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 N_1 = 1,198 of the 1,202 white respondents used in tutorials #1 and #2 and a supplemental sample of N_2 = 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

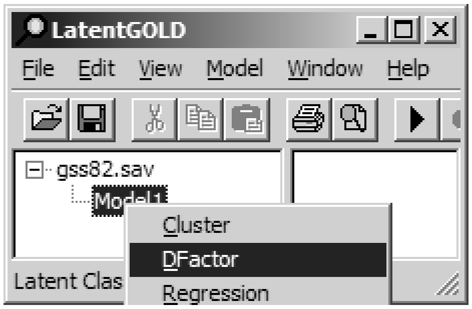
- From the Files of type drop down list, select SPSS System Files if this is not already the default listing.

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

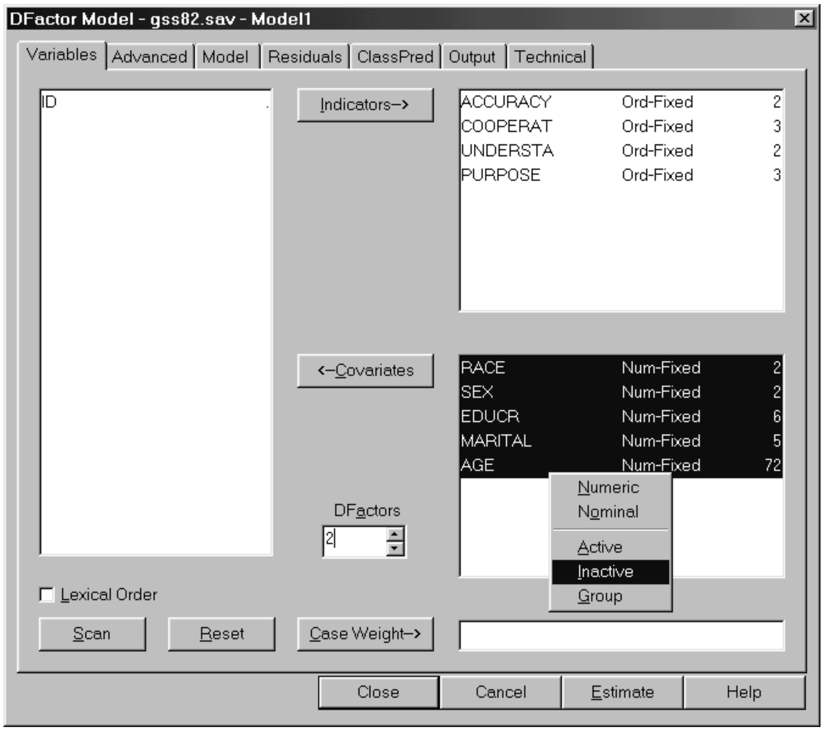
- Click Scan

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:



ID

ACCURACY	Ord-Fixed	2
COOPERAT	Ord-Fixed	3
UNDERSTA	Ord-Fixed	2
PURPOSE	Ord-Fixed	3

RACE	Num-Fixed	2
SEX	Num-Fixed	2
EDUCR	Num-Fixed	6
MARITAL	Num-Fixed	5
AGE	Num-Fixed	72

DFactors: 2

☐ Lexical Order

Figure 7-73. Making Covariates Inactive

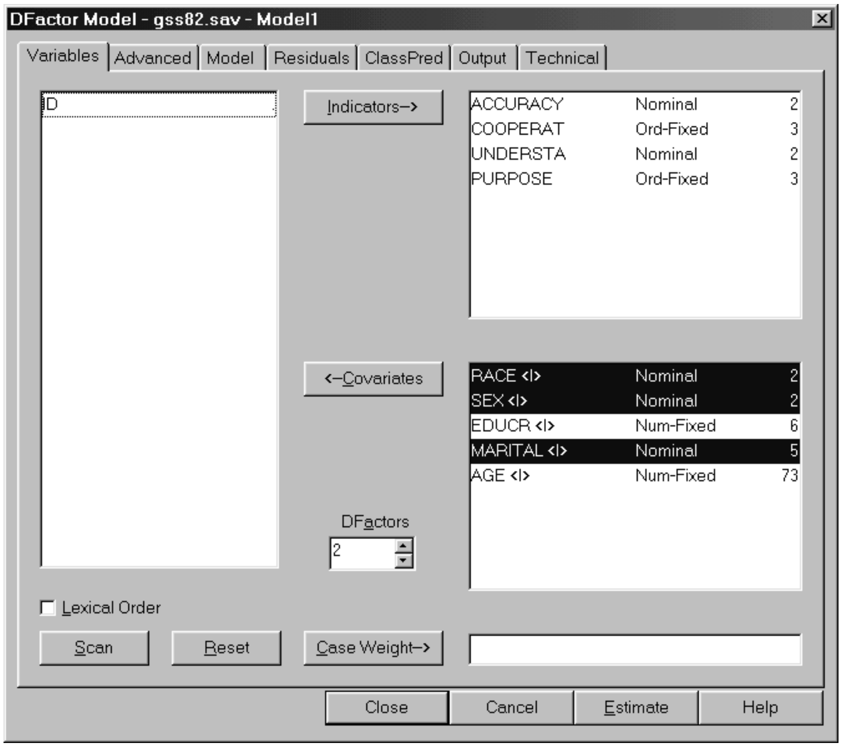
- ➢ Select Inactive

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:



```
ID

ACCURACY      Nominal      2
COOPERAT      Ord-Fixed    3
UNDERSTA      Nominal      2
PURPOSE       Ord-Fixed    3

RACE <I>      Nominal      2
SEX <I>       Nominal      2
EDUCR <I>     Num-Fixed    6
MARITAL <I>   Nominal      5
AGE <I>       Num-Fixed   73

DFactors: 2
☐ Lexical Order
```

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.)

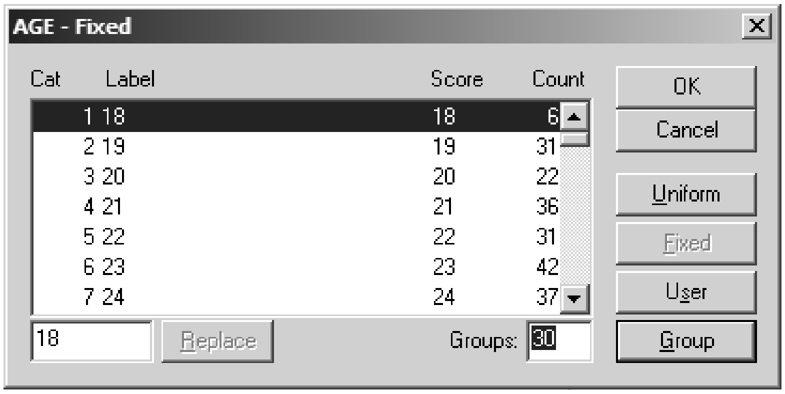
A limitation of SI-CHAID is that variables can have no more than 31 levels. SI-CHAID automatically reduces the number of levels to 15 for variables exceeding this limit. Here, we will illustrate the Group option in Latent GOLD to reduce the number of levels of AGE.

