CHAPTER 7. TUTORIALS

Sections 7.1 - 7.4 contain tutorials which introduces the basic ideas in traditional latent class modeling using 4 nominal categorical variables. Tutorial #1 introduces the LC Cluster model and also illustrates how the new conditional bootstrap feature in Latent GOLD 4.0 can be used to test for a significant improvement when increasing the number of classes. Tutorial #2 analyzes the same data using DFactor models, showing the close relationship between LC Cluster and DFactor models. Tutorial #3 introduces LC Regression models with rating-based conjoint data. Tutorial #4 focuses on profiling latent class segments using demographic and other exogenous covariates, and illustrates the CHAID option for developing such profiles.

Chapter 8 contains a listing and a brief description of some of the other tutorials that can be viewed on our website.

7.1. Tutorial #1: Using Latent GOLD 4.0 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 4.0 accepts data from a variety of formats including SPSS system files, and ASCII rectangular files. (Data saved in any of 80 additional file formats are available using the DBMS/Copy File/Import add-on). The following data illustrates the use of an SPSS .sav file containing N=1,202 cases and an optional case weight variable FRQ.

gss82white.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : frq 419

	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

Data View Variable View

SPSS Processor is ready

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

** 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

Outline Pane **Contents Pane**

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.

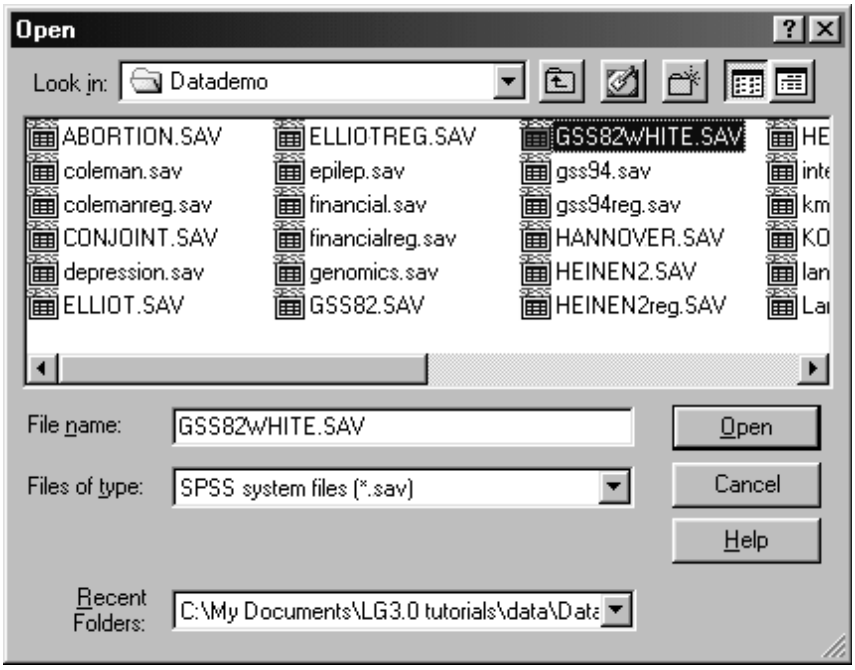


Figure 7-2. File Open Dialog Box

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

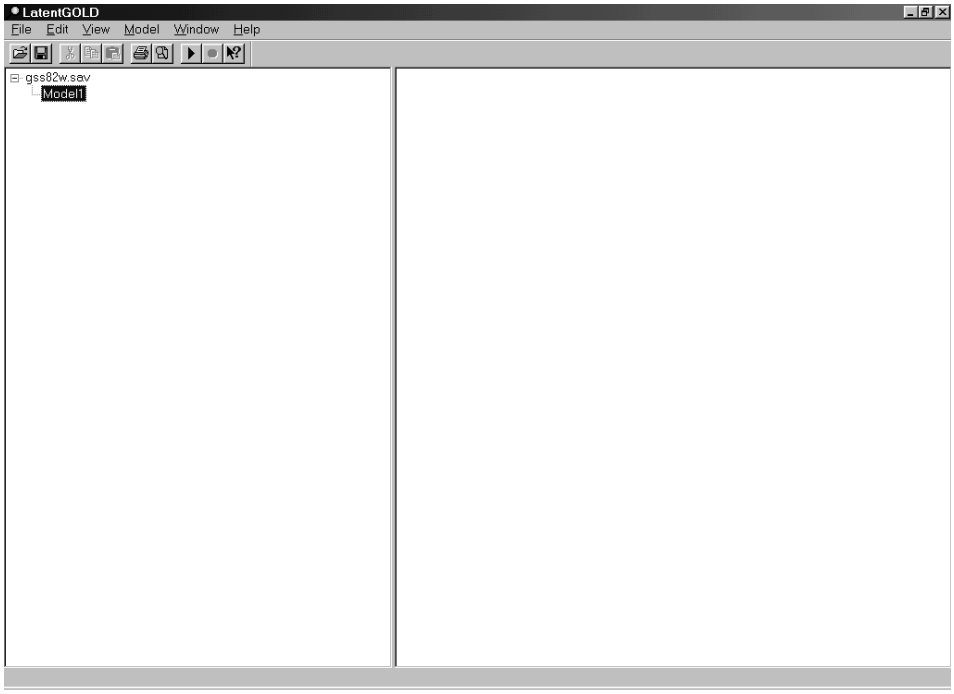


Figure 7-3. Viewer Window

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.

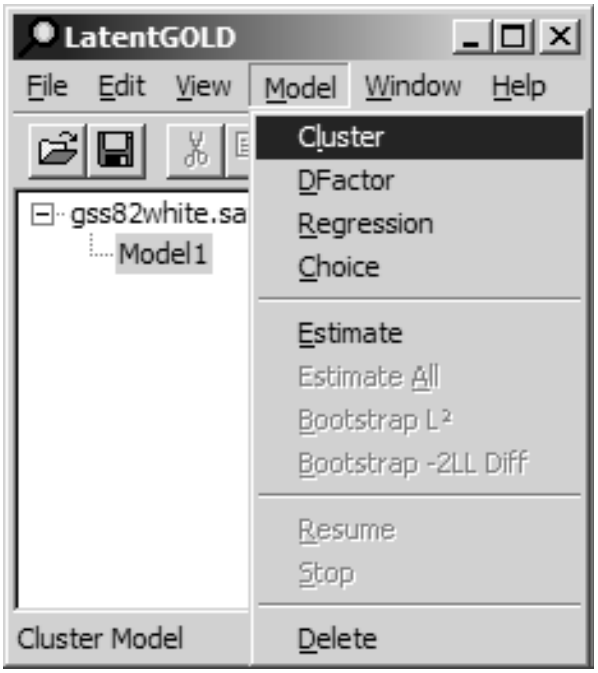


Figure 7-4. Model Selection Menu

▷ `Select Cluster`

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

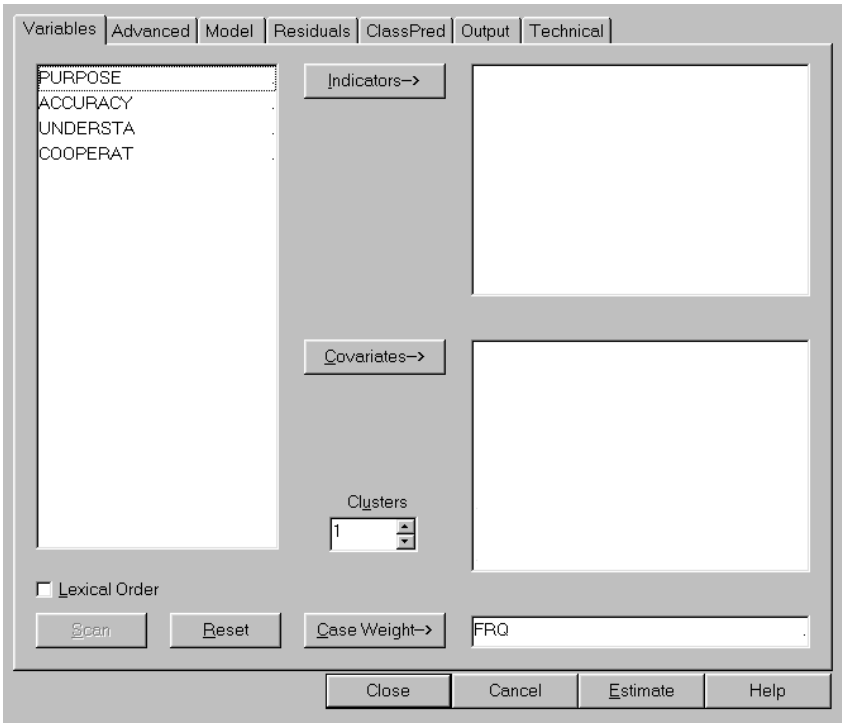


Figure 7-5. Analysis dialog box for LC Cluster Model

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').

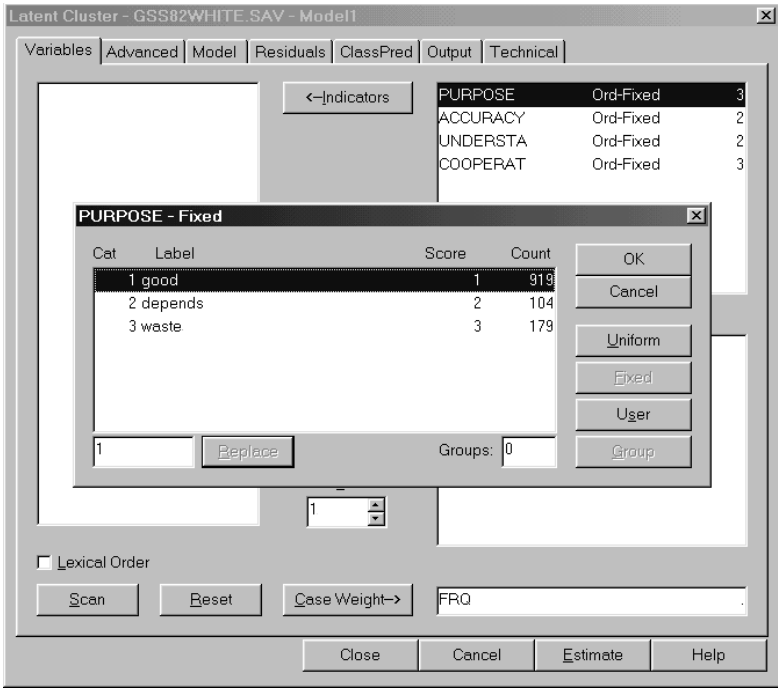


Figure 7-6. Category Information for the Variable PURPOSE

▷ 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.

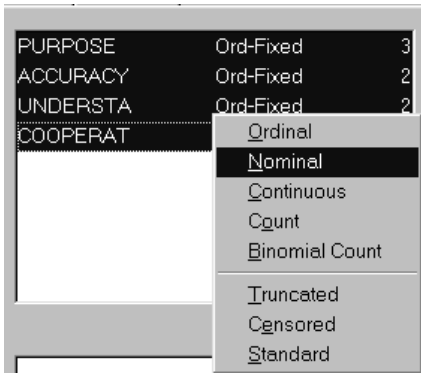


Figure 7-7. Pop-up Menu to Set Variable Scale Types and Subtypes

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:

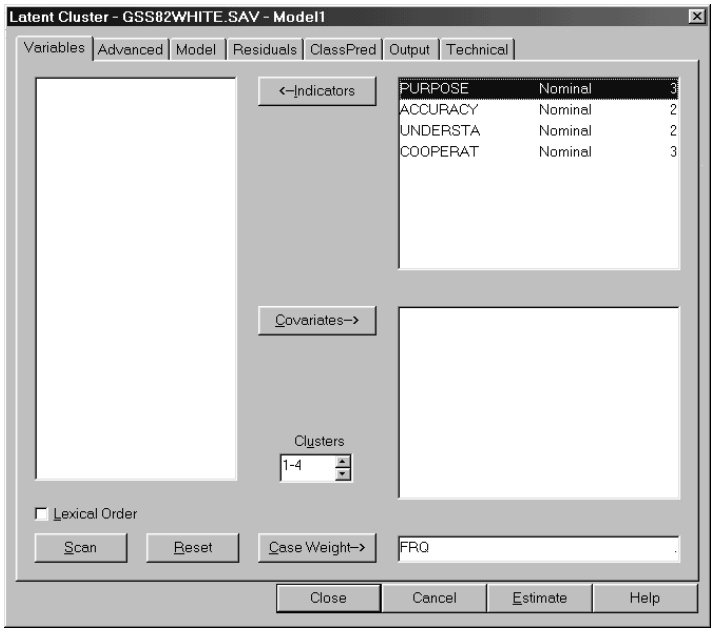


Figure 7-8. Analysis Dialog Box Prior to Model Estimation.

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 L², 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

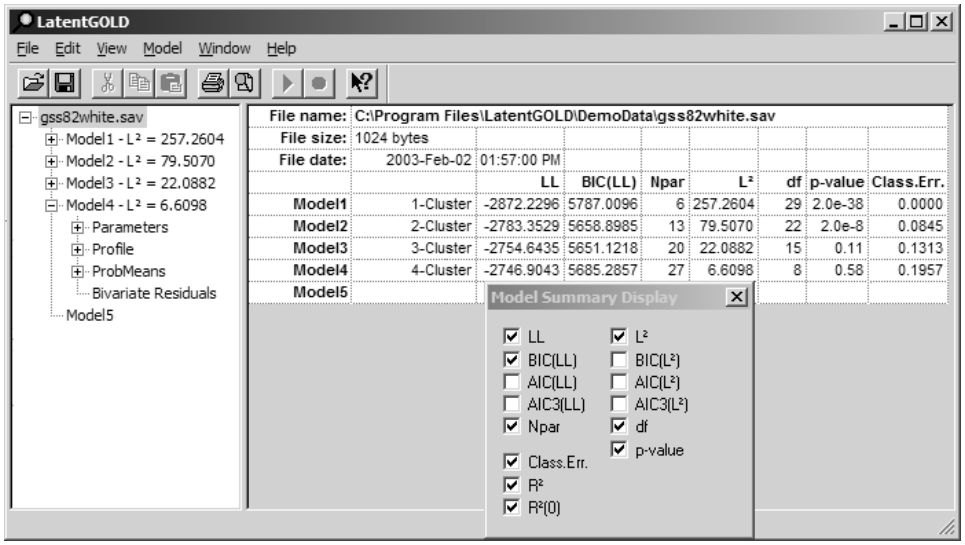


Figure 7-9: Model Summary Output and Model Summary Display

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

The model L² statistic, as shown in Figure 7-9 in the column labeled 'L² ', 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 p-value. This provides a more precise estimate by relaxing the assumption that the L^2 statistic follows a chi-square 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 L^2'

or alternatively,

- select '3-class'
- select 'Bootstrap L^2 from the Model Menu

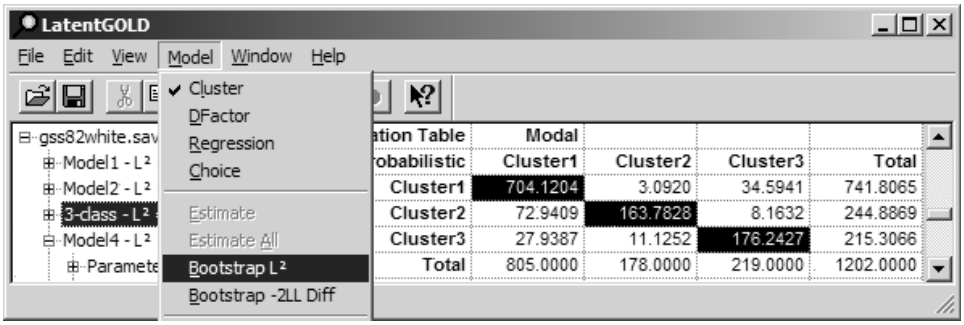


Figure 7-10: Model Menu Showing Option for Bootstrap of L^2

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.

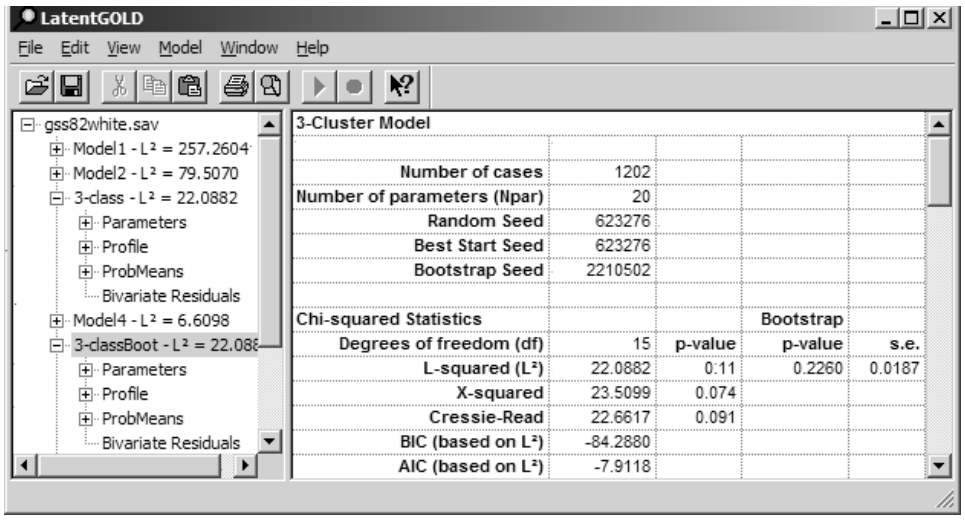


Figure 7-11: Results for the Bootstrap of L^2

Figure 7-11 shows the estimate of the p-value resulting from the bootstrap procedure. It is p = .226 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`

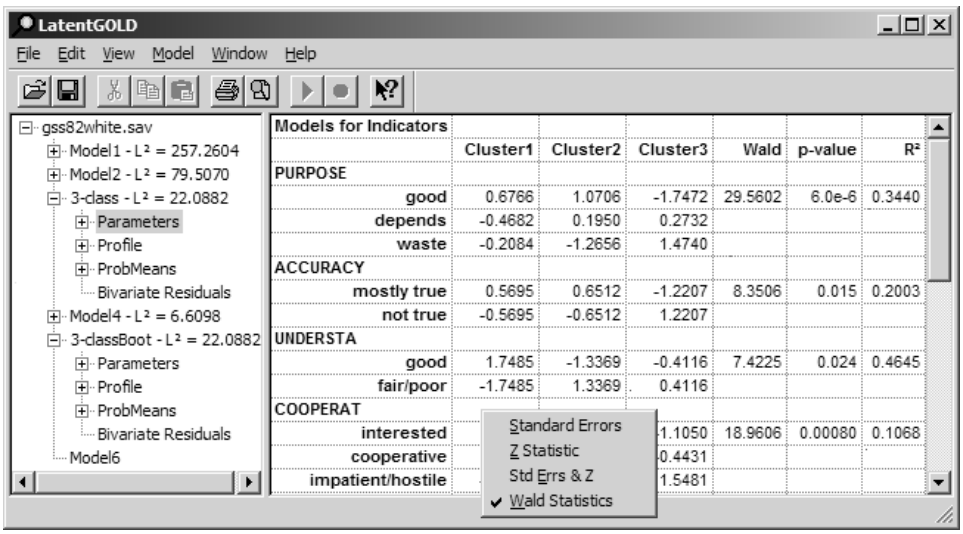
▷ `Click Parameters`

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.



Models for Indicators	Cluster1	Cluster2	Cluster3	Wald	p-value	R^2
PURPOSE						
good	0.6766	1.0706	-1.7472	29.5602	6.0e-6	0.3440
depends	-0.4682	0.1950	0.2732			
waste	-0.2084	-1.2656	1.4740			
ACCURACY						
mostly true	0.5695	0.6512	-1.2207	8.3506	0.015	0.2003
not true	-0.5695	-0.6512	1.2207			
UNDERSTA						
good	1.7485	-1.3369	-0.4116	7.4225	0.024	0.4645
fair/poor	-1.7485	1.3369	0.4116			
COOPERAT						
interested			1.1050	18.9606	0.00080	0.1068
cooperative			0.4431			
impatient/hostile			1.5481			

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

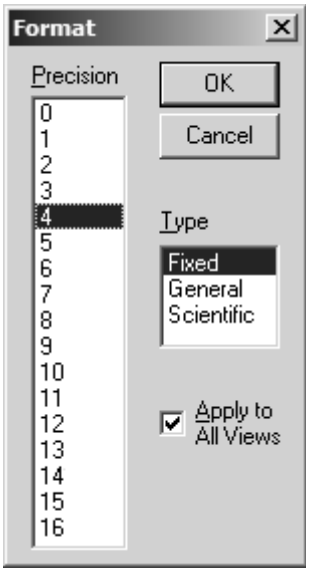


Figure 7-13. Format Control Options from View Menu

PROFILE OUTPUT AND ASSOCIATED PROFILE PLOT

To view the parameters re-expressed as conditional probabilities

▷ Click on Profile

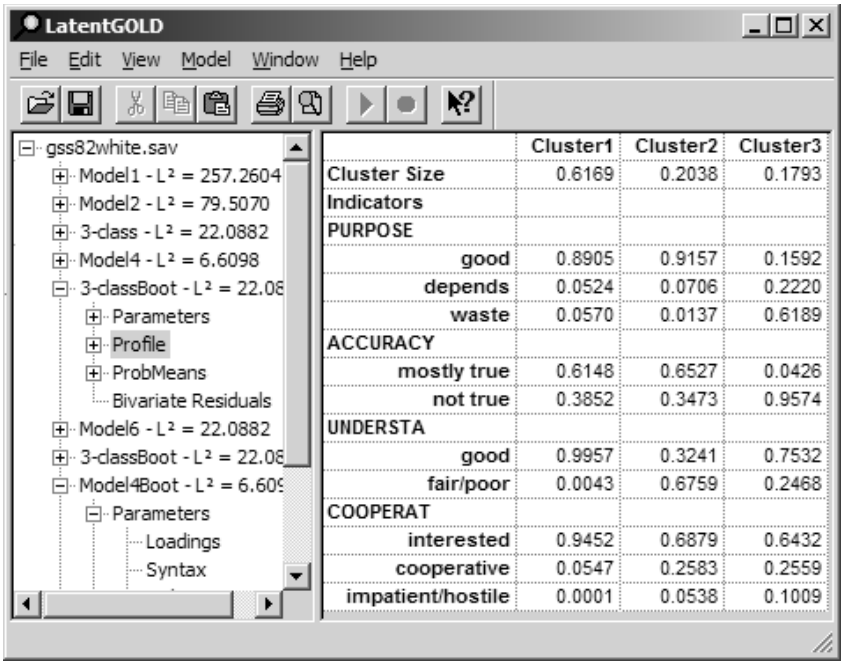


Figure 7-14. Profile Output for 3-cluster Model

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'

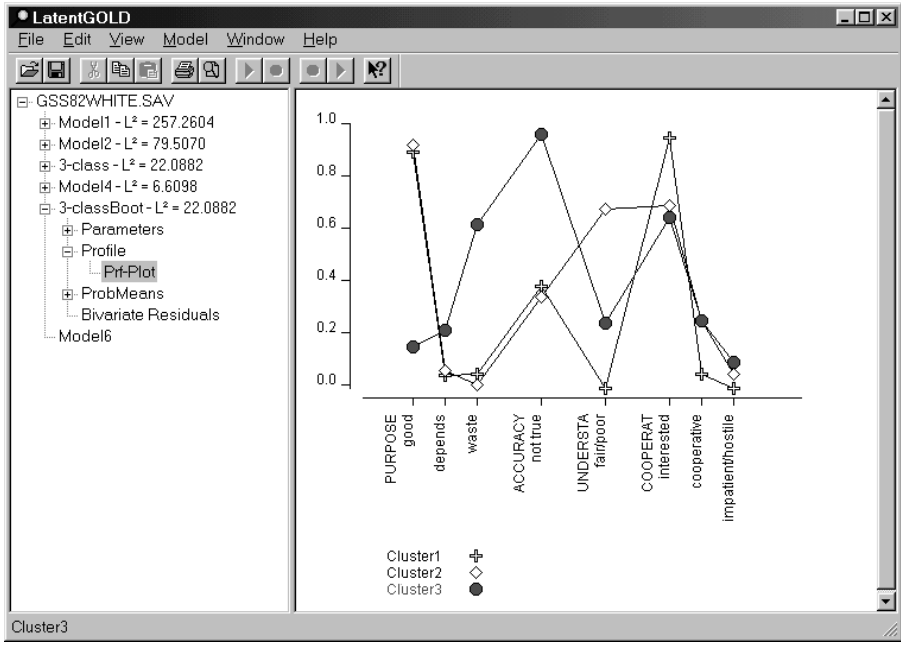


Figure 7-15. Profile Plot for 3-cluster Model

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.

