Middle-East Journal of Scientific Research 17 (4): 495-499, 2013
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2013.17.04.12048

A Comparison of Classical Linear Regression Approach with Spline procedure

Saima Mustafa and Sadia Rasheed

Department of Mathematics and Statistics,
PMAS University of Arid Agriculture, Rawalpindi, Pakistan

Abstract: Regression model is one of the most important and widely-used models in statistics which has proved its significance and application in every field of science. In this paper, we have used linear regression model and examined the relationship between continuous dependent (e.g., BMI) and independent (e.g., age) variables which may be separated into logical categories (e.g., age categories. Apart from this, we have used spline regression model which has provided a better fit, taking into consideration the variation in the relationship between the predictor variable and the response variable. Spline is very constructive function-type used in regression when the relationship between a response and a set of covariates is not known in advance. The analyses presented in this paper focuses on univariate regression splines. These functions provided a helpful and flexible basis for modeling relationships with continuous predictors. Comparisons of both techniques will be done by using real life data that will be collected from different fields.

Key words: Ordinary Least Square Regression · Splines · Univariate Splines · Natural spline

INTRODUCTION

Regression is an extensively explored branch in statistics and scientific areas. A meaningful relationship exists between response and predictor variable in the study of regression. It is generally used in intuitive level every day as well as for prediction and forecasting. Such as in medical a new medicine (dependent variable) was predicted on the base of body weight (independent variable), as for as in businesses it uses for prediction current exchange rates, future sales etc. Through the least square technique an appropriate model selecting and appropriate fitting is possible in which we forecast the one variable values on the basis of other. In this technique the model is best if the error sum of squares is least possible.

Numerous techniques are carrying out for regression analysis. Familiar used approaches are linear classical regression in parameters it is linear. It have unknown parameters in a finite number which are predictable from the data figures. Transformations of the response variable can improve the fit and may correct violations of model assumptions such as constant error variance. Greenland (1995) [1] suggests using spline regression (and fractional polynomial regression) as an alternative to categorical analysis for dose response and trend analysis. The term of splines is introduced by Pierre Bezier which is used for interpolating purposes in which a draftsman would draw a smooth curvature through a set of points on graph paper by imposing strip to pass over the points and discovered piecewise polynomials or splines could be used in place of polynomials occurred in the early twentieth century.

There are many types of splines and estimation procedures [2, 3]. The analyses presented in this paper focus on univariate splines in ordinary least squares regression. Knot selection (number and location of knots) can be accomplished by various methods. One can 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]. Predetermined knots are used in this paper.

The techniques of spline linear regression and linear piecewise regression are commonly used. Any degree of polynomial could be in use, but the cubic is convenient for most purposes, most progressive commonly use natural cubic splines.

Corresponding Author: Saima Mustafa , Department of Mathematics and Statistics,
PMAS University of Arid Agriculture, Rawalpindi, Pakistan.

MATERIALS AND METHODS

The model of SLR is an analytical technique which will use to describe the relation among explained and explanatory variable. The line of Simple linear regression is a straight line fitted to the data through the method of least squares [4]. We assume n sample data points and for this here we have two variables BMI and Age. The hypothesized relationship between BMI (body mass index) and Age may be written as:

$$b = \gamma_0 + \gamma_1 g + \varphi_j \quad (2.1)$$

In equ. (1) b and g symbolize as BMI and Age respectively, γ_0 represent a constant term, γ_1 is the coefficient of the variable Age and φ_j is the noise term reflecting other causes that effect BMI. Since BMI is the ratio of Weight in kg and Height in meter square [5]. ($BMI = \frac{Weight\ in\ Kg}{(Height\ in\ Meters)^2}$). The fitted line (1), which we calculated using the sample data points, is presented as:

$$\hat{b} = \hat{\gamma}_0 + \hat{\gamma}_1 g$$

The “hats” can be read as “estimator of” and the derivation of the jth value of b from its predicted value is:

$$b_j - \hat{b}_j = b_j - (\hat{\gamma}_0 + \hat{\gamma}_1 g_j);$$

$$i.e., \ \varphi_j = b_j - \hat{b}_j$$

The unsystematic error $\varphi_j = b_j - \hat{b}_j$ is there to present the change between the dependent variable predicted values by the model, $\hat{b}_j$ and the true value of the dependent variable b_j.

