Tutorial #1: Using Latent GOLD choice to Estimate Discrete Choice Models

In this tutorial, we analyze data from a simple choice-based conjoint (CBC) experiment designed to estimate market shares (choice shares) for shoes. In this tutorial you will:

- Set up an analysis
- Estimate choice models that specify different numbers of classes (segments)
- Explore which of these models provides the best fit to the data
- Utilize restrictions to refine the best fitting model
- Interpret results using our 'final' model
- Save results

In tutorial 2, the final model will be used to

- Predict future choices
- Simulate choices among additional products of interest

The Data

Latent GOLD Choice accepts data from an optional 1-file or its default 3-file structure from an SPSS .sav file, a Sawtooth .cho file, or ASCII rectangular file format. The current sample data utilizes 3 SPSS .sav files.

The 10 pairs of shoes included in this choice experiment differ on 3 attributes: Fashion (0=Traditional, 1= Modern), Quality (0=Standard, 1=Higher), and Price (5 equidistant levels coded from 1 to 5). Each of 8 choice sets offer 3 of these 10 possible alternative products to 400 individuals. The choice task posed to respondents is to assume that each set represents the actual options available for purchase and to select one of these alternatives from each set, with the response "none of the above" allowed as a fourth choice option.

The 11 alternatives (10 different products plus a 'None' option) are defined in terms of the 3 attributes plus the dummy variable NONE in the Alternatives File cbcALT11.sav:

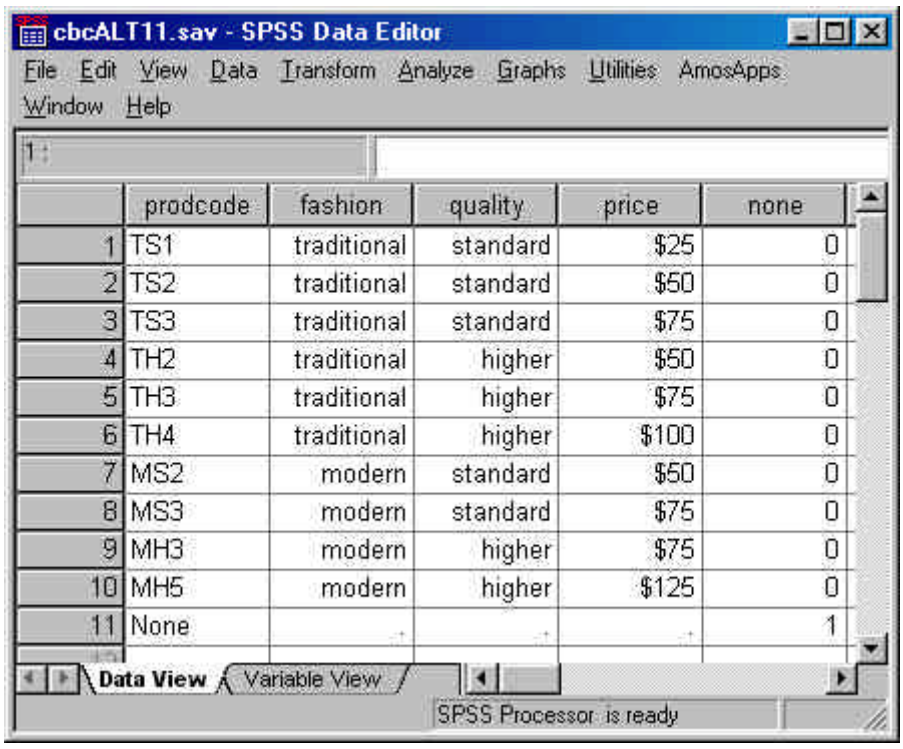
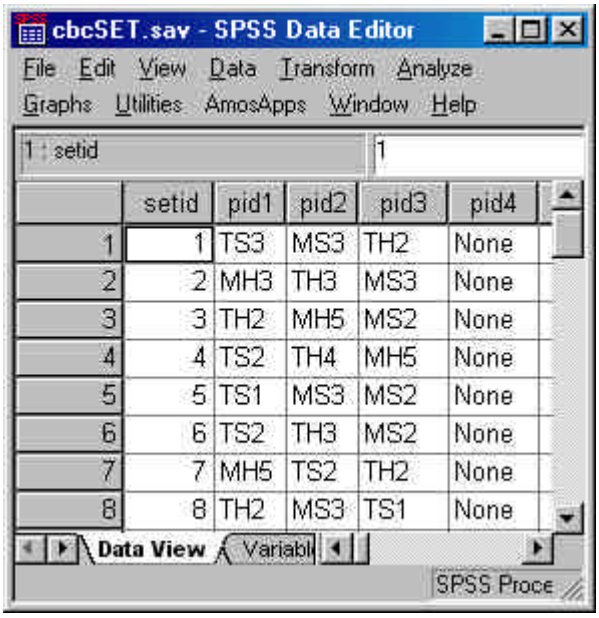


Figure 13. Alternatives File 'cbcALT11.sav' defining each of the 11 Alternatives

The specific 4-alternatives that constitute the 8 choice tasks are defined in the Sets File cbcSET.sav. For example, task 1 involves the choice between shoes TS3, MS3, TH2 or None of these.

Figure 14. Sets File 'cbcSET.sav' defining each of the 8 choice tasks



The universe is assumed to consist of 3 segments with different utility preferences. The first segment constitutes 50% of the Universe, the second and third each representing 25%.

Segment:	1	2	3
Segment Size	0.50	0.25	0.25

Part-worth Utility Parameters

Segment:	1	2	3
Modern fashion	3	0	1
Higher quality	0	3	1
PRICE	-0.5	-0.5	-0.5
NONE	1	0	-0.5

These utilities imply that all 3 segments are equally price sensitive. In addition, segment 1 (50% of the universe) prefers the more Modern style shoes and is indifferent regarding Higher quality, segment 2 (25%) prefers the Higher quality and is indifferent with regards to fashion. Segment 3 (25%) is influenced by both Modern fashion and Higher quality but to a lessor extent than the other segments. Overall, segment 1 is most likely and segment 3 least likely to choose None.

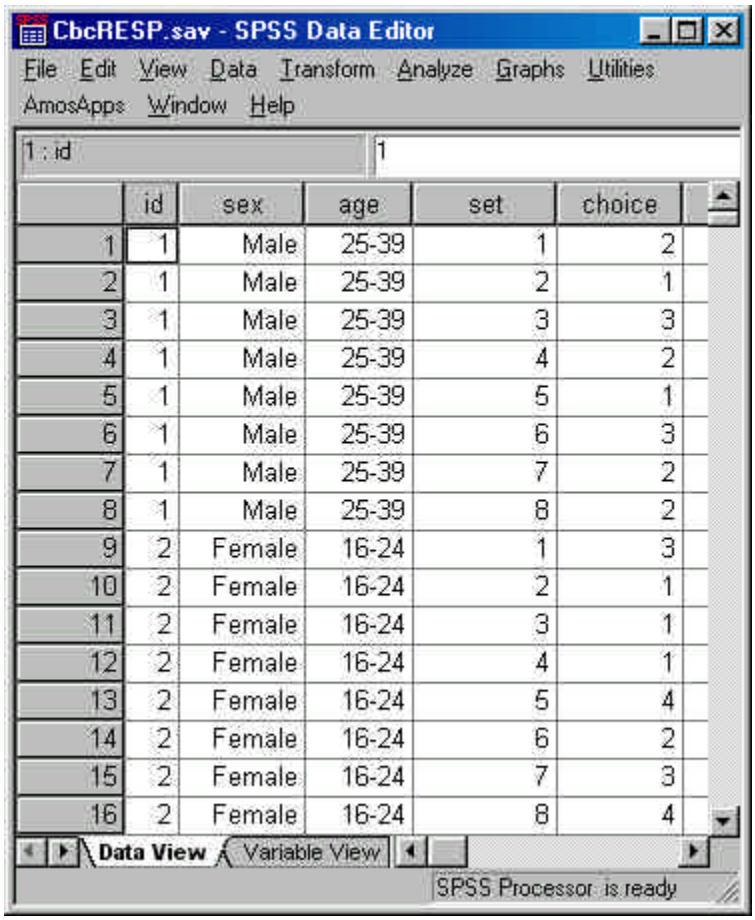
The segments differ on AGE and GENDER as follows:

Gender and Age distributions for each segment:

	Segment		
GENDER	1	2	3
Male	0.25	0.40	0.83
Female	0.75	0.60	0.17
AGE			
<25	0.69	0.28	0.19
25-39	0.14	0.15	0.43
40+	0.20	0.57	0.38

The choices obtained from 400 individuals, generated from this hypothetical Universe, are contained in the Response File cbcRESP.sav.

Figure 15. First 2 cases in the Response File 'cbcRESP.sav'



CbcRESP.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities
AmosApps Window Help

1 : id | 1

	id	sex	age	set	choice
1	1	Male	25-39	1	2
2	1	Male	25-39	2	1
3	1	Male	25-39	3	3
4	1	Male	25-39	4	2
5	1	Male	25-39	5	1
6	1	Male	25-39	6	3
7	1	Male	25-39	7	2
8	1	Male	25-39	8	2
9	2	Female	16-24	1	3
10	2	Female	16-24	2	1
11	2	Female	16-24	3	1
12	2	Female	16-24	4	1
13	2	Female	16-24	5	4
14	2	Female	16-24	6	2
15	2	Female	16-24	7	3
16	2	Female	16-24	8	4

Data View | Variable View

SPSS Processor is ready

The Goal

We wish to assess the ability of the Latent GOLD Choice program to unmix the data correctly to reflect the simulated structural relationships. (Since the design for this choice experiment has relatively low efficiency, it is unclear whether it is possible to uncover the 3 underlying segments.) The BIC statistic will be used to determine which model fits best.

Setting up the analysis

For this example, you can either open the setup file '1class.lgf' saved previously or go through the setup steps one at a time.

