Nethra Sambamoorthi, Ph.D

CRMportals Inc.,
11 Bartram Road, Englishtown, NJ 07726
nethra@crmportals.com
732-792-3945

Variable Selection: A Class of Natural (to marketing) Scatter Plots That Rank and Explain Independent Variables in Ordinary Least Squares, Logistic, Probit, and Tobit Regressions

(N-Plot)

Key words: OLS Regression, Logistic regressions, Probit Regressions, Tobit Regressions, Limited Dependent Variable Regressions, Response Lift, Natural (to marketing) Plots (N – Plot) .

Analysts and statisticians plot dependent variable (say, y) vs. independent variable (say, x) to see the univariate relationship between y and x. This is not doable, if you want to determine the relationship between a dichotomous y and an independent variable x. To over come this difficulty what we do is typically called a rank plot and are common among database marketing analysts.

Bin the x values into deciles and plot log(mean(y)/(1-mean(y)) vs. binned means of x.

There still is a problem with this, when you eye-ball the plots. Because plots have tendency to high light tight regressions but the lift explained is relatively small compared to somewhat less tightly expressed regressions, but that explains lot more lift in response;

To consistently rank variables that explains lot more lift compared to ones that explains less lift, I use lift in the response (y) among the binned x groups, that is find %change (lift) in the logits from the average logit for each of the binned x groups and plot the lift vs. binned x value averages. Note that bins are naturally the categories if x is a categorical variable. If you want one line output, (a common requirement in data mining situation) use range of the lift for plots that look nice in terms of mathematical relationship between lifts and the averages of (x) among the bins **(1)**. The plots with highest variation (assuming other things are not wacky with the variable) are the best variables. Do these plots for all relevant variables to understand the importance and strength of relationship between response and the independent variables. You can do lift in the response probabilities also, but it is not natural to the transformation we use.

Now, using the same principle, we can have a natural plot that explains Tobit regressions. For Tobit regressions, the values we want to plot are the following.

Probability [event with in the binned x]*(mean y within the binned x) vs. mean (x with in the bin) (as mentioned before, the binning works naturally with categorical variables). – let us call this as T-means (meaning tobit means conditional means).

The natural (natural to marketing) plot to look at is the plot of pairs of values of
(x mean, (tobit means/ whole sample mean)*100) for the binned groups. This works for categorical x variable also, and as before the bins are typically the categorical classes, unless the categories are too many and you want to combine categories in that situation.
Thus if we do demi-decile groupings for a continuous x, there will be 20 values to look at. **(2)**. Note that this is applicable in OLS also. Even if you look at the (x mean, tobit means) for all the bins, you will get a natural statistic relevant for the Tobit model, but some anomalies will creep in the selection of variables.

This is an integrated method to look at plot, in the event of trying to explain consistently the importance of a variable and the summary measures we print by a routine application of different procs such as Proc Logistic, Proc Reg, and Proc LifeReg.

The output of Proc LifeReg should be consistent with the eye-balling of the plot, if Minimum Chi-square method is used for estimation and for large sample sizes and reasonably well behaved likelihoods both Minimum chi-square and maximum likelihood methods should give the same results.

PS: For simplicity I used to call this in my presentation as N-Plot (natural lift plot), as people had so many plots in organizations; funnily people use to refer this as "using Nethra's plot" in some of my consulting work.

I think these plots take away the confusion in Tobit regressions, and everyone will be encouraged to do Tobit regressions with out guilty feelings of trying to explain complexity.

Exercises:

What is a natural plot for probit regression, hazard regression, … ?