Journal of Science Research (2013) Vol. 12: 299-304

Cubic Smoothing Spline on some Selected Blood Pressure Variables

Olubusoye, O.E., Alaba, O.O.* and Oranye, H.E.

Abstract

Generalized Additive Model has become an elegant and practical option in modelling non-linear and linear effects of covariates as well as the non-Gaussian response variable. This study considered modelling Blood Pressure (BP) using data with two levels of BP (abnormal and normal) and eight predictors which have both linear and non-linear effects. The non-parametric functions were estimated in a flexible manner using cubic smoothing spline in an iterative method called the Back-fitting algorithm. The Cubic smoothing spline was applied to the metrical covariates (Age and BMI), which gave significant results ($p < 0.0001$ and 0.0082 respectively) compared to the linear fit which was not significant. The empirical findings of this study have established that BMI and Age have significant non-linear effect while sex and cholesterol level have significant linear effect on BP.

Key words: Generalized Additive Model, Categorical Covariates, Metrical Covariates, Smoothing Spline, Blood Pressure.

Introduction

In recent years, many studies have focused on Blood Pressure (BP) because it is a useful indicator of the cardiovascular diseases, such as hypertension, heart attack and asthma [20]. Due to the high level of stress of the modern society, many people have hypertension whether young or elderly. Low or high BP can cause heart attack, stroke and other problems. Systolic Blood Pressure (SBP)/ Diastolic Blood Pressure (DBP) higher than 140/90 mmHg or lower than 90/60 indicates high or low BP respectively [1]. It is also a major cause of disability and an important risk factor for death, accounting for about 7.5 million deaths per year (13% of all deaths) [1].

Due to the population growth and ageing, the actual number of people with uncontrolled BP rose from 600 million in 1980 to nearly 1 billion in 2008. High-income countries achieved large reductions in uncontrolled BP, with the most impressive progress seen in women in Australia and men in North America [21]. [5] stated that being overweight, obese, having abnormal BP and high cholesterol are no longer Western problems or problems of wealthy nations, but their presence has shifted towards low and middle income countries, making them global problems. Literature has documented elevated BP as the leading cause of morbidity throughout the world, and also a major contributor to Cardiovascular Disease (CVD). The BP levels of a population have been seen to be influenced by interrelated factors such as biological, behavioral and socio-economic factors [4, 8, 9, 10, 12, 13]. Blood Pressure has become one of the top national health priorities, therefore it cannot be over-emphasized.

Model-based analyses are becoming important sources of global information on the prevalence of diseases, largely because of the absence of reliable national level empirical data, particularly in developing countries. Blood Pressure is associated with many factors including Body Mass Index (BMI), age, diet, exercise, weight, cholesterol level, pulse rate, alcohol, gland disorder, dehydration, medication e.t.c. The effects are often modelled parametrically. However, in real life, experience has shown that metrical covariates often have non-linear effects [3, 6, 19]. The underlying assumption (which are not often met in practice), of a parametric

Olubusoye, O.E., Alaba, O.O.* and Oranye, H.E.
** Department of Statistics, University of Ibadan*

linear predictor for assessing the effects of covariates on responses seems to be too strong and restrictive in realistically complex situations. It is therefore difficult to specify a parametric functional form for the non-linear effects of metrical covariates. Hence, it is necessary to relax the parametric linear assumption by a flexible method of estimating metrical covariates.

Generalized Additive Model (GAM) relaxes the assumption of linearity between the predictors and handles the categorical nature of responses. We shall explore the effects of the different covariates on the categorical responses of BP. Following the introductory section, section 2 is devoted to the methodology and data description. In section 3, we presented the analysis and discussion of results. We concluded the paper in section 4.

Methodology

The Generalized Additive Model unifies the family of Generalized Linear Models (GLM), by replacing the linear functional form by a sum of smooth functions enabling the discovery of a non-linear fit between a variable and a response [16, 17, 18]. It provides a powerful class of models for modelling non-linear effects of continuous covariates in regression models with non-Gaussian responses. A huge variety of competing approaches are now available for modelling and estimating non-linear functions of incorporating variables in a non-parametric way using smooth functions such as spline. Non-parametric estimation of metrical covariates assumes that the functions are unknown but smooth. There are several alternatives for estimating smooth functions; [17] assumes that smoothness of the unknown functions (f) is controlled by penalty terms. The fit of GAM is based on the local scoring algorithm; an extension of the Newton-Raphson algorithm used for fitting GLMs. The local scoring algorithm uses a Back-fitting algorithm that iteratively fits a smoothing function [2].

