FLEXIBLE REGRESSION MODELLING OF THE PURE PREMIUM USING GENERALIZED ADDITIVE MODELS WITH ISOTROPIC SMOOTHING

MAREK STREŽO[1]

Flexibilné regresné modelovanie rizikového poistného pomocou zovšeobecnených aditívnych modelov s isotrópnym vyhladením.

***Abstract**: In case of large number of independent variables mainly in insurance, many non-parametric methods do not perform well. The widely used Generalized Additive Model (GAM) is a flexible technique in which the usual linear relationships between the response and predictor variables are replaced by non-linear smooth. These semi-parametric models permit the response probability distribution to be a member of the exponential family of distributions with robust extensions of Logit, Poisson, Negative Binomial and other Generalized Linear Models. GAMs are represented using penalized regression splines and are estimated by penalized regression methods. Cross validation is using for estimation the degree of smoothness for the unknown functions in the linear predictor. The GAMs allow us to build a regression surface as a sum of lower-dimensional non-parametric term circumventing the curse of dimensionality. This paper discusses the non-life insurance pricing context in which this technique has been developed and several GAMs are compared, where the best model is selected using AIC, GAM UBRE score, deviations and explained deviation.*

***Keywords**: GAMs models, Smoothing splines, UBRE/GCV score, Pure Premium*

JEL Classification: C10, C21, G22

1. Introduction

Every non-life insurance pricing involves problem with continuous rating variables, like the age of the policyholder or the weight of the insured vehicle. In the Generalized Linear Model (GLM), continuous rating variables are categorized into intervals and all values within an interval are treated

[1] Ing. Marek Strežo, University of Economics in Bratislava, Slovak Republic, e-mail: marek.strezo@euba.sk

as identical. This method has advantage of being simple and often works well enough. However, an obvious disadvantage of categorization is that the premium for two policies with different but close values for the rating variable may get significantly different premiums if the values happen to belong to different intervals. Also, finding a good subdivision into intervals can be time consuming and tedious. The intervals must be large enough to achieve good precision of the price relativities, but at the same time they have to be small if the effect of the rating variable varies much. Sometimes both of these requirements cannot be met. With this in mind, an alternative modeling approach can be used. Hence, we introduce Generalized Additive Models (GAMs), which extend possibilities of GLMs by using splines as reparameterization tool. They were first proposed by Hastie and Tibshirani [6]. The difference is at the level of the linear predictor.

There exist several possible approaches that go under the name of smoothing splines. It utilizes results from the theory of splines, i.e. piecewise polynomials, which have their origin in the field of numerical analysis. [10] GAMs are suitable to study the behavior of the factors that influence the expected value of a response variable. They are especially useful when it is suspected that the relationship is not linear. On the other hand, since this is a nonparametric approach, GAMs let the data help choosing the functional forms (what is known as letting the data speak) and therefore allow going beyond the typical parametric relationship of a GLMs. However, GAMs are more complex and more difficult to interpret than GLMs. The classical references on these models are obviously Hastie and Tibshirani [5], where most of the explanations on the technical tools used in this paper can be found. Of course, it is also possible to use existing R packages, or SAS procedures for the application of GAMs to large databases. In the following sections, the data set used in our empirical application is presented and explained in detail.

1.1 Tariff Analysis

When an insurance company accepts new insurances or when the premiums of earlier accepted insurances have to be changed on renewal the company has to search for the factors that influence the premium and calculate the premium according to the values of these factors. A tariff is represented by a formula, by which the premium can be computed. The underlying work an actuary performs to obtain tariff is called a tariff analysis. The data material for a tariff analysis is historical data with information about policies and claims. Consequently, the pure premium is a product of the claim frequency and the claim severity. [10]

1.2 Rating Factors

Both claim severity and claim frequency vary between policies and can be estimated based on a set of a number of variables, the rating factors. A rating factor, also called as rating variable can be either continuous or categorical. In a tariff analysis, it is common to categorize continuous rating variables into intervals and to treat them as categorical rating variables. This is done to improve the significance of the statistical results. Policies within the same interval for each rating variable are said to belong to the same tariff cell and share the same premium. [10]

2. Isotropic smoothing – Thin plate regression splines

This section considers smooths of one or more variables, concretely Thin plate splines, in particular smooths that, in the multiple covariate case, produce identical predictions of the response variable under any rotation or reflection of the covariates. Our goal will be to model response variable as a smooth function of p covariates $f_{x_1,\dots,x_p}(x_1, \dots, x_p)$. This function is founded as a linear combination of some basis function, i.e.

$$f_{x_1,\dots,x_p}(x_1, \dots, x_p) = \sum_{i=1}^{k} \beta_i b_i\,(x_1, \dots, x_p), \tag{1}$$

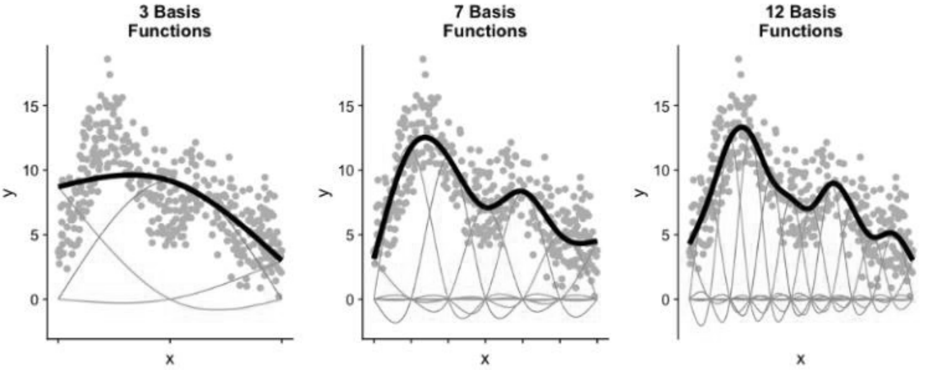
where β_i are unknown parameters. Consider the response variable satisfy following model as

$$Y_i = f_{x_1,\dots,x_p}(x_{1i}, \dots, x_{pi}) + \varepsilon_i\,,\ \ i = 1, \dots, n \tag{2}$$

and the $\varepsilon_i$termsare independent random variables such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i) = \sigma^2$.With construction of splines of multiple predictor variables ensues two main questions. First, we want to find the unknown basis functions $b_1(x_1, \dots, x_p), \dots, b_k(x_1, \dots, x_p)$. The factor that affects how wiggly a GAM function can be is the number of basis functions that make up a smooth function. In general, we want to balance two things when fitting a nonlinear model. We want a model that captures the relationship by being close to the data, but we also want to avoid fitting our model to noise, or over-fitting. How well the GAM captures patterns in the data is measured by a term called likelihood. Its complexity, or how much the curve changes shape, is measured by ‘wiggliness’. In Figure 1 we have plotted GAMs with 3, 7 and 12 basis functions all fit to the same data.

