In This Issue

1. Learn about limited dependent variables
2. Using Risk Simulator's Maximum Likelihood Models to identify which variables affect the default behavior of individuals

"How can we model situations involving dependent variables?"

Contact Us

Real Options Valuation, Inc.

4101F Dublin Blvd., Ste. 425,
Dublin, California 94568 U.S.A.

admin@realoptionsvaluation.com
www.realoptionsvaluation.com
www.rovusa.com

Theory

The term *limited dependent variables* describes the situation where the dependent variable contains data that are limited in scope and range, such as binary responses (0 or 1) and truncated, ordered, or censored data. For instance, given a set of independent variables (e.g., age, income, education level of credit card or mortgage loan holders), we can model the probability of default using maximum likelihood estimation (MLE). The response or dependent variable Y is binary; that is, it can have only two possible outcomes that we denote as 1 and 0 (e.g., Y may represent presence/absence of a certain condition, defaulted/not defaulted on previous loans, success/failure of some device, answer yes/no on a survey, etc.). We also have a vector of independent variable regressors X, which are assumed to influence the outcome Y. A typical ordinary least squares regression approach is invalid because the regression errors are heteroskedastic and non-normal, and the resulting estimated probability estimates will return nonsensical values of above 1 or below 0. MLE analysis handles these problems using an iterative optimization routine to maximize a log likelihood function when the dependent variables are limited.

A logit, or logistic, regression is used for predicting the probability of occurrence of an event by fitting data to a logistic curve. It is a generalized linear model used for binomial regression, and like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. MLE applied in a binary multivariate logistic analysis is used to model dependent variables to determine the expected probability of success of belonging to a certain group. The estimated coefficients for the logit model are the logarithmic odds ratios and cannot be interpreted directly as probabilities. A quick computation is first required and the approach is simple.

The logit model is specified as $Estimated\ Y = LN[P_i/(1 - P_i)]$ or, conversely, $P_i = EXP(Estimated\ Y)/(1 + EXP(Estimated\ Y))$, and the coefficients β_i are the log odds ratios. So, taking the antilog or $EXP(\beta_i)$, we obtain the odds ratio of $P_i/(1 - P_i)$. This means that with an increase in a unit of β_i, the log odds ratio increases by this amount. Finally, the rate of change in the probability is $dP/dX = \beta_i P_i(1 - P_i)$. The standard error measures how accurate the predicted coefficients are, and the t-statistics are the ratios of each predicted coefficient to its standard error and are used in the typical regression hypothesis test of the significance of each estimated parameter. To estimate the probability of success of belonging to a certain group (e.g., predicting if a smoker will develop chest complications given the amount smoked per year), simply compute the *Estimated Y* value using the MLE coefficients. For example, if the model is $Y = 1.1 + 0.005$ *(Cigarettes)* then someone smoking 100 packs per year has an *Estimated Y* of $1.1 + 0.005(100) = 1.6$. Next, compute the inverse antilog of the odds ratio by doing: $EXP(Estimated\ Y)/[1 + EXP(Estimated\ Y)] = EXP(1.6)/(1 + EXP(1.6)) = 0.8320$. Such a person has an 83.20% chance of developing some chest complications in his or her lifetime.

A probit model (sometimes also known as a normit model) is a popular alternative specification for a binary response model. It employs a probit function estimated using maximum likelihood estimation and the approach is called probit regression. The probit and logistic regression models tend to produce very similar predictions where the parameter estimates in a logistic regression tend to be 1.6 to 1.8 times higher than they are in a corresponding probit model. The choice of using a probit or logit is entirely up to convenience, and the main distinction is that the logistic distribution has a higher kurtosis (fatter tails) to account for extreme values. For example, suppose that house ownership is the decision to be modeled, and this response variable is binary (home purchase or no home purchase) and depends on a series of independent variables X_i such as income, age, and so forth, such that $I_i = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$, where the larger the value of I_i, the higher the probability of home ownership. For each family, a critical I^* threshold exists, where if it is

exceeded, the house is purchased; otherwise, no home is purchased, and the outcome probability (*P*) is assumed to be normally distributed, such that $P_i = CDF(I)$ using a standard normal cumulative distribution function (*CDF*). Therefore, use the estimated coefficients exactly like those of a regression model and using the *Estimated Y* value, apply a standard normal distribution (you can use Excel's *NORMSDIST* function or ***Risk Simulator's*** *Distributional Analysis* tool by selecting Normal distribution and setting the mean to be *0* and standard deviation to be *1*). Finally, to obtain a probit or probability unit measure, set $I_i + 5$ (because whenever the probability $P_i < 0.5$, the estimated I_i is negative, as a result of the normal distribution being symmetrical around a mean of zero).

