

## Latent Gold 4.0 User's Guide

Vermunt, Jeroen; Magidson, J.

*Publication date:*  
2005

[Link to publication](#)

*Citation for published version (APA):*  
Vermunt, J. K., & Magidson, J. (2005). Latent Gold 4.0 User's Guide. Belmont, MA: Statistical Innovations Inc.

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

### Take down policy

If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# **LATENT GOLD® 4.0** **USER'S GUIDE**

**Jeroen K. Vermunt**  
**& Jay Magidson**

For more information about Statistical Innovations Inc. please visit our website at  
**<http://www.statisticalinnovations.com>**

or contact us at

**Statistical Innovations Inc.**  
**375 Concord Avenue, Suite 007**  
**Belmont, MA 02478**  
**e-mail: [michael@statisticalinnovations.com](mailto:michael@statisticalinnovations.com)**

Latent GOLD® is a registered trademark of Statistical Innovations Inc.

Windows is a trademark of Microsoft Corporation.

SPSS is a trademark of SPSS, Inc.

Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

Latent GOLD® 4.0 User's Guide.

Copyright © 2005 by Statistical Innovations Inc.

All rights reserved.

No part of this publication may be reproduced or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission from Statistical Innovations Inc.

This document should be cited as "J.K. Vermunt and J. Magidson (2005) Latent GOLD 4.0 User's Guide. Belmont, Massachusetts: Statistical Innovations Inc."

**To our teacher**

**Leo A. Goodman**

# Preface to Latent GOLD 4.0

Since 1995, more significant books have been published on latent class (LC) and finite mixture models than any other class of statistical models. The recent increase in interest in latent class models is due to the development of extended algorithms which allow today's computers to perform LC analyses on data containing more than just a few variables, and the recent realization that the use of such models can yield powerful improvements over traditional approaches to segmentation, as well as to cluster, factor, regression and other kinds of analysis.

The development of Latent GOLD was a collaborative transcontinental effort that began in March 1998 in the Netherlands where we shared ideas for the first time. The initial version of the program, Latent GOLD 2.0 was released in May 2000. It provided a vehicle for developing modern LC models in a Windows-friendly environment. In February 2003 we released Latent GOLD 3.0 which contained several new features.

We believe that Latent GOLD 4.0 takes the program to a new level. Major interface improvements such as the ability to pause a model during estimation process to view preliminary output, should be especially useful when estimating models with many parameters. In addition, the Advanced version of Latent GOLD 4.0 includes multilevel modeling, accounting for complex sampling designs and the estimation of other kinds of models such as IRT (Item Response Theory) based latent trait models, and random-effects regression models.

This manual consists of two parts, Latent GOLD 4.0 User's Guide (this document) and a companion manual, Technical Guide for Latent GOLD 4.0: Basic and Advanced. Latent GOLD 4.0 User's Guide contains:

- General information regarding the program, its structure, and functions
- Four tutorials to get you up and running quickly
- A description of additional resources on the Statistical Innovations' website including many additional examples and associated .lgf files

Technical Guide for Latent GOLD 4.0: Basic and Advanced, contains:

- A formal technical introduction (including equations for all models and statistics)
- An extensive bibliography

Should you have any questions or comments, you can contact us by sending an email to Michael Denisenko at [Michael@statisticalinnovations.com](mailto:Michael@statisticalinnovations.com). Please be sure to include your serial number to expedite our reply.

Jay Magidson  
Belmont, Massachusetts

Jeroen Vermunt  
Tilburg, The Netherlands

April 2005

## **Compatibility**

Latent GOLD® is designed for computers running Windows 95, Windows 98, Windows 2000, Windows XP, Windows NT 4.0, or later

## **Registration Code**

Your registration code is your serial identification number with Statistical Innovations Inc. You will need this number when you contact us for information regarding support, payment, or upgrades. The registration code was provided with your Latent GOLD program.

## **Customer Service**

If you have any questions concerning your shipment or account, see Contacting Statistical Innovations. Please have your registration code ready for identification when calling.

## **Training Seminars**

We provide public and onsite training seminars on Latent GOLD. We also offer online courses. For information or to be placed on our mailing list, see Contacting Statistical Innovations or visit our website.

## **Tell Us Your Thoughts**

Your comments are important to us. Please write or e-mail us about your experiences with Latent GOLD. We especially like to hear about new and interesting applications using Latent GOLD. Consider submitting examples and application ideas for inclusion on our website.

## **Contacting Statistical Innovations**

To contact us or to be placed on our mailing list, visit our website at <http://www.statisticalinnovations.com> or write us at **Statistical Innovations Inc., 375 Concord Avenue, Belmont, MA 02478.**

You can also e-mail us at **[michael@statisticalinnovations.com](mailto:michael@statisticalinnovations.com)**.

# TABLE OF CONTENTS

CHAPTER 1. OVERVIEW .....	1
1.1 Latent Class, Finite Mixture Modeling, And Beyond .....	1
1.2 Kinds Of Latent Class Models .....	2
1.3 Regression Models with Random Effects And Parameter Restrictions ..	4
1.4 Interactive Use .....	5
1.5 New Features In Latent GOLD 4.0 .....	5
1.6 Optional Add-ons To Latent GOLD 4.0 .....	9
1.7 Additional Resources .....	11
1.8 Structure Of The Manual .....	11
CHAPTER 2. GENERAL PROGRAM STRUCTURE .....	13
<b>Windows</b> .....	<b>13</b>
To Copy Output from the Viewer .....	15
<b>Menus</b> .....	<b>15</b>
File .....	15
Edit .....	16
View .....	16
Model .....	17
Window .....	18
Help .....	18
<b>Toolbar</b> .....	<b>19</b>
To Show or Hide the Toolbar .....	19
To Move the Toolbar .....	19
<b>Status Bar</b> .....	<b>20</b>
<b>Dialog Boxes</b> .....	<b>20</b>
Analysis Dialog Box .....	20
Variables Tab .....	20
Model Tab .....	21

ClassPred Tab .....	21
Residuals Tab (Cluster and DFactor only) .....	21
Output Tab .....	21
Technical Tab .....	21
Advanced Tab (Optional) .....	21
Dialog Box Pushbuttons .....	22
Subdialog Boxes .....	22
Selecting Variables .....	22
Basic Steps in Model Estimation .....	23
<b>Getting Help .....</b>	<b>23</b>
 <b>CHAPTER 3. DATA FILES AND FORMATS .....</b>	<b>25</b>
 <b>Opening a Data File .....</b>	<b>25</b>
File Open Options .....	26
SPSS, Text and Array Files .....	27
<b>File Import Option .....</b>	<b>27</b>
Importing SAS, Excel, and other data file formats .....	27
Latent GOLD Files .....	28
Re-opening Data Files Quickly .....	28
<b>Data File Formats .....</b>	<b>29</b>
SPSS Files .....	29
Text Files .....	29
Arrays .....	31
Latent GOLD Save File (*.lgf) .....	32
<b>Saving Model Settings &amp; Output .....</b>	<b>32</b>
Save Definition .....	32
Save Results .....	34
<b>Closing a Data File .....</b>	<b>35</b>
 <b>CHAPTER 4. WORKING WITH OUTPUT .....</b>	<b>37</b>
 <b>Printing Output (File Menu) .....</b>	<b>37</b>
Print Preview (File Menu) .....	38
Print Setup (File Menu) .....	39
<b>Changing Fonts (Edit Menu) .....</b>	<b>40</b>
<b>Changing Numeric Format (Edit Menu) .....</b>	<b>41</b>
 <b>CHAPTER 5. BASIC STEPS FOR MODEL DEVELOPMENT .....</b>	<b>43</b>
 <b>STEP 1: LOAD YOUR DATA INTO LATENT GOLD .....</b>	<b>44</b>
<b>STEP 2: SELECT THE TYPE OF MODEL .....</b>	<b>44</b>

<b>STEP 3: SELECT VARIABLES FOR THE ANALYSIS (VARIABLES TAB)</b>	<b>47</b>
Specifying a Case Weight (optional)	47
Covariates (optional)	48
Indicators (Cluster and DFactor Models)	48
Dependent (Regression Models)	49
Predictors (Regression Models, optional)	50
Exposure (Regression Models for Counts, optional)	50
Regression Models with Repeated Measurements	50
Case ID Variable (Regression Models, optional)	50
Replication Weight (Repeated Measures Regression Models, optional)	51
<b>STEP 4: SCAN THE DATA FILE (VARIABLES TAB)</b>	<b>51</b>
Viewing Category Labels, Frequency Counts and Scores for a Variable	51
Missing Values	52
Reducing the Number of Categories for a Variable	52
<b>STEP 5: SET SCALE TYPES (VARIABLES TAB)</b>	<b>54</b>
<b>Scale Types</b>	<b>55</b>
Ordinal (indicators, dependent)	55
Nominal (indicators, covariates, dependent, predictors)	56
Numeric (covariates, predictors)	57
Active (covariates)	57
Inactive (covariates)	57
Advanced: Group (covariates)	57
Continuous (indicators, dependent)	57
Count (indicators, dependent)	57
Binomial Count (indicators, dependent)	57
<b>Scale Subtypes</b>	<b>58</b>
Standard (Indicators, Dependent)	58
Truncated (Count, Binomial Count, Continuous)	58
Censored (Continuous)	58
Zero Inflated (Regression only)	58
Overdispersed (Count and Binomial Count in Regression)	59
<b>STEP 6: SPECIFY THE NUMBER OF LATENT CLASSES (VARIABLES TAB)</b>	<b>59</b>
Specifying the Number of Clusters for LC Cluster Models	59
Specifying the Number of DFactors for DFactor Models	59
Specifying the number of Classes for LC Regression Models	59
Advanced. Specify the Number of CFactors, GClasses, GCFactors	60
<b>STEP 7: SET RESTRICTIONS AND OTHER MODEL OPTIONS</b>	<b>60</b>
<b>Model Tab: Applying (Or Relaxing) Parameter Restrictions</b>	<b>61</b>
<b>LC Cluster Model</b>	<b>61</b>
Changing the Number of Clusters	62
Imposing Zero Restrictions	62
Imposing Class Independent Restrictions	64
Estimating Order-Restricted Latent Classes	65
Equal effects across indicators	65

<b>DFactor Model</b>	<b>65</b>
Changing the Number of DFactors or DFactor Levels	66
Included Effects Box	67
Included Associations Box	68
Equal effects across indicators	68
<b>LC Regression Model</b>	<b>69</b>
Specifying Equality Restrictions Across Classes	70
Specifying Zero Restrictions	71
Specifying Order Restrictions	71
Specifying Fixed-Value Restrictions	71
Advanced: Including CFactor, GClasses, and GCFactor Effects in the Model for the Dependent Variable	72
<b>Classpred Tab: Restricting Cases Known (Not) To Belong To a Certain Class or Classes</b>	<b>72</b>
Known Class - Class Indicator	73
<b>Residuals Tab: Including Direct Effects in Model (Cluster and DFactor Models)</b>	<b>75</b>
<b>Advanced Tab (Requires The Advanced Version of Latent GOLD)</b>	<b>76</b>
Survey	77
Multilevel Model	78
<b>STEP 8: SET TECHNICAL OPTIONS (TECHNICAL TAB)</b>	<b>80</b>
Convergence Limits	80
Iteration Limits	81
Start Values	81
Bayes Constants	82
Missing Values	83
Bootstrap Options (Bootstrap $L^2$ , Bootstrap -2LL Diff)	83
Advanced. Continuous Factors	84
Default Options	84
<b>STEP 9: SET OUTPUT OPTIONS (OUTPUT TAB)</b>	<b>84</b>
Output Sections	85
Standard Errors and Wald	87
Prediction Type (Regression only)	87
Coding Nominal	88
Variance/Covariance Matrix	88
Default Options	89
<b>ClassPred Tab: Classification and Prediction Output to a file</b>	<b>89</b>
Classification Output To a File	89
Prediction Output To a File	92
<b>STEP 10: ESTIMATE THE MODEL, VIEW OUTPUT AND CONTINUE</b>	<b>93</b>
Stopping Model Estimation (Model Menu)	93
Bootstrap P-value ( $L^2$ ) (Model Menu)	95
Conditional Bootstrap (Bootstrap -2LL Diff) (Model Menu)	97
Output Files Created	99

Defining a New Model .....	99
Changing Model Names .....	100
Deleting Models .....	100

## CHAPTER 6. MODEL AND MODEL SUMMARY OUTPUT .....103

Ordering of Latent Classes in Output .....	105
Iteration Detail .....	105
<b>Data File Summary Output And Model Fit .....</b>	<b>107</b>
<b>Model Summary Output .....</b>	<b>108</b>
<b>Cluster .....</b>	<b>109</b>
General Information .....	109
LC Cluster Model Summary Output .....	110
Parameters Output (optional) .....	113
Profile Output (optional) .....	115
Profile Plot (for indicators and covariates) .....	117
GProfile Output (Advanced, optional) .....	119
ProbMeans Output (optional) .....	119
Uni-Plot .....	120
Tri-Plot .....	122
Frequencies and Residuals (optional) .....	124
Bivariate Residuals (optional) .....	124
Standard Classification (optional) .....	125
Covariate Classification (optional) .....	126
<b>DFactor .....</b>	<b>127</b>
General Information .....	127
DFactor Model Summary Output .....	128
Parameters Output (optional) .....	131
Profile Output (optional) .....	133
Profile Plot (for indicators and covariates) .....	136
GProfile Output (Advanced, optional) .....	138
ProbMeans Output (optional) .....	138
Uni-Plot .....	139
Bi-Plot .....	142
Frequencies And Residuals (Optional) .....	144
Bivariate Residuals (optional) .....	144
Standard Classification (optional) .....	144
Covariate Classification (optional) .....	146
<b>Regression .....</b>	<b>146</b>
General Information .....	146
LC Regression Model Summary Output .....	147
Parameters Output (optional) .....	151
Parameters Output Subcategory (Advanced) .....	153
Profile Output (optional) .....	153

Profile Plot (for dependent and covariates) .....	155
GProfile Output (Advanced, optional) .....	157
ProbMeans Output (optional) .....	157
Uni-Plot .....	157
Tri-Plot .....	158
Frequencies and Residuals (optional) .....	159
Standard Classification (optional) .....	160
Covariate Classification .....	161
Estimated Values (optional) .....	162
 CHAPTER 7. TUTORIALS .....	 163
 7.1. TUTORIAL #1: USING LATENT GOLD 4.0 TO ESTIMATE LC CLUSTER MODELS .....	  164
The Data .....	164
The Goal .....	165
Opening The Data File .....	165
Estimating LC Cluster Models .....	166
Selecting the Type of Model .....	166
Selecting the Variables for the Analysis .....	168
Specifying the Number of Clusters .....	169
Estimating the Model .....	170
Viewing Output and Interpreting Results .....	170
Assessing Model Fit Using the Bootstrap p-value .....	171
Viewing The Parameters Output .....	172
Profile Output and Associated Profile Plot .....	174
ProbMeans Output and Associated Tri-Plot .....	177
Classifying Cases Into Clusters Using Modal Assignment .....	177
Bivariate Residuals .....	179
Assessing Model Improvement Using the Conditional Bootstrap .....	180
 7.2. TUTORIAL #2: USING LATENT GOLD TO ESTIMATE DFACTOR MODELS .....	  184
The Goal .....	184
DFACTOR Analysis vs. Traditional Factor Analysis .....	184
Opening The Data File .....	185
Estimating a 2-D Factor Model .....	190
Restricting Loadings to Zero .....	195
Viewing Joint Profile Output for the 2-DFACTOR Model .....	198
Classifying Cases .....	199
Viewing the Bi-Plot Display for the 2-DFACTOR Model .....	200
 7.3 TUTORIAL #3: LC REGRESSION WITH REPEATED MEASURES .....	 202
The Data .....	202
The Goal .....	203

<b>Estimating an LC Regression Model</b>	<b>203</b>
Opening a Data File and Selecting the Type of Model	203
Selecting the Variables for the Analysis	205
Specifying the Number of Classes	205
Scanning the Data File	206
Estimating the Model	206
<b>Viewing Output And Interpreting Results</b>	<b>207</b>
Profile Output	208
Parameters Output	209
<b>Restricting Certain Effects To Be Zero Or Class Independent</b>	<b>210</b>
Viewing Output and Interpreting Results	212
Parameters Output	213
<b>Adding Covariates</b>	<b>213</b>
Viewing Output and Interpreting Results	214
Parameters Output	214
Classification Output	215
<b>7.4 TUTORIAL #4: PROFILING LC SEGMENTS</b>	
<b>USING THE CHAID OPTION</b>	<b>217</b>
<b>The Goal</b>	<b>217</b>
Including Covariates in the Models	218
<b>Opening The Data File</b>	<b>218</b>
Selecting the Variables for the Analysis	219
Specifying the Number of DFactors	219
Including Covariates	220
Estimating the Model	225
Viewing the DFactor Loadings	225
Viewing the Profile Output	225
Using the CHAID Option	228
<b>CHAPTER 8. ADDITIONAL TUTORIALS AND ASSOCIATED DATA SETS</b>	<b>239</b>
<b>Tutorial #5: Using Latent GOLD 4.0 With The Known Class Option</b>	<b>239</b>
<b>Tutorial #6: Estimating a Random Intercept Regression Model</b>	<b>239</b>
<b>DATA SETS AND EXAMPLE LGF FILES</b>	<b>240</b>
<b>8.1 Dichotomous, Nominal, Or Ordinal Indicators</b>	<b>240</b>
<b>8.2 Single Response Variable</b>	<b>244</b>
<b>8.3 Continuous, Count And Mixed-Scale Indicators</b>	<b>245</b>
<b>8.4 Latent Class And Random-Effects Regression Modeling</b>	<b>246</b>
<b>8.5 Multilevel Latent Class Models And Complex Surveys</b>	<b>253</b>



# CHAPTER 1. OVERVIEW

## 1.1 Latent Class, Finite Mixture Modeling, and Beyond

Latent classes are unobservable (latent) subgroups or segments. Cases within the same latent class are homogeneous on certain criteria, while cases in different latent classes are dissimilar from each other in certain important ways. Formally, latent classes are represented by  $K$  distinct categories of a nominal latent variable  $X$ . Since the latent variable is categorical, LC modeling differs from more traditional latent variable approaches such as factor analysis, structural equation models, and random-effects regression models that are based on continuous latent variables.

Latent class (LC) analysis was originally introduced by Lazarsfeld (1950) as a way of explaining respondent heterogeneity in survey response patterns involving dichotomous items. During the 1970s, LC methodology was formalized and extended to nominal variables by Goodman (1974a, 1974b) who also developed the maximum likelihood algorithm that serves as the basis for the Latent GOLD program. Over the same period, the related field of finite mixture (FM) models for multivariate normal distributions began to emerge, through the work of Day (1969), Wolfe (1965, 1967, 1970) and others. FM models seek to separate out or 'un-mix' data that is assumed to arise as a mixture from a finite number of distinctly different populations.

In recent years, the fields of LC and FM modeling have come together and the terms LC model and FM model have become interchangeable with each other. A LC model now refers to any statistical model in which some of the parameters differ across unobserved subgroups (Vermunt and Magidson, 2003a). It is the difference in model parameters that distinguishes cases in one latent class from cases in another. In the most basic forms of LC/FM analysis, these are the parameters defining the distributions of the response variables that depending on their scale types correspond to a) response probabilities, b) means, or c) means, variances, and covariances. In LC/FM regression analysis, the parameters that differ across latent classes are the coefficients in the regression model of interest.

Today's fast computers and efficient algorithms make it possible to estimate LC models with many cases, many observed responses (indicators), and many explanatory variables. Extensions and variants of the basic model have been developed to include:

- **response variables of mixed scale types, such as nominal, ordinal, (censored/truncated) continuous, and (truncated) counts**
- **several ordered categorical latent variables called discrete factors ( DFactors)**
- **discrete and continuous covariates predicting class membership**
- **predictors of a repeatedly observed response variable**
- **provisions to relax the local independence assumption**
- **tools for dealing with sparse tables (bootstrap p values), boundary solutions (Bayes constants), local maxima (multiple start sets), and other problems.**

The Advanced version of Latent GOLD 4.0 implements the following additional extensions: the option to take into account a complex sampling design

- **the option to specify LC models that contain one or more continuous latent variables called continuous factors (CFactors)**
- **multilevel extensions of the LC model, which involves inclusion of group-level latent classes (GClasses) and/or group-level continuous factors (GCFactors).**

The option to include continuous latent variables in a model extends Latent GOLD to a more general latent variable modeling program. It cannot only be used to estimate LC and FM models, but also factor analytic models, item response theory models, and random-effects regression models, including mixture and multilevel variants of these.

## 1.2 Kinds of Latent Class Models

Latent GOLD contains separate modules for estimating three different model structures - **LC Cluster models**, **DFactor models**, and **LC Regression models** - which are useful in somewhat different application areas. To better distinguish the output across modules, latent classes are labeled 'clusters' for LC Cluster models, 'classes' for LC Regression models and DFactor or joint DFactor 'levels' in DFactor models. In this manual we also occasionally use the term 'segments'.

**The LC Cluster Model:**

- Includes a K-category latent variable, each category representing a cluster.
- Each cluster contains a homogeneous group of persons (cases) who share common interests, values, characteristics, and/or behavior (i.e., share common model parameters).

Advantages over more traditional ad-hoc types of cluster analysis methods include model selection criteria and probability-based classification. Posterior membership probabilities are estimated directly from the model parameters and used to assign cases to the modal class - the class for which the posterior probability is highest.

**The DFactor Model:**

- Is a restricted form of the LC Cluster model which is often used for variable reduction or to define an ordinal attitudinal scale.
- Contains one or more DFactors which group together variables sharing a common source of variation.
- Each DFactor is either dichotomous (the default option) or consists of 3 or more ordered levels (ordered latent classes).
- Containing  $L > 1$  DFactors may be expressed in terms of cluster model parameters in the Profile Output. For example, a 3-DFactor model containing K1, K2 and K3 levels respectively can optionally be expressed in terms of the joint DFactor consisting of K1 x K2, x K3 'clusters'.

**The LC Regression Model:**

- Is used to predict a dependent variable as a function of predictor variables.
- Includes a K-category latent variable, each category representing a homogeneous subpopulation (segment) having identical regression coefficients
- Each case may contain multiple records (regression with repeated measurements).
- The appropriate model is estimated according to the dependent variable scale type
- Continuous - Linear regression (with normally distributed residuals)
- Dichotomous (specified as nominal, ordinal, or a binomial count) - Binary logistic regression
- Nominal (with more than 2 levels) - Multinomial logistic regression
- Ordinal (with more than 2 ordered levels) - Adjacent-category ordinal logistic regression
- Count: Log-linear Poisson regression
- Binomial Count: Binomial logistic regression model

For any of these three model types:

- **Diagnostic statistics are available to help determine the number of latent classes, clusters, or segments**
- **For models containing  $K > 1$  classes, covariates can be included in the model to improve classification of each case into the most likely segments.**

Further details on each of these model structures including explicit equations defining these models are given in the Technical Guide.

## 1.3 Regression Models with Random Effects and Parameter Restrictions

FM regression models provide an elegant approach for estimating models with discrete random effects. A coefficient  $b$  in a 3-class LC regression model, for example, may be assumed to take on the value  $b_1$  with probability  $p_1$ ,  $b_2$  with probability  $p_2$  and  $b_3$  with probability  $p_3$ . This describes a discrete distribution for the parameter  $b$ , yielding what is also referred to as nonparametric random-effects modeling: In contrast to continuous random-effects models that usually assume normality, no distributional assumption is made about the random effects (Vermunt and Magidson, 2003b). The Advanced version of Latent GOLD 4.0 contains the ability to include continuous random effects in a regression model, an option that can either be used in addition to or in instead of LC-based nonparametric random effects.

A primary goal of modeling is to achieve a parsimonious representation for the model of interest. That is, one that contains the fewest number of parameters needed to provide an adequate fit to the data. To help you accomplish this goal, Latent GOLD contains various ways to constrain the parameters, including restricting certain regression coefficients to be class independent. That is, selected regression coefficients may be restricted to be identical across all classes (e.g., in the example above,  $b_1 = b_2 = b_3$ ). Since these parameters are fixed at the same value for all cases regardless of their class membership, such estimates are referred to as fixed effects. Thus, mixed models containing both random and fixed effects can also be estimated in Latent GOLD.

In addition to the class independent restriction, several other types of restrictions may be applied including zero and other fixed-value restrictions, setting parameters to be equal in selected segments, and different ways of imposing order-restrictions.

## 1.4 Interactive Use

To help you obtain a good parsimonious model, Latent GOLD is designed to facilitate interactive use. The appearance of various output listings and plots may be customized in an interactive manner using different control panel options. The available options are listed in the View Menu and change based on the current output being viewed. Extensive model management allows comparison of estimated models, viewing of different sections of output associated with any model, and to estimate models on different input data files in the same session.

A conditional bootstrap ('Bootstrap -2LL Diff') has been added in Latent GOLD 4.0. Among many possible applications, it may be used to help determine the number of classes (DFactors, DFactor levels) to include in a model by assessing whether a less restrictive form of a model (e.g., one containing more classes), provides a significant improvement over a less restrictive form of the model. To use this option, you simply select the conditional bootstrap for any estimated model, and then choose the restricted form of the model from an eligibility list containing a subset of other models that have been estimated.

A common use of the program might be to estimate several models with different numbers of latent classes, examine various output listings, manipulate interactive graphs and then apply some parameter restrictions and re-estimate the model. You might use various statistical criteria, including the conditional bootstrap to help choose your final model.

## 1.5 New Features in Latent GOLD 4.0

Latent GOLD 4.0 comes in either a basic or advanced version. Latent GOLD 4.0 Advanced consists of Latent GOLD 4.0 Basic plus an Advanced Module. For details on obtaining the Advanced Module see our website. The primary interface improvements and new modeling features in Latent GOLD 4.0 are described below. For further technical details on the modeling features, see the Technical Guide.

### **LATENT GOLD 4.0 BASIC**

These are the new features included in the basic version of Latent GOLD 4.0. In Cluster, DFactor and Regression models:

#### **General interface additions and improvements:**

A **Pause** and **Resume** feature has been added allowing you to Pause during model estimation anytime prior to convergence, review model output and possibly change the output settings or certain

convergence options. You may then Resume estimation of the model. (see Chapter 5, Step 10)

Upon viewing model output, you can change the number of decimal places or significant digits that are displayed. (see Chapter 4)

For uniformity across all modules, two new Output Tabs are included:

- 1) The new **Model Tab** replaces the options previously contained in the Clusters Tab (LC Cluster Module), the Factors Tab (DFactor Module), and the Restrictions Tab (LC Regression Module). It contains various restriction options (new and old) and related specification options.
- 2) The new **ClassPred Tab** contains various new and old Output to a file options associated with prediction and classification that were previously included in the Output Tab and also contains the new **Known Class Indicator**. (See Chapters 2, 5)

**Known Class Indicator** - This feature allows more control over the segment definitions by pre-assigning selected cases (not) to be in a particular class or classes. (See Chapter 5, Step 7: ClassPred Tab.)

**Model difference bootstrap** can be used to formally assess the significance in improvement associated with adding additional classes, additional DFactors and/or an additional DFactor levels to the model, or to relax any other model restriction. (See chapter 5, Step 10.)

**Truncated and Censored scale sub-types** - for use with any model in which the dependent variable or indicator(s) are truncated or censored response variables.

**Dummy coding.** Parameters associated with Nominal latent, predictor, and response variables can be changed from the default Effects coding to be based on Dummy coding, with either the first or last category as reference category. (See Chapter 5, Step 9: Output Tab.) Additional Output includes:

New Dissimilarity Index  
Classification table (see Chapter 6: summary output)  
Covariate classification statistics  
Cook's distances

**Further controls over the production of specific output listings:**

- Covariate classification information and standard classification information are now separate output options
- Specific output file sections such as Parameters output can be specified to not be produced

**Technical improvements:**

- Increased speed of estimating models because of more efficient data handling and slight

improvements in the algorithms. This is especially noticeable in LC regression applications with large data sets

- Much more efficient memory use, making it possible to deal with much larger data sets, more predictors and more latent classes, especially in LC regression applications.
- Improved starting-values procedure using more disperse random starting values and including an option to change the convergence criterion
- Option to obtain robust standard errors

### In LC Cluster models:

**Order-restricted latent classes** - This feature restricts the resulting clusters to be ordered in a way that is less restrictive than estimating a DFactor model with a single DFactor. (See Chapter 5, Model Tab)

**Class Independent Variances** and **Covariances** are now separate options on Model Tab in Cluster (and are removed from the Technical Tab)

### In LC Cluster and DFactor models:

**Binomial count.** This additional indicator scale type has been added.

**Missing values.** The method of dealing with missing values on indicators when direct effects are included in the model has been improved.

**Equal effects.** The Cluster and DFactor effects on the indicators can be restricted to be equal across indicator of the same scale type. This restriction is available for LC cluster models, and also for one or more selected DFactors in DFactor models.

Subsections of Parameters Output includes:

- **Loadings** - these are (approximate) standardized linear regression coefficients for the Cluster-Indicator and DFactor-Indicator relationships,
- **Correlations** (DFactor only) - these are (approximate) DFactor-DFactor and DFactor-Indicator correlations,
- **Error Correlations** - when one or more continuous indicator is included in the model error correlations in addition to error covariances are provided.

## In LC Regression (and Choice) models:

**Zero-inflated models** - when specified, an additional class is automatically added for which the dependent variable takes on the value 0 with probability 1 (for continuous and counts), or takes on a specific value with probability 1 (for ordinal and nominal). (See Chapter 5, Step 3, Variables Tab)

**Offset restrictions** - can be used to restrict a regression coefficient for any numeric predictor equal to one (or to any value) when the dependent variable scale type is other than Nominal. (See Chapter 5, Model Tab)

Improved **"Order Restrictions"** for nominal dependent variables - restrictions are imposed on adjacent category logits now, yielding a "truly" ordinal regression model based on monotonicity constraints

**Additional output** section - Estimated Values provides the class-specific and overall estimated values for each predictor pattern. (See Chapter 6)

**Prediction** - Marginal mean in addition to posterior mean and HB-like prediction

## LATENT GOLD 4.0 ADVANCED

The following new features are included in the optional Advanced Module (requires the Advanced version) of Latent GOLD 4.0:

**Continuous latent variables** (CFactors) - an option for specifying models containing continuous latent variables, called CFactors, in a cluster, DFactor or regression model. CFactors can be used to specify continuous latent variable models, such as factor analysis and item response theory models, and regression models with continuous random effects. If included, additional information pertaining to the CFactor effects appear in the Parameters output and to CFactor scores in the Standard Classification, the ProbMeans, and the Classification Statistics output.

**Multilevel modeling** - an option for defining two-level data variants of any model implemented in Latent GOLD. Group-level variation may be accounted for by specifying group-level latent classes (GClasses) and/or group-level CFactors (GCFactors). In addition, when 2 or more GClasses are specified, group-level covariates (GCovariates) can be included in the model to describe/predict them. The multilevel option can also be used for specifying three-level parametric or nonparametric random-effects regression models.

**Survey options** for dealing with complex sampling data. Two important survey sampling designs are stratified sampling -- sampling cases within strata, and two-stage cluster sampling -- sampling within primary sampling units (PSUs) and subsequent sampling of cases within the selected PSUs. Moreover,

sampling weights may exist. The Survey option takes the sampling design and the sampling weights into account when computing standard errors and related statistics associated with the parameter estimates, and estimates the 'design effect' (see Chapter 6, Model Summary Output). The parameter estimates are the same as when using the weight variable as a Case Weight when this method is used. An alternative two-step approach ('unweighted') proposed in Vermunt and Magidson (2001) is also available for situations where the weights may be somewhat unstable.



Advanced features described in the manual are highlighted throughout the text using this symbol.

## 1.6 Optional Add-ons to Latent GOLD 4.0

The following optional add-on programs are available to link to Latent GOLD 4.0 in various ways:

### **LATENT GOLD CHOICE**

The Choice Module extends Latent GOLD 4.0 to estimate LC conditional logit models, useful in various discrete choice studies (with stated preference or revealed preference data). Dependent variable scale types include:

- *CHOICE* for modeling first choice data. Replication weights may be used to allow estimation of weighted choice/ allocation type models.
- *RANKING* for modeling full ranking, partial ranking, and best-worst data
- *RATING* for modeling ratings, such as those collected in rating-based conjoint studies.

An important application of Latent GOLD Choice is in LC conjoint analysis. It includes the capability to output simulated choices from an unlimited number of constructed scenarios of interest but for which no choice data has been collected (inactive sets), as well as the more usual kinds of output, and it extends the LC Regression Module output when used with rating-based conjoint data.

The Choice Module may also be used to define LC variants of log-linear models for frequency tables, such as models for network data and models for capture-recapture data.

The full LG Choice 4.0 manual is available at <http://www.statisticalinnovations.com/products/choicemanual.pdf>.

Latent GOLD Choice requires an annual license fee.

## SI-CHAID® 4.0

With this option, a CHAID (CHi-squared Automatic Interaction Detector) analysis may be performed following the estimation of any LC model in Latent GOLD 4.0, to profile the resulting LC segments based on demographics and/or other exogenous variables (Covariates). By selecting 'CHAID' as one of the output options, a CHAID input file is constructed upon completion of the model estimation, which can then be used as input to SI-CHAID 4.0.

This option provides an alternative treatment to the use of active and/or inactive covariates in Latent GOLD 4.0. In addition to standard Latent GOLD output to examine the relationship between the covariates and classes/DFactors, SI-CHAID provides a tree-structured profile of selected classes/DFactors based on the selected Covariates. In addition, chi-square measures of statistical significance are provided for all covariates (Latent GOLD does not provide such for inactive covariates). Either the standard (nominal) algorithm or the ordinal CHAID algorithm may be used to profile the classes, the latter useful with order-restricted classes or the levels of a DFactor to take into account the ordered nature of the classes (DFactor levels).

Whenever covariates are available to describe latent classes obtained from Latent GOLD 4.0, SI-CHAID 4.0 can be an especially valuable add-on tool under any of the following conditions:

- when many covariates are available and you wish to know which ones are most important
- when you do not wish to specify certain covariates as active because you do not wish them to affect the model parameters, but you still desire to assess their statistical significance with respect to the classes (or a specified subset of the classes)
- when you wish to develop a separate profile for each latent class
- when you wish to explore differences between 2 or more selected latent classes using a tree modeling structure
- when the relationship between the covariates and classes is nonlinear or includes interaction effects, or
- when you wish to profile order-restricted latent classes or DFactors

For an example of the use of CHAID, see Tutorial #4 in Chapter 7.

## DBMS/COPY INTERFACE

Latent GOLD 4.0 reads SPSS and ASCII text files for data input. The DBMS/Copy interface allows Latent GOLD 4.0 to directly open over 80 additional file formats, including Excel, SAS and HTML files. The full list of file formats is available at [http://www.statisticalinnovations.com/products/latentgold\\_80formats.html](http://www.statisticalinnovations.com/products/latentgold_80formats.html). For further details, see File Import Option in Chapter 3.

Check our website for pricing

## 1.7 Additional Resources

**Technical Guide.** This is the companion manual for Latent GOLD 4.0, an important work which provides a guide to the proper use of the program. It introduces the equations for all models, formulae for all statistics, describes all technical options, and discusses applications and proper interpretation of the output.

**Tutorials.** In addition to the 4 basic tutorials included in Chapter 7 of this manual to get you up and running quickly additional tutorials that illustrate various applications of Latent GOLD 4.0 are under development and will be available on our website. See Chapter 8 for a list of tutorials currently under development and check our website regularly at [www.statisticalinnovations.com](http://www.statisticalinnovations.com) for the addition of new tutorials. As LC modeling continues to evolve, we also wish to extend an invitation to interested users to contribute to the field by developing and submitting tutorials to illustrate their own applications of interest. User applications should be submitted as a .pdf file to [tutorials@StatisticalInnovations.com](mailto:tutorials@StatisticalInnovations.com). Tutorials published on our website will contain names and affiliations of the developers.

**Online courses.** Beginning in April 2005, on-line courses will be offered. These courses cover many of the topics discussed in our articles, tutorials, and publicly held courses (Statistical Modeling Week). Conducted over several weeks, structured assignments are provided for each weekly session. You will have an opportunity to ask questions by posting messages on a special discussion board and receive answers and comments from the instructor on a set schedule. Visit <http://www.statisticalinnovations.com/services/course.html> for more information.

**Demonstration data sets.** Chapter 8 also contains descriptions of several data sets which have been analyzed by LC models previously. They can be analyzed using the demo version of Latent GOLD 4.0 and some are the subject of the tutorials and online courses. They may be downloaded separately from our website at

[http://www.statisticalinnovations.com/products/latentgold\\_datasets.html](http://www.statisticalinnovations.com/products/latentgold_datasets.html).

## 1.8 Structure of the Manual

This manual has eight chapters that describe the functionality of Latent GOLD.

**The Overview (Chapter 1)** this chapter provides a general introduction to LC modeling, basic program features and modules, an overview of new modeling capabilities in version 4.0, and available add-on options.

**General Program Structure (Chapter 2)** provides a detailed overview of the program and all of its functions.

**Data Files and Formats (Chapter 3)** discusses file management, file formats and how to save specif-

ic model settings.

**Working with Output (Chapter 4)** shows how to print output files and how to copy and paste selected output into other programs.

**Basic Steps for Model Development (Chapter 5)** describes in detail the steps for estimating a model.

**Model and Model Summary Output (Chapter 6)** describes the various forms of output generated for a given model, including detailed descriptions of the various plots (Profile Plot, Uni-Plot, Bi-Plot and Tri-Plot).

**Tutorials (Chapter 7)** takes you step-by-step through building and estimating different types of models.

**Additional Tutorials and Associated Data Sets (Chapter 8)** contains descriptions of additional tutorials as well as data sets available from our website.

## CHAPTER 2. GENERAL PROGRAM STRUCTURE

### Windows

Latent GOLD contains a main window called the Viewer.

**Viewer.** When you estimate a model, all statistical results, tables and plots are displayed in the Viewer. In this window, you can navigate easily to whichever section of the output you want to view. You can also copy and paste selected portions of the output into different programs for reformatting and editing. The Viewer window opens automatically when you open the program.

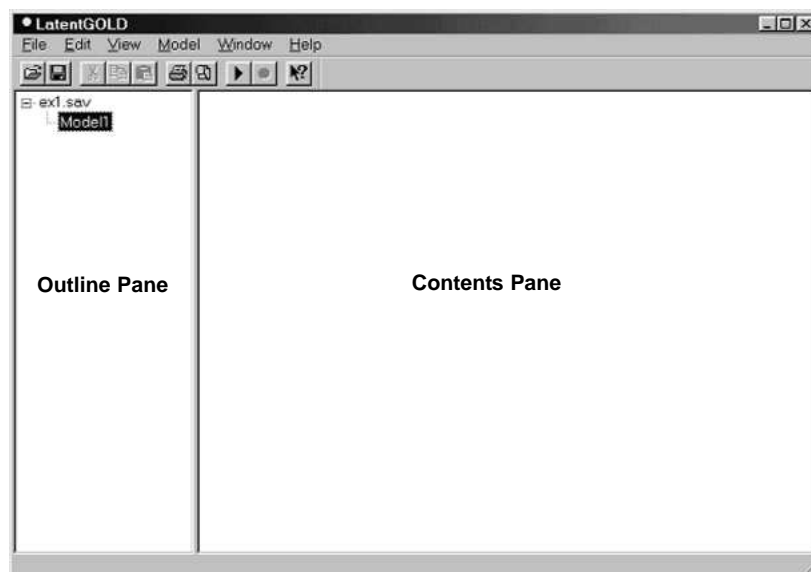


Figure 2-1. Viewer Window

The Viewer window contains 2 panes:

**The left pane, or Outline pane,** provides a hierarchical view of the contents, consisting of the names of data files opened (the outermost level of the hierarchy), the models estimated, and output files generated for each model. To expand/contract any level of the hierarchy, click the +/- icons. After a model has been estimated, the Outline pane will also contain a summary measure of performance (the  $L^2$  or LL statistic) for the model to the right of the model name.

**The right pane, or Contents pane,** contains text output, tables, or plots corresponding to that which is highlighted in the left pane. Many of the tables and plots may be modified interactively using the options shown on a Control Panel or the View Menu.

You can use the scroll bars or the up/down arrow keys to browse results, or you can click an item in the Outline pane to view the corresponding output in the Contents pane.

To change the width of the Outline pane, click and drag the right border or select Split from the Window menu.

The Outline pane of Latent GOLD contains 4 hierarchical levels:

1. The first (outermost) level lists the data files that have been opened.
2. The second level lists model names for 1) models that have been estimated for the associated data file, and 2) models that have not yet been estimated.
3. The third level lists output files produced for each model estimated.
4. The fourth level lists plots and other output subcategories associated with that output file.

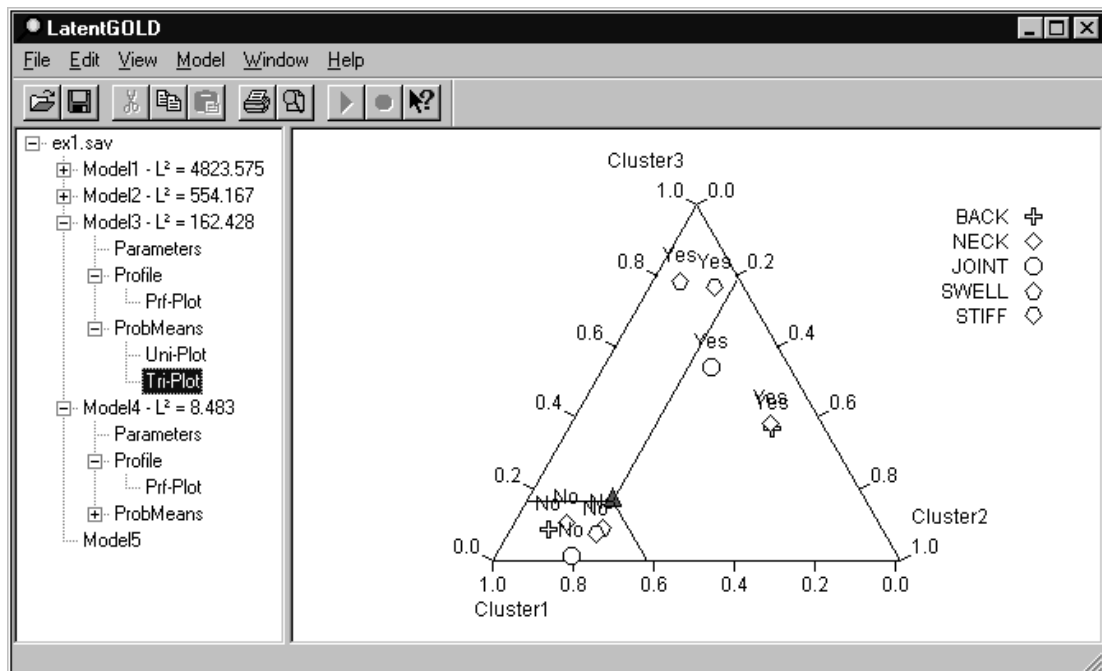


Figure 2-2. Outline Pane levels

For example, in Figure 2-2, the data file opened is ex1.sav (1st level), for which 4 models were estimated - named Model1, Model2, Model3, and Model4 (2nd level). Model5 is the name assigned to a New Model that has yet to be specified. Names associated with the 3 output files associated with Model3 appear in the Outline Pane - Parameters, Profile, ProbMeans (3rd level). The Tri-Plot, associated with the ProbMeans output, is currently highlighted (4th level). When an entry in the Outline pane is highlighted, its content is displayed in the Contents pane (text file or plot).

After one or more models are estimated for a data file, you may specify a new model by opening the analysis dialog box associated with 1) the new model default name (created automatically at the bottom of the list of model names), or 2) any model that has been estimated previously. For previously estimated models, you can specify a new model by modifying the settings for this model. For new models, the default settings correspond to the last model estimated. A detailed description of the 10 basic steps used to specify and estimate a model is given in Chapter 5.

### TO COPY OUTPUT FROM THE VIEWER



Click items in the Contents pane and highlight the portion of the output in the Contents pane you want to copy (you may click Edit, Select All to select the entire contents displayed, then click Ctrl-C or select Edit, Copy to copy the output to the clipboard to be pasted into other programs.

## Menus

The Menu Bar in Latent GOLD has 6 general menu options: File, Edit, View, Model, Window, and Help

### FILE

This file menu can be used to perform the following functions:

**Open.** Opens a data file. Latent GOLD accepts as input data an SPSS system file, an ASCII text (rectangular) file, a special 'array' file format for multi-way tables. In addition, you can Open a previously saved Latent GOLD definition (.lgf) file. For additional information on opening data files, see Chapter 3. Upon opening a data file, the data file name is listed in the Outline Pane (outermost level) and the default file name 'Model1' appears beneath the data file (second level), and may be used to specify and estimate a New Model.

**Close.** Closes the data file highlighted in the Outline Pane.

**Save Results.** Allows you to save your output to either a html or an ASCII (text) file. For more information, see Chapter 3.

**Save Definition.** This saves the analysis settings that have been specified for one or more models on a particular data file. For more information, see Chapter 3

**Print.** Prints output obtained after any model estimation See Chapter 4.

**Print Preview.** Preview printed output on screen. See Chapter 4.

**Print Setup.** Sets various printing options. See Chapter 4.

At the bottom of the File menu, recently opened data files are listed for easier access.

**Exit.** Exit the program. Prior to exiting, Latent GOLD will prompt you to save your Model definitions and results if you have not already done so.

## EDIT

**Copy.** Allows you to copy to the clipboard any output highlighted in the Contents Pane

**Select All.** Selects all output shown in the Contents Pane

**Plot Font.** Allows you to customize the plot font for the output plots in the Content Pane. Upon making a change to the font, this change goes into effect the next time the plot is opened in the Contents Pane.

**Text Style.** Allows you to customize the text style for the output in the Contents Pane. Upon making a change to the text style, this change goes into effect the next time the output listing is opened in the Contents Pane.

**Format.** Allows you to change the format (General, Fixed and Scientific) and number of digits for numeric values displayed in the output. For further details, see Chapter 4.

## VIEW

The options available in the View Menu change depending upon what is highlighted in the Outline Pane. For example, when a model name is highlighted, the options are:

**Toolbar.** Shows the shortcuts Toolbar. See Toolbar for more information

**Status Bar.** Shows the status bar. The status bar displays various information as the model is being estimated. See Status Bar.

**ProbChi.** Opens the ProbChi calculator. This calculator can be used to obtain a p-value for a given chi-square (or vice-versa) for a specified number of degrees of freedom, df. For more information, see Chapter 5.

When an interactive table or plot appears in the Contents pane, the View Menu lists the various options for changing the appearance of the associated output.

### MODEL

The Model Menu options are organized into 3 sections. The first section contains the options for specifying the type of model to be estimated:

**Cluster.** Specifies a Cluster model to be estimated.

**DFactor.** Specifies a DFactor model to be estimated.

**Regression.** Specifies a Regression model to be estimated.

If model names appear in the Outline pane for one or more models that have already been estimated:

If you click on a model name associated with a previously estimated model, the Model Menu contains a checkmark next to the type of model estimated.

If you click on the name for a new model (one that has not yet been estimated), there will be a checkmark next to the last type of model estimated.

Model options appearing in the second section of the Model Menu are

**Estimate.** Estimate the model. Select this option once your new model is fully specified (see Chapter 5)

**Estimate All.** This option may be used when the Outline Pane contains names for 1 or more models associated with a data file that have not yet been estimated. Upon selecting the data file name, the Estimate All menu entry becomes active in the Model Menu. Selection of this option causes all of the associated models that have not yet been estimated to be estimated sequentially beginning with the first such model.

**Note:** Multiple model names associated with models that have not yet been estimated can occur for a data file only if a previously saved definition (.lgf) file containing the setup for 2 or more models is opened. For details on saving .lgf files, see SAVE DEFINITION in the FILE menu options.

**Bootstrap L<sup>2</sup> and Bootstrap -2LL Diff.** The Bootstrap option can be used to estimate the p-value for certain estimated models. See Chapter 5, Step 9, for more information.

**Stop.** The Stop command may be used to pause the estimation prior to completion, or to abandon the estimation completely. See Chapter 5, Step 10, for more information.

**Resume.** If a model is paused (default names for paused models have the characters 'Paused' appended to the original model name -- e.g., 'Model4Paused'), the Resume command may be used to continue the estimation process.

**Delete.** This is used to delete the model name (and any associated output files) from the Outline pane.

## WINDOW

**Split** - Allows you to customize the window split between the Outline and Content Panes

## HELP

**Contents.** Lists all the Help topics available

**Help.** Displays context-sensitive help.

**Item Help.** Creates a help cursor that you can point to get help on any particular item in the program

**Register.** Displays your registration code.

**About Latent GOLD.** Provides general information about the program.

Many of the tasks you will want to perform with Latent GOLD utilize menu selections. Shortcuts for menu items are listed to the right of the item. For example, the shortcut for File, Open is Ctrl+O (on your keyboard, hold down the Ctrl key and then press the 'O' key).

In addition, a right click in the Contents Pane frequently causes a control panel or the appropriate menu options to appear. For example, a right click in a graphical display (such as the tri-plot) causes the (Tri-plot) Control Panel to appear which can be used to modify the appearance of the Plot in an interactive manner. A right click in the Parameters Output retrieves a pop-up menu containing the Options from the View Menu which allow you to change the appearance of the output in various ways such as adding a column for standard errors.

### Toolbar

The toolbar provides quick, easy access to common tasks. When you position the mouse pointer on a tool, ToolTips provides a brief description.



	File Open		File Print
	File Save		File Print Preview
	Edit Cut		Model Estimate / Resume
	Edit Copy		Stop Model Estimation Pause/Abandon/Continue
	Edit Paste		Get Help using mouse cursor



#### To Show or Hide the Toolbar

- ▷ From the menus choose: View → Toolbar



#### To Move the Toolbar

- ▷ Click anywhere in the tool bar outside the Tool Bar buttons.
- ▷ Drag the Toolbar to a new position.
  - Dragging the Toolbar to the left or right side of the window attaches the Toolbar vertically to that side.
  - Dragging the Toolbar to the top or bottom of the window attaches the Toolbar horizontally.
  - Dragging the Toolbar anywhere other than the window borders creates a detached, floating Toolbar.

## Status Bar

The status bar at the bottom left of the Latent GOLD application window provides:

- Current information about a menu option that is highlighted.
- Status of a current model estimation including number of iterations and warning messages.
- Basic descriptive information regarding variable labels and values associated with various objects selected on an active plot. Objects that can be selected include plot symbols, and reference lines. Simply left click an object to select it.



**To Show or Hide the Status Bar**

▷ From the menus choose: View → Status Bar

## Dialog Boxes

Most menu selections open dialog boxes which you will use to select variables and set analysis options. To manually move within a dialog box from one item to another, use the tab key. Alternatively, a shortcut can be used to reach dialog box items which contain an underlined letter as part of their descriptive name. Such items can be reached by typing the letter in conjunction with the Alt key. For example, in the Technical Tab portion of the Analysis Dialog Box, the shortcut to reach the Iteration box is Alt+I (on your keyboard, hold down the Alt key and then press the 'I' key).

### ANALYSIS DIALOG BOX

After you have selected from one of the three types of models, an Analysis dialog box appears that is appropriate for that model type. This dialog box has six basic tabs (plus an optional Advanced Tab) for Cluster and DFactor models and five basic tabs (plus an optional Advanced Tab) for Regression models. Some tabs are common to all model types:

#### Variables Tab

For each of the 3 modules, the first basic tab, called the Variables Tab, has a common structure.

The Source variable list provides a listing of all variables in the working data file. By default the variables are listed in the order they appear in the data file. To list them in alphabetical order, check the Lexical Order box.

Target variable list(s) specify how the variables are to be treated in the analysis. For example, a variable can be specified to be used as an indicator, dependent, predictor, case weight, replication weight or covariate, by selecting and moving it to the appropriate target list box. The target list boxes that are available depend upon the type of model selected.

For more specific information about the options related to a type of model, see Chapter 5.

### Model Tab

The second basic tab is the Model Tab. It contains options that affect the number of model parameters that will be estimated. For an LC Cluster or LC Regression model, this tab may be used to change the settings for the number of classes and to impose various model restrictions such as restricting certain effects to equal 0. For an DFactor model, this tab may be used to change the settings for the number of DFactors and DFactor levels, restrict to zero the effects of certain indicators on selected DFactors and to remove zero restrictions on correlations between selected DFactors.

### ClassPred Tab

The third basic tab is the ClassPred Tab. It allows you to specify a Known Class Indicator, or to use the results from an estimated model to score a data input file by appending selected prediction and/or classification information to it.

### Residuals Tab (Cluster and DFactor only)

The fourth basic tab is the Residuals Tab. This tab displays bivariate residuals (BVRs) and allows you to specify new models containing one or more "direct effect" parameters, each generally associated with a large BVR. A bivariate residual substantially larger than 1 indicates that the estimated model fails to account for all of the pairwise association associated with 2 indicators, or an indicator and an (active) covariate. See section 7.6 in the Technical Guide for further information about BVRs.

### Output Tab

The next basic tab is the Output Tab. This allows you to generate selected output file listings following the estimation of a model. For example, selection of the optional 'Standard Classification' generates a file listing containing the modal assignment and posterior membership probabilities for each case (or each grouped case if 2 or more cases have the same values on all observed variables). The Output Tab also contains options for specifying the methods to be used in the computation of standard errors (Hessian, outer-product, or robust), in the prediction of the dependent variable (posterior mean, HB-like, or marginal mean), and in the coding of nominal variables.

### Technical Tab

The last basic tab is the Technical Tab. It allows you to change the settings for various technical options before a model is estimated. These options include algorithmic settings such as convergence limits, iteration limits, start values, model settings such as Bayes constants, and the use of dummy coding as opposed to effects coding of parameters, bootstrap settings, treatment of missing values, and a control for the precision of CFactors, an optional Advanced feature.



### Advanced Tab (Optional)

This tab contains several advanced features. It requires the Advanced version of Latent GOLD 4.0. The Advanced Tab provides model options for handling complex survey data, multi-level data, and inclusions of up to 3 continuous factors (CFactors) in a model.

## DIALOG BOX PUSHBUTTONS

These buttons instruct the program to perform an action, such as run a procedure, display Help, or open a sub-dialog box to display additional options. Most dialog boxes contain several standard command pushbuttons:

**Close.** Close the current dialog box.

**Cancel.** Cancels any changes made to the dialog box settings since the last time it was opened, and closes the dialog box. Within a session, dialog box settings persist until you cancel or change them.

**Estimate.** Estimates the parameters of the current model. After you select your variables and set any analysis options, click Estimate to start the iterative estimation algorithm. This also closes the Analysis dialog box.

**Pause.** Pauses the model estimation procedure and produces output file listings that should be viewed as preliminary output.

**Resume.** Resumes the model estimation procedure for a paused model.

**Help.** Context sensitive help. This takes you to a standard Help window that contains information about the current dialog box. You can also get help on individual dialog box controls by clicking the control with the right mouse button.

**OK.** Runs the procedure. After you select your variables and set any additional specifications, click OK to run the procedure. This also closes the dialog box.

In addition, the Variables, ClassPred Output, and Advanced Tabs of the Analysis Dialog Box contains these additional command pushbuttons:

**Scan.** This option scans the data file and identifies each distinct category or value for variables specified as Indicators, Covariates, Dependent or Predictors. If this option is not selected before estimating a model, Latent GOLD will automatically scan the data file before estimating a model.

**Reset.** Deselects any variables in the target variable list(s). Note that this has no effect on specifications, such as Scale Type and number of classes.

## SUBDIALOG BOXES

Most procedures provide a great deal of flexibility. For simplicity in specifying models, the main dialog box usually contains only the basic options used for typical models. Other specifications are made in subdialog boxes.

Right-clicking on a variable also brings up additional menus which allow you to change variable properties.

## SELECTING VARIABLES

To select a single variable, highlight it on the source variable list and click the right arrow button next to the target variable list.

- For Cluster or DFactor models, you can also send individual variables to the Indicators box by double clicking them.
- For LC Regression models, the first variable you double-click will be sent to the Dependent box if that box is empty. Subsequent double clicks will send variables to the Predictors box.

You can select multiple variables simultaneously:

- To highlight multiple variables that are adjacent to each other on the variable list, click the first one and then Shift-click the last one in the group.
- To highlight multiple variables that are not all adjacent on the variable list, use the Ctrl-click method. Click the first variable, then Ctrl-click each of the other variables you want to select.

### BASIC STEPS IN MODEL ESTIMATION

Ten basic steps to specify and estimate a model are described in detail in Chapter 5. In general, you can follow these steps to estimate a model:

**Load your data into Latent GOLD.** You can open a previously saved SPSS system file, a rectangular ASCII text file, a special array file format or use a previously saved model definition (.lgf) file.

**Select the type of model.** Select one of the three analysis modules: Cluster, DFactor, Regression or a 4th Module -- LC Choice if you have also have a license for the Latent GOLD Choice add-on program).

**Select the variables for the analysis.** Select and move variables into the appropriate target list boxes and specify scale types.

**Specify the number of clusters/DFactors/classes to be estimated.**

**Scan the data file (optional).** This identifies each distinct category, value, and/or label on your data file for each variable selected for use in the specified model and applies various consistency checks. If you do not specifically request this, a file scan will automatically be performed by the program prior to estimating your model.

**Specify model options.** Specify scale types for your variables, apply any parameter restrictions, change any technical options, and request additional output files. After specifying your variables, by default the program chooses the options for you.

**Estimate the model and view the results interactively.** Results are displayed in textual and graphical form in the Viewer window. Options are available to customize the output in various ways.


Again, detailed information on these steps is presented in Chapter 5.

### Getting Help

Online Help is provided in several ways:

**Help Menu.** Every window has a Help menu on the menu bar. Topics provides access to the Contents, Index and Find tabs, which you can use to find specific help topics.

**Dialog box help.** Press the F1 key for help with the current dialog box.

**Toolbar Help option.** Click on the toolbar symbol  and then click on what you would like help on.



# CHAPTER 3. DATA FILES AND FORMATS

This chapter shows how to open or close a data file and how to save model setups. It also describes the four alternative data file formats that can be used.

## Opening a Data File

The standard program accepts input data:

- 1) saved as an SPSS (.sav) system file,
- 2) saved as an ASCII text (rectangular) file,
- 3) saved as a special array file format, or
- 4) referenced in a previously saved Latent GOLD (.lgf) file. For detailed descriptions about the formats for The DBMS/COPY option allows the program to import data from more than 80 additional file formats, such as Excel, SAS and HTML files, using the File Import menu option. For a full list of additional formats, see [http://www.statisticalinnovations.com/products/latentgold\\_80formats.html](http://www.statisticalinnovations.com/products/latentgold_80formats.html)

See below for details about the optional File Import menu option.



To open a data file using one of the 4 standard formats:

▷ From the menus choose: File → Open

or select the Open icon  from the toolbar.

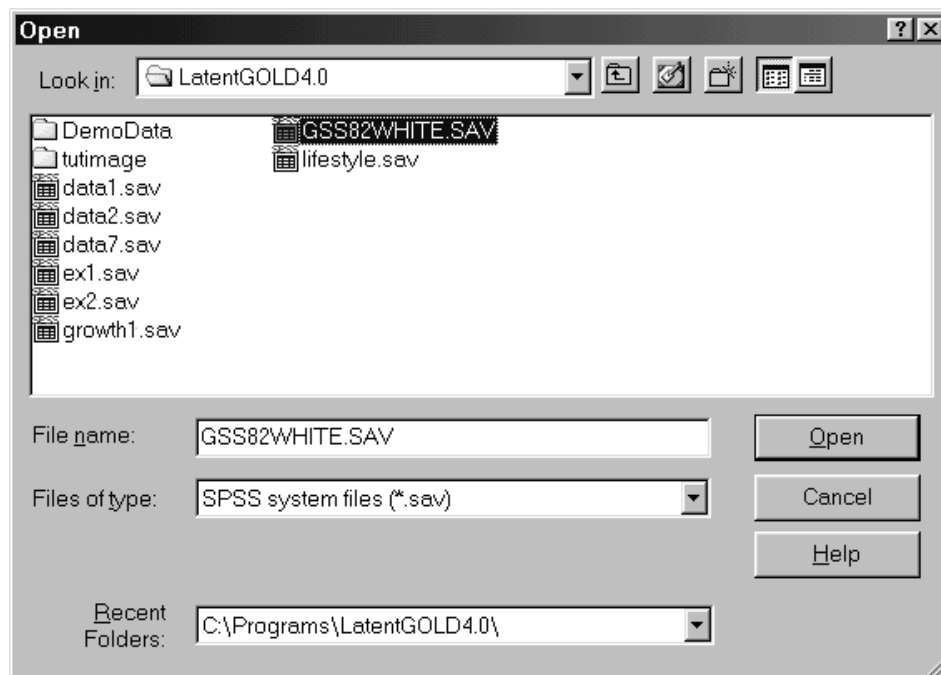


Figure 3-1: File Open Dialog Box

▷ In the Open dialog box, select the type and name of the file you want to open.

## FILE OPEN OPTIONS

**Look in:** Use this option to change drives and directories.

**File Name.** You can type in a filename, a path and filename or a wildcard search. By default, Latent GOLD looks for all files in the current directory with the same extension as the last file that was opened and displays them in the dialog box. If you type wildcards, all files that match the wildcard search in the directory will appear in the dialog box.

**Files of Type.** You may specify the data file format to be one of the following:

- Latent GOLD files (\*.lgf). Lists previously saved Latent GOLD analysis files.
- Arrays (\*.ar\*). Lists ASCII array files. These files contain frequency counts for each cell in a multi-way table.
- Text files (\*.dat, \*.csv, \*.txt). Lists text files with .dat, .csv or .txt extensions. These files are

often referred to as rectangular or ASCII text files.

- SPSS system files (\*.sav). Lists SPSS system files with .sav extensions.
- All Files (\*.\*). Lists all files in the current directory.

### SPSS, TEXT AND ARRAY FILES

For SPSS system files, Text files and Arrays, once you have selected a file, the data file name appears in the Outline pane with a default model name ('Model1') below it.

- ▷ Either double or right click 'Model1' to open the Model Selection menu.

### File Import Option

#### IMPORTING SAS, EXCEL, AND OTHER DATA FILE FORMATS

If you have licensed the DBMS/COPY add-on, an additional option called "Import" appears in the File menu. You can use this option to open data files saved in any of these 80 formats. Selection of File Import opens a dialog box as shown in Figure 3-2.

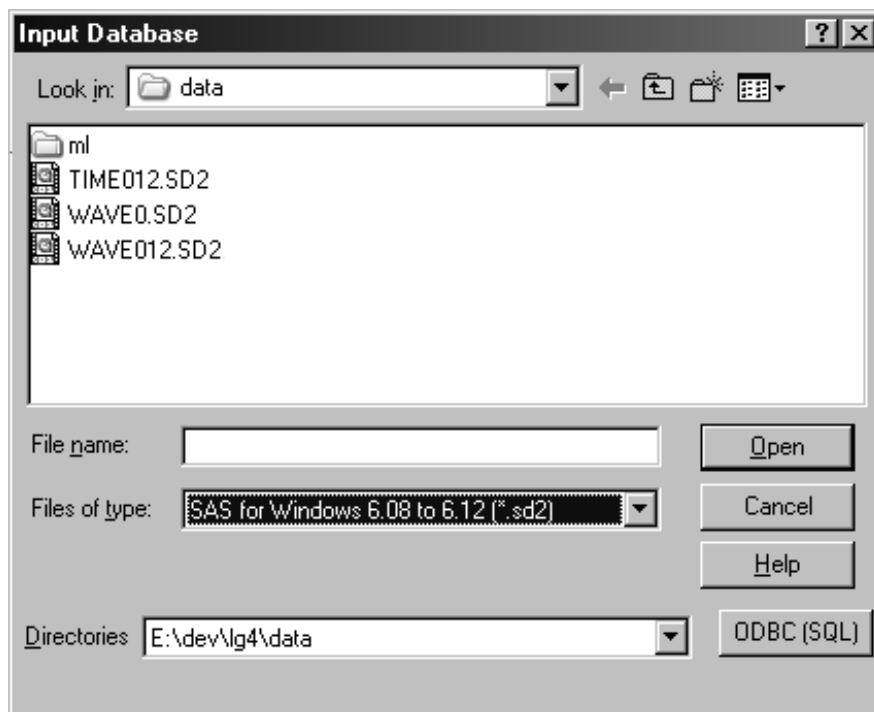


Figure 3-2: File Import Dialog Box

**Files of Type:** shows available file types. The database file type displayed by default is the type last used.

**Directories:** shows the last 10 directories that you have used. You can click on one and it becomes the current directory.

**ODBC (SQL):** this button invokes the Microsoft Open DataBase Connectivity/SQL (Structured Query Language) subsystem. This is how you get to Microsoft Access, Oracle and others. From here you will select an ODBC driver and use the Query Builder to create an SQL query.

**Note:** The option to output data to a file using the ClassPred Tab is not available when the data file was input using the File Import option.

## Latent GOLD Files

If you have previously saved the settings for one or more models in a Latent GOLD definition (.lgf) file using the File Save Definition menu option (see 'Saving Model Settings & Output' below), upon opening this file, a list of model names associated with the saved model settings appears in the outline pane.



**To estimate one of these models:**

- ▷ Double click a model name to open the Analysis dialog box for that saved model setup
- ▷ Make any desired changes to the settings (using the saved setup as the starting point).
- ▷ Click Estimate.

or

- ▷ Click  on the toolbar to estimate the model.

Alternatively, to estimate all models that have not yet been estimated.

- ▷ click on the data file name in the Outline Pane  
select 'Estimate All' from the Model Menu

## RE-OPENING DATA FILES QUICKLY

**Recent Folders:** the software retains a list of recently used folders that you have used: these can be quickly selected from this list.

Up to eight data files that have been opened most recently in Latent GOLD will be listed at the bottom of the File Menu (just above the Exit option). To re-open these data files quickly, simply select the data file name from the File Menu.

## Data File Formats

Latent GOLD will load data from input files in any of the following data formats:

- **SPSS system file (\*.sav)**
- **ASCII text file (\*.dat, \*.csv or \*.txt)**
- **ASCII array (array format) file (\*.arr)**
- **Latent GOLD save file (\*.lgf)**

Use of an SPSS .sav file (or the array format if all your variables are categorical) allows you to assign category labels to particular categories or values.

**Note:** The option to output data to a file using the ClassPred Tab is not available when the input data is an array format.

The Latent GOLD save file can be generated only after estimating a model.

