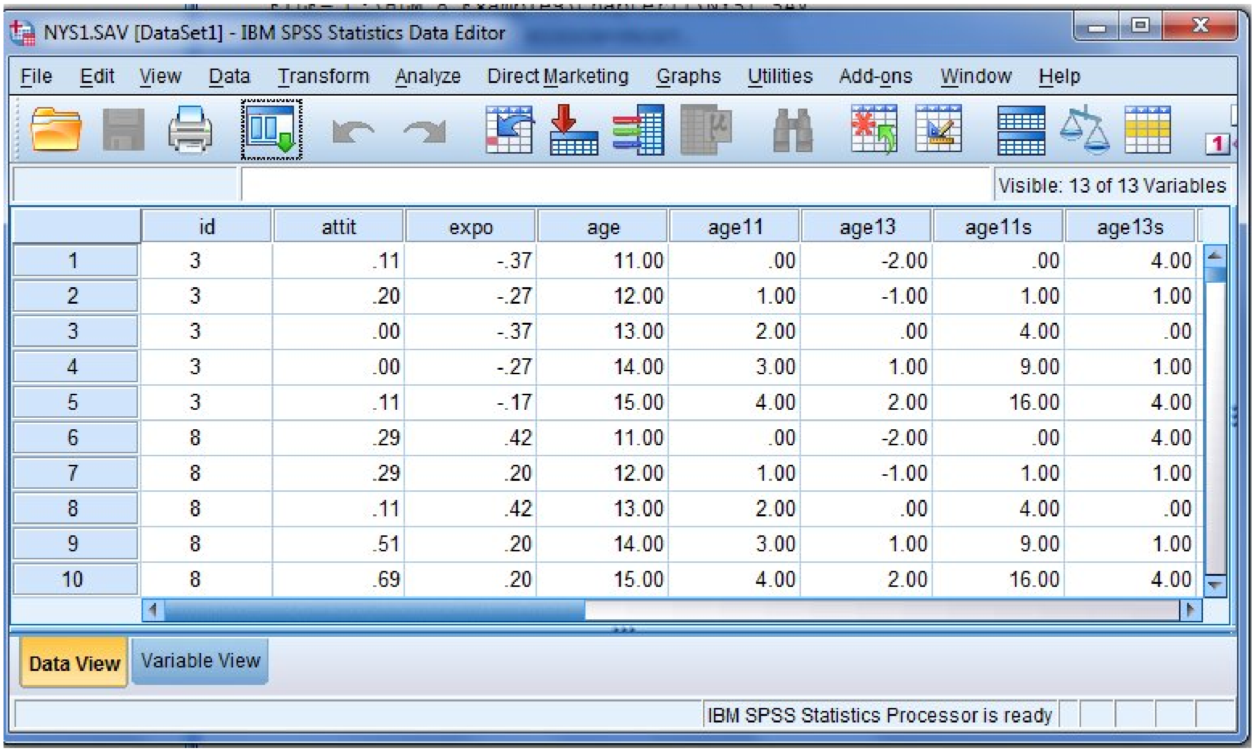
## Preparing data for use in HMLM or HMLM2

Prior to making an MDM file for HMLM or HMLM2, some data preparation is required. Both these models require the addition of indictor variables to the raw data.

As an example, we use the NYS data discussed in Chapter 10 of the HMLM manual. The data file **NYS1.SAV** contains observations collected from annual interviews of 11-year old youths beginning in 1976 and continuing for 5 years. For each subject, a maximum of 5 measurements, one per year, is present in the data. In the data (shown below)


NYS1.SAV [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Visible: 13 of 13 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

| | id | attit | expo | age | age11 | age13 | age11s | age13s |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | .11 | -.37 | 11.00 | .00 | -2.00 | .00 | 4.00 |
| 2 | 3 | .20 | -.27 | 12.00 | 1.00 | -1.00 | 1.00 | 1.00 |
| 3 | 3 | .00 | -.37 | 13.00 | 2.00 | .00 | 4.00 | .00 |
| 4 | 3 | .00 | -.27 | 14.00 | 3.00 | 1.00 | 9.00 | 1.00 |
| 5 | 3 | .11 | -.17 | 15.00 | 4.00 | 2.00 | 16.00 | 4.00 |
| 6 | 8 | .29 | .42 | 11.00 | .00 | -2.00 | .00 | 4.00 |
| 7 | 8 | .29 | .20 | 12.00 | 1.00 | -1.00 | 1.00 | 1.00 |
| 8 | 8 | .11 | .42 | 13.00 | 2.00 | .00 | 4.00 | .00 |
| 9 | 8 | .51 | .20 | 14.00 | 3.00 | 1.00 | 9.00 | 1.00 |
| 10 | 8 | .69 | .20 | 15.00 | 4.00 | 2.00 | 16.00 | 4.00 |