- To open the setup file, from the menus choose:

  File

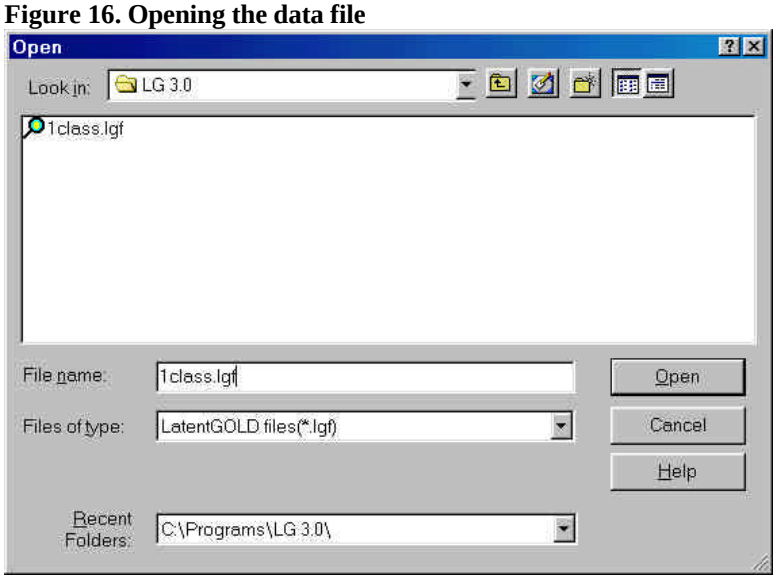
  Open

í From the Files of type drop down list, select 'LatentGOLD files (.lgf)' if this is not already the default listing.

All files with the .lgf extensions appear in the list (see Figure below)

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 16. Opening the data file



- Select '1class.lgf' and click Open.
- Double-click on Model1 to bring up the 1-class model
- In the Classes box, type '1-4' in place of 1 to request estimation of 4 models between 1 and 4 classes
- Skip ahead to *Estimating Models* (p. 11)

Alternatively, to go through the setup steps one at a time,

- from the menus choose:

File

Open

í From the Files of type drop down list, select SPSS system files (.sav) if this is not already the default listing.

All files with the .sav extensions appear in the list.

- Select 'cbcRESP.sav' and click Open

- Right click on 'Model1' and select Choice

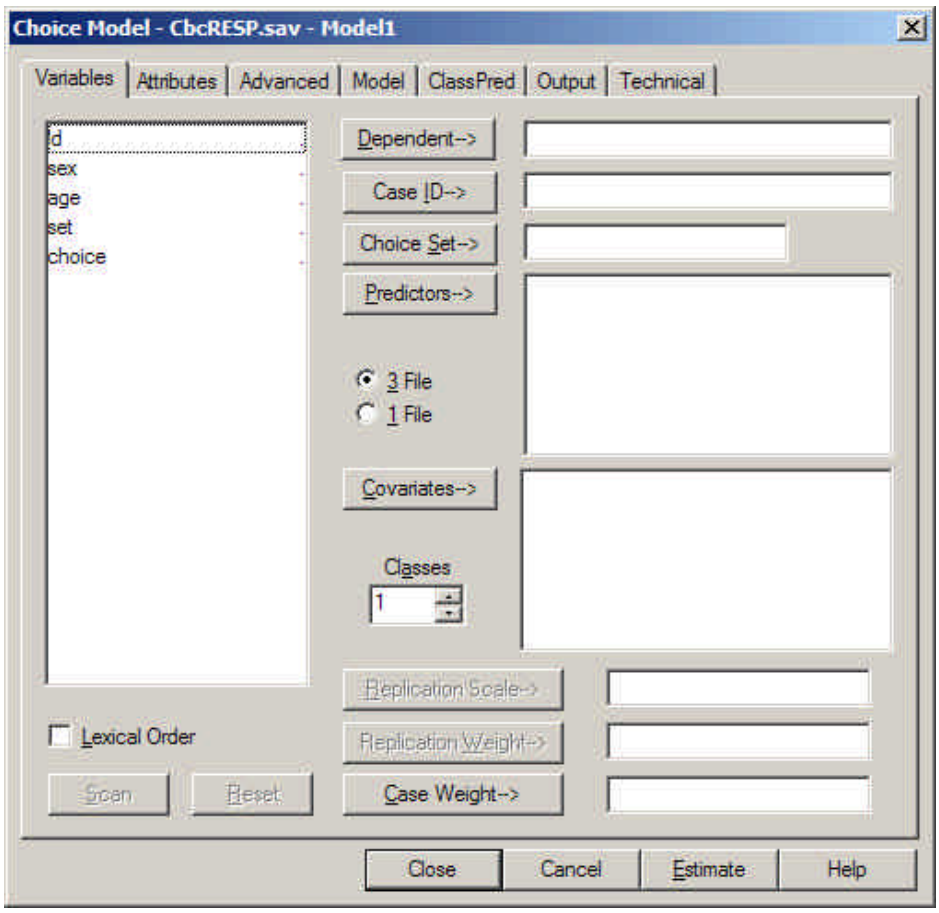


Figure 17. Model Analysis Dialog Box

- Select each variable and move it to the appropriate box by clicking the buttons to the left of these boxes.
  - ID to Case ID
  - SEX, AGE to Covariates
  - SET to Choice Set
  - CHOICE to Dependent
- In the Classes box, type '1-4' in place of 1 to request estimation of 4 models between 1 and 4 classes
- Right click on SEX and AGE and select 'Nominal' to change the scale type for these covariates.

Your Choice model setup should now look like this:

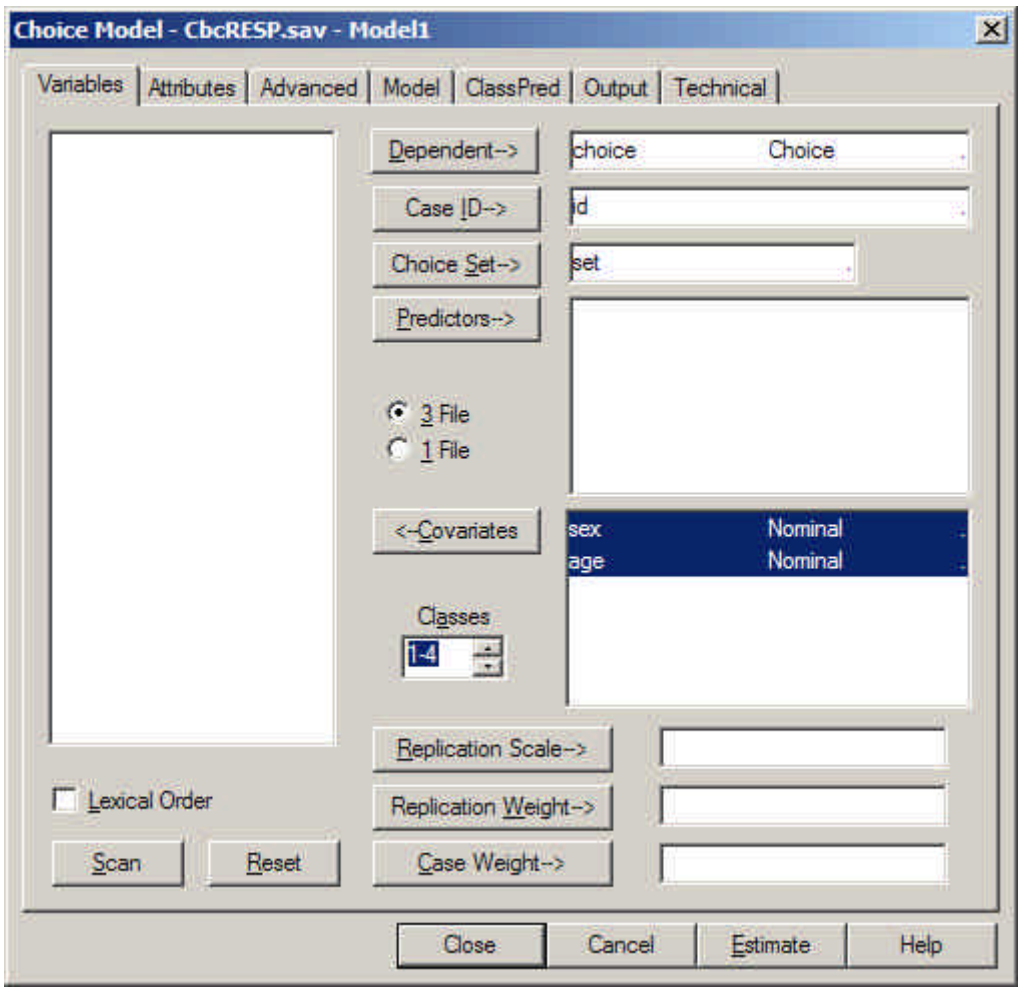


Figure 18. Model Analysis Dialog Box after setup

To complete the setup, we need to connect the Alternatives and Sets files to the Response File, and specify those attributes for which utilities are to be estimated.

- Open the Attributes tab by clicking on 'Attributes' at the top of the setup screen:

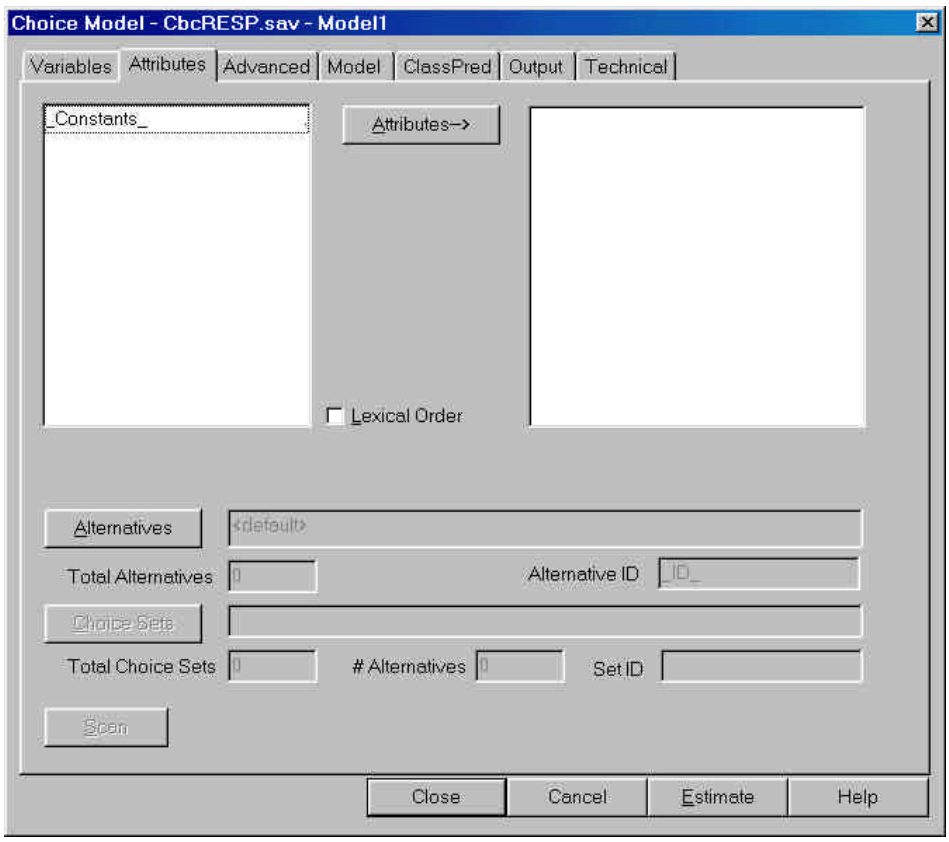


Figure 19: Attributes Tab

- Click the Alternatives button to display a list of files
- Select 'cbcALT11.sav' and click Open

In response to the prompt to select an ID variable,

- Select 'PRODCODE' and click OK

Figure 20. Select ID Variable prompt

The attribute variables from this file are now included in the variable list along with the choice specific constants variable _Constants_ , which is generated automatically by the program.

