Open Journal of Statistics, 2024, 14, 743-756
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Discovering the Best Choice for Spline's Knots and Intervals Using Order of Polynomial Regression Model

Farag Hamad[1]*, Najiah Younus[2], Mohamed Jaber[3]

[1]Department of Statistics, College of Arts and Science, University of Benghazi, Benghazi, Libya
[2]Department of Mathematics, College Arts and Science, University of Benghazi, Benghazi, Libya
[3]Department of Statistics, College of Science, University of Misurata, Misurata, Libya
Email: *farag.hamad@uob.edu.ly

How to cite this paper: Hamad, F., Younus, N. and Jaber, M. (2024) Discovering the Best Choice for Spline's Knots and Intervals Using Order of Polynomial Regression Model. *Open Journal of Statistics*, **14**, 743-756.
https://doi.org/10.4236/ojs.2024.146034

Received: November 14, 2024
Accepted: December 17, 2024
Published: December 20, 2024

Abstract

In this work, we seek the relationship between the order of the polynomial model and the number of knots and intervals that we need to fit the splines regression model. Regression models (polynomial and spline regression models) are presented and discussed in detail in order to discover the relation. Intrinsically, both models are dependent on the linear regression model. Spline is designed to draw curves to balance the goodness of fit and minimize the mean square error of the regression model. In the splines model, the curve at any point depends only on the observations at that point and some specified neighboring points. Using the boundaries of the intervals of the splines, we fit a smooth cubic interpolation function that goes through $(n + 1)$ data points. On the other hand, polynomial regression is a useful technique when the pattern of the data indicates a nonlinear relationship between the dependent and independent variables. Moreover, higher-degree polynomials can capture more intricate patterns, but it can also lead to overfitting. A simulation study is implemented to illustrate the performance of splines and spline segments based on the degree of the polynomial model. For each model, we compute the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to compare the optimal polynomial order for fitting the data with the number of knots and intervals for the splines model. Both AIC and BIC can help to identify the model that best balances fit and complexity, aiming to prevent overfitting by penalizing the use of excessive parameters. We compare the results that we got from applying the polynomial regression model with the splines model results in terms of point estimates, the mean sum of squared errors, and the fitted regression line. We can say that order five of the polynomial model may be used to estimate splines with five segments.

Keywords

Nonlinear Regression, Splines, Polynomial, Cross-Validation, Akaike Information & Bayesian Information Criterion

1. Introduction

Regression analysis is typically used by academics to examine the impact of several independent factors, or explanatory variables, on a single variable, or response variable. The regression equation is used by the investigators to explain how the response and explanatory variables relate to one another [1]. The regression analysis can be divided into linear regression and nonlinear regression [2]. A linear regression is easy to understand and simple to fit. The regression model is desirable because there are many techniques for testing the assumptions. However, in many cases, data are not linearly related. Therefore, it is not recommended to use linear regression. As previously mentioned, the traditional nonlinear regression model fits the model

$$y = f\left(X,\theta\right) + \varepsilon$$

However, in some situations, the structure of the data is so complicated that it is very difficult to find a function that estimates the relationship correctly. Other difficulties might emerge, such as the selection of good starting values and the suitable criterion to declare convergences. The general nonparametric regression model is written in a similar manner and f is left unspecified:

$$y = f\left(X\right) + \varepsilon = f\left(x_1, x_2, \cdots, x_p\right) + \varepsilon$$

where $f \in \left[a,b\right]$ is an unknown smooth function, $\left(y_i\right)_i^n$ are observation values of the response variable y, $\left(x_i\right)_i^n$ are observation values of the explanatory variable x and $\left(\varepsilon_i\right)_i^n$ are normal distributed random errors with zero mean and common variance σ^2, *i.e.* $\varepsilon_i \sim \text{NID}\left(0,\sigma^2\right)$. The basic goal of nonparametric regression is to estimate the regression function $f\left(.\right)$ directly, rather than to estimate parameters [3] [4]. Most nonparametric regression methods assume that $f\left(.\right)$ is smooth. The smooth function is a continuous function with first and second derivatives existence. This work aims to provide practical guidance for selecting the most appropriate order of the polynomial model that will fit the best model using spline regression. The study will explore the relationship through empirical experiments and by measuring model selection criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The optimal order of the polynomial model will be used to determine the optimal number of spline knots and intervals that would balance the spline model's goodness of fit.

2. Methodology

Many different techniques are used to fit the relationship between the variables when the relationship is nonlinear, such as polynomials and splines. The nonlinear models intrinsically involved linear models of the independent variable x that were either strictly increasing or strictly decreasing [5] [6]. In many cases, a theoretical scatter plot of the data suggests that the true regression function has one or more peaks or valleys, which is at least one relative minimum or maximum. In many situations, a polynomial function may provide a satisfactory approximation to the true regression function.

