Ilorin Journal of Science
Volume 10, Number 1, 2023, pp. 28 – 43 (Printed in Nigeria)
ISSN: 2408 – 4840 © 2023 Faculty of Physical Sciences, University of Ilorin
https://doi.org/10.54908/iljs.2023.10.01.003

ILORIN
JOURNAL OF SCEINCE

ILJS-23-004

Beyond Linearity: Harnessing Spline Regression Models to Capture Non-linear Relationships

Ajao*, I. O., Adaraniwon, A. O. and Ilugbusi, A. O.

Department of Statistics,The Federal Polytechnic, Ado-Ekiti, Nigeria

Abstract

This article explores the effectiveness of spline regression model in capturing non-linear relationships in data. A comparison of spline regression with other techniques, such as linear regression, polynomial regression, generalized additive, and log-transformed models, is conducted using simulated data. The performance metrics, including AIC, BIC, RMSE, MSE, MAE, and R-squared, are used to assess the goodness of fit for each model. The results indicate that the spline regression model outperforms other methods in accurately capturing non-linear relationships. The flexibility and smoothness provided by spline regression, through the incorporation of knots, result in better-fitted lines that closely match the data. This study recommends the use of spline regression for handling non-linear data and highlights its robustness and accuracy.

Keywords: Spline Regression, Non-linear Relationships, Goodness of Fit, Simulation, Comparison

1. Introduction

In traditional regression models, such as piecewise linear regression, when data exhibits non-linear patterns or multiple phases, the fitted line may show abrupt changes at the points where the segments meet. This sudden change in slope can lead to a discontinuous or jagged appearance in the regression line, which may not accurately represent the true underlying relationship between the variables. In contrast, spline regression models allow for smooth transitions at specific points called knots, ensuring a continuous and flexible fit without such abrupt jumps. This property makes spline regression a powerful tool for capturing complex non-linear relationships in data. In the context of the article, "jump" refers to a sudden, sharp change in the slope of the regression line when transitioning from one segment of the data to another.

Corresponding Author: Ajao, I. O.
Email: isaacoluwaseyiajao@gmail.com

The "jump" in combining the two regression lines is not possible in a spline regression model Wood (2003). In another lucid and uncomplicated presentation of spline regression models, (Wood, Pya and Säfken, 2016) gave detail information on smoothing parameter and model selection. In order to account for the various tax brackets in an income tax application utilizing a particular kind of spline, piecewise linear regression was utilized by Strawczynski and Paper (1944). As a result, the regression line for the relevant dependent variable (such as approval rating) may suddenly change its slope without creating a sharp "jump" in the line. In order to do this, a line must be allowed to kink (change in slope) without breaking, Wood (2003). Consequently, the slopes of the two distinct regression lines are allowed to differ at the place where they join (spline knot), forcing the lines to touch. Relationship "change points" or "knots" are places where the dynamic of the relationship shifts. Wood (2003) provides a succinct explanation of how to build and apply a basic spline model along with an example that shows how to use these techniques in a real-world setting. The straightforward adjustment strategy for spline models was created Baty *et al*. (2015). Numerous methods are described by Firth *et al.* (2020) for determining initial values for nonlinear fitting. When examining polynomial, Goldstein, Chatterjee and Price (1979) hypothesized that a second-order model may have fitted more effectively. In order to map soil salinity, both linear and nonlinear regression models are used, Garcia and Eldeiry (2020). Nonlinear models can handle multiphase, spline regressions, and multi-response data, Crosnier and Salmer (1994) whereas current nonlinear regression modules lack specialized diagnostic capabilities, Baty *et al.* (2015) offers users an enhanced toolbox of functions enabling a rigorous evaluation of nonlinear regression fits. A modification of the linear regression model is suggested by Fox and Weisberg (2018) in which the conditional mean of the response variable is not a linear function of the parameters. Max *et al*. (2022) has been working on a number of functions for training and visualizing classification and regression models. Wherever necessary, showing regression trees and classification and regression models. Kassambara (2021) creates a fantastic, versatile program for beautiful data visualization in nonlinear regression and other applications. In several instances of nonlinear data, a generalized nonlinear model has been presented Turner and Firth (2008). In nonlinear modeling of simulated and observed hydrological time series, goodness of fit metrics are required Mauricio and Bigiarini (2010). Implement metrics for issues with binary classification, classification, information retrieval, regression, time series, and other issues, Frasco *et al*. (2018). Tools are created to make it simpler for researchers to create ordinary least squares regression models, including thorough regression output, heteroskedasticity tests, diagnostics for collinearity and residuals, measurements of influence,

evaluation of model fit, and methods for variable selection. The nonlinear mixed-effects (NLME) model is discussed in Pinheiro and Bates (2000), which also worked on the linear mixed-effects (LME) model. Both linear and nonlinear regression can be used to fit models to biological data, Motulsky and Christopoulos (2003). There is a comprehensive discussion on fitting quadratic models, generalized linear models, and applications, Goldstein *et al.* (1979). According to Spiess and Neumeyer (2010), R-square cannot be used to demonstrate the effectiveness or validity of a nonlinear model. The foundation created by solving the thin plate spline smoothing problem, according to Wood (2003), can be easily changed and truncated to produce smoothers. According to Wood (2004), GAMs with ridge penalties offer a workable solution in such cases. To that end, a method for choosing numerous smoothing parameter combinations that works in the presence of such a penalty has been established.

