

7.1. Tutorial #1: Using Latent GOLD 5.1 to Estimate LC Cluster Models

DemoData = 'gss82white.sav'

In this tutorial, we use 4 categorical indicators to show how to estimate LC Cluster models and interpret the resulting output. For related analyses of these data, see McCutcheon (1987),

Magidson and Vermunt (2001) <http://www.statisticalinnovations.com/articles/SOME.pdf>, and

Magidson and Vermunt (2004) <http://www.statisticalinnovations.com/articles/sage11.pdf>.

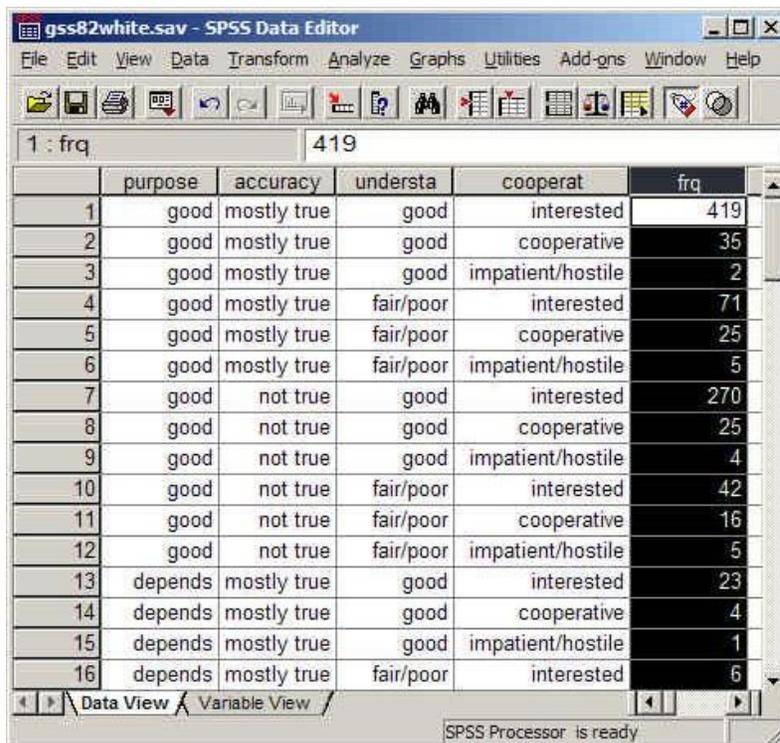
In this tutorial, you will:

- Open a data file
- Setup and estimate traditional latent class (cluster) models
- Explore which models best fit the data
- Generate and interpret output and interactive graphs
- Save results

The Data

Latent GOLD 5.1 accepts data from a variety of formats including SPSS system files, and ASCII rectangular files. The following data illustrates the use of an SPSS .sav file containing N=1,202 cases and an optional case weight variable FRQ.

Figure 7-1. SPSS data file gss82white.sav (first 16 records shown) *



	purpose	accuracy	understa	cooperat	frq
1	good	mostly true	good	interested	419
2	good	mostly true	good	cooperative	35
3	good	mostly true	good	impatient/hostile	2
4	good	mostly true	fair/poor	interested	71
5	good	mostly true	fair/poor	cooperative	25
6	good	mostly true	fair/poor	impatient/hostile	5
7	good	not true	good	interested	270
8	good	not true	good	cooperative	25
9	good	not true	good	impatient/hostile	4
10	good	not true	fair/poor	interested	42
11	good	not true	fair/poor	cooperative	16
12	good	not true	fair/poor	impatient/hostile	5
13	depends	mostly true	good	interested	23
14	depends	mostly true	good	cooperative	4
15	depends	mostly true	good	impatient/hostile	1
16	depends	mostly true	fair/poor	interested	6

* Source: 1982 General Social Survey Data National Opinion Research Center

The Goal

Identify distinctly different survey respondent types using two variables that ascertain the respondent's opinion regarding the purpose of surveys (PURPOSE) and how accurate they are (ACCURACY), and two additional variables that are evaluations made by the interviewer of the respondent's levels of understanding of the survey questions (UNDERSTAND) and cooperation shown in answering the questions (COOPERATE). More specifically, we will focus on different criteria for choosing the number of classes (clusters), and classify respondents into clusters.

Opening the Data File

For this example, the data file is in SPSS system file format.

- To open the file, from the menus choose:

File

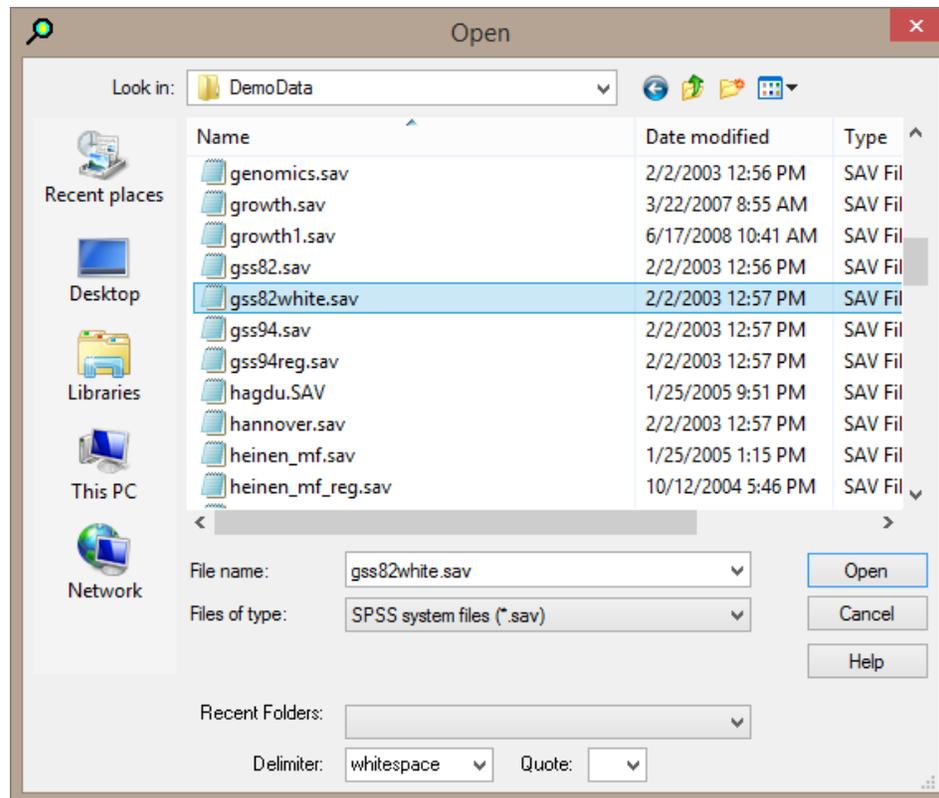
Open

- From the Files of type drop down list, select SPSS System Files if this is not already the default listing.