2.1. Polynomial Regression

Linear regression may not always be used to fit the relationship between the dependent variable and the independent variables. In many cases, the data pattern in the relationship may be a nonlinear relation that cannot be captured by a simple straight line [7]. Polynomial regression can easily capture nonlinear relationships. Moreover, the polynomial regression is useful when the data pattern indicates that the relationship between the dependent and independent is not a linear relation. Adding polynomial terms (e.g., x^2, x^3) into the simple regression model will increase the accuracy of the model and allow it to fit the complex data patterns. Higher-degree polynomials can capture more intricate patterns, but they can also lead to overfitting [8] [9]. Polynomial regression is also called the special case of multiple linear regression because polynomial regression is a linear model with some adjustments in order to increase the accuracy. The polynomial regression model is defined as:

$$P_k(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon = \sum_{j=0}^{k} \beta_j x^j + \varepsilon$$

2.2. Selecting Polynomial Regression Degree

For n polynomial data points, we can come up with an n^{th} degree polynomial that will ensure it goes through every single one of the points. Therefore, the polynomial degree is selected to optimize the target function for which the variance is minimum or we choose the degree of a polynomial when there is no significant decrease in the value of the variance as the degree of the polynomial increases [7] [10] [11]. In general, we need to balance the tradeoff between the bias and variance of the regression model. The estimated variance of the polynomial regression model is defined as:

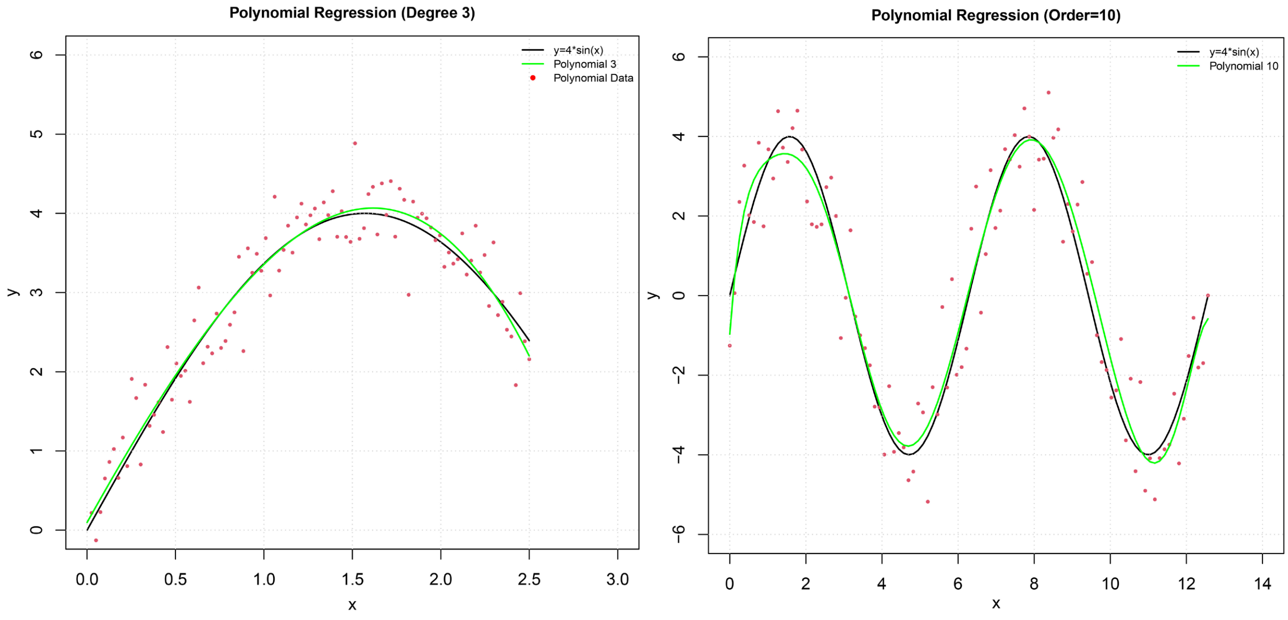
$$\hat{\sigma}^2 = \frac{Sr(k)}{n-k-1},$$

$Sr(k)$ sum square of the residuals for the k polynomial order;

k polynomial order;

n polynomial data points.

Figure 1 below shows different polynomial orders used to fit polynomial data.



(a) (b)

Figure 1. A plot of fitted nonlinear relationship by using different polynomial order.

3. Splines Regression

A spline is a piecewise function for which each segment is a polynomial function. Spline is designed to draw curves to balance the goodness of fit and minimize the mean square error of the model. We need to select certain breakpoints (called knots) to fit the regression model using splines. In splines, a linear or polynomial regression model is fitted between two knots. The degree of a polynomial and the number of knots must be determined to fit the splines regression model [11]. The piecewise regression between the knots is assumed to be a continuous polynomial function. Many different types of interpolating (splines) regression models can be used to fit the smoothing relationship between the explanatory variable and respond variables, such as cubic splines, B-splines, P-splines, natural splines, thin-plate splines, and smoothing splines [12] [13].

