Generalized Additive Model for Evaluation of Premium for Diabetic Patients

Harmanpreet Singh Kapoor, Kanchan Jain and Suresh Kumar Sharma

Department of Statistics, Panjab University, Chandigarh
Email: harman.pu.87@gmail.com, jaink14@gmail.com and ssharma643@yahoo.co.in

Abstract The widely used Generalized Additive Model (GAM) is a flexible and effective technique for conducting non-linear regression analysis. It relaxes the usual parametric assumptions and enables us to uncover structure in the relationship between the independent and dependent variable in exponential family that might be otherwise missed. In this paper, we describe the use of GAM procedure to determine the premium amount of diabetic patients in presence of predictors or covariates. The risk factors responsible for the cause of the diabetic patients have also been identified using Logistic GAM. The procedure has been applied to a real life data set of 134 diabetic patients by smoothening the effect of covariates.

Keywords: Generalized additive model, smoothing, premium, generalized cross validation, splines.

1 Introduction

In USA, nearly 26.8 million people suffer from diabetes whereas in India, the situation is even worse and 50.8 million suffer from this slow-killer. Type-2 diabetes (mellitus) is the most common form of diabetes in which, either the body does not produce enough insulin or the cells ignore the insulin. In the long run, high blood glucose levels may affect kidneys, nerves, eyes or even heart. Less than 2% people in India opt for medical insurance and most of the insurance companies do not provide any insurance to diabetic patients. Moreover, due to very high mortality and morbidity rate of diabetic patients, it is very expensive for the insurance companies to provide insurance cover to these patients. Hence, it is of utmost importance for an insurance company to provide insurance cover to diabetic patients and to determine the premium for such patients with high accuracy. Generalized Additive Model (GAM) proposed by Hastie and Tibshirani ([1] and [2]), plays a very important role in such situations.

If Y is the response variable and $X_1, X_2, \ldots, X_k$ are k predictors variables, a standard linear multiple regression model assumes the form

$$E[Y|X_1, \ldots, X_k] = \beta_0 + \sum_{i=1}^{k} \beta_i X_i$$

where β_0 is the intercept, $\beta_i, i = 1, \ldots, k$ are the regression coefficients and Y follows normal distribution.

Additive models (AM) extend the parametric form of predictors in the linear model to nonparametric forms. An additive model for response variable Y is defined as

$$E[Y|X_1, \ldots, X_k] = s_0 + \sum_{i=1}^{k} s_i(X_i)$$

where response variable has a probability density function belonging to the exponential family, s_0 is a constant used for smoothing and s_i's, $i = 1, 2, \ldots, k$ are the smoothers.

The response-predictor relationship is defined through a link function written as

$$g(E[Y \mid X_1, \ldots, X_k]) = \eta\beta_0 + \sum_{i=1}^{k} \beta_i X_i$$

The coefficients $\beta_i, i = 1, \ldots, k$ are estimated by a form of the iterative reweighted least square method, called the Fisher scoring procedure [3],[4] and [5].

Generalized linear models (GLM) are an extension of linear models to exponential family of distribution for which the probability function of Y has the form

$$f_y(y, \theta, \phi) = exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi))$$

where a, b and c are arbitrary functions, θ is a natural parameter used in the model for relating the response Y to the covariates and ϕ is a scale parameter.

GAM uses a link function to establish a relationship between the mean of the response variable and a smoothed function of the explanatory variable(s). Hence, it possesses the properties of both AM and GLM. Due to this feature, GAM has more applicability in different fields [6]. It has the ability to model data from exponential family of distributions [3]. Many statistical models like additive models for Gaussian data, logistic models for binary data, non-parametric log-linear models for Poisson data belong to this general class [7]. It is the most versatile method for handling non-parametric regression as it has more flexibility than the traditional parametric modeling tools.

Generalized Additive Model (GAM) is defined as

$$g(E[Y \mid X_1, \ldots, X_k]) = \eta = s_0 + \sum_{j=1}^{k} s_j(X_j)$$

where $s_0, s_1, \ldots, s_k$ are smooth functions and define the additive component.

To the best of our knowledge, no one has explored the prediction of the premium for diabetic patients using GAM. In our study, we use the GAM to evaluate the premium of diabetic patients based on the risk factors, that is, covariates listed in Table 1.

Table 1. List of covariates used in GAM.

