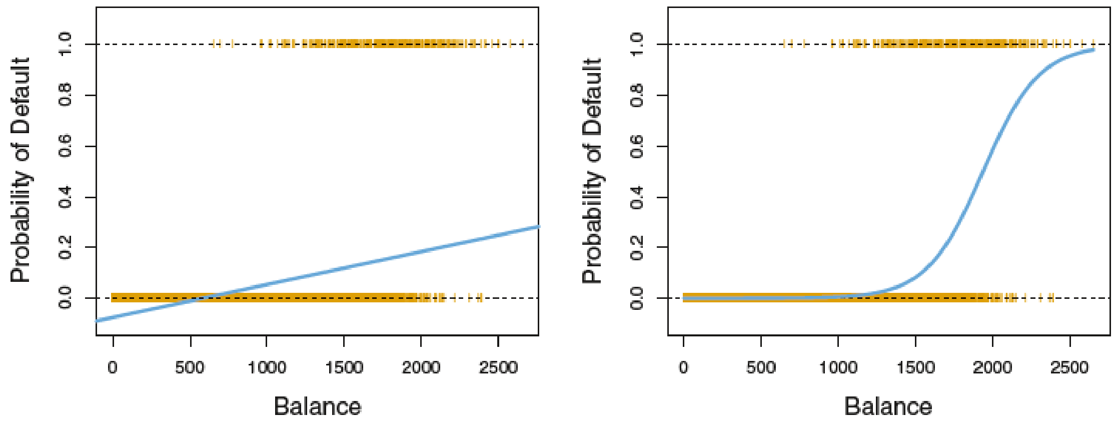
Logistic Regression

Logistic regression is technique borrowed by machine learning from the field of statistics. It's used for binary classification problems .Logistic regression models the probability that y belongs to a particular category rather than modelling the response itself.

$$p(X) = \beta0 + \beta1X.$$



If we use this approach to predict default=Yes using balance, then we obtain the model shown in the left-hand panel of Figure. Here we see the problem with this approach: for balances close to zero we predict a negative probability of default; if we were to predict for very large balances, we would get values bigger than 1. These predictions are not sensible, since of course the true probability of default, regardless of credit card balance, must fall between 0 and 1.

This problem is not unique to the credit default data. Any time a straight line is fit to a binary response that is coded as 0 or 1, in principle we can always predict p(X) < 0 for some values of X and p(X) > 1 for others (unless the range of X is limited).To fit the model , we use a method called maximum likelihood, which we discuss in the next section. The right-hand panel the fit of the logistic regression model to the Default data.

Notice that for low balances we now predict the probability of default as close to, but never below, zero. Likewise, for high balances we predict a default probability close to, but never above, one. The logistic function will always produce an S-shaped curve of this form, and so regardless of the value of X, we will obtain a sensible prediction.

To avoid this problem, we must model p(X) using a function that gives outputs between 0 and 1 for all values of X. Many functions meet this description. In logistic regression, we use the logistic function,

$$p(X) = \frac{e^{\beta0+\beta1X}}{1+e^{\beta0+\beta1X}}$$

Logistic regression uses the logistic function to ensure a prediction between 0 and 1.

$$\frac{p(X)}{1-p(X)} = e^{\beta0+\beta1X}$$

The left hand side of the above equation is called odds, can take on any value between 0 and ∞. Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta0 + \beta1X$$

The left-hand side is called the log-odds or logit. We see that the logistic regression model has a logit that is linear in X.Increasing X by one unit changes the log oddsby β1 , or equivalently it multiplies the odds by $e^{\beta1}$.