3.1. Basis Functions

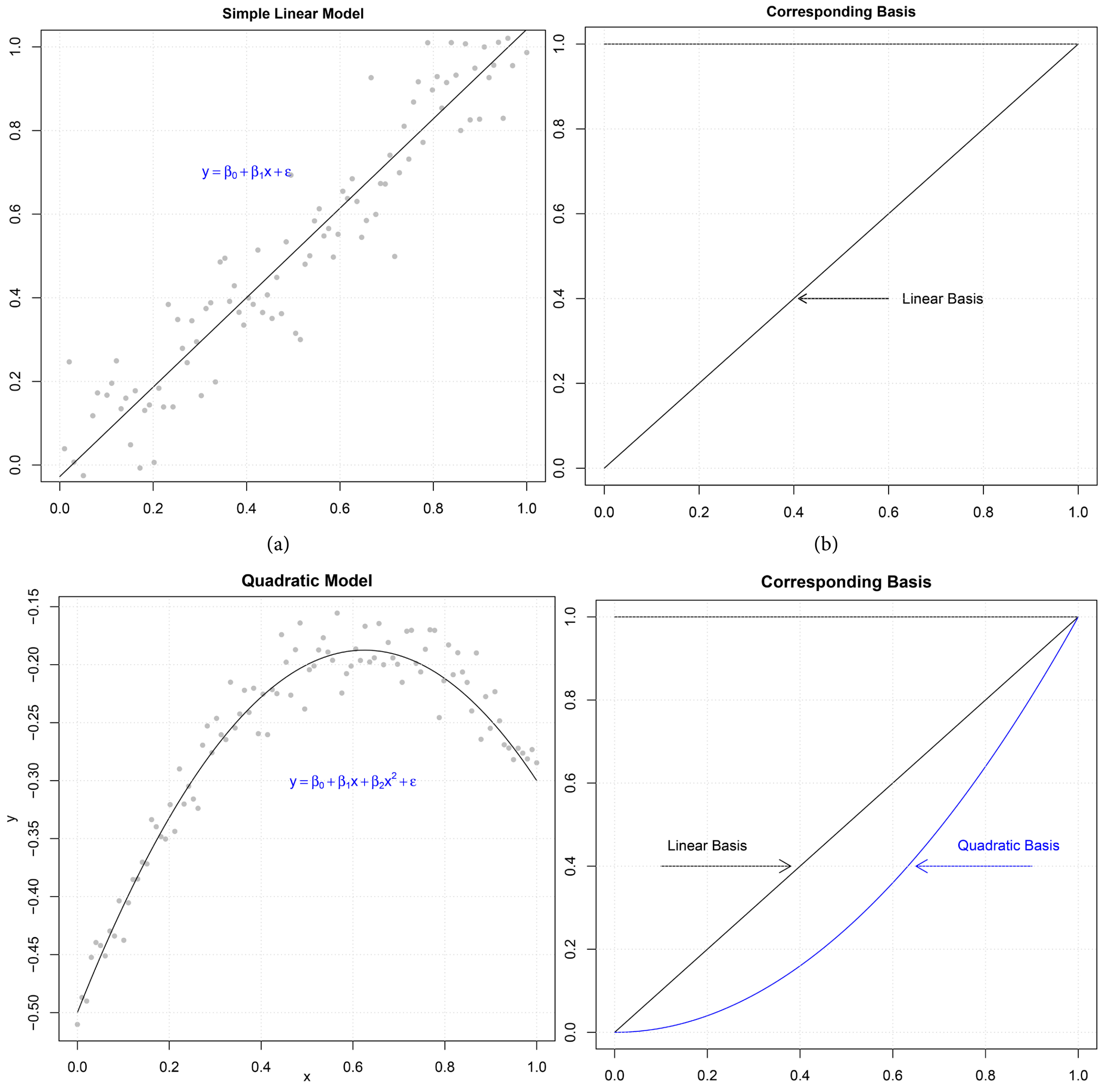
The basis function is defined as a set V of elements for which any elements of the space can be expressed uniquely as a linear combination of elements. Moreover, the regression model can be extended to accommodate nonlinear effects using some polynomials [14]. The most popular basis for the polynomial regression model will be introduced in this section. Basis functions for simple linear regression models are defined as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \ X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Using the matrix formula, we can obtain the vector of fitted values of $\hat{y}$ by:

$\hat{y} = X\left(X'X\right)^{-1}X'y = Hy$. The basis function for the quadratic model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,\ \ X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$



(a)

(b)

(c)

(d)

Figure 2. Top left: (a) Simple linear model representation; Top right: (b) Corresponding basis representation; Bottom left: (c) Quadratic model representation; Bottom right: (d) Corresponding basis representation.

The simple regression and quadratic regression model with their corresponding basis functions are illustrated in Figure 2(a) & Figure 2(b) for the simple regression model and Figure 2(c) & Figure 2(d) for the quadratic regression model. Moreover, the quadratic model is an extended simple linear regression model that accommodates and handles a different type of nonlinear structure.

