software review

A Big Step Forward in Latent Class Analysis

Syntax trumps GUI.

By Ken Deal

I don't usually write reviews on beta versions of software. However, after receiving a nearly final beta version of Latent GOLD 4.5 Syntax Module (LGS for short), I felt that this application is important enough to get the news out quickly. You just have to give a hand to folks like Jay Magidson and Jeroen Vermunt for hopping on new methodology and getting it to us in very short order in an easy-to-use package.

For those not familiar with Latent GOLD (LG), it is an application provided by Statistical Innovations (SI) for conducting latent class analysis. An extension, Latent GOLD Choice (LGC), provides latent class analysis for choice-based conjoint data. LG and LGC provide a broad range of features, extensive output, and excellent graphics. The authors of LGS, Jay Magidson of Statistical Innovations and Jeroen K. Vermunt of Tilburg University, have published many journal articles on this topic and are considered to be prominent authorities on latent class analysis.

One of the key benefits of LGC is that it will accept choice-based conjoint files in *.cho format directly from a study conducted using Sawtooth Software and provide extensive analyses, including segmentation. I've used Latent GOLD Choice for many years and have always been pleased with its functions and usability. LG was reviewed in the Winter 2000 issue of *Marketing Research*, and LGC was reviewed in the Winter 2003 issue.

When I heard that Statistical Innovations was going to introduce a syntax version of Latent GOLD, I felt fairly sure that I would review the software but, truthfully, I was not thinking that this would be a valuable or enjoyable experience. Having a syntax version of an established program might be considered by some to be a reversal on the normal order. I was not prepared to be impressed by LGS. After all, how could syntax improve on the already excellent Latent GOLD Windows interface? How wrong I was!

It seems that Vermunt and Magidson felt the need to add power to Latent GOLD and thought that flexibility could be better accessed through syntax rather than a more complicated GUI. This is almost like having a really sophisticated and powerful car, maybe a BMW M5, and then opening it up and adding even more power. One thing's for sure—the power is there for the use, and it's quite easy to program. In fact, the LGS scripting language is about as simple as one could find, and the effect is very powerful. I've found over the years that the folks at Statistical Innovations have the rare ability to make their applications look misleadingly clean and basic while providing exceptional statistical and graphical benefits.

Getting started in LGS is very simple and quick. Highly effective educators understand that students learn best through experience with examples. Magidson and Vermunt designed a very clever process whereby many syntax and GUI examples can be accessed through the Help menu. As shown in Exhibit 1, using that facility is as easy as slipping along a cascading menu stream until the desired example file is found and activated.

Also, a new or existing model can be specified using the conventional LG GUI in LG4.5 in the normal way by running a latent class analysis. After the calculations have stopped, pull down the Model menu and select Generate Syntax. That

This is almost like having a really sophisticated and powerful car, maybe a BMW M5, and then opening it up and adding even more power.

will provide the starting script for LGS, which includes a list of options, variable specifications, and equations for developing alternative models using syntax and running those models. Executing syntax simply involves having that code displayed in the output window and clicking an arrow in the menu bar.

Syntax format. The LGS syntax is comprised of three main sections: options, variables, and equations. The options section provides settings and parameters for the algorithm, starting values, Bayes components, Monte Carlo process, estimation commands, missing value treatment, output specification, and whether any resulting data should be saved to an output file.

The variables section specifies three types of ID variables (all are not always needed) and weighting variables; lists and characterizes the dependent and independent observed variables; identifies the latent variables and their attributes; specifies the sample design for complex samples and other options that provide flexibility over the inclusion of some of the values of variables; and identifies those cases to be used

as a holdout sample. In addition, stacked data produced by simulations and multiple imputations of missing values can be identified for analysis by using the commands "simulationid" and "imputationid," respectively.

The equations section defines the model that will be estimated. LGS provides such tremendous diversity for model creation that a full description would completely fill this magazine. Because latent class analysis is based on variations to the multinomial logit model, the equations are all of a regression type or similar forms. For example, the key script elements for specifying a two cluster regression is listed in Exhibit 2. The *|* indicates the conditional model where values of the independent variables are allowed to vary over the two clusters.

Facilities for imposing restrictions such as monotonicity constraints, equalities between parameters or specific levels of parameters, starting conditions, and variance conditioning are quite easy to apply.

