LG-Syntax Module Scoring Tutorial:

Using Saved Model Parameters to Score a File Containing New Cases[1]

DemoData files = ‘gss82white.sav’ and ‘gss82.newcases.sav’
.lgf file = ‘gss82white.lgf’
.lgs file = ‘3-cluster model.lgs’

The Goal:

In this scoring tutorial, we show how to use of the LG Syntax module to classify new cases into the most probable latent class (LC) and obtain posterior membership probabilities (scores) for these new cases. We illustrate the process with the LC model estimated in Tutorial 1: “Using Latent GOLD to Estimate LC Cluster Models”, a 3-class model estimated on the demo data file ‘gss82white.sav’ and saved in the usual GUI model (.lgf) file format ‘gss82white.lgf’.

Specifically, in this tutorial we will:

- save parameter estimates from the model in a syntax (.lgs) file *‘3-cluster model.lgs’*
- score new cases contained in a different data file ‘gss82.newcases.sav’

Note: You may wish to revisit Tutorial 1: “Using Latent GOLD to Estimate LC Cluster Models” to familiarize yourself with the 3-class model estimated before beginning this tutorial.

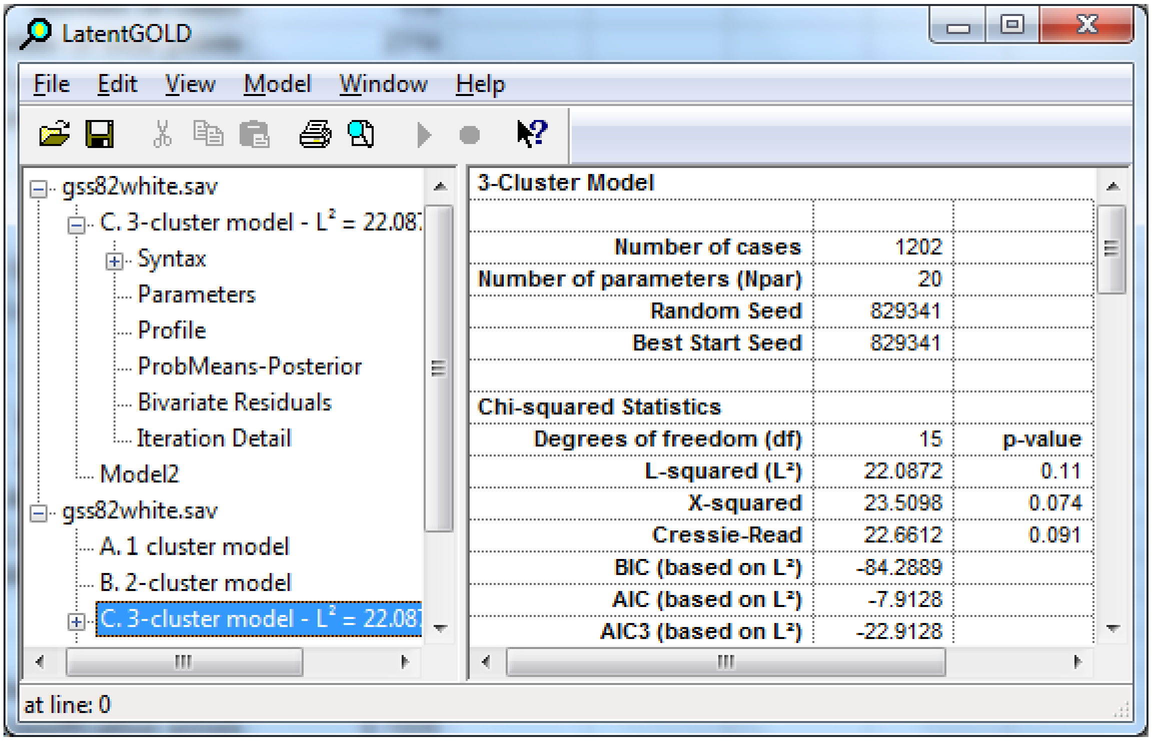
The Scoring Steps:

1. Open ‘gss82white.lgf’
   - From the “File” menu click on “Open”
   - From the demo data directory (e.g., c:\Program Files\LatentGOLD4.5\DemoData) select ‘gss82white.lgf’
2. Re-estimate this 3-cluster model in the usual way:
   - Highlight the model name ‘3-cluster model’ and click ▶
3. Right click on the name ‘3-cluster model’ and select “Generate Syntax”.
4. Re-estimate this model using the syntax module.