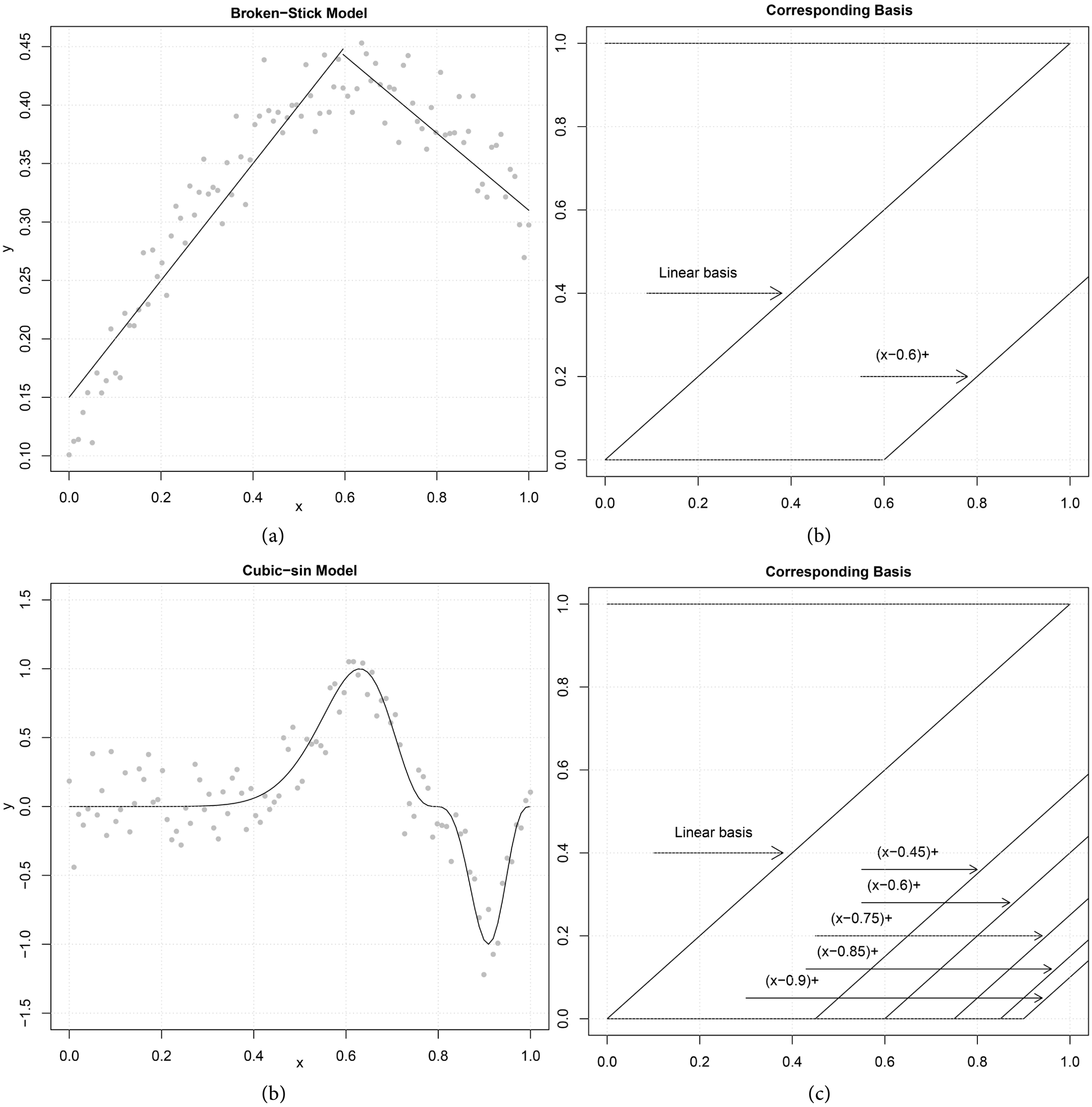


Figure 3. Top left: (a) Broken Stick model representation; Top right: (b) Corresponding basis representation; Bottom left: (c) Cubic model representation; Bottom right: (d) Corresponding basis representation.

The broken-stick regression model is a special case of the linear regression model with two differently sloped segment lines. The broken-stick regression model is proposed in order to make any complex and nonlinear function more suitable for modelling [15]. The broken-stick regression model demonstrated in

Figure 3(a) consists of two differently sloped lines that join together at $x = 0.6$. To introduce the basis function for broken-stick regression model, we need to find the slope for two lines that are connected at $x = 0.6$. Positive slop left of $x = 0.6$ and negative slop from $x = 0.6$ and onward. The new basis function of the broken stick model with two differently sloped lines can be expressed as $(x-0.6)_+$ (The positive part of the function $x - 0.6$).

$$y_i = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - 0.6)_+ + \varepsilon_i,\ \ X = \begin{bmatrix} 1 & x_1 & (x_1-0.6)_+ \\ \vdots & \vdots & \vdots \\ 1 & x_n & (x_n-0.6)_+ \end{bmatrix}$$

The cubic-sin model is more complicated than the broken stick model because there are several features, including peaks, troughs, and inflection points. Figure 3 panel (c) shows a cubic-sin model including straight lines and inflection points. The corresponding basis function for the cubic-sin model is demonstrated in Figure 3 panel (d). The cubic-sin model with associated basis function can be written by:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} b_k (x - k_k)_+,\ \ k = 0.45, 0.6, \cdots, 0.9,$$

$$X = \begin{bmatrix} 1 & (x_1-0.45)_+ & (x_1-0.6)_+ & \cdots & (x_1-0.9)_+ \\ & \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n-0.45)_+ & (x_n-0.6)_+ & \cdots & (x_n-0.9)_+ \end{bmatrix}$$

Because the function of the cubic-sin model contains multiple lines that are tied together at $x = k$. The value of k corresponding to the basis function is usually referred to knots [16]. Spline model can be estimated using a linear combination of basis functions $1, x, (x-k_1)_+, \cdots, (x-k_k)_+$ with multiple knots at $k_1, \cdots, k_k$.

3.2. Knots Placement and Numbers

In the spline regression model, the number of knots and their placement along the range of x must be determined by analysts. The analysts can override the default placement of knots, and most software packages place the knots in the data in either quartiles or quintiles. Even though the number of knots has an important effect on the spline fit, the analysts found that where the knots are placed matters less than how many knots are used [17]. A spline with two knots will be linear and globally smooth because there is only one piecewise function. By Increasing the number of knots, the number of piecewise functions for fitting the data will be increased. Moreover, selecting a large enough number of knots will control the number of piecewise fits and the amount of smoothing data [18]. Practically, evenly spaced intervals can be used as standard practice to place knots with each region of x to get a smooth fit for the data. The number of knots effectively acts as a span parameter for splines. Therefore, selecting the span parameter would go through tradeoffs. The estimated spline model using a small number of knots will be overly smooth with little variability and may be biased. Conversely, estimating the spline model using a high number of knots would imply a little bias but high variability, and the result may be overfitting the fit. The number of knots in a spline fit is not overly sensitive to the selected number of knots [12]. There are two existing methods for selecting a number of knots: visual trial and error, which involves adding or subtracting knots based on the fit, and Akaike's Information Criterion (AIC), which is less arbitrary and produces reasonable results. The choice of knots depends on sample size and sample size, with five knots for larger samples and three for smaller ones [18]. Knot selection can be complicated, but smoothing splines have made it easier to understand and compute.

