	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
17	0&0&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0

We can see that for the first case, **zygmiss** is equal to 0 and all the other missingness indicators are equal to 1, implying that the first case has **zygosity** observed but all other variables missing.

Counting the missingness patterns

We can now count up the number of cases with each pattern, and then condense the information so that we have one row per missingness pattern:

- From the **Data Manipulation** menu, select **Sort**
- Under **Key code columns** select **c130**
- Under **Input columns** select **c130**
- Under **Output columns** click **Same as input**
- Click **Add to Action List** and **Execute**



Action list (* = action executed)

Number of keys to sort on: 1

Key code columns: c130

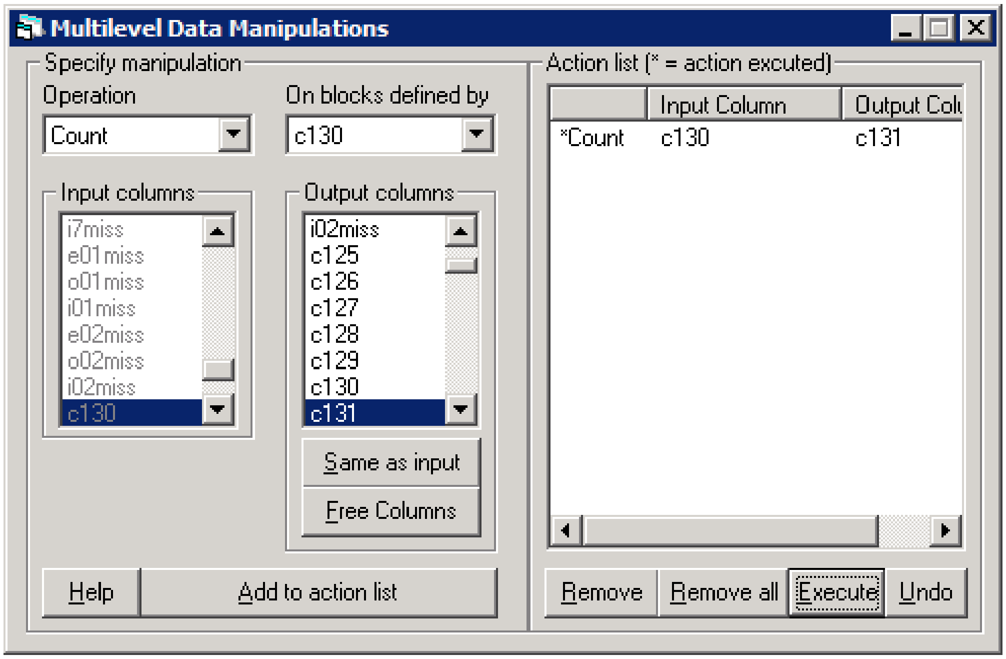
Input columns: i7miss, e01miss, o01miss, i01miss, e02miss, o02miss, i02miss, c130 (selected)

Output columns: o02miss, i02miss, c125, c126, c127, c128, c129, c130 (selected)

	Old Column	New Colur
*	c130	c130

This groups all the missingness patterns that are the same together.

- From the **Data Manipulation** menu, select **Multilevel Data Manipulations**
- From the **Operation** drop-down box select **Count**
- From the **On blocks defined by** drop-down box select **c130**
- Under **Output columns** select **c131**
- Click **Add to Action List** and **Execute**



Action list (* = action excuted)

Operation: Count

On blocks defined by: c130

Input columns: i7miss, e01miss, o01miss, i01miss, e02miss, o02miss, i02miss, c130

