SPLINE REGRESSION ANALYSIS WITH APPLICATION TO CLINICAL DATA: A REVIEW

DARE, W.M.* AND ADELEKE, K.A.
Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria.

ABSTRACT

Building reliable multivariable regression models is a major concern in many application areas, when one or more of the predictor(s) is/are continuous, the question arises of how to represent the relationship meaningfully following substantive background knowledge. In this paper, we review and present in epidemiological context polynomial regressions splines with applications in real-life clinical data. We also investigate if the added complexity of the spline regression models is justified by a significantly better fit (under certain conditions). Assumptions of constant variance, zero mean and many other properties were examined using residual versus fitted values plot. We imposed the assumption of no discontinuity at the spline knot respectively. Following the outcome of the scattered plot of survival time data for cancer patients, our data structures follow a quadratic relationship, based on all criteria. Also, the cubic polynomial regression model and cubic spline revealed that indeed the regression diagnostic tools such as the Normal Q-Q plot do not show a serious departure from the normality assumption. However, by assessing the contribution of the cubic effect parameter, the cubic model is inadequate to fit the data, Hence, employing regression diagnostics visualization plot/ techniques alone for assessing the quality of the fit of the data to a model is good but not sufficient enough to judge the suitability and adequacy of a model.

Keywords: Cubic Spline, Knots, Panel Data, Quadratic spline, Penalized spline, Spline regressions, Traditional regression
***Correspondence:** dareabdulwasiu99@gmail.com

INTRODUCTION

Polynomial regression is a higher order form of linear regression in which the relationship between the independent variable X and the dependent variable Y is modeled as an n^{th} order polynomial. Polynomial regression analysis consists of techniques for modeling the relationship between a dependent variable (also called response variable) and one or more independent variables (also known as explanatory variables or predictors) and the error term. The random error term represents variation in the dependent variable unexplained by the function of the dependent variables and coefficients [1]. Spline smoothing involves modeling a regression function as a piecewise polynomial with a high number of pieces relative to the sample size [2]. Splines are piecewise polynomials of order m, the joint points of the pieces are usually called knots (such as specific date, structural changes, e.t.c.). Strategies are required for knot selection and location which can be accomplished by various methods. One may use predetermined knots, natural division points, or visually inspect the data. There are also other (more complex) methods, such as nonlinear least squares methods, for knot selection [3]. Polynomial regression may be considered a special case of spline regression with no knots, [4, 5, 6]. If however, the piecewise linear model is not smooth at the knots, we replace the basis of linear functions with a basis of splines, that is a basis of functions with continuous first derivatives or still more regularity first and second derivatives continuous), [7]. The model is best used in situations where the analyst discovered that curvilinear effects are present in the true response function. Also, in approximating function for unknown and possible very complex nonlinear relationships. [8] suggested using spline regression (and fractional polynomial regression) as an alternative to categorical analysis for dose-response and trend analysis, stating that categorical analysis does not make use of within category information and is based on an unrealistic model for dose-response and trends. Univariate polynomial splines are piecewise polynomials in one variable of k- degree with function values and first k-1 derivatives that agree at the points where they join. The joint points (or abscissa values) that mark one transition to the next are referred to as break points, interior knots, or simply knots [3, 6]. Knots give the curve freedom to bend and more closely follow the data. Splines with few knots are generally smoother than splines with many knots; however, increasing the number of knots usually increases the fit of the spline function to the data. [9]. Spline functions have been applied to medical and epidemiological investigations such as survival analysis, linear dose-response problems, latency patterns, and data smoothing (to detect trends) as well as other studies. An Example of such is an assessment of mortality in colon cancer using survival analysis methods. [10] used restricted cubic splines to model time-by-covariate interactions. A penalized spline method was applied to a cohort study