3.3. Smoothing Splines

Smoothing splines are extensive techniques that minimize bias-variance tradeoffs, focusing on the solution to the penalized sum of squares. More details about smoothing splines can be found in [18] [19].

$$ss(f,\lambda)=\sum_{i=1}^{n}\left\{y_i-f(x_i)\right\}^2+\lambda\int_a^b f''(x)^2\,\mathrm{d}x$$

The residual sum of squares and roughness penalty are two key terms in smoothing splines. The first term is the residual sum of squares, while the second term is a roughness penalty, consisting of λ, a smoothing parameter, and the integrated squared second derivative of $f(x)$. The latter measures the rate of change of the slope for a function or curvature. As λ increases, the second derivative becomes constrained to zero, resulting in a smooth least squares fit. The penalty term ensures linearity and limits the approximate degrees of freedom [16].

3.4. Selecting the Smoothing Parameter

Local-polynomial regression and smoothing splines have adjustable smoothing parameters that can be selected by trial and error or cross-validation. Bulling functions from the ssanova library can be used to choose the smoothing parameter. The penalized spline model can be rewritten in matrix form and the penalty term can also be written as a quadratic form in β [16]. The matrix form of penalty term can be written as:

$$\int_a^b f''(x)^2\,\mathrm{d}x=\beta'D\beta,\ \ \beta=\left[\beta_0,\beta_1,\beta_{11},\cdots,\beta_{1k}\right]\ \text{and}\ D=\begin{bmatrix}0_{2\times2} & 0_{2\times k}\\ 0_{k\times2} & I_{k\times k}\end{bmatrix}_{(k+2)\times(k+2)}$$

where k denotes the number of knots. Therefore, the penalized spline regression model can be expressed in matrix notation as follows:

$$ss(f,\lambda)=\left\|y-X\beta\right\|^2+\lambda\beta'D\beta$$

A possible way to accommodate the penalized penalty term in the context of the standard linear regression matrix is by introducing the hat matrix [16]. A smoother matrix for splines can be derived by

$$S_\lambda=X\left(X'X+\lambda^2D\right)^{-1}X'$$

where λ^2 is penalty term scalar and it's multiplied by the matrix operator D. the optimal value of λ can be determined by using two approaches: Cross-Validation (CV) and Generalized Cross-Validation (GCV).

3.5. Cross-Validation

The fundamental concept of cross-validation is to leave the pair-point $\{x_i y_i\}_{i=1}^{n}$ out one at a time and choosing the smoothing parameter λ that minimizes the residual sum of squares. The squared residual for the function at point x_i estimates using the remaining $(n-1)$ data points. The cross-validation is given by:

$$CV(\lambda)=\sum_{i=1}^{n}\left\{y_i-\hat{f}_{-i}(x_i;\lambda)\right\}^2$$

where $\hat{f}_{-i}$ denotes the nonparametric regression that applied to the remaining data for which (x_i, y_i) were deleted. We choose the parameter λ that minimizes $CV(\lambda)$ and $\lambda \geq 0$. There are n-order algorithms for computation of $CV(\lambda)$ which is the most common smoothing technique [20]. For n versions of $\hat{f}_{-i}(x;\lambda)$ the vector of fitted values is defined by:

$$\begin{bmatrix}\hat{f}(x_1;\lambda)\\ \vdots \\ \hat{f}(x_n;\lambda)\end{bmatrix}=S_\lambda y \rightarrow \hat{f}(x_i;\lambda)=\sum_{i=1}^{n}S_{\lambda,ij}y_j$$