## SPSS FILES

For information on creating an SPSS system file, see your SPSS manuals (programs such as DBMS Copy may be used to convert SAS and other data formats to the SPSS format).

- Latent GOLD recognizes SPSS variable names and category (value) labels (less than 40 characters), but ignores SPSS variable labels and date variables.
- A variable used in SPSS to weight cases is automatically placed in the Frequency box.
- If an SPSS system file contains multiple user missing designations for a numeric variable, Latent GOLD will combine them into one aggregate 'missing' category. Latent GOLD does not recognize user missing specifications for non-numeric variables.
- String variables with lower and upper case letters, such as 'f' and 'F', are distinguished as separate categories.

## TEXT FILES

A text (rectangular) file can be created in any spreadsheet or word processor using the 'Save as..' command to save the data as a text file with either a .dat, .csv or .txt extension.

- The first line of the file needs to contain the variable names separated by blank spaces or a tab. The remainder of the file contains the data.
- For each record, the data for each variable should be separated by either a space or a tab (no other delimiters, such as " , " are allowed).

## LATENT GOLD® 4.0 USER'S GUIDE

- The data for any variable may be numeric (quantitative values only) or may be a string variable, containing some or all alphabetic characters (such as 'Female', 'Male'). For numeric data, do not use commas (such as 3,634). String variables containing lower and upper case letters, such as 'f' and 'F', are distinguished as separate categories.
- Missing data for a variable may be specified using '.'.
- Each data record should contain exactly the same number of data elements, one for each variable name.

Below is a partial listing of a text formatted data file. This file contains one record per case. Note that records 6 and 7 are identical (0 30 0 F).

IMPROVE	AGE	TREAT	GENDER
0	23	1	F
0	23	0	F
1	27	1	M
0	29	1	M
0	30	1	M
0	30	0	F
0	30	0	F
1	31	0	F
2	32	1	M
0	32	1	F
0	32	0	F
2	33	0	F
1	37	1	F
0	37	0	M

For data sets consisting of many observations that contain identical values on all variables, the inclusion of an optional integer value frequency count variable will reduce the number of physical records in the file. Below is a partial listing of another text formatted data file. For each data record, FREQ contains the count of observations having the specified values on each of the variables. For example, the value of FREQ for the first record, defined by the values specified as "1, 1, 1, 1, 1" is 3,634.

BACK	NECK	JOINT	SWELL	STIFF	FREQ
1	1	1	1	1	3634
1	1	1	1	2	73
1	1	1	2	1	87
1	1	1	2	2	10
1	1	2	1	1	440
1	1	2	1	2	89
1	1	2	2	1	106
1	1	2	2	2	75
1	2	1	1	1	295
2	2	2	1	2	162
2	2	2	2	1	44
2	2	2	2	2	176

## ARRAYS

In the case that all variables are categorical, a special array file format may be used. The array file can be created in any word processor or editor program. It contains frequency counts for each cell in the multi-way table. The general ASCII format is:

```

ARRAY
<d1> <d2> <d3> ...
LABELS
<var>/<cat1>/<cat2>/<cat3>/...
...
SCORES
<var>/<scr1>/<scr2>/<scr3>/...
...
DATA
f1...11 f1...12 f1...13 ...
f1...21 f1...22 f1...23 ...
... fn...n

```

The array below re-expresses the text formatted dataset shown above in array format with category labels (1 = 'No', 2 = 'Yes').

```

. ARRAY
. 2 2 2 2 2
. LABELS
. BACK/No/Yes
. NECK/No/Yes
. JOINT/No/Yes
. SWELL/No/Yes
. STIFF/No/Yes
. DATA
. 3634 73 87 10 440 89 106 75
. 295 25 15 5 137 42 35 39
. 489 37 23 7 255 116 71 65
. 306 48 16 11 229 162 44 176

```

The key words "**ARRAY**", "**LABELS**", "**SCORES**" and "**DATA**" must begin in column 1 and be written in upper case. Following the "**LABELS**" key word, exactly one label record must appear for each variable. The "**SCORES**" section of the file, which may be used to assign scores to the categories, is an optional section. If the keyword "**SCORES**" appears in the file, at least one score record must follow.

The frequency counts for each cell are entered on lines following the **DATA** keyword in free format, and must be entered in the order such that the first dimension varies least rapidly..., and the last dimension most rapidly. (The use of scientific notation for frequency counts is not supported.)

## LATENT GOLD SAVE FILE (\*.LGF)

This File, Save menu option can be used to save the variable selections and other option settings used for one or more previously defined models in estimating the model(s) in the format of a Latent GOLD save file (.lgf) for a particular model or series of models (see Saving Model Settings). This file can then be opened at a later time using the File, Open command. The advantage of an .lgf file is the ability to retrieve analysis settings for a particular model or series of models at a later time without having to re-specify these settings. You can also use a retrieved setting as a starting point for specifying a similar model on the same data. In addition, if a saved definition file

## Saving Model Settings & Output

Latent GOLD allows you to save your model settings for

- **an individual model**
- **a series of models estimated from one data file.**
- **model output**

The File / Save menu option incorporates two distinct save features -- **File / Save Definition**, and **File / Save Results**.

**File / Save Definition** creates an .lgf text file containing the model settings so that the model can be re-estimated at a later time. To save the settings for a model prior to that model being estimated, after selecting the desired model options click the Close button. Then, select Save Definition for the File menu. When creating an .lgf file for a model that has been estimated, the .lgf file contains the model settings together with the best start seed from the previously estimated model. When the best seed is obtained from Latent GOLD 4.0, this insures that re-estimation of the model using Latent GOLD 4.0 with this .lgf file will yield the same results as before. For further details on this see **Start Values Procedure**.

**File / Save Results** creates a ASCII text or HTML file containing selected output from a previously estimated model. Either output from all estimated models (Save All Views), all output from a single model (Save Model Views), or any one selected output section (Save View) can be saved.

## SAVE DEFINITION



This creates an .lgf text file containing the model settings for one or more models associated with a particular data file.

▷ Highlight a model name in the Outline pane,  
or to save the settings for all models associated with a data file

- ▷ Highlight the data file name in the Outline pane
- ▷ From the menus choose: File → Save Definition

(or select the Save icon  from the toolbar). The Save dialog box will open.

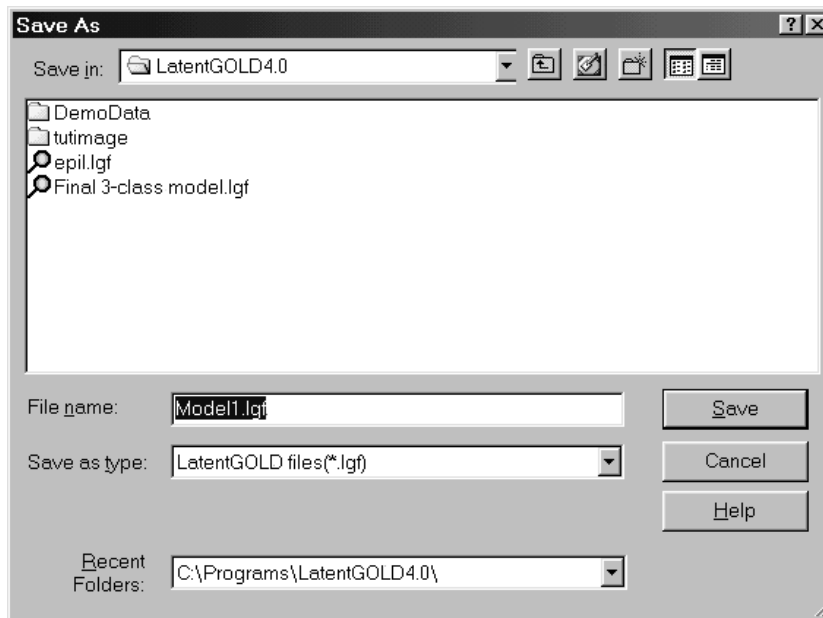


Figure 3-3: Save Definition Dialog Box

- ▷ In the Save dialog box, specify the directory you wish to save to and type in a filename (Latent GOLD will prompt you with a default file name consisting of the model name and .lgf extension.)
- ▷ Click Save. Save Options. File Name.  
Enter the name of the document (Latent GOLD will prompt you with a default name and a .lgf extension).  
Save as Type.  
Select Latent Gold files (\*.lgf).

The specifications for a particular model that are saved are:

All settings from the Analysis dialog box, such as

- Variable settings (variables selected as indicators, covariates, frequency, dependent or predictors) and scale types set under the Variables Tab.
- Any restrictions or other settings chosen in any of the other tabs.

## Save Results



This feature allows you to save the output file in either HTML or ASCII format for later viewings.

- ▷ Highlight the model name in the Outline pane (If you select File, Save without highlighting a particular model name, Latent GOLD will save the model that corresponds to the output appearing in the Contents pane.)
- ▷ Highlight the data file name in the Outline pane
- ▷ From the menus choose: File → Save Results



(or select the Save icon from the toolbar).  
The Save dialog box will open.

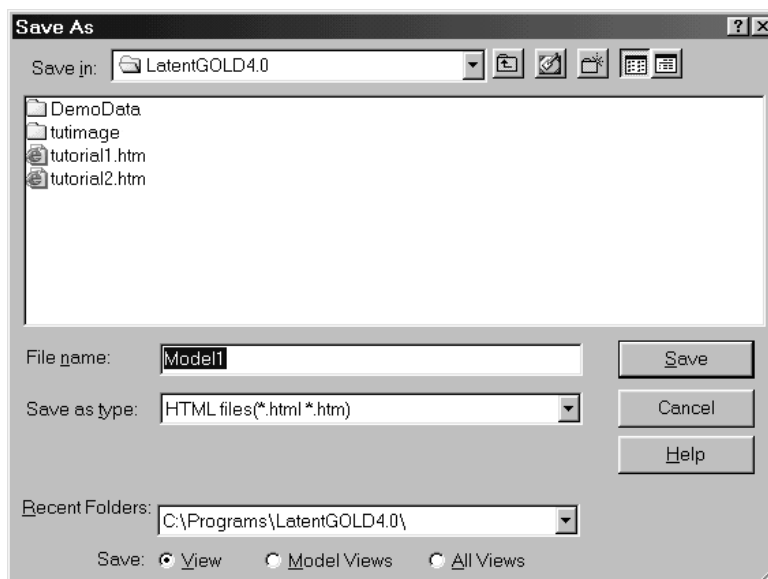


Figure 3-4: Save Results Dialog Box

At the bottom of the dialog box you have several options:

**Save: View:** Click to save only the selected output section from a model

**Save: Model Views:** Click to save all output from a single model

**Save: All Views:** Click to save all output from all estimated models

### Closing a Data File



To close a previously opened Latent GOLD data file,

- ▷ Click on the data file name listed in the Outline pane.
- ▷ From the menus choose: File → Close

Latent GOLD will close all windows and dialog boxes (including output files) associated with the data file.



## CHAPTER 4. WORKING WITH OUTPUT

Output may be printed directly from within Latent GOLD or saved as an HTML or ASCII file. It can also be customized in various ways.

### Printing Output (File Menu)

Use the File, Print option to print the output obtained after any model estimation. A line at the top of any printed output lists the output type and page number. A line at the bottom of each page lists the date, model name, model type and data file name.



**To print the contents of an output file or a plot listed in the Outline pane:**

- ▷ Select the output file from the Outline pane (the current output file will be highlighted).
- ▷ From the menus choose: File → Print



Or select the print icon from the toolbar. This opens the Print dialog box.

- ▷ Select the print settings you want.
- ▷ Click OK.

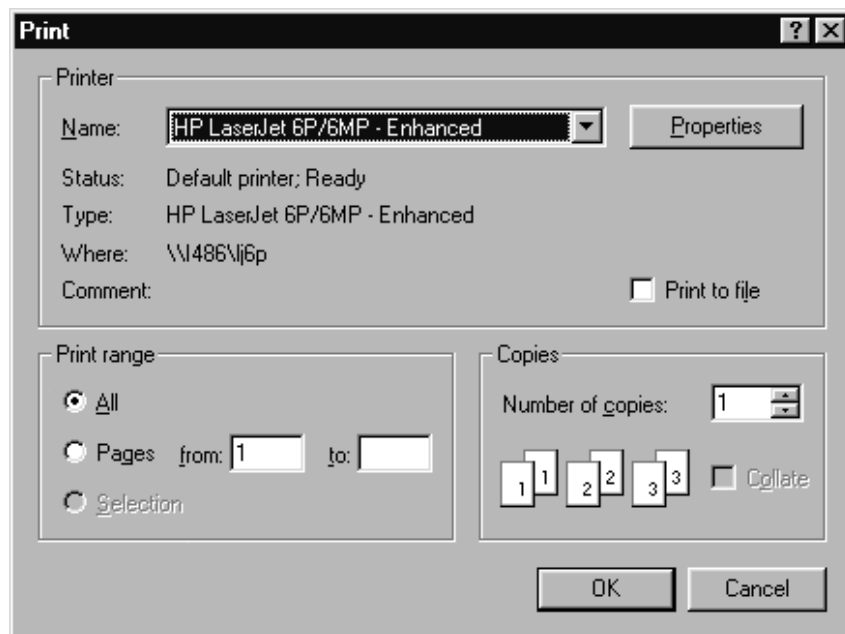


Figure 4-1. Print Dialog Box

The following options allow you to specify how the document should be printed:

**Printer.** This is the active printer and printer connection.

**Print Range.** Currently, all Latent GOLD output is printed on as many pages as needed. Each plot fits on one page.

**Copies.** Specify the number of copies to print.

**Properties.** Click Properties to display a dialog box containing additional print options, specific to the type of printer you have selected.

## PRINT PREVIEW (FILE MENU)

Use the File, Print Preview option to view the selected output prior to it being printed. When you choose this option, the main window is replaced by the print preview window in which the output is displayed one or two pages at a time.



To preview a particular type of output listed in the Outline pane:

- ▷ Select the output file from the Outline pane.
- ▷ From the menus choose: File → Print Preview

The print preview toolbar offers you the following options:

**Print.** Open the print dialog box.

**Next Page.** Preview the next page.

**Prev Page.** Preview the previous page.

**One Page / Two Page.** Select one or two pages at a time to preview.

**Zoom In.** Obtain a close-up view of the page.

**Zoom Out.** Obtain a global view of the page.

**Close.** Return from print preview to the Viewer window.

### PRINT SETUP (FILE MENU)

Use the File, Print Setup option to set various options for printing, such as printer selection, paper size, orientation and other options.

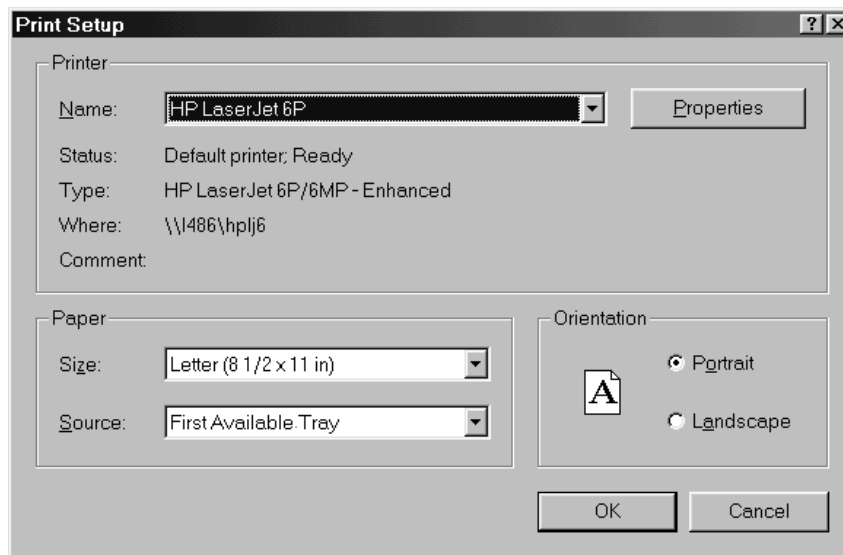


Figure 4-2: Print Setup dialog box

This allows you to set the following destination printer options:

**Name.** Select the printer you want to use. Either use the default printer or select one of the installed printers shown in the drop-down list. Use the Windows Control Panel to install printers and configure ports.

**Paper.** Select the size and source of the paper you are using.

**Orientation.** Select Portrait or Landscape.

**Properties.** Click Properties to display a dialog box in which you can make additional choices, specific to the type of printer you have selected.

### Changing Fonts (Edit Menu)



To change the font, font style or font size for your plots, use the **Edit, Plot Font** command (this command does not change the text or spreadsheet output).

- ▷ To change the font for a selected plot, from the menus choose  
Edit → Plot Font → Customize...

This will open the Font dialog box.

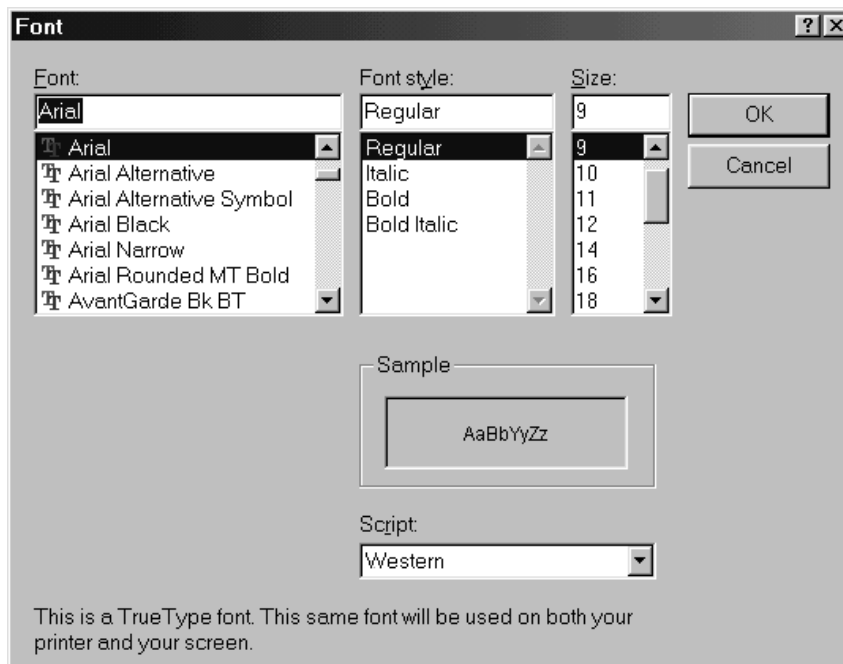


Figure 4-3: Edit Font dialog box

You may select from the font styles your computer supports. When finished, click OK to implement the changes.

- ▷ To change the font, font style, or font size for text or spreadsheet output, or to change other options relating to table borders and lines, from the menus choose  
Edit → Text Style → Customize...

This will open a dialog box with three tabs: tabs for Color and Border styles are present in addition to a Font tab.

## Changing Numeric Format (Edit Menu)

The dialog provides for control of the display of numeric values in any or all output listings associated with a selected model. For any selected output listing, the precision, type and scope aspects of the format can be set.

- ▷ To change the format for a selected output listing associated with an estimated model, from the menus choose

Edit → Format → Customize...

This will open the Format dialog box.

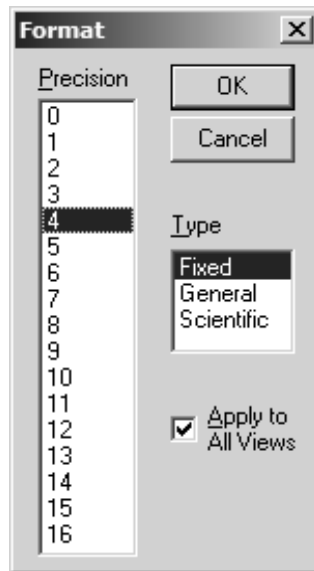


Figure 4-4. Numeric Format Dialog Box

### Precision

The interpretation of the selected quantity depends on the Type. For Fixed Type, the Precision sets the number of decimal places. For General and Scientific Type, the Precision sets the number of significant digits.

### Type

**Fixed.** A fixed number of digits are displayed to the right of the decimal point: 0.0646

**Scientific.** A fixed number of digits are displayed, followed by the exponent: 6.463e-002

**General.** The format of the number depends on the magnitude; in the 1.0e-4 to 1.0e4, a fixed point representation with <precision> significant digits is selected. Outside that range, a Scientific format is chosen, with <precision> digits. For example, 0.06463.

### Scope

**Apply to All Views.**

## LATENT GOLD® 4.0 USER'S GUIDE

The format selected will apply to all views of the selected model, all models corresponding to the selected file name (checked) or just the currently visible view (unchecked).

After selecting OK, the new formatting appears when the output file(s) are next opened.

# CHAPTER 5. BASIC STEPS FOR MODEL DEVELOPMENT

This chapter provides step by step instructions on how to define and estimate each of the three types of LC models (Cluster, DFactor or Regression) and also details the various model options. You need not follow all 10 steps in order to perform an analysis. For example, if you choose to maintain the default Technical and Output option settings, then you would skip Steps 8 and 9.

These steps are illustrated in Tutorial #1 (Chapter 7) for simple LC Cluster models.

## Step 1: Load Your Data into Latent GOLD



▷ From the menus choose: File → Open



Or select the Open icon from the toolbar.

▷ In the Open dialog box, select the type and name of the file you want to open.

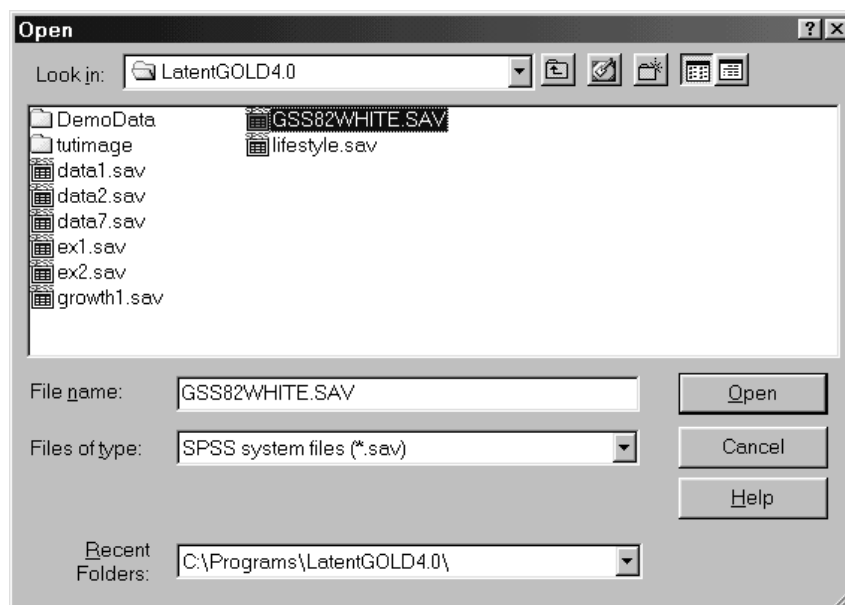


Figure 5-1. File Open Dialog Box

▷ Click Open.



**For more detailed information on File Open options and on the types of files Latent GOLD can read, see Chapter 3: Data Files and Formats.**

## Step 2: Select the Type of Model

If you open a data file, the Outline pane will contain the data file name plus the default model name ('Model1'). If you open a model definition (lgf) file, the settings for all models previously saved in this definition file will be loaded into the program and the Outline pane will contain a list of model names for each of these models.

Assuming that you have opened a data file and no models have yet been estimated for the correct data file, the next step is to use the Model Menu to choose from one of the three analysis modules:

## CHAPTER 5. BASIC STEPS FOR MODEL DEVELOPMENT

MODEL TYPE	ANALYSIS PERFORMED
Cluster	LC cluster analysis
DFactor	DFactor analysis
Regression	LC segmentation/regression analysis
Choice	LC choice analysis (requires Latent GOLD Choice license)



A quick way to open the Model Menu is to right click on a model name. If no models have yet been estimated for the current data file, the default name for the new model to be estimated (Model1) will be the only model name that appears in the Outline Pane.

▷ Right click on the model name and the Model Menu appears:

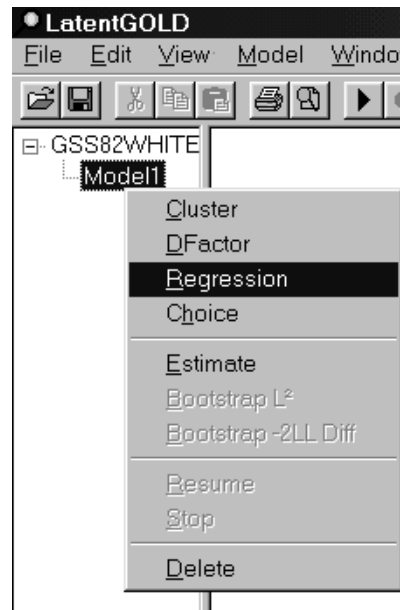


Figure 5-2. Model Menu

The Model Menu options are organized into 4 sections. Those used in step 2 are listed in the top section. The other options are discussed in step 10.

### The options for Model Type:

#### Model

#### Cluster/DFactor/Regression

▷ Choose from one of these to select the type of model to be estimated.

Once you select a model type, the Variables Tab of the appropriate Analysis Dialog Box opens. For example, Figure 5-3 shows the Variables Tab associated with the Cluster model type.

As an alternative to beginning your model setup from scratch, you may begin by altering the settings of a previously defined model. If one or more models have already been estimated for a data file, or if you opened a previously saved .lgf files you can right click on its name to display the Model Menu, which will contain a checkmark next to the type of model currently associated with that name. You may select that model type again or select one of the other model types. To maintain the model type, instead of a right click, you may simply double click on the name of the model and the Variables Tab corresponding to the associated model type opens with the current settings for that model.

If you right click on the name of a new model created automatically at the bottom of the list of model names, the checkmark appears next to the last model type estimated on the associated data file.

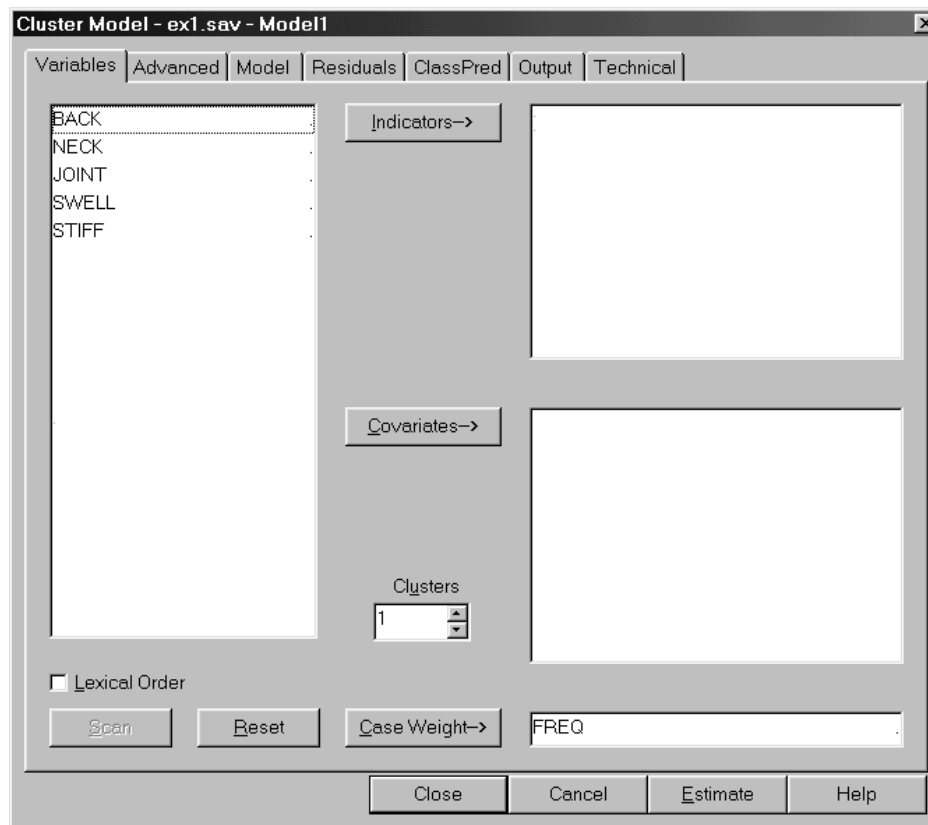


Figure 5-3: Variables Tab for an LC Cluster Model

The next several steps show how to specify the desired model using the various tabs of the Analysis Dialog Box. Specifically, use of the setup options in the Variables Tab are described in Steps 3, 4, 5 and 6, the Models, Residuals, and (optional) Advanced Tab in Step 7, the Technical Tab in Step 8 and the Output Tab in Step 9.

The appearance of the various tabs differs by Model Type. However, regardless of the type of model, the Title bar (above the main menu in any open Analysis dialog box) lists the model type, data file name and model name as identifying information. In addition, the following buttons appear below any Analysis Dialog Box.

**Close.** Closes the analysis dialog box but retains any changes you have made since opening the dialog box.

**Cancel.** Closes the analysis dialog box without retaining any changes you have made since opening the dialog box.

**Estimate.** Initiates estimation of the model based on the current model option settings

**Help.** Provides help associated with the options for the particular model type.

### Step 3: Select Variables for the Analysis (Variables Tab)

The appearance of the Variables Tab differs depending on the type of model you have selected. The primary differences are:

- **For LC Cluster and DFactor models, one or more variables must be selected from the variables list box to be used as Indicators, also known as Dependent Variables. For LC Regression models, one and only one Dependent Variable is used and one or more variables may be selected as Predictors.**
- **The Regression Module differs from the Cluster and DFactor Modules in that it can accommodate multiple records (repeated measures) for one or more cases. Thus, the use of a case ID and a Replication Weight is permissible in Regression.**

#### RESET button

For any model type, at any time during the model setup process, the Reset button can be used to restore the default settings (maintaining only the specified number of classes). Upon selecting Reset, all selected variables (including the Case Weight variable) are returned to their original position in the variables list box. The Reset Button is located in the lower left portion of the Variables Tab.

In addition, for any model type, the Variables Tab can be used to specify a variable as a Case Weight, and one or more variables as Covariates. The specification of variables as a Case Weight or Covariate(s) is optional.

#### SPECIFYING A CASE WEIGHT (OPTIONAL)

You may assign one variable to be used as a case weight (usually a frequency variable) in either a Cluster, DFactor or Regression model. See Tutorial #1 in Chapter 7 for an example that describes the use of a case weight.

For an SPSS .sav file in which a variable is designated as a weight variable (according to the SPSS data dictionary), this variable appears in the Case Weight box automatically upon opening the .sav file.



**To specify a case weight:**

- ▷ In the Variables Box, highlight the variable you want as your case weight.
- ▷ Click "Case Weight" to move this variable to the Case Weight Box.

## COVARIATES (OPTIONAL)

Covariates are variables that may be used to describe or predict (rather than to define or measure) the latent classes and if Active, to reduce classification error. For example, they are often used to profile the latent classes in terms of demographic or other exogenous variables. Covariates may be treated as Nominal or Numeric and may be Active or Inactive (see Step 5: Set Scale Types). Select any variables you want to use as covariates. For a formal distinction between covariates (zcov), predictors (zpred), indicators and dependent variables (y), see section 2.1 of the Technical Guide.



Group level covariates (GCovariates) If 2 or more group latent classes (GClasses) have been included in the model, any Active covariates can be specified as group level covariates (GCovariates) for describing these GClasses. For details, see the Advanced Modeling Options in Step 7.

## INDICATORS (CLUSTER AND DFACTOR MODELS)

Indicators are dependent variables that are used to define or measure the latent classes in a LC Cluster model, or latent variable(s) in a DFactor model. Indicators may be treated as Nominal, Ordinal, Continuous, Poisson Count, or Binomial Count (see Step 5: Set Scale Types).

Select one or more variables from the variables list box to be used as **Indicators** (required).

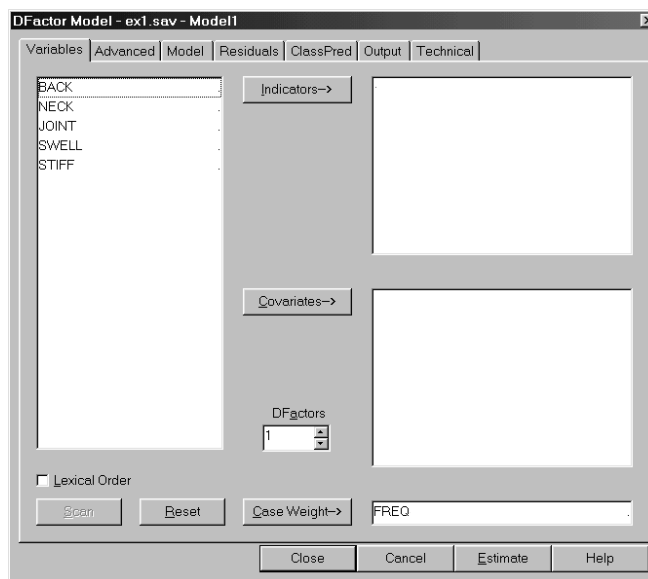


Figure 5-4: Variables Tab for a DFactor Model

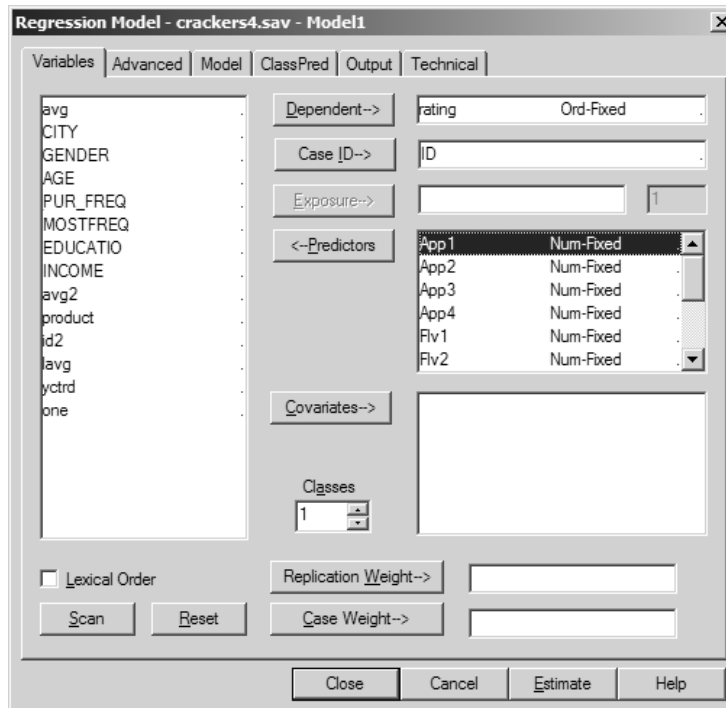


Figure 5-5: Variables Tab for LC Regression Model

## DEPENDENT (REGRESSION MODELS)

Assign one variable to be used as the dependent variable. Dependent variables may be specified as Nominal, Ordinal, Continuous, Count or Binomial Count (see Step 5: Set Scale Types). The Continuous scale type causes the normal distribution to be used resulting in a **linear** LC Regression model.

The appropriate model is estimated according to the dependent variable scale type:

**Continuous.** Linear regression (with normally distributed residuals)

**Dichotomous** (specified as nominal, ordinal, or a binomial count). Binary logistic regression

**Nominal** (with more than 2 levels). Multinomial logistic regression

**Ordinal** (with more than 2 ordered levels). Adjacent-category ordinal logistic regression

**Count.** Log-linear Poisson regression

**Binomial Count.** Binomial logistic regression model

The subtype **censored continuous** yields a tobit regression model. The subtypes **truncated continuous**, **truncated count**, and **truncated binomial** count yield truncated versions of the linear, log-linear Poisson, and binomial logistic regression model, respectively.

## **PREDICTORS ( REGRESSION MODELS, OPTIONAL)**

Select any variable(s) to be used as predictors of the dependent variable. Predictors may be treated as Nominal or Numeric (see Step 5: Set Scale Types) and as Class Independent or Class Dependent.

If no predictors are selected, the model will contain an intercept only.

## **EXPOSURE (REGRESSION MODELS FOR COUNTS, OPTIONAL)**

The Exposure button is active only if the scale type for the dependent variable has been specified to be Count or Binomial Count. (For other scale types, no exposure variable is used.)

For dependent variables specified as Count or Binomial Count, the exposure is specified by designating a variable as the exposure variable or, if no such variable is designated, by entering a value in the exposure constant box which appears to the right of the Exposure variable box. The use of an exposure variable allows the exposure to vary over cases.

By default, the value in the Exposure constant box is 1, a value often used to represent the Poisson exposure. To change the exposure constant, highlight the value in the exposure constant box and type in the desired value. If an exposure variable is selected, this option is not available.

When the scale type is specified as Binomial Count, the value of the dependent variable represents the number of 'successes' in N trials. In this case, the exposure represents the number of trials (the values for N), and hence should never take on a value lower than the value of the dependent variable and hence typically should be higher than the default constant of 1. Before the actual model estimation, Latent GOLD checks each case and will provide a warning message if this condition is not met for one or more cases. An exposure variable should be designated if the number of trials is not the same for all cases.

## **Regression Models with Repeated Measurements**

For LC Regression models where more than one record is included for at least one case, a Case ID variable must be specified and a Replication Weight may be used.

## **CASE ID VARIABLE (REGRESSION MODELS, OPTIONAL)**

If the data file contains only 1 record per case, no Case ID variable is required. For data files in which one or more cases have multiple records (e.g. repeated measures), a variable must be assigned as a Case ID variable to uniquely identify each case. For a regression example with repeated measures is given in Tutorial #3 (see Chapter 7).

To assign a variable as a Case ID, select that variable from the Variable List Box and click the Case ID button. The variable is moved to the Case ID box..

For data files containing multiple records per case:

- A case weight (if used) should take on the same value for each record associated with the same Case ID.
- Any covariates should take on the same value for each record associated with the same Case ID.
- If these conditions are not met, a warning message is produced during the file Scan (Step 4).

### REPLICATION WEIGHT (REPEATED MEASURES REGRESSION MODELS, OPTIONAL)

Once a variable is selected as a Case ID, the Replication Weight button becomes active. To assign a variable as a Replication Weight, select that variable from the Variable List Box and click the Replication Weight button. The variable is moved to the Replication Weight box. A common application of replication weights is in the estimation of certain kinds of allocation models, where respondents assign a fixed number of points to each of J alternatives. For each case, the assigned points are used as replication weights to weight each of J responses. A weighted multinomial logit model is estimated. See Section 5.1 of Technical Guide for further details about replication weights.

### Step 4: Scan the Data File (Variables Tab)

The next step is to ensure the integrity of the data by checking the number of cases, number of records, the observed distribution of each variable, any category labels that may be included on the file, and that certain consistency checks are satisfied. This is accomplished with the File Scan, which also issues warning messages if certain inconsistencies are found.

▷ Once the desired variables are selected, click Scan.

This option scans the data file and identifies each distinct category or value for all variables included in a model. Prior to a data file being scanned, only variable names will be listed in the target list box. After scanning, a message appears in the status bar indicating the number of records in the file, and the number of distinct categories (or values) appears to the right of the variable names in the target list boxes. If a Scan has not been selected before clicking Estimate, a scan will automatically be performed by Latent GOLD prior to beginning the estimation algorithm, and default settings will be used to scale the variables in the model.

Following a Scan, a "Replication Error" warning message appears if a covariate or case weight is not constant for all records corresponding to the same Case ID. The option to "Estimate anyway?" is given at the end of the error message. If Yes is selected, Latent GOLD uses the value on the first record as the desired constant.

### VIEWING CATEGORY LABELS, FREQUENCY COUNTS AND SCORES FOR A VARIABLE

After scanning, double click on a variable name in any of the target list boxes to open the Variables dialog box where category labels, frequency counts and any scores assigned to the variable are displayed. Any scores assigned to Nominal variables are used for descriptive purposes only. Scores associated with ordinal indicators may be changed. For further information about score options for Ordinal variables, see How to Specify Category Scores for an Ordinal Variable.

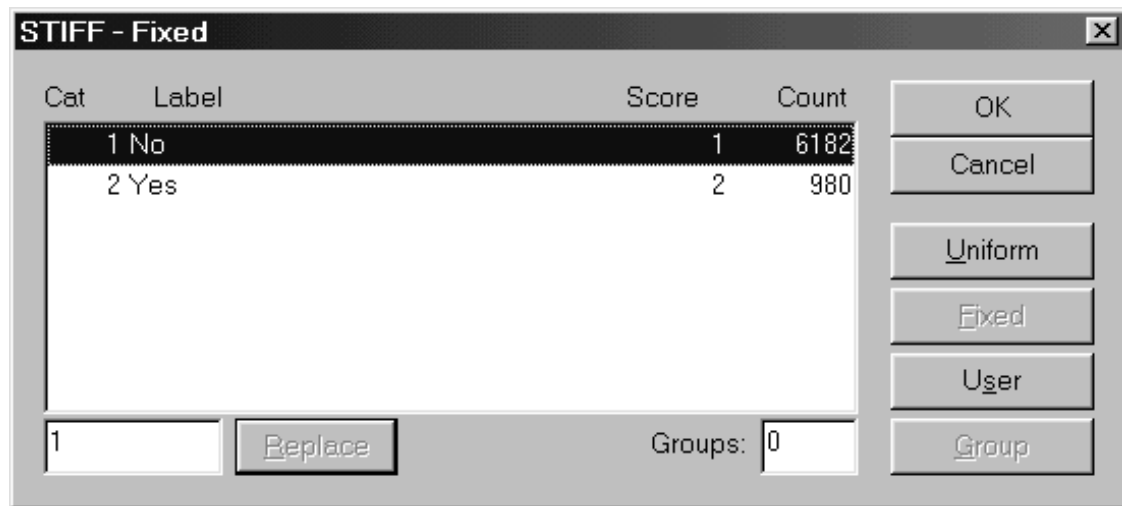


Figure 5-6. Category labels for the variable STIFF

## MISSING VALUES

Latent GOLD recognizes observations as missing on a particular variable if they were assigned a missing value code. In SPSS (for .sav formatted files) missing values may be specified as system missing or user missing. In text formatted files a missing value code is represented by '.'

During the Scan process, all records pertaining to such cases will be deleted. That is, by default, missing values are eliminated using list-wise deletion. After the scan, a message in the status bar indicates the number of records deleted because of missing values.

By default, cases containing missing values on any indicator, dependent variable, predictor or active covariate are excluded from the analysis. The treatment of missing values may be changed to allow cases containing missing values to be included in the analysis for the various options. To include such missing values in an analysis, see Missing Values in Step 8.

## REDUCING THE NUMBER OF CATEGORIES FOR A VARIABLE

After verifying the integrity of your data, you may wish to reduce the size of the data file to speed up estimation, especially if you are estimating a model with many parameters. The Group option can be used to reduce the number of values taken on by a variable by combining adjacent categories (in most cases, you would not want to group a Nominal variable in this manner). The score assigned to a group of adjacent categories is the average of the original (default) scores assigned to the original categories. The grouping algorithm groups adjacent categories that have fewer observations before grouping categories having larger number of observations towards the goal of having approximately equal sized groups.

The grouping used in Latent GOLD is identical to the RANK command used in SPSS with the number of groups equal to the NTILES subcommand.



## How To Group Variable Categories

- ▷ Following a file Scan, double click on a variable name to open the Variables dialog box.
- ▷ Type the number of desired groups into the Groups list box.
- ▷ Click Group.

The variables dialog box changes to show the new groups. The actual number of groups will be less than or equal to the number specified. After grouping, the label field contains the range for the original categories that have been grouped together, and the scores field contains the average of the scores originally assigned to cases in that group. Specifying 0 or a number equal to the number of original categories restores the variable to its original number of categories.

Cat	Label	Score	Count
1	1 - 8	30.866667	15
2	9 - 15	45.833333	12
3	16 - 22	54.588235	17
4	23 - 25	58.833333	12
5	26 - 30	62.785714	14
6	31 - 36	68.285714	14

Score: 30.866667    Replace    Groups: 6

Buttons: OK, Cancel, Uniform, Fixed, User, Group

Figure 5-7. Grouping the variable AGE

The 36-category AGE variable has been merged into 6 groups. New group #1 consists of the original AGE categories 1-8 (abbreviated as "1 1-8"), the 8 youngest ages. The score for this group (displayed in the score field) is 30.87, computed as the average of the original AGE scores assigned to the 15 cases in this group.



## To accept the category grouping and new scores

- ▷ Click OK

The new number of categories will now be listed next to the variable name in the target list along with the letter 'G' (signifying the variable has been grouped).

**Note:** For nominal and ordinal variables, the labels for the grouped categories consist of indices of the original categories. For continuous variables and counts, the label corresponds to the range of values.



## To Restore a Grouped Variable to its Default Setting

- ▷ Highlight the grouped variable name in the target list.
- ▷ Double click the variable name to open the Variables dialog box.
- ▷ Type '0' or the number of original categories in the Groups list box.
- ▷ Click OK.

### **WARNING!**

Use of the Group option will cause any previously set user scores to be ignored. User scores can be assigned to the new categories of a grouped variable to replace the average scores.

## Step 5: Set Scale Types (Variables Tab)

Scale Types determine the distributional assumptions and structural form of the model to be estimated. There are several different scale types (Nominal, Ordinal, Continuous, Count, Binomial Count, Numeric) and subtypes (Truncated, Censored, Standard, Zero-Inflated and Overdispersed) that may be assigned to a variable depending upon the type of variable (indicator, covariate, dependent, predictor). The default for indicators and dependent variables is Ordinal when the variable contains numeric (quantitative) values, and Nominal when the variable contains character codes (string variables). The default for covariates and predictors is Numeric when the variable contains numeric (quantitative) values, and Nominal when the variable contains character codes (string variables).



## To assign or change a scale type:

- ▷ Highlight the variable name(s) in a target list.
- ▷ Right click to open a menu with the selections for that variable type.
- ▷ Select a scale type.

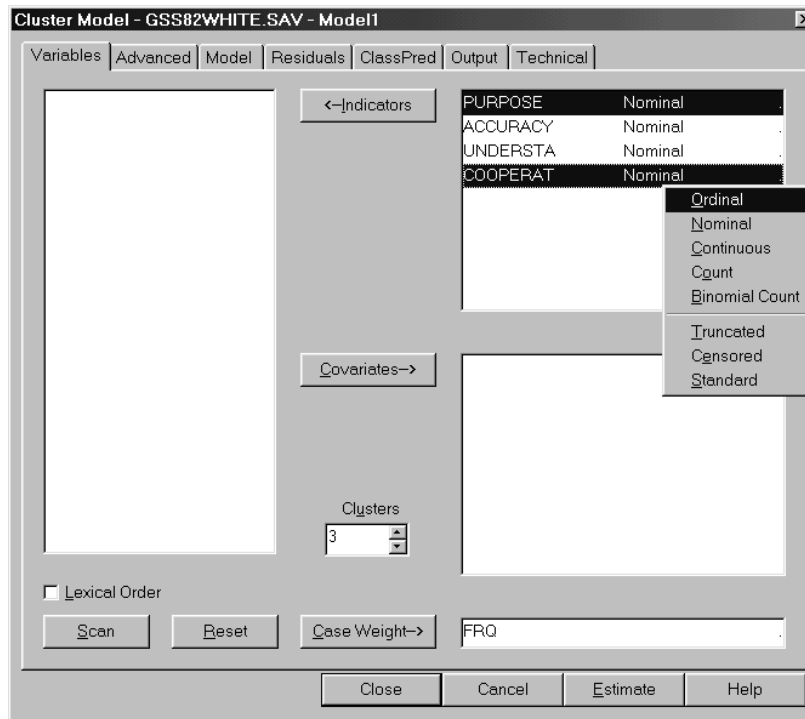


Figure 5-8. Selecting Scale Types

The selected scale type will appear to the right of the variable name. If Ordinal is selected, by default, the scores will be treated as Fixed at the values on the data file and the label "Ord-Fixed" appears to change one or more of the category scores to different values, see How to Specify Category Scores for an Ordinal Variable.

## Scale Types

Each scale type is listed below along with related details.

### ORDINAL (INDICATORS, DEPENDENT)

This setting should be used for categorical variables where the categories are ordered (either from high to low or low to high). If the dependent variable in an LC Regression model is set to Ordinal, the adjacent category logit model, also known as the baseline category logit model (see e.g. Magidson, 1996, 1998), is specified. For an Ordinal variable, the scale type will appear as 'Ord-Fixed', 'Ord-Uniform' or 'Ord-User' depending upon which scores are used (fixed, uniform or user scores):

- 1) **Fixed (default).** Either the original values contained in the data file or array, or, if no scores were specified, uniform scores are assigned to the categories.
- 2) **Uniform.** The variable is assigned fixed scores that are equidistant and have a low score of 0 and a high score of 1. For example, for a 3 category variable, the uniform scores assigned are 0, 0.5, and 1.
- 3) **User Scores.** The user specifies category scores.

## How to Specify Category Scores for an Ordinal Variable

- ▷ After scanning your data file, double click on the variable name to open the Variables dialog box.



### To specify uniform scores,

- ▷ Click Uniform.
- ▷ Click OK to accept the new scores and close the Variables dialog box.
- ▷ The label for the variable in the Variables Tab changes to "Ord-Uniform"



### To specify user scores,

- ▷ Double click on any score you wish to change.
- ▷ Type in a value for the new score.
- ▷ Click Replace.
- ▷ Click OK to accept the new scores and close the Variables dialog box.
- ▷ The label for the variable in the Variables Tab changes to "Ord-User"

If the user inputs scores, and then later resets them by clicking Uniform or Fixed, clicking User will reset scores to their most recent user setting.



### To restore the default setting:

- ▷ Click Fixed.
- ▷ Click OK to accept the new scores and close the Variables dialog box.
- ▷ The label for the variable in the Variables Tab is reset to "Ord-Fixed"

## NOMINAL (INDICATORS, COVARIATES, DEPENDENT, PREDICTORS)

This setting should be used for categorical variables where the categories have no natural ordering. When estimating an LC Regression model, if the dependent variable is set to Nominal, the multinomial logit model is used.

### **NUMERIC (COVARIATES, PREDICTORS)**

This setting should be used for an ordinal or continuous covariate or predictor.

### **ACTIVE (COVARIATES)**

Specifying a covariate as active influences the definition of the latent classes.

### **INACTIVE (COVARIATES)**

Specifying the covariate as **inactive** guarantees that its inclusion in the model has no influence on the model parameter estimates: An inactive covariate is not part of the specified model.



### **ADVANCED: GROUP (COVARIATES)**

If a multilevel model is specified, group level covariates (GCovariates) may also be included in the covariate box. To distinguish covariates at the group level from ordinary covariates, select the Group setting. A <G> appears next to the group level covariate. All group level covariates are **active**.

### **CONTINUOUS (INDICATORS, DEPENDENT)**

This setting should be used when the variable is continuous. When estimating an LC Regression model, if the dependent variable is set to Continuous, the normal linear Regression model is used.

### **COUNT (INDICATORS, DEPENDENT)**

This setting should be used when the variable represents Poisson counts. When estimating a LC Regression model, if the dependent variable is set to Count, the Poisson model is used and you can also specify an additional variable to be used as an exposure (see Exposure)

When estimating a LC Cluster or DFactor model, the exposure is set to 1 for all indicators specified as having the Count scale type.

### **BINOMIAL COUNT (INDICATORS, DEPENDENT)**

This setting should be used when the variable represents binomial counts. When estimating a LC Regression Model, if the dependent variable is set to Binomial Count, the binomial model is used and you can also specify a variable to be used as an exposure (see Exposure). During the scan, the program checks to make sure that the exposure, if specified, is larger than any observed count.

When estimating a LC Cluster or DFactor model, for any indicator(s) specified as having the Binomial Count scale type, the exposure is computed automatically by the program to be the largest count observed for the indicator(s).

**Note:** The exposure can easily be set to a value larger than the largest observed count by means of the following trick: Add one case with xvalues equal to the desired exposure values for all indicators. Include a very small case weight (say 1.0E-50) to the data file for this additional case.

## Scale Subtypes

Four scale subtypes are listed below along with related details.

### **STANDARD (INDICATORS, DEPENDENT)**

By default, this setting is used for indicators and dependent variables. This setting means that the standard scale type is used for this variable as opposed to one of the special subtypes described below.

### **TRUNCATED (COUNT, BINOMIAL COUNT, CONTINUOUS)**

Truncated indicates that only cases with a value larger than 0 are in the sample. It can be applied with an Indicator/Dependent that is a Poisson or Binomial Count, or a Continuous normally distributed variable. The likelihood function is then constructed using truncated variants of the Poisson, binomial, and normal distribution.

### **CENSORED (CONTINUOUS)**

Censored means that all cases with a "true" value smaller than 0 have the same observed value of 0; that is, the variable is assumed to be left censored at 0. Censored can be used with a Continuous Indicator/Dependent, in which case we use assume a left censored normal distribution for the variable concerned.

### **ZERO INFLATED (REGRESSION ONLY)**

Zero Inflated (ZIN) means that one additional latent class with a mean equal to zero is automatically added to the specified model. For Regression models with the dependent variable being a count, these specifications yield the well known Zero Inflated variants of the Poisson and binomial count regression models. In the case of a Continuous dependent variable, it yields a Zero Inflated variant for left censored normal regression, sometimes referred to as censored-inflated regression. In other words, for continuous dependent variables, Zero Inflated also implies Censored.

With Nominal and Ordinal dependent variables, (as well as with Choice, Rating and Ranking specification in Latent GOLD Choice), one ZIN class is added for each category of the Dependent. Each ZIN class responds with probability 1 into a particular category (and with probability 0 into the other categories). These classes are sometimes referred to as stayer classes (in a mover-stayer model) or brand-loyal classes (in a brand-switching model).

### OVERDISPERSED (COUNT AND BINOMIAL COUNT IN REGRESSION)

Overdispersion is a common phenomenon in count data. It means that, as a result of unobserved heterogeneity, the variance of the count variable is larger than estimated by the Poisson (binomial) model. The overdispersed option makes it possible to account for unobserved heterogeneity by assuming that the rates (success probabilities) follow a gamma (beta) distribution. This yields a negative-binomial model for overdispersed Poisson counts and a negative-binomial model for overdispersed binomial counts. Note that this option is conceptually similar to including a normally distributed random intercept in a regression model for a count variable.

The overdispersion option is useful if one wishes to analyze count data using mixture or zero-inflated variants of (truncated) negative-binomial or beta-binomial models (Agresti, 2000; Long, 1997; Simonoff, 2003). The negative-binomial model is a Poisson model with an extra error term coming from a gamma distribution. The beta-binomial model is a variant of the binomial count model that assumes that the success probabilities come from a beta distribution. These models are common in fields such as criminology, political sciences, medicine, biology, and marketing.

### Step 6: Specify the Number of Latent Classes (Variables Tab)

Use the Variables Tab to specify the number of classes for your model. In the Cluster Module, a class is called a Cluster. In the DFactor Module, the ordered latent classes associated with each DFactor are called 'levels'. A model may contain several DFactors, each of which may have a different number of levels.

### SPECIFYING THE NUMBER OF CLUSTERS FOR LC CLUSTER MODELS

The box labeled Clusters is located beneath the Indicators button in the Variables Tab.

### INDICATORS (CLUSTER AND DFACTOR MODELS)

Indicators are dependent variables that are used to define or measure the latent classes in a LC Cluster model, or latent variable(s) in a DFactor model. Indicators may be treated as Nominal, Ordinal, Continuous, Poisson Count, or Binomial Count (see Step 5: Set Scale Types).

Select one or more variables from the variables list box to be used as Indicators (required).

- ▷ Enter a number greater than 0. You may also specify a range to estimate several models. For example, enter "1-4" to estimate four different latent class models containing 1, 2, 3, and 4 clusters respectively.

## SPECIFYING THE NUMBER OF DFACORS FOR DFACTOR MODELS

The box labeled DFactors is located below the Covariates button in Figure 5-4. Enter a number greater than 0. (The range option is not available when specifying the number of DFactors.)

By default, each DFactor consists of 2 levels. To increase the number of levels for one or more DFactors, use the Model Tab and set the number of levels desired for each factor (See Step 7: Set Restrictions and Other Model Options). The maximum number of levels that can be specified for a DFactor is 20.

## SPECIFYING THE NUMBER OF CLASSES FOR LC REGRESSION MODELS

The box labeled Classes is located beneath the Covariates button in Figure 5-5. Enter a number greater than 0. Separate regression models will be estimated for each segment. If a range is specified for the number of classes, such as '1-4', separate sets of regression models will be estimated, the first representing a 1-class regression model (the traditional regression which assumes a single homogeneous population). For the other models, separate regressions are estimated for 2, 3 and 4 classes (segments) respectively.



### ADVANCED. SPECIFY THE NUMBER OF CFACORS, GCLASSES, GCFACORS

An advanced option allows for the inclusion of up to 3 continuous factors (CFactors) in a model. In addition, a multilevel extension of a model may be estimated which involves specifying a group-level ID variable. In the case of multilevel models, 2 or more group-level latent classes (GClasses) and up to 3 group-level CFactors (GCFactors) may be specified. For further details, see Step 7: Advanced Tab.

## Step 7: Set Restrictions and Other Model Options

Following Step 6, you may choose to estimate your model (Step 10), view the results and then impose (post-hoc) restrictions on the parameters to achieve a more parsimonious and interpretable model. Alternatively, you may choose to impose a priori restrictions and/or relax certain restrictions that are imposed by default. The most common restrictions are:

- zero restrictions:** restrict to zero any parameter estimate that is not statistically significant
- class independence restrictions:** restrict certain parameter estimates to be identical for each latent class
- offset:** specify certain parameters to be equal to one (Regression)
- equal effects:** equate parameters across indicators (Cluster and DFactor)

The additional effects that can be included are:

- associations between DFactors (DFactor model)
- CFactor1 affecting the predictor coefficients (Regression Advanced),
- GClasses and GCFactor effects on the indicators (Cluster/DFactor Advanced)
- GClasses and GCFactors affecting the intercept and the predictor coefficients (Regression)

Advanced),

- GClasses and GCFactors affecting the covariate effects in the regression model for the Clusters, Classes, or DFactors (Advanced)

These are a few of the kinds of restrictions that may be applied using the Model Tab. Some additional kinds of restrictions and other advanced options are available using other tabs:

**ClassPred Tab:** may be used to assign with certainty one or more designated cases to belong to a particular latent class or latent classes. This Known Class option amounts to indicating to which classes cases or subsets of cases may not belong.

**Residuals Tab** (Cluster and DFactor models only): may be used to include direct relationships between indicators (associations/covariances) and direct effects of covariates on indicators in a model.



**Advanced Tab** (available in the Advanced version of Latent GOLD): may be used to

- add group- level continuous latent variables (GCFactors) and/or a group level nominal latent variable (GClasses) in a multilevel extension of a model, as well as indicate which coefficients in the logistic regression model for the Clusters, DFactors, and Classes are random effects (differ across GClasses or are affected by GCFactors)
- add continuous factors (CFactors) to specify factor analytic, item response theory (IRT), or random-effects regression models for 2-level data
- set options to incorporate the sampling design if it deviates from simple random sampling.

This section details the use of each of these options.

### Model Tab: Applying (or Relaxing) Parameter Restrictions

The appearance of the Model Tab and the available options differ depending upon the type of model selected.

#### LC CLUSTER MODEL

The LC Cluster Model Tab allows you to change the number of clusters, and to restrict the following parameters in various ways:

- Beta effects of the latent variable on the selected indicators
- Gamma effects of selected (active) covariates on the latent variable
- Error variance parameters (for continuous variables only)
- Error covariance parameters (for continuous variables only)
- Advanced: Lambda effects of selected CFactors and GCFactors on selected indicators; beta effects of the GClasses on selected indicators; gamma effects of selected group covariates on the GClasses

The Name List Box appears in the left-most portion of the Model Tab. It contains the name 'Clusters' at the top, followed by the specific cluster names ('Cluster1', 'Cluster2', ...) for each cluster specified under Step 6. If the range option was used to specify a range of cluster models to be estimated, no specific cluster names will be listed in this box.



At the bottom of this list will be specific names for each CFactor specified ('CFactor1', 'CFactor2', 'CFactor3') if any CFactor effects have been specified in the Advanced Tab. For multilevel models, the list also contains GClasses and/or GCFactors.

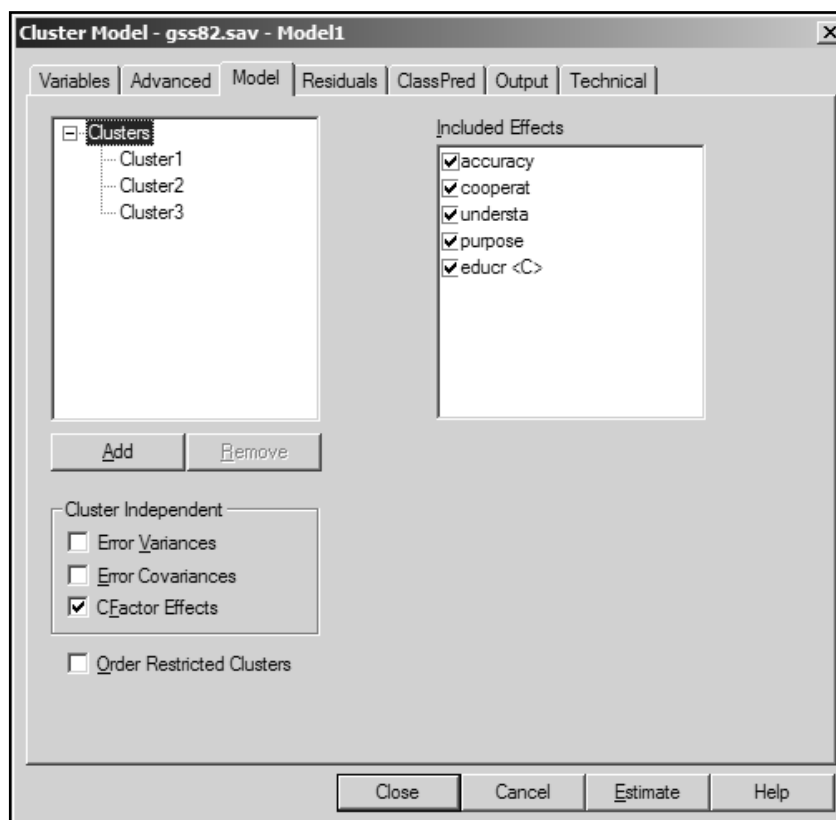


Figure 5-9. Model Tab for LC Cluster Model

## CHANGING THE NUMBER OF CLUSTERS

The Add and Remove buttons can be used as an alternative to the clusters box in the Variables Tab to set the number of clusters. To change the number of clusters (from that specified in Step 6), you can increase or decrease the number of clusters by selecting 'Clusters' (or one of the cluster names) and clicking Add to increase the number of clusters. You can also select one of the cluster names and click the Remove button to decrease the number of clusters.

**Note:** The Add and Remove buttons will not be active if the Range option was used in Step 6 to specify a range of clusters.

### IMPOSING ZERO RESTRICTIONS

The Included Effects box contains all indicators and active covariates that are included in the current model. Variable names for the active covariates are followed by the characters '<C>'. By default, each of these variable names is preceded by a check (check box equals on) to indicate that a set of beta effects will be estimated for the associated indicator or that a set of gamma effects will be estimated for the associated covariate. To restrict any of these sets of effects to zero, click in the check box and the check is removed (check box equals off).



If 1 or more CFactors is included in a model, names for these CFactors ('CFactor1', 'CFactor2', 'CFactor3') appear beneath the cluster names. By default, for each CFactor, lambda effects will be estimated for all indicators included in the model. For a selected CFactor, click on the name of that CFactor and the Included Effects box shows the lambda effects to be estimated for that CFactor. To restrict any of these lambda effect(s) to zero, click in the check-box of the indicator(s) and the check is removed.

**Note:** Since it is not possible to specify regression models for CFactors, the check-box for any Covariates is inactive for CFactor effects.



If GClasses or one or more CFactors is included in a model, names for these appear beneath the cluster names. By default, for GClasses and each GCFactor, beta(g) and lambda(g) effects will be excluded for all indicators included. For GClasses or a selected CFactor, click on its name and the Included Effects box shows the effects to be estimated for that latent variable. To add any of these beta/lambda effect(s), click in the check-box of the indicator(s) and the check is added.

**Note:** Since it is not possible to specify regression models for GCFactors, the check-box for any Covariates is inactive for GCFactor effects. Since Group Covariates can be used as predictors in the regression model for the GClasses, the check boxes for the corresponding gamma(g) effects are active and by default on (effect included).

#### For cluster effects:

Make sure that the name 'Clusters' (or one of the cluster names) is highlighted in the left-most box

- Removing a checkmark for selected indicators, restricts a set of beta effects to zero.
- Removing a checkmark for selected active covariates, restricts a set of gamma effects to zero.

**Note:** The set of beta (gamma) effects restricted to zero are associated with all clusters. Zero restrictions can not be used to restrict betas or gammas to zero for certain classes while allowing them to be estimated for other clusters.



#### For CFactor effects:

Make sure that the name of the CFactor (e.g., 'CFactor1') is highlighted in the left-most box

- Removing a checkmark for selected indicators restricts lambda effects to zero.

- The check-box is inactive for covariates.
- By default, effects for all indicators and active covariates are unrestricted..
- Variables with effects restricted to zero are still included in the calculation of the overall model statistics such as  $L^2$ .
- Restriction of the set of gamma effects to zero for a covariate causes that covariate to be inactive in the measurement of the latent variable. Setting to zero the effects of **all** covariates causes all covariates to be inactive in which case the estimates for the beta parameters will be identical to the estimates obtained if the covariates were excluded from the model, although the overall model  $L^2$  and related statistics will differ. (Although all covariate effects are set to zero, covariates specified as active are still used to form the overall multiway Table and hence affect the computation of the  $L^2$  statistic. On the other hand, covariates with scale type 'Inactive' affect neither the parameter estimates nor the statistics such as  $L^2$ . The choice as to whether to treat covariates as active (the default) or inactive is a matter of user preference).



## For GClasses and GCFactor effects:

Make sure that the name 'GClasses' or of the GCFactor concerned (e.g., 'GCFactor1') is highlighted in the left-most box

- Adding a checkmark for selected indicators removes the default zero restriction on the beta/lambda effects
- Removing a checkmark for selected group covariates, restricts a set of gamma effects to zero (GClasses only).

## IMPOSING CLASS INDEPENDENT RESTRICTIONS

Error variance and error covariance parameters are estimated for continuous indicators only, the latter occurring only when direct effects between 2 or more continuous indicators are included in a model. Following model estimation, these parameter estimates (sigmas) are shown in the Parameters Output. In addition, error correlation parameters may be viewed in the 'Error Correlation' subcategory of the Parameters Output. By default, separate estimates for these parameters are estimated for each cluster.