of autoworkers exposed to metalworking fluids to examine the linearity assumption for prostate and brain cancer mortality [11]. In a study to estimate longitudinal immunological and virologic markers in HIV patients with individual antiretroviral treatment strategies, Brown et al. [12] proposed univariate and bivariate cubic smoothing splines to fit CD4+ count and plasma viral load. Wu [13] applied spline regression to dose-response in epidemiology analysis in observational data, He however showed that spline can serve as a useful mechanism for viewing structural change in a continuous regression model, He also used restricted cubic spline to check nonlinearity and the overall trend in nutritional epidemiology. The research work by Fuller [14] discussed generally linear and quadratic polynomial in two variables and indicated how to test for the significance of spline terms. He illustrated their uses in a fertilizer experiment in which corn yield was written as a function of nitrogen and phosphorus in a 5 * 5 experimental design. Royston and Sauerbrei [15] research work also use spline regression to discuss issues in modeling a single risk variable, he adopted a cubic spline to fit a model on alcoholic consumption as a risk factor for oral cancer. Splines regression with fixed knot however is straightforward ordinary least square regression. Deciding on the number of knots and the degree of the polynomial pieces is still a problem from the statistical point of view. Wold [16], gave several examples of how splines with fixed knot(s) can be used as an advantage as a data fitting tool. He suggested that there should be as few knots as possible, with at least four or five data points per segment. Considerable caution should be exercised here because the great flexibility of spline functions makes it very easy to overfit the data, furthermore, he also suggested that there should be no more than one extreme point (maximum of minimum) and one point of inflection per segment. He restricted himself to cubic polynomial because of the lower degree. The basic cubic spline model can easily be modified to fit polynomials of different order such that in each segment one can impose different continuity restrictions at the knot. A potential disadvantage of this method is that the X'X matrix becomes ill-conditioned, [17]. Thurston et al. [11] applied penalized spline methodology to a cohort study of autoworkers exposed to metalworking fluids to examine the linearity assumption for prostate and brain cancer mortality. Sometimes a low-order polynomial provides a poor fit to the data, but increasing the order of the polynomial modestly does not seems to help, symptoms of this are the failure of the residual sum of squares to stabilize or residual plots that exhibit remaining unexplained structure, the problem may occur when the function behaves differently in different parts of the range of X, and as the order the polynomial increases, the matrix X'X becomes ill-condition, which means the matrix inversion calculation will be inaccurate and considerable error may be introduced into the parameter estimate [18], which could create strong multicollinearity between the different basis of X. In this paper, we review and present in epidemiological context polynomial regressions splines and also encourage good practice in the regression analyses. We aim to highlight the statistical problem that is caused if a regression model adequately reflects the true relationship between the response variable and independent variables. We also investigate if the added complexity of the spline regression models is justified by a significantly better fit (under certain conditions) than a regression model. Although, problem(s) may arise when a function behaves differently in different parts of the range of X (independent variable), as well as when the order of the polynomial increases.

MATERIALS AND METHODS

A function f:[a,b] – R is called a polynomial spline of degree $m \geq 0$ with knots a $=t_1 < t_2 < \ldots << t_n = b$, if it fulfills the following conditions :

1. f(x) is (m-1) – times it is continuously differentiable, the special case of m=1 corresponds to f(x) being continuous but not differentiable.
2. f(x) is a polynomial of degree m on the interval $(< t_j\ ,\ \ t_j + 1\ )$ define by knots.

Quadratic and cubic polynomial regressions

Consider a polynomial regression models with one variable,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

$$\& \qquad (1)$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

These models are called second-order and third-order models in one variable, sometimes called a quadratic and cubic regression model, y is called the dependent variable, $and\ x\ is$ the independent variable. β_0, when all explanatory variables are zero's, is the intercept of the regression line, β_1 the linear effect parameter, β_2 is the quadratic effect parameter, β_3 is the cubic effect parameter, and ε is the error term. It is usually assumed that error ε is normally distributed with $E(\varepsilon) = 0$ and a constant variance. Variance $(\varepsilon) = \sigma^2$ in the regression model.

Piecewise quadratic and piecewise cubic polynomial regressions

An important special case of practical interest involves fitting piecewise quadratic and cubic regression models. This can be treated using quadratic and cubic splines. For example,
suppose that there is a single knot t, say, with continuous first, second, and third derivatives.

The resulting quadratic spline model is given as

$$y = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_{21}(x-t)_+^2 + \varepsilon (2)$$

Now if x ≤ t, the quadratic spline model equals the quadratic polynomial regression model,

$$E(y) = \beta_0 + \beta_1 x + \beta_{(2)} x^2$$

and if x > t, the model is

$$E(y) = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_{21}(x-t)^2$$
$$= (\beta_0 - \beta_{21} t^2) + (\beta_1 + 2\beta_{21} t)x + (\beta_2 - \beta_{21})x^2 (3)$$

where:
$(\beta_0 - \beta_{21} t^2)$ is the new intercept,
$(\beta_1 + 2\beta_{21} t)$ is the new linear effect parameter,
$(\beta_2 - \beta_{21})$ is the new quadratic effect parameter.

The resulting cubic spline model is given as

$$y = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_3 x^3 + \beta_{31}(x-t)_+^3 + \varepsilon \quad (4)$$

Now if x ≤ t, the cubic spline model equals the cubic polynomial regression model,

$$E(y) = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_3 x^3$$

and if x > t, the model is

$$E(y) = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_{21}(x-t)^2$$
$$= (\beta_0 - \beta_{21} t^2) + (\beta_1 + 2\beta_{21} t)x + (\beta_2 - \beta_{21})x^2 (3)$$

where:
$(\beta_0 - \beta_{21} t^2)$ is the new intercept,
$(\beta_1 + 2\beta_{21} t)$ is the new linear effect parameter,
$(\beta_2 - \beta_{21})$ is the new quadratic effect parameter.

