Kellogg School of Management
Northwestern University

Predictive Analytics: Logistic Regression

Overview

Logistic regression is similar to ordinary multiple regression – except that logistic regression is used when the dependent variable is binary and assumes only two discrete values. Examples include 'yes-no' dependent variables such as whether a customer responded to a marketing campaign or not, whether a person is a homeowner or not, whether a business goes bankrupt or not, or whether a person votes guilty or not guilty. Like ordinary multiple regression, the predictor variables can be metric variables (e.g., age, income, or sales units) or categorical (e.g., gender, religion, or region). Indicator or 'dummy' variables are used to include categorical variables as predictors.

While the basic concepts are similar for multiple linear regression and logistic regression, the interpretation of the regression equation and the coefficients are somewhat different. In multiple regression, the dependent variable is a continuous or metric variable – sales or profits, for example – and can assume many values. The multiple regression coefficients are multiplied by the values of the predictor variables to yield the predicted value for the dependent variable.

In logistic regression, the observed values for the dependent variable take on only two values and are usually represented using a 0-1 dummy variable. The mean of a 0-1 dummy variable is equal to the proportion of observations with a value of 1 – and can be interpreted as a probability. The predicted values in a logistic regression will always range between 0 and 1 and are also interpreted as probabilities.

Suppose we are modeling home ownership (where 1 indicates a homeowner and 0 a non-owner) as a function of income. Each individual in the dataset is either a homeowner or not so

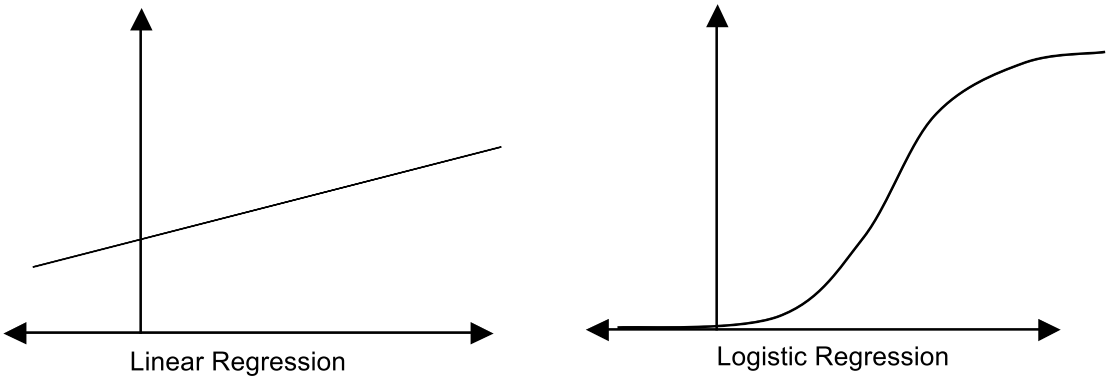
Professor Charlotte Mason from the University of North Carolina prepared the first version this note to provide material for class discussion rather than to illustrate either effective or ineffective handling of a business situation. This note was extended by Professor Florian Zettelmeyer.

the observed values for the dependent variable will be 0 or 1. The predicted value based on the model is interpreted as the probability that the individual is a homeowner. For example, for a person with an income of $35,000, the predicted probability may be .22 compared with a predicted probability of .95 for a person with a $250,000 income.

Like linear regression, once a logistic regression model has been estimated it can be used to make predictions for new observations. Assume a bank has data from a past marketing campaign promoting a ‘gold’ credit card – including whether the customer signed up for the offer or not (the dependent variable) as well as information on other bank services the customer used plus financial and demographic customer information (the predictor variables). These data can be used to estimate a logistic regression model. Then the bank could use this model to identify which additional customers to target with this or a similar offer. By inputting values for the predictor variables for each new customer – the logistic model will yield a predicted probability. Customers with high predicted probabilities may be chosen to receive the offer since they seem more likely to respond positively.

A difference between linear and logistic regression is the shape of the model as shown in Exhibit 1. The simple linear regression is represented by a straight line. For the linear regression, an increase of one unit in the predictor variable has a constant effect – equal to the slope of the line. In logistic regression, the relationship between the dependent variable and the predictor variables is assumed to be nonlinear. A logistic regression model with a single predictor is represented by an s-shaped curve. Moreover, the curve never falls below 0 or exceeds 1 – regardless of the values of the predictor variables. Thus, the predicted values can always be interpreted as probabilities.

Exhibit 1 Simple Linear Regression versus Logistic Regression



In logistic regression, the effect on the predicted probability of a one-unit increase in the predictor variable varies. At the extremes, a one-unit change has very little effect, but has a larger effect in the middle. In many situations, this is intuitive. For example, consider the effect that a $20,000 increase in income might have on the probability of home ownership. The difference in the likelihood that an individual owns a home may not change much as their income increases from $10,000 to $30,000 or from $1,000,000 to $1,020,000 – but may increase quite a bit if income increases from $50,000 to $70,000. Unfortunately, this nonlinearity complicates the interpretation of the regression coefficients. In a linear regression, the interpretation of the coefficient for X is straightforward: an increase of 1 unit in X results in a change in the expected value of Y equal to B (the coefficient for X). However, in a logistic regression, we cannot say that a 1 unit increase in X will result in an expected change in Y equal to B. Rather, it depends on where on the curve the value of X is located.

The Simple Logistic Regression Curve

Consider the simple case with a single predictor variable. For now, we assume that the predictor variable is a continuous variable. In a simple linear regression, the model would be:

$$Y = B_0 + B_1X$$

where B_0 is the intercept or constant term (equal to the predicted value of Y when X=0), and B_1 is the slope of the regression line.