Covariate name	Covariate Explanation
age	age of the patient
vitaminD	level of vitamin D
Hb	level of hemoglobin
calcium	level of calcium
crr	measurement of cardiorespiratory reflex
alb	level of albumin
phos	level of phosphate
BMDsp	Bone mineral density of spine
BMDnof	Bone mineral density measured in the neck of the femur
FBS	fasting blood sugar of patient (main risk factor)
premium	single premium amount for whole life assurance contract of 0.1 million (in Rs.) (dependent variable)

As non-linear relationship exists between response (premium) and covariates (age, vitaminD, Hb, calcium, crr, alb, phos, BMDsp, BMDnof, FBS), the GAM provides an estimate of premium, based on the main risk factor measured through FBS by smoothing the effect of other covariates.

Section 2 discusses the form of GAM and also determines the degrees of freedom for each covariate under consideration using R software. Using a real life data set, significant covariates have been determined in Section 3. Predicted plot for insurance premium is also presented. In Section 4, we discuss the GAM

procedure in SAS which is useful for predicting the diabetic status of an individual when the values of other significant covariates are known.

2 Development of GAM and Determination of Degrees of Freedom

We first identify those covariates / risk factors for diabetes used for model building and to form the base model. Smoothing splines [6] are used to estimate the unknown functions with appropriate degrees of freedom. Finally, the main variable is added to the base model to determine the premium of diabetic patients.

We fit a GAM on the data of 134 patients by taking into consideration the values of 9 independent covariates listed in Table 1 for each patient. The base GAM is written as

$$\begin{aligned} E[\log(\text{premium})] = {} & s(\text{age}) + s(\text{vitaminD}) + s(\text{Hb}) + s(\text{calcium}) + s(\text{crr}) \\ & + s(\text{alb}) + s(\text{phos}) + s(\text{BMDsp}) + s(\text{BMDnof}) \end{aligned}$$

where $s(.)$ denotes the smoother for a particular covariate. These smoothers are estimated in a non-parametric fashion.

Once the significant covariates have been determined, the effect of main variable FBS is added to the base model to determine the premium of diabetic patients. The final model has the form

$$\begin{aligned} E[\log(\textit{premium})] = {} & \text{FBS} + s(\text{age}) + s(\text{vitaminD}) + s(\text{Hb}) + s(\text{calcium}) + s(\text{crr}) \\ & + s(\text{alb}) + s(\text{phos}) + s(\text{BMDsp}) + s(\text{BMDnof}) \end{aligned}$$

The predicted amount of premium is based on the main risk factor (FBS) and smoothening the effect of covariates in the model. For carrying out the analysis, R software is used. The smoothers have a single smoothing parameter. Generalized Cross Validation function that approximates the expected prediction error, is used for choosing the smoothing parameters. The model selected by GCV function has the best prediction ability and the degrees of freedom are also specified for each individual smoothing component [8]. As an initial step for analyzing the predictor-response relationship, we find out the degrees of freedom for different covariates that minimize the Generalized Cross Validation (GCV) score of the model. There are many other smoothers available in literature, for example, bin smoothers, kernel smoothers, regression splines and cubic smoothing splines etc. The particular method used for smoothing is determined by smooth.spline function in R which fits a cubic smoothing spline to the data. This command also helps in determining the degrees of freedom for covariates. The degrees of freedom determined for each covariate are shown in Table 2.

Table 2. Degree of freedom for covariates.

Covariate	degree of freedom
age	5
Hb	3
crr	3
phos	4
calcium	1
alb	3
vitaminD	1
BMDsp	1
BMDnof	1

3 GAM Fitted to Real Life Dataset

In this section, we first determine the most significant covariates by fitting the base model without FBS using the log link function and quasi Poisson family. Values of covariates (age, vitaminD, Hb, calcium, crr, alb, phos, BMDsp, BMDnof), main risk factor FBS and dependent variable (premium) of 134 patients are used to fit GAM. The output shows the p-values that are used for determining the significant covariates.
The output of analysis is given below:
Family: quasipoisson
Link function: log
Formula: Base Model