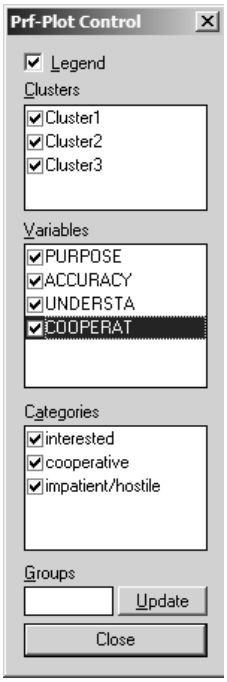


Figure 7-16. Profile Plot Control Panel

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

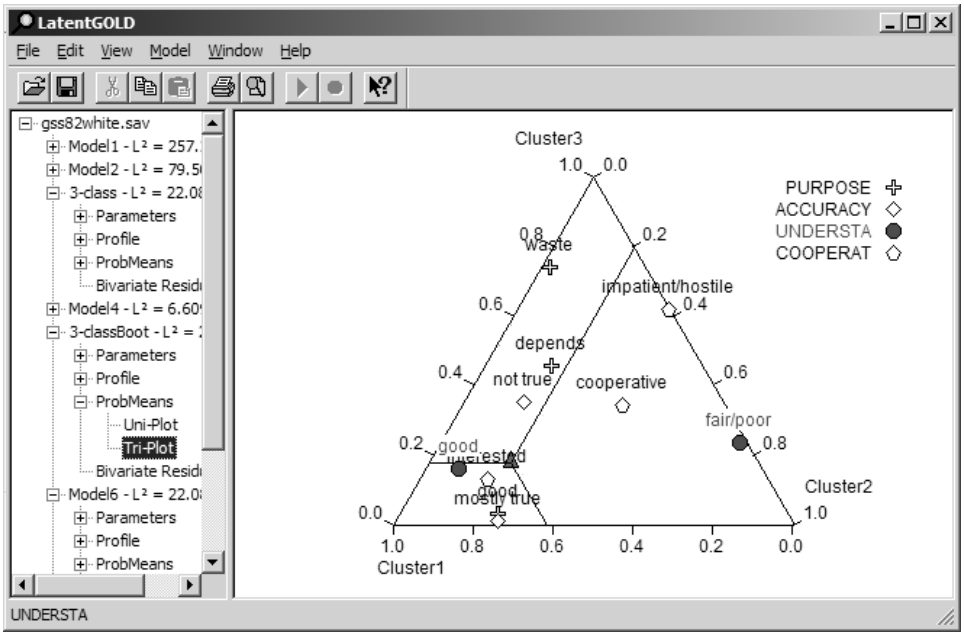


Figure 7-17. Tri-plot Display for 3-class Model

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 Standard Classification:

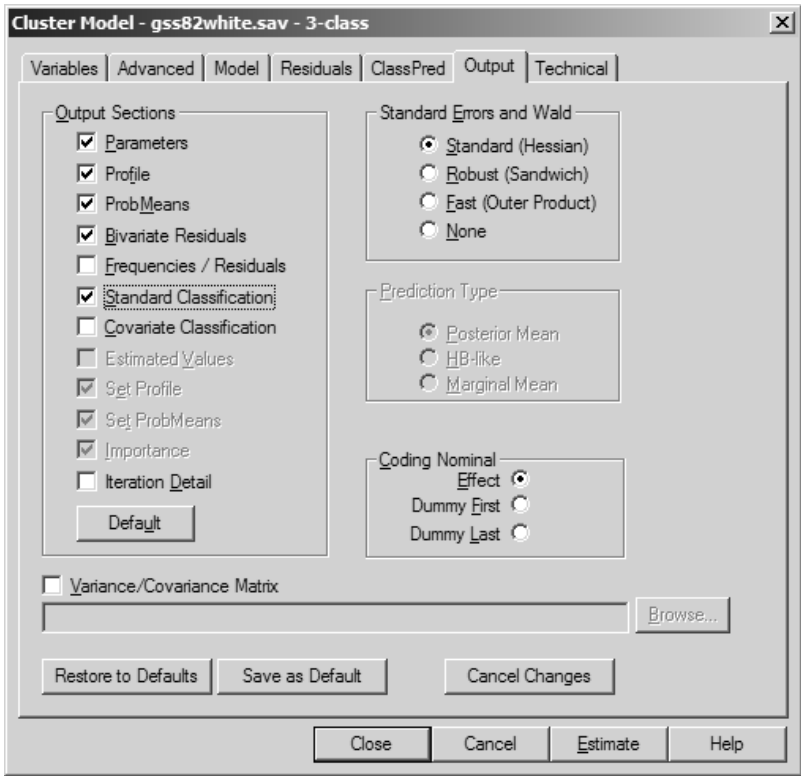
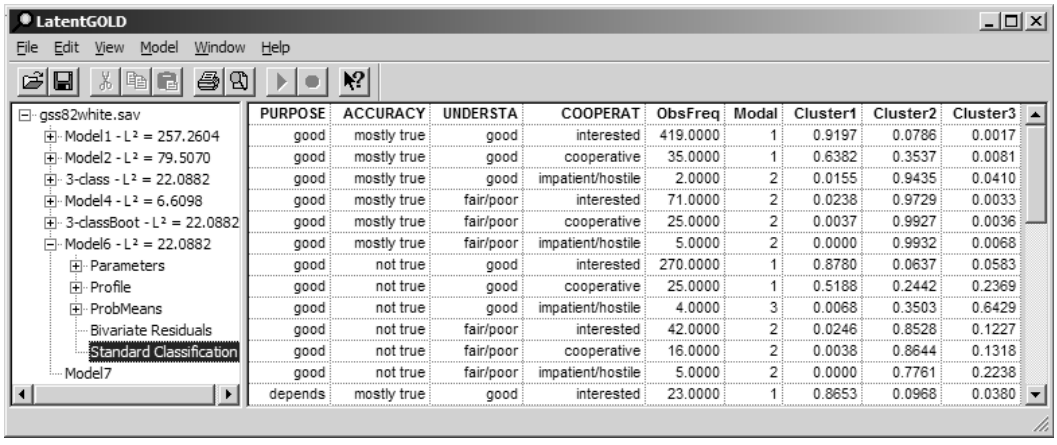


Figure 7-18. Requesting Standard Classification Output Listing in Output Tab

▷ `Click Estimate`

▷ `Under the new Model, click on 'Standard Classification' to view the Classification Output:`



PURPOSE	ACCURACY	UNDERSTA	COOPERAT	ObsFreq	Modal	Cluster1	Cluster2	Cluster3
good	mostly true	good	interested	419.0000	1	0.9197	0.0786	0.0017
good	mostly true	good	cooperative	35.0000	1	0.6382	0.3537	0.0081
good	mostly true	good	impatient/hostile	2.0000	2	0.0155	0.9435	0.0410
good	mostly true	fair/poor	interested	71.0000	2	0.0238	0.9729	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	0.8780	0.0637	0.0583
good	not true	good	cooperative	25.0000	1	0.5188	0.2442	0.2369
good	not true	good	impatient/hostile	4.0000	3	0.0068	0.3503	0.6429
good	not true	fair/poor	interested	42.0000	2	0.0246	0.8528	0.1227
good	not true	fair/poor	cooperative	16.0000	2	0.0038	0.8644	0.1318
good	not true	fair/poor	impatient/hostile	5.0000	2	0.0000	0.7761	0.2238
depends	mostly true	good	interested	23.0000	1	0.8653	0.0968	0.0380

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

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 (.9197). 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 Standard Classification on the ClassPred Tab:

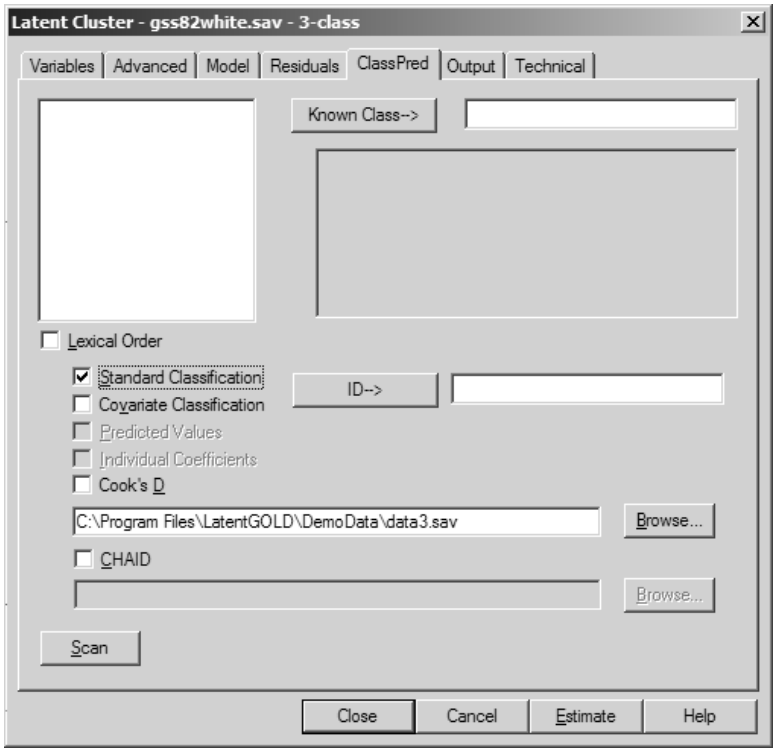


Figure 7-20. Requesting Output of Standard Classification Information to a Data File

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.

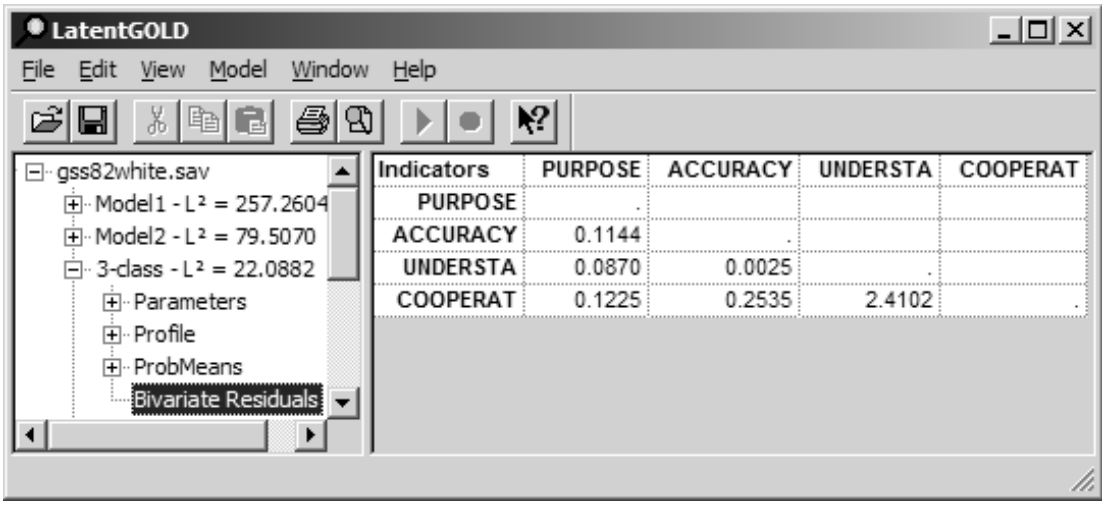


Figure 7-21. Bivariate Residuals Output for the 3-cluster Model

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.

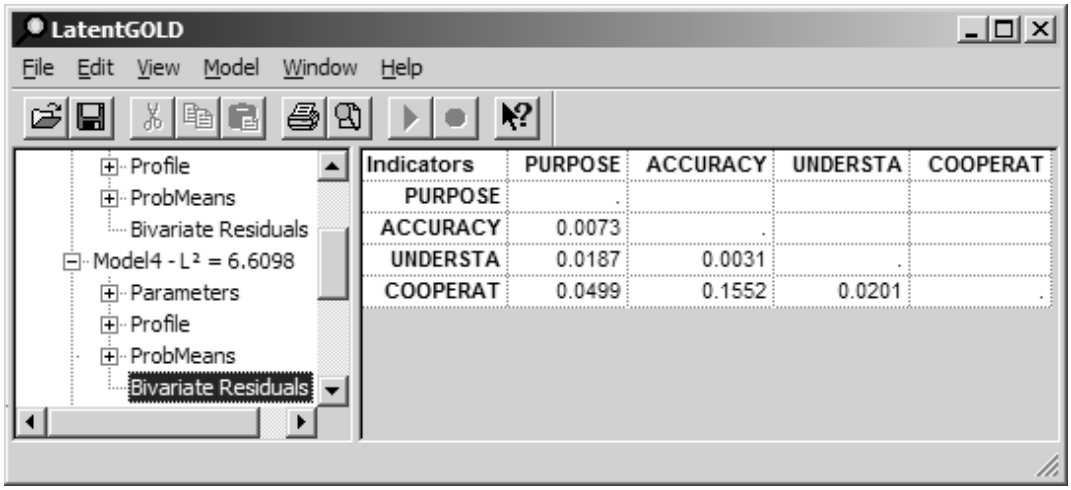


Figure 7-22. Bivariate Residuals Output for the 4-cluster Model

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.

$$L^2(\text{3-class}) - L^2(\text{4-class}) = 22.0882 - 6.6098 = 15.478.$$

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 4.0 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.

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

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:

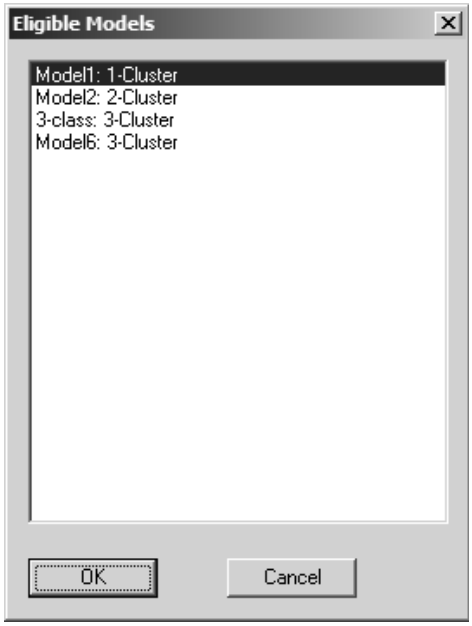


Figure 7-23. List of Eligible Reference Models for Conditional Bootstrap

- ▷ 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.

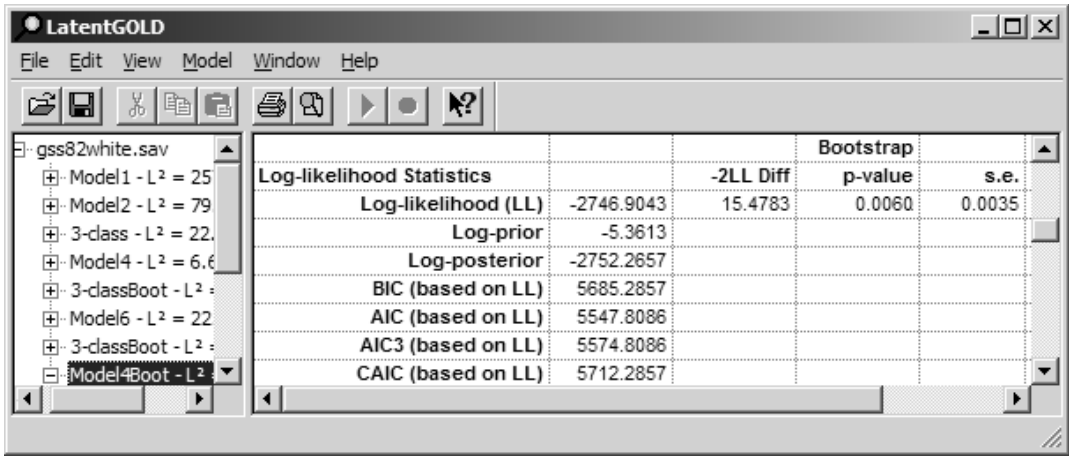


Figure 7-24. Conditional Bootstrap Output

We see that the estimated p-value associated with the increase in classes is 0.006 (with standard error of 0.0035). 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

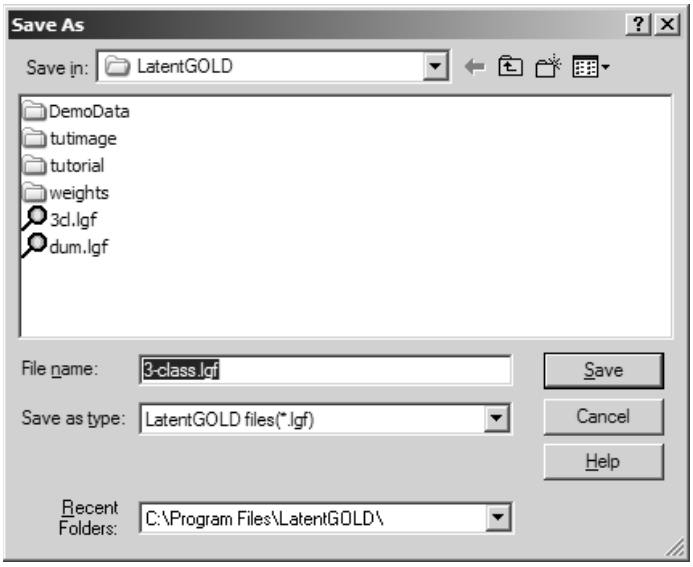


Figure 7-25. File Save Definition Dialog Box

▷ `Click Save`

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

7.2. Tutorial #2: Using Latent GOLD to Estimate DFactor Models

DEMODATA = 'GSS82WHITE.SAV'

The Goal

In this tutorial, we re-examine the results obtained from tutorial #1 using discrete factor (DFactor) models instead of LC Cluster models. We show how a 2-DFactor model consisting of 2 dichotomous factors can be viewed as a restricted form of the 4-cluster model and use the L^2 difference statistic to test whether the unrestricted 4-class model provides an improvement. In addition, this tutorial illustrates:

- **The use of the Ordinal scale type**
- **Estimating DFactor models**
- **Factor Loadings Output**
- **Restricting Factor Loadings to Zero**
- **Joint Profile output**
- **Classification Output**
- **The Bi-plot**

For these data the DFactor models provide additional insights into the different survey respondent types.

DFactor Analysis vs. Traditional Factor Analysis

In traditional factor analysis (FA), continuous observed variables are expressed as a linear function of 1 or more continuous latent factors (CFactors). DFactor analysis differ from FA in several respects:

- The observed variables may be of mixed scale types including nominal, ordinal, continuous and count.
- The latent variables are not continuous but discrete, containing 2 or more ordered categories (levels)
- The model is not linear
- Solutions need not be rotated to be interpretable (the indeterminacy issue of 'rotation' is unique to CFactors in a linear model).

In addition, a cross-tabulation of DFactors defines a set of clusters. For example, 2 dichotomous DFactors V and W, yields 4 latent classes (4 clusters).

	W=1	W=2
V=1	X=1	X=2
V=2	X=3	X=4

Figure 7-26. The 4 latent classes

As our starting point, we will re-estimate the 3- and 4-class Cluster models from tutorial #1.

Opening the data file

To retrieve the model setup for the 3-class model,

- Select File/Open '3-class.lgf'

Figure 7-27. Setup for '3-class' Model

- Double click on "3-class" to open the Analysis Dialog box
- Click Estimate to re-estimate this model

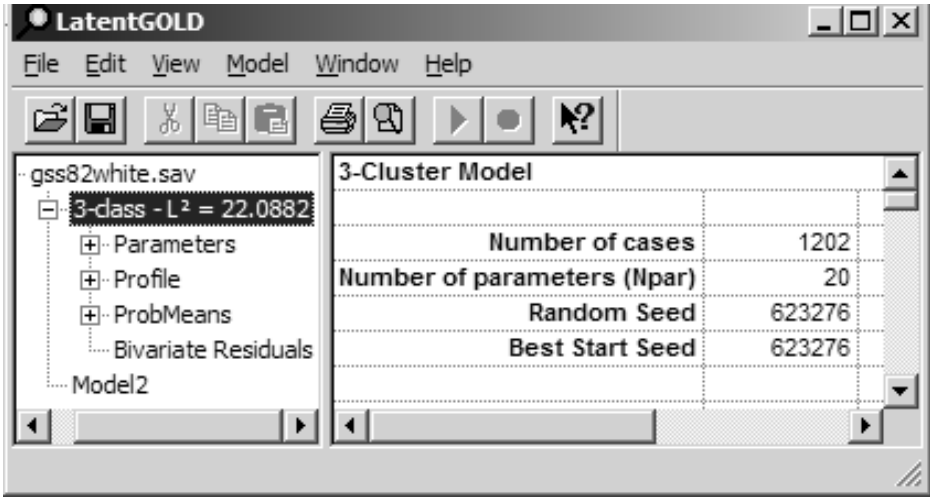


Figure 7-28. Model Summary Output for 3-class Model

To estimate the 4-class model,

- Double click on Model2 to re-open the Analysis Dialog box
- Change '3' to '4' in the Clusters box
- Click Estimate

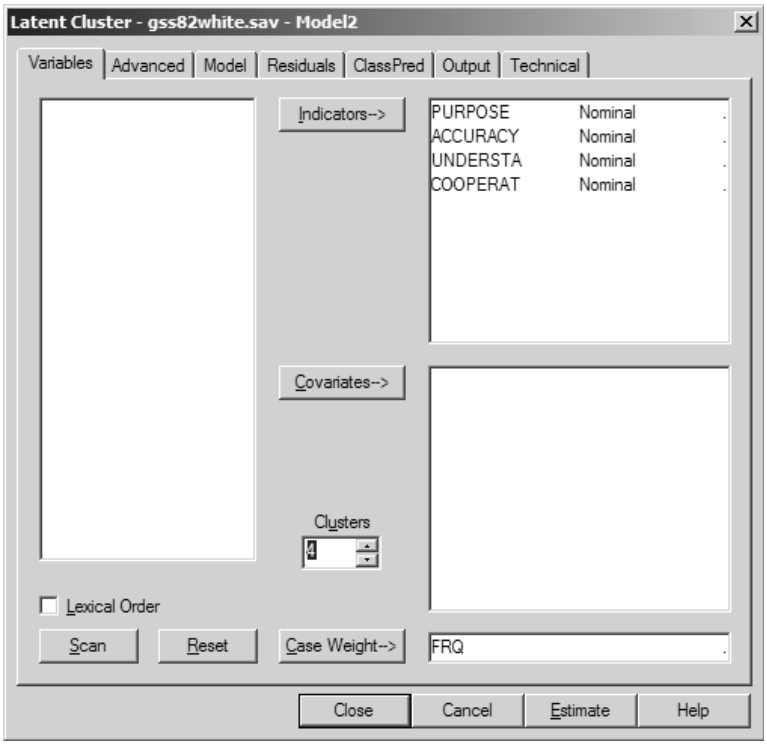


Figure 7-29. Estimating Model2

The 4-classmodel is named 'Model2'.

To change the name to '4-class'

- Click once on Model2 to select it
- Click once again on it to enter Edit mode
- Type '4-class'

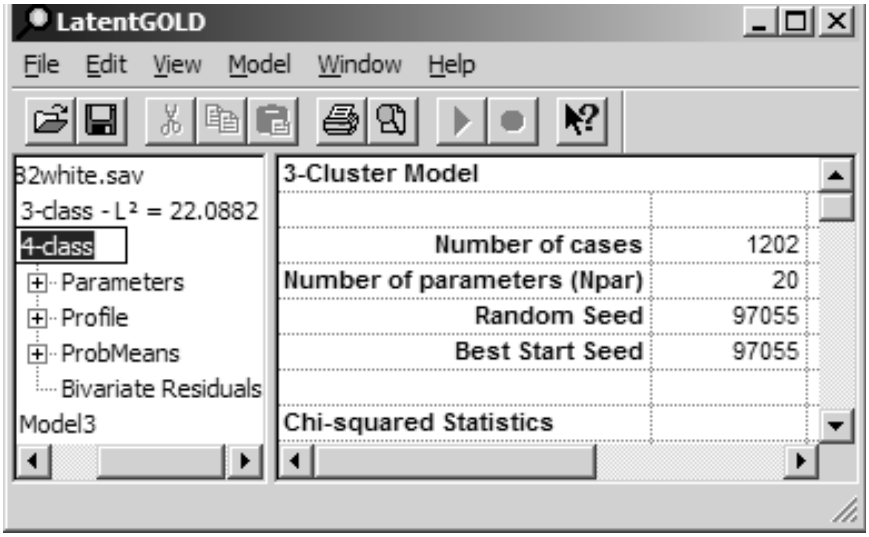
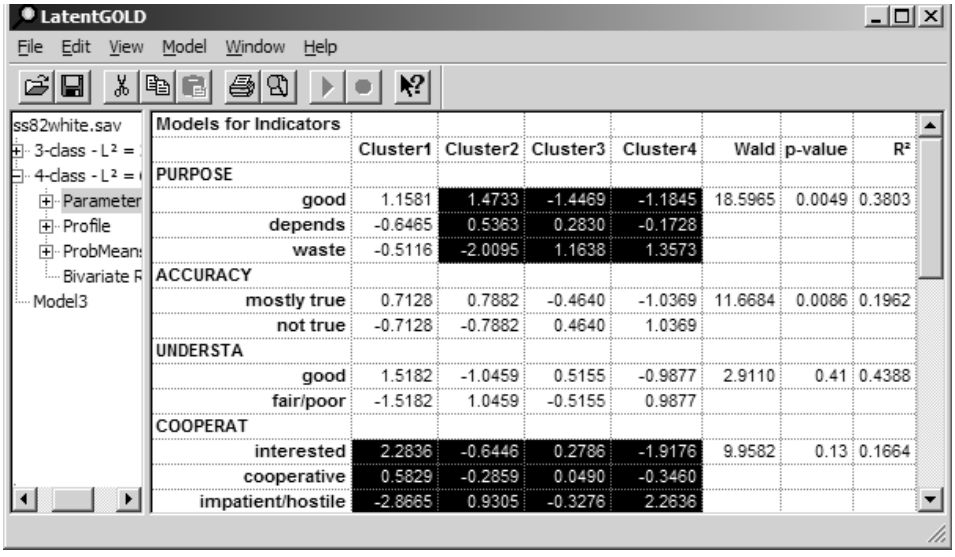


Figure 7-30. Editing the name of Model2

- Click Parameters to view the parameter estimates for the 4-class model



	Cluster1	Cluster2	Cluster3	Cluster4	Wald	p-value	R²
PURPOSE							
good	1.1581	1.4733	-1.4469	-1.1845	18.5965	0.0049	0.3803
depends	-0.6465	0.5363	0.2830	-0.1728			
waste	-0.5116	-2.0095	1.1638	1.3573			
ACCURACY							
mostly true	0.7128	0.7882	-0.4640	-1.0369	11.6684	0.0086	0.1962
not true	-0.7128	-0.7882	0.4640	1.0369			
UNDERSTA							
good	1.5182	-1.0459	0.5155	-0.9877	2.9110	0.41	0.4388
fair/poor	-1.5182	1.0459	-0.5155	0.9877			
COOPERAT							
interested	2.2836	-0.6446	0.2786	-1.9176	9.9582	0.13	0.1664
cooperative	0.5829	-0.2859	0.0490	-0.3460			
impatient/hostile	-2.8665	0.9305	-0.3276	2.2636			

Figure 7-31. Parameters Output for 4-class model.