- Click the Choice Sets button to display a list of files
- Select 'cbcSET.sav' and click Open

In response to the prompt to select an ID variable,
- Select 'setid' and click OK

Selecting the Variables for the Analysis

For this analysis, we will estimate main effects for all 4 attributes (FASHION, QUALITY, PRICE, and NONE) and a quadratic price effect by including the variable PRICESQ. To select these variables:

- Select all variables except for '_Constants_' and click Attributes to move them to the Attributes list box (or double click each variable that you want to move to the Attributes box and then click *Attributes*).

The designated variables now appear in the Attributes list box.

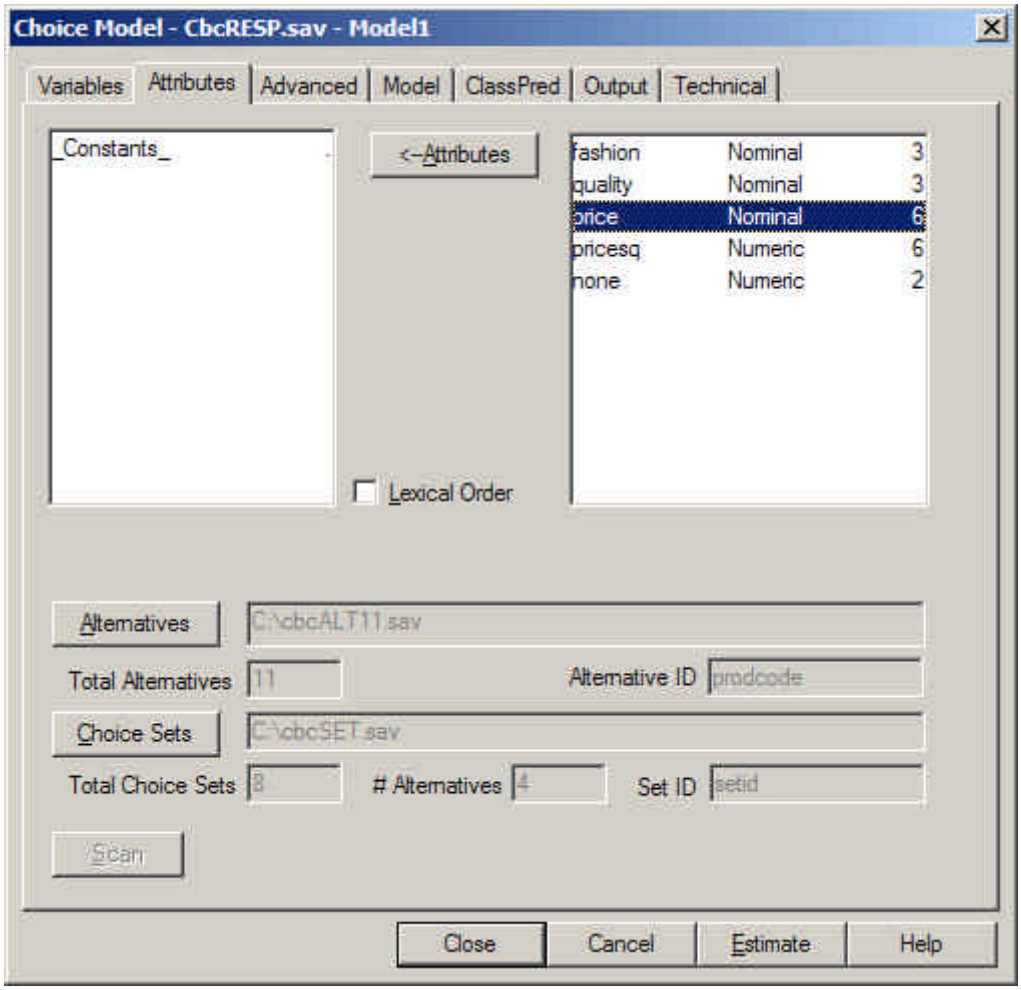


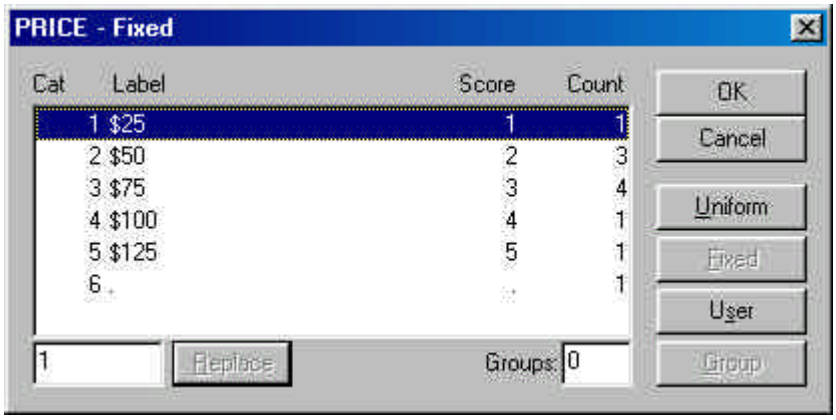
Figure 21. Selecting the Attributes

By default, character variables and variables containing consecutive integers (such as FASHION, QUALITY and PRICE) are treated as Nominal.

To view the coding for any variable, double-click on that variable.

- Double click on PRICE to view the consecutive integer scores assigned to the price levels.
- Click Cancel to close this Score window

Figure 22. PRICE Score window



To change the scale type of PRICE to numeric,

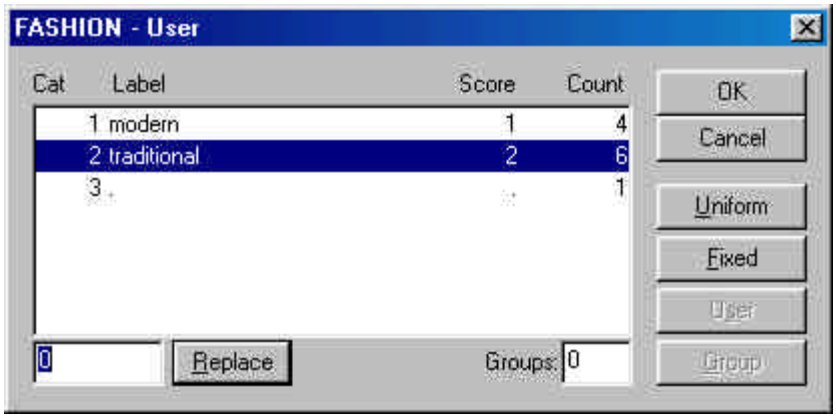
- Right click on PRICE, and select 'Numeric'

To change from effects coding (the default for Nominal variables) to '0,1' dummy coding for FASHION and QUALITY we can change the scale type to Numeric and change the scores to 0,1.

- Select FASHION and QUALITY
- Right click and select Numeric

- Double click on FASHION to retrieve the score box
- Double click on the second category
- Replace '2' with '0' in the Replace box
- Click Replace
- Click OK

Figure 23. FASHION Score window



Repeat he process for QUALITY

- Double click on QUALITY to retrieve the score box
- Double click on the second category
- Replace '2' with '0' in the Replace box
- Click Replace
- Click OK

Alternatively, dummy coding could be accomplished by maintaining the Nominal scale types for these 2 attributes, and selecting 'Dummy Last' in the 'Coding Nominal' section of the Output tab. Both approaches change to dummy coding, the only difference being in the display of the effect estimates in the Parameters Output. Only a single effect estimate is shown for each attribute using the former method, while the latter associates this effect estimate with the 1st category and displays a '0' effect for the 2nd category.

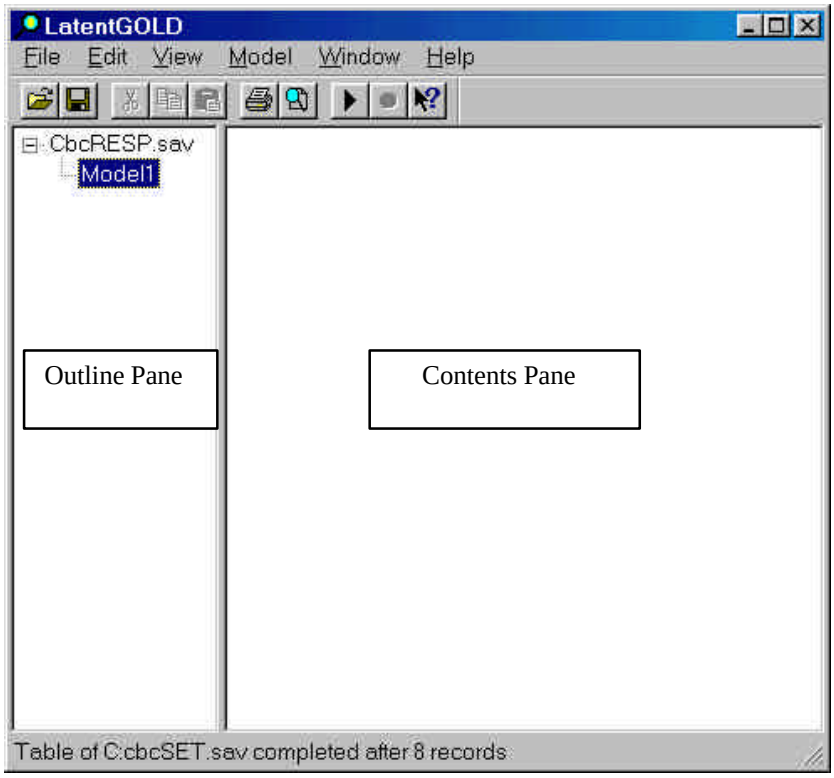
Estimating Models

Now that we have specified the models, we are ready to estimate these models.

