TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
1-AC/2013

DAMIAN KRUSZEWSKI*

NONPARAMETRIC MODELING OF MEDICAL SCHEME DATA

MODELOWANIE NIEPARAMETRYCZNE DANYCH MEDYCZNYCH

Abstract

The goal of this paper is to apply Generalized Additive Models to medical scheme data. The flexibility of the nonparametric approach is demonstrated based on a real-life empirical example that seeks to model hypertension and the interplay of determinants, such as physiological measurements, medical attributes, demographic and socioeconomic characteristics in predicting blood pressure. The assessment of nonlinear patterns in the response-predictor relationship and the strength of this association are investigated. The extended Generalized Additive Models allow for modeling not only location and scale, but also other distribution parameters, such as kurtosis and skewness.

Keywords: Generalized Additive Models, smoothing, hypertension, systolic/diastolic blood pressure

Streszczenie

Celem niniejszego artykułu jest aplikacja uogólnionych modeli addytywnych do danych medycznych. Elastyczność nieparametrycznych rozwiązań przedstawiono na przykładzie modelowania zmiennych determinujących poziom nadciśnienia tętniczego krwi, takich jak atrybuty zdrowotne, fizjologiczne, demograficzne czy charakterystyki społeczno-ekonomiczne. W artykule zbadano nieliniowe zależności (oraz ich siłę) pomiędzy zmiennymi objaśniającymi a nadciśnieniem tętniczym krwi. Rozszerzona wersja modelu pozwala wyznaczyć nie tylko parametry skali i położenia, lecz również inne parametry charakterystyczne rozkładu, takie jak kurtoza i skośność.

Słowa kluczowe: Uogólnione Modele Addytywne, wygładzanie, nadciśnienie tętnicze krwi, ciśnienie skurczone/rozkurczowe krwi

* Damian Kruszewski, M.Sc., PAREXEL International; Ph.D. Studies, Systems Research Institute, Polish Academy of Sciences.

1. Introduction

There are loads of methods and techniques for nonparametric/semiparametric regression, such as Locally Weighted Regression [2], Regression Splines, Smoothing Splines, B-Splines [16], P-Splines [4], etc. All of them are aimed at one problem – to make a precise prediction. Compared with standard parametric methods such as Linear/Binary/Logistic Regression Models or Generalized Linear Models (GLM), the methodology behind nonparametric modeling relaxes the assumption of linearity in the response-predictor relationship. It enables to uncover structural behavior of the response with the independent variables that may otherwise be missed. The notion of exploring data nonparametrically has been proven to be successful in the statistical modeling. Unfortunately, this success sometimes is accompanied with a weak interpretability and greater variance for greater dimensionality. Proposed by Hastie and Tibshirani [9], Generalized Additive Models (GAM) allow for multidimensional data and provide the ability to detect the nonlinear associations without any damaging repercussion on interpretability.

The aim of this paper is to use Generalized Additive Models (GAM) to predict hypertension by multiple independent variables whose effect is modeled nonparametrically. The presented results demonstrate the importance of nonparametric solutions, especially in the context of a real-life data set. The study on the example of hypertension shows that Generalized Additive Models (GAM) provide flexible statistical methods for identification of nonlinear regression effects and complex shapes in the relationship between the response and the predictors which are missed by standard parametric solutions. The models built for Systolic/Diastolic Blood Pressure account for scale, location, kurtosis and skewness of the continuous response distribution.

There is a plethora of studies trying to find and explain the factors influencing blood pressure. Unfortunately, most of them follow parametric assumptions [18], other allow for semiparametric inferences but strictly restricted to pre-specified response distribution belonging to the exponential family and thus, disregarding kurtotic or skewed distributions [8]. As presented on the example of Systolic/Diastolic Blood Pressure, the analyzed response variables do not follow exponential family features.

2. Rationale

Modeling hypertension with Generalized Additive Models (GAM) has by its nature interdisciplinary scope. Broadening the spectrum of medical applications of Generalized Additive Models (GAM) contributes to both IT and medicine. Hypertension is a chronic health condition prevalent in most developed nations. Its prevalence in the western population exceeds 20%. Untreated high blood pressure is a major risk factor for coronary heart disease, cardiovascular disease, stroke or diabetes. Thus, it is of crucial importance to develop models identifying potential markers for its prediction. Better knowledge of the blood pressure drivers supports the decision making process concerning hypertension diagnosis and its treatment.

The rationale behind incorporating Generalized Additive Models (GAM) are:

a) relaxing the assumptions of parametric models: 1) linear form of the relationship between response and predictors, 2) diagnostic checking of the residuals (normality and independence),

b) enabling the use of multidimensional data,
c) reasonably easy interpretation.

Generalized Additive Models (GAM) have greater flexibility than their parametric counterparts. Traditional methods, although attractively simple, often fail in many applied settings. In real-life, effects are generally not linear.

The rationale of relaxing the parametric assumptions of linearity and normality (a) enables to explore the data visually and uncovers structural behavior that may be otherwise missed. Of note is the fact, that these properties of Generalized Additive Models (GAM) are shared by other nonparametric solutions. What make them distinctive from other nonparametric models are their properties of multidimensionality and interpretability (b and c). In light of the necessity of including large number of explanatory variables in the real-life applications, the practical capabilities of the most commonly used nonparametric regression methods, such as Thin-Plate Smoothing or Local Regression Methods [2] are significantly restricted. In such circumstances, the sparseness of data results in the unacceptably large variance of estimates ("the curse of dimensionality").

The drawbacks of both parametric (linearity) and standard nonparametric solutions (multidimensionality and interpretability) are overcome by Generalized Additive Models (GAM). Their methodology allows for estimating the additive terms individually using a univariate smoother – each input is considered independently. This addresses the issue associated with "the curse of dimensionality". Additionally, individual term's estimate directly explains the relative contribution to the response changes and thus, Generalized Additive Models (GAM) are among the most interpretable statistical models.

3. Description of solutions

Generalized Additive Models (GAM) were first proposed by Hastie and Tibshirani [9]. Their fit allows to combine:

– the flexibilities of Generalized Linear Models (GLM) – an arbitrary function of dependent variable,
– the additive assumptions that enable to explore the data nonparametrically.

Generalized Additive Models (GAM) extend Linear Models (LM) and Generalized Linear Models (GLM) to include smooth functions of explanatory variables. They are an important step forward in the generalization of Generalized Linear Models (GLM). Generalized Additive Models (GAM) do not require any transformations of the predictors to improve the fit. The different regression models might be envisioned as being nested within each other, with linear regression being the most limiting case, and Generalized Additive Models (GAM) the most general. They combine the abilities to explore the data nonparametrically simultaneously with the distributional flexibilities of Generalized Linear Models (GLM). Instead of having a single estimation coefficient for each of the predictors, Generalized Additive Models (GAM) use an arbitrary nonparametric function to approximate the association between each of the predictors and the response.

The only underlying assumption made is that the nonparametric functions are additive and that the components are smooth. Generalized Additive Models (GAM), like Generalized Linear Models (GLM), apply a monotonic link function to establish a relationship (link) between the mean of the response variable and a "smoothed" function of the explanatory

variables. They are constructed through summing up all the functions which fit the data locally. The final model closely represents the behavior of the data (data driven approach). However, apart from the nature of the response-predictor relationship, the probability distribution of the response must still be specified. In this sense, Generalized Additive Models (GAM) are more aptly referred to as semiparametric models.

3.1. Fundamentals of Generalized Additive Models (GAM)

Generalized Additive Models (GAM) combine Generalized Linear Models (GLM) and Additive Models (AM):

Generalized Linear Models (GLM) extend the response distribution of the linear model into the exponential family. Providing that Y is a response random variable and $X_j, ..., X_p$ are explanatory variables, a standard linear regression model might be expressed as $E(Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$. Assuming that $E(Y) = \mu$ and $\eta = g(\mu)$ where $g(\bullet)$ is a smooth monotonic differentiable (up to third order) link function, the response-predictor relationship in Generalized Linear Models (GLM) is defined by $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$. The link function $g(.)$ describes how the expected value of Y is related to linear predictor $E(Y) = \mu$. Because the link function is a monotonic and invertible function, the mean can be expressed as the inversely linked linear predictor: $E(Y) \equiv \mu = g^{-1}(\eta)$ where $g^{-1}(.)$ is so called inverse link function. The form $\eta = g(\mu)$ emphasizes that Generalized Linear Models (GLM) use transformations of the mean (no transformation of the data). The second form, i.e. $\mu = g^{-1}(\eta)$ shows how predictions of the mean can be obtained following the estimation of η. The most commonly employed link functions are Normal, Exponential, Gamma, Inverse Gamma, Poisson and Binomial. For instance, for binary data, a common link function is the logit link: $g(t) = \log[t/(1 - t)]$. The mean function of Generalized Linear Model (GLM) with the assumed logit link function and one predictor can be written as: $\log\left\{\frac{\mu}{1-\mu}\right\} = \beta_0 + \beta_1 x$ and thus, $\mu = \frac{1}{1+\exp\{-\beta_0 - \beta_1 x\}}$. This is known as a logistic regression model. The response variable in Generalized Linear Models (GLM) is assumed to be a member of exponential family.

