

# Penalized Splines Fitting for a Poisson Response Including Outliers

Betül Kan-Kılınç  
Department of Statistics  
Eskisehir Technical University, Turkey  
Email: bkan@eskisehir.edu.tr

Huruy Debessay Asfha  
Department of Mathematics and Statistics  
Wichita State University, Wichita, KS, USA  
Email: asfhah32@gmail.com

## Abstract

There have been various studies in the literature on investigating the relationship between a count response and several covariates. Most researchers study count variables and use traditional methods (i.e. generalized linear models- GLM). However, GLM is limited when dealing with outliers and nonlinear relationships. Generalized Additive Models (GAM) is an extension of GLM, where the assumptions on the link functions and components are additive and smooth, respectively. Our aim is to propose a flexible extension of GLM and demonstrate the usefulness and performance of GAMs for the analysis of Poisson data set including outliers in the response variable through extensive Monte Carlo Simulations and using three applications.

**Keywords:** Poisson, Spline estimation, Deviance, Additive

## Introduction

The nature of the count data is that they appear positively, never zero, and always discrete number or integers within a period of time. There have been various studies in the literature on investigating the relationship between a count response and several covariates using traditional methods such as generalized linear models. Ordinary regression model can be extended to the family of generalized linear models when the conditional mean of the counts on a set of variables via a link function. This method, Poisson regression, is often lead to biased estimates of parameters due to its properties of equal mean and variance. Usually negative binomial regression is used to handle with overdispersion and underdispersion, but essentially it is not always the best choice for data.

Camarda et al. (2016) studied the conditional mean in Poisson regression to represent the response as a sum of smooth components and motivated their model in two applications and demonstrated the versatility of the approach. Currie et al. (2004) showed how p-splines can be used to model two-dimensional Poisson data. They introduced a model that described the expected values as a sum of components. Guisan et al. (2002) published a review paper that includes the application of GLM and GAM for ecological modelling. Sapra (2012, 2013) studied GAMs for several economic and business datasets. He compared the models in terms of AIC, deviances and  $R^2$ (adjusted).

In literature, different approaches of GAM estimating methods were proposed. Here, the smoothing spline basis approach which is a method incorporated in mgcv package is addressed (Wood 2018). One approach to estimate GAM models is by choosing a basis for the smoothing function and a wiggliness measure. In this approach, model estimation

implies estimation of smoothing parameter as well as model coefficients for a penalized likelihood maximization objective. The core point of the study is to build a flexible extension of generalized linear models and demonstrate the usefulness of GAMs for this analysis of count data set.

This paper is organized as follows: First we shortly introduce GLM and GAM. The penalized regression splines used for the estimation of GAMs are presented in the next. The first simulation study where the dependent variable suffers from outliers and regresses on one variate conducts an estimation scenario for a Poisson distributed response using GLM and GAMs fitted by three penalized splines, cubic regression splines, p-splines, thin-plate splines. Second simulation study is organized in the same manner except regressing the dependent variable on two covariates. Application part illustrates the comparisons of GLM and GAMs using commonly known datasets. Last section provides some concluding remarks.

### Generalized Linear Model and Generalized Additive Models

GAM is an extension of GLM, where the assumptions on the link functions and components are additive and smooth, respectively (Hastie et al. 1990, 2001). GLM is introduced to relax the assumptions of normality and homoscedasticity in ordinary linear regression. In GLM, distribution of response variable has to be one of the exponential family distributions such as normal, binomial, exponential, gamma and Poisson (Nelder et al. 1972, Montgomery et al. 2012). Also, the similarity in considerably several properties of the exponential family distributions allows us to use the same technique to estimate model coefficients using the likelihood concept of estimation. The general form of GLM is given by,

$$g(\mu_i) = \eta_i = \mathbf{X}_i \boldsymbol{\beta} = \beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_p \mathbf{X}_p \tag{1}$$