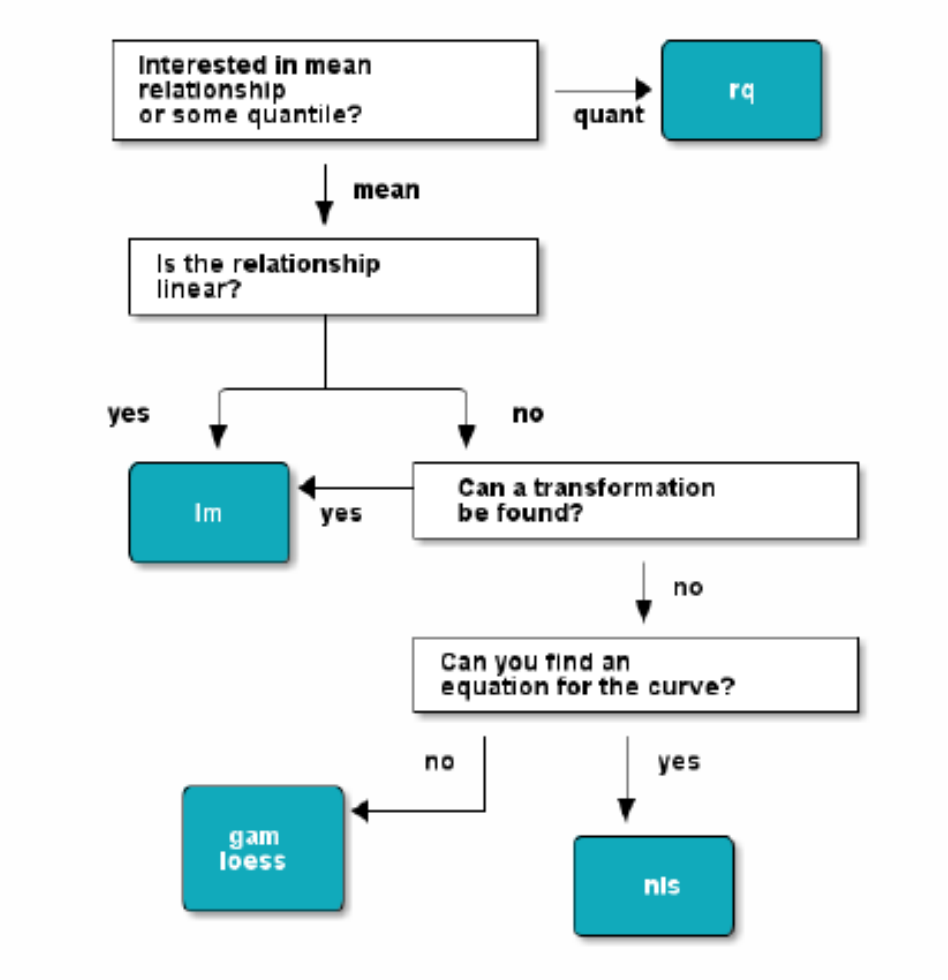

Figure 1: A simple flow chart showing how to pick a regression model

2. Materials and Methods

Linear regression model

Linear regression models the association between two variables when an equation that is linear in nature is fitted to the data. In this modeling approach, one variable is referred to as an independent variable, while the other is considered the dependent variable (Ruckstuhl, 2017).

$$Y = a + bX, \quad (1)$$

where X is the explanatory variable and Y is the dependent variable, is the equation of a linear regression line. According to Ruckstuhl (2017), b is the line's slope, and a is the intercept (the value of y when x = 0).

Polynomial Regression

One type of regression technique is polynomial regression that employs a non-linear function Wang and Yan (2021). By elevating each of the initial predictors to a power, polynomial regression adds additional predictors to the linear model. A cubic regression, for instance is a type of regression technique that models the relationship between variables using a cubic function, incorporating predictors raised to the power of three, allowing for more flexible and curved fits to the data.

Substituting the linear model with a function of the polynomial is the accepted method for broadening linear regression to a non-linear connection between the dependent and independent variables.

$$Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_p X_i^p \quad (2)$$

Strong oscillations may be seen in the resulting curve when we increase the value of power, which will produce overly flexible geometries. These curves cause over fitting. Over fitting in polynomial regression occurs when the model fits the data too closely, resulting in an overly complex polynomial function that perfectly matches the data points. As the degree of the polynomial increases, the model becomes increasingly flexible and capable of capturing even small fluctuations in the data. The problem with over fitting in polynomial regression is that the model may not generalize well to new data, leading to poor predictive performance. The excessively complex polynomial function might be overly sensitive to noise in the data, leading to inaccurate predictions on new data points.

Spline Regression Model

We can utilize an improved regression strategy that splits the information into many forms and fit each group (form) with a different model in order to overcome problem of over-fitting, which is a significant drawback of polynomial regression. Regression spline is a term used to describe such a method Marsh (2011); Wood *et al.* (2016). Knots are the locations where the division takes place. Piecewise functions are those that can be used to model each piece or bin. These individual bins can be fitted using a variety of piecewise functions. Compared to polynomial regression, regression splines frequently produce better results. This is due to the fact that splines introduce flexibility by increasing the number of knots while maintaining a fixed degree of order 3, unlike polynomials, which require a high degree polynomial to produce flexible fits (Wood, 2004).

The individual cubic spline equation with just one knot at a location g has the following shape:

$$y_i = \begin{cases} \beta_{01} + \beta_{11} z_i + \beta_{21} z_i^2 + \beta_{31} z_i^3 + \epsilon_i & \text{if } z_i < g \\ \beta_{02} + \beta_{12} z_i + \beta_{22} z_i^2 + \beta_{32} z_i^3 + \epsilon_i & \text{if } z_i \geq g \end{cases} \quad (3)$$

In other words, we used the component of the observations with zi ≥ g and the component of the observations with zi ≥ g to fit two separate polynomial functions to the data. The coefficients for the first polynomial function are $\beta_{01},..,\beta_{31}$; for the second, they are $\beta_{02},\ldots,\beta_{32}$. By applying least squares to the straightforward functions of the real regressor, each of these polynomial functions may be fitted.