The tobit model (censored tobit) is an econometric and biometric modeling method used to describe the relationship between a non-negative dependent variable Y_i and one or more independent variables X_i. A tobit model is an econometric model in which the dependent variable is censored; that is, the dependent variable is censored because values below zero are not observed. The tobit model assumes that there is a latent unobservable variable Y^*. This variable is linearly dependent on the X_i variables via a vector of β_i coefficients that determine their interrelationships. In addition, there is a normally distributed error term U_i to capture random influences on this relationship. The observable variable Y_i is defined to be equal to the latent variables whenever the latent variables are above zero and Y_i is assumed to be zero otherwise. That is, $Y_i = Y^*$ *if* $Y^* > 0$ and $Y_i = 0$ *if* $Y^* = 0$. If the relationship parameter β_i is estimated by using ordinary least squares regression of the observed Y_i on X_i, the resulting regression estimators are inconsistent and yield downward-biased slope coefficients and an upward-biased intercept. Only MLE would be consistent for a tobit model. In the tobit model, there is an ancillary statistic called sigma, which is equivalent to the standard error of estimate in a standard ordinary least squares regression, and the estimated coefficients are used the same way as a regression analysis.

Procedure

- Start Excel and open the example file *Advanced Forecasting Model*, go to the *MLE* worksheet, select the dataset including the headers, and click on *Risk Simulator | Forecasting | Maximum Likelihood*.
- Select the dependent variable from the drop-down list (see Figure 1) and click *OK* to run the model and report.

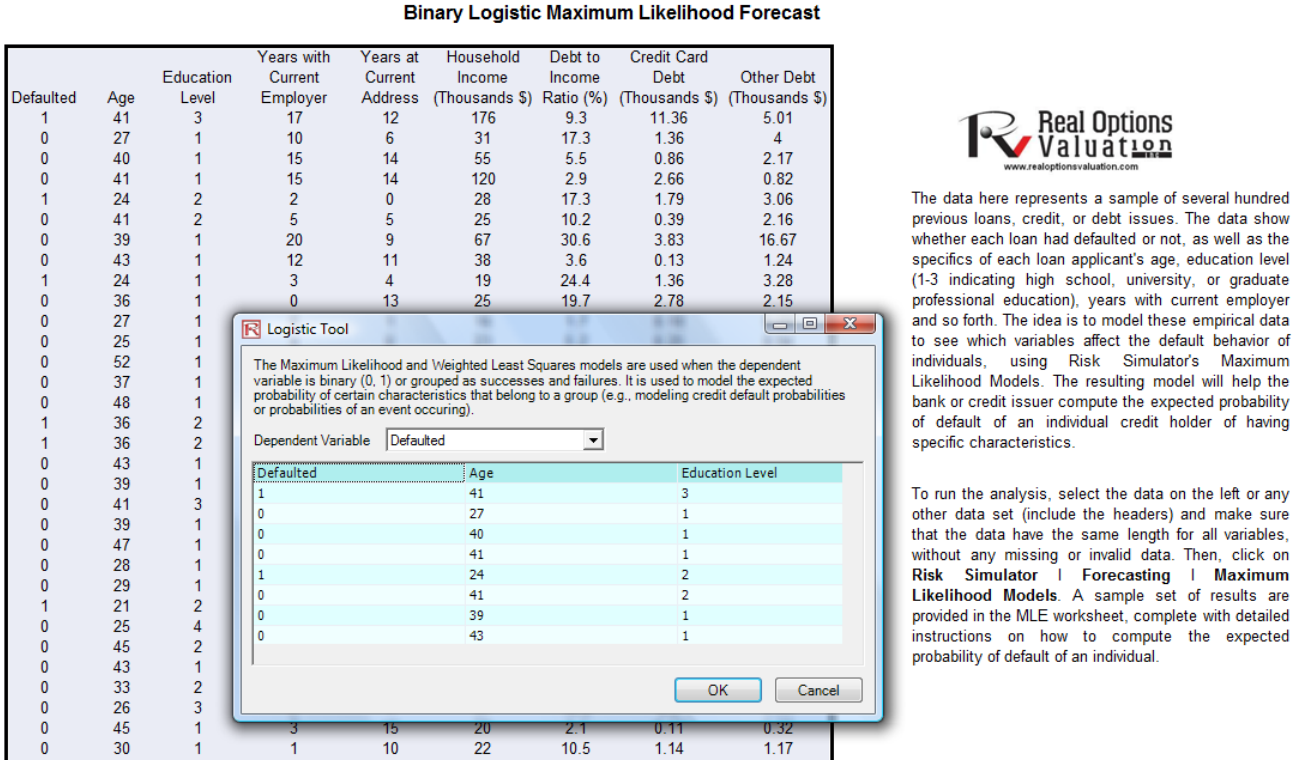

Figure 1. Logit, Probit, Tobit