Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics
Turkiye Klinikleri J Biostat. 2020;12(2):183-92

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2020-74122

Selection of Splines Models in the Presence of Multicollinearity and Outliers

Çoklu Bağlantı ve Aykırı Değer Varlığında Splayn Modellerin Seçimi

Betül KAN KILINÇ[a], Huruy DEBESSAY ASFHA[b]

[a]Eskişehir Technical University Faculty of Science, Department of Statistics, Eskişehir, TÜRKİYE
[b]Wichita State University, Department of Mathematics, Statistics and Physics, Wichita, Kansas, USA

ABSTRACT Objective: In this paper, the selection of the models evaluated using by three penalized regression splines; cubic splines, p-splines, and thin-plate splines are compared to linear models where the multicollinearity exists among covariates at different parameters and a response variable including outliers. **Material and Methods:** Generalized additive models (GAM) are extension of additive models as generalized linear models are to ordinary linear regression models. Different approaches of fitting these kinds of models that is the penalized regression techniques for representing generalized additive models are used in this study. **Results:** To examine the tolerance of the effect of multicollinearity and outliers and for the selection of these models and linear regression models, AIC and deviance are used. In all the situations, cubic splines regression models produced a smaller mean deviance in the presence of multicollinearity. On the other hand, cubic splayn regression models loses its dominance of producing smaller mean deviance when outliers are included to the data. It is remarkable that with the increase of sample size the number of times the p-splines method produced a smaller deviance. **Conclusion:** Results of the simulations showed that the GAMs fitted using these nonparametric regression techniques are less prone to multicollinearity and outliers compared to their parametric counterparts.

Keywords: Outlier; correlation; deviance; penalized

ÖZET Amaç: Bu çalışmada, yanıt değişkeninin aykırı değer ve açıklayıcı değişkenler arasında farklı düzeylerde çoklu bağlantının varlığı söz konusu olduğunda, üç farklı regresyon splaynlarının; kübik splayn, p-splayn ve ince tabakalı splayn tabanlı modellerin, doğrusal regresyon modelleri ile karşılaştırılması ve model seçimi üzerinde durulmuştur. **Gereç ve Yöntemler:** Doğrusal regresyon modellerinin genelleştirilmiş doğrusal modellerin bir uzantısı olması gibi genelleştirilmiş toplamsal modeller de toplamsal modellerin bir uzantısıdır. Genelleştirilmiş toplamsal modelleri oluşturmak için, cezalı regresyon splaynları bu tür modellerin veriye uyumu için kullanılan değişik yaklaşımlardan bazıları olup bu çalışmada kullanılmıştır. **Bulgular:** Çoklu bağlantı ve aykırı değerin modeller üzerindeki etkilerini incelemek ve model seçimi yapabilmek için AIC ve sapma ölçüleri kullanılmıştır. Bütün durumlarda kübik splayn regresyon modelleri, çoklu bağlantı varlığında, diğerlerine göre daha küçük sapma değerleri elde etmiştir. Öte yandan, kübik splayn regresyon modelleri, veriye aykırı değerler dahil edildiğinde, daha küçük sapmalı model etme başarısını gösterememiştir. P-splayn modellerinin örnek hacmi artırıldığında, daha küçük sapma değerleri veren modeller elde etmesi dikkat çekicidir. **Sonuç:** Benzetim çalışması sonuçları, parametrik olmayan regresyon tekniklerinin, doğrusal regresyon modellerine göre aykırı değer ve çoklu bağlantıdan daha az etkilendiğini göstermiştir.

Anahtar kelimeler: Aykırı değer; korelasyon; sapma; cezali

When examining a multivariate data, the interaction of several factors has to be recognized carefully and properly analysed. It is often assumed that multivariate data should follow a known specific statistical distribution. However, the complex structure of the data and the relationship among various factors may be too complicated to explain general patterns or to model dependency between variables.

