Int. J. Sci. R. Tech., 2026 3(4)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

Generalized Linear Models with Restricted Cubic Splines

C. Suresh[1], N. Jayalakshmi[2], Kalava Harish[1], B. Triveni[3], R. V. S. S. Nagabhushana Rao[4], P. Manohar[5], B. Sarojamma*[1]

[1]*Research scholar, Department of Statistics, S V University, Tirupati.*
[2]*Department of Computer Science, S.G.S Arts College, TTD, Tirupati.*
[3]*Department of Computer Science, Sri Govinda raja Swamy Arts College(A), Tirupati.*
[4]*Department of Statistics, Vikrama Simhapuri University, Nellore.*
[5]*Department of HAS, Sri Venkateswara College of Engineering & Technology, Chittoor.*

ABSTRACT
In this paper we will discuss Generalized Estimating Equations modelling with natural cubic splines application for an example discrete dataset using R software 'geepack' and 'splines' package. We are taken data of 298 points for binary variables like gender and age for either new treatment or drug control.
Keywords: Generalized Linear Models, Cubic Splines

INTRODUCTION

In medical research new medicine and drug control plays vital role according to age and their body immune system. A medicine have different association with 70 year old and 45 years old person with diabetes and blood pressure patients. For example, if a model suggests that an increase in fasting plasma glucose (FPG) leads to an increase in HbA1c, this increase is the same if FPG increase from 120 to 165 mg/dL or if FPG change from 180 to 220 mg/dL. Assumption of linearity would be true often than the assumption of dichotomising data. Many scenarios this assumption would not be true. In longitudinal randomized clinical trials collection of efficacy or safety data often occur off-schedule beyond protocol allowed visit windows, considering this data at scheduled visits introduce potential bias due to clinical observations being carried forward or backward to the closest planned visits. In this situation it might be advantageous to treat time as continuous (actual time from baseline) rather than category variable in analysis of repeated measures mixed modelling. In these scenarios alternative modelling strategy would assume and explore non-linear continuous associations. There are many regression spline models are available to explore, but here we focus use of natural cubic splines. Loic Desquilbet and Fralcoismariotti [1] in this paper non-linear dose response association are widely criticized and restricted cubic spline functions are powerful tools to characterize a dosce response association between a Continues exposer and an outcome. SAS software is used for fitting Linear, logistic and cox model as wall as linear and logistic generalized estimating equations and statistical tests for overall non-linear associations. Hansnyquist [2] Penalty function approach as iterate procedure for obtaining the estimates of generalize linear model under linear restrictions on parameters by using likelihood ratio test, Wald test and Lagrange's Multiple tests are considered as alternatives for testing hypothesis about linear restrictions on the parameters. Mary C. Meyar Et.al [3] explains about non-parametrical modeling using regression splines with Bayesian Approach to generalized partial linear regression model where Knots may be modeled as fixed or free the R code to implement the methods is described. Arisperoglou Et.al [4], in their paper the discussed about popular splines like cubic splines, B-splines, Penalized splines, natural cubic and cardinal splines and these splines where fitted and tested using R. Laralusa and CRT Ahlin[5], in their paper periodic restricted cubical splines (RCS) and cubical splines are fitted using R packages for different knots for 500 units of data and obtained simulation results, model accuracy

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

tested using AKAIKE information criteria. Ulrich Halekoh Et.al [6], explains about generalized estimates fact for R and its applications by using cluster binary data was listed. Chin-schang Li [7], explains about a penalized log-likelihood ratio test statistic is constructed for a null hypothesis of the non-parametric components of a semi parametric generalized linear model component estimates using cubic B splines. The knots for this splines is fixed and its limiting its null distribution is the distribution of a linear combination of independent chi-square random variables each with 1 degree of freedom. A real life data is used for practical use of problem.