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

The setup window now closes and estimation begins. When Latent GOLD Choice completes the estimation, the model L^2, which assesses how well the model fits the data, and a list of various output files appears in the Outline pane.

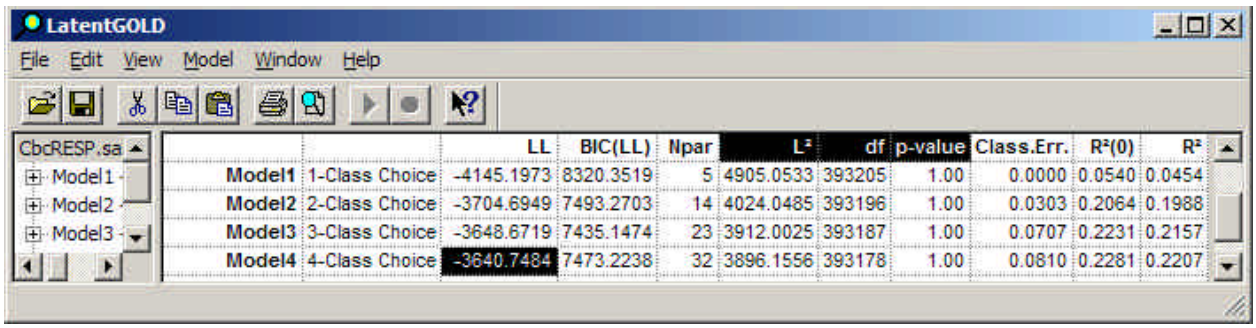
Figure 24. The two Panes



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

Viewing Output and Interpreting Results

- Highlight the data file name 'cbcRESP.sav' and a summary of all the models estimated on this file appears in the Contents pane.



		LL	BIC(LL)	Npar	L²	df	p-value	Class.Err.	R²(0)	R²
Model1	1-Class Choice	-4145.1973	8320.3519	5	4905.0533	393205	1.00	0.0000	0.0540	0.0454
Model2	2-Class Choice	-3704.6949	7493.2703	14	4024.0485	393196	1.00	0.0303	0.2064	0.1988
Model3	3-Class Choice	-3648.6719	7435.1474	23	3912.0025	393187	1.00	0.0707	0.2231	0.2157
Model4	4-Class Choice	-3640.7484	7473.2238	32	3896.1556	393178	1.00	0.0810	0.2281	0.2207

Figure 25. Model Summary Output

Note: For the 4-class model, you may obtain a local solution (LL = -3644.3868) rather than the global solution reported above (LL = -3640.7484). If this occurs, re-estimate the 4-class model.

For repeated measure applications such as this, data sparseness exists which implies that the L^2 statistic does not follow a chi-square distribution. In such cases the reported p-value is not correct and the number of degrees of freedom is not informative. Hence, we will exclude the L^2, df and associated p-value columns from our Model Summary output.

In addition, notice that 2 different R^2 measures are displayed. For this model, we use the standard R^2. (For other applications as in tutorial 3, it will be more appropriate to utilize $R^2(0)$ rather than R^2 to assess the goodness of model prediction. The baseline for $R^2(0)$ is a null model containing no predictors at all, not even the variable '*_Constants_*'. This latter null model predicts each alternative to be equally likely to be selected, while the former null model predicts each alternative to be selected with probability equal to the overall observed marginal distribution.)

To remove the undesired columns from the Model Summary:

- Right click in the Contents Pane to retrieve the Model Summary Display
- Click in the boxes to the left of the L^2, df and p-value to remove the checkmarks from these items
- Close the Model Summary Display (by pressing Escape or clicking the X in the upper right of the Display)

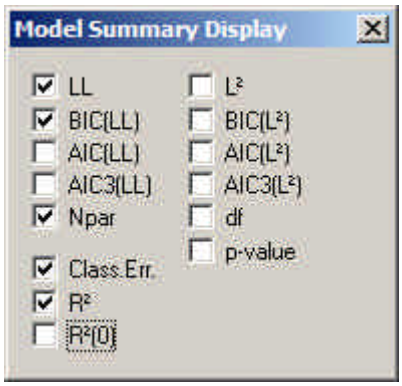


Figure 26. Model Summary Display

The Model Summary in the Contents Pane now looks like this:

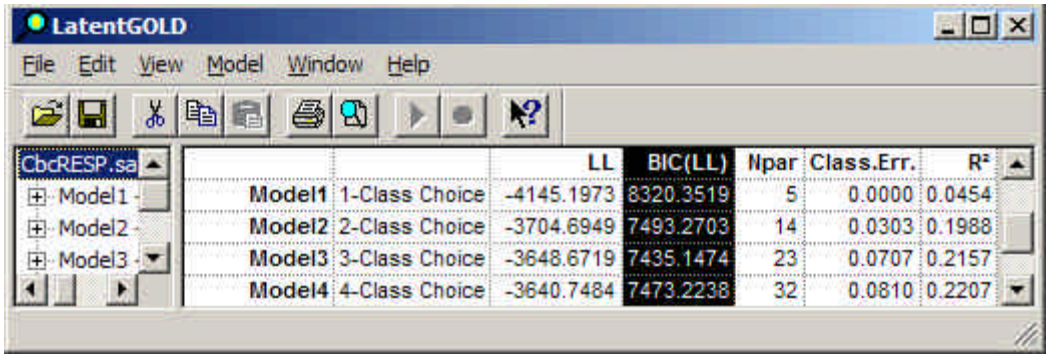


Figure 27. Updated Model Summary Display

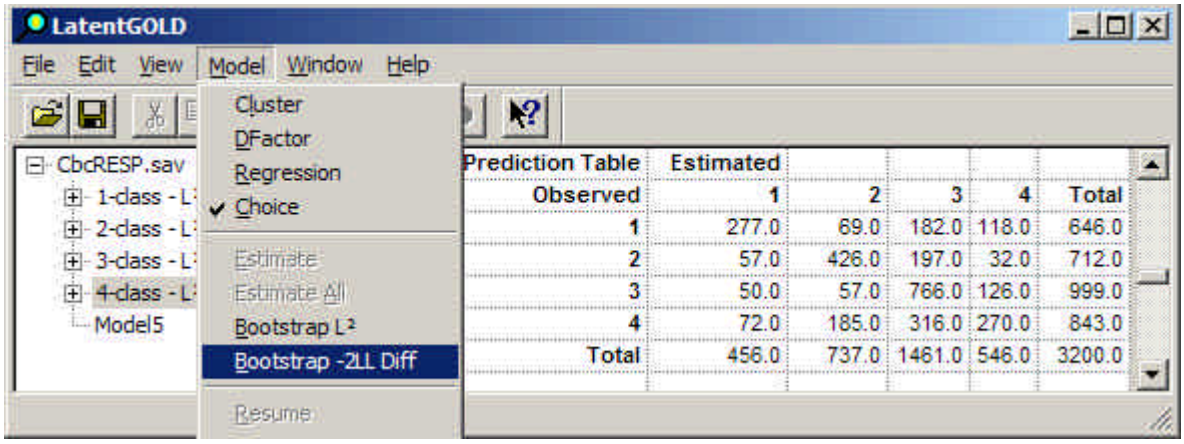
Notice that BIC criteria correctly identifies the 3-class model as best (lowest BIC). Notice also that the R^2 increases from .0454 for the standard aggregate (1-class) model to .2157 for the 3-class model. This R^2 statistic assesses the percentage variance explained in the dependent variable. In this example, the (nominal) dependent variable consists of 4 alternatives in each set (alt1, alt2, alt3 and alt4), variance is defined in terms of qualitative variance (Magidson, 1981), and the baseline or null model is the model that includes the alternative-specific constants '*_Constants_*' as the only predictor.

- 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.
- Click once on each of the other models and rename each of them as well.

As a formal assessment of whether the R^2 obtained from the 4-class model provides a significant increase over the 3-class model, we can estimate a p-value for this improvement using the Conditional Bootstrap. By default, the Bootstrap procedure utilizes 500 samples and thus may take a few minutes to complete on your computer. Feel free to skip this, or if it is taking too long you may Cancel the estimation procedure by clicking the Red Stop button to Pause the model and selecting 'Abandon'.

- Select the model '4-class' in the Outline Pane
- Select 'Bootstrap –2LL Diff' from the Model menu

Figure 28. Selecting Bootstrap –2LL Diff



A list of models appears that contain eligible nested base models.

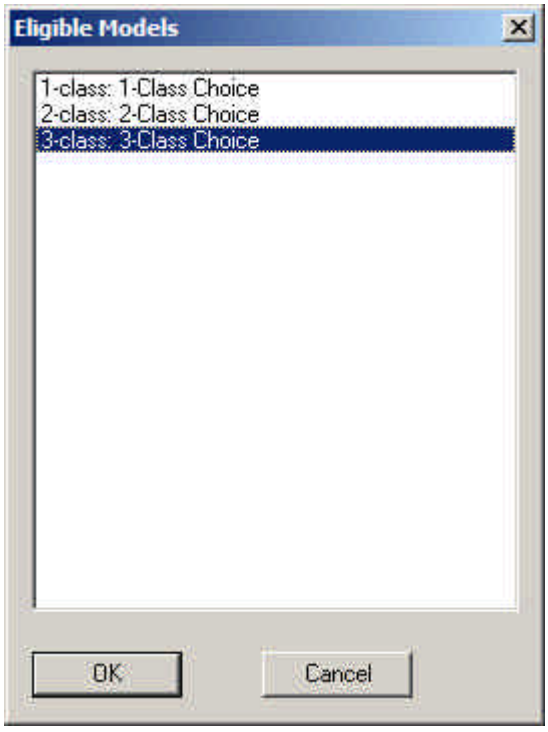


Figure 29. List of Eligible Models

- Select '3-class' from this list
- Select OK

When the procedure has completed, 2 additional model names appear in the Outline Pane – '3-classBoot' and '4classBoot'.