where  $\mathbf{X}_i$  is the  $i^{\text{th}}$  row of  $\mathbf{X}$  (data matrix),  $\boldsymbol{\beta}$  is the unknown parameter vector. Note that the expected response is given by,

$$E(y_i) = g^{-1}(\mu_i) = g^{-1}(\eta) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \tag{2}$$

where  $y_i \sim \text{Exponential family}(\mu_i, \phi)$ ,  $g$  is a link function, some known monotonic, which connects the systematic component,  $\eta$  (linear predictor), and the random component (response variable) of the model (Keele 2008). When the response variable follows normal distribution and an identity link function is used, the generalized linear model turns out to be an ordinary linear regression model.

This approach extends to additive models to form generalized additive models (GAM). When the linear predictor,  $\mathbf{X}\boldsymbol{\beta}$ , in GLM model are replaced with additive predictors (explanatory variables), the model is called generalized additive model (GAM), thus, GAMs are regarded as extensions of GLMs. In the GAM, the linear predictor becomes:

$$\beta_0 + \sum_{j=1}^p f_j(x_j) \tag{3}$$

where  $x_j$  is a covariate. The  $f_j$ 's in generalized additive models are smoothing functions which can be any of kernels, local regression (loess) or smoothing splines which are derived from the model. In this study we use cubic regression splines, p-splines and thin-plate regression splines. The model is obtained by precise characterisation of what is meant by a smooth function, in the form of measures of (wiggleness) spline penalty  $\int f''(x)^2 dx$ . The  $f_j$ 's is linear basis expansion

$$f_j(x) = \sum_{k=1}^K b_{jk}(x)\gamma_k \quad (4)$$

where the  $\gamma_k$  are the coefficients to be estimated and  $b_{jk}(x)$  are the basis functions (approximations). The shape of predictor functions is determined by the data. Hence there is no assumption on the specific link function for error distribution. The penalties become quadratic forms,  $\beta S_j \beta$  where the  $S_j$  are the matrices of known coefficients.

Specifically, for a dependent count variable with observation  $i$ , let  $Y_j$  represent the number of independent events that occur during a fixed time period.  $Y_j \sim \text{Poisson}(\lambda_j)$  where  $\lambda_j$  is the mean and the variance of this distribution. This parameter is given by

$$\lambda_i = \exp(\mathbf{X}_i \boldsymbol{\beta} + \sum_{j=1}^K f_j(Z_j)) \quad (5)$$

where  $X_i$  are the explanatory variables and  $f_j(Z_j)$  are the smooth terms for  $j = 1, \dots, K$  and  $\boldsymbol{\beta}$  is the vector of coefficients. Note that the Eq. (5) is the expected response.

To penalize overfit, estimation is made by penalized maximum likelihood method,

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) - \frac{1}{2} \sum_j \lambda_j \boldsymbol{\beta}^T S_j \boldsymbol{\beta} \quad (6)$$

Assuming the  $\lambda_j$ , smoothing parameters, are known, estimates of  $\boldsymbol{\beta}$ ,  $\hat{\boldsymbol{\beta}}$ , can be found by maximizing Eq. (6).

In GAM, while the assumption of standard linear models of the linear dependency of  $y$  on  $\mathbf{X}$  is relaxed, the additivity assumption still holds true and it is this additivity property that makes GAM easier to interpret than other algorithms such as support vector machines (SVM), neural networks, etc. (Keele 2008). The main difference between GLM and GAM is that in GAM, linear predictor is the sum of smoothing functions, not limited to being linear. GAM can handle non-linear, linear and nonmonotonic relationships between response and explanatory variables. Also, GLM emphasizes estimation and inference for the parameters of the model, while GAMs focus on exploring data nonparametrically (Guisan 2002).

Wood (2003, 2006) employs the approach of penalized regression spline to fit a model. By default, the degree of smoothness of the fit is chosen internally by the algorithm. Automatic selection of smoothing parameter is an advantage for reason being that it avoids the subjectivity and work of choosing it by the user. However, it can fail to obtain the best degree of smoothness and human intervention could sometimes be needed (Faraway 2006).