The GAM is given as

$$\eta_i = x_i'\beta + f_i'\gamma + \varepsilon_i \qquad (1)$$

where,

$f_i'\gamma$ are the smooth functions from the non-linear effect of the metrical covariates $x_i'\beta$ are the linear effect of the categorical covariates and ε_i is a noise variable such that $\varepsilon_i \sim N(0, \delta^2)$.

Estimation of (equation 1) above can be achieved in several ways. However, we consider the cubic smoothing spine in this study because it is an optimization technique. It is a piecewise polynomial fit and many studies have shown that k=3 is sufficient [7, 15].

Given, $(y_i, x_i, f_i)_{i=1,\dots,}$ on response y_i a vector $(x_1,\dots,x_p)$ of categorical covariates and a vector $(f_1,\dots,f_k)$ of metrical covariates. The simultaneous effect of the covariates on the response is modelled by a linear predictor

$$\hat{\eta}_i = x_i'\beta + f_i'\gamma \qquad (2)$$

However, because of the case of non-Gaussian response, where y_i becomes

$$y_{ir} \quad {\scriptstyle i=1,\dots,n \atop r=1,\dots,m}$$

$$\hat{\eta}_{ir} = x_i'\beta + f_i'\gamma \qquad (3)$$

So that,

$$\mu_i = g(\hat{\eta}_{ir}) \qquad (4)$$

The Penalized Maximum Likelihood Estimation (PLME) is used to avoid overfitting, which is characterized by a score function $\frac{\partial l(\eta, y)}{\partial \eta}$

The PLME for the non- Gaussian responses is obtained by

$$l(\eta, y) - \frac{\lambda}{2}\int (f^{''}(\gamma))^2 d(\gamma) \rightarrow \max_f \qquad (5)$$

where,
l is the log likelihood of the linear predictor and the second term penalizes the integrated squared curvature of the function $f(\gamma)$.

$\lambda > 0$ is the smoothing parameter that controls the trade-off between fit and smoothness.
To maximize (5) using B-splines through the local scoring algorithm, we obtained

$$g(\beta, f) = l(\eta, y) - \frac{\lambda}{2} f^{'}Mf \qquad (6)$$

where,
M:= $B(B^{'}B)^{-1}K(B^{'}B)^{-1}B^{'}$

K is the penalty matrix that penalizes too rough function f.

The solution method for finding $\hat{\beta}, \hat{f}$ by optimization technique is given by Hastie and Tibshirani (1990).
$f(\gamma)$ is estimated by repeatedly smoothing the adjusted dependent variable on X [14]
Given, $p_1 := X\beta$ *and* $p_2 := f(\gamma)$
Then,

$$h(\mu) = \eta(X, \gamma) = p_1 + p_2 \qquad (7)$$

where,
X is an $n\ x\ m$ matrix
p_1 & p_2 are vectors of $X\beta$ and $f(\gamma)$ respectively.
We maximized (5) over p_1 *and* p_2 by solving

$$\frac{\partial g(\beta, f)}{\partial p_1} = (\frac{\partial \eta}{\partial p_1})^{'} \frac{\partial l(\eta, y)}{\partial \eta} = 0$$
$$\frac{\partial g(\beta, f)}{\partial p_2} = (\frac{\partial \eta}{\partial p_2})^{'} \frac{\partial l(\eta, y)}{\partial \eta} - \lambda MP_2 = 0 \qquad (8)$$

which are non-linear in η *and* p_2.
We then linearized them around η^0 (current starting solution). We obtained (9) which is a Newton-Rapson method.

$$\frac{\partial l(\eta, y)}{\partial \eta} = \left.\frac{\partial l(\eta, y)}{\partial \eta}\right|_{\eta_0} + \left.\frac{\partial^2 l(\eta, y)}{\partial \eta \eta^{'}}\right|_{\eta_0} (5)(\eta - \eta_0) = 0 \qquad (9)$$

