China-USA Business Review, ISSN 1537-1514
January 2012, Vol. 11, No. 1, 31-37

A Note on the Predicted Responses in Cubic Splines*

Bahar Berberoglu
Anadolu University, Eskisehir, Turkey

One of the most used statistical methods in economic and business studies involving time series is related to predicted responses ($\tilde{y}$ values), which can be estimated with two different approaches, namely cubic spline regression method (CSR) and prediction sum of squares statistic (PRESS). This study aims to set and discuss the relation between these two approaches in estimation of predicted responses. In first approach estimated $\tilde{y}$ values are determined from the derived restricted model. According to the second approach, they are estimated with prediction sum of squares statistic (PRESS), and it argues that the use of this technique performs is better for cubic spline regression method (CSR). This study while introducing and discussing the relation between these approaches, also addresses to note the estimation of predicted $\tilde{y}$ values theoretically. The study concludes that same predicted responses can be received by employing both methods. For examining and testing this argument empirically real exchange rates data for Turkey in the period of 1987-2008 are used. Additionally another subject searched and discussed in the study was, how economic crises can be defined with spline methods. Because of the advantages provided by them in reaching minimum residual sum of squares, and achieving the result by using real economic data without changing their nature in time series analysis.

Keywords: cubic spline regression, predicted responses ($\tilde{y}$), prediction sum of squares statistic, real exchange rate

Introduction

Economists and econometricians are used to employ various statistical methods in their works related to time series analysis. One of them is estimating predicted responses ($\tilde{y}$ values), which is a quite known topic in statistics and widely used by economists and especially by econometricians. But these researchers need to transform or smooth the economic time series they use, because of their piece-wise nature. Time series containing structural breaks can be examined with real data (without any transforming or smoothing) by employing cubic spline regression. With spline regression which is a preferred method in interpolation, one can easily reach minimum residual sum of squares and achieve the result by using the real economic data without changing their nature. If an economist or an econometrician cares and takes this advantage into account, he/she will model the breaks in time series with splines and observe them better. Additionally, taking benefit from PRESS statistics for such researchers can also be advised.

Cubic splines are cubic polynomials in a single variable, which are joined together smoothly at known

* This study is the revised form of the paper presented in EconAnadolu 2, June 15-17, 2011, Eskisehir/Turkey.
Bahar Berberoglu, Ph.D., Open Education Faculty, Anadolu University.
Correspondence concerning this article should be addressed to Bahar Berberoglu, Open Education Faculty, Anadolu University, Yunusemre Kampusu, 26470, Eskisehir, Turkey. E-mail: bdirem@anadolu.edu.tr.

points, called "knot" points. The smoothness restrictions are such that, at the points where the cubic polynomials meet, their first and second order derivatives are also equal (Nyquist, 1991).

Buse and Lim (1977) followed up Poirier's paper (1973) and defined cubic splines as a special case of (RLS), and they proved the relation between RLS and cubic spline regression (CSR) methods mathematically, and Tarpey (2000a, 2000b) emphasized that using PRESS residuals in cubic spline models will bring better performance, and showed their mathematical proofs under both $R\beta = 0$ and $R\beta \neq 0$ restrictions.

Tarpey (2000a) argued that prediction sum of squares statistic (PRESS) shows better performance for restricted least squares method (RLS). Furthermore in another study Tarpey (2000b) pointed out some ways for calculating PRESS residuals for (RLS) in the existence of linear restrictions on regression parameters. In the same work, Tarpey (2000b) also denoted how predicted $\tilde{y}$ values over PRESS residuals could be estimated. In this study while introducing and discussing the relation between these works, it is also aimed to note the estimation of predicted $\tilde{y}$ values theoretically. For examining and testing this argument empirically, real exchange rates data for Turkey in the period of 1987-2008 is used.

In the present study before all else basic properties of the subject are specified, then the theorem which shows the equality of $\tilde{y}$ values derived with two cited approaches and afterwards this theorem is demonstrated on a case related to Turkish economy. The proof in the study is especially under the restriction $R\beta = 0$. And here, the mathematical proofs of how similar expected $\tilde{y}$ values were obtained with two different matrices solution in two separate studies namely Buse and Lim (1977) and Tarpey (2000a, 2000b) is also presented. At last in the conclusion, the models defining the structural breaks in economic data with cubic splines was produced by using two different approaches, and the argument that the same predicted responses can be received by employing both is highlighted.