Figure 1

Number of basis functions b_i



Source: processed by the author using statistical program R, 2019

As you can see, a smooth with a small number of basis functions is limited in its wiggliness, while one with many basis functions is capable of capturing finer patterns.

Second, it is necessary to identify a penalty function measuring the “wiggliness” $J_{md}(f)$. Let p is number of covariates and m is a some-order derivation considered in penalties. Let $\nu_1, \dots, \nu_p \in \{0,1,\dots,p\}$ [1]. Then the penalty function is defined as

$$J_{md} = \int \dots \int_{\Re^d} \sum_{\nu_1+\dots+\nu_d=m} \frac{m!}{\nu_1! \dots \nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1} \dots \partial x_d^{\nu_d}} \right)^2 \mathrm{d}x_1 \dots \mathrm{d}x_d. \tag{3}$$

Thin-plane splines were constructed directly from basic requirements for model of smoothing splines. λ is a smoothing parameter and $\mathbf{x} = (x_1, \dots, x_p)^{\mathrm{T}}$ is a vector of covariates. Further progress is only possible if m is chosen as $2m < p$, and in fact for ‘visually smooth’ results it is preferable that $2m < p + 1$.

Subject to the first of these restrictions, it can be shown that the function minimizing $\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \lambda J_{md}(f)$ has the form

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}(\|\mathbf{x} - \mathbf{x_i}\|) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}), \tag{4}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$are vectors of unknown parameters to be estimated. Vector $\boldsymbol{\delta}$ has also applied that $\mathbf{T^T\delta = 0}$, where $T_{ij} = \phi_{ij}(x_j)$ [2]. The function $\phi_1, \dots, \phi_M$, where $M = \binom{m+d-1}{d}$ are linearly independent polynomials spanning the space of polynomials in $\Re^d$ of degree less than m. Linear cover ϕ_i span the space of function, where $J_{md} = 0$. Basis functions used in (7) are defined as

$$\eta_{mp}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!\,(m-d/2)!} r^{2m-d} \log(r), & d \text{ is even} \\ \frac{\Gamma\left(\frac{d}{2} - m\right)}{2^{2m}\pi^{d/2}(m-1)!} \quad , & d \text{ is odd.} \end{cases} \tag{5}$$

Now defining matrix $\mathbf{E}_{n \times n}$ evaluated basis factors by $E_{ij} \equiv \eta_{md}(\|\mathbf{x_i} - \mathbf{x_j}\|)$. Then the matrix of evaluated basis functions $\mathbf{B}$for Thin-plate splines is defined as $\mathbf{B} = (\mathbf{T}\ \ \mathbf{E})$.Vector of unknown parameters $\boldsymbol{\beta}$ is defined as $\beta = (\alpha^{\mathrm{T}}, \delta^{\mathrm{T}})^{\mathrm{T}}$. Under the condition $\mathbf{T^T\delta = 0}$ results

$$J_{mp} = \left(\sum_{i=1}^{n} \delta_i \eta_{mp}(\|\mathbf{x} - \mathbf{x}_i\|)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \eta_{mp}(\|\mathbf{x} - \mathbf{x}_i\|). \tag{6}$$

Since, the basis functions $\phi_1(\mathbf{x}), \dots, \phi_M(\mathbf{x})$, are not penalized, then from (6) can be defined the penalized matrix $\mathbf{S}$ as

$$\mathbf{S} = \begin{pmatrix} \mathbf{0}_{M\times M} & \mathbf{0}_{M\times n} \\ \mathbf{0}_{n\times M} & \mathbf{E}_{n\times n} \end{pmatrix}.$$

Unknown parameters of Thin-plate splines can be fitted by minimalize of penalized least squared as

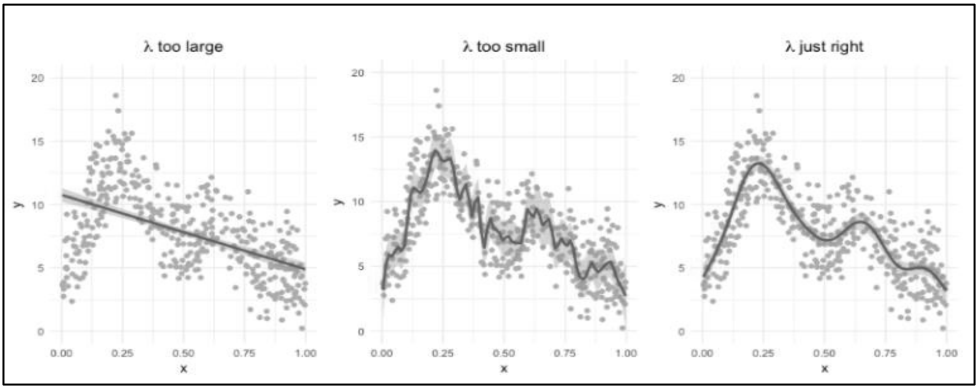
$$\mathbf{SS}_{pen}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta} = (\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}) + \lambda\boldsymbol{\delta}^{\mathrm{T}}\mathbf{E}\boldsymbol{\delta}, \tag{7}$$

under the condition $\mathbf{T}^{\mathrm{T}}\boldsymbol{\delta} = \mathbf{0}$.

