International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 4
ISSN: 2088-5334

Wavelet Estimation of Semi-parametric Regression Model

Ahmed Shaker Mohammed Tahir[a,1], Firas Monther Jassim[a]

[a] Statistics Department, College of Administration & Economics, Mustansiriyah University, Iraq
E-mail: [1]ahmutwali@uomustansiriyah.edu.iq

***Abstract*— The semi-parametric regression model combines parametric and nonparametric regression. However, non-parametric estimation may provide flexible solutions to the problems suffers by the regression model, but the problem of dimensionality that this estimator suffers, which occurs due to the increasing number of explanatory variables, still remain, this, in turn, may reduce the accuracy of the estimation process. Estimate the non-parametric part of the semi-parametric models that can be studied using conventional non-parametric methods such as the Spline Smoothing and Kernel Smoothing. However, there are other non-parametric methods that can be used, therefore, in this paper, the semi-parametric regression model was estimated by employing the wavelet estimate for the soft threshold, according to the "Speckman" method, and then comparing it with the two methods, Nadaraya-Watson and Local Linear, through the implementation of simulation experiments that included different sample sizes and threshold values. The parametric part estimation of the partially linear model according to the least-squares method was not identical to those estimates using the Speckman method, that is because the least-squares method was not appropriate for the uneven nature of the number of weekly work hours. Simulation experiments have demonstrated the efficiency of the wavelet estimation method and its superiority over other methods. The above estimation methods were applied to real data related to the study of the production value for the public industrial sector in Iraq, and some factors affect it, such as the value of industrial supplies, the total wages of workers, and the number of workers.**

***Keywords*— partially linear model; wavelet estimate; speckman method; nadaraya-watson smoothing; local linear smoothing.**

I. INTRODUCTION

Partially linear model (PLM) is a semi-parametric model which have a linear part represented by the parametric regression [1], [2], and a non-linear part represented by the non-parametric regression as a smooth function [3]. This model described mathematically with details in section 2. The importance of this model lies into two reasons, first reason is that it is more flexible than the parametric model because it combines both parametric and non-parametric components. Second reason is that it allows easier interpretation of the effect of each variable compared to a non-parametric regression that leads to overcoming the curse of dimensionality.

Numerous researches have been prepared in this aspect based on different smoothing methods and techniques, among them [4], in which spline smoothing was used, piecewise polynomials method [5], a local linear method [6], [7], based on profile least square method. In the preceding techniques, conditions for the function m(z) such as continuity or continuity of its derivatives may not be satisfied in some areas like economic series with discontinuous points, image processing, and signal. Wavelet technique is an active natural extension of the various non-parametric methods due to its adaptable ranges of unknown smoothness. Furthermore, it has many advantages practical, fast, and dull due to efficient algorithms [8].

In this paper, we aimed to find the best smoothing technique that can be applied under the Speckman method for estimating the partially linear model. To satisfy this aim, we have compared three different smoothing techniques: Nadaraya-Watson (NW), local linear (LL), and wavelet by using simulation. Furthermore, we have applied the three mentioned techniques to real data about the production value and some factors affecting it for the public industrial sector in Iraq [9], [10].

The paper is arranged as follows; in section 2 we described with details the mathematical formula of the partially linear model, Nadaraya-Watson and Local linear smoothing were shown in section 3, section 4 deals with the wavelet transformation, a simulation experiment was summarized in section 5, in section 6 an applied study was summarized using the estimator methods under research. Finally, section 7 shows the conclusions.

II. MATERIALS AND METHOD

The relation between the dependent variable and the explanatory variables can be described in the partially linear model

A. Wavelet Transformation

Wavelet transformation (WT) is a mathematical extension tool for Fourier method; it is one of the most advanced transformations in the field of signal processing, this transformation enables us to analyze the signal into a set of multiple levels solutions (Multiresolution) in both time and frequency [12]. The mechanism of (WT) can be summarized by using a variate width window to obtain the frequency changes throughout the wavelength. This variate window produces a limited length signal with zero average value called wavelet. The produced wavelet is compressed with two functions, the first is called mother function to get a set of coefficients called detailed coefficients, and the second is the measurement function (also called father function) to get the approximation coefficients. There are two types of (WT), Continuous Wavelet Transformation (CWT), and Discrete Wavelet Transformation (DWT). In this paper, we used the (DWT), [13], [14].

