Estimation of GDP in Turkey by nonparametric regression models

DURSUN AYDIN
Anadolu University- Eskişehir / TURKEY

Abstract:- Present study is about using of nonparametric models for GDP (Gross 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. 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 and parametric linear models.

Key words:- Estimation, semi-parametric models, additive models, smoothing spline, trend.

1 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 σ_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 ε_i's are assumed to be with zero mean and common variance σ_ε^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.

2 Semi-parametric 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 σ_v^2, and different from previous errors. D_{ki}^{*}'s are dummy variable that denotes the seasonal effects and β_k's are parametric coefficients. Dummy variables are denoted by $D_{ki}^{*}=D_{ki}-D_{ri}$ (where $D_{ki}=1$ if i correspond to the kth month of year, and $D_{ki}=0$ otherwise) for cancels the seasonal effects when a year is completed [1]. 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 $D^T=\left\{D_{ki}^{*}\right\}_{k=1,...,r-1}^{i=1,...,n}$, $\boldsymbol{\beta}=(\beta_1,...,\beta_{r-1})^T$, $\mathbf{y}=(y_1,...,y_n)^T$, $\mathbf{f}=(f(t_1),...,f(t_n))^T$, and $\mathbf{u}=(u_1,u_2,...,u_n)^T$.

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 $t_1,...,t_n$. 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 $t_1,...,t_n$ are estimated by minimizing the penalized residual sum of squares

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

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 $\hat{\mathbf{f}}=\left(\hat{f}(t_1),...,\hat{f}(t_n)\right)=S_\lambda\mathbf{y}$, where S_λ a known positive-definite (symmetric) smoother matrix that depends on λ and the knots $t_1,...,t_n$ (see, [2]; [3]; [4]).

For a pre-specified value of λ the corresponding estimators for $\mathbf{f}$ and $\boldsymbol{\beta}$ based on Eq. (5) can be obtained as follows [5], [6]: Given a smoother matrix S_λ, depending on a smoothing parameter λ,construct $\tilde{D}=(I-S_\lambda)D$. Then, by using penalized least squares, mentioned here estimator are given by

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

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

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

3 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:

$$s_i=g(t_i)+v_i,\ i=1,...,n \qquad (10)$$

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 σ_v^2. By substitution of the equations (3) and (10) in (2), it is obtained as

$$y_i=g(t_i)+f(t_i)+u_i,\ i=1,...,n, \qquad (11)$$

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 (11) 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 (11), it can be generalized the criterion (7) and (11) in an obvious way. Estimator of the model (11) is based on minimum of the penalized residual sum of squares [7]

$$PSS(\mathbf{f},\mathbf{g})=\sum_{i=1}^{n}\left\{y_i-f(t_i)-g(t_i)\right\}^2+\lambda_1\int_0^1\left(f^{(m)}(u)\right)^2du+\lambda_2\int_0^1\left(g^{(l)}(u)\right)^2du \qquad (12)$$

The first term in (12) denotes the residual sum of the squares (RSS) and this term penalizes the lack of fit. The second term multiplicand by λ_1 is denote the roughness penalty for the f and the third term multiplicand by λ_2 is denote the roughness penalty for g. Firstly, eq. (12) can be written as

$$PSS(\mathbf{f},\mathbf{g})=(\mathbf{y}-\mathbf{f}-\mathbf{g})^T(\mathbf{y}-\mathbf{f}-\mathbf{g})+\lambda_1\mathbf{f}^TK_f\mathbf{f}+\lambda_2\mathbf{g}^TK_g\mathbf{g} \qquad (13)$$

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:

$$\frac{PSS(\mathbf{f},\mathbf{g})}{\mathbf{f}}=-2(\mathbf{y}-\mathbf{f}-\mathbf{g})+2\lambda_1K_f\mathbf{f} \qquad (14)$$

$$\frac{PSS(\mathbf{f},\mathbf{g})}{\mathbf{g}}=-2(\mathbf{y}-\mathbf{f}-\mathbf{g})+2\lambda_2K_g\mathbf{g} \qquad (15)$$

