KMITL Sci. J. Vol. 5 No. 1 Feb. 2005

AN ALGORITHM TO APPROXIMATE THE FIXED-RESPONSE COVARIATES IN LOGISTIC REGRESSION USING SMOOTHING SPLINES

Nomchit Kittichotipanit[1] and Robert W. Jernigan[2]
[1]Department of Applied Statistics, King Mongkut Institute
of Technology Ladkrabang, Bangkok, Thailand.
[2]Department of Mathematics and Statistics, American University,
Washington D.C., USA.

ABSTRACT

We investigate the use of smoothing splines in logistic regression to estimate the covariate values that yield a fixed response probability, eg. LC50 or LD50. We develop an algorithm for a monotonic spline fit and approximate the resulting probability estimates and the fixed-response covariates. We illustrate our algorithm on data sets from studies of genetic spatial diversity.

KEYWORDS: smoothing splines, logistic regression.

1. INTRODUCTION

A problem in logistic regression using smoothing splines is investigated . In particular, we develop an algorithm to approximate the covariate values that yield a fixed response probability. Examples include the LC50, the lethal concentration of a carcinogen that results is a 50% death probability or the LD50, the lethal dose of a drug that results in 50% death probability.

Our motivation comes from a problem in evolutionary genetics in which the geographic prevalence of certain genetic markers is mapped, Jaarola et al. (1997). A first approach to such a mapping considers the prevalence of the marker as the probability of a response that is modeled by a function of a distance from a fixed point. This typically results in a monotonic S-shaped curve *f(x)* that defines the probability of the occurrence of the genetic marker in a population located at a distance *x* from the fixed point. Such a curve is called a cline in the biological literature, Brumfield et al. (2001). Of special interest to evolutionary biologists is the distance between the locations that result in 20% and 80% response probabilities. Such a distance is called the cline width, Brumfield et al. (2001). This cline width informs biologists of the extent of a zone where two species might hybridize, yielding key measures of the action of speciation.

Our statistical investigations will use smoothing splines in a logistic regression setting to estimate the covariate values that result in such fixed response probabilities.

The data are samples obtained from natural populations located at various distances from a fixed point. The probability of response or the prevalence of some genetic trait is related to this distance. Logistic regression is the standard approach used to model such genetic data. We will investigate the setting of logistic regression in Section 2. In Section 3 we consider smoothing splines as an alternative allowing the observed data to indicate nonparametrically the form of a smooth function of x to model prevalence.

2. LOGISTIC REGRESSION MODEL

Binary data, W_i, (1 or 0) indicating the presence or absence of a response, are often in the form of (X_i, W_i), for i=1,2,…,k for some covariate values X_i. The goal is often to find a function of the covariate that predicts the probability of a response:

$$P(W=1 \mid X=x)=\pi(x).$$

A typical form of this response function is the logistic equation:

$$\pi(x)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

where α and β are parameters to be estimated. This is one of many forms that guarantee that $\pi(x)$ always lies between zero and one. This equation can be transformed to be linear in the covariate,

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right]=\alpha+\beta x .$$

The terms $\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$ are called the logits and denoted $\operatorname{logit}(\pi(x))$, Agresti(1996).

Maximum likelihood estimates of α and β can be obtained by iteratively performing a weighted regression of these logits on the covariates. The weights are chosen inversely proportional to the variance of the logits, see Agresti (1990), McCullagh (1984).

In many settings, a problem with using logistic regression is the strong parametric straight-line assumption imposed on the logits. The tails of the logistic function behave in similar ways. For example, if β is positive, the response function's rate of rise from zero is a mirror image of its rate of approach to one as the covariate x increases. But many data sets exhibit different behaviors at each end of the covariate range. An alternative that allows the observed data to indicate, nonparametrically, the form of a smooth function of x to model the response is needed. Smoothing splines allow such an approach; see Hastie and Tibshirani (1990). Smoothing splines have been used in variety of settings in biology see Schluter (1988), Brumfield, et al.(2001), Munson and Jernigan (1989), and Culver, Jernigan, and O'Connell (1994).

3 SMOOTHING SPLINES

Smoothing splints are piecewise polynomials with additional smoothness properties ensuring continuity of first and second derivatives. Cubic smoothing splines have been the most use in the statistical literature, see Green and Silvermean (1994) and Hastie and Tibshirani (1990).