As we have shown above, the smoothing parameter λ controls the balance. Here are plots of three GAMs with different smoothing, or lambda values. One on the left smooths too much, creating a straight line through curved data. The one in the middle smooths too little, fitting noise rather than the trend. The one on the right is just right. It is lambda value balances over-and-under-fitting, see Figure 2.

Figure 2

Choosing the right smoothing parameter λ



Source: processed by the author using statistical program R, 2019

Normally when we fit a model with package mgcv and *gam()* function, we let this package does the work of selecting a smoothing parameter. However, we can fix the smoothing parameter to a value of our choosing. Instead if we allow R to do this work for us, the mgcv package offers several different methods for selecting smoothing parameters. I, and most GAM experts, strongly recommend that you fit models with the REML, or "Restricted

Maximum Likelihood" method. While different methods have their advantages, REML is most likely to give you stable results. [7]

3. Generalized Additive Models

Generalized Additive Models, also known with the acronym GAMs are nonparametric GLMs. These models are the extension of GLMs to a combination of a linear predictor and the sum of smooth functions of explanatory variables. While Gaussian models can be used in many statistical applications, for several types of problems they are not appropriate, like the case of non-life insurance pricing. [9]

GAMs are composed of a random component, an additive component and a link function. The response variable, Y, follows some exponential family distribution

$$f_{\theta}(y) = \exp\left[\frac{\{y\theta - b(\theta)\}}{a(\phi)} + c(y,\phi)\right], \tag{8}$$

where a, b and c are arbitrary functions, ϕ an arbitrary 'scale' parameter, and θ is known as the 'canonical parameter' of distribution. [9]

The mean$\mu = E(y)$characterized in GAMsis linked to nonlinear nonparametric predictor $\eta = g(\mu) = \alpha + \sum_{i=1}^{p} f_i(x_i)$, where $f_i(x_i)$ are smooth nonparametric functions. Relationship between the mean and η is defined by a link function $g(\mu) = \eta$. The most commonly used link function is canonical link function, where $\eta = \theta$. In order to fit GAMs to the data, we use basis expansions of smooth functions and penalized likelihood maximization for model estimation in which wiggly models are penalized more heavily than smooth models in a controllable manner, and degree of smoothness is chosen based on AIC. [9]

3.1 Fitting GAMs by penalized iterative re-weighted least squares (P-IRLS)

GAMs can be fitted by penalized likelihood maximization, and in practice this will be achieve by penalized iterative re-weighted least squares, as P-IRLS. For given smoothing parameters, the following steps are iterated to convergence:

1. Given the current linear predictor estimate, $\hat{\boldsymbol{\eta}}$, and corresponding estimated mean response vector, $\hat{\boldsymbol{\mu}}$, calculate:

$$w_i = \frac{1}{V(\hat{\mu}_i)g'(\hat{\mu}_i)^2} \text{and} z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i$$

where $\mathrm{var}(Y_i) = V(\mu_i)\phi$ and g is a link function.

2. Defining $\mathbf{W}$ as the diagonal matrix such that $W_{ii} = w_{ii}$, minimize

$$\left\|\sqrt{\mathbf{W}}\mathbf{z} - \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\beta}\right\|^2 + \lambda_1 \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{S_1} \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{S_2} \boldsymbol{\beta}$$

w.r.t. $\boldsymbol{\beta}$to obtain new estimate $\widehat{\boldsymbol{\beta}}$, and hence updated estimates $\widehat{\boldsymbol{\eta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ [9].

3.2 Smoothness selection criteria

In the previous section, we have described P-IRLS method for estimating $\boldsymbol{\beta}$ given the smoothing parameter $\boldsymbol{\lambda}$.The smoothing parameter balances between likelihood and wiggliness to optimize model fit. The question now is, how can be the parameter $\boldsymbol{\lambda}$ estimated? There two main approaches. The first one is used, when σ^2 is known and the second if this parameter is unknown. In the next sections the both approaches are described.

3.2.1 Known scale parameter: UBRE

This approach try to choose smooth parameters in order to value $\widehat{\boldsymbol{\mu}}$is a close as possible to the true value $\mu \equiv \mathbb{E}(\mathbf{y})$. That's the reason, why we need to minimize the expected mean square error (MSE) M of the model, such that

$$M = \mathbb{E}\left(\left\|\boldsymbol{\mu} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|^2/n\right) = \mathbb{E}(\|\mathbf{y} - \mathbf{Ay}\|^2)/n - \sigma^2 + 2\mathrm{tr}(\mathbf{A})\sigma^2/n. \qquad (9)$$

Minimizing (9) is the same as minimize of the un-biased risk estimator – UBRE, defined as

$$v_u(\boldsymbol{\lambda}) = \|\mathbf{y} - \mathbf{Ay}\|^2/n - \sigma^2 + 2tr(\mathbf{A})\sigma^2/n, \qquad (10)$$

Note that r.h.s of (10) depends on the smoothing parameters through $\mathbf{A}$. If σ^2 is known then estimating λ by minimizing V_u works well, but the problems arise if σ^2 has to be estimated [9]. For example, substituting the approximation

$$\mathbb{E}(||\mathbf{y}-\mathbf{Ay}||^2)=\sigma^2\{n-\mathrm{tr}(\mathbf{A})\} \qquad (11)$$

implied by estimating the variance defined as $\hat{\sigma}^2=\frac{||\mathbf{y}-\mathbf{Ay}||^2}{n-\mathrm{tr}(\mathbf{A})}$ into (9) yields

$$M=\mathbb{E}\left(\left\|\boldsymbol{\mu}-\mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2/n\right)=\frac{\mathrm{tr}(\mathbf{A})}{n}\sigma^2 \qquad (12)$$

where the MSE estimator $\widetilde{M}=\mathrm{tr}(\mathbf{A})\hat{\sigma}^2/n.$ Now consider comparison of 1 and 2 parameter unpenalized models using $\widetilde{M}$: the 2 parameter model has to reduce $\hat{\sigma}^2$ to less than half the one parameter σ^2 estimate before it would be judged to be an improvement. Clearly, therefore, $\widetilde{M}$, is not suitable basis for model selection. [8]