$$\begin{aligned}\text{premium} = {} & s(\text{age}, k = 5+1, fx = F, bs = \text{“}cr\text{”}) + s(\text{hb}, k = 3+1, fx = F, bs = \text{“}cr\text{”}) \\ & + s(\text{crr}, k = 3+1, fx = F, bs = \text{“}cr\text{”}) + s(\text{phos}, k = 4+1, fx = F, bs = \text{“}cr\text{”}) \\ & + s(\text{alb}, k = 3+1, fx = F, bs = \text{“}cr\text{”}) + s(\text{vitaminD}, k = 1+1, fx = F, bs = \text{“}cr\text{”}) \\ & + s(\text{BMDsp}, k = 1+1, fx = F, bs = \text{“}cr\text{”}) + s(\text{BMDnof}, k = 1+1, fx = F, bs = \text{“}cr\text{”}) \\ & + s(\text{calcium}, k = 1+1, fx = F, bs = \text{“}cr\text{”})\end{aligned}$$

Parametric coefficients:

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	11.27	$4.654e-05$	242104	< 2e-16***

Significant codes: *** 0.01 ** 0.05 * 0.1
Approximate significance of smooth terms:

Variable	edf	Ref.df	F	p-value
s(age)	4.993	5.000	8.842e+05	< 2e-16 ***
s(hb)	2.554	2.872	3.301e+00	0.02491**
s(crr)	1.000	1.000	1.000e-01	0.75199
s(phos)	2.754	1.792	2.397e+00	0.09953*
s(alb)	1.000	2.956	5.299e+00	0.00198 ***
s(vitaminD)	1.000	1.000	2.853e+00	0.09385*
s(BMDsp)	1.000	1.000	2.8001e-02	0.86668
s(BMDnof)	1.647	1.874	8.960e-01	0.39777
s(calcium)	1.000	1.000	4.000e-02	0.84146

R-sq.(adj) = 1 Deviance explained = 100%.

From the above output, it is evident that age, Hb, phos, alb and vitaminD are the significant covariates and age is the most significant covariate since the corresponding p-value is the lowest. The reason for adjusted R^2, being 1 for this model may be that age, the most significant variable, is influencing the premium to a large extent and hence, may suppress the effect of other covariates. This is further substantiated after fitting GAM by considering age as the only covariate. The results for this model are given below:
Family: quasipoisson
Link function: log
Formula:

$$\text{premium} = s(\text{age}, k = 5+1, fx = F, bs = \text{”}cr\text{”})$$

Parametric coefficients:

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	11.27	$7.597e-07$	14830495	$< 2e-16$ ***

Significant codes: *** 0.01 ** 0.05 * 0.1
Approximate significance of smooth terms:

Variable	edf	Ref.df	F	p-value
s(age)	19.7	19.98	1.059e+09	< 2e-16 ***

R-sq.(adj) = 1 Deviance explained = 100%.

Since with age as only covariate, $R^2(\text{adj}) = 1$, therefore the presence of age in the base model may suppress the effect of other covariates on premium. Thus, the GAM is again fitted using R software by excluding age in order to determine covariates that contribute significantly and the output is included below.

Family: quasipoisson
Link function: log
Formula:

$$\begin{aligned}\text{premium} = {} & s(\text{Hb}, k = 3+1, fx = F, bs = "cr") + s(\text{crr}, k = 3+1, fx = F, bs = "cr") \\ & + s(\text{phos}, k = 4+1, fx = F, bs = "cr") + s(\text{vitaminD}, k = 1+1, fx = F, bs = "cr") \\ & + s(\text{alb}, k = 3+1, fx = F, bs = "cr") + s(\text{BMDsp}, k = 1+1, fx = F, bs = "cr") \\ & + s(\text{BMDnof}, k = 1+1, fx = F, bs = "cr") + s(\text{calcium}, k = 1+1, fx = F, bs = "cr")\end{aligned}$$

Parametric coefficients:

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	11.270761	0.008446	1334	$< 2e-16$ ***

Approximate significance of smooth terms:

Variable	edf	Ref.df	F	p-value
s(Hb)	1.000	1.000	0.419	0.51892
s(crr)	1.000	1.000	0.488	0.48626
s(phos)	3.907	3.994	3.054	0.01949**
s(alb)	2.294	2.681	4.968	0.00428***
s(vitaminD)	1.333	1.555	0.706	0.44501
s(BMDsp)	1.000	1.000	3.439	0.06610*
s(BMDnof)	1.633	1.865	0.913	0.39058
s(calcium)	1.000	1.000	0.177	0.67462

R-sq.(adj) = 0.189 Deviance explained = 27.3%
GCV score = 837.93 Scale est. = 749.33 n = 134.

