# Comparison of regression models based on nonparametric estimation techniques: Prediction of GDP in Turkey

Dursun Aydın

***Abstract***- In this study, it has been discussed the comparision of nonparametric models based on prediction of GDP (Domestic Product) per capita prediction in Turkey. It has been considered two alternative situations due to seasonal effects. In the first case, it is discussed a semi-parametric model where parametric component is dummy variable for the seasonality. Smoothing spline and regression spline methods have been used for prediction of the semi-parametric models. In the second case, it is considered the seasonal component to be a smooth function of time, and therefore, the model falls within the class of additive models. The results obtained by semi-parametric regression models are compared to those obtained by additive nonparametric.

***Key words***- Prediction; nonparametric techniques; semi-parametric models; additive models; smoothing spline.

## I. INTRODUCTION

It is considered the following basic model

$$y(t_i) = s(t_i) + z(t_i) + e(t_i),\ i = 1,...,n \qquad (1)$$

where the $t_i$'s are uniformly spaced in [0,1], $s(t_i)$ denotes the seasonal component, $z(t_i)$ represents the trend, and $e(t_i)$ represents the terms of error with zero mean and common variance $\sigma_e^2$. The model mentioned here can be written as,

$$y_i = s_i + z_i + e_i,\ i = 1,2,...,n. \qquad (2)$$

It is assumed that the following structure for the trend:

$$z_i = f(t_i) + \varepsilon_i,\ i = 1,2,...,n \qquad (3)$$

where $f$ is a smooth function in [0,1], and $\varepsilon_i$'s are assumed to be with zero mean and common variance $\sigma_\varepsilon^2$, and different from $e_i$'s.

The basic aim is to estimate the functions $f$ and $s$. The function $f$ is estimated as a smooth function, but the estimation of the function $s$ is different due to seasonality. Therefore, it is considered two alternative models for the estimation of $s$. Firstly, it is treated a semi-parametric model where parametric component is dummy variable for the seasonality. Secondly, it is discussed the seasonal component to be a smooth function of time, and use a nonparametric method.

## II.SEMİPARAMETRİC ESTIMATION

It is assume that the seasonality is build as follows:

$$s_i = s(t_i) = \sum_{k=1}^{r-1} \beta_k D_{ki}^* + v_i,\ i = 1,...,n \qquad (4)$$

where $r$ is the number of annual observations ($r$=12) and $v_i$'s are assumed to be with zero mean and common variance $\sigma_v^2$, and different from the errors in (2) and (3). $D_{ki}^*$'s are dummy variable that denotes the seasonal effects and $\beta_k$'s are parametric coefficients. Dummy variables are denoted by $D_{ki}^* = D_{ki} - D_{ri}$ (where $D_{ki} = 1$ if $i.$ observation correspond to the $kth$ month of year, and $D_{ki} = 0$ otherwise) for cancels the seasonal effects when a year is completed [5]. By substitution equations (4) and (3) in (2), it is obtained as

$$y_i = \sum_{k=1}^{r-1} \beta_k D_{ki}^* + f(t_i) + u_i, \qquad (5)$$

where $u_i$'s are the sum of the random errors with zero means and constant variance $\sigma_u^2 = \sigma_e^2 + \sigma_\varepsilon^2 + \sigma_v^2$. Eq. (5) in vector-matrix form can be written

$$\mathbf{y} = D\boldsymbol{\beta} + \mathbf{f} + \mathbf{u} \qquad (6)$$

where $D$ is the $n \times (r-1)$ matrix, so that $D^T = \left\{D_{ki}^{*}\right\}_{k=1,\dots,r-1}^{i=1,\dots,n}$, $\boldsymbol{\beta} = \left(\beta_1,\dots,\beta_{r-1}\right)^T$, $\mathbf{y} = \left(y_1,\dots,y_n\right)^T$, $\mathbf{f} = \left(f(t_1),\dots,f(t_n)\right)^T$, and $\mathbf{u} = \left(u_1, u_2, \dots, u_n\right)^T$.

