Advances in the Theory of Nonlinear Analysis and its Applications 7 (2023) No. 4, 54-59.
https://doi.org/10.17762/atnaa.v7.i4.282
Available online at www.https://atnaea.org/
Research Article

Advances in the Theory of Nonlinear Analysis and its Applications

ISSN: 2587-2648

Peer-Reviewed Scientific Journal

A comparison Between Two Methods Of Estimating a Semi- parametric Regression Model in The Presence of The Autocorrelation Problem

Hassan .H.Razaq[a], Hayder .R. Talib[b], Reem .T. Kamil[c]

[a] *University of Thi-Qar. Gollege of Administration & Economics Department of Economics.*
[b] *University of Sumer. Gollege of Administration & Economics Department of Statistics.*
[c] *University of Baghdad. Gollege of Physical Education and sports Sciences for Girls. Department of Theoretical Sciences.*

Abstract

In this research, a comparison was made between two methods for estimating a semiparametric regression model with the presence of an autocorrelation problem, based on a semiparametric partial linear regression model, which contains a parametric component and a nonparametric component. The two components were estimated using two methods, the first methods is semi parametric generalized least squares estimators(SGLSE) , the second methods is least squares estimators method(LSEM) .Simulation study show the first method is beast than the second method by using mean squares of errors (MSE) .

Keywords: Semiparametric partial linear regression model , Semiparametric generalized least squares estimators , least squares estimators , Autocorrelation problem.
2010 MSC: 49J20, 49K20, 35Q35, 93C20.

1. Introduction

The problem of autocorrelation appears in most studies that take form time series data as well as research that depends on Cross- Section data , especially cross section data that take the form grouping

Email addresses: hasanhopop@utq.edu.iq (Hassan .H.Razaq), haider.raid@uos.edu.iq (Hayder .R. Talib), Reem.t@copew.uobaghdad.edu.iq (Reem .T. Kamil)

Received September 18, 2023; Accepted: November 24, 2023; Online: December 15, 2023.

of observations ,this problem may arise as a result of deleting some independent variable from the studied relationship which as a result of inaccurate diagnosis of the relation between response variable and the independent variables . In addition such a problem may arise as of making adjustments in the data or resorting to estimating the values of some observations based on the values of other observers, the processes of adjustment and estimation usually depend on taking the averages of the values of successive observations , which creates a relationship between the errors of those observations and thus effects the nature of their distribution. Parametric models have two parts , the first represents the random error term which is a random component , the second is a non-random part represents by the dependent variable or the response variable which is related to an independent variable or several independent variables . The dependent variable is the one to be estimated by one of the parametric or nonparametric methods . Parametric models are more common as they assume that the dependent variable has a functional formula which was obtained from pervious information about the regression function .Although it is one of the most important tools for analyzing data and estimating parametric because it is highly effective in interpreting the results , it was not able to be sufficient in some cases .As for the nonparametric regression model it focuses on estimating the regression function directly from the data and does require special assumptions to estimate the parameters of the linear model. Thus nonparametric regression is more flexible to detect data that may be missing or in some cases where no prior information is available about the data. Semiparametric regression models are constructed by mixing parametric and nonparametric models . The semiparametric models is often used when the parametric hypotheses are not specified and do not achieve consistency property or the nonparametric models works incompletely.

2. Semiparametric Partial Linear Model

The semi parametric partial linear regression model(PLM) is one of the most important and most widely used semi parametric models in the field of economics , medicine and environmental studies .This model has other names as it is called partial parametric model and because belong to one of the categories of the partial spline it is called partial spline model [3] . Partial linear model is semi parametric model since it contain both parametric and nonparametric components , it allows easier interpretation of the effect of each variable and many be preferred to a completely nonparametric regression because of the well-known curse of dimensionality [5] . This model (PLM) has been extensively studied by researchers and estimation it in many different ways such as Robinson estimator , spline estimator , kernel estimator , weighted partial spline ,difference based estimator and various other ways. Here , we will review the most important research and studies that relied on estimating the model (PLM) . [2] they estimation a model (PLM) by using semi-parametric two – step strategy , in the first step they used the nonparametric regression estimator (kernel) to estimate the nonparametric component , while the second step included weighted instrumental variables method to estimate the parametric component . [6] presented a paper that included the nearest neighbor estimator to estimate nonparametric part , as for the parametric part it was estimated using least square and they used data from a partial sample of the National Longitudinal Study (NLSY) for the year 1979 and the variable included earning , age and education . You , [**?**] they made a comparison between two semi parametric ordinary least squares method and the semiparametric general least squares method foe estimating the parametric component of the (PLM) model , as for the nonparametric component it was estimated using B- spline series and they concluded through the simulation results that the semiparametric general least squares estimators are more efficient than the ordinary least squares estimators . [10] they presented a paper that included a difference based estimator (DBE) to estimate the parametric component while the nonparametric component was estimated using kernel estimator , they used the price data for the year (1977) using five independent variables and one nonparametric variable , they concluded that the estimation of the parametric component is an efficient asymptotic estimator while the estimator of the nonparametric components is an ideally aligned estimator . [8] They estimated a model (PLM) using the double penalized least square method by using smooth spline to estimate the nonparametric component first and applying the shrinkage penalty to the parametric part secondly to achieve the estimation model ,they explained that