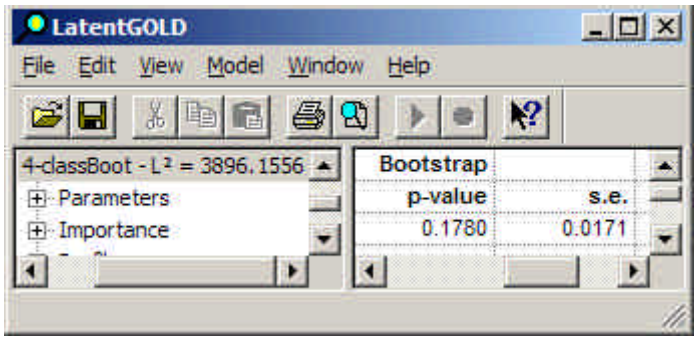


Figure 30. 4-classBoot Model Summary

Since the estimated p-value is larger than .05 (p=.178 with an estimated standard error of .-17 reported in the above Figure), the improvement by going to a 4th class is not statistically significant. The estimated p-value that you obtain should be close to this estimate.

Examining the Model Output files

- Click on the expand icon (+) next to the 1-class model and several output files appear.
- Click on Parameters to view the part-worth utility estimates in the Contents Pane

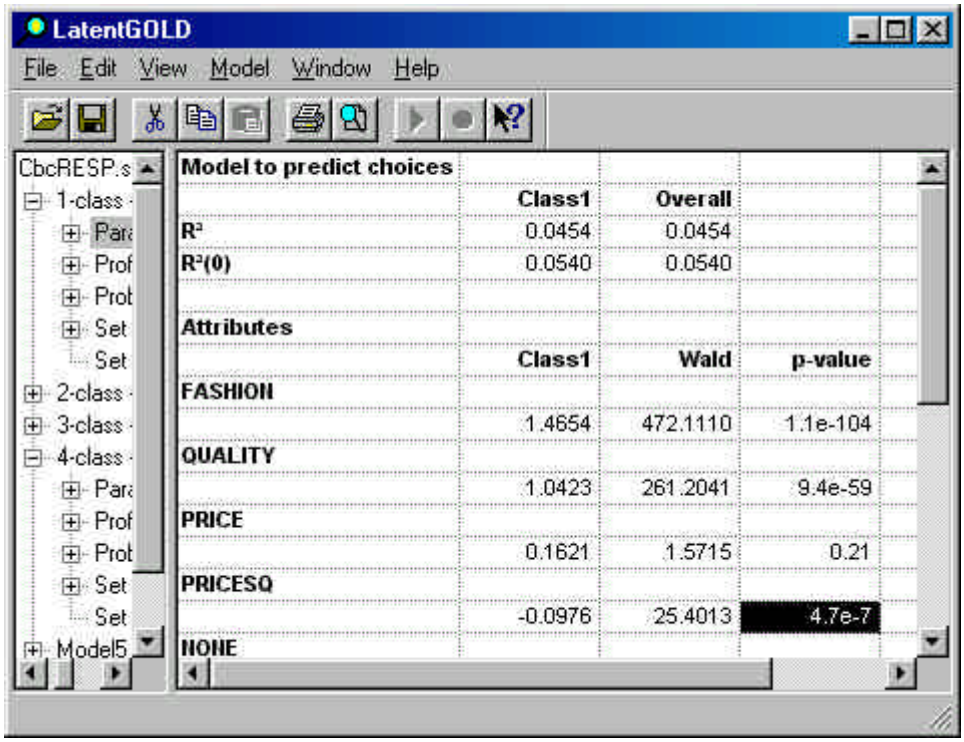


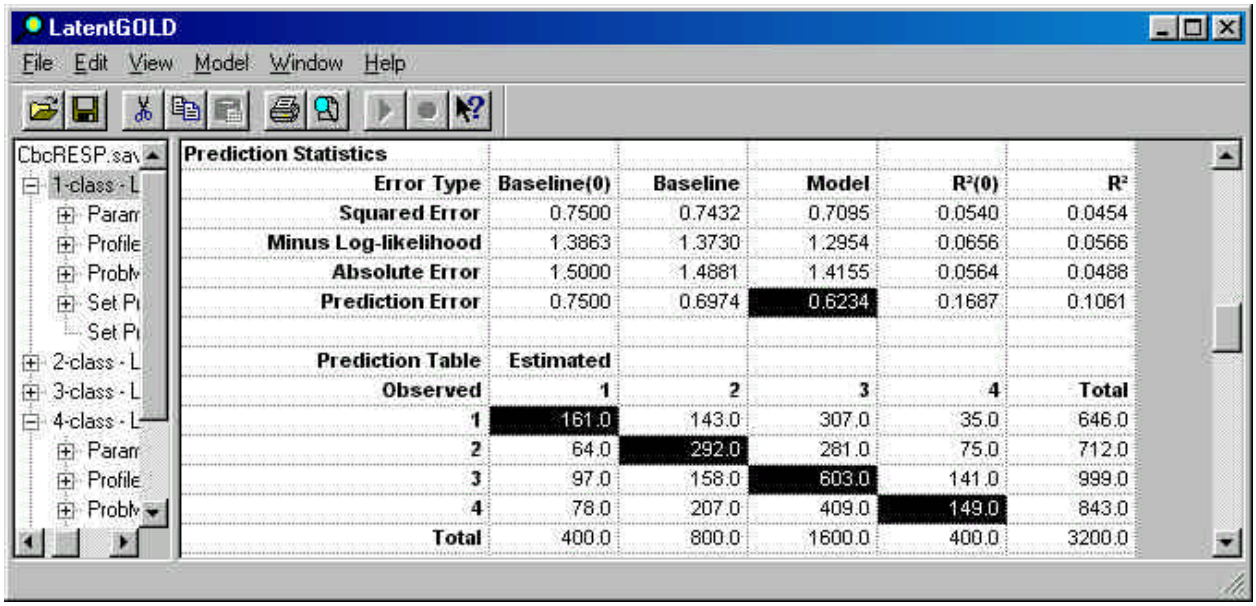
Figure 31. Parameters Output for 1-class Model

Notice that the effect of PRICESQ is highly significant ($p = 4.7 \times 10^{-7}$) in this model. Since the effect of PRICESQ is 0 in the true model, this example shows that the effects (and predictions) obtained from the

aggregate model can not be trusted. When choices are based on different utilities (in different population segments), estimates obtained under the aggregate model will typically be biased.

- Click on '1-class' and scroll down to the Prediction Statistics section

Figure 32. Prediction Statistics for 1-class Model



Prediction Statistics

Error Type	Baseline(0)	Baseline	Model	$R^2(0)$	R^2
Squared Error	0.7500	0.7432	0.7095	0.0540	0.0454
Minus Log-likelihood	1.3863	1.3730	1.2954	0.0656	0.0566
Absolute Error	1.5000	1.4881	1.4155	0.0564	0.0488
Prediction Error	0.7500	0.6974	0.6234	0.1687	0.1061

Prediction Table	Estimated				
Observed	1	2	3	4	Total
1	161.0	143.0	307.0	35.0	646.0
2	64.0	292.0	281.0	75.0	712.0
3	97.0	158.0	603.0	141.0	999.0
4	78.0	207.0	409.0	149.0	843.0
Total	400.0	800.0	1600.0	400.0	3200.0

The prediction table shows that this 1-class (aggregate) model correctly predicts only 161 of the 646 alt1 responses, only 292 of the 712 alt2 responses, 603 of the 999 alt3 responses and only 149 of the 843 alt4 responses. Overall, only 1,205 (161 + 292 + 603 +149) of the total 3,200 observed choices are predicted correctly – a 'hit rate' of only 37.66%. This represents a Prediction Error of 1-.3766 = .6234 (as reported in the row of the output labeled 'Prediction Error').

For comparison, we will look at the corresponding statistics under the 3-class model.

- Click on '3-class' and scroll down to the Prediction Statistics section

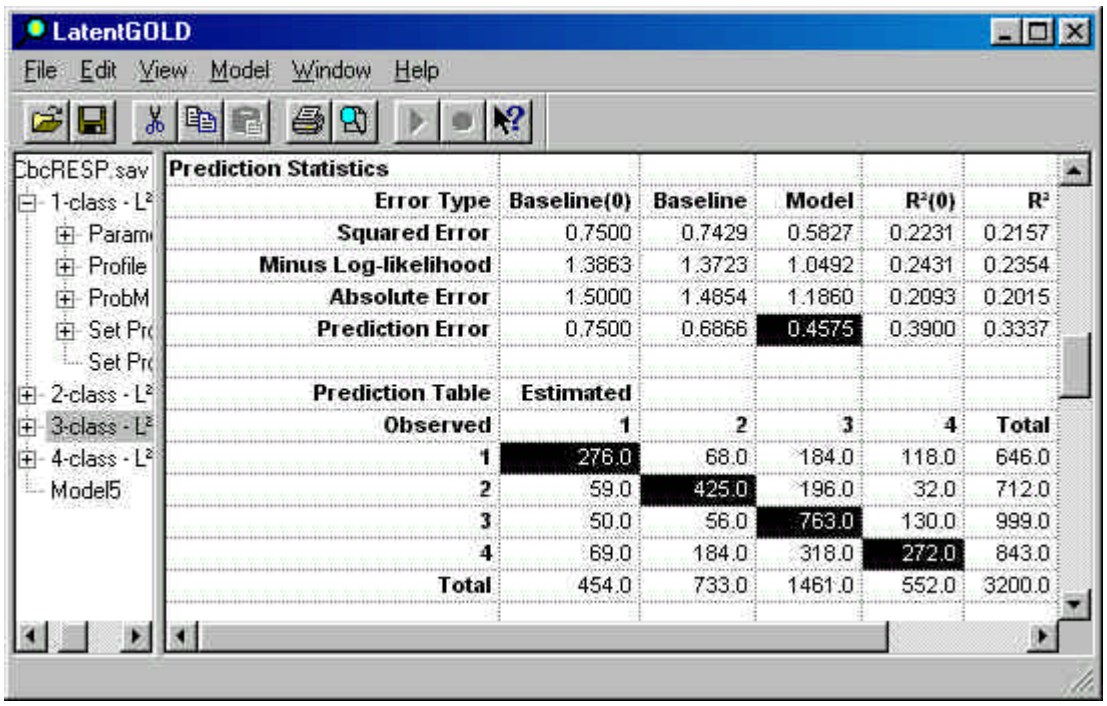


Figure 33. Prediction Statistics for 3-class Model

By accounting for the heterogeneity among the 3 segments, the prediction error has been reduced to .4575 by the 3-class model.