## Selection of GAM and model parameter estimation

Model evaluation plays a fundamental role in regression analysis; comparisons can be made between models to obtain the better of them. In linear regression, mean square error (MSE) is regarded as the building blocks of most model evaluation techniques and it measures how far the model estimations from the actual observations are. In GLM and GAM, it is necessary to have a quantity which is equivalent in importance and interpretation to residual sum of squares for ordinary linear modelling (McCullagh et al. 1989).

As minimizing MSE is to least square fits, in models fitted using maximum likelihood estimation (MLE), the quantity to be minimized is the deviance. Maximizing the likelihood in those models corresponds to minimizing the deviance of the model (Wood 2006). Model deviance is defined as twice the difference in log-likelihood between the saturated model and the full model (Montgomery et al.2012). Deviance of a model can be regarded as the lack of fit between the model and the data points. It is used for model adequacy checking; the smaller the deviance the better the model is. It is known that Akaike Information Criterion (AIC) and deviance can be used with GLM and GAM (Sakamoto et al. 1988; Guisan et al. 2002)

When large number of knots are used to represent the smoothers,  $f_j$  and the method of maximum likelihood is used to estimate  $\beta$ , the model parameters, then, there is possibility of over-fitting. This is the reason why penalized likelihood maximization is preferred over the ordinary likelihood maximization to estimate GAMs (Wood 2006).

## Monte Carlo Simulations

The simulations are performed to explore the performance of the proposed GLM and GAM fits in a variety of situations. Furthermore, the effects of the different proportions of outliers existing in response variable from Poisson family to the fitted models will be addressed.

Data generated with different mean functions and proportions of outliers or without outliers, using the procedures are discussed in the following sections, respectively. Generated data are used to evaluate the performance of the three smoothing splines; cubic regression, p-splines, thin plate splines and generalized linear model. The mean of AIC, residual deviance and explained deviance obtained for each model are provided for the goodness of fit. For the analysis and model building procedures, the *mgcv* package in R (R Development Core Team 2018) statistical software is performed.

## Poisson response with one covariate

The response variable was simulated from,  $Y/x \sim Pois(\lambda(x))$ , where  $\lambda(x) = g^{-1}(h(x))$  and  $h(x) = 4\cos(2\pi(1-x)^2)$ . Here,  $g$  is a log-link function. The covariate follows a uniform distribution,  $x \sim U(0,1)$ . The randomly selected  $Y$  values were multiplied by  $yu_1^{u_2}$  where  $u_1 \sim U(2, 5)$  and  $u_2 \sim U(-1, 1)$  and the result was rounded to the nearest integer. Five sample sizes of  $n$  were tested: 25, 50, 100, 250, and 500. For the univariate function, Wong et al. (2014) was followed.

For a specified outlier proportion value given by  $p$ : 0, 0.1, 0.2, a total of  $p100\%$  simulated data was randomly selected to be changed to outliers. The number of repetitions for each

scenario was 500 for all possible combinations of  $n$  and  $p$ . Each of the three smoothing splines were used to fit a GAM model for all the samples.

**Poisson response with two additive functions**

To illustrate the performance of the proposed methods, two covariates are considered by  $(x_1, x_2)$  for  $i = 1, \dots, n$  and response variable is generated from  $Y/(x_1, x_2) \sim Pois(\exp(t_1(x_1), t_2(x_2)))$ , where  $\lambda(x) = t_k(x)$ ,  $k = 1, 2$ ,  $t_1(x) = \exp(\sin(\frac{4x_1}{60}) + \cos(\frac{7x_1}{120}) + 1)$  and  $t_2(x) = \sin(180 - x_2^7\pi)^{11} + 29\exp(\sin(12x_2^3))$  where  $x_1 = i$  and  $x_2 \sim U(0,1)$ .

