Comparison of Fractional Splines with Polynomial Splines; An Application on under-five year's Child Mortality Data in Pakistan (1960-2012)

Saira Esar
College of Statistical and Actuarial Sciences
University of the Punjab Lahore, Pakistan
sairaesar88@gmail.com

Shahid Kamal
College of Statistical and Actuarial Sciences
University of the Punjab Lahore, Pakistan
kamal_shahid@yahoo.com

Rehan Ahmad Khan Sherwani
College of Statistical and Actuarial Sciences
University of the Punjab Lahore, Pakistan
rehan.stat@pu.edu.pk

Abstract

Cubic splines are commonly used for capturing the changes in economic analysis. This is because of the fact that traditional regression including polynomial regression fail to capture the underlying changes in the corresponding response variables. Moreover, these variables do not change monotonically, i.e. there are discontinuities in the trend of these variables over a period of time. The objective of this research is to explain the movement of under-five child mortality in Pakistan over the past few decades through a combination of statistical techniques. While cubic splines explain the movement of under-five child mortality to a large extent, we cannot deny the possibility that splines with fractional powers might better explain the underlying movement. Hence, we estimated the value of fractional power by nonlinear regression method and used it to develop the fractional splines. Although, the fractional spline model may have the potential to improve upon the cubic spline model, it does not demonstrate a real improvement in results of this case, but, perhaps, with a different data set.

1. Introduction

Regression is the study of dependency. Regression analysis is based upon the study and analysis of relationships among various variables. Obviously, it is often applied indiscriminately to data with no specific objective as a main priority. The classical linear model is represented by the following equation.

$$y_i = \gamma_0 + \gamma_1 x_{i1} + ...\gamma_p x_{ip} + \varepsilon_i \qquad i = 1, 2, ..., n \tag{1.1}$$

It is useful, however, to recognize two of the fundamental purposes for which linear regression is valuable. The main principle motivation behind regression is to give a summary and reduction of the observed data keeping in mind the end goal to investigate and present the relationship between the configuration variable x and the reaction variable y. The other main purpose of regression is to utilize the model for forecast. While, prediction is no doubt an essential part of regression, it is most likely a much more precise reflection of the statistical practice to consider regression fundamentally a model based system for data outline.

Among various statistical techniques, linear regression is preferred for analysis due to a number of reasons. Some of these include its simplicity, flexibility for the choice of deduction method and easy-to-understand techniques. It has widely been used in literature and has given reliable results. However, the method cannot be used if two or more variable are non-linearly related. The method cannot be applied to analyze varying time series data. To analyze such a data, a plot of data over time must be obtained and then be examined for non-linear trends. There are many approaches for estimating nonlinear trends. One of them most popular technique is piecewise polynomial regression splines.

The term “spline” originates from the tool utilized by the shipbuilders and drafters to manufacture smooth shapes having desired properties. Mathematically, a spline is a piecewise function represented by polynomials. The function has a high degree of smoothness at the nodes of polynomial functions. In computer science, the term spline more frequently points out a piecewise polynomial curve. Spline regression models which are also called piecewise or segmented line regression models or broken stick regression models framed of continual linear stages. Despite the fact that spline regression models may sound like something tricky, they are much same as dummy variable models with a couple of constraints set on them. For example, if a person is gaining weight over time, but suddenly decides to lose weight, then with liposuction, there is an instantaneous drop in his or her weight at that moment the decision is made to lose weight. The person’s weight could serve as dependent variable in a regression while time is an explanatory variable. There will be a gap between the regression line before and after liposuction. Using unrestricted dummy variables, the model after liposuction may have a different slope and intercept rather than the model before it.