All files with the .sav extensions appear in the list (see Figure 7-2)

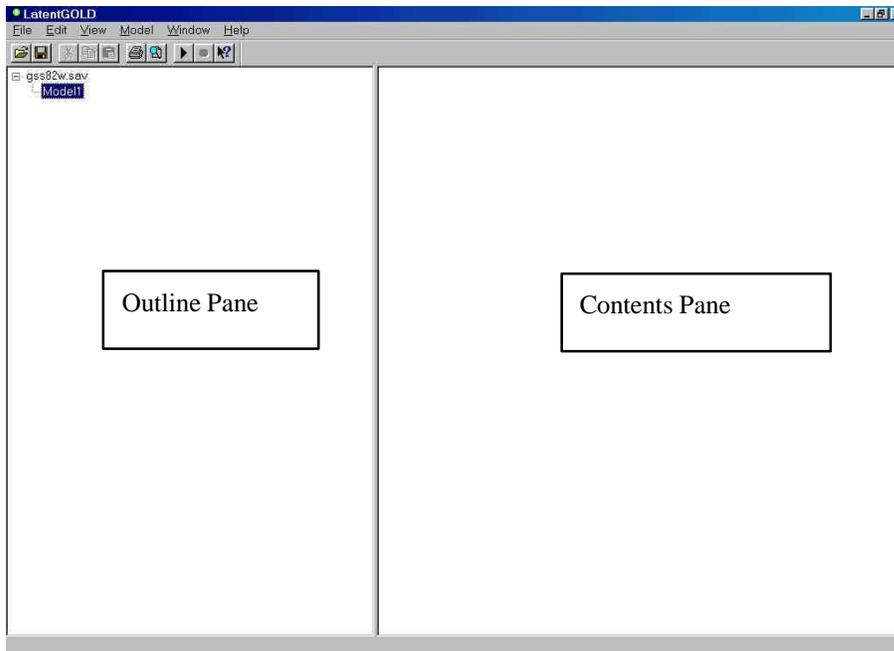
Note: If you copied the sample data file to a directory other than the default directory, change to that directory to retrieve the file.

Figure 7-2. File Open Dialog Box



- Select gss82white.sav and click Open to open the Viewer window, shown in Figure 7-3

Figure 7-3. Viewer Window



The Outline pane contains the name of the data file along with a list of any previously estimated models and their output. The Contents pane (currently empty) is where you will view the output from estimated models.

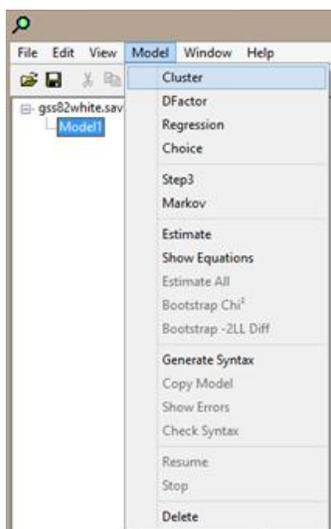
Estimating LC Cluster Models

Selecting the Type of Model

- Right click or double click on 'Model1' to open the Model Selection menu (see Figure 7-4).

Alternatively, you may also select the type of model from the Model Menu.

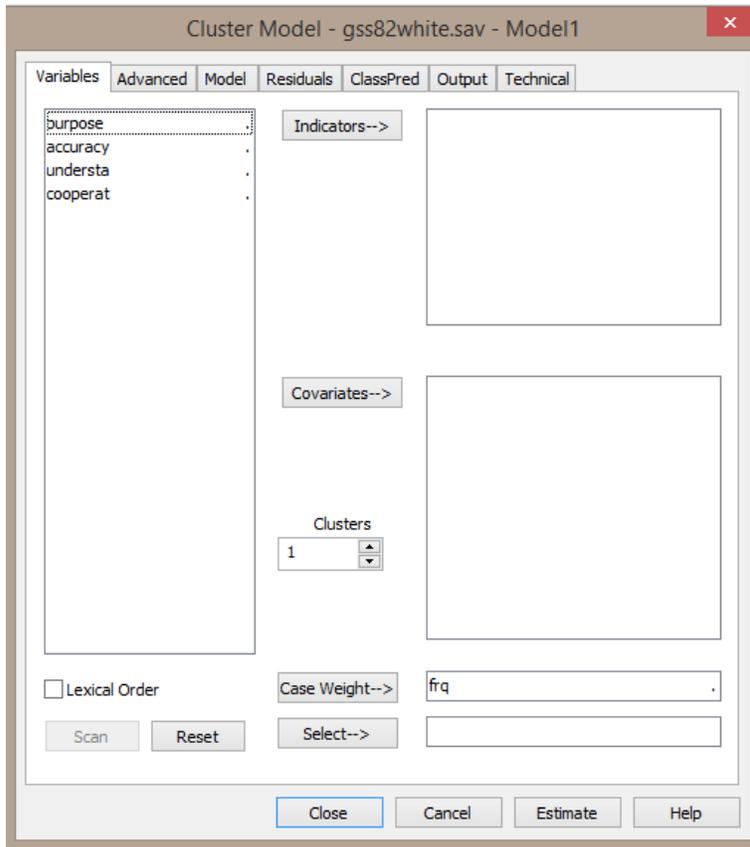
Figure 7-4. Model Selection Menu



➤ Select Cluster

The LC Cluster Analysis dialog box, which contains 7 tabs, opens (see Figure 7-5).

Figure 7-5. Analysis dialog box for LC Cluster Model



Selecting the Variables for the Analysis

For this analysis, we will be using all 4 variables (PURPOSE, ACCURACY, UNDERSTA, and COOPERAT) as indicators and the optional case weight variable FRQ. Since the .sav file already specified that the FRQ variable is used to weight the cases, FRQ is *automatically* placed in the Case Weight box.

To specify the other indicators:

- Use your mouse to select (highlight) the four variables in the Variables list box and click the *Indicators* button to move them to the Indicator list box.

The designated indicator variables now appear in the Indicators list box. To scan

the data file

- Click *Scan*.

You may now double click on any variable to view its categories, and the associated label, frequency count, and code for each category. The category scores may optionally be used in the model to fix the spacing of the categories by using the default Ordinal scale type (shown in Fig. 7-6 as ‘Ord-Fixed’).

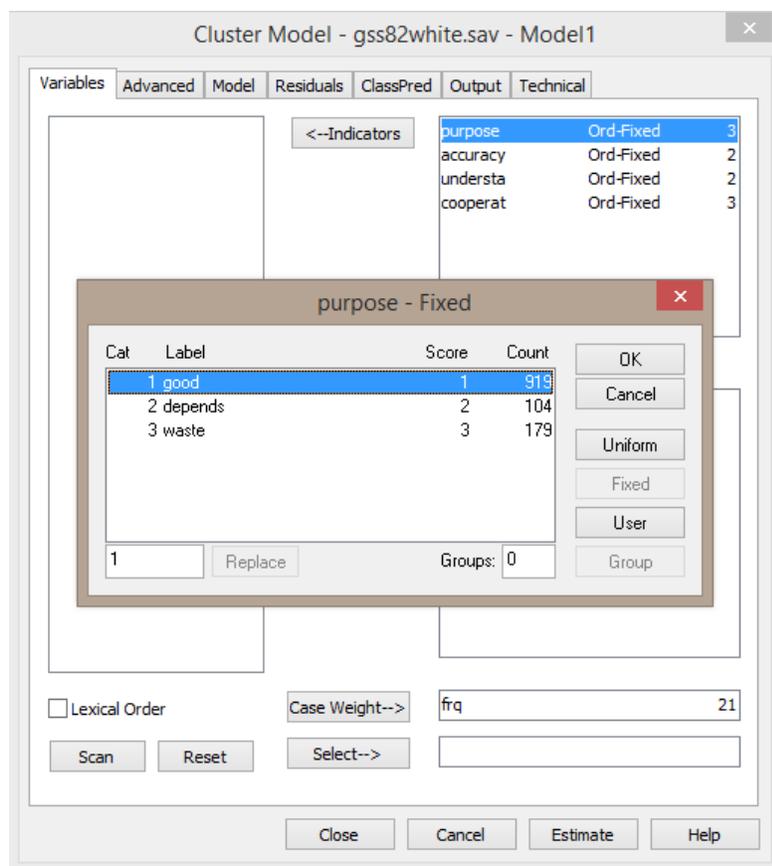
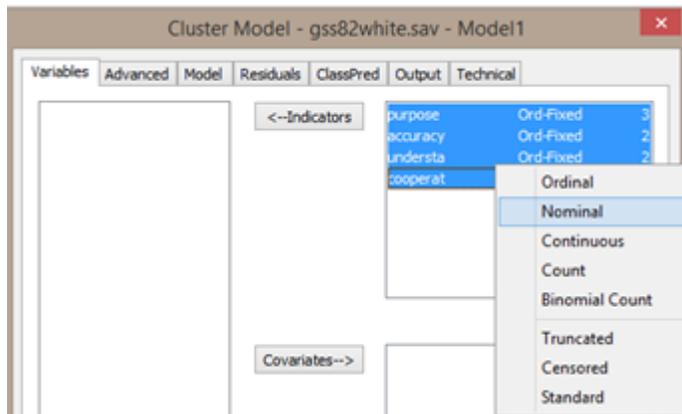


Figure 7-6. Category Information for the Variable PURPOSE

- Highlight all 4 indicator variables, right-click and select Nominal from the pop-up menu to change the scale type from ‘Ord-Fixed’ to ‘Nominal’ which causes any category scores to be ignored for the purpose of modeling.