Notice that for the trichotomous variables PURPOSE and COOPERATE, the estimate for the middle level in each class is approximately midway between the estimates for the end categories (with the single exception of PURPOSE for cluster 1). This suggests that treating these variables as ordinal rather than nominal may be justified, using the default equidistant category scores.

To change the scale type to Ordinal,

- Double click on Model3 to re-open the Analysis Dialog box
- Ctrl-click on PURPOSE and COOPERAT to select these variables
- Right click to retrieve the scale type settings menu
- Select Ordinal

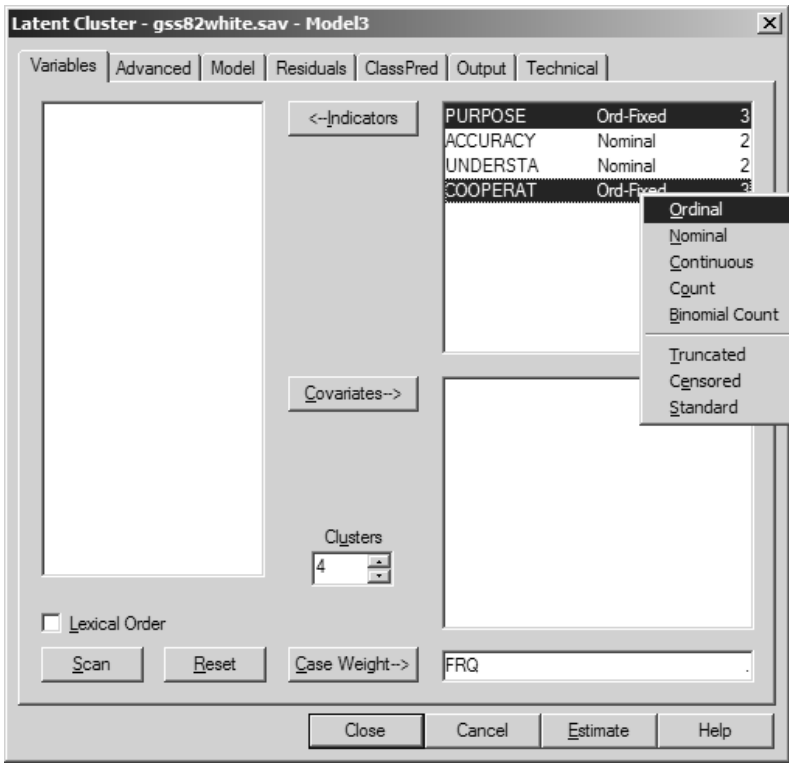


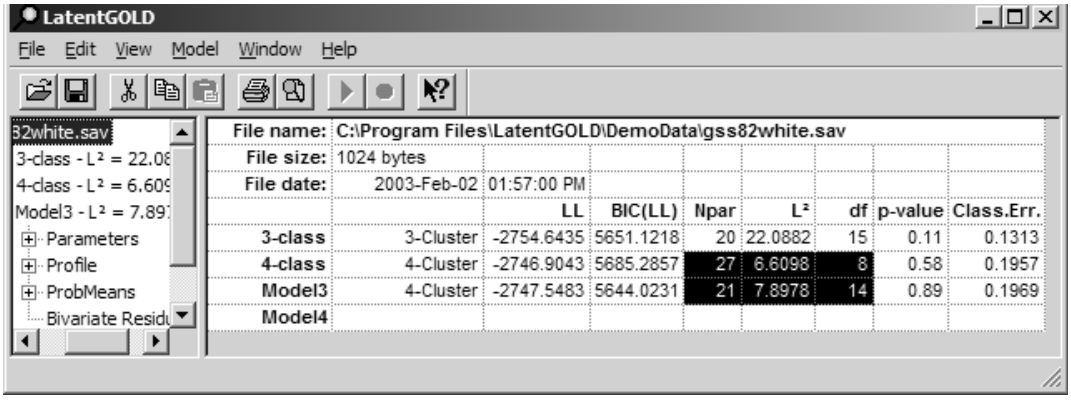
Figure 7-32. Changing PURPOSE and COOPERAT to ordinal variables

- Click Estimate to estimate the model

To compare the results from these models

- Click on the data file name in the Outline Pane

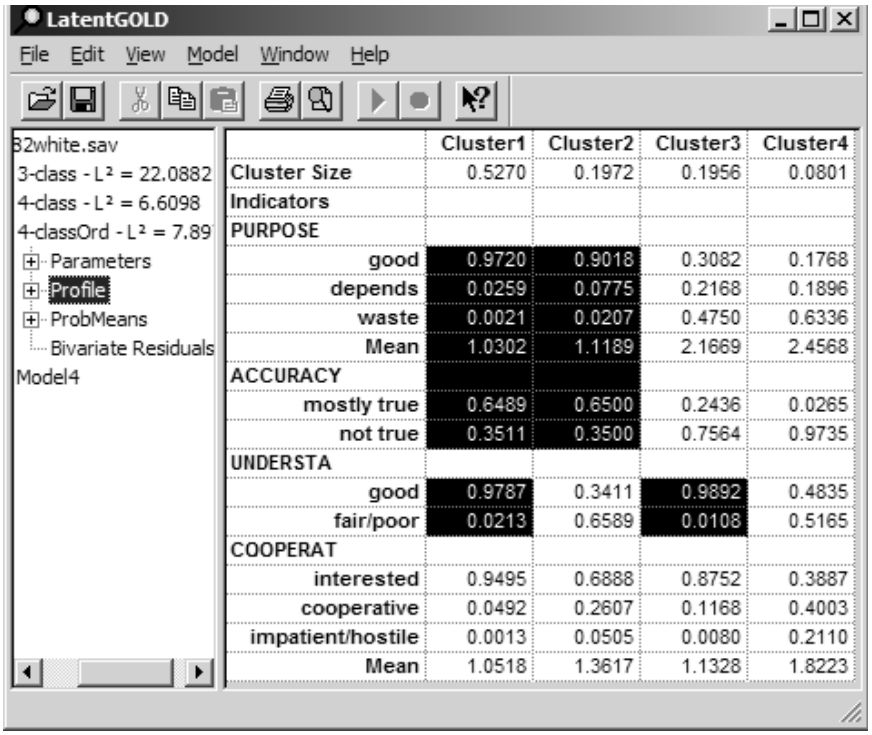
We see that imposing the ordinality restrictions increase L^2 from 6.6 to 7.9, a very small increase associated with the gain of 6 degrees of freedom. Thus, we choose Model3 over the unrestricted 4-class model, which results in a 4-class model with 6 fewer parameters.



		LL	BIC(LL)	Npar	L²	df	p-value	Class.Err.
3-class	3-Cluster	-2754.6435	5651.1218	20	22.0882	15	0.11	0.1313
4-class	4-Cluster	-2746.9043	5685.2857	27	6.6098	8	0.58	0.1957
Model3	4-Cluster	-2747.5483	5644.0231	21	7.8978	14	0.89	0.1969
Model4								

Figure 7-33. Model Summary Output for 4-class and Model3

- Change the name 'Model3' to '4-classOrd'
- Click Profile to view the Profile Output



	Cluster1	Cluster2	Cluster3	Cluster4
Cluster Size	0.5270	0.1972	0.1956	0.0801
Indicators				
PURPOSE				
good	0.9720	0.9018	0.3082	0.1768
depends	0.0259	0.0775	0.2168	0.1896
waste	0.0021	0.0207	0.4750	0.6336
Mean	1.0302	1.1189	2.1669	2.4568
ACCURACY				
mostly true	0.6489	0.6500	0.2436	0.0265
not true	0.3511	0.3500	0.7564	0.9735
UNDERSTA				
good	0.9787	0.3411	0.9892	0.4835
fair/poor	0.0213	0.6589	0.0108	0.5165
COOPERAT				
interested	0.9495	0.6888	0.8752	0.3887
cooperative	0.0492	0.2607	0.1168	0.4003
impatient/hostile	0.0013	0.0505	0.0080	0.2110
Mean	1.0518	1.3617	1.1328	1.8223

Figure 7-34. Profile Output for 4 class Ordinal model.

Notice that clusters 1 and 2 have similar response distributions associated with PURPOSE and ACCURACY, and the same is true for clusters 3 and 4. Also, notice that Clusters 1 and 3 are similar in their response distribution on UNDERSTAND, as is also true of Clusters 2 and 4. This pattern suggests that PURPOSE and ACCURACY may be associated with one DFactor, while UNDERSTAND may be associated with a second.

ESTIMATING A 2-DFACTOR MODEL

- ▷ Right click on Model4 to open the Model Selection menu.
- ▷ Select DFactor to open the DFactor Analysis dialog box.
- ▷ Change the number of DFactors

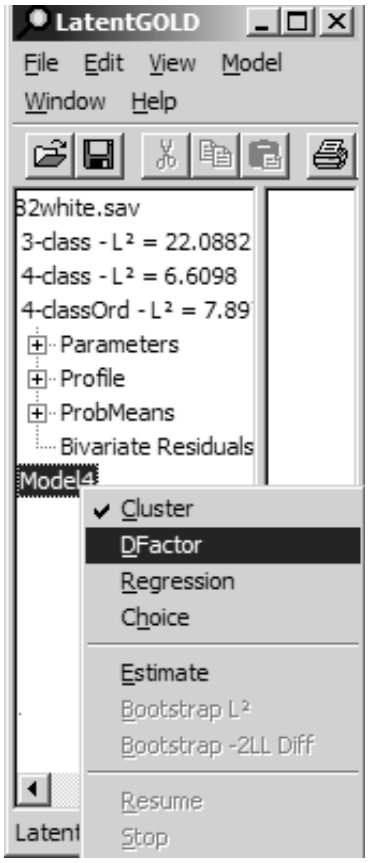


Figure 7-35. Selecting a DFactor Model

The DFactor Analysis Dialog Box opens, and the variable settings appear as before.

To estimate a 2-DFactor model.

- ▷ Set the number of DFactors in the DFactors Box to 2.

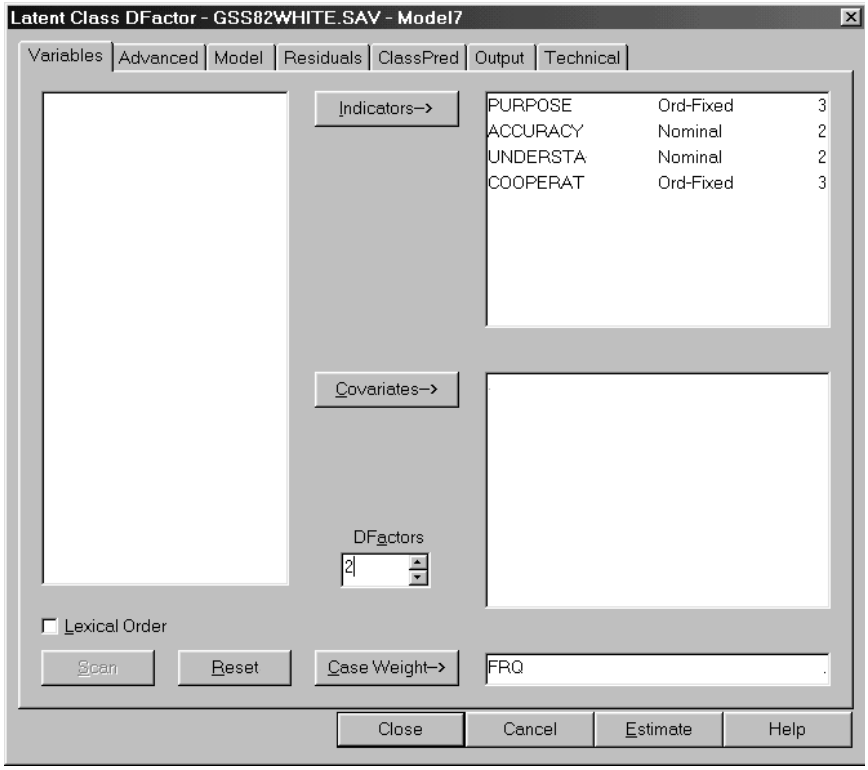


Figure 7-36. Changing number of DFactors

▷ `Click Estimate`

Now, highlight the file name gss82white.sav again to view the model comparisons

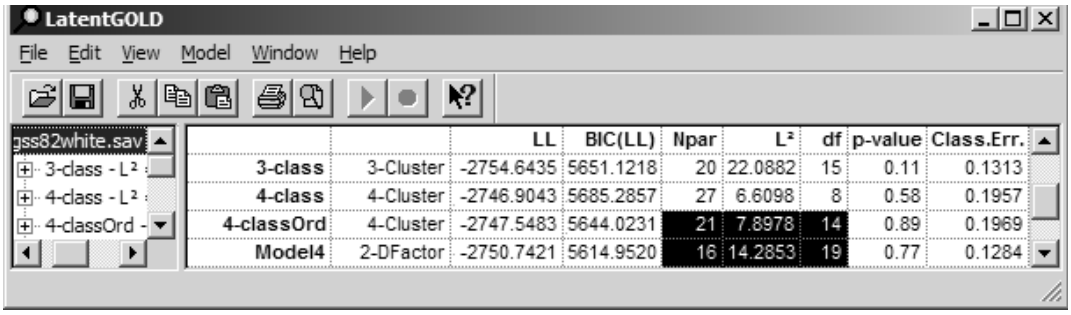


Figure 7-37. Model Comparison - 4-Cluster vs. 2-DFactor models

The 2-DFactor model applies further restrictions to the 4-class model, resulting in a model with 5 fewer parameters than model '4-classOrd' (resulting in a gain of 5 df). To test whether such restrictions are justified, we can test to see whether the increase in L^2 of 6.3875 (from 7.8978 to 14.2853) is statistically significant. We will test this in 2 ways.

First, we will use the chi-squared calculator which uses the chi-squared distribution to compute the p-value To open the chi-squared calculator

- Select View/ProbChi
- Enter 6.3875 in the Chi-Square box
- Enter 5 in the df box
- Click the Chi->p button

The p-value of .27 appears in the p-value box

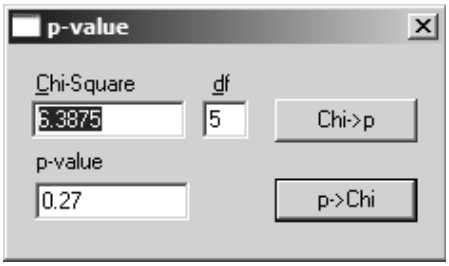


Figure 7-38. Chi-squared Calculator

Since .27 > .05, we fail to reject the restrictions, and so we will accept the 2-DFactor model.

The second way to estimate the p-value utilizes the conditional bootstrap which does not rely on any specific distribution. To use this, we would select model '4-classOrd' as the source and test whether it represents a significant improvement over the reference model 'Model4'.

- Click on '4-classOrd'
- Select 'Bootstrap -2LL Diff' from the Model Menu.

Alternatively, you can

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

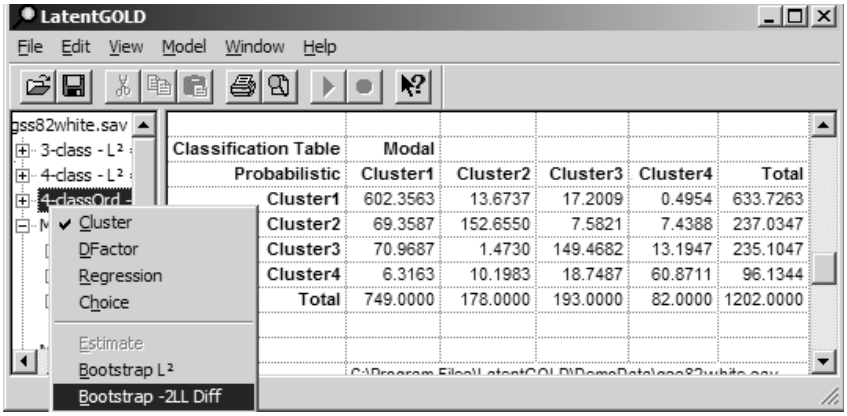


Figure 7-39. Estimating a Bootstrap -2LL Diff Model

Following this, a list of eligible reference models appears.

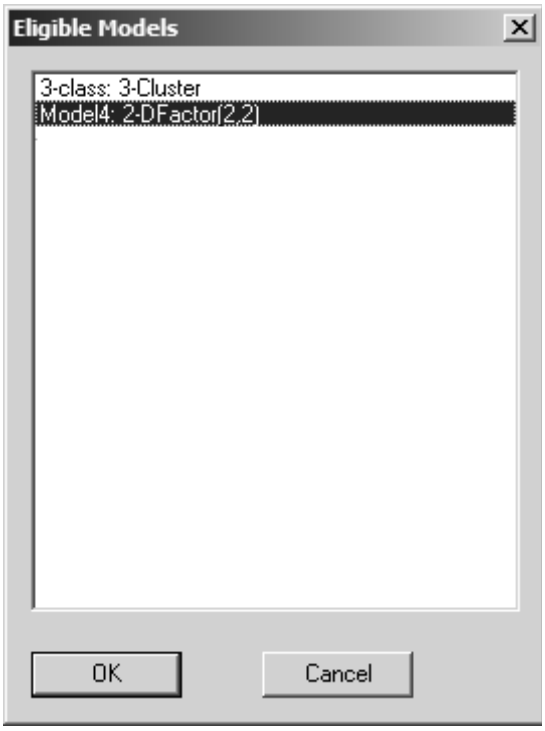


Figure 7-40. List of Eligible Reference Models for Conditional Bootstrap

- Select 'Model4', the 2-DFactor model
- Click OK

The conditional bootstrap procedure begins. Upon completion, 2 additional models named 'Model4Boot' and '4-ClassOrdBoot' appear in the Outline Pane. The results from the conditional bootstrap appear in the Outline Pane associated with model '4-ClassOrdBoot'. You may need to scroll down to see these results.

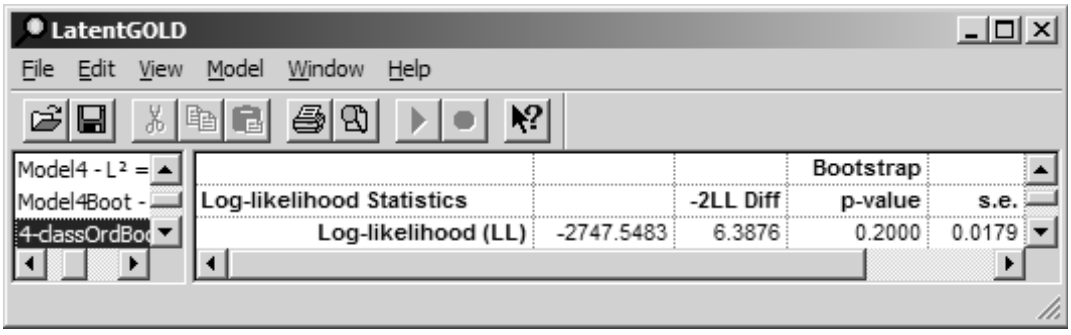


Figure 7-41. p-value Estimated using the Conditional Bootstrap

The p-value is estimated to be .20 with a standard error of about .02. This result is similar to what we obtained using the chi-squared approach.

The conditional bootstrap also provides a bootstrap estimate of the p-value associated with the reference model.

To view this,

▷ Click on 'Model4Boot'

The Contents Pane shows that the bootstrap estimate for the p-value associated with the 2-DFactor model ('Model4') is .866 with a standard error of .015. Again the results agree with the chi-squared based estimate of .77. (The assumptions underlying the use of the chi-squared based and the bootstrap estimates are both justified in this example.)

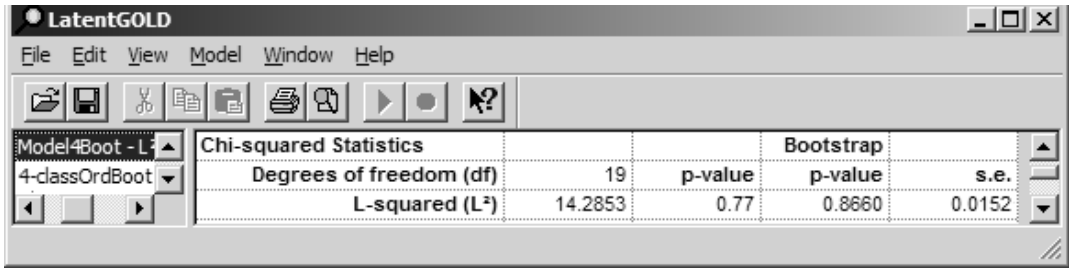


Figure 7-42. p-value Estimated by the Bootstrap for the Reference Model

Next, we will examine the output for the 2-DFactor model.

▷ Click on the expand/contract icon for Model4 to make the output listings visible

▷ Click on the expand/contract icon for Parameters to make the output subcategories visible

▷ Click 'Loadings' to view the DFactor loadings output

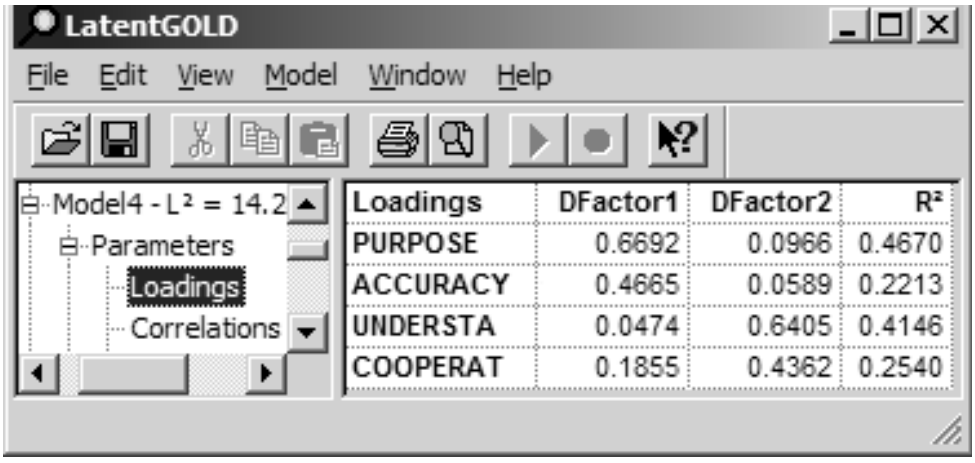


Figure 7-43. Loadings Output

This shows that PURPOSE and ACCURACY load primarily on DFactor1 (loadings of .67 and .47 on DFactor1 vs. loadings less than .1 on DFactor2), while UNDERSTAND loads primarily on DFactor2 (loading of .64 on DFactor2 vs. .05 on DFactor1).

RESTRICTING LOADINGS TO ZERO

We can use the Model Tab to restrict some of these loadings to zero.

- ▷ Double click Model4 to re-open the DFactor Analysis Box for this model
- ▷ Click on Model to open the Model Tab

By default, DFactor1 is highlighted, indicating that the effects in the Included Effects box pertain to this DFactor

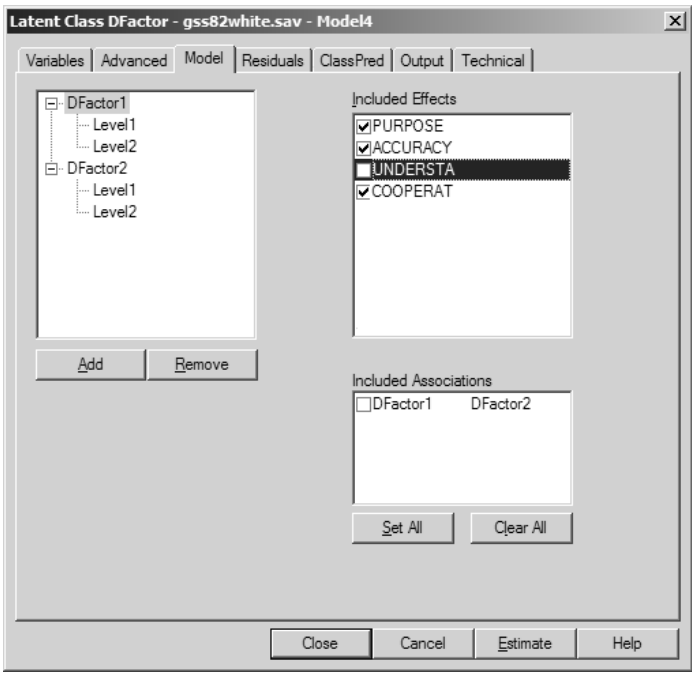


Figure 7-44. Included Effects Box

- ▷ Click the check-box preceding UNDERSTAND to set the loading on DFactor 1 to 0

To set loadings on DFactor2 to 0.