On the basis of p-values, it is observed that the covariates phos, alb and BMDsp are at 10% level of significance. Hence the output of the fitting of the base GAM (after excluding age) leaves us with these covariates and these are retained for further analysis. GAM is now fitted on the data set of 134 patients including the main risk variable FBS and the significant covariates obtained in base model. The final output is presented below:

Family: quasipoisson
Link function: log
Formula:
premium = FBS + s(phos, k = 4 + 1, fx = F, bs = "cr") + s(alb, k = 3 + 1, fx = F, bs = "cr") + s(BMDsp, k = 1 + 1, fx = F, bs = "cr")

Parametric coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	$1.122e+01$	$2.464e-02$	455.310	$< 2e-16$***
FBS	$4.256e-04$	$1.962e-04$	2.169	0.0320*

Approximate significance of smooth terms:

Variable	edf	Ref.df	F	p-value
s(phos)	3.908	3.994	3.089	0.01832 *
s(alb)	2.174	2.578	5.124	0.00404**
s(BMDsp)	1.600	1.839	7.246	0.00165**

R-sq.(adj) = 0.223 Deviance explained = 27.8%.

From the above output, it is evident that phos, alb and BMDsp are significant covariates along with main risk factor FBS. Hence, we can conclude that while determining the premium for diabetic insured, the insurer should take into account the FBS, phos, alb and BMDsp along with age.

Figure 1 displays the predicted plot for premium of all patients in the data set.

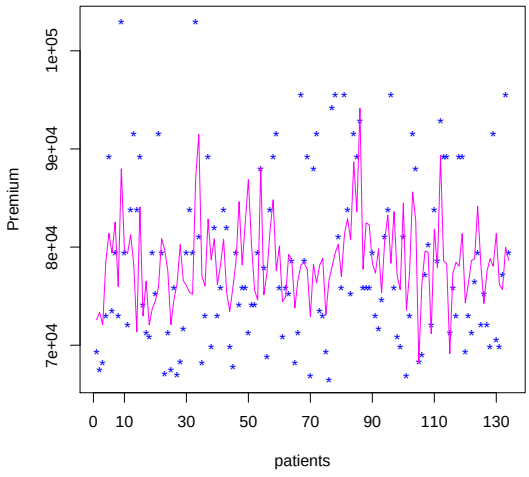

Figure 1. Predicted plot for insurance premium

From Figure 1, it is seen that GAM helps us to predict the premium amount for diabetic patients. As a result of a change in FBS level, the new predicted premium can be found as the product of Relative risk for FBS and predicted premium where Relative risk for FBS = exp(estimate of coefficient of FBS * (increase(+)/decrease(-) mg/dl (milligram/decilitre) in FBS level)).

From the final GAM output, the estimated coefficient of FBS level is 0.0004256. For example, if Rs. 78000/- is the predicted premium for a diabetic patient and there is an increase (a decrease) of 10mg/dl in FBS level, then new predicted premium is exp(+(-)0.0004256*10)*78000 which is equal to Rs. 78332.67 (Rs. 77668.73).

Hence, there is an increase (a decrease) of Rs. 332.67 (Rs. 331.26) in the premium amount with an increase (a decrease) of 10mg/dl in FBS level. In the next section, we fit a logistic GAM using the backfitting and local scoring algorithms and study those variables that are significant in modeling diabetic status of an individual using SAS software. These algorithms are general and iterative for fitting a GAM.

4 Predictive Model for Diabetic Patients Using Logistic GAM

Scatterplot matrix for all the covariates in the datset, obtained by using SAS and given in Figure 2 shows the linear relationship between variables under consideration. In this figure, depvar (dependent variable) represents the FBS status of an individual. It takes value 1 or 0 depending upon whether an individual is diabetic or not. Circles and plus symbols distinguish patients with positive and negative diabetes tests respectively. It is seen from the figure that BMDsp and BMDnof are positively associated.

The first step in analyzing the predictor-response relationship is to fit a GAM with all the predictors/covariates. We use PROC GAM to fit a logistic generalized additive model with binary dependent variable diabetes against all the covariates.