Correspondence: Betül KAN KILINÇ
Eskişehir Technical University Faculty of Science, Department of Statistics, Eskişehir, TÜRKİYE
E-mail: bkan@eskisehir.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 04 Feb 2020 ***Received in revised form:*** 23 Mar 2020 ***Accepted:*** 25 Mar 2020 ***Available online:*** 28 May 2020

A complication occurring frequently in multiple regression processes is the violation of the independency assumption among regressors. The near linear dependencies is a particular form of data weakness which is called as collinearity.[1] Generally, researchers either ignore this problem or eliminate one or more variables causing multicollinearity. It is, however, worth noting that by ignoring this problem causes incorrect findings. Some general approaches combating multicollinearity are given as collecting additional data, model respecification (create a new variable by using a function including nearly linearly dependent ones) and using some biased estimation methods (ridge regression, principal component regression etc).

There are also other causes of model disturbances one of which is the existence of abnormal observations in the dataset which distorts the model parameters. Furthermore, this in turn may result in an inflated estimate of σ^2, the residual sum of squares. In literature, some outlier-resistant GAM fitting techniques have been developed. Alimadad discussed a robust method of GAM fitting technique by using those derived from robust quasi-likelihood equations.[2] Wong et al. proposed an M-type robust estimating technique to fit a more robust generalized additive model in the presence of outliers.[3]

Nonparametric regression models allow one to fit a flexible nonlinear model to the data in order to represent the relationship between the response and predictor variables. As parametric models, nonparametric models are useful both for modeling and diagnosis of the nonlinear relationships.

Generally, GAMs are computationally expensive techniques compared to the linear models due to the fact that they build the model by using local fits. However, different algorithms have been developed to fit GAM models iteratively. The gam package was developed based on the work of Hastie to fit generalized additive models to the data of concern.[4,5] This gam function constructs GAMs by combining different smoothing methods using backfitting algorithm. Another package used to fit GAM models is the mgcv package of Wood which employs the approach of penalized regression splines to fit a model.[6]

Our aim is to investigate and compare the nonparametric and parametric models in terms of some model accuracy measures. Specifically, it is aimed to find the best fitting model by penalized regression splines and true regressors explaining the dependent variable and compare it with linear regression models.

This study is organized as follows: We introduce generalized linear models (GLM) and Generalized additive models (GAM), respectively. Three penalized splines are addressed in the following sub sections. Model adequency measures are given next. Then, a simulation study has been conducted and concluding remarks are offered in the final section.

MATERIAL AND METHODS

GENERALIZED LINEAR MODELS

Generalized linear models are introduced to relax the strict assumptions of normality and homoscedasticity in ordinary linear regression. In GLM, distribution of response variable has to be one of the exponential family distributions among which are normal, binomial, exponential, gamma and Poisson.[7,8]

The general form of GLMs is,[9,10]

$$g(\mu) = \eta = \boldsymbol{X}\beta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad (1)$$

where, g is a link function which connects the systematic component, η (called linear predictor), and the random component (response variable) of the model.[9-12] The expected response is then obtained as, $E(y) = g^{-1}(\eta) = g^{-1}(\boldsymbol{X}\beta)$. Here, if 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 (lm).

GENERALIZED ADDITIVE MODELS

When the linear predictors in GLM model given in Eq.1 are replaced with additive predictors, the model is called generalized additive model (GAM), thus, GAMs are regarded as extensions of GLMs.

$$g(\mu) = \eta = \beta_0 + f_1(x_1)\beta_1 + f_2(x_2)\beta_2 + \ldots + f_p(x_p)\beta_p + \varepsilon \quad (2)$$

The f_i's in generalized additive models are smoothing functions which can be any of kernels, local regression (loess) or smoothing splines. Due to the fact that GAMs can incorporate nonparametric models into parametric ones, they can sometimes be described as semiparametric regression models.[10]

In GAMs, while the assumption of standard linear models of the linear dependency of $\boldsymbol{y}$ on $\boldsymbol{X}$ is relaxed, the additivity assumption still holds true and it is this additivity property that makes GAMs easier to interpret than other algorithms such as supportive vector machines (SVM), neural networks, etc.[12]

Wood (2006) employs the approach of penalized regression spline to fit a model.[9] 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.[5]