Economic analysis regularly involves circumstances where one is required to investigate the effect of unexpected changes in the data**.** The 2010 earthquake in Chile and 2011 earthquake in Japan are two of the most important examples. In both the cases, an intermediate, persistent effect was observed for the product availability. Spline regression models searches for points in the data that would identify where these changes happen. These points are named as “knots”. Spline regression models give a method for capturing these changes smoothly and joining the segments without the usual break between the segments. Thus, in a spline model, a turning point in the product availability could be represented by a spline knot, which may join the upward regression line to the downward regression line. This type of spline model is often called as a piecewise regression model (Pindyck *et al*. 1998).

An undeniable inquiry is the reason not to utilize a polynomial regression model rather than splines? Interestingly, spline regression models or piecewise polynomials have considerably additional flexibility than polynomial regression models in low measurements and are mostly less liable to generate perfect multicollinearity in low extent. Other methods such as kernel regression can be used but Carroll (2000) reported that spline methods are generally more efficient than kernel methods.

The credentials of variations in the recent development are vital concerns in the analysis of a data related to mortality and incidence of a disease. Recently, Kim et al (2000) applied segmented line regression to illustrate the continuous changes in cancer mortality and incidence rates. We apply a join point regression model to describe such continuous changes of infant mortality rates data in Pakistan. Tracking the infant mortality rate over time is a good variable for using splines. The set of real numbers is much larger than the set of positive integers. One can allow coefficients and exponential power that is the fractional value (e.g 0.5 or whatever) for each spline knot.

In present research we find the number and location of spline knots where the regression line pivots by considering a new slope. In this regard, we initially fit the cubic spline and estimate the knots and then used these estimated knots in fractional splines (i.e. only degree of the polynomial is changed by some fractional number) to determine the better fit of the model. In last, we compare the fractional spline results with the polynomial splines.

Searching for the location of spline knots by creating a potential adjustment variable for every possible time period (54 in this case) allows one to statistically search for significant adjustment points by using an appropriate technique to locate one or more such points. The first step is to generate the spline adjustment variables. One can use the "+" functions or dummy variables to set up spline adjustment variables. The dummy variable, D, is needed to create a first derivative break at the point where X = K. This creates a kink in the line at X = K. We are not trying to change the overall slope of the line throughout the entire length of the line, but instead create a kink in the line at X = K. The dummy variable plays a critical role in accomplishing this.

For example, impact of year 1960 will be felt in the subsequent years and is being captured by the variable L1. For Year ≤ 1960 L1=0 and for Year > 1960 L1= Year-1960. Similarly the impact of each Year will be captured in the subsequent years. Since we are not sure of the degree of impact of each of the years on the outcome for the subsequent years, we have also taken Quadratic variables (Q1, Q2, Q3, … Q54) and also Cubic variables (C1, C2, C3 … C54).

$Q1 = (L1)^2$	$C1 = (L1)^3$
$Q2 = (L2)^2$	$C2 = (L2)^3$
$Q3= (L3)^2$	$C3= (L3)^3$
.	.
.	.
.	.
$Q54=(L54)^2$	$C54=(L54)^3$

We create linear, quadratic and cubic splines adjustment and also use year square and year cube for analysis. As we have so few observations it makes sense to restrict the model to fit more smoothly between points. Therefore, it makes sense to run a cubic spline model, which only allows third derivative adjustments. The problem here becomes even more complicated in that we do not even know the number of spline knots. Our general strategy for dealing with this problem is to create a large number of potential spline knots and then use stepwise regression to pick out those that are most statistically significant. Of course, we have to specify the significance level that determines when to stop adding additional spline knots to the stepwise regression.

To convert cubic spline results from the stepwise regression into fractional splines requires treating the exponent which is number "3" in cubic splines into an unknown parameter to be estimated by the nonlinear regression. For estimating fractional spline regressions, one can use any nonlinear regression estimation technique such as Gauss Newton or Newton-Raphson.

This study is conducted on the data of under-five mortality in Pakistan from the period 1960 to 2012. Data for this study has been obtained from secondary source, World Bank. Here is the link for data download: http://data.worldbank.org/country/pakistan.