The 'Cluster Independent' box can be used to impose the following class independence restrictions:

### Cluster Independent Error Variances (applies to all continuous indicators)

Selection of this option indicates that the error variances are restricted to be equal across classes (class independent) for all continuous indicators.

- ▷ To select this option, click in the check-box preceding 'Error Variances' (cluster independent check on).

**Note:** This option is only available for Cluster and Regression models (DFactor model variances are always class independent).

**Cluster Independent Error Covariances** (applies to pairs of continuous indicators for which direct effects have been included in the model - see Residuals Tab below)

This option indicates that the error covariances are restricted to be equal across classes (class independent).

- ▷ To select this option, click in the check-box preceding 'Error Covariances' (cluster independent check on).

**Note:** This option is only available for Cluster models.



### **Cluster Independent CFactor effects (default).**

By default, any CFactors that have been specified in the Advanced Tab are restricted to be the same across clusters (cluster independent check on). To allow separate CFactor effects for each cluster click to remove the check-mark (cluster independent check off).

## **ESTIMATING ORDER-RESTRICTED LATENT CLASSES**

By default, clusters are not order-restricted. Click this to indicate that the beta parameter effects and cluster-specific means of the indicators (reported in the Profile Output) should be restricted to be monotonically increasing. With nominal indicators this means that the local odds are assumed to be increasing. This option yields what is called an order-restricted LC model.

## **EQUAL EFFECTS ACROSS INDICATORS**

Make sure that the name 'Clusters' (or one of the cluster names) is highlighted in the left-most box. A right-click activates a popup menu that can be used to set the "Equal Effects" option. This option causes all included beta effects to be equal across indicators of the same scale type (nominal indicators should also have equal numbers of categories).



Equal effects for CFactors, GClasses, and GCFactors work in the same way as for Clusters.

DFACTOR MODEL

The DFactor Model Tab allows you to change the number of discrete factors (DFactors) and/or DFactor levels, and/or to restrict selected DFactor effects to zero.

The Factors Name List Box appears in the left-most portion of the Model Tab. It contains names for each DFactor specified in Step 6 ('DFactor1', 'DFactor2', etc.), followed by a name for each level of that DFactor. By default, each DFactor has 2 levels ('Level1', 'Level2').

**A** If any CFactor effects have been specified in the Advanced Tab, specific names for each continuous factor ('CFactor1', 'CFactor2', 'CFactor3') appears at the bottom of this list. In multilevel models, the list also contains GClasses and/or GCFactors.

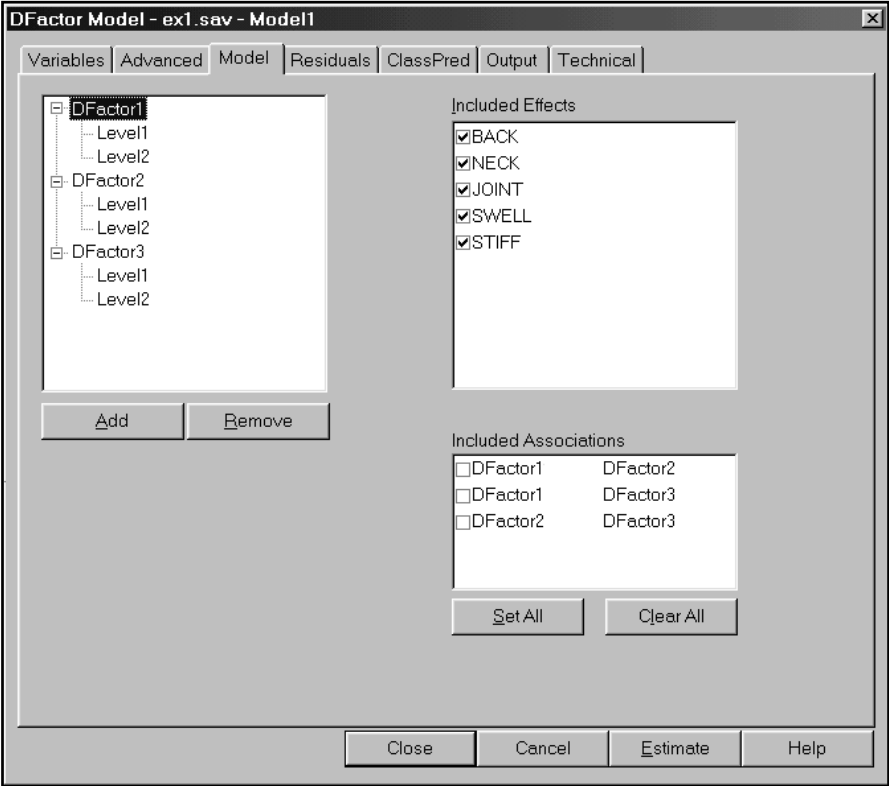


Figure 5-10. Model Tab for DFactor Model

CHANGING THE NUMBER OF DFACTORS OR DFACTOR LEVELS

To increase or decrease the number of DFactors, highlight a DFactor name (e.g., 'DFactor1') in the DFactors Name List Box and click Add to increase or Remove to decrease the number of DFactors. This option works the same as the DFactors box in the Variables Tab.

To change the number of levels of a DFactor, highlight a level name (e.g., 'Level1') and click Add to increase or Remove to decrease the number of levels. Each DFactor can be set to have any number between 2 and 20 levels.

After a DFactor model has been estimated, the number of levels for each DFactor is displayed in the first row of the model summary output file. Note: For each DFactor, the number of levels can be between 2 and 20.

### Editing DFactor Names



To customize the DFactor name for your model:

- ▷ Single click the DFactor name (for example, 'DFactor1') in the DFactors Name List Box to highlight it.
- ▷ Single click it again to edit the name (the name will be in an edit box).

Once you have an edit box open, you may also right click to open a menu that allows you to Cut, Copy, Paste, and Delete the DFactor name. This menu also allows you to Undo your last edit or select the whole name for editing (Select All).

### INCLUDED EFFECTS BOX

The Included Effects box contains all indicators and active covariates that are included in the current model. Variable names for the active covariates are followed by the characters '<C>'. By default, each of these variable names is preceded by a check (check box equals on) to indicate that for the selected DFactor, beta effects will be estimated for the associated indicator(s) and gamma effects will be estimated for the associated covariate(s). To restrict any of these effects to zero for a DFactor, select the DFactor from the Name List Box and click in the desired check box in the Included Effects box. The check is removed (check box equals off) to indicate that the associated effect is set to zero.



If 1 or more CFactors is included in a model, names for these CFactors ('CFactor1', 'CFactor2', 'CFactor3') appear beneath the DFactor names and DFactor level names. By default, for each CFactor, lambda effects will be estimated for all indicators included in the model. For a selected CFactor, click on the name of that CFactor and the Included Effects box shows the lambda effects to be estimated for that CFactor. To restrict any of these lambda effect(s) to zero, click in the check-box of the indicator(s) and the check is removed. Note: Since CFactor (lambda) effects are estimated only for indicators, the check-box for any Covariates is inactive for CFactor effects.



If GClasses or 1 or more CFactors is included in a model, names for these appear beneath the cluster names. By default, for GClasses and each GCFactor, beta(g) and lambda(g) effects will be excluded for all indicators included. For GClasses or a selected CFactor, click on its name and the Included Effects box shows the effects to be estimated for that latent variable. To add any of these beta/lambda effect(s), click in the check-box of the indicator(s) and the check is added. Note: Since it is not possible to specify regression models for GCFactors, the check-box for any Covariates is inactive for GCFactor effects. Since Group Covariates can be used as predictors in the regression model for the GClasses, the check boxes for the corresponding gamma(g) effects are active and by default on (effect included).

## For DFactor effects:

When 2 or more DFactors are included in a model, make sure that the name of the DFactor (e.g., 'DFactor1') is highlighted in the DFactor Name List Box

- Removing a checkmark for selected indicators, restricts the associated beta effects to zero.
- Removing a checkmark for selected active covariates, restricts the associated gamma effects to zero.
- By default, effects for all indicators and active covariates are unrestricted.
- Variables with effects restricted to zero are still included in the calculation of the overall model statistics such as  $L^2$ .
- Restriction of the set of gamma effects to zero for a covariate causes that covariate to be inactive in the measurement of the latent variable. Setting to zero the effects of all covariates causes all covariates to be inactive in which case the estimates for the beta parameters will be identical to the estimates obtained if the covariates were excluded from the model, although the overall model  $L^2$  and related statistics will differ. (Although all covariate effects are set to zero, covariates specified as active are still used to form the overall multiway table and hence affect the computation of the  $L^2$  statistic. On the other hand, covariates with scale type 'Inactive' affect neither the parameter estimates nor the statistics such as  $L^2$ . The choice as to whether to treat covariates as active (the default) or inactive is a matter of user preference).



## For CFactor effects:

When 2 or more CFactors are included in a model, make sure that the name of the appropriate CFactor (e.g., 'CFactor1') is highlighted in the left-most box

- Removing a checkmark for selected indicators restricts the corresponding lambda effects to zero.
- The check-box is inactive for covariates.
- By default, effects for all indicators are unrestricted for all CFactors.



## For GClasses and GCFactor effects:

Make sure that the name 'GClasses' or of the GCFactor (e.g., 'GCFactor1') is highlighted in the left-most box

- Adding a checkmark for selected indicators removes the default zero restriction on the beta/lambda effects
- Removing a checkmark for selected group covariates, restricts a set of gamma effects to zero (GClasses only).

### INCLUDED ASSOCIATIONS BOX

Included Associations refer to the associations between DFactors. By default, the DFactors are assumed to be uncorrelated (DFactor Association Check-box equals off). To include one or more factor association parameters in the model, click in the associated DFactor Association Check-box.

### EQUAL EFFECTS ACROSS INDICATORS

When 2 or more DFactors are included in a model, make sure that the name of the DFactor (e.g., 'DFactor1') is highlighted in the DFactor Name List Box. A right-click activates a popup menu that can be used to set the "Equal Effects" option. This option causes all included beta effects to be equal across indicators of the same scale type (nominal indicators should also have equal numbers of categories).



Equal effects for CFactors, GClasses, and GCFactors works in the same as for DFactors.

### LC REGRESSION MODEL

Various restrictions are available for intercepts and predictor effects. In addition, for models with continuous dependent variables, restrictions are available for error variances. The various restrictions include class independent restrictions, order restrictions, zero restrictions, fixed-value (offset) restrictions, and equality restrictions. By default, no restrictions are imposed.

Restrictions can be placed prior to estimating an initial model to incorporate prior knowledge. For example, the zero constraints make it possible to specify a different regression model -- with different predictors -- for each latent class based on a priori knowledge about the classes. A specific application for this is a model with a random responders class for which all predictor effects are zero. An application of the equality constraints is the possibility of defining a DFactor-like structure in which, for example, one DFactor influences the intercept and another the predictor effects. Ordering constraints are important if one has a priori knowledge on the direction of an effect. For example, the price effect on a product rating is usually assumed to be negative (or non-positive) in each latent class (segment). The estimates obtained with Regression will be constrained to be in agreement with this assumption if the price effect is specified to be **Descending**.

Alternatively, restrictions may be employed post-hoc to estimate a new model after viewing the results from an estimated model.

For Regression, when **K** classes have been specified in Step 6, the Model Tab contains **K** individual columns for each of the **K** latent classes (1,2, ...,K). Additional columns are labeled **Class Independent** and **Order Restriction**. When the range option is used in Step 6 to specify the estimation of models with different numbers of latent classes, class-specific columns are absent from the Model Tab and only the class independent and order restrictions can be applied. Such restrictions are applied to each of the models generated by the range option.

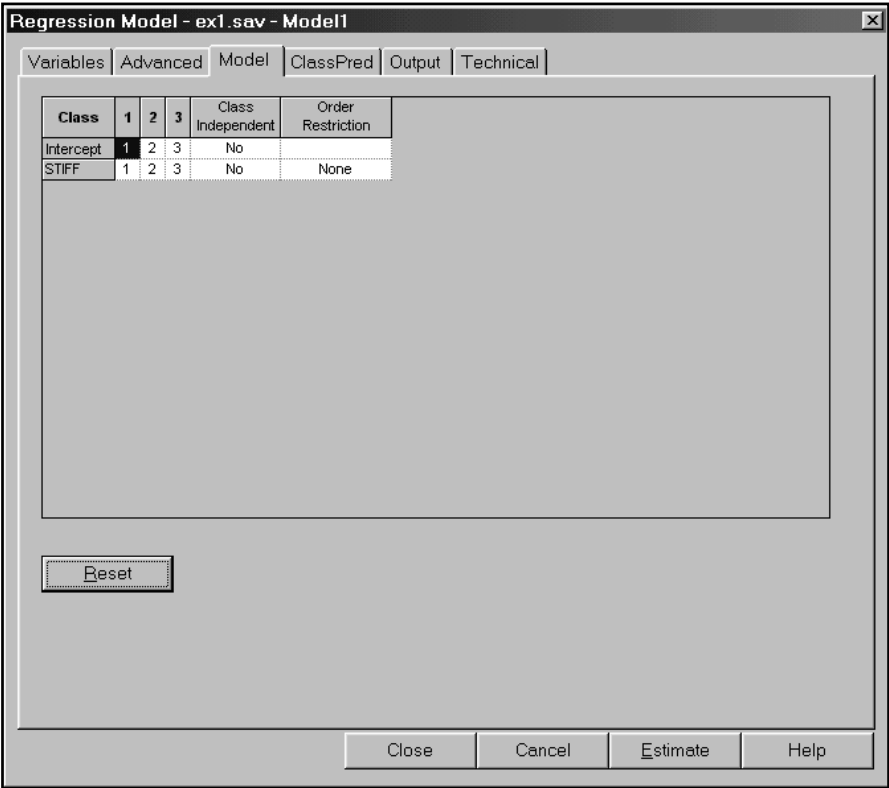



Figure 5-11. Model Tab for LC Regression Model.

Selected intercept and predictor effects can be set equal to zero for certain classes (No Effect), equated across two or more classes (Merge Effect), or equated across all classes (Class Independent). The effects of a predictor can be restricted to be ordered in each class (Ascending or Descending). Error variances can be equated between selected classes (Merge Effect) or between all classes (Class Independent). The effect of a numeric predictor on a dependent variable that has a scale type other than nominal can be fixed to 1 in one or more classes (Offset).

The first row consists of the label 'Intercept', followed by separate row for each predictor. Restrictions can be set separately for each row. By default, no restrictions are imposed. This is indicated by the unique integers (1,2, ...,K) that appear in each row, by the label 'No' indicating that the class independence restriction has not been imposed and the label 'None' indicating that no order restrictions have been selected for any predictors.

**SPECIFYING EQUALITY RESTRICTIONS ACROSS CLASSES**

The effects for a predictor may be specified as Class Independent (Fixed Effects) or Class Dependent (Random Effects, default). When the Class Independent restriction is applied, the regression coefficient for that predictor is restricted to be equal between each of the K latent classes (segments). When Class Independent is selected for a predictor, an '=' appears to the right of the predictor name in the Variables Tab. For a 1-class regression model (K=1), selection of this option has no effect.

 To select this option

- ▷ Right click on the desired cell(s) in the column labeled 'Class Independent' to retrieve the popup menu.

## CHAPTER 5. BASIC STEPS FOR MODEL DEVELOPMENT

- ▷ Select 'Yes'. The 'No' changes to 'Yes' in the selected cells, and the indices in the selected rows all change to '1' to indicate that the effects for classes 2, 3, etc. are all restricted to be equal to the corresponding effects for class 1.

Alternatively, the restriction of class independence can be imposed as follows:

- ▷ Select all the cells containing the indices for a chosen row
- ▷ Right click to bring up the pop-up menu
- ▷ Select 'Merge Effect'.
- ▷ To undo these restrictions, reselect the cells, right click and select 'Separate Effect'.

This alternative way of imposing the *class independence* restriction can also be used to specify equal effects between some but not all the classes. Simply select those class indices for which the across-class equality restriction is desired and select 'Merge Effect'. The indices selected will change to be equal. 'Merge Effect' can be used more than once for a given row, so that it is possible to specify that certain effects for classes 1 and 2 be restricted to be equal and that the effects for classes 3 and 4 be equal. After imposing these restrictions, the associated indices will appear as '1 1 3 3'.

**Note:** the numbers indicate to which class the effect for the class concerned is equated.

The regression intercept may also be restricted to be class independent. To set this option, right click on the row labeled 'Intercept' in the column labeled 'Class Independent' to retrieve the popup menu and select 'Yes'.

**Note:** For dependent variables having scale type 'Ordinal', a third option 'No - Simple' is also available. Rather than *complete* class independence ('Yes'), a 'simple' implementation of class independence analogous to a class independent intercept specification in the linear regression model (continuous scale type) is used where the (partial) dependent variable means are taken to be class independent rather than the entire marginal dependent variable distribution. For further details of this 'class-independent simple' specification for the intercept in ordinal regression models, see section 5.3 of Technical Guide

### SPECIFYING ZERO RESTRICTIONS

To restrict one or more effects to zero, select the desired effects, right click, and select 'No Effect'. This causes a '-' to appear in the selected cells which indicates that the effect(s) associated with these cells is now restricted to zero. These menu options can also be used in combination with each other to produce a desired result. For example, first selecting *class independence*, followed by selecting 'Separate Effect' in one of these cells causes the index for the selected cell to return to its default setting. However, the remaining cells in that row remain restricted to be equal and 'Yes' automatically changes to 'No' in the *Class Independent* column.

### SPECIFYING ORDER RESTRICTIONS

Order Restriction can be used to indicate that the regression coefficient is monotonically increasing (Ascending) or decreasing (Descending).

To impose an *ordering* restriction, right click on one or more cells containing a 'None' label and right click to retrieve the relevant pop-up menu. Select either *Ascending* or *Descending* to impose the desired ordering restriction on the selected predictor effect(s). This restriction is imposed across all classes. With a dependent variable of a scale type other than 'nominal', the usage and interpretation of this restriction is straightforward. In the case of a nominal dependent variable, the difference between parameters of adjacent categories of the dependent variable will be in agreement with the specified order restrictions (see section 5.3 of Technical Guide).

## SPECIFYING FIXED-VALUE RESTRICTIONS

With the *Offset* option one can indicate that the effect of a numeric predictor should be fixed to 1 in selected classes. The option is not available for Nominal dependent variables.

**Note:** It is possible to use the offset restriction to fix the effect to non-zero quantities other than 1 as well. For example, to restrict the effect of a numeric predictor to equal some desired constant  $c$ , you would first rescale that predictor by multiplying it by  $c$ . You would then use the rescaled version of the predictor in the model.

A specific application occurs in the case in which one would like to fix the response probability to 0 for certain predictor values for certain classes. Suppose that the dependent variable is 1='buy', 0='no buy' and that you suspected that one latent class existed that always responded 'no buy' when the predictor variables reflected a certain pattern (e.g., PRICE \$50). This option could be used to identify this latent class by creating a dummy predictor variable, say coded [-100, 0] where the value -100 refers to a situation which is suspected to always result in a 'no buy' for some class (fixing a logit coefficient to -100 amounts to fixing the probability of a 1='buy' response to 0).



## INCLUDING CFACOR, GCLASSES, AND GCFACOR EFFECTS IN THE MODEL FOR THE DEPENDENT VARIABLE

When included, CFactor, GClasses, and GCFactors appear as additional columns in the Model Tab.

Check box switches can be added or removed to indicate whether CFactors, GClasses, and/or GCFactors affect the intercept and/or the predictor coefficients,

By default:

- CFactor1: intercept checked and other terms unchecked.
- CFactor2 and CFactor3: all effects checked.
- GClasses and GCFactors: all effects unchecked.

**Note:** For all included effects the default is "Class Independent = Yes". This can be changed using the "Class Independent =No", "Merge Effects", and "No Effect" options.

Applications:

- random intercept regression model (1 CFactor, default specification)
- 3-level LC regression (GClasses)
- 3-level random-effects modeling (GCFactors combined with CFactors).

## ClassPred Tab: Restricting Cases Known (Not) to Belong to a Certain Class or Classes

With this option one can specify that one or more specific cases can belong to a certain class or certain classes only. To use this feature, select a variable from the list box in the ClassPred Tab to be used as the Known Class Indicator and click Known Class. The variable moves to the Known Class Indicator Box and the Assignment Table becomes active. For each category of the Known Class indicator you then specify to which classes the cases with that category code may belong (*or not belong*) using the Assignment Table. For example, Figure 5-12 illustrates a 4-Cluster model (4 columns) where the variable 'classind' is used as the Known Class Indicator. Cases for which 'classind=1' are allowed to be in cluster 1 only; those for which 'classind=2' are allowed to be in cluster 2 only; all other cases ('classind=3') may be assigned to any of the 4 clusters.

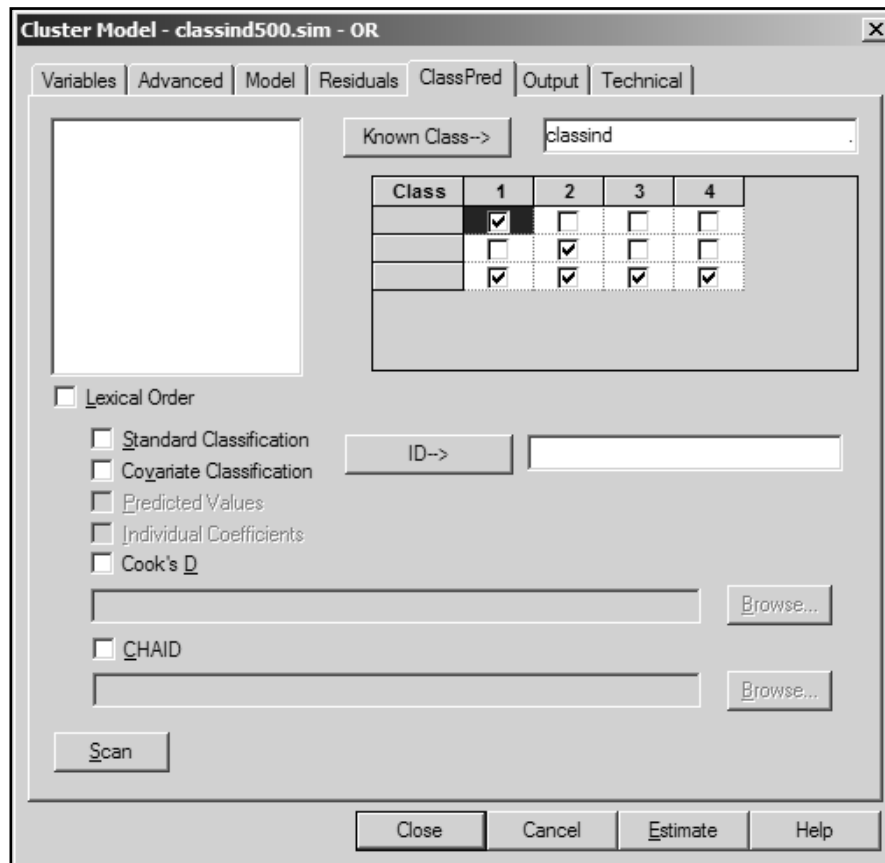


Figure 5-12. ClassPred Tab for Cluster Model with a Known Class Indicator

This option is useful if you have a priori class membership information for some cases (pre-assigned or pre-classified cases) or if membership to certain classes is very implausible for some combinations of observed scores.

### KNOWN CLASS - CLASS INDICATOR

In applications where a subset of the cases are known with certainty not to belong to a particular class, or particular classes, you can take advantage of this information to restrict their posterior membership probability to 0 for one or more classes and hence classify these cases into one of the remaining class(es) with a total probability = 1. This feature allows more control over the segment definitions to ensure that the resulting classes

are most meaningful. Common applications include:

- 1) using new data to refine old segmentation models while maintaining the segment classifications of the original sample
- 2) archetypal analysis - define class membership a priori based on extreme response patterns that reflect theoretical "archetypes"
- 3) partial classification -- high cost (or other factors) may preclude all but a small sample of cases from being classified with certainty. These cases can be assigned to their respective classes with 100% certainty, and the remaining would be classified by the LC model in the usual way
- 4) certain cases may be known to be "type 1 OR type 2" (e.g., 'clinically depressed' or 'troubled'). By excluding such cases from being in say class 3 = 'healthy', such cases can be pre-assigned to be in class 1 or 2, while additional cases may be freely classified into any class
- 5) post-hoc refinement of class assignment where modal assignment for certain cases is judged to be implausible based on the desired interpretation of the classes.

In addition, this option may also be used to specify multiple group models by including the group variable as both a Known Class Indicator and as an active covariate. For further details of this, see section 2.5 of Technical Guide.

**Note:** The Known Class option is not available in cluster and regression if the Range option has been used in the Variables Tab.

For DFactor models, this option applies only to levels of DFactor 1.



**To select known classes (clusters/classes/DFactor1 levels):**

- ▷ Select one variable from those appearing in Variables List Box (located in the upper left-hand portion of the ClassPred Tab). Variables appearing here are those that have not been previously selected as Indicators or Covariates.
- ▷ Click Known Class to move that variable to the Known Class Box and the class assignment window beneath the Known Class Box becomes active.
- ▷ A separate row appears for each category/code/value taken on by the known class indicator  
A separate column appears for each class.
- ▷ Click on the appropriate boxes to select or deselect the possible assignment of the categories to certain classes.

A checkmark off means that the posterior membership probability is restricted to zero for that class for cases in that category of the known class indicator.

By default, the checks are assigned as follows:

For a K-class model, a category with a code of K on the Known Class Indicator is assigned to only class

K. Categories coded less than 1, greater than K or missing are assigned to all classes (i.e., no restrictions). Missing values are not shown in the table.

**Note:** For the example in Figure 5-12 above, all cases are coded either 1, 2 or 3 on the variable 'classind' (i.e., no missing values). Those coded 'classind=1' and 'classind=2' are maintained at their default specifications on the table, while the default specification for cases coded 'classind=3' was changed (from 'cluster 3 only' to 'any cluster' -- all 4 cluster columns checked). This specification would be obtained by default if those coded '3' on the classind variable were instead coded as 'missing'. In this situation, the table would differ from that shown in Figure 5-12, in that the 3rd row of the table would not appear, since that category would be coded 'missing'.

For further information, see section 2.5 of the Technical Guide.

## Residuals Tab: Including Direct Effects in Model (**CLUSTER AND DFACTOR MODELS**)

From the Residuals Tab you may specify a direct effect to include in a model. All pairs of variables eligible for a direct effect parameter appear. To include a direct effect, click in the check-box and a check appears. Direct effect parameters will be estimated for the pairs of variables that have been so selected (direct effect check-box equals on). The inclusion of direct effects is one way to relax the assumption of local independence.

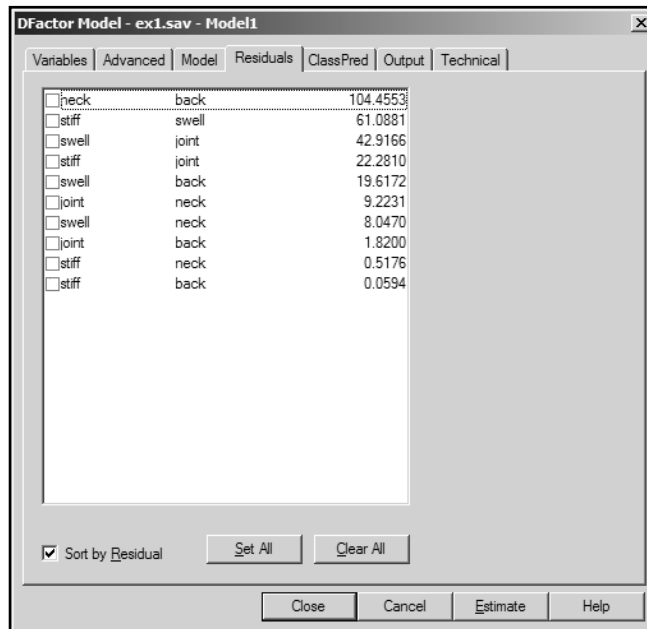


Figure 5-13. Residuals Tab for DFactor Model

Direct effects may be specified prior to estimating a model or post-hoc after examining the results of a model. To assist you in choosing pairs of variables for which direct effects may improve the fit of a model, bivariate residuals (BVR) may be output.

Bivariate residuals are listed in the Bivariate Residuals section of the Outline pane when this option has been selected in the Output Tab. They are useful for diagnostic purposes. Advanced users may wish to include one or more Direct Effects Parameters that are associated with large bivariate residuals in a model.

A direct effect between two indicators and direct effects of selected covariates on selected indicators can be

selected for inclusion in a model using the Residuals Tab on the Analysis dialog box. After estimation, these effects will be listed in the Parameters output in a section labeled 'Direct Effect Parameters'.

**Note:** If this is your first model for a data file, once you have specified the model in the Variables Tab, all pairs of indicators and covariates will be listed without any bivariate residuals in the Residuals Tab.

After you have estimated a model, the Residuals Tab will also contain a bivariate residual associated with each direct effect. Reopen the Analysis dialog box and now the Residuals Tab contains the residuals based on the model sorted from high to low according to the magnitude of the residual.

The example in Figure 5-13 shows that the direct effect between the variable pair NECK and BACK is associated with the largest residual is. This residual value (chi-square divided by its degrees of freedom) is 104.4553, which is much larger than 1, the reference value for these residuals.

In situations where there is only one large diagnostic value, a new model can be estimated by selecting the variable pair (a pair is selected when a checkmark appears in the box to the left), and clicking Estimate, thus adding the associated direct effect parameter to the current model.

**Note:** If there are several large residuals (as in Figure 5-13), a common strategy is to include the corresponding direct effects one at a time as needed, each time re-estimating the model, and checking the updated residuals after each new model is estimated before including additional direct effects. This is because once you have included a direct effect in a model, all of the residuals in that new model may be small.

After estimating, the bivariate residuals will be listed as 0 (or very close to 0) for any variable pair for which a parameter has been included in the model. Also note that if you estimate a new model that included residuals, Latent GOLD will retain all of your settings when estimating additional models, including the residuals selected for inclusion in the prior model.

## Check boxes:

**Sort by Residual.** By default, bivariate residuals are sorted from largest to smallest (sort residual check box equals on). If this option is not checked, the residuals will be listed in default order which is based on the order the variables were entered into the model in the Variables Tab.

**Set all.** Click this to include direct effects for all eligible pairs of variables.



## Advanced Tab (REQUIRES THE ADVANCED VERSION OF LATENT GOLD)

The Advanced Tab is divided into 3 areas according to the section labels 'Survey', 'Continuous Factors', and 'Multilevel Model'.

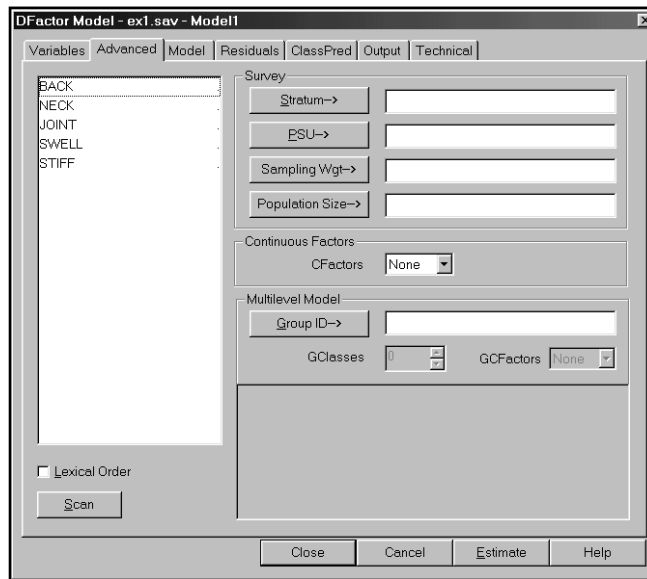


Figure 5-14. Advanced Tab for DFactor Model

The variables displayed in the variable list (left-most box of the Advanced Tab) are those that have not been specified previously for use as an indicator, dependent variable, predictor, covariate, known class indicator, case ID, case weight, or replication weight. These variables are eligible for use with any of these 3 advanced options.

## SURVEY

This advanced option can be used to specify information on the sampling design that was used to obtain your data. The program computes the design effect, as well as reports sampling design corrected standard errors and Wald statistics. Four aspects of the sampling design can be taken into account: stratification (Stratum), clustering (PSU), weighting (Sampling Wgt), and finite population size (Population Size).

For more details, see section 11 of Technical Guide.

## Stratum

The Stratum variable specifies the stratum to which a case belongs. When no Stratum variable is specified, it is assumed that all cases belong to the same stratum; that is, that there is only one stratum.

## PSU

The PSU(Primary Sampling Unit) variable is used for two-stage cluster samples. It specifies the (sampling) cluster to which a case belongs. PSUs are assumed to be nested within strata. When no PSU variable is specified, it is assumed that each case from a separate PSU.

## Sampling Wgt

The Sampling Wgt variable contains a sampling weight.

**Rescale (default) vs. No Rescale.** Upon selecting a variable to be used as a Sampling Wgt, the symbol <R> appears in the Sampling Wgt box to the right of the variable name to indicate that the weights will be rescaled. Rescaling of the original weights are accomplished by multiplying them by a constant

such that the sum of the weights equals the sample size.

- ▷ Right click on the variable name and select 'No Rescale' from the popup menu to maintain the weights without rescaling. Upon selection of 'No Rescale', the <R> symbol is removed.

**Active (default) vs. Inactive.** By default the sampling weights are used in the estimation to compute pseudo maximum likelihood estimates as indicated in Section 11.1 of Technical Guide. If the sampling weight were instead specified as a Case Weight in the Variables Tab, the resulting parameter estimates would be the same as when the Active option is used here but the standard errors are not correct.

The inactive option for sampling wgt employs an alternative 2-step estimation algorithm developed by Vermunt and Magidson (2001).

- ▷ Right click on the variable name and select 'Inactive' from the popup menu to select this option. Upon selection of 'Inactive', the <I> symbol appears in the Sampling Wgt box to the right of the variable name.

If the sampling weight variable were instead not used at all in the estimation (not specified as either a Case Weight nor a Sampling Wgt), the parameter estimates obtained would be the same as when the Inactive option is used here, but the sizes of the latent classes would be biased. The advantage of this method over the Active option is that the unweighted estimates may be more stable. See Section 11.2 of the Technical Guide for further information about these options.

## Population Size

The Population Size variable can be used to specify either the size of the population (# of PSUs in the Stratum concerned) or the population fraction. The variable is assumed to be a population fraction when it is smaller or equal to 1. This option can be use for finite population corrections.

## Continuous Factors

This advanced option can be used to include up to 3 continuous latent variables (CFactors) in an LC Cluster, DFactor or Regression model. In the Cluster and DFactor Modules, the use of CFactors yields (mixture variants of) factor-analytic models and various types of latent trait or IRT models. In the Regression (and Choice) Modules, it yields models with continuous random effects.

By default, the CFactors box is set to 'None'. To include CFactors in a model, click to open the drop down menu and select the number of CFactors to include in the model (1, 2, or 3). When 1 or more CFactors are included in the model, they appear on the Model Tab for further model specification.

By default, CFactors use 10 nodes to approximate normally distributed variables. To improve precision of the estimates, the number of nodes may be increased to a value as high as 50 (or reduced as low as 2). This change is made in the Continuous Factors section of the Technical Tab (see Step 8 for details).

**WARNING:** Inclusion of CFactors in a model may substantially increase the amount of time required to estimate the model. For example, inclusion of 2 CFactors, results in  $10 \times 10 = 100$  nodes used to approximate the bivariate normal distribution for these CFactors. Increasing the number of nodes to 50 results in  $50 \times 50 = 2500$  nodes, which will substantially increase the

amount of estimation time. For further details, see section 9.1 of Technical Guide.

When used with a 1-class cluster or 1-class regression model, the result is not a LC model. For example, a 2-CFactor 1-class model with continuous indicators is identical to a traditional linear factor analysis (FA) model. When one or more indicators is other than continuous, this model becomes an IRT variant of the FA model. For further details regarding the various kinds of applications with CFactors, see Section 9.2 of Technical Guide.

### MULTILEVEL MODEL

This advanced option is used to specify a multilevel extension to an LC Cluster, DFactor or LC Regression model which allows for explanation of the heterogeneity not only at the case level, but also at the group level.

Heterogeneity at the group level is explained by the inclusion of group-level classes (GClasses) and/or group-level CFactors (GCFactors) in a model.

#### Group ID

The Group ID variable indicates to which higher-level unit or group each case belongs. Upon selecting a variable as the Group ID, the Group Specification Box in the lower-right portion of the Advanced Tab is activated.

#### GClasses

This option assumes that groups belong to one of a set of latent classes of groups, the number of which is specified with GClasses (Group-level Classes). This yields the *nonparametric* variant of the multilevel LC model. By default, the GClass box is set to 1. To use this option to specify 2 or more GClasses

Click the up arrow in the drop-down box to increase the number of GClasses to 2 or more (up to 100). The GCFactors then appear in the Group Specification Box below.

#### GCFactors

This option assumes that groups differ with respect to their scores on one or more group-level continuous factors (GCFactors) or group-level random effects. This yields the *parametric* variant of the multilevel LC model. Click on the drop-down box to select the number of GCFactors. The GCFactors then appear in the Group Specification Box below.

GClasses and GCFactors may both be specified to combine the parametric and nonparametric approaches.

GClasses and GCFactors may

- affect the intercept and the covariate effects in the regression model for the Clusters, DFactors or Classes
- have direct effects on the indicators (see Model Tab in Cluster/DFactor)
- affect the intercept and the predictor effects in the model for the dependent variable (see Model Tab in Regression)

GClasses may themselves be affected by Group-level covariates (GCovariates).

When CGClasses or CGFactors are included in the model, they appear on the Model Tab for further model specification as described earlier; that is, to include effects on the indicators and the dependent variable. To include

GClass and/or GCFactor effects in the regression model for the Clusters, DFactors or Classes, use the Group Specification Box.

## Group specification box

The Group specification box at the lower-right of the Advanced Tab contains a column for GClasses and additional columns for each GCFactor specified. Click in the check boxes to allow estimation of desired parameters.

When GClasses and/or GCFactors are included, it is assumed that these affect the intercept in the regression model for the Clusters, Classes, or DFactors. This yields the standard multilevel latent class model in which class sizes are assumed to differ across groups by using a (parametric or nonparametric) random-intercept model for the latent classes.

GClasses and GCFactors may also be allowed to affect the covariate effects in the regression model for the Clusters, Classes, or DFactors. This is accomplished by checking the corresponding terms on the Advanced Tab.

GClasses play a role similar to the one of the Classes in a LC regression model: effects can be GClass independent (check off) or GClass dependent (check on). GCFactors play a role similar to CFactors in a random-effects regression model: effect can be assumed to be fixed (check off) or random (check on).

See Section 10 of Technical Guide details and application types for multilevel models.

## Step 8: Set Technical Options (Technical Tab)

The Technical Tab of the Analysis dialog box contains various technical options that are available to control the estimation process and need to be set before a model is estimated to take effect. In addition to the usual kinds of options affecting the number of iterations and convergence limits, the other options - Start Values, Bayes Constants, and treatment of missing data can affect the quality of the output results. For detailed descriptions of these options, see sections 6.2-6.7 of Technical Guide.

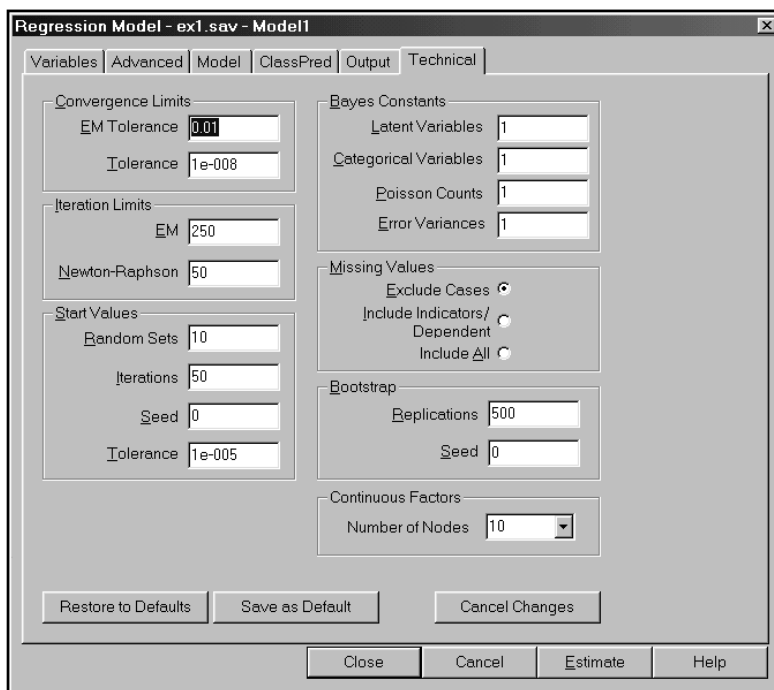


Figure 5-15. Technical Tab for Regression Model

### CONVERGENCE LIMITS

#### EM Tolerance

EM Tolerance is the sum of absolute relative changes of parameter values in a single iteration. It determines when the program switches from EM to Newton-Raphson (if the NR iteration limit has been set to  $> 0$ ). Increasing the EM Tolerance will switch faster from EM to NR. To change this option, double click the value to highlight it, then type in a new value. You may enter any non-negative real number. The default is 0.01. Values between 0.01 and 0.1 (1% and 10%) are reasonable.

#### Tolerance

Tolerance is the sum of absolute relative changes of parameter values in a single iteration. It determines when the program stops its iteration. The default is  $1.0 \times 10^{-8}$  which specifies a tight convergence criterion. To change this option, double click the value to highlight it, then type in a new value. You may enter any non-negative real number.

### ITERATION LIMITS

#### EM Iterations

Maximum number of EM iterations. The default is 250. If the model does not converge after 250 iterations, this value should be increased. You also may want to increase this value if you set NR iterations = 0. To change this option, double click the value to highlight it, then type in a new value. You may enter any non-negative integer.

#### Newton-Raphson

Maximum number of NR iterations. The default is 50. If the model does not converge after 50 iterations, this value should be increased. To change this option, double click the value to highlight it, then type in a new value. You may enter any non-negative integer. A value of 0 is entered to direct Latent GOLD to use only EM, which may produce faster convergence in models with many parameters or in models that contain continuous indicators or dependent variables.

### START VALUES

To reduce the likelihood of obtaining a local solution, the following options can be used to either increasing the number of start sets, the number of iterations per set, or both.

#### Random Sets

The default is 10 for the number of random sets of starting values to be used to start the iterative estimation algorithm. Decreasing the number of sets of random starting values for the model parameters reduces the likelihood of converging to a local (rather than global) solution. To change this option, double click the value to highlight it, then type in a new value. You may enter any non-negative integer. Using either the value 0 or 1 results in the use of a single set of starting values.

## Iterations

This option allows specification of the number of iterations to be performed per set of start values. Latent GOLD first performs this number of iterations within each set and subsequently twice this number within the best 10% of the start sets. For some models, many more than 20 iterations per set may need to be performed to avoid local solutions.

## Seed

The default value of 0 means that the Seed is obtained during estimation using a pseudo random number generator based on clock time. Specifying a non-negative integer different from 0, yields the same result each time.

If the current model setup was obtained by opening an .lgf file associated with a previously estimated model, 1) the Seed will not be 0 but will be the Best Start Seed for that model as specified in the .lgf file, and 2) the Random Sets parameter will be set equal to 0. This procedure assures that the model estimated is exactly the same model obtained when originally estimated, as long as the .lgf file was created using Latent GOLD 4.0 (see Warning below).

To specify a particular numeric seed (such as the Best Start Seed reported in the Model Summary Output for a previously estimated model), double click the value to highlight it, then type in (or copy and paste) a non-negative integer. When using the Best Start Seed, be sure to deactivate the Random Sets option (using Random Sets = 0). For further details see section 6.6 of Technical Guide.

**WARNING:** Due to improvements in this option in Latent GOLD 4.0, the random seed obtained from earlier versions of Latent GOLD will not necessarily reproduce the original model and has an increased chance of resulting in a local solution. Hence, if you open an .lgf file that was created using an earlier version of Latent GOLD, you should make sure to restore the default value of 0 and increase the value for Random Sets to the default value of 10 or some other desired quantity.

## Tolerance

Indicates the convergence criterion to be use when running the model of interest with the various start sets. The definition of this tolerance is the same as the one that use used for the EM and Newton-Raphson Iterations.

## BAYES CONSTANTS

The Bayes options can be used to eliminate the possibility of obtaining boundary solutions. You may enter any non-negative real value. Separate Bayes constants can be specified for three different situations:

### Latent Variables

The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used to prevent the occurrence of boundary zeroes in estimating the latent distribution. The number can be interpreted as a total number of added cases that is equally distributed among the classes (and the covariate patterns). To change this option, double click the value to highlight it, then type in a new value.

### Categorical Variables

The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used in estimating multinomial models with variables specified as Ordinal, Nominal or Binomial Count. The number can be interpreted as a total number of added cases to the cells in the models for the indicators (model for dependent) to prevent the occurrence of boundary zeroes. These pseudo cases are divided equally across classes and predictor/covariate patterns, and in accordance with the observed marginal distribution across categories of the indicator (dependent variable) concerned. To

change this option, double click the value to highlight it, then type in a new value.

### **Poisson Counts**

This prior is equivalent to adding a specified number of events to the data without changing the overall Poisson rate. In other words, the number of exposures is adjusted accordingly. This prior prevents boundary solutions in regression models where the Poisson count scale type is used. The default value for this Bayes constant is 1. To change this option, double click the value to highlight it, then type in a new value.

### **Error Variances**

The default is 1. Increase the value to increase the weight allocated to the inverse-Wishart prior which is used in estimating the error variance-covariance matrix in models for continuous dependent variables or indicators. The number can be interpreted as the number of pseudo-cases added to the data, each pseudo-case having a squared error equal to the total variance of the indicator (dependent variable) concerned. Such a prior prevents variances of zero from occurring. To change this option, double click the value to highlight it, then type in a new value.

## **MISSING VALUES**

The Missing Values option allows for the inclusion of records containing missing values on covariates and predictors as well as records containing missing values on the indicators. Including cases with missing values on covariates and predictors causes the mean to be imputed for the scale type numeric and the effect of the missing value category to be equated to zero for the scale type nominal. Missing values on indicators and dependent variables are handled directly in the likelihood function.

### **Exclude cases**

For Regression, selection of this option excludes all replications having missing values on the dependent variable or any of the predictors and all cases having missing values on any of the active covariates. For Cluster/DFactor, selection of this option excludes all cases having missing values on any of the indicators or active covariates.

### **Include indicators/dependent**

For Regression, selection of this option excludes replications having missing values on any of the predictors and cases having missing values on any of the active covariates. For Cluster/DFactor, selection of this option excludes cases having missing values on any of the active covariates. After exclusion of these cases, the remaining cases with missing values on the dependent variable (Regression) or on any indicator (Cluster and DFactor) are included in the analysis and handled directly in the likelihood function.

### **Include all**

Selection of this option includes all cases and replications in the analysis regardless of the presence of missing values. Cases or replications with missing values on the dependent variable (Regression) or on any indicator (Cluster and DFactor) are included in the analysis and handled directly in the likelihood function. Missing values on Predictors (Regression Module), or active covariates (Regression, Cluster and DFactor Modules) are imputed using Latent GOLD's imputation procedure

Inclusion in a model of covariates designated as inactive has no effect on which cases are excluded. Therefore, these missing values options have no effect with respect to the presence or absence of missing values on covariates specified to be inactive

## BOOTSTRAP OPTIONS (BOOTSTRAP $L^2$ , BOOTSTRAP -2LL DIFF)

The Technical Tab contains options for specifying the number of Replications and a Seed for both the Bootstrap  $L^2$  and the conditional bootstrap (Bootstrap -2LL Diff) procedures. Either of these bootstrap procedures can be requested from the Model Menu for an estimated model as described in Step 10.

### Replications

The default for the number of replication samples is 500. In most applications this number is large enough. The program also reports the Monte Carlo standard error of the p-value. By increasing this number, a more precise estimate of the p-value is obtained since the Monte Carlo error is reduced. With large models, to speed up the estimation you may consider reducing the number of replications.

### Seed

Seed can be used to specify the seed that is used to generate the replication data sets (the default value 0 means random seed) for either the Bootstrap of  $L^2$  or the Conditional Bootstrap ('Bootstrap -2LL Diff') procedures. Because of Monte Carlo simulation error, these bootstrap procedures yields a slightly different p-value each time that it is repeated, along with an estimate of the standard error. Specifying a particular seed guarantees the same result each time. By specifying the seed to be equal to the bootstrap seed reported in the Model Summary Output, one can replicate a previous run. In most bootstrap applications one will only use the Replications option.

If the Save Definition option in the File Menu is used to save a .lgf definition file for a model resulting from the Bootstrap, the Bootstrap Seed is saved. To reproduce results obtained from the Bootstrap of  $L^2$ , see the Note in Step 10 in the section on the Bootstrap p-value ( $L^2$ ). To reproduce results obtained from the Conditional Bootstrap, see Note2 in Step 10 in the section on the Conditional Bootstrap (-2LL Diff). For the Conditional Bootstrap, only the Bootstrap Seed associated with the source model is utilized.



## ADVANCED. CONTINUOUS FACTORS

### Number of Nodes

If 1 or more (group-level) continuous factors are specified in the Advanced Tab, this option determines the number of nodes used to approximate their normal distribution. By default, 10 nodes are used. Decreasing this number (minimum is 2 nodes) will speed up estimation time but reduce precision of the multivariate normality of the CFactors/GCFactors. For further details see Sections 9.1 and 10.1 of Technical Guide.

## DEFAULT OPTIONS

Click **Save as Default** to save the current technical settings as the new default values.

Click **Restore to Defaults** to restore the technical options to their last default values.

Click **Cancel Changes** to cancel any changes that have been made to the Technical options and not saved.

## Step 9: Set Output Options (Output Tab)

The Output Tab allows you to select various types of output listings which appear following estimation of your model(s). Each additional output selected for a model will be listed in the Outline pane. Chapter 6 provides a detailed specification of each of these output listings.

The ClassPred Tab allows you to output a data file containing selected classification and prediction information. This output file is only available if the input data is either an SPSS .sav file or an ASCII text file.

The Technical Guide provides related technical information, formulae and equations for all parameters and related statistics and other output.

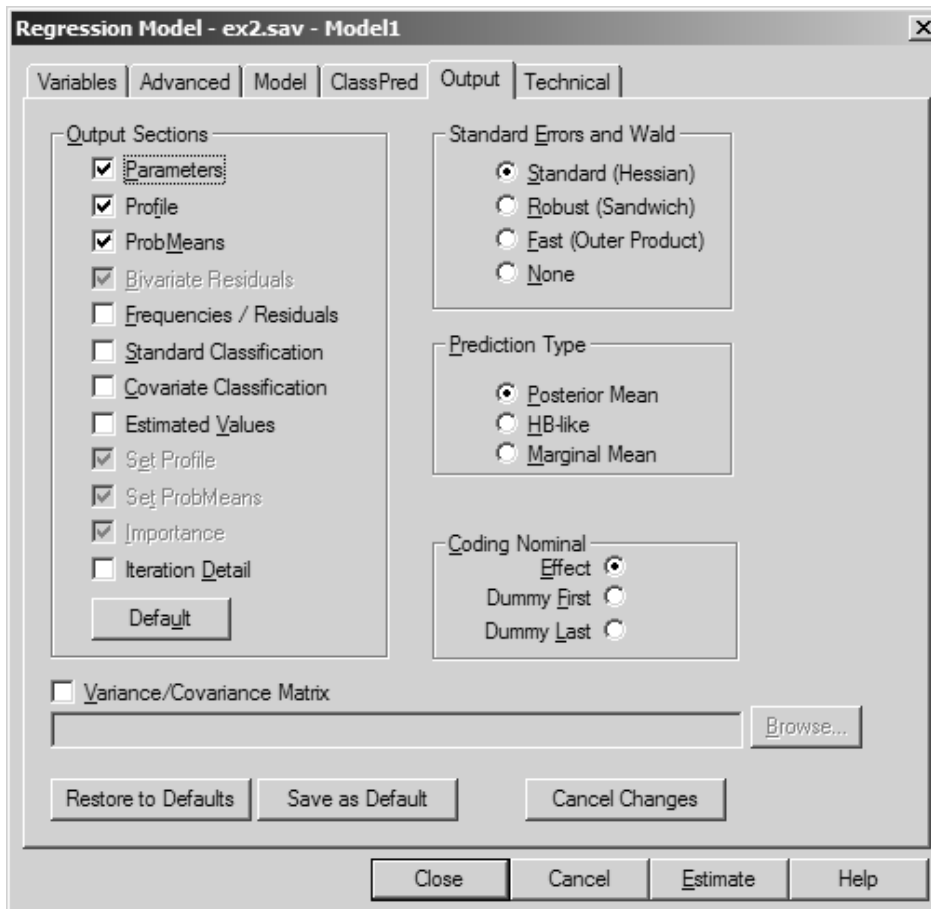


Figure 5-16. Output Tab for LC Regression Model

## OUTPUT SECTIONS

A checkmark indicates that the associated Output listings are produced. For details of these output files, see Chapter 6. By default the following are produced (checkmark equals on):

**Parameters.** Shows/hides Parameters in Output Window

**Profile.** Shows/hides Profile in Output Window

**ProbMeans.** Shows/hides ProbMeans in Output Window

They may be de-selected by clicking the check-box (Output check equals off) in which case this type of output will not appear.

The remaining output listings can be obtained by clicking the check-box (Output check equals on).

**Bivariate Residuals.** Produces a Table containing bivariate residuals. This output is not available with Regression Models. The output file will be listed as 'Bivariate Residuals' in the Outline pane.

**Frequencies/Residuals.** Produces an output file containing observed and estimated frequencies, and standardized residuals for each combination of variables. This output is not available if any variables in an analysis have been specified as Continuous or Count. It is also not available with Regression Models for which an ID variable has been specified (for repeated measures, for example). This output file will be listed as 'Freqs/Residuals' in the Outline pane.

## Classification Output (optional)

**Standard Classification.** Produces an output file listing containing posterior membership probabilities and other information used to classify cases into the appropriate cluster, latent class or DFactor level. This output file will be listed as 'Standard Classification' in the Outline pane.

Each row in the Standard Classification output corresponds to a distinct observed data pattern in the data file. For Cluster and Regression models, each of these rows contains the estimated probability of being in each class. For DFactor models, each row in the Standard Classification output contains the estimated probability of being in each DFactor level, the factor mean, and standard errors for these estimates.



For each CFactor and GCFactors, this file also contains the factor means and for GClasses the classification probabilities and the modal assignment.

**Covariate Classification.** Classification is usually performed based on all available information for a case (Standard Classification). However, it is also possible to compute the probability of being in a certain latent class (or a factor mean), given covariate values only. In fact these are model probabilities; that is,  $P(x|z)$  (see Technical Guide). These probabilities are useful for classifying new cases for which information on the dependent variable or indicators is not available.

Each row in the Covariate Classification output corresponds to a distinct pattern of active covariates that is observed in the data file. For Cluster and Regression models, each of these rows contains the estimated probability of being in each class. For DFactor models, each row in the Covariate Classification output contains the estimated probability of being in each DFactor level, the factor mean, and standard errors for these estimates. In multilevel models, it contains the GClass probabilities given group-level covariates.

**Note:** Inactive covariates do not influence the classification probabilities and hence have no effect on this output.

*Standard Classification* information as well as *Covariate Classification* can be viewed as Tabular output and/or can also be output to an external file. Selection of these from the Output Tab produces the Tabular output. Selecting *Standard Classification* and/or *Covariate Classification* from the 'Output to File' section of the ClassPred Tab produces the external files which contain the classification information as new variables appended to a copy of the input file used for estimation. See below:

**Estimated Values.** Shows/hides Estimated Values in Output Window (Regression only). This output section contains the class-specific and overall predicted values for each predictor pattern.

**Set Profile** (Available only for Choice Models) See Latent GOLD Choice Manual). Shows/hides Set Profile in Output Window

**Set ProbMeans** (Available only for Choice Models) See Latent GOLD Choice Manual). Shows/hides Set ProbMeans in Output Window

**Importance** (Available only for Choice Models) See Latent GOLD Choice Manual). Shows/hides Importance in Output Window

**Iteration Detail.** Shows/hides Iteration Detail in Output Window. If this output is not selected, it still will appear if any problems are encountered during model estimation.

### STANDARD ERRORS AND WALD

Choose one of four options:

The first three options specify different types of information matrices to be used in the computation of standard errors and Wald statistics. The fourth option suppresses such computations

#### Standard (Hessian)

The Standard method makes use of the second-order derivatives of the log-likelihood function called the Hessian matrix. This is the default option.

#### Robust (Sandwich)

The Robust method "sandwiches" the inverse of the outer-product matrix by the Hessian information. Standard errors and Wald statistics obtained by the Robust method are less affected by distributional assumptions about the indicators and the dependent variable.

#### Fast (Outer Product)

The Fast method approximates the information matrix using the outer-product of the first-order derivatives of the log-likelihood function. The Fast method may be used in models in which the other two methods are computationally intensive. In such cases, one can also suppress the computation of standard errors and Wald statistics.

#### None

This option suppresses the computation of standard errors and Wald statistics (option None). This option may be useful when estimating models containing an extremely large number of parameters, in which case computation of the second-order derivatives (used in Newton-Raphson, standard error computations and Wald statistics) may take a lot of time. By setting the *Newton-Raphson Iteration Limit* to 0 and setting *Standard Errors and Wald* to *off*, the estimation process for such large models is much quicker.

## PREDICTION TYPE (REGRESSION ONLY)

In regression, the program reports Prediction Statistics. It is also possible to write predicted values to a file. Predicted values can be computed in three different ways.

### Posterior Mean

Posterior Mean predicted values are defined as weighed averages of the class-specific predicted values using an individual's posterior membership probabilities as weights.

### HB-like

As in Hierarchical Bayes, the HB-like predicted values are based on the Individual Coefficients, which are weighted averages of the class-specific regression coefficients with the posterior membership probabilities as weights.

### Marginal Mean

Marginal Mean uses the prior membership probabilities as weights, which means that the observed values on the dependent variable are not used to generate the predictions.

Posterior Mean and HB-like prediction yield similar results. These methods give a good indication of within-sample prediction performance.

Marginal Mean prediction yields much lower R-sq values, but gives a better indication of out-of-sample prediction performance.

## CODING NOMINAL

**Effect** (default). By default, the Parameter Output contains effect coding for nominal indicators, dependent variable, active covariates and the latent classes (clusters). Use this option to change to dummy coding.

**Dummy Last.** Selection of this option causes dummy coding to be used with the last category serving as the reference category.

**Dummy First.** Selection of this option causes dummy coding to be used with the first category serving as the reference category.

## VARIANCE/COVARIANCE MATRIX

When the input data file is either an ASCII text file or an SPSS .sav file, this option outputs the variance-covariance matrix of the parameter estimates to an external file.

**Output Filename.** Upon selection of this option, a default filename appears in the box directly below the check-box. Use the browse button to change the filename and/or its save location. The format of the output file will be the same as that of the input file (ASCII or .sav).

## CHAPTER 5. BASIC STEPS FOR MODEL DEVELOPMENT

The body of the output file contains the variances and covariances of all model parameters. Each row in this output file corresponds to a parameter. The first variable (column) on this file is a string variable 'Location' which provides a unique name for the parameter such as 'r0001c01'. The 3 right-most columns in the output file are variables called 'se', 'param' and 'Label', serve to define the parameters. For example, for this parameter (row), the string variable called Label might contain a label such as 'purpose : 1 || Cluster1' which means that this is the parameter estimate for with the 1st category of the indicator PURPOSE, associated with cluster #1. The variables 'param' and 'se' correspond to the estimate and standard error for this parameter as reported in the Parameters Output.

The remaining variables on the file reproduce the parameter names provided in Location and contain the variance/covariance matrix. For example, the entry in row 1 (Location = 'r0001c01') and column 'r0001c01' is the variance of this parameter estimate. The entry in row 1 (Location = 'r0001c01') and column 'r0001c02' is the covariance associated with parameter estimates 'r0001c01' and 'r0001c02'.

**Note:** Most users will not need to use this option. These quantities are useful in computing the standard error of a particular function of the parameter estimates. For further details see Section 7.9 of Technical Guide.

### DEFAULT OPTIONS

Click **Default** to restore the Output options to their original program default values.

Click **Save as Default** to save the current output settings as the new default values.

Click **Restore to Defaults** to restore the Output options to their last saved default settings.

Click **Cancel Changes** to cancel any changes that have been made to the Output options and not saved.

For further details and examples, see section 2.4 of Technical Guide.

## ClassPred Tab: Classification and Prediction Output to a file

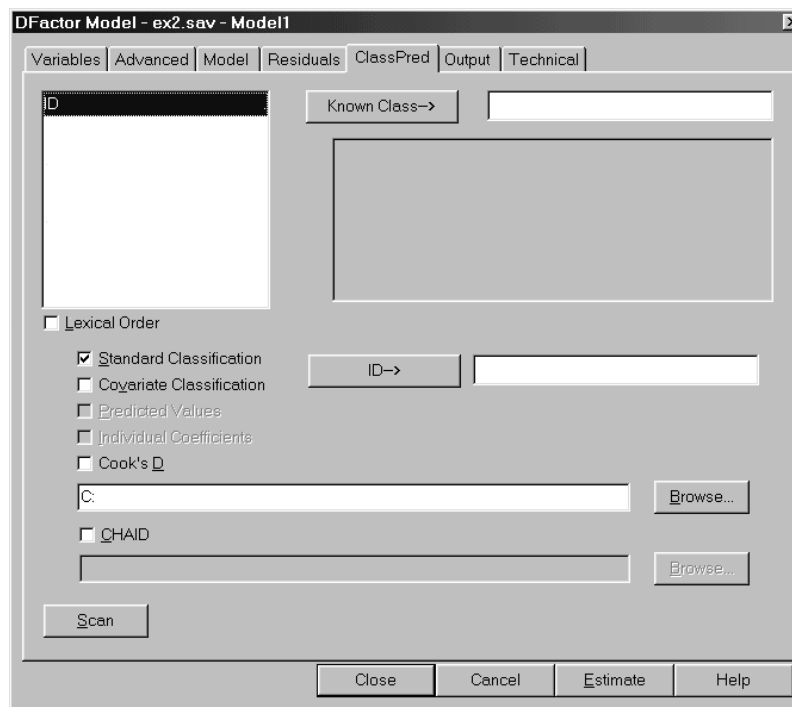
**CLASSIFICATION OUTPUT TO A FILE**

Figure 5-17. ClassPred Tab for DFactor Model

**Standard Classification (optional)**

When the input data file is either an ASCII text file or an SPSS .sav file, this option produces a new data file containing the standard classification information such as the probability of being in each cluster/class, or for DFactor models, in each level of each DFactor, together with any covariates and other variables specified in the Variables Tab and the variable included in the ID box of the ClassPred Tab (if any). The format of the output file will be the same as that of the input file (ASCII or .sav). This option is not available when using the range option in the Variables Tab to specify a range of models. In addition to these probabilities Latent GOLD also appends classification variables containing that cluster/ class/level into which the respondent should be classified (the one being the highest membership probability). For each case in the analysis file, the variables on the new data file consist of the model variables, the cluster/class/DFactor level probabilities, and class/DFactor classification (i.e., the index of the class/level containing the highest estimated probability). As an option, an ID variable can also be appended to the new file. For an example using this option, see Tutorial #3.

**Output Filename.** A default filename will appear in this box. Use the browse button to change the filename and/or its save location.

**ID Variable.** A single additional variable may be selected for inclusion (typically an ID variable or other key variable which provides a unique identification of each case on the file) to allow additional variables on the original data file that were not included in the analysis to be merged onto this file.

**Note:** The new file is created after the model has been estimated. After selecting this option, click Estimate to estimate the model and create the new file.

## CHAPTER 5. BASIC STEPS FOR MODEL DEVELOPMENT

**WARNING!** This setting is not preserved across models - it must be selected explicitly for each model estimated.



For each CFactor and GCFactor, the corresponding factor means are output; for GClasses, the classification probabilities and the modal assignment are output.

The order of the variables in the output file (and labels for a .sav formatted file) are as follows:

- the Known Class Indicator (if specified in the ClassPred Tab)
- any covariates included in the model
- other model variables specified in the Variables Tab
- the variable included in the ID box of the ClassPred Tab (optional)

For LC Cluster models:

- clu#1 (Cluster1) - posterior membership probability for cluster 1
- clu#2 (Cluster2) - posterior membership probability for cluster 2
- ..... (repeated for clusters 3,4, ...)
- clu# (Cluster modal) - the cluster number for the most likely (modal) cluster

For DFactor models:

- fac1#1: The probability of being in level 1 of factor 1
- fac1#2: The probability of being in level 2 of factor 1
- ... (repeated for levels 3,4, ...)
- fac1#: The level of factor 1 into which the case is classified
- fac1scr: The mean score for factor 1
- fac2#1: The probability of being in level 1 of factor 2
- fac2#2: The probability of being in level 2 of factor 2
- ... (repeated for levels 3,4, ...)
- fac2#: The level of factor 2 into which the case is classified
- fac2scr: The mean score for factor 2
- ... (sequence repeated for DFactors 3,4, ...)
- jfac#1 ('Joint DFactor 1 1')
- jfac#2 ('Joint DFactor 1 2')
- ..... (repeated for joint levels '1 3', ... , '2 1', ...)

For LC Segmentation/Regression models:

- clu#1 (Class1) - posterior membership probability for class 1
- clu#2 (Class2) - posterior membership probability for class 2
- ..... (repeated for classes 3,4, ...)
- clu# (Class modal) - the class number for the most likely (modal) class



Advanced:

- cfactor1 ('CFactor1')
- cfactor2 ('CFactor2')
- cfactor3 ('CFactor3')
- gclass1 ('GClass1')
- gclass2 ('GClass2')
- ...
- gcfactor1 ('GCFactor1')
- ...

## Covariate Classification

Classification based on covariates, as is the case with Standard Classification information can be output to an external file. Selecting Covariate Classification from the 'Output to File' section of the ClassPredTab produces the external files. The external file corresponding to the Covariate Classification information contains the new variables appended to a copy of the input file used for estimation.



For multilevel models, this output file also contains the GClass probabilities given group-level covariates.

## PREDICTION OUTPUT TO A FILE

### Predicted Values (Regression only)

Predicted values for the dependent variable can be output to an external file. The method used to determine the predicted values (pred\_dep in output file) depends on the Predicted Values setting on the Output Tab (Posterior Mean, HB-like, or Marginal Mean). The predicted value is a mean, except for nominal dependent variables for which it is a mode. For categorical dependent variables, the full estimated probability distribution for each replication is reported.