proposed estimator (DPLS) can be as effective as the Oracle estimator and they also studied and proved that alignment properties of the estimators when the effects of number of parameters diverge with the sample size . [4] has estimated the (PLM) model in the presence of the problem of Heteroscedasticity by using the weighted wavelet method to estimate parametric component and wavelet smooth method for estimating the nonparametric component , he also studied and demonstrated the alignment properties of the estimator under appropriate assumptions and proved that the wavelet estimator has weak consistent rates.

A partial linear model is given as follow:

$$Y_i = X_i'\beta + g\left(t_i\right) + \varepsilon_i \quad , \quad i = 1, 2, \ldots, n \tag{1}$$

Where $\mathrm{Y_i}$ are response , $X_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)'$, $t_i = \left(t_{i1}, t_{2i}, \ldots, t_{id}\right)$ are vectors of variable $(\mathrm{X_i}, \mathrm{t_i})$ are independent and identically distributed (i.i.d), $\beta = \left(\beta_1, \beta_2, \ldots, \beta_p\right)'$ is a vector of unknown parameters, $g(.)$ is a unknown function or smooth function, parameter vector and nonparametric function to be estimated, and the ε_i are unobservable random errors.

3. Estimation Methods

3.1. Semiparametric Generalized Least Squares Estimators

One of the basic assumptions that were relied upon in estimating parameter of the linear model is the lack of autocorrelation between the errors of observations in the sample, in other words :

$$\mathrm{E}\left(\varepsilon_t \varepsilon_{t-s}\right) = 0 \quad , \quad t = 1, 2, \ldots, n \tag{2}$$

But if the economic or social phenomenon includes a self-correlation between the errors of the studied observations, assumption (2) become as follows :

$$\mathrm{E}\left(\varepsilon_t \varepsilon_{t-s}\right) \neq 0 \tag{3}$$

We will assume that the distribution of errors in model (1) follows a first order autocorrelation :

$$\varepsilon_t = \rho \varepsilon_{t-1} + e_i \tag{4}$$

where ρ is unknown $((|\rho| > 0)$ and e_i it is Gaussian distribution with a zero mean and variance σ_ε^2. Assume that $w_{ni}(t) = \{w_{ni}\left(t_i, T_1, T_2, \ldots T_n\right\}$ where $W_{ni}(t)$ are the positive weighted function, for every given parameter β we define a nonparametric component estimator of g(.) is given by :

$$\hat{g}(t, \beta) = \sum_{i=1}^{n} W_{ni}(t)\left(Y_i - X_i'\beta\right) \tag{5}$$

In compensation $\hat{g}(t, \beta)$ into model (1), we get $Y_i = X_i'\beta + \hat{g}(t, \beta) + \varepsilon_i$, which can be rewritten as follows :

$$\tilde{Y}_i = \widetilde{X}_i'\beta + \widetilde{\varepsilon}_i \tag{6}$$

Where:

$$\widetilde{Y}_i = Y_i - \sum_{i=1}^{n} W_{ni}\left(t_i\right) Y_i \quad , \quad \widetilde{X}_i = X_i - \sum_{i=1}^{n} W_{ni}\left(t_i\right) X_i$$

$$\widetilde{\varepsilon}_i = g\left(t_i\right) - \sum_{i=1}^{n} W_{ni}\left(t_i\right) g\left(t_i\right) + \varepsilon_i - \sum_{i=1}^{n} W_{nj}\left(t_i\right) \varepsilon_i.$$

$W_{ni}(t)$ it is the decreasing distance function suggested by Nadarya - Watson 1964 it is calculated by the following formula [9] :

$$W_{ni}(t) = \frac{K\left(\frac{t-T_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{t-T_i}{h}\right)}$$