MEHODOLOGY:

Spline Regression Modelling

In general cubic splines are parabola curves fitted to data

$C(Y|X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$

Piece wise regression in basic form is linear function. Data is divided into parts, the intersection part is called knots. Generally, piece wise regression is as follows

$f(x) = \beta_0 + \beta_1 x + \beta_2 (x - a)_+ + \beta_3 (x - b)_+ + \beta_4 (x - c)_+$

where $(u)_+ = u,\ u > 0,$

$= 0,\ u \leq 0.$

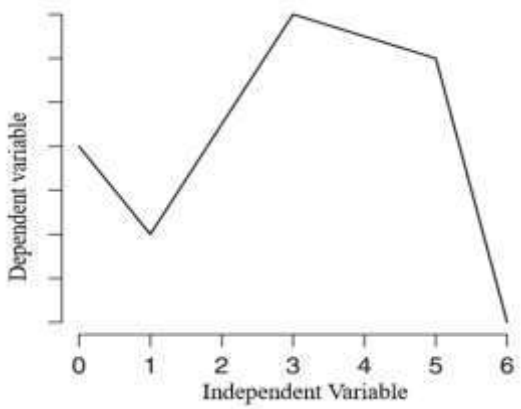
According to knots is increasing the function f(x) is as follows

$f(x) = \beta_0 + \beta_1 x,\ x \leq a$

$= \beta_0 + \beta_1 x + \beta_2 (x - a),\ a < x \leq b$

$= \beta_0 + \beta_1 x + \beta_2 (x - a) + \beta_3 (x - b),\ b < x \leq c$

$= \beta_0 + \beta_1 x + \beta_2 (x - a) + \beta_3 (x - b) + \beta_4 (x - c),\ c < x$



A linear spline function with knots at a=1, b=3, c=5

Cubic Splines (or) cubic piece wise regression:

In this situation cubic polynomial regression is fitted for each and every part of data. If data is divided into three knots is as follows

$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X - a)^3_+ + \beta_5 (X - b)^3_+ + \beta_6 (X - c)^3_+$

$= X\beta$

with constructed variables: $X_1 = X$, $X_2 = X_2$, $X_3 = X_3$, $X_4 = (X - a)^3_+$, $X_5 = (X - b)^3_+$, $X_6 = (X - c)^3_+$

Restricted Cubic Splines:

The restricted spline function with k knots $t_1,...,t_k$ is given by

$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{k-1} X_{k-1},$

where $X_1 = X$ and for $s = 1,...,m-2$, $X_{j+1} = (X - t_s)^3_+ - (X - t_{m-1})^3_+ (t_m - t_s) / (t_m - t_{m-1}) + (X - t_m)^3_+ (t_{m-1} - t_s)/(t_m - t_{m-1})$

It can be shown that X_j is linear in X for $X \geq t_m$. For numerical behaviour and to put all basis functions for X on the same scale, the above terms divide by $\tau = (t_k - t_1)^2$

Once $\beta_0,...,\beta_{k-1}$ are estimated, the restricted cubic spline can be restated in the form

$f(X) = \beta_0 + \beta_1 X + \beta_2(X - t_1)^3_+ + \beta_3(X - t_2)^3_+ + ... + \beta m_{+1}(X - t_m)^3_+$