The resulting cubic spline model is given as

$$y = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_3 x^3 + \beta_{31}(x-t)_+^3 + \varepsilon \quad (4)$$

Now if x ≤ t, the cubic spline model equals the cubic polynomial regression model,

$$E(y) = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_3 x^3$$

and if x > t, the model is

$$E(y) = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_3 x^3 + \beta_{31}(x-t)^3$$
$$= (\beta_0 - \beta_{31} t^3) + (\beta_1 + 3\beta_{31} t^2)\, x \quad + (\beta_2 - 3t\beta_{31})\, x^2 + (\beta_3 + \beta_{31})\, x^3 \quad (5)$$

where:
$(\beta_0 - \beta_{31} t^3)$ is the new intercept,
$(\beta_1 + 3\beta_{31} t^2)$ is the new linear effect parameter,
$(\beta_2 - 3t\beta_{31})$ is the new quadratic effect parameter,
$(\beta_3 + \beta_{31})$ is the new cubic effect parameter.

Generalization of quadratic and cubic splines

Suppose we extend quadratic and cubic spline regressions from a single knot with no continuity restriction at the knot to q knots, then a projection of y on its basis remains to estimate a piecewise quadratic and cubic models as follows,
$y = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \sum_{(i=1)}^{q} \beta_{2i}(x-t_i)_+^2 + \varepsilon$ and

$$y = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_3 x^3 + \sum_{(i=1)}^{q} \beta_{3i}(x-t_i)_+^3 + \varepsilon (6)$$

where ε is the random error term and

$$(x-t_i)_+ = \begin{cases} (x-t_i), if\,(x-t_i) > 0 \\ 0, otherwise \end{cases}$$

where $(x-t_i)_+ = \max(0, x-t)$, which means it measures the distance from the point $x = t$ on the right side of t and is set to zero for any value of x below the knot location.

Penalized quadratic and cubic splines

Let us consider the models

$$y = m(\mathrm{x}) + \varepsilon$$
$$y = \tau(\mathrm{x}) + \varepsilon \quad (7)$$

with

$$m(x) = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \sum_{i=1}^{q} \beta_{2i}(x-t_i)_+^2$$

(quadratic spline basis; q knots) and

$$\tau(x) = \beta_0 + \beta_1 x + \beta_{(2)} x^2 + \beta_3 x^3 + \sum_{i=1}^{q} \beta_{3i}(x-t_i)_+^3$$

(cubic spline basis; q knots), the ordinary least-squares can be written as
$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$minimizes $(y-\hat{y})^2$ with
$\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_{31}, \beta_{32}, \ldots, \beta_{3q})^{\boldsymbol{T}}$ with

$$X = \begin{pmatrix} 1 & x_1 & (x_1-t_1)_+ & \cdots & (x_1-t_q)_+ \\ 1 & x_2 & (x_2-t_1)_+ & \cdots & (x_2-t_q)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n-t_1)_+ & \cdots & (x_n-t_q)_+ \end{pmatrix}$$

Unconstrained estimation of$(\beta_{21}, \beta_{22}, \ldots, \beta_{2q})$ and $(\beta_{31}, \beta_{32}, \ldots, \beta_{3q})$leads to a wiggly fit.

For judicious choice of M, a constraint of the type $\sum_{i=1}^{q}\beta_{2i} < M$ and $\sum_{i=1}^{q}\beta_{3i} < M$may rectify this situation and lead to a smoother fit to the scatter plot. If we define the $(q+3)*(q+3)$ matrix for quadratic spline and $(q+4)*(q+4)$ for cubic spline.

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

and

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0}_{3*3}\mathbf{0}_{3*q} \\ \mathbf{0}_{q*3}\boldsymbol{I}_{\boldsymbol{q*q}} \end{pmatrix}, \qquad = \begin{pmatrix} \mathbf{0}_{4*4}\mathbf{0}_{4*q} \\ \mathbf{0}_{q*4}\boldsymbol{I}_{\boldsymbol{q*q}} \end{pmatrix},$$

the minimization problem can be written as

$\min(y-\hat{y})^2$
with respect to $(\boldsymbol{\beta^t D\beta})$
or
$(\sum_{i=1}^{q}{\beta_{2i}}^2 < M)$or $(\sum_{i=1}^{q}{\beta_{3i}}^2 < M)$

which is equivalent to choosing β to minimize

$$(y-\hat{y})^2 + \lambda\sum_{i=1}^{q}{\beta_{2i}}^2$$

or

$$(y-\hat{y})^2 + \lambda\sum_{i=1}^{q}{\beta_{3i}}^2$$

for some $\lambda \geq 0$. Thus having the solution of

$$\widehat{\beta_\lambda} = (\boldsymbol{X^T X\lambda D})^{-1}\boldsymbol{X^T y}.$$