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



Cat	Label	Score	Count
1	18	18	6
2	19	19	31
3	20	20	22
4	21	21	36
5	22	22	31
6	23	23	42
7	24	24	37

18

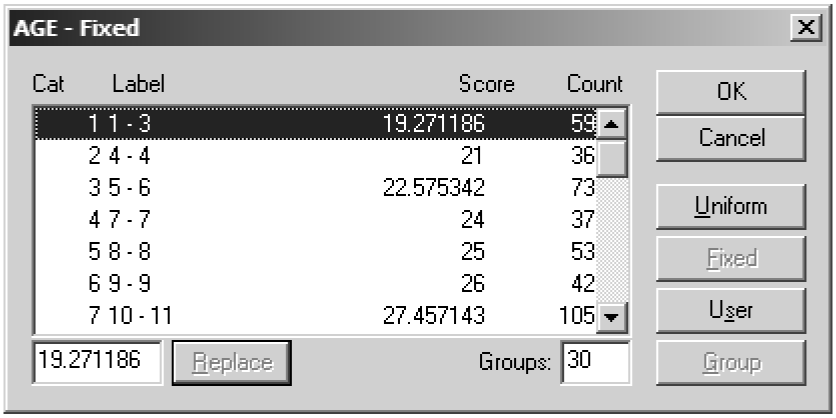
Groups: 30

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



Cat	Label	Score	Count
1	1 - 3	19.271186	59
2	4 - 4	21	36
3	5 - 6	22.575342	73
4	7 - 7	24	37
5	8 - 8	25	53
6	9 - 9	26	42
7	10 - 11	27.457143	105

19.271186

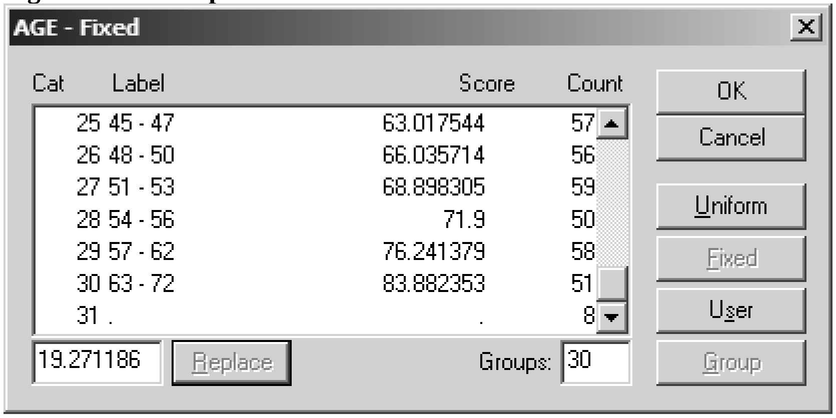
Groups: 30

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



Cat	Label	Score	Count
25	45 - 47	63.017544	57
26	48 - 50	66.035714	56
27	51 - 53	68.898305	59
28	54 - 56	71.9	50
29	57 - 62	76.241379	58
30	63 - 72	83.882353	51
31	.	.	8

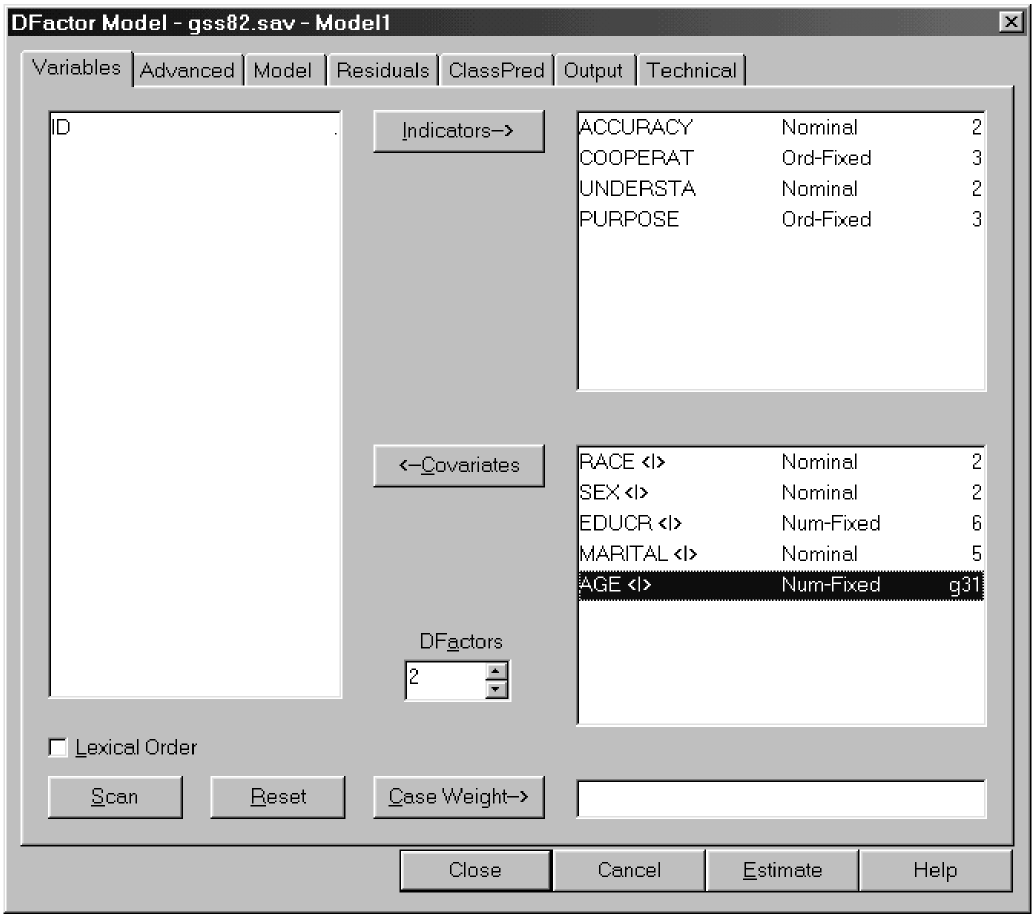
19.271186

Groups: 30

- 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



DFactor Model - gss82.sav - Model1

ID

ACCURACY	Nominal	2
COOPERAT	Ord-Fixed	3
UNDERSTA	Nominal	2
PURPOSE	Ord-Fixed	3

RACE <I>	Nominal	2
SEX <I>	Nominal	2
EDUCR <I>	Num-Fixed	6
MARITAL <I>	Nominal	5
AGE <I>	Num-Fixed	g31