If we put (9) in (8) and let $r = \frac{\partial l(\eta, y)}{\partial \eta}$ and $\nabla = -\frac{\partial^2 l(\eta, y)}{\partial \eta \eta^{'}}$
Then,

$$\begin{pmatrix} \nabla & \nabla \\ \nabla & \nabla + \lambda M \end{pmatrix} \begin{pmatrix} p_1^1 - p_1^0 \\ p_2^1 - p_2^0 \end{pmatrix} = \begin{pmatrix} r \\ r - \lambda M p_2^0 \end{pmatrix} \qquad (10)$$

where $(p_1^0, p_2^0) \rightarrow (p_1^1, p_2^1)$ is a Newton-Raphson step, ∇ and r are evaluated at η_0.

(6)

So that when,
$h := \eta^0 + \nabla^{-1} r$ *and* $S_B := (\nabla + \lambda M)^{-1}\nabla$
which is a weighted B-spline operator.
Then (equation 10) is

$$\begin{pmatrix} \nabla & \nabla \\ S_B & I \end{pmatrix} \begin{pmatrix} p_1^1 \\ p_2^1 \end{pmatrix} = \begin{pmatrix} \nabla \\ S_B \end{pmatrix} h \qquad (11)$$

is transformed to

$$\begin{pmatrix} p_1^1 \\ p_2^1 \end{pmatrix} = \begin{pmatrix} X\beta^{'} \\ f^{'} \end{pmatrix} = \begin{pmatrix} h - p_2^1 \\ S_B(h - p_1^1) \end{pmatrix} \qquad (12)$$

So that we estimate $\hat{\beta}$ *and* $\hat{f}$
Then,

$$\hat{p}_1 = X\hat{\beta} = X\{X^{'}\nabla(I - S_B)X\}^{-1}X^{'}\nabla(I - S_B)h \qquad (13)$$
$$\hat{p}_2 = \hat{f} = S_B(h - X\hat{\beta})$$

where
X is the regression matrix for the values x_i
S_B computes the weighted B-spline smoothing on the variable γ_i with weights given by $\nabla = -\frac{\partial^2 l(\eta, y)}{\partial \eta \eta^{'}}$ (14)
h is the adjusted dependent variable
From (equation 12), the Newton-Raphson updates are an additive model fit, which solves a weighted and penalized quadratic criterion which is an approximation of the penalized log-likelihood.

Data Description

Data used in this study are based on 69 patients in St. Anne's Anglican Hospital, Ibadan. The data were obtained at the laboratory unit of the Hospital, with two levels of BP (response variable). The SBP/DBP is coded as 0 if it falls within the range; 91 – 139/ 61 – 89mmHg or abnormal BP is coded as 1 if it falls outside the range specified above, whether low or high. The predictors include: age of patients in years, BMI, pulse rate (PR), respiration rate (RR), sugar level (SL), cholesterol level (ChL) and temperature (temp) which is coded as 0 for normal (between 35 – 37^0C) and 1 for abnormal (outside this range 35 – 37^0C). The Generalized Cross Validation (GCV) function is used as the criterion to choose the smoothing parameters.

Presentation and Discussion of Results

Age and BMI were estimated in the non-parametric manner (Fig. 1). The model converged immediately with 3.2958024E-9. The local scoring iteration was performed 14 times with value of 1.9096565E-9. The deviance of the final estimate is given as 27.190205628. The parametric predictor variables were not significant except sex which is 0.0052 is significant at $\alpha = 0.05$. The metrical covariates which are BMI and AGE when analysed with GLM were not significant. We then splined the metrical covariates using the cubic smoothing spline. From the analysis, the two metrical covariates (age and BMI), were significant when cubic smoothing spline was applied. The value of $log_{10}(n\lambda$) that minimized the GCV function is 0.479954. The final smoothing spline estimate is based on LOGNLAMBDA = 0.9895. The residual sum of squares is 6.2462, and the degree of freedom is 21.3716. The standard deviation is 0.3621.