– Additive Models (AM) extend the parametric form of predictors in the linear model to nonparametric forms. Additive Model (AM) is defined as: $E(Y) = s_0 + \sum_{j=1}^{p} s_j(X_j)$ where the smoothers s_i, $i = 1, ..., p$ are smoothing splines. Please note that smooth functions have to be constrained to have zero mean.

Combining Generalized Linear Models (GLM) and nonparametric Additive Models (AM), Generalized Additive Models (GAM) might be defined as:

$$\eta = s_0 + \sum_{j=1}^{p} s_j(X_j). \tag{1}$$

where the response variable has a probability density from the exponential family. Generalized Additive Models (GAM) extend Generalized Linear Models (GLM) by replacing the form $\beta_0 + \sum_{j=1}^{p} \beta_j X_j$ with the additive form $s_0 + \sum_{j=1}^{p} s_j(X_j)$.

The form and the nature of Generalized Additive Models (GAM) are dependent on the backfitting algorithm, the local scoring method, the specified smoothing parameters and the degrees of freedom (DF) used for their computation. All of these parameters are thoroughly discussed in the literature [19]. In this paper, only a brief summary is provided:

- **Backfitting algorithm:** The backfitting and the local scoring form an interactive method to estimate the smoothers s_i, $i = 1, ..., p$. The backfitting algorithm is an algorithm that enables to fit Additive Models (AM). It might be used with different smoothers such as univariate or bivariate splines. The iterative mechanism permits to estimate each of the smoothing functions $s_k(\ )$, given estimates $\{\hat{s}_j(\ ),\ j \neq k\}$ [6].
- **Selection of smoothing parameters:** Very different types of smoothing functions could be specified in Generalized Additive Models (GAM): Cubic Smoothing Spline, Local Regression, Thin-Plate Smoothing Spline, etc. A smoother is an operator for summarizing the trend and the variability of a response measurement Y as a nonparametric function of explanatory measurements $X_1, ..., X_p$. Smoothing methodology offers a way by which nonlinear and nonparametric relationships can be handled without the restrictions of parametric models. In Generalized Additive Models (GAM), each smoother has a single unique smoothing parameter. The most commonly used methods for the selection of smoothing parameters are Cross Validation (CV) function and Generalized Cross Validation (GCV) technique. For more details, please refer to Wahba [19].

4. Empirical results

4.1. Description of data

The data set used in this paper is obtained from the National Health & Nutrition Examination Survey (NHANES). NHANES is an ongoing program designed to assess the health status of patients in the United States. The NHANES collects, among others, demographic, health history and behavioral information. This paper uses blood pressure measurements and demographic characteristics data. Blood pressure measurements were assessed during physician examinations (taken in the mobile examination centers), whereas demographic characteristics were collected during personal interviews. For the analysis purposes, the data from 2003 to 2010 is pooled. Calculations are performed using SAS Base 9.2 and R[1].

[1] SAS = Statistical Analysis System (system software provided by SAS Institute Inc., 4GL language), R = programming software and language for statistical computing developed by Development Core Team (Robert Gentleman and Ross Ihaka).

The primary objective of this real-life empirical example is to investigate:

1) the usefulness and flexibility of Generalized Additive Models (GAM) for medical scheme data,
2) the dependence of hypertension on various medical factors,
3) the patterns of hypertension and the effects of the independent variables on the response,
4) the strength of the association between independent and dependent variables.

The response measurement is either continuous derived Mean Systolic/Diastolic Blood Pressure (referred to as Mean SBP/Mean DBP) or derived binary Hypertension/Borderline Hypertension level (Table 1). Mean SBP/Mean DBP is an average of three blood pressure readings taken during physician examinations. The binary Hypertension/Borderline Hypertension response is derived based on Mean SBP/Mean DBP and takes the value of 'Yes' (Hypertension/Borderline Hypertension) if Mean SBP is greater/equal 120 mmHg or Mean DBP greater/equal 80 mmHg and the value of 'No' (No Hypertension/Borderline Hypertension) otherwise. Its derivation is intended to account for both Mean SBP and Mean DBP. The histogram with an overlaid univariate kernel density estimate for continuous response variables is presented in Fig. 1. One-Way Frequencies for their binary counterparts are shown in Table 2.

Table 1

Response variables used for fitting Generalized Additive Models (GAM)

Variable Name	Variable Explanation
HYPERTFL	Hypertension/Borderline Hypertension (1=Yes, 0=No) [Derived]
mSBP	Mean Systolic Blood Pressure (mm Hg) [Derived]
mDBP	Mean Diastolic Blood Pressure (mm Hg) [Derived]

Table 2

One-Way Frequencies for binary Hypertension/Borderline Hypertension response

Hypertension/Borderline Hypertension (HYPERTFL)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
[1=Yes]	9905	44.61	9905	44.61
[0=No]	12299	55.39	22204	100.00

Table 3

Moments for Mean SBP and Mean DBP

Moments	Mean SBP	Mean DBP	Moments	Mean SBP	Mean DBP
N	22204	22204	**Sum Weights**	22204	22204
Mean	119.869	67.93787	**Sum Observations**	2661592	1508492
Std Deviation	17.521	11.81245	**Variance**	306.997	139.533
Skewness	1.136	0.04760	**Kurtosis**	**1.873**	**0.160**
Coeff Variation	14.617	17.38713	**Std Error Mean**	0.117	0.079

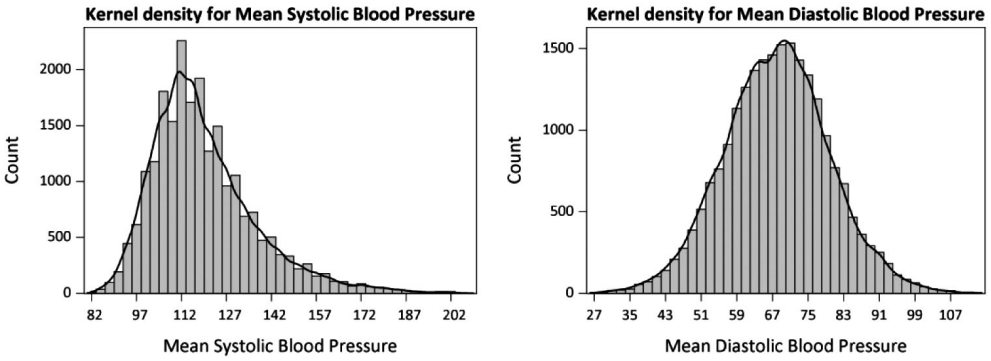


Fig. 1. Histogram for Mean SBP/Mean DBP

Rys. 1. Histogram dla Średniego Ciśnienia Skurczowego/Rozkurczowego Krwi

The predictor measurements are mainly variables representing physiological measurements, medical attributes as well as demographic and socioeconomic characteristics. Ratio of Income to Poverty compares a family's income with their appropriate poverty threshold[2]. The explanatory variables are listed in Table 4.

Table 4

Explanatory variables used for fitting Generalized Additive Models (GAM)

Variable Name	**Variable Explanation**
AGEYRS	Age at Screening (years)
GENDER	Gender (1=Male, 2=Female)
BMXBMI	Body Mass Index (kg/m²)
LBXSUA	Uric acid (mg/dL)
LBDHDDSI	HDL-cholesterol (mmol/L)
LBXSGTSI	Gamma Glutamyl Transferase (GGT) (U/L)
INDFMPIR	Family PIR (Ratio of Family Income to Poverty)
LBDSGLSI	Glucose (mmol/L)
LBDSCRSI	Creatinine (umol/L)

Model construction was preceded by outlier's detection. It involved removing extreme or missing values that might unduly influence the results of the analysis and potentially lead to incorrect conclusions. Extreme values were defined as values deviating from the expected range of 1st percentile and 99th percentile. Additionally, it is important to mention that in order to prepare the data for the analysis, it is recommended to apply the reduction of the high-dimensionality of the data set [12]. Reducing the number of variables under consideration

[2] Ratio of '1' means living right at the poverty line (income at 100% of poverty level), ratio above '1' indicates living above the official definition of poverty (i.e. a ratio of '1.5' means that income is 150% above the poverty threshold).

mitigates the effects of the curse of dimensionality and contributes to more accurate data analysis results [11].