DFactors 2

☐ Lexical Order

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

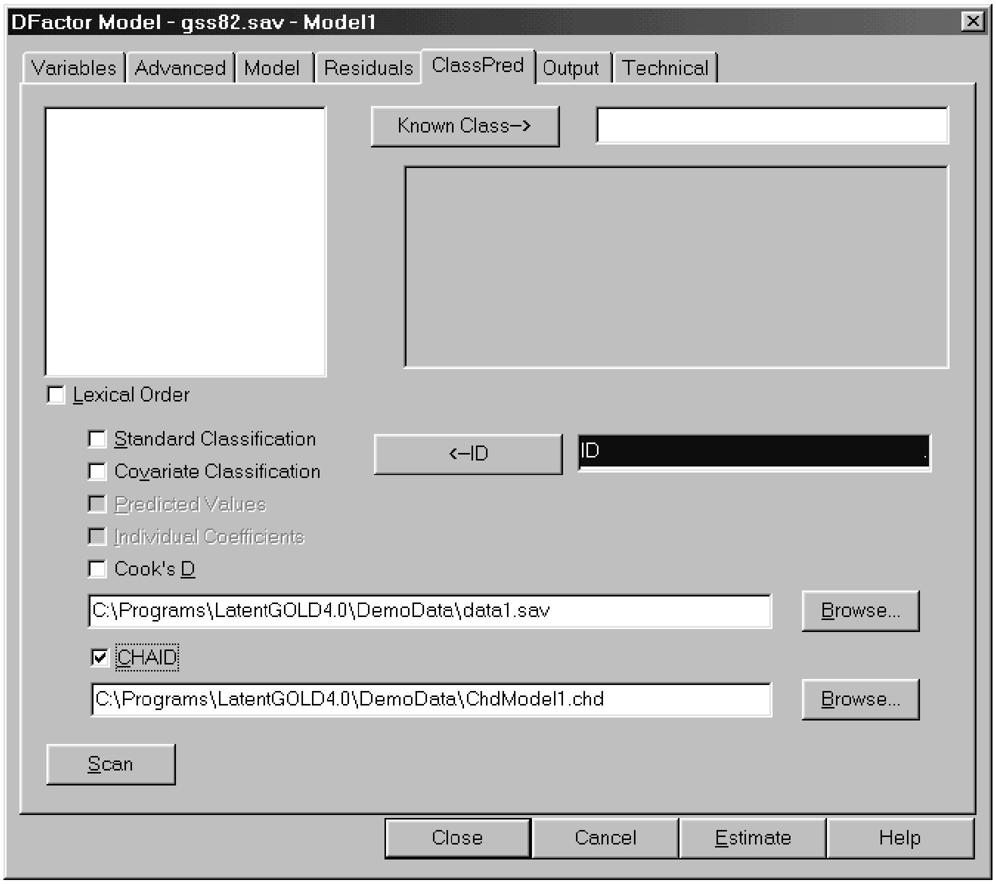
Default data file names containing the extensions .sav and .chd appear. The resulting .sav file will contain the standard classification information from this model (the same as produced when ‘Standard Classification’ information is requested in the ClassPred Tab). The .chd file contains the setup for 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



DFactor Model - gss82.sav - Model1

□ Lexical Order

□ Standard Classification

□ Covariate Classification

□ Predicted Values

□ Individual Coefficients

□ Cook's D

C:\Programs\LatentGOLD4.0\DemoData\data1.sav

☑ CHAID

C:\Programs\LatentGOLD4.0\DemoData\ChdModel1.chd

ID

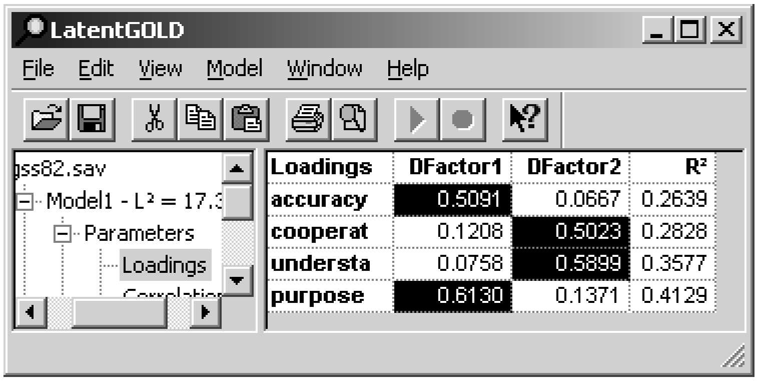
Estimating the Model

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



Loadings	DFactor1	DFactor2	R²
accuracy	0.5091	0.0667	0.2639
cooperat	0.1208	0.5023	0.2828
understa	0.0758	0.5899	0.3577
purpose	0.6130	0.1371	0.4129

Figure 7-80. DFactor Loadings Output

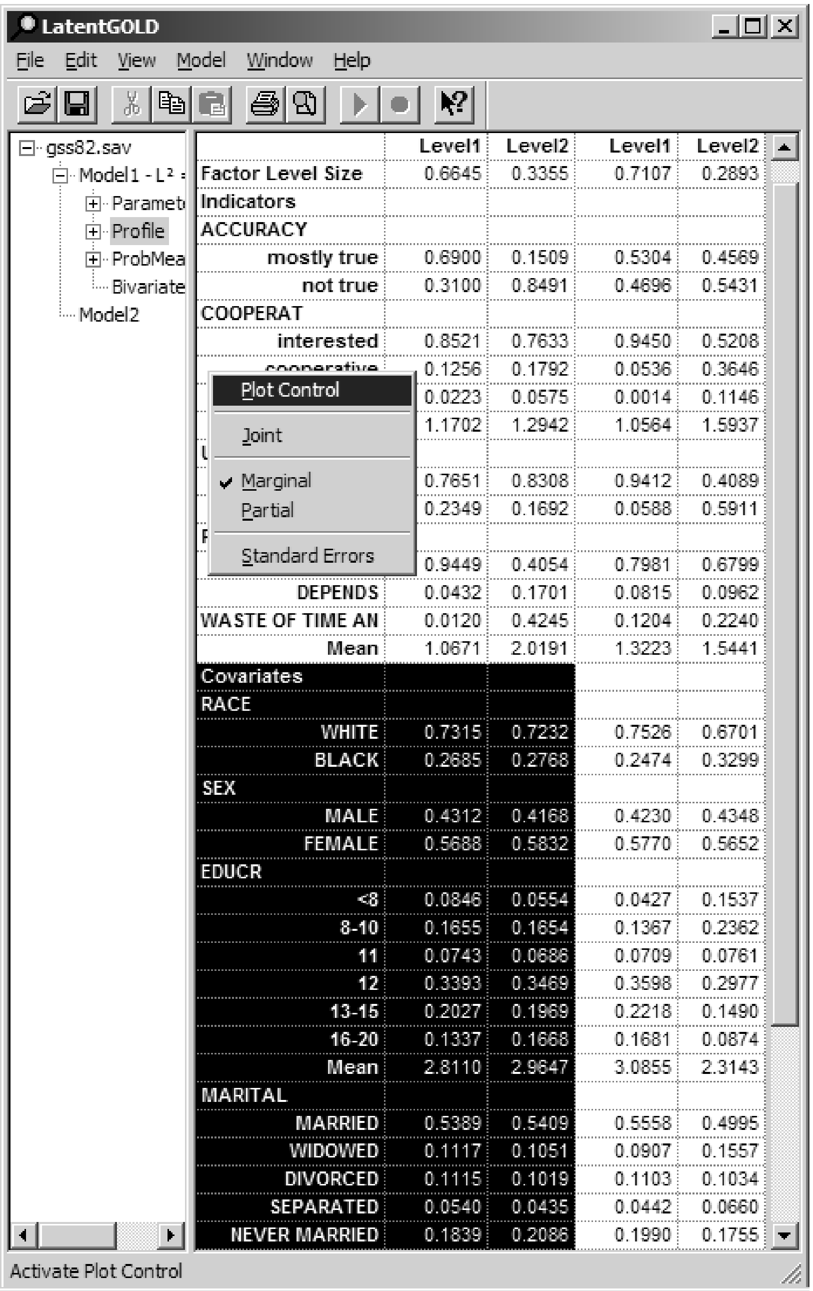
Similar to the results obtained in Tutorial #2 for the more restricted DFactor model, DFactor #1 is primarily associated with PURPOSE and ACCURACY and DFactor #2 is primarily associated with UNDERSTANDING and COOPERATION.