### Individual Coefficients (Regression only)

It is also possible to output posterior mean estimates for the Individual Coefficients to an external file. These are weighted averages of the class-specific effects, where the posterior membership probabilities of a case serve as weights. In the output file, the coefficients appear in the same order as in the Parameters Output and are labeled as b1, b2, b3, etc. Both the individual estimates (est\_b1, est\_b2, etc) and the individual standard deviations (std\_b1, std\_b2, etc) are provided.

### Cook's D

The Cook's Distance measure may be output to an external file. This measure is used to identify cases that have a large influence on the parameter estimates. A recommended cut-off point for Cook's distance is four times the number of parameters divided by the number of observations.

### CHAID (Requires a license to SI-CHAID 4.0)

This option creates a CHAID settings file (.chd file) from your model that can be then opened via the SI-CHAID 4.0 program. With this option, a CHAID (CHi-squared Automatic Interaction Detector) analysis may be performed following the estimation of any LC model in Latent GOLD 4.0. By selecting 'CHAID' as one of the output options, a CHAID input file will be constructed upon completion of the model estimation, which can then be used as input to SI-CHAID 4.0. For more information regarding CHAID, see section 1.6 in Chapter 1, and Tutorial #4 in Chapter 7.



*For additional information on the output to file options, see Section 7.9 of Technical Guide.*

## Step 10: Estimate the Model, View Output and Continue



Once you have specified your model, to estimate the model you may:

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

If your analysis window was originally opened from a new model (as opposed to a previously estimated model) you may estimate the model in other ways as well:

- ▷ From the Model Menu, select Estimate,

or

- ▷ Click 

If you opened a previously saved .lgf file and the Outline Pane contains names for 1 or more models that have not yet been estimated, to estimate all models that have not yet been estimated.

- ▷ click on the data file name in the Outline Pane
- ▷ select 'Estimate All' from the Model Menu

## STOPPING MODEL ESTIMATION (MODEL MENU)

Upon beginning a model estimation, the stop button on the toolbar becomes red (this may take several seconds or longer) which indicates that it is now active. You can stop a model estimation once it has begun to accomplish either 1) canceling the estimation, or 2) pausing the estimation to view preliminary results and/or make changes to the requested output options or change the iteration or convergence limits prior to resuming estimation.

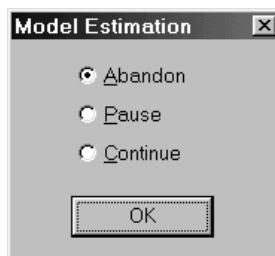


Once the stop button becomes active,

▷ To stop the estimation procedure select Stop from the Model Menu  
or

▷ click on the  button in the toolbar

and a popup-menu appears:

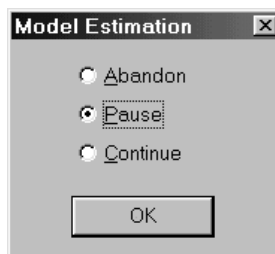


### Abandon/Cancel Estimation of a Model

▷ Select **Abandon** to cancel the model estimation

This option returns the program to its state prior to beginning the Estimation. Output is produced only for models that completed the estimation process without being terminated.

### Pause Model Estimation



The Pause option allows you to pause a model after the model estimation process has begun but prior to the estimation being completed to review preliminary Model Summary Output as well as any of the following Model

Output Sections that were requested in the Output Tab:

**Parameters**

**Profile**

**ProbMeans**

**Bivariate Residuals**

**Iteration Detail**

The Pause option is not available during estimation involving a bootstrap procedure.

If you are estimating a range of models, the pause option will pause the estimation process for the model currently being estimated and will cancel any further models. Depending upon how large the model is that is being estimated, it may take anywhere from one second to several minutes (or longer) to generate the preliminary output listings.

The Pause option does not cause a preliminary version of output to a data file to be created even if such was requested using the ClassPred Tab. If such has been requested, it will be produced only if the estimation is resumed and allowed to complete. See Resuming Estimation of a Paused Model below.

Prior to resuming model estimation, you may modify the Output options that were requested in the Output Tab or change the Iteration or Convergence Limits in the Technical Tab.

After a model is paused, the model name associated with that model appears in the Outline Pane with the characters 'Paused' appended to it. To change the Output options or the Iteration or Convergence limits, double click the model name to open the analysis dialog box and make the desired changes in the Output and/or Technical Tab. Note that the label 'Resume' replaces 'Estimate' on the Estimate button.

### Continue Model Estimation

If Stop was selected in error, click Continue to continue estimating the model.

### Resuming Estimation of a Paused Model (Model Menu)



**After viewing the preliminary output and making changes to the options as described above, to Resume the estimation of a paused model:**

- ▷ select Resume from the Model Menu
- or
- ▷ Click the Resume button at the bottom of the Analysis Dialog Box associated with the Paused model.
- ▷ To open the analysis dialog box for a Paused model double click the name of the paused model. In the Analysis Dialog box for a Paused model, the word 'Resume' replaces the word 'Estimate' on the Estimate button.

or, you may

- ▷ Click on the name of the Paused Model and select Resume from the Model Menu.

## BOOTSTRAP P-VALUE (L<sup>2</sup>) (MODEL MENU)

This option is only available if chi-squared statistics are available for the estimated model. Thus, it is not available:

- For any multilevel model
- For Cluster and DFactor models, if one or more indicators is specified as continuous.
- For Regression, if the dependent variable is specified as continuous.

With sparse data, the chi-squared based estimation for the p-value associated with L<sup>2</sup> cannot be trusted because these statistics do not follow a chi-squared distribution. When chi-squared statistics are available, a good alternative is to estimate the p-value by bootstrapping, or Monte Carlo simulation. The *Bootstrap L2* procedure involves generating a certain number of replication samples from the maximum likelihood solution and re-estimating the model with each replication sample. The bootstrap p-value is the proportion of replication samples with a higher L<sup>2</sup> than in the original sample.

A bootstrap p-value for the L<sup>2</sup> statistic can be obtained for a model after that model is estimated.



Once you have examined the results for an estimated model and wish to obtain a bootstrap p-value for it,

- ▷ right-click on the selected model to retrieve the model selection menu.

If chi-squared statistics are available for this model, the bootstrap option will be active on this menu.

- ▷ Select **Bootstrap L<sup>2</sup>**

The model will now be re-estimated using only a single start set and the best start seed, which guarantees that the bootstrap procedure starts with the previously estimated solution. Following the re-estimation of the model in the usual way, the estimation procedure for the bootstrap p-value begins. The number of Replications and the Seed used will be the ones provided in the Bootstrap section of the Technical Tab (see Bootstrap Options in Step 8).

After each replication of the procedure has completed, the bootstrap p-value is updated in the status bar displayed in the lower left portion of the screen.

Once the bootstrap procedure has completed, the name of the associated model together with the appended characters 'Boot' appears at the bottom of the Outline Tab. For example, if the estimated model is named 'Model3', the resulting model containing the bootstrap p-value will be named 'Model3Boot'. The resulting bootstrap p-value and standard error appear in the Model Summary Output in the columns to the right of the tradi-

tional p-value associated with this model name. All other output associated with the Boot model name is identical to that obtained earlier for the estimated model.

**Note:** Bootstrap information is provided only in the Model Summary Output listing. This information may be saved by copying it to the clipboard and pasting it elsewhere, or by using the Save Results option in the File Menu. If the Save Definition option in the File Menu is used to save a .lgf definition file for this model, the Bootstrap Seed is saved. Upon re-estimating the model following a File Open of the saved .lgf file, the model will again be estimated in the usual way but the bootstrap will not be performed. To reproduce the bootstrap, you will need to select Bootstrap  $L^2$  from the Model Menu for the re-estimated model. Use of the saved Seed guarantees that the same replication data sets will be generated and you will be able to reproduce the bootstrap results obtained previously.

### PROBChi Calculator (View Menu)

For assistance in assessing the improvement of a new model that imposes one or more testable restrictions on a previously estimated model, the ProbChi option allows you to obtain a p-value for a given chi-square (or vice-versa) for a specified number of degrees of freedom, df. For two models, Model2 being a more parsimonious (restricted) form of Model1, this tool tests whether the simplification is warranted using the difference between the  $L^2$  statistics for the two models,  $L^2(\text{Model2}) - L^2(\text{Model1})$ , with  $df = df(\text{Model2}) - df(\text{Model1})$ . Enter the  $L^2$  difference and the corresponding df and then click Chi->p to see if the difference between the models is significant. If not, you may use the more parsimonious model (Model2) without a significant loss of information.



To use this tool:

- ▷ From the menus choose View → ProbChi
- ▷ Enter a Chi-square value (or p-value) and number of degrees of freedom.
- ▷ Click on Chi->p (or p->Chi).

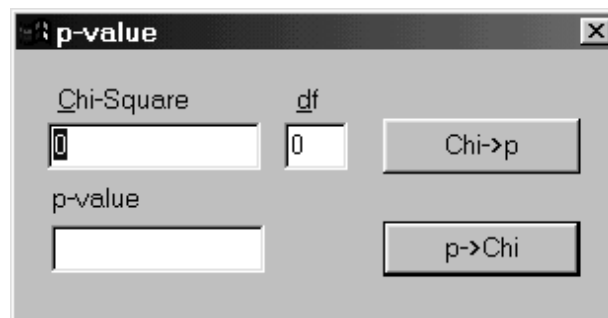


Figure 5-18: ProbChi dialog box

Given that the less restricted model is true, if the difference in model fit between the 2 models is not significant, the more restricted model may also be accepted as true and therefore preferred on grounds of being more parsimonious. On the other hand, if the difference is significant, the restricted model would be rejected in favor of the unrestricted model.

**Note:** To be valid, the chi-square value entered should only be an  $L^2$  value or a difference between two  $L^2$  values (as described above).

**WARNING:** The L-square difference is not a valid way to test the significance of adding 1 or more classes, DFactors, or levels associated with one or more DFactors to a model. To test these requires the more general conditional bootstrap approach.

## CONDITIONAL BOOTSTRAP (BOOTSTRAP -2LL DIFF) (MODEL MENU)

This option requires that you have already estimated at least 2 models, one of which is selected for the conditional bootstrap calculation. The second model, selected from a list of eligible models, must be a restricted version of (i.e., nested within) the selected model.

**Note:** This option is not available for multilevel models.

Similar to the use of the PROBChi calculator, the conditional bootstrap option (abbreviated as 'Bootstrap-2LL Diff') provides assistance in assessing the improvement of a model that imposes one or more testable restrictions on another estimated model by estimating a p-value. However, it is much more general than the -2LL difference test, and can be used to formally assess the statistical significance of imposing any set of model restrictions. Given that the less restricted model is true, if the difference in model fit between the 2 models is not significant, the more restricted model may also be accepted as true and therefore preferred on grounds of being more parsimonious. On the other hand, if the difference is significant, the more restricted model would be rejected in favor of the less restricted model.

A common application would be to test whether an additional class, DFactor or DFactor level provides a significant improvement in model fit. For example, since a 3 class model can be obtained from a 4-class model by restricting the size of the 4th class to be zero, the conditional bootstrap option can be used to test whether there is a significant difference in model fit between these 2 models. If the difference is not significant, the simpler 3-class model may be used in place of the 4-class model. If the difference is significant, you could conclude that the addition of the 4th class provides a significant improvement.

Another important application is in the testing of order restrictions, such as the ordered-restricted clusters (Cluster) and order-restricted predictor effects (Regression) assumptions. The test would then be between a model with and a model without the ordered restrictions of interest.



**To use this option to test the significance of the difference in fit between 2 model:**

- ▷ Estimate 2 or more models, at least one being a simpler, restricted form of another.
- ▷ After completion of the estimation, right click on the less restricted model.

## CHAPTER 5. BASIC STEPS FOR MODEL DEVELOPMENT

The available Model menu options appear.

▷ Select 'Boot -2LL Diff'

A list containing a subset of the other estimated models appears, containing possibly one or more simpler (nested) models that may be eligible to be used as reference models.

▷ Select the desired less restricted model

▷ Click OK

These steps are illustrated in Tutorial #1 in chapter 7 to test whether the improvement from a 3 to a 4-class model is statistically significant.

Once the conditional bootstrap procedure has completed, the names of 2 models -- the less restricted model and the reference model together with the appended characters 'Boot' -- appears at the bottom of the Outline Tab. The resulting conditional bootstrap p-value and associated standard error appears in the Model Summary Output in the columns to the right of the traditional p-value associated with the less restricted (selected) model. In addition, if chi-squared statistics are available, the bootstrap  $L^2$  p-value and standard error appear in the Model Summary Output in the columns to the right of the traditional p-value associated with the eligible reference model. All other output listings associated with these 2 'Boot' models is identical to that obtained earlier for the 2 models estimated originally.

**Note:** In order for the results of the conditional bootstrap to be valid, the reference model must be a restricted form of (i.e., nested within) the original model. You should be aware that the list of eligible models may well include some models that do not qualify as nested, and hence should not be selected. For example, you should make sure that both models include the same variables (dependent, exposure, caseid, indicators, covariates, predictors, attributes, case weights, replication weights, known class, etc.).

**Note 2:** Bootstrap information is provided only in the Model Summary Output listings. This information may be saved by copying it to the clipboard and pasting it elsewhere, or by using the Save Results option in the File Menu. If the Save Definition option in the File Menu is used to save a .lgf definition file for both the source model (say named 'Model2Boot') and the selected reference model (say 'Model1Boot'), the Bootstrap Seed is saved (for both models). To reproduce the bootstrap results, upon re-estimating both models following a File Open of the associated saved .lgf file(s), you will need to select 'Bootstrap -2LL Diff' from the Model Menu for the re-estimated source model and select the appropriate reference model again as described above. Use of the saved Seed guarantees that the same replication data sets will be generated and you will be able to reproduce the bootstrap results obtained previously. The 2 additional models estimated will have 'Boot' again appended to the original names, so the reproduced bootstrap models will be named for example 'Model1BootBoot' and 'Model2BootBoot'.

### OUTPUT FILES CREATED

Once a model has been Estimated or Paused, a number of output file listings are created which may be viewed, copied to the clipboard, or printed. Each file may be selected by highlighting the file name listed in the Outline pane. See Chapter 6 for specific details about the output produced. Following review of the output, you may elect

to estimate a new model.

## DEFINING A NEW MODEL

Once you have estimated at least one model, there are several ways to specify a new model on the same data file.

After estimating a model, Latent GOLD automatically creates a new model name in the Outline pane (it will be the last model name listed for a data file).



To setup a new model,

- ▷ Right click on the new Model<n> name in the Outline pane to open the Model Selection menu (which will have a checkmark next to the last type of model estimated) and proceed from Step 2 above.

When this method is used, the default new model settings will be the same as the last estimated model of that type, with the following exceptions:

- The Seed in the Start Values section of the Technical Tab is restored to '0' (random seed)
- The Random Sets value in the Technical Tab is restored to the original default
- Any output-to-file selections specified in the ClassPred Tab for the estimated model are removed
- ▷ Double click on the model name for a previously estimated model. The analysis dialog box will open and contain the settings for that model. Change whatever settings you wish to change, then click Estimate. The new model will appear at the bottom of the model list for that data file.
- ▷ Choose Define from the Model Menu after highlighting a previously estimated model. The analysis dialog box will open and contain the settings for that model. Change whatever settings you wish to change, then click Estimate. The new model will appear at the bottom of the model list for that data file.
- ▷ Highlight any model name and select the type of model from the Model Menu. A checkmark specifies either the type of model last selected (for new models) or the type of model selected for a particular model name (for previously estimated models).

After estimating several models, you may wish to delete some of these models, or rename models to be more informative names. The following model management options are available:

## CHANGING MODEL NAMES



To customize the model name for your model:

## CHAPTER 5. BASIC STEPS FOR MODEL DEVELOPMENT

- ▷ Single click the model name (for example, 'Model1') in the Outline pane to highlight it.
- ▷ Single click on it again to edit the name (the name will be in an edit box).

Model names can be edited at this time before or after the model is estimated

Once you have an edit box open, you may also right click to open a menu that allows you to Cut, Copy, Paste, and Delete the model name. This menu also allows you to Undo your last edit or select the entire name for editing (Select All).

### DELETING MODELS



**To delete a previously estimated model:**

- ▷ Single click the model name in the Outline pane to highlight it.
- ▷ Choose Delete from the Model Menu.

The model will be deleted from the Contents pane, and will no longer appear in the summary output when the corresponding data file is highlighted.



**To delete all models associated with a data set:**

- ▷ Single click the data file name in the Outline pane to highlight it.
- ▷ Choose File Close from the Model Menu.



# CHAPTER 6. MODEL AND MODEL SUMMARY OUTPUT

Following estimation of a model, various output file listings and graphical displays are produced. The specific output produced is controlled by the check-boxes on the Output Tab (see Step 9: Set Output Options in Chapter 5). In addition, case level information such as modal class assignments, posterior membership probabilities, regression predictions, and individual coefficients can be requested to be output to a data file using check-boxes on the ClassPred Tab (described at the end of Step 9 in Chapter 5).

This chapter describes each of the various output listings and plots that may be obtained using the Output Tab, and how to view this output. With the Advanced version of Latent GOLD, the Advanced Tab can be used to request the inclusion of CFactors, GClasses, and or GCFactors in a model. When such advanced features are used, additional information is included in the output. This additional information is also described in this chapter in sections labeled 'Advanced'. For further technical details and formulae see Part I (chapters 1-7) of Technical Guide for the various parameters and statistics associated with Latent GOLD Basic, and Part II (chapters 8-12) of Technical Guide for the various parameters and statistics associated with Latent GOLD Advanced.

In general, *all* output listings and displays are organized in sections using a hierarchical structure.

- For each data file opened, the Data File Summary Output contains summary information for all models estimated on these data.
- Within each data file, for each model estimated, the Model Summary Output contains information such as model specifications and performance statistics for that model.
- Within each model, several specific kinds of model output (such as Parameters Output) are listed in separate sections in the Outline pane beneath the model name.
- Within certain model output section listings, subsections containing specialized tables or plots are available as separate files.

The model output files Parameters, Profile, and ProbMeans are produced by default, regardless of the type of model estimated (Cluster, DFactor or Regression), and the Bivariate Residuals are also produced by default in Cluster and DFactor. The Profile and ProbMeans Output re-express certain parameters in the Parameters file in terms of conditional probabilities and/or means, and contain subcategories for plots (e.g., Prf-Plot). Another element common to the output for all three-model types is that output associated with active covariates is always presented in the bottom portion of Parameters Output. For Cluster models, this section is labeled Model for Clusters; for DFactor models, Model for DFactors, and contains the results for each factor separately; for Regression models, it is labeled Model for Classes. If no active covariates are included in the model, results in this section are given for the intercept only which relates to the size of the latent classes.

To view a particular output file, click on a name in the Outline pane and the associated output will be opened and displayed in the Contents pane (see Figure 6-1). To make visible the names of the specific kinds of output produced for an estimated model you will need to click on the expand/contract icon (+) to expand a model name or model output name. For example, to make the Prf-Plot visible you will need to expand the Profile output. When expanded, the expand/contract icon changes from (+) to (-).

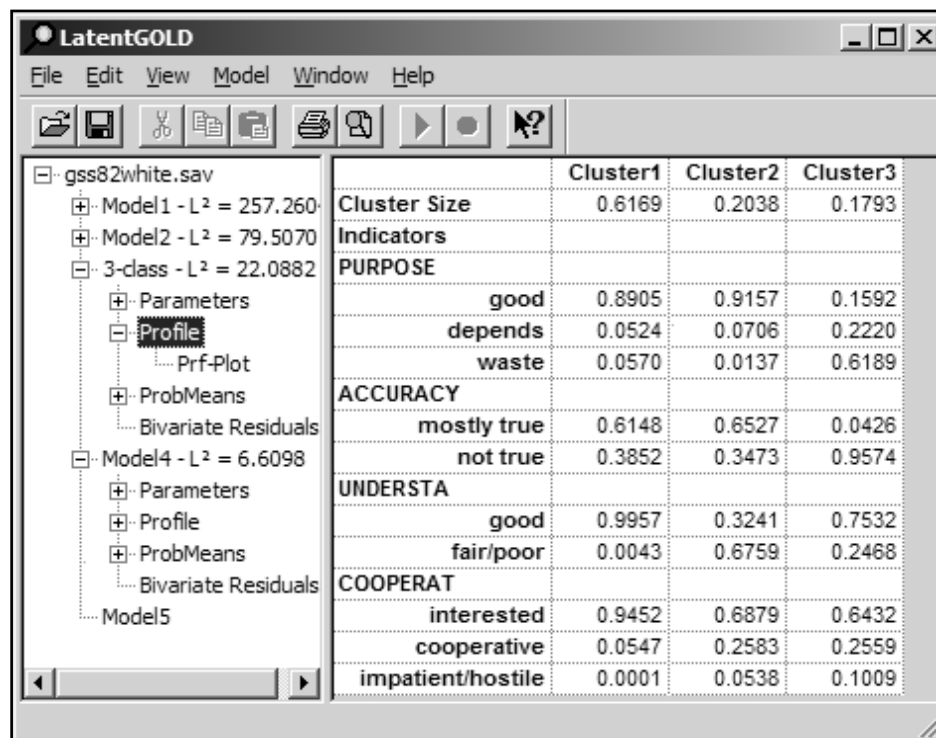


Figure 6-1. Expand/Contract Icon for 3-class Model changes to (-) which makes 'Prf-Plot' Visible. Expand/Contract Icon for 'Model4' remains as (+) so 'Prf-Plot' is Not Visible

Clicking on Prf-Plot causes this display to appear in the Contents Pane. For an example of this, see Figure 6-7.

The expanded output sub-categories may differ by type of model estimated. For example, for DFactor models, Parameters Output contains the subcategory listings Loadings, Correlations, and possibly also Error Correlations which re-express the parameters in a form similar to linear factor analysis, the latter subcategory available only when one or more indicators are continuous. For Cluster models, the Correlations subcategory does not appear and for Regression models, Parameters contains no subcategories. The use of the expand/contract icon is illustrated below.

### ORDERING OF LATENT CLASSES IN OUTPUT

Towards the goal of assuring comparability in the output when a model is re-estimated, Latent GOLD applies rules to establish a unique ordering of the classes. In general, classes are ordered by class size, from largest to smallest. However, if parameter restrictions are applied, this rule is ignored to allow for better correspondence between a model when estimated with and without such restrictions. For example, in an unrestricted model if a parameter associated with the largest class (class #1) is restricted to 0, the restricted model will show a 0 for that parameter for class #1, even if class #1 is no longer largest after imposing the restriction.

### ITERATION DETAIL

The model output file Iteration Detail produces similar kinds of information regardless of the type of model estimated (Cluster, DFactor or Regression). It is generated only if the Iteration Detail option is selected in the Output Tab before model estimation or if convergence problems or boundary solutions were encountered during the estimation.

This file contains technical information associated with the performance of the estimation algorithm, such as log-posterior and log-likelihood values at selected iterations:

- for each Random Set of starting values,
- for selected iterations of the EM algorithm,
- for selected iterations of the Newton algorithm.

When applicable, this file also contains warning messages concerning nonconvergence, unidentified parameters and boundary solutions, and additional iterations associated with an 'inactive sampling weight' when specified in the Advanced Tab.

### Estimation Warnings

#### **WARNING: negative number of degrees of freedom**

This warning indicates that the model contains more parameters than cell counts. A necessary (but not sufficient) condition for identification of the parameters of a latent class model is that the number of degrees of freedom is nonnegative. This warning thus indicates that the model is not identified. The remedy is to use a model with fewer latent classes (or DFactors)

**WARNING: # boundary or non-identified parameter(s)**

This warning is derived from the rank of the information matrix (Hessian or its outer-product approximation). When there are non-identified parameters, the information matrix will not be full rank. The number reported is the rank deficiency, which gives an indication of the number of non-identified parameters.

Note that there are two problems associated with this identification check. The first is that boundary estimates also yield rank deficiencies. In other words, when there is a rank deficiency, we do not know whether it is caused by boundaries or non-identified parameters. The Latent GOLD Bayes Constants prevent boundaries from occurring, which solves the first problem related to this message. However, a second problem is that this identification check cannot always detect non-identification when Bayes Constants are used; that is, Bayes Constants can make an otherwise non-identified model appear to be identified.

A more reliable method to check the identification of a LC model (when estimated with a particular data set) is as follows. Set all Bayes Constants, Random Sets, and both EM and Newton-Raphson Iteration Limits to 0, and select the Standard Errors and Wald method Fast (Outer-Product) prior to estimating the specified model. If no warning message appears, the model is identified. However, non-identification detected in this manner may be related to "true" non-identification (model is non-identified with any data set) or to the fact that there is not enough data (too few non-zero cells in the observed frequency table). The latter will lead to boundary estimates. These two explanations can easily be distinguished by adding one or more fictitious cases to the data set (with response patterns that are not already in the data set). With enough data, if the warning message still appears, the model is non-identified, irrespective of the data set.

**WARNING: # CFactor effect(s) should be excluded for identification**

For CFactor models with continuous indicators -- standard and mixture factor analysis - there is a special identification check for the factor structure. This may yield a warning indicating how many constraints should be imposed to get an identified factor structure. Usually, identification will be achieved by excluding one or more CFactor-Indicator effects as suggested by the message. Sometimes, one may instead wish to use the "Equal Effects" option to achieve identification.

**WARNING: maximum number of iterations reached without convergence**

This warning is provided if the maximum specified EM and Newton-Raphson iterations are reached without meeting the tolerance criterion. If the (by default very strict) tolerance is almost reached, the solution is probably ok. Otherwise, the remedy is to reestimate the model with a sharper EM tolerance and/or more EM iterations, which makes sure that the switch from EM to Newton-Raphson occurs later. The default number of 50 Newton-Raphson iterations will generally be more than sufficient.

**WARNING: estimation procedure did not converge (# gradients larger than 1.0e-3)**

This message may be related to the previous message, in which case the same remedy may be used. If the previous message is not reported, this indicates that there is a more serious non-convergence problem. The algorithms may have gotten trapped in a very flat region of the parameters space (a saddle point). The best remedy is to reestimate the model with other starting sets, and possibly with a larger number of Start Sets and more Iterations per set. If the problem repeats, the likely explanation is that the data is not informative enough about the parameter values.

## WARNING: # strata with only one PSU

A requirement for the complex survey estimator of the variance-covariance matrix of the model parameters is that each Stratum contains at least 2 PSUs. If a particular Stratum contains only one PSU, it is omitted from the computations of the variance-covariance matrix, which is clearly not the best solution. A better solution is for you to merge this stratum with another stratum, and re-estimate the model. The best remedy is, therefore, to perform the merging yourself.



*For further technical details regarding the estimation, see sections 6.3 - 6.8 in Technical Guide.*

The remainder of this chapter provides a description of each of the other output file listings and each output sub-category.

## Data File Summary Output and Model Fit

If you highlight a data file name in the Outline pane, the Contents pane lists general summary information regarding the data file (file name, location, size, date) and also lists names of all models that have been estimated in the current Latent GOLD session for this data file. By default, this output file reports the log-likelihood (LL), BIC based on LL, number of parameters (Npar), and the proportion of classification errors (Class.Err.) for all models. In addition, if chi-squared statistics are available, the likelihood-ratio statistic ( $L^2$ ), degrees of freedom (df), and the p-value are also reported.

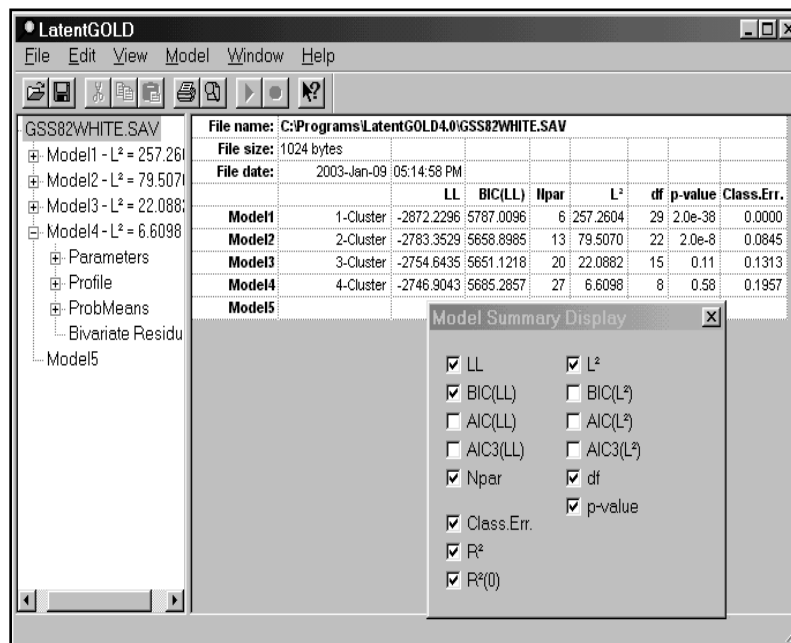


Figure 6-2. Data File Summary Output and associated Model Summary Display

The Model Fit likelihood ratio chi-squared statistic ( $L^2$ ) is one of several statistics that can be used to assess how well the model fits the data (how similar model-based estimated frequencies are to observed frequencies). In the context of latent class analysis,  $L^2$  can also be interpreted as indicating the amount of the observed relationship

between the variables that remains unexplained by a model; the larger the value, the poorer the model fits the data and the worse the observed relationships are described by the specified model. The associated p-value is a formal assessment of the extent to which the model fits the data (the null hypothesis of this test is that the specified model holds true in the population). It is obtained from a chi-squared table lookup with the reported number of degrees of freedom. Thus,  $p < .05$  indicates a poor fit.

As a general rule of thumb, a good fit is provided by a model when the  $L^2$  for that model is not substantially larger than the degrees of freedom which is the expected value for  $L^2$  under the assumptions that 1) the model is true and 2)  $L^2$  follows a chi-square distribution.

When dealing with a small sample size or sparse data, chi-square does not provide a good approximation to  $L^2$  and hence the p-value reported is not valid. Thus, the program also provides optional alternative statistics to assess model fit using the bootstrap procedure. The Bootstrap estimate of the p-value can be obtained when chi-square statistics are available and the conditional Bootstrap (Bootstrap -2LL Diff) can be obtained for any model to assess the significance of the improvement in fit due to an additional latent class, DFactor or DFactor level (see chapter 7 for examples of the use of these Bootstrap procedures).

In addition, information criteria such as the BIC may be used when the table is not sparse as well as when it is sparse. When chi-squared statistics are available such information criteria can be based on  $L^2$ , and when chi-squared statistics are not available, they can be based upon LL.

By selecting 'Summary Control' from the View Menu or with a right click from within the Contents Pane (right-hand) pane, the Model Summary Display menu becomes visible. This can be used to display the BIC and other items or to delete items to customize your summary table by clicking on the associated check-boxes in the Model Summary Display control panel.

Additional items that can be requested are AIC, AIC3, and BIC (based on  $L^2$  or LL) and BIC based on  $L^2$ . In addition to model fit, AIC, AIC3, and BIC take into account the parsimony of the model. They differ from one another according to how much weight is applied to penalize for each additional model parameter. When comparing models, the lower the value of the BIC (or AIC, AIC3), the better the model.



**See section 7.1.3 of Technical Guide for specific formulae for each of these statistics.**

## Model Summary Output

Once estimation of a model has been completed or you have paused the estimation prior to completion (by selecting Stop from the Model Menu or by clicking the Stop button), a Model Summary Output file is generated for that model and a name such as 'Model1', 'Model1Paused' (if the model was Paused), or 'Model1Boot' (if a bootstrap option is used) is assigned by the program and appears in the Outline pane. In addition, if one or more Output listings were requested prior to model estimation using the Output Tab, these listings are also generated and an expand/contract icon (+/-) appears to the left of the Model name. Opening the expand icon (+) by clicking it, changes its appearance to (-) and causes the names associated with these Output listings to appear beneath the model name.

To view the Model Summary Output, click on the model name and the Contents pane lists the contents of this output. Similarly, the contents of any other Model Output file can be viewed by clicking on the desired Output Section name. At the top of the Model Summary Output information appears which describes the type of model estimated (for example, '2-Cluster Model'). The next line contains any warning messages, followed by general model summary information. Additional information follows in sections labeled Chi-squared Statistics, Log-likelihood Statistics, Log-likelihood Statistics, Classification Statistics, Covariate Classification Statistics, and Variable Detail. These are described further below. When the Bootstrap  $L^2$  option is used, the Chi-squared Statistics section also contains the estimate for the associated bootstrap p-value. When the conditional bootstrap option is used, the Log-likelihood Statistics section contains the associated results. More specific information on the statistics including detailed formulae is provided in section 7 of Technical Guide.

Possible warning/alert messages include the following:

**Estimation Warnings!** This message appears if either boundary solutions, identification problems or convergence problems have been encountered during the estimation of the model. See **Iteration Detail** for more information on estimation warnings.

**Model Paused.** This message occurs whenever you cause the model to Pause prior to completion of the estimation. The output files in this case should be viewed as preliminary in this case. (To resume a paused model, you may simply right click on the model name of the paused model and select Resume from the pop-up menu.). See Step 10 in Chapter 5 for more information on pausing and resuming model estimation.

The specific contents of the Model Summary Output and other Model Output files depend upon the type of model estimated. In the remainder of this chapter we describe the various model output separately for each of the three modules - Cluster, DFactor and Regression.

## Cluster

### GENERAL INFORMATION

The Cluster Module is used to estimate LC Cluster models containing 1 or more latent classes (clusters). When 2 or more clusters are included in a model, as a general rule, the clusters are ordered in the output according to their size from the largest to the smallest. Exceptions may occur when parameter restrictions are applied. For example, when the Order Restricted Clusters option is used, Clusters are ordered according to the parameter values.

In addition to the Model Summary Output, additional specific model output file listings are generated following estimation of the model. The specific output sections that can be generated are those that appear as active in the Output Sections portion of the Output Tab. For Cluster models, these are the first 7 model Output Sections listed, plus the last one, Iteration Detail (see Figure 6-3 below).

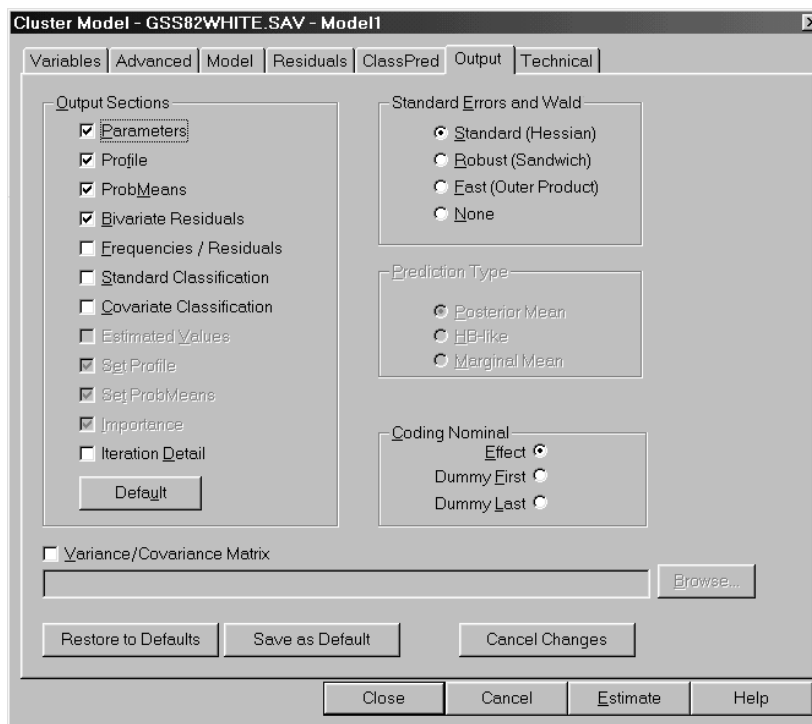


Figure 6-3. Output Tab for Cluster Models

Prior to estimating a Cluster model, click in the active check boxes to the left of the Output Sections to include or exclude that Section from being part of the model output files that are generated following the model estimation. A detailed description of each of the Model Summary Output as well as each specific Model Output Section file that can be output appears below.

## LC CLUSTER MODEL SUMMARY OUTPUT

If you click in the Outline pane on the name of any estimated Cluster model (e.g., 'Model1'), the Contents pane lists general summary information associated with this model including the type of model, any warning messages, the value of the Random Seed used in starting sets that can be used to reproduce exactly the same run, and the value of the best seed that can be used to reproduce the reported model results. See section 6.6 of Technical Guide for further details regarding the random seeds. If order-restricted models have been estimated, following the number of parameters, the number of activated constraints is provided.

Specifically, the general section of the Model Summary Output contains the following items:

**Type of Model.** For example, '3-Cluster Model' indicates that a 3-class Cluster model has been estimated.

**Warning messages** (generally not present). Estimation Warning message(s) appear here to indicate that boundary solutions, identification or convergence problems were encountered during the estimation of the model.

**Model Paused** (if you paused the model). If you Paused the model prior to estimation being completed this message appears to alert you that the output files should be viewed as preliminary.

**Number of cases.** This is the number of cases used in model estimation. This number may be less than the original number of cases on the data file if missing cases have been excluded.

**Advanced: Number of groups.** In multilevel latent class models, the program reports the number of groups used in model estimation.

**Number of parameters.** This is the number of distinct parameters estimated.

**Number of Activated Constraints** (appears only if order-restricted clusters were requested). This is the number of parameters that needed to be restricted in order to obtain ordered clusters plus the number of error variances that were fixed to the smallest allowed value. This number plus the number of parameters estimated equals the total number of model parameters in a model obtained without the order-restriction.

**Random Seed.** The seed required to reproduce this model.

**Best Start Seed.** The single best seed that can reproduce this model more quickly using the number of starting sets =0. This is the seed that is automatically inserted in a saved definition (.lgf) file.

**Advanced: Design Effect.** When the Survey option is used, the program reports the generalized design effect.

LatentGOLD

File Edit View Model Window Help

GSS82WHITE.SAV

- Model1 -  $L^2 = 257.261$
- Model2 -  $L^2 = 79.5071$
- Model3 -  $L^2 = 22.088$
- Model4 -  $L^2 = 6.6098$ 
  - Parameters
  - Profile
  - ProbMeans
  - Bivariate Residu
- Model5

3-Cluster Model

Number of cases	1202			
Number of parameters (Npar)	20			
Random Seed	836991			
Best Start Seed	2129580			
Chi-squared Statistics				
Degrees of freedom (df)	15	p-value		
L-squared ( $L^2$ )	22.0882	0.11		
X-squared	23.5099	0.074		
Cressie-Read	22.6617	0.091		
BIC (based on $L^2$ )	-84.2880			
AIC (based on $L^2$ )	-7.9118			
AIC3 (based on $L^2$ )	-22.9118			
CAIC (based on $L^2$ )	-99.2880			
Dissimilarity Index	0.0272			
Log-likelihood Statistics				
Log-likelihood (LL)	-2754.6435			
Log-prior	-4.8632			
Log-posterior	-2759.5067			
BIC (based on LL)	5651.1218			
AIC (based on LL)	5549.2869			
AIC3 (based on LL)	5569.2869			
CAIC (based on LL)	5671.1218			
Classification Statistics				
Classification errors	0.1313			
Reduction of errors (Lambda)	0.6570			
Entropy R-squared	0.5794			
Standard R-squared	0.6116			
Classification log-likelihood	-3224.8126			
AWE	6793.2949			
Classification Table				
Probabilistic	Cluster1	Cluster2	Cluster3	Total
Cluster1	704.1204	3.0920	34.5941	741.8065
Cluster2	72.9409	163.7828	8.1632	244.8869
Cluster3	27.9387	11.1252	176.2427	215.3066
Total	805.0000	178.0000	219.0000	1202.0000

Figure 6-4. Model Summary Output for Cluster Model

## Chi-squared Statistics

This section lists various chi-square based statistics related to model fit.

If the scale type for one or more indicators in a model has been set to 'continuous', no chi-squared statistics are available and this section is not displayed. The same applies when the multilevel option (Advanced) is used.

The information reported:

- Degrees of freedom (df). The degrees of freedom for the current model.
- L-squared ( $L^2$ ). The likelihood-ratio goodness-of-fit value for the current model. If the bootstrap p-value for the  $L^2$  statistic has been requested, the results will be displayed here.
- X-squared and Cressie-Read. These are alternatives to  $L^2$  that should yield a similar p-value according to large sample theory if the model specified is valid and the data is not sparse.
- BIC, AIC and CAIC (based on  $L^2$ ). In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC and CAIC value the better the model.

- **Dissimilarity Index.** A descriptive measure indicating how much the observed and estimated cell frequencies differ from one another. It indicates the proportion of the sample that needs to be moved to another cell to get a perfect fit.



*For more detailed information about Chi-squared statistics including the formulae for each, see section 7.1.1 of Technical Guide.*

## Log-likelihood Statistics

This section contains additional statistics related to the model fit that are especially useful when  $L^2$  and the other chi-squared statistics are not available. The statistics reported are:

- **Log-likelihood (LL).** If the conditional bootstrap (bootstrap -2LL Diff) has been requested, the results will be displayed here.
- **Log-prior** - this is the term in the function maximized in the parameter estimation that is associated with the Bayes constants. This term equals 0 if all Bayes constants are set to 0.
- **Log-posterior** - this is the function that is maximized in the parameter estimation. The value of the log-posterior function is obtained as the sum of the log-likelihood and log-prior values.

BIC, AIC, AIC3 and CAIC (based on LL) - these statistics (information criteria) weight fit and parsimony by adjusting the LL to account for the number of parameters in the model. The lower the value, the better the model. For example, according to the BIC values shown in Figure 6-4, the 3-class model is preferred over the 1-class, 2-class and 4-class models. Note: the same conclusions are obtained from information criteria based on LL or  $L^2$ . The numbers differ only by a constant that depends on the data set at hand.




*For more detailed information about Log-likelihood statistics including the formulae for each, see section 7.1.2 of Technical Guide*

## Classification Statistics

This information can be used to assess how well the model classifies cases into clusters. The statistics reported are:

- **Classification Errors.** When classification of cases is based on modal assignment (to the class having the highest membership probability), the proportion of cases that are estimated to be misclassified is reported by this statistic. The closer this value is to 0 the better.
- **Reduction of Errors (lambda), Entropy R-squared and Standard R-squared.** These pseudo R-squared statistics indicate how well one can predict class memberships based on the observed variables (indicators and covariates). The closer these values are to 1 the better the predictions.
- **Classification log-likelihood.** Log-likelihood value under the assumption that the true class membership is known.
- **AWE.** Similar to BIC, but also takes classification performance into account.
- **Classification Table.** The Classification Table cross-tabulates modal and probabilistic class

assignments.

-  **Standard R-squared** is reported for CFactors and GCFactors. For GClasses, as for Clusters, Classification Errors, Reduction of Errors, Entropy R-squared, and Standard R-squared values are reported.



*For more detailed information about the classification statistics including the formulae for each, see section 7.1.3 of Technical Guide.*

### Covariate Classification Statistics

If one or more active covariates are included in the model, additional statistics are provided as above but these are now based only on the active covariates. The statistics reported are:



The same information is obtained for classification into GClasses based on group-level covariates - **Classification errors, Reduction of Errors (lambda), Entropy R-squared and Standard R-squared**



*For more detailed information about the classification statistics including the formulae for each, see section 7.1.3 of Technical Guide.*

### Variable Detail

This section contains details about the variables entered into the model including variable names, scale types, number of categories and category labels and scores (if used).

### PARAMETERS OUTPUT (OPTIONAL)

For any estimated model, a Parameters file listing is generated by default. It contains estimates of parameters in the Model for Indicators (betas, sigmas), and in the Model for Clusters (gammas), as well as measures of significance for these estimates and the model  $R^2$  for each indicator.

The  $R^2$  indicates how well an indicator is explained by the model. For **ordinal**, **continuous**, and **counts**, these are standard  $R^2$  measures. For **nominal** variables, these are Goodman-Kruskal tau-b coefficients, representing a weighted average of separate  $R^2$  measures for each category treated as a separate dichotomous response variable. The  $R^2$  is similar to the explained variance in analysis of variance and to item communalities in factor analysis.



If CFactors have been included in the model, estimates and related information for additional parameters (lambdas) are included in the Models for Indicators output.



If GClasses and/or GCFactors have been included in the model, estimates and related information for additional parameters are included in the Models for Indicators and/or the Model for Clusters section of the Parameters output, and a separate section called Model for GClasses appears at the bottom of this output listing.

Models for Indicators		Cluster1	Cluster2	Cluster3	Wald	p-value	R <sup>2</sup>
<b>PURPOSE</b>							
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			
<b>ACCURACY</b>							
mostly true		0.5695	0.6512	-1.2207	8.3506	0.015	0.2003
not true		-0.5695	-0.6512	1.2207			
<b>UNDERSTA</b>							
good		1.7485	-1.3369	-0.4116	7.4225	0.024	0.4645
fair/poor		-1.7485	1.3369	0.4116			
<b>COOPERAT</b>							
interested		1.9589	-0.8539	-1.1050	18.9606	0.00080	0.1068
cooperative		0.6931	-0.2500	-0.4431			
impatient/hostile		-2.6520	1.1039	1.5481			

Figure 6-5. Parameters Output for Cluster Models

**Betas.** For Cluster models, the most important betas are that ones that indicate the strength of the effects of the clusters on the indicators. Other betas are the indicator intercepts, the direct associations between categorical (Nominal/Ordinal) indicators, and direct effects of covariates on indicators.

**Gammas.** Parameters of the multinomial logit model used to predict the clusters as a function of the covariates. Parameters include the intercept as well as the effects of each covariate.

**Sigmas.** Error variances and covariances for continuous indicators.

## Viewing Wald Statistics and Standard Errors

By default, Wald statistics are provided in the output to assess the statistical significance of a set of parameter estimates. For example, for each indicator the Wald statistic tests the restriction that each estimate in the set of beta parameter estimates associated with that indicator equals zero. A non-significant p-value associated with this Wald statistic means that the indicator does not discriminate between the clusters in a statistically significant way. For variables specified as Nominal, the set includes parameters for each category of the variable.

To view standard errors or related statistics associated with the parameter estimates, simply right click on the Parameters Output in the Contents Pane to retrieve a pop-up menu, and select/deselect the items to appear in various columns of this output listing. Alternatively, you can select the desired items from the View Menu. By default, Wald statistics appear in Parameters Output. You can suppress these Wald statistics, or replace them with standard errors, a Z Statistic, or both standard errors and the associated Z Statistic.

## Parameters Output Subcategories

Clicking on the + to the left of Parameters makes visible the Parameters Output Subcategories, which provide additional information regarding Parameters Output. The subcategories are Loadings and Error Correlations.

**Loadings.** The Loadings output contains the 2 columns labeled 'Clusters' and 'R<sup>2</sup>'. When only these 2 columns appear, the quantities reported under Clusters are simply correlations, representing the square root of the corresponding 'R<sup>2</sup>' or communality of the associated indicator. These numbers can be interpreted as standardized linear regression coefficients. The R<sup>2</sup> is the same as reported in the Parameters Output.



Additional columns appear immediately to the right of the 'Clusters' column when CFactors, GClasses and/or GCFactors are included in the model. The quantities reported in the columns to the left of the R<sup>2</sup> represent the decomposition of the R<sup>2</sup> into correlation components associated with each of these column effects. For further details, see section 12.2 of Technical Guide.

**Error Correlations** (available only when one or more indicators are continuous)-

The Error Correlations output provides the estimates of the within-cluster correlations between continuous indicators variables. Often, these correlations are easier to interpreted than covariances, which are the actual model parameters. (Note that the off diagonal entries are all 0 unless direct effects between continuous indicators have been included in the model)



*For more information, see section 7.2 in the Technical Guide.*

## PROFILE OUTPUT (OPTIONAL)

To view the profile table for a selected model, click Profile in the Outline pane. The Profile table contains probabilities or means associated with each Indicator or Dependent variable.

The screenshot shows the LatentGOLD software window. On the left is the 'Outline' pane with a tree view containing: GSS82WHITE.SAV, Model1 - L<sup>2</sup> = 257.2604, Model2 - L<sup>2</sup> = 79.5070, Model3 - L<sup>2</sup> = 22.0882, Model4 - L<sup>2</sup> = 6.6098 (expanded), Parameters, Profile (selected), ProbMeans, Freqs/Residuals, Bivariate Residuals, Standard Classification, Covariate Classification, Iteration Detail, and Model5. The main window displays a table with the following data:

	Cluster1	Cluster2	Cluster3	Cluster4
<b>Cluster Size</b>	0.5524	0.2237	0.1544	0.0695
<b>Indicators</b>				
<b>PURPOSE</b>				
good	0.9294	0.9131	0.1973	0.2366
depends	0.0347	0.0811	0.2523	0.1475
waste	0.0359	0.0058	0.5504	0.6159
<b>ACCURACY</b>				
mostly true	0.6282	0.6627	0.1383	0.0486
not true	0.3718	0.3373	0.8617	0.9514
<b>UNDERSTA</b>				
good	0.9916	0.4105	0.9405	0.4389
fair/poor	0.0084	0.5895	0.0595	0.5611
<b>COOPERAT</b>				
interested	0.9569	0.7081	0.8316	0.3340
cooperative	0.0430	0.2496	0.1628	0.3960
impatient/hostile	0.0001	0.0422	0.0056	0.2699

Figure 6-6. Profile Output for Cluster Model

- For a Cluster model, the first row of numbers shows how large each cluster is. For example, Figure 6-6 shows that Cluster 1 contains about 55% of the respondents (.5524) , Cluster 2 contains 22%, Cluster 3 contains 15% and Cluster 4 contains the remaining 7% .
- The body of the table contains (marginal) conditional probabilities that show how the clusters are related to the Nominal, Ordinal or Binomial Count indicator variables. These probabilities sum to 1 within each cluster (column). For example, Figure 6-6 shows that respondents in Cluster 1 have a 92.94% chance of responding that surveys serve a 'good' purpose.
- For indicators specified as Continuous or Count, the body of the table contains means (rates) instead of probabilities. For indicators specified as Ordinal, means are displayed in addition to the conditional probabilities.
- For covariates the Profile output contains rescaled ProbMeans output rather than model probabilities or means. These are aggregate class membership probabilities which are rescaled to sum to one within clusters. For Covariates specified as Numeric, means are displayed in addition to the conditional probabilities, and if the Covariate contains more than 5 distinct values, the probabilities are displayed for 5 grouped ranges. The number of such grouped ranges may be changed (increased or decreased) using Groups option in the Plot Control.
- The probabilities and means that appear in the Profile Output, are displayed graphically in the Profile Plot (see below for details).

## Viewing Standard Errors

To view the associated standard errors in a Profile table, from the View Menu select Standard Errors, or simply right click from within the Contents Pane to retrieve the pop-up menu and click Standard Errors. A column containing the standard errors appears to the right of each estimate. Standard Errors are not provided for Covariates.

## Partial Profile

Latent GOLD reports not only marginal but also partial probabilities/means for indicators in the Profile Table Output. These two differ only if direct effects between indicators or between a covariate and indicator are included in a model using the Residuals Tab (See Step 7, Chapter 5). As explained in section 7.3 of the Technical Guide, partial probabilities/means are obtained by conditioning on a certain value (mean) of the other variables involved in the model for the indicator concerned. *Marginal* probabilities/means, on the other hand, are obtained by collapsing over the categories of these "other" variables. Both have their advantages: the *Marginal Profile* output is somewhat easier to interpret and is displayed by default. *Partial Profile* reflects somewhat better the strength of the effects of the cluster on the indicators.

The distinction between Marginal and Partial is not relevant for Covariates.

To change from displaying Marginal to Partial probabilities/means, from the View Menu select Partial, or right click from within the Contents Pane to retrieve the pop-up menu and click Partial.



**For more detailed information about the Profile Output, see section 7.3 of Technical Guide.**

## PROFILE PLOT (FOR INDICATORS AND COVARIATES)

To view the Profile Plot, click the '+' icon to expand the Profile output and highlight Prf-Plot. The Profile Plot is constructed from the conditional probabilities for the nominal variables and means for the other indicators and covariates as displayed in the columns of the Profile table. Specific clusters (columns) are selected for display using the Plot Control pop-up menu. The quantities associated with the selected clusters are plotted and connected to form a line graph.

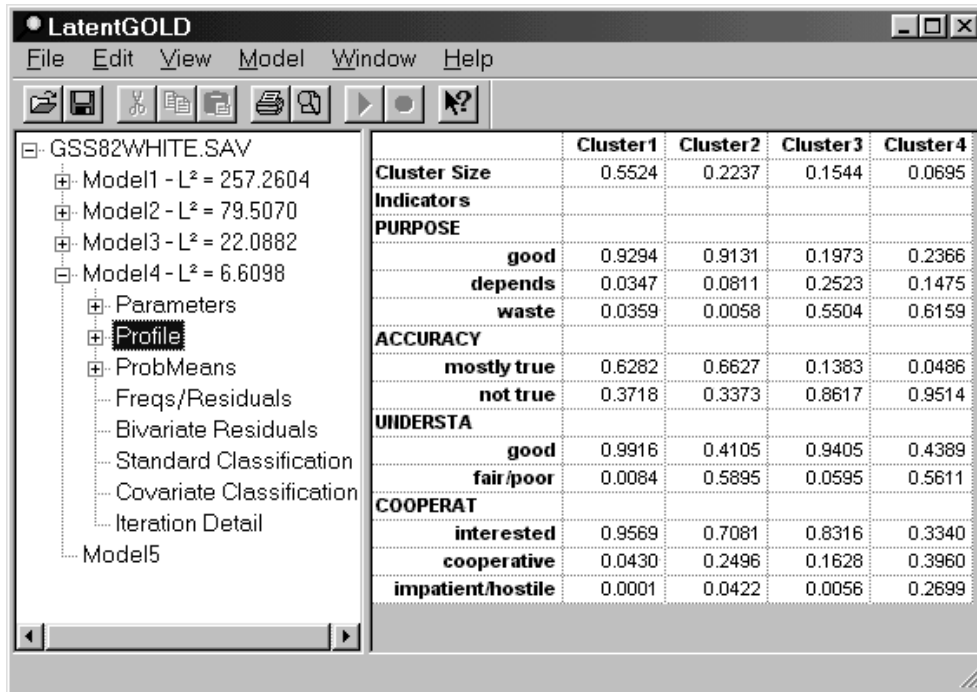


Figure 6-7. Profile Plot for Cluster Model

For the scale types **ordinal**, **continuous**, **count**, and **numeric covariate**, prior to plotting the class-specific means, they are re-scaled to always lie within the 0-1 range. Scaling of these "0-1 Means" is accomplished by subtracting the lowest observed value from the class-specific means and dividing the results by the range, which is simply the difference between the highest and the lowest observed value. The advantage of such scaling is that these numbers can be depicted on the same scale as the class-specific probabilities for **nominal** variables.

For **nominal** variables containing more than 2 categories, all categories are displayed simultaneously. For dichotomous variables specified as nominal, by default only the last category is displayed. The Plot Control (see Figure 6-8 below) can be used to display the first or both categories, as well as to specify the columns (Clusters), variables, and selected categories to appear in the display.

- A separate line is displayed for each cluster.
- Click on any variable symbol in the Profile Plot and the status bar describes it (variable name, cluster number, value). When 0-1 means are displayed, the status bar displays values for both the 0-1 mean and the original mean as shown in the Profile output.
- Click on any cluster name or symbol in the legend and Latent GOLD highlights all the symbols that refer to that cluster.

- When the contents of the Profile Table is changed from the default view to display Partial (instead of Marginal) probabilities, the points plotted change to reflect the current contents of the Profile Table.

By default, only the first 8 variables are displayed in the profile plot. Additional variables may be selected from the Plot Control (using a control-click or a shift click to select more than one at a time) and added or removed from the plot with a single click in the box to the left of the variable names.

## To Change Settings for a Profile Plot

To change the settings for a Profile Plot, from the View Menu select Plot Control, or right click within the Contents pane when a Profile Plot is displayed to open the Plot Control dialog box. To change the font type/size for a plot, see Chapter 2.

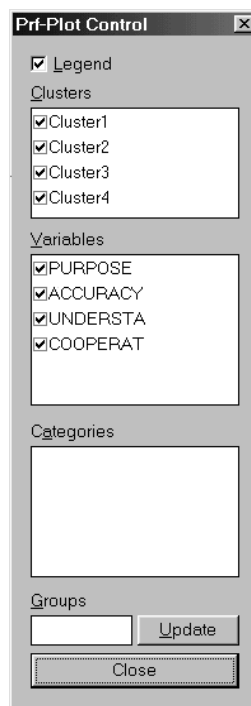


Figure 6-8: Plot Control for Profile Plots

The following plot settings are available for Profile Plots:

**Legend.** When this option is selected, a legend appears at the bottom of the Profile Plot.

**Clusters.** For Cluster models, a line will be drawn for each cluster selected. Those clusters with a checkmark in the checkbox are included in the plot.

**Variables.** Select which variables to include in the plot. Those with a checkmark in the checkbox are included in the plot.

**Categories.** For Nominal Indicators and Nominal Covariates, select which category of a variable to include in the plot. The category currently being plotted is listed in the plot beneath the variable name. To change the category that is plotted, highlight the variable name in the Variables box (the category currently being plotted will appear in the Category box), click the drop-down list to the right of the Categories box and select the category you wish to have plotted.

**Groups:** Select a variable for which range groups appear in the Profile Output, and the number of such

groups appears in the Groups Box. Replace this number with a value less than the total number of categories/ values for this variable and click Update. Certain adjacent categories will automatically be combined and the range groups for the selected variable change to the new number of groups in the Profile output and Profile Plot. Entering the value '0' causes each level to appear as a distinct value. The group option may be used with nominal/ordinal indicators and nominal/numeric covariates.

**Note:** A value of '1' appears in the Groups box for Continuous and Count Indicators when the Plot Control is opened from within the Profile Plot. This indicates that only one value is included in the Profile Output (the mean) and hence entering a different value in the Groups box will not alter the Output for such variables.



## GPROFILE OUTPUT (ADVANCED, OPTIONAL)

If GClasses and/or GCFactors are included in a model, this output file listing is generated. Similar to the Profile output, the top of this file contains the size of each GClass, followed by the probability of being in each cluster for each GClass. This section is followed by the Indicators section where means and (marginal) probabilities associated with each indicator are provided.

The View Menu options (which can be obtained by a right click on the GProfile output), can be used to switch between *Marginal* and *Partial* GProfile Output, or to obtain standard errors.

## PROBMEANS OUTPUT (OPTIONAL)

This table contains aggregated cluster membership probabilities for (ranges of) values of indicators and covariates which are displayed in the Uni-Plot and Tri-Plot.

To view the Probability/Mean table for a selected model, click ProbMeans in the Outline pane. To view a plot, click on the expand '+' icon to the left of ProbMeans to list the type of plots produced for a particular model. Highlight a plot type to view it in the Contents pane.

The screenshot shows the LatentGOLD software window. On the left is the Outline pane with a tree view containing 'GSS82WHITE.SAV', 'Model1 - L² = 257.2604', 'Model2 - L² = 79.5070', 'Model3 - L² = 22.0882', 'Model4 - L² = 6.6098', 'Parameters', 'Profile', 'Prf-Plot', 'ProbMeans' (highlighted), 'Freqs/Residuals', 'Bivariate Residuals', 'Standard Classification', 'Covariate Classification', and 'Iteration Detail'. The main window displays a table with the following data:

	Cluster1	Cluster2	Cluster3	Cluster4
<b>Overall</b>	0.5524	0.2237	0.1544	0.0695
<b>Indicators</b>				
<b>PURPOSE</b>				
<b>good</b>	0.6717	0.2672	0.0397	0.0214
<b>depends</b>	0.2212	0.2096	0.4506	0.1186
<b>waste</b>	0.1328	0.0084	0.5708	0.2879
<b>ACCURACY</b>				
<b>mostly true</b>	0.6676	0.2852	0.0409	0.0063
<b>not true</b>	0.4277	0.1571	0.2773	0.1380
<b>UNDERSTA</b>				
<b>good</b>	0.6719	0.1125	0.1782	0.0373
<b>fair/poor</b>	0.0250	0.7139	0.0496	0.2115
<b>COOPERAT</b>				
<b>interested</b>	0.6304	0.1889	0.1532	0.0276
<b>cooperative</b>	0.1794	0.4221	0.1900	0.2084
<b>impatient/hostile</b>	0.0011	0.3241	0.0295	0.6453

At the bottom of the window, the text 'Cluster3' is visible.

Figure 6-9. ProbMeans Output for Cluster Model

- The first row of the table contains the overall probability of being in a cluster (the size of each cluster), also reported in the first row of numbers in the Profile table.
- The body of the table contains conditional probabilities associated with each category of Nominal and Ordinal indicator variables (these probabilities sum to 100% across rows). For example, in Figure 6-9, for those respondents who responded that surveys serve a 'good' purpose, about 67% are classified as belonging in Cluster 1, 27% in Cluster 2, 04% in Cluster 3, and the remaining 2% in Cluster 4.
- At the bottom of the ProbMeans output, similar information is presented for each covariate.
- For any Indicator with scale type Continuous or Count, or for any numeric covariate, if more than 5 distinct values exists, the probabilities are displayed for 5 grouped ranges. The number of such grouped ranges may be changed (increased or decreased) using the Plot Control.



When the CFactor option is used, ProbMeans contains aggregated posterior CFactor means.

## UNI-PLOT

For a Cluster model, the membership probabilities in the body of the ProbMeans output are plotted to form a Uni-Plot. To view the Uni-Plot, click on the expand/contract icon (+) and highlight Uni-Plot.

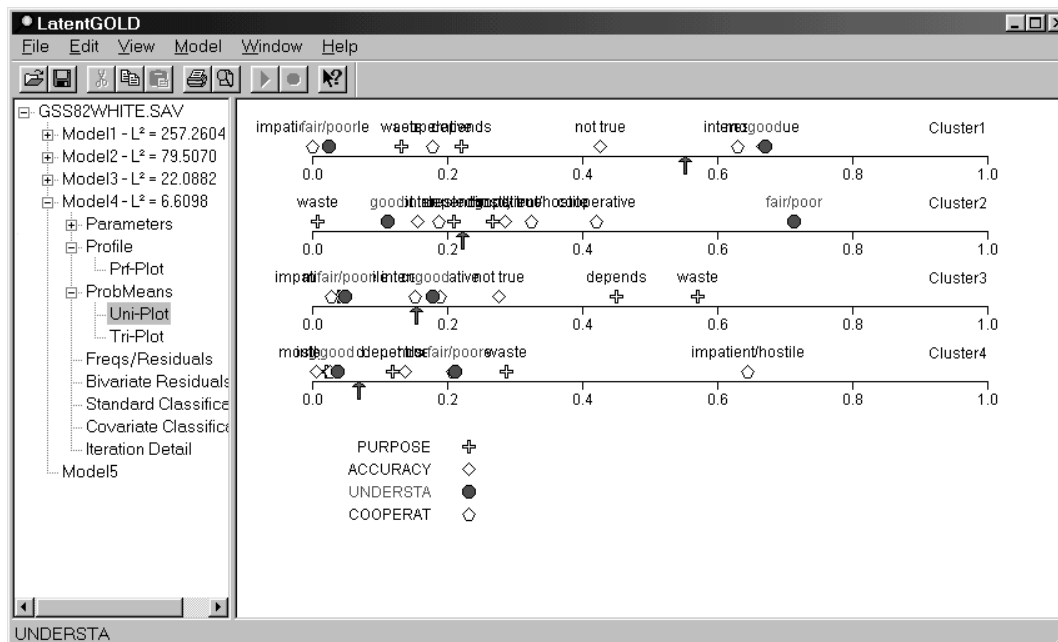


Figure 6-10. ProbMeans Uni-Plot for Cluster Model

- By default, a separate Uni-Plot is created for each cluster. Symbols appear in the plots for each value of each variable specified as an indicator or covariate in the model.
- The ↑ symbol marks the overall probability (the size) for a cluster.
- Click on any variable symbol in the Uni-Plot and the plot label will appear and the status bar will contain a description of the point (variable name and category value).
- Click on any variable name or symbol in the legend and Latent GOLD highlights points associated with that variable.

## To Change Settings for a Uni-Plot

To change the settings for a Uni-Plot, right click (or select Plot Control from the View Menu) within the Contents pane when a Uni-Plot is displayed to open the Plot Control. To change the font for a plot, see Chapter 2.

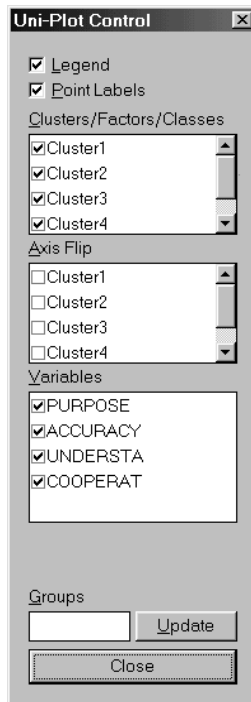


Figure 6-11: Plot Control for Uni-Plots

The following plot settings are available for Uni-Plots:

**Legend.** When this option is selected, a Legend appears at the bottom of the Uni-Plot.

**Point Labels.** When this option is selected, category labels for each variable are listed on the Uni-Plot next to the variable symbol.

**Clusters/DFactors/Classes.** Select which Clusters to include in the Uni-Plots. For each cluster selected (a checkmark in the checkbox) a separate Uni-Plot is displayed. By default, all clusters are selected.

**Axis Flip.** To flip (reverse) the axis for a Uni-Plot, select the corresponding cluster name. By default, the cluster probabilities range is from 0 to 1 (increasing). Selecting Axis flip for a cluster will reverse the axis to range from 1 to 0 (decreasing).

**Variables.** Select which indicators/covariates to include in the Uni-Plots. Selected variables are indicated by a checkmark in the checkbox. By default, the Uni-Plots contain the first 16 indicators/covariates included in the model. Additional variables may be selected from the Plot Control (using a control-click or a shift click to select more than one at a time) and added or removed from the plot with a single click in the box to the left of the variable names.

**Groups.** Select an Indicator with scale type Continuous or Count, or a numeric Covariate for which range groups appear in the ProbMeans Output, and the number of such groups appears in the Groups Box. Replace this number with an alternative number of groups and click Update. The range groups for the selected Covariate change to the new number of groups in the ProbMeans output, the UniPlot and Triplot. Entering the value '0' causes each level to appear as a distinct value.

Groups may be used more generally to produce range groups for the categories of any Nominal or Ordinal Indicator or any Covariate. Select any of these variables and enter a value less than the total number of categories/ values for this variable in the Groups box and click Update. Certain adjacent categories will automatically be combined and the output tables and plots are updated to present results for the associated grouped ranges.

