Using Logistic Regression Models

Matthew Curcio

Introduction

For individuals who have studied cell biology or biochemistry, logistic regression may be familiar as dose-response curves[1], enzyme kinetic curves[2], median lethal dose curve (LD-50)[3], a four parameters logistic regression, or even an exponential growth curve given limited resources[4].

The exponential growth curve given limited resources:

$$f(x) \ = \ \frac{M}{1 + Ae^{-r(x-x_0)}} \tag{1}$$

where:

- M is the carrying capacity, the curve's maximum value
- r is the Malthusian parameter[5], or the log growth rate
- x_0 is the midpoint of the curve
- A is the number of times that the initial population needs to double to reach the maximum value[6]

Outside of Biology Logistic Regression (Logit) is commonly used to assess credit or financial risk, profile customers for churn, and even political polling.

Logistic Regression is a supervised machine learning technique. It is commonly used to determine the relationship between predictor variables and a typically binary outcome. The input or predictor variables can be either categorical or continuous.

I will demonstrate on a data set from the well-known Framingham Longitudinal study that uses both categorical and continuous input variables and an outcome that gives a risk assessment on whether one may acquire cardiovascular disease.

Logit has several important features. Logit is not only helping us determine the relationship between variables. It helps one understand the importance of each variable to the whole.

For example, Logit does not simply tell the user if the loan should or should not be given. It will provide the probabilities associated with each variable or action. Additionally, if a borrower receives a value of 51%, it may indicate that the loan is given out. But by carefully inspecting the probabilities, we can turn those probabilities into odds. Finding the odds of an occurrence can provide an intuitive understanding of any predictor variables.

An advantage of using Logit is you avoid confounding effects by analyzing the association of all variables together. The procedure is similar to multiple linear Regression,

[1] https://en.wikipedia.org/wiki/Dose%E2%80%93response_relationship

[2] https://www.khanacademy.org/science/ap-biology/cellular-energetics/environmental-impacts-on-enzyme-function/a/basics-of-enzyme-kinetics-graphs

[3] https://en.wikipedia.org/wiki/Median_lethal_dose

[4] https://openoregon.pressbooks.pub/mhccmajorsbio/chapter/environmental-limits-to-population-growth/

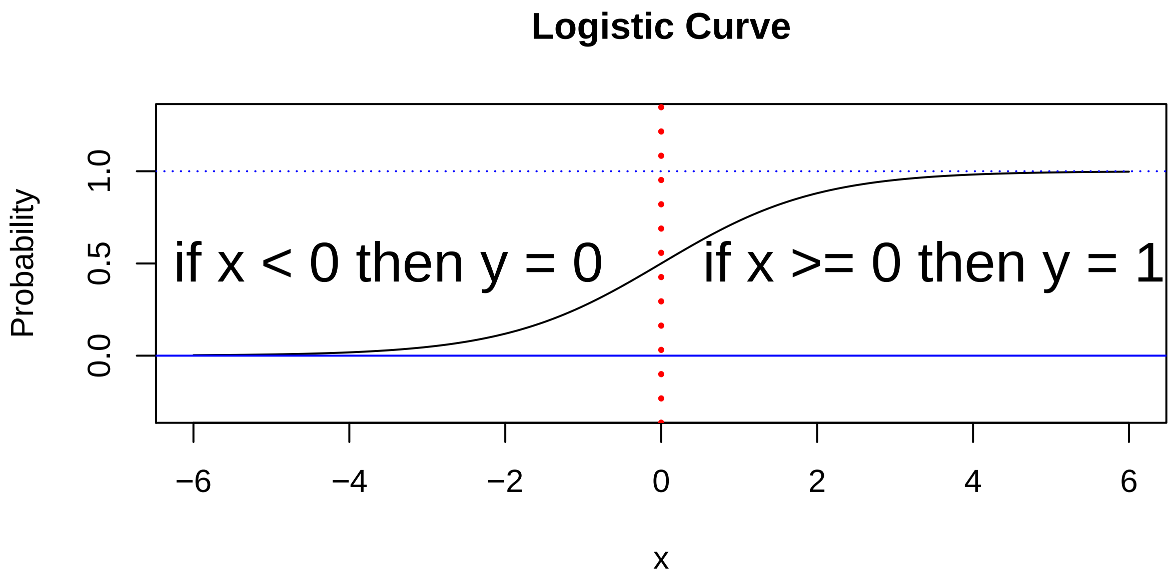
[5] https://en.wikipedia.org/wiki/Malthusian_growth_model

[6] https://en.wikipedia.org/wiki/Logistic_function

The result is the impact of each variable on the odds ratio of the observed event of interest.

Understanding Logit should be studied to understand neural networks better.

Figure #1 shows an idealized logistic curve.



In Figure #1, the function's independent variable(s), may be integers or continuous values. Interestingly, the dependent variable does not provide a 0 or 1 but gives a value between 0 and 1, the probability of that event will occur. In the figure, the *decision boundary* is $x\ = 0$, denoted by the *red dotted line.* At the mid-point the curves range changes from *zero* (the absence) to *one* (the presence) of a quality or item.

The Logit formula is simplified version of the exponential growth curve. It should be easy to see that the output is (0, 1) inclusive.

The Logistic Regression formula:

$$f(x) \ = \ \frac{1}{1 + e^{-(WX+b)}} \tag{2}$$

where:

- W is the matrix of the features
- b is a constant

Since the logistic equation is exponential, it is easier to work with the formula in terms of its odds, initially *log-odds.*

Odds are (generally) the probability of success over failure denoted as $\frac{p}{1-p}$ and more importantly, in this situation, log-odds are $ln\left(\frac{p}{1-p}\right)$.

$$Odds \ = \ \frac{Probability\ of\ Success}{Probability\ of\ Failure} \tag{3}$$

Simply by using log-odds, logistic regression may be more easily expressed as a set of linear equations in x.[7] Hence we can now go from linear regression to logistic regression.

Step #2: and change notation to summation on the right hand side;

$$ln\left(\frac{p}{1-p}\right) = \sum_{i}^{k} \beta_i x_i \tag{4}$$

[7] http://juangabrielgomila.com/en/logistic-regression-derivation/