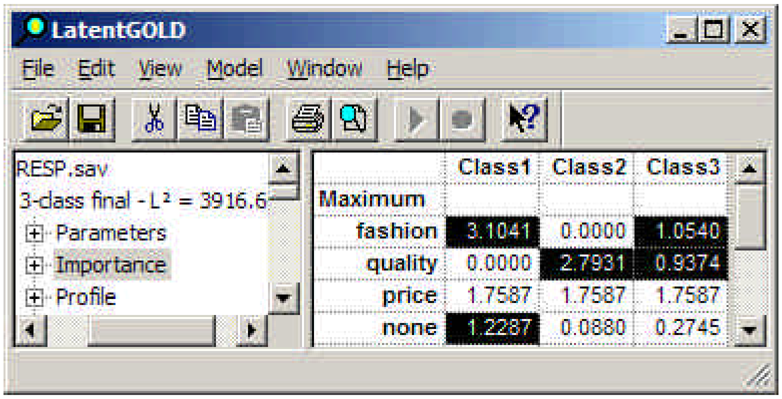
Tutorial 1A: Using CHAID to Profile Latent Class Segments

In Latent GOLD Choice Tutorial #1, 3 LC segments were obtained. The Importance output is one way to examine differences between these segments. As shown below, the first LC segment, mostly values FASHION over QUALITY, while the reverse is true for LC segment #2. LC segment #3 values both FASHION and QUALITY about the same, and all are equally price sensitive. In addition, segment #1 was most likely to choose the 'None' option, when presented with the different choice alternatives. These patterns adequately reflected the assumed utility structure that was used to generate the data.



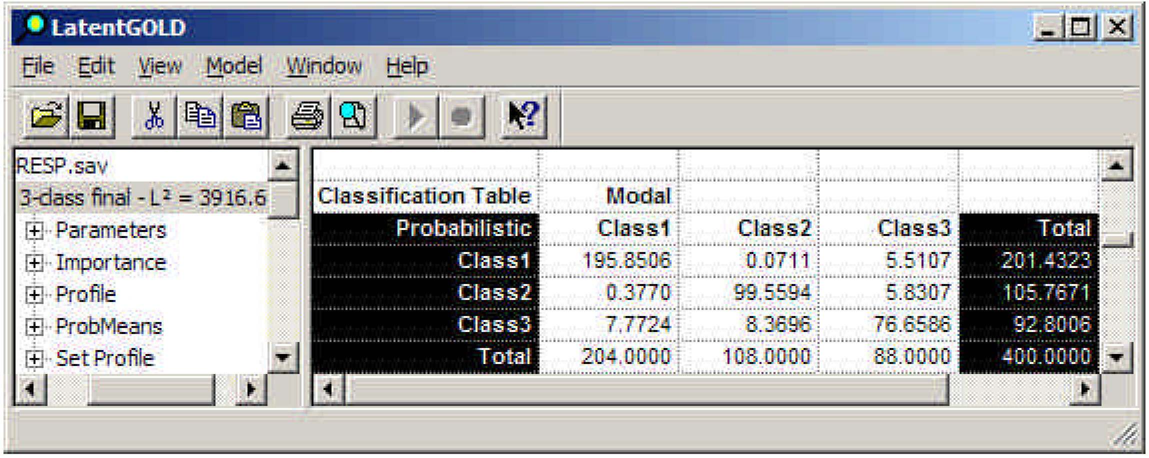
	Class1	Class2	Class3
Maximum			
fashion	3.1041	0.0000	1.0540
quality	0.0000	2.7931	0.9374
price	1.7587	1.7587	1.7587
none	1.2287	0.0880	0.2745

Figure 1. Importance Output

The Classification Table shows how well the modal assignment rule reproduces the actual (probabilistic) LC segments, the latter based on the model parameters. The large quantities in the diagonal table entries indicate that overall, the modal class assignment does quite well. Only a small number of misclassifications are shown in the off-diagonal cells.

Note that overall, the modal classifications do not reproduce perfectly the marginal distribution of the classes. For example, of the 400 cases, 201.4323 (50.36%) would be expected to be in class #1. Using the modal classification, 204 cases are so classified.

Figure 2. Classification Table



Classification Table	Modal			
Probabilistic	Class1	Class2	Class3	Total
Class1	195.8506	0.0711	5.5107	201.4323
Class2	0.3770	99.5594	5.8307	105.7671
Class3	7.7724	8.3696	76.6586	92.8006
Total	204.0000	108.0000	88.0000	400.0000

It is also useful to profile the LC segments in terms of exogenous variables (covariates). The CHAID option for doing this utilizes the posterior membership probabilities as weights and hence will reproduce the actual (Probabilistic, not Modal) classes. Thus, the CHAID approach has zero misclassification error.