Therefore,

$$D^T = \begin{bmatrix} 1 & 0 & . & . & . & 0 & -1 & 1 & 0 & . & . & . \\ 0 & 1 & . & . & . & 0 & -1 & 0 & 0 & . & . & . \\ & & . & & & & & & & & & \\ & & & . & & & & & & & & \\ & & & & . & & & & & & & \\ 0 & 0 & . & . & . & 1 & -1 & 0 & 0 & . & . & . \end{bmatrix}$$

Model (5) is called as a semi-parametric model due to consist of a parametric linear component and only a nonparametric component. The basic purpose, it is estimation of the parameter vector $\boldsymbol{\beta}$ and function $f$ at sample points $t_1,\dots,t_n$. For this aim, it is considered two estimation methods that called as smoothing spline, and regression spline.

***Estimation with smoothing spline*:** Estimation of the parameters of interest in equation (5) can be performed using smoothing spline. Mentioned here the vector parameter $\boldsymbol{\beta}$ and the values of function $f$ at sample points $t_1,\dots,t_n$ are estimated by minimizing the penalized residual sum of squares

$$PSS\left(\boldsymbol{\beta},\mathbf{f}\right) = \sum_{i=1}^{n}\left\{y_i - d_i^T\boldsymbol{\beta} - f(t_i)\right\}^2 + \lambda\int_0^1\left(f^{(m)}(u)\right)^2 du \qquad (7)$$

where $f \in C^2[0,1]$ and $d_i$ is the $ith$ row of the matrix $D$. When the $\boldsymbol{\beta} = 0$, resulting estimator has the form $\hat{\mathbf{f}} = \left(\hat{f}(t_1),\dots,\hat{f}(t_\mathrm{n})\right) = S_\lambda \mathbf{y}$, where $S_\lambda$ a known positive-definite (symmetric) smoother matrix that depends on $\lambda$ and the knots $t_1,\dots,t_n$ (see, [1]; [2]; [3]).

For a pre-specified value of $\lambda$ the corresponding estimators for $\mathbf{f}$ and $\boldsymbol{\beta}$ based on Eq. (5) can be obtained as follows [4]: Given a smoother matrix $S_\lambda$, depending on a smoothing parameter $\lambda$ ,construct $\hat{D} = \left(I - S_\lambda\right)D$. Then, by using penalized least squares, mentioned here estimator are given by

$$\hat{\boldsymbol{\beta}} = \left(D^T\hat{D}\right)^{-1}\hat{D}^T\mathbf{y} \qquad (8)$$

$$\hat{\mathbf{f}} = S_\lambda\left(\mathbf{y} - D\hat{\boldsymbol{\beta}}\right) \qquad (9)$$

Evaluate some criterion function (such as cross validation, generalized cross validation) and iterate changing $\lambda$ until it is minimized.

***Estimation with regression spline*:** Smoothing spline become less practical when $n$ is large, because they use $n$ knots. A more general approach to spline fitting is regression spline. Smoothing spline require that many parameters be estimated, typically at least at many parameter as observations. A regression spline is a piecewise polynomial function whose highest order nonzero derivative takes jumps at fixed “knots”. Usually regression splines are smoothed by deleting nonessential knots. When the knots have been selected, regression spline can be fit by ordinary least squares. For further discussion on selection of knots, see to Ruppert and Carroll, 2002.

It is approximated $f(t_i)$ in (5) by

$$f(t_i) = f(t_i,\gamma) = \gamma_0 + \gamma_1 t_i + \dots + \gamma_p t_i^p + \sum_{k=1}^{K} b_k\left(t_i - \kappa_k\right)_+^p,\ i = 1,\dots,n \qquad (10)$$

where $p \geq 1$ is an integer (order of the regression spline and usullay chosen a priori), $b_1,\dots,b_K$ are independently and identically distributed (i.i.d) with $N(0,\sigma_b^2)$, $(t)_+ = t$ if $t > 0$ and 0 otherwise and $\kappa_1 < \dots . < \kappa_k$ are fixed knots ($\min(t_i) < \kappa_1,\dots,< \kappa_K < \max(t_i)$).