The term

$\lambda\sum_{i=1}^{q}{\beta_{2i}}^2$ or

$$\lambda\sum_{i=1}^{q}{\beta_{3i}}^2$$

is called a roughness penalty because it penalizes fits that are too rough, thus yielding a smoother result. The amount of smoothing is controlled by (the smoothing parameterλ). Provided the knots cover the range of x_i's values reasonably well, their number and positioning do not do much difference to the result. However, λ has quite a big effect.

Q-Q plot

Q–Q plots are commonly used to compare a data set to a theoretical model. It provides an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary or comparing two theoretical distributions to each other. Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.

Analysis of variance test

Analysis of variance approach can be used to test the significance of regression, it is based on partitioning of the total variability in the response variable y.
To obtain this partitioning, we begin with the identity

$$(y_i - \overline{y}) = (\hat{y_i} - \overline{y}) + (y_i - \hat{y_i})$$

where
$y = observed\ value,$
$\bar{y} = overall\ average\ of\ y_i\ and$
$\hat{y_i} = fitted\ value\ of\ y_i$

Data presentation and analysis for quadratic and quadratic spline regression

Data on Survival time for six types of patients [19], types of cancers are stomach, colon, bronchus, rectum, bladder, and kidney. The assigned factor is Age and the dependent variable is mean survival time after all treatment ceased for the patient's 10g/day ascorbic vitamin C matched controls. The data can be found at https://bit.ly/3gknbAd. Judging from Figure 1, the graph is highly peaked (leptokurtic) in nature with normal, but still, gives a suspicion on the normality assumption. Also, from Figure 2, the curve is rightly skewed, an example of rightly skewed distribution is Exponential distribution, Weibull distribution, etc.

Position of knots

Without the imposition of continuity condition, then two knots were chosen to allow the function to vary on up to three segments for the data structure. The placement of the knots was at $t = 58$ and $t = 70$ for the Age-Mean survival data, and $t = Maryland(17.7)$ and $t = Tennessee\ (22.6)$ for the Oral-Mortality data, which creates nearly three equal intervals) between each data

set. This was done to keep the number of observations within each segment approximately the same.

RESULTS AND **DISCUSSION**

We may easily compare the quadratic spline model with the quadratic polynomial regression. Quadratic polynomial regression contains fewer parameters and would be preferred to the quadratic spline model if it provides a satisfactory fit. The quadratic polynomial regression is given as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

and the least-square fit is

$$\hat{y} = 65.77 - 2.83x - 0.04x^2$$

Quadratic spline model

The cubic spline model using two knots at $t_1 = 58$ and $t_2 = 70$ given as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_{21}(x - 58)_+^2 + \beta_{22}(x - 70)_+^2 + \varepsilon$$

And the least-squares fit is

$$\hat{y} = -50.98 + 2.20x - 0.02x^2 + 0.15\,(x - 58)_+^2 - 0.18(x - 70)_+^2$$

Regression diagnostics for quadratic polynomial and quadratic spline regression models

Regression diagnostics provide visualization techniques for assessing the quality of the fit of the data to a model. Residuals vs. Fitted plot help to identify patterns in residuals. Normal Q-Q plots help assess the normality of the residuals. A scatter diagram shown in Fig.3 of these data displays knowledge of the quadratic relationship and reasonably suggests that a quadratic model may adequately describe the relationship between the age and mean survival time data for the cancer patients. From quadratic polynomial regression, a plot of the residuals versus the fitted. Fig. 4 does not reveal a serious departure from assumptions, we do not seriously question the normality assumption, so we conclude that the quadratic model is adequate to fit the data.

The Normal Q-Q plot for the quadratic model in Figure 5 shows a slightly wiggle head, the normality assumption is likely to sound. For the objective approach, we also investigate the quadratic effect parameter by testing the null hypothesis $H_0: \beta_2 = 0$, using the extra sum of squares method. The regression sum of squares for the q spline is $SS_R\left(\frac{\beta_2}{\beta_1}, \beta_0\right) = 1906.3$ with one degree of freedom.

$$F_0 = \frac{SS_R\left(\frac{\beta_2}{\beta_1}, \beta_0\right)}{MS_{Res}}$$

$= \frac{1906.3}{31.8} = 59.95$, $F_{1,60,0.95} = 4.00$ on 1 and 60df. We reject the formulated belief that $\beta_2 = 0$, and conclude that the quadratic regression model provides a better fit.

