SPLINE FUNCTIONS FOR LOGISTIC REGRESSION MODELING

Thomas F. Devlin, Montclair State College
Barbara J. Weeks, CIBA-GEIGY Pharmaceutical

ABSTRACT

Stone's additive spline models for regression functions are described. We discuss their implementation within our SABRE macro IMACRO LBETRES for logistic regression modeling and provide an illustrative application of their use.

1. INTRODUCTION

Spline functions provide an interesting and illuminating way of looking at the relationship between a response variable and one or more explanatory or predictor variables. They have advantages over higher order polynomials and compete with regression smoothers in allowing the data to establish the regression relationship and in assisting the data analyst in judging the non-linearity of the relationship. Their use has been discussed by Eubank (1984), Smith (1977), Buse and Lai (1977), and Wood (1974) among others. Eubank (1984) indicates that splines have been utilized for such diverse purposes as analysis of response curves in agriculture, economics and pharmacokinetics, estimation of the liquidity trap in economies, determining the base temperature in heat accumulation models, calibration of nuclear materials processing tanks, and analysis and estimation of meteorological fields.

Recently, Stone (1985) has had success using restricted additive splines. He fits a cubic spline with fixed knots which is constrained to be linear in the tails. We have modified the SABRE macro statement macro IMACRO LBETRES (1984, 1985), which provides maximum likelihood fits and the associated diagnostics proposed by Pregibon (1981) for logistic regression models, to include an option for fitting Stone's additive splines. We describe this implementation and provide an example of its use.

2. STONE'S ADDITIVE SPLINES - A REVIEW

In classical linear regression we assume that a response variable Y is related to some predictor or explanatory variable X by some smooth unknown response function f(x). Usually f is approximated by some linear, quadratic, or higher order polynomial function since polynomials provide a rich class of approximating functions which are dense in the set of all continuous functions on a closed bounded interval. Polynomials, however, have some important shortcomings. Their global behavior is determined by their values in a small interval and they are not well suited for modeling phenomena whose structure changes as a function of the predictor variable. As an illustration, we have fit a 4th order and a 6th order polynomial to the titanium heat data (deBoor, 1978) in which a property of titanium is to be related to heat (Figure 11). The wavelike nature of the polynomials is inconsistent with the data. Compare these with the fit of Stone's additive spline with 5 knots which requires only the estimation of as many parameters as the quartic polynomial.

A spline is generally defined to be a piecewise polynomial of degree n whose function values and first n - 1 derivatives agree at the points where they join. The values of the regressor variable at these joint points are called knots. Polynomials may be considered a special case of splines with no knots. Commonly employed splines are piecewise constants, linear splines, quadratic splines, and cubic splines. Piecewise continuous cubic splines are especially attractive since modest discontinuities in the third derivative cannot be detected visually. Stone (1985) argues, however, that these are "too flexible in the tails in relation to the amount of noisy data available there." He requires these splines to be linear in the tails, i.e. outside the smallest and largest knots, and constrains them to satisfy the boundary condition that the second derivatives vanish at those knots. Such a spline requires that K parameters be fit where K is the number of knots.

Knots are assumed to be fixed not random. They may be specified a priori or after looking at a stem-and-leaf plot of the distribution of the regressor variable X. If the values of the response variable are used in knot selection, the standard error formulas for Stone's splines are invalid. When a regressor variable is symmetrically distributed and the sample size is relatively large, Stone recommends the use of 5 knots, putting knots at the 5th smallest and 5th largest values of the regressor variable and placing equally spaced between these. The resulting spline has 4 parameters for the regressor variable plus one for the constant term - just one more than a cubic polynomial.

What about more than one explanatory variable, say \( X_1, X_2, \ldots, X_p \)? Stone proposes the additive spline model with the response function f defined by

\[
 f(x_1, \ldots, x_p) = b_0 + \sum f_j(x_j)
\]

with the \( f_j \) assumed to be restricted splines as above. The \( f_j \) are called component functions. If \( K_j \) is the number of knots for the jth regressor then each component function has \( K_j - 1 \) parameters and Stone's additive spline model has \( 1 + \sum (K_j - 1) \) parameters.
3. STONE’S ADDITIVE SPLINES FOR LOGISTIC REGRESSION MODELING USING SAS

Logistic regression is a statistical tool for describing, controlling or predicting proportions, rates or probabilities. It has found application in such fields as medicine, marketing, ecology and industrial processes. For the data analyst, an appealing view of logistic regression is as a member of the class of generalized linear models, which includes classical linear regression, general linear and log-linear models as well. From this conceptual framework, the data analyst can employ modeling strategies and statistical tools on logistic data analogous to those used for classical linear models. These include maximum likelihood estimation, hierarchical model comparisons, residual plots for detection of ill-fitting points and index plots for detection and assessment of the impact of influential observations. The SAS® statement-style macro LGTREG documented in the Proceedings of Vol. 10 provides a convenient mechanism for implementing analyses from this point of view. LGTREG now includes an option to fit Stone’s additive splines.