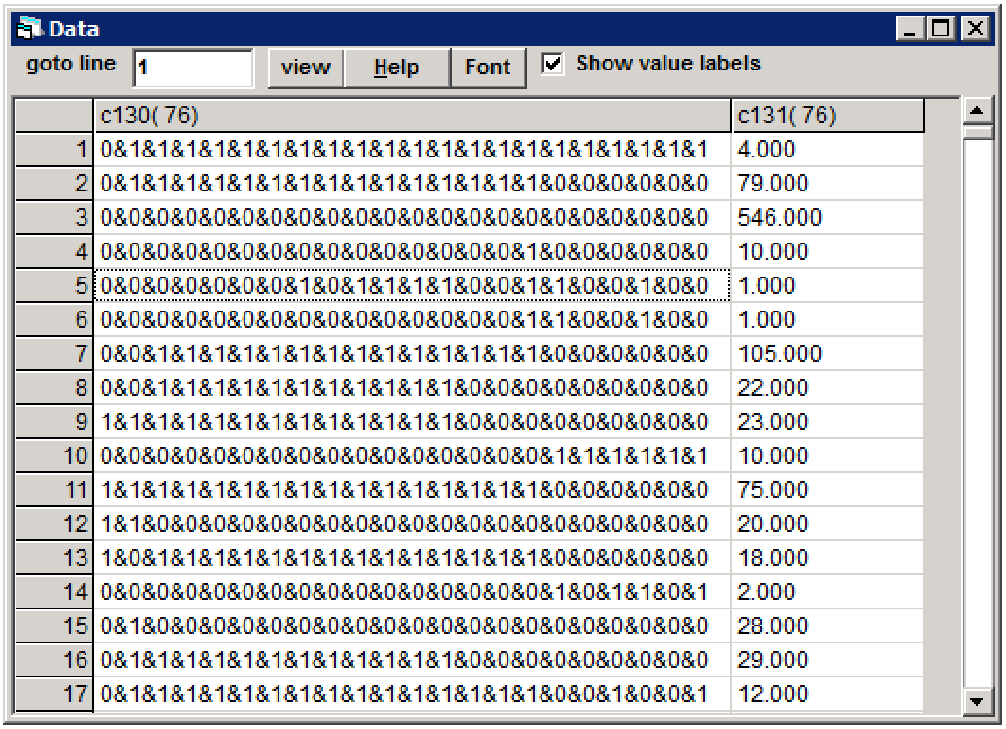
Output columns: i02miss, c125, c126, c127, c128, c129, c130, c131 (selected)

	Input Column	Output Colu
*Count	c130	c131

This counts how many times each pattern appears and stores the result in a new column, c131

- From the **Data Manipulation** menu, select **unreplicate**
- From the **Take first entry in blocks defined by** drop-down box select **c130**
- Under **Input columns** select **c130** and **c131**
- Under **Output columns** click **Same as input**
- Click **Add to Action List** and **Execute**

This keeps just one row of c130 and c131 for each missingness pattern, so that we have a list of all the missingness patterns in c130, each appearing just once, and a record of how many times that pattern appears in the data in c131:

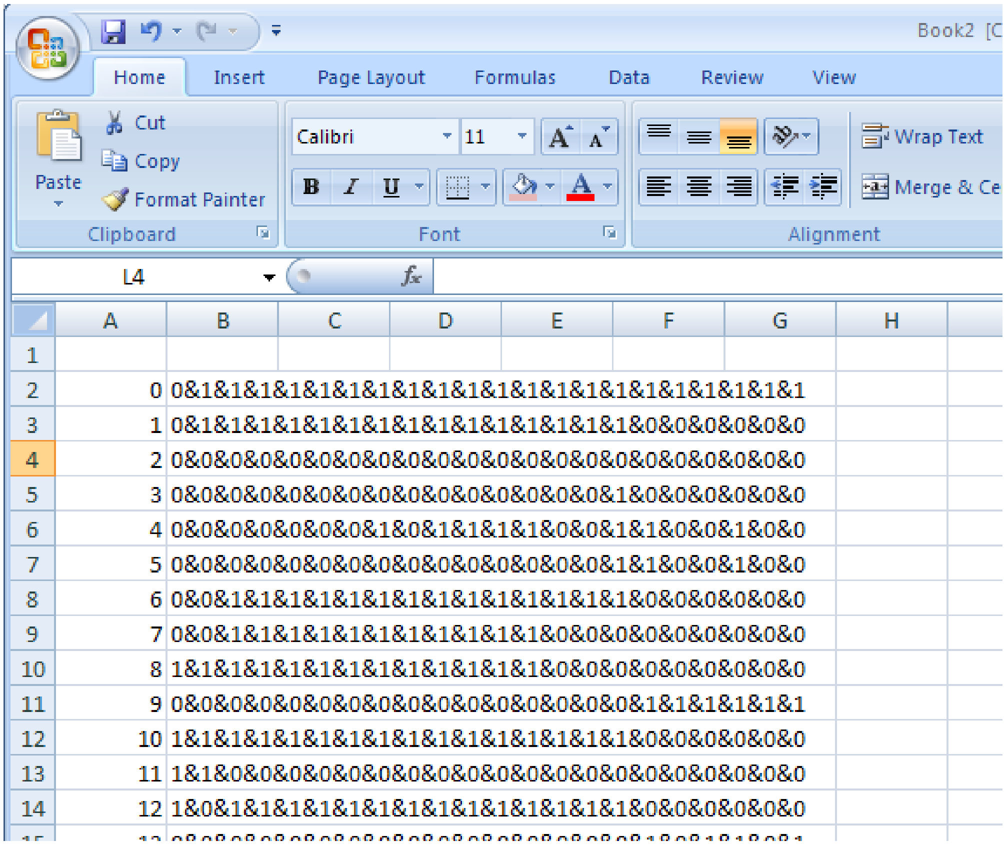


	c130(76)	c131(76)
1	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1	4.000
2	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0	79.000
3	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0	546.000
4	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0	10.000
5	0&0&0&0&0&0&0&1&0&1&1&1&1&0&0&1&1&0&0&1&0&0	1.000
6	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&0&0	1.000
7	0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0	105.000
8	0&0&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0	22.000
9	1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0	23.000
10	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1	10.000
11	1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0	75.000
12	1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0	20.000
13	1&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0	18.000
14	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&1&0&1	2.000
15	0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0	28.000
16	0&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0	29.000
17	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&1&0&0&1	12.000

Exporting the missingness patterns

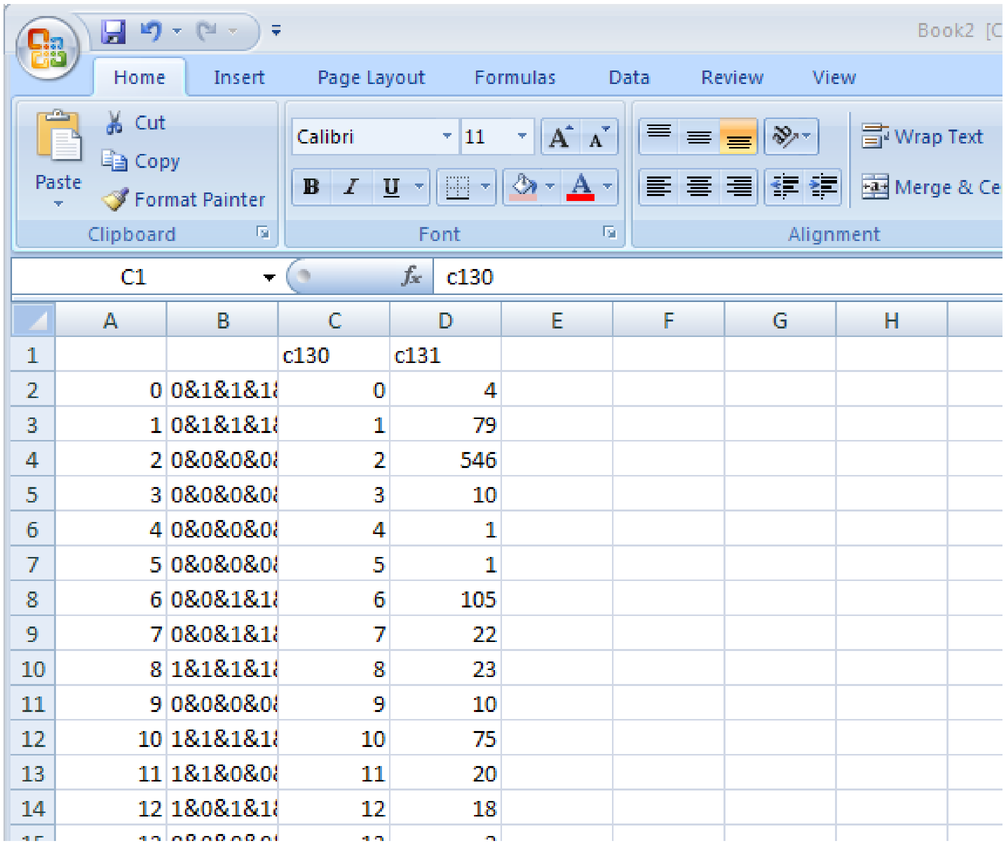
We can view the information we have generated in MLwiN, by highlighting **c130** and **c131** in the **Names** window and pressing **View** under **Data**. We can also export it from MLwiN, for example into Excel:

- In the **Names** window, highlight **c130**
- Click **Copy** under **Categories**
- In Excel, click on square A2 and paste



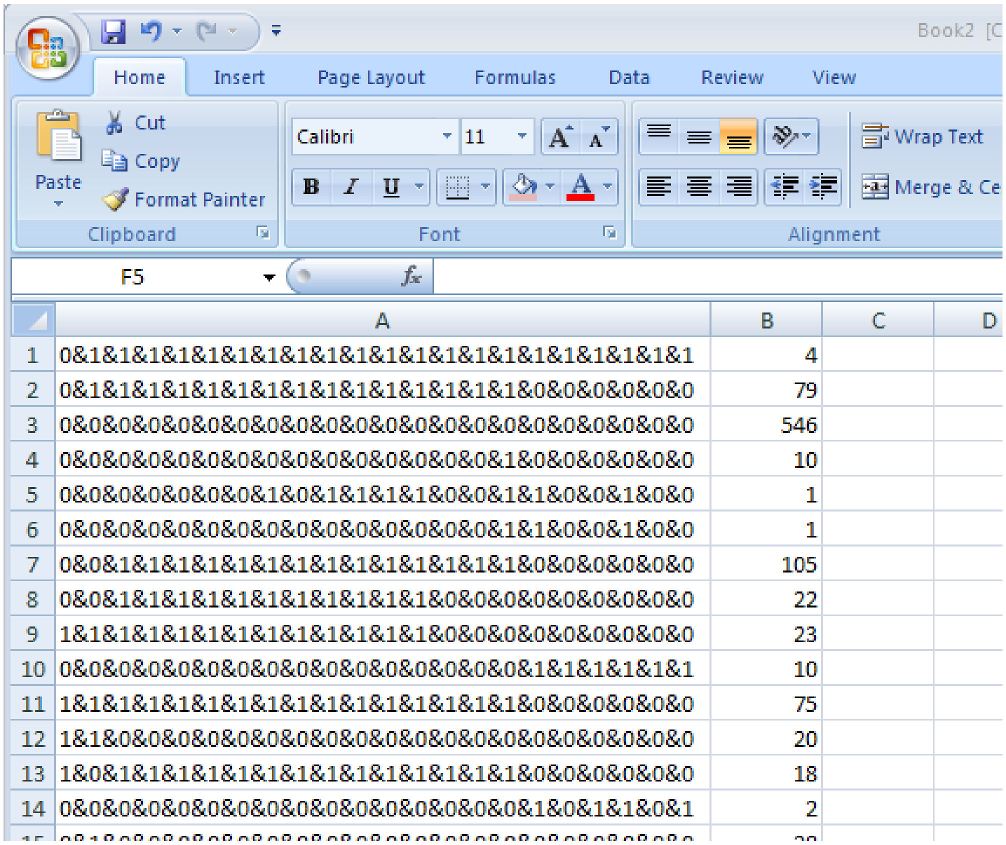
	A	B
1		
2	0	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1
3	1	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
4	2	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
5	3	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0
6	4	0&0&0&0&0&0&0&1&0&1&1&1&1&0&0&1&1&0&0&1&0&0
7	5	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&0&0
8	6	0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
9	7	0&0&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0
10	8	1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0
11	9	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1
12	10	1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0
13	11	1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
14	12	1&0&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0

- In MLwiN, in the **Names** window, highlight both **c130** and **c131**
- Click **Copy** under **Data**
- In Excel, click on square C1 and paste