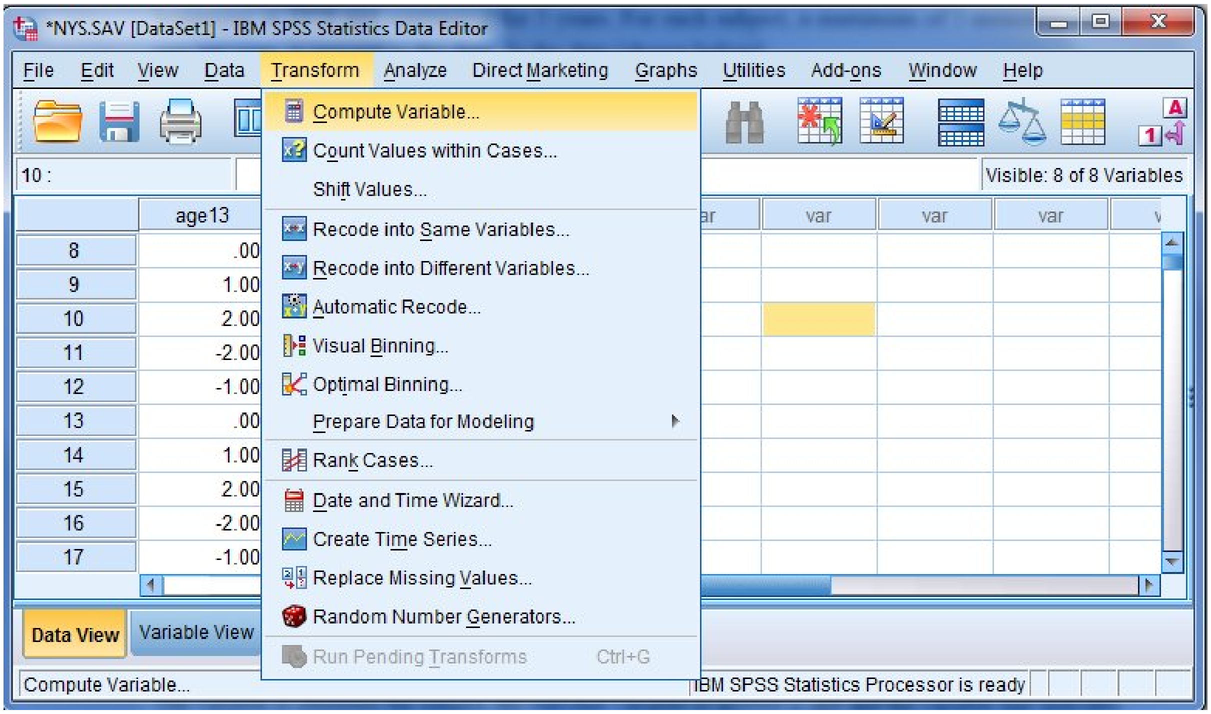
The variable id identifies the subject, the outcome variable of interest is attit and the variable age indicates the time of measurement, ranging between 11 and 15. As we have 5 measurements, 5 indicator variables have to be created to distinguish between the occasions of measurement.