Generalized Additive model: Univariate smooth functions

A generalized additive model is a generalized linear model with a predictor having a linearity property admitting the sum of smooth functions of variables (Hastie & Tibshirani, 1990). A standard polynomial regression can be used as a starting point for a generalized additive model, and it might work out adequately (Wood, 2004). We must first think about a basis to utilize, a space in which *f* is an element. In doing so, one selects a collection of bases functions *bj* with parameters *j* that are integrated to form *f(x):*

$$f(x) = \sum_{j=1}^{q} b_j(x)\gamma_j \quad (4)$$

The cubic polynomial can be used to express this better

$$b_1(x) = 1,\ b_2(x) = x,\ b_3(x) = x^2,\ b_4(x) = x^3 \quad (5)$$

which leads to

$$f(x) = \gamma_1 + \gamma_2 x + \gamma_3 x^2 + \gamma_4 x^3 \quad (6)$$

3. Data Analysis and Results

This data used in this study was obtained through simulation. R 4.1.3 was used for all analyses (R Core Team, 2022).

Simulation procedure

In R, skewed data are simulated using various probability distributions that exhibit skewness. One common distribution used for simulating skewed data is the gamma distribution. A specific function in R is utilized to generate random numbers from the gamma distribution with specified shape and scale parameters, allowing for control over the skewness of the data. To simulate positively skewed data, a gamma distribution with a shape parameter greater than 1 is used. On the other hand, for negatively skewed data, a gamma distribution with a shape parameter less than 1 is employed.

The following scatterplot is from the simulated non-linear dataset, it will be used throughout this paper

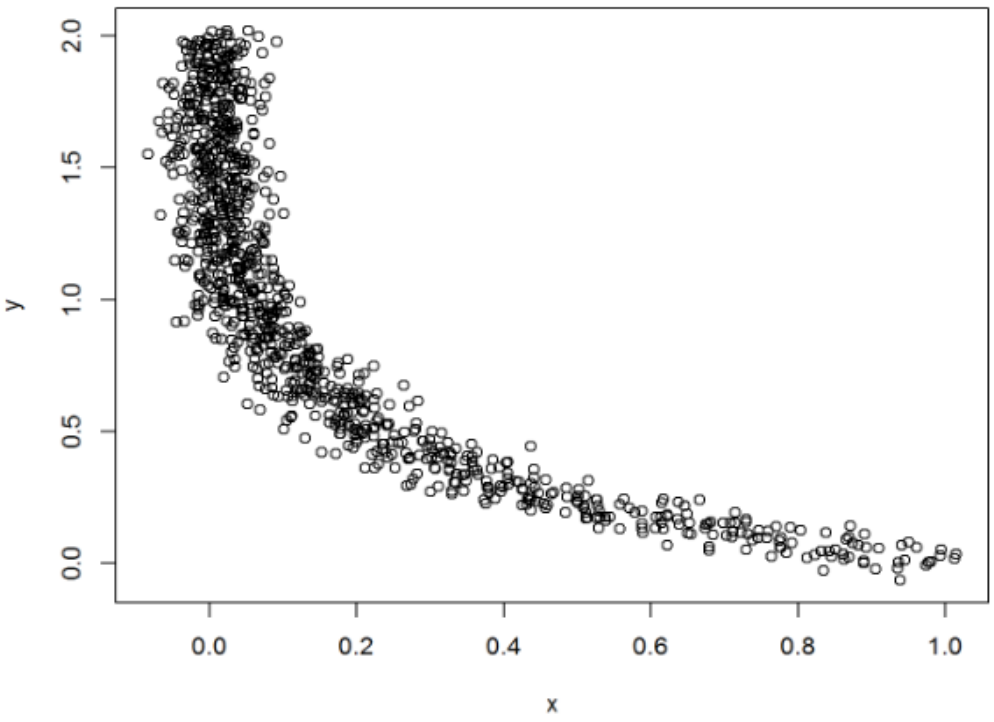

Figure 2: Scatter plot of the nonlinear relationship between y and x

Whenever reliability margins are really large, the fit is not likely to be helpful.

Table 1: Parameter estimate of the Linear regression fitted model

Term	Estimate	Std. error	P-Value
Intercept	1.3248	0.0132	0.0002
x	-1.9501	0.0455	0.0001

Forming the simple linear model,

$y = 1.32 - 1.95x$

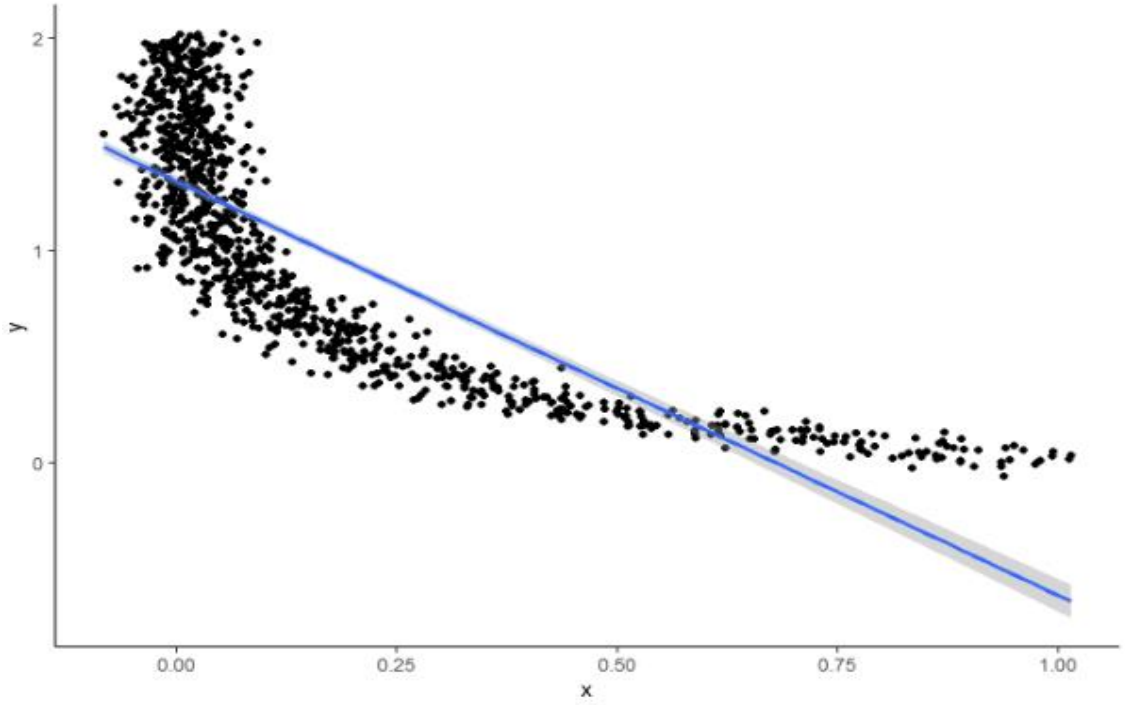

Figure 3: Non-linear relationship with linear regression fitted line

This fitted line in Figure 3 does not capture the actual data points, therefore linearity should not be assumed here.