## TRI-PLOT

For Cluster models, the cluster membership probabilities in the body of the ProbMeans output table are plotted to form a Tri-Plot. To view the Tri-Plot, click on the expand/contract icon (+) and highlight Tri-Plot. (Note: No Tri-Plot is produced for a 1- Cluster model; for a 2-Cluster model, the Tri-Plot reduces to the Uni-Plot.)

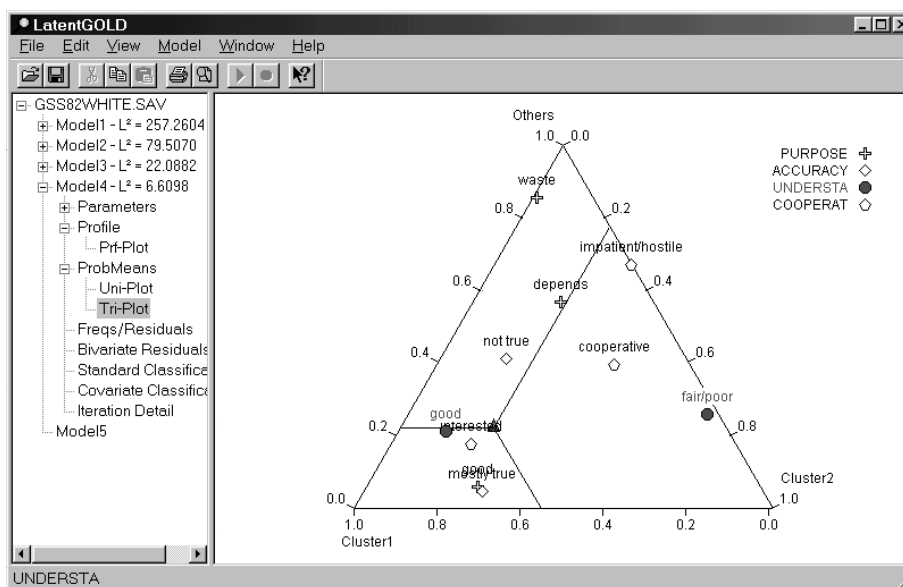


Figure 6-12. Tri-Plot for Cluster Model

- By default, Vertex A (left-most base vertex) is labeled 'Cluster 1', Vertex B (right-most base vertex) 'Cluster 2', and the third Vertex (the top point of the triangle) represents the aggregate of all other clusters. For a 3-Cluster model, by default, the third vertex will represent Cluster 3 and is labeled 'Cluster 3'. For a 4-or-more Cluster model, the third vertex is labeled 'Others'. For a 2-Cluster model, the cluster 3 membership probability is 0 and the Tri-Plot reduces to the Uni-Plot.
- The ▲ symbol marks the overall probabilities for the 3 clusters associated with the vertices. It represents the centroid of the triangle.
- Click on any variable symbol in the Tri-Plot and 1) the status bar contains a description of the point (variable name and category, cluster probabilities) 2) the category label appears next to that point on the plot and 3) lines emanate from that point to each side of the triangle, intersecting the side at the corresponding cluster probabilities value.
- Click on any variable symbol or name in the legend and all the symbols for that variable will be highlighted and their category labels listed in the Tri-Plot.

## To Change Settings for a Tri-Plot

To change the settings for a Tri-Plot, right click (or select Plot Control from the View Menu) within the Contents pane when a Tri-Plot is displayed to open the Plot Control dialog box. To change the font for a plot, see Chapter 2.

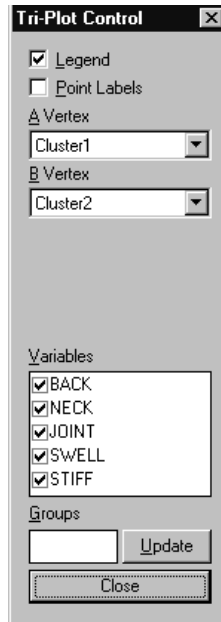


Figure 6-13: Plot Control for Tri-Plots

The following plot settings are available for Tri-Plots:

**Legend.** When this option is selected, a Legend appears to the right of the Tri-Plot.

**Point Labels.** When this option is selected, category labels for each selected variable are listed on the Tri-Plot next to the variable symbol.

**Vertices.** Latent GOLD allows you to select the base vertices in the Tri-Plot. The top vertex corresponds to the aggregate of the remaining clusters.

**A vertex.** The cluster currently used as the A vertex is listed in the drop down box. To select a different cluster, click on the down arrow to the right of the vertex box. A drop list containing all clusters will appear. Select the cluster to use as the A vertex.

**B vertex.** The cluster currently used as the B vertex is listed in the drop down box. To select a different cluster, click on the down arrow to the right of the vertex box. A drop list containing all clusters will appear. Select the cluster to use as the B vertex.

**Variables.** Select which variables to include in the Tri-Plot. Those with a checkmark in the checkbox are included in the plot. By default, the Tri-Plot contains the first 16 indicators/covariates that were input as part of the model. Additional variables may be selected from the Plot Control (using a control-click or a shift click to select more than one at a time) and added or removed from the plot with a single click in the box to the left of the variable names.

**Groups.** Select an Indicator with scale type Continuous or Count, or a numeric Covariate for which range groups appear in the ProbMeans Output, and the number of such groups appears in the Groups Box. Replace this number with an alternative number of groups and click Update. The range groups for the selected Covariate change to the new number of groups in the ProbMeans output, the UniPlot and Triplot. Entering the value '0' causes each level to appear as a distinct value.

Groups may be used more generally to produce range groups for the categories of any Nominal or Ordinal Indicator or any Covariate. Select any of these variables and enter a value less than the total number of categories/values for this variable in the Groups box and click Update. Certain adjacent categories will automatically be combined and the output tables and plots are updated to present results for the associated grouped ranges.



*For more detailed information about the ProbMeans Output, see section 7.4 of Technical Guide.*

## FREQUENCIES AND RESIDUALS (OPTIONAL)

This output appears only if the Frequencies/ Residuals option was selected in the Output Tab before model estimation.

This output is not available if one or more indicators is continuous or when the multilevel option (Advanced) is used.

Click Freqs/Residuals to view a table containing the observed and expected frequencies along with the standardized residuals for a model.

The screenshot shows the LatentGOLD software window with the 'Freqs/Residuals' tab selected. The table displays data for a cluster model, including observed and estimated frequencies, standardized residuals, and Cook's D values for various indicators and purposes.

	accuracy	cooperat	understa	purpose	Observed	Estimated	StdResid	Cook's D
Model 1 - L <sup>2</sup> = 40	mostly true	interested	Good	GOOD PURPOSE	535.0000	523.2405	0.5141	0.0026
	mostly true	interested	Good	DEPENDS	29.0000	45.4186	-2.4362	0.0220
	mostly true	interested	Good	WASTE OF TIME AND \$	32.0000	30.5503	0.2623	0.0657
	mostly true	interested	Fair/Poor	GOOD PURPOSE	105.0000	107.5559	-0.2464	0.0154
	mostly true	interested	Fair/Poor	DEPENDS	9.0000	7.7654	0.4430	0.0743
	mostly true	interested	Fair/Poor	WASTE OF TIME AND \$	4.0000	4.6829	-0.3156	0.1796

Figure 6-14. Freqs/Residuals Output for Cluster.



*For more detailed information about the Frequencies Output, see section 7.5 of Technical Guide.*

## BIVARIATE RESIDUALS (OPTIONAL)

This output only appears if the Bivariate Residuals option was selected in the Output Tab before model estimation

Click Bivariate Residuals to view a table containing the bivariate residuals (BVRs) for a model. A sorted view of this output is provided in the Residuals Tab for an estimated model. Double click on the name of an estimated

model to open the Analysis Dialog Box (or select an estimated model and from the Model Menu select Cluster) and open the Residuals Tab. The BVRs appear sorted from high to low. The Residuals Tab may be used to specify direct effect parameters in a model. Normally, a direct effect parameter associated with a large BVR would be included in a model to improve the fit of the model without the direct effect.

In addition to the  $L^2$  criterion, finding no significant residuals is another indication that a model provides a good fit to the data. In general, BVRs larger than 3.84 identify correlations between the associated variable pairs that have not been adequately explained by the model. (For 1 degree of freedom effects, bivariate residuals larger than 3.84 indicate statistical significance at the .05 level.)

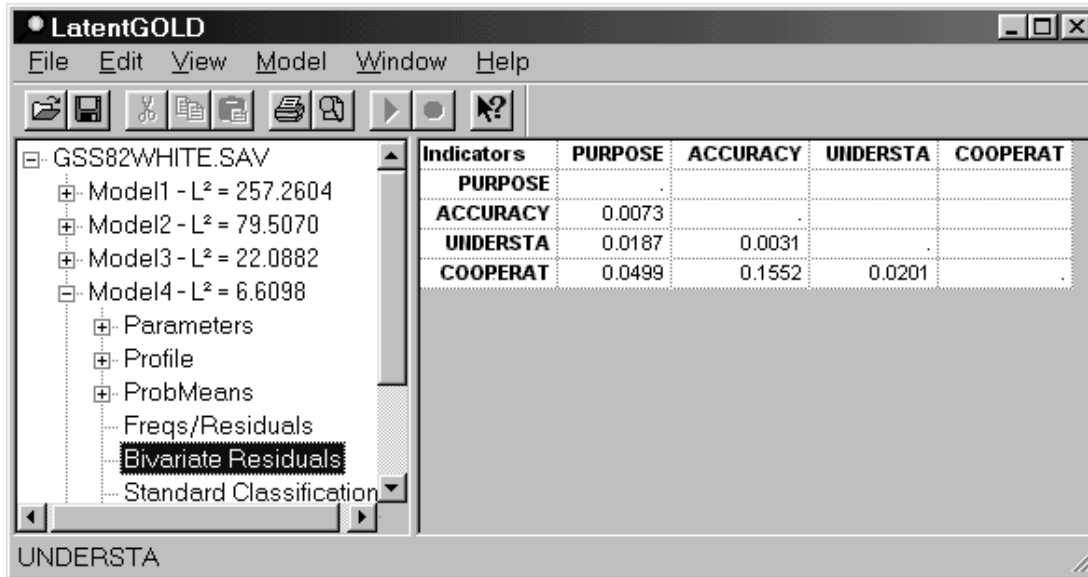


Figure 6-15. Bivariate Residuals Output in Cluster Model



For further details about Bivariate Residuals, see section 7.6 of Technical Guide.

### STANDARD CLASSIFICATION (OPTIONAL)

This output only appears if Standard Classification option was selected in the Output Tab before model estimation

Click Standard Classification to view a table containing the posterior membership probabilities and other classification information for a model..

For each observed response pattern, the classification output contains the frequency of occurrence ("ObsFreq"), the cluster for which the posterior membership probability is highest ('Modal'), the associated posterior membership probabilities of belonging to each cluster ('Cluster 1', 'Cluster 2', ...)

PURPOSE	ACCURACY	UNDERSTA	COOPERAT	ObsFreq	Modal	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
good	mostly true	good	interested	419.0000	1	0.8774	0.1128	0.0095	0.0003	
good	mostly true	good	cooperative	35.0000	2	0.4842	0.4882	0.0227	0.0049	
good	mostly true	good	impatient/hostile	2.0000	2	0.0088	0.9442	0.0089	0.0381	
good	mostly true	fair/poor	interested	71.0000	2	0.0438	0.9502	0.0035	0.0025	
good	mostly true	fair/poor	cooperative	25.0000	2	0.0058	0.9834	0.0020	0.0088	
good	mostly true	fair/poor	impatient/hostile	5.0000	2	0.0001	0.9649	0.0004	0.0346	
good	not true	good	interested	270.0000	1	0.8087	0.0894	0.0917	0.0102	
good	not true	good	cooperative	25.0000	1	0.3711	0.3217	0.1832	0.1239	
good	not true	good	impatient/hostile	4.0000	4	0.0041	0.3734	0.0432	0.5794	
good	not true	fair/poor	interested	42.0000	2	0.0446	0.8328	0.0376	0.0849	
good	not true	fair/poor	cooperative	16.0000	2	0.0050	0.7273	0.0182	0.2494	
good	not true	fair/poor	impatient/hostile	5.0000	4	0.0000	0.4190	0.0021	0.5788	
depends	mostly true	good	interested	23.0000	1	0.5946	0.1820	0.2196	0.0038	
depends	mostly true	good	cooperative	4.0000	2	0.1931	0.4637	0.3106	0.0326	
depends	mostly true	good	impatient/hostile	1.0000	2	0.0028	0.7027	0.0957	0.1989	
depends	mostly true	fair/poor	interested	6.0000	2	0.0177	0.9165	0.0487	0.0170	
depends	mostly true	fair/poor	cooperative	2.0000	2	0.0023	0.9136	0.0270	0.0572	
depends	not true	good	interested	43.0000	3	0.1865	0.0491	0.7249	0.0395	
depends	not true	good	cooperative	9.0000	3	0.0391	0.0807	0.6618	0.2183	
depends	not true	good	impatient/hostile	2.0000	4	0.0003	0.0737	0.1229	0.8031	
depends	not true	fair/poor	interested	9.0000	2	0.0094	0.4187	0.2722	0.2997	

Figure 6-16. Standard Classification for Cluster Model



If 1 or more CFactors have been included in the model, the associated CFactor scores ('Cfac1', 'Cfac2', 'Cfac3'), appear in the right-most portion of the standard classification output.

If 2 or more GClasses have been included in the model, posterior membership probabilities associated with these group classes are given in additional columns (GClass1, GClass2, ...) that appear in the right-most portion of the standard classification output.

If 1 or more GCFactors have been included in the model, the associated factor scores associated with these group level CFactors ('Gcfac1', 'Gcfac2', 'Gcfac3'), appear in the right-most portion of the standard classification output.

## COVARIATE CLASSIFICATION (OPTIONAL)

This output only appears if the Covariate Classification option was selected in the Output Tab before model estimation and 1 or more active covariates have been included in the model.

This table contains the estimated probabilities in the multinomial logistic regression model for the clusters,  $P(x|z)$ , as well as the modal cluster assignments based on these probabilities. The table has one row for each unique covariate pattern..



If 2 or more GClasses have been included in the model, estimated group-level class membership probabilities given group-level covariates are provided in additional columns (GClass1, GClass2, ...) that appear in the right-most portion of the covariate classification output.



**For more detailed information about the Classification Output, see sections 7.8 and 12.7 of Technical Guide.**

## DFactor

### GENERAL INFORMATION

The DFactor Module is used to estimate a type of restricted LC Cluster Model known as an DFactor model. DFactor models contain 1 or more discrete factors (DFactors), each of which may contain 2 or more ordered levels (ordered latent classes). A 1-DFactor model containing  $K > 2$  levels differs from an order-restricted LC Cluster analysis with  $K$  ordered clusters in that the levels of a DFactor are assumed to be equally spaced. For further details see sections 3.7 and 4 of Technical Guide.

When 2 or more DFactors are included in a model, DFactors are ordered according to the R-squared (indicating how well the model predicts the DFactor score). By default, the DFactor having the highest R-squared is listed first (DFactor 1). The levels of the DFactors are ordered such that the lowest level always contains at least as many cases as the highest level.

**Note:** The ordering of DFactors may not be according to the standard  $R^2$  (reliability) if any of the user options in the Model Tab are applied to increase the number of DFactor levels or restrict effects within one or more classes to zero.

In addition to the Model Summary Output file, additional model output files are generated following estimation of the model. The specific output sections that are available are those that appear as active in the Output Sections portion of the Output Tab. For DFactor models, these are the first 7 model output files listed, plus the last file (Iteration Detail).

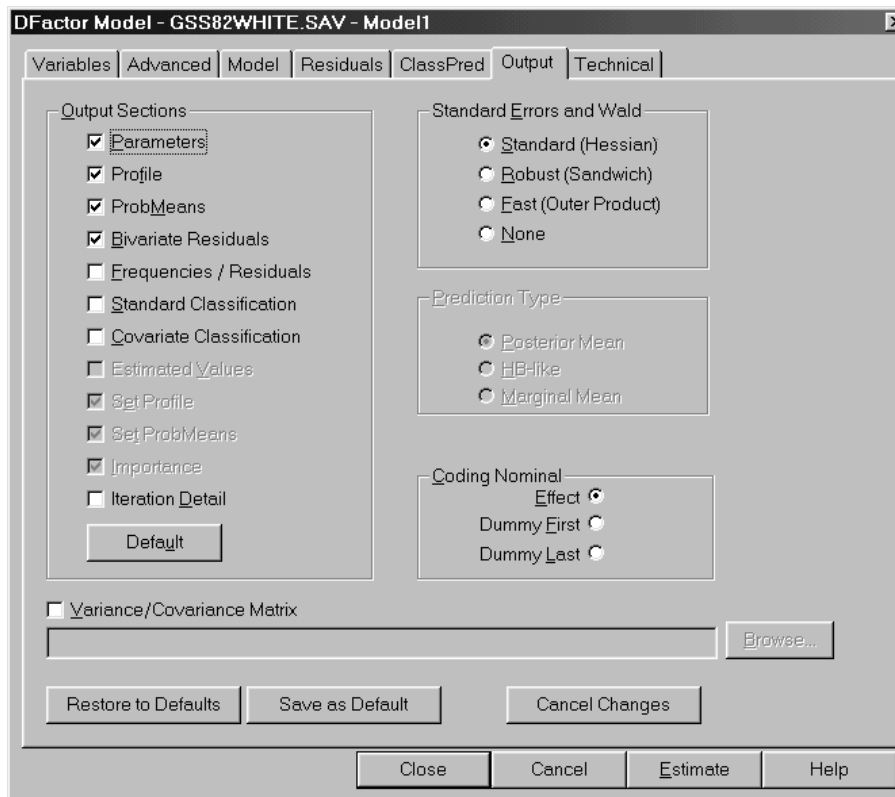


Figure 6-17. Output Tab for DFactor Models

Prior to estimating a DFactor model, click in the active check boxes to the left of the Output Sections to include or exclude that Section from being part of the model output files that are generated following the model estimation. A detailed description of each of the Model Summary Output as well as each specific Model Output Section file that can be output appears below.

## DFACTOR MODEL SUMMARY OUTPUT

Following the estimation of a DFactor model, if you click on the Model name in the Outline pane, the Model Summary Output for this model will be displayed the Contents Pane. The information presented is similar to that reported for the LC Cluster Models.

The general section of the Model Summary Output contains the following items:

**Type of Model.** For example '2-DFactor Model' describes a DFactor model containing 2 discrete factors.

**Warning messages** (generally not present). Estimation Warning message(s) appear here to indicate that boundary solutions, identification or convergence problems were encountered during the estimation of the model.

**Model Paused** (if you paused the model). If you Paused the model prior to estimation being completed this message appears to alert you that the output files should be viewed as preliminary.

**Number of cases.** This is the number of cases used in model estimation. This number may be less than the original number of cases on the data file if missing cases have been excluded.

**A Number of groups:** In multilevel latent class models, the program reports the number of groups used in model estimation.

**Number of parameters.** This is the number of distinct parameters estimated.

**Random Seed.** The seed required to reproduce this model.

**Best Start Seed.** The single best seed that can reproduce this model more quickly using the number of starting sets =0. This is the seed that is automatically inserted in a saved definition (.lgf) file.

**A Design Effect.** When the Survey option is used, the program reports the generalized design effect.

2-Factor Model		
	Factor1(2)	Factor2(2)
Number of cases	1202	
Number of parameters (Npar)	16	
Random Seed	1147811	
Best Start Seed	1424465	
<b>Chi-squared Statistics</b>		
Degrees of freedom (df)	19	p-value
L-squared ( $L^2$ )	14.2853	0.77
X-squared	11.5619	0.90
Cressie-Read	12.0806	0.88
BIC (based on $L^2$ )	-120.4578	
AIC (based on $L^2$ )	-23.7147	
AIC3 (based on $L^2$ )	-42.7147	
CAIC (based on $L^2$ )	-139.4578	
<b>Log-likelihood Statistics</b>		
Log-likelihood (LL)	-2750.7421	
Log-prior	-5.0898	
Log-posterior	-2755.8318	
BIC (based on LL)	5614.9520	
AIC (based on LL)	5533.4841	
AIC3 (based on LL)	5549.4841	
CAIC (based on LL)	5630.9520	
<b>Classification Statistics</b>		
	Factor1	Factor2
Classification errors	0.1284	0.1155
Reduction of errors (Lambda)	0.5715	0.5184
Entropy R-squared	0.4875	0.4272
Standard R-squared	0.5444	0.4980
Classification log-likelihood	-3500.1696	
AWE	7275.2750	

Figure 6-18. Model Summary Output for DFactor Model

## Chi-squared Statistics

This section lists various chi-square based statistics related to model fit.

If the scale type for one or more indicators in a model has been set to 'continuous' or when the multilevel option (Advanced) is used, no chi-squared statistics are available and this section is not displayed.

The information reported:

**Degrees of freedom (df).** The degrees of freedom for the current model.

**L-squared ( $L^2$ ).** The likelihood-ratio goodness-of-fit value for the current model. If the bootstrap p-value for the  $L^2$  statistic has been requested, the results will be displayed here.

**X-squared and Cressie-Read.** These are alternatives to  $L^2$  that should yield a similar p-value according to large sample theory if the model specified is valid and the data is not sparse.

**BIC, AIC and CAIC (based on  $L^2$ ).** In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model.

**Dissimilarity Index.** A descriptive measure indicating how much the observed and estimated cell

frequencies differ from one another. It indicates the proportion of the sample that needs to be moved to another cell to get a perfect fit.

## Log-likelihood Statistics

**Log-likelihood (LL).** If the conditional bootstrap (bootstrap -2LL Diff) has been requested, the results will be displayed here.

**Log-prior.** This is the term in the function maximized in the parameter estimation that is associated with the Bayes constants. This term equals 0 if all Bayes constants are set to 0.

**Log-posterior.** This is the function that is maximized in the parameter estimation. The value of the log-posterior function is obtained as the sum of the log-likelihood and log-prior values.

**BIC, AIC, AIC3 and CAIC (based on LL)** - these statistics (information criteria) weight fit and parsimony by adjusting the LL to account for the number of parameters in the model. The lower the value, the better the model. For example, according to the BIC values shown in Figure 6-18, the 3-class model is preferred over the 1-class, 2-class and 4-class models.

## Classification Statistics

This information can be used to assess how well the model classifies cases into clusters. The statistics reported are:

**Classification Errors.** When classification of cases is based on modal assignment (to the DFactor level having the highest membership probability), the proportion of cases that are expected to be misclassified is reported by this statistic. The closer this value is to 0 the better.

**Reduction of Errors (lambda), Entropy R-squared and Standard R-squared.** These statistics indicate how well the model predicts class memberships or DFactor scores. The closer these values are to 1 the better the predictions.

**Classification log-likelihood.** Log-likelihood value under the assumption that the true class membership is known.

**AWE.** Similar to BIC, but also takes classification performance into account.

**Classification Table.** The Classification Table cross-tabulates modal and probabilistic class assignments.



**Standard R-squared** is reported for CFactors and GCFactors. For GClasses, as for Clusters, Classification Errors, Reduction of Errors, Entropy R-squared, and Standard R-squared values are reported.



*For more detailed information about the classification statistics including the formulae for each, see section 7.1.3 of Technical Guide.*

### Covariate Classification Statistics

If one or more active covariates is included in the model, additional statistics are provided as above but these are now based only on the active covariates. The statistics reported are:

#### Classification errors, Reduction of Errors (lambda), Entropy R-squared and Standard R-squared



**Standard R-squared** is reported for CFactors and GCFactors. For GClasses, as for Clusters, Classification Errors, Reduction of Errors, Entropy R-squared, and Standard R-squared values are reported.



*For more detailed information about the classification statistics including the formulae for each, see section 7.1.4 of Technical Guide.*

### Variable Detail

This section contains details about the variables entered into the model including variable names, scale types, number of categories and category labels and scores (if used).

### PARAMETERS OUTPUT (OPTIONAL)

For any estimated DFactor model, a Parameters file listing is generated by default. It contains estimates of parameters in the Model for Indicators (betas, sigmas), followed by the estimates in the Model for DFactors (gammas), as well as measures of significance for these estimates and the model  $R^2$  for each indicator. The parameter estimates are arranged in separate columns corresponding to each DFactor included in the model.



If CFactors have been included in the model, estimates and related information for additional parameters (lambdas) are included in the Models for Indicators output.



If GClasses and/or GCFactors have been included in the model, estimates and related information for additional parameters are included in the Models for Indicators and/or Model for DFactors section of the output, and a separate section called Model for GClasses appears beneath this section.

Models for Indicators							
	Factor1	Wald	p-value	Factor2	Wald	p-value	R²
PURPOSE	2.5163	14.5193	0.00014	0.5760	0.9819	0.32	0.4670
ACCURACY							
mostly true	-1.1807	37.5215	9.0e-10	-0.1750	0.4814	0.49	0.2213
not true	1.1807			0.1750			
UNDERSTA							
good	0.2315	0.2448	0.62	-1.7988	2.0306	0.15	0.4146
fair/poor	-0.2315			1.7988			
COOPERAT							
	1.0344	5.8527	0.016	2.1478	4.7648	0.029	0.2540
Intercepts	Overall	Wald	p-value				
PURPOSE							
good	2.7286	48.0558	3.7e-11				
depends	-0.4982						
waste	-2.2305						
ACCURACY							
mostly true	0.4034	4.1121	0.043				
not true	-0.4034						

Figure 6-19. Parameters Output for DFactor Model

The Models for Indicators section contains:

**Betas.** For DFactor models, effects are organized separately for each DFactor. The most important betas are that ones that indicate the strength of the effects of the clusters on the indicators. Other betas are the indicator intercepts, the direct associations between categorical (Nominal/Ordinal) indicators, and direct effects of covariates on indicators.

**Sigmas.** Error variances and covariances for continuous indicators.

The Models for DFactors section contains:

**Gammas.** Parameters of the logit model used to predict each DFactor as a function of the covariates. These parameters include the intercepts, the associations between the DFactors, as well as the effects of each covariate on each DFactor.

## Viewing Wald Statistics and Standard Errors

By default, Wald statistics are provided in the output to assess the statistical significance of a set of parameter estimates. For example, for each indicator the Wald statistic tests the restriction that each estimate in the set of beta parameter estimates associated with that indicator equals zero. A non-significant p-value associated with this Wald statistic means that the indicator does not discriminate between the clusters in a statistically significant way. For variables specified as Nominal, the set includes parameters for each category of the variable.

To view standard errors or related statistics associated with the parameter estimates, simply right click on the parameters output and select/deselect the items that you want to appear. Alternatively, you can click on the appropriate item in the View Menu. In Parameters, you can suppress the Wald statistics, or replace the Wald with standard errors, a Z Statistic, or both standard errors and the associated Z Statistic for each parameter estimate.

### Parameters Output Subcategories

Clicking on the + to the left of Parameters makes visible the Parameters Output Subcategories, which provide additional information regarding Parameters Output. The subcategories are Loadings, Correlations, and Error Correlations.

**Loadings.** DFactor loadings for the indicators are given in separate columns for each DFactor, labeled 'DFactor1', DFactor2', ... followed by a column labeled 'R<sup>2</sup>' representing the communality of the indicator. The quantities reported in the DFactor columns are comparable to loadings (standardized regression coefficients) in traditional linear factor analysis. The R<sup>2</sup> is the same as reported in the Parameters Output.



Additional columns appear immediately to the right of the 'DFactors' columns when CFactors, GClasses and/or GCFactors are included in the model. The quantities reported in the columns to the left of the R<sup>2</sup> represent the decomposition of the R<sup>2</sup> into components associated with each of these column effects. For further details, see the Technical Guide.

**Correlations.** These are the correlation between the DFactors and the correlation between DFactors and indicators. Correlations between the DFactors are equal to 0 when neither DFactor associations nor Covariates are included in the model. In this case (and only in this case) DFactor-Indicator Correlations will be identical to the corresponding Loadings.

**Error Correlations** (available only when one or more indicators are continuous) - The Error Correlations output provides the estimates of the within-class correlations between continuous indicator variables. Often, these correlations are easier to interpreted than covariances, which are the actual model parameters. (Note that the off diagonal entries are all 0 unless direct effects between continuous indicators have been included in the model)



*For more detailed information about the Parameters Output, see section 7.2 of Technical Guide.*

### PROFILE OUTPUT (OPTIONAL)

To view the profile table for a selected model, click Profile in the Outline pane. The Profile table contains conditional response probabilities and/or means associated with each Indicator. Means are computed based on the category scores, using the conditional probabilities as weights. For Covariates the Profile table reports aggregated classification probabilities which are rescaled to sum to 1 within DFactor levels.

	Factor1	Factor2
	Level1	Level2
<b>Factor Level Size</b>	0.7003	0.2997
<b>Indicators</b>		
<b>PURPOSE</b>		
good	0.9457	0.3414
depends	0.0443	0.1852
waste	0.0100	0.4734
Mean	1.0643	2.1320
<b>ACCURACY</b>		
mostly true	0.6724	0.1637
not true	0.3276	0.8363
<b>UNDERSTA</b>		
good	0.8032	0.8434
fair/poor	0.1968	0.1566
<b>COOPERAT</b>		
interested	0.8801	0.7419
cooperative	0.1053	0.1952
impatient/hostile	0.0146	0.0629
Mean	1.1345	1.3210

Figure 6-20. Profile Output for DFactor Model

- By default, columns of the Profile table display information separately for each level of each DFactor
- At the top of this output table, the size of each level of each DFactor is indicated (for each DFactor these numbers sum to 100%). For example, regarding DFactor #1, 70% are in level 1, the remaining 30% in level 2 of this DFactor.
- The body of the table contains (marginal) conditional probabilities that show how the DFactor levels are related to the indicator variables and covariates. Within each level, these probabilities sum to 1. For ordinal indicators and numeric covariates, cluster-specific means are also reported beneath the probabilities.
- For indicators specified as Continuous or Count, the body of the table contains means instead of probabilities.
- Optionally, when 2 or more DFactors are included in a model, the columns may be changed to correspond to joint levels of the DFactors (see the section on Joint Profile below). The joint view makes clear that the DFactor model is an LC Cluster model. For example, Figure 6-21 below shows the 4 clusters corresponding to the joint DFactor levels (1,1), (1,2), (2,1), and (2,2).

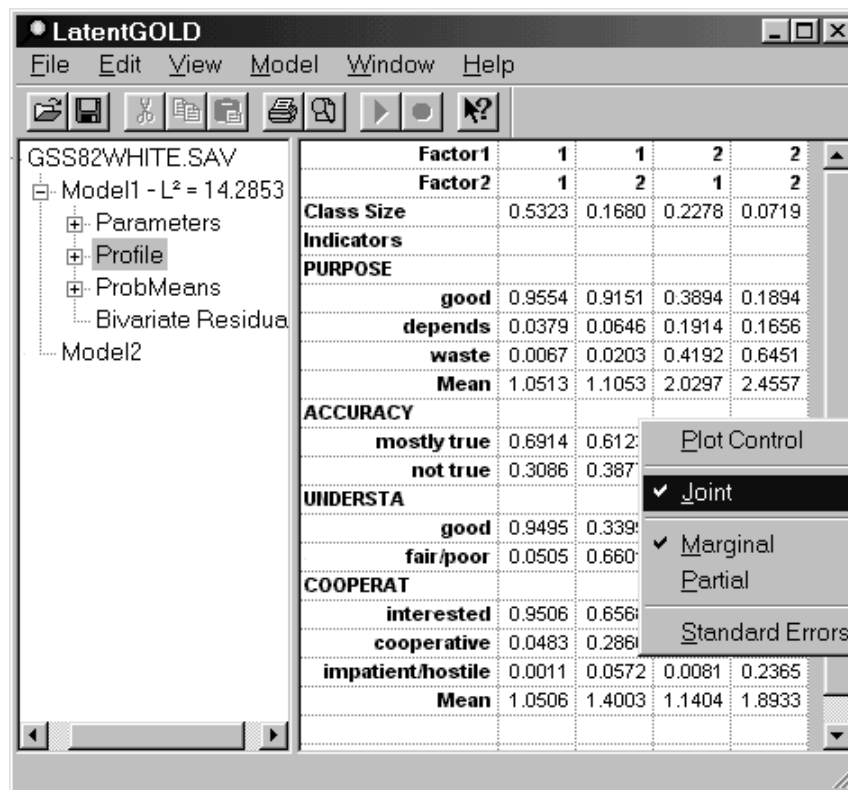
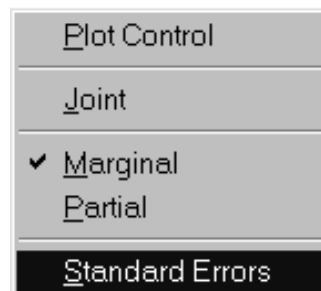


Figure 6-21. Joint Profile Output for DFactor Model



The View Menu options (which can be obtained by a right click on the Profile output), can be used to obtain the Joint Profile Output, to switch between Marginal and Partial Profile Output, or to obtain standard errors.

## Partial Profile

Latent GOLD 4.0 reports not only marginal but also partial probabilities/means for indicators in the Profile Table Output. These two differ in models with more than a single DFactor or if local dependencies or effects of covariates on indicators are included using the Residuals Tab. As explained in section 7.3 of the Technical Guide, partial probabilities/means are obtained by conditioning on a certain value (mean) of the other variables involved in the model for the indicator concerned. The marginal probabilities/means, on the other hand, are obtained by collapsing over the categories of these "other" variables. Both have their specific advantages: the Marginal Profile output is somewhat easier to interpret and is displayed by default. Partial Profile reflects somewhat better the strength of the effects of the cluster variable on the indicators.

## Joint Profile

A *Joint Profile* amounts to treating each DFactor combination as a cluster. For example, with 3 DFactors having 2 levels each, we have  $2 \times 2 \times 2 = 8$  clusters, represented as (1,1,1), (1,1,2), (1,2,1), (1,2,2), (2,1,1), (2,1,2), (2,2,1), and (2,2,2).

The Joint Profile corresponds to the actual model probabilities if no direct effects are included in the model. Joint Profile probabilities can be Partial or Marginal (the default). As in Cluster, these only differ if direct effects are included in a model (direct effects are added to a model using the Residuals Tab - see Chapter 5 - Residuals Tab).

## PROFILE PLOT (FOR INDICATORS AND COVARIATES)

To view the Profile Plot, click the '+' icon to expand the Profile output and highlight Prf-Plot. The Profile Plot is constructed from the conditional probabilities for the nominal variables and from the means for the other indicators and covariates as displayed in the columns of the Profile table. Specific columns (DFactor levels) are selected for display using the Plot Control pop-up menu. The quantities associated with the selected columns are plotted and connected to form a line graph.

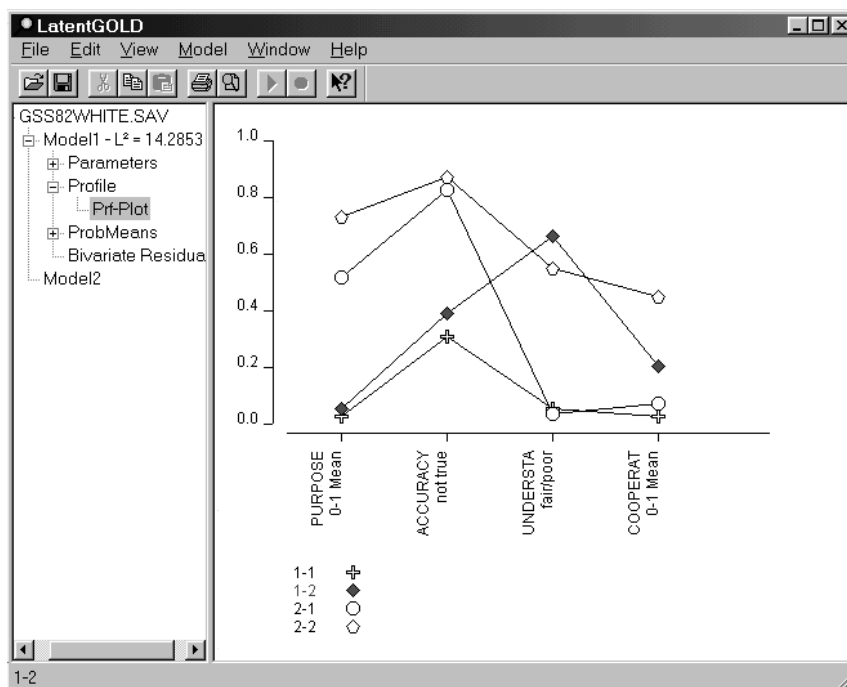


Figure 6-22. Profile Plot for DFactor Model

For the scale types **ordinal**, **continuous**, **count**, and **numeric** covariate, prior to plotting the class-specific means, they are re-scaled to always lie within the 0-1 range. Scaling of these "0-1 Means" is accomplished by subtracting the lowest observed value from the DFactor-level-specific means and dividing the results by the range, which is simply the difference between the highest and the lowest observed value. The advantage of such scaling is that these numbers can be depicted on the same scale as the corresponding probabilities for **nominal** variables.

For nominal variables containing more than 2 categories, all categories are displayed simultaneously. For dichotomous variables specified as nominal, by default only the last category is displayed. The Plot Control can

be used to display the first or both categories, as well as to specify the columns, variables, and selected categories to appear in the display.

By default, all levels associated with DFactor #1 are displayed, or when the Joint Profile table view has been selected, all joint factor levels are displayed.

Each column is associated with either a DFactor level, or a joint DFactor level. When the Profile Table has been changed to the Joint Profile view, the associated probabilities (means) within each selected joint DFactor level displayed graphically in the Profile Plot and connected to form a line graph. the profile information is displayed.

When the contents of the Profile Table is changed from the default view to display Partial (instead of Marginal) probabilities, or to the Joint Profile view, the points plotted change to reflect the current contents of the Profile Table.

- A line is displayed for each level of each selected DFactor (or each joint DFactor level if the Joint Profile has been selected).
- Click on any variable symbol in the Profile Plot and the status bar describes it (variable name, DFactor level number, and value).
- Click on any DFactor level name or symbol in the legend and Latent GOLD highlights all the symbols that refer to that DFactor level.

By default, only the first 8 variables are displayed in the profile plot. Additional variables may be selected from the Plot Control (using a control-click or a shift click to select more than one at a time) and added or removed from the plot with a single click in the box to the left of the variable names.

### To Change Settings for a Profile Plot

To change the settings for a Profile Plot, right click (or select Plot Control from the View Menu) within the Contents pane when a Profile Plot is displayed to open the Plot Control dialog box. The appearance of the Profile Plot Control is somewhat different depending upon whether the Joint Profile view has been selected.

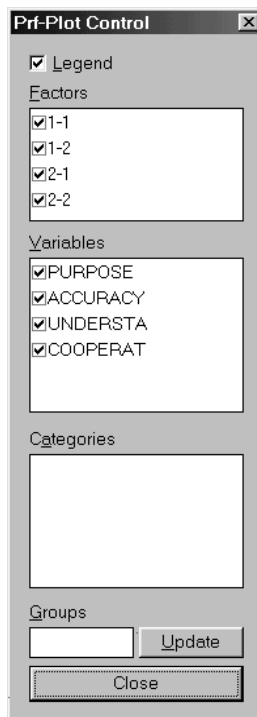


Figure 6-23: Plot Control for Profile Plots

The following plot settings are available for Profile Plots:

**Legend.** When this option is selected, a legend appears at the bottom of the Profile Plot.

**DFactors.** For DFactor models, a line will be drawn for each level of each DFactor selected. Those DFactors with a checkmark in the checkbox are included in the plot.

**Variables.** Select which variables to include in the plot. Those with a checkmark in the checkbox are included in the plot.

**Categories.** Select which category of a variable to include in the plot. The category currently being plotted is listed in the plot beneath the variable name. To change the category that is plotted, highlight the variable name in the Variables box (the category currently being plotted will appear in the Category box), click the drop-down list to the right of the Categories box and select the category you wish to have plotted.

**Groups:** Click Update once you have specified a new number of groups.



**For more detailed information about the Profile Output, see section 7.3 of Technical Guide.**



## **GPROFILE OUTPUT (ADVANCED, OPTIONAL)**

If GClasses and/or GCFactors are included in a model, this output file listing is generated. Similar to the Profile output, the top of this file contains the size of each GClass, followed by the probability of being in each DFactor level for each GClass. This section is followed by the Indicators section where (marginal) means and (marginal) probabilities associated with each indicator are provided.

The View Menu options (which can be obtained by a right click on the *GProfile* output), can be used to switch between *Marginal* and *Partial* GProfile Output, or to obtain standard errors.

## **PROBMEANS OUTPUT (OPTIONAL)**

For DFactor models, this table contains aggregated DFactor means for (ranges of) indicator and covariate values which are displayed in the Uni-Plot and Bi-Plot (when there are 2 or more DFactors).

To view the Probability/Mean table for a selected model, click ProbMeans in the Outline pane. To view a plot (not available for Regression models), click on the expand '+' icon to the left of ProbMeans to list the type of plots produced for a particular model. Highlight a plot type to view it in the Contents pane.

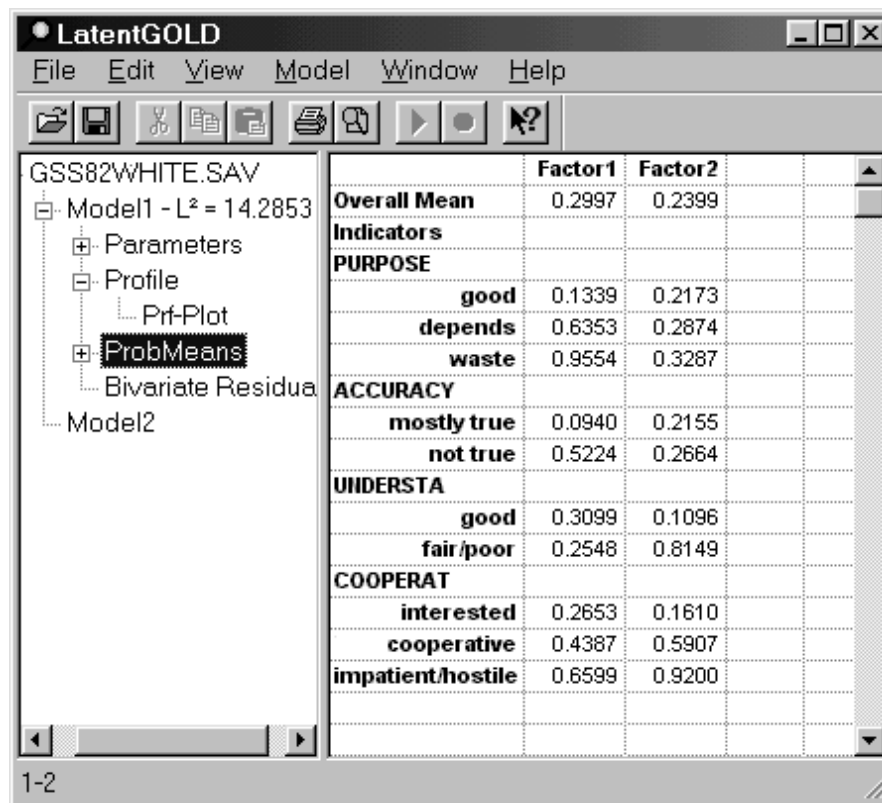


Figure 6-24. ProbMeans Output for DFactor Model

For a DFactor model, the first row contains the mean for each DFactor. For example, if DFactor 1 contained 3 levels, this value is the mean of those three levels computed using equidistant (uniform) scores between 0 and 1 (the first level of the DFactor is scored 0, the last level 1). In the case of a dichotomous DFactor the mean equals the probability of being at level 2. The body of the table contains (partial) DFactor means that show how the indicator variables are related to the DFactors.



When the CFactor option is used, ProbMeans contains aggregated posterior CFactor means.

## UNI-PLOT

For a DFactor model, the membership probabilities in the body of the ProbMeans output are plotted to form a Uni-Plot. To view the Uni-Plot, click on the expand/contract icon (+/-) to list the ProbMeans plots and highlight Uni-Plot.

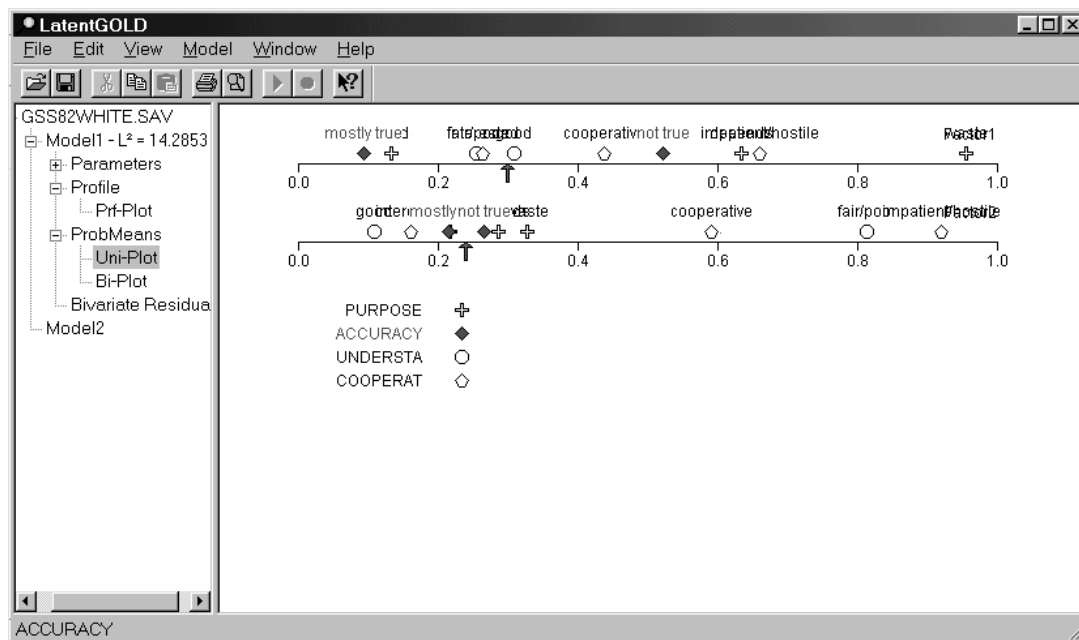


Figure 6-25. ProbMeans Uni-Plot for DFactor Model

The larger the distance (range) between points belonging to a particular variable, the stronger the variable is related to the latent variable.

- By default, a separate Uni-Plot is created for each DFactor. Symbols appear in the plots for each value of each variable specified as an indicator or covariate in the model.
- The ↑ symbol marks the overall mean for a DFactor (corresponding to the first row of the table).
- Click on any variable symbol in the Uni-Plot and the plot label will appear and the status bar will contain a description of the point (variable name and category value).
- Click on any variable name or symbol in the legend and Latent GOLD will highlight all the points that refer to that variable.

## To Change Settings for a Uni-Plot

To change the settings for a Uni-Plot, right click (or select Plot Control from the View Menu) within the Contents pane when a Uni-Plot is displayed to open the Plot Control dialog box.

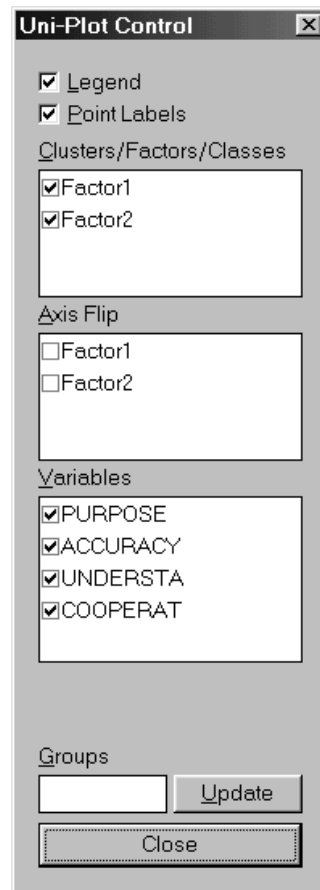


Figure 6-26: Plot Control for Uni-Plots

The following plot settings are available for Uni-Plots:

**Legend.** When this option is selected, a Legend appears at the bottom of the Uni-Plot.

**Point Labels.** When this option is selected, category labels for each variable are listed on the Uni-Plot next to the variable symbol.

**DFactors.** Select which DFactors to include in the Uni-Plots. For each DFactor selected (a checkmark in the checkbox) a Uni-Plot will be displayed. By default, all DFactors are selected.

**Axis Flip.** To flip (reverse) the axis for a Uni-Plot, select the corresponding DFactor name. By default, the DFactor mean range is from 0 to 1 (increasing). Selecting Axis flip for a DFactor will reverse the axis to range from 1 to 0 (decreasing).

**Variables.** Select which indicators/covariates to include in the Uni-Plots. Selected variables are indicated by a checkmark in the checkbox. By default, the Uni-Plots contain all the indicators/covariates included in the model.

**Groups.** Use the grouping option to reduce the number of categories for variable, click Update once you have specified a new number of groups (see Chapter 5, Step 4 for further details on the grouping option).

## Bi-Plot

For a 2-or-more DFactor model, the DFactor means for any 2 selected DFactors are plotted to form a Bi-Plot. A symbol in the Bi-Plot is displayed for each category of indicators and covariates for the two DFactors selected. To view the Bi-Plot, click on the expand/contract (+/-) icon to list the ProbMeans plots and highlight Bi-Plot.

- By default, DFactor1 is used as the horizontal axis and DFactor2 as the vertical axis.
- For each DFactor, a reference line representing the overall mean is displayed. The intersection of the reference lines represents the origin of the plot.
- Click on any variable symbol in the Bi-Plot and the status bar will contain a description of the point (variable name and category, values of the DFactor means plotted).
- Click on any variable name or symbol in the legend and Latent GOLD will highlight all the symbols that refer to that variable.

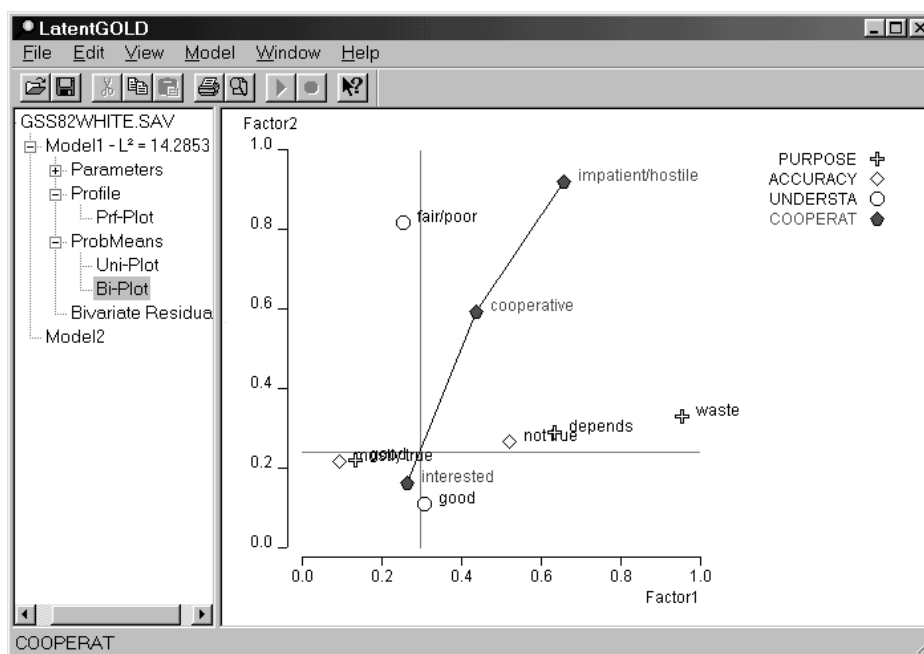


Figure 6-27: Bi-Plot with Lines Option

## To Change Settings for a Bi-Plot

To change the settings for a Bi-Plot, right click (or select Plot Control from the View Menu) within the Contents pane when a Bi-Plot is displayed to open the Plot Control dialog box.

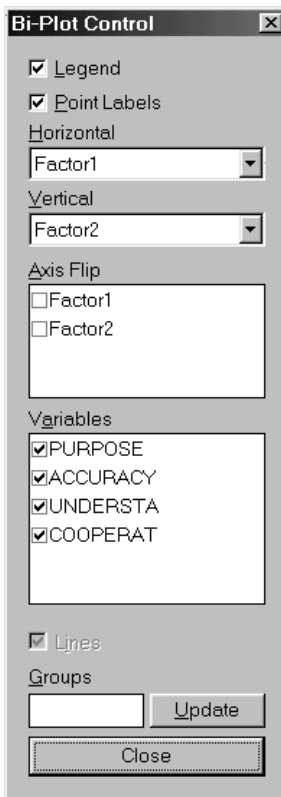


Figure 6-28: Plot Control for Bi-Plots

The following plot settings are available for Bi-Plots:

**Legend.** When this option is selected, a Legend appears to the right of the Bi-Plot.

**Point Labels.** When this option is selected, category labels for each variable are listed on the Bi-Plot next to the variable symbol.

**Horizontal.** Select the DFactor to be used as the horizontal axis. The DFactor currently used as the horizontal axis is listed in the drop-down box. To change the DFactor, click on the downward arrow to the right of the drop-down box. A drop list containing the possible DFactors will appear.

**Vertical.** Select the DFactor to be used as the vertical axis. The DFactor currently used as the vertical axis is listed in the drop-down box. To change the DFactor, click on the downward arrow to the right of the drop-down box. A drop list containing the possible DFactors will appear.

**Axis Flip.** To flip either or both axes for a Bi-Plot, select the corresponding DFactor associated with that axis. By default, the DFactor means range from 0 to 1 (increasing in the likelihood of being at a high DFactor level). Selecting Axis Flip for a DFactor (adding a checkmark in the checkbox), will reverse the axis to range from 1 to 0 (decreasing) for that DFactor.

**Variables.** Select which variables to include in the plot. Those with a checkmark in the checkbox are included in the plot. By default, the Bi-Plots contain all the indicators/covariates that were input as part of the model.

**Groups.** Click Update once you have specified a new number of groups.

**Lines.** Highlight the variable(s) and click Lines to suppress (or include, the default) lines that connect the category points for the selected variable(s).



*For more detailed information about the ProbMeans Output, see section 7.4 of Technical Guide.*

## FREQUENCIES AND RESIDUALS (OPTIONAL)

This output appears only if the Frequencies/ Residuals option was selected in the Output Tab before model estimation.

Click Freqs/Residuals to view a table containing the observed and expected frequencies along with the standardized residuals for a model. In addition, for any model in which  $L^2$  has not been calculated (models where the dependent variable has been specified as Continuous or for multilevel models), this output will not be produced even if selected.



*For more detailed information about the Frequencies Output, see section 7.5 of Technical Guide.*

## BIVARIATE RESIDUALS (OPTIONAL)

This output only appears if the Bivariate Residuals option was selected in the Output Tab before model estimation

Click Bivariate Residuals to view a table containing the bivariate residuals for a model. As an advanced option, users may include parameters associated with the bivariate residuals in a model to improve the fit of an estimated model.

In addition to the  $L^2$  criterion, finding no significant residuals is another indication that a model provides a good fit to the data. In general, those that are larger than 3.84 identify correlations between the associated variable pairs that have not been adequately explained by the model. (For 1 degree of freedom effects, bivariate residuals larger than 3.84 indicate statistical significance at the .05 level.



*For more detailed information about the Bivariate Residuals Output, see section 7.6 of Technical Guide.*

## STANDARD CLASSIFICATION (OPTIONAL)

This output only appears if the Standard Classification option was selected in the Output Tab before model estimation

Click Standard Classification to view a table containing posterior DFactor means, as well as posterior DFactor-level membership probabilities and DFactor modal assignments for each DFactor.

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 6-29. Standard Classification Output for a DFactor Model

For DFactor models, each unique data pattern corresponds to a row in the output., Columns consist of the following information associated with the pattern:

- observed frequency (ObsFreq)
- factor (mean) scores (DFactor1, DFactor2, etc) -- assigning 0 to the first level of the DFactor, and 1 to the last, the mean score can be computed, using the probabilities of being in the DFactor levels as weights. For example, in Figure 6-29 above, for cases in the first row, their scores on Dfactor1 and DFactor2 are .0183 and .0763 respectively.
- the modal level for each DFactor (Modal1, Modal2) -- that DFactor level corresponding to the modal probability (the one that is most likely).
- the posterior membership probabilities corresponding to each level of each DFactor. For example, with a dichotomous DFactors, the posterior probabilities for level 2 (labeled 'DFactor1\_2' and 'DFactor2\_2' in Figure 6-29) are equal to the (mean) score of the corresponding DFactor.
- **A** If 1 or more CFactors have been included in the model, the associated CFactor scores ('Cfac1', 'Cfac2', 'Cfac3'), appear in the right-most portion of the standard classification output.
- **A** If 2 or more GClasses have been included in the model, posterior membership probabilities associated these group classes are given in additional columns (GClass1, GClass2, ...) that appear in the right-most portion of the standard classification output.
- **A** If 1 or more GCFactors have been included in the model, the associated factor scores associated with these group level CFactors ('GCfac1', 'GCfac2', 'GCfac3'), appear in the right-most portion of the standard classification output.

The classification information shown in the classification output can be appended to your data file using the Standard Classification Output option located in the ClassPred Tab. This option must be set before estimating your model.



**For more information on this option, see Step 9 in Chapter 5.**

## COVARIATE CLASSIFICATION (OPTIONAL)

This output only appears if the Covariate Classification option was selected in the Output Tab before model estimation and 1 or more active covariates have been included in the model.

This table contains the estimated probabilities in the multinomial logistic regression model for each DFactor,  $P(x|z)$ , as well as the modal cluster assignments for the joint DFactor based on these probabilities. The table has one row for each unique covariate pattern.



If two or more GClasses have been included in the model, estimated group-level class membership probabilities given group-level covariates are provided in additional columns (GClass1, GClass2, ...) that appear in the right-most portion of the covariate classification output.



*For more detailed information about the Classification Output, see sections 7.8 and 12.7 of Technical Guide.*

## Regression

### GENERAL INFORMATION

The Regression Module is used to estimate LC Regression models containing 1 or more latent classes .

The appropriate Regression model is estimated according to the dependent variable scale type.

**Continuous** - Linear regression (with normally distributed residuals)

**Dichotomous** (specified as nominal, ordinal, or a binomial count) - Binary logistic regression

**Nominal** (with more than 2 levels) - Multinomial logistic regression

**Ordinal** (with more than 2 ordered levels) - Adjacent-category ordinal logistic regression

**Count** - Log-linear Poisson regression

**Binomial Count** - Binomial logistic regression model

The subtype *censored continuous* yields a tobit regression model. The subtypes *truncated continuous*, *truncated count*, and *truncated binomial count* yield truncated versions of the linear, log-linear Poisson, and binomial logistic regression model, respectively.

In addition to the Model Summary Output file, additional model output files are generated following estimation of the model. The specific output sections that are available are those that appear as active in the Output Sections portion of the Output Tab. For Regression models, these are the same as those available for Cluster and DFactor models with 2 exceptions: Bivariate Residuals output is not available for Regression, Estimated Values is available.

In addition, unlike the Cluster and DFactor models, the Prediction Type section of the Output Tab is also active for Regression models.

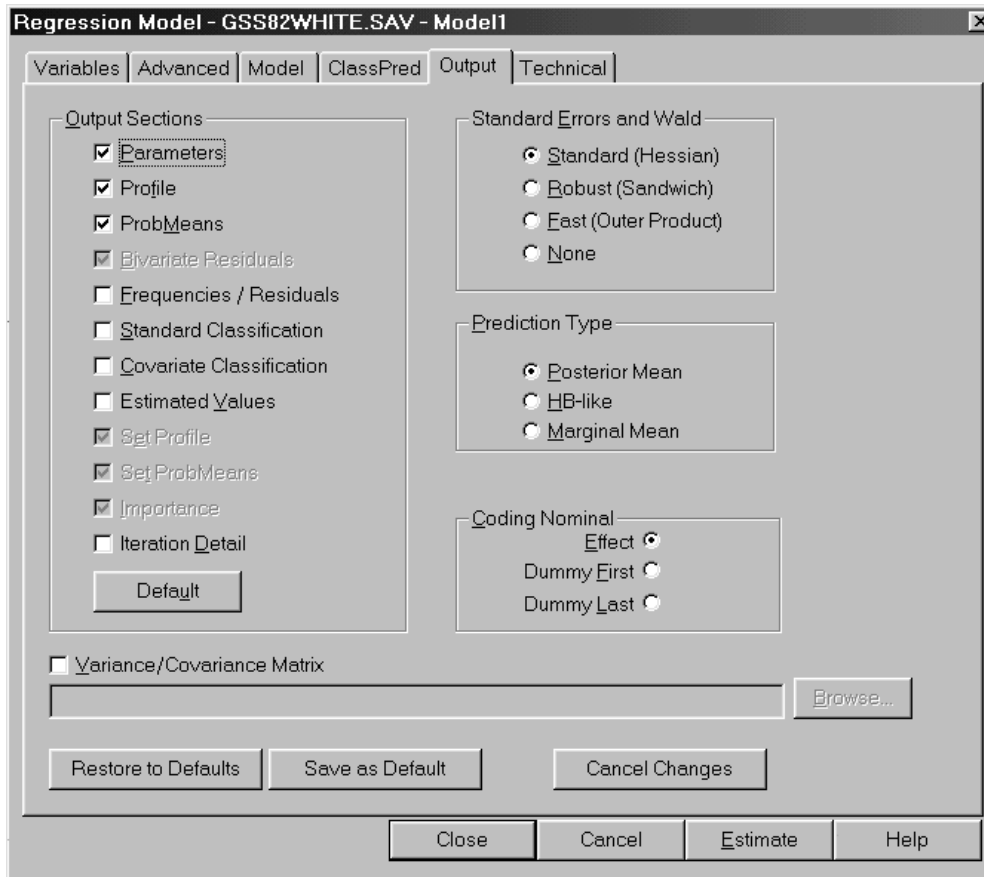


Figure 6-30. Output Tab for Regression Models

Prior to estimating a DFactor model, click in the active check boxes to the left of the Output Sections to include or exclude that Section from being part of the model output files that are generated following the model estimation. A detailed description of each of the Model Summary Output as well as each specific Model Output Section file that can be output appears below.

## LC REGRESSION MODEL SUMMARY OUTPUT

By default, Latent GOLD 4.0 reports the log-likelihood (LL), number of parameters (Npar), BIC based on LL, and the proportion of classification errors (Class.Err.) for all models. In addition, if chi-squared statistics are available, we also report the likelihood-ratio statistic ( $L^2$ ), degrees of freedom (df), and the p-value. The table also contains an overall  $R^2$  based on mean squared error. For more information on these statistics, see the LC Cluster section above.

Specifically, the general section of the Model Summary Output contains the following items:

**Type of Model.** For example, '3-Cluster Model' indicates that a 3-class Cluster model has been estimated.

**Warning messages** (generally not present). Estimation Warning message(s) appear here to indicate that boundary solutions, identification or convergence problems were encountered during the estimation of the model.

**Model Paused** (if you paused the model). If you Paused the model prior to estimation being completed this message appears to alert you that the output files should be viewed as preliminary.

**A Number of groups:** In multilevel latent class models, the program reports the number of groups used in model estimation.

**Number of cases.** This is the number of cases used in model estimation. This number may be less than the original number of cases on the data file if missing cases have been excluded.

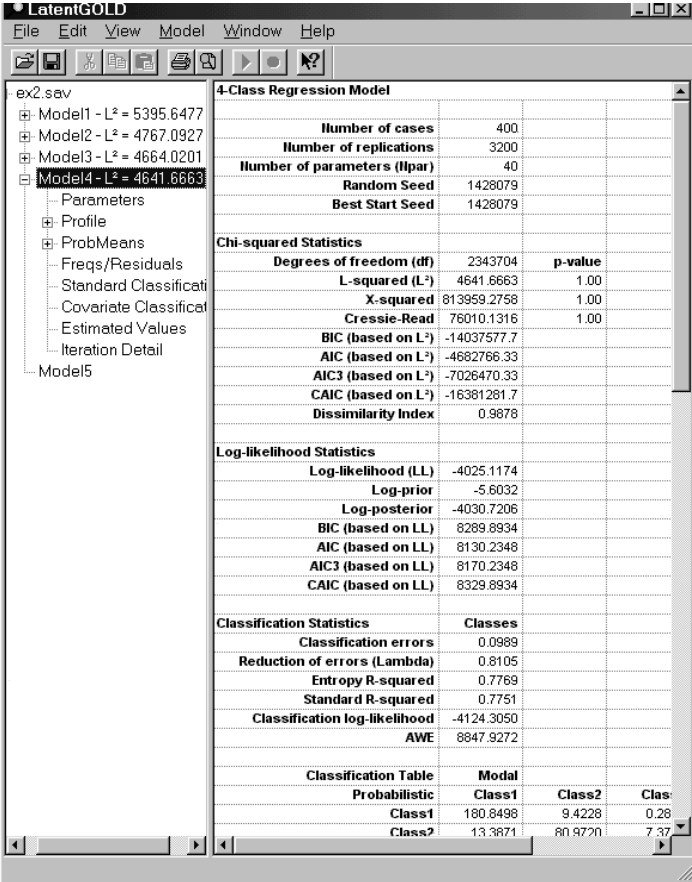
**Number of replications.** If an ID is used (for repeated measure Regression models) the number of replications (records) is provided.

**Number of parameters.** This is the number of distinct parameters estimated.

**Random Seed.** The seed required to reproduce this model.

**Best Start Seed.** The single best seed that can reproduce this model more quickly using the number of starting sets =0. This is the seed that is automatically inserted in a saved definition (.lgf) file.

**A Design Effect.** When the Survey option is used, the program reports the generalized design effect.



The screenshot shows the LatentGold 4.0 software interface. On the left, a tree view shows the model structure for 'ex2.sav', including Model1 through Model5, with Model4 selected. The right pane displays the '4-Class Regression Model' summary output, which includes various statistical measures categorized into Chi-squared, Log-likelihood, Classification, and a Classification Table.