3.2.2 Unknown scale parameter: Cross validation

As we have presented in previous section, minimize the average square error in model predictions of $\mathbb{E}(\mathbf{y})$ will not work well when ϕ is unknown. An alternative is to base smoothing parameter estimation on mean square prediction error, which is readily shown to be

$$P=\sigma^2+M. \qquad (13)$$

where M denotes the same as in (9). The direct dependence on σ^2 tends to mean that criteria based on P are much more resistant to over-smoothing, which would inflate the σ^2 estimate, than are criteria based on M. [8]

To estimate P we use the most obvious way – cross validation. There is always omitted one response,y_i, from the model fitting process and then is fitted again on the remaining data. P estimate by Ordinary Cross Validation – OCV is given by

$$v_0 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i^{[-i]}\right)^2, \tag{14}$$

where $\hat{\mu}_i^{[-i]}$ denotes the prediction $\mathbb{E}(y_i)$ obtained from the fitted model to all data except y_i. The advantage is, that it is not necessary to calculate v_0 by performing n models fits, to obtain the n terms $\hat{\mu}_i^{[-i]}$. [9] When we to find i^{th} term in OCV score, we have to focus on minimizing the penalized least squares objective. We have that

$$\sum_{\substack{j=1\\ j\neq i}}^{n}\left(y_j - \hat{\mu}_j^{[-i]}\right)^2 + Penalties. \tag{15}$$

For simplify (15) we can add the term $\left(\hat{\mu}_j^{[-i]} - \hat{\mu}_j^{[-i]}\right)^2$, which is equal to zero. Clearly adding zero to this objective will leave the estimates that minimize it completely unchanged. So we obtain

$$\sum_{j=1}^{n}\left(y_j{}^{*} - \hat{\mu}_j^{[-i]}\right)^2 + Penalties, \tag{16}$$

where $\mathbf{y}^{*} = \mathbf{y} - \bar{\mathbf{y}}^{[i]} + \bar{\boldsymbol{\mu}}^{[i]}$: $\bar{\mathbf{y}}^{[i]}$and$\bar{\boldsymbol{\mu}}^{[i]}$are vectors of zeroes except for their i^{th} elements, which are y_i and $\hat{\mu}_i^{[-i]}$, respectively. Minimizing itself obviously results i^{th} prediction of $\hat{\mu}_i^{[-i]}$ and also in an influence matrix $\mathbf{A}$ for the model with whole data. So, considering the i^{th} prediction, we have that, $\hat{\mu}_i^{[-i]} = \mathbf{A}_i\mathbf{y}^{*} = \mathbf{A}_i\mathbf{y} - A_{ii}y_i + A_{ii}\hat{\mu}_i^{[-i]} = \hat{\mu}_i - A_{ii}y_i + A_{ii}\hat{\mu}_i^{[-i]}$, where $\hat{\mu}_i$ comes from the fit to full $\mathbf{y}$. After rearrangement then yields $y_i - \hat{\mu}_i^{[-i]} = (y_i - \hat{\mu})/(1 - A_{ii})$, so that the OCV score becomes

$$v_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2}, \tag{17}$$

what can clearly be calculated from a single fit of the original model. [9]

OCV is reasonable way of estimating smoothing parameters, but suffers from two potential drawbacks. Firstly, it is computationally expensive to minimize in the additive model case, where there may be several smoothing parameters. Secondly, it has a slightly disturbing lack of invariance. For solving these problems Woods shows, that the ordinary cross validation score can be written

$$\upsilon_G = \frac{n\|\mathbf{y} - \widehat{\boldsymbol{\mu}}\|^2}{[n - tr(\mathbf{A})]^2}, \tag{18}$$

which is known as the Generalized Cross Validation score – GCV [3].

When in each iteration a penalized reweighted least squares problem is solved, and the smoothing parameters of that problem are estimated by GCV or UBRE. Eventually, both regression coefficients and smoothing parameter estimates converge.

3.2.3 A distributional assumptions and testing hypotheses about β

In previous section, we have shown how to estimate parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$. In practice, we have to be able to determine their significance and find the confidence intervals of parameters $\boldsymbol{\beta}$. It is advisable to showshow reliable our estimates are. There are two ways to determine the quality of estimates. Firstly, it is by defining $S = H + \sum_i \lambda_i S_i$, recalling that parameter estimates are given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{y}, \tag{19}$$

where data or pseudo data have covariance matrix $\mathbf{W}^{-1}\phi$.We have then

$$\mathbf{V}_{\boldsymbol{e}} = (\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-\mathbf{1}}\phi. \tag{20}$$

From the normality or asymptotical normality as distributional results we have

$$\widehat{\boldsymbol{\beta}} \sim N\big(\mathbb{E}(\widehat{\boldsymbol{\beta}}), \mathbf{V}_e\big). \tag{21}$$

Generally, it does not hold $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$**.** However, if $\boldsymbol{\beta} = \mathbf{0}$ then $\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{0}$, with the same validity approximately for same subsets of $\boldsymbol{\beta}$ [13]. Therefore, this result can be used for testing significance of regression parameters. And so, based on distributional results we can perform Wald tests for significance of regression coefficients $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ in the following form

$$H_0\colon \boldsymbol{\beta_j} = \mathbf{0} \qquad vs. \qquad H_1\colon \boldsymbol{\beta_j} \neq \mathbf{0}.$$

Tests for two nested models from which we want to choose the preferable one are performed by ANOVA for GAM. [4]

To summarize it, we introduce GAM as extension of GLM in which the linear predictors can also partly depend linearly on some unknown smooth functions. Regression coefficients are estimated by a penalized version of the method used to fit GLM, where an extra criterion has to be optimized to find the smoothing parameters.