Basic Properties

This study is concerned with the linear model:

$$y = X\beta + \varepsilon \tag{1}$$

which summarizes the dependence of the response y on the carriers $X_1, X_2, \ldots, X_p$ in terms of the data values y_i and $x_{i1}, \ldots, x_{ip}$ for $i = 1, \ldots, n$. In fitting the model (1) by least squares (assuming that X has rank p and that $E(\varepsilon) = 0$ and Var $(\varepsilon) = \sigma^2 I_n$), usually the fitted or predicted values are obtained from $\hat{y} = X\hat{\beta}$, where $\hat{\beta} = (X'X)^{-1}X'y$. From this it is simple to see that:

$$\hat{y} = X(X'X)^{-1}X'y \tag{2}$$

To emphasize the fact that (when X is fixed) each $\hat{y}_i$ is a linear function of the y_j, and it is possible to write equation (2) as:

$$\hat{y} = Hy \tag{3}$$

where, $H = X(X'X)^{-1}X'$. The $n \times n$ matrix H is known as the hat matrix simply because it maps y into $\hat{y}$. Geometrically, if the data vector y and the columns of X are presented as points in Euclidean n space, then the points $X\beta$ constitute a p dimensional subspace. The fitted vector $\hat{y}$ is the point of that subspace nearest to y, and it is also the perpendicular projection of y into the subspace. Thus H is a projection matrix (Hoaglin & Welsch, 1978).

Allen (1974) computes the PRESS (prediction sum of squares residuals) without having to refit the model for each of the observations is to note that:

$$e_{(i)} = y_i - \hat{y}_i/1 - h_{ii}$$

where h_{ii} is the ith diagonal element of the hat matrix $H = X(X'X)^{-1}X'$. When X has full column rank, $P_X = H = X(X'X)^{-1}X'$ which is known as the hat matrix (Tarpey, 2000a, 2000b).

For the data analyst, the element h_{ij} of H has a direct interpretation as the amount of leverage or influence exerted on $\hat{y}_i$ by y_j (regardless of the actual value of y_j, since H depends only on X). Thus, a look at the hat matrix can reveal sensitive points in the design, points at which the value of y has a large impact on the fit (Huber, 1975; Hoaglin & Welsch, 1978).

In some applications, β will be restricted by some linear constraint:

$$R\beta = r \tag{4}$$

where R is an $r \times p$ matrix ($r \leq p$). When $rank(X) = p$ and $rank(R) = r$, the restricted least squares estimator β_R of β is given by:

$$\beta_R = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}\left(r - R\hat{\beta}\right) \tag{5}$$

where $\hat{\beta} = (X'X)^{-1}X'y$ is the usual unrestricted least squares estimator of β (Draper & Smith, 1981).

For the restricted model, let $\tilde{y}$ denote the vector of fitted values; that is, $\tilde{y} = X\beta_R$. If $r = 0$ in the restriction (4), then $\tilde{y}$ corresponds to the projection of y onto a subspace of the column space of X. In particular, when $rank(X) = p$ and $rank(R) = r$, it follows from equation (5) that:

$$\tilde{y} = (H - J)y$$

where $J = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$ is the projection matrix $P_{X(X'X)^{-1}R'}$.

In a leave-one-out analysis for the restricted least squares model, for each observation i, one can compute the PRESS residual $\tilde{e}_i = y_i - \tilde{y}_i$, where $\tilde{y}_i$ is the predicted response for the ith observation from the restricted model fit with all the data except for the ith observation and $\tilde{e}_i$ is the corresponding PRESS residual from the restricted model. Then, the leave-one-out residuals are given by:

$$\tilde{e}_{(i)} = y_i - \tilde{y}_i/1 - h_{ii} + j_{ii}$$

where h_{ii} denotes the ith diagonal element of the hat matrix H and j_{ii} denotes the ith diagonal element of the matrix J (Tarpey, 2000a).