In this case, outliers were included into the response variable by using (Alimadad et al 2011). Let  $(\tilde{y}_j, x_j)_{j=1}^n$  be the data and  $x_j$  is ordered where  $x(m)$  is the  $m^{th}$  order statistics. Next,  $y_j = (1 - z_j)\tilde{y}_j + z_jw_j$  where  $z_j \sim B(1, p)$  and  $w_j = Pois(30)$ . The number of outliers in each sample is random.

It is important to note that, in each case, sample sizes of 20, 50, 100, 200, and 500 were simulated and a total of 500 repetitions were generated. Each of the three smoothing splines were used to fit a GAM model for all the samples.

**Results**

Mean of deviances and AIC of the models using GLM, and GAMs fitted by cubic regression spline (cr), p-spline (ps) and thin-plate spline (tp) bases for one covariate are presented in Table 1. Additionally, the percentage of explained deviance for each model is provided in the same table.

**Table 1: Mean deviance, AIC, (%) of explained deviance with/without outliers-univariate case**

p	Dev				AIC				Dev.Exp.(%)			
	glm	cr	ps	tp	glm	cr	ps	tp	glm	cr	ps	tp
<b>n=25</b>												
0.	367.13	13.06	13.28	12.74	449.85	106.78	106.78	106.31	0.46	0.98	0.98	0.98
0									4	1	0	1
0.	461.37	49.95	47.11	47.93	543.82	143.82	141.28	142.76	0.42	0.94	0.94	0.94
1									7	4	7	6
0.	600.21	100.52	94.39	95.90	681.91	195.05	189.03	191.03	0.39	0.90	0.90	0.90
2									3	4	9	8
<b>n=50</b>												
0.	797.47	33.21	33.01	33.08	958.25	207.08	206.57	207.37	0.42	0.97	0.97	0.97
0									9	6	6	6
0.	1096.3	187.22	181.54	183.95	1257.2	361.93	356.33	359.16	0.37	0.90	0.90	0.90
1					7				3	0	3	2
0.	1302.3	325.57	316.41	320.13	1462.0	499.22	490.33	494.44	0.35	0.84	0.85	0.84
2	8				1				6	6	1	9
<b>n=100</b>												
0.	1654.4	74.63	72.95	74.43	1972.4	407.03	405.29	407.18	0.41	0.97	0.97	0.97
0	3				8				6	4	4	4
0.	2249.0	456.72	443.83	451.23	2567.0	789.23	776.49	784.26	0.36	0.87	0.88	0.87
1	9				4				8	7	0	8
0.	2762.5	797.04	777.25	788.55	3078.1	1127.2	1107.7	1119.3	0.33	0.81	0.81	0.81
2	8				5	6	9	1	6	3	7	5

n=250												
0.	4224.0	201.86	195.93	201.57	5009.9	1002.8	996.92	1002.9	0.40	0.97	0.97	0.97
0	4				7	1	6	8	8	2	3	2
0.	5700.7	1233.9	1217.8	1230.1	6485.4	2033.7	2017.8	2030.3	0.35	0.86	0.86	0.86
1	2	3	6	5	3	8	6	8	5	3	5	3
0.	7052.3	2226.7	2195.7	2218.1	7841.2	3024.9	2994.0	3016.6	0.32	0.79	0.79	0.79
2	4	8	4	4	9	2	8	7	8	0	3	1

  

n=500												
0.	8484.0	414.72	402.13	414.58	1053.7	1999.8	1987.2	2000.0	0.40	0.97	0.97	0.97
0	2				7	5	7	5	6	1	2	1
0.	11484.	2609.7	2583.9	2605.9	13051.	4192.3	4166.5	4188.8	0.35	0.85	0.85	0.85
1	68	8	7	6	81	0	9	2	5	5	6	5
0.	14288.	4690.6	4653.8	4683.4	15853.	6270.9	6234.2	6264.0	0.32	0.78	0.78	0.78
2	90	2	7	5	74	2	8	4s	7	0	2	0