4. The practical part – empirical study of Motor TPL insurance portfolio

In this chapter, we will use some public insurance data to demonstrate usage of models which have been introduced in the previous chapters of this paper. We will show calculation and comparison the results of different models. GAMs have various applications in all fields related with statistics. The nonlife insurance is no exception where advanced regression models are considered to be the best market practice in the pricing and the reserving. We will be using the R software to calculate and to analyze the results from different models. As we mentioned above, it is used package 'mgcv' and function *gam()*.

4.1 Data structure

Using the package 'insuranceData' in R [11], we can load the free available dataset 'dataCar'. These data contain one-year vehicle insurance policies in years 2004 and 2005, where is 67,856 number of policies. These observations have the following 11 rating factors:

- **veh_value** vehicle value , in 10 000 dollars,
- **exposure** contains the measure between 0 and 1,
- **clm** has the binary data – occurrence of claims (0 = no, 1 = yes),
- **numclaims** represents the number of claims,
- **clmcosts0** has the claim cost information (also with 0 value if there is no claims),
- **veh_body** vehicle body, coded as BUS CONVT COUPE HBACK HDTOP MCARA MIBUS PANVN RDSTR SEDAN STNWG TRUCK UTE

- **veh_age** contains values from 1 to 4, where in the tariff class 1 are the youngers vehicles,
- **gender** may include one of the two values F or M,
- **area** a factor which include the levels A, B, C, D, E, and F,
- **agecat** with levels 1, 2, 3 and 4, where in the level 1 are the younger drivers,
- **X OBSTAT** a factor with levels 01101 0 0 0.

4.2 Model construction, Testing and the Results

The final selected GAM model formula is expressed as

$$g(\mu_i) = \beta_0 + \beta_1 agecat + \beta_2 area + \beta_3 veh_{body} + \beta_4 gender + \beta_5 vehbody + f(veh\_value) \quad (22)$$

where the parameter β_0 denotes the mean and the other parameters β_i are the weights of the explanatory variables. In equation (22), is the function $f(\cdot)$ smoothing splines function of the continuous rating variable *veh_body*. In this paper, we analysed three different models for frequency and 3 different models for severity as well, in term of GAMs. Here are presented the following models, which will be analysed in this paper:

MODEL 1:uses these variables, *agecat, area, veh_body*that have the linear effects and variable *veh_value* has a non-linear effect.

MODEL 2: uses these variables *agecat, area, veh_body* and *veh_age* that follow linear effects, while variable*veh_value* has a non-linear effect.

MODEL 3: uses these variables*, agecat, area, veh_body,veh_age* and *gender* that have linear effects and the variable *veh_value* has a non-linear effect.

4.2.1 Model for Claim Frequency

This section will be presented the results of the different models presented in the previous section. Here, we aim at modeling the expected frequency $\lambda > 0$ such that it allows us to incorporate structural difference (heterogeneity) between different underlying risks. In this paper, we assume that the claims count random variable N has a Poisson distribution with expected frequency, where $\lambda > 0$.

The estimation results of the GAMs are presented in Table 1. (The signs ***, ** and * represent that the results are significant under the 1%, 5% and 10% confidence intervals respectively). The results demonstrate that the variable gender has an insignificant effect on variable *clm*, so we should remove this variable. For select the best model for frequency, we will use the Akaike's information criterion (AIC). The AIC is an approach to model selection in which models are selected to minimize an estimate of the expected

Kullback-Leibler divergence between the fitted model and the 'true model'. The criterion is expressed as

$$AIC = -2l + 2p, \tag{23}$$

where l is the maximized log likelihood of the model and p the number of model parameters that have had to be estimated. The model with the lowest AIC is selected.

Table 1

Estimation of parametric and non-parametricresults of GAMs for claim frequency models

Parametric coefficients	**MODEL 1**			**MODEL 2**			**MODEL 3**		
Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
β_0	-2.56***	0.0422	<2e-16	-2.68***	0.0557	<2e-16	-2.68***	0.0563	<2e-16
agecat1	0.19***	0.0529	0.0001	0.20***	0.0529	0.0001	0.20***	0.0528	0.0001
agecat2	0.03	0.0430	0.5095	0.03	0.0431	0.5247	0.03	0.0430	0.5258
agecat4	-0.02	0.0412	0.5567	-0.02	0.0412	0.5802	-0.02	0.0412	0.5900
agecat5	-0.22***	0.0490	7.20e-06	-0.22***	0.0491	8.08e-06	-0.22***	0.0491	9.37e-06
agecat6	-0.21***	0.0590	0.0003	-0.21***	0.0590	0.0004	-0.21***	0.0591	0.0005
areaA	-0.01	0.0389	0.8552	-0.01	0.0389	0.0014	-0.01	0.0389	0.8539
areaB	0.06	0.0406	0.1702	0.05	0.0406	0.1616	0.06	0.0406	0.1616
areaD	-0.12*	0.0512	0.0187	-0.12*	0.0512	0.0199	-0.12*	0.0513	0.0187
areaE	-0.04	0.0563	0.4654	-0.04	0.0563	0.0048	-0.04	0.0563	0.4536
areaF	0.08	0.0648	0.2044	0.08	0.0649	0.0471	0.07	0.0650	0.2543
veh_bodyBUS	0.99**	0.3177	0.0017	0.97**	0.3181	0.0023	0.97**	0.3182	0.0022
veh_bodyCONVT	-0.61	0.5874	0.3007	-0.68	0.5886	0.0445	-0.68	0.5887	0.2424
veh_bodyCOUPE	0.26*	0.1187	0.0268	0.23*	0.1198	0.0476	0.24*	0.1198	0.0466
veh_bodyHBACK	-0.02	0.0377	0.6232	0.01	0.0390	0.0294	0.01	0.0391	0.9503
veh_bodyHDTOP	0.07	0.0905	0.4295	0.03	0.0929	0.0385	0.03	0.0929	0.7186
veh_bodyMCARA	0.46 .	0.2603	0.0784	0.38	0.2628	0.0458	0.38	0.2628	0.1435
veh_bodyMIBUS	-0.21	0.1515	0.1620	-0.26 .	0.1542	0.0861	-0.27.	0.1542	0.0851
veh_body_PANVN	0.18	0.1240	0.1505	0.16	0.1242	0.0913	0.17	0.1247	0.1731
veh_bodyRDSTR	0.25	0.5802	0.6620	0.23	0.5804	0.0850	0.24	0.5804	0.6792
veh_bodySTNWG	-0.06	0.0430	0.1537	-0.10*	0.0479	0.0313	-0.10*	0.0479	0.0338
veh_bodyTRUCK	-0.07	0.0932	0.4491	-0.09	0.0946	0.3076	-0.09	0.0952	0.3502
veh_bodyUTE	-0.27***	0.0667	4.58e-05	-0.29***	0.0683	1.54e-05	-0.29***	0.0691	2.86e-05
veh_age2	-	-	-	0.16***	0.0446	0.0002	0.16***	0.0446	0.0002
veh_age3	-	-	-	0.13**	0.0503	0.0083	0.13**	0.0503	0.0077
veh_age4	-	-	-	0.19**	0.0658	0.0034	0.19**	0.0660	0.0029
genderM	-	-	-	-	-	-	-0.02	0.0301	0.5068
Non-parametric (smooth) coefficients	**MODEL 1**			**MODEL 2**			**MODEL 3**		
Variable	**edf**	**Ref.df**	**p-value**	**edf**	**Ref.df**	**p-value**	**edf**	**Ref.df**	**p-value**
s(veh_value)	2.982	3.761	1.99e-09	3.152	3.986	2.98e-07	3.162	3.997	2.42e-07

