

# Logistic Regression

Logistic regression is a classification technique. It is an example of a *Generalised Linear Model*, which form one of the corner stones of modern (computationally intensive) statistics. We suppose that our data comes in records of the form

$$(\mathbf{x}, y) = (x_1, x_2, \dots, x_k, y)$$

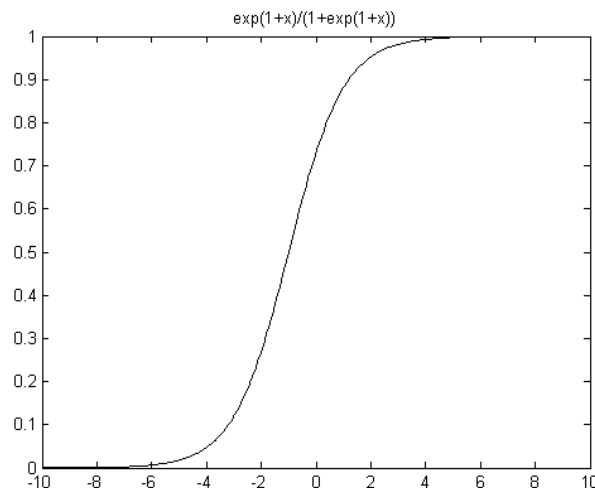
where  $k$  is fixed, the  $x_i$ 's are all numerical (ordinal or continuous) and  $y$  is binary. We suppose that the value of the response variable  $y$  is a function of the inputs  $\mathbf{x}$ , subject to some randomness. Logistic regression imposes a statistical model for the relationship between  $\mathbf{x}$  and  $y$ , viz

$$P(y = 1) = \frac{\exp(b_0 + b_1 x_1 + \dots + b_k x_k)}{1 + \exp(b_0 + b_1 x_1 + \dots + b_k x_k)}$$

or equivalently

$$\text{logit}(P(y = 1)) := \log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right) = b_0 + b_1 x_1 + \dots + b_k x_k .$$

If  $k = 1$  then we can plot  $P(y = 1)$  as a function of  $x = x_1$ , obtaining the famous *logistic curve*



Once we have  $\mathbf{b} = (b_0, b_1, \dots, b_k)$  we can classify a new observation of  $\mathbf{x}$  using the model. For  $0 < p < 1$  we put  $\hat{y} = 1$  if  $P(y = 1) > p$ , that is

$$b_0 + b_1 x_1 + \dots + b_k x_k > \log\left(\frac{p}{1-p}\right)$$

A reasonable choice for  $p$  is 0.5, but other choices may be better depending on the application. A larger  $p$  gives a classifier that is more likely to be correct when it gives  $\hat{y} = 1$ , but less likely to be correct when it gives  $\hat{y} = 0$ . The choice of  $p$  can be informed by an ROC curve (see below).

## Categorical data

If we have a variable  $x$  which is categorical rather than numerical, we can still apply logistic regression by converting it into a number of binary variables. Suppose  $x$  takes on values  $\{a,$

$b, c\}$ . We replace  $x$  by two variables  $x_1 = \{x = a\}$  and  $x_2 = \{x = b\}$ . Note that we don't need  $x_3 = \{x = c\}$  since  $x_1 = 0$  and  $x_2 = 0$  implies  $x_3 = 1$ . In general if  $x$  takes on  $m$  possible values we will need  $m - 1$  binary variables to replace it.

### Fitting the Model

The model is fitted using *maximum likelihood*. Suppose we have  $n$  records, given by  $(\mathbf{x}(i), y(i))$  for  $i = 1, \dots, n$ . The likelihood of the sample is the probability of observing it, for a given value of the parameters  $\mathbf{b}$ . That is

$$L(b_0, b_1, \dots, b_k) = \prod_{i:y(i)=1} \frac{\exp(b_0 + b_1 x_1(i) + \dots + b_k x_k(i))}{1 + \exp(b_0 + b_1 x_1(i) + \dots + b_k x_k(i))} \\ \times \prod_{i:y(i)=0} \frac{1}{1 + \exp(b_0 + b_1 x_1(i) + \dots + b_k x_k(i))}$$

Note that this assumes that we have *independent* observations. The maximum likelihood fit of  $\mathbf{b}$  is that value which maximises the likelihood. In practice it is usually easier, and computationally more tractable, to maximise the log likelihood

$$\log L(b_0, b_1, \dots, b_k) \\ = \sum_{i:y(i)=1} (b_0 + b_1 x_1(i) + \dots + b_k x_k(i) - \log(1 + \exp(b_0 + b_1 x_1(i) + \dots + b_k x_k(i)))) \\ - \sum_{i:y(i)=0} \log(1 + \exp(b_0 + b_1 x_1(i) + \dots + b_k x_k(i)))$$

There is no analytic solution to the problem of maximising  $\log L$ , so we use numerical methods. In this case we have an unconstrained non-linear optimisation problem, so most gradient-based local-search techniques are appropriate. Note that local-search methods are not guaranteed to find the global optimum, just a local optimum. The method most commonly used is an iterative weighted least-squares approach developed by Nelder & Wederburn (See for example McCullagh & Nelder 1989).