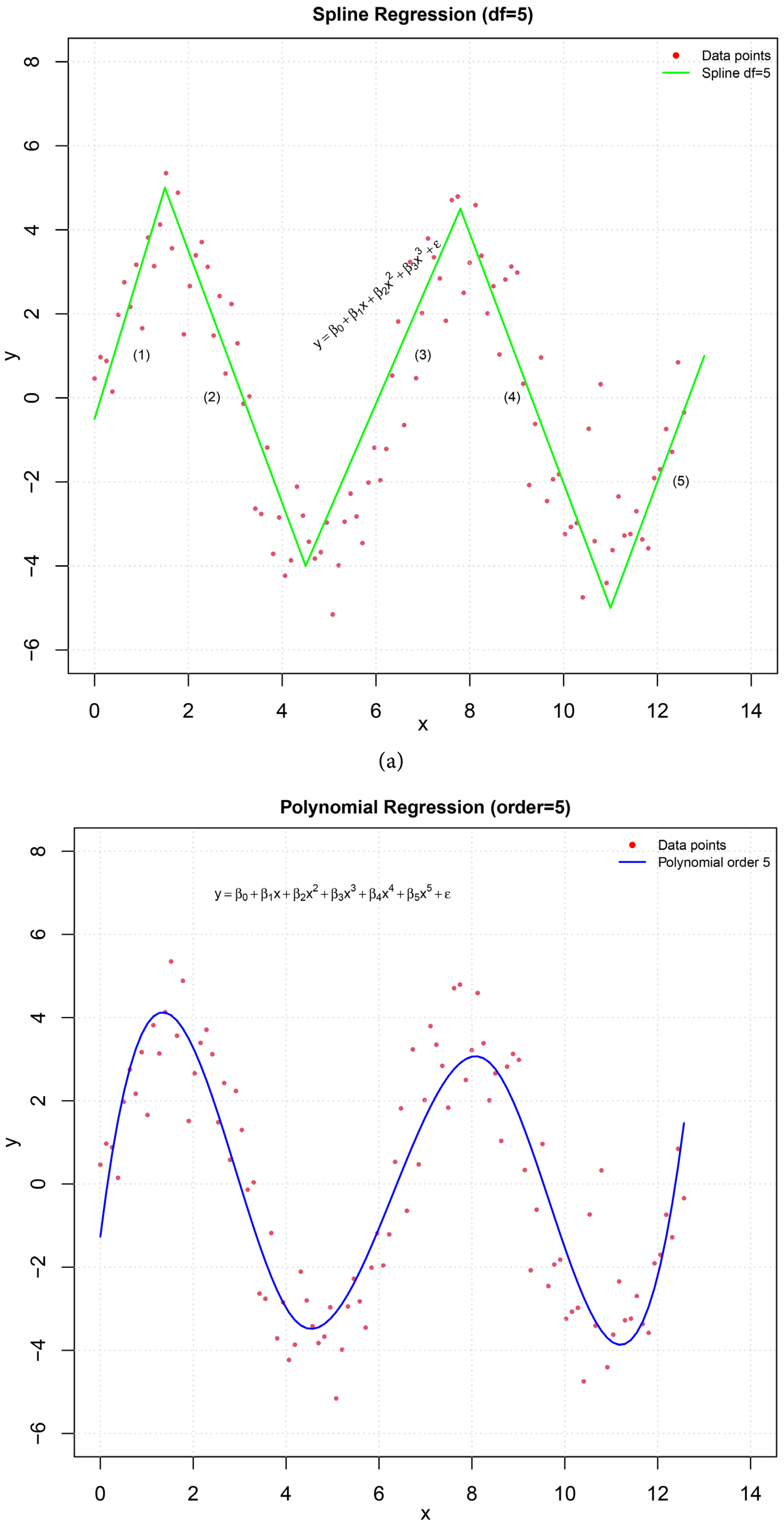
where $S_{\lambda,ij}$ is the (i,j) entry of S_λ. For many smoothers, $\hat{f}_{-i}(x_i;\lambda)=\frac{\sum_{j\neq i}S_{\lambda,ij}y_j}{\sum_{j\neq i}S_{\lambda,ij}}$. Even that the expression does not hold exactly, it usually holds approximately [21]. Also, we can use $\hat{f}_{-i}(x_i;\lambda)$ expression for cross-validation definition [22]. All smoothers used the sensible property that $y_i \equiv 1$ then $\hat{y}_i \equiv 1$, which implies that $\sum_{j=1}^{n}S_{\lambda,ij}=1$ for all i. Moreover, the denominator in the cross-validation expression is equal to $1-S_{\lambda,ii}$. Using cross-validation expression, we can show that

$$CV(\lambda)=\sum_{i=1}^{n}\left(\frac{y_i-\hat{f}(x_i;\lambda)}{1-S_{\lambda,ii}}\right)^2=\sum_{i=1}^{n}\left(\frac{\{(I-S_\lambda)y\}_i}{1-S_{\lambda,ii}}\right)^2=\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{1-S_{\lambda,ii}}\right)^2$$

Therefore, cross-validation can be computed using only ordinary residuals and the diagonal elements of the smoother matrix [23].

4. Simulation Study

The study was carried out to estimate the nonlinear model using the higher polynomial regression model and the splines regression model. This simulation study was also implemented to determine the relationship between the polynomial order and the number of knots that need to fit the model using splines. RStudio was used to generate a dataset for estimating nonlinear relations which the true equation was $y=\theta * \sin(\theta x)$. We used a sequence for independent variable x, fixed value for $\theta = 4$, and random noises from normal distribution to set a dataset that we would use for estimating a regression model. The regression model was fitted to the generated dataset using two proposed methods as shown in **Figure 4**. More results can be seen in **Figure 4** and **Table 1**.



(a)

(b)

Figure 4. It shows the association between polynomial order and the number of knots for estimating the nonlinear model.

Table 1. Regression model results using splines and polynomials.

	Splines model				Polynomial model			
Coefficients	Estimate	SE	t value	Pr (>|t|)	Estimate	SE	t value	Pr (>|t|)
β_0	−1.8889	0.6177	−3.058	0.00290**	−0.086	0.14197	−0.603	0.548
β_1	14.4120	1.1857	12.154	<2e−16***	−11.829	1.41970	−8.333	6.37e−13***
β_2	−13.7853	0.8107	−17.004	<2e−16***	−0.813	1.41970	−0.573	0.568
β_3	17.3196	1.0914	15.870	<2e−16***	−9.378	1.41970	−6.605	2.35e−09***
β_4	−10.2133	0.8905	−11.469	<2e−16***	−0.990	1.41970	−0.697	0.487
β_5	2.6959	0.8980	3.002	0.00343**	23.143	1.41970	16.302	2e−16***
MSE	0.01248936				0.01510638			
R^2	0.8643				0.8015			
R^2_{adj}	0.8571				0.791			