Source: processed by the author, using statistical program R 2019

In model comparative analysis, the results of AICs are presented in Table 2.

Table 2

AIC’s criterion for proposed frequency models

	MODEL 1	MODEL 2	MODEL 3
Degrees of freedom	25.98	29.15	30.16
AIC	36 072.57	26 061.95	36 063.47

Source: processed by the author, using statistical program R 2019

Model 2 has the lowest AIC, as can be seen from the Table 2, so the MODEL 2 is more accurate than others.

There is another method to compare our models by using the explained deviation. The explained deviations to the models are 0.681% for MODEL 1, 0.745% for MODEL 2 and 0.747% for MODEL 3. Both models MODEL 2 and MODEL 3 have higher explained deviations, which mean that their results are more accurate than the other model.

Comparing the models can be provided also by analysing the deviation using the function ANOVA in R. We can use this method to present the residual deviations and χ^2 of degrees of freedom of the models. Table 3 presents these results. The presence of ’**’ means they reject the null hypotheses under the 0.1% confidence interval.

Table 3

Analysis of Deviations in proposed models of claim frequency

	Residual df.	Residual Deviation	Difference df.	Difference deviation	$Pr(> \chi^2)$
MODEL 1	67 829	26 586	-	-	-
MODEL 2	67 826	26 569	3.2252	16.9520	0.0009 ***
MODEL 3	67 825	26 568	1.0107	0.5044	0.4818

Source: processed by the author, using statistical program R 2019

From the results declared in Table 3, we can see that MODEL 2 is better than MODEL 1 and MODEL 3, because it has the lowest p-value. Finally, we selected MODEL 2 as the best model for fitting the claim frequency. We can select the best GAM model, when we take a look at their UBRE scores. The lower UBRE score means the best model. The UBRE scores for our models are shown in the next Table 4.

Table 4

UBRE scores of claim frequency models

MODEL	UBRE score
MODEL 1	-0.60744
MODEL 2	-0.60759
MODEL 3	-0.60757

Source: processed by the author, using statistical program R 2019

From the Table 4 we can see, that the most accurate model for fitting the frequency is MODEL 2 again.

4.2.2 Model for Claim Severity

The average amount of claims is the quantity of interest in this part. Here, we aim at modeling with gamma distribution. Similar derivations will be done for individual claim size modeling, for detail we refer to Ohlsson, Johansson [10]. Severity is modeled with gamma distribution with log link function.

As in previous part, we can take look on the AIC among different models. These criterions are shown in the Table 5.MODEL 1 and MODEL 2 has the lowest AIC (their values are very similar), as can be seen from the Table 6, so the MODEL 2 is more accurate than others in our empirical study.

Table 5

AIC's criterion for proposed severity models

	MODEL 1	MODEL 2	MODEL 3
Degrees of freedom	26.26	29.47	30.19
AIC	78 881.86	78 884.02	78 861.34

Source: processed by the author, using statistical program R 2019

We can compare different severity models by comparing the explained deviation of these models. MODEL1 has 2.49% explained deviation. For MODEL 2 it is equal to 2.57% and for MODEL 3 2.98%. Both models MODEL 2 and MODEL 3 have higher explained deviations, which mean, that these models are enough.

The estimation results of the GAMs for severity are presented in Table 6. (There are also the signs ***, ** and * which represent that the results are significant under the 1%, 5% and 10% confidence intervals respectively).

Table 6

Estimation of parametric and non-parametric results of GAMs for claim severity models