- Click on the expand icon (+) next to the 3-class model and select Parameters.

The part-worth utility estimates appear for each class.

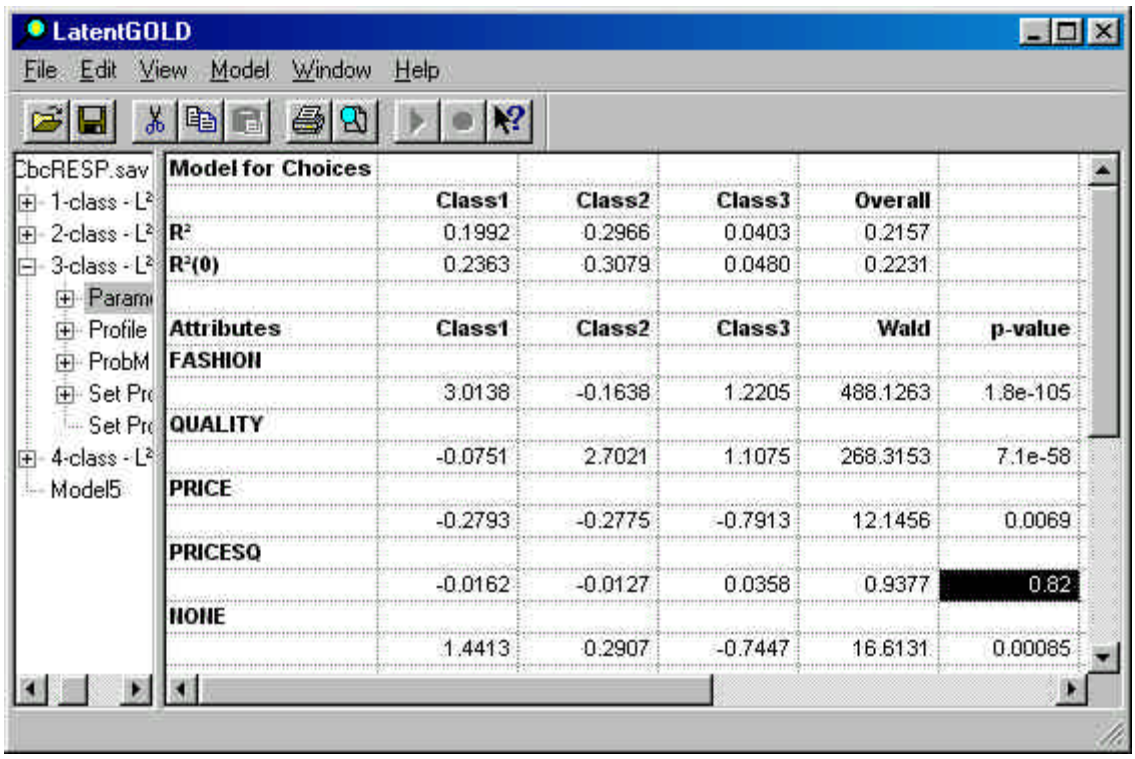


Figure 34. Parameters Output for 3-class Model

Using Restrictions to Refine the 3-Class Model

Under the 3-class model, the estimates for PRICESQ are no longer significant, which reflects the true structure of this model. The associated p-value is greater than 0.05 (p = .82).

- Double click on '3-class' to re-open the model setup window.
- Click Attributes to open the attributes tab
- Select PRICESQ and click the Attributes button to remove this effect from the model.
- Click Estimate to re-estimate the model.
- After the estimation has completed click Parameters to view the new parameters estimates.

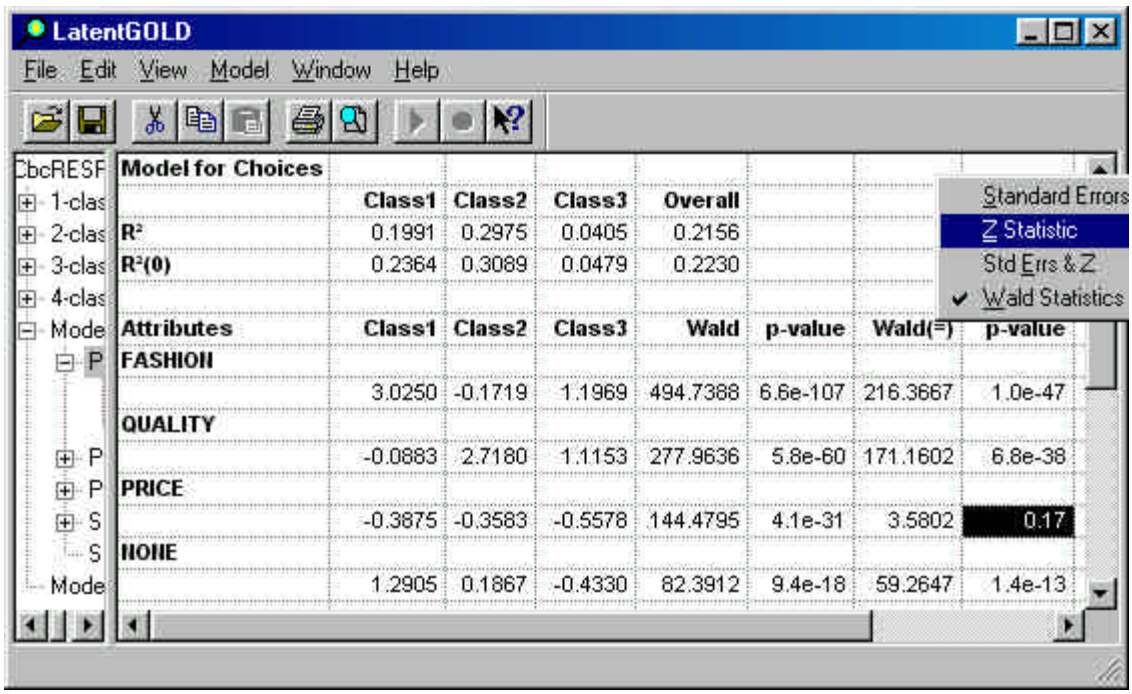


Figure 35. New Parameters Output

Notice that the within-class estimates for PRICE are close to each other. The formal test of equality is given by the Wald(=) statistic, for which the p-value (0.17) is not significant. Thus, we will restrict the effect of PRICE to be equal across all 3 segments.

To view whether any of the parameter estimates are non-significant:

- Right click on the Contents Pane and select Z Statistic from the Popup menu.

The Z-statistics show that for class 1 the effect of higher quality, and for class 2 the effect of modern fashion are not significant.

To implement these parameter restrictions

- Double click on the last model ('Model5' here) to re-open the model setup screen.
- Click on 'Model' to Open the Model tab.
- Right click under column 1 in the QUALITY row and select 'No Effect' from the pop-up menu to restrict the class 1 effect for QUALITY to zero.
- Right click under column 2 in the FASHION row and select 'No Effect' from the pop-up menu to restrict the class 1 effect for FASHION to zero.
- Right click under Class Independent column in the PRICE row and select 'Yes' to restrict the effect of PRICE to be class independent.

