

Mining for Gold gets easier and a lot more fun!

By Ken Deal

Marketing researchers develop and use scales routinely. It seems to be a fairly common procedure when analyzing survey data to assume that a seemingly well-developed scale has equal intervals between the scale values. Whether using a Likert-type scale, an importance scale, satisfaction, likelihood of purchase or a myriad of other scales for measuring respondents' attitudes, we tend to assign equal-spaced integer values to the scale, whether numbers were used on the questionnaire or not.

Students taking basic marketing research courses are typically told to scrutinize the scales to ensure that the intervals appear to be equi-distant between anchored points. Rarely are these scales tested to determine whether in fact respondents agree with the assessment of equal appearing intervals. Devoted statisticians continue to be concerned with this question. However, those concerns have largely not been communicated to practicing marketing researchers in understandable and convincing dialogues. And since practitioners need to turn around data analysis in order to satisfy timely marketing decision making requirements, if the software is not available to handle these seemingly esoteric concerns of statisticians, the normal techniques for analyzing data, i.e., cross-tabs, regression, etc., using consecutive integers to mark categories have seemed to be adequate.

Jay Magidson, owner of Statistical Innovations, has developed and commercialized GOLDMineR 2.0 (GMR), a statistical application that allows marketing researchers who are concerned about the assumptions used with survey data to follow a more cautious approach to analyzing data sets that include ordinal dependent variables and predictor variables having a variety of information properties.

GMR was developed to provide an easily usable application for general regression analysis. This means that GMR can be used in situations where the dependent variable is either qualitative or quantitative and where the predictors are also of either information level. Both dependent and independent variables can be dichotomous, ordinal or grouped continuous. This objective makes it necessary to use a general model of logit estimation that will also treat conventional linear regression needs.

An innovative feature of GOLDMineR is that the interval properties that are often assumed for seemingly well-developed scales do not need to be imposed on the data analysis. Magidson's rendering of GMR based on monotonic regression alleviates many of the potential problems that can be forced on data when using linear regression or logistic regression. Specifically, instead of forcing the integer scaling of 1, 2, 3, 4, 5 to an intention to buy scale, the spacing of the scale responses can be estimated simultaneously with the values of the parameters of the predictor variables. Also, predictor variables rarely have ratio properties and often the assumed interval properties are not met. Ordinal characteristics have sometimes been ignored in favor of assuming the interval characteristics that are more directly manageable with general purpose statistical packages.

While GOLDMineR operates in a less restrictive environment, there are still constraints on what is permissible analysis within the application. The assumption of monotonicity and the statistical requirements of maximum likelihood estimation of a log odds model are central to the process.

It is a pleasure to find applications built on solid technical foundations that also produce graphical output that is attractive, instructive and flexible. Statistical Innovations has certainly proved to be creative and resourceful in providing a tremendous range of flexibility in the graphing in GMR and in its sister product, Latent GOLD Choice 3.0. I've remarked in earlier reviews of Latent GOLD and Latent GOLD Choice that the outputs from those applications, and especially the graphical output, provide outstanding opportunities to investigate modeled data in great detail. GOLDMineR 2.0 also provides an extensive range of numerical output.

Statistical Innovations has ensured that analysts can easily use the output. GM graphs can be seamlessly be copied to other applications, such as PowerPoint and Word, problem free. And it is particularly nice to be able to copy and paste the tabular output directly into Excel without having to format the output so that each number fits into its own individual cell.

Now, what are the benefits of using GOLDMineR compared to linear regression or logistic regression or multinomial logit analysis? Since linear regression and logistic regression are included in most statistical packages, why learn something new and have to pay a substantial additional fee?

To illustrate these benefits, data from a survey conducted for Firestone Canada Inc. approximately twenty years ago, and released for publication purposes, has been analyzed by GOLDMineR after

having been analyzed originally using linear multiple regression. The dependent variable is intention to return to Firestone for automotive service (V088) as measured on a five point scale from definitely will not return (1) to definitely will return (5) with each point anchored. The three predictors are “all services were done correctly” (V075), “the service manager was trustworthy” (V079), and “they provided prompt attention” (V087). The predictors were measured on a five point scale from strongly disagree (1) to strongly agree (5) with all points anchored.

The data was in SPSS format and many analysts will highly value the benefit of importing SPSS data directly into GMR. Data of other formats and in array format can be entered into GMR quite easily.