Parametric coefficients	MODEL 1			MODEL 2			MODEL 3		
Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
β_0	7.41***	0.0769	0.0769	7.34***	0.0996	<2e-16	7.29***	0.0998	<2e-16
agecat1	0.31*	0.0959	0.0959	0.31**	0.0961	0.0010	0.29**	0.0953	0.0025
agecat2	0.13	0.0785	0.0785	0.12	0.0786	0.1093	0.11	0.0779	0.1385
agecat4	0.05	0.0753	0.0752	0.05	0.0754	0.5044	0.03	0.0748	0.6690
agecat5	-0.06	0.0891	0.0891	-0.06	0.0892	0.4819	-0.07	0.0886	0.3883
agecat6	0.03	0.1073	0.1073	0.04	0.1074	0.7321	-0.01	0.1068	0.9715
areaA	-0.09	0.0714	0.0714	-0.09	0.0715	0.2084	-0.09	0.0709	0.1938
areaB	-0.09	0.0738	0.0738	-0.09	0.0739	0.2029	-0.10	0.0733	0.1639
areaD	-0.12	0.0930	0.0930	-0.12	0.0931	0.1960	-0.10	0.0924	0.2776
areaE	0.10	0.1027	0.1027	0.09	0.1029	0.3522	0.09	0.1021	0.3595
areaF	0.32**	0.1195	0.1196	0.31**	0.1201	0.0088	0.32	0.1191	0.0057
veh_bodyBUS	-0.33	0.5912	0.5911	-0.35	0.5921	0.5487	-0.42	0.5873	0.4725
veh_bodyCONVT	0.20	1.0279	1.0269	0.08	1.0324	0.9320	0.17	1.0232	0.8616
veh_bodyCOUPE	0.32	0.2199	0.2199	0.28	0.2214	0.1923	0.27	0.2195	0.2060
veh_bodyHBACK	0.09	0.0689	0.0689	0.11	0.0715	0.1149	0.12	0.0708	0.0738
veh_bodyHDTOP	0.16	0.1635	0.1635	0.13	0.1674	0.4275	0.08	0.1659	0.6087
veh_bodyMCARA	-0.83 .	0.4769	0.4769	-0.90	0.4823	0.0622	-0.98	0.4785	0.0388
veh_bodyMIBUS	0.47 .	0.2735	0.2735	0.43 .	0.2796	0.1239	0.41	0.2771	0.1311
veh_body_PANVN	0.21	0.2291	0.2291	0.19	0.2297	0.3928	0.13	0.2282	0.5644
veh_bodyRDSTR	-1.40	1.2524	1.2524	-1.43	1.2542	0.2526	-1.40	1.2433	0.2594
veh_bodySTNWG	0.08	0.0793	0.0793	0.05	0.0882	0.5647	0.02	0.0874	0.7892
veh_bodyTRUCK	0.31 .	0.1708	0.1708	0.29 .	0.1730	0.0891	0.22	0.1726	0.2060
veh_bodyUTE	0.18	0.1219	0.1218	0.16	0.1247	0.1744	0.09	0.1247	0.4468
veh_age2	-	-	-	0.07	0.0815	0.3616	0.07	0.0808	0.3612
veh_age3	-	-	-	0.08	0.0916	0.3359	0.08	0.0907	0.3535
veh_age4	-	-	-	0.14	0.1202	0.2404	0.12	0.1188	0.3076
genderM	-	-	-	-	-	-	0.17	0.0545	0.0011
Non-parametric (smooth) coefficients	**MODEL 1**			**MODEL 2**			**MODEL 3**		
Variable	**edf**	**Ref.df**	**p-value**	**edf**	**Ref.df**	**p-value**	**edf**	**Ref.df**	**p-value**
s(veh_value)	3.258	4.128	0.174	3.469	4.407	0.431	3.19	4.087	0.383

Source: processed by the author, using statistical program R 2019

Using ANOVA function in R we can analyze and compare all presented models. Table 7 shows the results of the residual deviation and χ^2 of degrees of freedom of the models. Bind in mind that presence of '**' means they reject the null hypotheses under the 0.1% confidence interval.

Table 7

Analysis of Deviations in proposed models of claim severity

	Residual df.	Residual Deviation	Difference df.	Difference deviation	$\mathbf{Pr}(> \chi^2)$
MODEL 1	4 596.9	7 155.6	-	-	-
MODEL 2	4 592.6	7 119.8	3.2787	30.4549	0.0008 ***
MODEL 3	4 593.9	7150.2	0.6799	5.3917	0.6727

Source: processed by the author, using statistical program R 2019

From the results shown in Table 7, we can see that MODEL 2 is better than MODEL 1 and MODEL 3, because it also has the lowest p-value. Finally, it was selected MODEL 2 as the best model for fitting the claim severity.

There is also another way to judge which model is the best. This comparison is provided by GCV score. The lower GCV score means the best model is. We present this score for our models in the next Table 8.

Table 8

GCV scores of claim severity models

MODEL	GCV score
MODEL 1	1.5652
MODEL 2	1.5600
MODEL 3	1.5662

Source: processed by the author, using statistical program R 2019

4.2.3 Pricing Application and Pure Premium

The tariff is given by the model of the pure premium. Fitting model for frequency and severity can provide a better understanding of the way in which factors affect the cost of claims. This more easily allows the identification and removal (via smoothing) of certain random effects from one element of the experience. Ultimately, however, these underlying models generally need to be combined to give an indication of the pure premium relativities. In the case of multiplicative models for a single claim type, the calculation is straightforward – the frequency multipliers for each factor can simply be multiplied by the severity multipliers for the same factors (which is analogous to adding the parameter estimates when using a log link function). To analyze the premium directly, one might consider using a Tweedie models. The reason for the separation claim frequency and claim severity analyze is:

- The claim frequency is usually more stable than claim severity and often much of the power of rating factors is related to claim frequency – these factors can then be estimated with greater accuracy;

- A separate analysis gives more insight into how a rating factor affects the pure premium. [10]

A great advantage of the GLMs with multiplicative link is that it is easy to use it in practice and the interpretation is very intuitive. In the previous part of this paper we have calculate a reference level of premium and we have applied relativities to adjust down or up the premium based on the risk properties of insured.

The typical model form for modelling insurance claim counts of frequencies is a multiplicative Poisson. As well as being a commonly assumed distribution for claim numbers, the Poisson distribution also has a particular feature which makes it intuitively appropriate in that it is invariant to measures of time. In other words, measuring frequencies per month and measuring frequencies per year will yield the same results using a Poisson distribution. This is not true of some other distributions such as gamma.

Specifically, the claim frequency model is represented by the following equation:

$$\hat{y}_f = e^{-2.6878} \cdot (e^{0.2024})^{agecat1} \cdot (e^{0.0274})^{agecat2} \ldots \cdot (e^{0.1927})^{veh\_age4} \cdot\cdot \left(e^{f_{freq}(x)}\right)^{veh\_value}$$

We assume that we want to apply price segments as a function of vehicle value. In practice in the case of claim frequencies the prior weights are typically set to be the exposure of each record. In the case of claim counts the offset term is set to be the log of the exposure.