The model for linear spline regression is in part a special case of the piecewise regression model where we have only one independent variable. The main difference in the linear spline model and the piecewise linear model is that, in the former, the adjacent regression lines are required to intersect at the knots or change points [6]. Let we have sample size n at that point for the *ith* sample point we assume b_j as the response variable and g_j as explanatory variable. Then we have a model $b_j = m_j + \varphi_j$ where,

$$m_j = b_0^* + \left\{ {b_1^* - b_0^*} \Big/ {g_1^* - g_0^*} \right\} \left(g_j - g_0^* \right), \qquad g_0^* \le g_j < g_1^*, \qquad j = 1, n_1^*$$

$$m_j = b_1^* + \left\{ {b_2^* - b_1^*} \Big/ {g_2^* - g_1^*} \right\} \left(g_j - g_1^* \right), \qquad g_1^* \le g_j < g_2^*, \qquad j = n_1^* + 1, n_1^* + n_2^*$$

$$m_j = b_{k-1}^* + \left\{ {b_k^* - b_{k-1}^*} \Big/ {g_k^* - g_{k-1}^*} \right\} \left(g_j - g_{k-1}^* \right), \qquad g_{k-1}^* \le g_j < g_k^*, \qquad j = n_1^* + \cdots + n_{k-1}^* + 1, n$$

In equ. (2.1) φ_j have mean zero as well as constant S.D. Here we have k straight line interconnecting segments.so the *ith* line (i = 1... k) is connected with n_j^* sample points. These segments are defined through knots $(g_0^*, b_0^*; g_j^*, b_j^*; ... ^*; g_k^*, b_k^*)$.

In this study, our main concerned is with splines models which are linear, quadratic and cubic. These models have been analyzed by using one of the spline techniques which is called natural spline. Natural splines confirm the typical interpolating restrictions. A spline procedure specified degree, name of variable, list of abscissa and ordinate etc [7]. This type of splines used a list of polynomial with valid interval of each polynomial. Natural splines are another type of flexible polynomial-based function that starts with a cubic spline and then imposes the constraint that the function for the mean is to be linear (rather than cubic) beyond some boundary points usual the min and max of independent variable [8].

Let we have (m+1) points and g_0, g_1,...,g_m are the knots which are equally spaced, we use g and b variables which are equal to Age and BMI respectively, we wish to construct a piecewise cubic polynomials [9].

$$S(g)=\begin{cases} s_1(g) & if \;\; g_1 \le g < g_2 \\ s_2(g) & if \;\; g_2 \le g < g_3 \\ \vdots & \\ s_{m\text{-}1}(g) & if \;\; g_{m\text{-}1} \le g < g_m \end{cases}$$

Then in general, a function s is called a spline of degree k. Let us assume a cubic spline $S(g)$ is a piecewise-defined function:

$$s_j(\mathrm{g}) = A_j(\mathrm{g}-\mathrm{g}_j)^3 + B_j(\mathrm{g}-\mathrm{g}_j)^2 + C_j(\mathrm{g}-\mathrm{g}_j) + \mathrm{D}_j, \quad j = 1,2,\cdots\cdots m-1 \tag{2.3}$$

Data Collection: The cross sectional data of 250 adult (aged 15 years or above) people, both males and females were taken from Rawalpindi district. The data on dependent and explanatory variables were collected from different secondary sources to originate the models. The sample was taken by convenient sampling, from Arid Agriculture University, different clinics and different hospitals, etc [10]. Data analysis have been done in the software in SPSS and R.

RESULTS

In current section, we have presented the statistical approach in which we apply the OLS regression with diagnostic tests for assumptions of linear models, then we apply spline techniques [11].

Linear models are based on some assumptions (normality, linearity and homoscedasticity etc.) and when these assumptions are fulfilled then it became BLUE (Best Linear Unbiased Estimator). Most common assumptions for OLS regression analysis have been checked. For this purpose several tests have been performed. The normality of our data is checked by using histograms for response variable (BMI). As shown in Fig. 1.1, standardized residuals of dependent Variable (BMI) are normally distributed. Normal distributions of error term can be conformed from the Q-Q Plot of residuals, shown in the Figure 1.2.