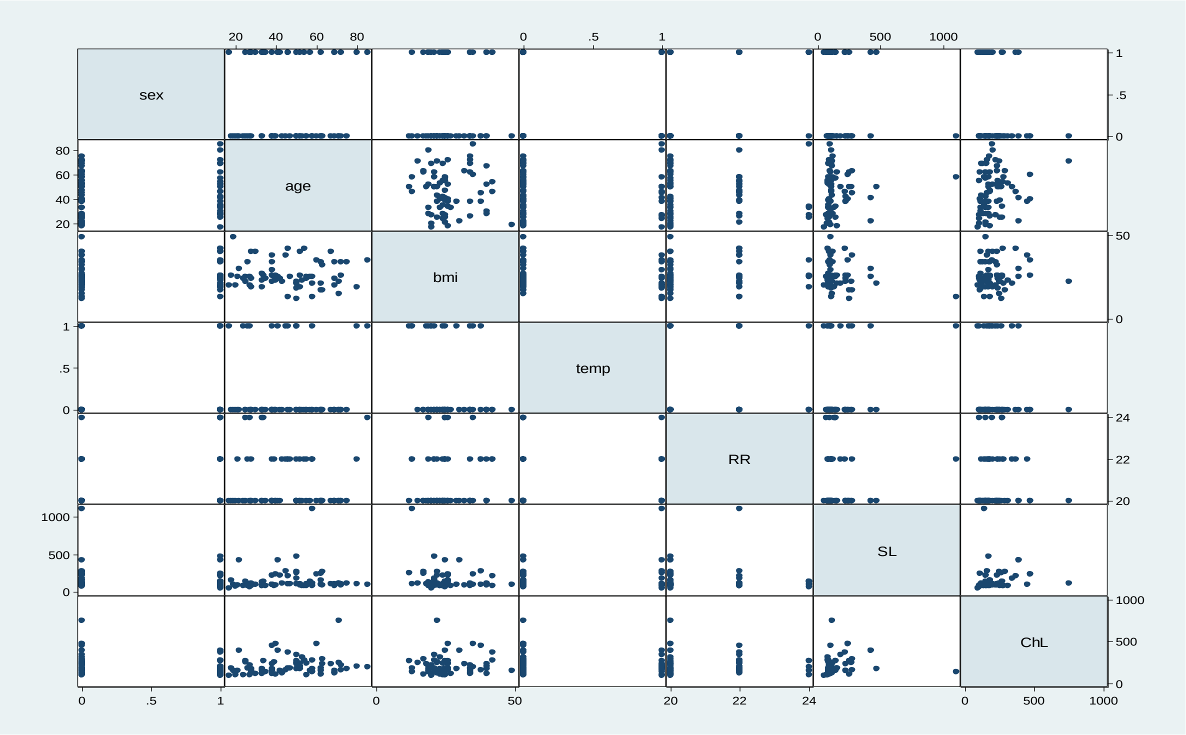

Fig. 1: Scatter plot of the selected variables.

Table 1: Iteration Summary and Fit Statistics

Number of local score iterations	14
Local score convergence criterion	1.9096565E-9
Final Number of Backfitting Iterations	1
Final Backfitting Criterion	3.2958024E-9
The Deviance of the Final Estimate	27.190205628

The local score algorithm converged.

Regression Model Analysis
Parameter Estimates

	Parameter	Standard		
Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	-14.27437	18.05014	-0.79	0.4325
sex	-3.17061	1.08952	-2.91	0.0052
temp	-0.82507	1.22996	-0.67	0.5052
PR	0.38994	0.25560	1.53	0.1329
RR	-0.48416	0.46095	-1.05	0.2982
SL	0.00932	0.00533	1.75	0.0862
ChL	-0.00220	0.00577	-0.38	0.0014
Linear(bmi)	-0.31565	0.17860	-1.77	0.0828
Linear(age)	-0.00710	0.03361	-0.21	0.8334

Smoothing Model Component(s): spline(bmi) spline(age)

Smoothing Model Analysis
Analysis of Deviance

Sum of Source	DF	Squares	Chi-Square	Pr > ChiSq
Spline (bmi)	3.00000	27.378388	27.3784	<.0001
Spline (age)	3.00000	11.770351	11.7704	0.0082

Conclusion

The Back-fitting iteration is better compared to the Local Scoring criterion. Conclusively, in this study we have successfully considered both the linear and non-linear effects of the covariates that affect Blood Pressure simultaneously. Also, the categorical responses were adequately taken care of.