However, in quadratic spline regression, the plot of residuals against the fitted value in Fig.6 is mildly disturbing, it revealed a not-too-serious away from assumption compared to quadratic polynomial regression. The Normal QQ-plot of the residuals in Fig.7 shows a slight deviation from normality, namely hardly any tail events. This is not as serious as the Normal QQ-plot of the residuals for quadratic polynomial regression Figure 5.

The result of our investigation on the contribution of the quadratic spline effect to the model shows that

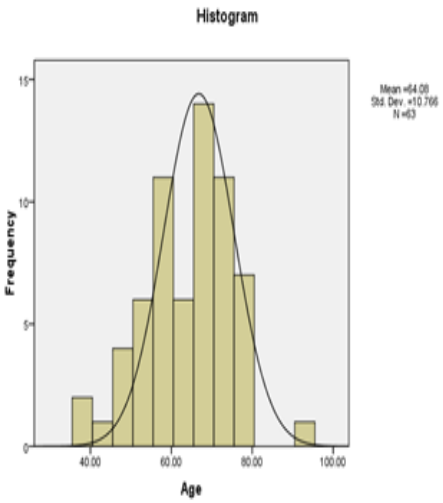

Fig 1: A histogram showing the age of the cancer patients.

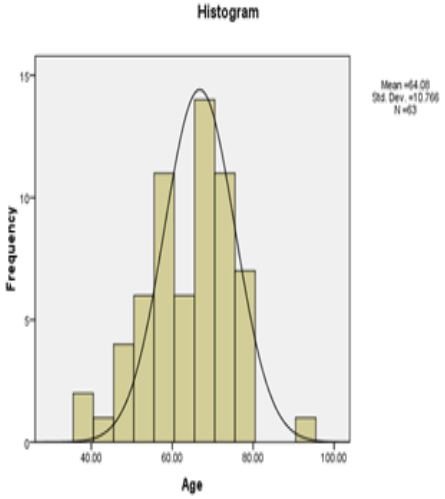

Fig 2 : A histogram showing the mean survival time after all treatment ceased for the patient's 10g/day ascorbic vitamin c matched control of the cancer patients

(i.e $H_0: \beta_{21} = \beta_{22} = 0$, using the extra sum of square method). The regression sum of squares for the quadratic spline is $SS_R\left(\frac{\beta_{21},\beta_{22}}{\beta_0}, \beta_1, \beta_2\right) = 789.5$ with two degrees of freedom.

Since $F_0 = \frac{SS_R\left(\frac{\beta_{21},\beta_{22}}{\beta_0},\beta_1,\beta_2\right)}{MS_{Res}} = \frac{789.5}{19.2} = 41.12$, and $F_{2,58,0.95} \approx 4.00$,

We reject the null hypothesis that: $\beta_{21} = \beta_{22} = 0$, and conclude that the quadratic spline regression provides a better fit.

Data presentation and analysis for cubic and cubic spline regression

Table B shows the Oral-Mortality rate (https://bit.ly/3aSyQW4) through cigarette consumption by states for the year 2006 in the United States, 44 states were taken into consideration and a task free on the risk factor was listed first. From Figure 8, the graph is highly peaked (leptokurtic) in nature with a normal shape (mesokurtic), a suspicion of normality assumption is embedded within.

Cubic polynomial regression

We may easily compare the cubic spline model with the cubic polynomial regression. Cubic polynomial regression contains fewer parameters and would be preferred to the cubic spline model if it provides a satisfactory fit.

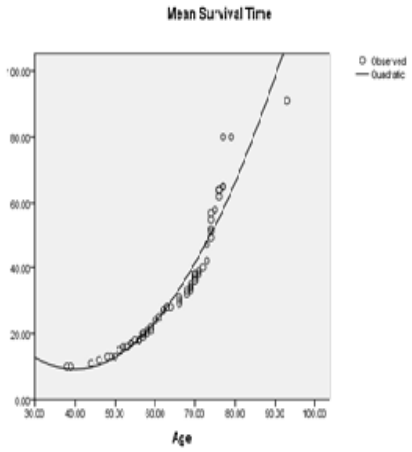

Fig. 3. Scatterplot of Age-Mean survival time data for the cancer patients.