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
U				
	0.7651	0.8308	0.9412	0.4089
	0.2349	0.1692	0.0588	0.5911
P				
	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:

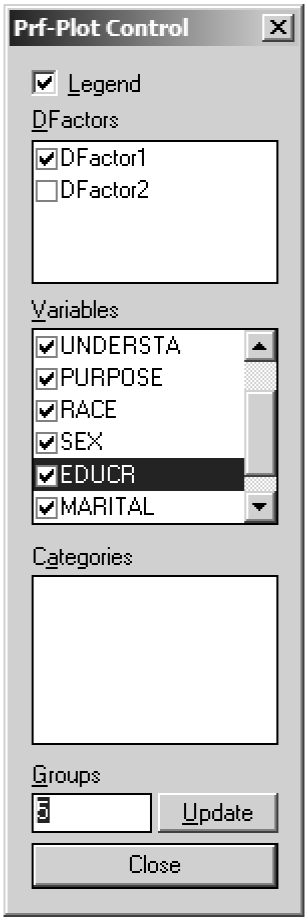
From the View Menu

- 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



Prf-Plot Control

☑ Legend

DFactors

☑ DFactor1
☐ DFactor2

Variables

☑ UNDERSTA
☑ PURPOSE
☑ RACE
☑ SEX
☑ EDUCR
☑ MARITAL

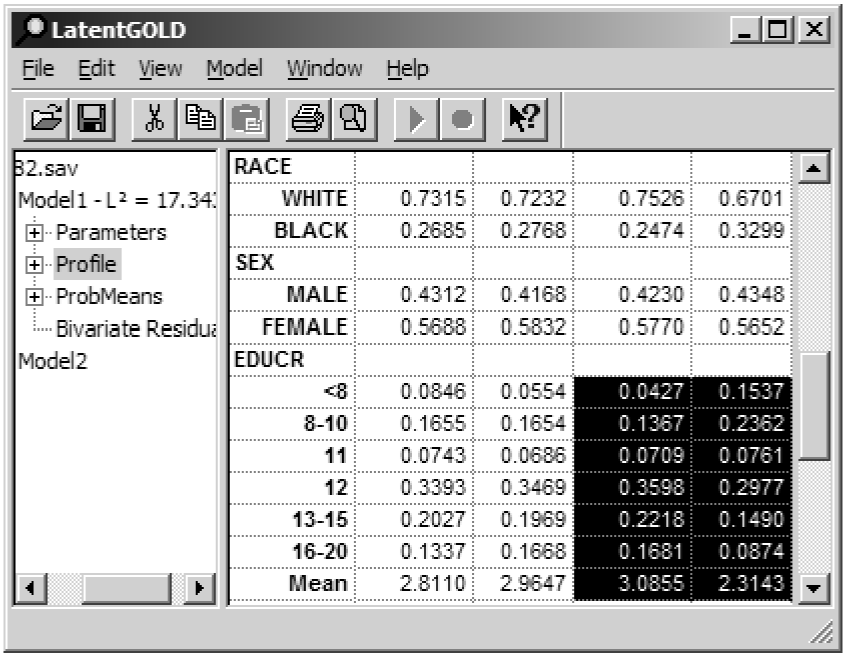
Categories

Groups

5

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

Figure 7-83. New Profile Output



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

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’

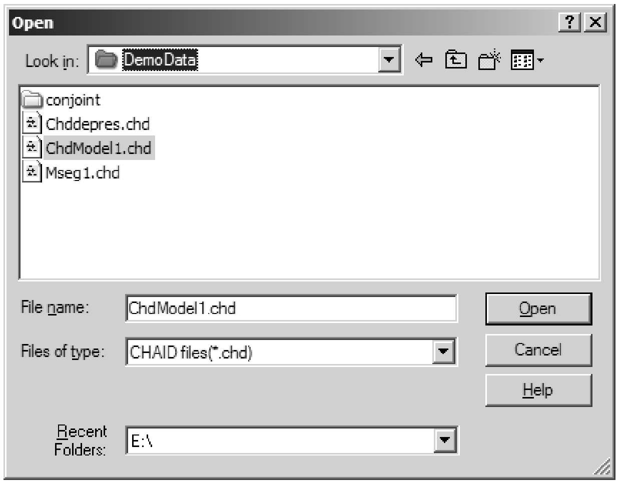
- On your computer, go to the directory where you saved the file ‘data1.sav’, that was generated by Latent GOLD in this tutorial. By default, this should be the Latent GOLD 4.0/DemoData directory – (see Figure 7-79).
- Delete the file ‘data1.sav’.
- Locate the file by the same name, ‘data1.sav’, that 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



Look in: DemoData

File name: ChdModel1.chd

Files of type: CHAID files(*.chd)

Recent Folders: E:\

Figure 7-84. Open Dialog Box

Alternatively,

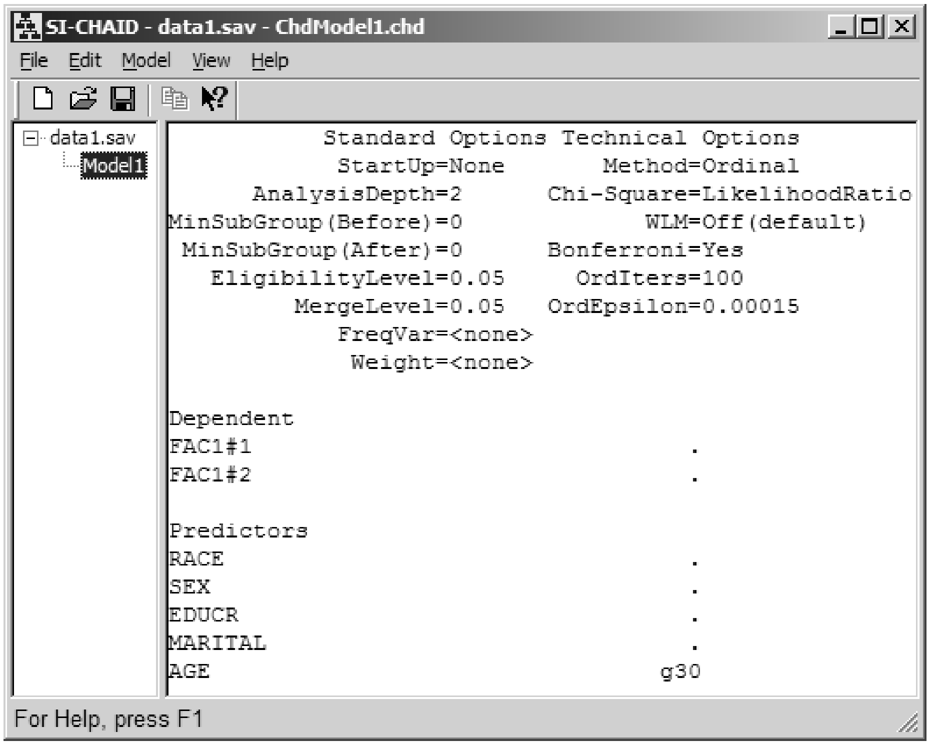
- Double click on the CHAID definition file ‘ChdModel1.chd’