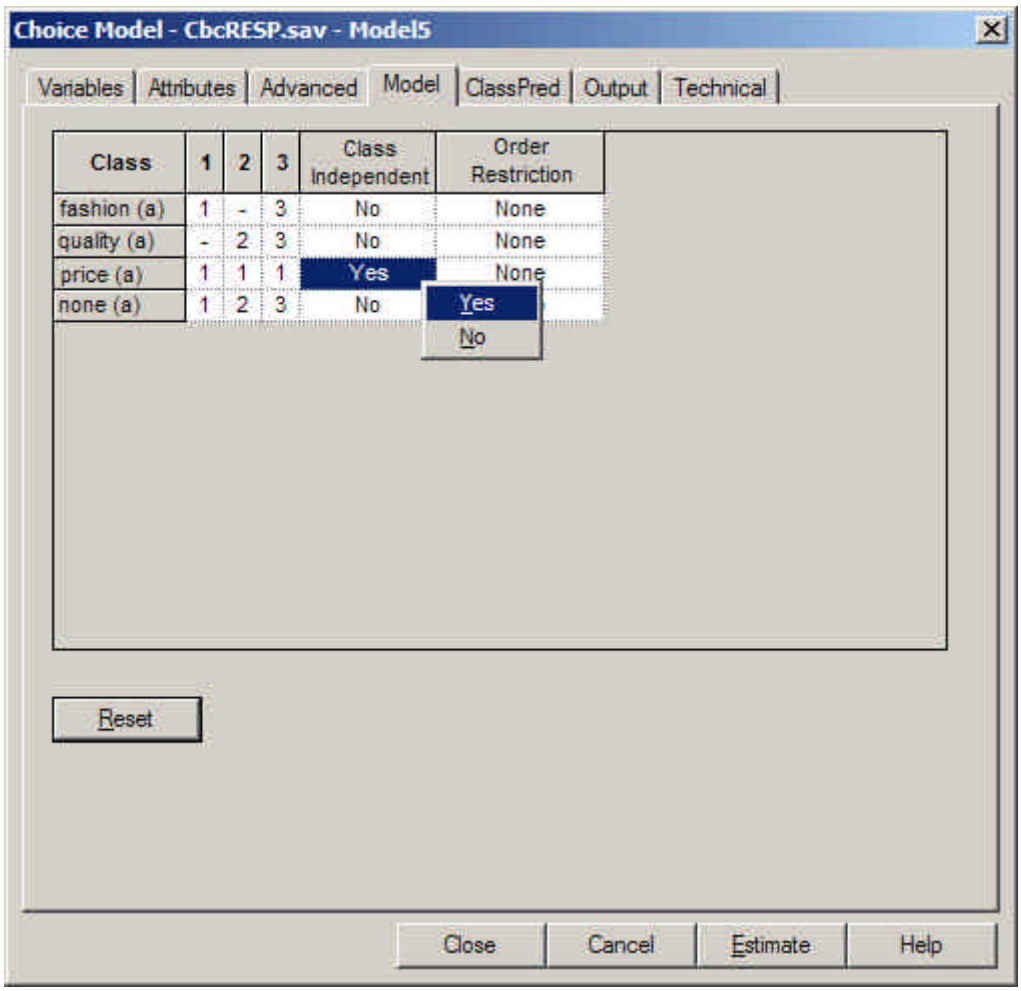
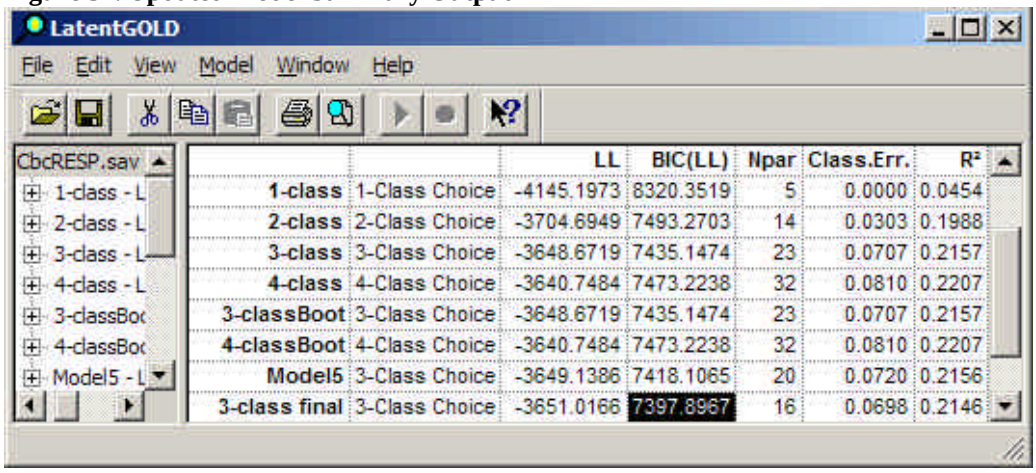


Figure 36. Implementing Parameter Restrictions

- Click Estimate to re-estimate the model.
- Re-name this model '3-Class Final'

Notice that the BIC for this final model is now the best.

Figure 37. Updated Model Summary Output



		LL	BIC(LL)	Npar	Class.Err.	R^2
1-class	1-Class Choice	-4145.1973	8320.3519	5	0.0000	0.0454
2-class	2-Class Choice	-3704.6949	7493.2703	14	0.0303	0.1988
3-class	3-Class Choice	-3648.6719	7435.1474	23	0.0707	0.2157
4-class	4-Class Choice	-3640.7484	7473.2238	32	0.0810	0.2207
3-classBoot	3-Class Choice	-3648.6719	7435.1474	23	0.0707	0.2157
4-classBoot	4-Class Choice	-3640.7484	7473.2238	32	0.0810	0.2207
Model5	3-Class Choice	-3649.1386	7418.1065	20	0.0720	0.2156
3-class final	3-Class Choice	-3651.0166	7397.8967	16	0.0698	0.2146

To remove the Bootstrap models,

- Select '3-classBoot'
- Select Delete from the Model menu
- Repeat this for '4-classBoot'

Profile Output: viewing Re-scaled Parameters

- Click on the expand icon (+) next to 3-class final model to display the output files
- Click on Profile to display the Profile Output in the Contents Pane

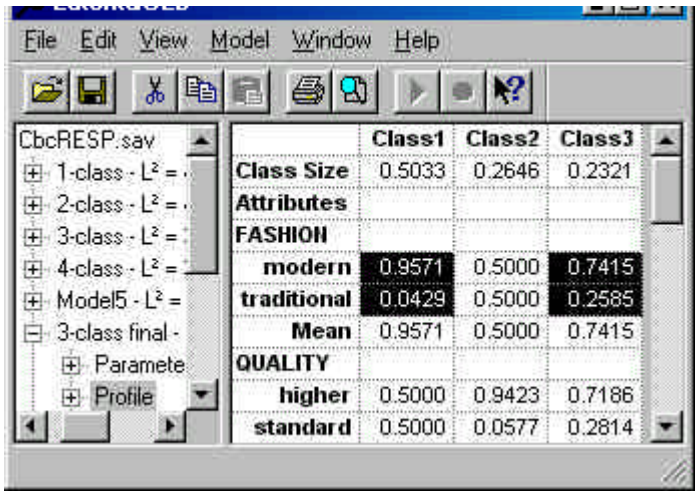


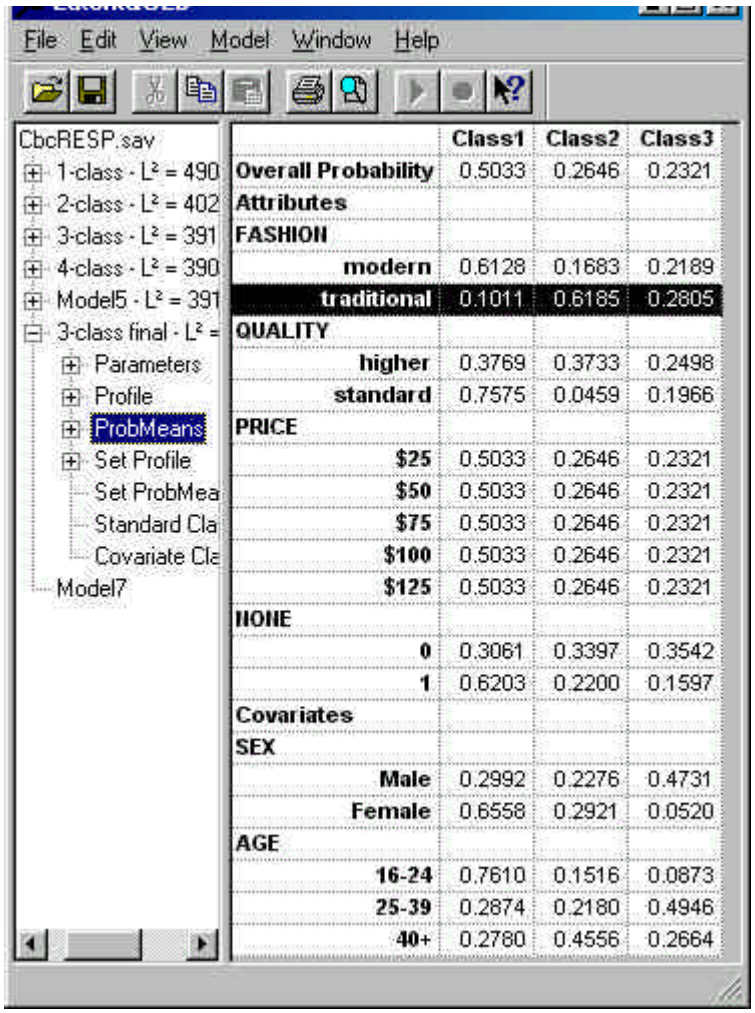
Figure 38. Final 3-class Model Profile Output

The Profile output consists of the segment size estimates together with re-scaled parameter estimates that correspond to column percentages. Notice that the segment size estimates of .50, .26 and .23 are very close to the true universe parameters.

These re-scaled parameters for FASHION and QUALITY have a nice probabilistic interpretation. The probability of choosing the Modern over the Traditional style when faced with just these 2 alternatives where the QUALITY and PRICE were identical is estimated to be .9571 for persons in segment 1, .5 for persons in segment 2, and .7415 for persons in segment 3. A similar interpretation is available for QUALITY. When faced with a choice between 1) a pair of Higher quality and 2) a pair of Standard quality shoes – both having the same price and same style (Modern or Traditional), the probability of choosing the Higher quality shoes is .5 for segment 1, .9423 for segment 2 and .7186 for segment 3. While the re-scaled parameters for PRICE can not be interpreted in same way, they still convey information about which segment(s) are more price sensitive.

A second benefit of the re-scaled parameters is that when presented as row percentages (as opposed to column percentages) they become the ProbMeans output, which yields an informative display in the dimension of the segments.

- Click on ProbMeans to display the corresponding row percentages.



	Class1	Class2	Class3
Overall Probability	0.5033	0.2646	0.2321
Attributes			
FASHION			
modern	0.6128	0.1683	0.2189
traditional	0.1011	0.6185	0.2805
QUALITY			
higher	0.3769	0.3733	0.2498
standard	0.7575	0.0459	0.1966
PRICE			
$25	0.5033	0.2646	0.2321
$50	0.5033	0.2646	0.2321
$75	0.5033	0.2646	0.2321
$100	0.5033	0.2646	0.2321
$125	0.5033	0.2646	0.2321
NONE			
0	0.3061	0.3397	0.3542
1	0.6203	0.2200	0.1597
Covariates			
SEX			
Male	0.2992	0.2276	0.4731
Female	0.6558	0.2921	0.0520
AGE			
16-24	0.7610	0.1516	0.0873
25-39	0.2874	0.2180	0.4946
40+	0.2780	0.4556	0.2664

Figure 39. Final 3-class Model ProbMeans Output

The row percentages highlighted above state that persons choosing Traditional vs. Modern style shoes (where price and quality are the same), are most likely (posterior probability = .6185) to be in segment 2. These row percentages can be used to position each category in an informative 2-dimensional barycentric coordinate display. Choices can also be appended to this plot.

- Click on the expand icon (+) next to ProbMeans
- Click on Tri-Plot to display the Plot
- Right click in the Contents Pane to display the Plot Control Panel
- Click on PRICE to remove PRICE from the plot
- Click on the drop down set list and select set #4
- Click on the Choice Set symbol to highlight the appended choices on the plot

Each vertex of the tri-plot corresponds to a segment. From the plot it can be seen that Class 2 (corresponding to the lower right vertex) is the segment most likely to prefer Higher Quality and Traditional style shoes. This segment is also tends to be the oldest (higher proportion of persons 40+) of the segments.

