How and when to use Binary Logistics Regression

M.M. Shankar

Introduction

In the research tools section, in the previous issue, Exploratory Factory Analysis was demonstrated. Since it is an exploratory method it cannot be help in the conclusions without doing further explanatory analysis such as regression or any other Multivariate dependence techniques. In this section, Binary Logistic Regression is demonstrated which is quiet useful especially when the dependent variable is Binary or Dichotomous.

In last Research tool section, three important Caveats were discussed. The first caveat is thinking of techniques before data collection, the second caveat is the scale of measurement and the third caveat is using interval scale as measurement on all variables except demographic. However, while designing the questionnaire or doing secondary research sometimes using nominal scale may be inevitable.

Example 1.a

Will you recommend this brand to other?

□ Yes □No

Will you prefer this brand for future use?

□ Yes □No

Example 1.b

The Bank has developed a statistical model to decide whether or not to provide or not to provide loan (Binary criteria) for customers based on customer's monthly salary, Age, Marital status etc.

In Example 1, the researcher can use a rating scale of 1 to 5 but options like "Yes or No" may be easier for the respondent. From the above examples it is understood that sometimes the dependent variable can be on a nominal scale. But in such a situation the researcher cannot not use traditional linear regression, generally used to estimating the dependent variable based on Independent variables.

To solve this problem, Binary Logistic Regression is used and it does not demand linear relationship between the variables like Discriminant Analysis. Refer Annexure 1 details of the difference between General Linear Regression and Binary Logistic Regression.

Executing Binary Logistic Regression

This section deals with the demonstration of binary logistic regression with help of SPSS software by assuming a hypothetical scenario of measuring customer satisfaction on various attributes of restaurant service such as food quality, food variety, bill handling, music etc., on a 5 point scale. It is assumed that the researcher runs factor analysis and names the factors. There are three factors: Factor 1: Service, Factor 2: Cuisine and Factor 3: Ambience. There is one more item in the Questionnaire which measures the intensity of the respondent's recommendation to others ('1' is recommend, '0' is not recommend). The data set is shown in SPSS window (in Figure 1). The dataset set contains three factor scores for three variables and one dichotomous Recommend variable. To run the Binary logistic regression in SPSS select Analyze→Regression→Binary→logistic. In Figure 2, the variable Recommend is kept dependent and other three variables placed in covariates before click OK.

The Results of Binary Logistic Regression

SPSS generates more tables, but here only the important ones are shown which are listed below:

1. Dependent variable encoding
2. Block 0: Beginning Block
3. Omnibus Tests of Model Coefficients
4. Model Summary
5. Classification Table
6. Variables in the Equation

Table 1 shows the internal value (coded value) for the given dependent variable encoding. Table 2 shows Block 0: Beginning Block, Classification of observed and predicated values only, with constant, but not any predictor variables. Estimate is made of without Independent variables, what the classification percentage is between observed and predicted value, which is measured by overall percentage of correct classification, that is 68.5 per cent . Table 5 is the Classification Table after Including Predictors, which can be compared with Table 2 Block 0: Beginning Block (classification with only constant). If the latter model, correct classification percentage has gone up to 92.7 per cent after inclusion of predictor variables. It clearly shows 24.2 per cent (92.7 per cent from 68.5 per cent) as incremental correct classification because of inclusion of predictor variables such as Service, Cuisine and Ambience. This classification percentage is analogous to R square in linear regression.

Table 3 shows the omnibus test to know the fit of the model. From this, the given model is significant at 0.01 level and so we can say that on the whole the model is predicting display rule understanding significantly better than it was with the inclusion of only the constant.

In Table 4 of Model summary, The -2Log-likelihood is based on summing the probabilities associated with the predicted and actual outcomes. It is similar to the residual sum of squares in Multiple Regression. In simple terms, it is an indicator of how much unexplained information is there after the model has been fitted. It is understood that larger the values of log likelihood, the statistical model is a poor fit. At this stage of the analysis the value of –2*log-likelihood @ 39.215 is a small value and the model is good fit. Similarly to R Square in Multiple regression, there is Cox and Snell and Nagarkelke which show a better value for binary logistic regression. However the accuracy of the model is based on the classification rule rather than available R square like Cox and Snell and Nagarkelke.