4-Class Regression Model			
Number of cases	400		
Number of replications	3200		
Number of parameters (Npar)	40		
Random Seed	1428079		
Best Start Seed	1428079		
<b>Chi-squared Statistics</b>			
Degrees of freedom (df)	2343704	p-value	
L-squared (L <sup>2</sup> )	4641.6663		1.00
X-squared	813959.2758		1.00
Cressie-Read	76010.1316		1.00
BIC (based on L <sup>2</sup> )	-14037577.7		
AIC (based on L <sup>2</sup> )	-4682766.33		
AIC3 (based on L <sup>2</sup> )	-7026470.33		
CAIC (based on L <sup>2</sup> )	-16381281.7		
Dissimilarity Index	0.9878		
<b>Log-likelihood Statistics</b>			
Log-likelihood (LL)	-4025.1174		
Log-prior	-5.6032		
Log-posterior	-4030.7206		
BIC (based on LL)	8269.8934		
AIC (based on LL)	8130.2348		
AIC3 (based on LL)	8170.2348		
CAIC (based on LL)	8329.8934		
<b>Classification Statistics</b>			
Classification errors	0.0989		
Reduction of errors (Lambda)	0.8105		
Entropy R-squared	0.7769		
Standard R-squared	0.7751		
Classification log-likelihood	-4124.3050		
AWE	8847.9272		
<b>Classification Table</b>			
Probabilistic	Class1	Class2	Class3
Class1	180.8498	9.4228	0.28
Class2	13.3871	80.9720	7.37

Figure 6-31. Model Summary Output for Regression Model

## Chi-squared Statistics

This section lists various chi-square based statistics related to model fit.

If the scale type for the dependent variable has been set to 'continuous', no chi-squared statistics are available and this section is not displayed.

The information reported:

**Degrees of freedom (df).** The degrees of freedom for the current model.

**L-squared ( $L^2$ ).** The likelihood-ratio goodness-of-fit value for the current model. If the bootstrap p-value for the  $L^2$  statistic has been requested, the results will be displayed here.

**X-squared and Cressie-Read.** These are alternatives to  $L^2$  that should yield a similar p-value according to large sample theory if the model specified is valid and the data is not sparse.

**BIC, AIC and CAIC (based on  $L^2$ ).** In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC and CAIC value the better the model.

**Dissimilarity Index.** A descriptive measure indicating how much the observed and estimated cell frequencies differ from one another. It indicates the proportion of the sample that needs to be moved to another cell to get a perfect fit.



*For more detailed information about Chi-squared statistics including the formulae for each, see section 7.1.1 of Technical Guide.*

## Log-likelihood Statistics

This section contains additional statistics related to the model fit that are especially useful when  $L^2$  and the other chi-squared statistics are not available. The statistics reported are:

**Log-likelihood (LL).** If the conditional bootstrap (bootstrap -2LL Diff) has been requested, the results will be displayed here.

**Log-prior --** this is the term in the function maximized in the parameter estimation that is associated with the Bayes constants. This term equals 0 if all Bayes constants are set to 0.

**Log-posterior -** this is the function that is maximized in the parameter estimation. The value of the log-posterior function is obtained as the sum of the log-likelihood and log-prior values.

**BIC, AIC, AIC3 and CAIC (based on LL) -** these statistics (information criteria) weight fit and parsimony by adjusting the LL to account for the number of parameters in the model. The lower the value, the better the model. For example, according to the BIC values shown in Figure 6-31, the 3-class model is preferred over the 1-class, 2-class and 4-class models.



*For more detailed information about Log-likelihood statistics including the formulae for each, see section 7.1.2 of Technical Guide.*

## Classification Statistics

This information can be used to assess how well the model classifies cases into clusters. The statistics reported are:

**Classification Errors.** When classification of cases is based on modal assignment (to the class having the highest membership probability), the proportion of cases that are expected to be misclassified is reported by this statistic. The closer this value is to 0 the better.

**Reduction of Errors (lambda), Entropy R-squared and Standard R-squared.** These statistics indicate how well the model predicts class memberships or DFactor scores. The closer these values are to 1 the better the predictions.

**Classification log-likelihood.** Log-likelihood value under the assumption that the true class membership is known.

**AWE.** Similar to BIC, but also takes classification performance into account.

**Classification Table.** The Classification Table cross-tabulates modal and probabilistic class assignments.

**A** **Standard R-squared** is reported for CFactors and GCFactors. For GClasses, as for Clusters, Classification Errors, Reduction of Errors, Entropy R-squared, and Standard R-squared values are reported.

## Covariate Classification Statistics

If one or more active covariates is included in the model, additional statistics are provided as above but these are now based only on the active covariates. The statistics reported are:

**Classification errors, Reduction of Errors (lambda), Entropy R-squared and Standard R-squared**

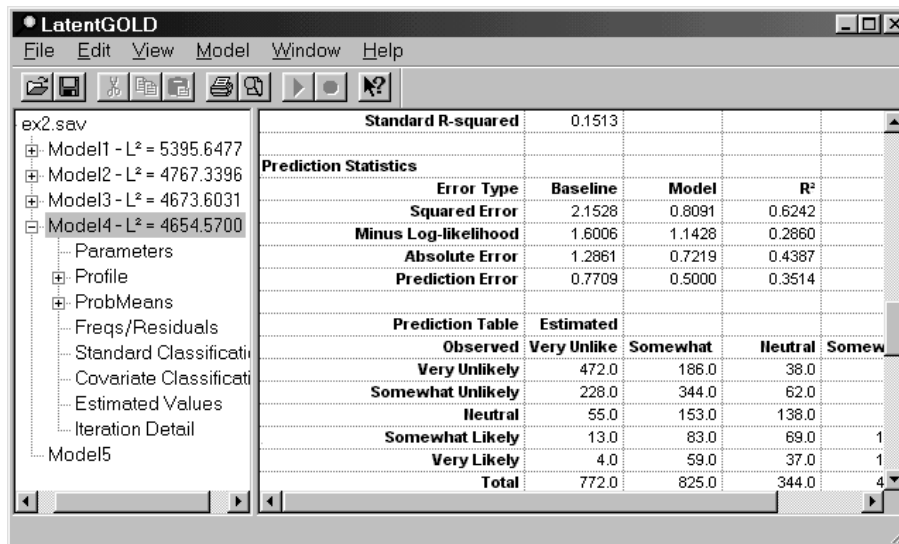
**A** **Standard R-squared** is reported for CFactors and GCFactors. For GClasses, as for Clusters, Classification Errors, Reduction of Errors, Entropy R-squared, and Standard R-squared values are reported.



*For more detailed information about the classification statistics including the formulae for each, see section 7.1.3 of Technical Guide.*

## Prediction Statistics

The prediction statistics are based on the comparison between observed and predicted responses.



Standard R-squared: 0.1513

Prediction Statistics

Error Type	Baseline	Model	R <sup>2</sup>
Squared Error	2.1528	0.8091	0.6242
Minus Log-likelihood	1.8006	1.1428	0.2860
Absolute Error	1.2861	0.7219	0.4387
Prediction Error	0.7709	0.5000	0.3514

Prediction Table

Observed	Estimated	Very Unlikely	Somewhat	Neutral	Somewhat Likely	Very Likely
Very Unlikely	472.0	186.0	38.0			
Somewhat Unlikely	228.0	344.0	62.0			
Neutral	55.0	153.0	138.0			
Somewhat Likely	13.0	83.0	69.0			
Very Likely	4.0	59.0	37.0			
Total	772.0	825.0	344.0			

Figure 6-32. Prediction Statistics

This information can be used to assess prediction performance of the model. It is also possible to output predicted values for the **dependent** variable to an external file (**For more information on this option, see the section on the ClassPred Tab in Step 9 of Chapter 5.**)

*Prediction Statistics*, contains the following measures of Prediction error: **mean squared error** (MSE), **mean absolute error** (MAE), **minus mean log-likelihood** (-MLL), and for ordinal/nominal dependent variables, the proportion of predictions errors under modal prediction (PPE). For each error measure, we provide the prediction error of the baseline, or intercept-only model, the prediction error of the estimated model, and a R<sup>2</sup> value, which is the proportional reduction of errors in the estimated model compared to the baseline model. For nominal and ordinal dependent variables, a prediction table that cross-classifies observed and against estimated values is also provided.

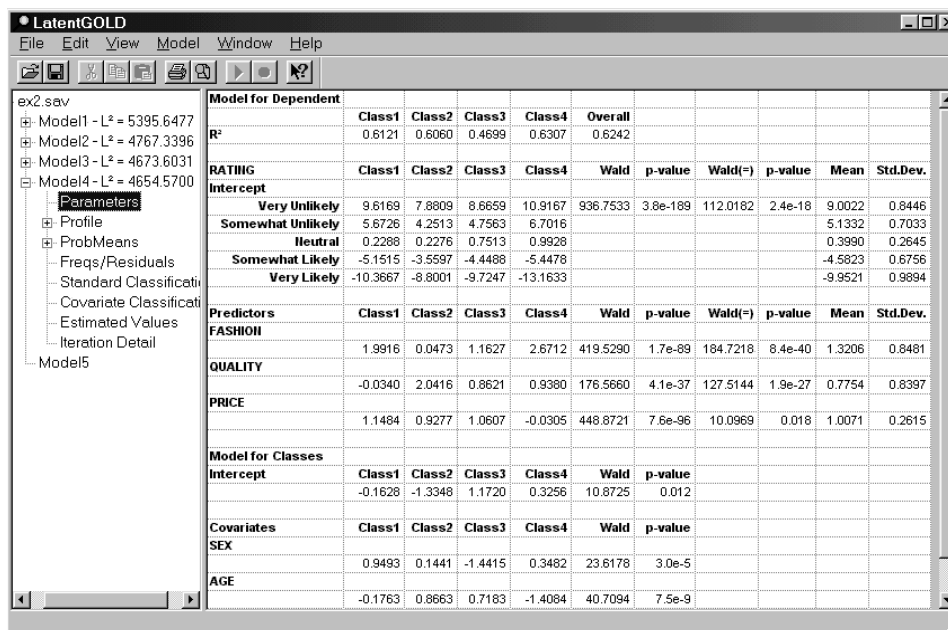


**For more technical information on Prediction Statistics, see section 7.1.4 in the Technical Guide.**

**Variable Detail** This section contains details about the variables entered into the model including variable names, scale types, number of categories and category labels and scores (if used).

## PARAMETERS OUTPUT (OPTIONAL)

For any estimated model, click Parameters and Latent GOLD displays a table containing class-specific parameter estimates (betas, gammas, sigmas), class-specific and overall R<sup>2</sup> values, and measures of significance for these estimates. In addition, the right-most columns contain means and standard deviations for the regression coefficients. In the case of a 1-class model, the mean is identical to the corresponding coefficient and the standard deviation is 0 which indicates that this is a fixed effect. For K>1 class models and unequal coefficients across classes, these are the means and the standard deviations for the (discrete) random effects. The size of the standard deviation indicates the extent of the between-class variation in the coefficient concerned.



	Class1	Class2	Class3	Class4	Overall					
<b>Model for Dependent</b>										
R <sup>2</sup>	0.6121	0.6060	0.4699	0.6307	0.6242					
<b>RATING</b>										
Intercept										
Very Unlikely	9.6169	7.8809	8.6659	10.9167	936.7533	3.8e-189	112.0182	2.4e-18	9.0022	0.8446
Somewhat Unlikely	5.6726	4.2513	4.7563	6.7016					5.1332	0.7033
Neutral	0.2288	0.2276	0.7513	0.9928					0.3990	0.2645
Somewhat Likely	-5.1515	-3.5597	-4.4488	-5.4478					-4.5823	0.6756
Very Likely	-10.3667	-8.8001	-9.7247	-13.1633					-9.9521	0.9894
<b>Predictors</b>										
FASHION										
	1.9916	0.0473	1.1627	2.6712	419.5290	1.7e-89	184.7218	8.4e-40	1.3206	0.8481
QUALITY										
	-0.0340	2.0416	0.8621	0.9380	176.5660	4.1e-37	127.5144	1.9e-27	0.7754	0.8397
PRICE										
	1.1484	0.9277	1.0607	-0.0305	448.8721	7.6e-96	10.0969	0.018	1.0071	0.2615
<b>Model for Classes</b>										
Intercept										
	-0.1628	-1.3348	1.1720	0.3256	10.8725	0.012				
<b>Covariates</b>										
SEX										
	0.9493	0.1441	-1.4415	0.3482	23.6178	3.0e-5				
AGE										
	-0.1763	0.8663	0.7183	-1.4084	40.7094	7.5e-9				

Figure 6-33. Parameters Output for Regression Model

**Betas.** For Regression models, in addition to the class-specific regression intercept, betas indicate the class-specific effect of each predictor on the dependent variable.

**Gammas.** Parameters of the multinomial logit model used to predict the latent distribution as a function of the covariates. Parameters include the intercept as well as the effects of each covariate on the latent variable(s).

**Sigmas.** Error variances for a continuous dependent variable.

The Parameters output reports class-specific  $R^2$  values and the overall  $R^2$  value. These are reduction of error measures based on mean squared error. The overall  $R^2$  indicates how well the dependent variable is overall predicted by the model (same measure as appearing in Prediction Statistics). For ordinal, continuous, and (binomial) counts, these are standard  $R^2$  measures. For nominal dependent variables, these can be seen as weighted averages of separate  $R^2$  measures for each category treated as a separate dichotomous response variable. For specific formulae for the overall and class-specific  $R^2$ , see section 7.2 of Technical Guide.



If CFactors have been included in the model, estimates and related information for additional parameters (lambdas) are included in the Model for Dependent output.



If GClasses and/or GCFactors have been included in the model, estimates and related information for additional parameters are included in the Model for Dependent and/or the Model for Classes output, and a separate section called Model for GClasses appears at the bottom of the listing.

### Viewing Wald Statistics and Standard Errors

By default, Wald statistics are provided in the output to assess the statistical significance of the set of parameter estimates associated with a given variable. Specifically, for each variable, the Wald statistic tests the restriction that each of the parameter estimates in that set equals zero (for variables specified as Nominal, the set includes parameters for each category of the variable). For Regression models, by default, two Wald statistics (Wald, Wald(=)) are provided in the table when more than 1 class has been estimated. For each set of parameter estimates, the Wald(=) statistic considers the subset associated with each class and tests the restriction that each parameter in that subset equals the corresponding parameter in the subsets associated with each of the other classes. That is, the Wald(=) statistic tests the equality of each set of regression effects across classes.

To view standard errors or related statistics associated with parameter estimates, simply right click on the parameters output and select/deselect the items that you want to appear. Alternatively, you can click on the appropriate item in the View Menu. In Parameters, you can alter the output view to obtain standard errors, a Z Statistic, and can suppress the Wald statistics.



*For more general information about the Parameters Output, see section 7.2 of Technical Guide.*



### PARAMETERS OUTPUT SUBCATEGORY (ADVANCED)

If 1 or more CFactors have been included in the model, clicking on the + to the left of Parameters makes visible the Parameters Output Subcategory called Random Effects.

**Random Effects.** This table reports the CFactor effects in the first columns (labeled 'CFactor1', 'CFactor2', etc.) and provides the variance-covariance matrix for the continuous random effects in the columns to the right of these CFactor effects (labeled 'eff1', 'eff2', etc.). For further information see Section 9.2.4 of Technical Guide.

### PROFILE OUTPUT (OPTIONAL)

To view the profile table for a selected model, click Profile in the Outline pane. The Profile table contains probabilities or means associated with the Dependent variable and covariates.

	Class1	Class2	Class3	Class4
<b>Class Size</b>	0.4405	0.2577	0.2487	0.0532
<b>Dependent</b>				
<b>RATING</b>				
Very Unlikely	0.2838	0.1796	0.1778	0.2553
Somewhat Unlikely	0.2821	0.1466	0.1818	0.2697
Neutral	0.1239	0.1240	0.2641	0.1942
Somewhat Likely	0.1117	0.2696	0.1859	0.1966
Very Likely	0.1986	0.2802	0.1904	0.0842
Mean	2.6592	3.3242	3.0292	2.5848
<b>Covariates</b>				
<b>SEX</b>				
Male	0.2463	0.4120	0.7762	0.3727
Female	0.7537	0.5880	0.2238	0.6273
Mean	1.7537	1.5880	1.2238	1.6273
<b>AGE</b>				
16-24	0.6417	0.2615	0.2617	0.8944
25-39	0.1750	0.1950	0.3216	0.0964
40+	0.1832	0.5435	0.4167	0.0092
Mean	1.5415	2.2819	2.1551	1.1147

Figure 6-34. Profile Output for Regression Model

- For a Regression model, the first row of numbers shows how large each class is. The body of the table contains the (marginal) conditional probabilities that show how the classes are related to the dependent variable, specified as Nominal, Ordinal or Binomial Count. The probabilities within each class (column) sum to 1. For dependent variables specified as Ordinal and Continuous, the body of the table contains means, and for Counts these are rates.
- For covariates the Profile output contains rescaled ProbMeans output rather than model probabilities or means. These are aggregate class membership probabilities which are rescaled to sum to one within clusters. For Covariates specified as Numeric, means are displayed in addition to the conditional probabilities, and if the Covariate contains more than 5 distinct values, the probabilities are displayed for 5 grouped ranges. The number of such grouped ranges may be changed (increased or decreased) using Groups option in the Plot Control.

## Viewing Standard Errors

To view the associated standard errors in a Profile table, simply right click on the parameters output to retrieve the pop-up menu and click Standard Errors. Column containing the standard errors appear to the right of each estimate.

## Partial Profile

Latent GOLD 4.0 reports not only marginal but also partial probabilities/means for the dependent variable in the Profile Table Output. These two differ if predictors have been included in the model. As explained in section 7.3 of the Technical Guide, partial probabilities/means are obtained by conditioning on a certain value (mean) of the

predictors. The marginal probabilities/means, on the other hand, are obtained by aggregating over the categories of these predictors. Both have their specific advantages: the Marginal Profile output is somewhat easier to interpret and is displayed by default.

By a right click on the Profile output, you can retrieve a pop-up menu that can be used to switch between Partial to Marginal and to obtain standard errors. (The partial/marginal switch has no effect on covariate information).

### PROFILE PLOT (FOR DEPENDENT AND COVARIATES)

To view the Profile Plot, click the '+' icon to expand the Profile output and highlight Prf-Plot. The Profile Plot is constructed from the conditional probabilities for the nominal variables and means for the other indicators and covariates as displayed in the columns of the Profile Table. Specific classes (columns) are selected for display using the Plot Control pop-up menu. The quantities associated with the selected clusters are plotted and connected to form a line graph.

For the scale types ordinal, continuous, count, and numeric covariate, prior to plotting the class-specific means, they are re-scaled to always lie within the 0-1 range. Scaling of these "0-1 Means" is accomplished by subtracting the lowest observed value from the class-specific means and dividing the results by the range, which is simply the difference between the highest and the lowest observed value. The advantage of such scaling is that these numbers can be depicted on the same scale as the class-specific probabilities for nominal variables.

For nominal variables containing more than 2 categories, all categories are displayed simultaneously. For dichotomous variables specified as nominal, by default only the last category is displayed. The Plot Control can be used to display the first or both categories, as well as to specify the columns, variables, and selected categories to appear in the display.

When the contents of the Profile Table is changed from the default view to display Partial (instead of Marginal) probabilities, the points plotted change to reflect the current contents of the Profile Table.

- A separate line is displayed for each class.
- Click on any variable symbol in the Profile Plot and the status bar describes it (variable name, Class number, and value).
- Click on any Class name or symbol in the legend and Latent GOLD highlights all the symbols that refer to that Class.

By default, only the first 8 variables are displayed in the profile plot. Additional variables may be selected from the Plot Control (using a control-click or a shift click to select more than one at a time) and added or removed from the plot with a single click in the box to the left of the variable names.

### To Change Settings for a Profile Plot

To change the settings for a Profile Plot, right click (or select Plot Control from the Model Menu) within the Contents pane when a Profile Plot is displayed to open the Plot Control dialog box. To change the font type/size for a plot, see Chapter 2.



Figure 6-35: Plot Control for Profile Plots

The following plot settings are available for Profile Plots:

**Legend.** When this option is selected, a legend appears at the bottom of the Profile Plot.

**Classes.** For Regression models, a line will be drawn for each class selected. Those classes with a checkmark in the checkbox are included in the plot.

**Variables.** Select which variables to include in the plot. Those with a checkmark in the checkbox are included in the plot.

**Categories.** Select which category of a variable to include in the plot. The category currently being plotted is listed in the plot beneath the variable name. To change the category that is plotted, highlight the variable name in the Variables box (the category currently being plotted will appear in the Category box), click the drop-down list to the right of the Categories box and select the category you wish to have plotted.

**Groups:** Click Update once you have specified a new number of groups.



*For more detailed information about the Profile Output, see section 7.3 of Technical Guide.*



### **GPROFILE OUTPUT (ADVANCED, OPTIONAL)**

If GClasses and/or GCFactors are included in a model, this output file listing is generated. Similar to the Profile output, the top of this file contains the size of each GClass, followed by the probability of being in each Class for each GClass. This section is followed by the Dependent section where means and (marginal) probabilities associated with the dependent variable are provided.

The View Menu options (which can be obtained by a right click on the GProfile output), can be used to switch between Marginal and Partial GProfile Output, or to obtain standard errors.

### **PROBMEANS OUTPUT (OPTIONAL)**

For Regression models, this table contains aggregated class membership probabilities for (ranges of) dependent and covariates values which are displayed in the Uni-Plot and Tri-Plot.

To view the Probability/Means table for a selected model, click ProbMeans in the Outline pane. To view a plot, click on the expand '+' icon to the left of ProbMeans to list the type of plots produced for a particular model. Highlight a plot type to view it in the Contents pane.

For Regression models, the first row of the table contains the overall probability of being in a class (the size of each class, as also reported as the first row of numbers in the Profile Table). The body of the table contains conditional class probabilities associated with each category of categorical dependent variables (these probabilities will sum to 100% across rows).

### **UNI-PLOT**

For a Regression model, the membership probabilities in the body of the ProbMeans output are plotted to form a Uni-Plot. To view the Uni-Plot, click on the expand/contract icon (+/-) to list the ProbMeans plots and highlight Uni-Plot.

The larger the distance (range) between points belonging to a particular variable, the stronger the variable is related to the latent variable.

- By default, a separate Uni-Plot is created for each class. Symbols appear in the plots for each value of each variable specified as an indicator or covariate in the model.
- The ↑ symbol marks the overall probability (the size) for a class.
- Click on any variable symbol in the Uni-Plot and the plot label will appear and the status bar will contain a description of the point (variable name and category value).
- Click on any variable name or symbol in the legend and Latent GOLD will highlight all the points that refer to that variable.

## To Change Settings for a Uni-Plot

To change the settings for a Uni-Plot, right click (or select Plot Control from the View Menu) within the Contents pane when a Uni-Plot is displayed to open the Plot Control dialog box. To change the font for a plot, see Chapter 2.

The following plot settings are available for Uni-Plots:

**Legend.** When this option is selected, a Legend appears at the bottom of the Uni-Plot.

**Point Labels.** When this option is selected, category labels for each variable are listed on the Uni-Plot next to the variable symbol.

**Classes.** Select which Classes to include in the Uni-Plots. For each class selected (a checkmark in the checkbox) a Uni-Plot will be displayed. By default, all classes are selected.

**Axis Flip.** To flip (reverse) the axis for a Uni-Plot, select the corresponding class name. By default, the class probabilities range is from 0 to 1 (increasing). Selecting Axis flip for a class will reverse the axis to range from 1 to 0 (decreasing).

**Variables.** Select which indicators/covariates to include in the Uni-Plots. Selected variables are indicated by a checkmark in the checkbox. By default, the Uni-Plots contains all the indicators/covariates included in the model.

**Groups.** Use the grouping option to reduce the number of categories for a variable, click Update once you have specified a new number of groups (see Chapter 5, Step 4 for further details on the grouping option).

## TRI-PLOT

For Regression models, the class membership probabilities in the body of the ProbMeans output table are plotted to form a Tri-Plot. To view the Tri-Plot, click on the expand/contract icon (+/-) to list the ProbMeans plots and highlight Tri-Plot. (Note: No Tri-Plot is produced for a 1-class model; for a 2-class model, the Tri-Plot reduces to the Uni-Plot.)

- By default, Vertex A (left-most base vertex) is labeled 'Class 1', Vertex B (right-most base vertex) 'Class 2', and the third Vertex (the top point of the triangle) represents the aggregate of all other classes. For a 3-class Regression model, by default, the third vertex will represent Class 3 and is labeled 'Class 3'. For a 4-or-more class Regression model, the third vertex is labeled 'Others'. For a 2-class Regression model, the class 3 membership probability is 0 and the Tri-Plot reduces to the Uni-Plot.
- The ▲ symbol marks the overall probabilities for the 3 classes associated with the vertices. It represents the centroid of the triangle.
- Click on any variable symbol in the Tri-Plot and 1) the status bar will contain a description of the point (variable name and category, class probabilities) 2) the category label will appear next to that point on the plot and 3) lines emanate from that point to each side of the triangle, intersecting the side at the corresponding class probabilities value.
- Click on any variable symbol or name in the legend and all the symbols for that variable will be highlighted and their category labels listed in the Tri-Plot.

### To Change Settings for a Tri-Plot

To change the settings for a Tri-Plot, right click (or select Plot Control from the View Menu) within the Contents pane when a Tri-Plot is displayed to open the Plot Control dialog box. To change the font for a plot, see Chapter 2.

The following plot settings are available for Tri-Plots:

**Legend.** When this option is selected, a Legend appears to the right of the Tri-Plot.

**Point Labels.** When this option is selected, category labels for each variable are listed on the Tri-Plot next to the variable symbol.

**Vertices.** Latent GOLD allows you to select the base vertices in the Tri-Plot. The top vertex corresponds to the aggregate of the remaining classes.

**A vertex.** The class currently used as the A vertex is listed in the drop down box. To select a different class, click on the down arrow to the right of the vertex box. A drop list containing all classes will appear. Select the class to use as the A vertex.

**B vertex.** The class currently used as the B vertex is listed in the drop down box. To select a different class, click on the down arrow to the right of the vertex box. A drop list containing all classes will appear. Select the class to use as the B vertex.

**Variables.** Select which variables to include in the Tri-Plot. Those with a checkmark in the checkbox are included in the plot. By default, the Tri-Plot contains all the indicators/covariates that were input as part of the model.

**Groups.** Click Update once you have specified a new number of groups.



*For more detailed information about the ProbMeans Output, see section 7.4 of Technical Guide.*

### FREQUENCIES AND RESIDUALS (OPTIONAL)

This output appears only if the Frequencies/ Residuals option was selected in the Output Tab before model estimation.

Click Freqs/Residuals to view a table containing the observed and expected frequencies along with the standardized residuals for a model. In addition, for any model in which  $L^2$  has not been calculated (models where a dependent variable has been specified as Continuous or in multilevel models), this output will not be produced even if selected.



*For more detailed information about the Frequencies Output, see section 7.5 of Technical Guide.*

STANDARD CLASSIFICATION (OPTIONAL)

This output only appears if the Standard Classification Information option was selected in the Output Tab before model estimation.

Click Standard Classification to view a table containing the posterior membership probabilities and other classification information for a model.

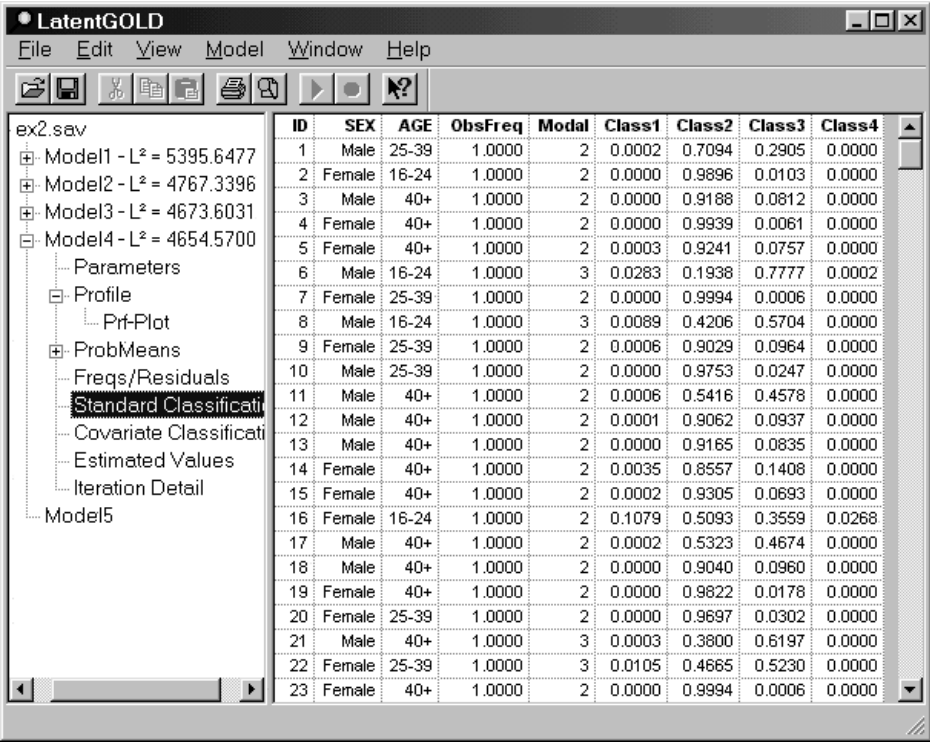


Figure 6-36. Standard Classification Output for Regression Model

Each row contains grouped cases corresponding to responses to covariates, predictors, and dependent (in that order). If an ID variable is specified, it is included in the first column. If an ID is specified, a frequency count of 2 or more means that some cases have been combined (included in the same row of the classification output) because their data response patterns are identical. In this case, the ID of the first case only is displayed. (An output data file consisting of separate records for each case can be obtained using the ClassPred Tab. This option must be set before estimating your model. For more information on this option, see the section on the ClassPred Tab in Step 9 of Chapter 5).

For each data response pattern, the classification output contains separate columns for frequencies ('ObsFreq'), modal class ('Modal'), posterior probabilities of belonging to each class ('Class 1', 'Class 2', ...), and CFactor scores ('Cfac1', 'Cfac2', 'Cfac3'). The classification variable 'Modal' contains that class number associated with the modal class.

**A** If 1 or more CFactors have been included in the model, the associated CFactor scores ('Cfac1','Cfac2','Cfac3'), appear in the right-most portion of the standard classification output.

**A**

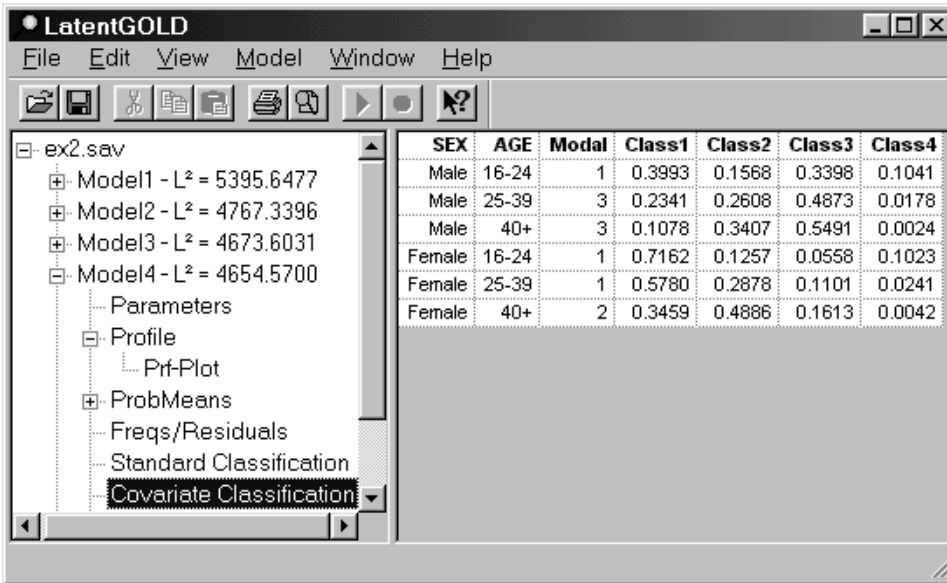
If 2 or more GClasses have been included in the model, posterior membership probabilities associated these group classes are given in additional columns (GClass1, GClass2, ...) that appear in the right-most portion of the standard classification output.

**A**

If 1 or more GCFactors have been included in the model, the associated factor scores associated with these group level CFactors ('GCfac1','GCfac2','GCfac3'), appear in the right-most portion of the standard classification output.

## COVARIATE CLASSIFICATION (OPTIONAL)

This output only appears if the Covariate Classification option was selected in the Output Tab before model estimation and if 1 or more active covariates have been included in the model.



SEX	AGE	Modal	Class1	Class2	Class3	Class4
Male	16-24	1	0.3993	0.1568	0.3398	0.1041
Male	25-39	3	0.2341	0.2608	0.4873	0.0178
Male	40+	3	0.1078	0.3407	0.5491	0.0024
Female	16-24	1	0.7162	0.1257	0.0558	0.1023
Female	25-39	1	0.5780	0.2878	0.1101	0.0241
Female	40+	2	0.3459	0.4886	0.1613	0.0042

Figure 6-37. Covariate Classification Output for Regression Model

This table contains the estimated probabilities in the multinomial logistic regression model for the classes,  $P(x|z)$ , as well as the modal cluster assignments based on these probabilities. The table has one row for each unique covariate pattern.

**A**

If 2 or more GClasses have been included in the model, estimated group-level class membership probabilities given group-level covariates are provided in additional columns (GClass1, GClass2, ...) that appear in the right-most portion of the covariate classification output.

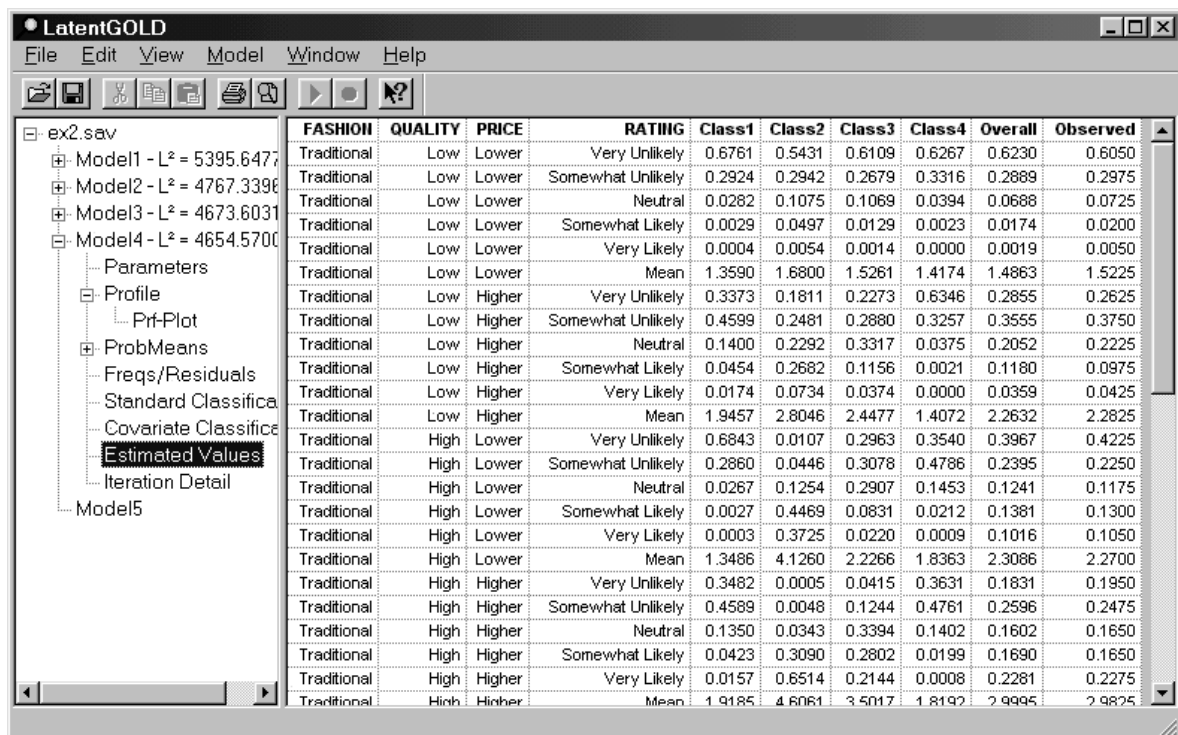


**For more detailed information about the Classification Output, see sections 7.8 and 12.7 of Technical Guide.**

## ESTIMATED VALUES (OPTIONAL)

This output only appears if the Estimated Values option was selected in the Output Tab before model estimation.

The Estimated Values output section reports the estimated class-specific, the estimated overall, and the observed means (probabilities) for each unique predictor pattern.



FASHION	QUALITY	PRICE	RATING	Class1	Class2	Class3	Class4	Overall	Observed
Traditional	Low	Lower	Very Unlikely	0.6761	0.5431	0.6109	0.6267	0.6230	0.6050
Traditional	Low	Lower	Somewhat Unlikely	0.2924	0.2942	0.2679	0.3316	0.2889	0.2975
Traditional	Low	Lower	Neutral	0.0282	0.1075	0.1069	0.0394	0.0688	0.0725
Traditional	Low	Lower	Somewhat Likely	0.0029	0.0497	0.0129	0.0023	0.0174	0.0200
Traditional	Low	Lower	Very Likely	0.0004	0.0054	0.0014	0.0000	0.0019	0.0050
Traditional	Low	Lower	Mean	1.3590	1.6800	1.5261	1.4174	1.4863	1.5225
Traditional	Low	Higher	Very Unlikely	0.3373	0.1811	0.2273	0.6346	0.2855	0.2625
Traditional	Low	Higher	Somewhat Unlikely	0.4599	0.2481	0.2880	0.3257	0.3555	0.3750
Traditional	Low	Higher	Neutral	0.1400	0.2292	0.3317	0.0375	0.2052	0.2225
Traditional	Low	Higher	Somewhat Likely	0.0454	0.2682	0.1156	0.0021	0.1180	0.0975
Traditional	Low	Higher	Very Likely	0.0174	0.0734	0.0374	0.0000	0.0359	0.0425
Traditional	Low	Higher	Mean	1.9457	2.8046	2.4477	1.4072	2.2632	2.2825
Traditional	High	Lower	Very Unlikely	0.6843	0.0107	0.2963	0.3540	0.3967	0.4225
Traditional	High	Lower	Somewhat Unlikely	0.2860	0.0446	0.3078	0.4786	0.2395	0.2250
Traditional	High	Lower	Neutral	0.0267	0.1254	0.2907	0.1453	0.1241	0.1175
Traditional	High	Lower	Somewhat Likely	0.0027	0.4469	0.0831	0.0212	0.1381	0.1300
Traditional	High	Lower	Very Likely	0.0003	0.3725	0.0220	0.0009	0.1016	0.1050
Traditional	High	Lower	Mean	1.3486	4.1260	2.2266	1.8363	2.3086	2.2700
Traditional	High	Higher	Very Unlikely	0.3482	0.0005	0.0415	0.3631	0.1831	0.1950
Traditional	High	Higher	Somewhat Unlikely	0.4589	0.0048	0.1244	0.4761	0.2596	0.2475
Traditional	High	Higher	Neutral	0.1350	0.0343	0.3394	0.1402	0.1602	0.1650
Traditional	High	Higher	Somewhat Likely	0.0423	0.3090	0.2802	0.0199	0.1690	0.1650
Traditional	High	Higher	Very Likely	0.0157	0.6514	0.2144	0.0008	0.2281	0.2275
Traditional	High	Higher	Mean	1.9185	4.6061	3.5017	1.8192	2.9995	2.9825

Figure 6-38. Estimated Values Output for Regression Model



For further information about Estimated Values, see section 7.7 of Technical Guide.

# 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.

	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

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.

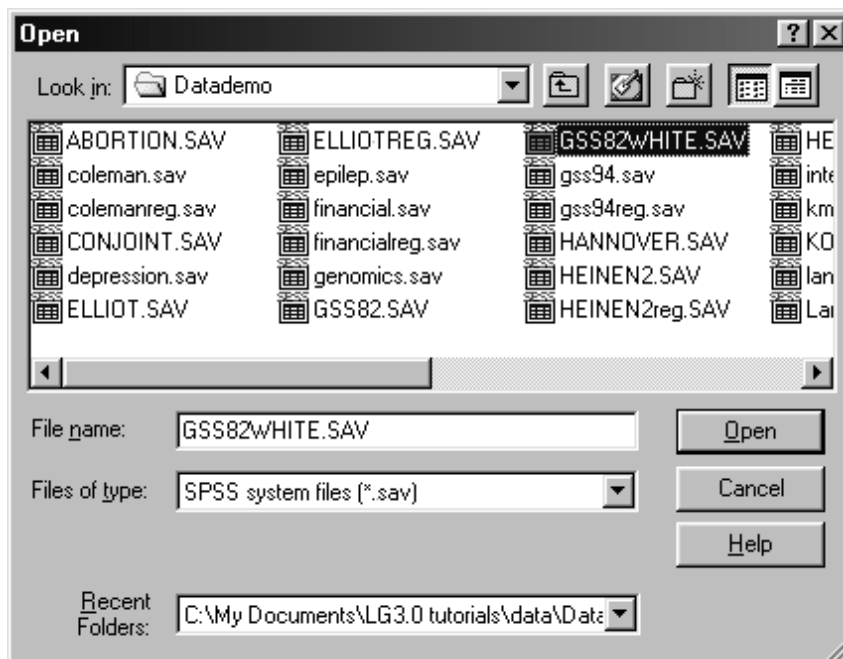


Figure 7-2. File Open Dialog Box

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

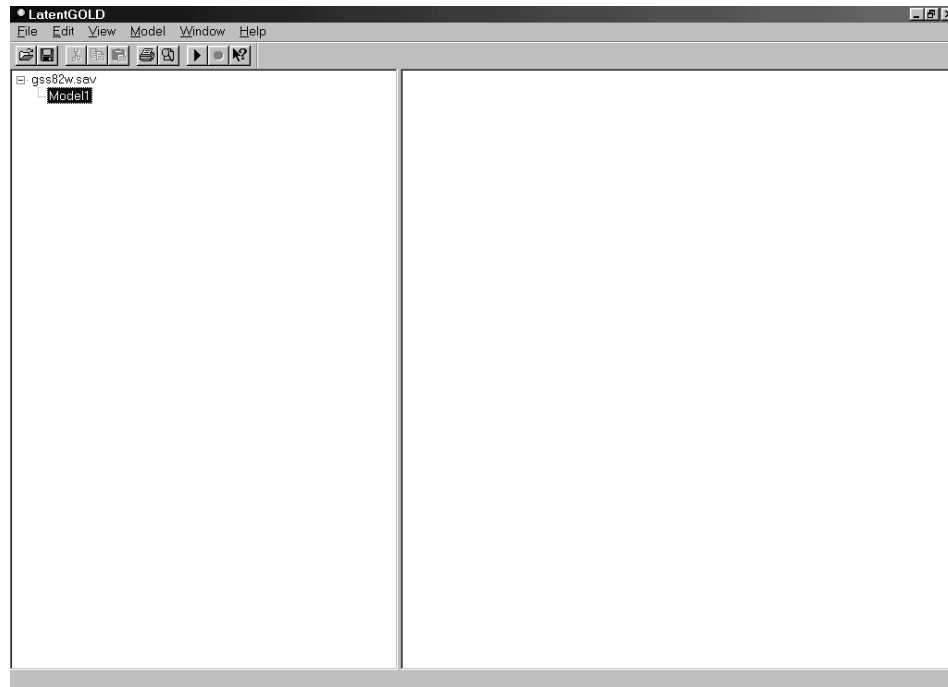


Figure 7-3. Viewer Window

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.

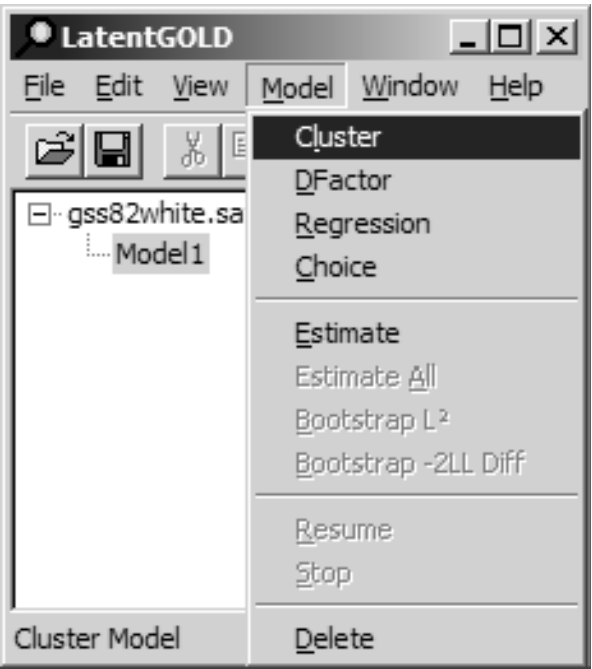


Figure 7-4. Model Selection Menu

▷ Select Cluster

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

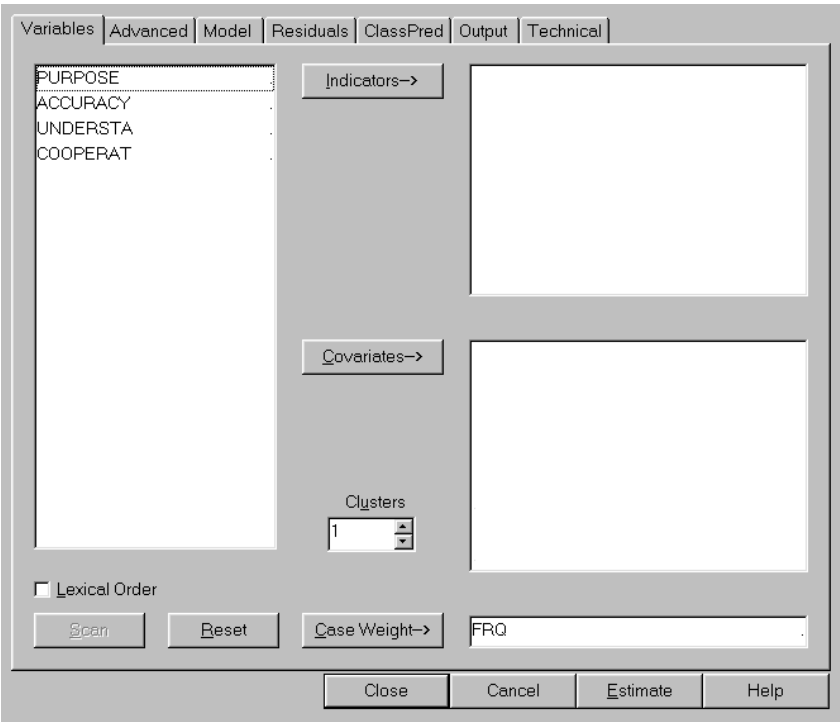



Figure 7-5. Analysis dialog box for LC Cluster Model


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').

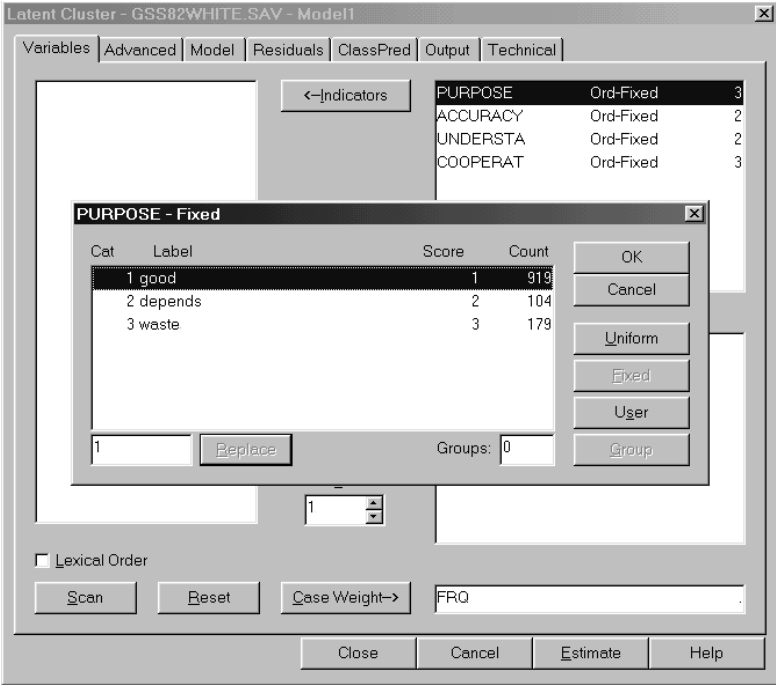


Figure 7-6. Category Information for the Variable PURPOSE

- ▷ 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.

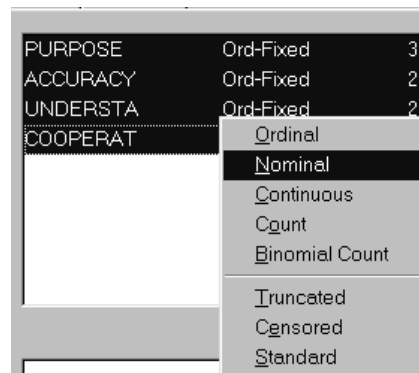


Figure 7-7. Pop-up Menu to Set Variable Scale Types and Subtypes

## 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:

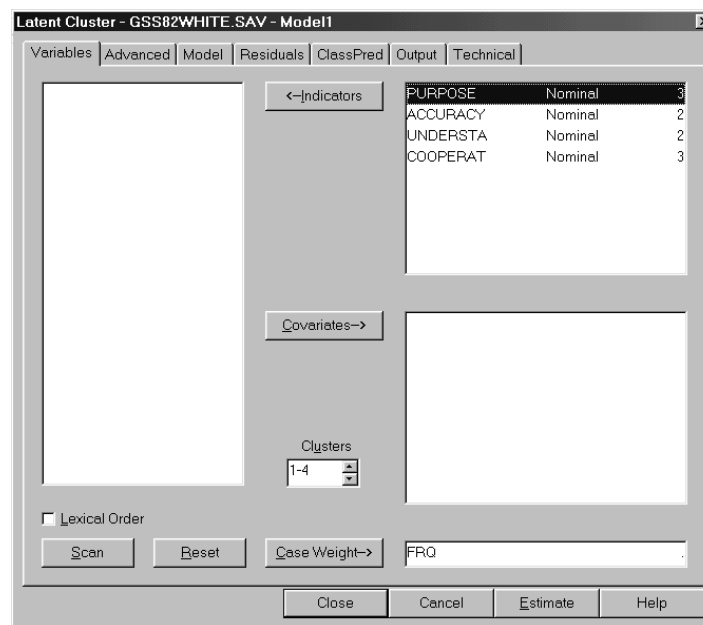


Figure 7-8. Analysis Dialog Box Prior to Model Estimation.

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

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

VIEWING OUTPUT AND INTERPRETING RESULTS

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

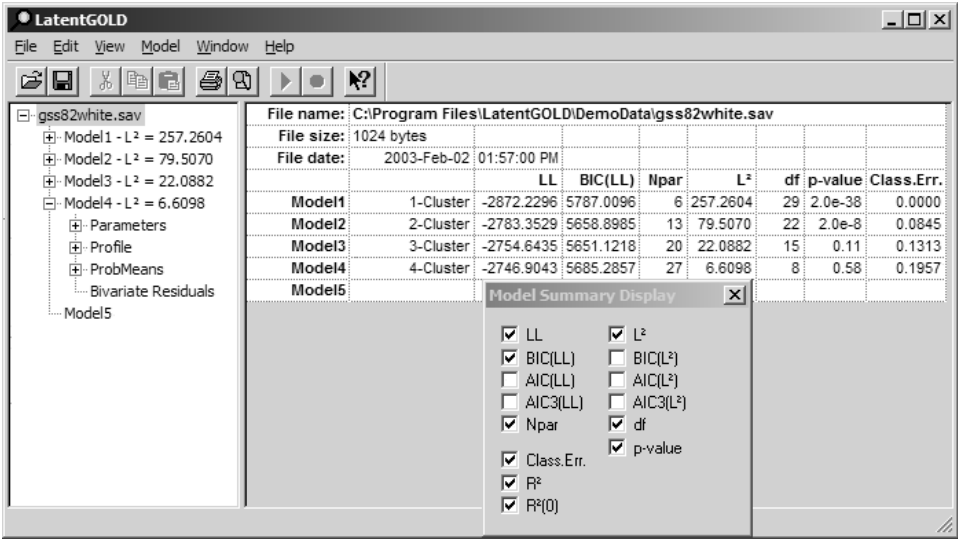


Figure 7-9: Model Summary Output and Model Summary Display

Additional statistics can be displayed by clicking on the associated check-box in the Model Summary Display. The model  $L^2$  statistic, as shown in Figure 7-9 in the column labeled ' $L^2$ ', indicates the amount of the association among the variables that remains unexplained after estimating the model; the lower the value, the better the fit of the model to the data. One criteria for determining the number of clusters is to look in the 'p-value' column which provides the p-value

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

## ASSESSING MODEL FIT USING THE BOOTSTRAP P-VALUE

Latent GOLD offers an alternative option to assess your model using the bootstrap of  $L^2$  to estimate the 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

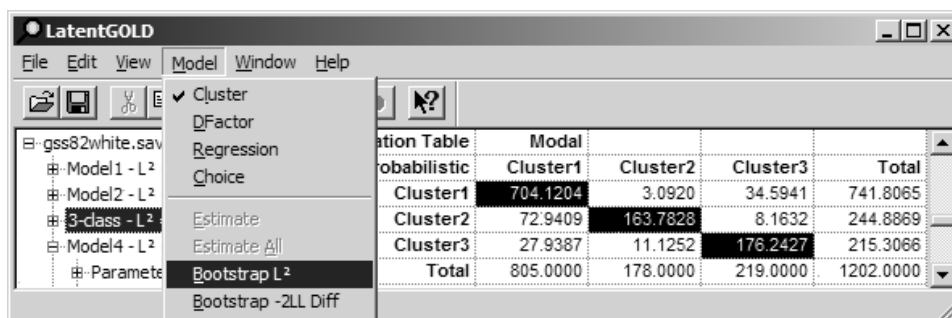


Figure 7-10: Model Menu Showing Option for Bootstrap of  $L^2$

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.

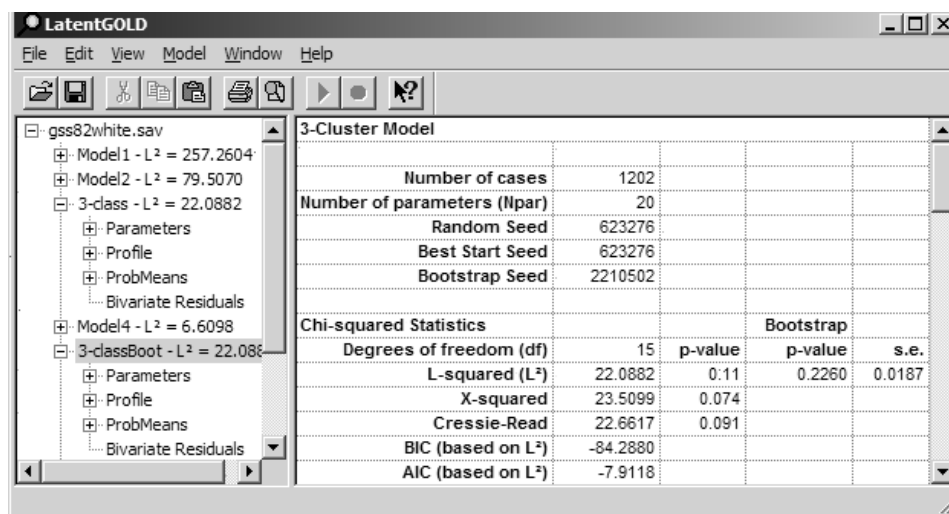


Figure 7-11: Results for the Bootstrap of  $L^2$

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.

LatentGOLD

File

Edit

View

Model

Window

Help

gss82white.sav

Model 1 - L<sup>2</sup> = 257.2604

Model 2 - L<sup>2</sup> = 79.5070

3-class - L<sup>2</sup> = 22.0882

Parameters

Profile

ProbMeans

Bivariate Residuals

Model 4 - L<sup>2</sup> = 6.6098

3-classBoot - L<sup>2</sup> = 22.0882

Parameters

Profile

ProbMeans

Bivariate Residuals

Model 6

Models for Indicators

	Cluster1	Cluster2	Cluster3	Wald	p-value	R <sup>2</sup>
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			

Standard Errors

Z Statistic

Std Errs & Z

Wald Statistics

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<sup>2</sup> 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

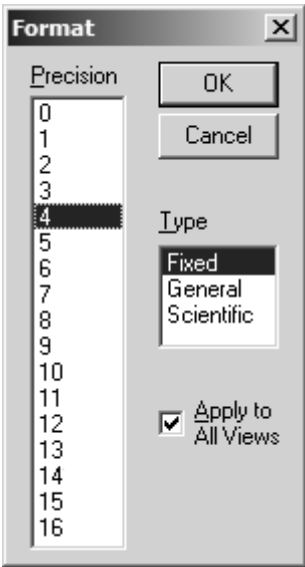


Figure 7-13. Format Control Options from View Menu

PROFILE OUTPUT AND ASSOCIATED PROFILE PLOT



To view the parameters re-expressed as conditional probabilities

- Click on Profile

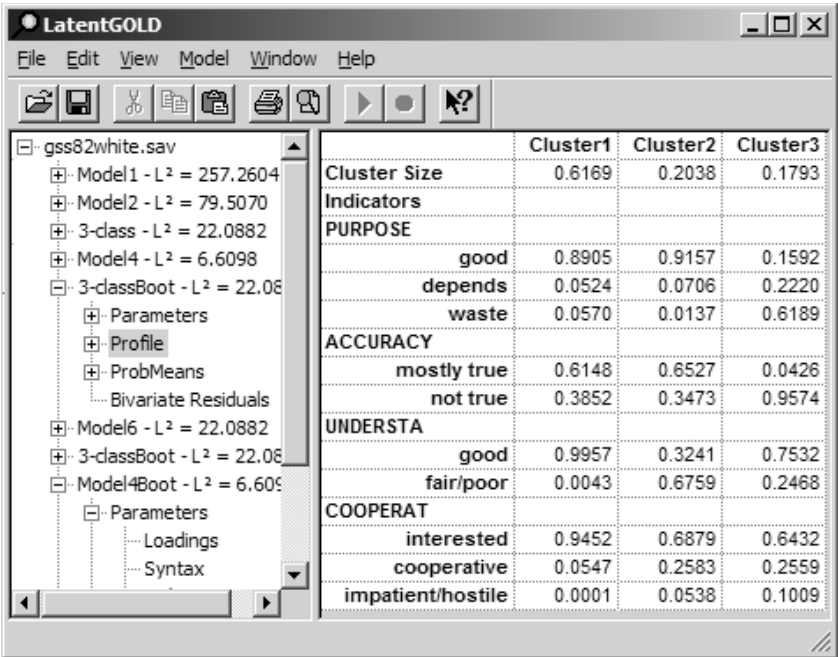


Figure 7-14. Profile Output for 3-cluster Model

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 (ACCURACY = '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 Cluster3) at the bottom of the plot. For example, to highlight the profile for cluster3

- ▷ Click the symbol next to 'Cluster 3'

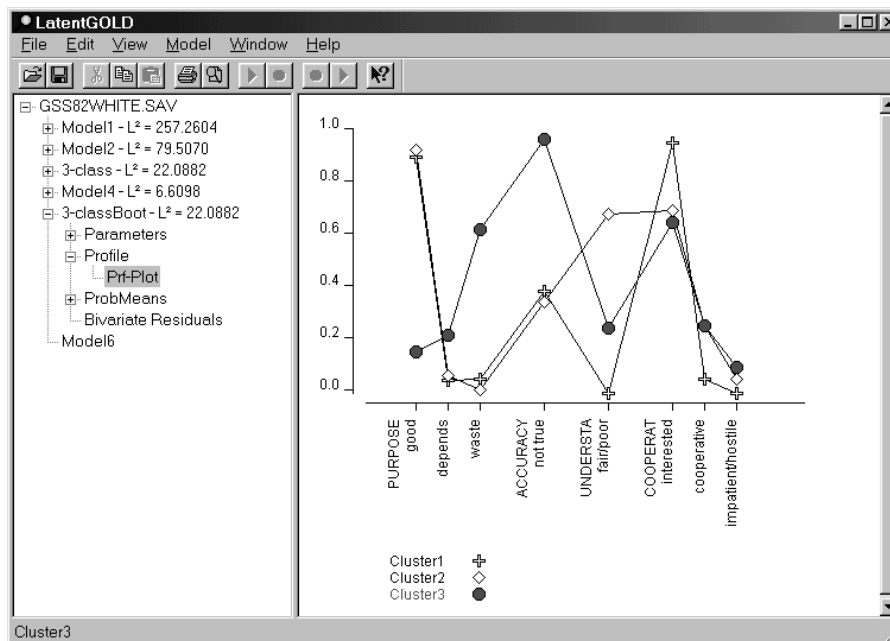


Figure 7-15. Profile Plot for 3-cluster Model

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.

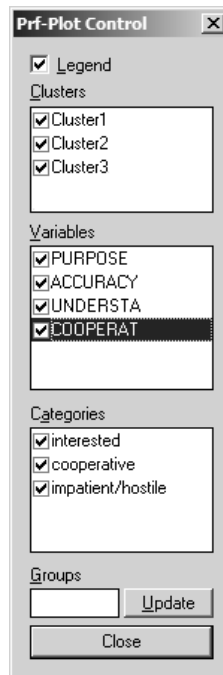


Figure 7-16. Profile Plot Control Panel

## 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

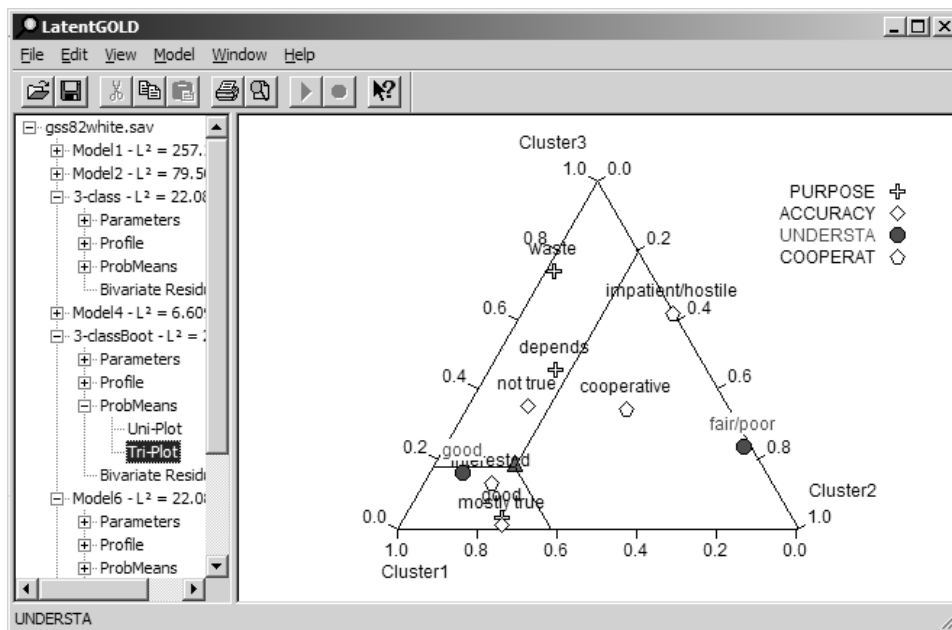


Figure 7-17. Tri-plot Display for 3-class Model

## 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:

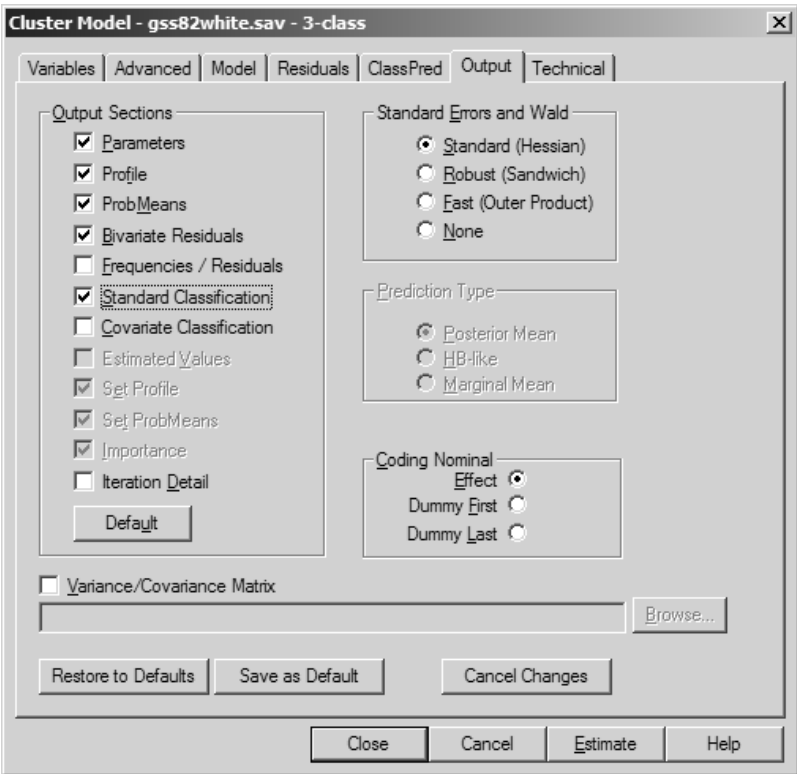


Figure 7-18. Requesting Standard Classification Output Listing in Output Tab

- 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.

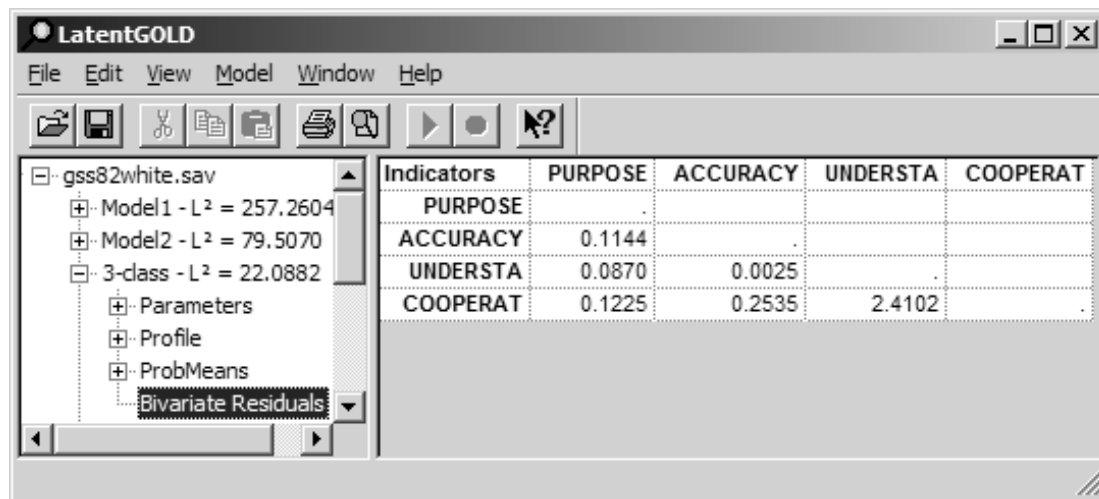


Figure 7-21. Bivariate Residuals Output for the 3-cluster Model

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.