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

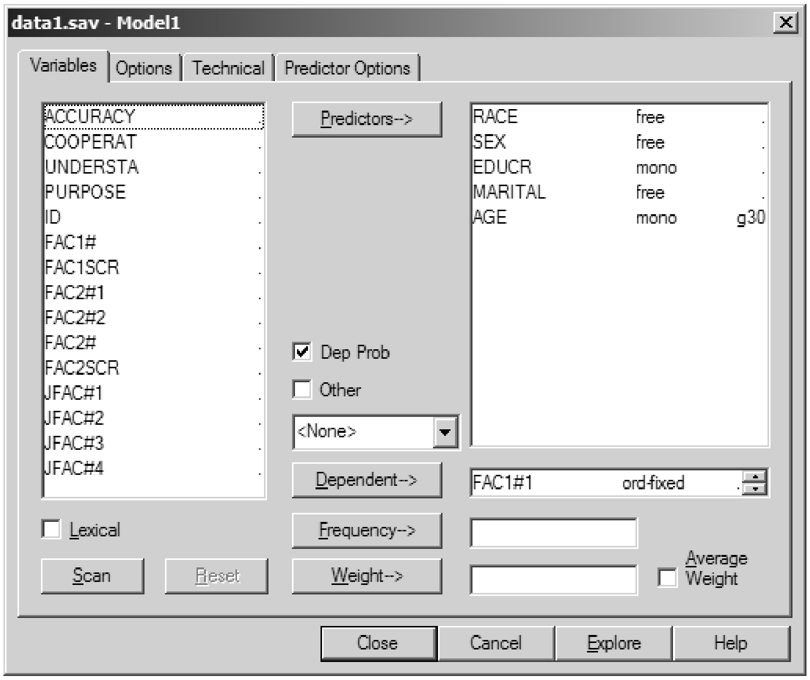

SI-CHAID - data1.sav - ChdModel1.chd
File Edit Model View Help
data1.sav
Model1
For Help, press F1

```
             Standard Options Technical Options
              StartUp=None       Method=Ordinal
      AnalysisDepth=2      Chi-Square=LikelihoodRatio
MinSubGroup(Before)=0             WLM=Off(default)
 MinSubGroup(After)=0      Bonferroni=Yes
   EligibilityLevel=0.05     OrdIters=100
         MergeLevel=0.05   OrdEpsilon=0.00015
            FreqVar=<none>
             Weight=<none>

Dependent
FAC1#1                                .
FAC1#2                                .

Predictors
RACE                                  .
SEX                                   .
EDUCR                                 .
MARITAL                               .
AGE                                 g30
```

- ➢ 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


data1.sav - Model1

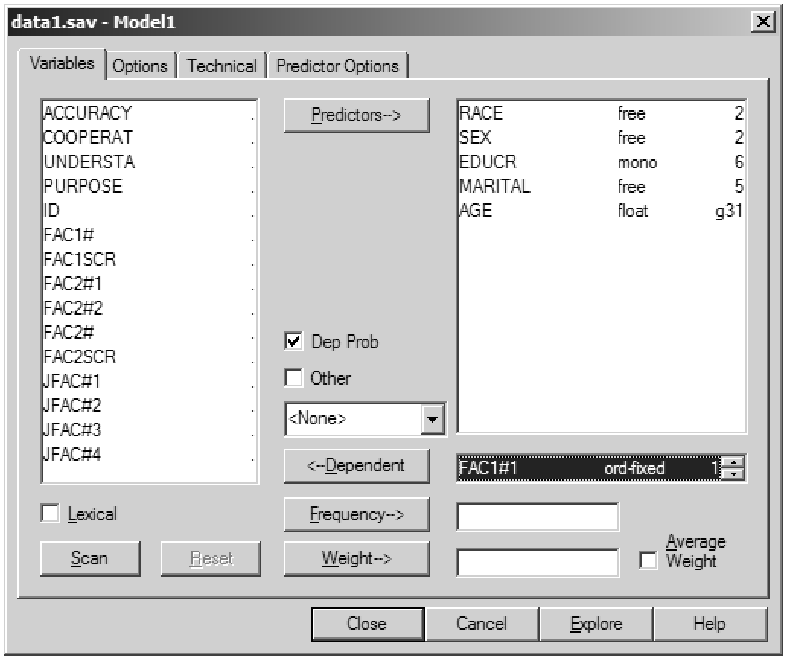
```
RACE      free     .
SEX       free     .
EDUCR     mono     .
MARITAL   free     .
AGE       mono     g30
```

```
FAC1#1    ord-fixed   .
```

☑ Dep Prob
□ Other
<None>
□ Lexical
□ Average Weight

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


data1.sav - Model1

```
RACE      free     2
SEX       free     2
EDUCR     mono     6
MARITAL   free     5
AGE       float    g31
```

```
FAC1#1    ord-fixed   1
```

☑ Dep Prob
□ Other
<None>
□ Lexical
□ Average Weight

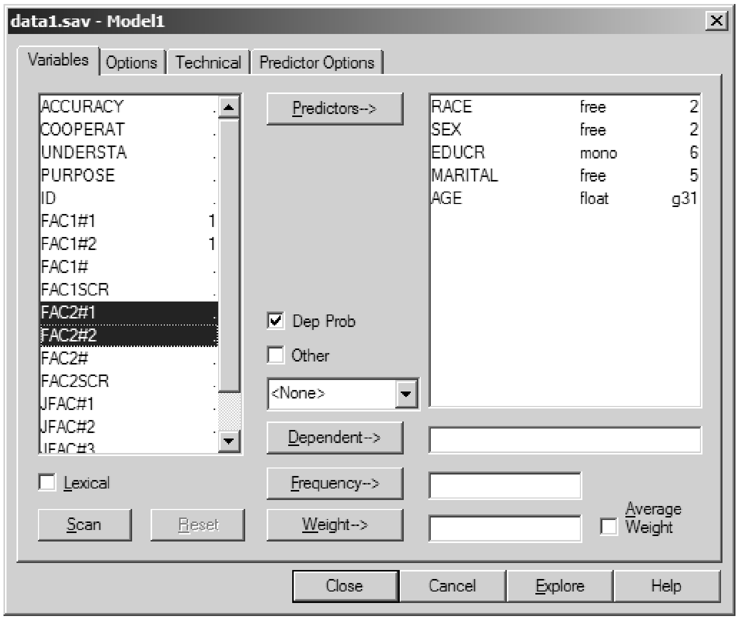
Figure 7-87. Analysis Dialog Box after Scanning