Figure 7-7. Pop-up Menu to Set Variable Scale Types and Subtypes



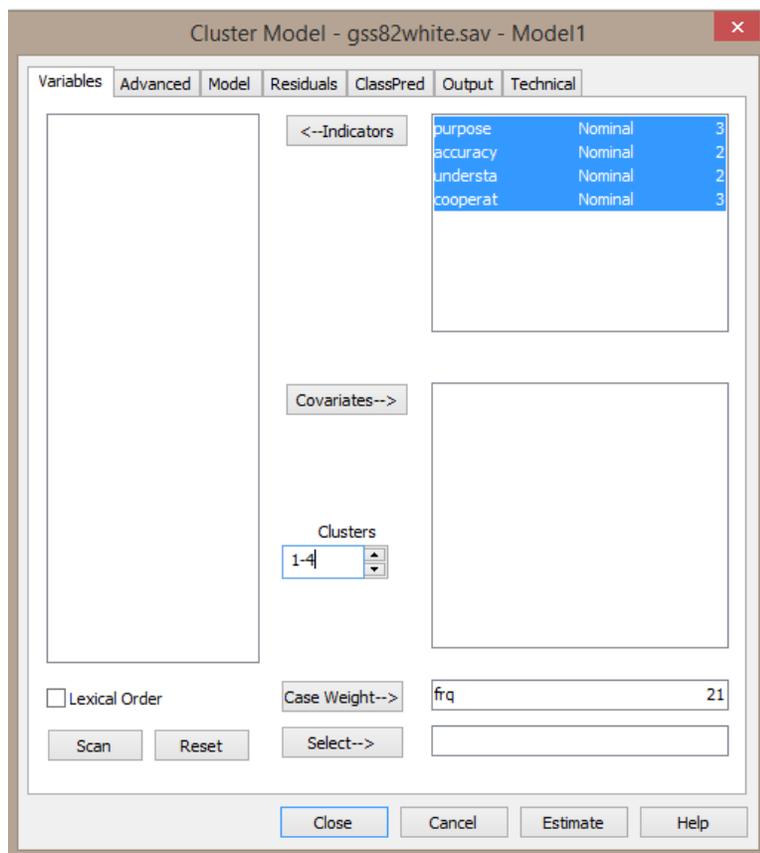
Specifying the Number of Clusters

To determine the number of clusters we will estimate 4 different cluster models, each specifying a different number of clusters. As a general rule of thumb, a good place to start is to estimate all models between 1 and 4 clusters.

- In the Variables Tab, in the box titled Clusters (below the Indicators pushbutton) type '1-4' to request the estimation of 4 models – a 1-cluster model, a 2-cluster model, a 3-cluster model and a 4-cluster model.

Your Analysis Dialog Box should now look like this:

Figure 7-8. Analysis Dialog Box Prior to Model Estimation.



Estimating the Model

Now that we have selected our variables and specified the models, we are ready to estimate these models.

- Click Estimate (located at the bottom right of the analysis dialog box).

When Latent GOLD completes the estimation, the model L^2 , which assesses how well the model fits the data appears in the Outline pane to the right of the name assigned to each model estimated. Several kinds of output are available, organized in a hierarchical fashion. To view an output listing, you may click on the name of the data file, a model associated with the data file, or a specific output listing associated with a model, and the selected output appears in the Contents pane.

Following the estimation, the expand/contract [+/-] icon is expanded for the last model estimated ('Model4'), so that names for the output listings for that model becomes visible.

Viewing Output and Interpreting Results

- Highlight the data file name gss82white.sav and a summary of all the models estimated on that data appears in the Contents pane.
- Right click in the Contents Pane to retrieve the Model Summary Display

Figure 7-9: Model Summary Output and Model Summary Display

The screenshot shows the LatentGOLD software interface. The main window displays a list of models estimated for the file 'gss82white.sav'. The models are:

- Model1 - $L^2 = 257.2604$
- Model2 - $L^2 = 79.5079$
- Model3 - $L^2 = 22.0872$
- Model4 - $L^2 = 6.6142$

The 'Model Summary Display' dialog box is open, showing a table of model statistics:

File name:	C:\Users\Margot\Documents\LatentGOLD5.1\DemoData\gss82white.sav								
File size:	1024 bytes	33 records							
File date:	2003-Feb-02	1:57:00 PM							
		LL	BIC(LL)	Npar	L^2	df	p-value	Class.Err.	
Model1	1-Cluster	-2872.2296	5787.0096	6	257.2604	29	2.0e-38	0.0000	
Model2	2-Cluster	-2783.3533	5658.8993	13	79.5079	22	2.0e-8	0.0846	
Model3	3-Cluster	-2754.6430	5651.1209	20	22.0872	15	0.11	0.1314	
Model4	4-Cluster	-2746.9065	5685.2900	27	6.6142	8	0.58	0.1959	
Model5	0-Cluster								

The 'Model Summary Display' dialog box has the following checked options:

- LL
- BIC(LL)
- AIC(LL)
- AIC3(LL)
- CAIC(LL)
- SABIC(LL)
- BIC(LL,N)
- CAIC(LL,N)
- SABIC(LL,N)
- L^2
- BIC(L^2)
- AIC(L^2)
- AIC3(L^2)
- CAIC(L^2)
- SABIC(L^2)
- Npar
- Class.Err.
- df
- R^2
- p-value
- $R^2(0)$

Additional statistics can be displayed by clicking on the associated check-box in the Model Summary Display.

The model L^2 statistic, as shown in Figure 7-9 in the column labeled ' L^2 ', indicates the amount of the association among the variables that remains unexplained after estimating the model; the lower the value, the better the fit of the model to the data.

One criteria for determining the number of clusters is to look in the 'p-value' column which provides the p-value for each model under the assumption that the L^2 statistic follows a chi-square distribution. Generally, among models for which the p-value is greater than 0.05 (provides an adequate fit), the one that is most parsimonious (fewest number of parameters -- Npar) would be selected. Using this criteria, the best model is given by Model 3, the 3-cluster model (p-value of 0.11, and Npar = 20).

Assessing Model Fit Using the Bootstrap p-value

Latent GOLD offers an alternative option to assess your model using the *bootstrap* of L^2 to estimate the pvalue. This provides a more precise estimate by relaxing the assumption that the L^2 statistic follows a chisquare distribution.

- In the Outline Pane, click once on Model 3 to select it and click again to enter Edit mode and rename it '3-class' for easier identification.