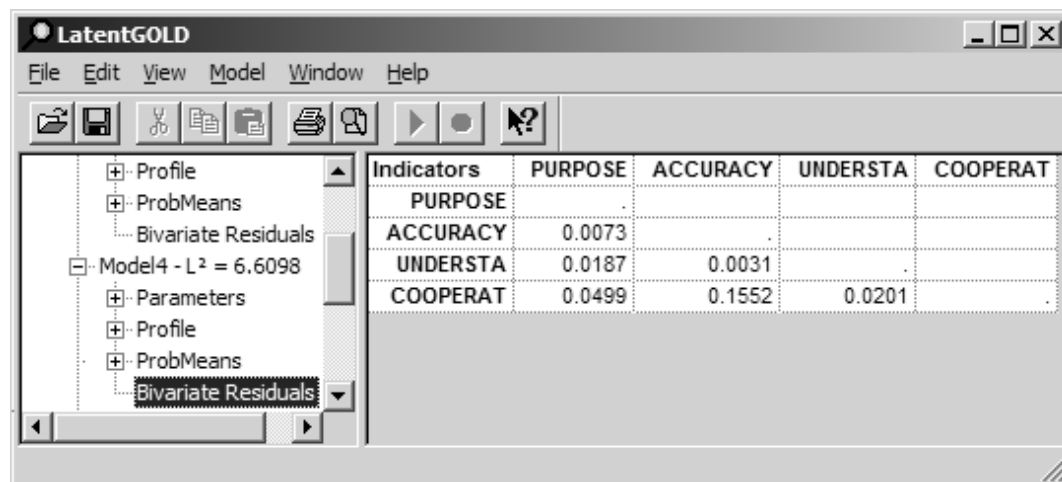


Figure 7-22. Bivariate Residuals Output for the 4-cluster Model

## 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.

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

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.

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

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

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

Alternatively, you can

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

Following this, a list of eligible reference models appears:

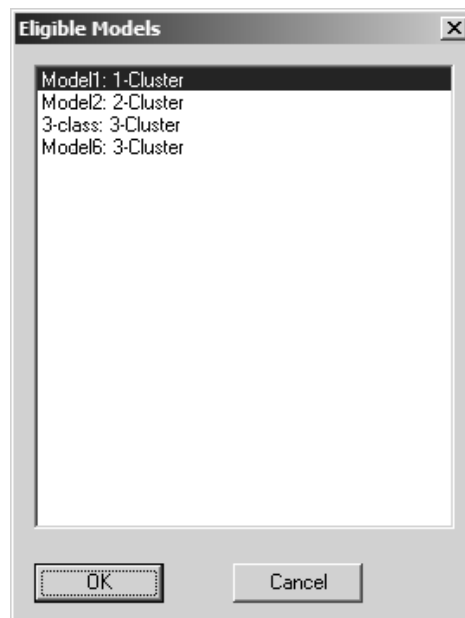


Figure 7-23. List of Eligible Reference Models for Conditional Bootstrap

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

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

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

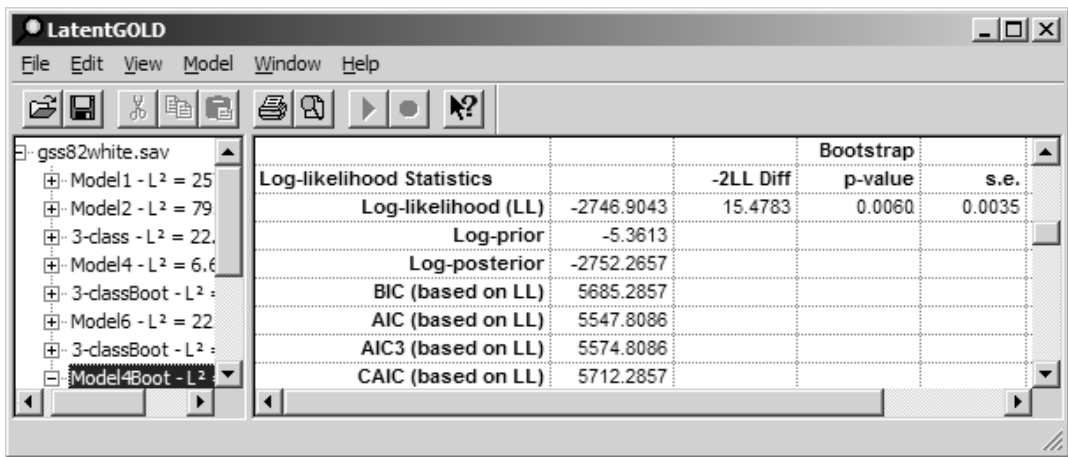


Figure 7-24. Conditional Bootstrap Output

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

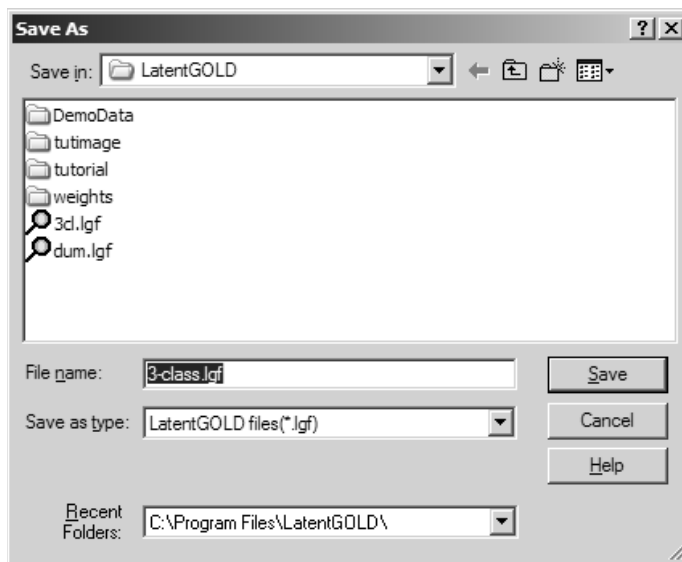


Figure 7-25. File Save Definition Dialog Box

► 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'

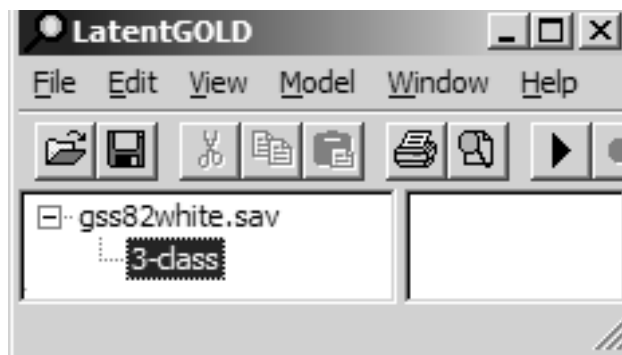


Figure 7-27. Setup for '3-class' Model

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

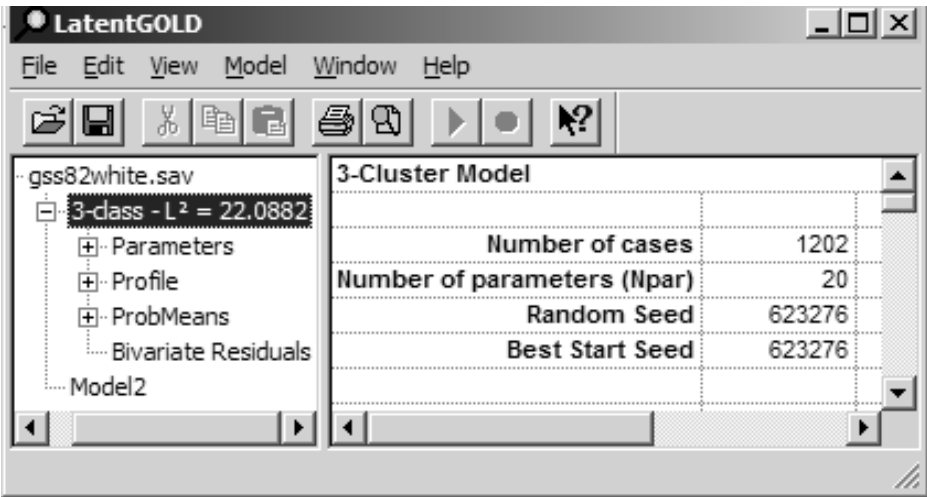


Figure 7-28. Model Summary Output for 3-class Model



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

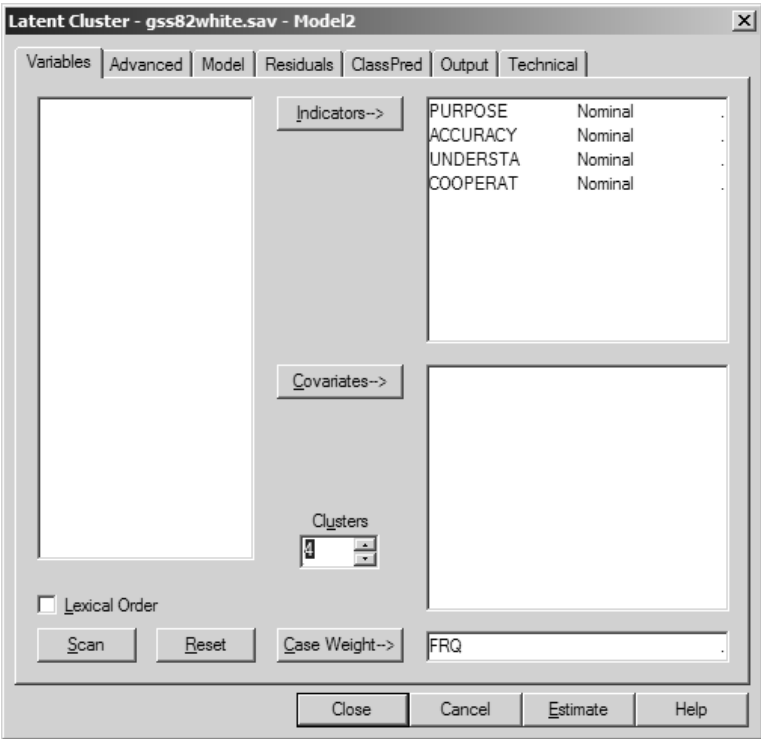


Figure 7-29. Estimating Model2

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'

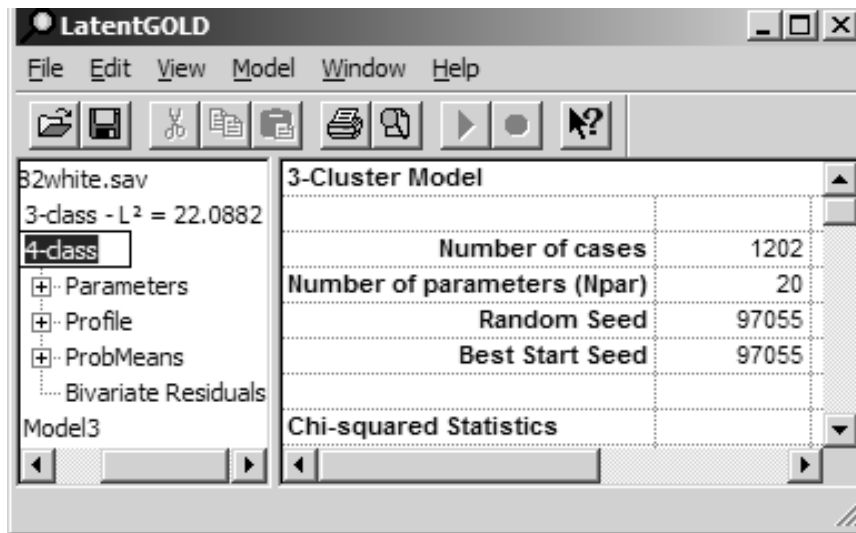


Figure 7-30. Editing the name of Model2

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

	Cluster1	Cluster2	Cluster3	Cluster4	Wald	p-value	R <sup>2</sup>
<b>PURPOSE</b>							
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			
<b>ACCURACY</b>							
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			
<b>UNDERSTA</b>							
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			
<b>COOPERAT</b>							
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

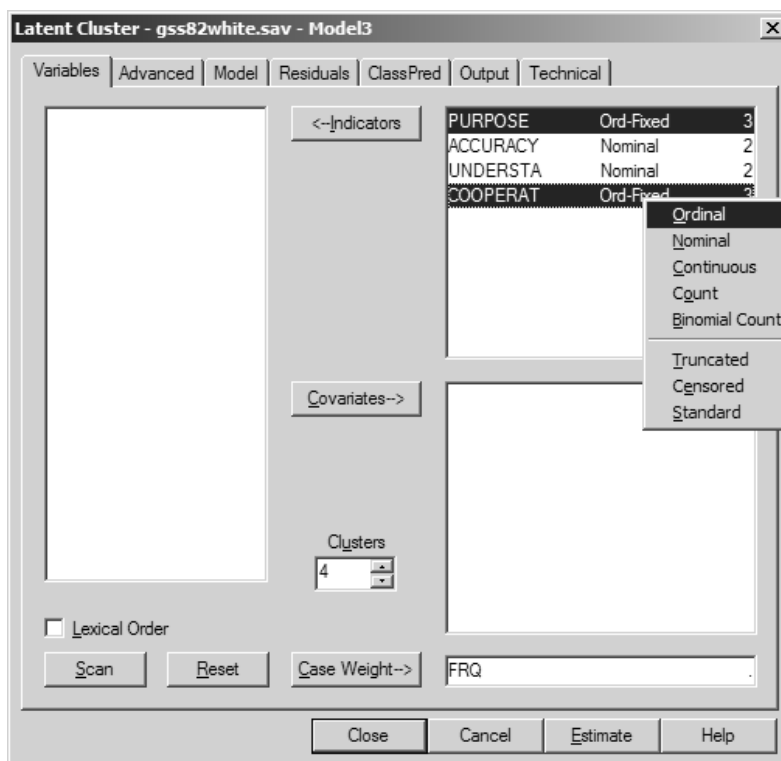


Figure 7-32. Changing PURPOSE and COOPERAT to ordinal variables

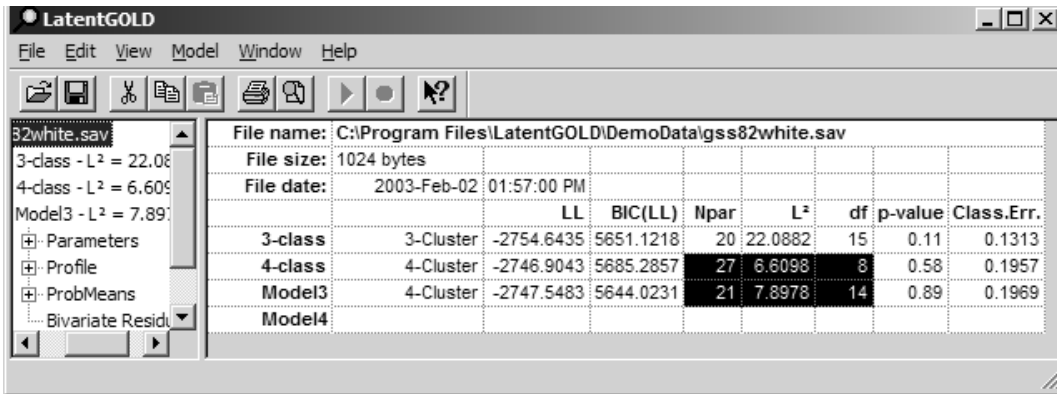
- ▷ Click Estimate to estimate the model



**To compare the results from these models**

- ▷ Click on the data file name in the Outline Pane

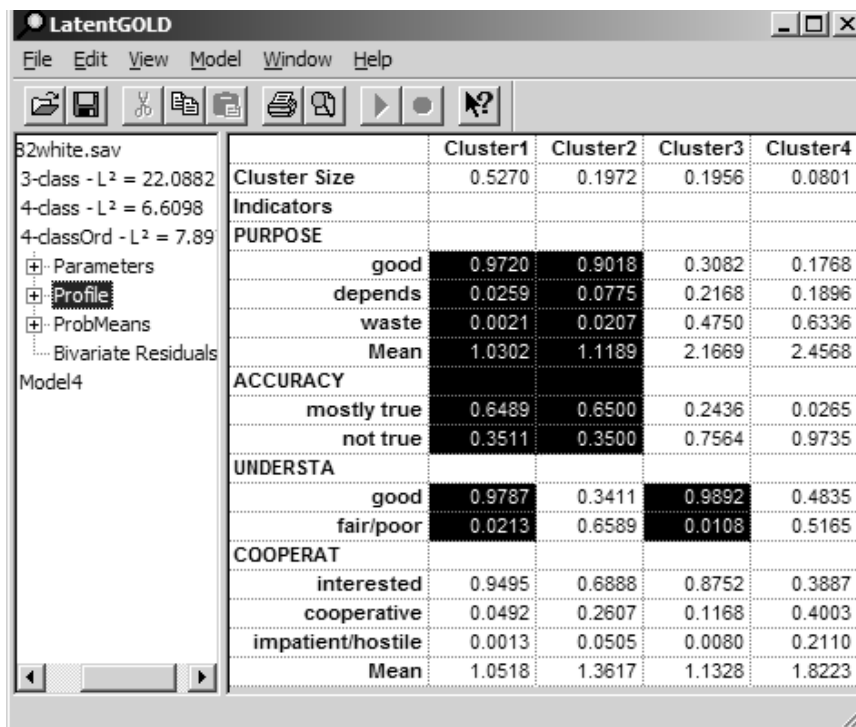
We see that imposing the ordinality restrictions increase  $L^2$  from 6.6 to 7.9, a very small increase associated with the gain of 6 degrees of freedom. Thus, we choose Model3 over the unrestricted 4-class model, which results in a 4-class model with 6 fewer parameters.



		LL	BIC(LL)	Npar	$L^2$	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

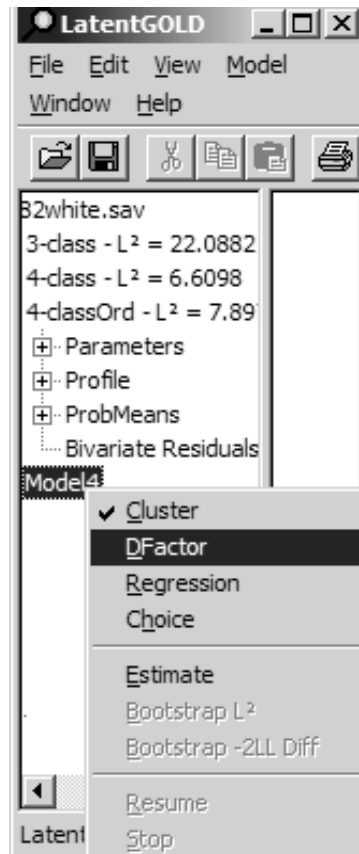


Figure 7-35. Selecting a DFactor Model

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.

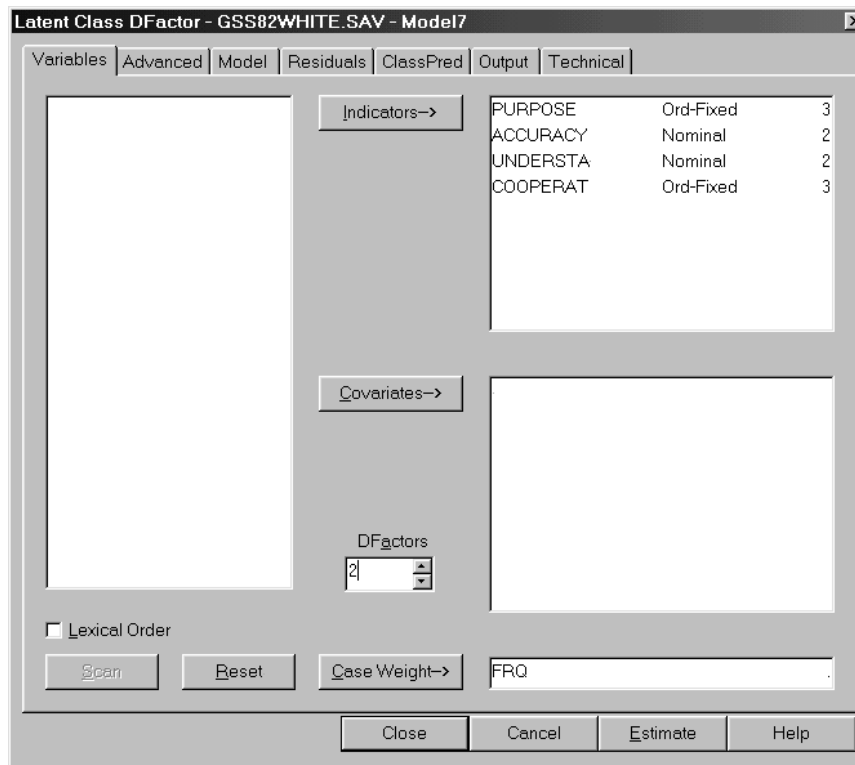


Figure 7-36. Changing number of DFactors

► Click Estimate

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

			LL	BIC(LL)	Npar	L <sup>2</sup>	df	p-value	Class.Err.
gss82white.sav	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	-2750.7421	5614.9520	16	14.2853	19	0.77	0.1284

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

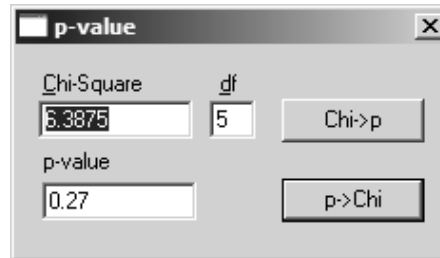


Figure 7-38. Chi-squared Calculator

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.

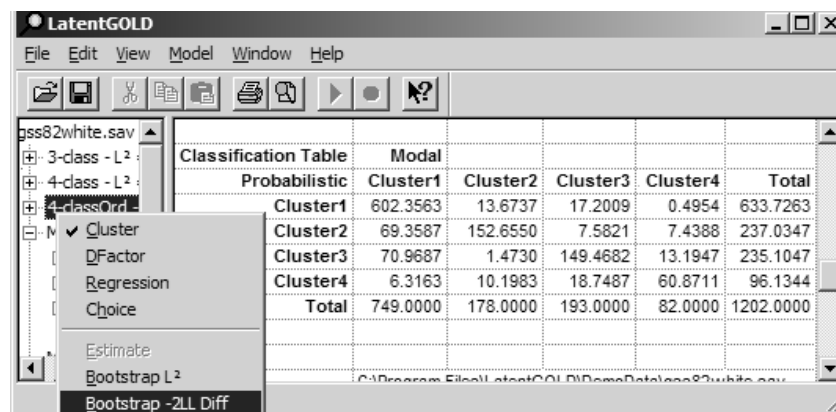


Figure 7-39. Estimating a Bootstrap -2LL Diff Model

Following this, a list of eligible reference models appears.

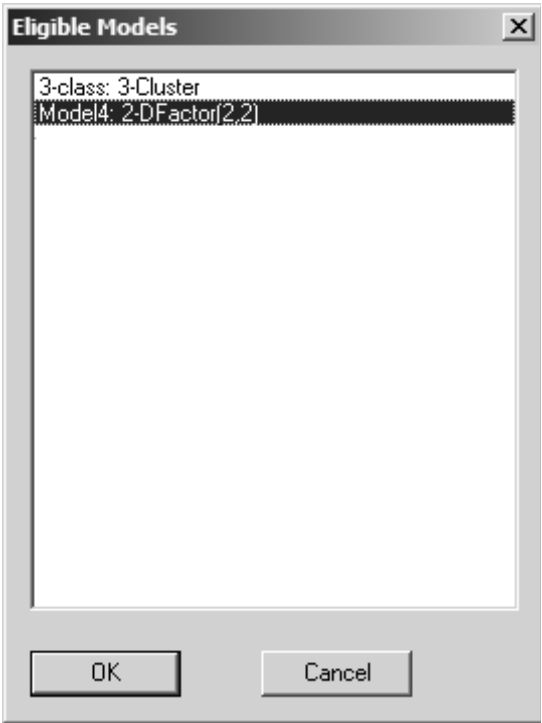


Figure 7-40. List of Eligible Reference Models for Conditional Bootstrap

- ▶ 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.

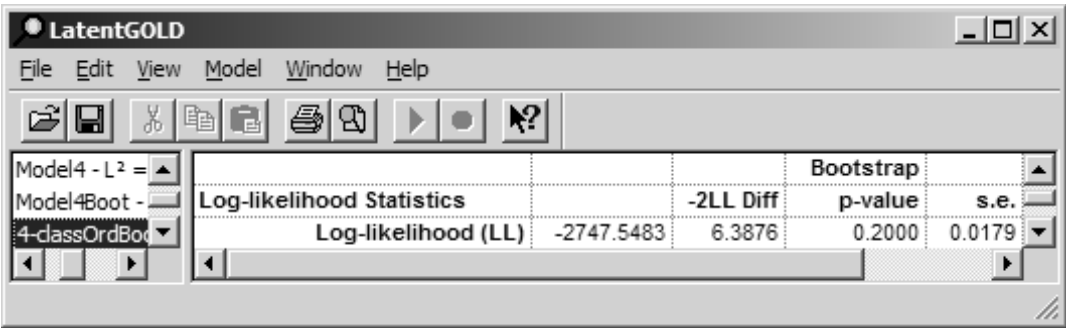


Figure 7-41. p-value Estimated using the Conditional Bootstrap

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.)

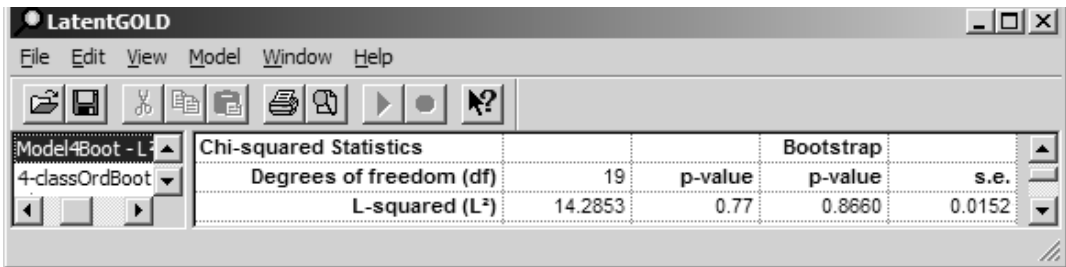


Figure 7-42. p-value Estimated by the Bootstrap for the Reference Model

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

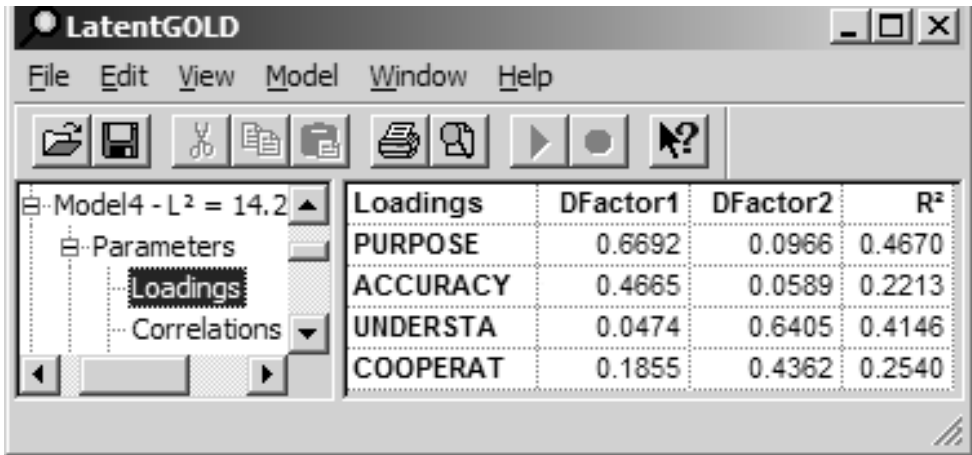


Figure 7-43. Loadings Output

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

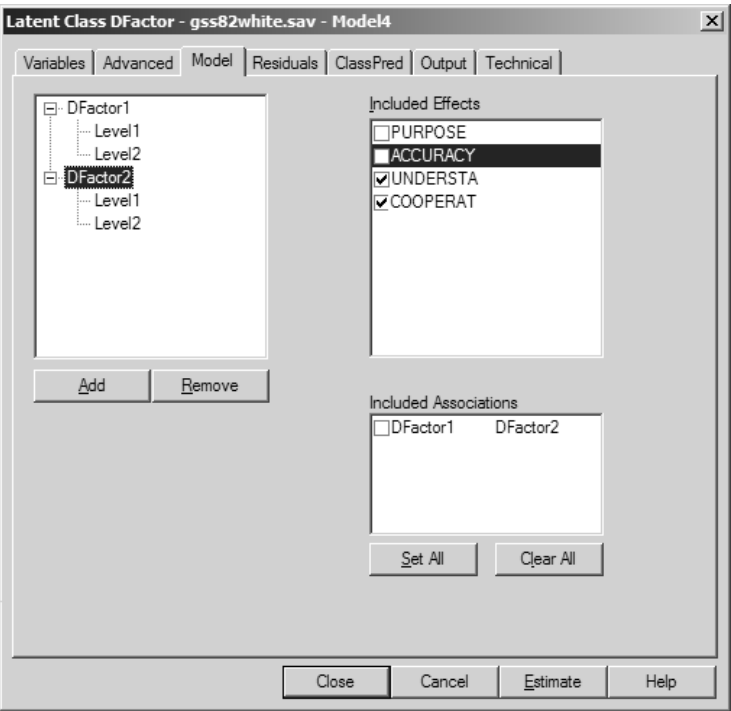


Figure 7-45. Restricting Effects to Zero

➤ Click Estimate

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

➤ Click the data file name in the Outline Pane

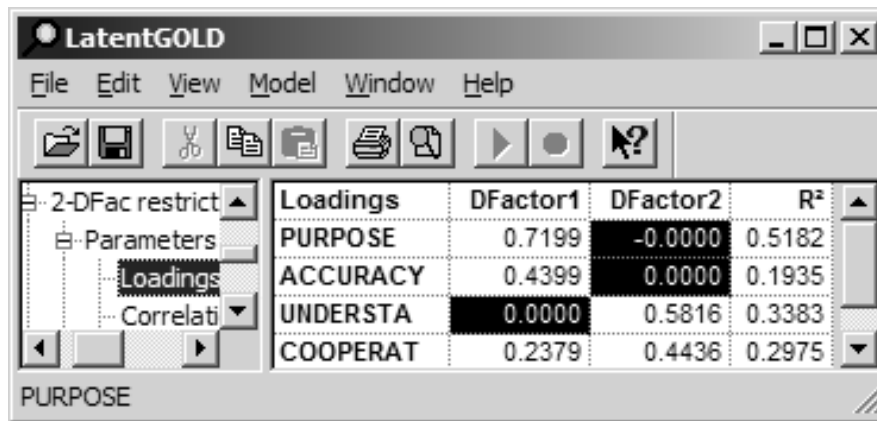
			LL	BIC(LL)	Npar	L <sup>2</sup>	df	p-value	Class.Err.
3-class	3-Cluster		-2754.6435	5651.1218	20	22.0882	15	0.11	0.1313
4-class	4-Cluster		-2746.9043	5685.2857	27	6.6098	8	0.58	0.1957
4-classOrd	4-Cluster		-2747.5483	5644.0231	21	7.8978	14	0.89	0.1969
Model4	2-DFactor(2,2)		-2750.7421	5614.9520	16	14.2853	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<sup>2</sup> 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'



The screenshot shows the LatentGOLD software window. On the left, a tree view shows '2-DFac restrict' expanded, with 'Parameters' and 'Loadings' visible. The main window displays a table of factor loadings for four parameters: PURPOSE, ACCURACY, UNDERSTA, and COOPERAT, across two factors: DFactor1 and DFactor2. The R-squared values are also shown. The table is as follows:

Loadings	DFactor1	DFactor2	R <sup>2</sup>
PURPOSE	0.7199	-0.0000	0.5182
ACCURACY	0.4399	0.0000	0.1935
UNDERSTA	0.0000	0.5816	0.3383
COOPERAT	0.2379	0.4436	0.2975

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).

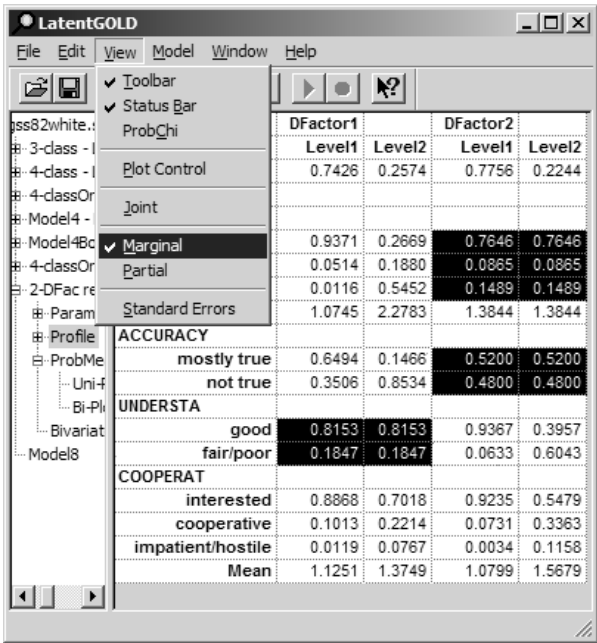


Fig. 7-48. Marginal Profile Output

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

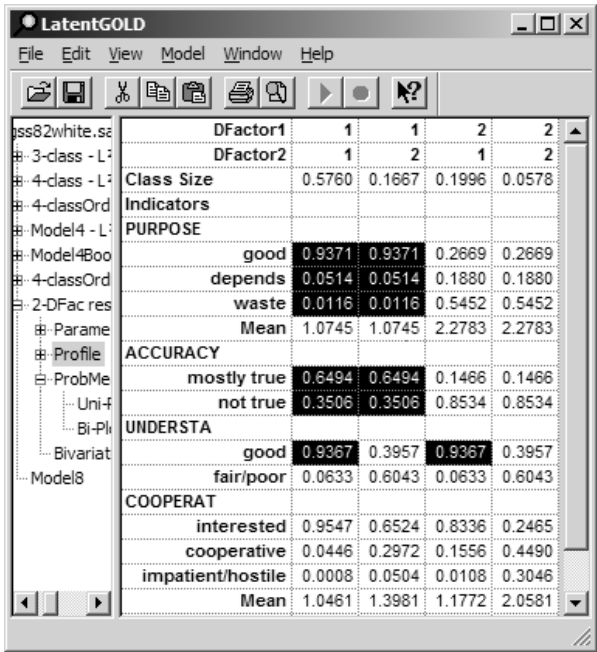


Figure 7-49. Joint Profile Output

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.

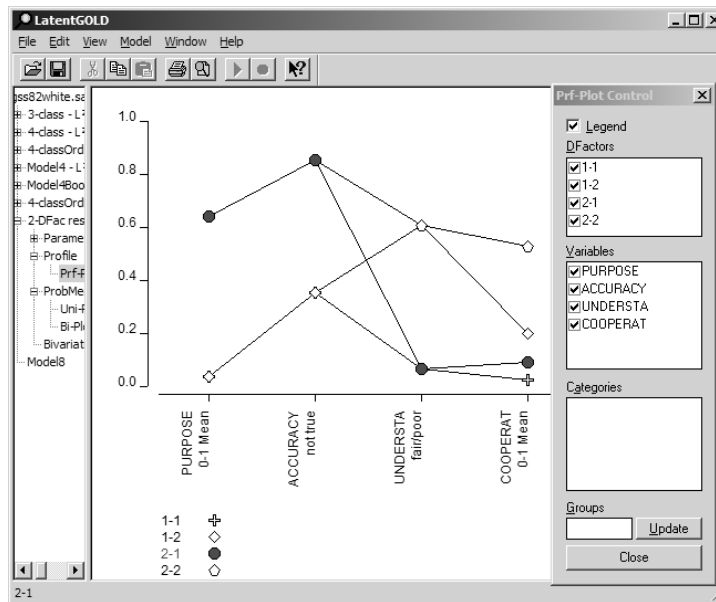


Figure 7-50. Profile Plot

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 Figure 7-51). 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.

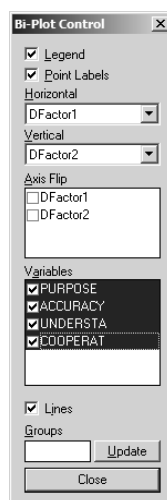


Figure 7-52. Plot Control Menu

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

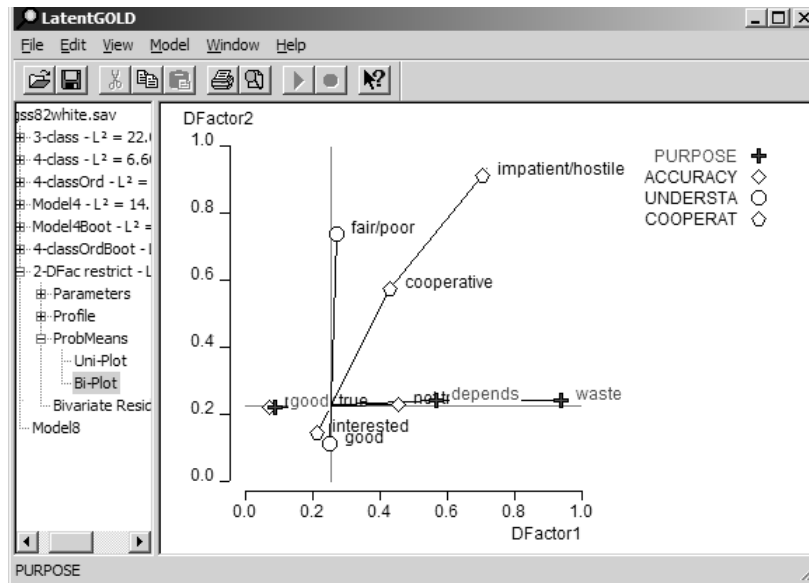


Figure 7-53. Bi-Plot

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.

data4.sav - SPSS Data Editor													
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help													
1 : jfac#1 0.906050102873919													
	frq	fac1#1	fac1#2	fac1#1	fac1#2	fac2#1	fac2#2	fac2#1	fac2#2	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.

	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

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.

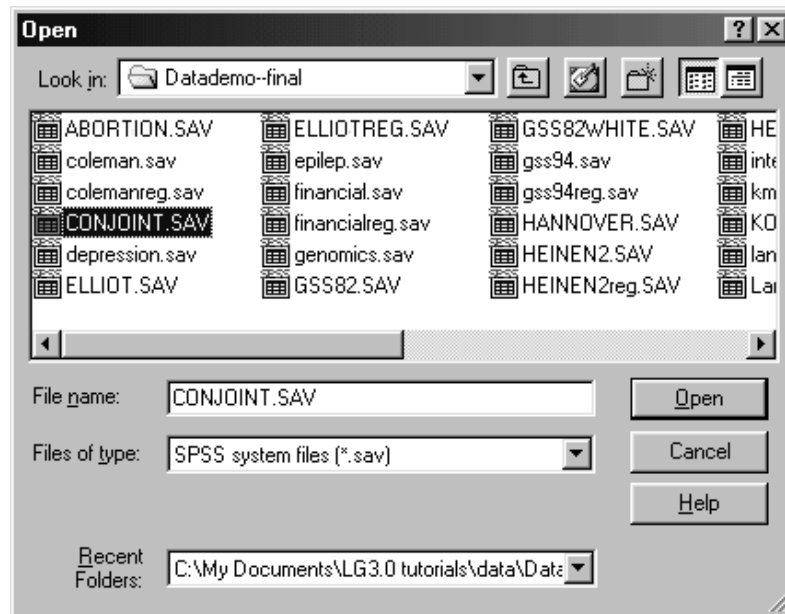


Figure 7-56: Open Dialog Box

- ▷ 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).

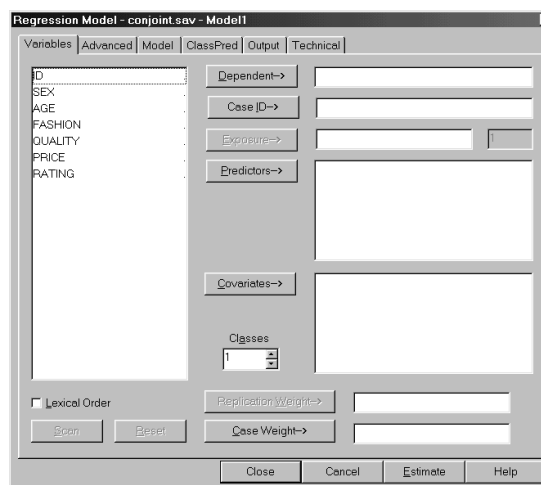


Figure 7-57: Analysis Dialog Box for LC Regression Model

## 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.

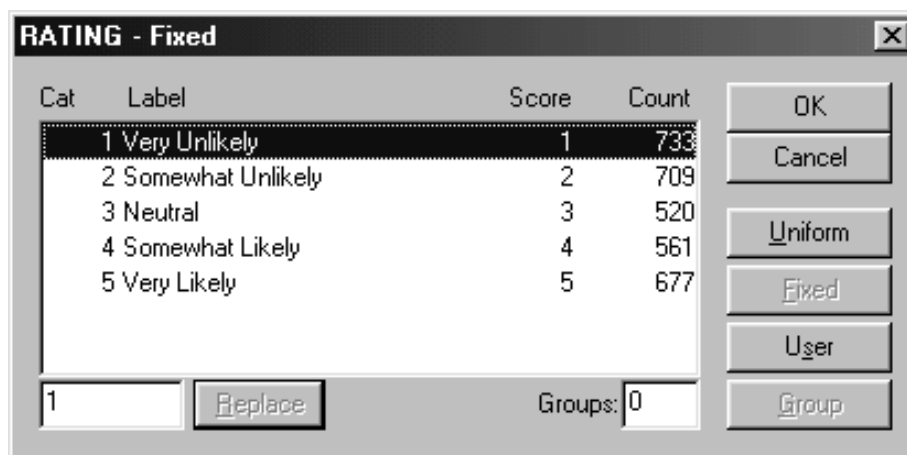


Figure 7-58: Variables box for RATING

- ▷ 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.

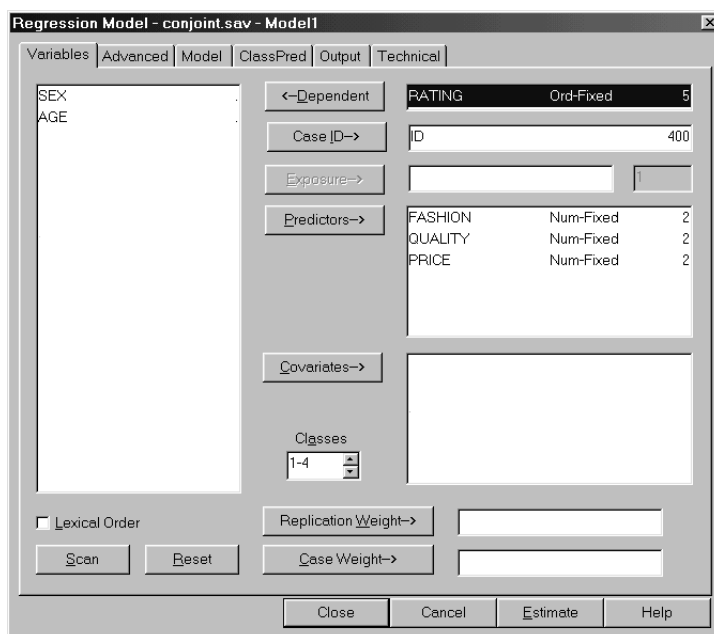


Figure 7-59: Regression Analysis Dialog Box with Initial Settings

- 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.

File name: C:\My Documents\LG3.0 tutorials\data\Datademo--final\CONJOINT.SAV										
File size: 39420 bytes										
File date: 2003-Jan-09 17:11:42										
		LL	BIC(LL)	lpar	L <sup>2</sup>	df	p-value	Class.Err.	R <sup>2</sup>	
<b>Model1</b>	1-Class Regression	-4402.1081	8846.1565	7	4027.6801	390617	1.00	0.0000	0.3726	
<b>Model2</b>	2-Class Regression	-4114.3853	8318.6425	15	3452.2344	390609	1.00	0.0399	0.5879	
<b>Model3</b>	3-Class Regression	-4087.1265	8312.0566	23	3397.7168	390601	1.00	0.1224	0.6143	
<b>Model4</b>	4-Class Regression	-4075.9222	8337.5798	31	3375.3083	390593	1.00	0.1246	0.6216	
<b>Model5</b>										

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).

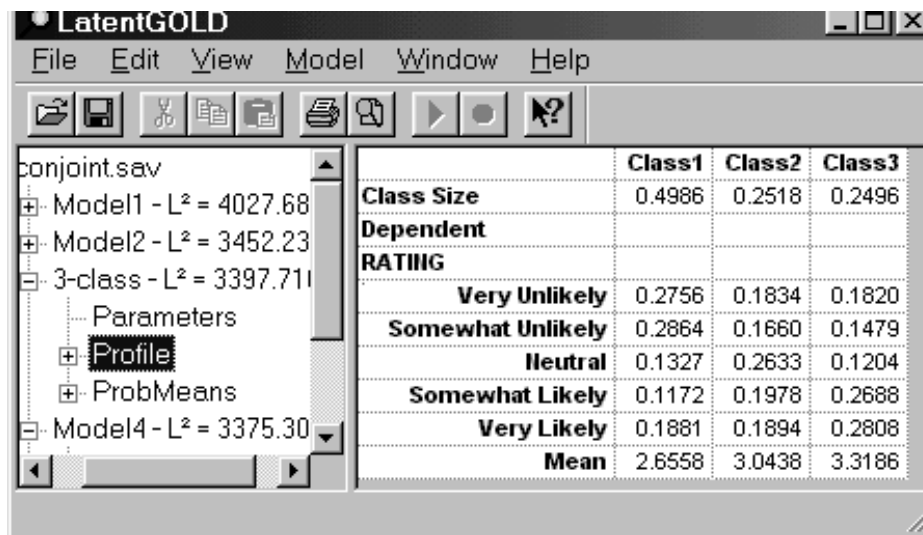
**Note 2:** Notice that the p-values based on the model  $L^2$  and reported degrees of freedom (df) are 1.0 under each model. These are not valid assessments of fit because we are dealing with sparse data.

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).



	Class1	Class2	Class3
<b>Class Size</b>	0.4986	0.2518	0.2496
<b>Dependent RATING</b>			
Very Unlikely	0.2756	0.1834	0.1820
Somewhat Unlikely	0.2864	0.1660	0.1479
Neutral	0.1327	0.2633	0.1204
Somewhat Likely	0.1172	0.1978	0.2688
Very Likely	0.1881	0.1894	0.2808
Mean	2.6558	3.0438	3.3186

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

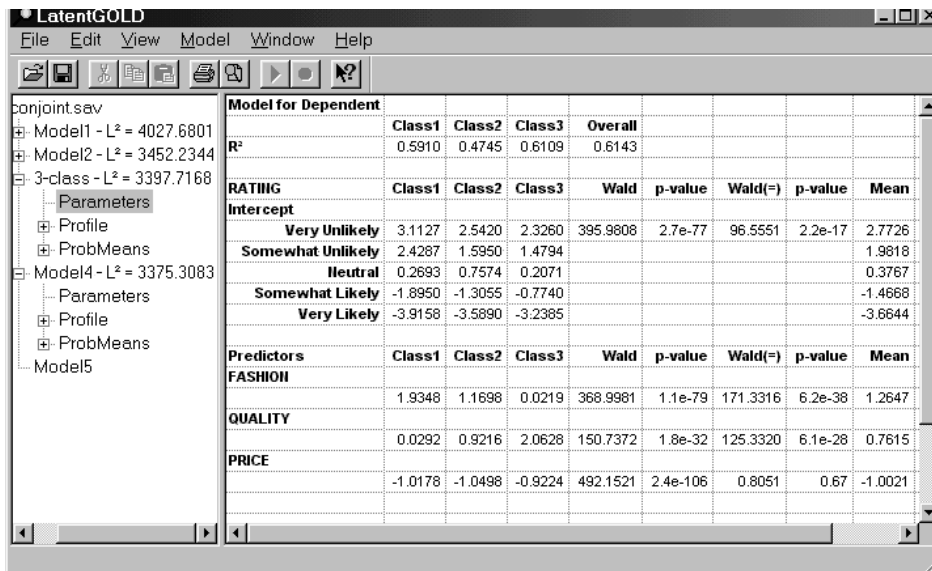
The classes are always ordered from high to low according to their size. It can be seen from the first row of the table that segment 1 contains about 50% of the subjects (.4986), segment 2 contains about 25% and segment 3 contains the remaining 25%.

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				
R²	0.5910	0.4745	0.6109	0.6143				
<b>RATING</b>								
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
<b>Predictors</b>								
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), high quality (beta = .9216) and higher priced (beta = 1.0498). Members of segment 3 prefer high quality (beta = 2.0628) and the higher priced (beta = .9224) product choices, but 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 =  $6.2 \times 10^{-36}$ .

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	
<b>R<sup>2</sup></b>	0.5910		0.4745		0.6109		0.6143	
<b>RATING</b>								
<b>Intercept</b>								
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
<b>Predictors</b>								
<b>FASHION</b>								
	1.9348	17.8803	1.1698	8.0618	0.0219	0.2000	1.2647	0.7821
<b>QUALITY</b>								
	0.0292	0.4085	0.9216	6.1328	2.0628	11.4726	0.7615	0.8345
<b>PRICE</b>								
	-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'.

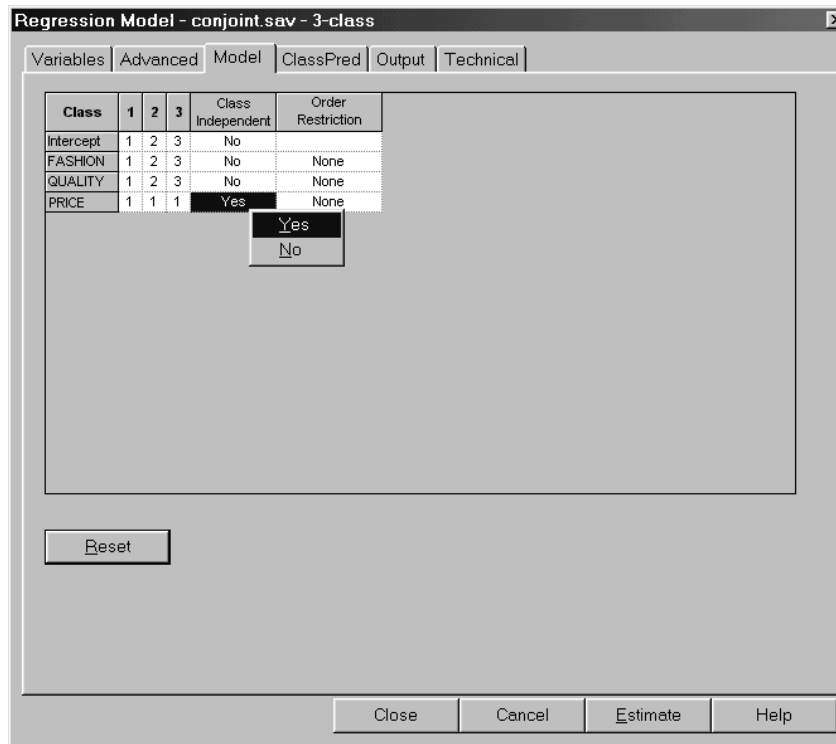


Figure 7-64. Setting the Class-Independent effects for the variable PRICE



To specify the betas to be zero,

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

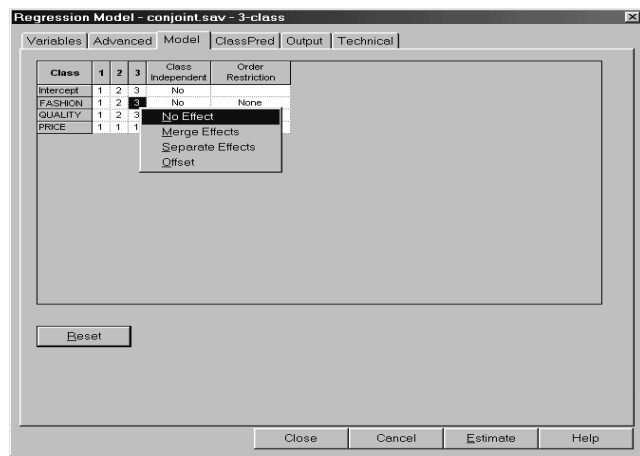


Figure 7-65. Specifying betas to 0 for variable FASHION

- 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.

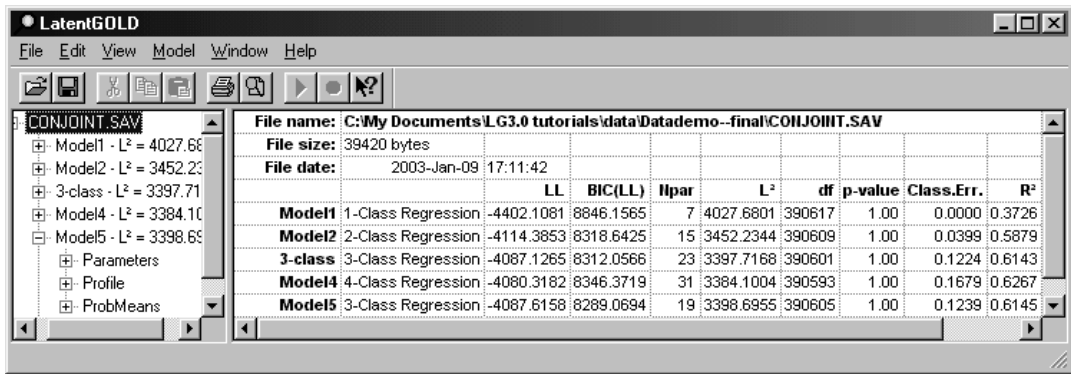


Figure 7-66: Summary Output

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.

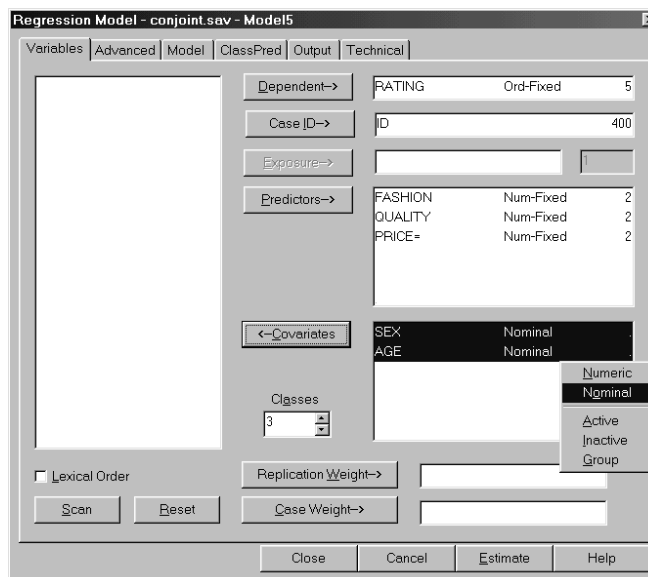


Figure 7-67: Specification of Model with Covariates

- ▷ 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.

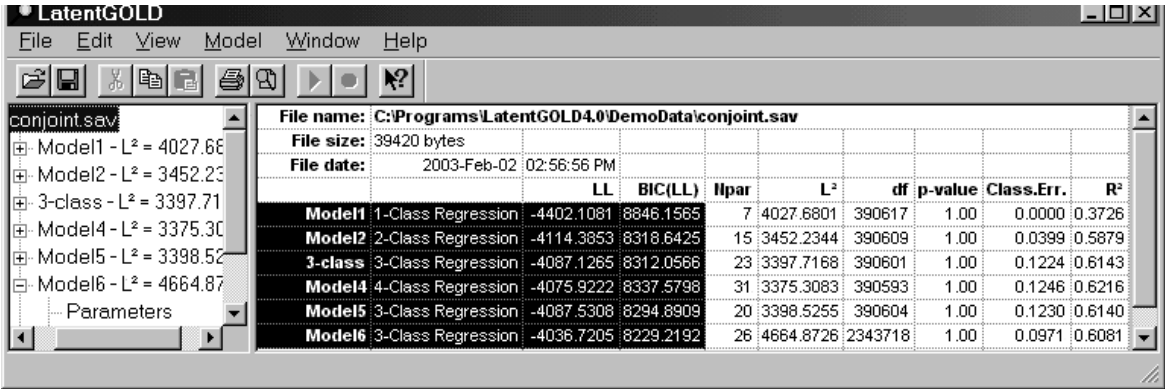


Figure 7-68: Summary Output

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
R²	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**

	Class1	Class2	Class3	Wald	p-value
Intercept	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.



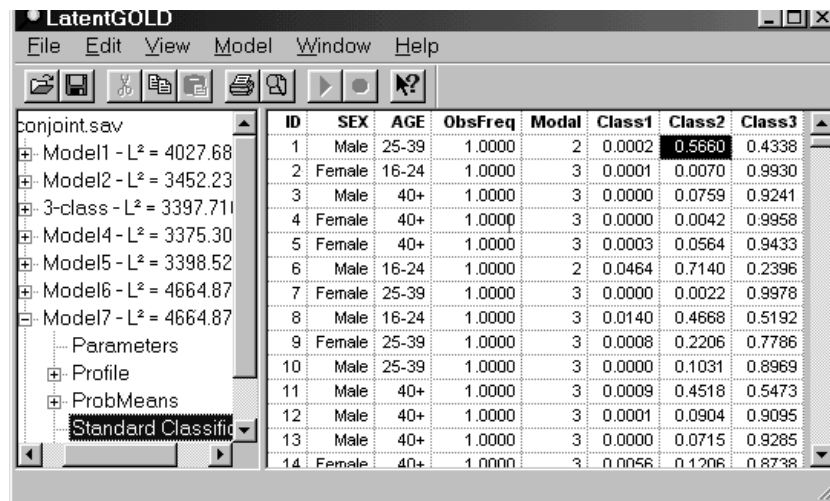
To view the standard classification output for Model 6,

- ▷ 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.



ID	SEX	AGE	ObsFreq	Modal	Class1	Class2	Class3
1	Male	25-39	1.0000	2	0.0002	0.5660	0.4338
2	Female	16-24	1.0000	3	0.0001	0.0070	0.9930
3	Male	40+	1.0000	3	0.0000	0.0759	0.9241
4	Female	40+	1.0000	3	0.0000	0.0042	0.9958
5	Female	40+	1.0000	3	0.0003	0.0564	0.9433
6	Male	16-24	1.0000	2	0.0464	0.7140	0.2396
7	Female	25-39	1.0000	3	0.0000	0.0022	0.9978
8	Male	16-24	1.0000	3	0.0140	0.4668	0.5192
9	Female	25-39	1.0000	3	0.0008	0.2206	0.7786
10	Male	25-39	1.0000	3	0.0000	0.1031	0.8969
11	Male	40+	1.0000	3	0.0009	0.4518	0.5473
12	Male	40+	1.0000	3	0.0001	0.0904	0.9095
13	Male	40+	1.0000	3	0.0000	0.0715	0.9285
14	Female	40+	1.0000	3	0.0056	0.1206	0.8738

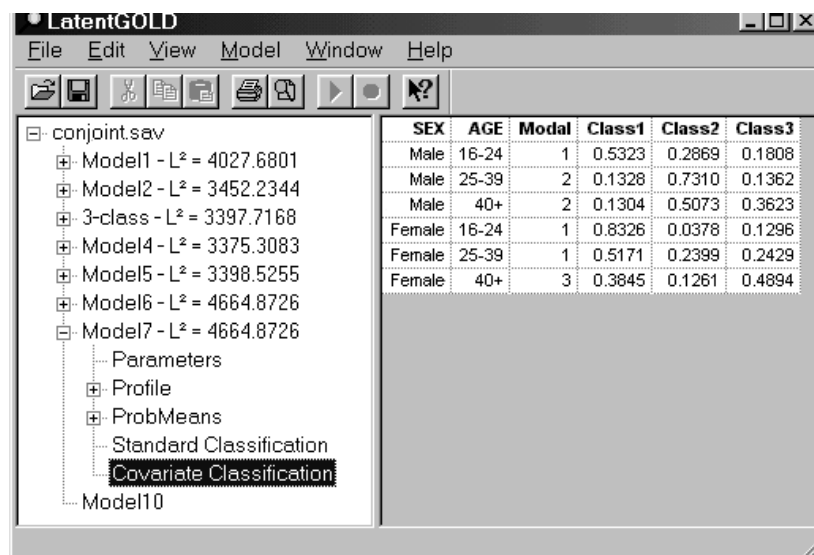
Figure 7-70: Classification Output for Model 7 (partial listing)

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 your 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.



SEX	AGE	Modal	Class1	Class2	Class3
Male	16-24	1	0.5323	0.2869	0.1808
Male	25-39	2	0.1328	0.7310	0.1362
Male	40+	2	0.1304	0.5073	0.3623
Female	16-24	1	0.8326	0.0378	0.1296
Female	25-39	1	0.5171	0.2399	0.2429
Female	40+	3	0.3845	0.1261	0.4894

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 'Modell1' 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:

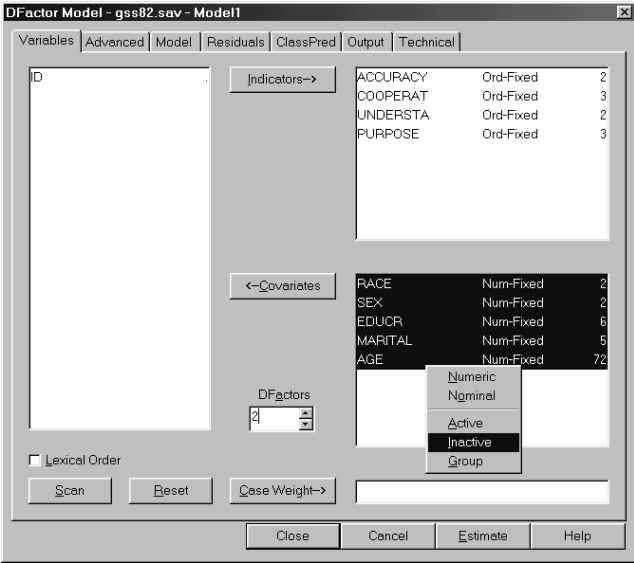


Figure 7-73. Making Covariates Inactive

- ▷ 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:

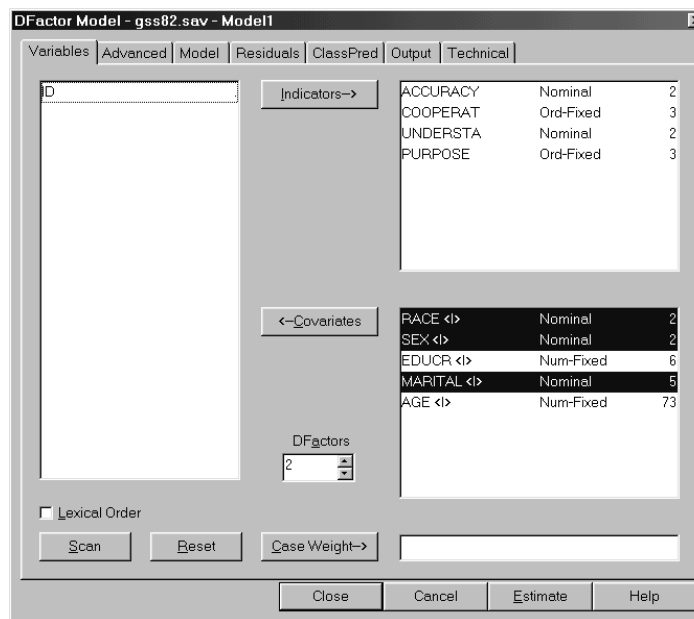


Figure 7-74. Analysis Dialog Box after adding covariates

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.

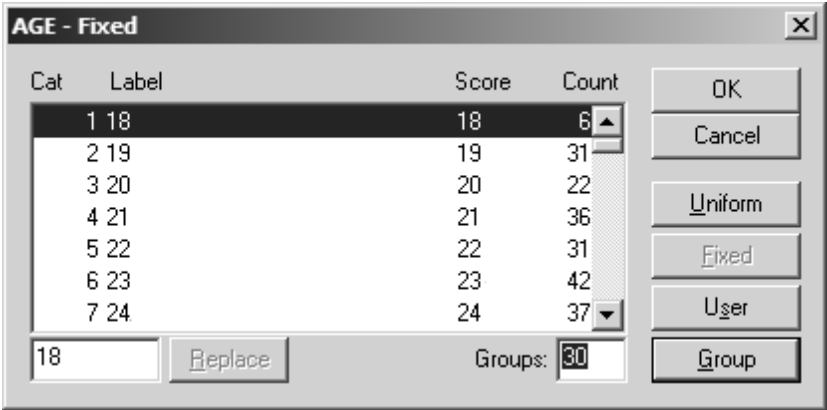


Figure 7-75. Levels for the variable AGE



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.

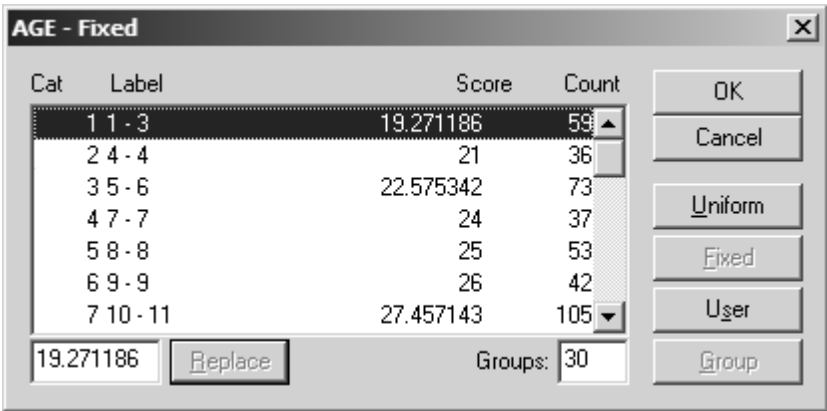


Figure 7-76. Grouped AGE variable

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.