Some new features of IN macro LGTREG are:

- Stone’s additive spline models may be easily requested. Requests can be global or model specific. Knots can be input directly by giving the name of a SAS® data set containing their abscissas or by specifying the number of knots desired. If only the number of knots desired is specified (or if nothing is specified), knots are equally spaced between the 5th smallest and 5th largest values of each regressor variable. By default, five knots are provided as recommended by Stone. An output data set is generated containing the knots used.

- The model matrix is constructed starting from a truncated power basis (see appendix for details) and, as before, models are fit by maximum likelihood estimation and inferential statistics for model comparison common to the class of generalized linear models are generated.

- For each regressor variable, plots of estimates of the centered component functions ± one standard error are provided.

- Values of the components of fit, i.e., the estimated “centered” component functions, and their estimated standard errors are automatically stored in output SAS® data sets.

LGTREG is invoked by the following statements:

```
INCLUDE macro source code reference;
LGTREG Y=y Men CARRIER='X1 X2 ... Xn' /model options' optional parameters)
```

Here y and n are the variable names for the binomial responses and the corresponding number of trials respectively. The parameter N= may be omitted if the data are Bernoulli. X1, X2, ..., Xn are the regressor variable names and the model options available are 2 to request diagnostics and S to request Stone’s additive splines. Optional parameters which may appear in the LGTREG statement and which are documented in Devlin and Meeks (1985) and Devlin (1984) are DATA=, MODEL1=,...,MODEL9=, OUTFV=, OUTBBB=, OUTSDB=, MAXITR=, EPSILON=, and DGNs=. New optional parameters follow.

**SPLINE** Global request for Stone’s additive splines for all models.

NK= Number of knots (Default is 5).

**KNOTS**= SAS dataset name for user-defined knots. Variable names for each regressor must be the same as in the dataset specified by DATA=. When omitted, knots are equally spaced between 5th smallest and 5th largest values of each regressor.

OUTKNOTS=Output dataset name for knots used (default is OUTKNOTS).

**SASGRAPH** Requests SAS/GRAPH® Procedure GPLOT be used for component of fit plots.

**VPOS**= Controls VPOS option for all plots (default is 20).

**HPDS**= Controls HPDS option for all plots (default is 40).

OUTCOMP= Output data set name for components of fit (default is OUTCOMP).

OUTC SE= Output data set name for standard errors of components of fit (default is 40).

OUTBBB S= Output data set name for Pregibon’s basic building blocks if splines are requested (default is OUTBBB S).

OUTSDB S= Output data set name for diagnostic measures which assess the impact of influential points if splines are requested (default is OUTSDB S).

4. AN EXAMPLE

A study conducted between 1958 and 1970 at the University of Chicago’s Billings Hospital concerned the survival of patients after surgery for breast cancer (Haberman, 1973; Stone and Koo, 1985). There were 306 observations on four variables:

SURV= 1 if patient survived five or more years 0 otherwise

AGE= age of patient at time of operation

YEAR= year of operation (minus 1900)

NODES= number of positive auxiliary nodes detected in patient

We can fit a model which relates the log odds of survival to an additive spline function of the regressors AGE, YEAR, and NODES as well as one which relates the logit linearly and request diagnostics with the following statement:

```
LGTREG Y=SURV CARRIER='AGE YEAR NODES' DGNS
   SPLINE KNOTS=NOKEKNOT;
```

Since the empirical distribution of the regressor NODES is highly skewed, five knots for NODES are chosen so that the middle knots k2, k3, k4 are equally spaced on the log(1+k) scale between the 5th smallest k1 and 5th largest k5 values. These are stored
in the dataset NODEKNOT. Other knots are chosen as Stone recommends. Part of the output is presented in Figures 2 and 3. The component of fit plots (Figure 2) clearly indicate the non-linearity of the systematic component. This is confirmed by comparing the difference in the deviances of the two models (328.256-299.798)= 28.46 to the chi-squared distribution with (302-293)=9 degrees of freedom. The diagnostic plots (Figure 3) indicate several observations which are outlying in the response or model space and which deserve further attention.

Pursuing the analysis further, we fit additive spline models for all two and one carrier subsets of the regressors ABE, YEAR, and NODES by including additional MODEL= parameters, e.g. MODEL=N+NODES ABE', in the HBTRREG statement. The resulting summary of fits are presented in Table 2. Had Stone's additive spline functions not been used, the analyst would be led to believe a model in NODES alone is sufficient. A look at the spline models suggests otherwise.