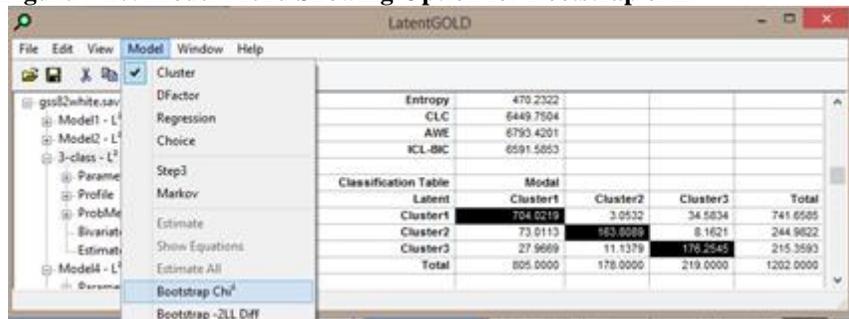
To estimate the bootstrap p-value for the '3-class' model:

- right-click on this model and select 'Bootstrap Chi²'

or alternatively,

- select '3-class'
- select 'Bootstrap Chi²' from the Model Menu

Figure 7-10: Model Menu Showing Option for Bootstrap of L^2



Latent GOLD then performs 500 iterations to estimate the p-value. When completed, the model name '3classBoot' appears in the Outline Pane and the Bootstrap p-value along with its standard error appears in the Contents Pane.

Figure 7-11: Results for the Bootstrap of L^2

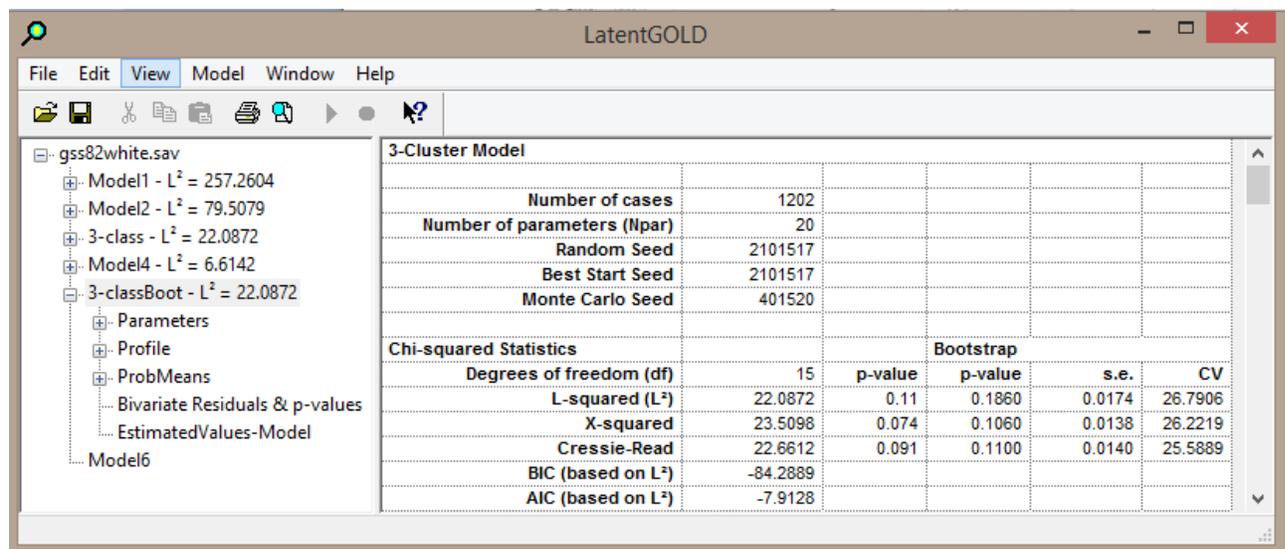


Figure 7-11 shows the estimate of the p-value resulting from the bootstrap procedure. It is $p = .1860$ with a standard error of about 0.02. Hence, the earlier estimate of the p-value based on the chi-squared approach ($p = .11$) appears to be somewhat understated.

Note. Since the bootstrap p-value is estimated from the generated random sample of size 500, the results you may get for the estimated p-value will be somewhat different because your random sample will be different (unless you utilize the specific Bootstrap Seed reported above as a Technical option prior to requesting the bootstrap).

Viewing the Parameters Output

Next, for the 3-class model we will assess the significance associated with each indicator.

- Click on the expand icon (+) next to the '3-class' model to show the available output listings
- Click Parameters

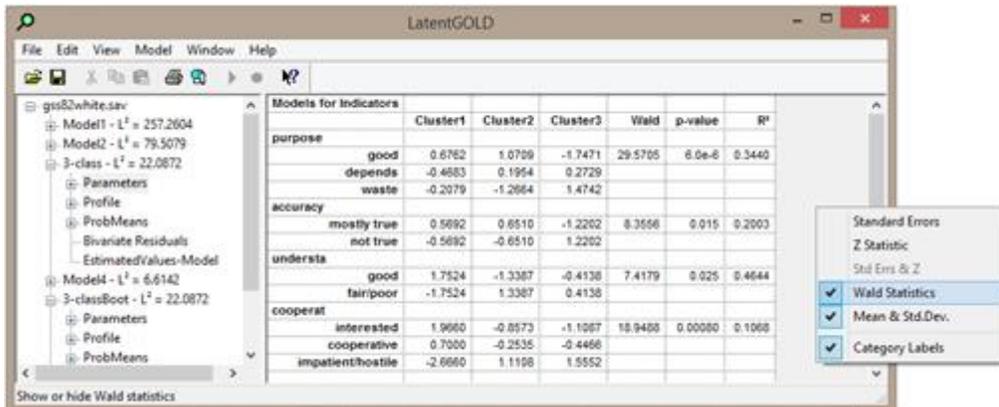
or alternatively, these *same log-linear parameter estimates* may be viewed using the re-estimated model given the name '3-classBoot',

- click Parameters under the name '3-classBoot'

In the Content Pane, a summary of Parameter estimates and related statistics appears.

These log-linear parameters utilize effect coding, the default option (the parameters can alternatively be based on dummy coding). Effect coding means that for each indicator the estimates sum to zero over the categories of that indicator (columns). Since effect coding is also used for the clusters, the effect estimates also sum to zero across the clusters (rows). To utilize dummy coding instead of effect coding, the Nominal Coding option would be changed in the Output Tab prior to estimating the models.

Figure 7-12. Parameters Output and View Menu Customization Options



For each indicator, the p-value is shown to be less than .05, indicating that the null hypothesis stating that all of the effects associated with that indicator are zero would be rejected. Thus, for each indicator, knowledge of the response for that indicator contributes in a significant way towards the ability to discriminate between the clusters.

The R^2 values are in the right-most column of the table indicating how much of the variance of each indicator is explained by this 3-cluster model. For example, we see that 34.4% of the variance of the PURPOSE variable is explained.

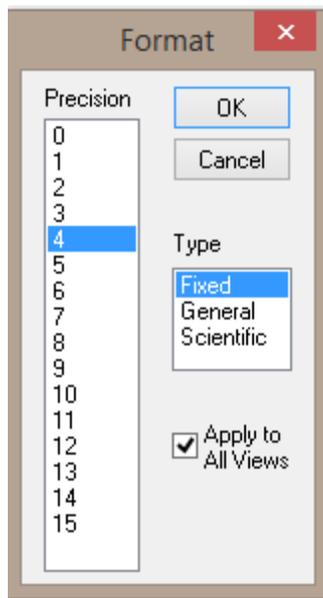
Standard errors and Z-statistics can be added to the output using the model display menu.

- Right click in the Contents Pane to display this menu (shown above).