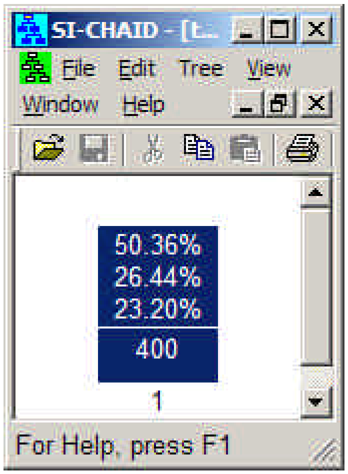
Growing the CHAID Tree

SI-CHAID consists of 2 programs, called ‘CHAID Define’ and ‘CHAID Explore’. Typically, the Define program is used first to set the analysis options and then the Explore command is executed to perform the CHAID analysis.

In this tutorial, we will first use the Explore program with the default settings in the .chd file generated in tutorial #1, to examine demographic profiles associated with all 3 segments, in a way that is somewhat analogous to a 3-group discriminant analysis. We will then use the Define program to examine demographic profiles associated with each segment (vs. the others), one at a time.

- ➢ Open SI-CHAID Explore

- ➢ File Open ‘Tutor 1a.chd’



50.36%
26.44%
23.20%
400
1

For Help, press F1

Figure 3. Root Node for Tutor1a.chd

The root node of the tree appears, with the overall sizes of the 3 LC segments.

- ➢ From the Tree menu, select Auto.