2. Results and Discussion

The first objective of the study is to search for the number and location of spline knots (join points). For our study, we are using the pattern recognition approach. Pattern recognition creates explanatory variables that just fit the pattern of data as a function of time. The research problem here becomes more complicated in that we do not even know the number and also the location of spline knots. We deal with this problem by creating a large number of spline knots and then we use stepwise regression to pick out those which are most statistically significant.

We create a set of dummy variables by using R statistical package. The following mathematical explanation to understand the dummy variables is given as under.

$$\{D1=0\,if\;X\le knot1\,and\;D1=1\,if\;X>knot1\}$$
$$\{D2=0\,if\;X\le knot2\,and\;D2=1\,if\;X>knot\,2\}$$
$$\{D3=0\,if\;X\le knot3\,and\;D3=1\,if\;X>knot\,3\}$$
$$\{D4=0\,if\;X\le knot4\,and\;D4=1\,if\;X>knot\,4\}$$
$$\{D5=0\,if\;X\le knot5\,and\;D5=1\,if\;X>knot\,5\}$$
$$\vdots$$
$$\{D54=0\,if\;X\le knot54\,and\;D54=1\,if\;X>knot\,54\}$$

We define linear, quadratic and cubic spline adjustments as follows.

$$\{L1=D1*(X-knot1)\}\qquad\{Q1=D1*(X-knot1)**2\}\qquad\{C1=D1*(X-knot1)**3\}$$
$$\{L2=D2*(X-knot2)\}\qquad\{Q2=D2*(X-knot2)**2\}\qquad\{C2=D2*(X-knot2)**3\}$$
$$\vdots\qquad\qquad\vdots\qquad\qquad\vdots$$
$$\{L54=D54*(X-knot54)\}\qquad\{Q54=D54*(X-knot54)**2\}\qquad\{C54=D54*(X-knot54)**3\}$$

It makes sense to run a cubic spline model, which only allows third derivative adjustments. Therefore, we have considered year, year squared, year cubed and cubic spline adjustment variables. Given the large number of variables (57 variables), we have done stepwise regression to select only significant variables that have impact on the under-five deaths. The stepwise regression process starts with an empty model (Null Hypothesis that under-five mortality is not dependent on any of the variables). From this step, it will include any independent variable that has a significance level of below 0.50 (entry criteria). Once a variable is inside the model, the variable will stay inside as long as the significance level is below 0.15 (stay criteria). So at the end of the Stepwise Regression, we have a set of variables (including the knots) that have dependence on under-five child mortality. The stepwise procedure comes to an end at step 23. After this stepwise regression analysis with cubic spline adjustments, now we need to plot a couple of them to get a sense of how realistic they are. We Plot the fitted equation against the observed sample points using the results from different steps of stepwise output.

By comparing the models of different steps of stepwise regression procedure, we have come to the conclusion that model of step 10 looks adequate. The output for step 10 is shown in table 1 and 2. The F-value is greater than as in step 4 and R squared is now over 0.99 and all the regression coefficients are highly significant. There are now 7 spline adjustment variables in the model. Notice that the coefficients alternate in sign. This means that both upward and downward movements have a tendency to go too far. This tendency can complicate the problem of forecasting with spline models beyond the scope of the original data. Our predicted regression curve is represented among the actual values of under-five child deaths in the graph in Figure 3. The predicted values are approximately near the actual values. This graph shows considerably more sensitivity to the data and therefore greater flexibility. A higher step might be needed if we are fitting a 3 dimensional spline to search for caves in a mountain terrain but for simple policy purposes, there is no need to go beyond step 10. Determining the proper fit for each problem depends very much on the purpose of the analysis. Step 10 shows a much flexible model and in comparison with step 23, it only uses 7 spline adjustment variables.