- ▷ Click DFactor2
- ▷ In the Included Effects Box, click to remove the checks for PURPOSE and ACCURACY

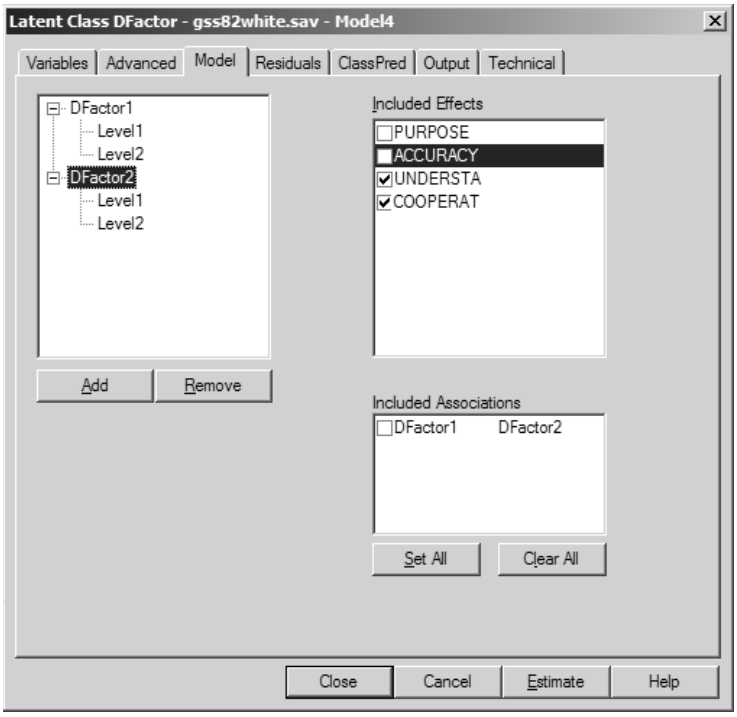
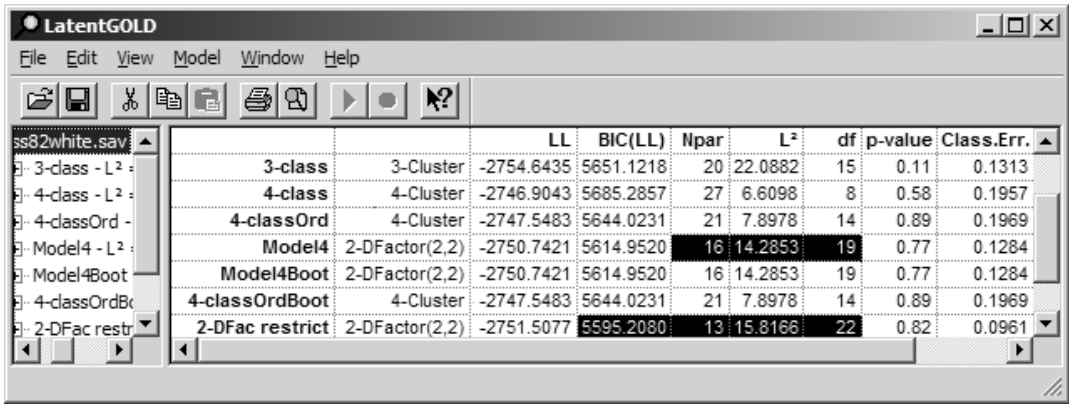


Figure 7-45. Restricting Effects to Zero

▷ `Click Estimate`

When the estimation is completed, rename the model to '2-DFac restrict'

▷ `Click the data file name in the Outline Pane`



		LL	BIC(LL)	Npar	L²	df	p-value	Class.Err.
3-class	3-Cluster	-2754.6435	5651.1218	20	22.0882	15	0.11	0.1313
4-class	4-Cluster	-2746.9043	5685.2857	27	6.6098	8	0.58	0.1957
4-classOrd	4-Cluster	-2747.5483	5644.0231	21	7.8978	14	0.89	0.1969
Model4	2-DFactor(2,2)	-2750.7421	5614.9520	16	14.2853	19	0.77	0.1284
Model4Boot	2-DFactor(2,2)	-2750.7421	5614.9520	16	14.2853	19	0.77	0.1284
4-classOrdBoot	4-Cluster	-2747.5483	5644.0231	21	7.8978	14	0.89	0.1969
2-DFac restrict	2-DFactor(2,2)	-2751.5077	5595.2080	13	15.8166	22	0.82	0.0961

Figure 7-46 Comparing Model4 to 2-Dfac restrict

Comparing the restricted model with the unrestricted 2-DFactor model, we see that the number of parameters has been reduced by 3 due to the 3 parameters that we set to zero, and L^2 increased only slightly. The restricted model is also preferred according the BIC criteria (lowest BIC).

The parameters for this model may be viewed in several different forms. We will look at the factor loadings, and the associated (marginal) conditional probabilities. To view the loadings:

▷ Click on the expand/contract icon for Parameters to make the output subcategories visible

▷ Click 'Loadings'

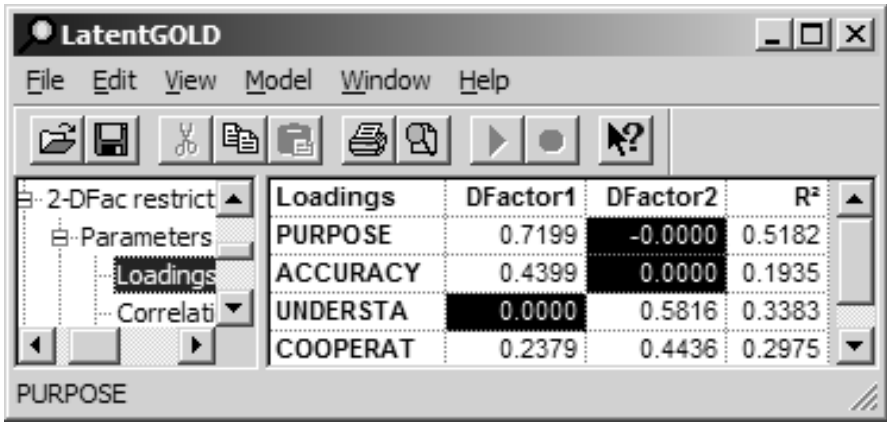


Figure 7-47. Loadings Output for Model '2-DFac restrict'

Note that the factor loadings associated with the 3 parameter restrictions are zero.

To view the model parameters as (marginal) conditional probabilities:

▷ Click on Profile

The parameters associated with each DFactor are shown in separate columns. Notice that for DFactor2, the conditional probabilities associated with PURPOSE and ACCURACY are identical for each factor level, indicating no effect. The same is true for DFactor1, regarding the effect of UNDERSTAND.

Note: This zero effect pattern would not be seen with marginal conditional probabilities if the DFactors were allowed to be correlated in the model. In the correlated situation, partial conditional probabilities would show this same pattern. (You may select Partial from the View menu to replace the marginal probabilities with partial probabilities. When the DFactors are restricted to be uncorrelated, both probability options show this zero-effect pattern).

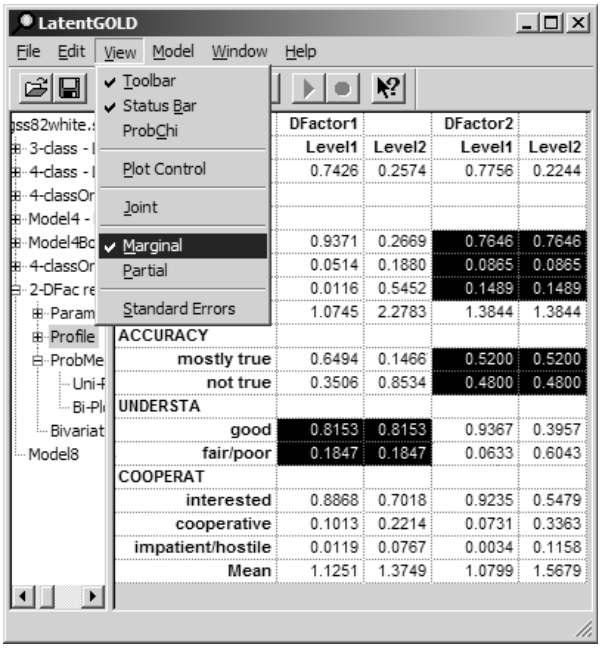


Fig. 7-48. Marginal Profile Output

VIEWING JOINT PROFILE OUTPUT FOR THE 2-DFACTOR MODEL

The Joint Profile View re-expresses the parameters in a form comparable to the corresponding cluster model. For this example, there are 4 joint categories formed by cross-tabulating the 2-DFactors, which correspond to 4 clusters.

▷ `Select Joint from the View menu`

The table now displays the Joint Profile Output

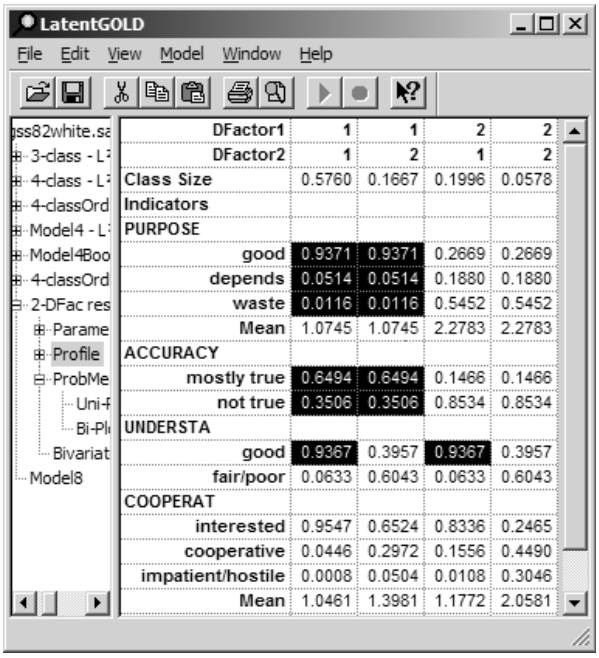


Figure 7-49. Joint Profile Output

Note the similarity between this Profile output view and the standard Profile output view obtained earlier for the '4-classOrd' cluster model (recall Fig. 7-34).

This information is plotted in the Joint View of the Profile Plot

▷ `Click the expand icon (+) to the left of Profile and click on Prf-Plot.`

The Profile Plot for the Joint Profile Output appears.

▷ `Right-click on the Plot to view the Plot Control Dialog Box.`

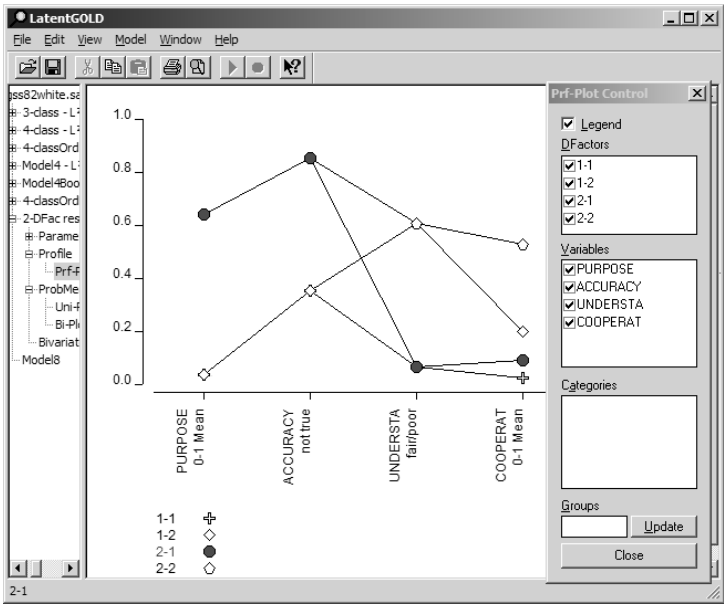
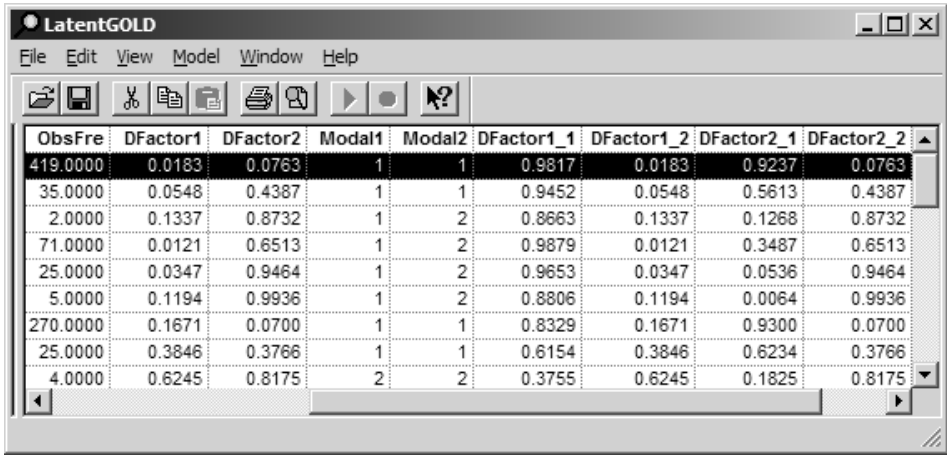


Figure 7-50. Profile Plot

You may use the Plot Control to select/ deselect the joint DFactor levels and variables to be shown on the plot.

CLASSIFYING CASES

Standard Classification Information may be requested from the Output Tab prior to estimating the model. The classification information is presented in the right-most columns, as shown in Figure 7-51 below.



ObsFre	DFactor1	DFactor2	Modal1	Modal2	DFactor1_1	DFactor1_2	DFactor2_1	DFactor2_2
419.0000	0.0183	0.0763	1	1	0.9817	0.0183	0.9237	0.0763
35.0000	0.0548	0.4387	1	1	0.9452	0.0548	0.5613	0.4387
2.0000	0.1337	0.8732	1	2	0.8663	0.1337	0.1268	0.8732
71.0000	0.0121	0.6513	1	2	0.9879	0.0121	0.3487	0.6513
25.0000	0.0347	0.9464	1	2	0.9653	0.0347	0.0536	0.9464
5.0000	0.1194	0.9936	1	2	0.8806	0.1194	0.0064	0.9936
270.0000	0.1671	0.0700	1	1	0.8329	0.1671	0.9300	0.0700
25.0000	0.3846	0.3766	1	1	0.6154	0.3846	0.6234	0.3766
4.0000	0.6245	0.8175	2	2	0.3755	0.6245	0.1825	0.8175

Figure 7-51. Standard Classification Output

For each DFactor, this information includes the posterior probability of belonging to each level of that DFactor (e.g., for DFactor 1, DFactor1_1 =.98 and DFactor1_2 = .02), and the corresponding modal levels. For example, the first row contains 419 observations with the response pattern shown in the left-most columns (not visible in Figure7-51). Using the modal assignment rule, these cases would be classified into level 1 of DFactor1 ('Modal1 = 1) and level 1 of DFactor2 ('Modal2 = 1).

DFactor scores are also provided. Assigning 0 to the first level of the DFactor, and 1 to the last, the mean score can be computed, using the corresponding posterior probabilities as weights. Thus, for cases in the first row, their scores on DFactor1 and DFactor2 are .0183 and .0763 respectively, which correspond to the posterior probability of being in level 2 of each DFactor.

This classification information will also be appended to your data file if requested from the ClassPred Tab. Such output, which also contains posterior probabilities associated with the joint DFactor is illustrated in Figure 7-54.

VIEWING THE BI-PLOT DISPLAY FOR THE 2-DFACTOR MODEL

Note that the DFactor mean scores -- DFactor1 and DFactor2 -- can be plotted in a 2-dimensional space. While plotting respondents may not be of interest, plotting Indicator categories in a bi-plot display as in Correspondence Analysis may provide useful insights. Each such category can be positioned at a point whose coordinates are aggregated mean DFactor scores obtained for all cases responding in this category. Demographics and other covariate levels can also be appended to this plot. For the exact formula for producing the bi-plot coordinates, see Section 7.4 of Technical Guide.

The DFactor mean scores are summarized in the ProbMeans output.

To view the bi-plot of this information:

- Click on the expand/contract icon for ProbMeans to make the output subcategories visible
- Click Bi-Plot
- Right click on the bi-plot to retrieve the Plot Control
- Select all the variables and click in the Lines checkbox, to connect the categories of each variable with lines.

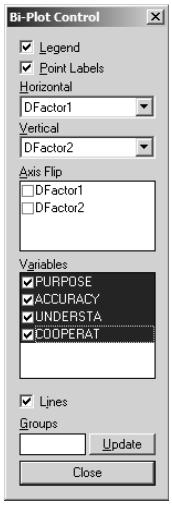


Figure 7-52. Plot Control Menu

▷ Click on the + symbol in the plot to highlight the categories of PURPOSE

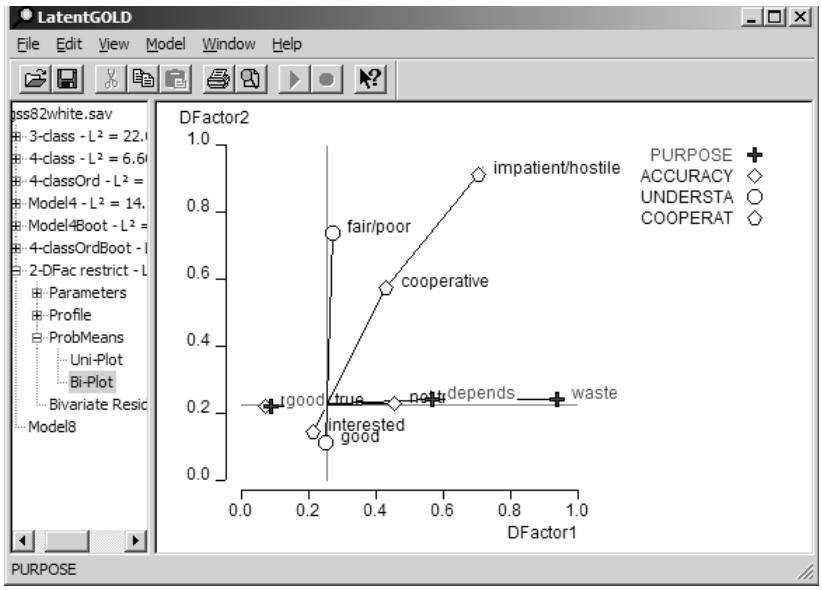
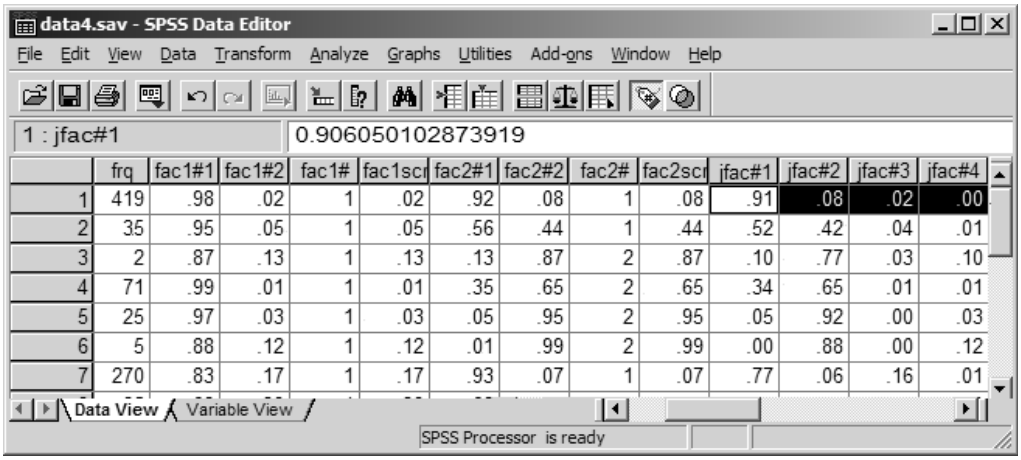


Figure 7-53. Bi-Plot

As can be seen, the categories of PURPOSE (and ACCURACY) vary along the horizontal axis associated with DFactor1 but not the vertical axis associated with DFactor2. Similarly, the categories for UNDERSTAND (denoted by the ❍ symbol), vary only with respect to the DFactor2 axis.

The bi-plot can help you interpret the DFactors and can also serve a diagnostic function prior to restricting DFactor loadings to zero by plotting any 2 DFactors to help determine what restrictions to make.

Figure 7-54 shows the standard classification output as appended to an SPSS .sav file.



	frq	fac1#1	fac1#2	fac1#	fac1scr	fac2#1	fac2#2	fac2#	fac2scr	jfac#1	jfac#2	jfac#3	jfac#4
1	419	.98	.02	1	.02	.92	.08	1	.08	.91	.08	.02	.00
2	35	.95	.05	1	.05	.56	.44	1	.44	.52	.42	.04	.01
3	2	.87	.13	1	.13	.13	.87	2	.87	.10	.77	.03	.10
4	71	.99	.01	1	.01	.35	.65	2	.65	.34	.65	.01	.01
5	25	.97	.03	1	.03	.05	.95	2	.95	.05	.92	.00	.03
6	5	.88	.12	1	.12	.01	.99	2	.99	.00	.88	.00	.12
7	270	.83	.17	1	.17	.93	.07	1	.07	.77	.06	.16	.01

Figure 7-54. Standard Classification Output appended to an SPSS .sav file

This file contains the same information shown in Figure 7-51 plus posterior probabilities associated with all DFactors jointly. For example, for the 419 observations shown in the first row of the file, jfac#1, jfac#2, jfac#3 and jfac#4 contain the posterior probabilities associated with the joint DFactor (labeled within SPSS 'Joint DFactor 1 1', 'Joint DFactor 1 2', 'Joint DFactor 2 1', 'Joint DFactor 2 2' respectively). The most likely joint level for these cases is (1,1) since the probability of being in this level is .9817.

7.3 Tutorial #3: LC Regression with Repeated Measures

DEMODATA = 'CONJOINT.SAV'

This tutorial shows how to develop Latent Class (LC) Regression models using the sample data file "conjoint.sav". You will learn how to:

- **Select the dependent variable and specify its scale type**
- **Distinguish predictors from covariates**
- **Impose restrictions on the predictor effects**
- **Specify covariates as active or inactive**
- **Determine the number of latent classes (i.e., segments)**
- **Examine R^2 and various other information related to model prediction**

In addition, this example illustrates several advanced options in the LC Regression Module. You will learn how to:

- **Use the optional case ID variable to specify repeated observations**
- **Explore the Profile and ProbMeans output**
- **Use demographic variables as covariates to predict segment membership**
- **Obtain predictions based solely on the covariates**
- **Classify cases into latent segments**

The Data

The data for this example are obtained from a hypothetical conjoint marketing study involving repeated measures where respondents were asked to provide likelihood of purchase ratings under each of several different scenarios. A partial listing of the data is shown in Figure 7-55.

CONJOINT.SAV - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities AmosApps Window Help

1 : id 1

	id	sex	age	fashion	quality	price	rating
1	1	Male	25-39	Traditional	Low	Higher	Very Unlikely
2	1	Male	25-39	Traditional	Low	Lower	Neutral
3	1	Male	25-39	Traditional	High	Higher	Neutral
4	1	Male	25-39	Traditional	High	Lower	Very Likely
5	1	Male	25-39	Modern	Low	Higher	Somewhat Unlikely
6	1	Male	25-39	Modern	Low	Lower	Somewhat Unlikely
7	1	Male	25-39	Modern	High	Higher	Very Likely
8	1	Male	25-39	Modern	High	Lower	Very Likely
9	2	Female	16-24	Traditional	Low	Higher	Somewhat Unlikely
10	2	Female	16-24	Traditional	Low	Lower	Neutral
11	2	Female	16-24	Traditional	High	Higher	Very Likely
12	2	Female	16-24	Traditional	High	Lower	Very Likely

Data View Variable View

SPSS Processor is ready

Figure 7-55: Partial Listing of Conjoint Data

As suggested in Figure 7-55, there are 8 records for each case (there are 400 cases in total); one record for each cell in this 2x2x2 complete factorial design of different scenarios for the purchase of a product:

- FASHION (1 = Traditional; 2 = Modern)
- QUALITY (1 = Low; 2 = High)
- PRICE (1 = Lower; 2 = Higher)

The dependent variable (RATING) is a rating of purchase intent on a five-point scale. The three attributes listed above will be used as predictor variables in the model.

We will also include the two demographic variables as covariates, in a second model.

- SEX (1 = Male; 2 = Female)
- AGE (1 = 16-24; 2 = 25-39; 3 = 40+).

The Goal

Use Latent GOLD to identify latent segments differing with respect to the estimate of importance attached to each of the three attributes, which influence an individual's purchase decision. The LC regression model allows for the fact that these estimates may differ for different segments. That is, for one segment, price and only price may influence the decision, while a second segment may be influenced by quality and modern appearance, but is price insensitive. We will treat RATING as an ordinal dependent variable and compare several models to determine the number of segments. We will then show how to describe the demographic differences between these segments and to classify each respondent into that segment which is most likely.