[1] This scoring process can be used with any model estimated using Latent GOLD or LG Syntax

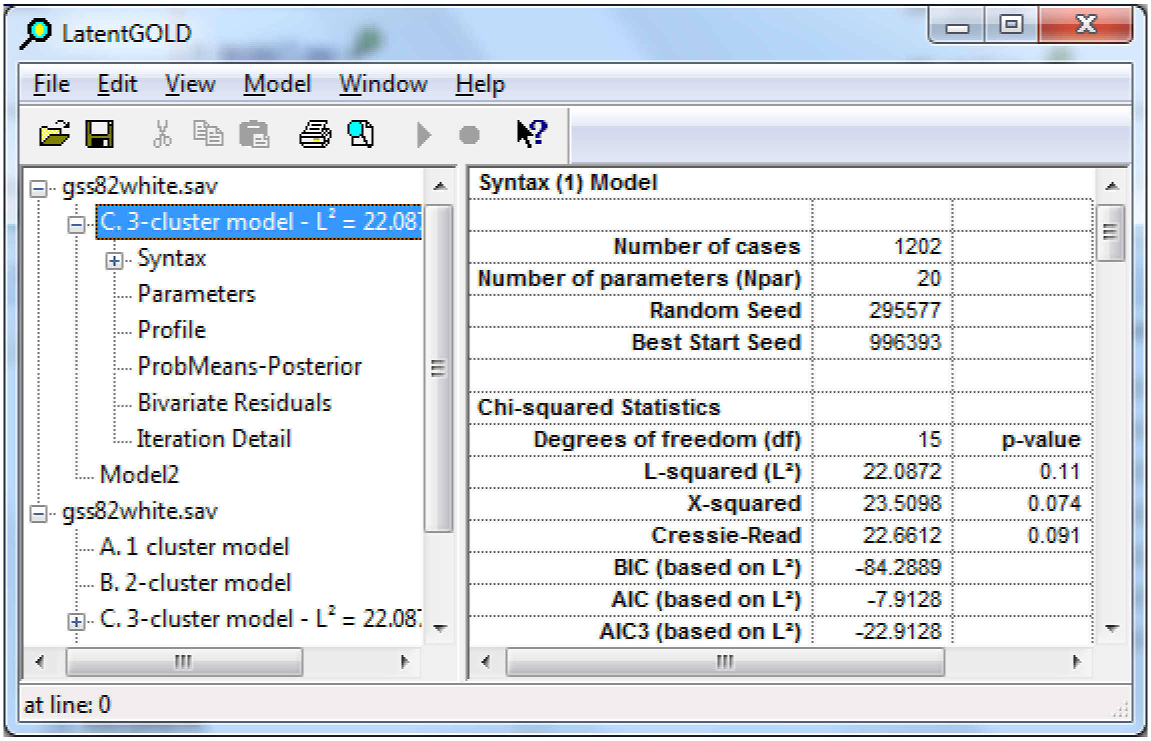
- Click ▶

The resulting summary output (Fig. 1B) is identical to that obtained in step 2 (Fig. 1A). However, the latent class ordering will not generally be preserved in the output sections.