Figure 1. Graph of Actual/Predicted Under-five Deaths at Step 10

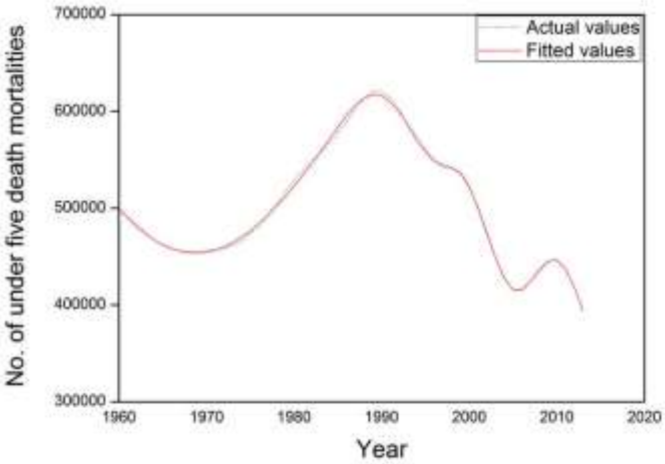

Table 1: Stepwise Regression Output at Step 10

Analysis of Variance					
Source	**df**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr>F**
Model	9	2.05866E11	22874093099	3065.531	< 0.0001
Error	44	328339113	7462253		
Corrected Total	53	2.061952E11			

RMSE = 2731.7129; R2=0.9984; Adj. R2=0.9981; Dep-Mean=502551; Coeff. Var=0.54357

Table 2: Parameter estimates of cubic spline regression model Parameter Estimates

Variable	DF	Parameter	Standard	t Value	Pr > |t|
Intercept	1	1544195401	47754294	32.34	<.0001
Year	1	-1176166	36361	-32.35	<.0001
Year3	1	0.10115	0.00312	32.39	<.0001
C13	1	-12.54399	1.55019	-8.09	<.0001
C25	1	-158.90631	8.37328	-18.98	<.0001
C31	1	456.13705	19.07716	23.91	<.0001
C38	1	-1809.90784	81.38916	-22.24	<.0001
C40	1	2236.84067	95.45595	23.43	<.0001
C46	1	-1583.32887	71.74118	-22.07	<.0001
C51	1	1468.42910	353.15434	4.16	0.0001

The second main objective of our research is to fit a fractional spline regression model. To convert cubic splines into fractional splines, we consider that the number of knots and their location is already known. We have estimated the knots by considering it a nonlinear regression problem from stepwise regression procedure in the first objective of our research.

The problem here is to deal with the estimation of degree of the polynomial which is “3” in cubic spline regression model. We estimate this by using Newton Raphson method which have implemented using SAS Programming language. We use year 1972, 1984, 1990, 1997, 1999, 2005 and 2010 as the knot years already known which we have estimated from cubic spline stepwise regression procedure. Thus, we created a SAS code for this.

In the code, the first nonlinear least squares regression starts out with the initial value a=3 representing cubic splines, but allows the nonlinear estimation to search for a fractional power that replaces the "3" with some fractional number that provides a better nonlinear fit to the data. Thus, fractional splines are estimated by turning the exponents into parameters to be estimated as part of the regression equation.

In nonlinear regression, Taylor Series Expansion has an important role to play. A typical nonlinear regression estimation method is to expand the nonlinear regression equation in a Taylor series around initial guesses for the value of the parameters. Then, right after the linear term chop off the rest of the Taylor series, and then run ordinary least squares estimation on the linear part to estimate new estimates of the parameter. Next expand the Taylor series around the new parameter estimates and chop off the higher order terms and run ordinary least squares again on the linear part. Repeat this until convergence. That's how Newton Raphson method works by Taylor series expansion.