Where $W_{ni}(t)_{i=1}^{n}$ represents the weight series whose sum is equal to one, K(.) is kernel function and h_n represents bandwidth parameter. By writing equation (6) in matrix form:

$$\widetilde{Y} = \widetilde{X}\beta + \widetilde{\varepsilon} \tag{7}$$

Where $\tilde{Y} = \left(\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_n\right)', \tilde{X} = \left(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n\right)', \widetilde{\varepsilon} = (\widetilde{\varepsilon}_1, \widetilde{\varepsilon}_2, \ldots, \widetilde{\varepsilon}_n)'$ From (7) a semi parametric least squares estimators for parameter component is:

$$\hat{\beta}_{SLSE} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\tilde{Y} \tag{8}$$

Substituting (8) into (6) the estimated residuals can be obtained as :

$$\widetilde{\varepsilon}_i = \widetilde{Y}_i - \widetilde{X}_i\hat{\beta}_{SLSE} \tag{9}$$

Therefore, the autocorrelation coefficient can be estimated as follows:

$$\hat{\rho}_n = \left(\sum_{i=1}^{n} \widetilde{\varepsilon}_i^2\right)^{-1} \sum_{i=1}^{n-1} \widetilde{\varepsilon}_{i+1}\widetilde{\varepsilon}_i \tag{10}$$

By (4) we have :

$$E\left(\varepsilon\varepsilon'\right) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{bmatrix} = \sigma^2\Omega$$

$$\therefore \Omega^{-1} = \frac{adj\Omega}{|\Omega|} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho & 0 & 0 & 0 & \ldots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & 0 & \ldots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 & \ldots & 0 & 0 \\ 0 & 0 & -\rho & 1+\rho^2 & -\rho & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & 0 & \ldots & -\rho & 1 \end{bmatrix}$$

Under the assumption of the autocorrelation, the sum of square errors can be put into the model (7) in the following form :

$$\widetilde{\varepsilon}'\Omega^{-1}\widetilde{\varepsilon} = (\tilde{Y} - \widetilde{X}\beta)'\Omega^{-1}(\widetilde{Y} - \widetilde{X}\beta) \tag{11}$$

By taking the first partial derivative with respect to the parameter vector to be estimated, we get :

$$\hat{\beta}_{SGLSE} = \left(\tilde{X}'\Omega^{-1}\tilde{X}\right)^{-1} \tilde{X}'\Omega^{-1}\tilde{Y} \tag{12}$$

As for the non-parametric part , it can be estimated by substituting the parametric component in the equation (5) and it is Nadarya – Watson estimator :

$$\hat{g}(t,\beta) = \sum_{i=1}^{n} W_{ni}(t)\left(Y_i - X_i'\hat{\beta}_{\mathrm{SGLSE}}\right)$$

$$\hat{g}(t,\beta) = \frac{\sum_{i=1}^{n} K\left(\frac{t-T_i}{h}\right)\left(Y_i - X_i'\hat{\beta}_{\mathrm{SGLSE}}\right)}{\sum_{i=1}^{n} K\left(\frac{t-T_i}{h}\right)} \tag{13}$$

3.2. Least Squares Estimators Method

By transformation method for model (6) the form will be as follows [7]:

$$\tilde{Y}_i - \rho\tilde{Y}_i = \beta\left(\tilde{X}_i - \rho\tilde{X}_i\right) + \tilde{\varepsilon}_i - \rho\tilde{\varepsilon}_i \tag{14}$$

The above model can be written as follows:

$$Y_i^* = X_i^*\beta + e_i \tag{15}$$

Write (15) in matrix form as:

$$Y^* = X^*\beta + E \tag{16}$$

We get least squares estimates by minimizing the sum of the squares of the random error as follow:

$$\begin{aligned}\min \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left(Y_i^* - X_i^*\beta\right)^2 \\ &= \min\left(E'E\right) = \left(Y^* - X^*\beta\right)'\left(Y^* - X^*\beta\right)\end{aligned}$$

By taking the first partial derivative of the parameter vector to be estimated, we get :

$$\hat{\beta}_{\text{LSEM}} = \left(X^{*'}X^*\right)^{-1} X^{*'}Y^* \tag{17}$$

As for the non-parametric part, it is estimated according to the previous method replacing the parameter part only, we get :

$$\hat{g}(t) = \sum_{i=1}^{n} W_{ni}(t)\left(Y_i^* - X_i^*\hat{\beta}_{\text{SLEM}}\right) \tag{18}$$

Sins the random errors are distributed normally with mean zero and constant variance σ_ε^2, and define the estimator of σ_ε^2 by:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left(X_i'\hat{\beta} + \hat{g}\left(t_i\right)\right)^2\right. \tag{19}$$

Noting the replacement of the parameter and non-parameter component in (19) according to the estimation methods mentioned previously.