To compare these models in a way that balances goodness of fit (deviance) and the degree of model complexity (p), we use a natural generalization of Mallows' Cₚ defined as follows:

\[ C_p = D_p / \hat{\sigma}^2 - (N-2p) \]

where N is the number of binomial responses, \( \hat{\sigma}^2 \) is the deviance of the full model divided by its degrees of freedom (299.798/293=1.023 here), and D_p is the deviance of a p parameter model. The following table displays these for the models with the smallest deviance for p = 2, 3, 4, 5, 9, and 13. It is apparent that NODES is an important variable which is related to the logit of survival in a non-linear way and also that an additive spline function of NODES and AGE is adequate.

<table>
<thead>
<tr>
<th>Model*</th>
<th>Deviance</th>
<th>DF</th>
<th>p</th>
<th>C_p</th>
</tr>
</thead>
<tbody>
<tr>
<td>FULL(S)</td>
<td>299.80</td>
<td>295</td>
<td>13</td>
<td>13.00</td>
</tr>
<tr>
<td>N+AGE(S)</td>
<td>307.41</td>
<td>297</td>
<td>9</td>
<td>12.44</td>
</tr>
<tr>
<td>NODES(S)</td>
<td>319.09</td>
<td>301</td>
<td>5</td>
<td>15.82</td>
</tr>
<tr>
<td>FULL</td>
<td>328.26</td>
<td>302</td>
<td>4</td>
<td>22.81</td>
</tr>
<tr>
<td>N+AGE</td>
<td>328.31</td>
<td>303</td>
<td>3</td>
<td>20.85</td>
</tr>
<tr>
<td>NODES</td>
<td>330.72</td>
<td>304</td>
<td>2</td>
<td>21.20</td>
</tr>
</tbody>
</table>

*FULL=Age+Year+Nodes
(S)=Stone's additive splines

The Tukey sum-difference plots (Cleveland, 1985) in Figure 4 provide a visual summary of the practical differences between the fitted probabilities of pairs of models. Fitted probabilities of survival for the additive spline model with the regressors ABE and NODES differ from those of the spline model for NODES only by more than .20 in a fair number of observations.

5. CONCLUSIONS

Stone's additive spline functions provide a rich class of models which allow the data more freedom to speak for itself and which allow for a powerful test against a broader class of alternatives to linearity. We hope that the implementation of this methodology within a SAS® statement style macro can provide data analysts with easier access to additive spline models. We also hope that this implementation will enable further empirical investigation of the properties of these models. We think it desirable that REB procedure have an option for fitting Stone's additive splines.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the State of New Jersey, MSC-SBR 5540-100-129-51030.

REFERENCES


APPENDIX

Suppose x is a regressor variable and f is Stone's spline function with KN knots t₁,..., t KN. The truncated power basis with terms \((x-t_k)\) can be used as a starting point for representing f. Note that \(u_x^\alpha = u_x^\alpha \) if \(x=0\) and it equals 0 otherwise. Then f is given by

\[
 f(x) = B_0 + B_1 x + \sum B_{k\alpha} \left( x-t_k \right)\alpha \\
 = (x-t_{k-1})\alpha (t_k-t_{k-1})/\Delta \\
 + (x-t_{k})\alpha (t_{k+1}-t_k)/\Delta \\
 \text{where } \Delta = (x-t_{k+1})
\]

where \(\alpha = (x-t_{k+1})\)

APPENDIX

Suppose x is a regressor variable and f is Stone's spline function with KN knots t₁,..., t KN. The truncated power basis with terms \((x-t_k)\) can be used as a starting point for representing f. Note that \(u_x^\alpha = u_x^\alpha \) if \(x=0\) and it equals 0 otherwise. Then f is given by

\[
 f(x) = B_0 + B_1 x + \sum B_{k\alpha} \left( x-t_k \right)\alpha \\
 = (x-t_{k-1})\alpha (t_k-t_{k-1})/\Delta \\
 + (x-t_{k})\alpha (t_{k+1}-t_k)/\Delta \\
 \text{where } \Delta = (x-t_{k+1})
\]

where \(\alpha = (x-t_{k+1})\)