By default, the smoothing parameter for each B-spline term is chosen to yield four degrees of freedom where one is taken up by the linear portion and three by the nonlinear spline portion of the fit. Each term is fitted using a B-spline smoother where B-spline is a method implemented by GAM procedure for univariate smoothing components. The output from PROC GAM is displayed in Tables 3-5. Table 3 provides the analytical information about the fitted model, including parameter estimates for the linear portion of the model. Fit summary for smoothing components is given in Table 4 and the approximate analysis of deviance table is given in Table 5.

In the last column of Table 5, * (asterisk) represents those values that are significant at 10% level.

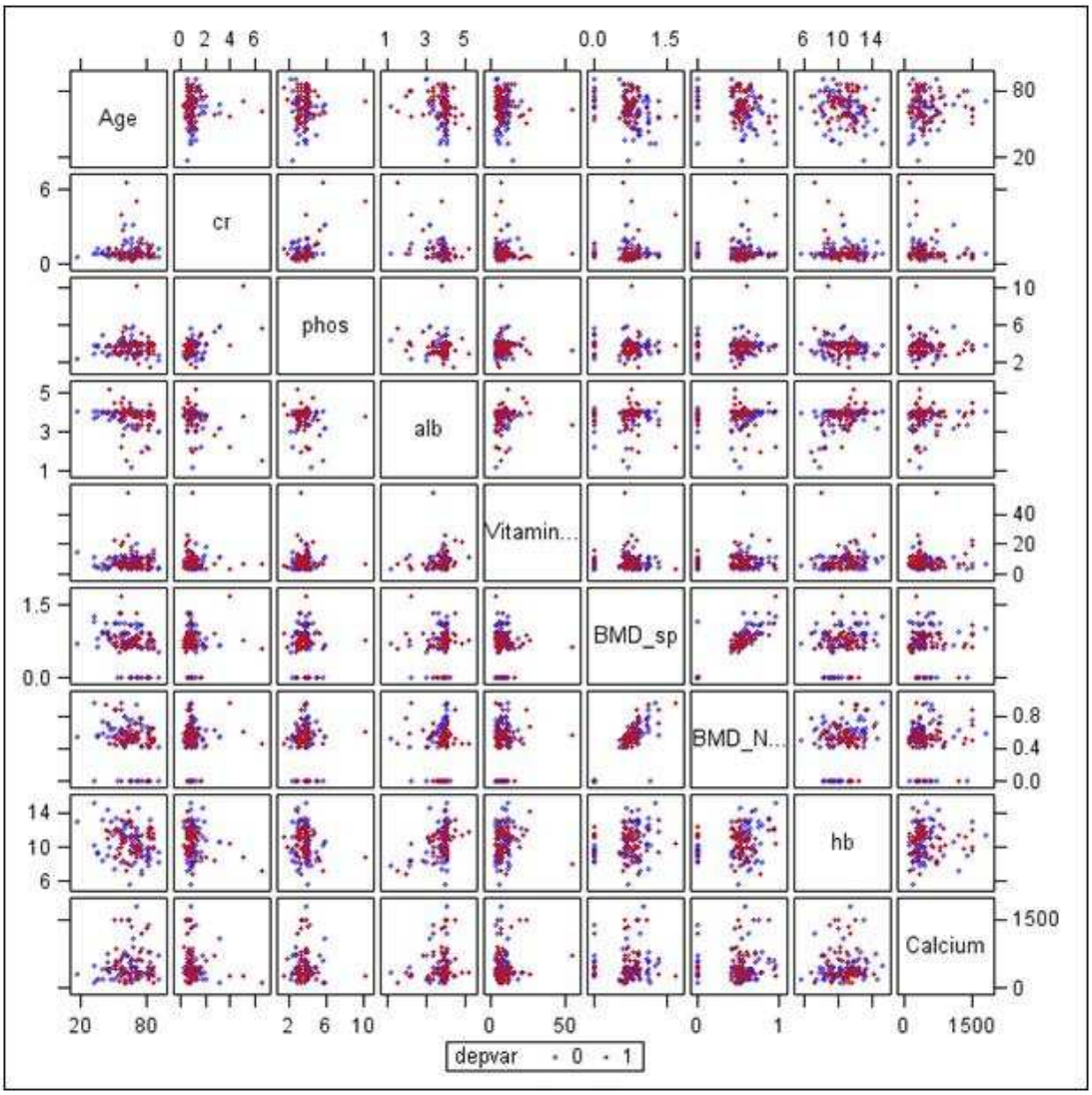

Figure 2. Scatterplot matrix of covariates in the diabetes data set

Table 3. Regression model analysis parameter estimates.