Estimating an LC Regression Model

OPENING A DATA FILE AND SELECTING THE TYPE OF MODEL

For this example, the data file being used is an SPSS system file.

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 .sav extensions appear in the list (see Figure 7-56).

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

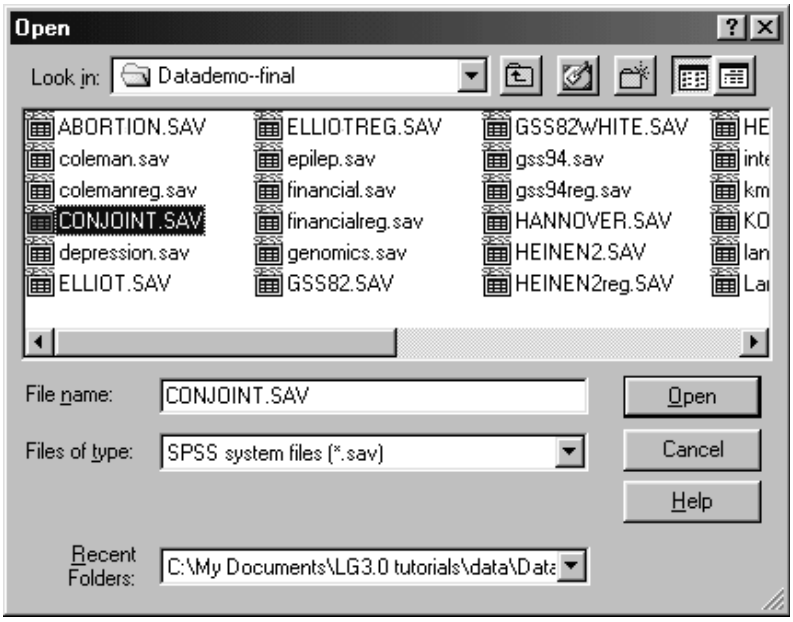


Figure 7-56: Open Dialog Box

▷ Select conjoint.sav and click Open to open the Viewer window.

▷ Highlight 'Model1' if it is not already highlighted.

▷ Right click to open the Model Selection menu (you may also double click the model name to open this menu or select the type of model from the Model Menu).

▷ Select Regression and the LC Regression analysis dialog box, which contains 3 tabs, will open (see Fig. 7-57).

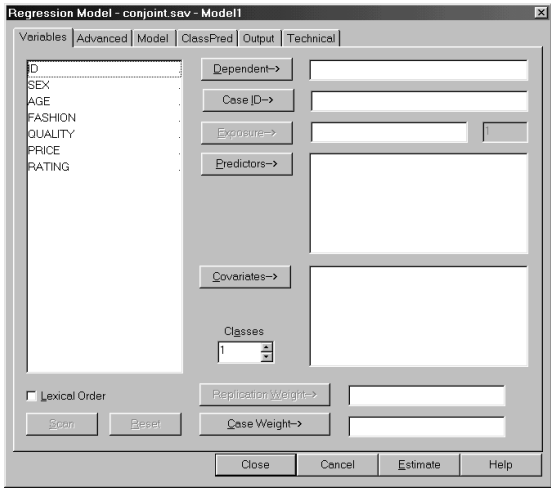


Figure 7-57: Analysis Dialog Box for LC Regression Model

SELECTING THE VARIABLES FOR THE ANALYSIS

For this analysis, RATING will be the dependent variable.

▷ Select RATING in the Variables List and click Dependent to move the variable to the Dependent box.

We also need to indicate the dependent variable scale type. For this example, we will use the default scale type (Ordinal-Fixed) which takes into account the natural ordering between the 5 levels of purchase intent. By default, the fixed scores on the data file (1, 2, 3, 4 and 5) are used which order the levels and establish equal distance between adjacent levels.

As explained above, the data contains repeated observations for each respondent (case). Therefore, we need to indicate which records belong to each case. This is accomplished using a Case ID variable, which contains a unique identification number for each case. All records belonging to the same case are assigned the same unique ID.

▷ Select ID in the Variables list and click Case ID to move the variable into the Case ID box.

Next, we will select the Predictors. Predictors are used as independent variables in the regression model. In the current example, we use the product attributes FASHION, QUALITY and PRICE as predictors.

▷ Select FASHION, QUALITY and PRICE in the Variables list and click Predictors to move the variables into the Predictors box.

SPECIFYING THE NUMBER OF CLASSES

The LC regression model simultaneously estimates a separate regression model for each class. A 1-class model estimates only a single regression model. It makes the standard homogeneity assumption that a single regression model holds true for all cases. In the current example, we will start by estimating a 1-class model and obtain a log-likelihood statistic to be used as a base. We will then estimate additional models, which successively increment the number of classes by 1 and assess the significance of each additional class.

One assessment consists of a check of whether the change in the log-likelihood for each pair of successive models fails to decrease by a significant amount as determined by the BIC statistic. (The model having the lowest BIC might then be selected.) A second assessment is to utilize the p-value associated with the L^2 fit statistic.

▷ In the box titled Classes (located below the Covariates pushbutton) type '1-4' to request the estimation of 4 different LC Regression models - a 1-class model, a 2-class model, a 3-class model and a 4-class model.

SCANNING THE DATA FILE

▷ Click Scan (located in the lower left of the Analysis dialog box) to scan the data file.

The number of distinct categories (or values) along with the scaling option appears next to each variable in the Dependent and Predictors boxes.

To view category labels, frequency counts and any scores assigned to any scanned variable, double click on the variable name in the Dependent or Predictor list box. The Variables dialog box will open (see Figure 7-58).

▷ Double click the dependent variable, RATING.

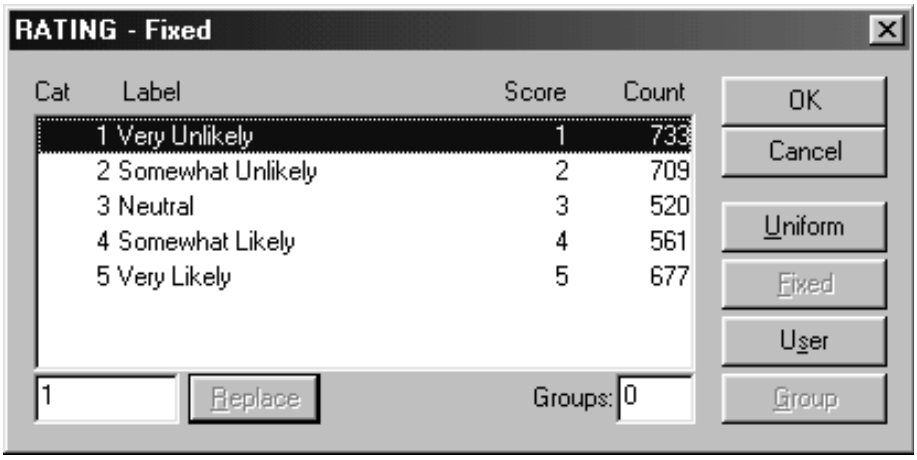


Figure 7-58: Variables box for RATING

▷ Click OK to close the Variables dialog box and return to the Regression Analysis dialog box.

ESTIMATING THE MODEL

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

Your analysis dialog box should look like Figure 7-59.

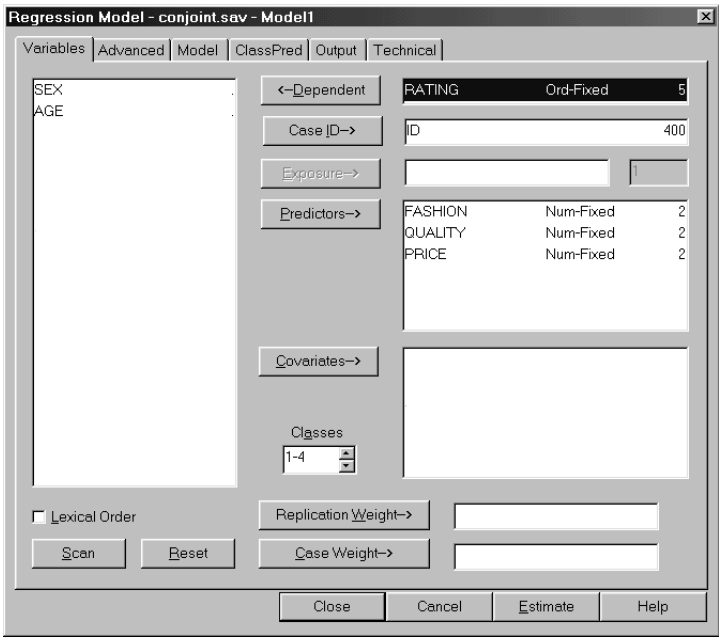


Figure 7-59: Regression Analysis Dialog Box with Initial Settings

▷ Click Estimate (located at the bottom right of the Analysis dialog box).

Viewing Output and Interpreting Results

For LC Regression models, several output files are produced. To view a summary of the models estimated (Figure 7-60),

▷ Click on the data file name, conjoint.sav in the Outline pane.

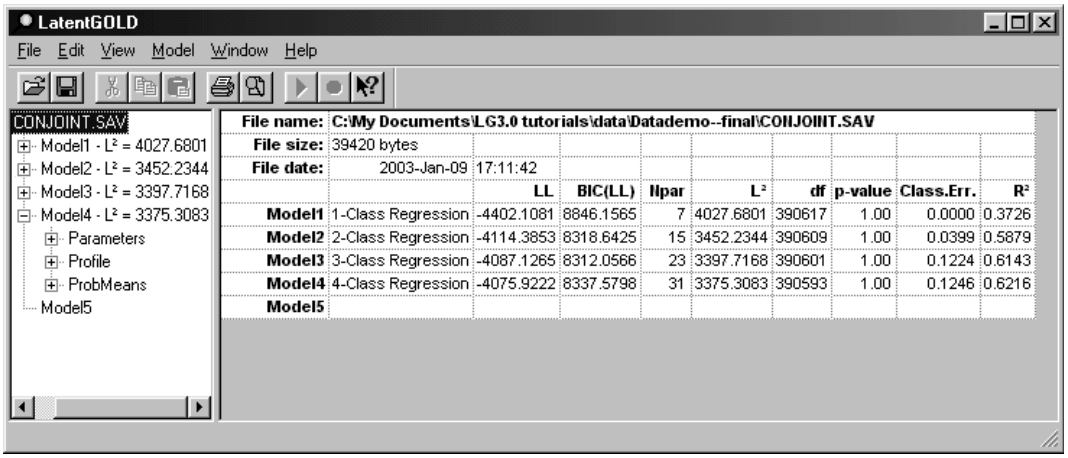


Figure 7-60: Summary of Models Estimated

This output reports statistics that will assist you in determining the correct number of classes -- the log-likelihood (LL) values, the BIC values, and the number of parameters in the estimated models. It is important to determine the right number of classes because specifying too few ignores class differences, while specifying too many may cause the model to be unstable. While the log-likelihood increases each time the number of classes is increased, the minimum BIC value occurs for Model3, suggesting that the 3-class solution is the best of the four estimated models. The R^2 increases from .37 for the 1-class model to .61 for the 3-class regression.

Note 1: Occasionally, you might obtain a local (suboptimal) solution. For these data, it is possible to obtain a local solution for the 4-class model, obtaining LL = -4080.318 instead of -4075.922. If this occurs, click Estimate to re-estimate the 4-class model (see section 6.6 in the Technical Guide for a discussion of preventing local solutions).

Note 2: Notice that the p-values based on the model L^2 and reported degrees of freedom (df) are 1.0 under each model. These are not valid assessments of fit because we are dealing with sparse data.

We will now examine the detailed output for the 3-class solution.

PROFILE OUTPUT

- ▷ Rename Model3 to "3-class" by clicking on its name.
- ▷ Click the + icon next to '3-class' in the Outline pane to expand the listing of output for this model.
- ▷ Click Profile.

The Profile output contains information on the class sizes, the class-specific (marginal) probabilities and means of the dependent variable (see Figure 7-61).

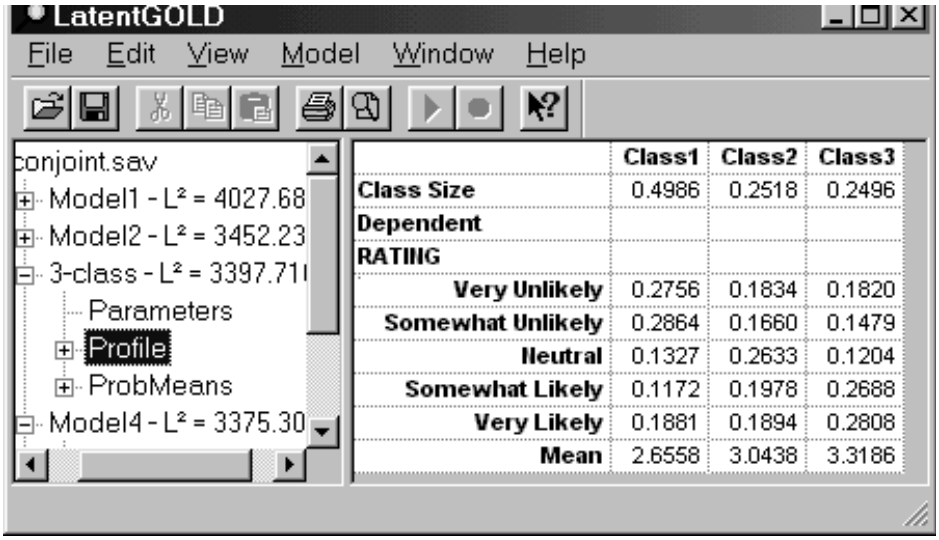


Figure 7-61: Profile Output for 3-Class Model

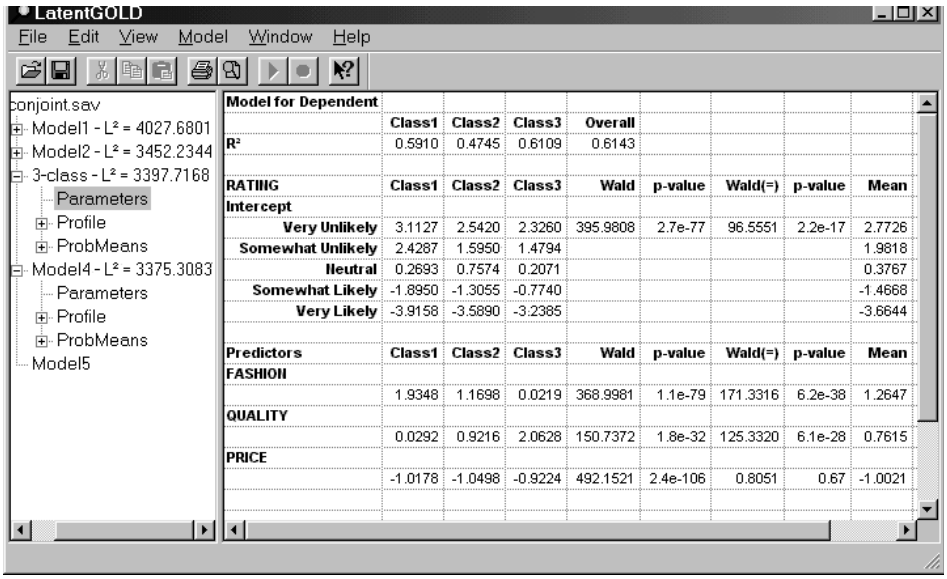
The classes are always ordered from high to low according to their size. It can be seen from the first row of the table that segment 1 contains about 50% of the subjects (.4986), segment 2 contains about 25% and segment 3 contains the remaining 25%.

Examination of class-specific probabilities shows that overall, segment 1 is least likely to buy (only 27.56% are Very Unlikely to buy) and segment 3 is most likely (28.08% are Very Likely to buy). Later in this tutorial, we will show how to classify each case into the most appropriate segment.

PARAMETERS OUTPUT

Next, we will view the Parameters output (see Figure 7-62).

▷ For the '3-class' model, click Parameters.



Model for Dependent								
	Class1	Class2	Class3	Overall				
R^2	0.5910	0.4745	0.6109	0.6143				
RATING	Class1	Class2	Class3	Wald	p-value	Wald(=)	p-value	Mean
Intercept								
Very Unlikely	3.1127	2.5420	2.3260	395.9808	2.7e-77	96.5551	2.2e-17	2.7726
Somewhat Unlikely	2.4287	1.5950	1.4794					1.9818
Neutral	0.2693	0.7574	0.2071					0.3767
Somewhat Likely	-1.8950	-1.3055	-0.7740					-1.4668
Very Likely	-3.9158	-3.5890	-3.2385					-3.6644
Predictors	Class1	Class2	Class3	Wald	p-value	Wald(=)	p-value	Mean
FASHION								
	1.9348	1.1698	0.0219	368.9981	1.1e-79	171.3316	6.2e-38	1.2647
QUALITY								
	0.0292	0.9216	2.0628	150.7372	1.8e-32	125.3320	6.1e-28	0.7615
PRICE								
	-1.0178	-1.0498	-0.9224	492.1521	2.4e-106	0.8051	0.67	-1.0021

Figure 7-62: Parameters Output for 3-Class Model

The beta parameter for each predictor is a measure of the influence of that predictor on RATING. The beta effect estimates under the column labeled Class 1 suggest that segment 1 is influenced in a positive way by products for which FASHION = Modern (beta = 1.9348) and PRICE = Higher (beta = 1.0178), but not by QUALITY (beta is approximately 0). We also see that segment 2 is influenced by all 3 attributes, having a preference for those product choices that are modern (beta = 1.1698) and higher quality (beta = .9216), but their preference decreases as a function of price (beta = 1.0498). Members of segment 3 prefer higher quality products (beta = 2.0628), but their preference decreases as a function of price (beta = .9224), and they are not influenced by FASHION.

Note that PRICE has more or less the same influence on all three segments. The Wald (=) statistic indicates that the differences in these beta effects across classes are not significant (the p-value = .67 which is much higher than .05, the standard level for assessing statistical significance). This means that all 3 segments exhibit price sensitivity to the same degree. This is confirmed when we estimate a model in which this effect is specified to be class-independent (see next section). The p-value for the Wald statistic for PRICE is 2.4x10-106 indicating that the amount of price sensitivity is highly significant.

With respect to the effect of the other two attributes we find large between-segment differences. The predictor FASHION has a strong influence on segment 1, a less strong effect on segment 2, and virtually no effect on seg-

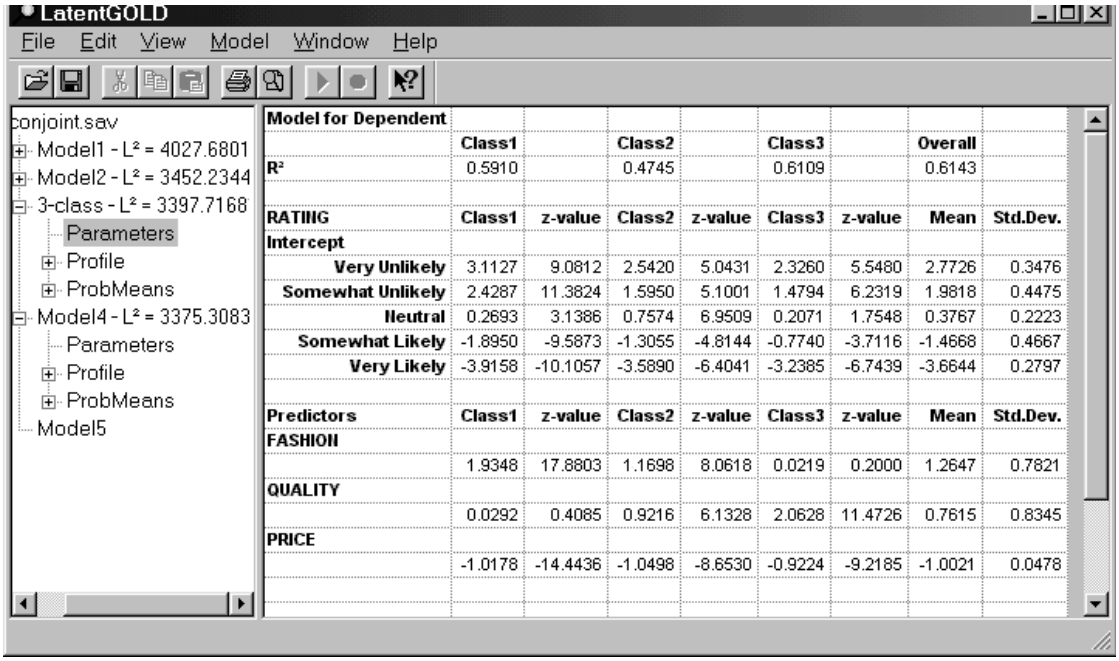
ment 3. QUALITY has a strong effect on segment 3, a less strong effect on segment 2, and virtually no effect on segment 1. The fact that the influence of FASHION and QUALITY differs significantly between the 3 segments is confirmed by the significant p-values associated with the Wald(=) statistics for these attributes. For example, for FASHION, the p-value = 6.2x10-36.

In summary, segment 1 could be labeled the "Fashion-Oriented Segment", segment 3 the "Quality-Oriented Segment", and segment 2 is the segment that takes into account all 3 attributes in their purchase decision.

To test each individual class-specific beta for statistical significance we can append standard errors, Z-statistics, or both to the output.

Right click on the output in the Contents Pane and choose Z Statistic:

The Wald statistics are replaced by the Z-statistics for the betas. Notice that the absolute values of the z-score associated with QUALITY for class 1 and with FASHION for class 3 fall under 2, and hence are not significant at the .05 level.



Model for Dependent								
	Class1		Class2		Class3		Overall	
R^2	0.5910		0.4745		0.6109		0.6143	
RATING	**Class1**	**z-value**	**Class2**	**z-value**	**Class3**	**z-value**	**Mean**	**Std.Dev.**
Intercept								
Very Unlikely	3.1127	9.0812	2.5420	5.0431	2.3260	5.5480	2.7726	0.3476
Somewhat Unlikely	2.4287	11.3824	1.5950	5.1001	1.4794	6.2319	1.9818	0.4475
Neutral	0.2693	3.1386	0.7574	6.9509	0.2071	1.7548	0.3767	0.2223
Somewhat Likely	-1.8950	-9.5873	-1.3055	-4.8144	-0.7740	-3.7116	-1.4668	0.4667
Very Likely	-3.9158	-10.1057	-3.5890	-6.4041	-3.2385	-6.7439	-3.6644	0.2797
Predictors	**Class1**	**z-value**	**Class2**	**z-value**	**Class3**	**z-value**	**Mean**	**Std.Dev.**
FASHION								
	1.9348	17.8803	1.1698	8.0618	0.0219	0.2000	1.2647	0.7821
QUALITY								
	0.0292	0.4085	0.9216	6.1328	2.0628	11.4726	0.7615	0.8345
PRICE								
	-1.0178	-14.4436	-1.0498	-8.6530	-0.9224	-9.2185	-1.0021	0.0478

Figure 7-63: Parameters Output with Z-values

Restricting Certain Effects to be Zero or Class Independent

In the Parameters output above we saw that the beta estimates associated with PRICE are approximately equal for all 3 classes. To test the null hypothesis of equality, we used the Wald(=) statistic. The low value of .67 was too small to reject this null hypothesis. We also showed that 2 of the betas were not significantly different from 0. We will now show how to obtain a more parsimonious model by imposing zero restrictions on the 2 betas and by restricting the betas associated with PRICE to be equal across segments. This is accomplished using the Model Tab.

To specify the PRICE effects to be class independent,

- Double click '3-class' to open the Analysis Dialog Box for this model
- Click on the Model Tab
- In the row for PRICE, right click on the Class Independent column and select 'Yes'.

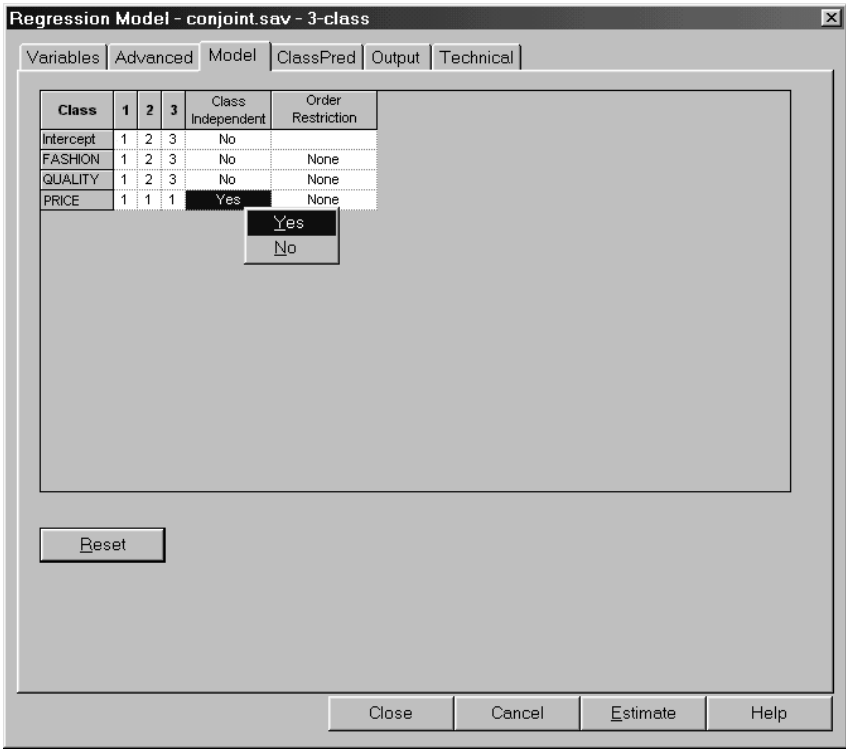


Figure 7-64. Setting the Class-Independent effects for the variable PRICE

To specify the betas to be zero,

- Right click on the cells corresponding to the betas to be set to zero
- Select No Effect

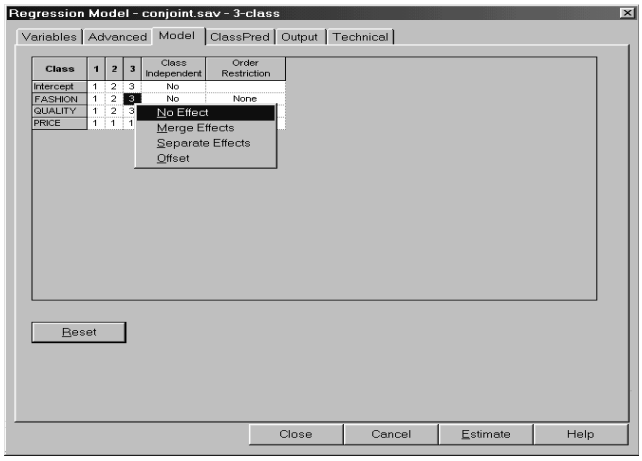


Figure 7-65. Specifying betas to 0 for variable FASHION

▷ Click on Variables to return to the Variables Tab.