Let $x_1 < x_2 < ... < x_k$ be a set of k ordered distinct covariates. Let n_i be the size of a random sample of Y_i values at covariate X_i. Let the probabilities $\pi(x_1), \pi(x_2), ..., \pi(x_k)$ be a corresponding probabilities of response. The probability of response at covariate x is given by

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

where *f(x)* is a smooth function of the covariate x. To fit this model, the logistic transformation is used to obtain;

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = f(x).$$

A penalized maximum likelihood approach is employed to estimate the smooth function *f(x)*. Following Schluter (1988), the log-likelihood function is represented by

$$l(f) = \sum_{i=1}^{n} l(x_i, f) = \sum_{i=1}^{k} \left\{ y_i \log[\pi(x_i)] + (n_i - y_i) \log[1 - \pi(x_i)] \right\}.$$

A penalized log-likelihood function is used, that is, we seek to minimize the negative penalized log-likelihood function:

$$-\sum_{i=1}^{k} l(x_i, f) + n\lambda \int [f''(x)]^2 \, dx$$

The integral measures how "rough" the chosen function *f* is. A rough function has a rapidly changing slope. The parameter λ is a nonnegative constant called the smoothing parameter. It controls the roughness penalty on the function *f*. The roughness of the smoothing function *f* can also be indexed by equivalent degrees of freedom. These degrees of freedom, *df*, indicate the number of parameters needed to specify a function of the desired smoothness. For example, as $\lambda \to \infty$ and the smooth function f approaches a straight line, the degrees of freedom approach 2 indicating that two parameters, (a slope and an intercept) are needed to specify the smooth function *f*. As $\lambda \to 0$, the degrees of freedom can grow until a non-monotonic interpolation cubic spline polynomial is reached.

In many settings, researchers believe that a smooth monotonic function is the best function describing their data, see Culver, et al (1994). Monotonicity can be used to choose values for the smooth parameter λ or the equivalent degrees of freedom *df*. As the smoothing parameter λ decreases from ∞ to 0 or the degrees of freedom *df* increase from 2 to ∞, our fit goes from being a straight (monotonic) line to being a relatively rough (non-monotonic) function that interpolates the data. The smallest value of the smoothing parameters or the largest value of the equivalent degrees of freedom that specifies this transition from monotonic to non-monotonic defines a lower bound for our selection of the smoothing parameter λ or upper bound for our selection of the degrees of freedom df, respectively. However, this smallest smoothing parameter or the largest degree of freedom that gives monotonicity will not necessary result in the best fit as measured by the deviance, a standard goodness-of-fit criterion given by

$$D = 2\sum\left\{ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right\}$$

(3.2)

where y_i is observed values, and $\hat{y}_i$ is the fitted values, see Gill (2000). Since The value of *D* is positive and the smaller the value of deviance, the better the model fits the data, so we choose our model as the one with smallest deviance under the restriction of monotonicity.

4. FITTING THE MODEL

Suppose we are given *k* distinct design points $x_1, x_2, \ldots, x_k$ satisfying $x_1 < x_2 < \ldots < x_k$ and the relative frequencies $p(x_1), p(x_2), \ldots, p(x_k)$ which correspond to the k distinct design points, respectively, based on sample sizes $n_i, i = 1, 2, \ldots, k$. For all i=1,2,...,k, the mean of $p(x_i)$ is $\pi(x_i)$, the probability of response, and the variance of $p(x_i)$ is $\frac{\pi(x_i)(1-\pi(x_i))}{n_i}$. For all i=1,2,...k, we define the sample logits as, $\ln\left(\frac{p(x_i)}{1-p(x_i)}\right)$ which are estimates of the true logits $\varepsilon(x_i)=\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)$. We have

$$E\left[\log\left(\frac{p(x_i)}{1-p(x_i)}\right)\right]\cong\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)+\frac{1}{2}\left(\frac{2\pi(x_i)-1}{n_i\pi(x_i)(1-\pi(x_i))}\right).$$

Since the sample logits are undefined for $p(x_i)=0$ or 1, a constant value c is used to define the empirical logits, that is,

$$\hat{\varepsilon}(x_i)=\log\left(\frac{p(x_i)+c}{1-p(x_i)+c}\right)=\log\left(\frac{y_i+c}{n_i-y_i+c}\right), i=1,2,...,k\,,$$

for $p(x_i)=\frac{y_i}{n_i}$ and some constant $c>0$. The bias due to the second term is reduced by choosing the constant c=1/2, see Agresti (1990), Anscombe (1956), Cox and Snell (1989) and Gart, Pettigrew and Thomas (1985). Davis (1985) recommended that a linear function of the reciprocal of the number of design points is the best choice for the constant c to reduce bias.