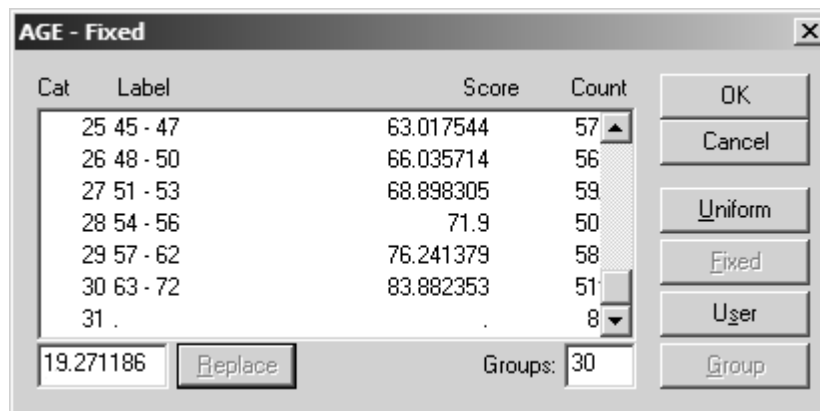


Figure 7-77. Grouped AGE variable

- ▷ 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.

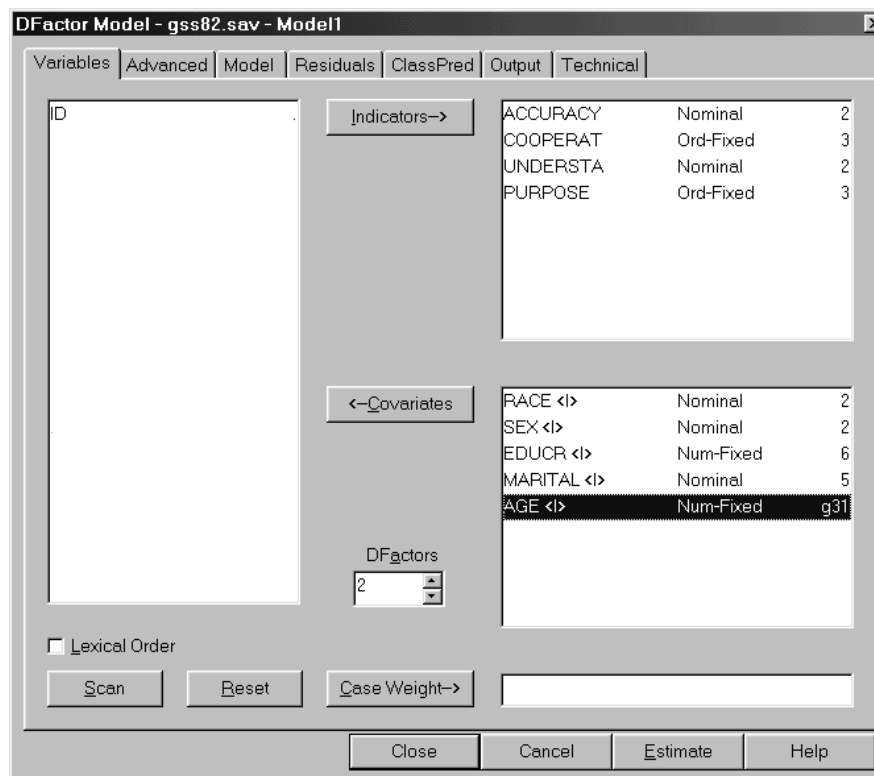


Figure 7-78. Analysis Dialog Box after Grouping Age



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:

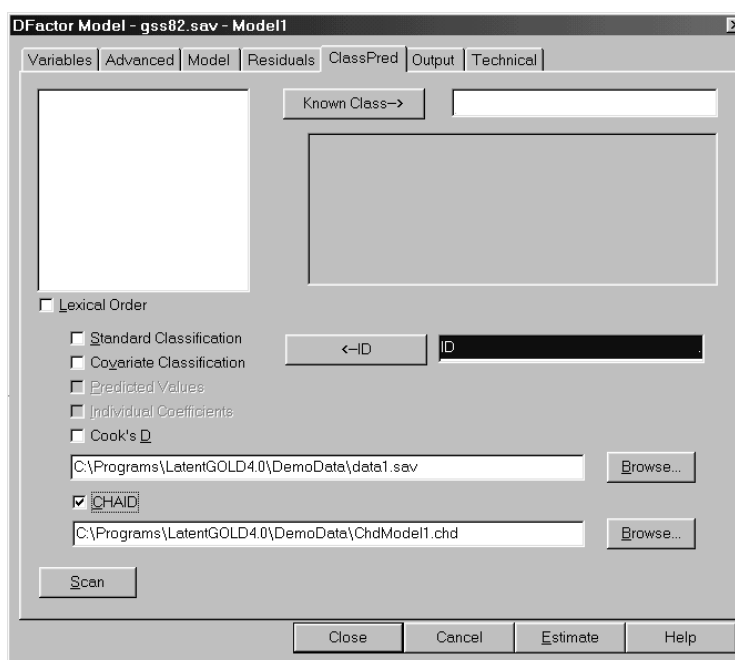


Figure 7-79. ClassPred Tab

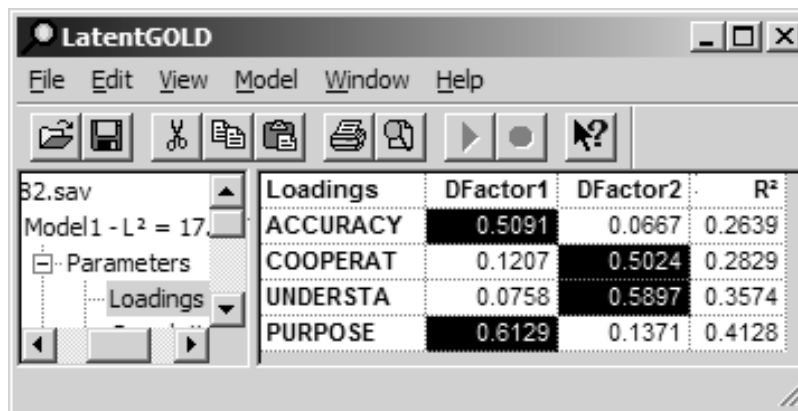
## ESTIMATING THE MODEL

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



Loadings	DFactor1	DFactor2	R²
ACCURACY	0.5091	0.0667	0.2639
COOPERAT	0.1207	0.5024	0.2829
UNDERSTA	0.0758	0.5897	0.3574
PURPOSE	0.6129	0.1371	0.4128

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

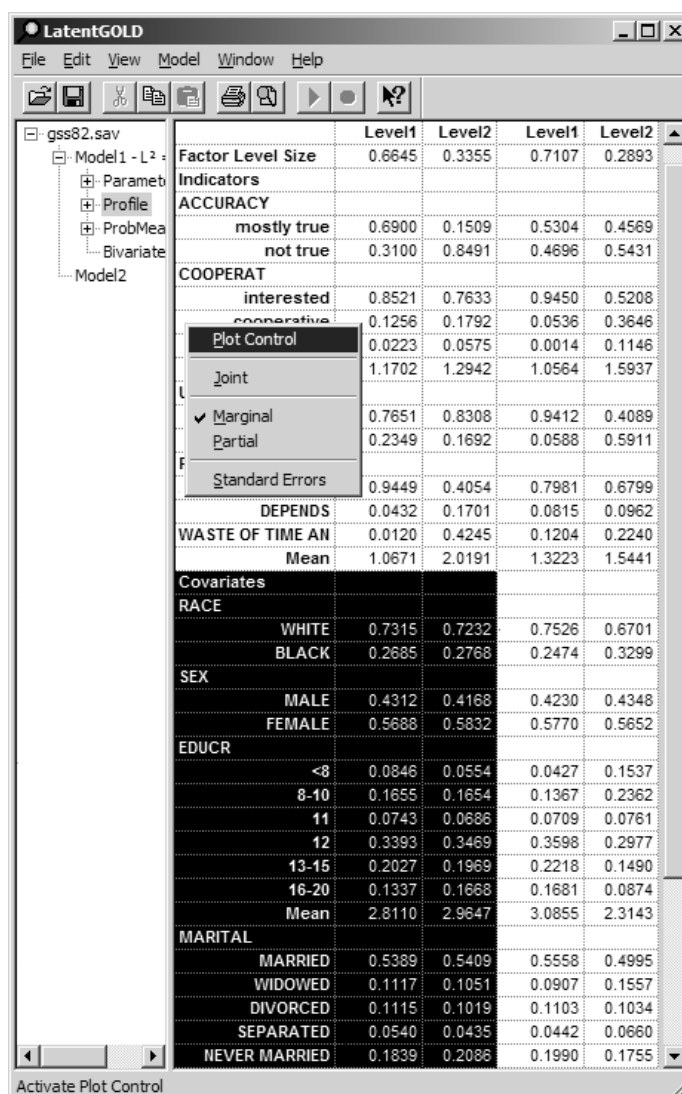


Figure 7-81. Marginal Profile Output

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

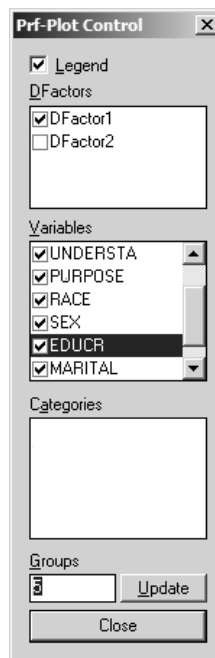


Figure 7-82. Profile Plot Control

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

LatentGOLD					
File Edit View Model Window Help					
B2.sav					
Model1 - L <sup>2</sup> = 17.34					
Parameters					
Profile					
ProbMeans					
Bivariate Residuals					
Model2					
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	0.0688	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	

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.

- ▷ Open the CHAID Define program

From the File Menu

- ▷ Select Open

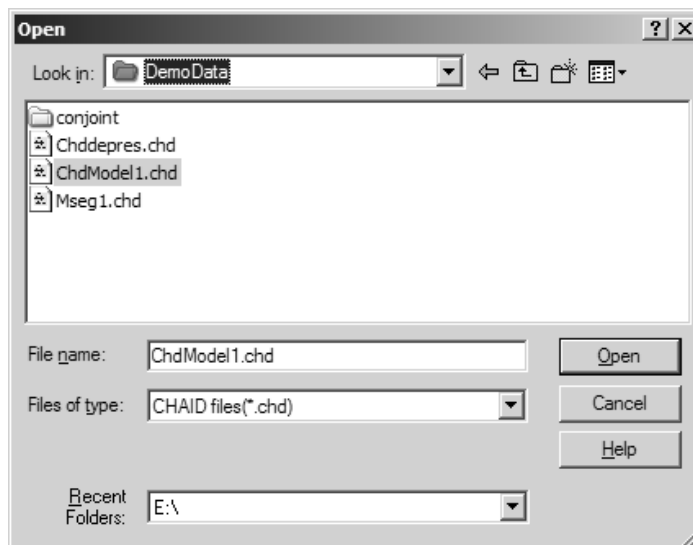


Figure 7-84. Open Dialog Box

Alternatively,

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

The Define program opens. 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.

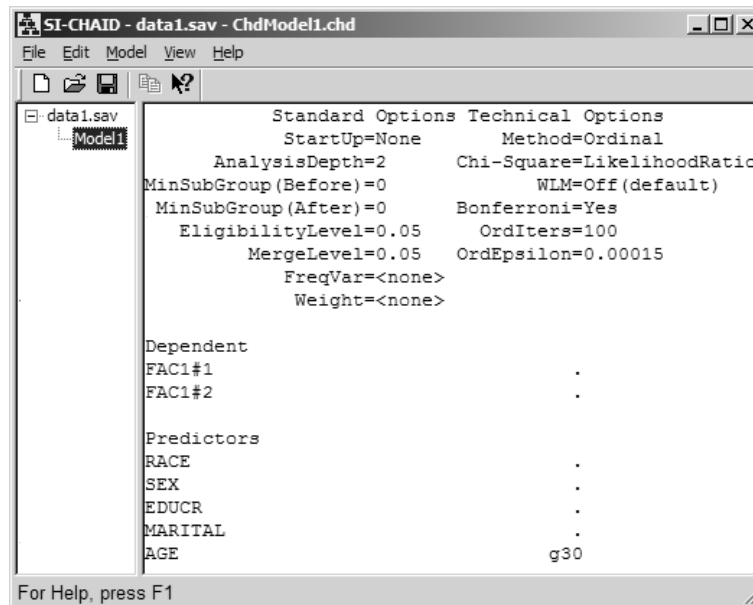


Figure 7-85. Summary for Model1 in Chaid Define

- 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.)

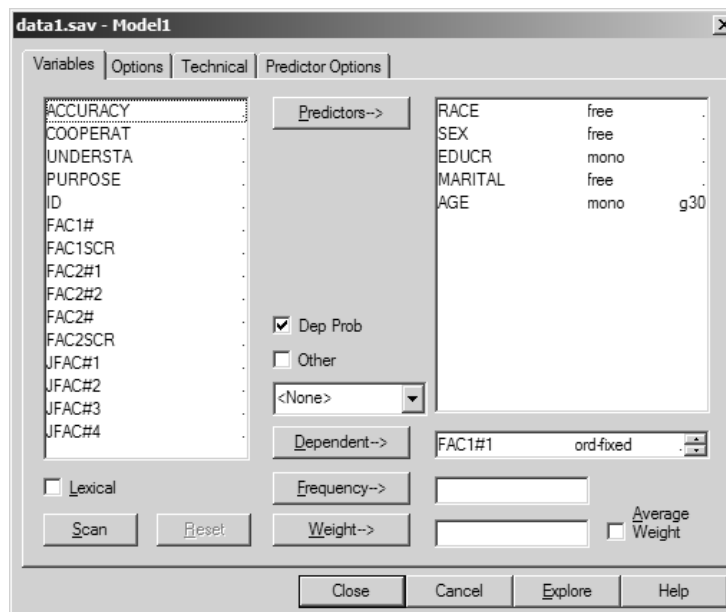


Figure 7-86. The Analysis Dialog Box

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

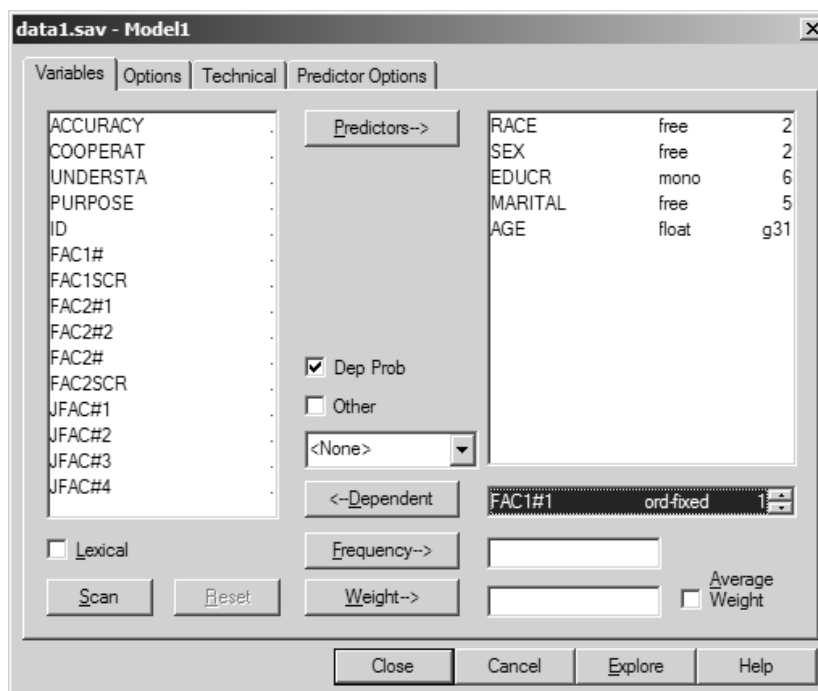


Figure 7-87. Analysis Dialog Box after Scanning

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

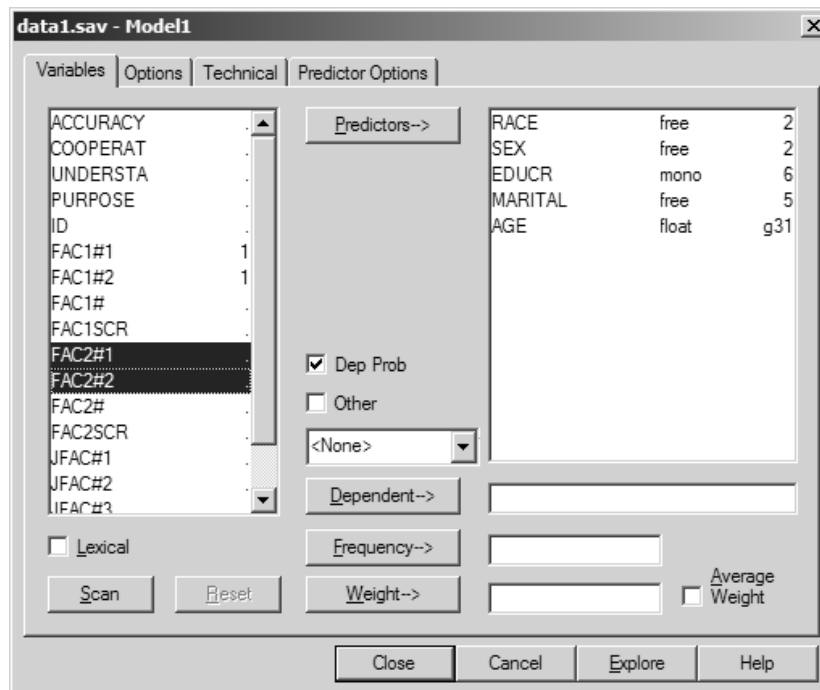


Figure 7-88. Moving FAC2#1 and FAC2#2 to Dependent Box

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

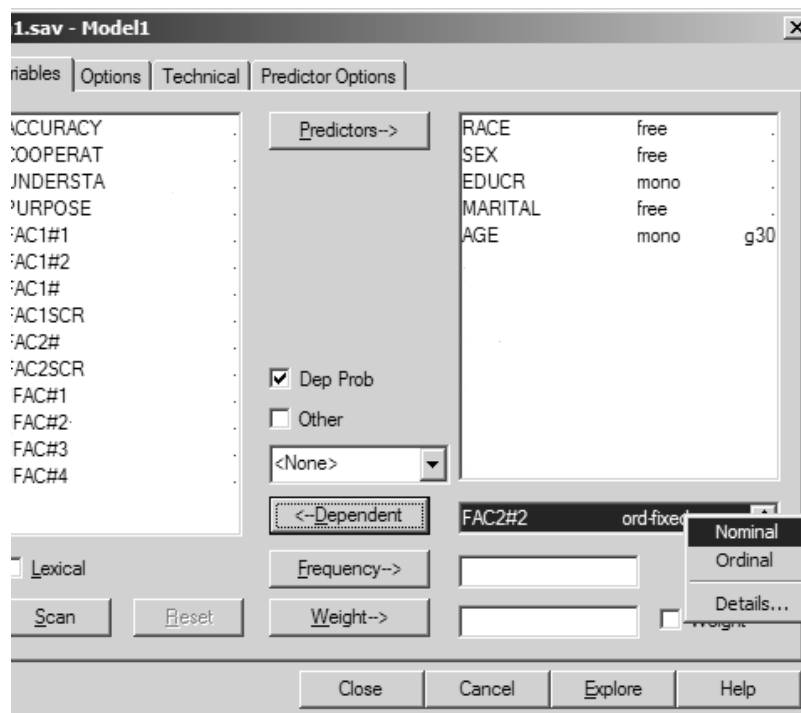


Figure 7-89. Changing FAC2#1 and FAC2#2 to Nominal

## LATENT GOLD® 4.0 USER'S GUIDE

- ▷ Click Options to open the Options Tab
- ▷ Select Auto to start in automatic mode

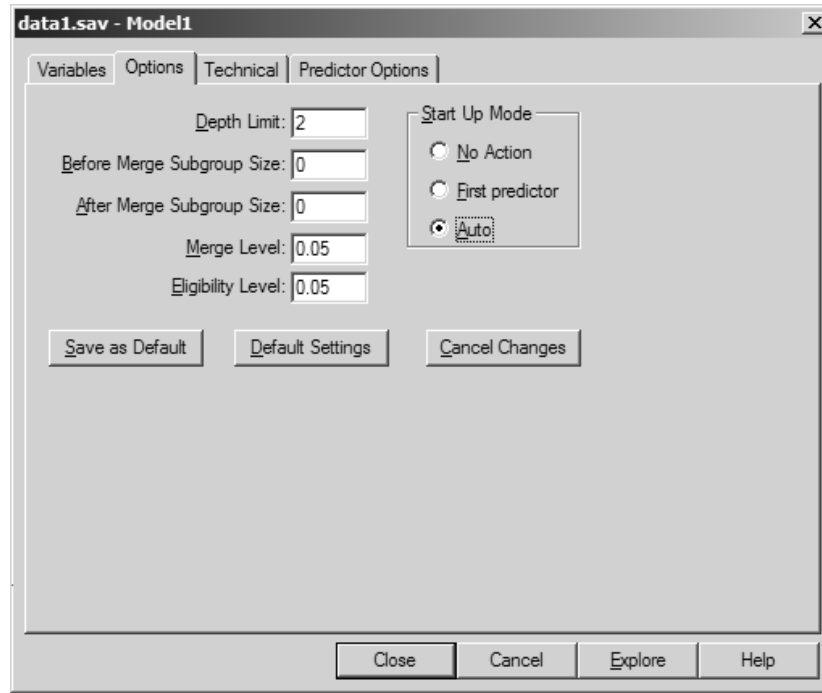


Figure 7-90. Options Tab

- ▷ Click the Explore button

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

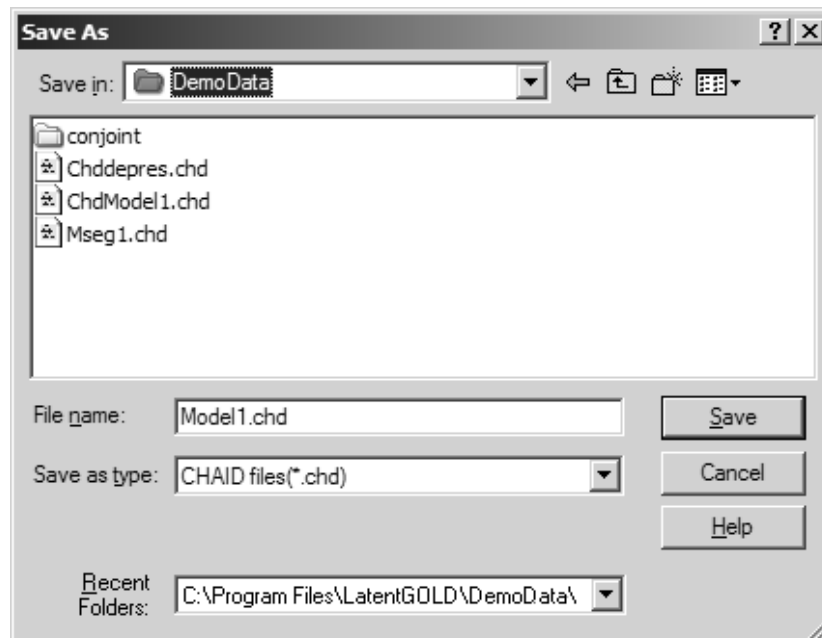


Figure 7-91. Save Definition Dialog Box

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.

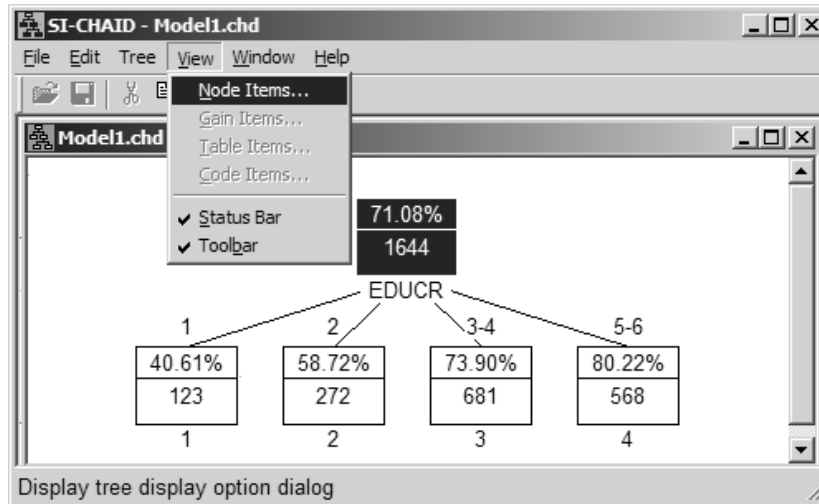


Figure 7-92. CHAID Segmentation Tree

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

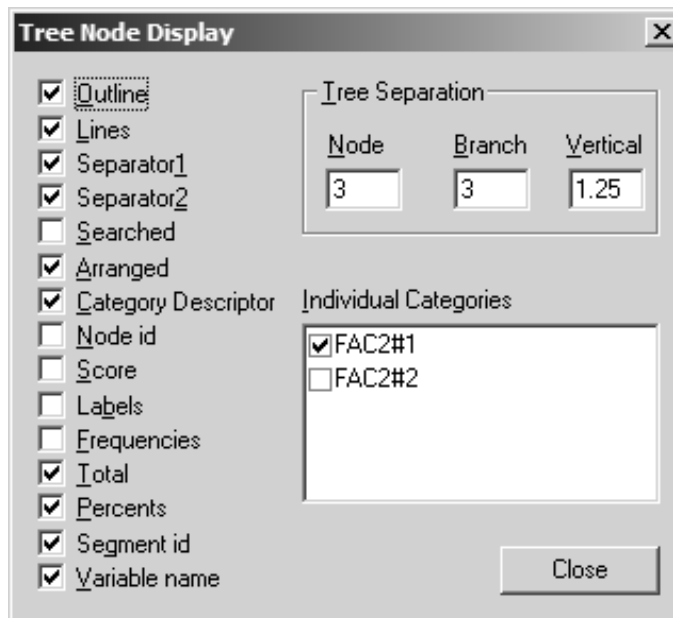



Figure 7-93. Node Items Dialog Box

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  $1.3 \times 10^{-19}$  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

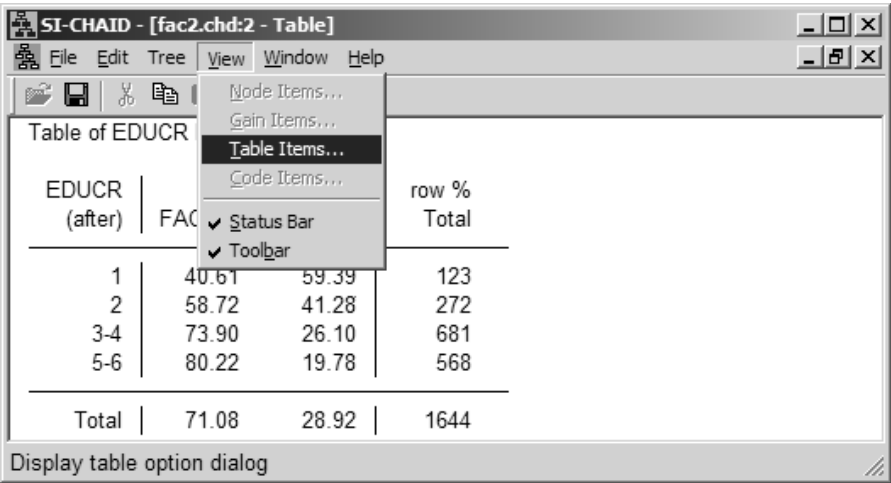


Figure 7-94. Table Display

From the Table Display

- ▷ Select Before Merge
- ▷ Select Column Percents

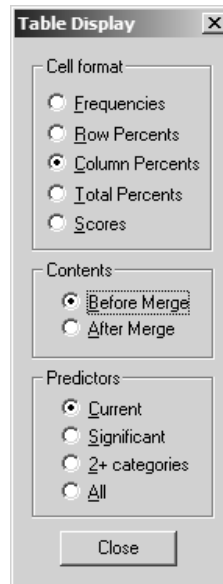


Figure 7-95. Table Display Dialog Box

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

SI-CHAID - [fac2.chd:2 - Table]

File Edit Tree View Window Help

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.

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.

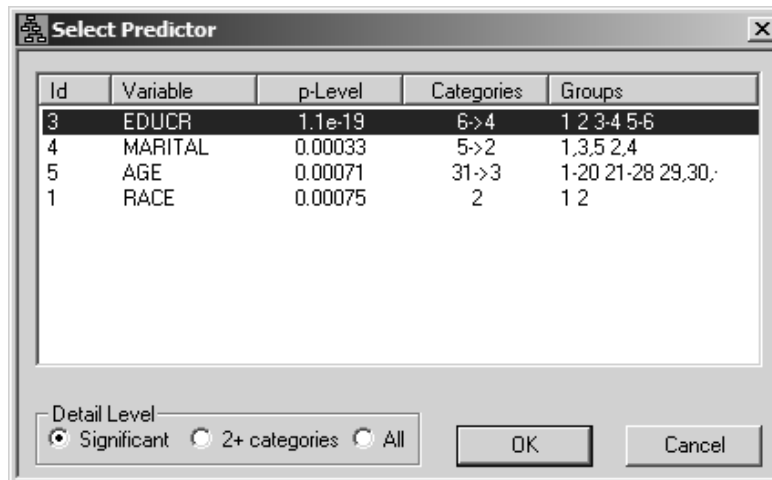


Figure 7-98. Select Predictor Dialog Box

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

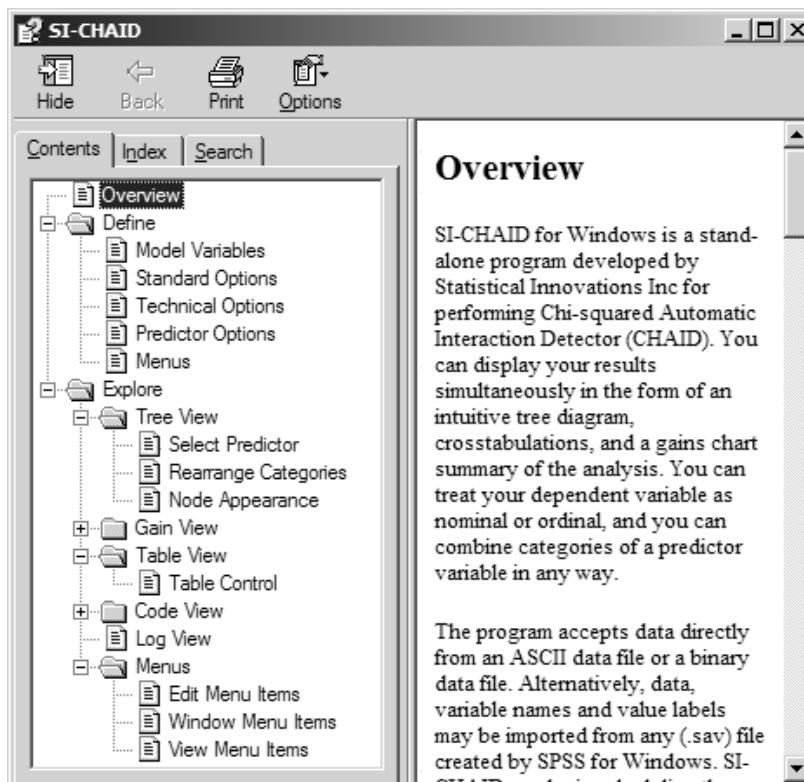


Figure 7-99. SI-CHAID Help Contents



## CHAPTER 8. ADDITIONAL TUTORIALS AND ASSOCIATED DATA SETS

Several additional tutorials are under development and when completed will be accessible from our website. These tutorials include:

### Tutorial #5: Using Latent GOLD 4.0 with the Known Class Option

**DEMADATA = 'DEPRESS2.SAV'**

In this tutorial, we illustrate the use of the 'known class' feature in Latent GOLD 4.0 to take into account additional information on a subset of cases which allows us to classify them into a particular class with probability one. In this case, the information comes from a physician's diagnosis of the patient as 'Depressed' or merely 'Troubled', corresponding to 2 of the 3 latent classes.

### Tutorial #6: Estimating a Random Intercept Regression Model

**DEMADATA = 'CRACKERS.SAV'**

(SOURCE: KELLOGG COMPANY STUDY)

In this tutorial, we illustrate the use of continuous factors (CFactors) to control for the 'level effect' in ratings data. A latent class regression model is estimated where the dependent variable is ratings of 15 crackers on taste, and 12 predictors correspond to different attributes of the crackers. Different classes are identified that show different taste preferences, controlling for their overall rating level. These data are based on a paper by Popper et. al. The use of CFactors requires the Advanced version of Latent GOLD 4.0.

## DATA SETS AND EXAMPLE LGF FILES

Below are descriptions of 80 data sets that have been analyzed using Latent GOLD 4.0. Some of these are the subjects of tutorials, whereas others are used in our courses and workshops. These data sets are included in the demo version of Latent GOLD 4.0. For some data sets we also prepared lgf files illustrating the most important Latent GOLD 4.0 Basic and Advanced features. The data and lgf files are also available separately on our website at [http://www.statisticalinnovations.com/products/latentgold\\_datasets.html](http://www.statisticalinnovations.com/products/latentgold_datasets.html).

### 8.1. Dichotomous, Nominal, or Ordinal Indicators: Cluster and DFactor Models, as well as Models with Continuous Factors (IRT Models)

#### Dichotomous indicators

##### 1. hannover.sav:

- 5 dichotomous indicators
- survey data on pain related to rheumatic arthritis
- cluster or DFactor model
- used in Kohlman and Formann (1997), Magidson and Vermunt (2001), and in the Latent GOLD 2.0 user's manual (Vermunt and Magidson, 2000a)

##### 2. political.sav

- 5 dichotomous indicators on political involvement and tolerance
- 3 (nominal) covariates
- 3-cluster model, 2-DFactor model, or 2-cluster model with a local dependency
- data from Political Action Survey
- used in Hagenaars (1993) and Vermunt and Magidson (2000a, 2000b)

##### 3. landis77.sav

- dichotomous rating (presence/absence of carcinoma in the uterine cervix) of 118 slides by 7 pathologists
- see also landisreg.sav for other data structure
- 3-cluster, 2-DFactor model, CFactor model (2PLM), CFactor model with equal effects (Rasch), or combination of 2-cluster and Rasch model
- sparse table: use bootstrap p value
- used as illustration in Agresti (2002), Magidson and Vermunt (2003a, 2004), and Vermunt and Magidson (2005a). Original data in Landis and Koch (1977).

### 4. heinen2.sav

- 5 dichotomous indicators of gender roles (male sample)
- same data set in other format in heinen2reg.sav
- 3-cluster, 3-level 1-DFactor model, and various types of IRT models
- used as illustration in Heinen (1996)

### 5. heinen\_mf.sav

- same data as heinen2.sav but now for males and females
- gender can be used as a covariate, possibly affecting indicators (item bias)
- data can also be used for unrestricted multiple group analysis (with female2 as known-class indicator)
- see also SMABS 2004 workshop transparencies

### 6. vdheijden.sav

- 3 dichotomous indicators of youth delinquency
- ethnic group and age group are covariates
- used by Van der Heijden et al. (1992) to illustrate logit-restricted latent budget analysis, which is a LC cluster model with covariates

### 7. depression.sav

- 5 depression indicators and covariate sex
- 3-cluster, 3-level 1-DFactor, or 2-cluster model with a CFactor model
- used in Magidson and Vermunt (2001) and Schaeffer (1988)

### 8. knowclass.sav

- simulated data set based on the 3-cluster solution obtained with the depression.sav data set
- information on known class membership generated using 3 mechanism: MCAR, MAR (depending on the sum of all item responses), NMAR (depending class membership itself)
- in the NMAR case known-class yes/no should be used as covariate

### 9. lcamis.dat

- 5 dichotomous indicators
- example of LC model with missing data on indicators
- simulated data set

### 10. lifestyle.sav

- data on a large set of lifestyle activities (dichotomous indicators) and a few covariates (source: The Polk Co.)
- demo data set in Latent GOLD 2.0 and used in Magidson and Vermunt (2003b)

## 11. store.sav

- 5 dichotomous items related to consumer behavior
- standard LC cluster model
- used in Dillon and Kumar (1994)

## 12. coleman.sav

- classical data set of Coleman
- 2 indicators, membership of and attitudes toward leading crowd, measured at two occasions
- 2-DFactor model (unrestricted or restricted)
- analysed by Goodman (1974) and Agresti (2002, table 12.8)

## 13. gss94.sav

- data from the 1994 General Social Survey
- 3 attitudes toward abortion indicators, and covariate gender
- 2-cluster model, LC Rasch (two-cluster with equal effects), parametric Rasch (CFactor with equal effects)
- data taken from Agresti (2002, table 10.13)

## 14. financial.sav

- data on ownership of 4 financial products
- taken from Paas (2002)

## 15. hadgu.sav

- 6 measures (tests) for diagnosing chlamydia trachomatis (most common sexually transmitted disease), where one test (culture) is a gold standard and can therefore be used a Known-Class indicator
- 2-class model with local dependencies modeled with a CFactor (with equal effects across tests)
- used by Hadgu and Qu (1998) with the purpose is to determine the sensitivity and specificity of the various tests

## Polytomous indicators

## 16. judges.dat

- trichotomous ratings of three judges, that can be treated as ordinal
- 3-cluster, 3-level 1-DFactor model, and CFactor/IRT (partial credit) model, possibly with equal effects across indicators
- used in Dillon and Kumar (1994), and in the Latent GOLD version 2.0 user's manual (Vermunt and Magidson, 2000a)

### 17. gss82white.sav

- 2 dichotomous and 2 trichotomous indicators that can be treated as nominal or ordinal
- data from General Social Survey '82, white sample
- the purpose of the analysis is to construct a typology of survey respondents
- 3-cluster or 2-factor model
- used in McCutcheon (1987), Magidson and Vermunt (2001, 2004), and Vermunt and Magidson (2005a)

### 18. gss82.sav

- same indicators as in gss82white.sav, but for full sample (whites and non-whites)
- several covariates that can be treated as active or inactive
- used in Magidson and Vermunt (2004)

### 19. elliot.sav

- marijuana use of children (13 years of age in 1976) in 5 consecutive years (trichotomous ordinal response variable)
- see also elliotreg.sav for other data structure
- standard cluster model with time-specific indicators and sex as covariate
- use bootstrap p value because of sparseness
- references to data set: Elliot et al. (1989), Vermunt and Hagenaaars (2004), and Vermunt, Rodriguez and Ato (2001)

### 20. heinen3.sav

- 5 trichotomous indicators of gender roles that can be treated as nominal or ordinal
- cluster, order-restricted cluster, DFactor, and various types of IRT models.
- bootstrap p value
- used as illustration in Heinen (1996)

### 21. environment.dat

- 6 trichotomous items measuring attitudes towards environmental issues
- there are two underlying dimensions: willingness (item 1-3) and awareness (items 4-6)
- data used by Croon (2002)

### 22. internet99.sav

- data on internet use (source: Mediamark Research Inc. 1999)
- relationship between internet usage and several demographic covariates
- used in Magidson and Vermunt (2003b)

## **23. USelection2000.sav**

- National Election Studies election survey data set 2000.T
- relationship between vote and ratings of Bush and Gore
- Source: Burns et. al. (2001), ICPSR Study Number: 3131

## **8.2. Single Response Variable: Mixture of Univariate Distributions (using Cluster or Regression Module)**

### **24. galaxy.dat**

- velocities of 82 galaxies diverging away from our own galaxy
- mixture of univariate normals
- set bayes constants off and increase number of start sets to reproduce results reported by McLachlan and Peel (2000)

### **25. enzyme.dat**

- enzymatic activity in the blood among a group of 245 individuals
- mixture of univariate normals
- used as illustration in McLachlan and Peel (2000)

### **26. acidity.dat**

- acidity index measured in a sample of 155 lakes in north-central Wisconsin
- mixture of univariate normals
- used as illustration in McLachlan and Peel (2000)

### **27. candy.dat**

- single count variable: number of packages of hard candy purchased in a week
- example of simple mixture model
- can be specified with regression or cluster with number of packages as count
- used in Dillon and Kumar (1994), Magidson and Vermunt (2004), and in Latent GOLD 2.0 user's manual Vermunt and Magidson (2000a).

### **28. candy\_trunc.txt**

- single truncated count variable: number of packages of hard candy purchased in a week among consumers
- example of simple mixture model for truncated counts
- can be specified with regression or cluster

- used in Dillon and Kumar (1994)

### 29. nov2002.sav

- results of statistics exam November 2002
- a 2 class binomial (exposure equal to 20) or normal mixture separates perfectly the students who pass and the ones that do not pass the exam.

### 30. sids.dat

- data of 100 counties in north Carolina concerning children suffering from sudden infant death syndrome: number of deaths and population at risk
- mixture of Poisson rates
- example used by Böhning (2000) to illustrate disease mapping.

## 8.3 Continuous, Count and Mixed-Scale Indicators: Cluster Models and Models with Continuous Latent Variables (Factor Analysis and Generalized IRT)

### 31. iris.dat

- 4 continuous indicators: measures taken on 150 irises
- LC cluster model or mixture model clustering
- true specie is known and can be compared with cluster solution (use true as inactive covariate)
- illustrates different specifications of within cluster variance-covariance matrix
- classical data set from Fisher

### 32. kmeans.sav

- simulated data set to illustrate LC clustering with continuous variables and compare it with K-means clustering
- different specification of the error variances
- used in Magidson and Vermunt (2002a, 2002b)

### 33. diabetes.sav

- 3 continuous indicators
- example of LC clustering
- clinical classification can be compared with LC cluster classification
- used in Fraley and Raftery (1998), Vermunt and Magidson (2002), and Magidson and Vermunt (2004)

## 34. cancer.dat

- clustering based on pre-trial "covariates" collected before for a prostate cancer clinical
- eight are treated as continuous and four as categorical indicators
- example of LC clustering with mixed mode data
- used as illustration in Hunt and Jorgensen (1999), McLachlan and Peel (2000), and Vermunt and Magidson (2002)

## 35. srcddata.txt

- continuous outcome variable read# (child's reading recognition) measured at 4 occasions (many missing data) on 405 children
- covariates: child's gender (male=1), mother's age in years at Time 1, child's age in years at Time 1, child's cognitive stimulation at home, and child's emotional support at home
- longitudinal data for specifying growth model: cluster model with one or two CFactors
- data used in Vermunt and Magidson (2005c) and made available at <http://www.duke.edu/curran/>
- file also contains an ordinal outcome variable anti# (child's antisocial behavior) measured at four time points, which was however not used in Vermunt and Magidson (2005c)

## 36. abortion\_cluster.sav

- same data as abortion.sav example (see below), but here in standard rectangular data format instead of repeated measures format
- indicators are binomial counts: number of agrees out of 7 abortion situations measured at 4 occasions
- source: McGrath and Waterton (1986)

## 8.4. Latent Class and Random-effects Regression Modeling

### Mixture Regression Models for Single Response

## 37. follman.dat

- effect of poison on survival
- dichotomous dependent variable survival can be treated as nominal, ordinal, or binomial count, since all are equivalent for dichotomous variables
- using logdose as a numeric class-independent predictor yields a non-parametric random effects logistic regression model
- used in by Follmann and Lambert (1989), Formann (1992), and Agresti (2002) to illustrate non-parametric random-effects logistic regression

## 38. fabric.dat

## CHAPTER 8. ADDITIONAL TUTORIALS AND ASSOCIATED DATA SETS

- number of faults in a bolt of fabric of a certain length
- random-effects Poisson regression with log length as predictor
- data used by Aitkin (1996) and McLachlan and Peel (2000)

### 39. beta.dat

- meta analysis of 22 clinical trials of beta-blockers for reducing mortality after myocardial infarction
- dependent is a binomial count
- observations within a clinic are dependent
- LC regression model with random intercept (3 classes) and fixed treatment effect
- data used by Aitkin (1999) and McLachlan and Peel (2000)

### 40. dmft.sav

- dental health trial on prevention of tooth decay among 797 Brazilian children
- dependent variable: # of decayed, missing or filled teeth (DMFT)
- explanatory variables are: Treatment (1 = no treatment; 2 = oral health education; 3 = school diet enriched with rice bran; 4 = mouth rinse with 0.2% NaF solution; 5= oral hygiene; 6 = all four treatments), Ethnic group (1= brown; 2 = white; 3 = black) and Gender (1 = male; 2 = female)
- Poisson or binomial count regression with overdispersion, using a LC regression, a zero-inflated regression, or random-intercept regression model
- data analyzed by Skrondal and Rabe-Hesketh (2004, section 11.2)
- see also SMABS 2004 workshop transparencies

### 41. cace.sav

- to illustrate "complier average causal effect" model using Known-Class option
- compliance is known for the treatment group but unknown (latent) for the control group
- LC regression in which treatment has an effect in the compliance class, and in which compliance (yes/no) is predicted using covariates
- data analyzed by Skrondal and Rabe-Hesketh (2004) and can be obtained from the ICPSR website (under JOB #2739)

### 42. long1.sav

- continuous dependent: firstjobcens0 or firstjobtrunc0 is the prestige of the first academic job (minus 1 to get the censoring/truncation at 0 instead of 1)
- various predictors
- censored normal regression, censored-inflated normal regression, or truncated normal regression
- data used by Long (1997, chapter 8)

### 43. long2.sav

- count dependent: number of articles in last 3 years of PhD
- various predictors (two copies of each in file)

- Poisson regression, zero-inflated Poisson regression, random-intercept Poisson regression, and zero-inflated random-intercept Poisson regression
- data used by Long (1997, chapter 9)

### **44. runshoes.dat**

- count dependent: number of running shoes for a sample of runners
- predictors: runs per week, miles run per week, distance runner
- truncated Poisson count regression model
- used in textbook "Analyzing Categorical Data" by Jeffrey S. Simonoff

## Two-level and Multiple Response Data Sets

### **45. bang.txt**

- contraceptive use (dichotomous outcome)
- data from 1989 Bangladesh Fertility Survey (Huq and Cleland 1990)
- women nested within districts
- predictors: number of children, age in years (centered), and urban (0=rural)
- data obtained from multilevel modeling website

### **46. snijdersbosker.sav**

- performance of pupils on a language test (continuous outcome)
- data taken from the Snijders and Bosker (1999) book on multilevel analysis
- children nested within school
- pupil-level predictors: IQ, SES (both overall centered)
- school-level predictors: school\_IQ, school\_SES, groupsize (centered), combination classes (yes/no)
- see also SMABS 2004 workshop transparencies

### **47. conjoint.sav**

- rating-based conjoint example
- simulated data
- full factorial design (2\*2\*2) with 8 replications
- LC regression with ordinal dependent, 3 predictors (product attributes) and 2 covariates (individual characteristics)
- used in Magidson and Vermunt (2003b) and in Latent GOLD 2.0 user's manual (Vermunt and Magidson 2000a)

### **48. crackers.sav**

## CHAPTER 8. ADDITIONAL TUTORIALS AND ASSOCIATED DATA SETS

- data from a consumer taste study sponsored by the Kellogg company, where consumers rated their liking of 15 crackers on a nine-point liking scale.
- an independent trained sensory panel evaluated the same crackers in terms of their sensory attributes (e.g. saltiness, crispness, thickness, etc.), yielding ratings on 12 flavor, texture, and appearance dimensions
- LC regression analysis with a random intercept

### 49. USselection2000reg.sav

- National Election Studies election survey data set 2000.T
- same data as USselection2000.sav, but now in regression format
- Source: Burns et. al. (2001), ICPSR Study Number: 3131

Various Cluster and IRT Data Set in the Form of Multiple Records per Case

### 50. landisreg.sav

- dichotomous rating (presence/absence of carcinoma in the uterine cervix) of 118 slides by 7 pathologists
- dependent variable "rating" can be treated as nominal, ordinal, or binomial count since all are equivalent for dichotomous variables.
- LC regression model with rater as nominal predictor. Specifying the rater effect as class independent yields a LC Rasch model. Class dependent yields a standard LC model.
- variable "sumscore" can be used as inactive covariate to see how the latent classification is related to the sum of the ratings.
- the file contains dummies for the raters to change the coding scheme.
- a copy of the predictor rater (rater\_) is included to specify a two-dimensional model (LC factor model).
- sparse table: use bootstrap p value
- used as illustration in Agresti (2002) and Magidson and Vermunt (2003a, 2004) and Vermunt and Magidson (2005a). Original data in Landis and Koch (1977).

### 51. heinen2reg.sav (male sample)

- 5 dichotomous indicators of gender roles
- same data as heinen2.sav, but other data structure
- 3-class regression model: item effect class-independent yields a LC Rasch model; item effect class-dependent yields a standard LC model
- used as illustration in Heinen (1996)

### 52. heinenreg\_mf.sav

- same as heinenreg2.sav but now for males and females (also the same as heinen2\_mf.sav but in other format)
- gender can be used as covariate, predictors, or in gender-item interaction (item bias)
- standard LC, restricted LC, LC Rasch, and IRT models

- see also SMABS 2004 workshop transparencies

### **53. colemanreg.sav**

- same data as coleman.sav but in a different format
- item characteristics are included as predictors to test several assumption
- predictors: item, member, attitude, time1, time2, member1, member2, attitude1, and attitude2
- best model is a 2-factor like structure with a member and a attitude factor
- analysed by Goodman (1974) and Agresti (2002, table 12.8)

### **54. gss94reg.sav**

- same data as gss94.sav but in a different format

### **55. financialreg.sav**

- same data as financial.sav: ownership of 4 financial products
- taken from Paas (2002)

## Longitudinal Data for LC Growth Modeling

### **56. abortion.sav**

- data from the British Social Survey (McGrath and Waterton, 1986)
- the dependent "number of times that one agrees with abortion out of 7 situations" should be treated as binomial count
- year is a class-dependent (random, level-1) predictor and religion a class-independent (fixed, level-2) predictor
- the data file contains dummies for the time and religion categories to use dummy instead of default effects coding
- the data file also contains an incremental coding of the time categories and time squared to play with the time effect
- used by Vermunt and Van Dijk (2001) to illustrate the connection between LC regression and random-coefficients, mixed, hierarchical, or multilevel models, as well as in Magidson and Vermunt (2004).

### **57. elliotreg.sav**

- marijuana use of children (13 years of age in 1976) in 5 consecutive years (trichotomous ordinal response variable)
- LC growth model with time as nominal/ascending/class-dependent predictor and sex as covariate (see Vermunt and Hagenaars, 2004)

- possible to include random intercept (CFactor)
- references to data set: Elliot et al. (1989) and Vermunt et al. (2001)
- see also SMABS 2004 workshop transparencies

### 58. rats.dat

- grow of rats in first weeks
- LC growth model for continuous outcome variable
- reference to data set: Gelfand et al. (1990)

## Event History and Transition Data

### 59. jobchange.dat

- LC regression model for event history data (piece-wise exponential survival model)
- data from 1975 Social Stratification and Mobility Survey Japan (see, Yamaguchi, 1991)
- the event of interest is first inter-firm job change
- event should be treated as Poisson count with an exposure variable
- time, categorized in 3 intervals, is a class-independent nominal predictor
- single covariate firm size (either nominal or linear with extra dummy for government)
- used as illustration in Vermunt (2002a)

### 60. empltran.dat

- discrete-time event history or survival model with multiple outcomes
- two predictors/covariates: cohort and sex
- exposure time should be used as replication weight
- used in Blossfeld and Rohwer (1995)

### 61. dropout.dat

- school drop-out of brothers at two school levels
- modelled as discrete-time event history model with unobserved heterogeneity to capture dependence between respondent and brother (family effect)
- brother and time (school level) are predictors; father's education can serve as predictor or as covariate
- used as illustration in Mare (1994) and Vermunt (1997)

### 62. land.sav

- duration time to first serious delinquency
- 411 males from working-class area of London followed from ages 10 through 31
- dependent "first" can be treated as Poisson count or as binomial count. If treated as Poisson count, the exposure can be set to one or one half for the time point at which the event occurs.
- variable "tot" is a risk index that can be used either as predictor or as covariate
- the duration effect (age effect) can be modelled by a quadratic function
- data used as illustration by Land et al. (2001).

### 63. poulsen.sav

- transitions in brand preference (brand A or other brand) between 5 occasions
- example of mixture transition or mixed Markov model
- predictors are time0 (whether record corresponds to the initial state), ylag\_a (previous time point equals brand A), and ylag\_oth (previous time point equals brand A). Either the intercept or time0 should be omitted from the model
- data used as illustration by Poulsen (1982)

### 64. vinken.sav

- timing of four events related to first experience with relationships
- used in Vermunt (2002a)
- Cox model for correlated events
- see also SMABS 2004 workshop transparencies

## Longitudinal Data from Repeated Measures Clinic Trials

### 65. koch.sav

- repeated measures clinical trial with outcome normal (1) or not normal (0)
- time is a class-dependent predictor, severity a class-independent predictor and treatment is a covariate; this yields a LC growth model in which treatment has an effect on the type of growth curve that one follows.
- an alternative is to use time, severity, treatment, and the treatment-time interaction as class-independent predictors, yielding a standard non-parametric random-effects model.
- used in Agresti (2002) to illustrate random-effects logistic regression. Original data are in Koch et al. (1977).

### 66. epilep.sav

- randomized controlled trial comparing a new drug with placebo
- outcome variable y is the number of epileptic seizures during the two weeks before each of 4 clinic

visits (Poisson count)

- 4 replications per case (4 visits)
- class-independent numeric predictors: treatment, log baseline, log age, visit number, dummy for fourth visit, and treatment log base interaction
- data from Thall and Vail (1990), also used by Rabe-Hesketh et al. (2002)

### 67. aspartame.dat

- multiple period (5 weeks) crossover trial to test the side effect of aspartame
- the dependent variable is a binomial count; that is, the number of days with a headache out of a total of 7 days (a week).
- the total number of days exposed in a period may be smaller than 7 and the total number of periods may be less than 5 because of drop out.
- predictors are week and aspartame (1= aspartame; 0=placebo)
- covariate: believe as to whether drug will cause and headache
- data used by McKnight and Van Den Eeden (1993) to illustrate models for correlated binomial counts. Can also be modeled as Poisson counts (see Hedeker, 1998).

### 68. genomics.sav

- multi-visit follow-up of 7 rheumatoid arthritis patients diagnosed as unstable during first visit and assigned to new drug therapy
- blood sample taken during each visit to obtain genetic expressions
- drug effects assessed using IndexZ to see if levels approach those of normals
- source of IndexZ (Source Precision Medicine, Inc.) - patents pending

### 69. schizophrenia.sav

- effect of drug on severity of schizophrenia
- dichotomious or ordinal dependent variable "severity" measured at 7 occasions (with many missing values)
- can be used for random-effect logistic regression, LC logistic regression, and LC logistics regression with a random intercept
- data used by Hedeker and Gibbon's (1996)

## Two-level Cluster and DFactor Models

### 70. miero\_socmeth.txt

- 5 dichotomous items measuring task variety
- missing values on items (some of which were caused by a mistake made in the recoding of the items)
- employees nested with teams
- simplest variant of multilevel LC model with either GClasses or GCFactors affecting the clusters
- data set taken from dissertation from Van Mierlo (2003), and used by Vermunt (2003)

### 71. miero\_mbr.sav

- same data set as miero\_socmeth.dat, but without the mistake in the recoding (results are therefore slightly different)
- in addition, 4 individual-level covariates: year of birth (4 levels), number of years in the current job (3 levels), number of working hours per week (3 levels), and gender. The 57 cases with missing values on items and/or covariates can be retained in the analysis (using the include missing all option).
- random-intercept model for the clusters using a GCFactor
- used in Vermunt (2005)

### 72. cito.dat

- data on mathematical skills on pupils: 18 mathematics test items (correct/incorrect) administered to 2157 pupils
- pupils are nested within 97 schools
- three individual-level covariates (SES, IQ and Gender) and one school-level covariate (CITO)
- multilevel variant of a DFactor model
- used by Fox and Glas (2001) and by Vermunt (2003)

### 73. meulders.sav

- three-mode three-way data from a psychological "experiment" 101 1st year psychology students to indicate whether when angry at someone they would display 8 behaviors ( fly off the handle, quarrel, leave, avoid, pour up ones heart, tell one's story, make up, clear up the matter) in 6 situations (like the other, dislike the other, unfamiliar with the other, other has higher status, other has lower status, and other has equal status other.
- situations are nested within persons
- situations are non exchangeable, therefore use situation as covariate affecting class membership
- GClasses (of persons) affecting intercept of and situation effect on clusters (of persons in situations)
- used in Meulders et al. (2002, Journal of Classification) paper on LC models for three-mode data

### Three-level Regression Models

#### 74. immunization.sav

- complete immunization of children in Guatemala (binary response variable)
- individuals (children) nested within families, and families nested within communities
- three-level binary logistic regression using parametric or nonparametric random effects
- 4 individual, 5 family and 2 community level predictors (some are dummies)
- used by Rodriguez and Goldman (2001).

#### 75. tvsfp.sav

- ordinal outcome variable: the tobacco and health knowledge scale (THKS) score defined as the number of correct answers to seven items on tobacco and health knowledge (collapsed into our ordinal categories).
- schools were randomized into one of four conditions combining the factors TV (a television intervention, 1=present, 0=absent) and CC (a social-resistance classroom curriculum, 1=present, 0=absent)
- classes are nested within schools and pupils are nested within classes
- data are from the Television School and Family Smoking Prevention and Cessation Project (TVSFP) and used by Hedeker and Gibbons (1996)

#### 76. socatt.txt

- same data as abortion.sav file, but now with district number and some extra covariates
- repeated measures nested within cases and cases nested within districts
- three-level binomial count regression using either parametric or nonparametric random effects
- source: McGrath and Waterton (1986) and multilevel modeling webpage
- used in Vermunt (2002c) and Vermunt (2004)

#### 77. zugugl.sav

- three well-being items (zufrieden, gut, glücklich) measured at three occasions
- responses are both in 3-point scale and 5-point scale format
- 3-level regression or 2-level IRT model
- data used by Steyer and Partchev (2001) to illustrate their state-trait model for ordinal variables, which is a 2-level IRT model

#### 78. tob3vote.sav

- response variable: voting pro-tobacco by members of the Congress from 1997-2000
- predictors/covariates: party, amount of money member received from tobacco industry (money), and the number of harvest acres in the member's state in 1999 (acres)
- votings/bills nested within members and members nested within states
- 3-level random-effects regression model and 2-level LC model
- used by Luke (2004) in his Sage textbook "Multilevel Modeling"

## Complex Survey Options

### 79. **patterson.sav**

- standard LC model for 4 dichotomous response variables (vegetable consumption at 4 occasions)
- stratum, PSU, and weight variable
- totvgt1-totvgt4 variables were used by Patterson, Dayton, and Graubard's (2002) in their article on LC analysis of complex sampling survey data
- the variables v1-v6 were used by Vermunt (2002b), who made use of the fact that the 4 occasions were actually 6 different time points with missing values on at least two time points
- data set does also contain some covariates, as well as information on fruit consumption (was not used by the above authors)

### 80. **pattersonreg.sav**

- regression format data set based on patterson.sav
- contains the variables totvgt1-totvgt4 used Patterson, Dayton, and Graubard's (2002)
- LC growth model

### 81. **pattersonreg2.sav**

- regression format data set based on patterson.sav
- contains variables v1-v6 (vegetables) used in Vermunt (2002b), as well as f1-f6 (fruit)
- can be used to specify a LC growth model for vegetables or for fruit, or a multilevel LC model in which vegetables and fruit consumption are used as indicators of a time-specific latent variable

# Index

## A

adjacent category logit model, 55

### AIC

based on L2. See chi-squared statistics

based on LL. See log-likelihood statistics

Analysis dialog box, 20

arrays, 27, 31

## B

baseline category logit model, 55

Bayes constants, 82

betas, 114

### BIC

based on L2. See chi-squared statistics

based on LL. See log-likelihood statistics

binomial count scale type, 57

bi-plot, 142

bivariate residuals, 144

specifying, 75

Bootstrap options, 84, 96

boundary solutions

warning message, 106

## C

### CAIC

based on L2. See chi-squared statistics

based on LL. See log-likelihood statistics

case id variable. See ID variable

case weight. See weights

categories

displaying labels for 51

reducing the number before estimation 52

CFactors, 72

CHAID option, 10, 93, 217

chi-square calculator, 97

chi-squared statistics, 111

class independent restriction, 64, 68

classification

appending to output file, 90

example for cluster model, 125

example for factor model, 144

example for regression model, 160

output, 86

classification statistics, 112, 130, 150

AWE, 112, 130, 150

classification errors, 112, 130, 150

classification log-likelihood, 112, 130, 150

entropy R-squared, 112, 130, 150

reduction errors, 112, 130, 150

standard R-squared, 112, 130, 150

ClassPred Tab

Classification and Prediction output, 90

Known Class option, 73

closing

data files, 35

Cluster Model, 3, 59, 109, 164

Contents pane, 13-14

continuous factors. See CFactors

continuous scale type, 57

convergence limits, 81

copying and pasting output, 15

count scale type, 57

covariates

- active vs. inactive, 57

- for cluster and factor models, 48

- for regression models, 213

Cressie-Read. See Chi-square statistics

## D

data files

- arrays, 31

- closing, 35

- defining a new model, 100

- example of opening, 165

- Latent GOLD save files, 32

- opening, 25

- saving model settings, 32

- scanning, 51

- SPSS files, 29

- summary and listing of models output, 107

- text (rectangular) files, 29

dependent variable, 47

DFactor Model, 3, 66, 127, 190

- changing the number of factor levels, 66

- changing the number of factors, 66

- including factor correlations, 66

DFactor Loadings, 133, 195

Dialog Boxes, 20

- Pushbuttons, 22

- Subdialog Boxes, 22

direct effects

- including in cluster and factor models. See Residuals tab

## E

error variances, 83

- specifying class dependent/independent, 64

estimation, 47

- cancelling (stopping, pausing), 94

- warning messages, 105

exposure variable, 50

## F

fixed scale type. See ordinal scale type

fonts

- changing for plots, 40

formats

- input data formats, 26

- output formats, 40

- changing numeric font, 41

- changing fonts, 40

frequencies and residuals output, 124, 144, 159

- specifying, 86

frequency counts

- viewing for a variable, 51

frequency variable, 31, see also case weight

## G

gammas, 114

grouping option, 53

group ID, 79

group level, 79

GClasses, 79

GCFactors, 79

## H

Help, 18

## I

ID variable

- Case ID, 50, 205

- appending to file, 90

identification

- warning message, 106

indicators, 48

iteration detail, 105

- specifying output, 87

iteration limits, 81

## L

Latent Classes  
 definition of, 1  
 ordering of, 105  
 Latent GOLD Choice, 9  
 Latent GOLD save files, 32  
 LL. See log-likelihood statistics  
 log-likelihood (LL). See log-likelihood statistics  
 log-likelihood statistics, 112  
 log-posterior. See log-likelihood statistics  
 log-prior. See log-likelihood statistics  
 L-squared. See chi-squared statistics

## M

Menus, 15  
 missing values, 52  
 options, 83  
 model fit. See Chi-squared statistics  
 model selection menu, 45  
 model summary output  
 example of for cluster model, 110  
 example of for factor model, 128  
 example of for regression model, 147  
 model type  
 selection of, 44  
 models  
 defining, 44  
 deleting, 101  
 editing names, 100  
 multilevel model, 79  
 multinomial logit model  
 specifying for a regression analysis, 56

## N

nominal scale type, 56  
 nonconvergence  
 warning message, 106  
 numeric scale type, 57

## O

opening  
 data files, 25  
 dialog boxes, 20  
 ordering  
 of clusters/factors/classes, 105  
 ordinal scale type, 55  
 fixed, 55  
 uniform, 55  
 user scores, 55  
 Outline pane, 13-14  
 editing model names, 100  
 output  
 changing plot fonts, 40  
 copying and pasting, 16  
 listings, 84  
 options. See Output Tab  
 print preview, 38  
 print setup, 39  
 printing, 37  
 saving, 32  
 output to external files  
 bivariate residuals output. See bivariate residuals  
 CHAID input, 93  
 classification output. See classification  
 classification prediction information, 90  
 frequencies and residuals output. See frequencies  
 and residuals output  
 iteration detail. See iteration detail  
 paramaters output. See parameters output  
 probmeans. See probmeans output  
 profile output. See profile output  
 variance and covariance matrix, 88  
 Output tab, 21, 85

## P

parameters output, 113  
 betas, 114  
 example of for regression model, 151  
 gammas, 114  
 restrictions, 60-76, 195, 210  
 sigmas, 114

viewing standard errors, 114

viewing Wald statistic, 114

pause estimation, 94

plot control dialog box

for bi-plots, 142

for profile plot, 118

for tri-plots, 123

for uni-plots, 120

plots

bi-plot. See bi-plot

profile. See profile plot

tri-plot. See tri-plot

uni-plot. See uni-plot

Poisson model, 57

specifying an exposure, 50

posterior mean, 88

predictor variable, 50

printing, 37

changing plot fonts, 40

preview, 38

setup, 39

probmeans output, 119, 138, 157

profile output, 115, 133, 153

profile plot, 117, 136, 155

changing settings for, 118

PSU, 77

## R

random sets. See start values

rectangular files. See text files

Regression Model, 3, 69, 146, 202

repeated measures, 51, 202

replication weight. See weights

residuals

bivariate. See bivariate residuals

standardized. See frequencies and residuals output

Residuals tab, 21, 75

resume estimation, 95

## S

sampling weights. See weights

saving, 32

model settings, 32

model results, 32

scale type, 54

binomial count, 57

continuous, 57

count, 57

nominal, 56

numeric, 57

ordinal, 55

scanning a data file, 51, 206

scores

specifying, 56

viewing, 51

scoring a data file, 55, 97

seed. See start values

selecting variables, 22

SI-CHAID. See CHAID option

sigmas, 114

spss files, 29

standard errors

computation of, 87

in parameters output, 114

in profile output, 116

start values, 81

location of random seed value in output, 114

Status Bar, 20

stop estimation, 94

Stratum, 77

## T

Technical tab, 21, 80

Bayes constants, 82

convergence limits, 81

error variances, 83

iteration limits, 81

missing values options, 83

start values, 81

text files, 29

Tolerance, 81, 82

Toolbar, 19  
tri-plot, 122

## U

uniform scale type. See ordinal scale type  
uni-plot, 120  
    changing settings for, 121  
user scores. See ordinal scale type

## V

variables  
    detail, 131  
    types of, 47  
Variables tab, 20, 47  
variance covariance matrix, 88  
Viewer window, 13  
    Contents pane, 13-14  
    Outline pane, 13-14

## W

Wald statistic, 114  
    computation of, 87  
Wald(=) statistic, 153, 210  
weights  
    case weight, 47  
    replication weight, 51  
    sampling weight, 77

## X

X-squared. See chi-squared statistics