References

(1) Arthur. 2010. Effects of High Blood Pressure. American Heart Association Article.

(2) Berg, D. 2007. Bankruptcy Prediction by Generalized Additive Models. The Norwegian Computing Center and University of Oslo.

(3) Borgoni, R. and Billari, F.C. 2001. Spatial Discrete – Time Event History models: An Application to Home Leaving. Rostock D-18057, Germany.

(4) Ezike, C., Ugwu, C., Ezeanyika, L., Olayemi, A. 2008. BP Patterns in Relation to Geographic Area of Residence: A Cross Sectional Study of Adolescents in Kogi State, Nigeria. BioMed Central Ltd.

(5) Ezzati, M. 2011. Country-by-Country Trends in Obesity, Cholesterol and BP since 1980. Published online. First by the Lancet.

(6) Fahrmeir, L. and Lang, S. 2000. Bayesian Semiparametric Regression Analysis of Multicategorical Time-space Data. *Ann. Inst. Statist. Math*. 53, 10-30.

(7) Green, P.J. and Silverman, B.W. 1994. Non-parametric Regression and Generalized Linear Models. A Roughness Penalty Approach. Monographs Statistics and Applied Probability 58, Chapman and Hall, London.

(8) Kusuma, Y.S., Abu, B.V. and Naidu, J.M. 2009. BP levels among Cultural Populations

of Visakhapatnam District, Andhra Pradesh India. *Anna Hum Boil*; 29: 502-510.

(9) Murphy, M. 2010. Low BP and preserved systolic function in elders with Heart failures. University of Utah, UT, USA. *J. Cardiovasc. Nurc*. 2010 Sept. – Oct., 25(5); 405 – 10.

(10) Osondu, N., Jewase, E., Barnabas 2009. Sex Differences and Relationship between BP and Age among the Ibos of Nigeria. *The Internet Journal of Biological Anthropology* 2009. Vol. 3 number 2.

(11) Pickering, T.G. 2005. Recommendations for blood pressure measurement in humans and experimental animals: part 1: blood pressure measurement in humans: a statement for professionals from the subcommittee of professional and public education of the American heart association council on High Blood Pressure research," *Journal of the American Heart Association*, Vol. 45, pp. 142{161, 2005.

(12) Rohrscheib. 2009. Age related BP Patterns and BP Variabitly among Haemodialysis Patients. *Clinical Journal of American Soc Nephrol*, 3:1407-1414.

(13) Sanya, A.O., Ogwumike, O.O., Ige, A.P., Ayanniyi, O.A. 2009. Relationship of Waist-Hip Ratio and Body Mass Index to Blood Pressure of Individuals in Ibadan North Local Government. Department of Physiotherapy, College of Medicine, University of Ibadan.

(14) Taylan, P., Weber, G. and Urgan, N.N. 2003. On the Parameter Estimation for Generalized Partial Linear Models with B–Splines and Continuous Optimization.

(15) Terzi, E. 2009. Using of Generalized Additive Model for model selection in multiple passion regression for air pollution. Dept. of Statistics, University of Ondokuz Mayis, Scientific Research & Essay Vol. 4(9) 867 – 871.

(16) Hastie, T. and Tibshirani, R. 1986. Generalized Additive Models. Statistical Science, Vol. 1. No. 3, 297-318.

(17) Hastie, T. and Tibshirani, R. 1987. Generalized Additive Models, Cubic Splines and Penalized Likelihood. Technical Report No. 390, May 22, 1987. Department of Statistics, Stanford University. Stanford, California.

(18) Hastie, T. and Tibshirani, R. 2002. Generalized Additive Models. Department of Statistics, Stanford University. Stanford, California.

(19) Wagner, S. and Jerak, A. 2003. Estimating Probabilities of EPO Patent Oppositions in a Bayesian Semiparametric Regression Framework.

(20) World Health Organization 2007. Prevention of Cardiovascular Disease: Pocket Guidelines for Assessment and Management of Cardiovascular Risk: (WHO/ISH Cardiovascular Risk Prediction Charts for the African Region). Geneva: *World Health Organization Press*.

(21) World Health Organization 2012. Global Health Observatory.