Put  $x = (x_0, x_1, \dots, x_n)$  with  $x_0 = 1$  and  $b = (b_0, b_1, \dots, b_k)$ , then the logistic regression model can be written  $P(y = 1) = \frac{1}{1 + \exp(-\langle x, b \rangle)} =: f(\langle x, b \rangle)$ . The log-likelihood is then

$$l(b) = \sum_{i=1}^n (y(i) \log f(\langle x(i), b \rangle) + (1 - y(i)) \log(1 - f(\langle x(i), b \rangle))).$$

It is easy to check that  $f'(u) = f(u)(1 - f(u))$ , whence

$$\frac{\partial l(b)}{\partial b_j} = \sum_{i=1}^n (y(i) x_j(i) (1 - f(\langle x(i), b \rangle)) - (1 - y(i)) x_j(i) f(\langle x(i), b \rangle)) \\ = \sum_{i=1}^n x_j(i) (y(i) - f(\langle x(i), b \rangle))$$

Putting  $\frac{\partial l(b)}{\partial b_j} = 0$  for each  $j$  gives a system of non-linear equations for  $b$ . Unfortunately

we don't have an analytical solution to this system, so we look for a *numerical* scheme to maximise  $l$ .

### Newton's method

Newton's method is an iterative technique. To find a local stationary point of  $g(x)$  for  $x \in \mathfrak{R}^{k+1}$ , we start with some  $x(0)$  then put

$$x(t+1) = x(t) - H^{-1} \nabla g(x(t))$$

where

$$\nabla g(x) = \begin{pmatrix} \frac{\partial g(x)}{\partial x_0} \\ \vdots \\ \frac{\partial g(x)}{\partial x_k} \end{pmatrix}$$

$$H = \left( \frac{\partial^2 g(x)}{\partial x_i \partial x_j} \right)_{i,j=0}^n$$

To derive this rule we start with the Taylor series approximation

$$g(x) \approx g(x(t)) + (x - x(t))' \nabla g(x(t)) + \frac{1}{2} (x - x(t))' H (x - x(t))$$

then take the gradient of both sides to get

$$\nabla g(x) \approx \nabla g(x(t)) + H(x - x(t)).$$

Put  $\nabla g(x) = 0$  to get Newton's method, as above.

To apply Newton's method to the log-likelihood we need  $\frac{\partial l(b)}{\partial b_j}$  as above and  $H$  given by

$$H_{h,j} = \frac{\partial^2 l(b)}{\partial b_h \partial b_j} = - \sum_{i=1}^n x_h(i) f(< x(i), b >) (1 - f(< x(i), b >)) x_j(i)$$

Put

$$X' = (x(1) | \dots | x(n))$$

$$Y' = (y(1), \dots, y(n))$$

$$P'(b) = (f(< x(1), b >), \dots, f(< x(n), b >))$$

$$W(b) = \text{diag}(f(< x(i), b >)(1 - f(< x(i), b >)))_{i=1}^n$$

then  $\nabla l(b) = X'(Y - P(b))$  and  $H = -X'W(b)X$ , and Newton's method for the log-likelihood becomes

$$b(t+1) = b(t) + (X'W(b(t))X)^{-1} X'(Y - P(b(t)))$$

$$= (X'W(b(t))X)^{-1} X'W(b(t))(Xb(t) + W^{-1}(Y - P(b(t))))$$

$$= (X'W(b(t))X)^{-1} X'W(b(t))Z(b(t))$$

where  $Z(b(t)) = Xb(t) + W^{-1}(Y - P(b(t)))$ .

$X$  is  $n \times (k+1)$  and  $W$  is  $n \times n$ , so it takes  $O(nk^2)$  operations to calculate  $H$  and  $O(k^3)$  operations to solve for  $b(t+1)$  given  $b(t)$ . Thus assuming  $k < n$  we get  $O(nk^2)$  operations for each iteration of the algorithm, which is computationally feasible.