The Search option in GMR was used to identify and select those variables that could be considered as significant predictors of intention to return to Firestone. Furthermore, in addition to operating in a fashion similar to a stepwise procedure, Search also ranks the variables based on their unique p-values. Search can also be used for evaluating predictors one-at-a-time. Two additional significant predictors were removed to make the following output more manageable for presentation.

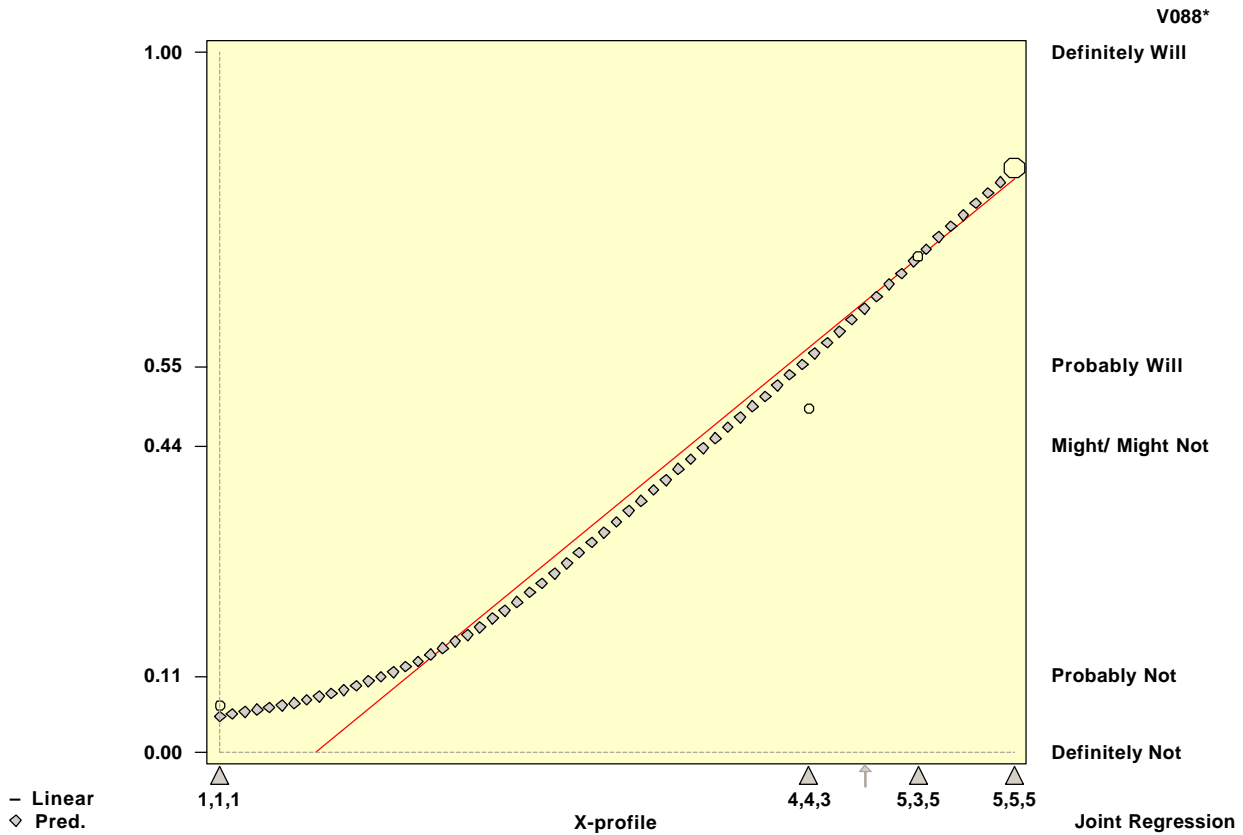
Exhibit 1 The statistical table from GOLDMineR analysis of scaled data.

y: V088 (Free)	Definitely Not	Probably Not	Might/ Might Not	Probably Will	Definitely Will
Y-scores	0	0.11	0.44	0.55	1
Alpha(j)	0	-0.67	-3.28	-3.9	-10.23
	L²(Y)	df	p-value	Beta	exp(Beta)
V075 (Free)	152.08	4	7.30E-32	7.26	1424.05
V087 (Free)	55.72	4	2.30E-11	4.49	89.35
V079 (Free)	34.84	4	5.00E-07	4.37	79.11
Association Summary	L²	df	p-value	R²	phi
Explained by Model	463.87	15	2.40E-89	0.413	1.1053
Residual	308.09	349	0.94		
Total	771.96	364	1.80E-31		