Theorem:

Under the restriction as:

$$R\beta = 0$$

it must be $\tilde{y} = X\beta_R$ and $\tilde{y} = (H - J)y$ if $X\beta_R = (H - J)y$ if, $\beta_R = Ay$ and $A = [(X'X)^{-1}X' - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X']$, then it should be $XAy = (H - J)y$ and then $XA = (H - J)$. (Proof see Appendix A).

Example 1:

In Figure 1, the real exchange rates in Turkey in the period of 1987 and 2008 are shown. For obtaining data besides a 1 USD + 1.5 EUR basket, consumer prices for Turkey (1987 JAN. = 100) are used in the relative price calculations. Seyidoğlu (2003) denoted that, economic crises which are experienced in the years of 1994 and 2001 caused important structural breaks in real exchange rates (Seyidoğlu, 2003). For this reason, in modeling these break points cubic spline method is used. In the model, t_1: 8 knot corresponds to the year 1994, and t_2: 15 knot to 2001.

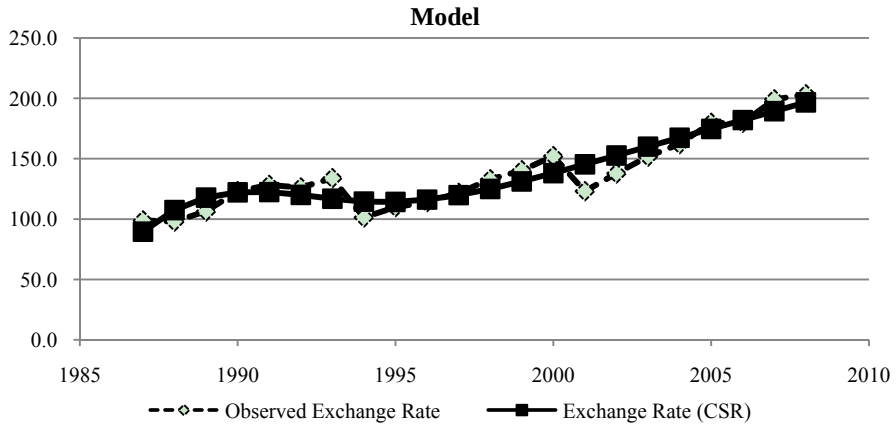

Figure 1. CSR model (model IV) for Turkey's real exchange rate in 1987-2008.

$$E(y) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \beta_{13}(x - t_1)_+^3 + \beta_{23}(x - t_2)_+^3$$

where *y* denotes the real exchange rate, and *x* denotes the year, and:

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

After this demonstration the following cubic spline regression model (restricted model) it can be obtained:

Here the continuous spline is:

$$E(y_i) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \beta_{13}(x - 8)_+^3 + \beta_{23}(x - 15)_+^3$$

$$E(y) = 62.400 + 32.333x - 5.467x^2 + 0.28x^3 - 0.339(x - 8)_+^3 + 0.059(x - 15)_+^3$$

s.e.:	(17.133)	(10.317)	(1.677)	(0.08)	(0.092)	(0.013)
t:	3.642	3.134	-3.260	3.493	-3.681	4.497

F = 1.427 and the model and the coefficients of the model are significant at 95% levels.

Under the restriction of $R\beta = 0$, the values of $\tilde{y} = X\beta_R$ and $\tilde{y} = (H - J)y$ are similar with the $\tilde{y}$ values which are obtained from the model. Because of these values are all overlapping they are seen as a single curve in Figure 2. Furthermore these values are given by years in Appendix B. The slight difference between the $\tilde{y}$ values is due to the use of three digits after the comma.