Based on raw data and before building Generalized Additive Models (GAM), a positive association with Mean SBP/Mean DBP for Age and Body Mass Index (BMI) is noticed.

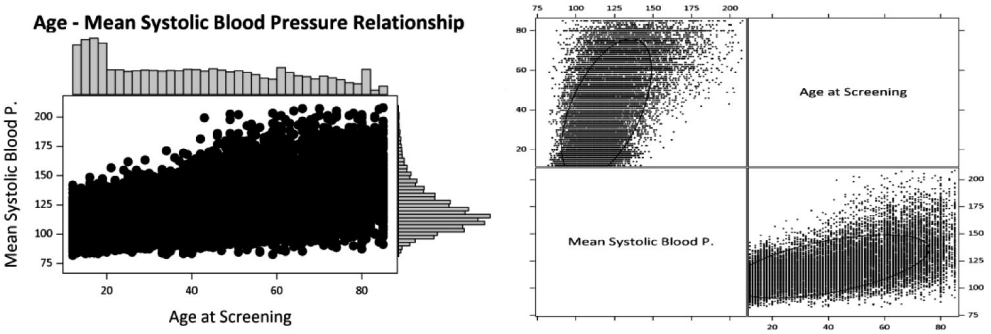


Fig. 2. Marginal Scatter Plot for Age – Mean SBP Relationship

Rys. 2. Brzegowy Wykres Rozrzutu dla Relacji Wiek – Średnie Ciśnienie Skurczowe Krwi

Knowing the positive response-predictor relationship for both explanatory variables, Age at Screening is classified into 4 Age Cohorts: 12–<23 years, 23–<40 years, 40–<57 years and 57–<85 years, and Body Mass Index (BMI) into 3 BMI Groups: 14–<25 kg/m², 25–<30 kg/m², 30–<45 kg/m². Please note that Body Mass Index (BMI) in the range of 25–<30 kg/m² could be an indicator of being overweight and Body Mass Index (BMI) > 30 kg/m² an indicator of being obese. Age Cohort, BMI Group and Gender are treated as the classification variables and constitute the parametric part of Generalized Additive Models (GAM).

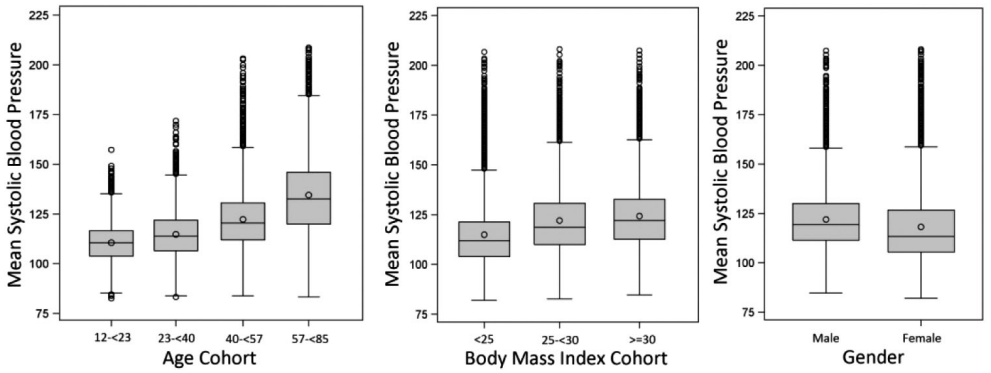


Fig. 3. Box plots of Mean SBP for classification variables

Rys. 3. Wykresy pudełkowe Średniego Ciśnienia Skurczowego dla poszczególnych zmiennych klasyfikujących

Alike the classification variables, the continuous predictors are intended to account for the nonparametric inferences. To examine the trends between the explanatory variables, all of them are put on one panel (Fig. 4). Only a weak correlation exists between explanatory variables. Thus, the impact of multicollinearity (concurvity) on parameter estimates is not a major issue. Appendix 1 presents bivariate kernel density estimates for selected explanatory variables, with contour and surface plots, in which density function is averaged across the observed data points to create a smooth approximation.

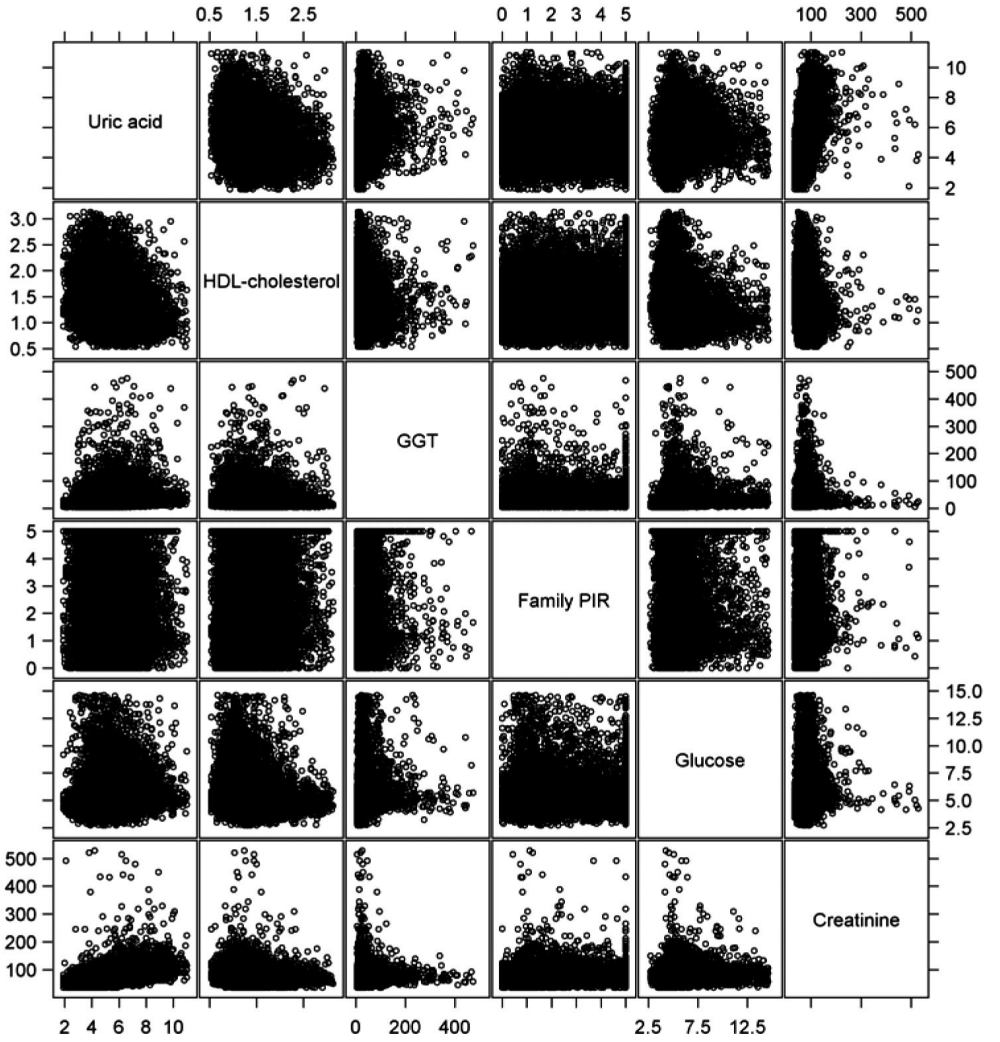


Fig. 4. Scatter Plot Matrices for Continuous Explanatory Variables

Rys. 4. Wykres Rozrzutu dla Ciągłych Zmiennych Objaśniających

4.2. Modeling binary response variable

As assessed by the scatter plot of explanatory variables (Fig. 4), there are striking features in the analyzed data set. It is really hard to determine from the plot whether the relationship between explanatory variables and the response is linear or not. Another issue has to do with the response distribution. Although histograms for continuous response variables, particularly for Mean DBP, seem to look pretty symmetric (Fig. 1), they are not of normal distribution and more importantly, do not follow exponential family features (as verified by Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling Tests).

To account for both Mean DBP and Mean SBP, as the starting point, Generalized Additive Model (GAM) for binary Hypertension/Borderline Hypertension is built (Table 5). It depends on additive predictors through a 'Logit' link function. An additive model with univariate cubic smoothing splines is requested for all continuous explanatory predictors (Uric acid, HDL-cholesterol, Gamma Glutamyl Transferase (GGT), Family PIR, Glucose and Creatinine).

Parameter Estimates for the parametric (linear) part of the model (Table 5) indicate high significance of the linear trends for each of the explanatory variables, with p-values much lower than the assumed significance level of 0.05. The Analysis of Deviance (Table 6) presents a χ^2 test comparing the deviance between the fully specified model (all explanatory variables with parametric and nonparametric part) and the model without the nonparametric component of a given variable (omitting nonlinearity). The nonparametric effects are concluded to be highly significant for each univariate smoothing splines introduced into the model.