We remark that the weighted least squares solution of  $Ax = \beta$ , with diagonal weight matrix  $W$ , is that  $x$  which minimises  $(Ax - \beta)'W(Ax - \beta)$ . Setting the gradient to 0 we get the solution  $x = (A'WA)^{-1} A'W\beta$ . Thus  $b(t+1)$  is the weighted least-squares solution of  $Xb = Z(b(t))$ , with weights  $W(b(t))$ .

Finally we note that a logistic regression model can sometimes be improved by including *higher order* and *interaction* terms. That is we augment the set of independent variables by including derived variables of the form  $x_i^2$ ,  $x_i^3$ ,  $x_i x_j$  etc.

## Model Selection

Often the number of parameters  $k$  is large. Parsimonious models are easier to interpret, more reliable predictors (less prone to over fitting) and more robust (less influenced by extreme values). In this context, a more parsimonious model will not use as many independent variables. If  $b_i = 0$  then variable  $x_i$  has no influence; we say that it has been excluded from the model. As with all classification techniques, we can compare two or more alternatives by looking at their performance on a validation data set or by using cross-validation with the training data set. However when determining which variables to include in a logistic regression we can use a *likelihood ratio test* or the *AIC* instead, which are much quicker.

### Likelihood Ratio

Suppose  $A$  and  $B$  are nested models, that is  $A$  is a sub-model of  $B$ . If  $L_A$  is the maximum likelihood of model  $A$  (evaluated using the estimated parameter values) and  $L_B$  is the maximum likelihood of model  $B$ , then  $L_B > L_A$  and  $\lambda = -2\log(L_A / L_B) \approx \chi_d^2$  where the degree of freedom  $d$  is the difference in the number of parameters between the two models.<sup>1</sup> We test for the significance of  $\lambda$  in the usual way. That is, at the 95% significance level, we reject the hypothesis that  $A$  and  $B$  are equivalent (which is to say we prefer  $B$  to  $A$ ) if  $\lambda$  is larger than  $\chi_d^2(0.95)$ , the 0.95 percentage point of a  $\chi_d^2$  distribution.

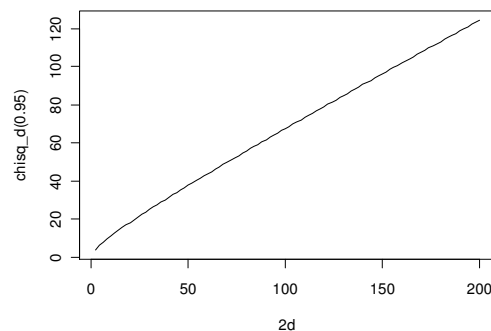
### Akaike's Information Criterion

The AIC is a model selection tool based on the Kullback-Leibler (K-L) information and tends to produce better models than likelihood ratio tests. The goal is to choose, from some given set of models, that model  $A$  which minimizes  $-2\log L_A + 2k$ , where  $k$  is the number of parameters used to specify model  $A$ .

Suppose  $A$  is a sub-model of  $B$ . Using the AIC we prefer  $A$  to  $B$  if

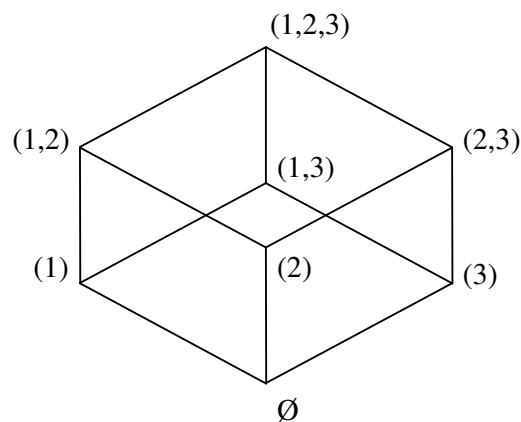
$$-2\log L_A + 2k_A < -2\log L_B + 2k_B \Leftrightarrow -2\log(L_A / L_B) < 2d$$

where  $d$  is the difference in the number of parameters between the two models. Compare this to the likelihood ratio test, which prefers  $A$  to  $B$  if  $-2\log(L_A / L_B) < \chi_d^2(0.95)$ . The figure below plots  $2d$  against  $\chi_d^2(0.95)$  for  $d = 1, \dots, 100$ . As  $2d > \chi_d^2(0.95)$  we see that the AIC will tend to produce smaller models than using the likelihood ratio.



<sup>1</sup> Strictly speaking we are comparing two model classes, given by two nested parameter spaces. The distribution of  $-2\log(L_A / L_B)$  is  $\chi_d^2$  in the limit as the sample size tends to infinity, provided that the parameterisation of the true model is in the interior of both spaces.