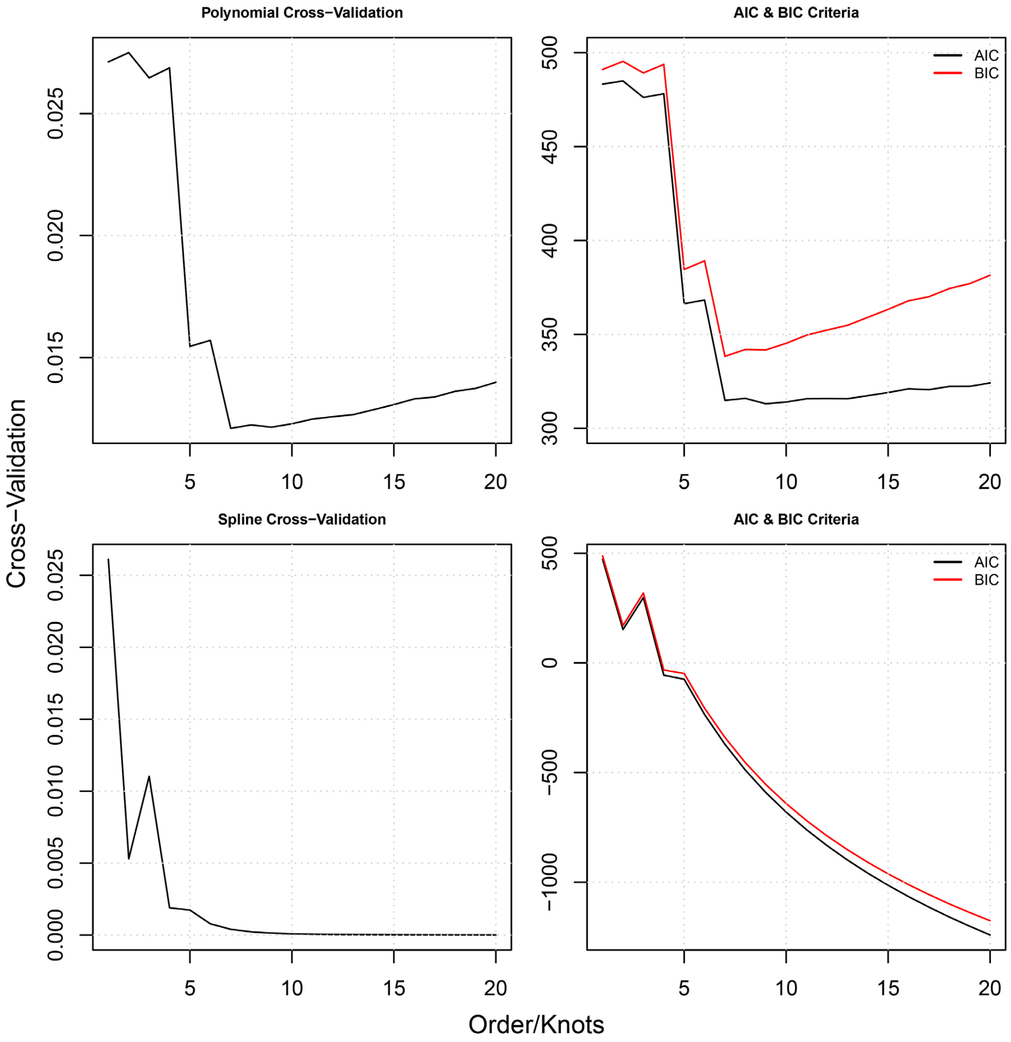


Figure 5. Top left: cross-validation of polynomial model based on different orders.; Top right: AIC&BIC for polynomial model; Bottom left: cross-validation of splines model based on different knots; Bottom right: AIC&BIC for splines model.

From the estimated model result, we can observe that both methods performed well in estimating nonlinear relationships. Moreover, by comparing the values of the MSE that are estimated using the proposed methods, we can realize that the splines model is more accurate than the polynomial model (assuming the polynomial order and splines knots are fixed with 5). In our study object, not only would we like to compare those estimated models, but we would also like to use the results of the polynomial model to determine the number of knots to fit the splines model. The estimated values of MSE, R^2, and R_{adj}^2 using the splines model, are 0.01248936, 0.8643, and 0.8571 respectively and compared with those estimated values using order five polynomial model 0.01510638, 0.8015, and 0.791. Even though some higher orders of polynomial coefficients are not significant, the estimated values of the MSE and R^2 of the model are still comparable.

The above plots show a comparison between the cross-validation and the model goodness of fit coefficients that were estimated using two models. From the plots, we can observe that the polynomial regression model, particularly with high-degree polynomials, tends to overfit and show excellent fit with order 5. Based on the results of the regression model and goodness of fits (AIC and BIC), we can see that the order 5 for the polynomial model also minimizes the model coefficients. A large number of knots for the splines model would decrease the model coefficients. Therefore, we can say that the best number of knots for modeling the data using the spline model is 5 knots, which also minimizes the spline model's goodness of fits coefficients. (**Figure 5**)

5. Conclusion

The first important step in building a realistic regression model is to understand the differences among these models (splines and polynomial regression). As we know, the polynomial regression model is smooth and fits with data wiggly; this is probably due to the high degree of freedom. The first-order polynomial model is a straightforward generalization of simple linear regression. Splines are more flexible and smoother than polynomial regression techniques. It is better in terms of extrapolation and is smoother. The complexity of using the spline model starts to increase when the number of knots and the number of intervals is increased. Therefore, we need to seek more to determine the best choice for the number of knots and the number of intervals first to build a regression model using splines. This study aimed to compare the results of the polynomial regression model and the splines regression model when the number of segments is equal to the polynomial degree. We compared the results that we got by using the polynomial regression model with the results that we got by using the splines regression model in terms of point estimates. The mean sum of squared errors and the fitted regression line show that the splines regression model improves well when the number of segments is equal to the polynomial degree of freedom. In our simulation study, order five of the polynomial model was the best choice for the spline segments.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Green, P.J. and Silverman, B.W. (1993) Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. CRC Press. https://doi.org/10.1201/b15710

[2] Sykes, A.O. (1993) An Introduction to Regression Analysis. Coase-Sandor Institute for Law & Economics Working Paper No. 20.