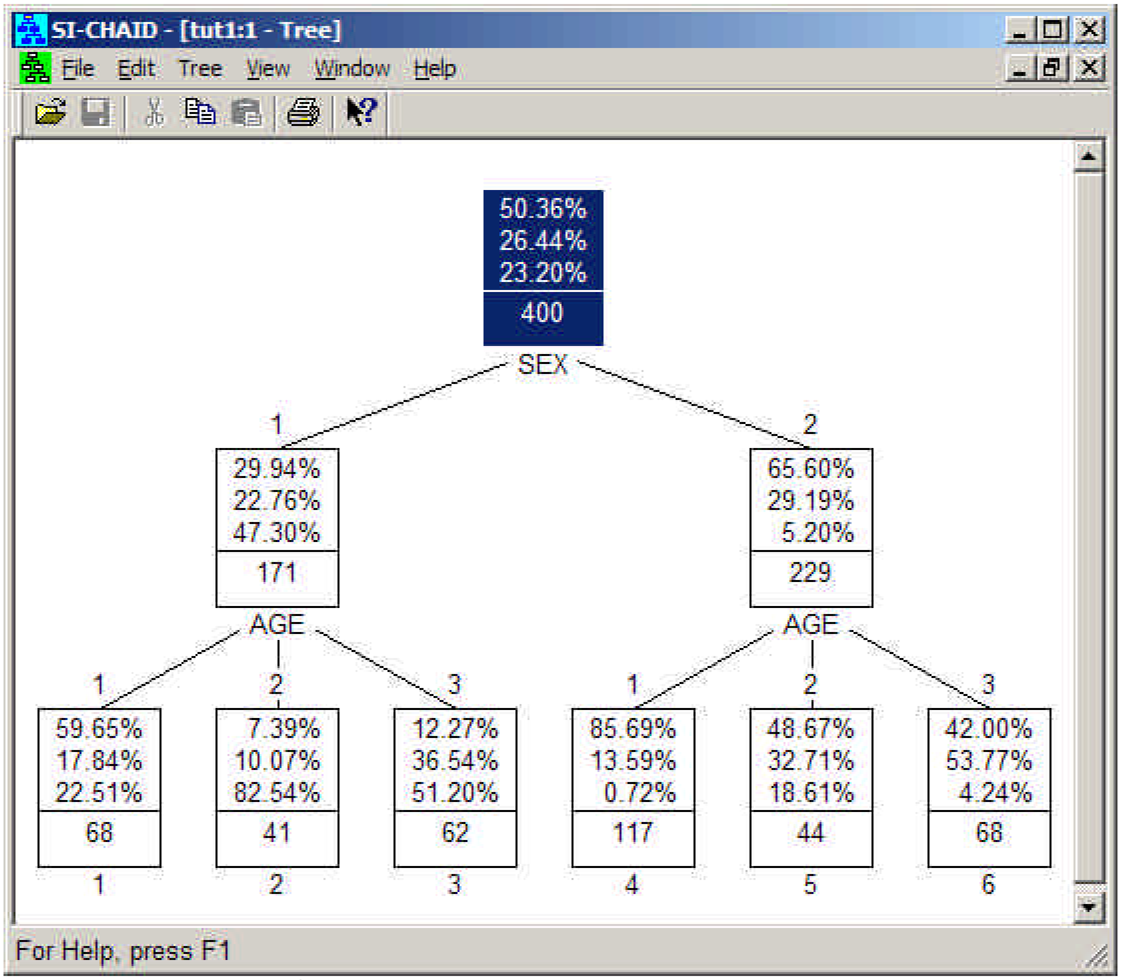


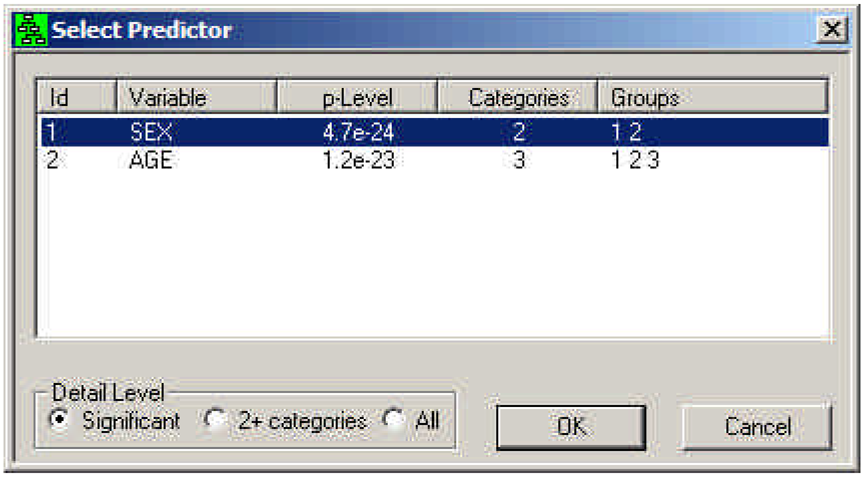
Figure 4. Tree Diagram after Auto

The CHAID analysis first splits on SEX, then on AGE, resulting in 6 demographic segments.

To examine the significance associated with these predictors at the root node of the tree,

- ➢ From the Tree menu, choose Select

Figure 5. Select Predictor Dialog Box



Id	Variable	p-Level	Categories	Groups
1	SEX	4.7e-24	2	1 2
2	AGE	1.2e-23	3	1 2 3

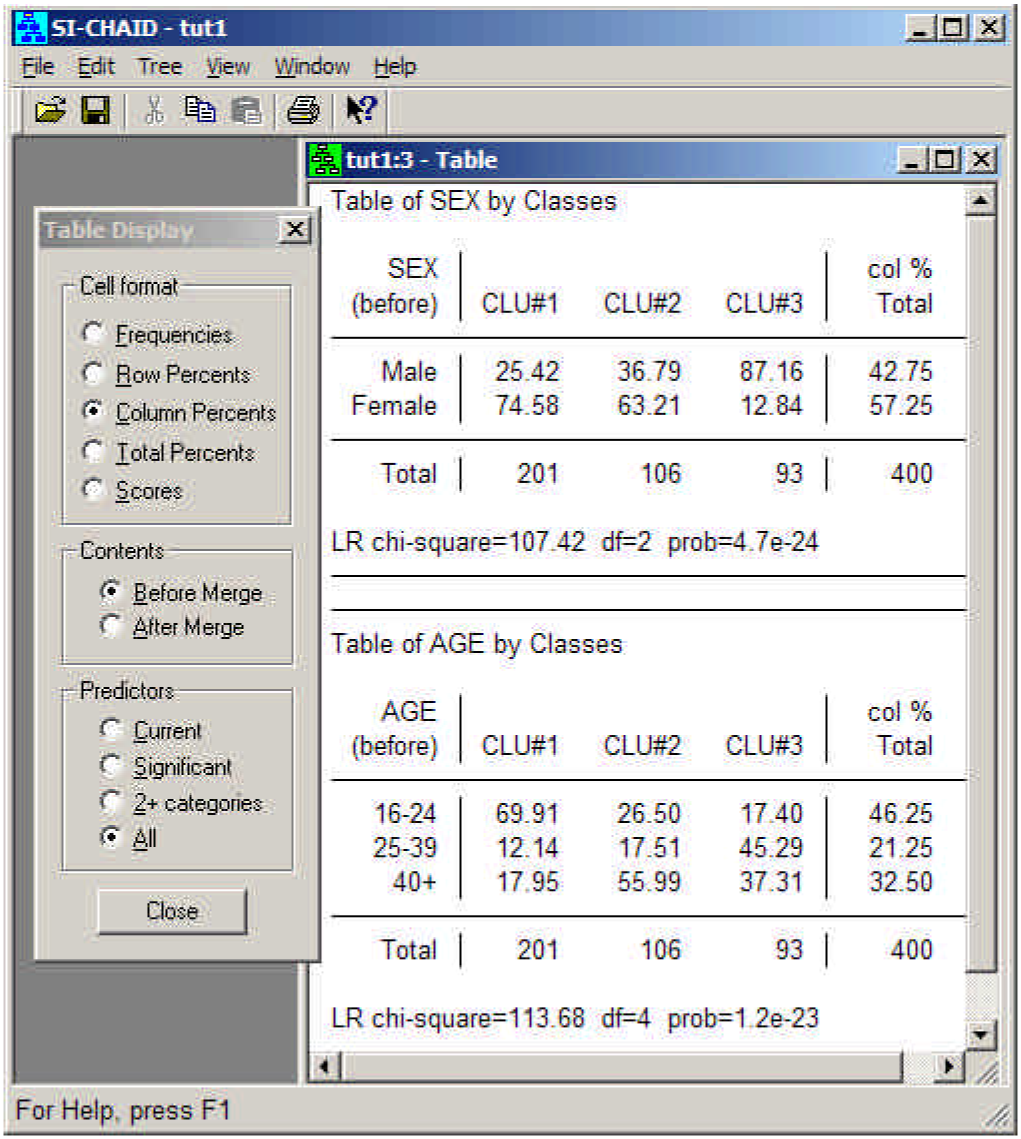
Detail Level
● Significant ○ 2+ categories ○ All