An '=' will appear to the left of the variable PRICE to indicate the class independent restriction.

▷ Click Estimate to re-estimate this model.

The output for this new model will be listed as Model5.

VIEWING OUTPUT AND INTERPRETING RESULTS

To view the summary output,

▷ Click on the data file name, conjoint.sav in the Outline pane.

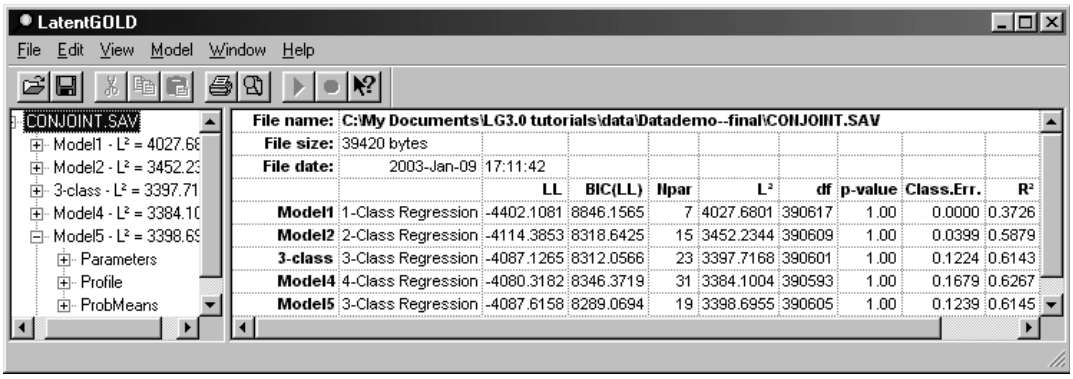


Figure 7-66: Summary Output

The new model estimated has been added to the bottom of the list as Models 5. For Model 5, the fit is almost identical to Model 3 and we obtain a lower (better) BIC value.

PARAMETERS OUTPUT

- Click Parameters for Model5 to view the results containing the desired restrictions.

For PRICE, which we specified as class independent, note that the Wald(=) is now zero because the betas have been restricted to be exactly equal to each other across classes.

Adding Covariates

There is one important topic left with respect to the specification of LC regression models; that is, the use of covariates. In the Covariates list box, we can specify variables that we want to use to predict class membership. For this example we will re-estimate the 3-class model, this time including SEX and AGE as covariates.

To estimate the model specifying SEX and AGE as covariates,

- In the Outline pane, double click Model5 to open the Analysis dialog box for this model. Latent GOLD has maintained our previous settings (we will keep PRICE set as class independent).
- Select SEX and AGE in the Variables list box.
- Click Covariates to move these variables to the Covariates box.
- Right click on the variable names and select scale type Nominal.

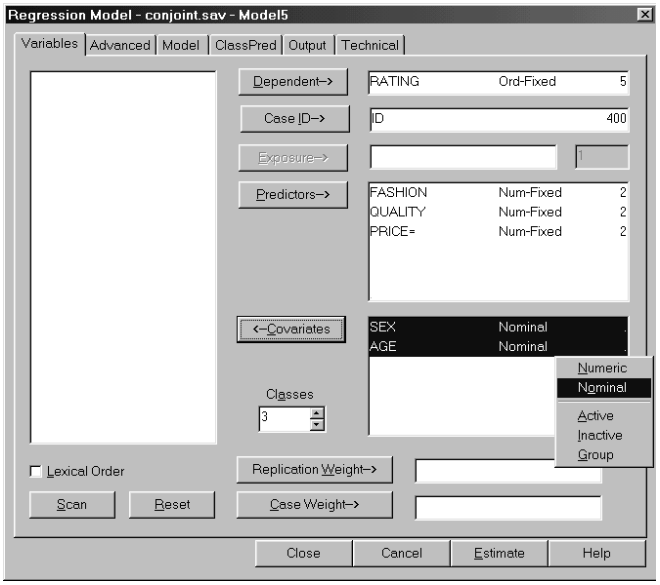


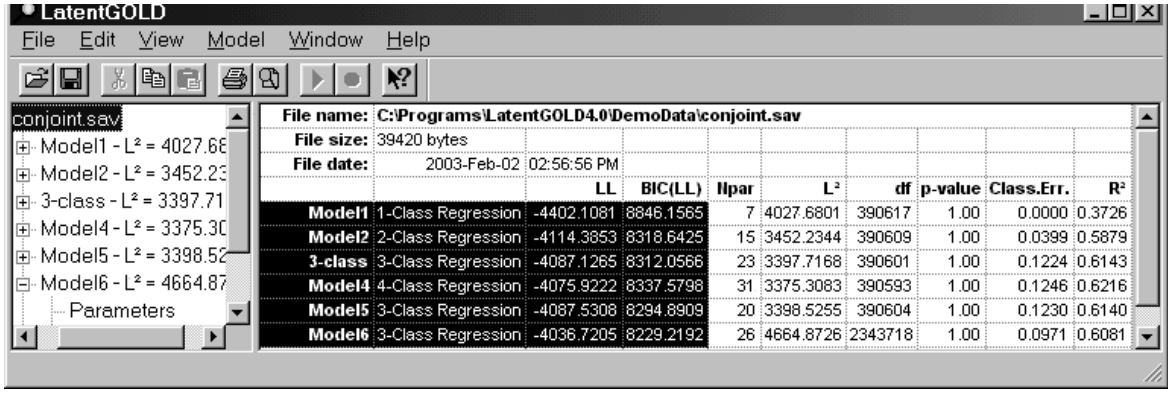
Figure 7-67: Specification of Model with Covariates

▷ Click Estimate. The output for this new model will be listed as Model6.

VIEWING OUTPUT AND INTERPRETING RESULTS

To view the summary output,

▷ Click on the data file name, conjoint.sav in the Outline pane.



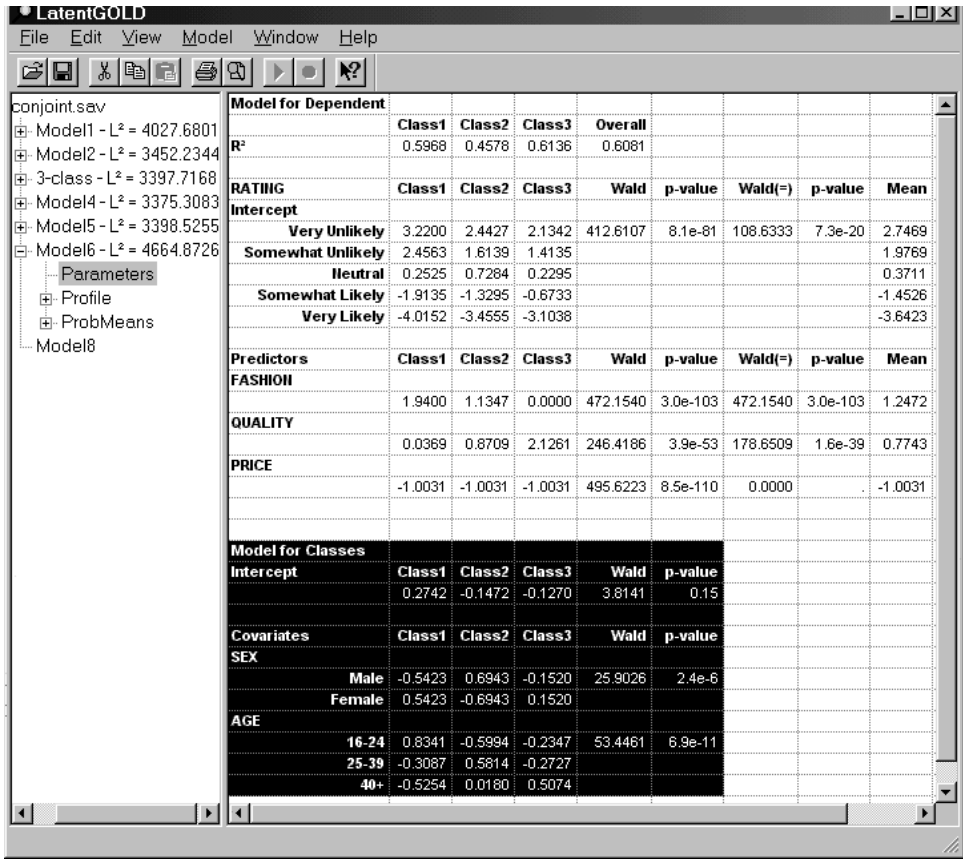
		LL	BIC(LL)	Npar	L²	df	p-value	Class.Err.	R²
Model1	1-Class Regression	-4402.1081	8846.1565	7	4027.6801	390617	1.00	0.0000	0.3726
Model2	2-Class Regression	-4114.3853	8318.6425	15	3452.2344	390609	1.00	0.0399	0.5879
3-class	3-Class Regression	-4087.1265	8312.0566	23	3397.7168	390601	1.00	0.1224	0.6143
Model4	4-Class Regression	-4075.9222	8337.5798	31	3375.3083	390593	1.00	0.1246	0.6216
Model5	3-Class Regression	-4087.5308	8294.8909	20	3398.5255	390604	1.00	0.1230	0.6140
Model6	3-Class Regression	-4036.7205	8229.2192	26	4664.8726	2343718	1.00	0.0971	0.6081

Figure 7-68: Summary Output

The new model has been added to the bottom of the list as Model 6. Note that this model is even better than our previous 3-class models as indicated by the lower BIC value.

PARAMETERS OUTPUT

▷ Click Parameters for Model6 to view the results reported in Figure 7-69.



Model for Dependent

	Class1	Class2	Class3	Overall
R^2	0.5968	0.4578	0.6136	0.6081

RATING	Class1	Class2	Class3	Wald	p-value	Wald(=)	p-value	Mean
Intercept								
Very Unlikely	3.2200	2.4427	2.1342	412.6107	8.1e-81	108.6333	7.3e-20	2.7469
Somewhat Unlikely	2.4563	1.6139	1.4135					1.9769
Neutral	0.2525	0.7284	0.2295					0.3711
Somewhat Likely	-1.9135	-1.3295	-0.6733					-1.4526
Very Likely	-4.0152	-3.4555	-3.1038					-3.6423

Predictors	Class1	Class2	Class3	Wald	p-value	Wald(=)	p-value	Mean
FASHION								
	1.9400	1.1347	0.0000	472.1540	3.0e-103	472.1540	3.0e-103	1.2472
QUALITY								
	0.0369	0.8709	2.1261	246.4186	3.9e-53	178.6509	1.6e-39	0.7743
PRICE								
	-1.0031	-1.0031	-1.0031	495.6223	8.5e-110	0.0000	.	-1.0031

Model for Classes

Intercept	Class1	Class2	Class3	Wald	p-value
	0.2742	-0.1472	-0.1270	3.8141	0.15

Covariates	Class1	Class2	Class3	Wald	p-value
SEX					
Male	-0.5423	0.6943	-0.1520	25.9026	2.4e-6
Female	0.5423	-0.6943	0.1520		
AGE					
16-24	0.8341	-0.5994	-0.2347	53.4461	6.9e-11
25-39	-0.3087	0.5814	-0.2727		
40+	-0.5254	0.0180	0.5074		

Figure 7-69: Parameters Output for 3-Class Model with Covariates

First, note that the beta parameter estimates for the 3-class model with covariates (see Figure 7-69) are similar to those in the original 3-class model (see Figure 7-61).

The gamma parameters of the model for the latent distribution appear at the bottom of the Parameters output in Figure 7-69 under the heading 'Model for Classes'. The p-values associated with the Wald statistic shows that overall, both effects are significant. The betas associated with SEX = Female (0.5423, -0.6943, 0.1520) suggest that females are more likely than males of belonging to the "Fashion-Oriented Segment (segment 1), and much less likely to belong to segment 2. The AGE effects show that the youngest age group is more likely than other respondents to be in the "Fashion-Oriented Segment" while the oldest age group is more likely to be in the "Quality-Oriented Segment".

CLASSIFICATION OUTPUT

To obtain the Classification output, you need to specify it as an option in the Output Tab before estimating your model.

To view the standard classification output for Model 6,

- Double click on Model6 in the Outline pane to open the Analysis dialog box for this model.

▷ Click on the Output Tab.

▷ Click in the checkbox next to 'Standard Classification' and 'Covariate Classification' to select this output.

▷ Click Estimate to re-estimate the model.

The new model estimated has been added to the bottom of the list as Model 7 (it is the same as Model6 except for the additional output file selected).

▷ In the Outline pane, for Model7, click Standard Classification.

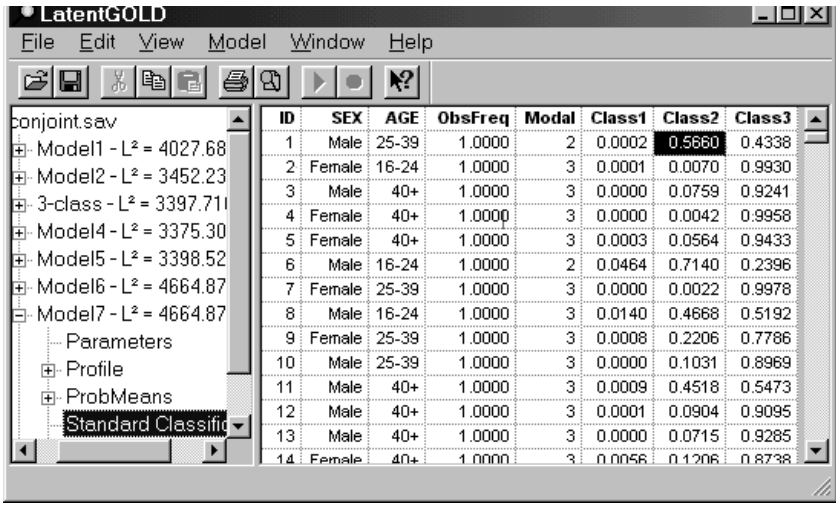


Figure 7-70: Classification Output for Model 7 (partial listing)

We can see that the first respondent (ID=1) would be classified into segment 2, since segment 2 has the highest membership probability (.5660) for that respondent (see Figure 7-70).

(For information on how to append classification scores to you original data file, see Step 9 in Chapter 5.)

Now, suppose that you wish to classify new data into the appropriate classes but you only had covariate information on these cases. You would use the Covariate Classification output for this purpose.

To view this output, click on Covariate Classification.

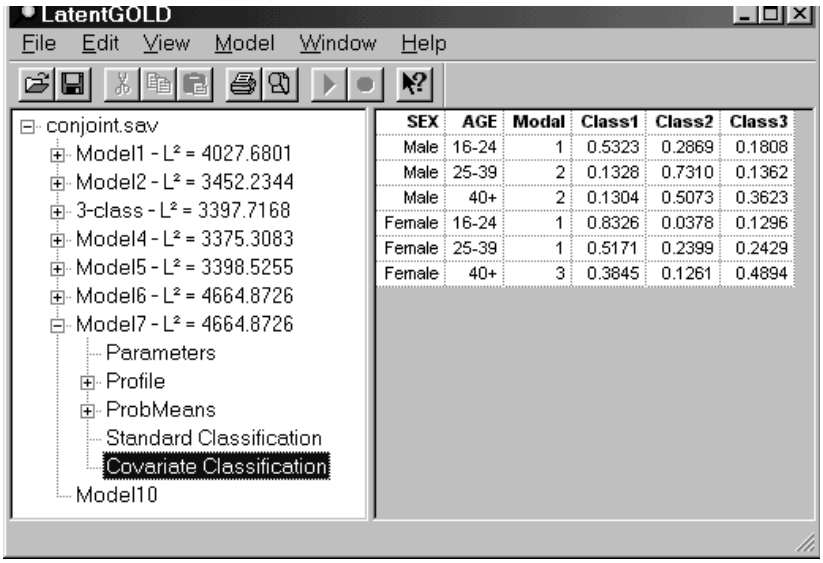


Figure 7-71. Covariate Classification Output for Model7

7.4 Tutorial #4: Profiling LC Segments Using the CHAID Option

DEMODATA = 'GSS82.SAV'

After an LC model is estimated, it is often desirable to describe (profile) the resulting latent classes in terms of demographic and/or other exogenous variables (covariates). Traditionally, a 2-step approach has been used to do this. In step 1, cases are scored by appending the Standard Classification output to a data file. The ClassPred Tab is used to do this. In step 2, cross-tabulation, regression, discriminant analysis or some other procedure is used to relate the modal classifications to the covariates.

The disadvantage of modal classifications is that they contain misclassification error which biases the relationship between the covariates and the true (latent) classes. This bias in the cross-tabulations can be eliminated through the use of posterior membership probabilities instead of the modal assignments to construct the tables, which take into account the uncertainty of the classification. In this tutorial, two options for attaining such bias-free profiles are illustrated:

1) Inclusion of Inactive Covariates in a model

Since no additional parameters are estimated when covariates are specified as Inactive, any number of inactive covariates can be included in a model with only a modest increase in the model estimation time. When inactive covariates are included in a model, column and row percentages showing the relationship of such to the latent classes appear in the Profile and ProbMeans output tables respectively. In DFactor models, tables relate the covariates to the levels of each DFactor separately, as well as to the levels of the joint DFactor.

2) Use of the CHAID option (requires the SI-CHAID 4.0 program)

The CHAID (CHi-squared Automatic Interaction Detector) analysis option can be used to assess the statistical significance of each Covariate in its relationship to the latent classes, as well as to develop detailed profiles of these classes, based on the relationships in 3- and higher-way tables. For example, in this tutorial, a CHAID analysis shows that while RACE and EDUCATION are both significantly related to the levels of DFactor2, once the education effect is taken into account, the race effect is no longer significant. Thus, the relationship between RACE and DFactor2 may be spurious, explained by the fact that the blacks in the sample had lower education levels than the whites. As such, the differences between levels 1 and 2 of DFactor2 may simply be interpreted as educational differences.

The Goal

In this tutorial, we obtain further insights into the latent class segments obtained from tutorials #1 and #2 using additional variables (covariates) to profile these segments in terms of respondent demographics - gender (SEX), education (EDUCR), marital status (MARITAL), and age (AGE).

This tutorial illustrates:

- **Use of 'inactive' covariates feature to describe LC segments**
- **Use of the SI-CHAID add-on program to obtain additional descriptive profiles and tests of significance**

In addition, it illustrates

- **Use of the Grouping option to reduce the number of categories of a variable**

INCLUDING COVARIATES IN THE MODELS

It is possible to examine the relationship between exogenous variables and LC segments obtained from LC Cluster, DFactor and LC Regression models, by specifying the exogenous variables as active or inactive covariates. In this tutorial, we focus on the inactive covariate feature and LC segments obtained from a DFactor model.

We will be using a case level data file called gss82.sav which contains covariates on N1 = 1,198 of the 1,202 white respondents used in tutorials #1 and #2 and a supplemental sample of N2 = 446 black respondents.

Opening the Data File

For this example, the data file is in the SPSS system file (.sav) 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.

▷ Select gss82.sav and click Open to open the Viewer window

▷ Right click 'Model1' in the Outline Pane to open the Model Selection menu (you may also double click the model name to open this menu or select the type of model from the Model Menu), and select DFactor from the pop-up menu

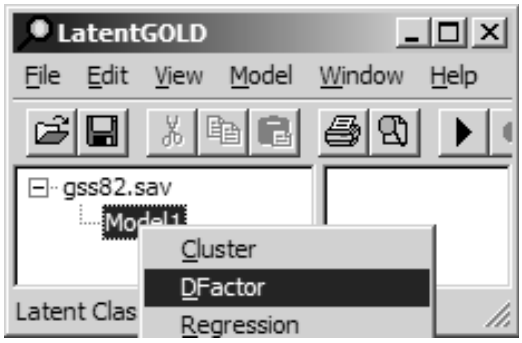


Figure 7-72. Selecting the DFactor Model

The DFactor Analysis Dialog Box will open.

SELECTING THE VARIABLES FOR THE ANALYSIS

For this analysis, we will be using the 4 variables as indicators (PURPOSE, ACCURACY, UNDERSTA, COOPERAT) as in our earlier tutorials. To select the indicator variables:

- Select PURPOSE, ACCURACY, UNDERSTA, COOPERAT in the Variables list
- Click Indicators to move them to the Indicator list box.

These variables now appear in the Indicators list box.

SPECIFYING THE NUMBER OF DFACTORS

To specify a 2-DFactor model as in Tutorial #2:

- In the Variables Tab, in the box titled DFactors select or type '2'.

INCLUDING COVARIATES

To include the demographic variables as covariates

- Select RACE, SEX, EDUCR, MARITAL, AND AGE in the Variables list.
- Click Covariates to move them to the Covariates Box.

To scan the file

- Click Scan

Following the Scan, the number of levels is reported to the right of the variable names. Note for example in Figure 7-73 that AGE shows 72 levels.

To make the covariates Inactive so that they do not influence the estimation of the model parameters

- Select RACE, SEX, EDUCR, MARITAL, AND AGE in the Covariates list box.
- Right click to retrieve the covariate scale type menu

Your Analysis Dialog Box should now look like this:

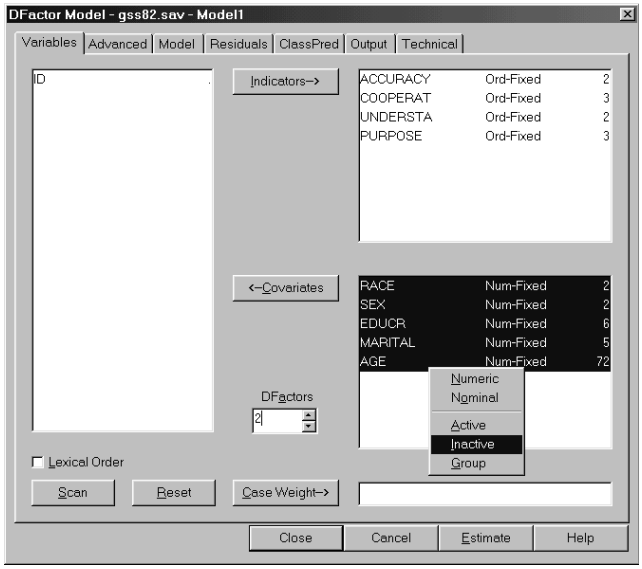


Figure 7-73. Making Covariates Inactive

- Select Inactive

The symbol < I > now appears to the right of each covariate name to indicate the Inactive setting.

Change the scale type for MARITAL to Nominal, and for improved table formatting, do the same for the dichotomous variables:

- Select ACCURACY and UNDERSTA
- Right click to retrieve the Indicators scale type menu
- Select Nominal
- Select RACE, SEX and MARITAL
- Right click retrieve the Covariate scale type menu
- Select Nominal
- Click Scan again

Your Analysis Dialog Box should now look like this:

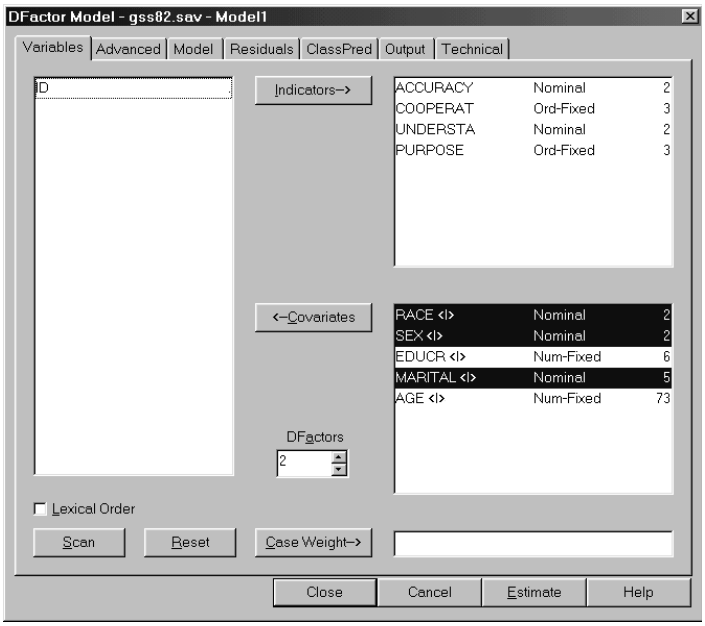


Figure 7-74. Analysis Dialog Box after adding covariates

Note that after scanning the file again, the number of levels for AGE changes from 72 (recall Figure 7-73) to 73. The last (73rd) level now contains the 8 cases for which AGE is missing. (Prior to making the Covariates Inactive, the default treatment for missing values was to exclude the 8 cases from the analysis during the Scan.)