Table 6, (Variables in the Equation), is important because it tells us the estimates for the coefficients for the predictors included in the model. It shows the coefficient and statistics for the variables that have been included in the model at this point (namely, Services, Cuisine and Ambience and Constant). The b-value interpretation is the change in the logit of the outcome variable associated with a one-unit change in the predictor variable. The logit of the outcome is simply the natural logarithm of the odds of Y occurring (Not Recommended or Recommended). But except, Ambience the other two variables, namely Service and Cuisine alpha value are less than .05 and it is statistically significant @ 5% and it shows the expected positive sign. To generate the predicted Y in SPSS, click save button and choose probabilities and group membership from Predicted values. By default, the probability cut off value is 0.5 to determine particular cases as belonging to 0 (Not Recommended) or 1(Recommended).

Summary

When the dependent variable is on a nominal or dichotomous scale, to estimate logarithm relationship with other Independent variables, the binary logistic regression is suitable. The accuracy of the model is based on correct classification percentage. The significance of the overall model is based an on

omnibus test of model coefficient. Individual independent variables significance is shown in significant value in the Table of Variables in the Equation and to generate predicted Y value and groups, select Analyze→Regression→Binary→Logistic Regression→Save→Predicted values and select choose probability and group membership in SPSS.

Apart from the above explanation, other things have to be taken care of by the researcher while doing Binary Logistic Regression such as Hosmer-Lemeshow for goodness of fit, Classification plot and categorical covariates if the independent variable is on categorical scale. These aspects are not covered in this article.

References:

Andy Field, (2005), "Logistic Regression", *Discovering Statistics Using SPSS,* Sage Publications, 2nd edition, P 219-240.

Hair, Black, Babin, Anderson & Tatham, (2007), "Multiple Discriminate Analysis and Logistic Regression", *Pearson education*, 6th edition, page 379-402.

Table 1
Dependent Variable Encoding

Original Value	Internal Value
Not Recommended	0
Recommended	1

Table 2
Block 0: Beginning Block

Observed		Predicted		
		Recommendation		
		Not Recommended	Recommended	Percentage correct
Step 0	Not Recommended	85	0	100
	Recommended	39	0	.0
	Overall Percentage			68.5

Table 3
Omnibus Test of Model Coefficients

		Chi-square	d	Sig.
Step 1	Step	115.207	3	.000
	Block	115.207	3	.000
	Model	115.207	3	.000

Table 4
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	39.215	.605	.850

Table 5
Classification Table (a)

Observed			Predicted		
			Recommendation		Percentage Correct
			Not Recommended	Recommended	Not Recommended
Step 1	Recommendation	Not Recommended	80	5	94.1
		Recommended	4	35	90.0
	Overall Percentage				92.8

Table 6
Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp (B)	95.0% C.I.for EXP(B)	
							Lower	Upper
Services	3.093	1.562	3.922	1	.048	22.0	1.033	470.9
Cuisine	4.651	2.066	5.069	1	.024	105	1.826	6005
Ambience	-2.055	1.431	2.062	1	.151	.128	.008	2.116
Constant	-27.331	5.838	21.92	1	.000	.000		