Natural Cubic Spline (Cr)

Natural cubic splines are a special case of cubic splines where the second derivative at the two end points are constrained to have zero value. They are called natural due to being a solution of an optimization problem.[12,13] In general, a natural cubic spline satisfies the following:

- It interpolates the points (x, y) i.e. $g(x_i) = y_i$
- Its second derivative at the two end points is zero;
- $g''(x_1) = g''(x_n) = 0$.
- Natural cubic spline is the smoothest interpolator. If $f(x)$ is any continuous function on the interval $[x_1, x_n]$ and has continuous first and second derivatives, and interpolates the points (x_i, y_i), then the natural cubic spline $g(x)$ is smoothest when it comes to minimize the roughness measure.

$$\int_{x_1}^{x_n} g''(x_1)^2 dx \leq \int_{x_1}^{x_n} f''(x_1)^2 dx, \quad \forall f \in C^2 \quad (3)$$

Green (1995) demonstrated the smoothness of a cubic spline in his work.[13]

P-Splines (Ps)

For (natural) cubic spline, there is high tendency for the columns of the model matrix $\boldsymbol{X}$, to be correlated for they are in someway a transformed version of the predictor variable(s).[12] This dependency may cause multicollinearity or concurvity which may result in numerical instability and imprecision in the spline fit.[10,12] To somehow get rid off this problem, a B-spline basis which is a refined form of a cubic spline, can be employed. This kind of splines can be used to represent cubic splines as well as higher order splines.

B-spline basis, a strictly local type of spline is non-zero only on the intervals between m + 3 adjacent knots where $m + 1$ is the order of the basis (i.e. $m = 2$ for cubic spline).[9] In B-spline basis, m+1 knots are added on two sides of the specified knots so that totally there will be $(m + 1) + k + (m + 1)$ knots. The spline is however, defined only on the interval $[x_{m+2}, x_k]$ which implies that the first $m + 1$ and last $m + 1$ knots are arbitrary. Any spline of order $m + 1$ can then be represented as:

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\beta_i \quad (4)$$

where the B-splines can recursively be written as:

$$B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x), \quad i = 1,2,\ldots,k \quad (5)$$

P-splines, a penalty incorporated B-splines, are proposed by Eilers as a more stable version of the B-spline bases particularly for lower rank smoothing.[14] They are generally defined on an equidistant knots and used a difference penalty applied to adjacent coefficients, β_i, directly. For example, as given in Wood, if a squared difference of adjacent parameters is to be used as penalty measure, then the following is obtained:[9]

$$P = \sum_{i=1}^{k-1}(\beta_{i+1} - \beta_i)^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 - 2\beta_2\beta_3 + 2\beta_3^2 + \ldots + \beta_k^2 \qquad (6)$$

By increasing the differences parameter, a higher order penalty can be produced. The advantage of p-spline is that they are easy to set up and use. Additionally, they are flexible in the sense that any order of penalty can be incorporated to any order of B-spline basis.

Thin Plate Splines (Ts)

For an arbitrary spaced data (x_i, y_i), thin plate spline, $f(x_i, y_i)$, is a two-dimensional interpolation scheme which is an extension of the natural cubic spline for one dimensional data.[15] Splines of these types are good solutions for the smoothing function problem of more than one predictor variables.[9] In thin plate spline, the problem of estimating a smoothing function, f, is an estimation of a surface, while in natural cubic spline, it is a curve estimation problem.[13]

Green put for the general properties of the extended cubic spline in order to develop a methodology (thin- plate interpolant) for bivariate (for simplicity a bivariate case is used) case.[13] The properties of the roughness penalty, J, for data points (x_1, x_2) can be summarized as following:

1. If the second derivatives of f are square-integrable over $\Re^2$, J is finite.
2. If f has high local curvature, J will be large resulting in a large second derivative. Intuitively, it can be seen that J measures the wiggliness of f.
3. Rotating the coordinates in $\Re^2$ does not affect J.
4. The wiggliness penalty, J, is zero if and only if f is a linear function.