In matrix notation model (5) can be written as

$$\mathbf{y} = D\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{\eta} \qquad (11)$$

Where

$$D^T = \begin{bmatrix} 1 & 0 & . & . & . & 0 & -1 & 1 & 0 & . & . & . & 1 & t_1 & . & . & . & t_n \\ 0 & 1 & . & . & . & 0 & -1 & 0 & 1 & . & . & . & 1 & t_1^{p-1} & . & . & . & t_n^{p-1} \\ & & . & & & & & & & & & & . & . & & & & . \\ & & & . & & & & & & & & & . & . & & & & . \\ & & & & . & & & & & & & & . & . & & & & . \\ 0 & 0 & . & . & . & 1 & -1 & 0 & 0 & . & . & . & 1 & t_1^p & . & . & . & t_n^p \end{bmatrix}$$

$$Z = \begin{bmatrix} (t_1 - \kappa_1)_+^p & . & . & . & (t_1 - \kappa_K)_+^p \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ (t_n - \kappa_1)_+^p & . & . & . & (t_n - \kappa_K)_+^p \end{bmatrix}$$

$\mathbf{b} = \left(b_1,\dots,b_K\right)^T$ is vector of coefficients and $\boldsymbol{\eta} = \left(\eta_1,\dots,\eta_n\right)^T$ is a vector of the random error. Predicted value of $\hat{y}$ in (11) is given by

$$\hat{\mathbf{y}}=\hat{\gamma}_0+\hat{\gamma}_1 t_i+...+\hat{\gamma}_p t_i^p+\hat{\beta}_1 D_1+...+\hat{\beta}_{r-1} D_K+\left(Z_1,...,Z_K\right)\left(\hat{b}_1,...,\hat{b}_K\right)^T. \quad (12)$$

Regression spline estimators $\left(\hat{\boldsymbol{\beta}}=\left(\hat{\gamma}_0,\hat{\gamma}_1,...,\hat{\gamma}_p,\hat{\beta}_1,...,\hat{\beta}_{r-1}\right)^T, \hat{\mathbf{f}}=(\hat{b}_1,...,\hat{b}_k)\right)$ of $\left(\boldsymbol{\beta}, \mathbf{f}\right)$ are defined as the minimizer of

$$PSS(\boldsymbol{\beta}, \mathbf{f})=\sum_{i=1}^{n}\left\{y_i-d_i^T\boldsymbol{\beta}-f(t_i)\right\}^2+\lambda\sum_{k=1}^{K}b_k^2\,, \quad (13)$$

where $\lambda > 0$ is a smoothing parameter as in (7). As $\lambda \to \infty$, the regression spline converges to a $pth$ degree polynominal fit. As $\lambda \to 0$, the regression spline converges to the ordinary least squares (OLS) fitted spline. For a pre-specified value of $\lambda$ the corresponding estimators for $\boldsymbol{\beta}$ and $\mathbf{f}$ based on Eq. (12) can be obtained as follows [6]:

$$\hat{\boldsymbol{\beta}}=\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X\hat{\Sigma}^{-1}\mathbf{y}\,, \quad (14)$$

where, $\hat{\Sigma}=ZZ^T\hat{\sigma}_b^2+diag(\hat{\sigma}_{\eta i}^2), \quad i=1,2,...,n$,

$$\hat{\mathbf{f}}=(\hat{b}_1,...,\hat{b}_k)^T=\hat{\sigma}_b^2 Z^T\hat{\Sigma}^{-1}\left(\mathbf{y}-D\hat{\boldsymbol{\beta}}\right) \quad (15)$$

The smoothing parameter (penalty parameter $\lambda$ ) and the number of knots $K$ must be selected in implementing the regression spline. However, $\lambda$ plays a more important role. See Ruppert (2002) for a detailed discussion of the knot selection. The solution can be obtained by S-Plus software.

## III. NONPARAMETRIC ESTIMATION

In the previous section it was used semi-parametric model for estimation of the parameters in (5). However, there are situations in which a dummy variable specification does not capture all fluctuations because of the seasonal effects. For this reason, in this section it is considered a more general case for seasonal component as follows:

$$s_i=g(t_i)+v_i,\; i=1,...,n \quad (16)$$

where $g$ is an [0,1] and $g\in C^2[a,b]$, $v_i$'$s$ are denote the terms of random error with zero mean and common variance $\sigma_v{}^2$. By substitution of the equations (3) and (16) in (2), it is obtained as

$$y_i=g(t_i)+f(t_i)+u_i,\; i=1,...,n\,, \quad (17)$$

where $u_i$'$s$ are the terms of random error with zero mean and constant variance $\sigma_u^2=\sigma_e^2+\sigma_\varepsilon^2+\sigma_v^2$.