Afterwards, by making (14) and (15) equal to zero, the estimators of $\mathbf{f}$ and $\mathbf{g}$ are defined by

$$\hat{\mathbf{f}}=(I+\lambda_1K_f)^{-1}(\mathbf{y}-\mathbf{g})=S_{\lambda_1}(\mathbf{y}-\mathbf{g}) \qquad (16)$$

$$\hat{\mathbf{g}}=\left(I+\lambda_2K_g\right)^{-1}\left(\mathbf{y}-\mathbf{f}\right)=S_{\lambda_2}\left(\mathbf{y}-\mathbf{f}\right) \qquad (17)$$

4 An application: Estimation of 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].

4.1 Empirical results

Firstly, it is discussed a semi-parametric regression model where parametric components are dummy variables for the seasonality. Secondly, it has been treated nonparametric additive models and finally, discussed linear parametric regression. Results obtained with these models are given in Table 1, Table 2 and Table 3 respectively.

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 99.85 %. An R^2 value of 0.9985 means that only 98.85 % of variability in GDP is predictable using the semi-parametric model.

As shown the Table 2, the effects of interaction seasonality with time on GDP are significance in statistical. So, curvilinear effects are significance and similarly to semi-parametric model, 96.38 % of variability in GDP is predictable by nonparametric model. Furthermore, a one-unit increase in time corresponds to mean increase of 79.121 GDP. This model is account for 98.85 % of the variability in response variable.

According to Table 3, most of the coefficients in parametric linear model are not significance in statistical. This model has a smaller R^2 value and a bigger RSS value than the other model. For this reason, estimates obtained by using ordinary least square in parametric linear model are unfavourable.

The variable in nonparametric part of semi-parametric model can be only displayed graphically, because it can't be expressed as parametric. While Figure 1 (**a)** and **(b)** shows the estimates (solid) and the 95% confidence intervals (dashed) for nonparametric techniques, Figure 2 shows the estimates and 95% confidence intervals for parametric linear regression. As shown Figure-1 (**a)** and **(b)**, shape of the effects of trend on GDP is appears as a curve.

Estimated values in these figures (**a-b**) are following true regression curves very closely. This situation indicates that estimated values are very good. However, estimated values in Figure 2 are not following the regression curve closely. So, estimated values by this model aren't good results.

Table1. Results obtained by semi-parametric regression

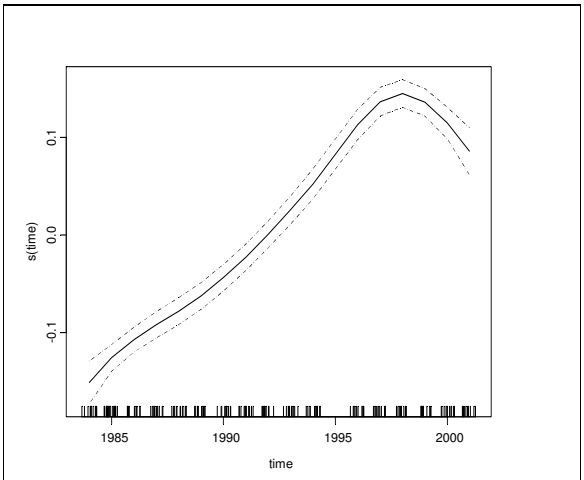
	Parametric Part			
	Estimate	St. Error	t value	Pr(>|t|)
(Intercept)	-17.352	1.84e-01	-94.35	6.38e-16
S(time,15)	0.020	9.23e-05	220.90	4.74e-23
D1	0.019	1.59e-03	11.711	3.61e-24
D2	-0.073	1.59e-03	-46.235	5.60e-10
D3	0.019	1.59e-03	11.711	3.61e-24
D4	-0.014	1.59e-03	-8.935	3.65e-16
D5	0.019	1.59e-03	11.711	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 = 0.9985; MSE=0.238				