We see that both SEX and AGE are highly significant. For example, for SEX, $p=4.7 \times 10^{-24}$. Also, for AGE, the 3 categories are all significantly different from each other, as indicated by a space between each ("1 2 3"), and thus none of these categories would be combined in growing the first level of the tree.

Note that these p-values indicate that the relationships between these covariates and the segments are stronger than those indicated by the Wald tests in the Parameters output. That is because the LC variable is treated as an observed variable in CHAID; the posterior membership probabilities are treated as fixed weights, with no error.

To reproduce the Profile output tables obtained from Latent GOLD Choice

- From the Windows menu, select New Table
- Right click on the table to retrieve the Table Display Panel
- From the Cell format section select 'Column Percents'
- From the Contents section select 'Before Merge'
- From the Predictors section, select 'All' to display both tables



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

Contents
● Before Merge
○ After Merge

Predictors
○ Current
○ Significant
○ 2+ categories
● All

```
Table of SEX by Classes

     SEX |                           |  col %
(before) |  CLU#1    CLU#2    CLU#3  |  Total
------------------------------------------------
    Male |  25.42    36.79    87.16  |  42.75
  Female |  74.58    63.21    12.84  |  57.25
------------------------------------------------
   Total |    201      106       93  |    400

LR chi-square=107.42  df=2  prob=4.7e-24

Table of AGE by Classes

     AGE |                           |  col %
(before) |  CLU#1    CLU#2    CLU#3  |  Total
------------------------------------------------
   16-24 |  69.91    26.50    17.40  |  46.25
   25-39 |  12.14    17.51    45.29  |  21.25
     40+ |  17.95    55.99    37.31  |  32.50
------------------------------------------------
   Total |    201      106       93  |    400

LR chi-square=113.68  df=4  prob=1.2e-23
```

Figure 6. Tables for SEX and AGE with Table Display Panel

We see for example, that segment 1 (labeled 'CLU#1), has a higher percentage of the youngest (AGE = 16-24) and the highest percentage female of the 3 segments.

Next, we will use the SI-CHAID Define program to change the default options to develop a separate profile for this segment vs. the other 2.

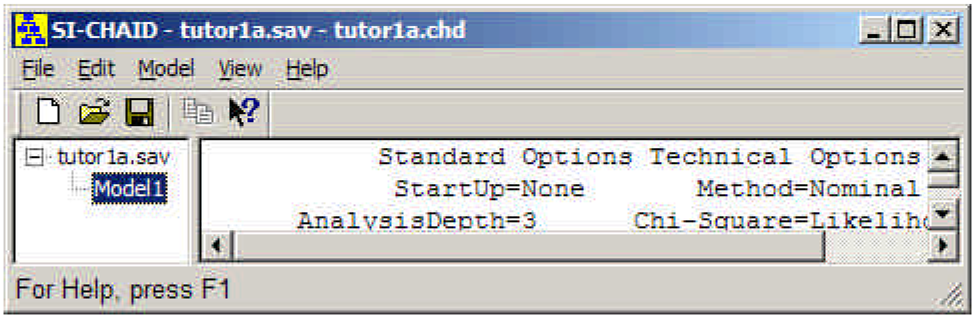
Using CHAID to Profile Segment 1 vs. Others

➢ Open the CHAID Define program.