	A	B	C	D
1			c130	c131
2	0	0&1&1&1	0	4
3	1	0&1&1&1	1	79
4	2	0&0&0&0	2	546
5	3	0&0&0&0	3	10
6	4	0&0&0&0	4	1
7	5	0&0&0&0	5	1
8	6	0&0&1&1	6	105
9	7	0&0&1&1	7	22
10	8	1&1&1&1	8	23
11	9	0&0&0&0	9	10
12	10	1&1&1&1	10	75
13	11	1&1&0&0	11	20
14	12	1&0&1&1	12	18

- Delete column A and column C and the first row



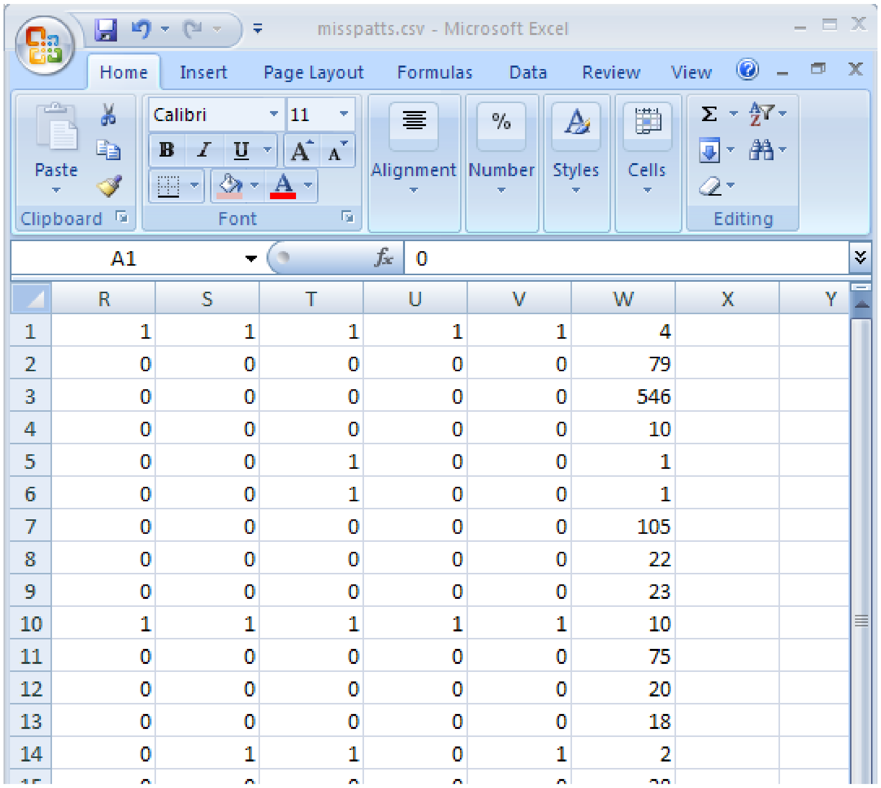
	A	B	C	D
1	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1	4		
2	0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0	79		
3	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0	546		
4	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0	10		
5	0&0&0&0&0&0&0&1&0&1&1&1&1&0&0&1&1&0&0&1&0&0	1		
6	0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&0&0	1		
7	0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0	105		
8	0&0&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0	22		
9	1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0	23		
10	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1	10		
11	1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0	75		
12	1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0	20		
13	1&0&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0	18		
14	0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&1&0&1	2		

We could now, for example, sort to list the patterns in descending order of frequency if so desired.

Bonus: creating a separate column for each variable

Something else we might like to do is to separate out the missingness patterns so that instead of a single entry containing something like `0&1&1&1&1&1&1&1&1&1&1&1 &1&1&1&1&1&1&1&1&1` we have columns corresponding to the missingness indicators we created, i.e. a column with a 0, then a column with a 1, then a column with a 1 and so on.