Parameter	Parameter Estimate	Standard Error	t Value	Pr $> \|t\|$
Intercept	-0.58774	2.81982	-0.21	0.8353
Linear(age)	0.02118	0.01845	1.15	0.2537
Linear(crr)	-0.57580	0.43641	-1.32	0.1901
Linear(phos)	-0.06296	0.30761	-0.20	0.8383
Linear(alb)	-0.47664	0.47719	-1.00	0.3204
Linear(vitaminD)	0.00892	0.05576	0.16	0.8732
Linear(BMDsp)	-0.18408	1.63243	-0.11	0.9104
Linear(BMDnof)	0.89532	2.32833	0.38	0.7014
Linear(Hb)	0.11818	0.15405	0.77	0.4448
Linear(calcium)	0.00008	0.00063	0.13	0.8988

Table 4. Smoothing model analysis fit summary for smoothing components.

Component	Smoothing Parameter	DF	GCV	Number of unique obs
Spline(age)	0.999853	3.000000	78.799930	40
Spline(crr)	0.998329	3.000000	31.066097	34
Spline(phos)	0.998972	3.000000	31.380623	41
Spline(alb)	0.995309	3.000000	21.948220	22
Spline(vitaminD)	1.000000	3.000000	80176	105
Spline(BMDsp)	0.999972	3.000000	42.331083	90
Spline(BMDnof)	1.000000	3.000000	80995	83
Spline(Hb)	1.000000	3.000000	115763	55
Spline(calcium)	0.989844	3.000000	3.511094	24

Table 5. Smoothing model analysis, analysis of deviance.

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(age)	3.00000	5.483135	5.4831	0.1397
Spline(crr)	3.00000	7.209273	7.2093	0.0655*
Spline(phos)	3.00000	3.522797	3.5228	0.3178
Spline(alb)	3.00000	7.398008	7.3980	0.0602*
Spline(vitaminD)	3.00000	5.292984	5.2930	0.1516
Spline(BMDsp)	3.00000	3.542647	3.5426	0.3153
Spline(BMDnof)	3.00000	1.005551	1.0056	0.7999
Spline(Hb)	3.00000	11.799933	11.7999	0.0081*
Spline(calcium)	3.00000	4.040213	4.0402	0.2572

For each smoothing effect in the model, Table 5 provides χ^2-values comparing the deviance between the full model and model without the corresponding covariate. These values indicate that the all smoothing terms except crr, alb and Hb (at 10% level of significance) are insignificant.

Without any prior knowledge, it is hard to specify appropriate values for the degrees of freedom. An alternative is to use the GCV option to choose smoothing parameters.

Figure 3 shows two panels of smoothing component plot produced by using GAM with the GCV option specified in PROC GAM procedure in SAS.

The use of common vertical axis in Figure 3 enables us to see that all variables exhibit non-linear trends.

The analytical information for significant predictors using GAM are listed in Tables 6 and 7. The Analysis of Deviance results in Table 8 indicate that there are significant non-linear contributions from three variables crr, alb and Hb. These nonlinearities are also observed in Figure 4.

Table 6. Regression model analysis parameter estimates of significant covariates.

Parameter	Parameter Estimate	Standard Error	t value	Pr > $\lvert t \rvert$
Intercept	1.42536	1.82332	0.78	0.4359
Linear(alb)	-0.41074	0.43726	-0.94	0.3494
Linear(crr)	-0.47708	0.36228	-1.32	0.1904
Linear(Hb)	0.04558	0.13396	0.34	0.7342