A limitation of SI-CHAID is that variables can have no more than 31 levels. SI-CHAID automatically reduces the number of levels to 15 for variables exceeding this limit. Here, we will illustrate the Group option in Latent GOLD to reduce the number of levels of AGE.

To open the Grouping and Recoding Dialog Box

▷ Double click on AGE

Figure 7-75 shows that 6 cases are at the first age level, 18 years of age; 31 cases are aged 19; 22 cases are aged 20; and so on.

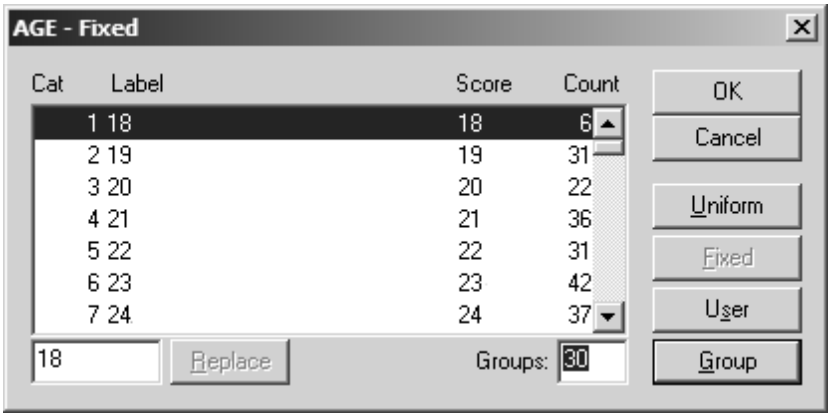


Figure 7-75. Levels for the variable AGE

To reduce the number of levels to 30

▷ Enter 30 in the Groups box

▷ Click the Group button

The result is a 'grouped AGE' variable.

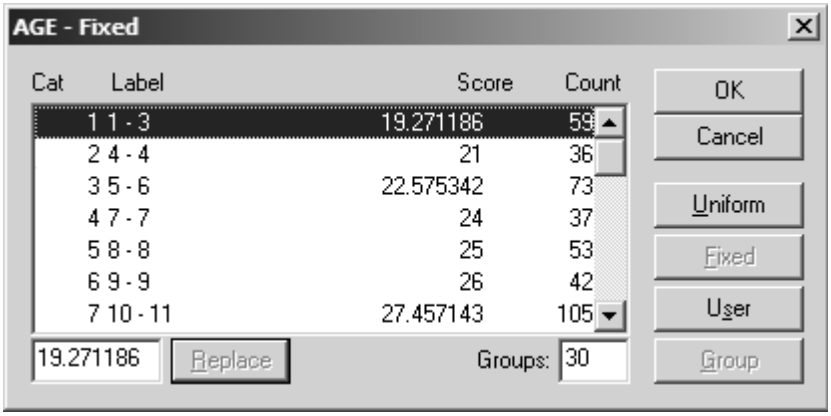


Figure 7-76. Grouped AGE variable

The new grouped level 1, labeled : '1-3' is comprised of the first 3 original age levels. From Figure 7-76, we see that this new age group consists of 6 + 31 + 22 = 59 cases aged 18-20. The Score column in Figure 7-76 shows that the average age for this group is 19.27, which is the Score now associated with all cases in grouped level 1.

▷ `Scroll to the bottom`

Figure 7-77 shows the 25th-30th grouped levels plus a 31st level for the 8 cases containing no AGE information.

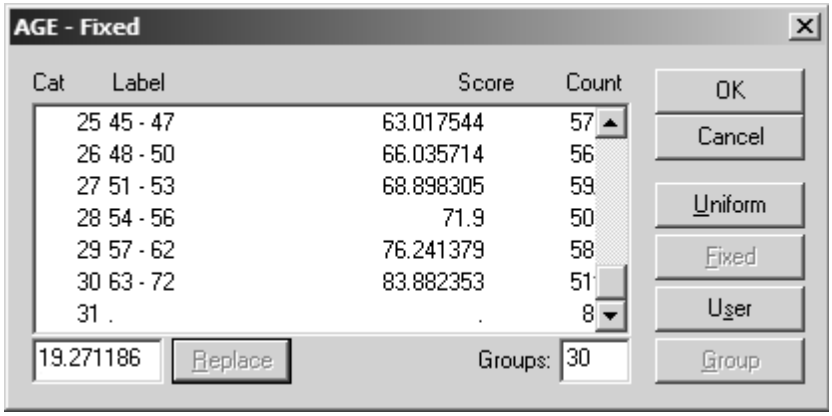


Figure 7-77. Grouped AGE variable

▷ `Click OK to accept the grouping`

As shown in Figure 7-78, the number of AGE levels now shows 'g31' indicating that this new variable has been reduced to 30 grouped age levels plus an additional category (the 31st level) that contains the 8 cases missing AGE information.

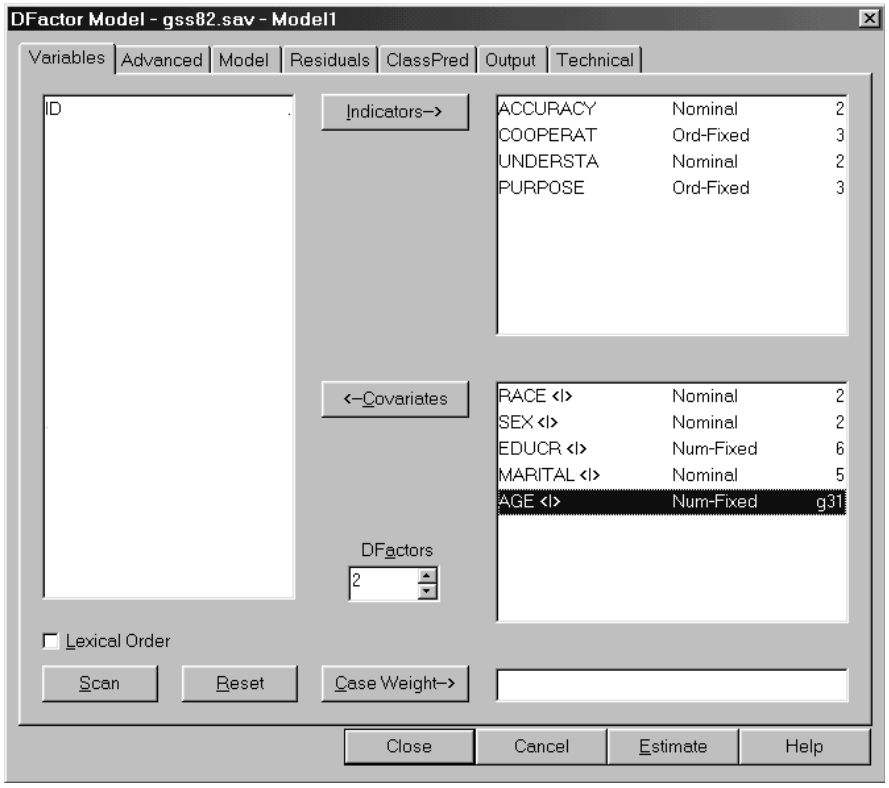


Figure 7-78. Analysis Dialog Box after Grouping Age

To request that a CHAID data file be created following the estimation of this model

▷ Click on 'ClassPred' to open this tab

From the ClassPred Tab

▷ Select CHAID

Default data file names containing the extensions .sav and .chd appear. The resulting .sav file will contain the standard classification information from this model (the same as produced when 'Standard Classification' information is requested in the ClassPred Tab). The .chd file contains the setup for the CHAID analysis. You may change these data file names but be sure to maintain the extensions .sav and .chd.

To include a case ID on each of these output files

▷ Select the variable ID from the list box

▷ Click the ID button to move it to the ID box

Your ClassPred Tab should now look like this:

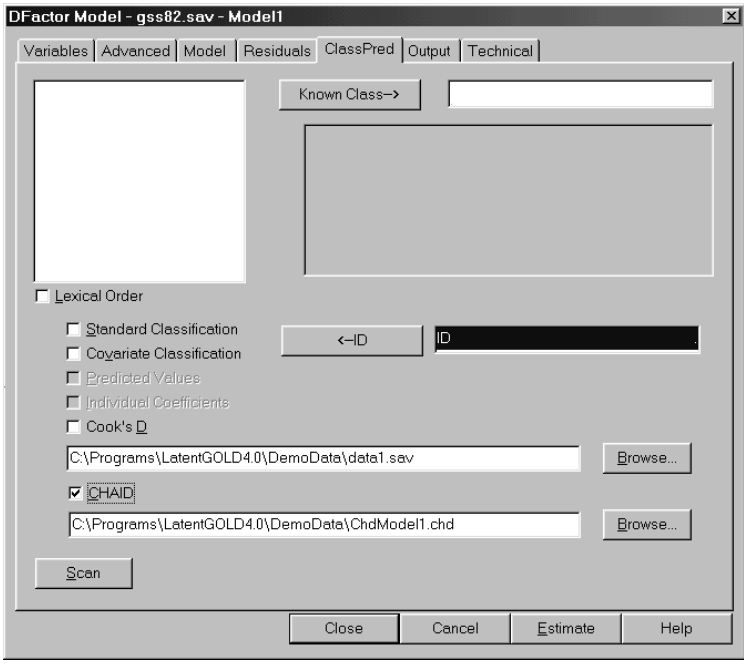


Figure 7-79. ClassPred Tab

ESTIMATING THE MODEL

Now that we have selected our variables and requested a CHAID file, we are ready to estimate the model.

- ▷ Click Estimate

VIEWING THE DFACTOR LOADINGS

- ▷ Click on the expand/contract icon for Parameters to make the output subcategories visible
- ▷ Click 'Loadings' to view the DFactor loadings output

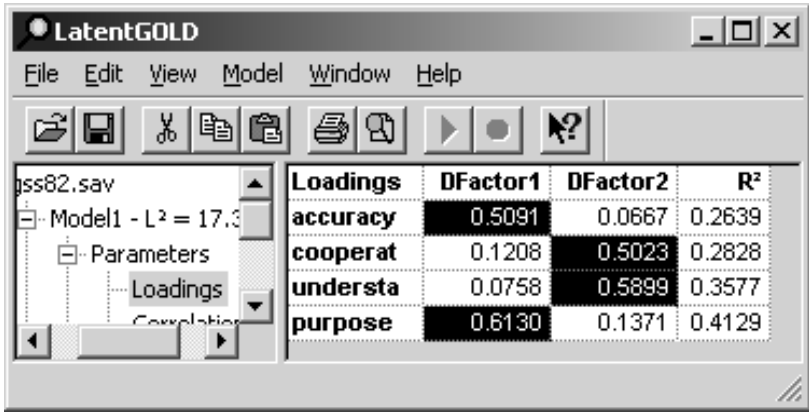


Figure 7-80. DFactor Loadings Output

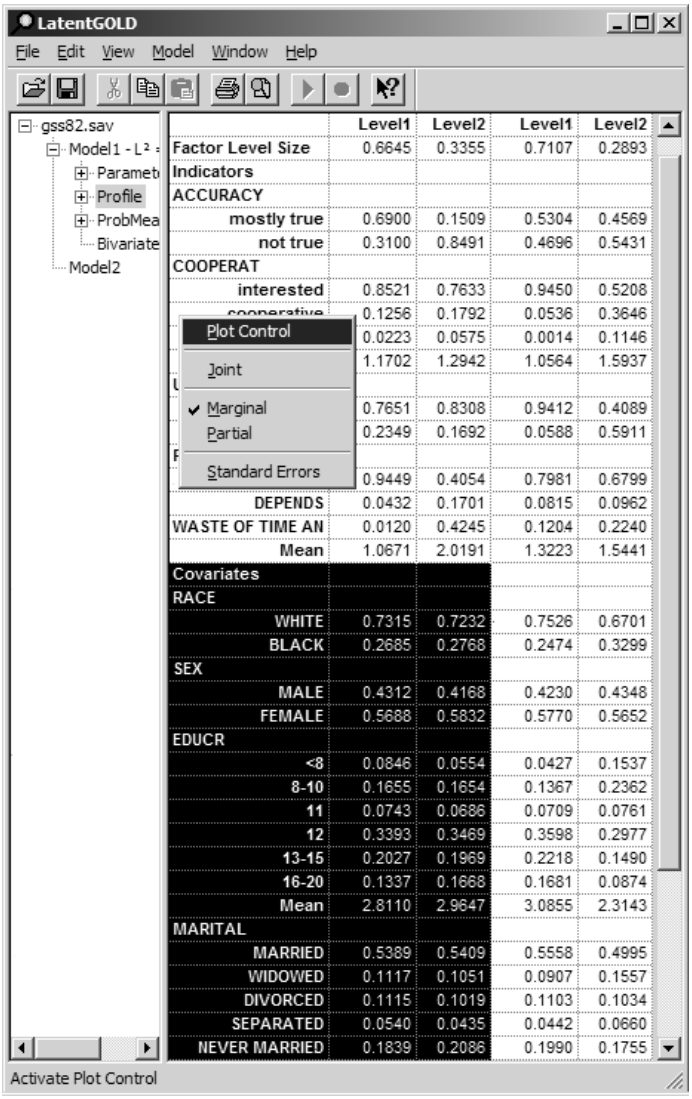
Similar to the results obtained in Tutorial #2 for the more restricted DFactor model, DFactor #1 is primarily associated with PURPOSE and ACCURACY and DFactor #2 is primarily associated with UNDERSTANDING and COOPERATION.

VIEWING THE PROFILE OUTPUT

- ▷ In the Outline Pane, click on Profile

The Profile output is displayed in the Contents Pane:

- ▷ Right click on the Contents Pane to display the View Menu



	Level1	Level2	Level1	Level2
Factor Level Size	0.6645	0.3355	0.7107	0.2893
Indicators				
ACCURACY				
mostly true	0.6900	0.1509	0.5304	0.4569
not true	0.3100	0.8491	0.4696	0.5431
COOPERAT				
interested	0.8521	0.7633	0.9450	0.5208
cooperative	0.1256	0.1792	0.0536	0.3646
	0.0223	0.0575	0.0014	0.1146
	1.1702	1.2942	1.0564	1.5937
	0.7651	0.8308	0.9412	0.4089
	0.2349	0.1692	0.0588	0.5911
	0.9449	0.4054	0.7981	0.6799
DEPENDS	0.0432	0.1701	0.0815	0.0962
WASTE OF TIME AN	0.0120	0.4245	0.1204	0.2240
Mean	1.0671	2.0191	1.3223	1.5441
Covariates				
RACE				
WHITE	0.7315	0.7232	0.7526	0.6701
BLACK	0.2685	0.2768	0.2474	0.3299
SEX				
MALE	0.4312	0.4168	0.4230	0.4348
FEMALE	0.5688	0.5832	0.5770	0.5652
EDUCR				
<8	0.0846	0.0554	0.0427	0.1537
8-10	0.1655	0.1654	0.1367	0.2362
11	0.0743	0.0686	0.0709	0.0761
12	0.3393	0.3469	0.3598	0.2977
13-15	0.2027	0.1969	0.2218	0.1490
16-20	0.1337	0.1668	0.1681	0.0874
Mean	2.8110	2.9647	3.0855	2.3143
MARITAL				
MARRIED	0.5389	0.5409	0.5558	0.4995
WIDOWED	0.1117	0.1051	0.0907	0.1557
DIVORCED	0.1115	0.1019	0.1103	0.1034
SEPARATED	0.0540	0.0435	0.0442	0.0660
NEVER MARRIED	0.1839	0.2086	0.1990	0.1755

Figure 7-81. Marginal Profile Output

The size of each level of each factor is given in the top row. For example, for DFactor2, about 71% are in level 1, the remaining 29% in level 2.

The remainder of the Profile Output is divided into 2 sections; the first section contains tables for the Indicators, the second for the Covariates. For interpretation of the tables pertaining to the Indicators, see Tutorial #2. Here, we will focus on the section pertaining to the Covariates (see columns highlighted in Figure 7-81). The body of the tables contain probabilities for each variable category conditional on the levels for DFactor1 and DFactor2 (column percentages). Beneath these probabilities, means are displayed for the Numeric variables (not the Nominal variables).

The Joint view of the Profile output contains similar information for the levels of the Joint DFactor (1,1), (1,2), (2,1) and (2,2), where (1,1) refers to those classes at level 1 on DFactor #1 and level 1 on DFactor #2. The Joint view for a restricted form of the DFactor model was illustrated in Tutorial #2.

By default, covariates such as EDUCR that contain more than 5 levels are grouped into 5 levels in the Profile output. To restore the original education levels for EDUCR:

From the View Menu

- Select Plot Control

The Control Panel for the Profile Output and Associated Plot appears (see Figure 7-82)

- Select the variable EDUCR
- Change the number '5' to '0' in the Groups box
- Click Update

Figure 7-82. Profile Plot Control

The table for EDUCR changes as shown in Figure 7-83

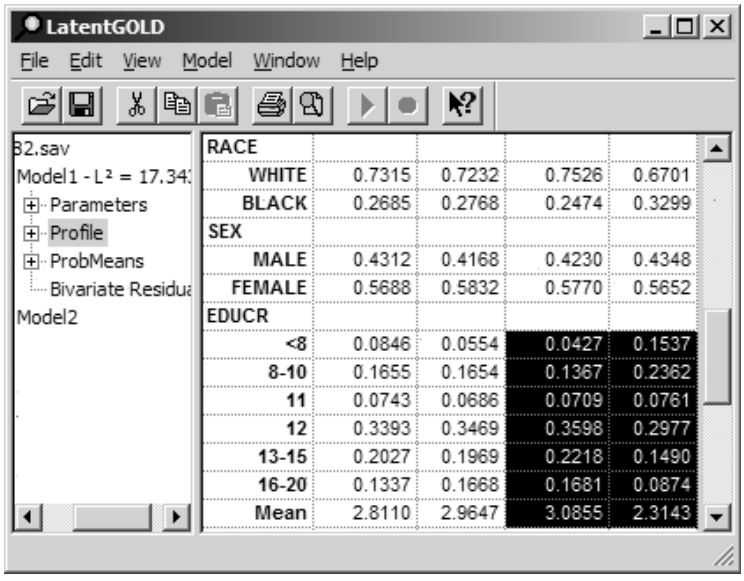


Figure 7-83. New Profile Output

Notice that the levels of DFactor #1 do not appear to differ with respect to race, gender or educational attainment, while DFactor #2 shows strong differences with respect to race and education. For example, cases in level 1 of DFactor #2 have higher levels of education -- 22.2% have some college ('13-15' years of education), and an additional 16.8% have a college degree (completed '16-20' years) -- than cases in level 2 (14.9% and 8.7% respectively).

We will now show how to use the CHAID option to assess the statistical significance of these and other Covariate x Latent Class relationships.

USING THE CHAID OPTION

The SI-CHAID program actually consists of 2 programs, called 'CHAID Define' and 'CHAID Explore' both of which utilize a .chd file as input. Typically, the Define program is used first to set the analysis options and then the Explore command is executed to perform the CHAID analysis. However, if the default settings are adequate, the Explore program may be used immediately to perform the CHAID analysis.

The default .chd file generated by Latent GOLD ('ChdModel1.chd') based on DFactor models, defines the dependent variable to be identical to DFactor #1. Thus, the Explore program can be used immediately to profile the levels of DFactor1, or the dependent variable and other default settings may be changed first (using CHAID Define).

For the current model, a CHAID analysis based on the default specification finds that none of the demographics are significant. This suggests that the levels of DFactor1 which reflect either a favorable or unfavorable attitude towards the purpose and accuracy of surveys are not related to any of our demographic variables. Thus, we will show how to use the Define program to change the default settings to re-define the dependent variable (which specifies the latent classes to profile) to DFactor2.

NOTE: If you are using the Demo version of SI-CHAID, follow Step A below. If you have a regular SI-CHAID license, you may skip this step.

STEP A (For users of SI-CHAID demo version only). Replace the file 'data1.sav'

- ▷ On your computer, go to the directory where you saved the file 'data1.sav', that was generated by Latent GOLD in this tutorial. By default, this should be the Latent GOLD 4.0/DemoData directory - (see Figure 7-79).
- ▷ Delete the file 'data1.sav'.
- ▷ Locate the file by the same name, 'data1.sav', that came with your SI-CHAID demo program.
- ▷ Copy it and paste it into the Latent GOLD 4.0/DemoData directory (or whatever directory the 'data1.sav' was saved).
- ▷ Continue with the steps below.

END OF STEP A

- ▷ Open the CHAID Define program

From the File Menu

▷ Select Open

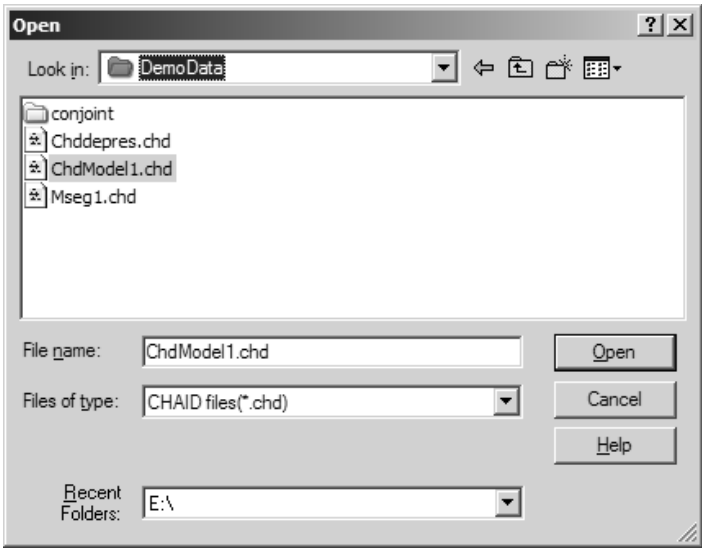


Figure 7-84. Open Dialog Box

Alternatively,

▷ Double click on the CHAID definition file 'ChdModel1.chd'

The Define program opens.

> **NOTE**: If you are receiving an error message at this point, it may be because you are using the Demo version of SI-CHAID and you did not complete Step A above. Go back to Step A and follow it prior to proceeding.

The Outline Pane shows that it is ready for you to define 'Model1' associated with the Standard Classification data file ('data1.sav') that was generated by Latent GOLD. The Contents Pane contains the current (default) settings for Model1. StartUp = None means that the program will begin in interactive as opposed to automatic mode.

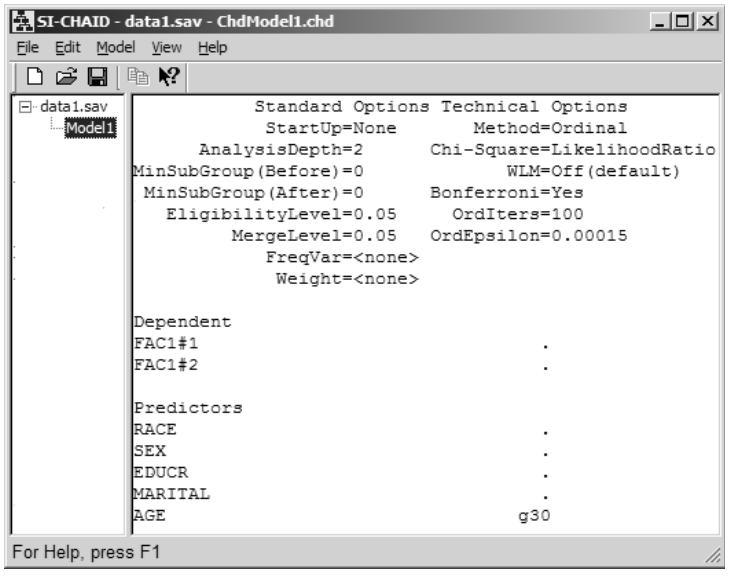


Figure 7-85. Summary for Model1 in Chaid Define

▷ Double click on Model1 to edit the current settings

The Analysis Dialog box opens. (This dialog box can also be opened by selecting Edit from the Model Menu.)