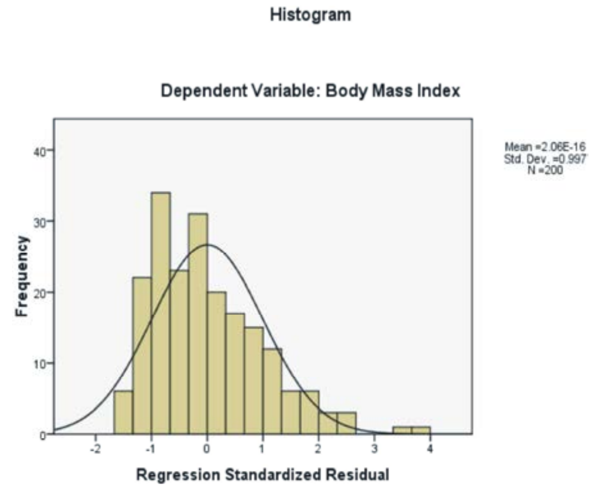

Fig. 1.1: Histogram of BMI variable

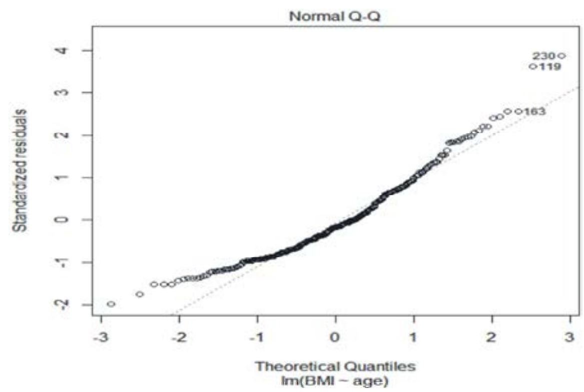

Fig. 1.2: Q-Q Plot for Normality of error term

Results of OLS Regression: The results of the OLS regression have been summarized in Table 1.1 for the BMI with one independent variable. Here R-square is 0.65610 which interprets that only 65% of the variation is described by the independent variable. It tells that there is a strong relationship between numbers of age and BMI. In each point of linear model gives an estimate. The value of γ_1 (=0.08452) which is a slope of linear model shows as age increases on the average as BMI increases about 0.08452. The value of γ_0 (=20.5958) is an intercept of our linear model show the average level of BMI. The p-value of the variable (is less than 0.001 alpha levels with positive coefficient of age which indicate age is positively related to BMI and analysis of variance is also done Table 1.2 display the significant results at alpha level 0.001.

Main Results by Spline Techniques: In this section results based on spline regression model. First its general model is given then after numerical results. We have use spline techniques with different degree and knots selection and then comparison is made to check which model is best.

Table 1.1: OLS Regression Summary

	Coeff.Value	t value	Std. Error	Pr(>|t|)
Intercept	20.59580	27.739	0.74249	< 2e-16 ***
Age	0.08452	4.173	0.02026	4.16e-05 ***

Residual std. errs. 4.301 on 248 DF

Multiple R-sq. 0.6561, Adjusted R-sq. 0.6184

p-val. 4.163e-05, F-statistic: 17.41 on 1 and 248 DF

Note: **=p<0.01, *=p<0.05, ***= p<0.001.

Table 1.2: ANOVA table of regression model

	Df	Sum Sq.	Mean Sq.	F value	Pr(>F)
Regression	1	322.1	322.15	17.413	4.163e-05 ***
Residuals	248	4588.0	18.50	---	---
Total	249	4910.1	---	---	---

Table 3.1: Natural splines results with degree 3 and knots (8 inner and 2 boundary knots)

	Estimate	Std. Error	t value	Pr(>|t|)
γ_0	40.871	8.272	4.941	1.46e-06 ***
γ_1	-18.759	6.562	-2.859	0.00463 **
γ_2	-15.509	8.539	-1.816	0.07060.
γ_3	-17.221	8.345	-2.064	0.04013 *
γ_4	-16.094	8.567	-1.879	0.06151.
γ_5	-14.539	8.398	-1.731	0.08468.
γ_6	-13.700	8.699	-1.575	0.11661
γ_7	-18.391	8.154	-2.256	0.02500 *
γ_8	-9.884	5.654	-1.748	0.08169.
γ_9	-40.675	18.635	-2.183	0.03002 *
γ_{10}	NA	NA	NA	NA

Note: ***= p<0.001, **=p<0.01, *=p<0.05, ='.'p<0.1

Residual std. errs. 4.086 on 240 df

Multiple R-sq. 0.1839, Adj. R-sq. 0.1533

P-val. 1.362e-07, F-stats. 6.008 on 9 and 240 DF

The general form of the fitted spline model:

$$b = \gamma_0 + \gamma_1(g - g_1) + \gamma_2(g - g_2) + \cdots\cdots + \gamma_m(g - g_m) + \varphi \quad (3.1)$$

where b is the BMI and γ_0, γ_1,.........,γ_m are the coefficients, g_1, g_2,........,g_m are the so-called knots (A knot is the internal breakpoints that define the spline), g_1 is the age at which the first growth period starts; therefore, equals zero. The actual model that was fitted to the data was as follows:

$$b = \gamma_0 + \gamma_1 g + \gamma_2(g-15) + \gamma_3(g-20) + \gamma_4(g-25) + \gamma_5(g-30) + \gamma_6(g-35) + \gamma_7(g-40) + \gamma_8(g-45) + \gamma_9(g-50) + \gamma_{10}(g-55) + \gamma_{11}(g-60) + \varphi \quad (3.2)$$

Now the coefficients of the model (3.2) are estimated by using natural splines technique in r package.