Because I've reviewed all of the products of Statistical Innovations, I'm familiar with their valuable inclusion of helpful facilities while providing a very clear palette. When building a model in the output window, right-clicking produces a floating menu from which keywords, variables, files and models can be selected. For example, when adding variables to the syntax, right-clicking brings forth a menu of variable names that can be clicked for inclusion in the equation rather than typing the names. The Files menu allows immediate browsing for files, and Models provides a list of eligible models for use—all very thoughtful and handy.

Clustering and segmenting. Latent GOLD and Latent GOLD Choice provide the most comprehensive integrated options among software applications for grouping respondents. Now LGS surpasses both of its predecessors with a span of clustering options so broad that some users may never explore all of the possibilities. The likely result of this flexibility is that LGS will find its way into the lexicon of methodology for almost any analyst who needs to cluster.

Clustering in the repertoire of LGS ranges from basic latent class cluster analysis through latent class regression, factor analysis, IRT, latent growth models, and into the more esoteric areas of D(iscrete) Factor and C(ontinuous) Factor models and path models. In fact, all of the GUI models available in LG and LGC have been expanded with additional options and with new facilities for imposing restrictions.

Models can be specified using a series of regression equations—for latent variables and for dependent variables, independent variables, covariates, and grouping variables. This focus provides wide flexibility for analysts because rather complex relationships can be described by the syntax quickly and with fewer opportunities for errors than in other syntax-based applications. And those syntax models can contain any combination of nominal, ordinal, and continuous latent variables.

All the other models that have been available in Latent GOLD can be modeled in LGS. Plus, many other features are available with greater flexibility than when using the GUI alone. For example, DFactor latent variables can be nominal in addition to dichotomous or ordinal, as in the GUI DFactor model. Also, single, exploratory, confirmatory, and path model DFactor models can be modeled very easily.

Regression models. The LG facility with regression has been expanded in LGS to include interaction terms and a wide assortment of restrictions. The most valuable extension of regression is the opportunity to have multiple dependent variables that uses the stacking feature of LGS.

Markov models. A brand new feature is the latent Markov model, a dynamic latent class model where latent class membership can vary over the time points. This interesting development may provide an excellent platform for investigating dynamic patterns of segment membership. Parametric and non-parametric bootstrapping using Monte Carlo simulation

Exhibit 1 Accessing ready-made syntax

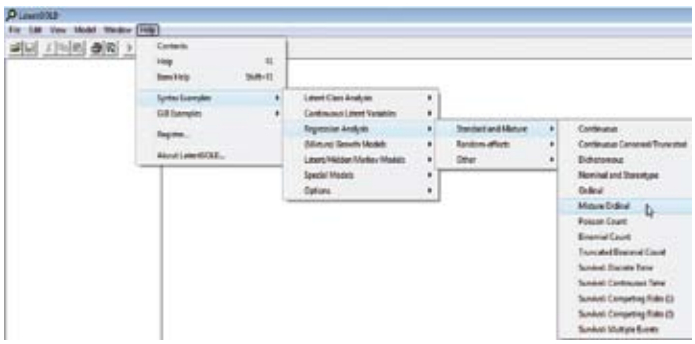

Exhibit 2 LGS Syntax for a two-cluster regression model

Variables

```
independent noisy, arr, way, modern, contrib, rules, frnd,
busy, know, appr, hours, impper, impcar;
dependent satis;
```

Latent

```
Cluster nominal 2;
```

Equations

```
Cluster <- 1;
satis <- 1 + noisy|Cluster + arr|Cluster + way |Cluster + modern|Cluster
+ contrib|Cluster + rules|Cluster + frnd|Cluster + busy|Cluster
+ know|Cluster + appr|Cluster + hours|Cluster + impper|Cluster +
impcar|Cluster;
```

and other new options are included. Time homogeneous and time heterogeneous transitions are allowed, and grouping variables and covariates can be modeled.

Multiple observations. For panel-type data, where each case has multiple observations (i.e., repeated measures or stacked data), multiple dependent variables can be specified, thereby greatly expanding the regression capabilities of Latent GOLD.