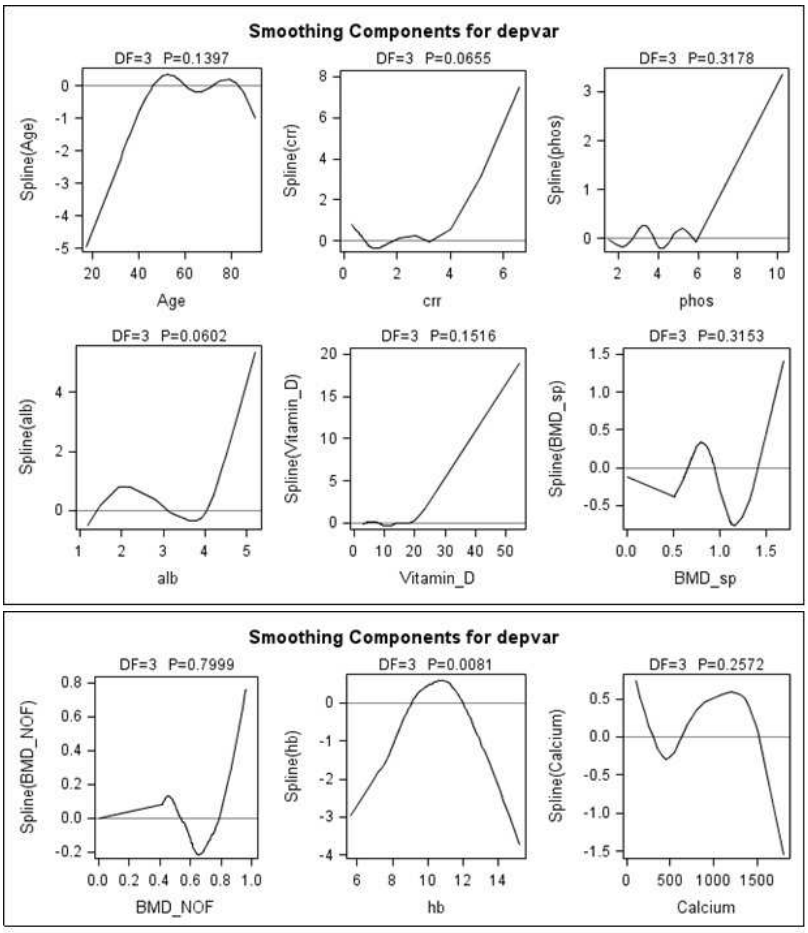

Figure 3. Smoothing component plot

Table 7. Smoothing model analysis, fit summary for smoothing components of significant covariates.

Component	Smoothing Parameter	DF	GCV	Num unique obs
Spline(alb)	0.999568	3.000000	319.765397	22
Spline(crr)	0.999722	3.000000	288.061015	34
Spline(Hb)	0.999981	3.000000	209.624887	55

Table 8. Smoothing model analysis, analysis of deviance of significant covariates.

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(alb)	3.00000	10.087506	10.0875	0.0178*
Spline(crr)	3.00000	12.781607	12.7816	0.0051*
Spline(Hb)	3.00000	10.543108	10.5431	0.0145*

In the last column of Table 8, * (asterisk) represents those values that are significant at 5% level.