Model (17) mentioned above has a fully nonparametric model because of the parametric component is missing. These models are called additive nonparametric regression models. In order to estimate model (17), it can be generalized the criterion (7) and (17) in an obvious way. Estimator of the model (17) is based on minimum of the penalized residual sum of squares [7]

$$PSS(\mathbf{f}, \mathbf{g})=\sum_{i=1}^{n}\{y_i-f(t_i)-g(t_i)\}^2+\lambda_1\int_0^1\left(f^{(m)}(u)\right)^2du+\lambda_2\int_0^1\left(g^{(l)}(u)\right)^2du \quad (18)$$

The first term in (18) denotes the residual sum of the squares (RSS) and this term penalizes the lack of fit. The second term multiplicand by $\lambda_1$ is denote the roughness penalty for the $f$ and the third term multiplicand by $\lambda_2$ is denote the roughness penalty for $g$. Firstly, eq. (18) can be written as

$$PSS(\mathbf{f},\mathbf{g})=(\mathbf{y}-\mathbf{f}-\mathbf{g})^T(\mathbf{y}-\mathbf{f}-\mathbf{g})+\lambda_1\mathbf{f}^T K_f\mathbf{f}+\lambda_2\mathbf{g}^T K_g\mathbf{g} \quad (19)$$

Here $K_f$ is a penalty matrix for $\mathbf{f}$ and $K_g$ is a penalty matrix for $\mathbf{g}$. Then, by differentiating according to $\mathbf{f}$ and $\mathbf{g}$, it is obtained as follow:

$$\frac{PSS(\mathbf{f},\mathbf{g})}{\mathbf{f}}=-2(\mathbf{y}-\mathbf{f}-\mathbf{g})+2\lambda_1K_f\mathbf{f} \quad (20)$$

$$\frac{PSS(\mathbf{f},\mathbf{g})}{\mathbf{g}}=-2(\mathbf{y}-\mathbf{f}-\mathbf{g})+2\lambda_2K_g\mathbf{g} \quad (21)$$

Afterwards, by making (20) and (21) equal to zero, the estimators of $\mathbf{f}$ and $\mathbf{g}$ are defined by

$$\begin{aligned}\hat{\mathbf{f}}&=(I+\lambda_1K_f)^{-1}(\mathbf{y}-\mathbf{g})=S_{\lambda_1}(\mathbf{y}-\mathbf{g})\\ \hat{\mathbf{g}}&=\left(I+\lambda_2K_g\right)^{-1}\left(\mathbf{y}-\mathbf{f}\right)=S_{\lambda_2}\left(\mathbf{y}-\mathbf{f}\right)\end{aligned} \quad (22)$$

## IV. AN APPLICATION: PREDICTION OF TIME SERIAL GDP IN TURKEY

For the purpose of illustration let us analyze a data set, known as the GDP for Turkey. Data related to variables used in this study consists of monthly time series which starts January, 1984 and ends December 2001, comprising $n=216$ observations. Mentioned here variables are defined as follow:

**gdp** : Gross Domestic Product ( TL )

**time** : Data monthly from January 1984 up to December 2001

$\boldsymbol{D}_{k=1}^{r-1}$ : Dummy variables that denotes the effects seasonality

The main idea of this application presented here is to estimate time series and compare the nonparametric regression models in section 2 and 3. Semi-parametric regression results obtained using smoothing spline with $m=l=2$, which for the method presented section three are very similar to nonparametric regression ones obtained using the same method. The solution can be obtained by S-Plus and R software [8].

### A. EMPRICAL RESULTS

It is discussed a semi-parametric regression model where parametric components are dummy variables for the seasonality. Results obtained with this model are given in Table 1.