Multiple imputation of missing values. The imputation of missing values using methodologies that produce multiple data sets, thereby recognizing the probabilistic nature of the estimation process, has made great strides over the past 10 to 15 years. The process of using Amelia and AmeliaView to produce multiple imputations in the R language and then using the R package Zelig to perform analyses using those several data sets was reviewed in the Summer 2007 issue of this magazine. Also, SOLAS and NORM, two other multiple imputation applications, were reviewed in the Fall 2004 issue.

LGS now includes a two-step process for multiple imputation and analysis of the multiple data sets; while this is two stages, everything is done within LGS. The non-parametric bootstrap procedure generates multiple versions of the missing values, with the default being 10 data sets. The really inventive addition is where those multiple data sets are stacked and made available for estimation of any of the models available in LGS.

Simulations. Monte Carlo simulation of data sets is achieved by specifying a population model, supplying an example data set in a *.dat or spss file, and writing the appropriate syntax.

Due to the tremendous flexibility of LGS and our limited space, still other very valuable features will just not fit in this review. Perhaps the best way is to just add a few words about some more features and leave the remainder up to your investigation.

The scale factor. Multinomial logit and related approaches often assume that variances are independently and identically distributed (iid). However, during the past 15 to 20 years, several prominent researchers have stated that iid is a simplifying assumption that leads to the estimation of model parameters that may include scale factors in addition to the "real parameters." The scale factor has been shown to be proportional to the inverse of the standard deviation (i.e., the parameter is confounded with the error variability). This topic was addressed at the 2006 Sawtooth Software Conference by Louviere and Eagle and at the 2007 Conference by Magidson and Vermunt. LGS provides a fairly straightforward procedure for estimating the scale factor and, for some, this ability will fully justify the price of this application.

Other features. With all of the variables, clusters, factors, stacking, and restrictions that are available in LGS, keeping track of all the parameters can be challenging with large models. SI has implemented a clever pop-up "right-click facility" through which the internal number associated with each parameter is listed in the left column of the parameters output screen. LGS allows the syntax to be saved "with parameters" (i.e., the parameter values are saved along with the other model settings). Those parameters are saved in a list contained within curly parentheses {.} and can be used as starting values to begin the iterative analysis once again after having stopped the calculation. The index values in the parameter output window provide a convenient, almost vital, glossary for identifying the parameters tied to the parameter values. There are other more detailed benefits to the saving of parameter values.

In addition to all of this, SI provides a clever little sequence of steps where the parameters and variable definitions from a clustering estimation, for example, can be saved, altered in a very minor way, read back into LGS, and used to score a holdout sample or customer database.

Conclusion

Those who are relatively new to Latent GOLD or who are not statisticians by training and have a strong interest in expanding their analysis options will enjoy being able to start with a GUI study, generate the syntax, and expand that base script to investigate other models. Those who have good to excellent statistical education and experience may find that LGS provides a very broad and sturdy environment for conducting much of their multivariate analyses. Plus, the method developed for providing examples is very easy to use, and those examples can provide many hours of statistical pleasure.

Some users of LG and LGC will find that the format of some of the output is not as easy to use, and probably all users of LG will miss its excellent graphical facilities. None of the graphics from LG have been adapted to LGS, although SI intends to add the graphics at some later time, perhaps by the time this review appears in print. While I understand the purpose for the new arrangement of the output screen of parameter values and related statistical tests, that format makes it more difficult to move the data into Excel for graphing and further investigation. Perhaps the Latent GOLD Formatter of Stats Wizards will be adapted to move the output from LGS into Excel as it so cleverly does for LG output. (See the Spring 2008 issue.)

LGS is available for downloading and testing from the SI Web site (www.statisticalinnovations.com). The incremental price of $295 over the cost of LG (a tentative price at the time of writing) seems well designed to draw substantial attention, especially since it is a one-time fee. A substantial amount of support material is available for downloading from the site, including many articles that range from the basics to highly technical papers published in academic journals. ●

Ken Deal is in strategic market leadership and health services management at the DeGroote School of Business, McMaster University, in Hamilton, Ontario. He is also president of marketPOWER research inc. in Winona, Ontario, and St. Joseph-du-Moine, Cape Breton Island, Nova Scotia. He may be reached at kendeal@marketpowerresearch.com.