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


data1.sav - Model1

Variables | Options | Technical | Predictor Options

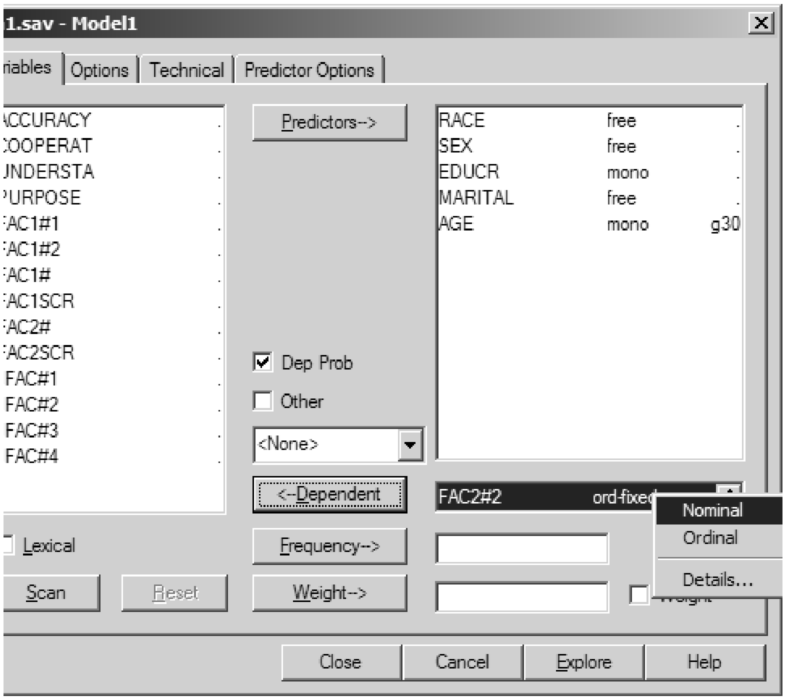
ACCURACY	.
COOPERAT	.
UNDERSTA	.
PURPOSE	.
ID	.
FAC1#1	1
FAC1#2	1
FAC1#	.
FAC1SCR	.
FAC2#1	.
FAC2#2	.
FAC2#	.
FAC2SCR	.
JFAC#1	.
JFAC#2	.
JFAC#3	

RACE	free	2
SEX	free	2
EDUCR	mono	6
MARITAL	free	5
AGE	float	g31

☑ Dep Prob
☐ Other
<None>
☐ Lexical
☐ Average Weight

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



ACCURACY
COOPERAT
UNDERSTA
PURPOSE
FAC1#1
FAC1#2
FAC1#
FAC1SCR
FAC2#
FAC2SCR
FAC#1
FAC#2
FAC#3
FAC#4

RACE	free	.
SEX	free	.
EDUCR	mono	.
MARITAL	free	.
AGE	mono	g30

☑ Dep Prob

□ Other

<None>

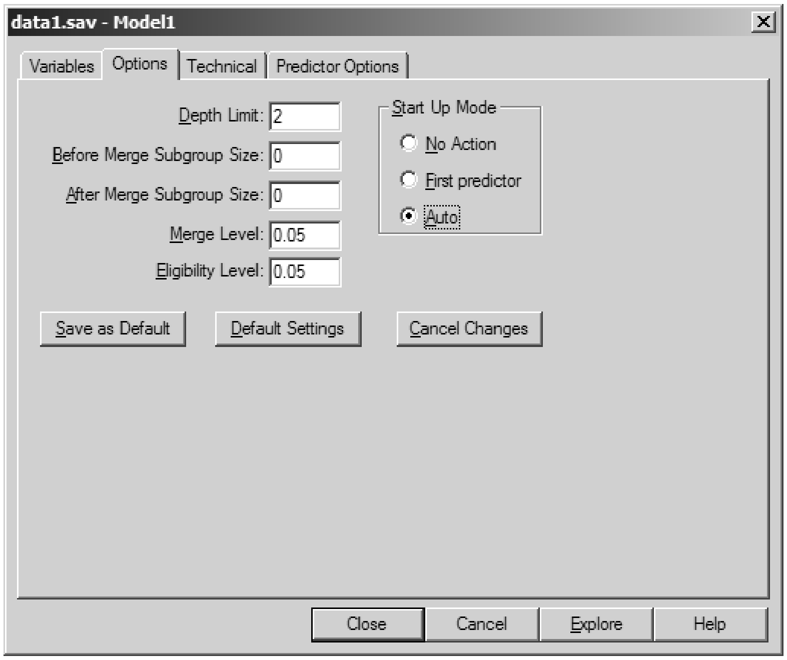
FAC2#2 ord-fixed

□ Lexical

□ Weight

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



Depth Limit: 2

Before Merge Subgroup Size: 0

After Merge Subgroup Size: 0

Merge Level: 0.05

Eligibility Level: 0.05

Start Up Mode

○ No Action

○ First predictor

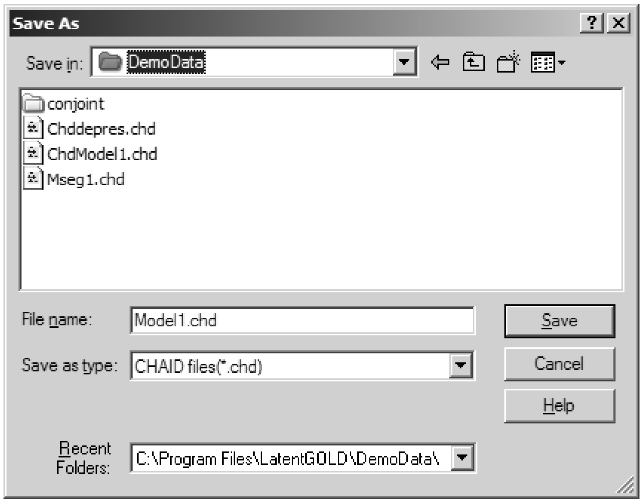
● Auto

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



Save in: DemoData

conjoint
Chddepres.chd
ChdModel1.chd
Mseg1.chd

File name: Model1.chd

Save as type: CHAID files(*.chd)