- Save the Excel file as a CSV (comma separated value; listed in Excel as ‘comma delimited’) file
- Close it
- Open it with a text editor (e.g. WordPad or Notepad)
- Replace every `&` with a `,`
- Save the file
- Open with Excel



A1 fx 0

	R	S	T	U	V	W	X	Y
1	1	1	1	1	1	4		
2	0	0	0	0	0	79		
3	0	0	0	0	0	546		
4	0	0	0	0	0	10		
5	0	0	1	0	0	1		
6	0	0	1	0	0	1		
7	0	0	0	0	0	105		
8	0	0	0	0	0	22		
9	0	0	0	0	0	23		
10	1	1	1	1	1	10		
11	0	0	0	0	0	75		
12	0	0	0	0	0	20		
13	0	0	0	0	0	18		
14	0	1	1	0	1	2		

The columns can easily be given names to indicate which variable in the dataset each refers to.

Special bonus for LaTeX users

It is very easy to create a nice looking chart in LaTeX from the missingness patterns exported into Excel.

First, in the preamble of your `.tex` document, create two new macros, one which specifies what to display for a missing observation, and one which specifies what to display for an observed observation. There are many possibilities, e.g. simply the text 'miss' for missing and 'obs' for observed, but here we will use pale red squares for missing and larger, darker green squares for observed:

```
\usepackage{tikz}
\definecolor{pred}{RGB}{255,174,194}
\definecolor{dgreen}{RGB}{0,113,99}

\newcommand{\miss}{
\begin{tikzpicture}
\fill[pred] (0,0) rectangle (0.1,0.1);
\end{tikzpicture}
}

\newcommand{\pres}{
```

```
\begin{tikzpicture}
\fill[dgreen] (0,0) rectangle (0.2,0.2);
\end{tikzpicture}
}
```

Now, beginning at the point where you have just deleted columns A and C and the first row (i.e. don't follow the instructions in the previous section):

- Save the file in Excel as a CSV (comma separated value; listed in Excel as 'comma delimited') file
- Close it
- Open it with a text editor (e.g. WordPad or Notepad)
- Replace every `,` with a `&`
- Replace every `0&` with `\pres&` and every `1&` with `\miss&`
- Add \\ to the end of every line

You now have the body of a `tabular` environment, and can add appropriate code above and below before including it in your `.tex` document, e.g.

```
\begin{tabular}{cccccc@{\hspace{0.5cm}}r}
\rotatebox{90}{zygosity}&\rotatebox{90}{age}&\rotatebox{90}{sex}
&\rotatebox{90}{full scale IQ}&\rotatebox{90}{verbal IQ}
&\rotatebox{90}{performance IQ}&\\
\hline

\pres&\miss&\miss&\miss&\miss&\miss&4\\
\miss&\miss&\miss&\miss&\miss&\miss&79\\
\pres&\pres&\pres&\pres&\pres&\pres&546\\
\pres&\pres&\pres&\pres&\pres&\miss&1\\
\pres&\pres&\miss&\miss&\miss&\miss&10\\
\pres&\pres&\miss&\pres&\pres&\pres&105\\
\pres&\miss&\miss&\pres&\miss&\miss&1\\
\pres&\pres&\miss&\miss&\miss&\miss&22\\
\miss&\miss&\miss&\miss&\miss&\pres&2\\
\pres&\pres&\pres&\pres&\miss&\pres&3\\
\miss&\miss&\miss&\miss&\pres&\pres&1\\

\hline
\end{tabular}
```

zygosity	age	sex	full scale IQ	verbal IQ	performance IQ	
🟩	▪	▪	▪	▪	▪	4
▪	▪	▪	▪	▪	▪	79
🟩	🟩	🟩	🟩	🟩	🟩	546
🟩	🟩	🟩	🟩	🟩	▪	1
🟩	🟩	▪	▪	▪	▪	10
🟩	🟩	▪	🟩	🟩	🟩	105
🟩	▪	▪	🟩	▪	▪	1
🟩	🟩	▪	▪	▪	▪	22
▪	▪	▪	▪	▪	🟩	2
🟩	🟩	🟩	🟩	▪	🟩	3
▪	▪	▪	▪	🟩	🟩	1

References

Lawlor, J.P., Gladen, E., Dhavale, D., Tamagoglu, D., Hardy, J.B., Duggan, A.K., Eaton, W.W. and Torrey, E.F. (2005) Modernization and enhancement of the collaborative perinatal project (1959-74). Technical report, The Johns Hopkins University, Funded by the Stanley Medical Research Institute, Bethesda MD and NIMH grant 070333