The number of decimal places can be changed in any of the output file listings using the Format Control. To display the format control for the current output listing

- Click Edit from within the Contents Pane
- Select Format

Figure 7-13. Format Control Options from View Menu



Profile Output and Associated Profile Plot

To view the parameters re-expressed as conditional probabilities

- Click on Profile

Figure 7-14. Profile Output for 3-cluster Model

	Cluster1	Cluster2	Cluster3
Cluster Size	0.6168	0.2039	0.1793
Indicators			
purpose			
good	0.8905	0.9157	0.1593
depends	0.0524	0.0706	0.2220
waste	0.0570	0.0137	0.6188
accuracy			
mostly true	0.6148	0.6528	0.0426
not true	0.3852	0.3472	0.9574
understa			
good	0.9957	0.3243	0.7532
fair/poor	0.0043	0.6757	0.2468
cooperat			
interested	0.9452	0.6880	0.6432
cooperative	0.0547	0.2582	0.2559
impatient/hostile	0.0001	0.0538	0.1009

Overall, cluster 1 contains 62% of the cases, cluster 2 contains 20% and the remaining 18% are in cluster 3. The conditional probabilities show the differences in response patterns that distinguish the clusters. For example, cluster 3 is much more likely to respond that surveys are a waste of time (PURPOSE = 'waste') and that survey results are not true (ACCUARACY = 'not true') than the other 2 clusters.

To view these probabilities graphically

- Click expand icon (+) next to Profile
- Click Prf-Plot.

The Profile Plot for the 3-cluster model now appears

The profile for any particular cluster may be highlighted by clicking on the symbol next to any of the 3 Clusters (Cluster1, Cluster2, or Cluster 3) at the bottom of the plot. For example, to highlight the profile for cluster3

- Click the symbol next to 'Cluster 3'

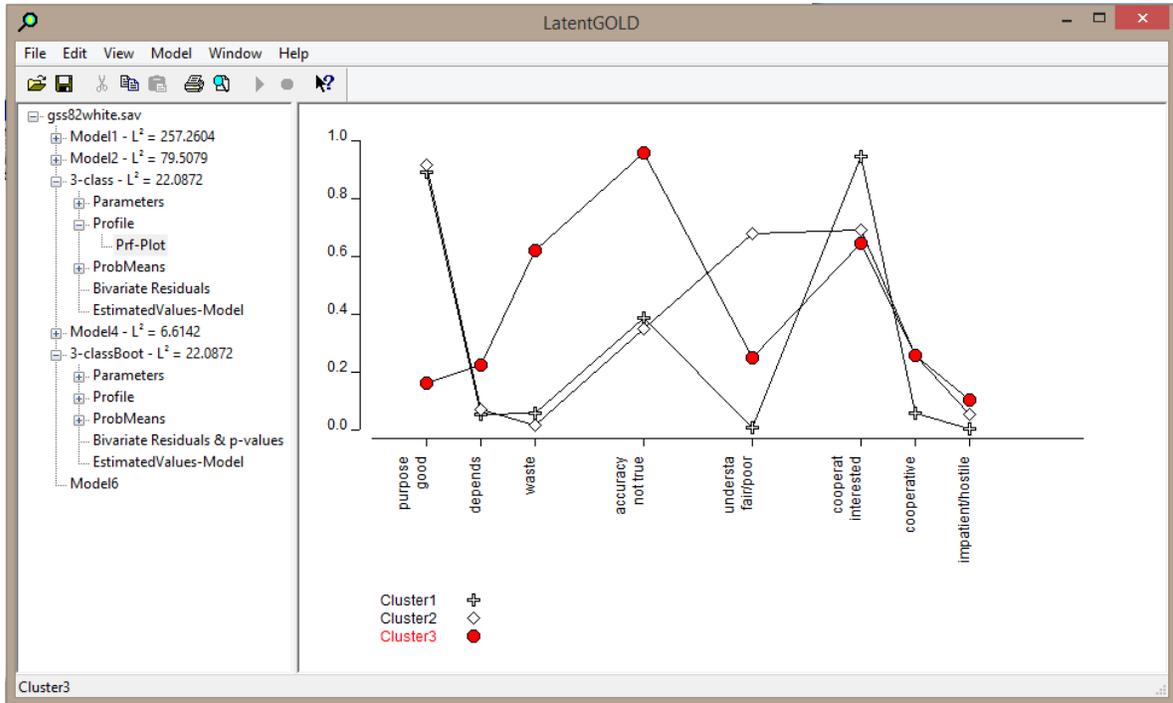


Figure 7-15. Profile Plot for 3-cluster Model

The labels appear vertically, allowing display of all categories of each nominal variable (such as ‘good’, ‘depends’, or ‘waste’). By default, the last category for dichotomous variables and all categories for other nominal variables are displayed.

To customize the variables and categories to appear in the display, the plot control panel may be used. To retrieve the plot control panel

- Right click on the plot

Alternatively, the plot control panel may be selected from the View Menu.



Figure 7-16. Profile Plot Control Panel

ProbMeans Output and Associated Tri-Plot

The ProbMeans output re-expresses the parameters in terms of row percentages rather than column percentages. This has the advantage of yielding a barycentric coordinate display of the categories of all indicators, where the vertices of the triangle represent the 3 clusters.

To view the tri-plot

- Click the expand icon (+) next to ProbMeans
- Click on Tri-plot

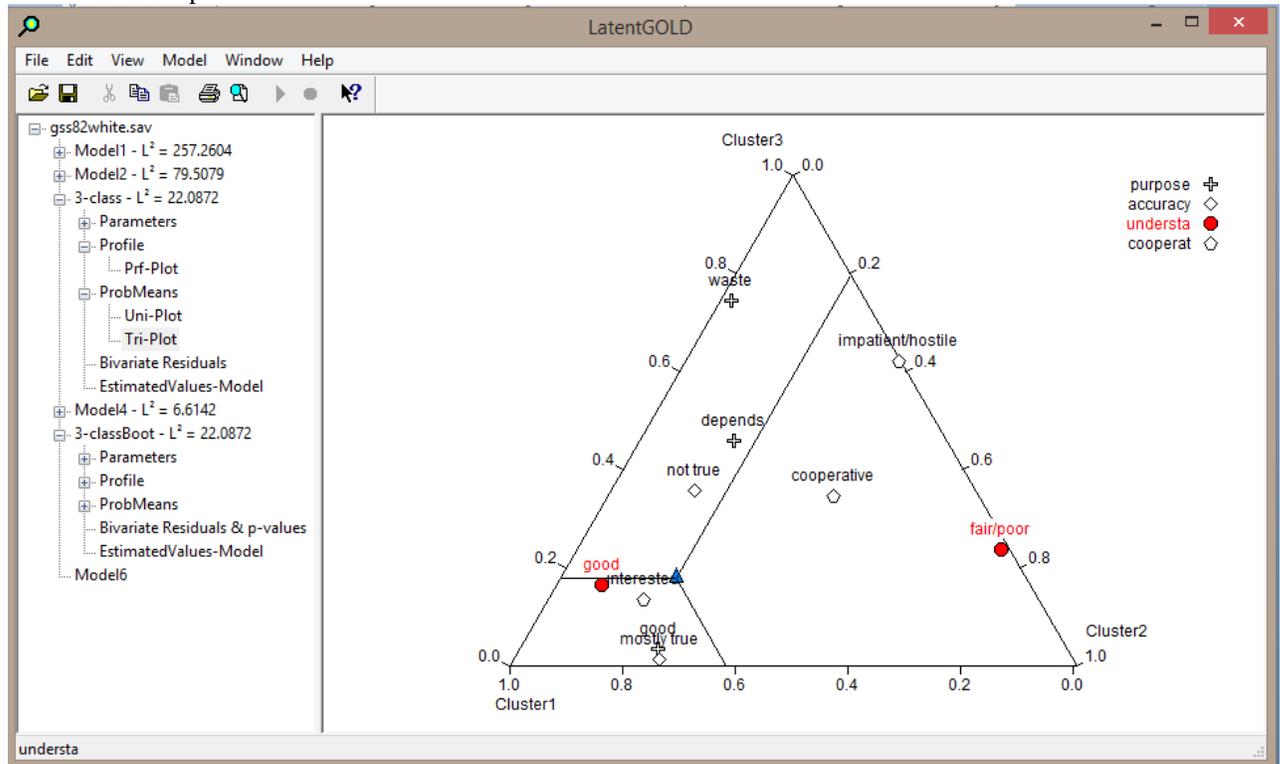


Figure 7-17. Tri-plot Display for 3-class Model

Classifying Cases into Clusters using Modal Assignment

Additional output such as classification output can be obtained from the Output Tab.