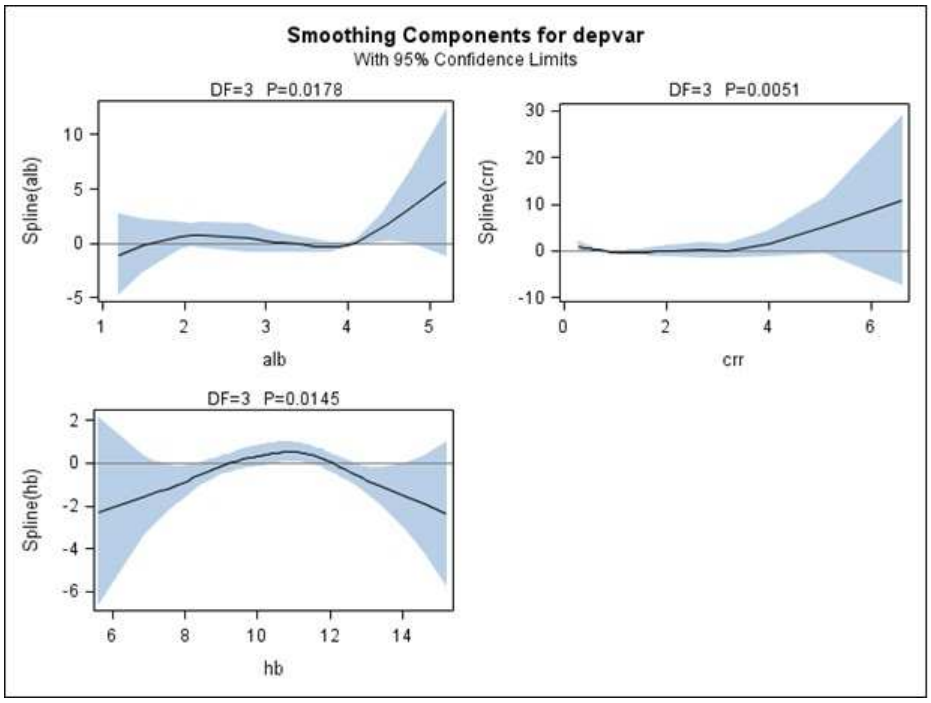

Figure 4. Smoothing component plots with 95% confidence limits

Figure 4 shows the 95% confidence intervals for alb, crr and Hb as depicted by shaded area. The general trends in the smoothing component plot for alb and Hb suggest possible quadratic dependence on this variable and the trend in crr resembles a polynomial. Linear and quadratic transformations make sense physically for these variables and can provide insights into their characteristics. After discovering an appropriate form of the dependence of diabetes on the independent variable, one can use PROC GENMOD procedure to assess these transformations in a parametric model.

In Table 9, the partial output of Logistic Generalized Additive Model with predicted Logit and predicted probability for specific alb, crr and Hb level and diabetic status is presented. The sixth and seventh columns of Table 9 represent the predicted Logit value and predicted probability of suffering from diabetes. The predicted Logit is not of direct interest and inverse of link function is used to convert Predicted Logit to predicted probability by using the relationship

$$\text{Predicted Probability} = \frac{1}{1 + e^{-\text{Predicted Logit}}}$$

For example, if the fifth observation (patient) has Hb 10.5, crr 3.2, alb 3.2 and is non-diabetic, then according to Logistic GAM fitted to the given data, the predicted probability of suffering from diabetes is 0.60300.

5 Conclusions

GAM is fitted using R and SAS softwares for evaluating the premium of the diabetic patients in presence of covariates (age, Hb, vitaminD, albumin etc.) by considering FBS as the main risk factor. Most of the insurance companies consider only age and do not provide any insurance cover, particularly, to diabetic patients. This model will help the insurance companies to take a decision about fixing premium amount for diabetic patients and also taking care of significant factors affecting the status of diabetes. The significant

Table 9. Partial output of logistic generalized additive model.

Obs.	Hb	crr	alb	Diabetes	Pred Logit	Pred Probability
1	11.3	0.9	3.8	0	0.46322	0.61377
2	11.2	0.8	3.9	0	0.52960	0.62939
3	12.2	0.4	4.1	0	0.69681	0.66748
4	9.9	0.8	4.1	0	0.53559	0.63078
5	10.5	3.2	3.2	0	0.41799	0.60300
6	11.2	0.8	4	1	0.55421	0.63511
7	10.1	0.6	3.4	1	0.70408	0.66909
8	7.3	6.6	1.5	1	0.99927	0.73091
9	10.3	0.8	4	1	0.52347	0.62795
10	8.4	2.7	2.8	1	0.35801	0.58855

factors responsible for causing diabetes have also been identified. With the help of Logistic generalized additive model, prediction probability of being diabetic, based on alb, crr and Hb is given. The techniques discussed in this study can also be applied with some modifications to other types of insurance contracts and also for those individuals who suffer from cancer or other chronic diseases.

Acknowledgments. The first author is grateful to University Grants Commission, Government of India, for providing financial support for this work.

References

1. T. J. Hastie and R. J. Tibshirani, "Generalized Additive Models", *Statistical Science*, vol. 3, pp. 297-318, 1986.
2. T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Chapman and Hall, USA,1990.
3. J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models", *Journal of the Royal Statistical Society*, Series A, vol .135, pp. 370-384, 1972.
4. P. J. Green, " Iteratively reweighted least square for maximum likelihood estimation and some robust and resistant alternatives (with discussion)", *Journal of Royal Statistical Society*, Series B, vol. 46, pp. 149-192, 1984.
5. P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Second Edition, Chapman and Hall, UK, 1989.
6. S. N. Wood, *Generalized Additive Models: An Introduction with R*, Texts in Statistical Science, Chapman and Hall, USA, 2006.
7. C. J. Stone, "Additive Regression and Other Nonparametric Models", *Annals of Statistics*, vol. 13, pp. 689 -705, 1985.
8. P. Craven and G. Wahba, "Smoothing Noisy Data with Spline Functions :Estimating the Correct Degree of Smoothing by the Method of Generalized Cross Validation", *Numerische Mathematik*, vol. 31, pp. 377-403, 1979.