From the File Menu

➢ Select 'Open' 'Tutor 1a.chd'

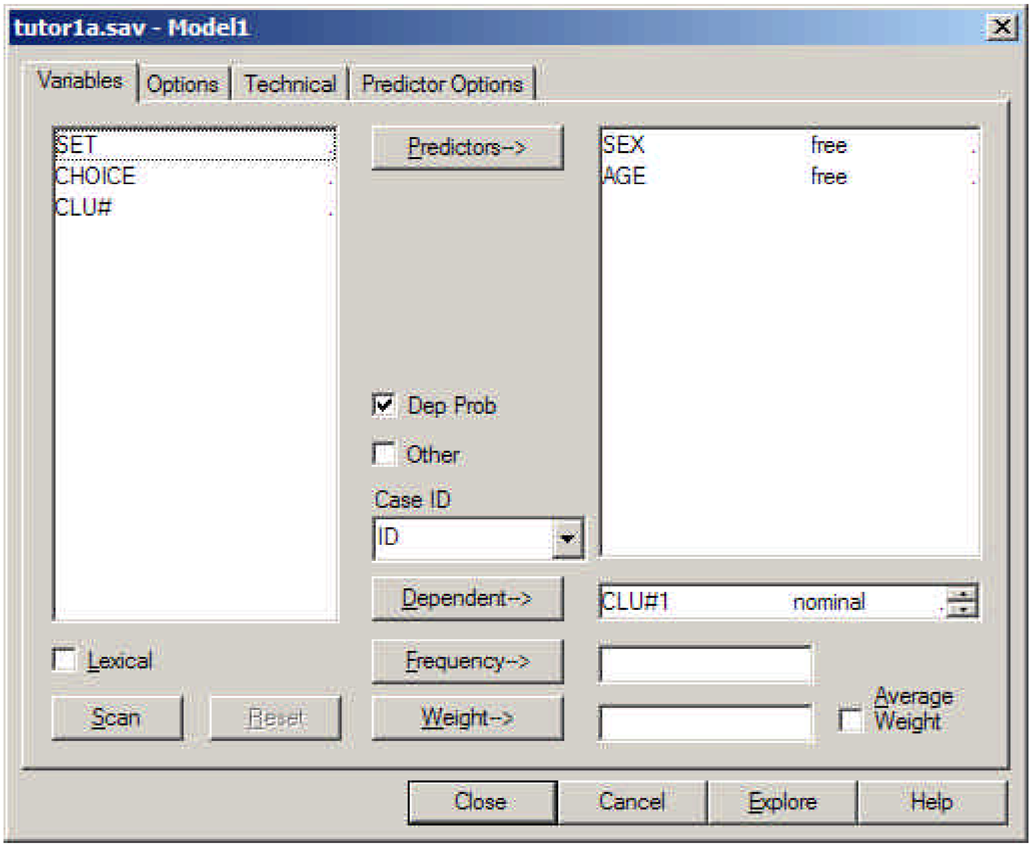
The default settings are displayed in the Contents Pane



```
Standard Options Technical Options
     StartUp=None        Method=Nominal
AnalysisDepth=3     Chi-Square=Likelih
```

Figure 7. Default Settings for Tutor 1a.chd in the CHAID Define program.