- Double-click on '3-class' in the Outline Pane to re-open the Analysis Dialog Box
- Click the Output Tab
- In the Output Tab, check the box for 'Classification – Posterior':

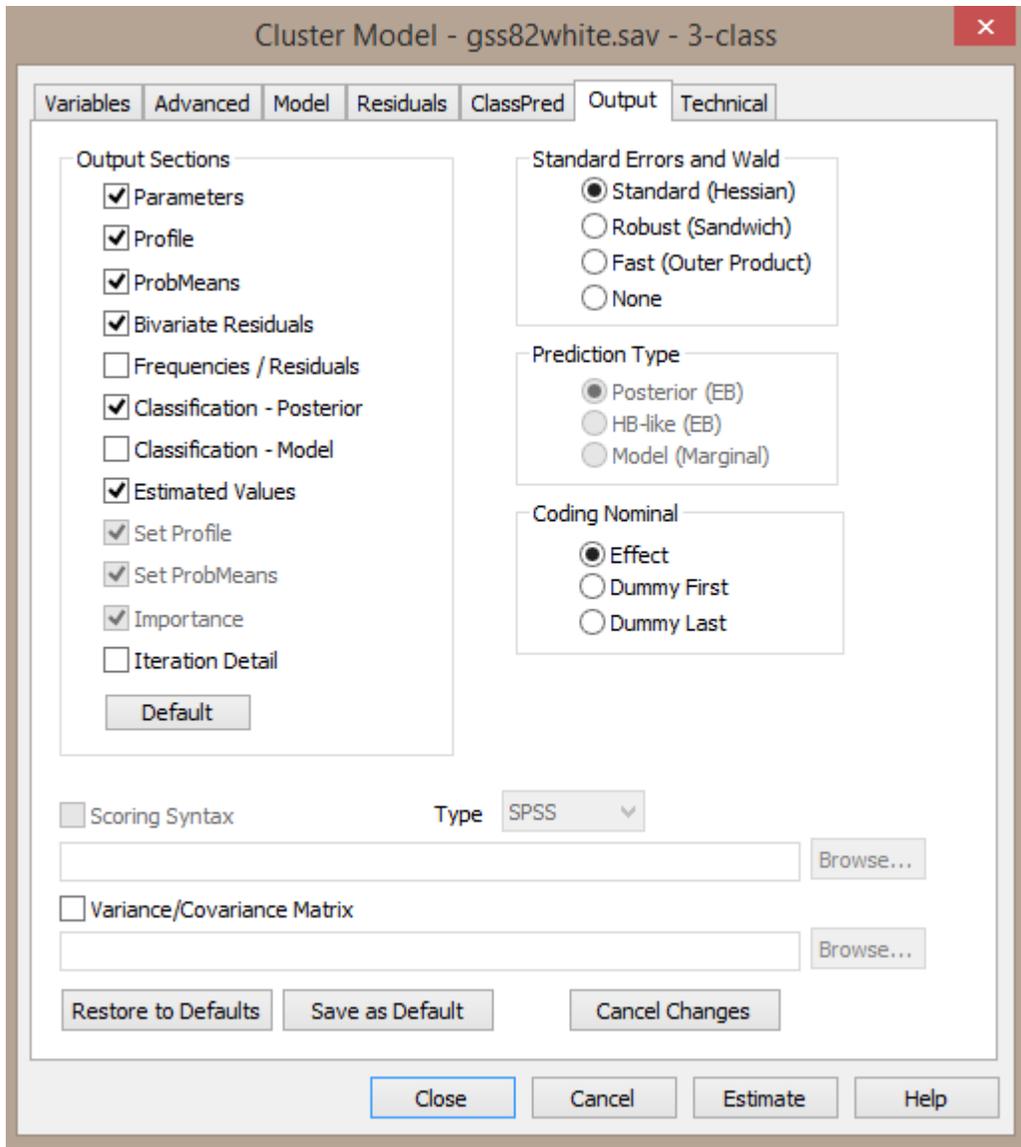


Figure 7-18. Requesting Classification - Posterior Output Listing in Output Tab

- Click Estimate
- Under the new Model, click on ‘Standard Classification’ to view the Classification Output:

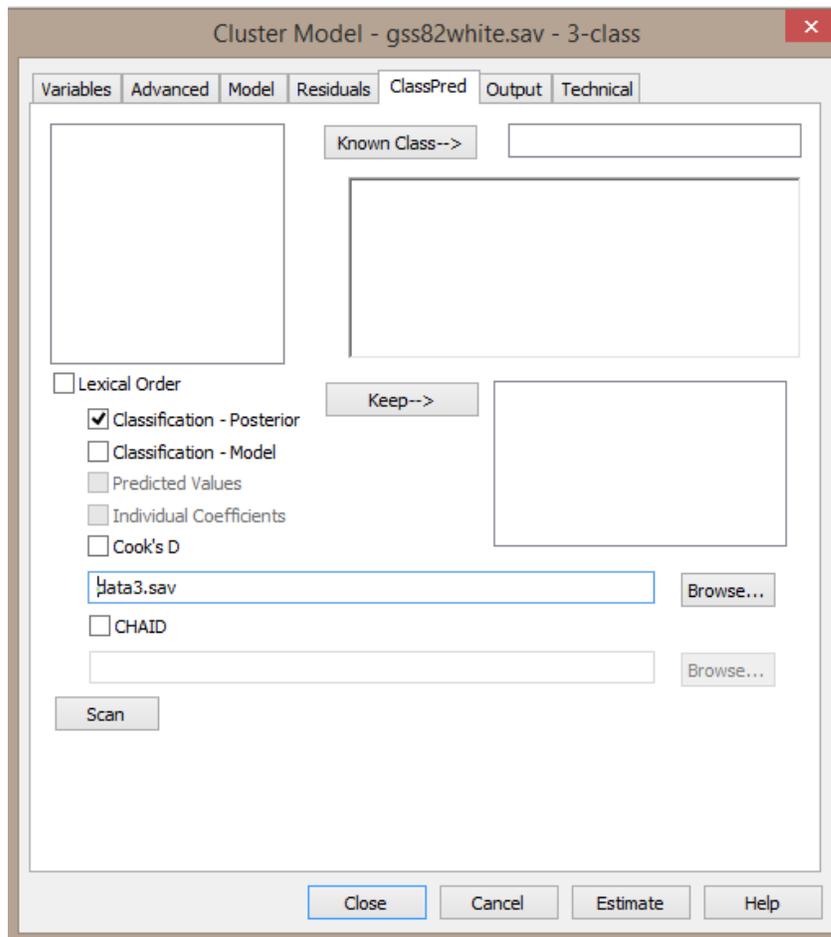
Figure 7-19. Standard Classification Output Listing for 3-cluster Model

purpose	accuracy	understa	cooperat	ObsFreq	Modal	Cluster1	Cluster2	Cluster3
good	mostly true	good	interested	419.0000	1	0.9196	0.0787	0.0017
good	mostly true	good	cooperative	35.0000	1	0.6378	0.3540	0.0081
good	mostly true	good	impatient/hostile	2.0000	2	0.0150	0.9439	0.0411
good	mostly true	fair/poor	interested	71.0000	2	0.0235	0.9732	0.0033
good	mostly true	fair/poor	cooperative	25.0000	2	0.0037	0.9927	0.0036
good	mostly true	fair/poor	impatient/hostile	5.0000	2	0.0000	0.9932	0.0068
good	not true	good	interested	270.0000	1	0.8778	0.0638	0.0584
good	not true	good	cooperative	25.0000	1	0.5185	0.2444	0.2371
good	not true	good	impatient/hostile	4.0000	3	0.0066	0.3503	0.6431
good	not true	fair/poor	interested	42.0000	2	0.0243	0.8529	0.1228
good	not true	fair/poor	cooperative	16.0000	2	0.0038	0.8643	0.1319
good	not true	fair/poor	impatient/hostile	5.0000	2	0.0000	0.7759	0.2240
depends	mostly true	good	interested	23.0000	1	0.8651	0.0969	0.0380

The first row of the Classification Output shows that the 419 respondents have the response pattern (PURPOSE = good, ACCURACY =mostly true, UNDERSTA = good, and COOPERAT = good) are classified into Cluster 1 because the probability of being in this class is highest (.9196). Under the column labeled 'modal', they have the value 1 to indicate this classification.