In smoothing procedure, J is used as roughness penalty and in interpolation, subjected to the interpolation conditions, it is used to find the natural thin-plate interpolator.[13] Now, consider the smoothing function, $f(\boldsymbol{x})$, estimation problem in Eq.7

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i \qquad (7)$$

where ε_i is the random error term and $\boldsymbol{x}$ is a d−vector from n (≥ d) observations (x_i, y_i). In this case, thin plate can be used to estimate the smoothing function, f, of the data points$(\boldsymbol{x}_i, y_i)$., where $i = 1, 2, \ldots, n$ (n ≥ d) by finding the function g which minimizes

$$\| \boldsymbol{y} - \boldsymbol{g} \|^2 + \lambda J_{md}(g)$$

where, $\boldsymbol{g} = (g(x_1), g(x_1), \ldots, g(x_1))'$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. J_{md} is the penalty function which measures the wiggliness of the smoother, g, whereas, λ is the smoothing parameter. Here, the roughness penalty function is given by Eq. 8

$$Jmd = \int \ldots \int_{\Re^d} \sum_{v_1+v_2+\cdots+v_d=m} \frac{m!}{v_1!\ldots v_d!} \left(\frac{\partial^m g}{\partial_{x1}^{v1}\ldots\partial_{xd}^{vd}}\right)^2 d_{x_1} \ldots d_{x_d} \qquad (8)$$

Here, it is important to note that thin plate splines can be used for any number of predictors.[9] In addition, there is no need of specifying knot positions. On the other hand, the disadvantage of these kind of smoothers is their being computational expensive; there are as many parameters to be estimated as there are data points.[15]

SELECTION OF GAM

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 inferences made for it measures how far the model estimations from the actual observations are. In GLMs and GAMs, it is necessary to have a quantity which is equivalent in importance and interpretation to residual sum of squares for ordinary linear modeling.[9]

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.[16] Model deviance is defined as twice the difference in log-likelihood between the saturated model and the full model (model of interest).[8,17-19]

Model deviance is defined as twice the difference in log-likelihood between the saturated model and the full model:

$$D = 2\left[l\left(\hat{\beta}_{max}\right) - l\left(\hat{\beta}\right)\right]\phi \qquad (9)$$
$$= \sum_{i=1}^{n} 2w_i\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right]$$

where $l\left(\hat{\beta}_{max}\right)$ is maximum likelihood of the saturated model: the model which have separate parameter for each observation aand a perfect fit $\hat{\boldsymbol{\mu}} = \mathbf{y}$. $\tilde{\theta}_i$ and $\hat{\theta}_i$ respectively the maximum likelihood estimated of canonical parameters. 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.

MONTE CARLO SIMULATIONS

In this section, a simulation study is conducted in order to illustrate the performance of three generalized additive models using thin-plate spline, cubic spline and p-spline and a linear model (lm) when fitted to data set suffering from multicollinearity and outliers. To achieve the desired degrees of collinearity, the explanatory variables are generated in the following order.[19,20] First, independent standard normal pseudo random numbers, z_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, 5$, were generated. Then, Eq. (10) was used to generate a total of four covariates with a specified degree of linear relationship.

$$x_{ij} = (1 - \rho^2)^{1/2} z_{ij} + \rho z_{i5}, \qquad i = 1, \ldots, n; j = 1, \ldots, 5. \qquad (10)$$

The value of ρ is specified so that the correlation between any two covariates will approximately be equal to ρ^2. Three different degrees of multicollinearity (ρ = 0.9, 0.99, 0.999) were considered (McDonald 1975).[19] The response variable is then generated as follows:

$$y_i = 2 + 5X_1 + 3X_2 + 4X_3 + 8X_4 + \varepsilon_i \qquad (11)$$

where ε is the error term which is i.i.d. generated from N(0,1). The inclusion of outliers was achieved by randomly selecting 10% of response values and multiplying them by 20 in order to inflate them in their absolute values.

In order to explore the behavior of the models, 500 samples were simulated for each of sample sizes of 50, 100, and 500 and the average deviance and AIC values for each model was recorded. It is important to note that a sample size less than 50 could not be used since the number of parameters of the generalized additive models would exceed the sample size.