Table 5

Parameter Estimates (Binary Regression, Link Function = Logit)

Parameter	Par. Estimate	Standard Error	t Value	Pr > |t|
Intercept	–3.28	0.132	–24.9	< 0.0001**
Age Cohort (23–<40 vs 12–<23 years)	0.60	0.048	12.4	< 0.0001**
Age Cohort (40–<57 vs 12–<23 years)	1.42	0.050	28.5	< 0.0001**
Age Cohort (57–<85 vs 12–<23 years)	2.26	0.052	43.5	< 0.0001**
Gender (Male vs Female)	0.35	0.040	8.9	< 0.0001**
BMI Group (25–<30 vs 14–<25 kg/m²)	0.25	0.040	6.1	< 0.0001**
BMI Group (30–<45 vs 14–<25 kg/m²)	0.57	0.044	12.9	< 0.0001**
LINEAR(Uric acid)	0.11	0.015	7.6	< 0.0001**
LINEAR(HDL-cholesterol)	0.20	0.045	4.5	< 0.0001**
LINEAR(Gamma Glutamyl Transferase)	< 0.01	< 0.001	9.8	< 0.0001**
LINEAR(Family PIR)	–0.03	0.001	–2.7	0.0066*
LINEAR(Glucose)	0.06	0.014	4.6	< 0.0001**
LINEAR(Creatinine)	< 0.01	0.001	4.1	< 0.0001**

Table 6

Smoothing Model Analysis – Analysis of Deviance

Cubic Spline (CS)	Generalized Cross Valid.	Sum of Squares	Chi-Square	Pr > ChiSq
CS(Uric acid)	0.538	19.550	19.550	0.0002**
CS(HDL-cholesterol)	0.394	11.222	11.222	0.0106*
CS(Gamma Glutamyl Transferase)	5.346	113.702	113.702	< 0.0001**
CS(Family PIR)	0.625	13.693	13.692	0.0034**
CS(Glucose)	3.085	27.985	27.985	< 0.0001**
CS(Creatinine)	4.717	26.990	26.990	< 0.0001**

To allow the visual judgment of the relative nonparametric effect sizes, a curvewise Bayesian confidence interval (standard-error band) to each smoothing component is used [9]. Smoothing Components Plot (Fig. 5) demonstrates the estimated smoothing spline functions with the linear effect subtracted out. It gives an idea where significant

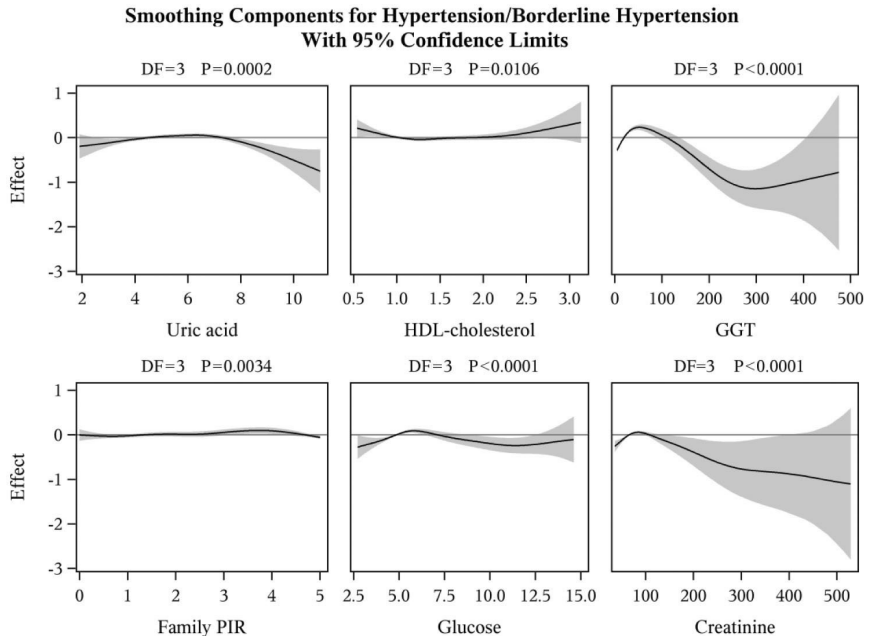


Fig. 5. Smoothing Components Plot for Hypertension/Borderline Hypertension

Rys. 5. Wykres Komponentów Wygładzania dla Nadciśnienia/Podwyższonego Ciśnienia Krwi

nonlinearities occur. The small p-values indicate that the data exhibits significant nonlinear structure. All variables are nonlinear predictors of Hypertension/Borderline Hypertension. They have a pretty pronounced complicated structure with a quadratic pattern for Uric acid and HDL-cholesterol, and even more curved pattern for Gamma Glutamyl Transferase (GGT) or Creatinine. It highlights the ability of Generalized Additive Models (GAM) to uncover nonlinear relationships and their potential in identifying patterns which are missed by standard parametric approaches.

The estimate of Generalized Additive Model (GAM) is just the sum of individual predictors' estimates plus a constant. Fig. 6 shows the partial prediction and the entire prediction effects of individual predictors (derived as the sum of the estimated linear terms – parametric part and the respective nonlinear partial predictions – nonparametric part). It further reveals the nature of the data and the overall shape of the relationship between the predictors and the response variable.

Concluding, the estimated Generalized Additive Model (GAM) for binary Hypertension/ Borderline Hypertension response reveals pronouncedly complex nonlinear patterns in the response-predictor relationships for all explanatory variables. For more in-depth investigation, the estimation of continuous Mean SBP/Mean DBP is needed. However, as verified by Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling Tests, both Mean SBP and Mean DBP do not belong to the class of exponential family distributions. Excess kurtosis is negative for both variables (Table 3) indicating platykurtic distributions. Thus, the estimation of continuous response variables requires going beyond exponential family distributions. This is accomplished by extending the standard Generalized Additive Models (GAM) to Generalized Additive Models (GAM) for Location Scale and Shape.

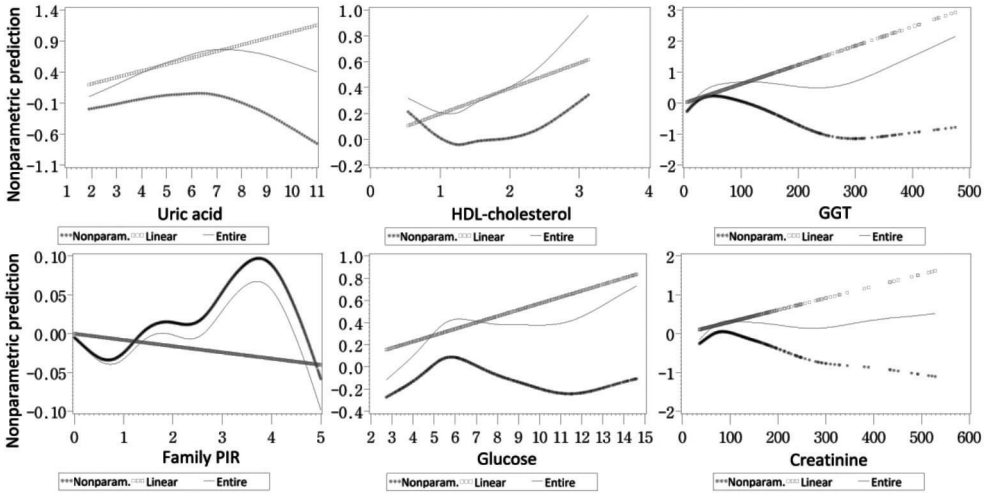


Fig. 6. Partial predictions for explanatory variables – predictors with and without linear terms

Rys. 6. Prognozy częściowe dla zmiennych objaśniających – predyktory z oraz bez składowych liniowych

4.3. Modeling continuous response variables

In Generalized Additive Models (GAM) for Location Scale and Shape the probability density function $f(y_i \mid \theta^i)$ is conditional on distribution parameter vector $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \upsilon_i, \tau_i)$ for $i = 1, 2, ..., n$. Each of the parameters $(\mu_i, \sigma_i, \upsilon_i, \tau_i)$ may be a function of the predictors. The first two distribution parameters μ_i and σ_i are referred to as location and scale distribution parameters, whereas υ_i and τ_i are referred to as shape distribution parameters (skewness and kurtosis). The formulation of Generalized Additive Models (GAM) for Location Scale and Shape goes as follows [14]:

$$g_k(\theta_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{2}$$

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = h_2(\mathbf{X}_1 \boldsymbol{\beta}_1) + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}) \tag{3}$$

$$g_2(\boldsymbol{\mu}) = \boldsymbol{\eta}_2 = h_2(\mathbf{X}_2\, \boldsymbol{\beta}_2) + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2}) \tag{4}$$

$$g_3(\boldsymbol{\mu}) = \boldsymbol{\eta}_3 = h_3(\mathbf{X}_3\, \boldsymbol{\beta}_3) + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3}) \tag{5}$$

$$g_4(\boldsymbol{\mu}) = \boldsymbol{\eta}_4 = h_4(\mathbf{X}_4\, \boldsymbol{\beta}_4) + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4}) \tag{6}$$

where:

$\mathbf{y}^T = (y_1, y_2, ..., y_n)$	– the vector of response variables,
θ_k for $k = 1, 2, 3, 4$	– the distribution parameter vector,
$\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\upsilon}$, $\boldsymbol{\tau}$ and $\boldsymbol{\eta}_k$	– vectors of length n,
$\boldsymbol{\beta}_k^T = (\beta_{1k}, \beta_{2k}, ...\beta_{J_k^{'}k})$	– a parameter vector of length $J_k^{'}$,
$\mathbf{X}_k$	– a known design matrix of order $n \times J_k^{'}$,
$\mathbf{x}_{jk}$ for $j = 1, 2, ..., J_k$	– vectors of length n,
$g_k(.)$	– monotonic link functions relating the distribution parameters $(\mu_i, \sigma_i, \upsilon_i, \tau_i)$ to explanatory variables,
h_{jk}	– an unknown smoothing function of explanatory variables x_{jk},
$\mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$	– a vector evaluating the function h_{jk} at $\mathbf{x}_{jk}$,
h_k	– nonlinear parametric function of explanatory variables.

If equation (2) does not include any of the additive terms in any of the distribution parameters ($J_k = 0$) then the model defined by (2) reduces to the nonlinear parametric model $g_k(\theta_k) = \eta_k = h_k(\mathbf{X}_k\, \boldsymbol{\beta}_k)$. If additionally $h_k(\mathbf{X}_k\, \boldsymbol{\beta}_k) = \mathbf{X}_k^T\, \boldsymbol{\beta}_k$ then model (2) reduces to the linear parametric one $g_k(\theta_k) = \eta_k = \mathbf{X}_k \boldsymbol{\beta}_k$.

Equation (2) allows for modeling the distribution parameters as linear/nonlinear parametric function ($h_k(\mathbf{X}_k\boldsymbol{\beta}_k)$) and nonparametric smooth function $\left(\sum_{j=1}^{J_k} h_{jk}(x_{jk})\right)$ of explanatory variables.

The form of the response distribution $f(y_i|\mu_i,\sigma_i,\upsilon_i,\tau_i)$ in Generalized Additive Models (GAM) for Location Scale and Shape may be very general. Table 7 compares the goodness of fit based on the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for 28 different continuous distributions fitted to continuous response variables. For details about these distributions, please refer to Johnson, Kotz and Kemp [10].

Table 7

Continuous distributions applied to Mean SBP and Mean DBP data – Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

	Mean SBP			Mean DBP		
	Df	**AIC**	**BIC**	**Df**	**AIC**	**BIC**
Box-Cox Power Exponential	16	**177773**	**177901**	16	167818	167946
Box-Cox-t	16	**177809**	**177936**	16	167825	167953
Inverse Gaussian	14	177895	178007	14	169376	169488
Zero Adjusted IG	15	177897	178017	15	169378	169498
Generalized Beta Type 2	16	177904	178032	16	167940	168068
Box-Cox Cole and Green	15	178230	178350	15	167880	168000
Generalized Gamma	15	178366	178486	15	167916	168036
Generalized Inverse Gaussian	15	178484	178603	15	168446	168566
Log Normal	14	178619	178732	14	169051	169163
Log Normal (Box-Cox)	14	178620	178732	14	169051	169163
Gamma	14	178896	179008	14	168444	168556
Shash	16	178896	179024	16	167811	167939
Johnson's SU (the mean)	16	178930	179058	16	167764	167893
Skew t Type 1	16	178986	179113	16	167767	167895
Johnson's Original SU	16	179076	179204	16	168173	168301
Skew t Type 2	16	179091	179220	16	167766	167894
Skew Power Exponential Type 1	16	179152	179280	16	167764	167892
Skew Power Exponential Type 2	16	179298	179427	16	167764	167892
Generalized y	16	179765	179893	16	167762	167890
t Family	15	179768	179888	15	167764	167884
Power Exponential	15	179911	180032	15	**167761**	**167881**
Reverse Gumbel	14	180144	180257	14	171709	171821
NET	14	180166	180278	14	169169	169281

	Mean SBP			Mean DBP		
	Df	**AIC**	**BIC**	**Df**	**AIC**	**BIC**
Normal	14	181593	181705	14	167829	167942
Weibull	14	184068	184180	14	168911	169023
Gumbel	14	190253	190365	14	171757	171869
Exponential	13	256838	256942	13	231641	231745

With both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), the Box-Cox Power Exponential (BCPE) and the Box-Cox-t (BCT) distributions come as the best ones in approximating Mean DBP. For Mean SBP, both AIC and BIC favor the Power Exponential distribution. The Box-Cox Power Exponential (BCPE) and the Box-Cox-t (BCT) are continuous four parameter distributions (μ, σ, υ, τ) They generalize the Box-Cox Cole and Green distribution (BCCG) to allow for modeling kurtosis and skewness [3]. The Power Exponential distribution requires three distribution parameters (μ, σ, υ). For details about probability density functions of the Box-Cox-t/Box-Cox Power Exponential and the Power Exponential distributions please refer to Rigby and Stasinopoulos [14].

The estimation of continuous response variables assuming the Box-Cox Power Exponential distribution (BCPE) for Mean DBP and the Power Exponential distribution for Mean SBP is preceded by the selection of explanatory variables. The first distribution parameter μ (the mean of the response) is modeled assuming the inclusion of all the explanatory variables defined at the very beginning (Table 4). Table 8 checks whether the model can be simplified by potential dropping any of the terms in μ.

Table 8

Single term deletions for μ

	Mean DBP			Mean SBP		
	AIC	**LRT**	**Pr(Chi)**	**AIC**	**LRT**	**Pr(Chi)**
Age Cohort (years)	170573	2760	< 0.0001**	184047	4141	< 0.0001**
Gender (1=Male, 2=Female)	167833	17	< 0.0001**	180078	169	< 0.0001**
BMI Group (kg/m²)	167883	69	< 0.0001**	180219	312	< 0.0001**
Uric acid (mg/dL)	167861	45	< 0.0001**	179984	75	< 0.0001**
HDL-cholesterol (mmol/L)	167816	<0	0.6537	179933	23	< 0.0001**
Gamma Glut. Trans. (U/L)	167978	161	< 0.0001**	180048	139	< 0.0001**
Family PIR	167839	22	< 0.0001**	179956	47	< 0.0001**
Glucose (mmol/L)	167843	27	< 0.0001**	179962	52	< 0.0001**
Creatinine (umol/L)	167841	25	< 0.0001**	179969	60	< 0.0001**

Based on the Chi square test, HDL-cholesterol (mmol/L) does not contribute significantly to Mean DBP (no significant reduction of Akaike Information Criterion (AIC) as assessed by Likelihood-Ratio Test (LRT)). For Mean SBP, no terms can be left out, all of them will contribute to the final model.

Modeling the distribution parameters (μ, σ, υ) and τ of continuous response variables and thus selecting the best distributions for Mean DBP/Mean SBP is performed based on the linear parametric functions of the predictors (Table 7). For fitting nonlinear and nonparametric smooth functions, as the next step, additive term functions are applied and checked for the goodness of fit. Note that in this paper the modeling of the Box-Cox Power Exponential distribution (BCPE) for Mean DBP and the Power Exponential distribution for Mean SBP as the nonparametric smooth terms is restricted to Cubic Smoothing Spline Functions. Alternative additive terms, such as: Penalized Splines [4], Thin-Plate Smoothing Splines, Local Regression Splines [2], Fractional Polynomials, Power Polynomials [15], Random effects, Random coefficients [1], although very attractive, are not employed and not compared in this paper. Of particular notice are Random effects and Generalized Additive Mixed Models (GAMM) which pose very different and flexible approach to estimating Generalized Additive Models (GAM) [13].

Cubic Smoothing Spline Functions are curves which are made up of sections of joined cubic polynomials so that the functions $h(x)$ in model (2) are continuous in value and twice continuously differentiable. They are extensively covered in the literature [7, 9], thus their derivation is omitted.