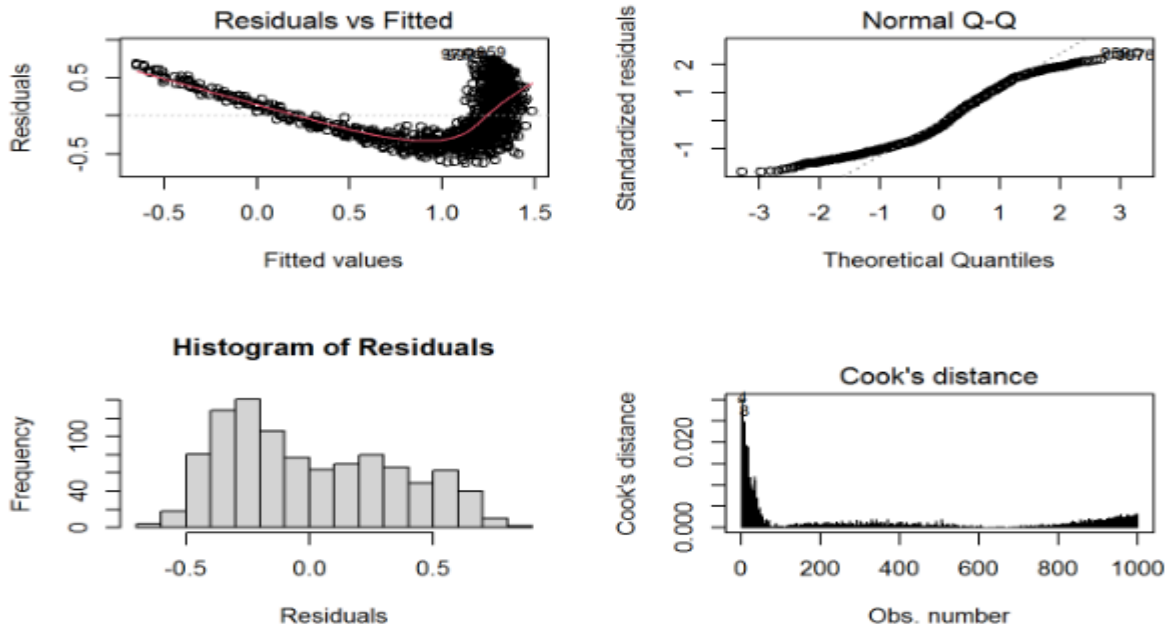

Figure 4: Model diagnostic for Linear regression model

Table 2: Parameter estimates of the Spline regression fitted model

Term	Estimate	Std. error	P-Value
Intercept	1.7509	0.1587	0.0001
bs(x, knots = knots)$_1$	-0.4030	0.2162	0.0626
bs(x, knots = knots)$_2$	0.0252	0.1510	0.8677
bs(x, knots = knots)$_3$	-1.0027	0.1669	0.0001
bs(x, knots = knots)$_4$	-1.6698	0.1701	0.0002
bs(x, knots = knots)$_5$	-1.5890	0.1976	0.0001
bs(x, knots = knots)$_6$	-1.7682	0.1750	0.0002

Which produces the following model:

$$y = 1.75 - 0.4(bs(x, knots = knots)_1) + 0.03(bs(x, knots = knots)_2) - 1(bs(x, knots = knots)_3)$$
$$-1.67(bs(x, knots = knots)_4) - 1.59(bs(x, knots = knots)_5) - 1.77(bs(x, knots = knots)_6)$$

