Statistical
Innovations
Thinking outside the brackets!™

LG-Syntax™
User's Manual

LG Choice Modeling Extensions
via the LG-Syntax Module

This document should be cited as: Vermunt, J.K., and Magidson, J. (2011). LG-Syntax User's Guide: LG Choice Modeling Extensions via the LG-Syntax Module, Belmont, MA: Statistical Innovations Inc.

For more information about Statistical Innovations Inc. please visit our website at http://www.statisticalinnovations.com or contact us at

Table of Contents

1 Latent GOLD Choice Modeling Extensions via LG-Syntax™

The LG-Syntax™ module extends the capabilities of the Advanced versions of Latent GOLD[1] and Latent GOLD Choice in several important ways, including the ability to perform simultaneous analyses of *multiple* dependent variables, the inclusion of *multiple* latent variables, *dynamic* (time-varying) latent variables, flexible parameter restrictions, and options for monte carlo simulation, validation and hold out, saving a model with parameter estimates for rapid re-estimation and scoring of other data files, and submitting runs in batch mode.

This document describes features of the LG-Syntax that pertain to Latent GOLD Choice. For more complete details of all the features available in LG-Syntax see LG-Syntax User's Guide.

Some important special features described in this manual include the following extensions of latent class (LC) choice, MaxDiff and other ranking models:

- **Inclusion of scale factors**. Choice responses elicited from persons with lower scale factors exhibit higher error variance which is indicative of higher amounts of uncertainty. Inclusion of scale factors in LC choice models provides better separation and clearer interpretation of the latent class segments[2].
- **Estimation of fused models with a single common LC segmentation**. Data from a discrete choice design may be analyzed in conjunction with one or more ratings or other (non-choice) dependent variables to obtain a common LC segmentation that applies to both the choice and non-choice dependent variables.
- **Estimation of fused models with 2 or more latent variables** to investigate the relationship between a single LC segmentation for the discrete choice data and a second for explaining heterogeneity with respect to the other dependent variables.
- **Numerous other model extensions**

The remainder of this document provides a general overview of LG Choice syntax features, an example of the syntax for a typical choice model (Chapter 2.1), an extended data structure to accommodate the estimation of fused models, and 2 examples of data fusion (Chapter 2.2).

---

[1] Latent GOLD is a separate program from Latent GOLD Choice, but for users who license both, the programs are provided in a single integrated package. LG-Syntax works with both Latent GOLD and Latent GOLD Choice, and needs to be licensed only once for each copy of Latent GOLD Choice or each copy of the integrated Latent GOLD/Latent GOLD Choice package.

[2] Magidson, J., and Vermunt, J.K. (2007b). Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference. October 2007 <u>Sawtooth Software Conference Proceedings</u>,

2 Choice Models

This section describes the estimation of choice models using LG-Syntax. The primary benefit of specifying choice models using the Syntax rather than the GUI is that more general models are available such as those based on multiple dependent variables and/or multiple latent variables. In addition, more general parameter restrictions, monte carlo simulation options, estimation of separate models on selected data subsets, as well as other extended capabilities are available.

When estimating a model containing multiple dependent variables, a restriction is that only the last dependent variable can be of the type choice, ranking, or rating. However, a given case ID may have different kinds of choice sets associated with it so that a single dependent variable may in fact refer to different dependent variable types (e.g., 1st choice, MaxDiff, etc.). Section 4 contains an example of a 4-class fused model where respondents selected the most and least important attributes from each of 16 sets of 5 attributes in a MaxDiff (best-worst scaling) experiment, and also provided importance ratings for all 36 attributes.

As mentioned above, the LG-Syntax contains commands that are unique to Latent GOLD Choice in that they pertain only to choice models. These include commands for specifying:

- the one- or 3-file structure
- a choice set id
- the scale type of the dependent variable
- attributes of the choice alternative
- a replication scale variable
- output options that are specific to choice

One thing that has to be specified when using a choice-type dependent variable is the file structure, which can be a 1- or a 3-file structure. When using a 3-file structure, the names and location of the alternatives and sets files needs to be specified.

A choicesetid can be specified to define the choice sets in response file. In the 3-file format this id connects the choices in the response file with the choice set defined in the choicesets file. In the 1-file format the choicesetid connects the multiple records that pertain to a choice set.

As in the GUI Choice module, the dependent variable corresponding to a choice variable can be defined to be scale type choice, ranking, or rating.

In addition to independent variables that are case-level predictor variables, the model may contain 'attributes' which are alternative-specific predictors that affect the choice, ranking, or rating dependent variable. Also, as in the Choice GUI, the keyword '_Constants_' is a reserved attribute name.