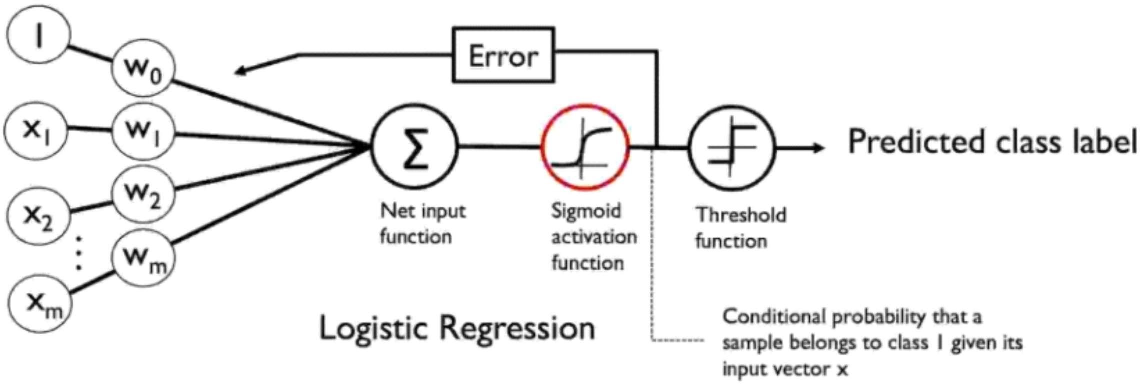
Logistic Regression Continued

Logistic Function

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Where e is the base of the natural logarithms (Euler's number) and 'x' is the actual numerical value that you want to transform.



Logistic regression is used in weather forecasting, for example, not only to predict if it will rain on a particular day but also to report the chance of rain. Similarly, logistic regression can be used to predict the chance that a patient has a particular disease given certain symptoms, which is why logistic regression enjoys great popularity in the field of medicine.

Logistic Regression- Parameter Estimation

Parameter Estimation

The goal of logistic regression is to estimate the unknown parameters β0, β1.
This is done with maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is greatest. The maximum likelihood equation is derived from the probability distribution of the dependent variable. Since each yi represents a binomial count in the ith population, the joint probability density function of Y is

$$f(\boldsymbol{y}|\beta) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

The joint probability density function in equation expresses the values of y as a function of known, fixed values for β. (Note that is related to π by Equation). The likelihood function has the same form as the probability density function, except that the parameters of the function are reversed.
The likelihood function expresses the values of in terms of known, fixed values for y.
Let π_i be a column vector also of length N with elements π_i = P(Zi = 1|i), i.e., the probability of success for any given observation in the ith population.

$$L(\beta|\boldsymbol{y}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

The critical points of a function (maxima and minima) occur when the first derivative equals 0. If the second derivative evaluated at that point is less than zero, then the critical point is a maximum. Thus, finding the maximum likelihood estimates requires computing the first and second derivatives of the likelihood function.

First, note that the factorial terms do not contain any of the π_i. As aresult, they are essentially constants that can be ignored: maximizing the equation without the factorial terms will come to the same result as if they were included.
Second, note that since $a^{x-y} = a^x/a^y$, and after rearranging terms, the equation to be maximized can be written as:

$$\prod_{i=1}^{N} \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} (1 - \pi_i)^{n_i}$$

Taking e on both sides

$$\left(\frac{\pi_i}{1 - \pi_i}\right) = e^{\sum_{k=0}^{K} x_{ik}\beta_k}$$

Solving for π_i

$$\pi_i = \left(\frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}}\right)$$

Substituting π_i for the first term in top equation

$$\prod_{i=1}^{N} (e^{\sum_{k=0}^{K} x_{ik}\beta_k})^{y_i} \left(1 - \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}}\right)^{n_i}$$

Use $(a^x)^y = a^{xy}$ to simplify the first product, rewrite the equation

$$\prod_{i=1}^{N} (e^{y_i \sum_{k=0}^{K} x_{ik}\beta_k})(1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k})^{-n_i}$$

Logistic Regression- Parameter Estimation…

$$\prod_{i=1}^{N} (e^{y_i \sum_{k=0}^{K} x_{ik}\beta_k})(1+e^{\sum_{k=0}^{K} x_{ik}\beta_k})^{-n_i}$$

This is the kernel of the likelihood function to maximize. However, it is still cumbersome to differentiate and can be simplified a great deal further by taking its log.
Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log likelihood function and vice versa.