3-Cluster Model		
Number of cases	1202	
Number of parameters (Npar)	20	
Random Seed	829341	
Best Start Seed	829341	
Chi-squared Statistics		
Degrees of freedom (df)	15	p-value
L-squared (L^2)	22.0872	0.11
X-squared	23.5098	0.074
Cressie-Read	22.6612	0.091
BIC (based on L^2)	-84.2889	
AIC (based on L^2)	-7.9128	
AIC3 (based on L^2)	-22.9128	

Figure 1A: Summary Output for the 3-class model as obtained in step 2



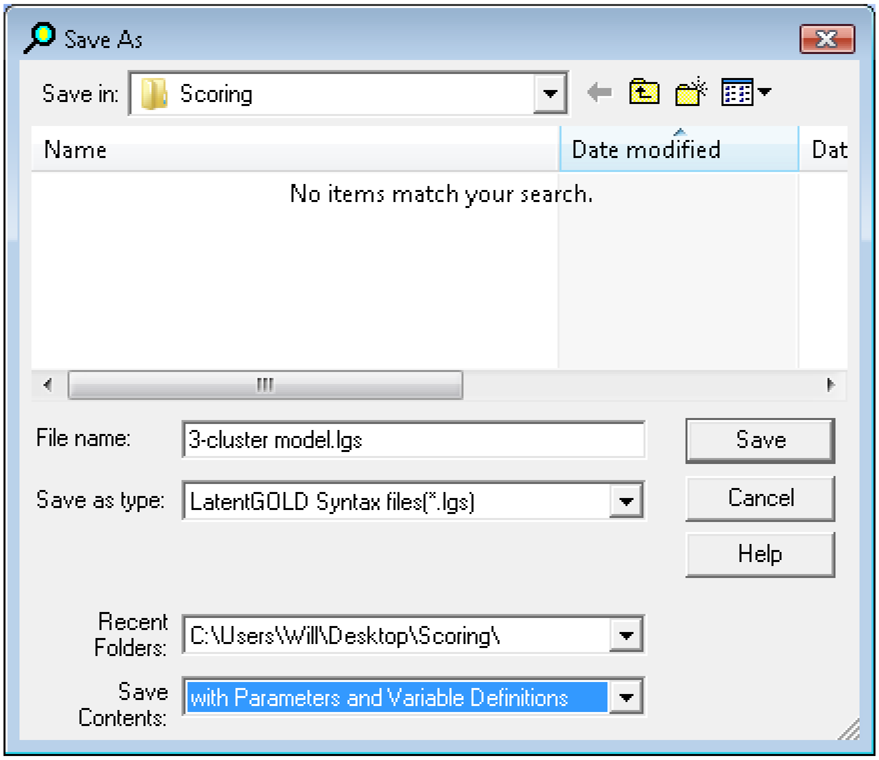
Syntax (1) Model

Number of cases	1202	
Number of parameters (Npar)	20	
Random Seed	295577	
Best Start Seed	996393	
Chi-squared Statistics		
Degrees of freedom (df)	15	p-value
L-squared (L^2)	22.0872	0.11
X-squared	23.5098	0.074
Cressie-Read	22.6612	0.091
BIC (based on L^2)	-84.2889	
AIC (based on L^2)	-7.9128	
AIC3 (based on L^2)	-22.9128	

Figure 1B: Summary Output from the 3-class model estimated with the syntax module

5. Save the parameter estimates and variable definitions in a .lgs file:
   - Highlight the model name (‘3-cluster model’).
   - From the File Menu select ‘Save Syntax’ and a ‘Save As’ dialog box opens.
   - From the ‘Save Contents’ Pane at the bottom of the ‘Save As’ box select ‘Syntax with Parameters and Variable Definitions’.
   - In the ‘File Name’ pane enter the desired file name for the resulting .lgs file. In our example, save the file as *‘3-cluster model.lgs’*.
   - Click on the ‘Save’ button.

In step 5 if the syntax model is saved ‘with Parameters’ instead of ‘with Parameters and Variable Definitions’, if any cases on the new data file to be scored contain values for categorical variables that were not present on the original analysis sample (gss82white.sav), you will get the error message ‘mismatch in the number of internal parameters’ in Step 7 when attempting to run the saved .lgs file. If ‘with Parameters *and Variable Definitions*’ is selected in step 5, there will be no error message and *only* the records containing values in the valid range will be scored (assuming that the default option missing = ’exclude all’ is specified). If ‘missing=includeall’ is specified, cases with invalid labels (values) will be treated as missing, and scored accordingly under the Missing At Random (MAR) assumption.