**Table1.** Results obtained by smoothing spline

| | Parametric Part | | | |
|---|---|---|---|---|
| | Estimate | St. Error | t value | Pr(>\|t\|) |
| (Intercept) | -17.352 | 1.84e-01 | -94.35 | 6.38e-16 |
| S(time,15) | 0.020 | 9.23e-05 | 220.90 | 4.74e-23 |
| D1 | 0.019 | 1.59e-03 | 11.711 | 3.61e-24 |
| D2 | -0.073 | 1.59e-03 | -46.235 | 5.60e-10 |
| D3 | 0.019 | 1.59e-03 | 11.711 | 3.61e-24 |
| D4 | -0.014 | 1.59e-03 | -8.935 | 3.65e-16 |
| D5 | 0.019 | 1.59e-03 | 11.711 | 3.61e-24 |
| D6 | -0.014 | 1.59e-03 | -8.935 | 3.65e-16 |
| D7 | 0.019 | 1.59e-03 | 11.711 | 3.61e-24 |
| D8 | 0.019 | 1.59e-03 | 11.711 | 3.61e-24 |
| D9 | -0.014 | 1.59e-03 | -8.935 | 3.65e-16 |
| D10 | 0.019 | 1.59e-03 | 11.711 | 3.61e-24 |
| D11 | -0.014 | 1.59e-03 | -8.935 | 3.65e-16 |
| | Nonparametric Part | | | |
| | Df Npar | Df | Npar F | Pr(F) |
| S(time1) | 1 | 14 | 766.07 | 2.2e-16 |
| Response: log(gdp); Deviance=0.009; $R^2$ = 0.998; MSE=1.767633e-28 | | | | |

According to Table 1, it is shown that both parametric and nonparametric coefficients are significance. So, we can say that GDP is under the effect of months. On the other hand, for example, a one-unit increase in time corresponds to mean increase of 0.020 GDP. As shown Table 1, the $R^2$ value is 99.8 %. An $R^2$ value of 0.998 means that only 98.8 % of variability in GDP is predictable using the semi-parametric model.

As shown the Table 2, results obtained by regression spline is similar to results of the smoothing spline. Semi-parametric model is significance because of the parametric and nonparametric components. As Table 2, GDP is under of the effects of months. Aspect of the relation between dummy variables and GDP are same for both methods (see the estimate columns of the Table 1 and 2). On the other words, while there is a positive relation between GDP and D1, D3, D5, D7, D8 and D10, there is a negative relation between GDP and other remain variables. Furthermore, models obtained by smoothing spline and regression spline have a very small deviance.

According to Table 3, the effects of interaction seasonality and time on GDP are significant in statistical. So, curvilinear effects are significant. Analogously to semi-parametric models, 96.3 % of variability in GDP is predictable by nonparametric model. As such in semi-parametric models, additive model has a small deviance too.