Table 3: SAS output of NLIN Procedure using Newton Raphson Method

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	ApproxPr > F
Model	9	2.053E11	2.281E10	1100.33	<.0001
Error	44	9.121E8	20729500		
Corrected Total	53	2.062E11			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits Lowe	Upper
b0	1.5666E9	75840423	1.4138E9	1.7195E9
byear	-1191739	56490.7	-1305588	-1077890
byear3	0.0964	.-	.	.
b13	-14.0343	2.3283	-18.7267	-9.3419
b25	-134.7	15.4475	-165.8	-103.5
b31	366.1	32.4040	300.8	431.4
b38	-966.0	105.9	-1179.4	-752.5
b40	839.0	109.6	618.1	1059.9
b46	13828.8	1778.9	10243.6	17414.0
b51	-10069.0	2763.2	-15637.9	-4500.1
a	3.0077	0.00598	2.9957	3.0198

In above Table 3, first table displays the analysis of variance table for the model. The table displays the degrees of freedom, sum of squares and mean squares along with the model F-test. Second section of Table 3 displays the estimates for each parameter, the associated asymptotic standard error and the upper and lower values for the asymptotic 95% confidence interval. Thus, the estimated fractional degree "a" is 3.007. Note that these values are linear approximations (based on normality assumption). PROC NLIN also gives the asymptotic correlations between the estimated parameters. Since nonlinear regression may result in dependency among the parameters. Very high correlations (>> .99) may indicate that the model form is inappropriate for the data. In our model, there are correlations between "a" & b0, byear, b13 and b25 that are high.

Table 4 displays another analysis of variance table. After nonlinear regression analysis by Newton Raphson method, we fit the model with the degree of polynomial “3.007”. There is not much difference between results of cubic spline model and the model with this fractional degree estimated by NLIN. Moreover, the values of R square and adjusted R square are same as 0.998 in both models. The values of parameter estimates of fractional spline model are just slightly different from the cubic spline model as shown in the figures below. An overlay graph is also shown above these Tables in Figure 2 which depicts the trend of actual values and values fitted by fractional spline model. This graph is very much similar to as one made by cubic spline model (step10) in Figure 2.

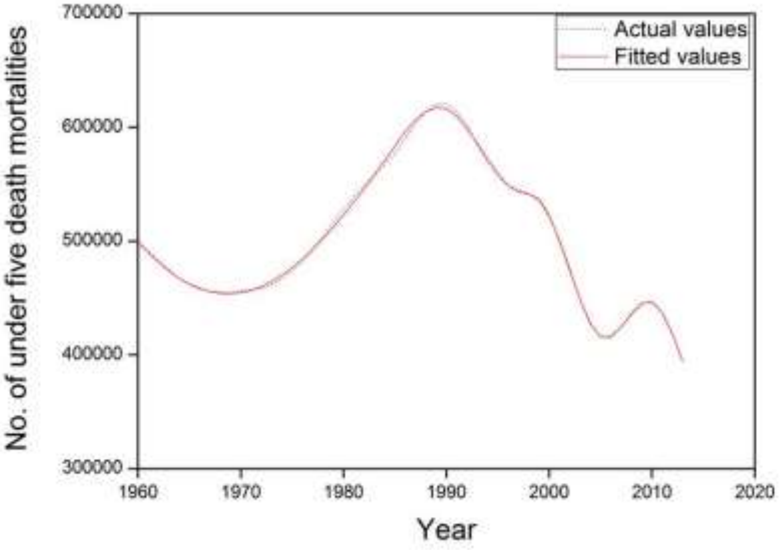

Figure 2: Graph of Actual/ Predicted values of Fractional Spline Model

Table 4(A): Fractional Spline Regression Model Output

Analysis of Variance

Source	**df**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	9	2.05871E11	22874556999	3104.85	<.0001
Error	44	324164016	7367364		
Corrected Total	53	2.061952E11			