For the two subjects whose data are shown above, complete data is available. Both were measured every year. For both these cases, the five indicator variables IND1, IND2, …IND5 should have the form

| *IND*1 | *IND*2 | *IND*3 | *IND*4 | *IND*5 |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 |

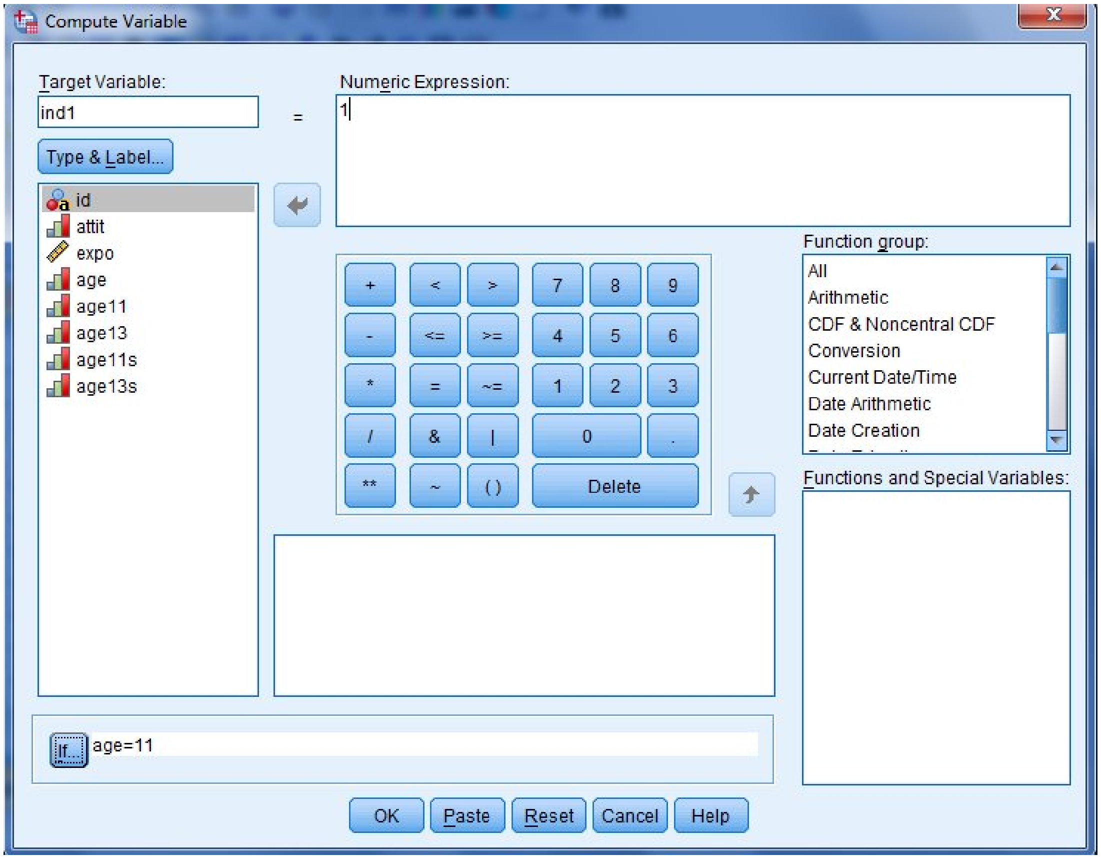
IND1 has a value of 1 at age = 11, and only then. Similarly, IND2, associated with the measurement at age 12, only assumes the value of 1 in the second row.

We use SPSS to illustrate the creation of the indicator variables. Start by selecting the **Compute Variable** option from the **Transform** menu.



| | age13 |
|---|---|
| 8 | .00 |
| 9 | 1.00 |
| 10 | 2.00 |
| 11 | -2.00 |
| 12 | -1.00 |
| 13 | .00 |
| 14 | 1.00 |
| 15 | 2.00 |
| 16 | -2.00 |
| 17 | -1.00 |

On the **Compute Variable** dialog box, create a new variable named IND1 and set the value to 1 (top of dialog box). Now add a condition to it (bottom of dialog box). The completed dialog box requests the creation of the new variable IND1 that will assume a value of 1 if age = 11, that is at the first measurement occasion.