by dividing $\beta_2,...,\beta_{k-1}$ by τ and computing

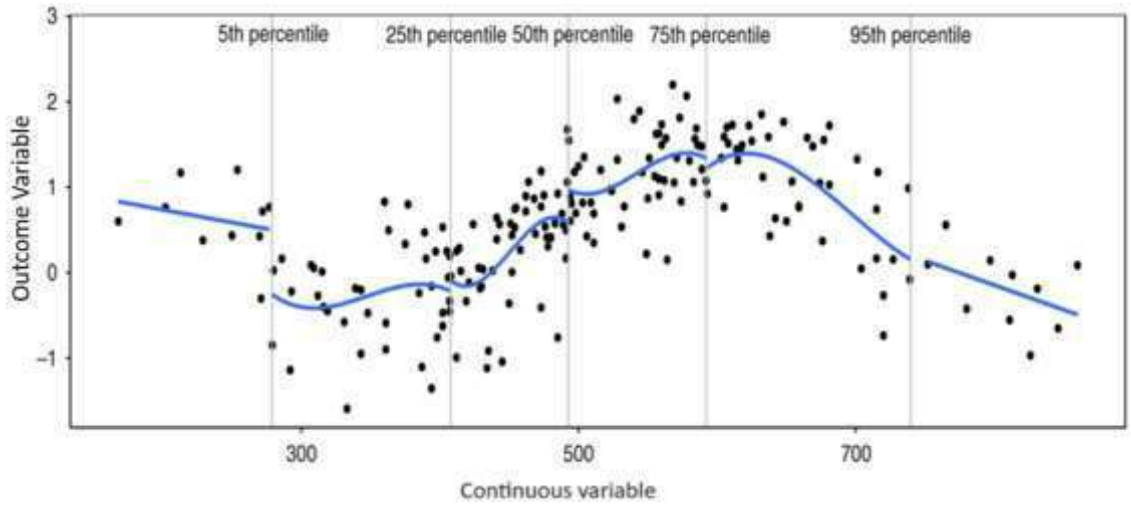
$\beta_k = [\beta_2(t_1 - t_m) + \beta_3(t_2 - t_m) + ... + \beta_{m-1}(t_{m-2} - t_m)]/(t_m - t_{m-1})$

$\beta_{k+1} = [\beta_2(t_1 - tm_{-1}) + \beta_3(t_2 - t_{m-1}) + ... + \beta_{m-1}(t_{m-2} - t_{m-1})]/(t_{m-1} - t_m)$

A test of linearity in X can be obtained by testing H_0: $\beta_2 = \beta_3 = ... = \beta_{m-1} = 0$

Default values for knots

m	Quantiles
3	0.10 0.5 0.90
4	0.05 0.35 0.65 0.95
5	0.05 0.275 0.5 0.725 0.95
6	0.05 0.23 0.41 0.59 0.77 0.95
7	0.025 0.1833 0.3417 0.5 0.6583 0.8167 0.975



In above figure, the scatter plot with five knots included along with x-axis, within first and last portions linear regression lines and interior portions cubic polynomials fit to the data. Restricted cubic spline will be achieved with the curve to be continuous and smooth at the knots.

Generalized Linear Models (GLM):

In bio statistic the predicted number of hypoglycaemic episodes in diabetes insulin first time use patients is following a Poisson distribution and a log link, while case of predicted probability of occurrence of hypoglycaemic events would follows Bernoulli distribution or binomial distribution and or logit link function.

Generalized Estimating Equations (GEEs)

Y_{ij} is binomial with mean $\mu_{ij} = \pi_{ij}$, and the logit link would be used for g. If predictors/covariates say drug (treatment/placebo), follow-up time (years), gender and age are added, we have x_{ij}^{l}=$(1, drug_i, gender_i, age_i, followup\ year_{ij})$ and

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + drug_i\,\beta_1 + gender_i\beta_2 + age_i\,\beta_3 + year_{ij}\beta_4$$

Application:

An example dataset with safety laboratory parameter results over several follow-up years for each subject and whether results are normal or abnormal categorized as binary outcome used for analysis implementation, along with demographic variables

like sex and age data for about 298 subjects who were received either new treatment or control drug. Based on similar research it was presumed that the log odds of abnormal lab result may evolve nonlinearly during follow-up period; It is also plausible that the log odds evolutions over time are different between males and females, and between control drug and new treatment subjects. Additionally, age is an important risk factor, and the effect of age may be modified by sex. Above questions were translated into a suitable GEE model for the log odds of abnormal lab results, by using different working correlation matrices and for the nonlinear term (follow-up time in years) use natural cubic splines with 2, 3 or 4 degrees of freedom. Results were compared for various number of knots and correlation matrices; Data are not available over same range of follow-up years for all subjects (i.e., number of repeated measures differing significantly among all subjects), thus unstructured matrix was not chosen; Exchangeable matrix was chosen over autoregressive and independence matrices, considering sustained correlation over years. After checking QIC for different modelling choices, below GEE model is suggested to better fit for the data with nonlinear time using natural cubic splines, also adjusting for treatment, gender and age.