The followings present the results from simulation studies implemented in R Software (packages: ridge, mgcv) with only multicollinearity and with multicollinearity and outliers, respectively.[9,21,22]

RESULTS

PERFORMANCE OF MODELS IN THE PRESENCE OF MULTICOLLINEARITY

Table 1 presents the VIF values of the explanatory variables including for four different values of ρ. For ρ = .8, it is shown that the simulated data do not suffer from multicollinearity, hence, higher values of ρ were used.

TABLE 1: VIF values when n=100

ρ	X_1	X_2	X_3	X_4
0.8	2.3	2.3	2.1	2.0
0.9	5.5	4.7	6.4	5.4
0.99	40.0	49.6	37.9	47.7
0.999	347.5	341.8	396.6	309.7

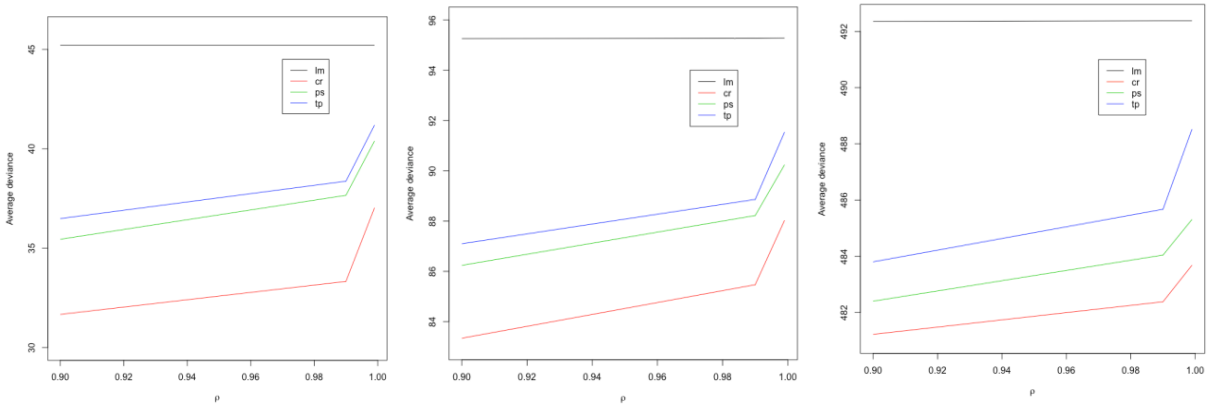
Deviance and AIC values of models for the three different degrees of collinearity are presented in Table 2. The results in all the cases show that linear model produced a higher deviance and AIC values compared to models fitted using the three spline techniques. From the results of this simulation experiment, it is clear that generalized additive models are more tolerant to the effects of multicollinearity than linear models.

TABLE 2: Average deviance and AIC values of models fitted to data with multicollinearity.

	Criteria	lm	cr	ps	tp
n=50	AIC	147.82	140.84	142.85	143.08
ρ =0.9	Deviance	45.21	**31.67**	35.45	36.52
n=50	AIC	147.82	141.88	144.27	144.52
ρ =0.99	Deviance	45.21	**33.33**	37.66	38.37
n=50	AIC	147.82	142.74	144.91	145.71
ρ =0.999	Deviance	45.21	**37.03**	40.39	41.19
n=100	AIC	289.99	286.59	286.98	286.88
ρ =0.9	Deviance	95.26	**83.34**	86.24	87.10
n=100	AIC	290.00	287.56	287.98	288.02
ρ =0.99	Deviance	95.27	**85.47**	88.22	88.86
n=100	AIC	290.01	287.71	288.40	288.73
ρ =0.999	Deviance	95.28	**88.03**	90.24	91.54
n=500	AIC	1422.01	1419.89	1419.83	1419.5
ρ =0.9	Deviance	492.36	**481.22**	482.40	483.8
n=500	AIC	1422.03	1420.64	1420.62	1420.44
ρ =0.99	Deviance	492.38	**482.38**	484.04	485.67
n=500	AIC	1422.03	1421.35	1421.19	1421.01
ρ =0.999	Deviance	492.38	**483.68**	485.31	488.52