The general model becomes:

$$\hat{\varepsilon}(x_i)=\log\left(\frac{p(x_i)+c}{1-p(x_i)+c}\right)=f(x_i)+e_i, i=1,2,...,k$$

where $f(x_i)$ is a suitably smooth, but unknown, function of the covariate and the term e_i represents random errors satisfying $E(e_i)=0, E(e_i^2)=\sigma_i^2$ and $E(e_i, e_j)=0$ for all i≠j. Based on the asymptotic normality of the empirical logits, for large n_i, the variance of the empirical logits can be approximated by:

$$Var(\hat{\varepsilon}(x_i))\cong\frac{1}{n_i\pi_i(1-\pi_i)},\quad i=1,2,...,k$$

Since these variances are not constant, weighted smoothing splines are used. The weights are the inverse of the estimated variances of the empirical logits, $Var(\hat{\varepsilon}(x_i))$, that is, for all $i = 1, 2, \ldots, k$, the chosen weights, w_i, are

$$w_i = n_i p(x_i)(1 - p(x_i))$$

We seek a smooth function, $\hat{f}$, under the restriction of monotonicity, to estimate the true smooth function, *f*. Algorithmically, as the degrees of freedom are increased from a value of 2 to a larger value, the resulting cubic spline fit goes from being a monotonic straight line to a non-monotonic interpolating function in a smooth and continuous fashion. The largest *df* that results in a monotonic fitting function is the upper bound for our selection of the degrees of freedom, *df*. This largest *df* that gives monotonicity will not necessary result in the best fit as measured by the deviance. We choose our fit as the one with smallest deviance under the restriction of monotonicity.

Iterative algorithm

As with standard logistic regression, the algorithm to minimize the negative penalized log-likelihood works with the empirical logits of the relative frequencies, $\hat{\varepsilon}(x_i) = \log\left(\dfrac{p(x_i) + c}{1 - p(x_i) + c}\right), i = 1, 2, \ldots, k$. Given a vector of these empirical logits, denoted $[\hat{\varepsilon}(x_i)]$, we smooth them with a smoothing matrix, *S*, defined by a weighted smoothing spline. This results in a vector of estimated smoothed empirical logits $[\hat{\varepsilon}^*(x_i)] = S[\hat{\varepsilon}(x_i)], i = 1, 2, \ldots, k$ that are then transformed back to form estimated smooth response probabilities by $p^*(x_i) = \dfrac{e^{\hat{\varepsilon}^*(x_i)}}{1 + e^{\hat{\varepsilon}^*(x_i)}}$, the inverse of the logistic transform.

As with standard logistic regression, this approach requires an iterative algorithm involving the following steps:

1. Sort the data by covariates to achieve a decreasing curve for the response probabilities.

2. Set $df = 2$, that is, we begin with the straight line.

3. Smooth the vector of empirical logits, $[\hat{\varepsilon}(x_i)]$ as a function of the covariate points, $\{x_i\}$, weighted by the inverse of the estimated variances of the empirical logits,

$$[w_i]=[n_i p(x_i)(1-p(x_i))],$$

A smoothing spline fit results in the first iteration of estimated smoothed empirical logits,

$$[\hat{\varepsilon}^*(x_i)]=S[\hat{\varepsilon}(x_i)], i=1,2,...,k.$$

4. Transform the estimated smoothed empirical logits $\hat{\varepsilon}^*(x_i)$ back to the form of estimated smooth response probabilities, $p^*(x_i)$, by

$$p^*(x_i)=\frac{e^{\hat{\varepsilon}^*(x_i)}}{1+e^{\hat{\varepsilon}^*(x_i)}}, i=1,2,...,k.$$

5. Update the initial empirical logits by

$$\hat{\varepsilon}^*(x_i)=\log\left(\frac{p^*(x_i)+c}{1-p^*(x_i)+c}\right)+\frac{p(x_i)-p^*(x_i)}{(p^*(x_i)+c)(1-p^*(x_i)+c)}, i=1,2,...,k$$

which represents the first-order Taylor's series approximation and update the weights of the estimated variances of $\hat{\varepsilon}^*(x_i)$, that is,

$$w^*_i=n_i p^*(x_i)(1-p^*(x_i)), i=1,2,...,k,$$

and repeat from step 3 for another iteration. New estimated smoothed empirical logits, $\hat{\varepsilon}^*(x_i)$, are obtained by performing weighted smoothing splines of $\hat{\varepsilon}^*(x_i)$ onto the covariates x_i with weights w^*_i. The new $p^*(x_i)$, which are different from the ones in the previous iteration are obtained by

$$p^*(x_i)=\frac{e^{\hat{\varepsilon}^*(x_i)}}{1+e^{\hat{\varepsilon}^*(x_i)}}, i=1,2,...,k.$$

6. Continue until convergence. Examine the maximum absolute value of the difference between $p^*(x_i)$ of the present iteration and the previous iteration and compare to the tolerance value (we use 10^{-4}).