*geefit <- geeglm(binary_outcome ~ ns(year, 3) * (drug + sex) + age + age : sex, family = binomial(), data = InputData, id = id, corstr = "exchangeable")*

Output

```
 Coefficients:
                              Estimate   Std.err  Wald Pr(>|W|)
(Intercept)                   1.197296  0.416738  8.25  0.00407 **
ns(year, 3)1                 -0.694651  0.549222  1.60  0.20595
ns(year, 3)2                 -1.446637  0.497662  8.45  0.00365 **
ns(year, 3)3                 -2.768619  0.771634 12.87  0.00033 ***
drugTreatment                 0.177416  0.236152  0.56  0.45248
sexmale                      -0.421826  1.068230  0.16  0.69293
age                          -0.012170  0.008129  2.24  0.13437
ns(year, 3)1:drugTreatment    0.491051  0.733645  0.45  0.50328
ns(year, 3)2:drugTreatment   -0.985646  0.707515  1.94  0.16359
ns(year, 3)3:drugTreatment   -2.416412  1.011570  5.71  0.01690 *
ns(year, 3)1:sexmale          1.074291  1.015696  1.12  0.29020
ns(year, 3)2:sexmale          1.022894  0.989692  1.07  0.30135
ns(year, 3)3:sexmale          2.927860  1.320684  4.91  0.02663 *
sexmale:age                  -0.000482  0.018276  0.00  0.97897
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this data modelling we found the log odds of abnormal results evolving nonlinearly during follow-up years. The log odds abnormal evolutions over time are different between males and females, and between control and treatment subjects. Effect of age on results are not modified by sex.

CONCLUSIONS:

In this paper we are progress on both theoretical and computational fronts for cubic splines for bio statistics variables like age and gender with treatment. As compared with general cubic splines and restricted cubic splines, second one plays significant role in fitting of data. Restricted cubic splines provide a useful tool for the analysis of the effect of a continuous predictor or an outcome.

REFERENCE

1. Loic Desquilbet and Francois Mariotti, Statistics in Medicine, 2010, 29, 1034-1057.
2. Hans Nyquist, Restricted Estimation of generalized linear Models, Appl. statist, 19, 91,40 (1), pp 133-141
3. Mary C Meyer, Amber J. Hanstadt and Jennifer A-Hoeting, Bayesian estimation and inference for generalized partial linear models using shape-restricted Splines, Journal of Nonparametric Statistics, Vol23(4) December 11 pp 867-884
4. Aris Perperogloa, Welli Sauerbrei, Michal Abrahnowica and Matthias Schmis, A review of spline function procedures in R, BMC Medical Research Methodology, 2019, 19:46, PP 1 to 16.

5. Lara Lusa and Crt Ahlin (2020) Restricted Cubic Splines for modeling periodic data, PLOS ONE 15 (10), pp 1 to 17.
6. Ulrich Halekoh, Soren Hojsgaard and Jun Yan (2006), The R package geepack for Generalized Estimating Equations, Journal & Statistical Software, Vol 15 (2), 1-11
7. Chin-Shang Li (2012), Lack-of-Fit Tests for generalized Linear Modls via splines, communications in statistics theory and Methods, 41, 4240-4250.

HOW TO CITE: C. Suresh, N. Jayalakshmi, Kalava Harish, B. Triveni, R. V. S. S. Nagabhushana Rao, P. Manohar, B. Sarojamma*, Generalized Linear Models with Restricted Cubic Splines, Int. J. Sci. R. Tech., 2026, 3 (4), 258-262. https://doi.org/10.5281/zenodo.19479817