The Y variable and the three X variables were specified as free in this run and were held as fixed in a parallel run of the model. The first row of the fixed model listed the Y-scores on the original scale of 1, 2, 3, 4, and 5. By setting Y to be free, GMR fixed the end points at 0 and 1 by default and freely estimated the intermediate Y-scores simultaneously with the other components of the model. Rather than following the equally spaced scale of consecutive integers, the Y-scores are spaced differently in the free estimation model, with the greatest distance being between definitely will return and probably will return to Firestone, i.e., 0.55 and 1.00, respectively. (See line 2, Y-scores, of Exhibit 1.)

Exhibit 2 shows graphically the spacing of the original integer values (SPSS value labels shown here) of V088 on the right vertical scale and the values estimated by GM on the left vertical scale. This is a joint regression so the horizontal scale attempts to represent the estimated spacing of the three independent variables. Since the value labels overlapped, they were not used for the X-variables in these graphs. Partial regressions to investigate the relationship of each independent variable can be graphed in seconds. The red linear regression line in Exhibit 2 follows the logit line closely for the higher values of the dependent variable V088, but diverges in the lower range of values.

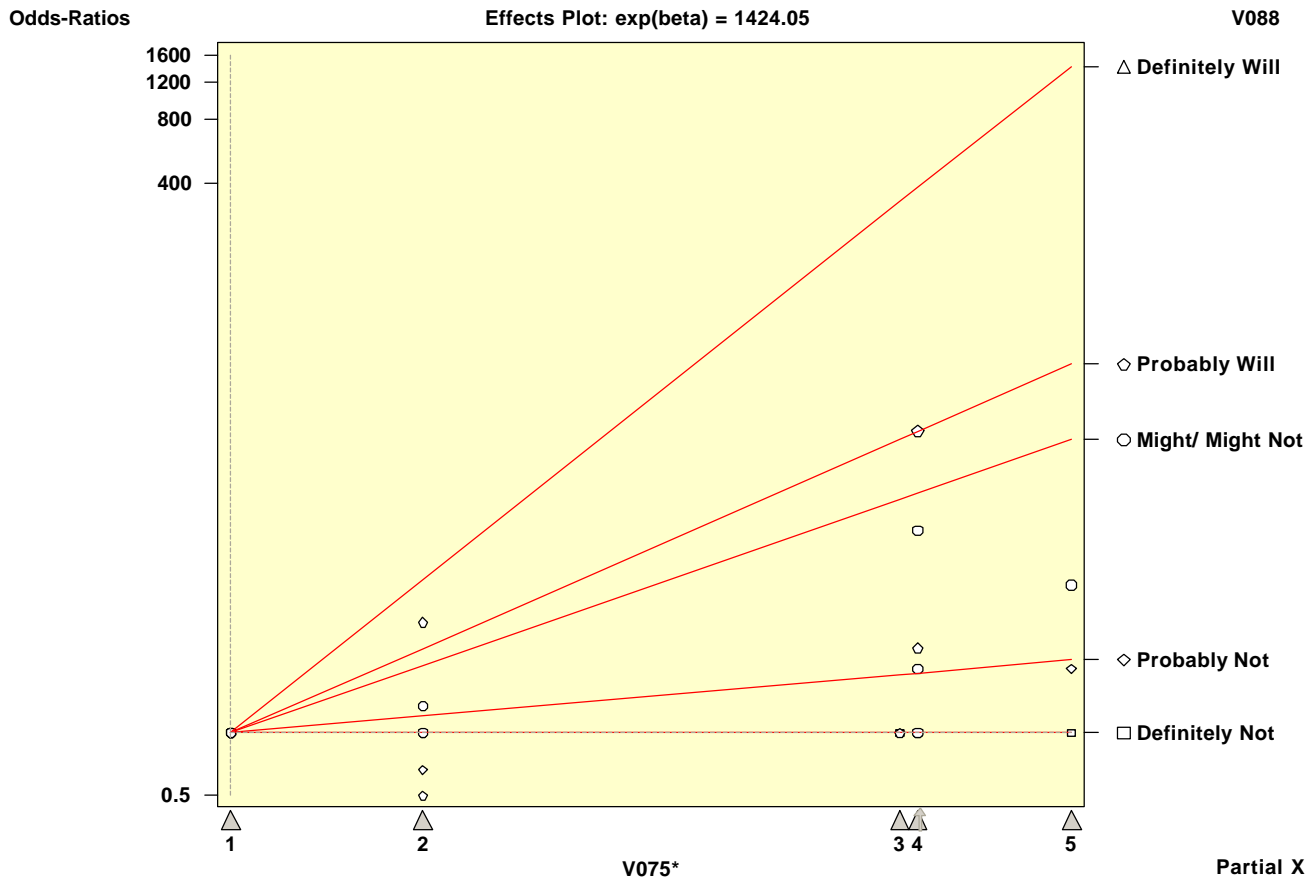
Exhibit 2 The joint regression graph for freely estimated dependent and three independent variables.