Save in: Scoring

No items match your search.

File name: 3-cluster model.lgs

Save as type: LatentGOLD Syntax files(*.lgs)

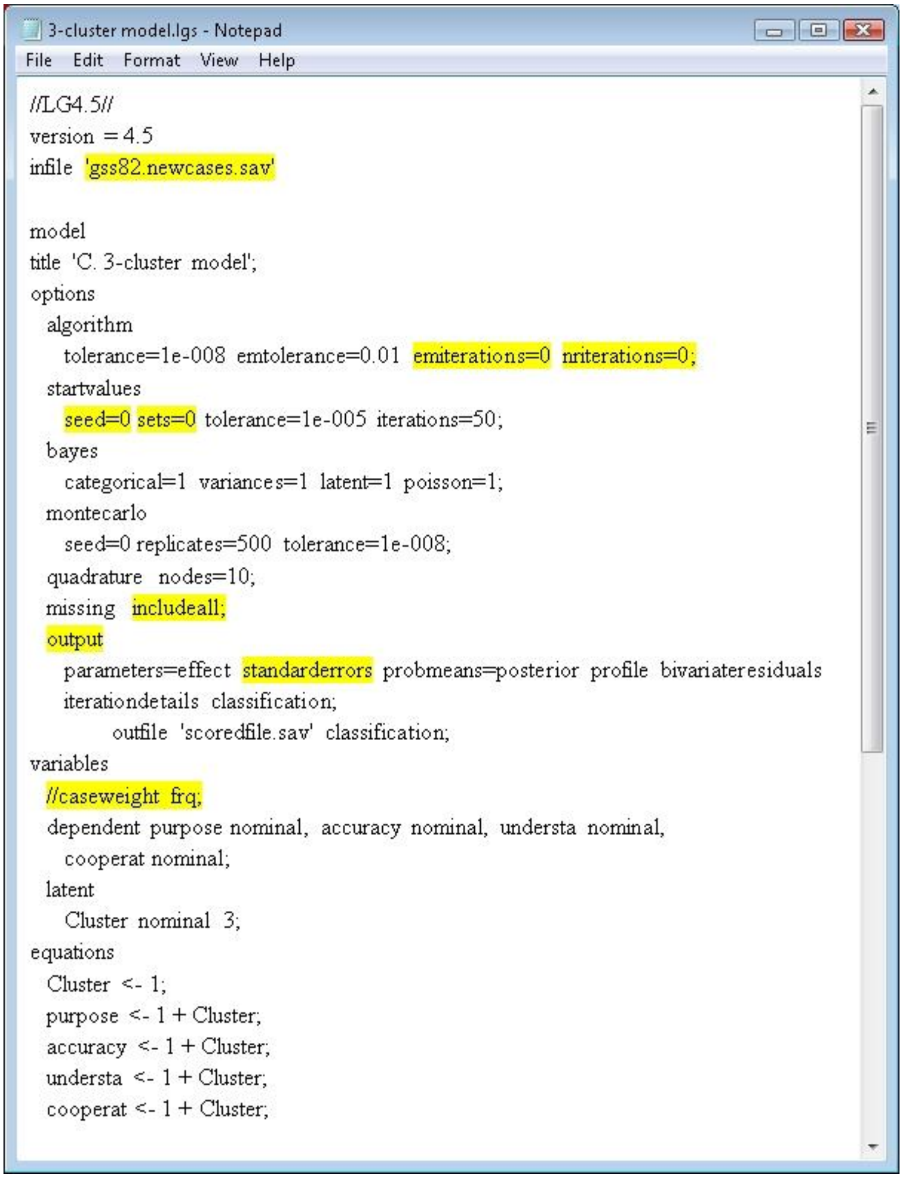
Recent Folders: C:\Users\Will\Desktop\Scoring\

Save Contents: with Parameters and Variable Definitions

Figure 2: Saving Syntax with ‘Parameters and Variable Definitions’