If we have  $k$  independent variables then there are  $2^k$  possible choices for the set of variables to include in the model. For  $k$  large it is not feasible to check all possibilities (the curse of dimensionality). In this case we can use an iterative procedure to select a model. The commonly used procedures are forward, backward and stepwise regression. Forward regression starts with no variables in the model, then adds the single variable that most improves the likelihood. If this results in a *significant* improvement (or a reduced AIC) then you continue adding variables, otherwise you stop. Backward regression starts with all variables in the model, then removes the single variable that causes the smallest reduction in likelihood, provided this is *not* significant (or does not increase AIC). Stepwise regression is a mixture of forward and backwards regression, where single variables are added or removed at each stage. These methods do not in general give the same model, or the best model, but should produce a good model.



A representation of the model classes available with three parameters/input variables. Each vertex of the hypercube corresponds to a subspace where only certain variables are included. Stepwise model selection procedures move along edges of the hypercube looking for the best model.

**Parameter estimates**

For any given model, provided the sample size is large enough, the maximum likelihood estimates for the parameters  $\mathbf{b}$  are approximately normally distributed, with covariance matrix given by the inverse of the Fisher information matrix. That is

$$-\left(\frac{\partial^2 \log L}{\partial b_i \partial b_j}\right)^{-1}$$

This allows us to calculate confidence intervals for the  $b_i$  using normal percentage points. It also allows us to formally test the hypothesis  $b_i = 0$ : let  $\sigma(b_i)$  be the variance of  $b_i$ , then

$$(b_i / \sigma(b_i))^2 \approx \chi_1^2. \text{ This is called the Wald Statistic.}$$

Testing for the significance of individual parameters can guide the model selection procedure. However note that the parameters are not independent, so removing one will effect the significance of the others.

### Classification

Call the event  $y = 1$  a positive outcome. For either the learning or validation data set we form a *confusion matrix*

Observed	Predicted		Total
	+ve Outcomes	-ve Outcomes	
+ve Outcomes	a	b	a+b
-ve Outcomes	c	d	c+d
Total	a+c	b+d	a+b+c+d

$a + d$  records are correctly classified,  $b + c$  are incorrectly classified. The following terminology is commonly used:

$$\text{Sensitivity} = a/(a + b)$$

$$\text{Specificity} = d/(c + d)$$

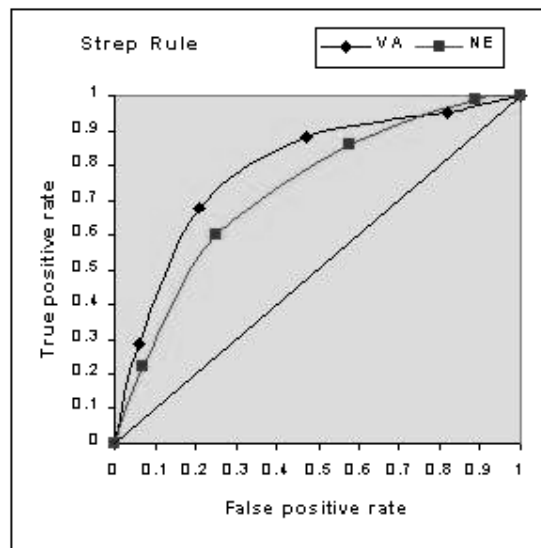
$$\text{False positives} = c/(c + d) = 1 - \text{Specificity}$$

$$\text{False negatives} = b/(a + b) = 1 - \text{Sensitivity}$$

### ROC curves

We classify an observation as positive if  $P(y = 1) > p$ . For every  $0 < p < 1$  we can form a confusion matrix. If we plot the pair (Sensitivity, False Positives) for several  $p$  we get an *ROC curve*. The ROC curve can be used to choose  $p$ , and can also be used to compare models. The further above and to the left of the diagonal the ROC curve lies, the better the model.

ROC curves can be used whenever classification is performed by setting a threshold for some quantitative measure. In medicine they are often used to judge the efficacy of diagnostic tests. ROC curves were developed for measuring the performance of radar operators in WWII; ROC stands for receiver operating characteristic.



Another method for judging the effectiveness of a classifier is a *lift chart*.

## **References**

McCullagh, P. & Nelder, J.A., *Generalized Linear Models: Second Edition*. Chapman & Hall, 1989.

Burnham, K. P. & Anderson, D. R., *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*, 2nd ed. Springer-Verlag, 2002.

Kuhnert, P. and Venables, B., *An Introduction to R: Software for Statistical Modelling & Computing*. CSIRO Mathematical and Information Sciences, 2005.