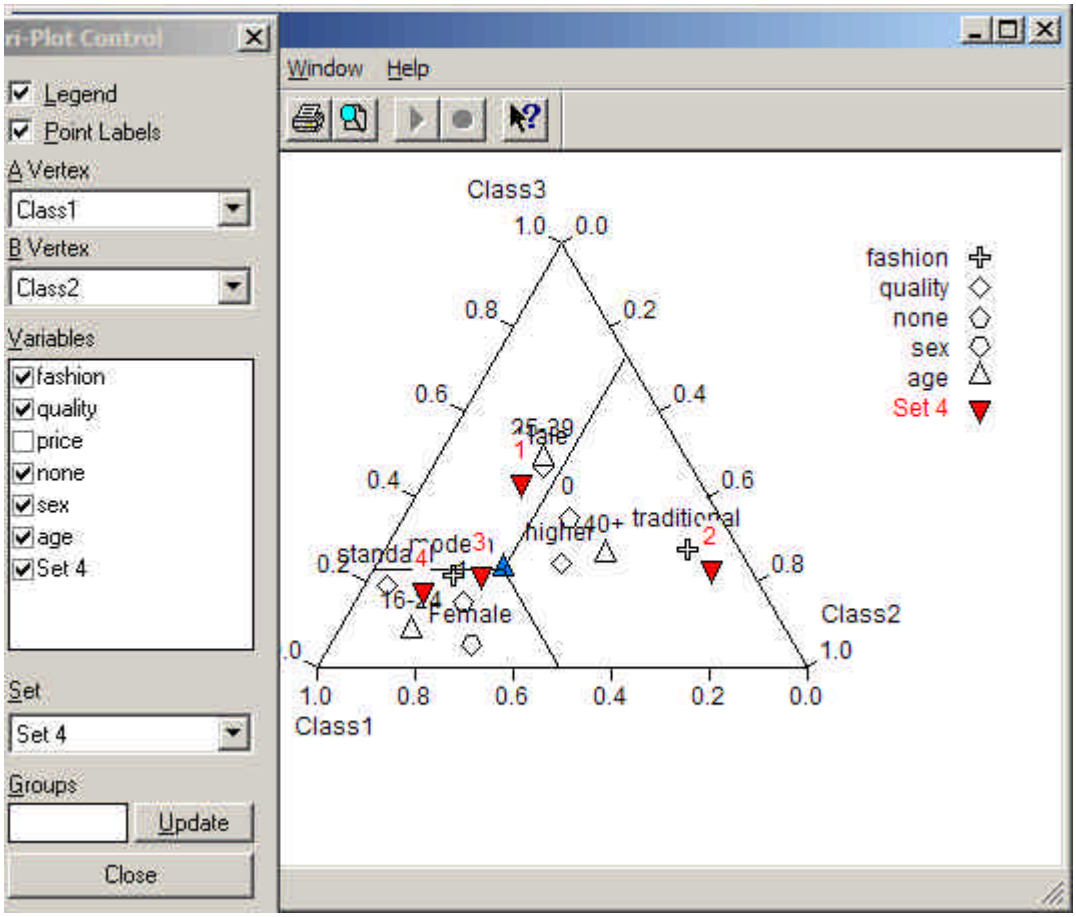


Figure 40. Tri-Plot

Set #4 consists of 1:TS2 2:TH4 3:MH5 4:None. It is clear from the plot that persons who choose alternative #2:TH4 from this set are highly likely to be in segment 2.

Additional Output Options

- Latent GOLD Choice offers many additional output options.

Double click on '3-class final' to re-open the model setup screen

- Click on Output to Open the Output Tab

To obtain output to see how to classify each person into the most likely segment

- Click on Standard Classification
- Click on Covariate Classification

To append this information and predicted choices to the response file

- Click on ClassPred to Open the ClassPred Tab
- Click on Standard Classification
- Click on Covariate Classification
- Click on Predicted Values

To append individual HB-like coefficients to the response file

- Click on Individual Coefficients
- Replace the default output filename if desired

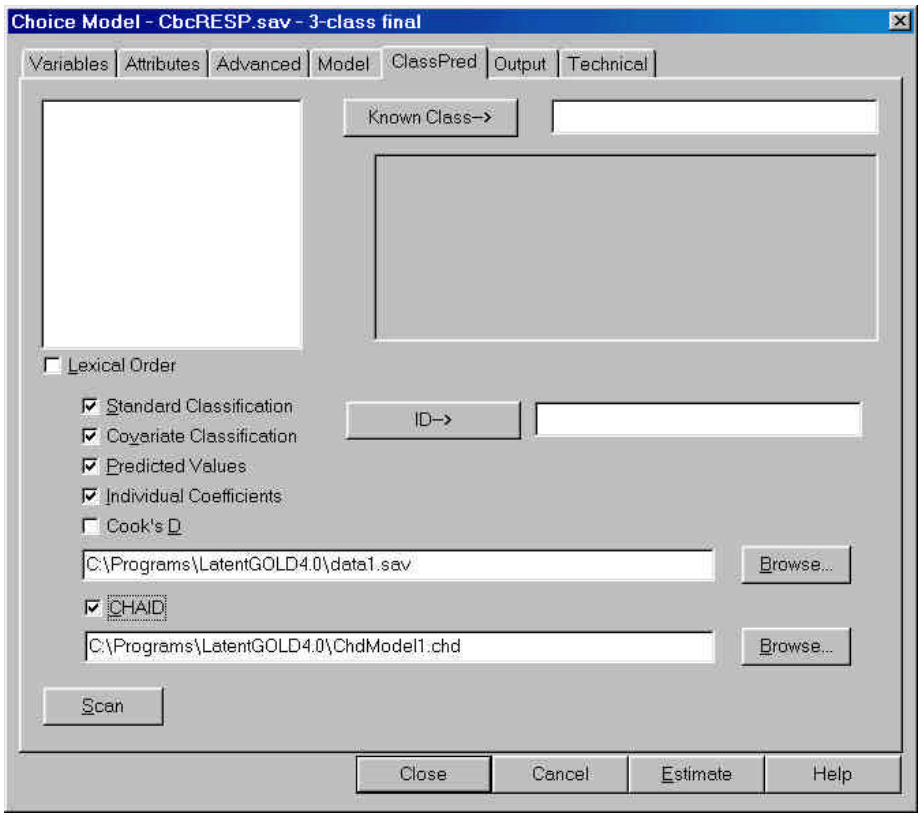


Figure 41. ClassPred Tab

Note that we also requested a CHAID input file to be created. We will illustrate this option later in tutorial 1A.

- Click Estimate

To open the standard classification output,

- Click Standard Classification in the Outline Pane

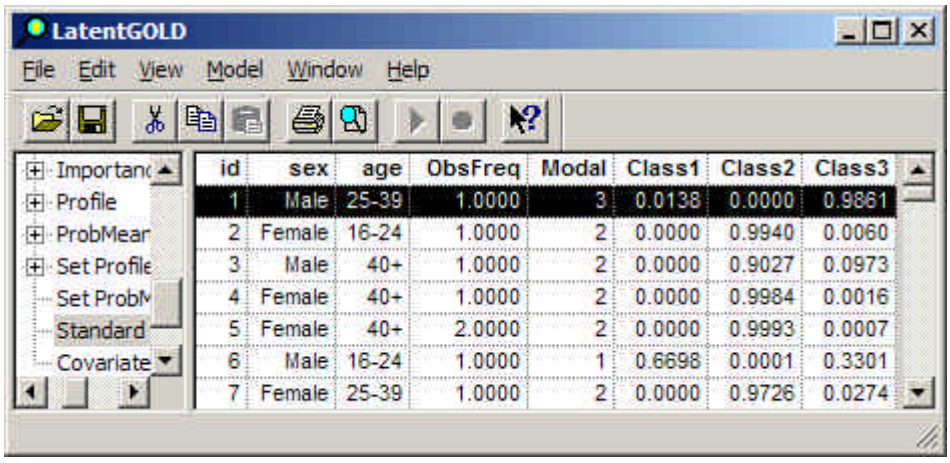


Figure 42. Standard Classification Output

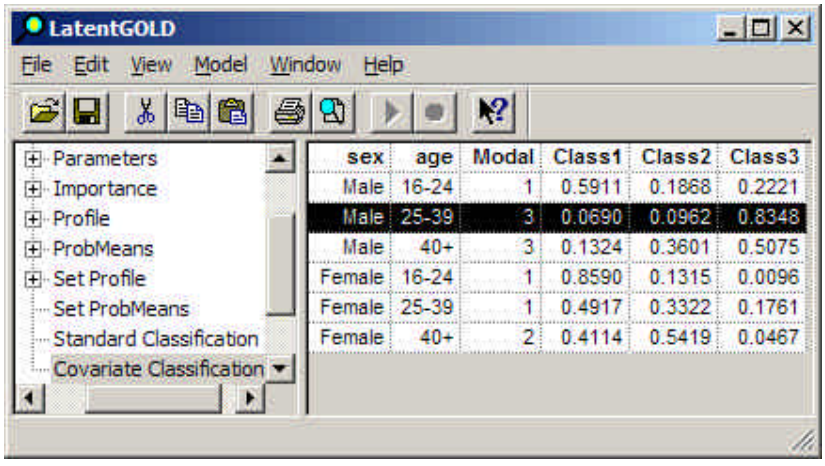
This output shows that case 1 is very likely (posterior probability = .9861) a member of segment 3.

The covariate classification information is useful for classifying new cases, which did not participate in the conjoint experiment.

- Click Covariate Classification

This table shows that based solely on this person's demographics, he would still be classified into segment 3 although with somewhat less certainty (probability = .8348). We will explore the relationship between the covariates and the classes more extensively in Tutorial 1A using the SI-CHAID 4.0 option.

Figure 43. Covariate Classification Output



Latent GOLD Choice was found to be successful in uncovering the 3-segment structure in this example. In tutorial #2 we will use our final model to generate choice predictions.

To save the analysis file for future work

- Click on the data file name cbcRESP.sav at the top of the outline pane
- From the File Menu, select Save Definitions
- Click Save