[3] Hamad, F. and Kachouie, N.N. (2018) A Hybrid Method to Estimate the Full Parametric Hazard Model. *Communications in Statistics—Theory and Methods*, **48**, 5477-5491. https://doi.org/10.1080/03610926.2018.1513149

[4] Elmesmari, N., Hamad, F. and Abdalla, A. (2021) Parameters Estimation Sensitivity of the Linear Mixed Model to Alternative Prior Distribution Specifications. *Scholars Journal of Physics*, *Mathematics and Statistics*, **8**, 166-170. https://doi.org/10.36347/sjpms.2021.v08i09.001

[5] Hamad, F., Younus, N., Muftah, M.M. and Jaber, M. (2023) Specify Underlining Distribution for Clustering Linearly Separable Data: Normal and Uniform Distribution Case. *Journal of Data Acquisition Processing*, **38**, 4675.

[6] Milliken, G.A., Bates, D.M. and Watts, D.G. (1990) Nonlinear Regression Analysis and Its Applications. *Technometrics*, **32**, 219-220. https://doi.org/10.2307/1268866

[7] Kim, Y. and Oh, H. (2021) Comparison between Multiple Regression Analysis, Polynomial Regression Analysis, and an Artificial Neural Network for Tensile Strength Prediction of BFRP and GFRP. *Materials*, **14**, Article 4861. https://doi.org/10.3390/ma14174861

[8] Theil, H. (1950) A Rank-Invariant Method of Linear and Polynomial Regression Analysis. *Indagationes Mathematicae*, **12**, 173.

[9] Jaber, M., Hamad, F., Breininger, R.D. and Kachouie, N.N. (2023) An Enhanced Spatial Capture Model for Population Analysis Using Unidentified Counts through Camera Encounters. *Axioms*, **12**, Article 1094. https://doi.org/10.3390/axioms12121094

[10] Hamad, A., Abdulkarim, S. and Hamad, F. (2023) First Order Harmonic Flow of Heavy Quarks Using a Hybrid Transport Model. *Scholars Journal of Physics*, *Mathematics and Statistics*, **10**, 49-52. https://doi.org/10.36347/sjpms.2023.v10i02.001

[11] Griggs, W. (2013) Penalized Spline Regression and Its Applications. Whitman Coll. https://www.whitman.edu/Documents/Academics/Mathematics/Griggs.pdf

[12] Keele, L. (2007) Semiparametric Regression for the Social Sciences. Wiley. https://doi.org/10.1002/9780470998137

[13] Jaber, M., Breininger, R.D., Hamad, F. and Kachouie, N.N. (2024) Spatiotemporal Bayesian Machine Learning for Estimation of an Empirical Lower Bound for Probability of Detection with Applications to Stationary Wildlife Photography. *Computers*, **13**, Article 255. https://doi.org/10.3390/computers13100255

[14] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. and Schmid, M. (2019) A Review of Spline Function Procedures in R. *BMC Medical Research Methodology*, **19**, Article No. 46. https://doi.org/10.1186/s12874-019-0666-3

[15] van Buuren, S. (2023) Broken Stick Model for Irregular Longitudinal Data. *Journal of Statistical Software*, **106**, 1-51. https://doi.org/10.18637/jss.v106.i07

[16] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) Semiparametric Regression. Cambridge University Press. https://doi.org/10.1017/cbo9780511755453

[17] Stone, C.J. and Koo, C.Y. (1985) Additive Splines in Statistics. *Proceedings of the American Statistical Association*, **45**, 45-49.

[18] Eilers, P.H.C. and Marx, B.D. (1996) Flexible Smoothing with B-Splines and Penalties. *Statistical Science*, **11**, 89-121. https://doi.org/10.1214/ss/1038425655

[19] Hastie, T. and Tibshirani, R. (1990) Exploring the Nature of Covariate Effects in the Proportional Hazards Model. *Biometrics*, **46**, 1005-1016. https://doi.org/10.2307/2532444

[20] Reinsch, C.H. (1967) Smoothing by Spline Functions. *Numerische Mathematik*, **10**, 177-183. https://tlakoba.w3.uvm.edu/AppliedUGMath/auxpaper_Reinsch_1967.pdf

[21] Wood, S.N. and Augustin, N.H. (2002) Gams with Integrated Model Selection Using Penalized Regression Splines and Applications to Environmental Modelling. *Ecological Modelling*, **157**, 157-177. https://doi.org/10.1016/s0304-3800(02)00193-x

[22] Hutchinson, M.F. and de Hoog, F.R. (1985) Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, **47**, 99-106. https://doi.org/10.1007/bf01389878

[23] Craven, P. and Wahba, G. (1978) Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik*, **31**, 377-403. https://doi.org/10.1007/bf01404567