B. Wavelet Shrinkage

Wavelet shrinkage is a way to reduce the signal noise, proposed a non-linear wavelet estimator for non-parametric function by reconstructed wavelet coefficients and scaling coefficients. The wavelet reduced by the threshold to transform the low-frequency signal to zero and keep the high-frequency signal close to zero. Mainly, there are two threshold types [15], [16]:

1) Hard threshold: employed to reduce the wavelet coefficients that are smaller than the threshold value to zero, and keep the values that are greater than the threshold.

2) Soft threshold function: different from the hard threshold by shrinking the values of the wavelet coefficients that are higher than the threshold. Threshold value must be chosen carefully, since the larger threshold value is caused by fuzzy transformation, and the smaller threshold value leads to no noise reduction. There are several methods to find the threshold value, among them the universal threshold [17], [18].

III. Result and Discussion

A. Discrete Wavelet Transformation (DWT)

The Discrete Wavelet Transformation is a linear process performed on noisy data through two filters, the low-frequency filters (scaling filter) and high-frequency filters (wavelet filter). The main points of the (DWT) can be described through Daubechies theorem. It is summarized by finding an accurate formula for the non-parametric function m(z) which is produced from both scale (low-frequency filter $\tilde{g}$) ω(z) and wavelet (high-frequency filter $\tilde{h}$) ∅(z) functions. It is done based on the vanishing moments which gives the approximation properties of wavelengths for these two functions, where many the vanishing moments will give better approximation functions. Also, the estimation that basis on a specified number of non-zero coefficients is better than the estimation for all coefficients [19][20], introduced a fast algorithm for (DWT) require the sample size n to be 2^J for some integer J, That is, double filters, Thus we start with function $\emptyset_0$ (z) until reaching to the $\emptyset_n$ (z) as in the following formula, figure (1) shows the quick wavelength transform start with the scaling,[12].

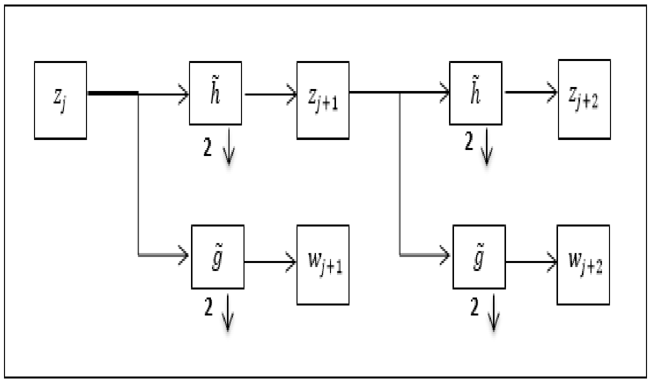

Fig.1 The Filters with Mallat DWT

B. Simulation

Figure (2) shown these m(z)^, s functions. For the DWT we based on the Daubechies orthonormal compactly supported wavelet with 8 vanishing moments, also two threshold rules universal and cross-validation were carried out. Note that in the wavelets simulation both signal and noise must be measured at the same or equivalent points in a system, and within the same system bandwidth, so we select σ^2 to satisfy a fixed signal-to-noise ratio (SNR), it is merely the ratio of the sample standard deviation of the signal to the standard deviation of the added noise.