The classification information can be appended to your data file by selecting Classification - Posterior on the ClassPred Tab:

Figure 7-20. Requesting Output of Classification- Posterior Information to a Data File



Notice that when cases are classified into clusters using the modal assignment rule, a certain amount of misclassification error is present. The expected misclassification error can be computed by cross-classifying the modal classes by the actual probabilistic classes. This is done in the Classification Table, shown in the Contents Pane in **Figure 7-10** for the 3-class model. For this model, the modal assignment rule would be expected to classify correctly 704.1204 cases from the true cluster 1, 163.7828 from cluster 2 and 176.2427 from cluster 3 for an expected total of 1,044.146 correct classifications of the 1,202 cases. This represents an expected misclassification rate of 13.13% ($1 - 1,044.146/1,202$).

Notice also that the expected sizes of the clusters are not reproduced by modal assignment classification. The Classification Table in Figure 7-10 shows that 67.0% of the total cases (805 of the 1,202) are assigned to cluster 1 compared to 61.7% expected to be in this cluster.

Bivariate Residuals

In addition to various *global* measures of model fit, *local* measures called bivariate residuals (BVR) are also available to assess the extent to which the 2-way association(s) between any pair of indicators are explained by the model.

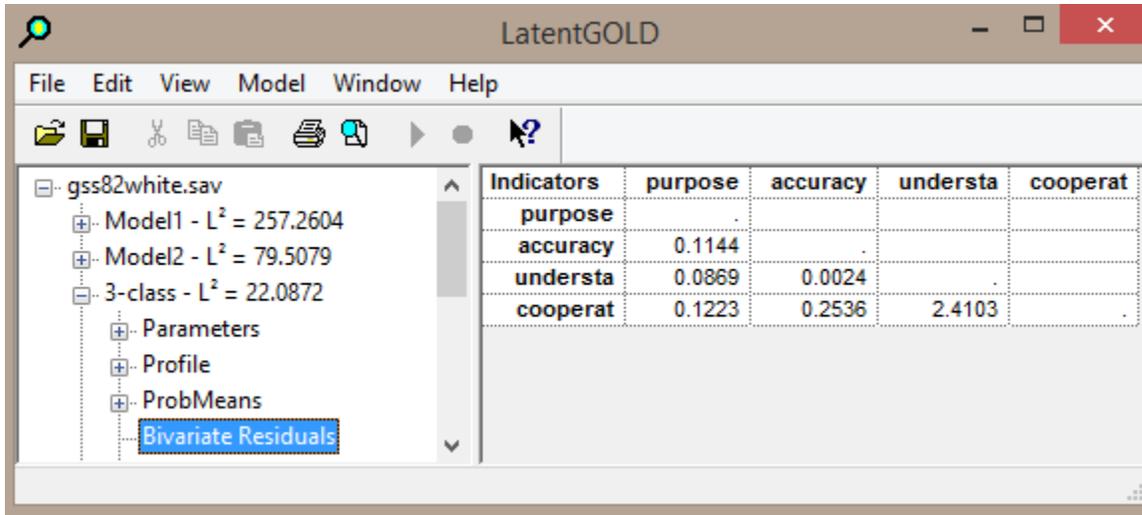
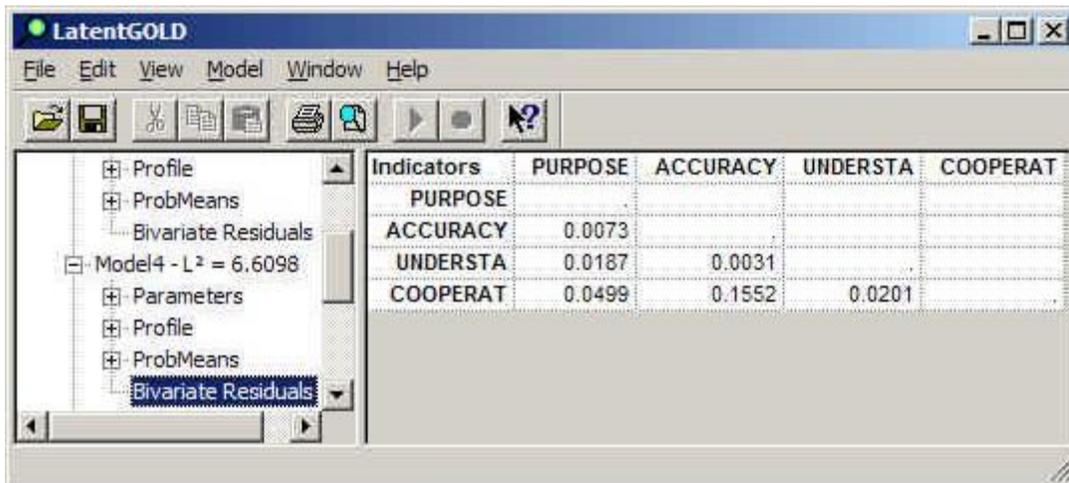


Figure 7-21. Bivariate Residuals Output for the 3-cluster Model

The BVR corresponds to a Pearson chi-squared divided by the degrees of freedom. The chi-square is computed on the observed counts in a 2-way table using the estimated expected counts obtained from the estimated model. If the model were true, BVRs should not be substantially larger than 1. The BVR of 2.4 in Figure 7-21 above suggests that the 3-cluster model may fall somewhat short in reproducing the association between COOPERATE and UNDERSTAND.

In contrast, the BVRs associated with 4-cluster model (shown below) are all less than 1. This suggests that the 4-cluster model may provide a significant improvement over the 3-cluster model in model fit.

Figure 7-22. Bivariate Residuals Output for the 4-cluster Model



Assessing Model Improvement Using the Conditional Bootstrap

The difference in L^2 between the 3- and 4-cluster models is a measure of the amount of fit improvement associated with the 4-cluster model over the 3-cluster model.

$$L^2(3\text{-class}) - L^2(4\text{-class}) = 22.0872 - 6.6142 = 15.473.$$

In general, the L^2 difference associated with nested models (where the nested model is a restricted form of the other model) can be tested using chi-square, with the degrees of freedom (df) being equal to the difference in df associated with both models. However, this test is not valid when the restriction involves setting the probability of a class membership to zero (e.g., the 3-class model can be formed by restricting the size of the 4th cluster to be zero). However, in such cases, a conditional bootstrap must be used to assess the significance of the difference in the L^2 statistics associated with the 3 and 4-class models.