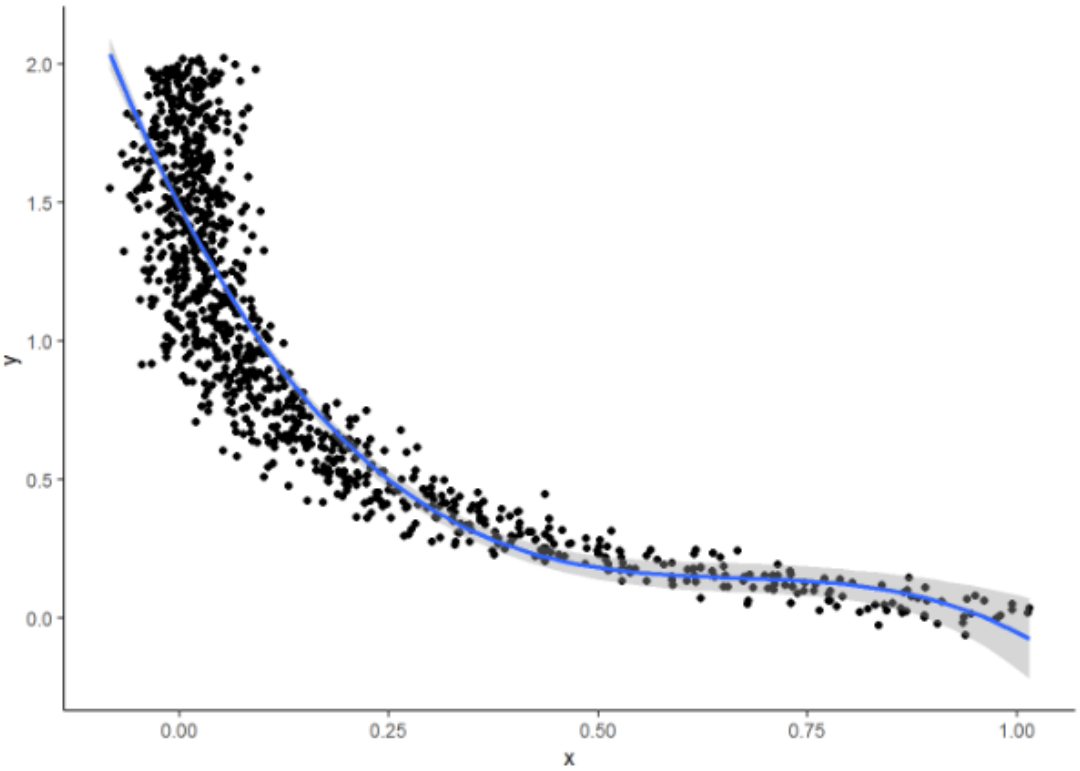

Figure 5: Non-linear relationship with Spline regression

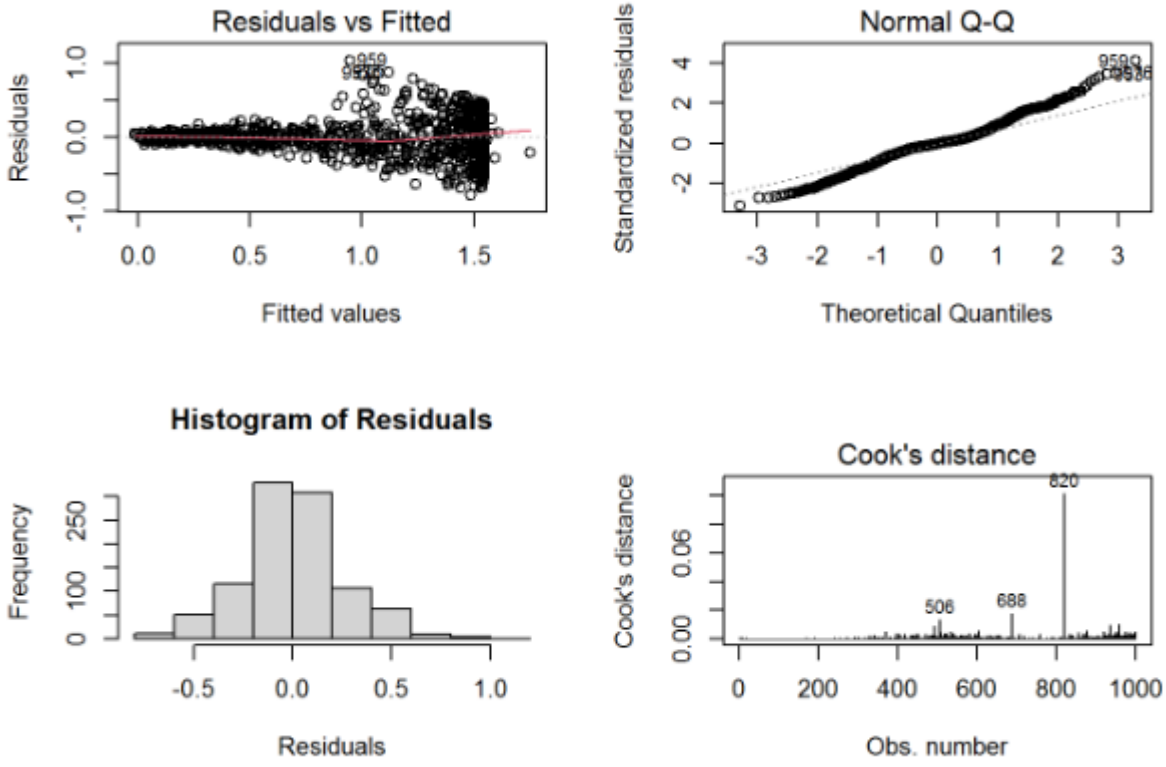

Figure 6: Model diagnostics for Spline model

Table 3: Parameter of the estimates of the Polynomial regression fitted model at fourth order

Term	Estimate	Std. error	P-Value
Intercept	1.48448	0.0121	0.0002
poly(x, 4, raw = TRUE)1	-5.64981	0.2814	0.0001
poly(x, 4, raw = TRUE)2	7.23395	1.6386	0.0011
poly(x, 4, raw = TRUE)3	-1.56802	3.0766	0.6100
poly(x, 4, raw = TRUE)4	-1.59548	1.7695	0.3670

From the output above, it can be seen that polynomial terms beyond the second order are not significant. So, a second polynomial regression model as follows:

Table 4: Parameter of the estimates of the Polynomial regression fitted model at second order

Term	Estimate	Std. error	P-Value
Intercept	1.4532	0.01187	0.0002
poly(x, 4, raw = TRUE)1	-4.39671	0.10887	0.0001
poly(x, 4, raw = TRUE)2	3.29459	0.13821	0.0001

The visualization of the second polynomial regression line as follow:

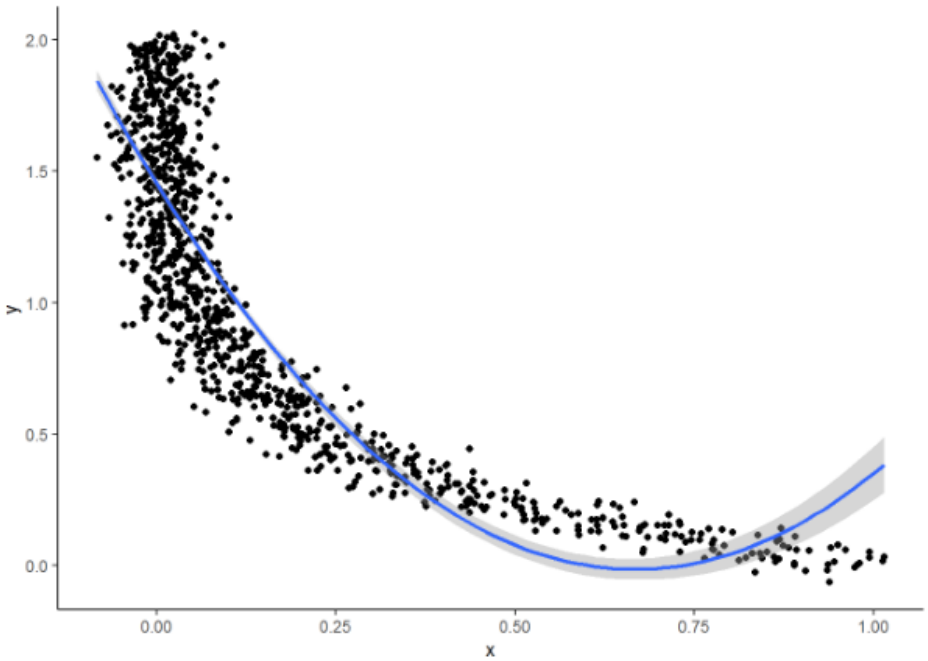

Figure 7: Non-linear relationship with Polynomial regression

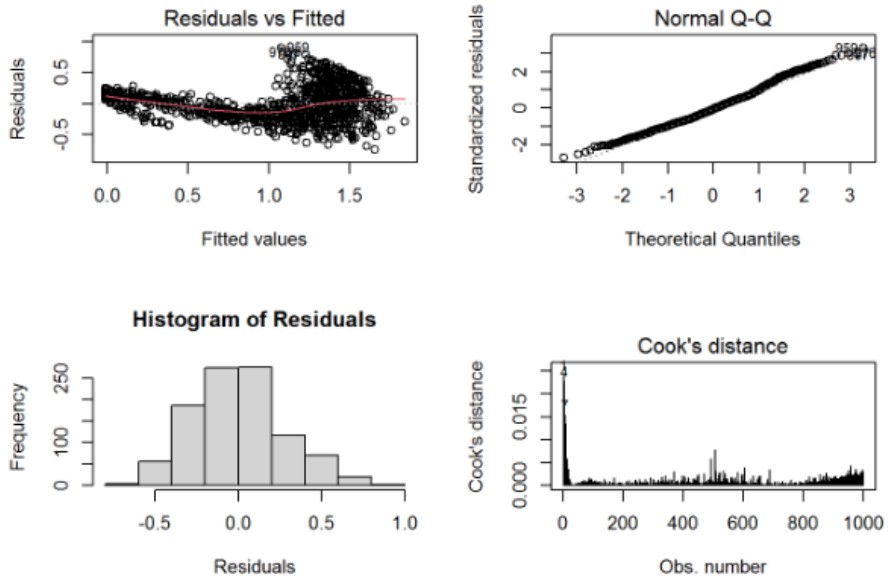

Figure 8: Model diagnostic for Polynomial

Table 5: Fitted Generalized Additive Model (GAM)