The simulation results presented in Table 2 show that, regardless of the sample sizes or ρ, cubic splines produced the smallest deviance and AIC values. For n=50 the mean deviance is obtained by cubic splines model among others. This results holds for all models when ρ increseas. The impact of doubling the sample size produced the cubic splines model with least mean deviance for all ρ. Also for =500

and high multicollinearity, the cubic splines model obtained the lowest mean deviance. This is explicitly demonstrated in Figure 1a-c.



a) n=50 b) n=100 c) n=500

FIGURE 1: Average deviances when only multicollinearity exists.

PERFORMANCE OF MODELS IN THE PRESENCE OF MULTICOLLINEARITY AND OUTLIERS

A randomly selected 10% of the y values of the simulated data used in section 4 were multiplied by 20 so that they will be extreme values. It is clear from Figure 2 that the data includes some outliers.

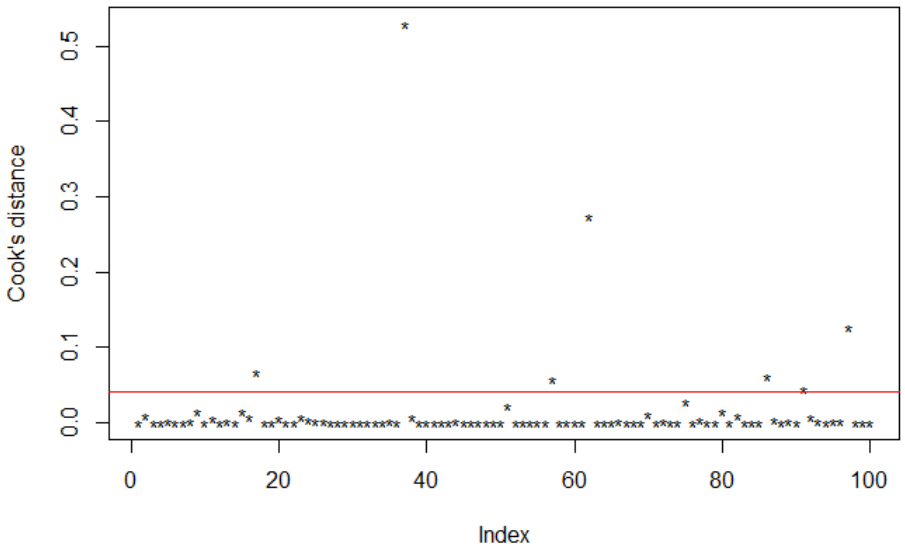


FIGURE 2: Influential observations by Cook's distance when ρ = 0.999 and n=100

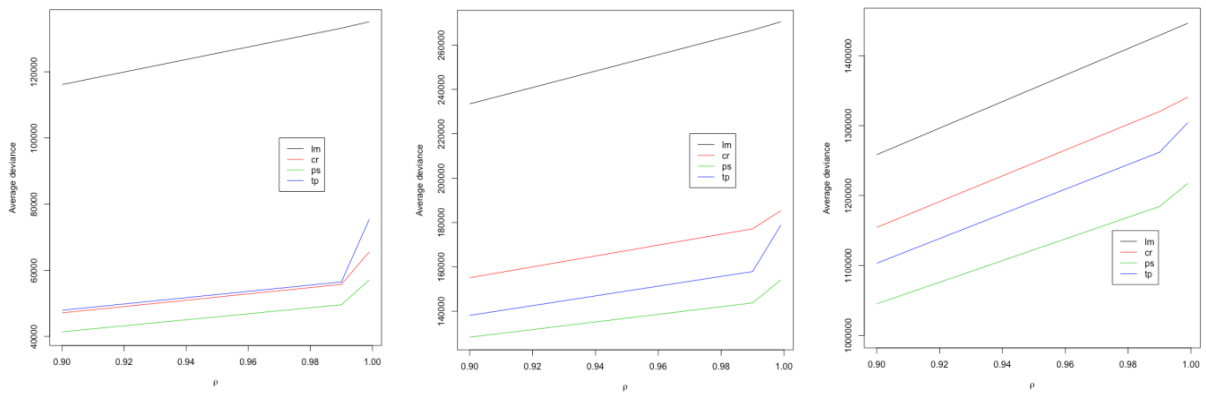
The average deviance and AIC values of the fitted models from simulations that suffers from multicollinearity and outliers are provided in Table 3.