In a simple logistic regression, the model is:

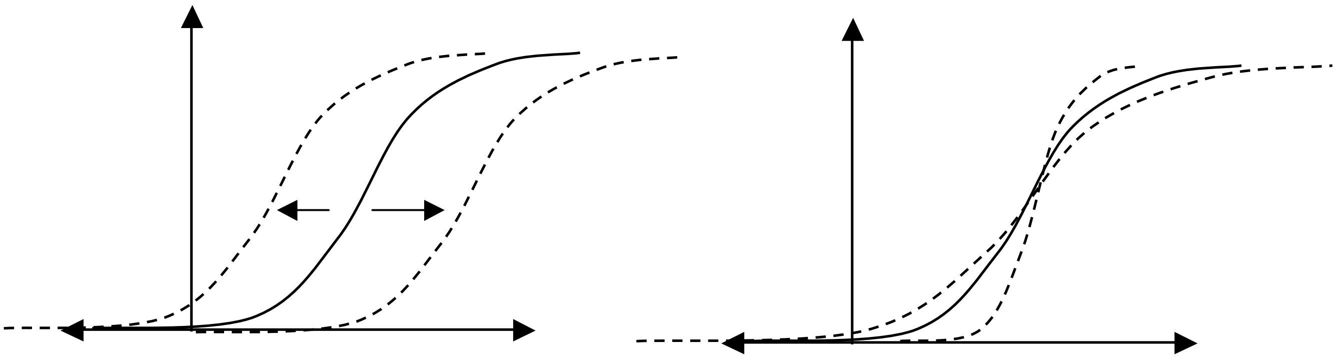
$$\text{Prob}(Y = 1) = \frac{e^{B_0 + B_1X}}{1 + e^{B_0 + B_1X}}$$

which can be also be written as:

$$\text{Prob}(Y = 1) = \frac{1}{1 + e^{-(B_0 + B_1X)}}\,.$$

Thus, as in linear regression, there are two coefficients, B_0 and B_1, in a simple logistic regression. These coefficients determine the specific shape of the curve. The B_0 coefficient (also referred to as the constant) determines the location of the logistic curve along the X axis. As the constant increases, the logistic curve shifts left on the X axis. The B_1 coefficient determines the steepness and direction of the curve. A positive B_1 means the curve will increase as X increases. If B_1 is negative, the curve decreases as X increases. Larger values for B_1 indicate a steeper curve.

Exhibit 2 Logistic Regression Coefficients Control the Shape of the Curve



B_0 coefficient 'shifts' the curve

B_1 coefficient controls the steepness

Interpreting Logistic Regression Coefficients

In simple linear regression, interpretation of the coefficients is straightforward. The constant term estimates the value of Y when X=0. The B_1 coefficient estimates the change in Y for a one unit increase in X. Because of the nonlinear nature of the logistic regression model, interpretation of the coefficients is more complex. To interpret logistic regression, we start with a discussion of probabilities, odds and odds ratios.

A *probability* is the likelihood of an event and is bounded between 0 and 1. If the weather forecast says the probability of rain is 0.25, then there is a 25% chance of rain. *Odds* are the ratios of two probabilities: the probability that the event will occur divided by the probability that the event will not occur. If the probability of rain is 0.25, then the odds are:

$$\text{Odds} = \frac{\text{Prob (event)}}{\text{Prob (no event)}} = \frac{0.25}{0.75} = \frac{1}{3} = .333$$

Since odds are the ratio of two probabilities, odds are always positive, but may be greater than one. In fact, odds can range from 0 to infinity. When the odds are less than 1, the probability of the event (say, rain) is lower than the probability of no event (no rain). Conversely, odds greater than 1 indicate the probability of the event is greater than the probability of no event. Odds of 1 indicate equal (that is, .50) probabilities of event and no event – meaning that both outcomes are equally likely.

Finally, an *odds ratio* is the ratio of two odds. In logistic regression, the odds ratio for the predictor variable X indicates the expected change in the odds that Prob(Y=1) for a one unit increase in X. The odds ratio is particularly important in logistic regression because, unlike linear regression, the 'slope' of the curve is not constant. However, the odds ratio for a predictor variable is constant. The odds ratio for the predictor variable is computed by raising e to the power B_1, or e^{B_1}. For example, consider a logistic regression to predict the probability of purchase of a newly released movie DVD (the dependent variable) using the value (in dollars) of an 'instant coupon'. If the B_1 coefficient is 0.7, we know – since the coefficient is positive - that increasing the value of the instant coupon increases the probability of purchase. However, because of the s-shaped curve, the magnitude of the increase will depend on whether the increase is, say, from \$1 to \$2 or from \$4 to \$5. Since $e^{0.7} = 2.01$, we can say that for every dollar increase in the instant coupon value, the odds of purchase increase by a factor of 2.01. That is, the odds of purchase are twice a large for a \$5 coupon compared with a \$4 coupon.

Summary

Logistic regression is used when the dependent variable is binary. Like linear regression, the predictor variable can be metric or categorical. In a logistic regression, the predicted values are bounded between 0 and 1 and are interpreted as the probability that the dependent variable equals one. Like linear regression, the coefficients and statistical tests will indicate whether the predictor variables are statistically significant – and whether they have a positive or negative effect on the probability that the dependent variable is one. However, unlike linear regression, the effect of a one-unit change in X on Y is not linear – rather it depends on the value of X. The odds ratios is used to interpret the effects of individual predictor variables.