6. Use a **text editor** to edit the saved .lgs file, make the following changes:
    - Change the ‘Infile’ statement (typically this will be the 3rd line in the .lgs file) to reflect the file name of the data file that you will score. In our example, replace ‘gss82white.sav’ with ‘gss82.newcases.sav’ (on the infile statement).
    - In the Output subsection, include an ‘Outfile’ statement containing the name to be used for the scored file (‘scoredfile.sav’). This name needs to differ from the file name specified in the ‘Infile’ statement. For our example, since we want to score the file with the ‘posterior membership probabilities and classifications’, following the ‘;’ in the Output section include the statement “outfile ‘scoredfile.sav’ classification;”.
    - In the Algorithm subsection, change the values for ‘emiterations’ and ‘nriterations’ so that it reads ‘emiterations=0’ and ‘nriterations=0’ (since the parameter estimates are saved as starting values to be used the next time this model is estimated, this change will assure that the starting values for the parameters that these starting values are final values (no iterations are performed in the estimation algorithm).
    - In the Startvalues subsection, change the values for ‘seed’ and ‘sets’ so that it reads ‘seed=0’ and ‘sets=0’ (this will speed up the scoring run).

- In the Missing subsection, include the keyword ‘includeall’ if your data file to be scored contains values for categorical variables that differ from those on the analysis file. (Recall Warning Message associated with Step 5 above).
- If present in the Output subsection, remove the keyword ‘standarderrors’ (this will speed up the scoring run).
- In the Output subsection, include the keyword ‘classification’.
- If present in the Variables section, comment out (or remove) the ‘caseweight’ statement.

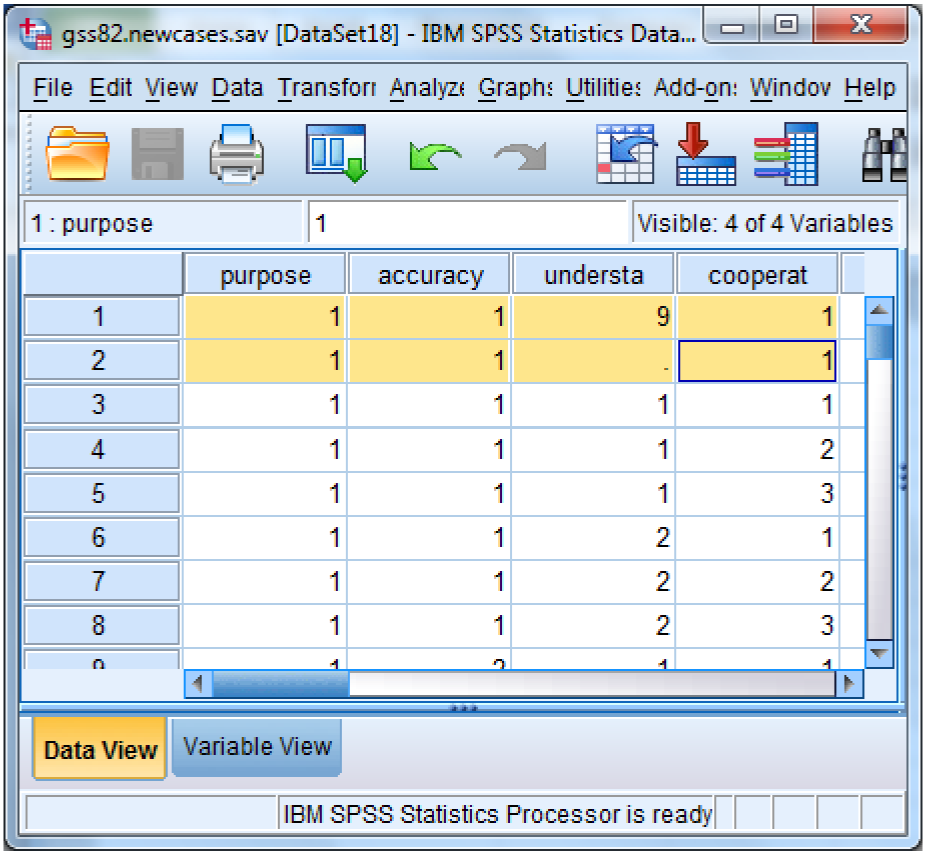


```
//LG4.5//
version = 4.5
infile 'gss82.newcases.sav'

model
title 'C. 3-cluster model';
options
  algorithm
    tolerance=1e-008 emtolerance=0.01 emiterations=0 nriterations=0;
  startvalues
    seed=0 sets=0 tolerance=1e-005 iterations=50;
  bayes
    categorical=1 variances=1 latent=1 poisson=1;
  montecarlo
    seed=0 replicates=500 tolerance=1e-008;
  quadrature  nodes=10;
  missing  includeall;
  output
    parameters=effect standarderrors probmeans=posterior profile bivariateresiduals
    iterationdetails classification;
          outfile 'scoredfile.sav' classification;
variables
  //caseweight frq;
  dependent purpose nominal, accuracy nominal, understa nominal,
    cooperat nominal;
  latent
    Cluster nominal 3;
equations
  Cluster <- 1;
  purpose <- 1 + Cluster;
  accuracy <- 1 + Cluster;
  understa <- 1 + Cluster;
  cooperat <- 1 + Cluster;
```