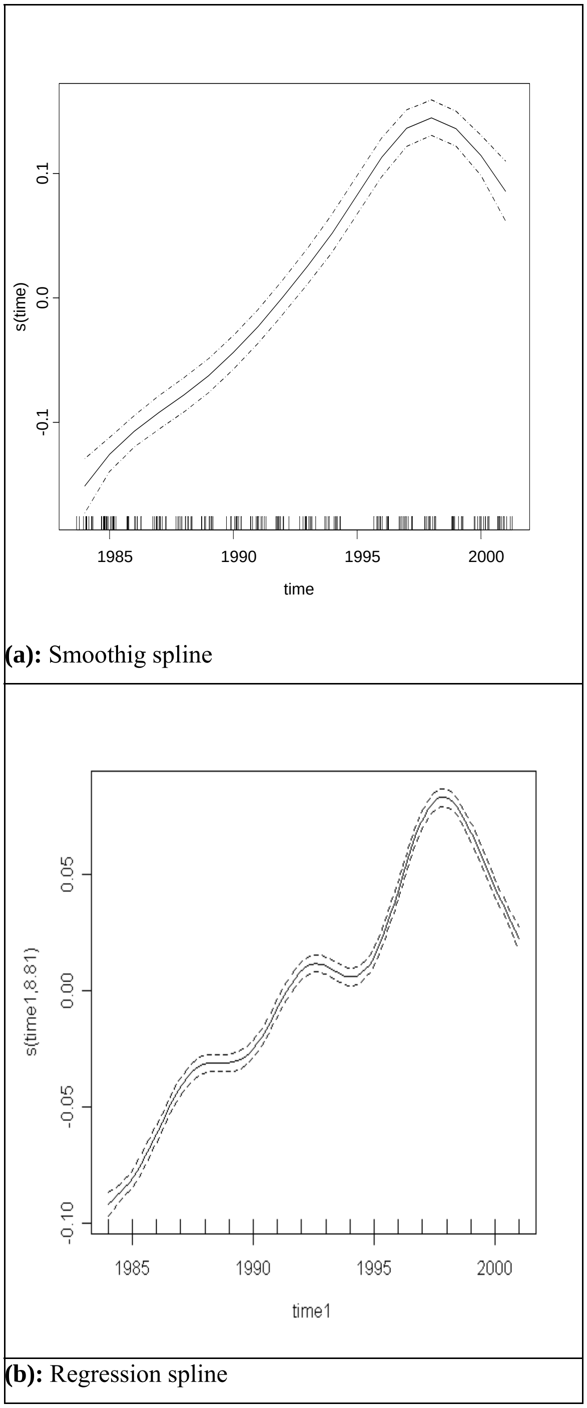
The variable in nonparametric part of semi-parametric model can be only displayed graphically, because it can't be expressed as parametric. Figure 1 (**a**) and **(b)** shows the estimates (solid) and the 95% confidence intervals (dashed) for smoothing spline and regression spline methods. As shown Figure 1 (**a**) and **(b)**, shape of the effects of trend on GDP is appears as a curve. Figure 2 shows the estimates and 95% confidence intervals for additive regression model. As such in Figure 1 (**a**) and **(b),** the shape of the effects of trend on GDP is appears as a curve in Figure 2. Mentioned here curve are significant (see Table 3).

**Table2.** Results obtained by regression spline

| | Parametric Part | | | |
|---|---|---|---|---|
| | Estimate | St. Error | t value | Pr(>\|t\|) |
| (Intercept) | 10.108 | 0.000670 | 15087 | 2e-16 |
| D1 | 0.0080 | 0.002222 | 3.635 | 0.000355 |
| D2 | -0.0318 | 0.002222 | -14.353 | 2e-16 |
| D3 | 0.0080 | 0.002222 | 3.635 | 0.000355 |
| D4 | -0.0061 | 0.002222 | -2.774 | 0.006083 |
| D5 | 0.0080 | 0.002222 | 3.635 | 0.000355 |
| D6 | -0.0061 | 0.002222 | -2.774 | 0.006083 |
| D7 | 0.0080 | 0.002222 | 3.635 | 0.000355 |
| D8 | 0.0080 | 0.002222 | 3.635 | 0.000355 |
| D9 | -0.0061 | 0.002222 | 2.774 | -0.00608 |
| D10 | 0.0080 | 0.002222 | 3.635 | 0.000355 |
| D11 | -0.0061 | 0.002222 | -2.774 | -0.00608 |
| | Nonparametric Part | | | |
| | Df Npar | Df | Npar F | Pr(F) |
| S(time) | 8.812 | 9 | 617.7 | 2e-16 |
| Response: log(gdp); Deviance=0.100304; $R^2$ = 0.965; MSE= 7.40571e-27 | | | | |

**Table 3.** Results obtained by nonparametric regression

| | Df | Npar Df | Npar F | Pr(F) |
|---|---|---|---|---|
| s(time) | 1 | 3 | 79.121 | 2.2e-16 |
| s(time×D1) | 1 | 3 | 9.457 | 8.281e-06 |
| s(time×D2) | 1 | 3 | 9.739 | 5.845e-06 |
| s(time×D3) | 1 | 3 | 9.831 | 5.216e-06 |
| s(time×D4) | 1 | 3 | 9.455 | 8.303e-06 |
| s(time×D5) | 1 | 3 | 9.235 | 1.091e-05 |
| s(time×D6) | 1 | 3 | 9.225 | 1.105e-05 |
| s(time×D7) | 1 | 3 | 9.184 | 1.164e-05 |
| s(time×D8) | 1 | 3 | 9.182 | 1.166e-05 |
| s(time×D9) | 1 | 3 | 9.217 | 1.117e-05 |
| s(time×D10) | 1 | 3 | 9.182 | 1.166e-05 |
| s(time×D11) | 1 | 3 | 9.217 | 1.117e-05 |
| Response : log (gdp); Deviance = 0.224; $R^2$ = 0.963; MSE= 4.733165e-30 | | | | |