➢ Double click on Model1 to open the Analysis Dialog box



SET
CHOICE
CLU#

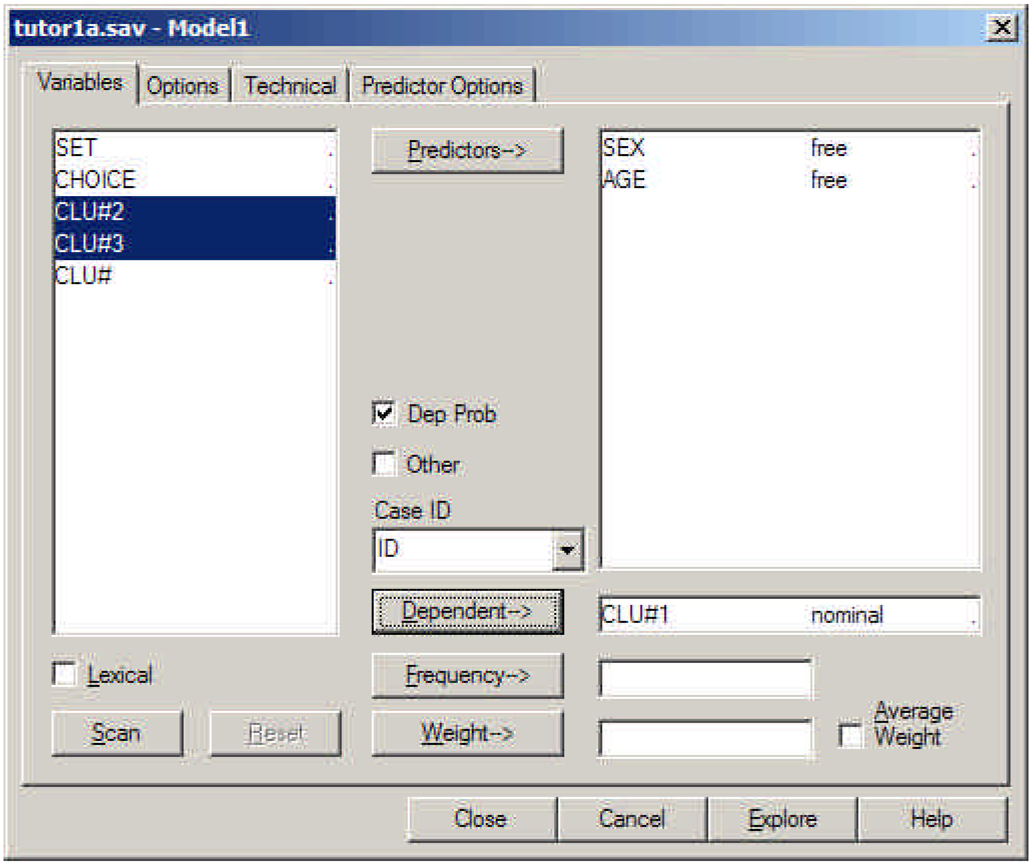
SEX	free
AGE	free

☑ Dep Prob
□ Other
Case ID: ID
CLU#1 nominal
□ Lexical
□ Average Weight

Figure 8: The Analysis Dialog Box

Notice that a checkmark appears next to the 'Dep Prob' box. This indicates that posterior membership probabilities are used to specify the categories of the dependent variable. For this 3-class model, the posterior probabilities obtained from Latent GOLD Choice are labeled 'CLU#1', 'CLU#2' and 'CLU#3'. Only CLU#1 is visible in the Dependent box.

- ➢ In the Dependent box, scroll down to display all three variables
- ➢ Select 'CLU#2' and 'CLU#3' and click the Dependent Box to restore them to the Variables List box



SET
CHOICE
CLU#2
CLU#3
CLU#

SEX	free
AGE	free

☑ Dep Prob

□ Other

Case ID: ID

CLU#1	nominal

□ Lexical

□ Average Weight

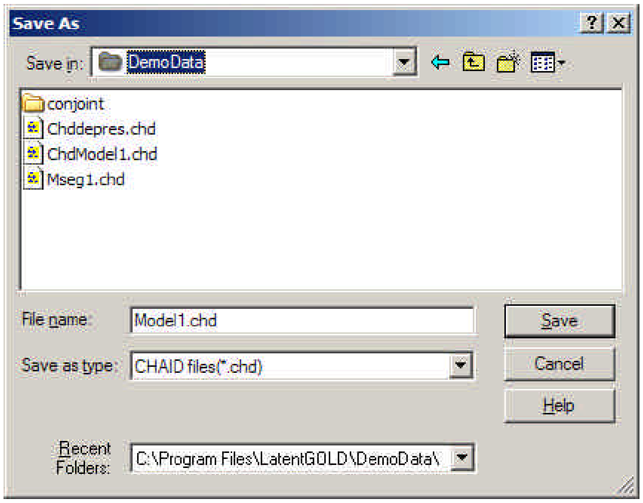
Figure 9. CLU#2' and 'CLU#3 restored to the Variables List Box

- ➢ Click to add a Checkmark to the 'Other' box. This tells SI-CHAID to create a second category for the dependent variable consisting of all other categories – in this case, categories for classes 2 and 3.