Figure 3: Syntax in text editor with changes highlighted
(Note: for simplicity, syntax shown *without* definitions and parameters)

The file to be scored (gss82.newcases.sav) consists of 38 new cases. The first 2 cases in the file illustrate the scoring that results from incomplete or invalid data. The 1st case contains an invalid value '9' for the variable 'understa' and the 2nd case contains a missing value for this variable. The remaining 36 cases illustrate all 3x2x2x3 = 36 valid response patterns, which include some which did not occur at all in the observed data file 'gss82white.sav' such as case 20, (purpose, accuracy, understa, cooperat) = (2, 1, 2, 3).


gss82.newcases.sav [DataSet18] - IBM SPSS Statistics Data...
File Edit View Data Transforı Analyzı Graphı Utilitieı Add-onı Windov Help
Data View
Variable View

1 : purpose | 1 | Visible: 4 of 4 Variables

	purpose	accuracy	understa	cooperat
1	1	1	9	1
2	1	1	.	1
3	1	1	1	1
4	1	1	1	2
5	1	1	1	3
6	1	1	2	1
7	1	1	2	2
8	1	1	2	3
9	1	2	1	1

IBM SPSS Statistics Processor is ready

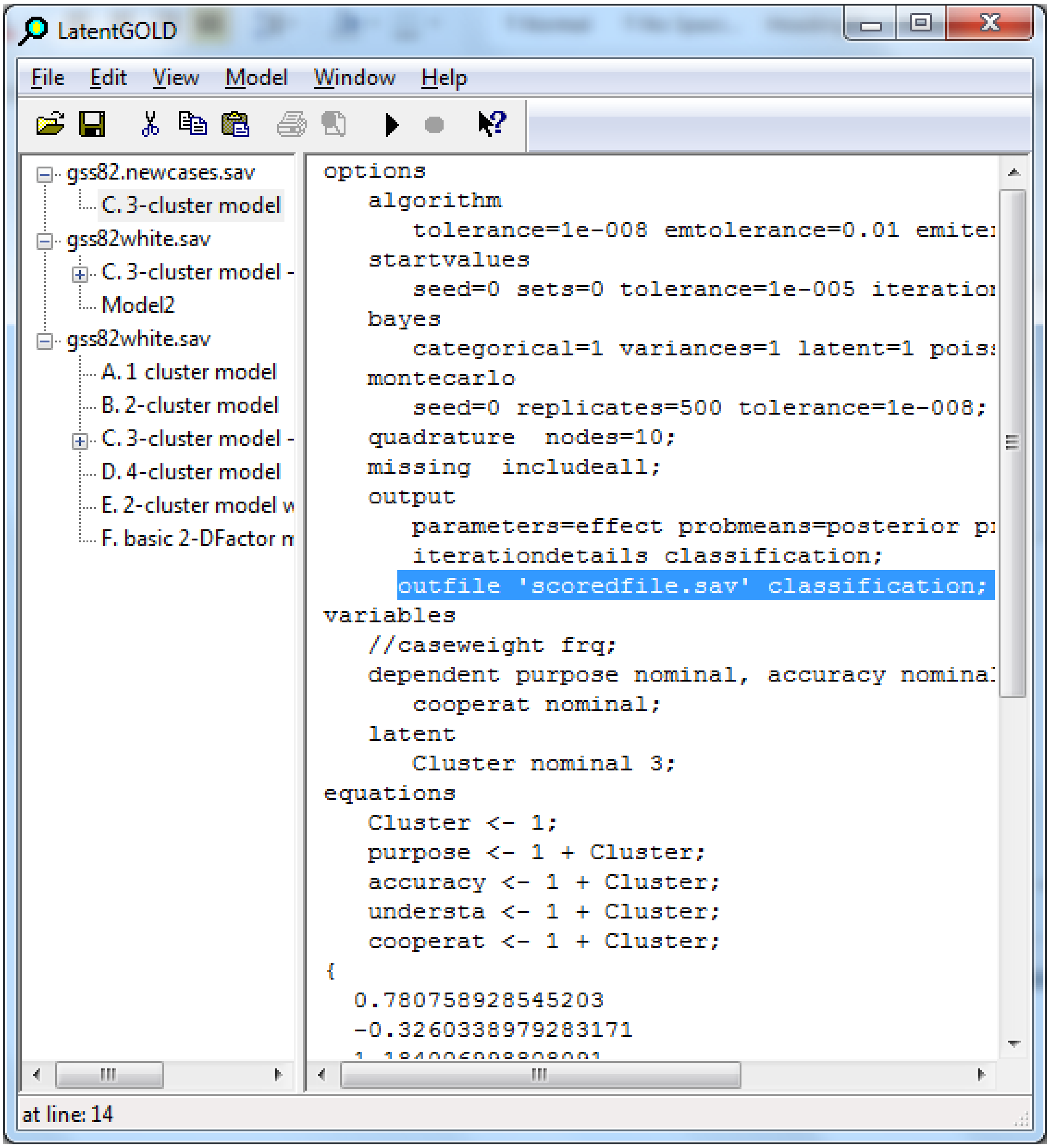
Figure 4: The first 8 records containing the 38 new cases to be scored

7. Open and Run the edited .lgs file
   - From the "File" menu click on "Open"
   - Select '3-cluster model.lgs'

Fig. 5 shows the syntax including the first few saved parameters at the bottom of the file. These parameters are consistent with the original class ordering where the largest class is class 1, and the smallest is class 3.