The results in Table 1 show that an increase in the proportion of outliers included in the response variable has inflated the mean deviances and AIC of all the models. To illustrate this, one sample size (e.g,  $n = 25$ ) can be considered for comparison of the outcomes when different number of outliers are included in the data. In cases where an outlier is not included, the mean deviances of GLM, cubic,  $p$ -spline and thin plate splines are respectively 367.13, 13.06, 13.28, and 12.74, however, when some outliers ( $\% p = 0.1$ ) are introduced in the response variable their respective mean deviances are 461.37, 49.95, 47.11, and 47.93 respectively. Similarly, for  $\% p = 0.2$  as well, the mean deviances are seen to be increasing in all models. Similarly, results are obtained using  $n (=50, 100, 250, 500)$  support this argument. This comparison indicates that GAM fitted by  $p$ -splines outperformed GLM.

Results from Table 1 demonstrate that for all sample sizes  $n$ ,  $p$ -splines based models produced a smaller mean deviance than others at  $\% p (=0.1, 0.2)$ . In addition to this, the models with a smaller mean AIC are produced by using  $p$ -splines in all combinations of outlier proportion ( $p$ ) and sample sizes except for  $n = 25$  and  $\% p = 0.0$ . Furthermore, the mean of the proportion of explained deviance scored by each method shows that GAMs fitted by  $p$ -splines outperformed the others.

Results from scenario 2 given in Table 2, demonstrate that GLM and GAMs fitted by cubic splines,  $p$ -splines and thin-plane splines for two covariates.

**Table 2: Mean Deviance, AIC, (%) of Explained Deviance with/without outliers- two covariates**

p	Dev				AIC				Dev.Exp.(%)			
	glm	cr	ps	tp	glm	cr	ps	tp	glm	cr	ps	tp
n=25												
0.	210.38	14.76	13.66	14.04	358.08	181.24	179.26	179.89	0.07	0.93	0.93	0.93
0									9	1	8	5
0.	215.72	20.91	21.17	20.90	362.21	190.48	190.20	190.14	0.07	0.90	0.90	0.90
1									7	7	6	6
0.	213.36	27.35	29.05	27.73	358.70	198.09	198.71	198.06	0.07	0.88	0.87	0.88
2									0	0	0	0
n=50												
0.	491.86	55.99	53.97	51.96	776.44	361.28	359.30	357.03	0.06	0.89	0.89	0.89
0									0	0	5	9

**Penalized Splines Fitting for a Poisson Response Including Outliers**

0.	477.83	89.06	88.37	86.73	760.72	395.98	395.02	393.40	0.05	0.81	0.82	0.82
1									6	9	2	5
0.	456.20	118.40	119.93	116.63	737.36	425.79	426.53	423.89	0.05	0.74	0.74	0.75
2									3	8	6	3
<b>n=100</b>												
0.	1063.9	137.54	137.63	130.21	1620.2	717.51	717.19	710.01	0.03	0.87	0.87	0.88
0	9				8				1	3	4	0
0.	1029.9	222.46	224.77	217.67	1583.5	802.06	803.90	797.38	0.03	0.78	0.78	0.79
1	4				1				1	8	6	3
0.	970.43	293.19	295.86	289.66	1521.2	871.46	873.33	867.84	0.02	0.70	0.70	0.70
2					2				9	3	1	7
<b>n=250</b>												
0.	2775.7	409.48	410.57	396.45	4167.1	1831.5	1831.3	1818.2	0.00	0.85	0.85	0.85
0	9				4	0	6	5	8	3	3	8
0.	2664.6	678.64	674.64	666.15	4048.6	2093.2	2088.1	2080.6	0.00	0.74	0.74	0.75
1	5				6	5	8	0	8	6	8	1
0.	2536.9	887.22	883.06	877.38	3913.1	2293.9	2288.7	2284.0	0.00	0.65	0.65	0.65
2	5				0	6	4	0	8	1	3	5
<b>n=500</b>												
0.	5625.0	981.79	1049.6	977.08	8388.5	3777.0	3842.6	3773.0	0.00	0.82	0.81	0.82
0	9		9		5	8	7	1	6	6	4	7
0.	5338.3	1485.7	1535.6	1480.3	8087.5	4266.6	4314.2	4260.8	0.00	0.72	0.71	0.72
1	4	1	1	3	4	9	3	7	6	3	3	4
0.	5043.9	1884.7	1922.2	1881.5	7779.0	4651.5	4686.7	4648.0	0.00	0.62	0.62	0.62
2	3	6	9	4	4	3	5	1	5	8	0	8