APPENDIX

Suppose x is a regressor variable and f is Stone's spline function with KN knots t₁,..., t KN. The truncated power basis with terms \((x-t_k)\) can be used as a starting point for representing f. Note that \(u_x^\alpha = u_x^\alpha \) if \(x=0\) and it equals 0 otherwise. Then f is given by

\[
 f(x) = B_0 + B_1 x + \sum B_{k\alpha} \left( x-t_k \right)\alpha \\
 = (x-t_{k-1})\alpha (t_k-t_{k-1})/\Delta \\
 + (x-t_{k})\alpha (t_{k+1}-t_k)/\Delta \\
 \text{where } \Delta = (x-t_{k+1})
\]

where \(\alpha = (x-t_{k+1})\)

APPENDIX

Suppose x is a regressor variable and f is Stone's spline function with KN knots t₁,..., t KN. The truncated power basis with terms \((x-t_k)\) can be used as a starting point for representing f. Note that \(u_x^\alpha = u_x^\alpha \) if \(x=0\) and it equals 0 otherwise. Then f is given by

\[
 f(x) = B_0 + B_1 x + \sum B_{k\alpha} \left( x-t_k \right)\alpha \\
 = (x-t_{k-1})\alpha (t_k-t_{k-1})/\Delta \\
 + (x-t_{k})\alpha (t_{k+1}-t_k)/\Delta \\
 \text{where } \Delta = (x-t_{k+1})
\]
FIGURE 1. deBoor Titanium Data

SPLINES VS. POLYNOMIALS
TITANIUM DATA FROM DEBOOR (P. 222)

FIGURE 2. ADDITIVE SPLINES: BILLINGS DATA SET

MODEL = FULL(SPLINE)
COMPONENTS OF THE FIT (LOGIT SCALE)
FIGURE 2. (Cont'd.)
ADDITIVE SPLINES: BILLINGS DATA SET

FIGURE 3. Diagnostic Plots
STONE'S ADDITIVE SPLINES
SURVIVAL AFTER SURGERY FOR BREAST CANCER
DETECTING INFLUENCING OBSERVATIONS

FIGURE 4.
TUKEY SUM-DIFF PLOT OF FITTED VALUES (PROBABILITY SCALE)
SURVIVAL AFTER SURGERY FOR BREAST CANCER
TABLE 1. SAS® Listing for CARRIERS-'AGE YEAR NODES'

STONE'S ADDITIVE SPLINES
SURVIVAL AFTER SURGERY FOR BREAST CANCER

MODEL: FULL(SPLINE)

CONVERGENCE REACHED

CINFO  EPSILON  STEP
ROW1  6.00000000  6.00000000

COEFFICIENT ESTIMATES

COEFF  INTERCPT  AGE  Z113  Z123  Z133  YEAR
ESTIMATE  -31.1471  -0.3504  0.0006  0.7967  -0.0378  0.1076

STDERR  0.1572  0.0003  0.0003  0.0394  0.0308  0.0472

ASYMPTOTIC COVARIANCE MATRIX

COV  INTERCPT  AGE  Z113  Z123  Z133  YEAR
INTERCPT  334.9644  -0.0495  0.0116  -0.0296  0.0940
AGE  -0.0495  0.0247  -0.0001  -0.0001  -0.0001
Z113  0.0116  -0.0001  0.0001  0.0001  0.0001
Z123  -0.0296  -0.0001  0.0001  0.0001  0.0001
Z133  0.0940  -0.0001  0.0001  0.0001  0.0001
YEAR  -0.0296  -0.0001  0.0001  0.0001  0.0001

ANALYSIS OF DEVIANCE.... SUMMARY OF FITS
MODELS  DEVIANCE  CHI-SQ  DF
FULL  328.256  309.422  302.000
N+A  328.311  309.348  302.000
N+Y  330.713  316.747  303.000
A+Y  352.241  305.820  303.000
NODES  330.724  306.857  304.000
AGE  352.277  305.494  304.000
YEAR  353.681  306.061  304.000

ANALYSIS OF DEVIANCE.... SUMMARY OF SPLINE FITS
MODELS  DEVIANCE  CHI-SQ  DF
FULL  299.798  327.012  293.000
N+A  307.411  326.206  297.000
N+Y  341.765  312.041  297.000
A+Y  338.339  319.041  297.000
NODES  344.444  319.914  291.000
AGE  344.444  319.914  291.000
YEAR  344.444  319.914  291.000

SAS DATASET: ADULTMOT

TABLE 2. Summary of Fits - Many Models

ANALYSIS OF DEVIANCE.... SUMMARY OF FITS
MODELS  DEVIANCE  CHI-SQ  DF
FULL  328.256  309.422  302.000
FULL(SPLINE)  328.256  309.422  302.000

SAS DATASET: ADULTMOT

OBS KNOT1 AGE YEAR NODES
1 31 70 90 0.0000
2 25 90 70 1.0000
3 70 90 30 1.0000
4 70 70 10 1.0000
5 70 70 30 1.0000

651