RESEARCH QUESTIONS RELATED TO PREDICTING PROBABILITY OF SUCCESS

BY: Togar Alam Napitupulu, Ph.D

Regression also can be used to predict the value of Y given the value of Xs. In such case we are actually building a regression model or a "machine", a learning machine by estimating the parameters of the regression. During the estimation of the parameters of the regression, in the jargon of data mining, we say that the machine is learning from the data. It is sometimes called "supervised learning", because the parameters are forced (supervised) to fit the pairs of Y and the Xs.

Consider for Example ones want to build a "machine" (a regression) that is able to distinguish non-performing loan vs. well-performing loan customers. The Y then would be loan performance of the customers having values of 1 for well-performing borrower and 0 for non-performing borrower. The Xs would be the variables or characteristics of the borrowers that we think might be related to or to cause performance of the borrowers. Notice however, because Y is binary (0,1), it violates the normality requirements of the traditional Ordinary Least Square Method of estimation. To resolve this, we take the regression of the logarithm (base 2) of the "odds ratio" against the Xs, following this logistic model:

$$\ln(\text{odds ratio}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_n X_{ki} + \varepsilon_i \quad \ldots \quad (1)$$

Where

$$\text{odds ratio} = \frac{\text{probability of an event of interest}}{1 - \text{probability of an event of interst}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

Suppose an event of interest is well-performing loan. Probability of a well-performing loan (or the probability of an event of interest) then is the frequency of performing loan in the sample. The estimated regression equation is:

$$\ln(\widehat{\text{odds ratio}}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots + \hat{\beta}_n X_k \quad \ldots \quad (3)$$

Odds ratio then (the estimated one) is

$$\text{odds ratio} = e^{(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \ldots\ldots\ldots\ldots\ldots + \hat{\beta}_n X_k)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

Having known the estimated odds ratio, we can now substitute it back to equation (8) to get the (estimated) probability of an event of interest as follows:

$$\text{Probability of an event of interest} = \frac{\text{estimated odds ratio}}{1 + \text{estimated odds ratio}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

The learning process stops right at equation (9) the "machine". Next, this machine can now be used to predict probability of a new particular customer whether he is going to be a well performing borrower using equation (11).

This is actually a special case of “Discriminant analysis”, where the possible value of Y not only 1 or 0; but could be more than two categories.