It shows that for  $n = 25$  cubic splines based models produced a smaller mean deviance than others at  $p = 0.2$ . For sample size of  $n = 50, 100, 250, 500$  however, GAMs fitted by thin-plate splines performed better in all combinations of outlier proportion ( $p$ ). In addition to this, the models with a smaller mean AIC are produced by using thin-plate splines in all combinations of outlier proportion ( $p$ ) and sample sizes except for  $n = 25$ . Furthermore, the mean of the proportion of explained deviance scored by each method is another evidence that GAMs fitted by thin-plate splines outperformed the others. Compared to all models, the GAM using thin-plate regression splines produced the highest explained deviances than the rest of models. This comparison also indicates that GLM produced poor goodness of fit (Dev.Exp. % 0.01-0.08) while GAMs produced more desirable results when the response has some proportion of outliers when regressing on two covariates.

On the contrary, results from the univariate case do not show the same trend as that of the case of two covariates, rather, GAMs fitted by p-splines seems to be outperformed by both cubic and thin plate splines regardless of the sample sizes and the outlier proportion except for  $n = 25$  and  $p = 0.0$ . The percentage of the explained deviance is almost equal to each other for splines base model.

**Comparison on Data Sets**

In this section, we apply the provided fitting procedure to several datasets from literature. Each data consists of a response from Poisson distribution with/without outliers. The description of the data sets is given in Table 3.

**Table 3: The description of data sets**

	<b>Ships</b>	<b>Frisk</b>	<b>AIDS</b>
Number of observations	40	900	570
Number of predictors	2	2	1
Number of outliers	5	many	many

The Table 4 presents the results of the models from GLM and GAMs. In order to obtain the goodness of fits of the models, deviance and AIC are evaluated for each model sand summarized in the table.

The ships data concerning a type of damage caused by waves to the forward section of cardio-carrying vessels (McCullagh et al, 1989). The variables are incidents, service, and year. This data suffers from outlier problem. To investigate whether the expected number of incidents per aggregate months of service makes sense on year and service, using GLM and GAM fitted by cubic splines, p-splines and thin- plate splines are compared. Notice that AIC (447.75) and deviance (343.88) values obtained by GLM are quite larger which is not preferred than any GAMs. Compared to all models, the GAM using cubic regression splines produced the highest explained deviance as 92.9% than the rest of models.

The frisk data is from Gelman and Hill (2006)’ study on police stops in New York City. The data give the counts of police stops for past arrests. The data also suffers from outliers. The model obtained from GLM produced higher values of AIC and deviance (64820.00, 63109.70) whereas these values from GAM models show better performance. Among the GAMs, the p-splines GAM models produced the smallest AIC (38088.04) and highest explained deviance (70%).

Another data set that includes the reported cases of AIDS diagnosed in 1983 and 1992 (De Angelis et al 1994). The data are the time delay in reporting of diagnosis and the number of AIDS cases. The data suffers from outlier in the response variable. Considering the GLM fit, it performs the data poorly. However, GAMs fitted by p- splines produced the smallest AIC (4834.453) and highest explained deviance as (57.6%).