Term	Estimate	Std. error	P-Value
Intercept	1.0000	0.0080	0.0002

Table 6: Rough smooth terms significance

Term	edf	Ref.df	F	P-Value
s(x)	8.631	8.961	480.4	0.0001

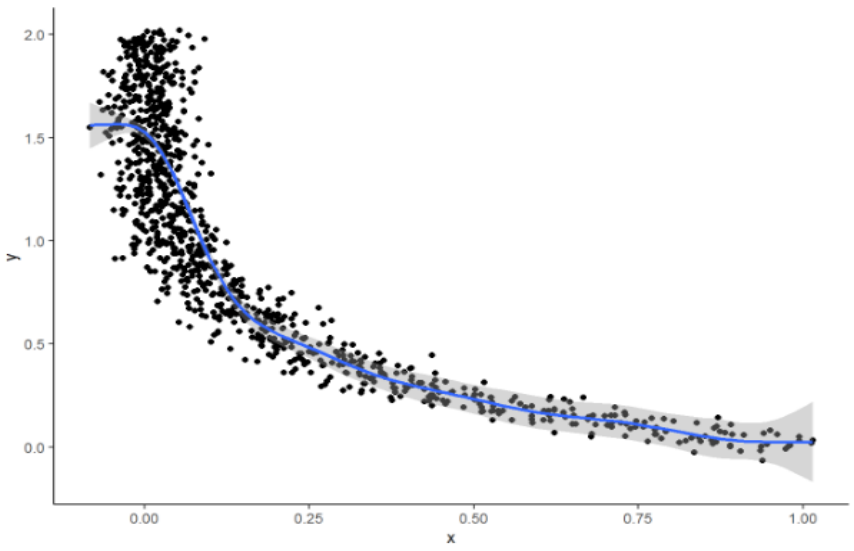

Figure 9: Non-linear relationship with GAM

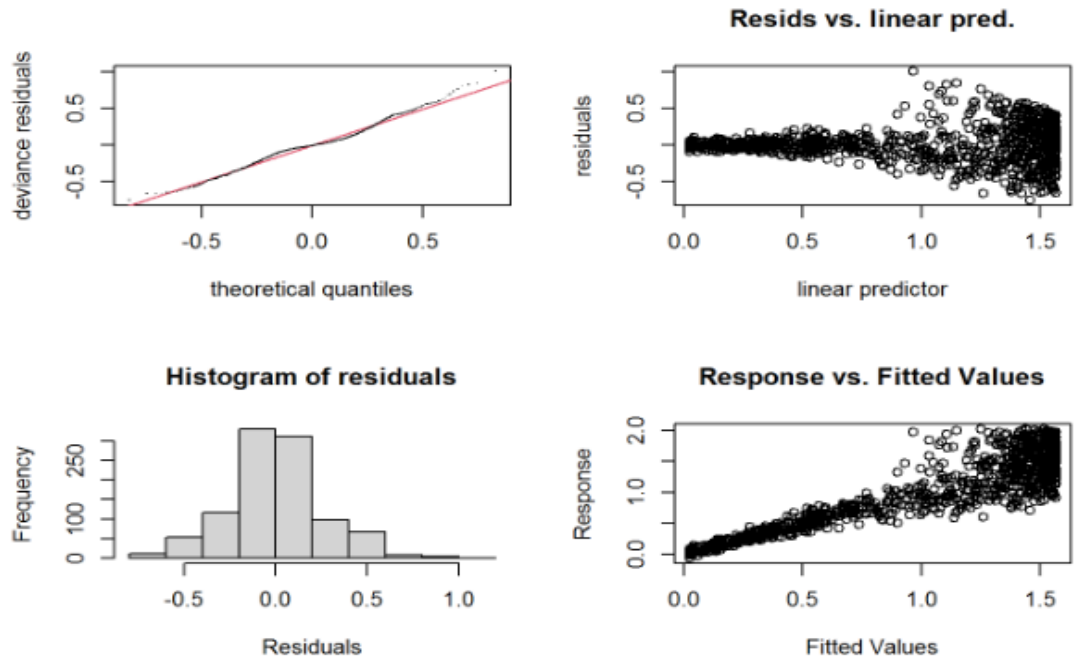

Figure 10: Model diagnostics for GAM

Table 7: Log-transformed fitted model

Term	Estimate	Std. error	P-Value
Intercept	0.0880	0.0132	0.0003
log(x)	-0.3337	0.0070	0.0001

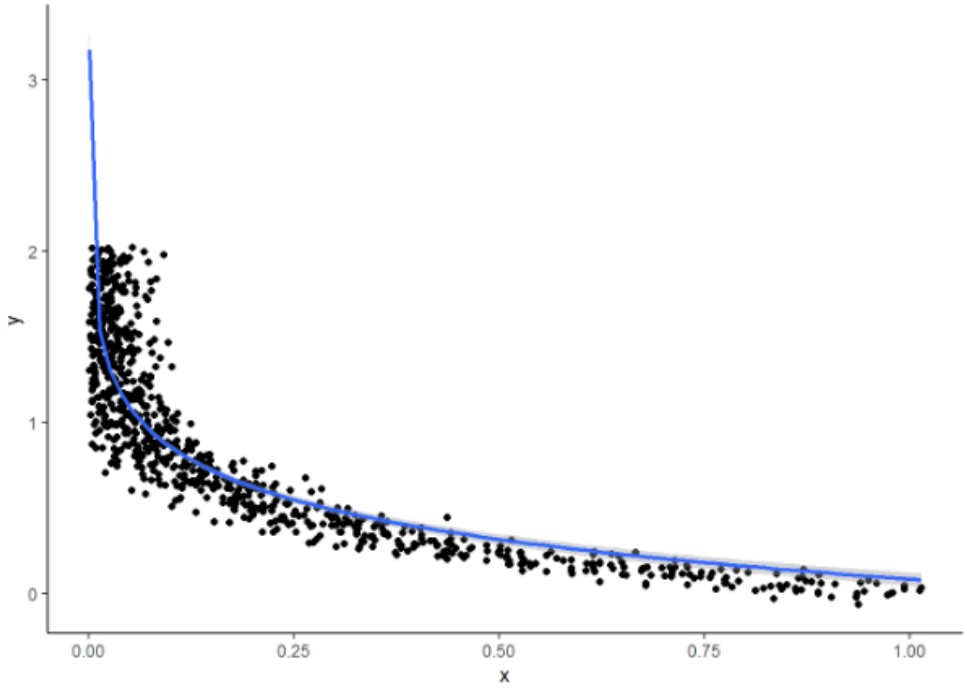

Figure 11: Non-linear relationship with Log-transformation

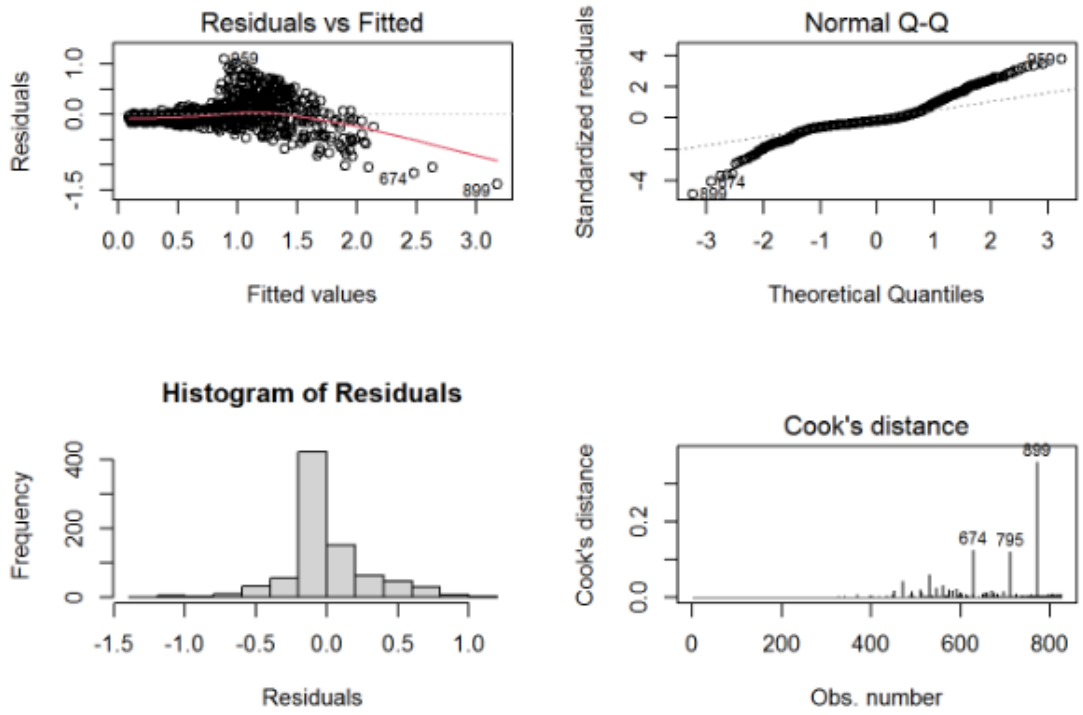

Figure 12: Model diagnostic for Log-transformation

Goodness of fit

The R-square is typically poor indicator of the goodness of fit in non-linear models Spiess and Neumeyer (2010). The standard deviation of the residuals can be generally understood as the RMSE ('approximately' because it is slightly adjusted for the number of fitted components).