Two output sections that are unique to Latent GOLD Choice are SetProfile and SetProbMeans, which are the same as in GUI. Except for the 'Importance' output, all of the output sections currently available in the GUI version of Latent GOLD Choice, are retained as LG-Syntax output sections. The Importance section does not generalize to many of the extended models that can be

obtained in the LG-Syntax and thus is not available in the LG-Syntax. However, a new output section, BivariateResiduals, is provided when a Choice model is estimated with the syntax. This output provides some diagnostics as to whether certain model assumptions are supported such as testing whether a covariate is conditionally independent of the choices.

2.1   Examples

To see what the LG-Syntax looks like in the case of a choice model, one can simply open a Syntax choice example from the Examples Menu, or open a GUI choice example (from the Examples Menu or a previously saved .lgf file) and from the Model Menu select 'Generate Syntax' to view the syntax statements in the Syntax Editor. For the Coffee example, the choice responses are identified by the dependent variable flag named 'ch' in a 1-file format, or 'choice' in the 3-file format. The following syntax statements appear following those in the Options section:

 (For the 1-file format)

```
    choice = 1;
    variables
      caseid id;
      choicesetid set ;
      dependent ch choice;
      attribute brand, cap, price, filter, therm;
      latent Class nominal 3 ;
    equations
      Class <- 1 ;
      ch <- brand | Class + cap | Class + price | Class + filter | Class + therm | Class ;
```

Note that the first statement 'choice = 1;' refers to the 1-file format.

Aside from replacing the variable name 'ch' with the corresponding name 'choice', the only other change required with a 3-file structure is to replace the statement 'choice = 1' with the following statements which refer to the 3-file structure and provide the file names for the alternatives and sets files:

```
choice = 3
  alternatives 'coffee_alternatives.sav'
  id=altid
  choicesets 'coffee_sets.sav'
  id=setid ;
```

2.2 Data Structure for Choice plus Other Dependent Variables

When a choice model is combined with other types of dependent variables, the data file contains one or more additional records with missing values in the choice fields (choice set id #, choice, etc.) to accommodate the other dependent variables. Similarly, the records pertaining to the choice information contain missing values in the fields that pertain to the other dependent variables, including fields that may define independent variables that pertain only to one or more of these other dependent variables.

For example, assume respondent #1 (caseid = 1) provides responses to 4 choice sets, the choices being recorded in dependent variable 'dep1', and responses are also provided to 2 other dependent variables -- ordinal dependent variable 'dep2', and continuous dependent variable 'dep3'. In addition, suppose that there are 3 independent variables. The first, 'indep1', is used as a covariate (i.e., as a predictor of the latent classes), and the others, 'indep2' and 'indep3', as predictors of the dependent variables 'dep2' and 'dep3'.

caseid	setid	indep1	indep2	indep3	dep1	dep2	dep3
1	1	1	.	.	2	.	.
1	2	1	.	.	1	.	.
1	3	1	.	.	1	.	.
1	4	1	.	.	3	.	.
1	.	1	2	2	.	4	3.24

Caseid #1 provides responses (dep1) to 4 choice sets, selecting alternative #2 to set 1, alt. #1 to sets 2 and 3, and alt. #3 to set 4.

The last record contains additional responses (dep2 and dep3) to non-choice related survey questions.

Figure 2-1: Response file for the 3-file data structure

The syntax for the 3-file structure:

choice = 3
  alternatives 'alt.dat'
  id=altnr
  choicesets 'set.dat'
  id=setnr;
variables
  caseid caseid;
  choicesetid setid ;
  dependent dep2, dep3 continuous, dep1 choice;
  independent indep1, indep2, indep3;
  attribute att1, att2, att3;
  latent
    Class nominal 2;
equations
  Class <- 1 + indep1;
  dep2 <- 1 | Class + indep2 | Class + indep3 | Class;
  dep3 <- 1 | Class + indep2 | Class + indep3 | Class;
  dep1 <- att1 | Class + att2 | Class + att3 | Class;

'choice = 3' refers to the 3-file format

the relationship between the alternatives and attributes are provided in the alternatives file 'alt.dat';

the alternatives appearing in each choice set are given in the sets file 'set.dat'.