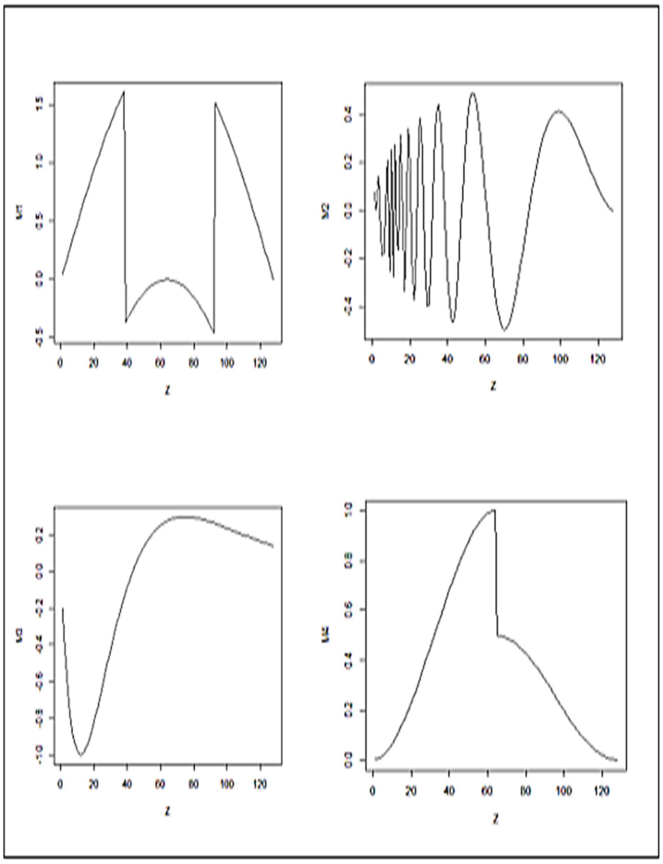

Fig. 2 Nonparametric test functions m(z)^,s

TABLE I
ROOT SQUARE ERROR (RMSE) FOR ESTIMATED PLM'S

M	n	SP_{NWE}	SP_{NWG}	SNR			
				3		6	
				SP_{UV}	SP_{CV}	SP_{UV}	SP_{CV}
M_1	128	0.3556	0.4774	0.2234	0.1558	0.1417	0.1272
	256	0.2986	0.4085	0.1828	0.1158	0.1147	0.0704
	512	0.2488	0.3381	0.1432	0.0936	0.0893	0.0547
M_2	128	0.2073	0.2524	0.1265	0.1043	0.0918	0.0910
	256	0.1765	0.2252	0.1043	0.0729	0.0699	0.0611
	512	0.1480	0.1948	0.0805	0.0522	0.0545	0.0342
M_3	128	**0.1951**	0.2236	0.0879	0.0854	0.0678	0.0580
	256	**0.1561**	0.1907	0.0710	0.0676	0.0480	0.0429
	512	**0.1211**	0.1532	0.0531	0.0483	0.0346	0.0312
M_4	128	**0.0881**	0.1195	0.0518	0.0466	0.0349	0.0291
	256	**0.0727**	0.0899	0.0452	0.0377	0.0298	0.0231
	512	**0.0604**	0.0695	0.0401	0.0316	0.0254	0.0187

Through table (1) above, for all test functions and sample sizes, we found that the estimated values of RMSE for the wavelets smoothing were lower than of those values for the kernel smoothing, and its clearly that the lowest of those values were in the case of function M_4. Also we can notice that Speckman with cross-validation threshold (〖SP〗_CV) had the lowest values for RMSE, followed by the Speckmea with universal threshold (〖SP〗_UV). Whereas for the kernel smoothing at the M_1 and M_2 functions we can say that the values of RMSE according to the 〖SP〗_LLRE were lower than these values for the 〖SP〗_NWE, while at cases of M_3 and M_4 the 〖SP〗_NWE was presented lower estimates for RMSE. Note that the RMSE values for all experiments decrease with increasing sample size.

Furthermore, the RMSE values had decrease behavior with the increasing of SNR values. Figure (3) is represented the real and estimated wavelets curves for the best test function M_4 at n=128,SNR=3,9, we can be seen through it a little differences in smooth between UV,CV thresholds rules when SNR=9, but at SNR=3 we can clearly noticed the preference smoothing of the CV especially in the top and in the right lower in each curve. Through figure (4) which showing comparison between the real M_4 curves and their kernel estimators at n=128, the real and estimated differences are expanded, but overall the M ̂_NWE, M ̂_LLRG are looking nearly to the real curve, while the smoothing looking far in the left bottom, top and in the middle of the right side in each curve.