Is there a difference between the estimation of free or fixed variables? Yes, but marginally. The L^2 -residual for the fixed model declined from 351.86 (p-value of 0.63) to 308.09 (p-value 0.94) for the free model. (See line 3 of the association summary in Exhibit 1.) The R^2 values for the free and fixed models in GMR and for the linear regression estimated in SPSS are very similar.

The Betas in Exhibit 1 are interpretable as log-odds ratios and $\exp(\text{Beta})$ as odds ratios. Odds ratios and 95% confidence intervals are tabled for each fixed X variable and odds ratios and confidence intervals are provided for each category of each free variable. While more statistical findings are produced by GMR, these are not provided here to save space. Linear regression of this data did not disagree directionally with any of the results produced by GMR.

Exhibit 3 The partial regression effects plot of intention to return and all services were done correctly (vertical axis represents odds ratios)



Partial regression effects plots are very helpful for interpreting relationships. For example, Exhibit 3 shows the partial effects plot for V075 (all services done correctly) on intention to return (V088). The vertical axis shows the odd ratios (log-odds ratios can be displayed alternatively) for intention to return to Firestone. The odds of replying “definitely will return to Firestone” are 1424 times higher if a respondent agrees strongly that “all services were done correctly” than if a person disagreed strongly with that statement. To compare the ratio of the odds of stating “definitely return” if the respondent strongly agrees that all services were done correctly versus if a person said that they neither agree nor disagree with that statement is very simple and quick: double clicking on the triangle for neither agree nor disagree (value 3) below the horizontal axis in Exhibit 3 causes the reference point to change to that central scale point. The odds ratios can then be read off of the graph. In addition, a pull-down menu allows the analyst to select “new statistics” to have the statistical table recalculated immediately with reference to the new base. Exhibit 3 is a partial X graph. A partial Y graph is also available. One has to work with these graphs to fully appreciate the wide range of features integrated into this mechanism.

A table coinciding with the plot can be produced to show observed or expected frequencies, probabilities, odds or odds ratios. Of course, the partial effects plot can be produced for each of the other independent variables while holding the other two independent variables at their default base values of disagree strongly, or at any other desired reference point.

Since GOLDMineR 2.0 is part of the Data Mining Advanced Kit offered by SI, it should have some strong mining features and it does, in addition to those mentioned above. A key GMR mining feature is the ability to save the SPSS code that will score an external file based on the estimated parameter values derived during a GMR session. This code can then be executed within SPSS to calculate the values of the predicted dependent variable for those cases that have not been included in the analysis stage.

GMR is a highly sophisticated and fully debugged application. Absolutely no problems were encountered during fairly extensive investigation of its features. However, it would be nice to have the ability within GMR to split the data set into an estimation component and a hold out sample. While this can be done outside of GMR, and the scoring feature of producing SPSS code helps this function, it would be good to a sample splitting feature directly within the application.

At first, GOLDMineR seemed to me to be a special purpose program with a limited audience. However, after testing the application on several data sets, I find that its usefulness can be fairly extensive for analysts who are involved regularly in regression-based analysis of survey data sets. This is especially true when wanting to very carefully investigate relationships based on survey data that lead to marketing actions for large numbers of customers residing in electronic company files. I've tried to highlight the main features of GMR in this article but be aware that there is still an extensive array of benefits that have not been mentioned here.

A demo of GMR can be downloaded from the Statistical Innovations web site, www.statisticalinnovations.com. GMR can be purchased by itself for US\$695 for one user, US\$595 for each additional user and there is a US\$200 annual fee for maintenance, support and updates. Academic copies are priced at US\$395 and site licenses for student class usage are available. GMR can also be purchased as part of the Data Mining Advanced Kit for US\$1,495.