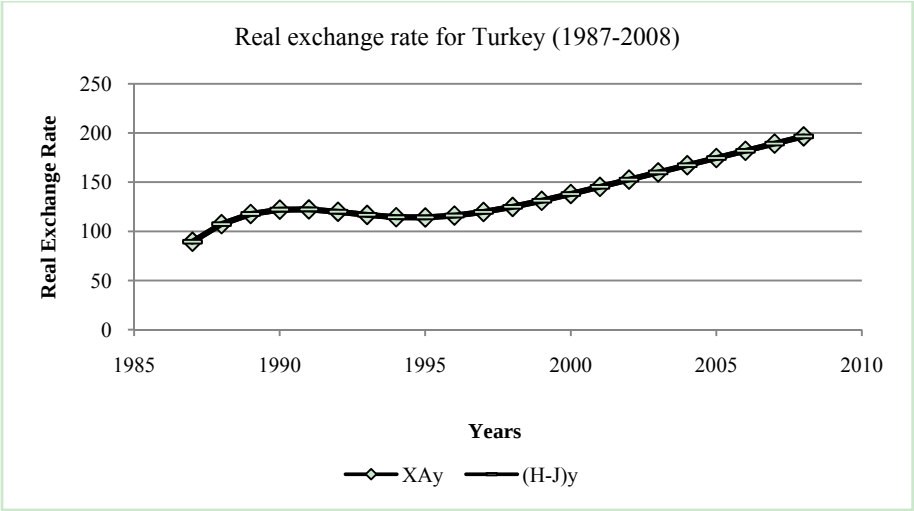

Figure 2. CSR model for Turkey's real exchange rate 1987-2008.

Also, the PRESS statistics for the unrestricted and restricted models are 4,384.508 and 4,230.006 respectively, which indicates that in terms of PRESS, the restricted model performs better than the unrestricted model.

Conclusion

In this work Tarpey's (2000a) findings and arguments on PRESS statistics are supported. In the example used here, the values of PRESS statistics both in restricted and unrestricted models are found. But in Tarpey's (2000a) work, while putting forward some proofs he did not mention the proof of Buse and Lim (1977) on cubic splines. In this work, the relation between the works of Buse and Lim (1977) and Tarpey's (2000a, 2000b) was set and discussed, and it was emerge that similar $\tilde{y}$ values can be estimated in both approaches. That means of course, in the estimation of $\tilde{y}$ values same results can be obtained both with starting from restricted coefficients and from PRESS statistics. Here the choice will be related to the perspective of the researcher. An economist, a statistician or an econometrician will decide and choose the technique of estimating $\tilde{y}$ values under the $R\beta = 0$ constraint according to his/her preference.

Buse and Lim (1977) anyway showed the equality of restricted least squares and cubic spline regression before, in this work, the proof that the values which were estimated with PRESS statistics fit $\tilde{y}$ values in the case of $R\beta = 0$, $\tilde{y}$ is given. As explained in the appendix, estimating similar $\tilde{y}$ values in both approaches may be important for statisticians and econometricians. Additionally, taking benefit from PRESS statistics can also be advised especially to econometricians.

Economic crises even they are global or only in a country size will cause significant breaks in time series. In 1994, the devaluation of TL which is known as 5th April Decisions and the political instability and economic crisis in February 2001 caused significant structural breaks in the real exchange rate values of Turkey. This structural breaks can only be examined with real data (without any transforming or smoothing) with Cubic Spline Regression. According to spline regression theory, spline regression is a preferred method in interpolation. With this method, one can easily reach minimum residual sum of squares and achieve this result by using the real economic data without changing their nature. If an economist cares and takes this advantage into account, he/she will model the breaks in time series with splines and observe the deep impacts of crises better.

When the real exchange rates of Turkey are examined, it can be seen that the rate was 133.8 in 1993 and declined to 101.3 in 1994. This decline might be accepted as a requested result of devaluation in the same year. But also in 2001 Crisis which started suddenly in February, although an economic stability program based on exchange rate was implemented in 1999, similar changes occurred and real exchange rate declined to 123.1 in 2001, although it was 152 in 2000. In conclusion it can be said that, the argument of this work on spline regression is supported exactly with the results.

References

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics, 16*(1), 125-127.

Buse, A., & Lim, L. (1977). Cubic splines as a special case of restricted least squares. *Journal of the American Statistical Association, 72*(357), 64-68

Draper, N. R., & Smith, H. (1981). *Applied regression analysis*. New York: Wiley, 122.

Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in a regression and ANOVA. *The American Statistician, 32*(1), 17-22.

Huber, P. J. (1975). Robustness and designs. In J. N. Srivastava (Ed.), *A survey of statistical design and linear models.* Amsterdam: North-Holland Publishing Co..

Nyquist, H. (1991). Restricted estimation of generalized linear models. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 40*(1), 133-141.

Poirier, D. J. (1973). Piecewise regression using cubic splines. *Journal of the American Statistical Association, 68*(343), 515-524.