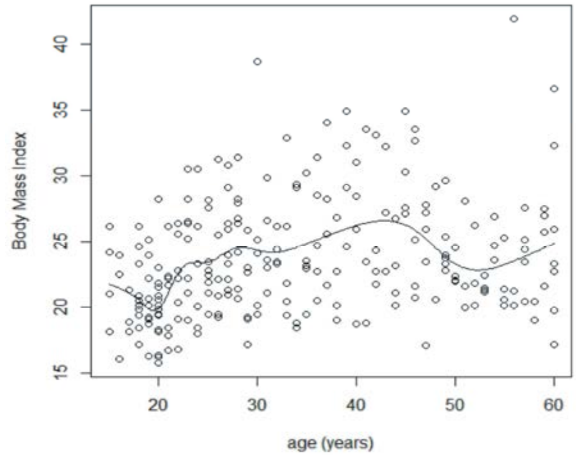

Fig. 3.1: Graph of Cubic Spline (with 10 inner and 2 outer knots)

In Table 3.1 we have results of natural cubic splines with different knots between 15-60 with 5 point difference, we have 10 inner knots and two outer knots g<15, 15 = g<20, 20 = g<25, 25 = g<30, 30 = g<35, 40 = g<45, 50 = g<55, at the first knot results are highly significant at alpha level p=0.001 but at the last knot it does not define because of singularities. The degree of freedom and degree is equal to three and Figure 3.1 shows the visual display of spline at different knots. The Fig. 3.1 displays the overall graphical illustration of the spline regression model. It can be seen from the above figure the model fits the data very well and the line nicely approximate the data. By applying natural spline technique, we have estimated the unknown parameters in eq (3.2) which are summarized in the following table.

CONCLUSION

This research was mainly concerned with the estimation of parameters by linear regression and natural spline technique. Comparisons of both techniques are also obtained firstly we have applied linear regression and obtained the unknown parameters. The constant behavior of linear regression model was essentially found correlated with the estimated values of the parameter. It has been observed, that parameter variability in linear regression model is based on an analysis of residuals. Secondly, this study was focused on spline technique which is based the spline regression methodology. For this purpose, cubic spline was used for analyzing data, especially natural splines which provides piecewise regression functions. After comparison of both techniques, spline regression gave more reliable and efficient results. Using the spline technique, the estimated values of the parameters gave the clear picture of the model and minimized residuals as compared to simple regression model.

REFERENCES

1. Greenland, S., 1995. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology, 6(4): 356-365.
2. Gu, Chong, 2002. Smoothing Spline ANOVA Models. New York: Springer-Verlag New York, Inc, pp: 2-3, 30-52, 111-142.
3. Eubank, R.L., 1999. Nonparametric Regression and Spline Smoothing, 2nd Ed. New York: Marcel, Dekker, Inc, pp: 1-23, 27-37, 119-144, 291-307.
4. Ahlberg, J.H., E.N. Nilson and J.L.Walsh, 1967. The Theory of Splines and Their Applications. New York: Academic Press, pp: 1-74, 109-152.
5. De Boor, C., 2001. A Practical Guide to Splines. New York: Springer-Verlag New York, Inc, pp: 17-37, 69-76, 79-86, 207-224.
6. Schumaker, Larry L., 1981. Spline Functions Basic Theory. New York: John Wiley and Sons, Inc, pp: 1-11, 108-134,309-316.
7. Smith, P.L., 1979. Splines as a useful and convenient statistical tool. J. Amer. Stat. Assoc., 2(33): 57-62.
8. Wegman, E.J. and L.W. Wright, 1983. Splines in statistics. J. Amer. Stat. Assoc., 382(78): 351-365.
9. Muhammad Azam, Sallahuddin Hassan and Khairuzzaman, 2013. Corruption, Workers Remittances, Fdi and Economic Growth in Five South and South East Asian Countries: A Panel Data Approach Middle-East Journal of Scientific Research, 15(2): 184-190.
10. Sibghatullah Nasir, 2013. Microfinance in India: Contemporary Issues and Challenges, Middle-East Journal of Scientific Research, 15(2): 191-199.
11. Mueen Uddin, Asadullah Shah, Raed Alsaqour and Jamshed Memon, 2013. Measuring Efficiency of Tier Level Data Centers to Implement Green Energy Efficient Data Centers, Middle-East Journal of Scientific Research, 15(2): 200-207.