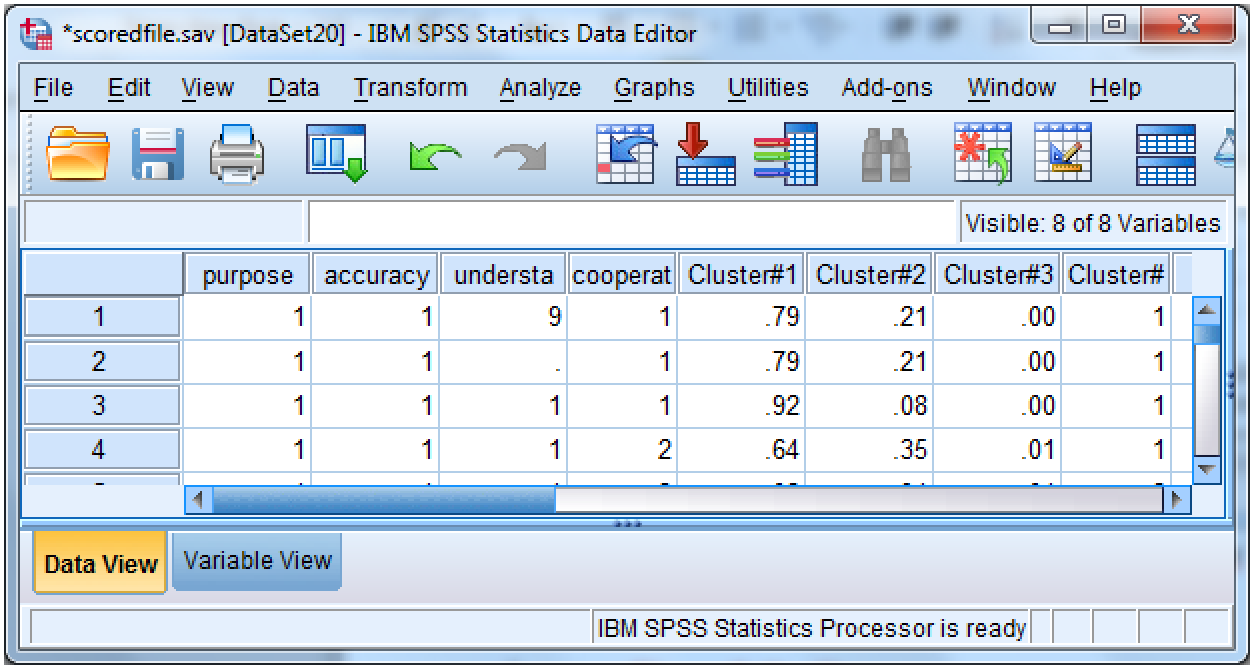
- Click ▶ to run

Fig. 6 shows the resulting scored datafile ‘scoredfile.sav’. Notice that the posteriors for the first 2 cases are identical, both responding (purpose, accuracy, cooperat) = (1, 1, 1), and the value for ‘understa’ being treated as missing.



```
options
   algorithm
      tolerance=1e-008 emtolerance=0.01 emite
   startvalues
      seed=0 sets=0 tolerance=1e-005 iteratio
   bayes
      categorical=1 variances=1 latent=1 pois
   montecarlo
      seed=0 replicates=500 tolerance=1e-008;
   quadrature  nodes=10;
   missing  includeall;
   output
      parameters=effect probmeans=posterior p
      iterationdetails classification;
     outfile 'scoredfile.sav' classification;
variables
   //caseweight frq;
   dependent purpose nominal, accuracy nomina
      cooperat nominal;
   latent
      Cluster nominal 3;
equations
   Cluster <- 1;
   purpose <- 1 + Cluster;
   accuracy <- 1 + Cluster;
   understa <- 1 + Cluster;
   cooperat <- 1 + Cluster;
{
  0.780758928545203
  -0.3260338979283171
```

Figure 5: Edited Syntax containing an ‘Outfile’ statement to create the scored file



*scoredfile.sav [DataSet20] - IBM SPSS Statistics Data Editor

Visible: 8 of 8 Variables

	purpose	accuracy	understa	cooperat	Cluster#1	Cluster#2	Cluster#3	Cluster#
1	1	1	9	1	.79	.21	.00	1
2	1	1	.	1	.79	.21	.00	1
3	1	1	1	1	.92	.08	.00	1
4	1	1	1	2	.64	.35	.01	1

IBM SPSS Statistics Processor is ready

Figure 6. Scored file showing the posterior probabilities and class assignment

If one or more model variables are not included in the file to be scored, when attempting to open the .lgs file in Step 7 will result in an error message indicating the variables not found on the file. To score the file you must add these variable names to the file even if all records have missing values on these variables.

Note: If the grouping option is used to reduce the number of values for one or more numeric variables in your saved model, to use this model to score a new file correctly and avoid the 'mismatch in the number of internal parameters' warning message, you should implement the precise grouping on the data file to be scored. (To assist you in this process, the summary output from Latent GOLD shows the cutpoints used to define the grouping).

A much simpler method for scoring new cases is to use 'Step3 Scoring', an advanced option in Latent GOLD 5.0. For details, see the Step 3 Scoring tutorial.