Table 8: Model selection

Model	AIC	BIC
Linear	704.4970	719.2200
Spline	82.6670	121.9290
GAM	87.6756	139.8760
Polynomial	255.4550	275.0860
Log-Transformed	300.2970	314.4470

The minimum values of information criteria according to Table 8 are achieved in regression spline model, which suggest it's robustness and sensitivity to non-linear data.

Table 9: Summary of models performance

Model	RMSE	MSE	MAE	R^2
Linear	0.3431	0.1177	0.2970	0.6481
Spline	0.2502	0.0626	0.1815	0.8129
GAM	0.2501	0.0625	0.1818	0.8131
Polynomial	0.2738	0.0750	0.2196	0.7758
Log-Transformed	0.2892	0.0836	0.1995	0.7314

Table 8 shows that Regression Spline model and GAM have the minimum values of the error metrics, this means GAM also performs well in term of model performance.

4. Summary of findings

None of the fitted lines in Figures 3, 7, 9, and 11 capture the actual values of the non-linear dataset accurately as the Regression spline fitted line in Figure 5. It is also obvious from the residual diagnostics as shown in Figures 4, 6, 8, 10, and 12 carried out on linear, spline, polynomial, GAM, and log-transformed models respectively that residual and qq-plots did not show the features of good models in any of the methods as that of the spline regression model. Tables 1 – 7 show the parameter estimates in each model which produced the various model equations, while the values of the information criteria and model performance in Tables 8 and 9 respectively establish the robustness of the regression spline model.

5. Conclusion and Recommendation

The results from the Tables and figures in this paper have proved that the regression spline model outperforms the other models considered. It is therefore recommended that the model be employed when dealing non-linear data.

References

Baty, F., Ritz, C., Charles, S., Brutsche, M., Flandrois, J. P. and Delignette-Muller, M. L. (2015). A toolbox for nonlinear regression in R: The package nlstools. *Journal of Statistical Software, 66*(5), 1–21. https://doi.org/10.18637/jss.v066.i05

Crosnier, Y. and Salmer, G. (1994). Non-linear Modelling. *Gallium Arsenide Technology in Europe*, 243–260. https://doi.org/10.1007/978-3-642-78934-2_18

Firth, D., Ripley, B., Venables, B., Bates, D. M. and Turner, M. H. (2020). *Package ' gnm .'*

Fox, J. and Weisberg, S. (2018). Nonlinear Regression, Nonlinear Least Squares, and Nonlinear Mixed Models in R. *An R Companion to Applied Regression*, 31.

Frasco, M. M., Ben, A., Frasco, M. and Ledell, E. (2018). *Package ' Metrics .'*

Garcia, L. A. and Eldeiry, A. A. (2020). *Evaluating Linear and Nonlinear Regression Models in Mapping Soil Salinity*. *7*(3), 21–34.

Goldstein, M., Chatterjee, S. and Price, B. (1979). Regression Analysis by Example. In *Journal of the Royal Statistical Society. Series A (General)* (Vol. 142).https://doi.org/10.2307/2982566

Hastie, T. J. and Tibshirani, R. J. (1990). Hastie T.J., Tibshirani R.J. - Generalized Additive Models (CRC,1990)(175d)0412343908.pdf. *New York: Chapman and Hall.*, Vol. 1, pp. 297–310.

Kassambara, A. (2021). *Ggpubr: "Ggplot2" Based Publication Ready Plots.* Retrieved from https://cran.r-project.org/package=ggpubr

Marsh, L. C. (2011). Spline Regression Models. *Journal of Applied Business Research (JABR)*, *19*(2). https://doi.org/10.19030/jabr.v19i2.2166

Mauricio, A. and Bigiarini, Z. (2010). *Package ' hydroGOF .'*

Max, A., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Kuhn, M. M. (2022). *Package ' caret ' R topics documented :*

Motulsky, H. and Christopoulos, A. (2003). Fitting Models to Biological Data Using Linear and Nonlinear Regression.Fitting curves with GraphPad Prism. *GraphPad Prism Software Inc., San Diego,California.*, 296–297.

Pinheiro, J. and Bates, D. (n.d.). *Bates_2000_Mixed_Effect_Models.pdf*.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ruckstuhl, A. (2017). *Introduction to Linear Regression*. 140–147. https://doi.org/10.4018/978-1-68318-016-6.ch006

Spiess, A. N. and Neumeyer, N. (2010). An evaluation of R2as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC Pharmacology*, *10*(June). https://doi.org/10.1186/1471-2210-10-6

Strawczynski, M. and Paper, D. (1944). Social insurance. *British Medical Journal*, *2*(4370), 474–475. https://doi.org/10.1136/bmj.2.4370.474

Turner, H. and Firth, D. (2008). Generalized nonlinear models in R : An overview of the gnm package Generalized linear models Preamble. *Social Research*, 1–55.

Wang, W. and Yan, J. (2021). Shape-Restricted Regression Splines with R Package splines2. Journal of Data Science, 498–517. https://doi.org/10.6339/21-jds1020

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686. https://doi.org/10.1198/016214504000000980

Wood, S. N., Pya, N. and Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986