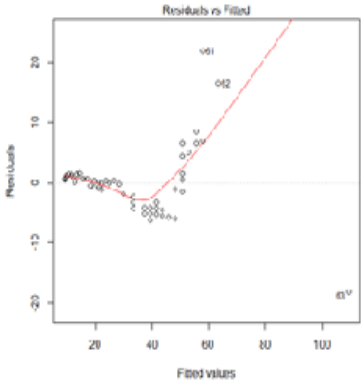

Fig.4: Plot of residual e_i versus fitted value y_i for the quadratic polynomial model.

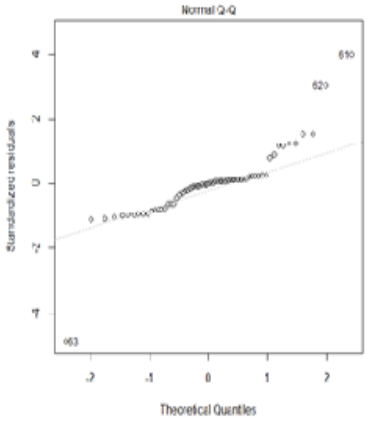

Fig.5 : Normal Q-Q plot for the quadratic polynomial regression.

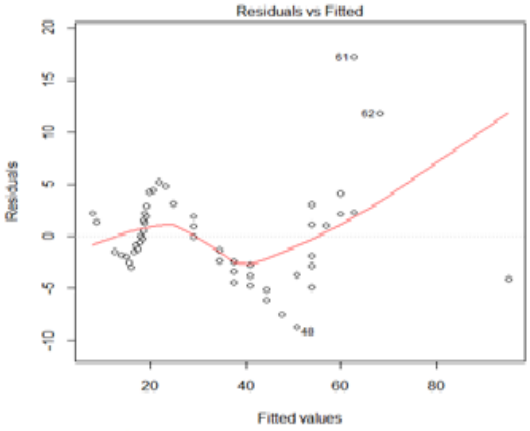

Fig. 6 : Plot of residual e_i versus fitted value y_i for the quadratic spline regression model.

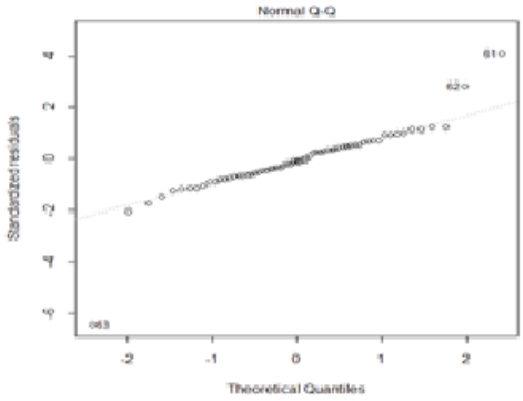

Fig. 7 :Normal Q-Q plot for the quadratic spline regression model.

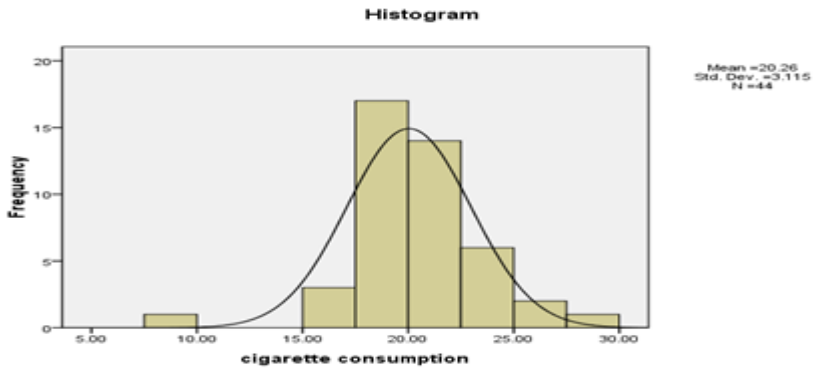

Fig 8: A histogram showing the cigarette consumption for the 44 states.

The cubic polynomial regression is given as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

and the least-square fit is

$$\hat{y}=0.6989 + 0.3632x - 0.0242x^2 + 0.0005x^3$$

Cubic spline model

The cubic spline model using two knots at $t_1 = 17.7$ and $t_2 = 22.6$ given as

$$y = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_{31}(x - 17.7)_+^3 + \beta_{32}(x - 22.6)_+^3 + \varepsilon$$

and the least-squares fit is

$$\hat{y} = -4.3713 + 1.40x - 0.09x^2 + 0.0019x^3 - 0.0016(x - 17.7)_+^3 - 0.0013(x - 22.6)_+^3$$

Regression diagnostics for cubic polynomial and cubic spline regression models

A scatter diagram shown in Fig.9 of this data display a knowledge of the cubic relationship and reasonably suggests that a cubic model may adequately describe the relationship between cigarette consumption by states and oral mortality rate, it revealed an S shape if viewed in a North-South direction. For the cubic polynomial regression model, a plot of the residuals versus the fitted in Fig.10 revealed a serious departure from assumptions, we seriously question the normality assumption, so we conclude that the cubic polynomial is an inadequate model for the Oral-Mortality data.