4. Simulation Studies

Simulation experiments were carried out using four sample size 20, 30, 60, 100 , and replicates 1000 for each simulation experiments. We achieved the model $y_i = x_i\beta + g\left(t_i\right) + \varepsilon_i$ with $\varepsilon_i = \rho\varepsilon_{i-1} + e_i, \mathrm{i} = 1, 2, \ldots, \mathrm{n}$, where $g\left(t_i\right) = \sin\left(t_i\right)$, $\beta = (1, 5)'$, the error distribution generated from standard normal distribution N(0, 1).For the autoregressive coefficient we considered three cases ($\rho = 0.2$, $\rho = 0.4$, $\rho = 0.6$). The independent variable x_i, t_i are generated from uniform distribution with $(0, 1)$.

For the purpose of making a comparison between the estimation methods, We first calculated the parametric component β and nonparametric component $g\left(t_i\right)$ using both SGLSE and LSEM .For the weighted function, we use the Gaussian kernel for calculated K(.) :

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-u^2\right) \quad , \quad u \in (-\infty, \infty)$$

The bandwidth parameter (h) is selected by using Cross -Validation (CV):

$$CV_h = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left(X_i'\hat{\beta} + \hat{g}\left(t_i\right)\right)^2\right.$$

$$\therefore \hat{h} = \arg\cdot\min\left(CV_h\right)$$

5. Simulation Results

In this section we calculated the mean squares of error for SGLSE and LSEM for the purpose of comparing the two methods , the results were as in the table below :

Table 1: The mean squares of error(MSE) for SCLSE, LSME.

n	Methods	ρ		
		0.2	0.4	0.6
20	SGLSE	0.0099	0.0100	0.0095
	LSEM	0.0328	0.02436	0.0235
30	SGLSE	0.0412	0.0493	0.0554
	LSEM	0.0420	0.0558	0.0541
60	SGLSE	0.0249	0.0267	0.0397
	LSEM	0.0313	0.0461	0.0549
100	SGLSE	0.9428	0.9613	0.9430
	LSEM	0.9756	0.9903	0.9660

6. Summary and Conclusion

Through the simulation results showed that semiparametric generalized least squares estimators (SGLSE) is better than the least squares estimators method (LSEM) because it has the least mean squares of error for all sample size . This study can be made as a basis for expanded future studies , in the event that there is a problem of heterogeneity of error variance or in the event of a linear multiplicity between the explanatory variables. Expand the presentation for other semiparametric model methods that are not used in this research such as Semiparametric binary response model and the semiparametric single – index model .

References

[1] Ahn,H. and Powell.j. (1993).” Semi parametric estimation of censored selection models with a nonparametric selection mechanism”. Journal of Econometrics, vol.58, pp.2-29.
[2] Chen, H.,(1988).” Convergence rates for parametric components in a partially Linear Models”. Annals of Statistics, 16, , 136-146.
[3] Cheng,W.,(2012).”Weighted Wavelet Estimate in Semi parametric Models with Heteroscedastic Errors.” School of Mathematics and Statistic, Hubei Normal University, Huangsh,China,435002. cwj113ok@163.com.
[4] Hardle W., Liang H., and Gao J.(2000).” Partially linear models”. Physica-Verlag, Heidelberg.
[5] Imbens,G. and Porter,j.(2005).” Bias- adjusted Nearest Neighbor Estimation For The Partial Linear Model.” UC Berkeley and NBER, University Of Wisconsin- Madison.
[6] Kadhim ,A.H and Muslim B.S (2002). Advance Econometrics measurements theory and practice .
[7] Ni,X.,Zhang,H.H.,and Zhang, D.,(2007).” Automatic Model Selection For Partially Linear Models.” xni@stat.ncsu.edu, hzhang2@stat.ncsu.edu, zhang@stat.ncsu.edu , Department of Statistics, North Carolina State University.
[8] Speckman, P. (1988)”. Kernel Smoothing in Partially Linear Models”. Journal of Royal Statistical Society-B 50, 413-436.
[9] Wang,L. Brown,L.D. and Cai, T.(2006).”A difference Based Approach to Semi parametric Partial Linear Model.” Depatment of Statistics, The Wharton School University of Pennsylvania.
[10] You,j. and Zhou.X.(2005).”Bootstrap Of Semi parametric Partially Linear Model With Autoregressive Errors.” The Hong Kong Palytechnic University Statistics Sinica 15,117-113.