Figure 1
The Data Set in SPSS Window

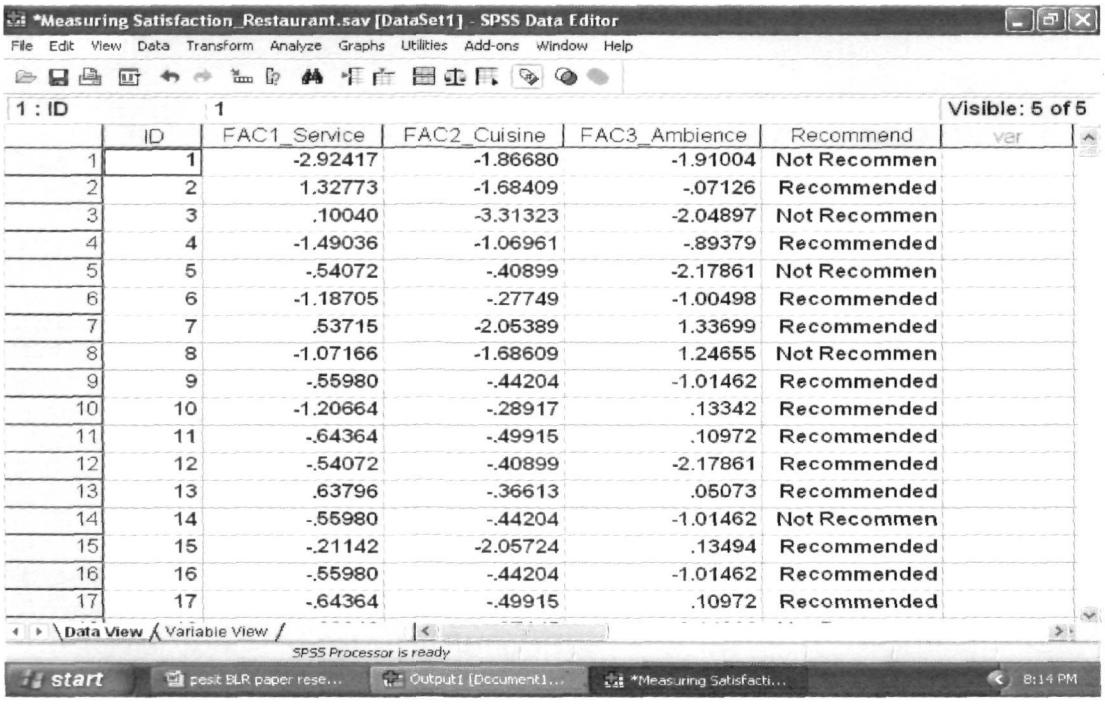
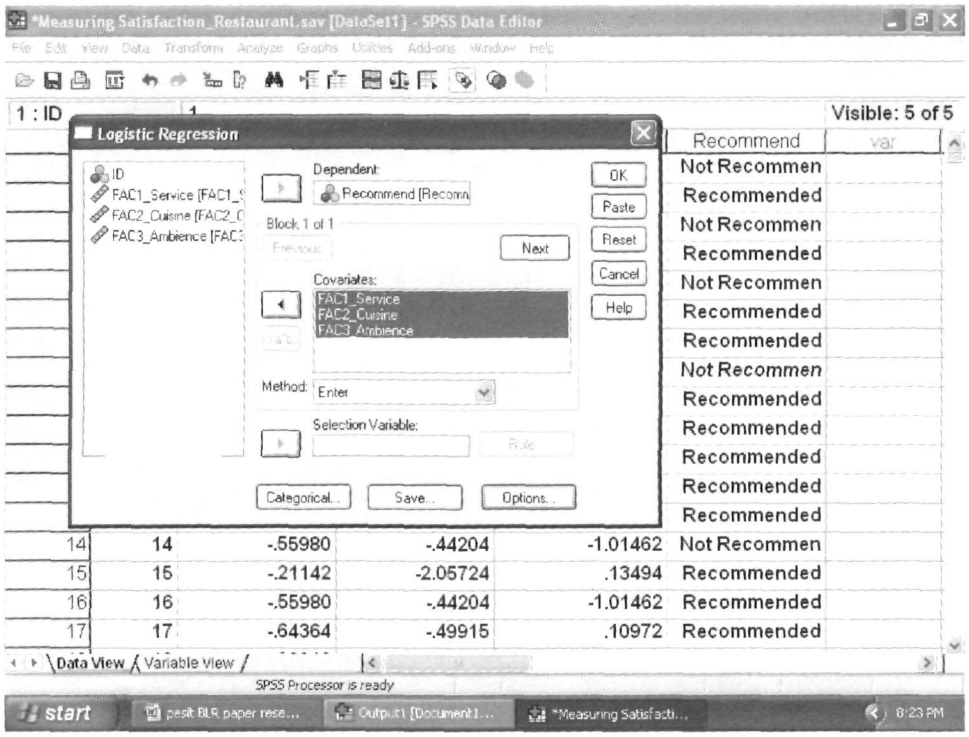


Figure 2
Operation of SPSS Tool



Annexure 1 Difference between Linear Regression and Binary Logistic Regression		
Feature	**Linear Regression**	**Binary Logistic Regression**
Nature of Dependent Variable	Continuous Scale (Ratio or Interval Scale)	Categorical Data (only two outcome) or Dichotomous value
Basic Equation	$Y = b_0 + b_1x_1 + \varepsilon_i$	Probability of Y occurring $P(Y)= 1/1 + e^{-(b0 + b1x1+\varepsilon i)}$
Relationship between variables	The relationship between dependent variable and independent variable is linear. The curve looks symmetrical.	The relationship between dependent variable and independent variable is logarithmic. The curve is 'S' Shaped.
Indicator of Influencing relationship in Model	R Square close to 1, fit of the model is good.	Corrected classification percentage compared with Block 0 Beginning. Closer to 100% of correct classification indicates a good fit. Log-Likelihood, larger values, indicates unexplained relationship. Lesser value is better for the model.
Individual contribution of Predictors	Regression coefficient and their standard errors to compute t-statistic	Wald statistics, Most useful is exp b (Exp(B) change in odds resulting from a unit change in predictor.

About Author:
M.M. Shankar is associated with Mark-How Consultancy, Bangalore as a Research Consultant. He can be reached at mmshankar@gmail.com