A common model form in this paper for modelling severities follows gamma distribution. As well as often being appropriate because of its general form, the gamma distribution also has an intuitively attractive property for modelling claim amounts since it is invariant to measures of currency. In other words, measuring severities in dollars and measuring it in cents will yield the same results. This is not true of some other distributions such as Poisson.

The similar form can be expressed for severity model which predicts the claim costs per policy where the various properties of policyholder are taken into consideration

$$\hat{y}_s = e^{7.3448} \cdot (e^{0.31548})^{agecat1} \cdot (e^{0.1259})^{agecat2} \ldots \cdot (e^{0.1412})^{veh\_age4} \cdot\cdot \left(e^{f_{sev}(x)}\right)^{veh\_value}$$

Although it is not as simple as in the case of GLMs, we can replicate a classical system of relativities in GAM structure. The pure premium which is mentioned above in this section can be generally expressed as

$$PURE\ PREMIUM = Frequency\ \times Severity.$$

5. Summary, Conclusion and Discussion

Nowadays, we can see a revolution in the general non-life insurance world (i.e., mostly a car insurance or a property insurance). The new technology is going to transform our daily life, and probably change the face of insurance in the medium term.

The aim of this paper is to provide some answers to the question of whether it is possible to change the way of calculating car insurance premiums. A free available database of an insurer from package 'insuranceData', has been used. The GAMs approach has been used to measure the impact of the *veh_value* on the risk of claims in automobile insurance.In the previous section we presented some conclusions regarding the use of three models to predict claim frequency and claim severity to produce pure premium.

We have also compared a pricing model widely used in practice with GAM approaches for proposed three models for both risk and severity models for select the best one. Initially, the GAMs approach based on independent Thin plate splines highlighted the existence of a non-proportional relationship between of the vehicle value and number of claims. GAMs are often difficult to interpret but in practice they offer more flexibility than other alternatives. Future work could determine the actual value (benefits, improved customer satisfaction, etc.) to implement pricing systems based on GAMs.

Finally, to highlight some of the benefits of GAMs, it should be noted that we have proposed some models with price structure based on their results. In that sense, we have considered a conventional price structure in which a reference premium is multiplied by relativities that have been obtained from the combination of the parametric part, that is, estimated parameters, known as price relativities and the effect of the vehicle value. The contribution of GAMs could be relevant in future research where more telemetric information could be introduced in the pricing system, such as sudden accelerations or braking, without including necessarily the moment and location of driving, which is, for many drivers, a privacy concern. The dependence between different types of claims could also be studied, in order to add the contracts of the same insuree. In such situations, the Generalized Additive Models for the Location, Scale and Shape (GAMLSS) proposed by Rigby and Stasinopoulos [12] could be considered.

The main aim of this paper is to start finding ways to correct the premium dynamically where are known characteristics of the driver and non-linear relationship is present. This point of view allows to analyze also the driving style of a driver. This fact improves pricing techniques in non-life insurance.

Acknowledgement

The paper was supported by a grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA no. 1/0120/18– *Modern risk management tools in the internal models of insurance companies in Solvency II.*

The paper was supported by a grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA no. 1/0647/19 – *Modern tools for managing and modelling of risks in non-life insurance.*

This contribution is the result of the internal grant project I-19-102-00 of the University of Economics in Bratislava for young pedagogical staff, scientific and PhD students entitled *Modern stochastic methods applied in tourism in Slovak Republic*.

References

[1] DUCHON, J. 1977.Splines minimizing rotation-invariant semi-norm in Sobolev spaces. In *Journal of the Royal Statistical Society: Series C (Applied Statistics).* Vol. 571, 1977, pp. 85-100. DOI 10.1007/BFb0086566.

[2] FREES, E. W. 2010. *Regression Modeling with Actuarial and Financial Applications,* Cambridge University Press, 2010. ISBN-13: 978-0521135962.

[3] FOX, J. 2016. *Appliedregressionanalysis and generalizedlinearmodels*(3rd ed.). Los Angeles: SAGE, 2016. ISBN-13: 978-1452205663.

[4] GU, C. 2013. *Smoothing spline ANOVA models*. New York, NY: Springer, 2013. ISBN 978-1-4614-5369-7.

[5] HASTIE, T.J. – TIBSHIRANI, R.J. 1990. *Generalized Additive Models*. Chapman and Hall/CRC, 1990. ISBN 9780412343902-CAT# C4390

[6] HASTIE, T.J. – TIBSHIRANI, R.J. 1986. Generalized Additive Models. *Statist. Sci.*Vol. 1, Number 3, pp.297-310. DOI 10.1214/ss/1177013604.

[7] WOOD, S.N. 2010. Fast stable restricted maximum likelihood and margine likelihood estimation of semiparametric generalized linear models. In *Journal of the Royal Statistical Society.* Vol. 73, Issue 1, pp.3-36. DOI 10.1111/j.1467-9868.2010.00749.x.

[8] WOOD, S. N. 2003.Thin plate regression splines. In *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Vol. 65, Issue 1, pp. 95-114. DOI 10.111/1467-9868.00374.

[9] WOOD, S. N. 2017. *Generalized additive models: An introduction with R*(2nd ed.). Boca Raton: CRC Press – Taylor &Francis Group, 2017. ISBN 978-1-4987-2833-1.

[10] OHLSSON, E. – JOHANSSON, B. 2015. *Non-life insurance pricing with generalized linear models*. Heidelberg: Springer, 2015. ISBN 978-3-642-

10790-0.

[11] R core team. 2019. *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.[online]. Available at:<http://www.R-project.org/>.

[12] RIGBY – STASINOPOULOS. 2005. Generalized additive models for location, scale and shape. *Statistics of Human Growth.* Vol. 54, Issue 3, pp.507-554. DOI 10.1111/j.1467-9876.2005.00510.x

[13] ŠOLTÉS, E. 2008. *Regresná a korelačná analýza.* Bratislava: IuraEdition spol. s.r.o., 2008. ISBN 978-80-8078-167-7.