Given $X = x$, Mean DBP is modeled by the Box-Cox Power Exponential distribution denoted as $BCPE(\mu, \sigma, \upsilon, \tau)$ where the distribution parameters (μ, σ, υ) and τ are modeled as smooth nonparametric functions of x, i.e.: $Y \sim BCPE(\mu, \sigma, \upsilon, \tau)$ where $g_1(\mu)$, $g_2(\mu)$, $g_3(\upsilon)$, $g_4(\tau)$ are defined by (3)–(6), respectively and for $k = 1, 2, 3, 4$, $g_k(.)$ are known link functions. The similar approach applies to Mean SBP (the Power Exponential distribution), i.e. $Y \sim PE(\mu, \sigma, \upsilon)$.

In order to establish whether smoothing terms are needed in the μ model defined by (3), all possible combinations of linear and cubic spline functions to the data are fitted. For each of the estimated models, Akaike Information Criterion (AIC) is assessed (stepwise model selection). The selection process is very time-consuming and its outputs very extensive. Thus, the full selection process of smoothing cubic splines is not included in this paper. Table 9 and Table 10 are brief summaries of the results.

Table 9

Summary of the selection process for Mean DBP – the first distribution parameter μ

From	To	Deviance	Resid. Df	Resid. Dev	AIC
LINEAR(Creatinine)	CS(Creatinine)	–181	22186	167606	167641
LINEAR(GGT)	CS(GGT)	–143	22183	167462	167504
LINEAR(Glucose)	CS(Glucose)	–70	22180	167392	167440
CLASS(Gender)		< 0	22181	167393	167439

Table 10

Summary of the selection process for Mean SBP – the first distribution parameter μ

From	To	Deviance	Resid. Df	Resid. Dev	AIC
LINEAR(GGT)	**CS(GGT)**	–141	22186	179740	179776
LINEAR (Creatinine)	**CS(Creatinine)**	–47	22183	179693	179735
LINEAR(Glucose)	**CS(Glucose)**	–31	22180	179662	179710
LINEAR(Uric acid)	**CS(Uric acid)**	–27	22177	179636	179689
LINEAR (HDL-cholesterol)	**CS(HDL-cholest.)**	–17	22174	179618	179678
LINEAR (Family PIR)	**CS(Family PIR)**	–16	22171	179603	179668

Having the model for μ determined, models for variance, skewness and kurtosis are searched for. The model selection procedure for these distribution parameters comprised of:

- Choosing link functions: for the Box-Cox Power Exponential distribution (Mean DBP), the default identity link functions are chosen for μ and υ, and log link functions are chosen for σ and τ;
- Selecting linear terms influencing the given distribution parameter: this is achieved by fitting models with linear terms;
- Applying cubic spline functions $h(x)$ to explanatory variables exhibiting nonparametric relation to the given distribution parameter: this is verified based on the Akaike Information Criterion (AIC);
- Selecting appropriate level of the "smoother" for each of the predictors modeled nonparametrically, i.e. choosing the effective degrees of freedom (DF) for smooth cubic spline functions $h(x)$ and denoted as df_μ, df_σ, df_υ and df_τ respectively: this is achieved by employing numerical optimization function to minimize the Generalized Akaike Information Criterion $GAIC(\#) = -2\hat{\ell} + \# df$ over hyper-parameters df_μ, df_σ, df_υ and df_τ in the Box-Cox Power Exponential model for Mean DBP, where $\hat{\ell}$ is the maximized log-likelihood function, # denotes penalty, *df* refers to the effective degrees of freedom (DF) and $-2\hat{\ell}$ is referred to as the global deviance. Due to limited space for this paper, the process of finding and searching for the best fit is not presented here. Note that the best model is found for hyper-parameters corresponding to penalty #= 2 and results in selecting a vector of hyper-parameters minimizing *GAIC* (2). For more details please refer to Rigby and Stasinopoulos [14].

Given that *CS* is Cubic Spline function and *df* denotes degrees of freedom (DF), the model for Mean DBP defined by four distribution parameters μ, σ, υ and τ of the Box-Cox Power Exponential *BCPE*(μ, σ, υ, τ) (for patients from Age Cohort: 57–<85 years and with Body Mass Index (BMI) Group: 30–<45 kg/m²) is given by:

$$\begin{cases} \mu = 58.09 + 6.05*Age\ Cohort + 1.33*BMI\ Group + 0.25*Uric\ acid \\ \quad + 0.04*CS(Gamma\ Glutamyl\ Transf., df = 6.48) - 0.25*Family\ PIR \\ \quad - 0.21*CS(Glucose, df = 7.25) + 0.03*CS(Creatinine, df = 9.49) \\ \log(\sigma) = -1.92 + 0.02*Age\ Cohort + 0.01*CS(Uric\ acid, df = 0.89) \\ \quad - 0.02*Family\ PIR + 0.01*CS(Glucose, df = 3.20) \\ \upsilon = 1.41 - 0.20*Age\ Cohort + 0.06*Family\ PIR \\ \quad - 0.003*CS(Creatinine, df = 2.76) \\ \log(\tau) = 0.71 - 0.13*Age\ Cohort + 0.09*Gender \end{cases}$$

Applying similar model selection procedure to Mean SBP for the first three distribution parameters, the model assuming the Power Exponential distribution PE(μ, σ, υ) is given by:

$$\begin{cases} \mu = 96.23 + 2.75*Age\ Cohort + 1.46*Gender + 4.59*BMI\ Group \\ \quad + 0.40*CS(Uric\ acid,\ df = 2.18) + 0.93*CS(HDL\text{-}cholest., df = 3.90) \\ \quad + 0.05*CS(Gamma\ Glutamyl\ Transf.,\ df = 8.11) \\ \quad - 0.20*CS(Family\ PIR,\ df = 2.39) + 0.79*CS(Glucose,\ df = 6.52) \\ \quad + 0.06*CS(Creatinine,\ df = 3.54) \\ \log(\sigma) = 2.03 + 0.76*Age\ Cohort - 0.04*Gender \\ \quad + 0.04*CS(HDL\text{-}cholest., df = 0.90) \\ \quad + 0.0008*CS(Gamma\ Glutamyl\ Transf., 2.55) - 0.02*Family\ PIR \\ \quad + 0.01*CS(Glucose, df = 1.34) + 0.001*Creatinine \\ \log(\upsilon) = 0.53 - 0.09*Age\ Cohort + 0.06*Gender \end{cases}$$

5. Discussion

The real-life example of Mean DBP/Mean SBP demonstrates the need for going beyond the exponential family distributions and thus, the usefulness of Generalized Additive Models (GAM) for Location, Scale and Shape. This overcomes the shortcomings associated with standard Generalized Additive Models (GAM) and Generalized Linear Models (GLM). In Generalized Additive Models (GAM) for Location, Scale and Shape the assumption of exponential family distributions is relaxed. It allows for modeling not only the mean associated with the location but also other distribution parameters of the response variable as additive nonparametric smoothing functions of explanatory variables. In the case of Mean DBP/Mean SBP data, the usage of the Box-Cox Power Exponential/Power Exponential distributions within Generalized Additive Models (GAM) allowed for modeling kurtotic distributions.

The estimated Generalized Additive Model (GAM) for binary Hypertension/ Borderline Hypertension response indicates that all the explanatory variables influence hypertension: Age Cohort, Gender, Body Mass Index (BMI) Group, Uric Acid, HDL-cholesterol, Gamma Glutamyl Transferase (GGT), Family PIR, Glucose and Creatinine. All explanatory variables are statistically significant with *p*-values much

lower than the assumed significance level of 0.05. It applies also to Generalized Additive Models (GAM) for Location, Scale and Shape estimated for Mean SBP. The results suggest that Mean DBP (which refers to the pressure when the heart is resting between beats) does not depend on Gender and HDL-cholesterol.