**Table 4: The comparison of GLM and GAM**

<b>Criteria</b>	<b>Ships</b>	<b>Frisk</b>	<b>AIDS</b>
<b>GLM</b>			
AIC	447.75	12134.56	5137.492
Deviance	343.88	37448533	271136.2
% Deviance	0.529	0.11	0.258
<b>GAM (cr)</b>			
AIC	159.18	60207.71	4957.391
Deviance	38.92	54803.79	192311.7
% Deviance	0.946	0.699	0.474
<b>GAM (tp)</b>			
AIC	172.09	60120.81	4956.201
Deviance	50.47	54716.6	191896.1
% Deviance	0.917	0.699	0.475
<b>GAM (ps)</b>			



AIC	177.83	59955.86	4834.453
Deviance	59.18	54555.2	154933.8
% Deviance	0.918	0.705	0.576

## Conclusion

In this paper, we compare the performance of the methods GLM and GAM for a Poisson response variable with/without outliers when regressing on one variate and two covariates. Considering the GAM fits, three penalized regression spline smoothers are evaluated where the first is cubic splines, another is the p-splines, and the third is the thin-plate splines. Three common datasets have been also used to compare the performance of GLM and GAM on the Poisson response including outliers. Using these datasets, the comparison of the performance of GLM and GAM which employs a penalized smoothing spline approach are investigated. Our simulation studies and the illustrations using three data sets from literature reflect the parallel results to each other. The results show that GAMs are more resistant to outliers for a response variable from Poisson family.

## References

1. Alimadad, A. and Salibian-Barrera, M. (2011). An outlier-robust fit for generalized additive models with applications to disease outbreak detection, *Journal of the American Statistical Association*, 106(494), 719-731.
2. Camarda, G.C., Eilers, P.H.C. and Gampe, J. (2016). Sums of the smooth exponentials to decompose complex series of counts, *Statistical Modeling*, 16(4), pp. 279-296.
3. Currie, I.D., Durban, M. and Eirlers P.H.C. (2004). Smoothing and forecasting mortality rates., *Statistical Modelling*, 4(4), pp. 279-298.
4. De Angelis, D. and Gilks, W.R. (1994) Estimating acquired immune deficiency syndrome accounting for reporting delay. *Journal of the Royal Statistical Society, A*, **157**, 31–40
5. Faraway, J.J. (2006). *Extending the Linear Model with R-Generalized Linear, Mixed Effects and Nonparametric Regression Models*, New York, NY: Taylor & Francis Group, LLC.
6. Gelman A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New York: NY, USA
7. Guisan, A., Thomas, C., Edwards, J. and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene, 157, 89-100
8. Hastie, T.J., Tibshirani, R.J., 1990. *Generalized Additive Models*. Chapman & Hall.
9. Hastie, T.J., Tibshirani, R.J., Friedman, J., 2001. *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.
10. Keele, L. (2008). *Semiparametric Regression for the Social Sciences*, England: John Wiley & Sons Ltd.
11. Nelder, J.A., Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, 3, Series A 135, 370-384.

12. McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2<sup>nd</sup> ed. Chapman & Hall, London.
13. Montgomery, D.C., Peck, E.A, Vining, G.G. (2012). Introduction to Linear Regression Analysis. 5th ed. New Jersey, John Wiley & Sons, Inc.
14. Sakamoto, Y., Ishiguro, I., Kitagawa, G. (1986). Akaike Information Criterion Statistics. KTK Scientific Publishers. Japan.
15. Sapra, K. (2013). Generalized Additive Models in Business and Economics, International Journal of Advanced Statistics and Probability, 1(3), 64-81.
16. Sapra, K. (2012). A Generalized Additive Poisson Regression Model: An Application to Target Management Resistance, Empirical Economics Letters, 11(11), 1105-1114.
17. Wong, R.K.W., Yao, F., Lee, T.C.M. (2014) “Robust Estimation of Generalized Models. Journal of Computational and Graphical Statistics, 23 (1), 270-289.
18. Wood, S.N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1), 95-114, Wiley Online Library.
19. Wood, S.N. (2006). Generalized Additive Models; An Introduction with R. New York, NY: Chapman & Hall/CRC.
20. Wood, S.N.(2018). Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R Package *mgcv*. Version 1.8-24.
20. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>