7. Calculate the deviance goodness of fit statistic using equation (3.1) where $\hat{y}_i = n_i p^*(x_i)$.

8. Check monotonicity by calculating the maximum value of $p^*(x_{j+1}) - p^*(x_j)$ for all $j = 1, 2, ..., k-1$. If this measure is negative, the weighted smoothing splines is monotone in x_i. If this measure is positive, stop.

9. If monotonicity persists increase the value of df by 1 and repeat steps 3 through 8.

The *df* chosen is the one that results in monotonicity of the estimated smooth response probabilities and smallest deviance. Therefore, $p^*(x_i)$ can be found by

$$p^*(x_i) = \frac{e^{\hat{\varepsilon}^*(x_i)}}{1+e^{\hat{\varepsilon}^*(x_i)}}, i = 1, 2, ..., k.$$

We have applied our technique for fitting model to data sets that examine genetic markers in field vole (*Microtus agrestis*), see Jaarola et al. (1997). Data were collected from 270 voles for a genetic marker named LUYA and 156 voles for another marker named JAAROLA The results of fitted curves shown in Figure 1 indicatethat the monotonic smoothing splines result in clines that fit the data well.

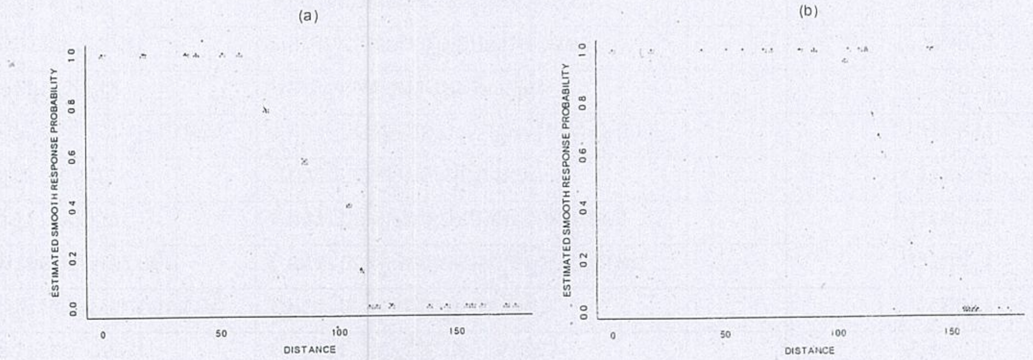

Figure 1 The fitted curves for (a) JAAROLA genetic marker and (b) LUYA genetic marker. The asterisk denotes the observations.

REFERENCES

[1] Agresti, Alan (1990), Categorical Data Analysis, John Wiley and Sons.

[2] Agresti, Alan (1996), An Introduction to Categorical Data Analysis, John Wiley and Sons.

[3] Anscombe, F.J. (1956), " On estimating binomial response relations ", *Biometrika*, Vol.43(3), pp461-464.

[4] Brumfield, Robb T., Jernigan, Robert W., McDonald, David B. and Braun, Michael J. (2001), "Evolutionary implications of divergent clines in an avian (Manacus:Aves) hybrid zone", *Evolution*, Vol. 55(10), pp. 2070-2087.

[5] Cox, D.R. and Snell, E.J. (1989), Analysis of binary data, 2nd edition, Chapman and Hall.

[6] Culver, David C., Jernigan, Robert W. and O'Connell; Julie (1994), "The geometry of natural selection in cave and spring populations of the amphipod *Gammarus minus* Say (Grustacca: Amphipoda)", *Biological Journal of the Linnean Society*, Vol. 52, pp. 49-67.

[7] Davis, Linda June (1985), "Modification of the empirical logit to reduce bias in simple linear logistic regression", *Biometrika,* Vol. 72(1), pp. 199-202.

[8] Gart, J. J., Pettigrew, H. M., and Thomas, D. G. (1985), "The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses", *Biometrika*, Vol. 72, pp 179-190.

[9] Gill, Jeff (2000), *Generalized Linear Models: A Unified Approach*, Sage University papers, Sage Publications.

[10] Hastie, T.J. and Tibshirani, R.J. (1990), Generalized Additive Models, Chapman and Hall.

[11] Jaarola, Maarit , Tegelstrom, Hakan and Fredga, Karl (1997), "A Contact

Zone with Noncoincident Clines for Sex-Specific Markers in the Field Vole (Microtus Agresti)", *Evolution*, Vol.51(1), pp.241-249.

[12] McCullagh, P. and Nelder, J.A. (1989), Generalized Linear Models, Chapman and Hall.

[13] Munson, Peter J. and Jenigan, Robert W. (1989), "A cubic spline extension of the Durbin-Watson test", *Biometrika*, Vol. 76(1), pp. 39-47.

[14] Schluter, D. (1988), "Estimating the Form of Natural Selection on a Quantitative Trait", *Evolution*, Vol. 42(5), pp. 849-861.