The conditional bootstrap implemented in Latent GOLD is based on the log-likelihood (LL) rather than the L^2 statistic and hence can be used much more generally to compare restricted (i.e., nested) models, even in situations when chi-square statistics are not available. The reduction in L^2 can be expressed exactly in terms of twice the increase in LL associated with the increase in number of classes from 3 to 4.

$$L^2(3\text{-class}) - L^2(4\text{-class}) = -2LL(3\text{-class}) - -2LL(4\text{-class}).$$

To test whether the 4-class model (the source model) provides a significant improvement over the 3-class model (the nested reference model) you would select the 4-cluster model as the *source model*

- Click on 'Model4'
- Select 'Bootstrap -2LL Diff' from the Model Menu

Alternatively, you can

- Right-click on 'Model4'
- Select 'Bootstrap -2LL Diff' from the pop-up menu.

Following this, a list of eligible *reference models* appears:

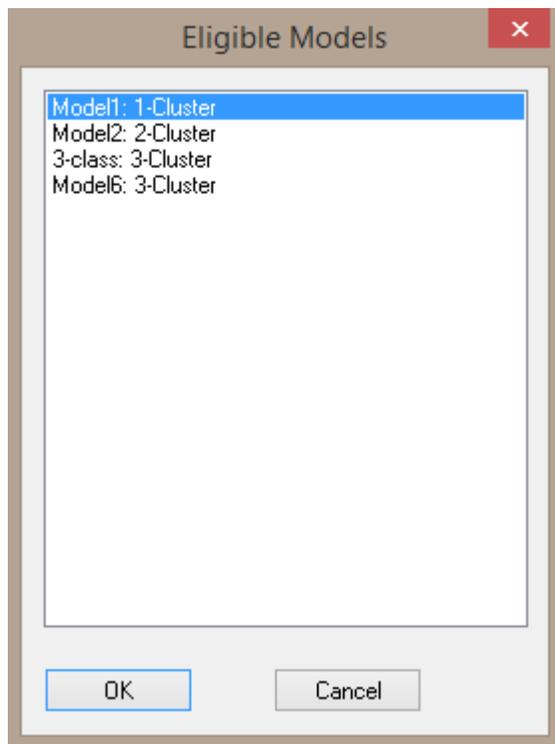


Figure 7-23. List of Eligible Reference Models for Conditional Bootstrap

- Select '3-Class' as the reference model
- Click OK

The conditional bootstrap procedure begins. Upon completion, 2 additional models named '3-ClassBoot' and 'Model4Boot' appear in the Outline Pane. The model labeled '3-ClassBoot' reproduces the earlier Bootstrap result we obtained.

To see the results from the conditional bootstrap, click on Model4Boot.

Figure 7-24. Conditional Bootstrap Output

4-Cluster Model					
Number of cases	1202				
Number of parameters (Npar)	27				
Random Seed	1323014				
Best Start Seed	1323014				
Monte Carlo Seed	492616				
Chi-squared Statistics		Bootstrap			
Degrees of freedom (df)	8	p-value	p-value	s.e.	CV
L-squared (L ²)	6.6142	0.58	0.8520	0.0159	17.0092
X-squared	6.1078	0.64	0.8220	0.0171	15.9461
Cressie-Read	6.0582	0.64	0.8400	0.0164	15.8399
BIC (based on L ²)	-50.1198				
AIC (based on L ²)	-9.3858				
AIC3 (based on L ²)	-17.3858				
CAIC (based on L ²)	-58.1198				
SABIC (based on L ²)	-24.7087				
Dissimilarity Index	0.0085		0.9660	0.0081	0.0213
Total BVR	0.2536		0.7200	0.0201	1.2478
Log-likelihood Statistics		-2LL Diff	p-value	s.e.	CV
Log-likelihood (LL)	-2746.9065	15.4731	0.0180	0.0059	13.5798
Log-prior	-5.3439				
Log-posterior	-2752.2503				
BIC (based on LL)	5685.2900				
AIC (based on LL)	5547.8129				
AIC3 (based on LL)	5574.8129				
CAIC (based on LL)	5712.2900				
SABIC (based on LL)	5599.5274				

We see that the estimated p-value associated with the increase in classes is 0.0180 (with standard error of 0.0059). Since $p < 0.05$, this means that the 4-Class Model *does* provide a significant improvement over the 3-class Model.

Note. Since the bootstrap p-value is estimated from the generated sample of size 500, the results you may get for the estimated p-value may be somewhat different because your sample will be different.

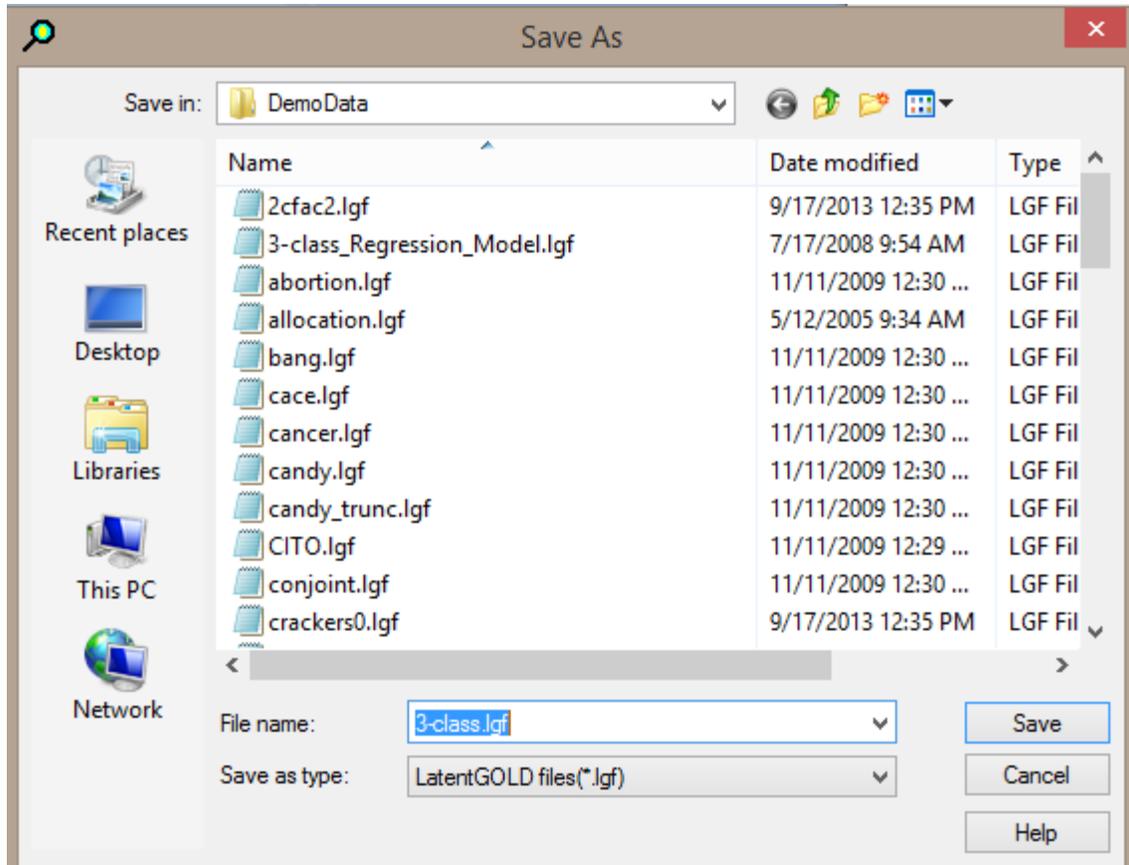
In Tutorials #2 and #3, we will explore the analyses of these data further. To

save the 3-class model settings for use in these future tutorials:

- Select '3-class'
- Select the 'Save Definition' option from the File Menu

The save dialog box appears

Figure 7-25. File Save Definition Dialog Box



- Click Save

You may also save any or all output using the Save Results option from the File Menu.