**(a):** Smoothig spline

**(b):** Regression spline

**Figure 1:** Estimates (solid) and the 95 % confidence intervals (dashed) for semi-parametric regression

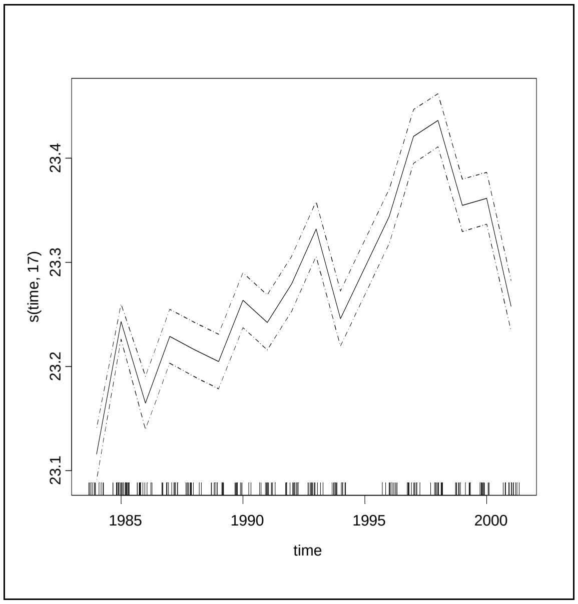


**Figure 2:** Estimates (solid) and the 95 % confidence intervals (dashed) for additive regression

## V. CONCLUSION

In this paper it has been discussed two alternative models based on nonparametric regression techniques for estimation in time series including trend and seasonality. Results obtained with these two models have been compared to additive model. Some of the performance criteria associated with these models have been given the following Table:

| | Performance criteria of the models | | |
|---|---|---|---|
| | MSE | Deviance | $R^2$ |
| Smoothing spline | 1.76763e-28 | 0.009 | 0.998 |
| Regression spline | 7.40571e-27 | 0.100 | 0.965 |
| N.parametric model | 4.733165e-30 | 0.224 | 0.963 |

These results emphasize that estimates based on smoothing spline method is very better than the regression spline and additive traditional methods, like a parametric linear regression. However, estimates obtained by semi-parametric regression model using smoothing spline are better than results obtained by regression spline and additive model.

## REFERENCES

[1] Eubank, R. L., Nonparametric Regression and Smoothing Spline, Marcel Dekker Inc., 1999.

[2] Wahba, G., Spline Model For Observational Data. Siam, Philadelphia Pa., 1990.

[3] Green, P.J. and Silverman, B.W., Nonparametric Regression and Generalized Linear Models, Chapman & Hall, 1994.

[4] Green, P., Yandell, B.S., Semi-parametric generalized linear models. Proceedings of the second International GLIM conference, Lancaster, Lectures Notes in Statistics 32, Springer, New York, pp.44-55.

[5] Eva, F., Vicente, N, A., Juan, R, P., Semi-parametric approaches to signal extraction problems in economic time series, Computational Statistics & Data Analysis, Vol: 33, 2000, pp:315-333.

[6] Ruppert, D., Wand, M.P., Carrol, R.J., Semi-parametric regression, Cambridge university press, Cambridge, 2003.

[7] Hastie, T.J. and Tibshirani, R.J., Generalized Additive Models, Chapman & Hall /CRC, 1999.

[8] Chambers, J, H., Hastie, T, J., Statistical Models in S. wadsworth & Books / Cole, Pacific Grove

[9] Ruppert, D., Selection the number of knots for penalized spline, Journal of Computational and Graphical Statistics, vol: 11, pp:735-757, 2002.

[10] Ruppert, D., Carrol, R., Spatially-adaptive penalties for penalized spline, New Zeland J.statist, vol:42, pp:205-224, 2000.