The Normal Q-Q plot for cubic polynomial in Fig. 11 shows a slightly wiggle head and tail normality assumption is likely to sound.

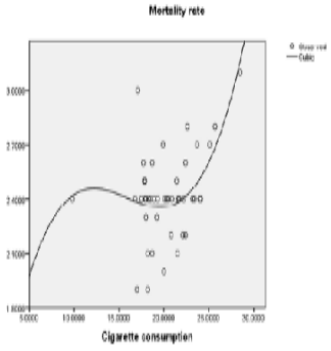

Fig. 9: Scatterplot of Oral-Mortality cancer data.

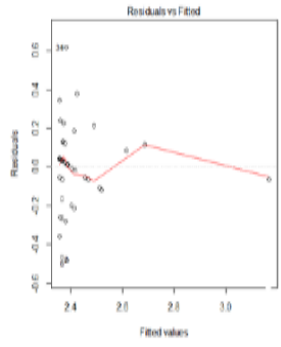

Fig. 10: Plot of residual e_i versus fitted value y_i for th polynomial regression model.

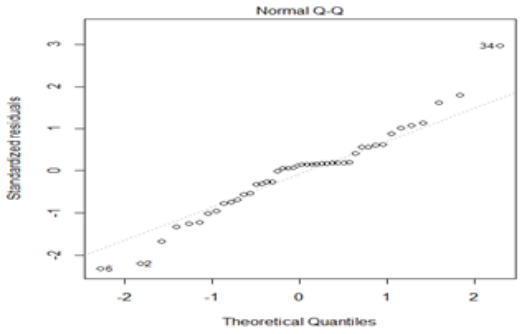

Fig. 11: Normal Q-Q plot for the cubic polynomial regression model.

For the objective approach, we also investigate the cubic effect parameter by testing the null hypothesis $H_0: \beta_3 = 0$, using the extra sum of the square method. The regression sum of squares for the cubic polynomial regression model is

$SS_R\left(\frac{\beta_3}{\beta_0}, \beta_1, \beta_2\right) = 0.1314$ with 1 degree of freedom.

$$F_0 = \frac{SS_R\left(\frac{\beta_3}{\beta_0}, \beta_1, \beta_2\right)}{MS_{Res}} = \frac{0.1314}{0.0468} = 2.808\,,$$

$F_{1,40,.95} \approx 4.08$ on 1 and 40 DF, we do not reject the formulated belief that $\beta_3 = 0$ and conclude that the cubic polynomial regression does not provide a better fit. However, in cubic spline regression, the plot of residuals against the fitted value in Fig.12 has more observations outside zero, revealing a serious departure from the assumption, hence, cubic spline also does not provide an adequate fit to the data.

The Normal Q-Q plot for cubic spline in Fig. 13 shows a slight deviation from normality, namely hardly any tail events.

The regression sum of squares for the cubic spline is 0.02393 with 2 degrees of freedom. Since

$$F_0 = \frac{SS_R\left(\frac{\beta_{31}, \beta_{32}}{\beta_0}, \beta_1, \beta_2, \beta_3\right)}{2} / MS_{Res} = \frac{0.01197}{0.04861} = 0.2463$$

, $F_{2,38,0.95} = 3.23$ on 2 and 38 DF, which would be referred to the $F_{2,38}$ distribution, We do not reject the null hypothesis that $\beta_{31} = \beta_{32} = 0$. We conclude that the cubic spline does not spline a better fit.

SUMMARY

For any regression procedure, it is desirable to use models that closely fit the data. This study critically looked into modeling the polynomial and spline regression model and its traditional counterpart, (polynomial regression on real-life data).

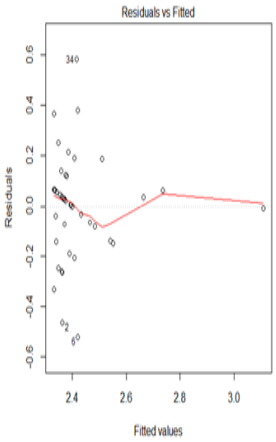

Fig.12 : Plot of residual e_i versus fitted value y_i for the cubic spline regression model.