Systolic Blood Pressure seems to have stronger relationship with physiological and medical attributes than Diastolic Blood Pressure. All explanatory variables are nonlinear predictors of Systolic Blood Pressure. The models estimated for Hypertension/Borderline Hypertension and Mean DBP/Mean SBP suggest that:

- The blood pressure is higher among older subjects and subjects with elevated Body Mass Index (BMI). The risk of Hypertension/Borderline Hypertension is higher among population groups with overweight and obesity, particularly in Body Max Index Group >=30 kg/m^2. A similar trend prevails for Age Cohorts. The slopes of blood pressure are significantly higher in men than women (Table 5).
- People with elevated Uric Acid levels are at greater risk of hypertension. It is backed by other studies on Uric Acid [5]. The researchers conclude that Uric Acid may be an independent driver of high blood pressure and a marker of its prediction. A simple blood test can determine how much of it is present in the body. Effective drugs already exist which lower the level of Uric Acid and thus, offer a potential remedy for high blood pressure prevention.
- HDL-cholesterol levels influence Systolic Blood Pressure response. This is due to the association between high HDL-cholesterol and atherosclerosis (accumulation of HDL-cholesterol on the walls of arteries – hardening of the arteries). High HDL-cholesterol accelerates the progression of atherosclerosis which is thought to contribute to hypertension. This link is not significant for Diastolic Blood Pressure readings.
- There is a positive link between Gamma Glutamyl Transferase (GGT) levels, a marker of oxidative stress and blood pressure. Although the underlying mechanism of this association is still unclear, some studies confirm that higher Gamma Glutamyl Transferase (GGT) is implicated in the pathogenesis and progression of hypertension [17].
- People in low socioeconomic status environments as assessed by Income-Poverty Ratio are more susceptible to illnesses. Those with lower income tend to be at greater risk of hypertension. These findings are pretty worrying, especially in light of the fact that most researches to date are concentrated on hypertension in developed urban countries. As a result, very little is known about the problems and barriers to treatment and diagnosis outside high-income areas. This issue is even more acute knowing that the epidemic of cardiovascular disease occurring in low-income nations is largely driven by the increasing prevalence of high blood pressure.
- Glucose is able to induce Systolic Blood Pressure. Elevated Glucose level increases the likelihood of having diabetes which leads to higher Systolic Blood Pressure and heart diseases. The relationship with Diastolic Blood Pressure seems to be reversed.
- Higher blood pressure is associated with elevated Serum Creatinine level (an indicator of chronic renal disease). Creatinine draws water into the muscle what increases body weight and muscle volume. Retaining more water in the body impacts the blood volume and thus, blood pressure. Of note is the fact that half of the body's Creatinine is created naturally (produced in the liver and kidney) whereas the rest comes from the diet (the consumption of red meat and poultry). It confirms that appropriate diet plays its role in the treatment of blood pressure.

6. Conclusions

In this paper, the underlying methodology for Generalized Additive Models (GAM) has been introduced. The real-life data set example has demonstrated how one can use the nonparametric approach to model medical scheme data. This was achieved by investigating the dependencies and patterns of Hypertension/Borderline Hypertension and Mean DBP/Mean SBP on various physiological measurements, medical attributes and Income to Poverty Ratio. It is concluded that Generalized Additive Models (GAM) are a very powerful and flexible tool in an exploratory analysis, especially when you have little prior information about the data or you want to find new features that parametric tools ignore. Its nonparametric nature does not require much prior information and can also shed light on underlying parametric relationships. Generalized Additive Models (GAM) help to avoid model misspecification and provide information that might not be revealed by standard modeling techniques. The presented Generalized Additive Models (GAM) revealed pronouncedly complex nonlinear patterns in the response-predictor relationships for all predictors entered into the model. These nonlinear associations have been handled without the restrictions of parametric models, without sacrificing the interpretability and without the bias associated with the "curse of dimensionality". The built Generalized Additive Models (GAM) seem to represent the behavior of the data closer than the parametric counterparts. It underlines the importance of this class of models in detecting nonlinear dependencies and suggests potential failure of parametric solutions in capturing important features of the medical scheme data.

References

[1] Chambers J.M., Hastie T.J., *Statistical Models in S*, Chapman & Hall, London 1992.

[2] Cleveland W.S., Grosse E., Devlin S.J., *Regression By Local Fitting*, Journal of Econometrics, vol. 37, 1988, 87-114.

[3] Cole T.J., Green P.J., *Smoothing reference centile curves: the LMS method and penalized likelihood*, Statistical Modeling, vol. 11, 1902, 1305-1319.

[4] Eilers P.H., Marx B.D., *Flexible smoothing with B-splines and penalties*, Statistical Science, vol. 11, 1996, 89-121.

[5] Feig D.I, Kang D.H., Nakagawa T., Mazzali M., Johnson R.J., *Uric acid and hypertension*, Curr Hypertens Rep., vol. 8(2), 2006, 111-115.

[6] Friedman J.H., Stuetzle W., *Projection Pursuit Regression*, Journal of the American Statistical Association, vol. 76, 1981, 817-823.

[7] Green P.J., Silverman B.W., *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall, London 1994.

[8] Gurven M., Blackwell A., Rodríguez A., Stieglitz J., Kaplan H., *Does Blood Pressure Inevitably Rise With Age?*, Longitudinal Evidence Among Forager-Horticulturalists, Hypertension, vol. 60(1), 2012, 25-33.

[9] Hastie T.J., Tibshirani R.J., *Generalized Additive Models*, Chapman & Hall, London 1990.

[10] Johnson N.L., Kotz S., Kemp A.W., *Univariate Discrete Distributions*, Wiley, New York 2005.

[11] Lee J.A., Verleysen M., *Nonlinear Dimensionality Reduction*, Springer, New York 2007.
[12] Łukasik S., Kulczycki P., *Using Topology Preservation Measures for High-Dimensional Data Analysis in a Reduced Feature Space*, Technical Transactions, vol. 1-AC/2012, Cracow University of Technology Press, 5-15.
[13] Pinheiro J.C., Bates D.M., *Mixed-Effects Models in S and S-PLUS*, Springer-Verlag, New York 2000.
[14] Rigby R.A., Stasinopoulos D.M., *Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis*, Statistical Modeling, vol. 6, 2006, 209-229.
[15] Royston P., Altman, D.G., *Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling*, Appl. Statist., vol. 43, 1994, 429-467.
[16] Stone C.J., Hansen M., Kooperberg C., Truong Y.K., *Polynomial splines and their tensor products in extended linear modeling*, Annals of Statistics, vol. 25, 1997, 1371-1470.
[17] Shankar A., Li J., *Association between serum gamma-glutamyltransferase level and prehypertension among US adults*, Circ J., vol. 71(10), 2007, 1567-1572.
[18] Tomera J.F., Harakal C., *Multiple linear regression analysis of blood pressure, hypertrophy, calcium and cadmium in hypertensive and nonhypertensive states*, Food and Chemical Toxicology, vol. 35(7), 1997, 713-718.
[19] Wahba G., *Bayesian Confidence Intervals for the Cross Validated Smoothing Spline*, Journal of the Royal Statistical Society, vol. 45, 1983, 133150.
[20] Webpage of National Health & Nutrition Examination Survey (source of the data set used for the analysis): http://www.cdc.gov/nchs/nhanes.htm.

Table 11

Distribution parameters of Box-Cox Power Exponential distribution (BCPE) – Mean DBP

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
μ – Mu Coefficients (Mu link function: identity)				
(Intercept)	58.09	0.481	120.7	< 0.0001**
Age Cohort (23–<40 vs 12–<23 years)	5.76	0.212	27.1	< 0.0001**
Age Cohort (40–<57 vs 12–<23 years)	10.81	0.222	48.7	< 0.0001**
Age Cohort (57–<85 vs 12–<23 years)	6.05	0.240	25.2	< 0.0001**
BMI Group (25–<30 vs 14–<25 kg/m²)	< 0.01	0.184	< 0.1	0.1986
BMI Group (30–<45 vs 14–<25 kg/m²)	1.33	0.202	6.6	< 0.0001**
LINEAR(Uric acid)	0.25	0.065	3.8	< 0.0001**
CS(Gamma Glutamyl Transferase, df = 6.48)	0.04	0.003	13.1	< 0.0001**
LINEAR(Family PIR)	–0.25	0.046	–5.3	< 0.0001**
CS(Glucose, df = 7.25)	–0.21	0.067	–3.1	< 0.0001**
CS(Creatinine, df = 9.49)	0.03	0.004	6.2	< 0.0001**

	Estimate	Std. Error	t value	Pr(>|t|)
σ – Sigma Coefficients (Sigma link function: log)				
(Intercept)	–1.92	0.031	–62.5	< 0.0001**
Age Cohort (23–<40 vs 12–<23 years)	–0.07	0.015	–4.7	< 0.0001**
Age Cohort (40–<57 vs 12–<23 years)	–0.19	0.015	–12.6	< 0.0001**
Age Cohort (57–<85 vs 12–<23 years)	0.02	0.015	1.4	0.1503
CS(Uric acid, df = 0.89)	0.01	0.004	3.7	0.0002**
LINEAR(Family PIR)	–0.02	0.003	–5.9	< 0.0001**
CS(Glucose, df = 3.20)	0.01	0.004	2.8	0.0049**
υ – Nu Coefficients (Nu link function: identity)				
(Intercept)	1.41	0.117	12.0	< 0.0001**
Age Cohort (23–<40 vs 12–<23 years)	–0.39	0.101	–3.8	< 0.0001**
Age Cohort (40–<57 vs 12–<23 years)	–0.67	0.113	–5.5	< 0.0001**
Age Cohort (57–<85 vs 12–<23 years)	–0.20	0.094	–2.2	0.0246*
LINEAR(Family PIR)	0.06	0.022	2.6	0.0082*
CS(Creatinine, df = 2.76)	< –0.01	0.001	–2.9	0.0032**
τ – Tau Coefficients (Tau link function: log)				
(Intercept)	0.71	0.040	17.8	< 0.0001**
Age Cohort (23–<40 vs 12–<23 years)	–0.19	0.042	–4.5	< 0.0001**
Age Cohort (40–<57 vs 12–<23 years)	–0.18	0.043	–4.2	< 0.0001**
Age Cohort (57–<85 vs 12–<23 years)	–0.13	0.042	–3.2	0.0014**
Gender (Male vs Female)	0.09	0.030	2.9	0.0034**
Degrees of Freedom for the fit: 68.90, No. of observations in the fit: 22204				