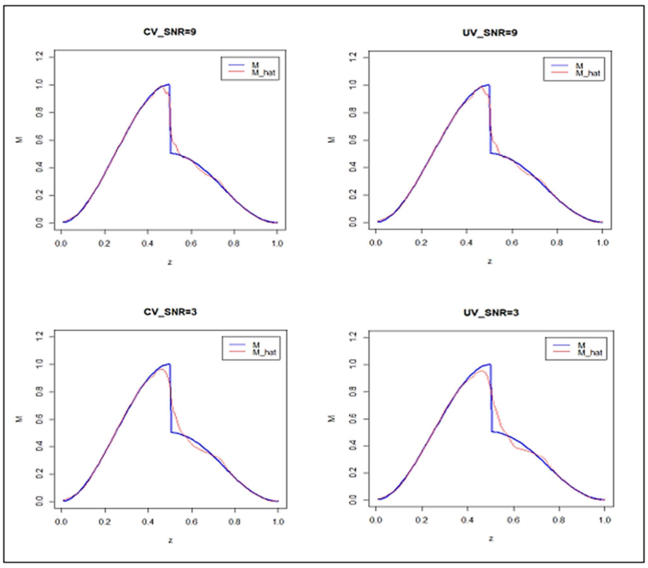

Fig.3 Wavelets smoothing for M_4 function when n=128, SNR=3,9

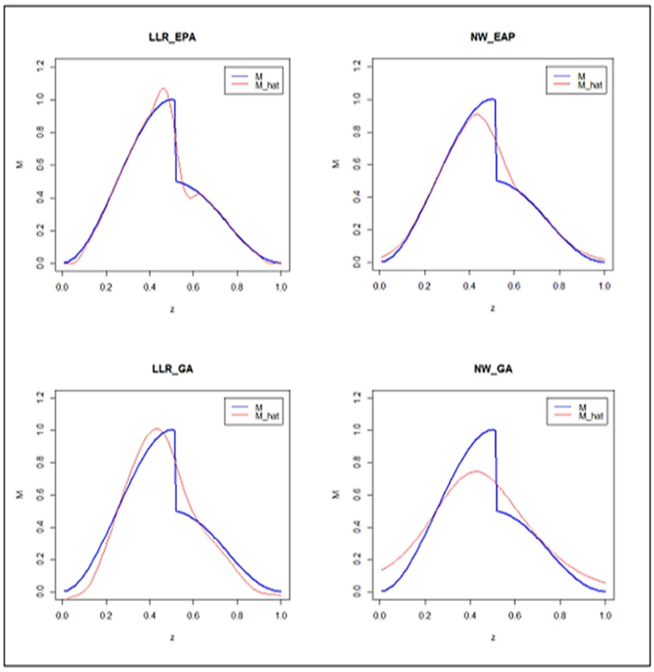

Fig.4 LL & NW smoothing for M_4 function when n=128

C. Application

In this part, we used the PLM in modeling the relationships between the industrial production value for the public sector companies in Iraq, as a response variable (Yi), and three explanatory variables, The value of production inputs (X1i), the number of employees (X2i), and the number of weekly working hours (Zi). Note that the two variables X1i and X2i represented the parametric part and, Zi represent the nonparametric part. Depending on the real data about the variables under research we applied the two wavelets shrinkage approaches (UV,CV). All the variables were transforming into standardize form because they were measured in different units. To estimate the model, we must first find the preliminary estimates for β_1, β_2 by using ordinary least square, we found that β_1=-0.116, β_2=0.113.

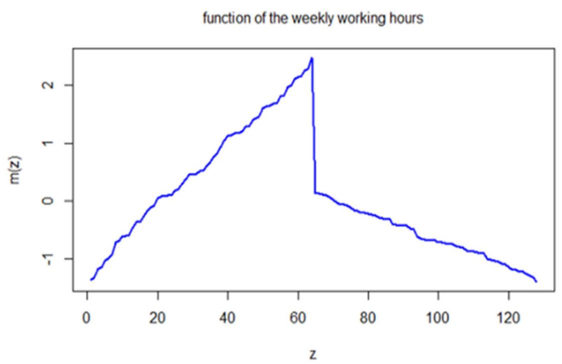

Fig.5 Function of the weekly working hours

Then by using M_4 test function and each 〖SP〗_CV, 〖SP〗_UV, 〖SP〗_NWE and 〖SP〗_LLG, thus we get the following parametric estimators for PLM in table (2). Also we plotted the mˆ〖(z_i)〗^,s for the four Speckman estimators as in figure (6), which show us that 〖m ̂(z_i)〗_CV gives a very close smoothing to the real test function, followed respectively by 〖m ̂(z_i)〗_UV, 〖m ̂(z_i)〗_NWE and 〖m ̂(z_i)〗_LLG.