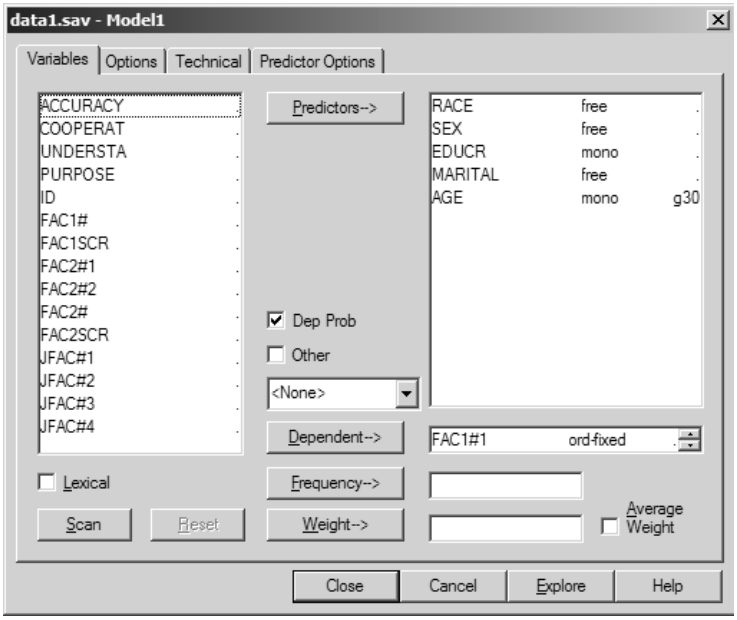


Figure 7-86. The Analysis Dialog Box

Note that the demographic variables that were entered into Latent GOLD as Covariates are now included in the SI-CHAID Predictors box, with their current CHAID setting listed to the right of the variable name. By default, Covariates that were specified as 'Nominal' are set to 'free' and those that were specified as 'Numeric' to either 'mono' or 'float' depending upon whether or not missing values are present. 'Free' means that CHAID is free to combine any of its categories that are not significantly different with respect to the dependent variable, while 'mono' means that only adjacent categories may be combined. The 'float' scale type setting means that the predictor is treated as 'mono' except for the last ('floating') category (generally containing missing values) which is 'free' to combine with any category.

▷ Click Scan

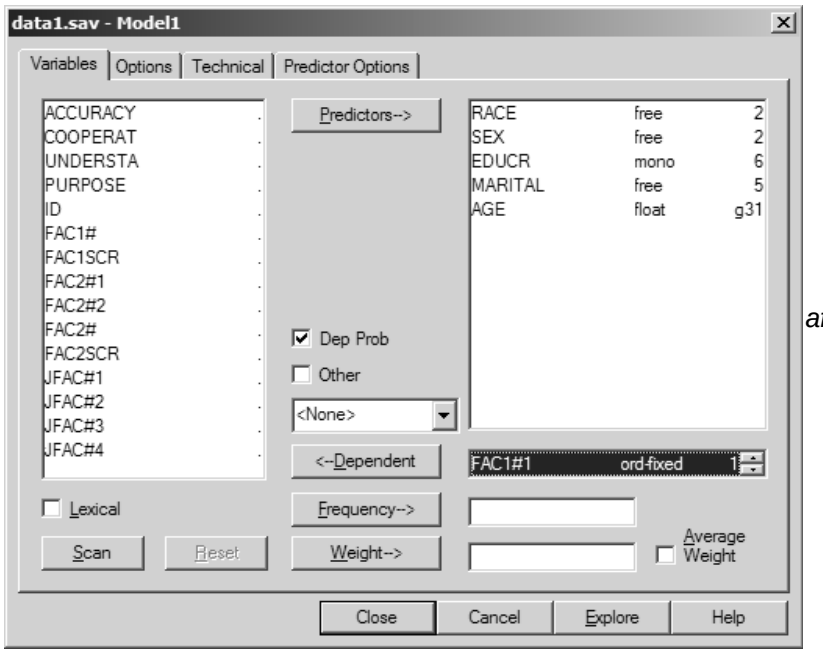


Figure 7-87. Analysis Dialog Box after Scanning

Note that the setting for AGE has been changed to 'float' and 'g31' replaces 'g30'. This change is because the scan detected the 31st level for AGE as containing missing values.

Notice that the 'Dep Prob' box is checked, which indicates that the posterior membership probabilities are used to weight the dependent variable, which by default is DFactor1. The variables FAC1#1 and FAC1#2 contain the posterior membership probabilities for levels 1 and 2 of DFactor1. Both are included in the Dependent Box (only the first is visible).

To change the dependent variable from DFactor1 to DFactor2

- Select both variables FAC1#1 and FAC1#2 in the Dependent Box
- Click Dependent to remove them
- Select the variables FAC2#1 and FAC2#2 from the Variables List box
- Click the button labeled 'Dependent->' to move these variables to the Dependent box

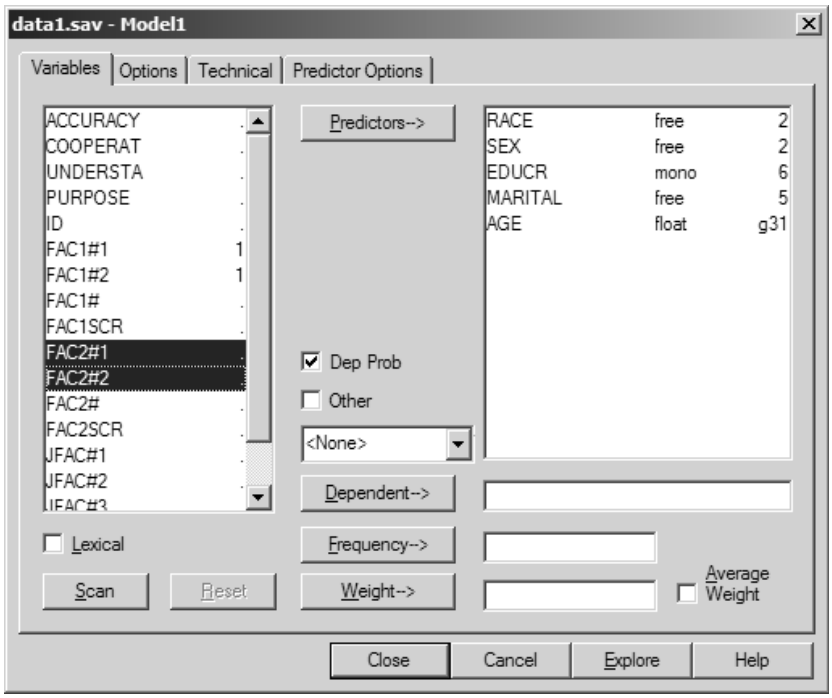


Figure 7-88. Moving FAC2#1 and FAC2#2 to Dependent Box

- Right click in the Dependent box
- Select 'Nominal' to use the Nominal CHAID algorithm

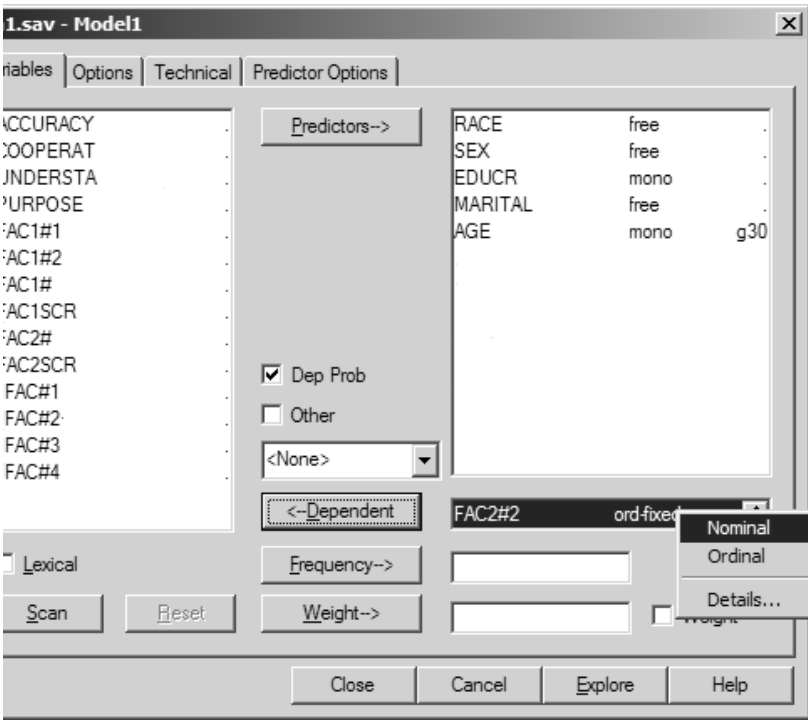


Figure 7-89. Changing FAC2#1 and FAC2#2 to Nominal

▷ Click Options to open the Options Tab

▷ Select Auto to start in automatic mode

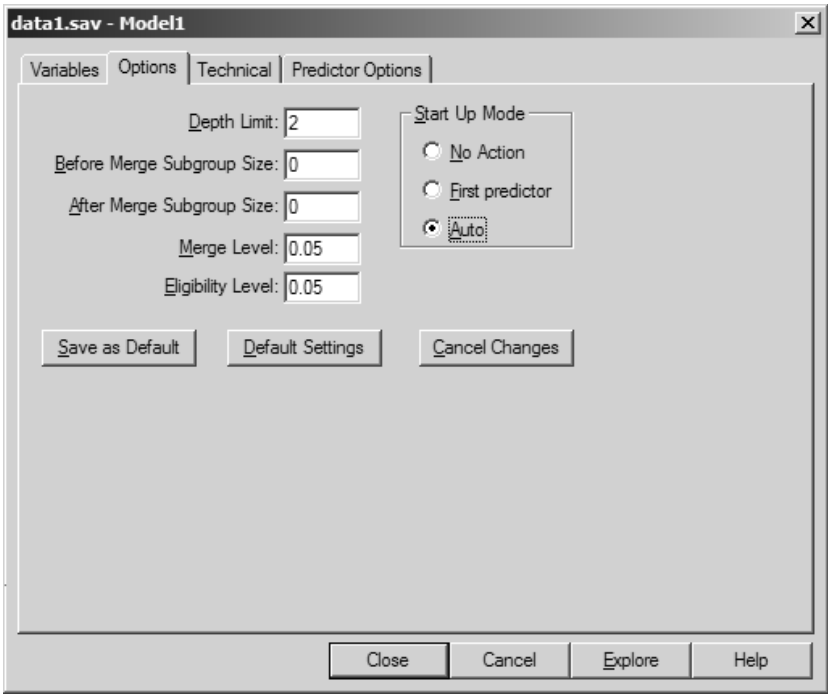


Figure 7-90. Options Tab

▷ Click the Explore button

CHAID prompts you to save the updated definition file named Model1.chd

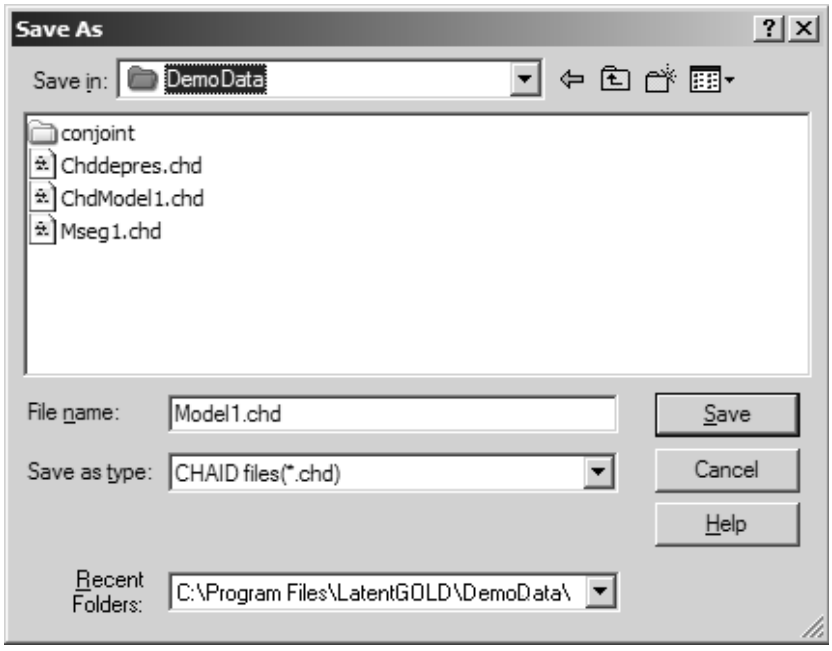


Figure 7-91. Save Definition Dialog Box

You may change the name of this file and the directory where it will be saved

▷ `Click Save to save the definition file and open the CHAID Explore program`

CHAID Explore opens and displays the resulting segmentation tree.

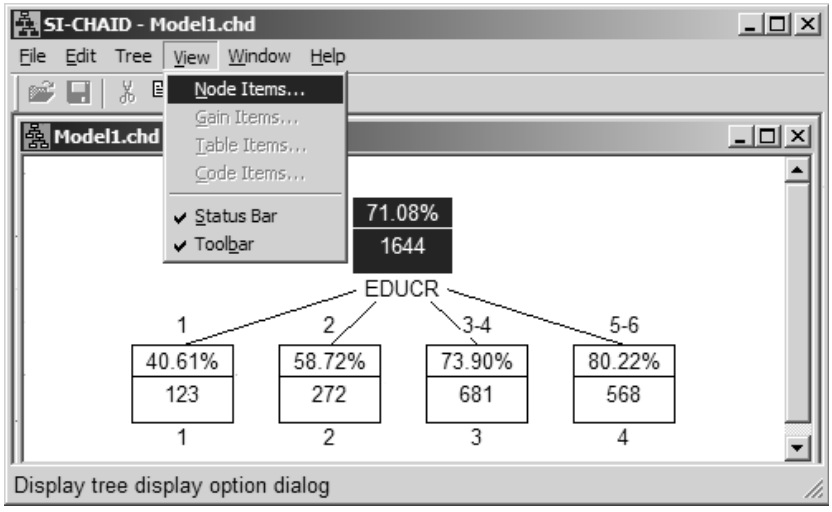


Figure 7-92. CHAID Segmentation Tree

Note that the resulting segmentation tree is based only on the education variable. That is, none of the 4 nodes splits further on any other variable. This means that given one's education level, the levels of DFactor2 are not related to RACE, SEX, MARITAL or AGE.

To verify what is displayed in each node:

▷ Select Node Items from the View Menu

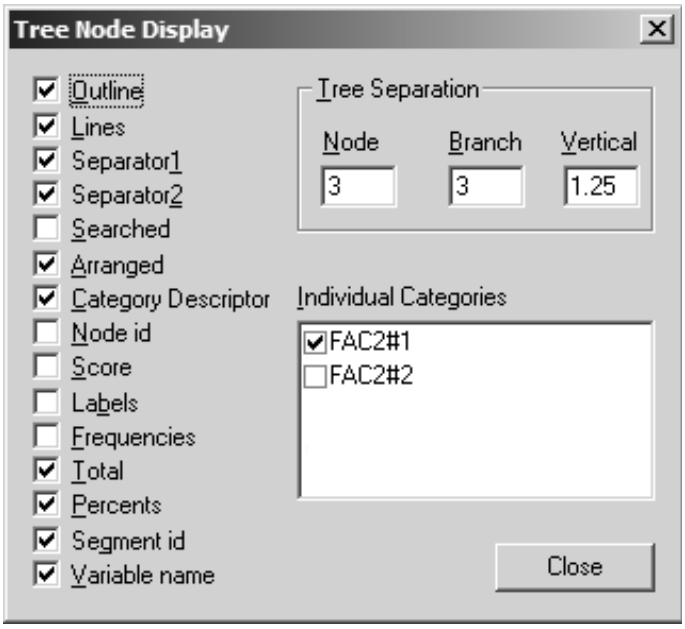


Figure 7-93. Node Items Dialog Box

The items displayed in each node are indicated by checkmarks. The bottom four items checked are:

Total - the total sample size (displayed in the lower portion of each node)

Percents - In the upper portion of each node, (row) percentages are displayed for the selected categories of the dependent variable (see Individual Categories box)

Segment id - a sequential id number appears below each node

Variable name - name of the predictor variable(s) whose categories defines the nodes (e.g., 'EDUCR' in Figure 7-92).

In the Individual Categories box, a check mark appears next to 'FAC2#1' only. This means that the percentages that are displayed in the node, correspond to the 1st category of the dependent variable only -- level 1 of DFactor2. In the root node, we see that overall about 71% of the 1,644 cases are in level 1 of the dependent variable. This agrees with the Profile output shown in Figure 7-81.

The tree grows by splitting on the grouped categories of EDUCR. We see that as the education level changes from category 1 ('< 8 years') to category 2 ('8-10 years') to categories '3-4' ('11-12 years') to categories '5-6' ('>12 years'), the percentage in level 1 of DFactor2 increases from 40.6% to 58.7% to 73.9% to 80.2%.

To obtain a cross-tab of the dependent variable by EDUCR and an assessment of the statistical significance

▷ From the Window Menu, select New Table

At the bottom of the table, we see that the p-value is 1.3 x 10-19 for this variable.

To open the Table Display control panel from which you can alter the table view,

▷ From the View Menu, select Table Items

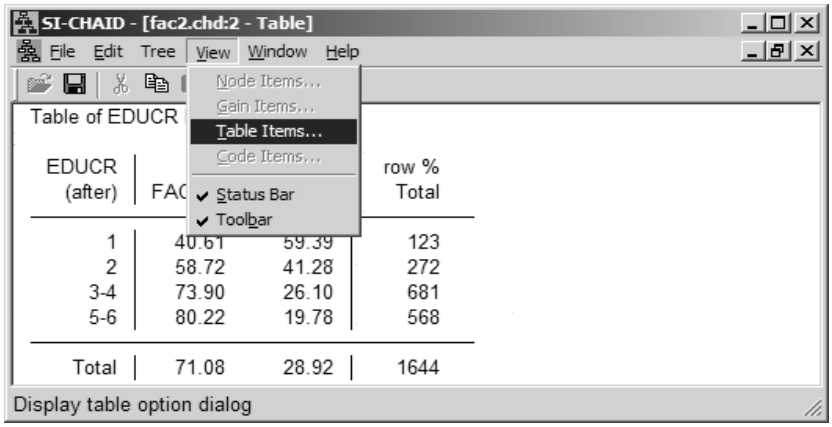


Figure 7-94. Table Display

From the Table Display

▷ Select Before Merge

▷ Select Column Percents

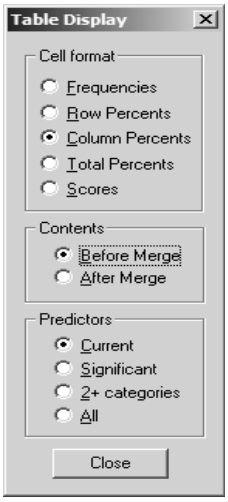


Figure 7-95. Table Display Dialog Box

The table changes to column percentages for the original EDUCR categories and matches the Profile table (recall Figure 7-83).

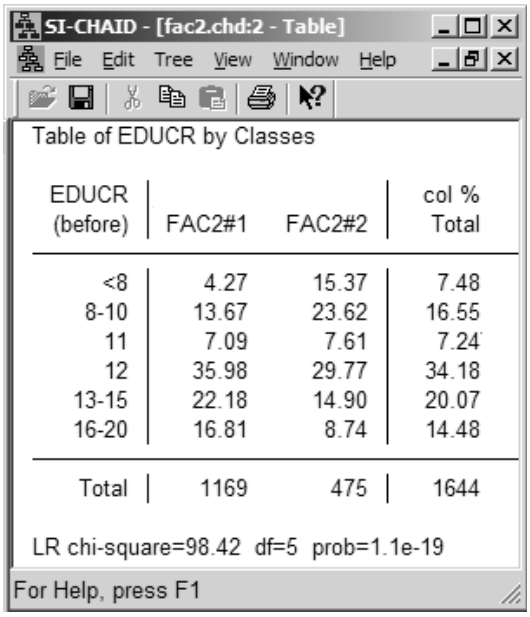


Figure 7-96. New Table

To obtain tables for all predictors,

▷ From the Predictors section of the Table Display select 'All'

▷ Scroll down to view the table for RACE.

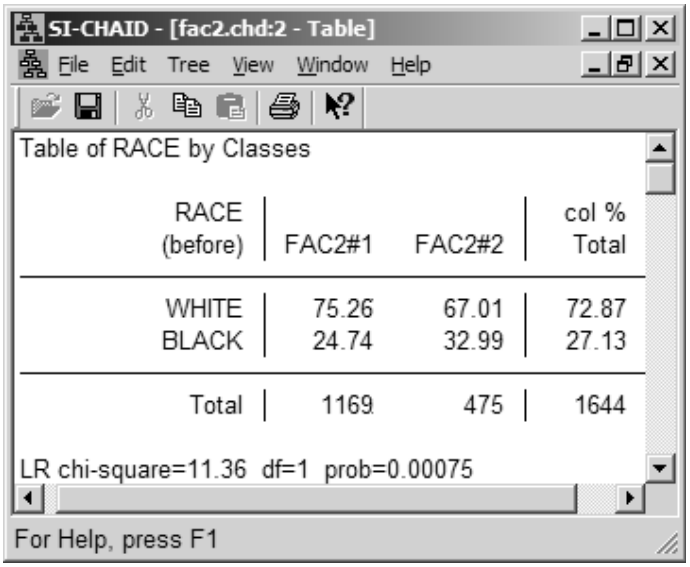


Figure 7-97. Table for RACE

Information regarding the statistical significance associated with a predictor, provided at the bottom of the table, shows that RACE is significantly related to DFactor2 overall. However, the terminal nodes in the CHAID tree are defined solely in terms of EDUCR (i.e., these nodes do not split further on RACE), which means that the RACE effect is no longer significant once education is taken into account. Thus, this relationship can be viewed as spurious, being explainable by the fact that the blacks in the sample had lower levels of educational attainment.

In summary, we found that none of the demographics are significantly related to DFactor1 and that EDUCR is the most important descriptor for profiling DFactor2. Given the results from tutorial #2 that showed that DFactor1 is related to the dependent variables PURPOSE and ACCURACY, while DFactor2 is related to the dependent variable UNDERSTAND, we might expect the following relationships to exist between the demographic variables and these indicators:

1) No demographics are significantly related to PURPOSE or ACCURACY.
2) EDUCR is the most important variable related to UNDERSTAND.

CHAID Explore can also be run in interactive mode. For example, to grow a tree interactively beginning at the root node,

▷ Click the root node

▷ Choose 'Select' from the Tree menu

By default, the variable selection chart lists the predictors that are significant at this overall level of the tree.

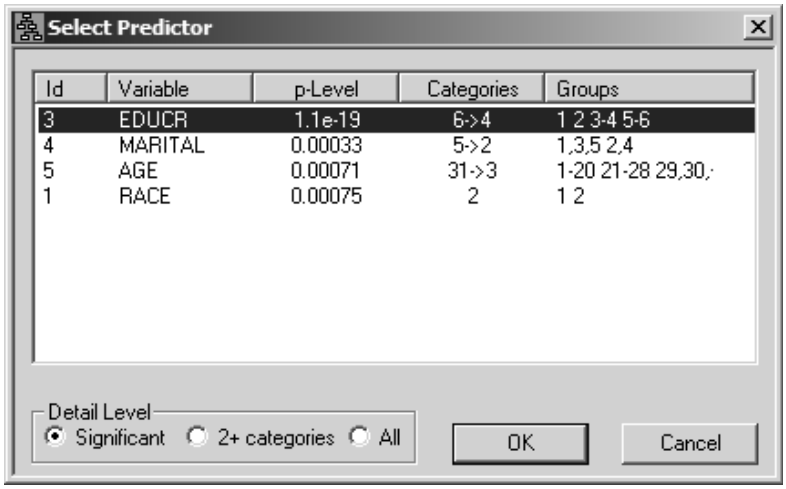


Figure 7-98. Select Predictor Dialog Box

While EDUCR is most significant, you may select any predictor and select OK to grow the tree based on that predictor. SI-CHAID allows you to select this or any other variable to grow the tree.

Select Contents from the Help Menu to display detailed information on the SI-CHAID program

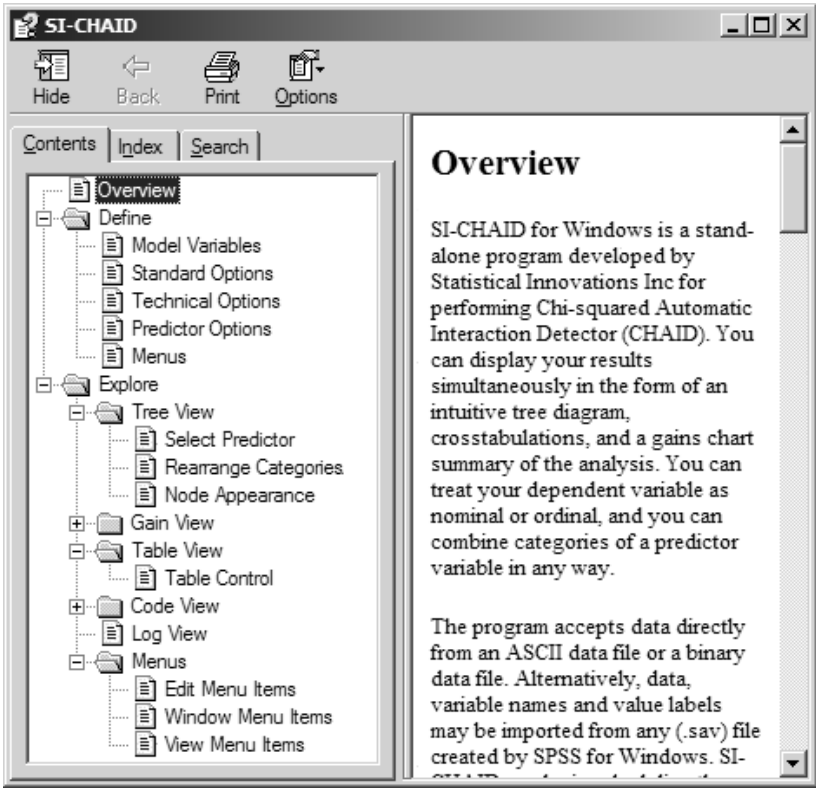


Figure 7-99. SI-CHAID Help Contents