The corresponding 1-file structure contains not only the responses but also information regarding alternatives, attributes, and choice sets. To accomplish this, separate rows are required for each alternative in each choice set.

caseid	setid	att1	att2	att3	indep1	indep2	indep3	dep1	dep2	dep3
1	1	1	1	2	1	.	.	0	.	.
1	1	2	2	1	1	.	.	1	.	.
1	1	3	2	2	1	.	.	0	.	.
1	1	1	1	1	1	.	.	0	.	.
1	2	3	2	2	1	.	.	1	.	.
1	2	3	3	3	1	.	.	0	.	.
1	2	2	1	2	1	.	.	0	.	.
1	2	3	2	3	1	.	.	0	.	.
1	3	3	3	3	1	.	.	1	.	.
1	3	2	1	2	1	.	.	0	.	.
1	3	3	3	2	1	.	.	0	.	.
1	3	2	3	1	1	.	.	0	.	.
1	4	3	2	4	1	.	.	0	.	.
1	4	2	3	4	1	.	.	0	.	.
1	4	3	2	4	1	.	.	1	.	.
1	4	2	3	4	1	.	.	0	.	.
1	.	.	.	.	1	2	2	.	4	3.24

Unlike the 3-file format, the choice response variable 'dep1' is now a dummy variable coded '1' or '0', where '1' flags the selected alternative.

Figure 2-2: Response for the 1-file data structure

The corresponding syntax for the 1-file format:

choice = 1;
variables
  caseid caseid;
  choicesetid setid ;
  dependent dep2, dep3 continuous, dep1 choice;
  independent indep1, indep2, indep3;
  attribute att1, att2, att3;
  latent
    Class nominal 2;
equations
  Class <- 1 + indep1;
  dep2 <- 1 | Class + indep2 | Class + indep3 | Class;
  dep3 <- 1 | Class + indep2 | Class + indep3 | Class;
  dep1 <- att1 | Class + att2 | Class + att3 | Class;

Data Fusion Case Study Example

N=305 lab managers responded to MaxDiff (most-least scaling) questions involving the importance of 36 attributes posed in sets of 5 at a time, followed by records pertaining to ratings (5-point scale) for each of these 36 attributes.

A Data Fusion syntax model for the extended 3-file structure:

```
choice = 3
   alternatives 'alt.sav'
   id=altn
   choicesets 'SetLM.sav'
   id=setid2;
variables
   caseid ID;
   repscale sweight;
   choicesetid setid2;
   dependent RATING ordinal, Q5 ranking;
   independent Index nominal, set nominal inactive;
   attribute attr nominal;
   latent
      Class nominal 4, sClass nominal 2, Scale continuous, CFactor1 continuous;
equations
   Class <- 1;
   sClass <- 1;
   Scale <- (s) 1 | sClass;
   (0) Scale;
   (1) CFactor1;
   Q5 <- attr scale | Class;
   RATING <- 1 + Class + CFactor1 + Index | Class;
```

For the fused model, we combined the ratings random intercept model with the MaxDiff model for all 36 attributes. We estimated both models simultaneously with one discrete latent factor underlying both models.

The best fitting fused model was four classes with a Scale Factor:

		LL	BIC(LL)	Npar	L²	df	p-value	Class.Err.
class2	Syntax (1)	-23842.21	48536.75	149.00	45040.43	156.00	3.8e-9559	0.03
class3	Syntax (2)	-23577.97	48420.13	221.00	44511.94	84.00	1.2e-9537	0.05
class4	Syntax (3)	-23356.04	48388.12	293.00	44068.08	12.00	2.4e-9550	0.04
class5	Syntax (4)	-23164.99	48417.90	365.00	43685.99	-60.00	.	0.02

Table 2-1

We didn't find a significant correlation between the Scale Factor and Classes. The BIC was also higher for this model and the change in L^2 between the two models was not significant given the additional parameters estimated:

		LL	BIC(LL)	Npar	L²	df	p-value	Class.Err.
class4	Syntax (1)	-23354.96	48403.14	296.00	44065.93	9.00	2.2e-9555	0.04

Table 2-2

4 Class (Scale Factor Correlation with and without Class)

	L²	Npar	p-value
Without Correlation	44068.08	293.00	
With Correlation	44065.93	296.00	
Difference	2.14	3.00	0.54

Table 2-3

Below are the tables for the two different coefficients for the fused 4 class model:

	Class				sClass			
	1	2	3	4	1	2		
Item	22.1%	25.0%	27.6%	25.4%	72.9%	27.2%	Average	Variance
1	4.75	2.66	1.30	2.28	1	.5929	2.65	2.11
3	2.54	2.21	1.25	2.52			2.10	.37
2	1.85	2.36	1.79	2.00			2.00	.07
4	.61	2.69	1.32	.95			1.41	.83
5	1.03	1.01	.77	1.59			1.09	.12
6	1.69	.61	.83	1.12			1.04	.22
7	1.16	.48	.50	1.46			.88	.24
8	.16	.77	.35	1.24			.64	.23
9	1.11	.61	.40	.42			.62	.11
10	.79	.94	.18	.42			.57	.12
11	.31	.48	.16	.60			.39	.04
12	.53	-.24	.30	.57			.28	.14
13	.95	-.02	.15	.04			.26	.21
14	-.08	-.03	-.15	.96			.18	.28
15	.98	-.17	.15	-.37			.12	.35
16	-.18	.65	-.21	-.65			-.10	.29
17	.36	-.46	-.08	-.26			-.12	.12
18	-.54	.31	.55	-.95			-.13	.50
19	-.74	-.30	-.55	.82			-.18	.49
20	-.84	-.04	.74	-.83			-.20	.57
21	.03	-.52	-.52	-.22			-.32	.07
22	-.55	-.67	-.31	-.02			-.38	.08
23	-.49	-.29	-.70	-.10			-.40	.07
24	-.84	1.27	-.72	-1.31			-.40	1.31
25	.06	-.86	-.38	-.72			-.49	.17
26	-.58	-.30	-.25	-1.03			-.53	.13
27	-.89	-.50	-.14	-.78			-.56	.11
28	-1.19	-.88	-.11	-1.15			-.81	.25
29	-1.55	-1.80	-.23	-.06			-.87	.80
30	-1.76	-.94	-.35	-.80			-.92	.34
31	-1.27	-.70	-.97	-.86			-.94	.06
32	-1.62	-1.56	-.14	-1.07			-1.06	.47
33	-1.09	-1.45	-1.02	-1.42			-1.24	.05
34	-1.17	-1.54	-1.33	-1.40			-1.36	.02
35	-1.70	-1.64	-1.15	-1.29			-1.43	.07
36	-1.87	-2.14	-1.42	-1.70			-1.77	.09
							Average	.32

Table 2-4: Ratings Coefficients

Profile							
Class				**sClass**			
1	**2**	**3**	**4**	**1**	**2**		
22.1%	25.0%	27.6%	25.4%	72.9%	27.2%	**Overall**	
.01	.02	.01	.02	.02	.02	.02	**Not Important**
.07	.10	.07	.09	.08	.08	.08	**2**
.26	.29	.30	.29	.29	.29	.29	**3**
.37	.36	.41	.36	.38	.38	.38	**4**
.29	.23	.20	.23	.24	.24	.24	**Extremely Important**
3.85	3.68	3.72	3.70	3.73	3.73	3.73	**Mean**

Table 2-5: The overall profile of each class ratings

Item	Class 1 (22.1%)	Class 2 (25.0%)	Class 3 (27.6%)	Class 4 (25.4%)	sClass 1 (72.9%)	sClass 2 (27.2%)	Average	Variance
2	2.44	2.92	2.65	2.90	1	.5929	2.73	.05
1	3.11	2.93	1.10	2.36			2.32	.83
3	2.69	2.19	1.62	2.34			2.18	.20
7	1.82	.96	1.74	2.75			1.82	.54
5	1.33	1.47	1.32	1.97			1.52	.09
4	.48	2.77	1.47	.80			1.41	1.03
8	.49	1.52	.34	1.68			1.01	.48
10	1.05	1.17	.82	.83			.96	.03
6	1.77	.21	.77	.89			.88	.42
12	1.56	.14	.75	.97			.83	.34
13	2.06	.03	.56	.28			.69	.83
9	1.06	.35	.11	.25			.42	.18
17	1.24	-.42	.37	-.30			.20	.57
11	-.20	.54	.16	.01			.14	.10
14	-.54	.19	-.39	1.19			.12	.62
15	.88	-.42	.18	-.38			.04	.37
20	-.53	-.15	.86	-.38			-.02	.39
23	-.62	.57	-.56	.54			-.02	.44
25	.57	-.67	-.57	-.33			-.28	.32
16	-.22	.64	-.67	-1.04			-.33	.52
18	-1.28	.37	.26	-1.22			-.43	.82
24	-.89	1.73	-1.14	-1.78			-.53	2.39
22	-.66	-.97	-.12	-.44			-.53	.13
21	.11	-.88	-.60	-.76			-.55	.20
19	-1.27	-.95	-1.23	1.09			-.58	1.27
27	-.78	-.55	-.32	-.81			-.60	.05
26	-.51	-.25	-.33	-1.49			-.64	.33
29	-1.37	-2.15	-.20	.25			-.83	1.20
32	-1.96	-1.47	.52	-.71			-.84	1.17
30	-1.76	-.77	-.66	-.49			-.89	.33
31	-1.76	-.60	-1.20	-1.06			-1.14	.23
28	-1.30	-1.51	-.38	-1.68			-1.19	.33
34	-1.35	-1.80	-1.58	-1.99			-1.69	.08
33	-1.38	-2.14	-1.70	-1.87			-1.78	.10
35	-1.91	-2.15	-1.79	-1.79			-1.90	.03
36	-2.36	-2.86	-2.14	-2.59			-2.48	.09
							Average	.47

Table 2-6: MaxDiff Coefficients