TABLE II
SPECKMAN PLM ESTIMATION

Method	β_1	β_2
SP_{CV}	-0.1018	0.0959
SP_{UV}	-0.1131	0.0912
SP_{NWE}	-0.0822	0.1246
SP_{LLG}	-0.1054	0.1657

The four 〖PLM〗^,s Speckman estimators of the industrial production variable are explained in figure (7), those estimators plots refers to a great match between 〖SP〗_CV and real variable, this matching is declines gradually according to the remaining three estimation methods as following order 〖SP〗_UV, 〖SP〗_NWE and 〖SP〗_LLG, where 〖SP〗_NWE and 〖SP〗_LLG are gives noisy smooth curves.

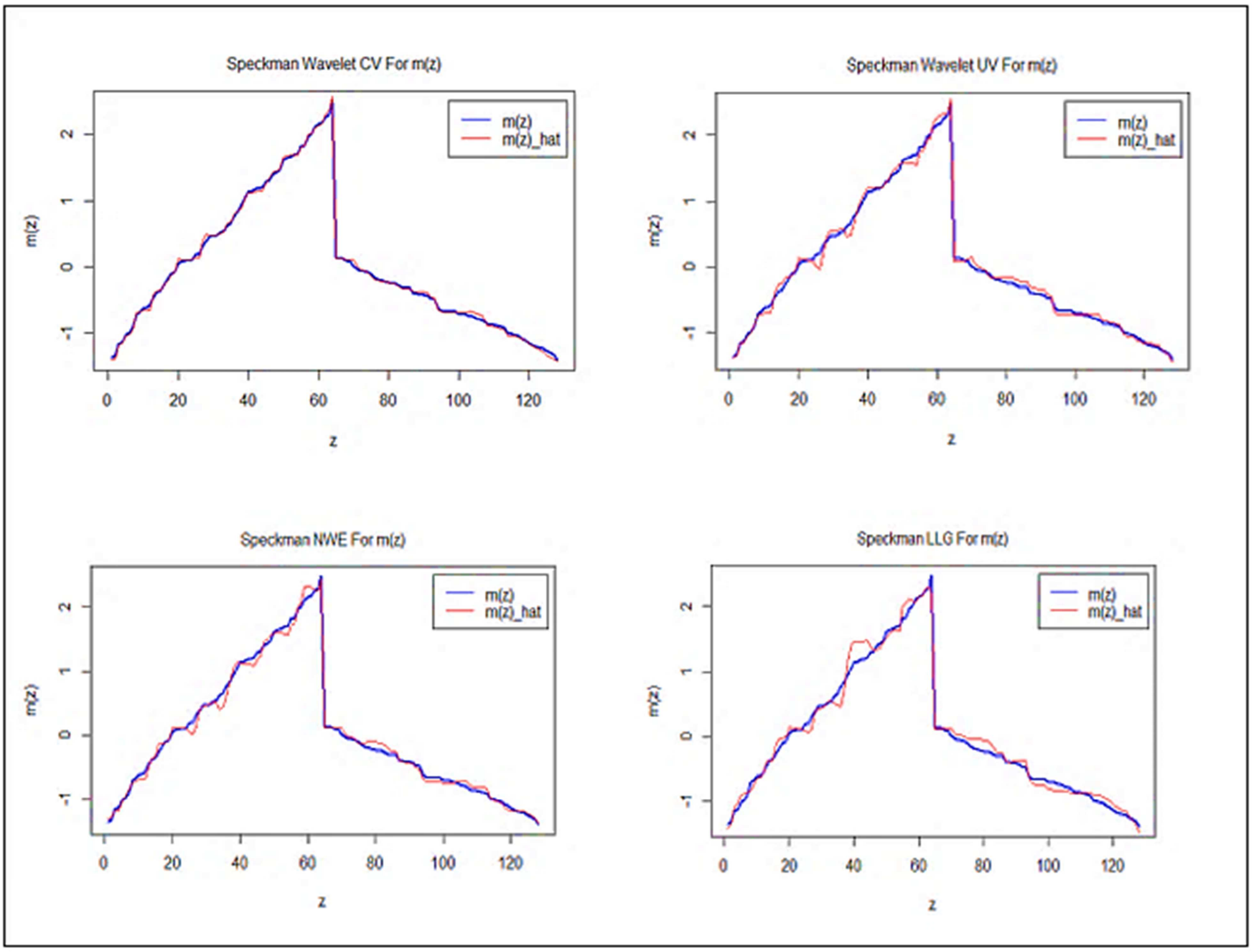

Fig.6 Weekly working hours

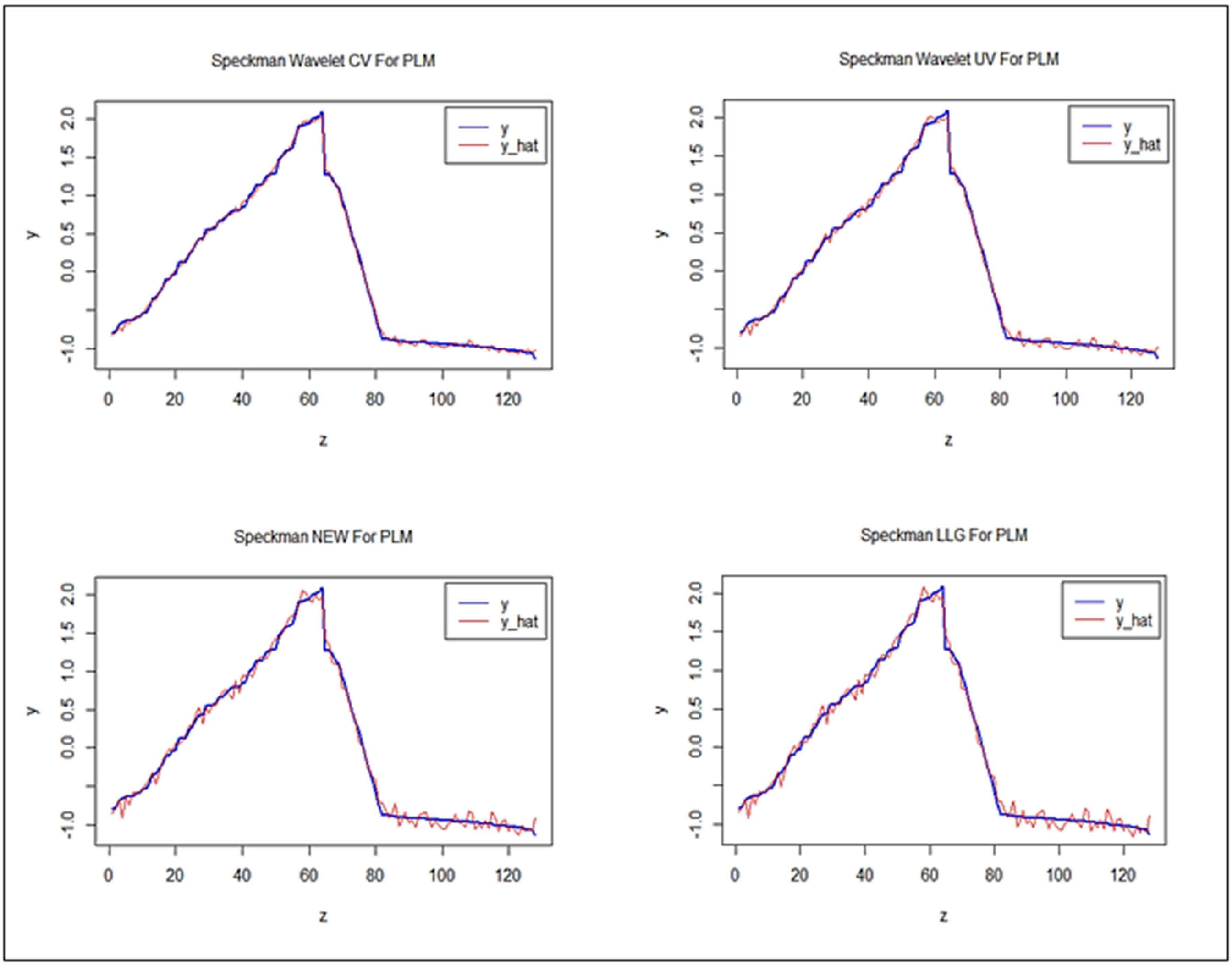

Fig.7 Speckman estimators for the industrial production value

IV. CONCLUSIONS

The wavelet smoothing approach for the partially linear model works well under suitable signal to noise ratios. The piecewise polynomial test function gives an accurate description of asymmetric spaces data. The non-parametric variables have a most considerable role in the estimating of a partially linear model the nonlinear functions have the primary effect that is controlling the nature of the response. In estimating the partially linear model specifically is the cause of the case of nonlinear functions that did not belong to the wavelet family. Nadaraya-Watson method and the Gaussian kernel function respectively should provide more accurate estimates than the local polynomial method, vice versa in the case of signal functions that belong to the wavelet family, that is, the local polynomial regression and Epanechnicov kernel function give more efficient estimates of Ndaraya-Watson and the Gaussian kernel function respectively.