Compute Variable

Target Variable: ind1

Numeric Expression: 1

If... age=11

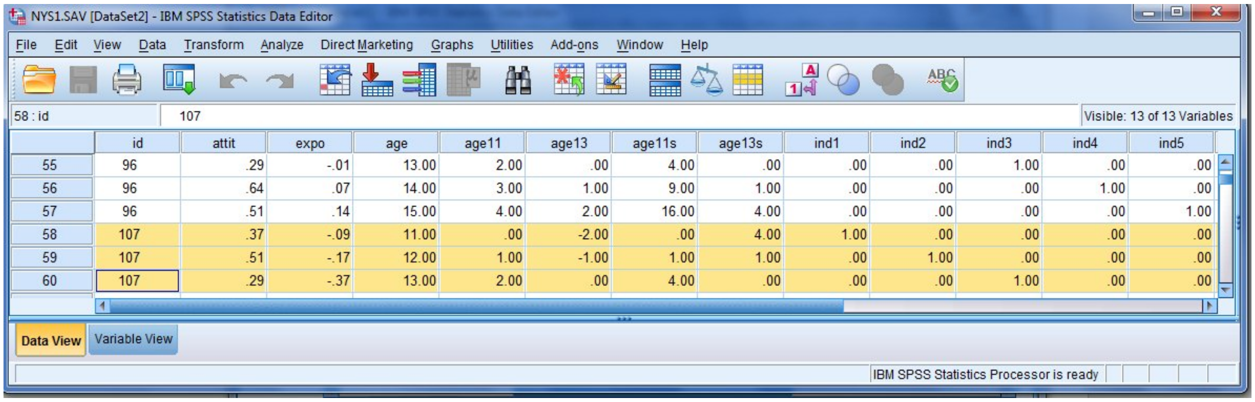
The other indicator variables are created in the same way. Note that within each subject's data, each indicator variable can only be equal to 1 only once. If after completion, a subject's data contains multiple 1's in a column as shown below the program will produce an error message

| *IND*1 | *IND*2 | *IND*3 | *IND*4 | *IND*5 |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 |

"Actual number of occasions exceeds the specified number of occasions"

The program will also complain if the number of indicator variables is larger than the actual number of measurements. For example, suppose the data only contains data on these subjects for four years, even though the study stretched over 5 years. In that case, only 4 indicator variables should be created as the data only contains data for 4 measurement occasions.

When dealing with data collected over time, it frequently happens that the data collected for a subject may be incomplete: a subject may have missed one measurement during the time period, or perhaps simply dropped out. Subject 107 is such a case: data were collected for the first 3 years, but the last two measurements are missing.



58 : id 107

Visible: 13 of 13 Variables

| | id | attit | expo | age | age11 | age13 | age11s | age13s | ind1 | ind2 | ind3 | ind4 | ind5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 55 | 96 | .29 | -.01 | 13.00 | 2.00 | .00 | 4.00 | .00 | .00 | .00 | 1.00 | .00 | .00 |
| 56 | 96 | .64 | .07 | 14.00 | 3.00 | 1.00 | 9.00 | 1.00 | .00 | .00 | .00 | 1.00 | .00 |
| 57 | 96 | .51 | .14 | 15.00 | 4.00 | 2.00 | 16.00 | 4.00 | .00 | .00 | .00 | .00 | 1.00 |
| 58 | 107 | .37 | -.09 | 11.00 | .00 | -2.00 | .00 | 4.00 | 1.00 | .00 | .00 | .00 | .00 |
| 59 | 107 | .51 | -.17 | 12.00 | 1.00 | -1.00 | 1.00 | 1.00 | .00 | 1.00 | .00 | .00 | .00 |
| 60 | 107 | .29 | -.37 | 13.00 | 2.00 | .00 | 4.00 | .00 | .00 | .00 | 1.00 | .00 | .00 |

IBM SPSS Statistics Processor is ready

Note that for this subject, the indicator variables IND4 and IND5 show a value of 0 throughout.

HMLM and HMLM2 do not require complete data at all time points, so “missing data” as in the case shown above, can be incorporated in a model while still allowing the analyst to consider level-1 random effects beyond the $\sigma^2$ estimated in HLM2, HLM3 etc. This is under the assumption that any missing data is “missing at random”. As HMLM and HMLM2 allow estimation of multivariate normal models from incomplete data, HMLM and HMLM2 do not have the usual **Missing Data** options on the **Make MDM** dialog box. The data preparation of the level-1 data file for HMLM2 is done in the same way.

It should also be kept in mind that the number of random effects in a HMLM/HMLM2 model cannot exceed the number of timepoints. If you have 4 measurements indicated by 4 indicators, 3 random effects can be estimated. Attempting to estimate more will lead to the display of the message

> The number of random effects cannot equal or exceed the number of timepoints!

For a comparison of models with different level-1 covariance structures for the NYS data, see Raudenbush & Bryk (Sage, 2nd Ed.) pp. 190-199.