TABLE 3. Average deviance and AIC values of models fitted to data with multicollinearity and outliers.

	Criteria	lm	cr	ps	tp
n=50	AIC	607.08	578.92	566.75	573
ρ =0.9	Deviance	517272	208930.1	**184011.1**	213666
n=50	AIC	614.39	584.38	573.26	586.08
ρ =0.99	Deviance	593280.6	248307.2	**220846.3**	282187.8
n=50	AIC	615.25	589.04	577.19	588.39
ρ =0.999	Deviance	601828.5	291836.2	**253975.4**	335633.1
n=100	AIC	1211.6	1189.14	1172.27	1179.48
ρ =0.9	Deviance	1039693	690920.8	**571711.2**	615236.8
n=100	AIC	1225.08	1203.33	1183.77	1192.56
ρ =0.99	Deviance	1187865	788311.3	**639734.5**	701328.9
n=100	AIC	1226.51	1207.05	1188.04	1199.26
ρ =0.999	Deviance	1204589	825424.8	**685864.1**	795739
n=500	AIC	6085.46	6062.31	6024.89	6047.53
ρ =0.9	Deviance	5605144	5141729	**4651736**	4913665
n=500	AIC	6149.24	6129.21	6087.86	6113.98
ρ =0.99	Deviance	6366311	5880103	**5271732**	5620942
n=500	AIC	6155.27	6137.6	6099.1	6125.21
ρ =0.999	Deviance	6443169	5971687	**5421672**	5809956

According to the performance results, the models fitted by using cubic splines were found to be the least advantageous whereas the models fitted by using p-splines produced a smaller average deviance and AIC under different samples sizes and some degree of linear relatisonship. In other words, the smallest deviance and AIC values were obtained from the models fitted by p-splines.

Figure 3 clearly demonstrates the performance differences of the four models considered in this paper. Figure 3a represents the results for n=50 whereas Figure 3b and Figure 3c show the results for n=100 and n=500, respectively. More essentially, it is demonstrated that the mean deviance of all the models increased with the increase of the degree of linear relationship provided when sample size n is kept fixed.



a) n=50 b) n=100 c) n=500

FIGURE 3: Average deviance when data suffers from both mutlicollinearity and outliers.

It is worth noting that with the increase of sample size, the dominance of the p-spline method increased. On the other hand, cubic regression loses its dominance of producing smaller average deviance when outliers are included to the data.

CONCLUSION

In this paper, we apply GAM models which were more resistant to outliers for the data sets suffering from multicollinearity and outliers. First, three penalized regression spline smoothers are evaluated in fitting generalized additive models for simulated datasets which contain outliers in the response variable and have predictor variables with linear relationship among them. The first is cubic splines, a curve made up of sections of cubic polynomials which are joined together and are continuous up to the second derivatives. Another is the p-splines, which is fitted using B-splines with penalty. With cubic or p-splines, on top of defining the basis functions, knots have to be specified in order to operate the fitting procedure. The third is the thin-plate splines which avoids the selection of basis functions and specifying knots positions. The results show that cubic regression outperformed both p-spline and thin plate spline in the sense of producing smaller mean deviance when multicollinearity exist among covariates. However, the models fitted using p-splines produced smaller average deviance and AIC when multicollinearity and outliers both exist.

This paper also presents explanations of parametric and nonparametric statistical evaluations for the selection of the best fit models. The results of this research are a valuable tool to make the predictions from a generalized additive model using p-splines when the data set is suffering from outlier and multicollinearity.

By way of conclusion, the results would seem to demonstrate that penalized smoothing splines can be used instead of generalized linear models when multicollinearity and outliers are present in the dataset.