RMSE = 2714.3; R^2=0.9984; Adj. R^2=0.9981; Dep-Mean=502551; Coeff. Var=0.54010

Table 4(B): Parameter estimates of fraction spline regression model

Variable	DF	ParameterEstimate	Standard Error	t Value	Pr > |t|
Intercept	1	1541250570	47183975	32.66	<.0001
Year	1	-1172422	35881	-32.68	<.0001
year3	1	0.09486	0.00290	32.72	<.0001
C13	1	-12.36844	1.50001	-8.25	<.0001
C25	1	-154.76042	8.15907	-18.97	<.0001
C31	1	447.40371	18.64046	24.00	<.0001
C38	1	-1784.43884	79.71715	-22.38	<.0001
C40	1	2206.84975	93.55335	23.59	<.0001
C46	1	-1571.42689	70.47752	-22.30	<.0001
C51	1	1505.66049	348.74501	4.32	<.0001

We have also tried to fit some other fractional spline models with different degrees as from 3.1 to 3.9. In result, as we increase the degree of model by 0.1, F value decreases and it becomes 492.19 for the fractional degree 3.9. But, there is not much difference appeared in R Square and adjusted R square. It is approximately the same for all these different degrees.

3. Conclusion

Selection a good proper model, this is one of the most difficult and contentious issues in statistics. By comparing the fractional and cubic spline regression models, we conclude that there is not much difference in the results of both models. Consequently, the fractional spline approach offers a much greater range of flexibility than the traditional polynomial splines. Fractional splines work in the same context that polynomial splines do, but with much greater flexibility and better fit. Consequently, fractional splines can produce a smaller error sum of squares and a higher R-squared than polynomial splines.

In general, spline models work well for interpolation within the range of the observed data. Neither polynomial nor fractional splines are useful for forecasting outside of that range without first transforming the underlying data series. The example provided in this research hopefully has given the reader an understanding of how spline models are more appropriate than traditional linear and polynomial regression models and how these models are set up and estimated and how to choose a proper model among various alternative spline models. We have tried to fit a fractional spline model also. In this case, fractional splines are not depicting a very good result as compared to cubic splines. But, they may work better with a larger data set such as stock market data. As we have worked with only 54 observations and there is an oscillation in our mortality data, therefore it is possible that fractional spline may work more accurately with big data and with data having cycles. Further investigation to reveal the circumstances where fractional splines significantly outperform than traditional splines will require a different model, and, perhaps, a different data set.

References

1. Buse, A., & Lim, L. (1977). Cubic Splines as a Special Case of Restricted Least Squares. *Journal of the American Statistical Association, 72*, 64-68.

2. Cahill, N., Rahmstorf, S., & Parnell, A. (2015). Change points of global temperature. *Environ. Res. Lett. Environmental Research Letters, 10(8),* 084002-084002

3. Khan, A., Kinney, M. V., Hazir, T., Hafeez, A., Wall, S. N., Ali, N., ... & Pakistan Newborn Change and Future Analysis Group. (2012). Newborn survival in Pakistan: a decade of change and future implications. *Health policy and planning*, *27*(suppl 3), iii72-iii87.

4. Kim, H. J., Fay, M. P., Yu, B., Barrett, M. J., & Feuer, E. J. (2004). Comparability of segmented line regression models. *Biometrics*, *60*(4), 1005-1014.

5. Lawn, J. E., Blencowe, H., Darmstadt, G. L., & Bhutta, Z. A. (2013). Beyond newborn survival: the world you are born into determines your risk of disability-free survival. *Pediatric research*, *74*(S1), 1-3.

6. Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., ... & Child Health Epidemiology Reference Group of WHO and UNICEF. (2012). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *The Lancet*, *379*(9832), 2151-2161.

7. Marsh, L. C. (1983). On estimating spline regressions. Proceedings of SAS Users Group International, 8, 723-728.

8. Marsh, L. C. (2011). Estimating the number and location of knots in spline regressions. *Journal of Applied Business Research (JABR),* 2(3), 60-70.

9. Marsh, L. C., Maudgal, M., & Raman, J. (1990). Alternative methods of estimating piecewise linear and higher order regression models using SAS software. In *Proceedings of SAS User's Group International* (Vol. 15, pp. 523-527).

10. Marsh, L., & Cormier, D. (2001). *Spline regression models*. Thousand Oaks, Calif.: Sage Publications.