To grow the tree:

- ➢ Click 'Explore'

CHAID prompts you to save the updated definition file named Model1.chd (the default name)



Save in: DemoData

File name: Model1.chd

Save as type: CHAID files(*.chd)

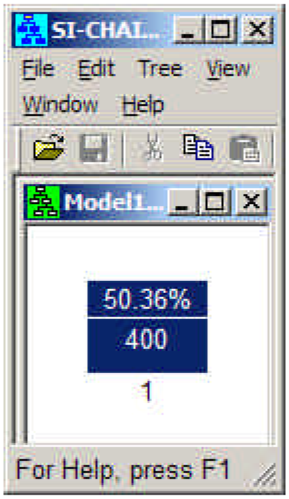
Recent Folders: C:\Program Files\LatentGOLD\DemoData\

Figure 10. Save File 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

Control passes back to the Explore program and the root node appears as follows:



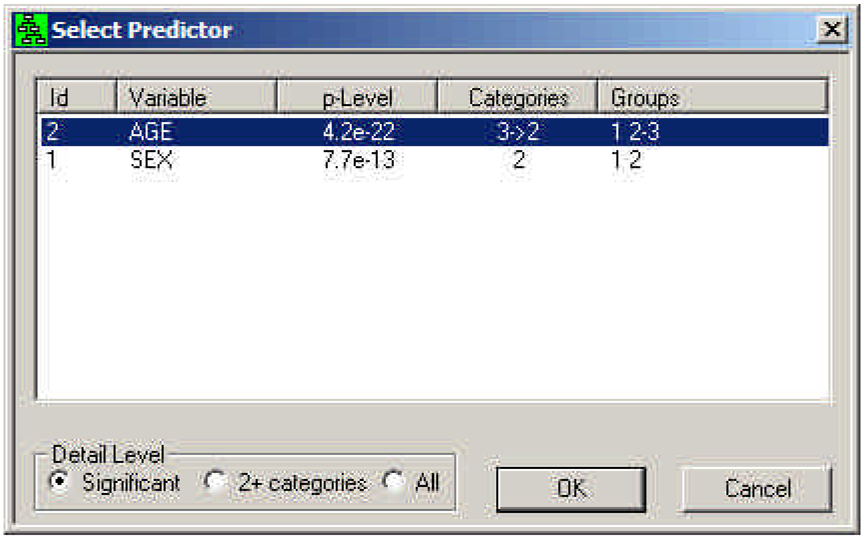
For Help, press F1

Figure 11: Segmentation Tree Nodes Showing the % in each Latent Class

.By default when the dependent variable contains only 2 categories, only the percentage associated with category 1 is displayed in each tree node.

To begin to grow the tree interactively,

- From the Tree menu choose ‘Select’



Id	Variable	p-Level	Categories	Groups
2	AGE	4.2e-22	3->2	1 2-3
1	SEX	7.7e-13	2	1 2

Detail Level: ● Significant ○ 2+ categories ○ All

Figure 11. Select Predictor Dialog Box

Now we see that AGE is more significant than SEX for distinguishing LC segment 1 from the others. In addition, we see that CHAID has combined age levels 2 and 3 because they are not significantly different in predicting the probability of being in LC segment 1.

➢ Click the OK button to use AGE for the first split

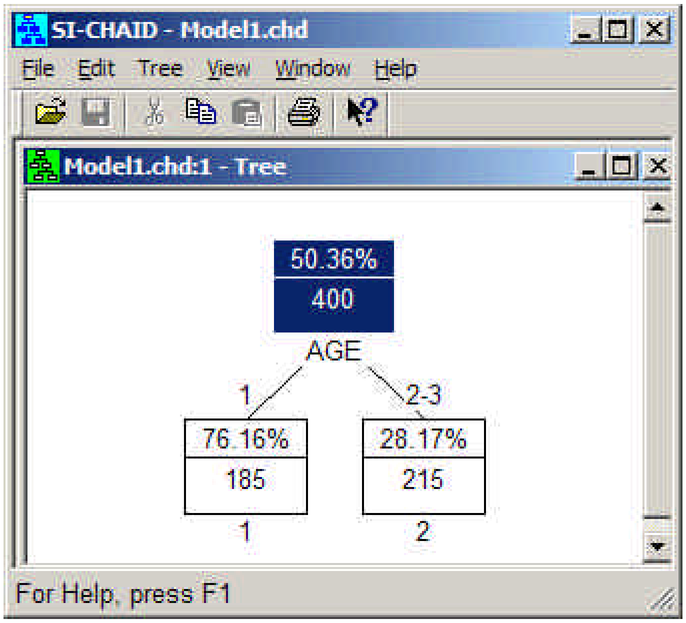


Figure 12. Tree Diagram after selecting AGE as first split

As was suggested earlier, the primary difference between segment 1 and the others is that this segment contains a much higher percentage of the youngest age range (aged 16-24). The tree indicates that among this youngest group (node 1), 76% are in segment 1, compared to only 28% of the older respondents (node 2).