Acknowledgement

This work was supported by Anadolu University Scientific Research Projects Commission under the grant no 1610F671.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Betül Kan Kılınç;* ***Design:*** *Betül Kan Kılınç;* ***Control/Supervision:*** *Betül Kan Kılınç;* ***Analysis and/or Interpretation:*** *Betül Kan Kılınç, Huruy Debessay Asfha;* ***Literature Review:*** *Betül Kan Kılınç, Huruy Debessay Asfha;* ***Writing the Article:*** *Betül Kan Kılınç;* ***Critical Review:*** *Betül Kan Kılınç, Huruy Debessay Asfha.*

REFERENCES

1. Belsley DA. A guide to using the collinearity diagnostics. Computer Science in Economics and Management. 2011;4(1):33-50.
2. Alimadad A, Salibian-Barrera M. An outlier-robust fit for generalized additive models with applications to disease outbreak detection. J Am Stat Assoc. 2011;106(494):719-31. [Crossref]
3. Wong RKW, Yao F, Lee TCM. Robust estimation for generalized additive models. J Comput Graph Stat. 2013;23(1):270-89. [Crossref]
4. Hastie TJ, Tibshirani RJ. Generalized Additive Models (Chapman & Hall/CRC Monographs on Statistics and Applied Probability). 1st ed. Routledge: Chapman and Hall/CRC; 1990. p.352.
5. Faraway JJ. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. 1st ed. Ann Arbor: Taylor and Francis Group, LLC; 2006. p.312.
6. Wood SN. Mixed GAM Computation Vehicle with Automatic Smoothness Estimation, R package version 1.8-31. 2018. [Link]
7. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Series A, 3. 1972;135(3):370-84. [Crossref]
8. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. 5th ed. Hoboken, New Jersey: Wiley; 2012. p.672.
9. Wood SN. Generalized Additive Models; An Introduction with R. 2nd ed. Routledge: Chapman and Hall/CRC; 2006. p.154-5.
10. McCullagh P, Nelder JA. Generalized Additive Models. 2nd ed. Routledge: Chapman and Hall/CRC; 1989. p.532. [Crossref] [PMC]
11. Eberly D. Ridges in Image and Data Analysis. 1st ed. Netherland: Kluwer Academic Publishers; 1996. p.215. [Crossref]
12. Keele L. Semiparametric Regression for the Social Sciences. 1st ed. Chichester, England; Hoboken, NJ: Wiley; 2008. p.230. [Crossref]
13. Ruppert D, Wand MP, Caroll RJ. Semiparametric Regression. 1st ed. Cambridge: Cambridge University Press; 2003. p.386. [Crossref]

14. Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. 1st ed. U.S.A.: Chapman and Hall/CRC; 1994. p.184. [Crossref]
15. Eilers PHC, Marx DB. Flexible smoothing with *B*-splines and penalties. Stat Sci. 1996;11(2):89-121. [Crossref]
16. Wood SN. Thin plate regression splines. J R Stat Soc: Series B (Statistical Methodology). 2003;65(1):95-114. [Crossref]
17. Agresti A. Foundations of Linear and Generalized Linear Models. 1st ed. Hoboken, New Jersey: John Wiley & Sons Inc.; 2015. p.480.
18. Myers RH, Montgomery DC, Vining GG, Robinson T. Generalized Linear Models: with Applications in Engineering and the Sciences. 2nd ed. Hoboken, N.J.: Wiley; 2010. p.496. [Crossref]
19. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002. p.498. [Crossref]
20. McDonald GC, Galarneau DI. A Monte Carlo evaluation of some ridge-type estimators. J Am Stat Assoc. 1975;70(350):407-16. [Crossref]
21. Månsson K, Shukur G, Sjölander P. A New Ridge Regression Causality Test in the Presence of Multicollinearity. HUI Working Papers 37, HUI Research; 2010.
22. R Development Core Team, R: A Language And Environment For Statistical Computing Vienna, Austria, R Foundation for Statistical Computing. 2013. [Link]