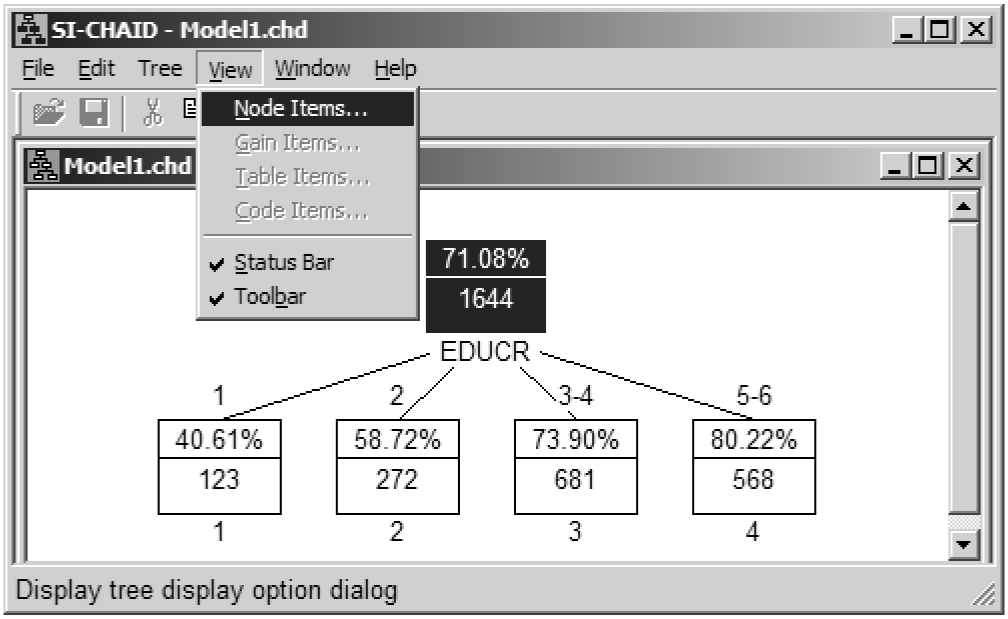
Recent Folders: C:\Program Files\LatentGOLD\DemoData\

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

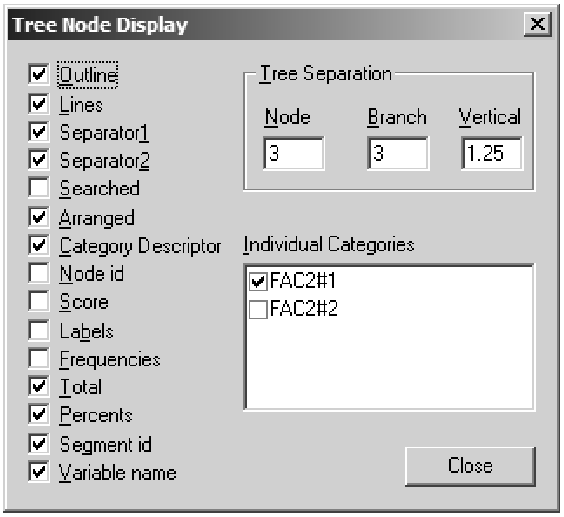


Display tree display option dialog

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



Tree Node Display

☑ Outline
☑ Lines
☑ Separator1
☑ Separator2
☐ Searched
☑ Arranged
☑ Category Descriptor
☐ Node id
☐ Score
☐ Labels
☐ Frequencies
☑ Total
☑ Percents
☑ Segment id
☑ Variable name

Tree Separation

Node	Branch	Vertical
3	3	1.25

Individual Categories

☑ FAC2#1
☐ FAC2#2

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

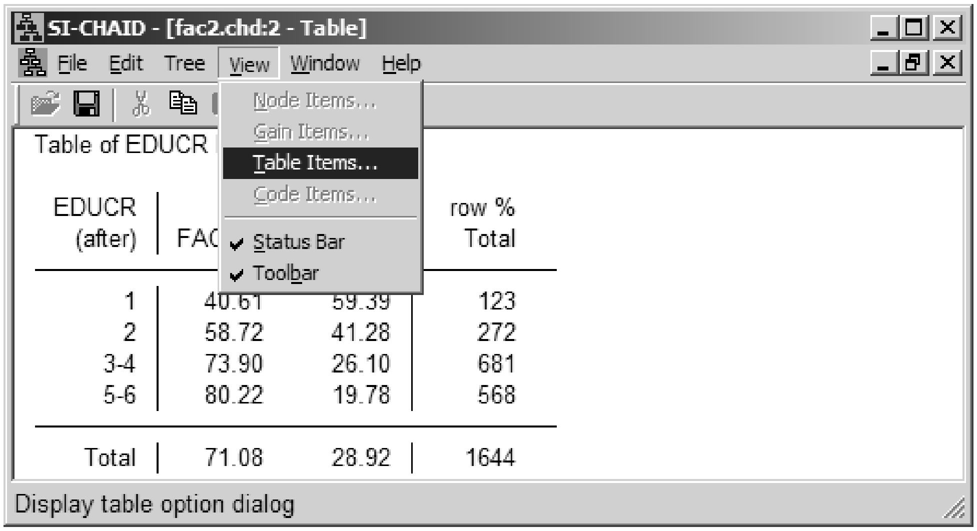


Table of EDUCR

EDUCR (after)	FAC		row % Total
1	40.61	59.39	123
2	58.72	41.28	272
3-4	73.90	26.10	681
5-6	80.22	19.78	568
Total	71.08	28.92	1644

Display table option dialog

Figure 7-94. Table Display

From the Table Display

- ➢ Select Before Merge
- ➢ Select Column Percents

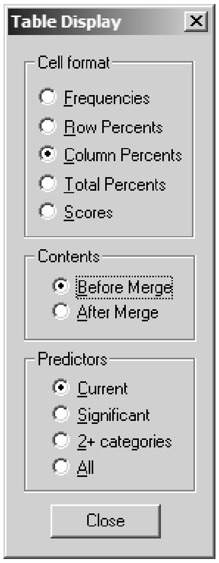


Table Display

Cell format
○ Frequencies
○ Row Percents
● Column Percents
○ Total Percents
○ Scores

Contents
● Before Merge
○ After Merge

Predictors
● Current
○ Significant
○ 2+ categories
○ All

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).

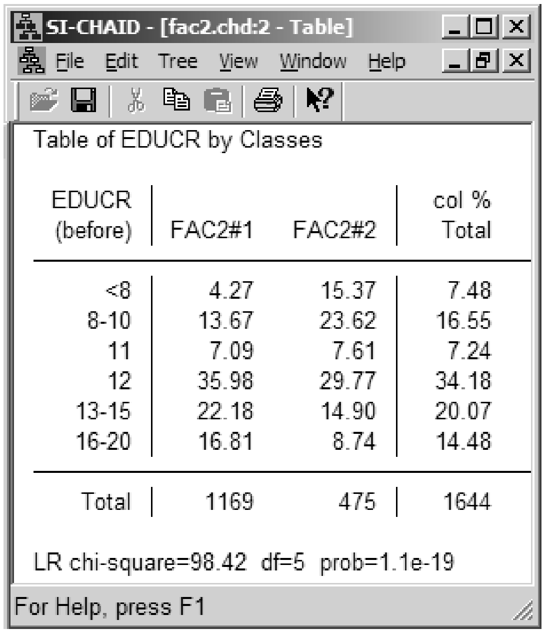


Table of EDUCR by Classes

EDUCR (before)	FAC2#1	FAC2#2	col % Total
<8	4.27	15.37	7.48
8-10	13.67	23.62	16.55
11	7.09	7.61	7.24
12	35.98	29.77	34.18
13-15	22.18	14.90	20.07
16-20	16.81	8.74	14.48
Total	1169	475	1644

LR chi-square=98.42 df=5 prob=1.1e-19

For Help, press F1

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.