➢ To continue the analysis of the youngest respondents in interactive mode:

- Click on node 1 to make it active
- From the Tree menu choose ‘Select’
- Select OK to split on SEX

Your tree now looks like this:

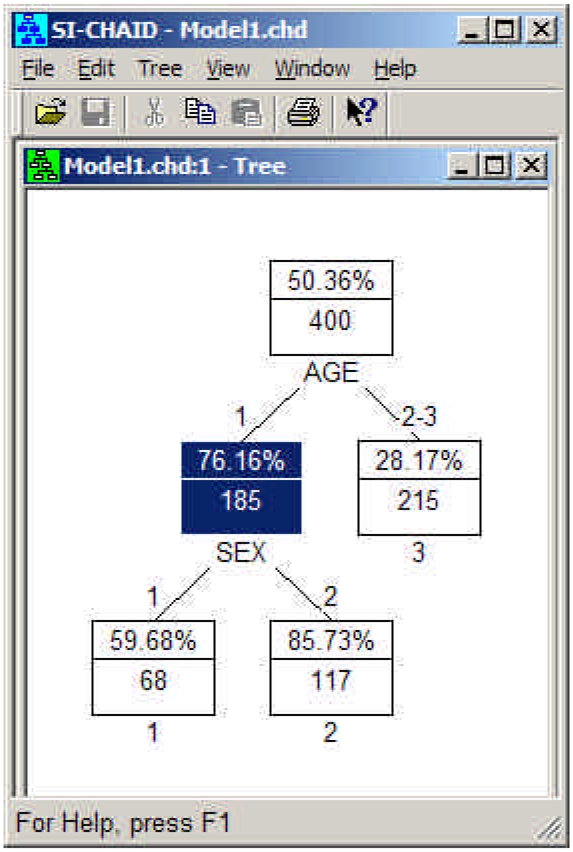


Figure 13. Tree Diagram after selecting SEX in Node 1

We see that 85.73% of young females (AGE = 1, SEX = 2) are in LC segment 1.

- Click on the node corresponding to the other categories of AGE (node 3) to make it active
- From the Tree menu choose ‘Select’
- Select OK to split on SEX

Our final tree now looks like this:

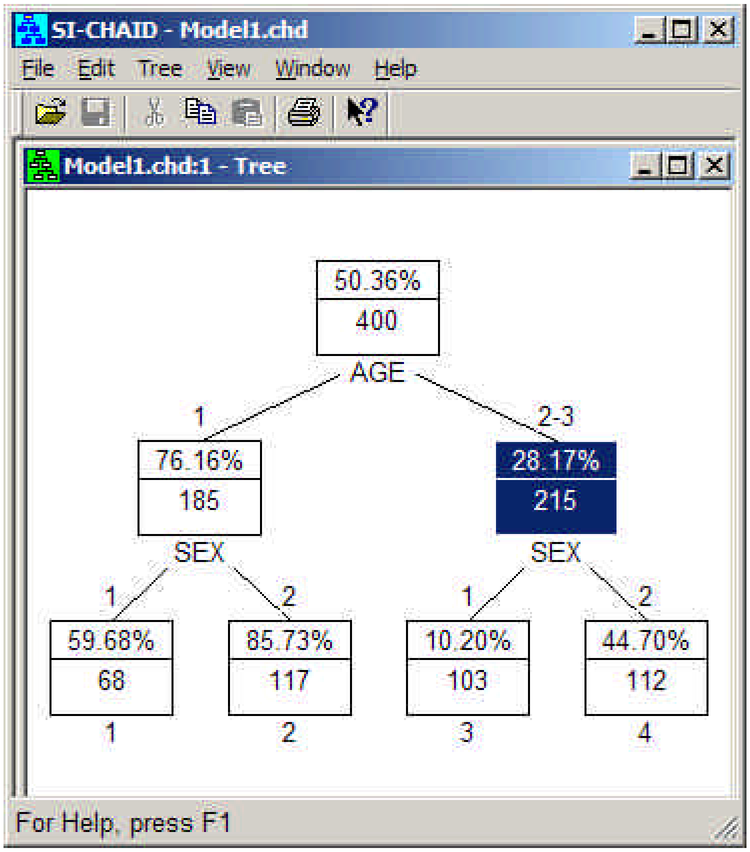


Figure 14. Tree Diagram after selecting SEX in Node 3

Thus, only 4 demographic segments are required to profile segment 1 vs. the others.

We can repeat these steps to obtain separate profiles for LC segments 2 and 3.