The behavior of the industrial production value for the studied public sector companies in Iraq have an upward slope in the first weeks; then it declined sharply in the middle of the period, then to be stable at the lowest level, this is the reason that prompted to use the piecewise polynomial which depends on the dividing the period into three individually modeled regions. The results of the estimation showed the negative effect on the value of production inputs, while the positive effect of the number of workers on the value of production. The parametric part estimation of the partially linear model according to the least-squares method was not identical to those estimates using the Speckman method, that is because the least-squares method was not appropriate for the uneven nature of the number of weekly work hours.

REFERENCES

[1] S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] Khyade, V. B., & Eigen, M. (2019). Key Role of Statistics for the Fortification of Concepts in Agricultural Studies. International Academic Journal of Psychology and Educational Studies, 6(1), 20-34.

[3] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., Yidris, N., & Fattahi, A. (2020). Experimental and numerical investigation of the mechanical behavior of full-scale wooden cross arm in the transmission towers in terms of load-deflection test. Journal of Materials Research and Technology, 9(4), 7937-7946.

[4] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., & Yidris, N. (2020). Conceptual design of the cross-arm for the application in the transmission towers by using TRIZ–morphological chart–ANP

methods. Journal of Materials Research and Technology, 9(4), 9182-9188.

[5] Johari, A. N., Ishak, M. R., Leman, Z., Yusoff, M. Z. M., Asyraf, M. R. M., Ashraf, W., & Sharaf, H. K. (2019). Fabrication and cut-in speed enhancement of savonius vertical axis wind turbine (SVAWT) with hinged blade using fiberglass composites. In Proceedings of the Seminar Enau Kebangsaan (pp. 978-983).

[6] Asyraf, M. R. M., Ishak, M. R., Sapuan, S. M., Yidris, N., Johari, A. N., Ashraf, W., ... & Mazlan, R. (2019). Creep test rig for full-scale composite cr ossarm: simulation modelling and analysis. In Seminar Enau Kebangsaan (pp. 34-38).

[7] Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., ... & Suveges, D. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research, 47(D1), D1005-D1012.

[8] Zhang, Y., Qi, G., Park, J. H., & Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nature genetics, 50(9), 1318-1326.

[9] Weissbrod, O., Flint, J., & Rosset, S. (2018). Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. The American Journal of Human Genetics, 103(1), 89-99.

[10] Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. The annals of applied statistics, 11(4), 2027.

[11] Cichonska, A., Rousu, J., Marttinen, P., Kangas, A. J., Soininen, P., Lehtimäki, T., ... & Ripatti, S. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics, 32(13), 1981-1989.

[12] Mahmoudi, M. R., & Abbasalizadeh, A. (2019). How statistics and text mining can be applied to literary studies?. Digital Scholarship in the Humanities, 34(3), 536-541.

[13] Shelke, M., Deshpande, S. S., & Sharma, S. (2020). Quinquennial Review of Progress in Degradation Studies and Impurity Profiling: An Instrumental Perspective Statistics. Critical Reviews in Analytical Chemistry, 50(3), 226-253.

[14] Worster, A., & Haines, T. (2004). Advanced statistics: understanding medical record review (MRR) studies. Academic Emergency Medicine, 11(2), 187-192.

[15] Perer, A., & Shneiderman, B. (2008, April). Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 265-274).

[16] Brazzale, A. R., Davison, A. C., & Reid, N. (2007). Applied asymptotics: case studies in small-sample statistics (Vol. 23). Cambridge University Press.

[17] Schaid, D. J., & Sommer, S. S. (1994). Comparison of statistics for candidate-gene association studies using cases and parents. American journal of human genetics, 55(2), 402.

[18] Morellato, L. P. C., Alberti, L. F., & Hudson, I. L. (2010). Applications of circular statistics in plant phenology: a case studies approach. In Phenological research (pp. 339-359). Springer, Dordrecht.

[19] Holey, E. A., Feeley, J. L., Dixon, J., & Whittaker, V. J. (2007). An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. BMC medical research methodology, 7(1), 52.

[20] Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., ... & Suveges, D. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research, 47(D1), D1005-D1012.