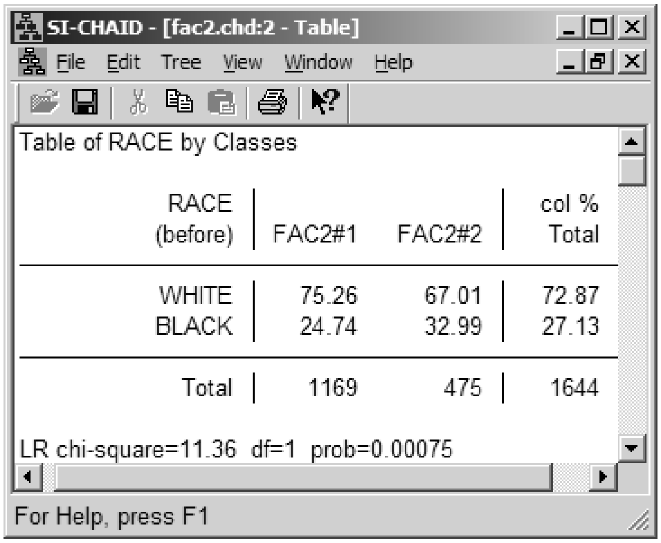


Table of RACE by Classes

RACE (before)	FAC2#1	FAC2#2	col % Total
WHITE	75.26	67.01	72.87
BLACK	24.74	32.99	27.13
Total	1169	475	1644

LR chi-square=11.36 df=1 prob=0.00075

For Help, press F1

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.

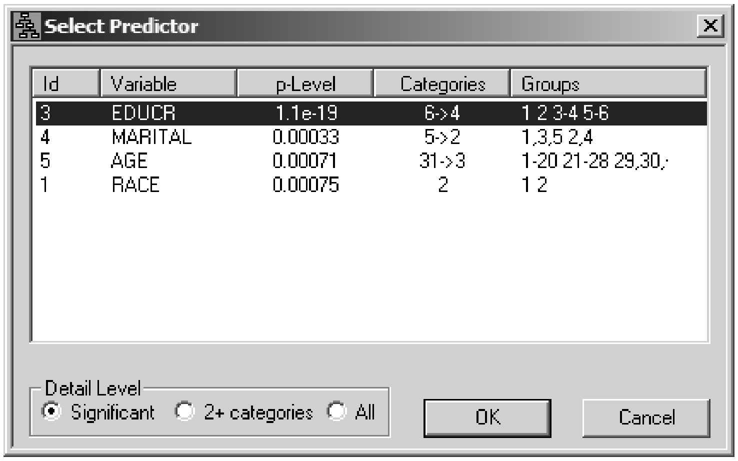
In summary, we found that none of the demographics are significantly related to DFactor1 and that EDUCR is the most important descriptor for profiling DFactor2. Given the results from tutorial #2 that showed that DFactor1 is related to the dependent variables PURPOSE and ACCURACY, while DFactor2 is related 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.



Select Predictor

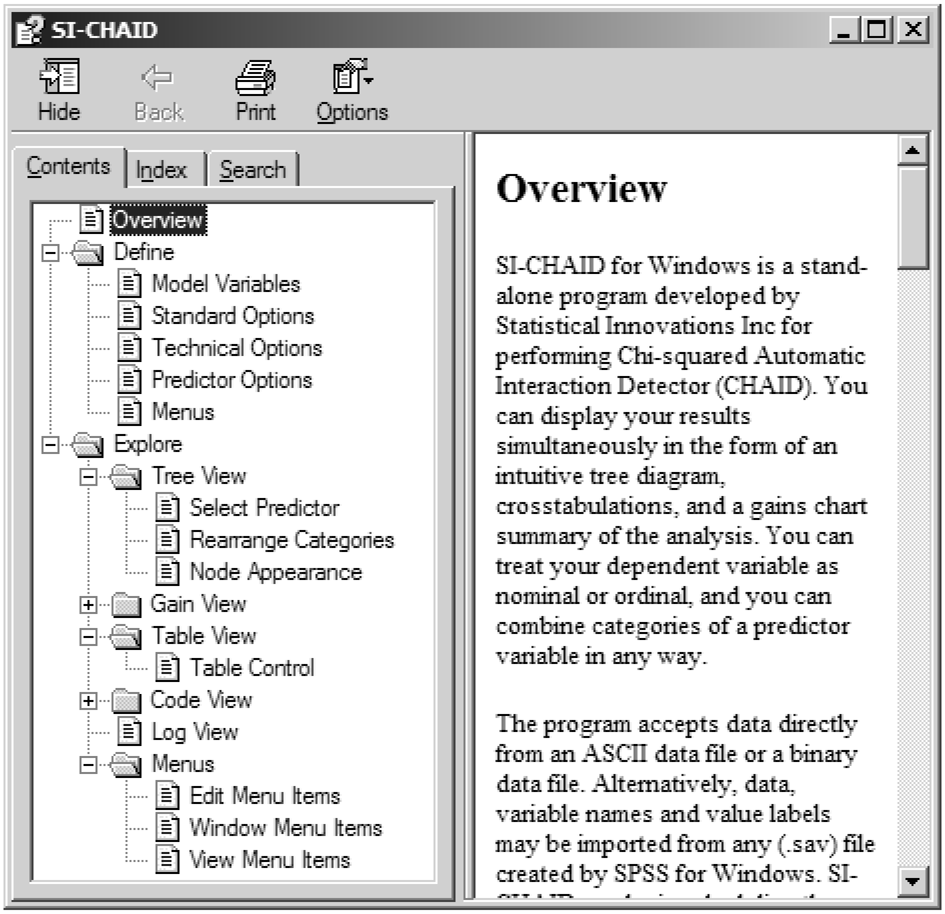
Id	Variable	p-Level	Categories	Groups
3	EDUCR	1.1e-19	6->4	1 2 3-4 5-6
4	MARITAL	0.00033	5->2	1,3,5 2,4
5	AGE	0.00071	31->3	1-20 21-28 29,30,·
1	RACE	0.00075	2	1 2

Detail Level
● Significant ○ 2+ categories ○ All

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



Overview

SI-CHAID for Windows is a stand-alone program developed by Statistical Innovations Inc for performing Chi-squared Automatic Interaction Detector (CHAID). You can display your results simultaneously in the form of an intuitive tree diagram, crosstabulations, and a gains chart summary of the analysis. You can treat your dependent variable as nominal or ordinal, and you can combine categories of a predictor variable in any way.

The program accepts data directly from an ASCII data file or a binary data file. Alternatively, data, variable names and value labels may be imported from any (.sav) file created by SPSS for Windows. SI-

Figure 7-99. SI-CHAID Help Contents