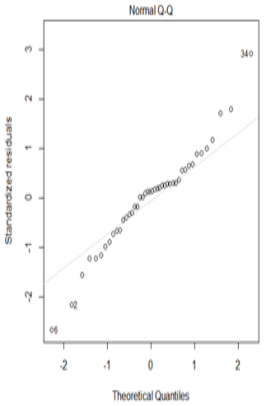

Fig. 13: Normal Q-Q plot for the cubic spline model.

CONCLUSION

Following the outcome of the scattered plot of survival time data for cancer patients, our data structures follow a quadratic relationship, based on all criteria. This means that it is necessary to employ all regression diagnostics visualization plots/techniques for assessing the quality of the fit of the data to a model which is not alone sufficient. Reviewing each structure independently, it is possible to draw some general conclusions about the quadratic polynomial structure and the quadratic spline with two knots location and to determine when to use the more complicated spline models. If the data between all knot locations and endpoints (all partitions) has a quadratic data structure, is continuous, and has different slopes, then the quadratic spline is the most appropriate modeling tool. Finally, when a data structure follows a more purely polynomial shape (e.g. quadratic), then quadratic polynomial regression (that is the representative of the overall respectively) is the best model. For the purely polynomial structures, both the polynomial model and the spline model are "correct" models for this type of structure. However, the traditional regression models do not require knowledge of knot location and variation in the data. Furthermore, for the cubic polynomial regression model, likewise cubic spline we can see that indeed the regression diagnostic tools such as the Normal Q-Q plot do not reveal a serious departure from normality assumption, however, by assessing the contribution of the cubic effect parameter, the cubic model is inadequate to fit the data, hence, a further exercise is required in checking whether a particular model fits a data.

ACKNOWLEDGMENTS

Special thanks to those who contributed in one way or the other to this work and the anonymous reviewers.

REFERENCES

1. YAN, X. & SU, X. (2009). *Linear regression analysis: theory and computing*. World Scientific.
2. WAND, M.P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics*, **15**(4): 443-462.
3. EUBANK, R.L. (1999). *Nonparametric regression and spline smoothing*. CRC press.
4. ANDREWS, D.F. & HERZBERG, A.M. (2012). *Data: a collection of problems from many fields for the student and research worker*. Springer Science & Business Media.
5. MONTGOMERY, D.C., PECK, E.A. & VINING, G.G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.
6. POIRIER, D.J. (1976). *econometrics of structural change, with special emphasis on spline functions*. North-Holland Pub. Co.
7. RUPPERT, D., WAND, M.P. & CARROLL, R.J. (2003). *Semiparametric p[=regression* (No. 12). Cambridge university press.
8. GREENLAND, S. (1995). Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, 356-365.
9. HANSEN, M.H. & KOOPERBERG, C. (2002). Spline adaptation in extended linear models (with comments and a rejoinder by the authors. *Statistical Science*, **17**(1): 2-51.
10. BOLARD, P., QUANTIN, C., ABRAHAMOWICZ, M., ESTEVE, J., GIORGI, R., CHADHA-BOREHAM, H. & FAIVRE, J. (2002). Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *Journal of Cancer Epidemiology and Prevention*, **7**(3): 113-122.
11. THURSTON, S.W., EISEN, E.A., & SCHWARTZ, J. (2002). Smoothing in survival models: an application to workers exposed to metalworking fluids. *Epidemiology*, **13**(6): 685-692.

12. BROWN, E.R., MAWHINNEY, S., JONES, R.H., KAFADAR, K. & YOUNG, B. (2001). Improving the fit of bivariate smoothing splines when estimating longitudinal immunological and virologic markers in HIV patients with individual antiretroviral treatment strategies. *Statistics in medicine*, **20**(16): 2489-2504.
13. WU, W. (2009). An application of spline regression to dose-response analysis in observational study. *Cancer Biostatistics Center, Preston Building Nashville US*.
14. FULLER, W.A. (1969). Grafted polynomials as approximating functions. *Australian Journal of Agricultural Economics*, **13**(1): 35-46.
15. ROYSTON, P. & SAUERBREI, W. (2007). Multivariable modeling with cubic regression splines: a principled approach. *The Stata Journal*, **7**(1): 45-70.
16. WOLD, S. (1974). Spline functions in data analysis. *Technometrics*, **16**(1): 1-11.
17. HUDSON, D.J. (1966). Fitting segmented curves whose join points have to be estimated. *Journal of the American statistical association*, **61**(316): 1097-1129.
18. FORSYTHE, G.E. (1957). Generation and use of orthogonal polynomials for data-fitting with a digital computer. *Journal of the Society for Industrial and Applied Mathematics*, **5**(2): 74-88.
19. CAMERON, E. & PAULING, L. (1978). Supplemental ascorbate in the supportive treatment of cancer: reevaluation of prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Sciences*, **75**(9): 4538-4542.