Seyidoğlu, H. (2003). Uluslararası Mali Krizler, IMF Politikaları, Az Gelişmiş Ülkeler, Türkiye ve Dönüşüm Ekonomileri, *Doğuş Üniversitesi Dergisi, 4*(2), 141-156.

Tarpey, T. (2000a). Spline bottles. *The American Statistician, 54*(2), 129-135.

Tarpey, T. (2000b). A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician, 54*(2), 116-118.

T. R. Ministry of Development. Economic and Social Statistics (Data File) (2011). Retrieved January 24, 2011, from www.dpt.gov.tr

Appendix A: Proof of Theorem

Buse and Lim (1977) show that the coefficients of the polynomials obtained directly by the restricted least squares (RLS) method are equal to those obtained indirectly from the cubic spline regression (CSR) method. This proof proceeds by showing that $\beta_R = Ay$ and $\beta_S = By$ then $A = B$.

Where:

β_R : RLS estimator and β_S : CSR estimator.

If in RLS method linear restrictions are denoted:

$$R\beta = r$$

$$\beta_R = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})$$

And if $r = 0$,

$$\beta_R = \hat{\beta} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R\hat{\beta}$$

And $\hat{\beta}$: OLS estimator and $\hat{\beta} = (X'X)^{-1}X'y$,

$$\beta_R = \left[(X'X)^{-1}X' - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'\right]y$$

And where if $\beta_R = Ay$,

Then,

$$A = \left[(X'X)^{-1}X' - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'\right]$$

For this reason, Buse ve Lim (1977, pp. 64-68) used such an A matrice in estimating the coefficients of the restricted model, and found the estimated $\tilde{y}$ values with $\tilde{y} = X\tilde{\beta}_R$ formula.

Then, it can be said that:

$$XA = X\left[(X'X)^{-1}X' - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'\right]$$

For the restricted linear regression, there also exists a formula for computing PRESS residuals. Tarpey (2000b) showed a simple way to compute the prediction sum of squares (PRESS) residuals for the restricted least squares fit without having to recompute the fitted model for each of the n observations. For the linear restrictions $R\beta = r$, when the case $r = 0$. Then, the leave-one-out residuals are given by:

$$\tilde{e}_{(i)} = y_i - \tilde{y}_i/1 - h_{ii} + j_{ii}$$

Where h_{ii} denote the ith diagonal element of the hat matrix H and j_{ii} denote the ith diagonal element of the matrix J. And,

$$H = X(X'X)^{-1}X'$$

$$J = X(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'$$

$$\tilde{y} = (H - J)y$$

and

$$\tilde{y} = \left[X(X'X)^{-1}X' - X(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'\right]y$$

$$\tilde{y} = X\left[(X'X)^{-1}X' - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'\right]y$$

$$A = \left[(X'X)^{-1}X' - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'\right]$$

$$\tilde{y} = XAy$$

As seen, $XA = (H - J)$ and then, $XAy = (H - J)y$.

According to the proof which was realized here, both of Buse and Lim's y value and Tarpey's $\tilde{y}$ value fits each other exactly. This fitting or coincidence was shown in Figure 2.

Appendix B

Table B1

Observed and Estimated Real Exchange Rate for Turkey

Years	Observed exchange rate	$\tilde{y}$=XAy	$\tilde{y}$=(H-J)y
1987	98.8	89.5459	89.5459
1988	98.0	107.437	107.437
1989	106.4	117.751	117.751
1990	123.1	122.168	122.168
1991	128.2	122.367	122.367
1992	126.1	120.027	120.027
1993	133.8	116.826	116.826
1994	101.3	114.444	114.444
1995	110.0	114.22	114.22
1996	114.0	116.137	116.137
1997	121.6	119.837	119.837
1998	132.9	124.965	124.965
1999	140.3	131.163	131.163
2000	152.0	138.075	138.075
2001	123.1	145.343	145.343
2002	137.7	152.67	152.671
2003	151.7	159.998	159.999
2004	162.1	167.325	167.327
2005	179.6	174.653	174.655
2006	179.8	181.98	181.983
2007	199.0	189.307	189.311
2008	203.2	196.634	196.639

Note. Source: Retrieved from http://www.dpt.gov.tr.