$$l(\beta) = \sum_{i=1}^{N} y_i \left(\sum_{k=0}^{K} x_{ik}\beta_k \right) - n_i \cdot \log(1+e^{\sum_{k=0}^{K} x_{ik}\beta_k})$$

To find the critical points of the log likelihood function, set the first derivative with respect to each equal to zero.

$$\frac{\partial}{\partial \beta_k} \sum_{k=0}^{K} x_{ik}\beta_k = x_{ik}$$

since the other terms in the summation do not depend on β_k and can thus be treated as constants.
In differentiating the second half , take note of the below partial derivative

$$\frac{\partial}{\partial x} \log y = \frac{1}{y}\frac{\partial y}{\partial x}.$$

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta_k} &= \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot \frac{\partial}{\partial \beta_k}\left(1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}\right) \\
&= \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot e^{\sum_{k=0}^{K} x_{ik}\beta_k} \cdot \frac{\partial}{\partial \beta_k} \sum_{k=0}^{K} x_{ik}\beta_k \\
&= \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot e^{\sum_{k=0}^{K} x_{ik}\beta_k} \cdot x_{ik}
\end{aligned}$$

Substituting π_i in the above equation

$$\pi_i = \left(\frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \right)$$

$$= \sum_{i=1}^{N} y_i x_{ik} - n_i \pi_i x_{ik}$$

The maximum likelihood estimates for can be found by setting each of the K + 1 equations equal to zero and solving for βk.

Each such solution, if any exists, specifies a critical point- either a maximum or a minimum. The critical point will be a maximum if the matrix of second partial derivatives is negative definite; that is, if every element on the diagonal of the matrix is less than zero.
Another useful property of this matrix is that it forms the variance-covariance matrix of the parameter estimates. It is formed by differentiating each of the K + 1 equations a second time with respect to each element of βk.

Logistic Regression continued…

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_{k'}} &= \frac{\partial}{\partial\beta_{k'}}\sum_{i=1}^{N} y_i x_{ik} - n_i x_{ik}\pi_i \\ &= \frac{\partial}{\partial\beta_{k'}}\sum_{i=1}^{N} -n_i x_{ik}\pi_i \\ &= -\sum_{i=1}^{N} n_i x_{ik}\frac{\partial}{\partial\beta_{k'}}\left(\frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}}\right)\end{aligned}$$

To solve the above equation we will make use of two general rules for differentiation. First, a rule for differentiating exponential functions

$$\frac{\mathrm{d}}{\mathrm{d}x}e^{u(x)} = e^{u(x)}\cdot\frac{\mathrm{d}}{\mathrm{d}x}u(x)$$

In our case, let u(x) $\sum_{k=0}^{K} xik$ βk. Second, the quotient rule for differentiating the quotient of two functions:

$$\left(\frac{f}{g}\right)'(a) = \frac{g(a)\cdot f'(a) - f(a)\cdot g'(a)}{[g(a)]^2}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\frac{e^{u(x)}}{1+e^{u(x)}} &= \frac{(1+e^{u(x)})\cdot e^{u(x)}\frac{\mathrm{d}}{\mathrm{d}x}u(x) - e^{u(x)}\cdot e^{u(x)}\frac{\mathrm{d}}{\mathrm{d}x}u(x)}{(1+e^{u(x)})^2} \\ &= \frac{e^{u(x)}\frac{\mathrm{d}}{\mathrm{d}x}u(x)}{(1+e^{u(x)})^2} \\ &= \frac{e^{u(x)}}{1+e^{u(x)}}\cdot\frac{1}{1+e^{u(x)}}\cdot\frac{\mathrm{d}}{\mathrm{d}x}u(x)\end{aligned}$$

$$\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_{k'}} = -\sum_{i=1}^{N} n_i x_{ik}\frac{\partial}{\partial\beta_{k'}}\left(\frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}}\right)$$

$$\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_{k'}} = -\sum_{i=1}^{N} n_i x_{ik}\pi_i(1-\pi_i)x_{ik'}$$