Table 2. Results obtained by nonparametric regression

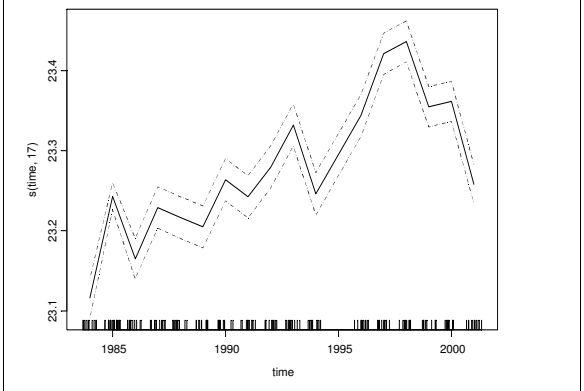
	Df	Npar Df	Npar F	Pr(F)
s(time)	1	3	79.121	2.2e-16
s(time×D1)	1	3	9.457	8.281e-06
s(time×D2)	1	3	9.739	5.845e-06
s(time×D3)	1	3	9.831	5.216e-06
s(time×D4)	1	3	9.455	8.303e-06
s(time×D5)	1	3	9.235	1.091e-05
s(time×D6)	1	3	9.225	1.105e-05
s(time×D7)	1	3	9.184	1.164e-05
s(time×D8)	1	3	9.182	1.166e-05
s(time×D9)	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 = 0.9638; MSE=0.367				

Table3. Results obtained by parametric linear regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Int.)	-12.3517	1.6899	-7.3091	0.0000
time	0.0179	0.0008	21.0822	0.0000
D1	0.0186	0.0153	1.2194	0.2241
D2	-0.0734	0.0153	-4.8145	0.0000
D3	0.0186	0.0153	1.2194	0.2241
D4	-0.0142	0.0153	-0.9303	0.3533
D5	0.0186	0.0153	1.2194	0.2241
D6	-0.0142	0.0153	-0.9303	0.3533
D7	0.0186	0.0153	1.2194	0.2241
D8	0.0186	0.0153	1.2194	0.2241
D9	-0.0142	0.0153	-0.9303	0.3533
D10	0.0186	0.0153	1.2194	0.2241
D11	-0.0142	0.0153	-0.9303	0.3533
Response: log(gdp); RSS = 0.540; $R^2 = 0.7021$; MSE=289				



(a): For semi-parametric regression



(b): For nonparametric regression

Figure 1: Estimates (solid) and the 95 % confidence intervals (dashed)

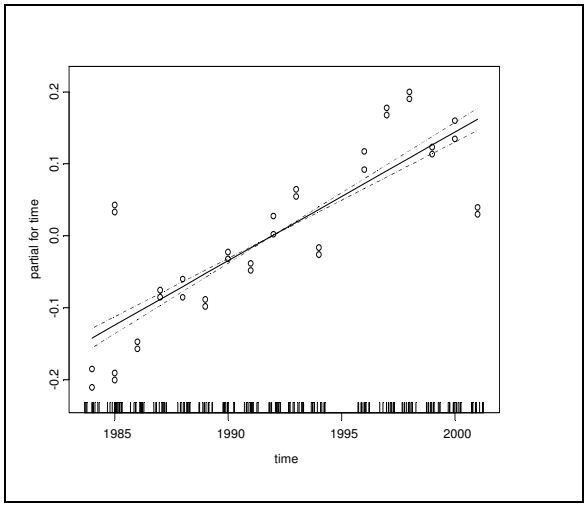


Figure 2: Estimates (solid) and the 95 % confidence intervals (dashed) for parametric linear regression

5 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 parametric linear 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
Semi-parametric model	0.238	0.009	0.9985
Nonparametric model	0.367	0.224	0.9638
Parametric linear model	0.289	(RSS) 0.540	0.7021

In brief, from a closer inspection of the empirical results, the following observations were made:

- The semi-parametric model has the smallest MSE and Deviance. On the other hand, 99.85 % of variability in GDP is predictable using the semi-parametric model.
- The results of the nonparametric model are close to semi-parametric ones. However, MSE value of the nonparametric model is higher than MSE values of the two other models.
- RSS and R^2 values of the linear parametric

regression are worse from those of the semi-parametric and nonparametric additive model.

These results emphasize that estimates based on nonparametric regression techniques are very better than the traditional methods, like a parametric linear regression.. However, estimates obtained by semi-parametric regression model are better than nonparametric ones.

References

[1] 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**,** pp:315-333, 2000.

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

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

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

[5] 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, 1985.

[6] Schimek, G. Michael, Estimation and İnference in Partially Linear Models with Smoothing Splines, Journal of Statistical Planning and Inference, 91, 525-540, 2000.

[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, 1992.