Table 12

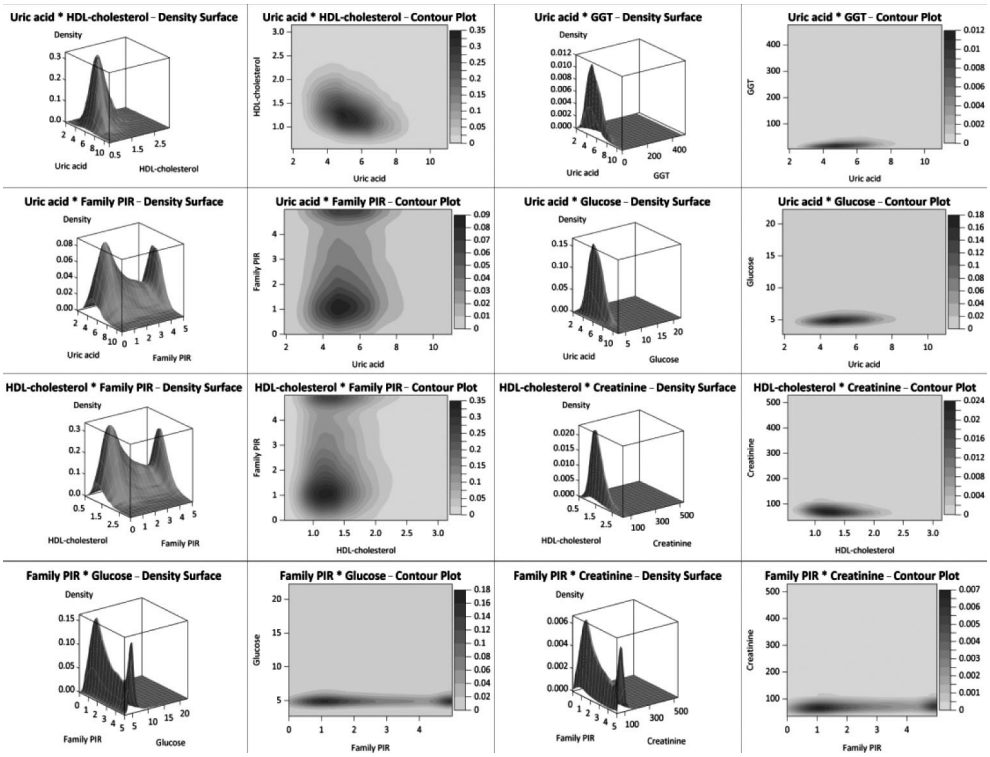
Distribution parameters of Power Exponential distribution (PE) – Mean SBP

	Estimate	Std. Error	t value	Pr(>|t|)
μ – Mu Coefficients (Mu link function: identity)				
(Intercept)	96.23	0.696	138.2	< 0.0001**
Age Cohort (23–<40 vs 12–<23 years)	6.18	0.250	24.7	< 0.0001**
Age Cohort (40–<57 vs 12–<23 years)	18.23	0.319	57.2	< 0.0001**

	Estimate	Std. Error	t value	Pr(>|t|)
Age Cohort (57–<85 vs 12–<23 years)	2.75	0.198	13.9	< 0.0001**
Gender (Male vs Female)	1.46	0.191	7.6	< 0.0001**
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
BMI Group (25–<30 vs 14–<25 kg/m²)	2.20	0.194	11.3	< 0.0001**
BMI Group (30–<45 vs 14–<25 kg/m²)	4.59	0.219	20.9	< 0.0001**
CS(Uric acid, df = 2.18)	0.40	0.077	5.2	< 0.0001**
CS(HDL-cholesterol, df = 3.90)	0.93	0.230	4.1	< 0.0001**
CS(Gamma Glutamyl Transferase, df = 8.11)	0.05	0.004	13.5	< 0.0001**
CS(Family PIR, df = 2.39)	–0.20	0.045	–4.2	< 0.0001**
CS(Glucose, df = 6.52)	0.79	0.085	9.4	< 0.0001**
CS(Creatinine, df = 3.54)	0.06	0.005	12.2	< 0.0001**
σ – Sigma Coefficients (Sigma link function: log)				
(Intercept)	2.03	0.042	48.1	< 0.0001**
Age Cohort (23–<40 vs 12–<23 years)	0.12	0.015	7.9	< 0.0001**
Age Cohort (40–<57 vs 12–<23 years)	0.48	0.016	29.2	< 0.0001**
Age Cohort (57–<85 vs 12–<23 years)	0.76	0.016	48.7	< 0.0001**
Gender (Male vs Female)	–0.04	0.013	–3.3	< 0.0001**
CS(HDL-cholesterol, df = 0.90)	0.04	0.015	3.1	< 0.0001**
CS(Gamma Glutamyl Transferase, df = 2.55)	< 0.01	< 0.001	4.1	< .0001**
LINEAR(Family PIR)	–0.02	0.003	–6.7	< 0.0001**
CS(Glucose, df = 1.34)	0.01	0.004	2.9	< 0.0001**
LINEAR(Creatinine)	< 0.01	<0.001	4.1	< 0.0001**
υ – Nu Coefficients (Nu link function: log)				
(Intercept)	0.53	0.065	8.4	<0.0001**
Age Cohort (23–<40 vs 12–<23 years)	–0.19	0.041	–4.7	<0.0001**
Age Cohort (40–<57 vs 12–<23 years)	–0.32	0.041	–7.8	<0.0001**
Age Cohort (57–<85 vs 12–<23 years)	–0.09	0.042	–2.0	0.0430*
Gender (Male vs Female)	0.06	0.028	2.1	0.0350*
Degrees of Freedom for the fit: 63.59, No. of observations in the fit: 22204				

Appendix 1

Kernel Density for selected Explanatory Variables: Density Surface and Contour Plot



Appendix 2

Probability density functions of Box-Cox-t, Box-Cox Power Exponential and Power Exponential distributions

The probability density function $f_T(y_i|\mu_i,\sigma_i,\upsilon_i,\tau_i)$ of the Box-Cox-t distribution (BCT) is given by:

$$f_T(y\,|\,\mu,\sigma,\upsilon,\tau)=\frac{y^{\upsilon-1}f_T(z)}{\mu^{\upsilon}\sigma F_T\left(\dfrac{1}{\sigma\,|\,\upsilon\,|}\right)} \tag{7}$$

for $y > 0$, where Y is a positive random variable of the Box-Cox t distribution, $\mu > 0$, $\sigma > 0$, and $-\infty < \upsilon < \infty$ and Z is the transformed random variable given by:

$$Z = \begin{cases} \dfrac{1}{\sigma\upsilon}\left[\left(\dfrac{Y}{\mu}\right)^{\upsilon} - 1\right], & \text{if} \quad \upsilon \neq 0 \\ \dfrac{1}{\sigma}\log\left(\dfrac{Y}{\mu}\right), & \text{if} \quad \upsilon = 0 \end{cases} \tag{8}$$

where Z follows a truncated t distribution with degrees of freedom (DF), $\tau > 0$. $f_T(t)$ and $F_T(t)$ are respectively the probability density function and the cumulative distribution function of a random variable T. T has a standard t distribution with degrees of freedom (DF) $\tau > 0$.

The probability density function of the Box-Cox Power Exponential distribution (BCPE) is given by (7) where Z follows a truncated standard Power Exponential distribution with power distribution parameter, $\tau > 0$. The Power Exponential distribution requires three parameters. The probability density function of the Power Exponential family distribution is defined by:

$$f_Y(y \mid \mu, \sigma, \upsilon) = \frac{\upsilon \exp\left[-\left|\dfrac{z}{c}\right|^{\upsilon}\right]}{2c\sigma\Gamma\left(\dfrac{1}{\upsilon}\right)} \tag{9}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\upsilon > 0$ and where $z = (y - \mu)/\sigma$ and $c^2 = \Gamma\left(\dfrac{1}{\upsilon}\right)\left[\Gamma\left(\dfrac{3}{\upsilon}\right)\right]^{-1}$. In this parameterization, $E(Y) = \mu$ and Var $(Y) = \sigma^2$.