Using A Semi-Parametric Regression Model to Estimate A Function of The Quantity of Date Production In Iraq

Kareem K. Aazer [*1]

[1] Department of Statistics, University of Sumer, Iraq
Emails: Kareemalataby28@gmail.com

Abstract

The semi-parametric referral model is one of the important developments that used the analysis of their independent effect on other variables, which produces prediction issues. It is known that the semi-parametric referral model combines direct referral models, whose variables are linear, with analmic conversion models, whose variables are non-linear. In this, it was done. The research presents the production function of the quantity of dates, which is affected by the multiplicities. Some of them control parameters such as heart rate and settings of fruiting palm trees, and some of them behave nonlinearly, such as humidity, temperature, wind, and others, and we take the variable to determine the air temperature. It has been observed that the mean square error of the semi-parametric regression model is less than the mean square error of the parametric regression model, which assumes that all variables behave linearly. This proves the validity of the results in the case of using sample sizes for a set of data generated through simulation experiments, which showed that the regression models are semi-parametric.

Keywords: Parametric regression; nonparametric regression; semi-parametric regression; date production quantity.

1. Introduction

The Semiparametric Regression method has gained wide popularity recently, due to its advantage in combining parametric regression functions with nonparametric regression functions in one, which is a feature that made the new model bypass the problem of dimensionality in the case of completely nonparametric models, and also provided a larger environment. The application is one of those in the case of parametric regression models because the latter may be affected by some independent (explanatory) variables that do not have a known parametric distribution. Also, sometimes the function under study may not be fully represented because some of the variables behave parametrically and others Non-parametrically. Also, there are often doubts about the homogeneity of error variances, especially in those cases where variables are represented spatially or temporally, in addition to not knowing the distribution of errors (the error distribution is unknown). The semi-parametric model has two cases, the first is called the partial linear model, which depends on Both the parametric regression model and the nonparametric regression model. In the second case, the regression model is called the combined regression model, which combines a parametric regression function with a nonparametric regression function in the presence of a merging parameter, which gives added weight to both the parametric and nonparametric models. The combined model depends on the value of the merging parameter. It is between zero and one, which determines whether the model is correct, incorrect, or close to correct [1,2].

Doi: https://doi.org/10.54216/PMTCS.030104
Received: June 19, 2023 Revised: September 25, 2023 Accepted: December 27, 2023

2. The theoretical side [3–6]

The semi-parametric regression model, which was proposed by researchers Robinson in 1988 and Paul Speckman in 1988, is a combination of the parametric regression function with the nonparametric regression function. Therefore, it is a special case of the additive regression model. Because it consists of two functions, which is called the partial linear regression function. The explanatory variables in the parametric particle are binary or discrete explanatory variables, and they are often referred to as 〖X′〗_i=(x_i1,x_i2, ... ,t_i2,…,t_ip) and that the nonparametric particle is estimated according to one of the following methods:

1. Kernel regression .
2. Spline approximation .
3. Piecewise polynomial .
4. Local polynomial techniques .

In this aspect, the multiple linear regression function was taken in the parametric part, and in the nonparametric part, a local polynomial smoother (Local polynomial) was chosen. The average partial linear regression is written in the following formula:

$$E(Y_i/X_i, T_i) = X_i^{'}\beta + g(T_i) \qquad \dots(2-1)$$

Y_i:Dependent variable

$X_i^{'} = (x_{i1}, x_{i2}, \dots, x_{ip})\ T_i = (t_{i1}, t_{i2}, \dots, t_{ip})$ Independent variables

$\beta = (\beta_1, \beta_2, \dots, \beta_p)^{'}$: Estimated feature vector

$g(T_i)$:Anonymous bootstrap function

(1-2) Local polynomial smoother

The local polynomial smoother is one of the best common methods for estimating the kernel, as it is used with both fixed and random designs. The local linear regression smoother is also highly efficient compared to various other smoothers. When choosing the kernel function and bandwidth, it is 100% suitable among All available or possible graders [7].

Assuming that (P+1)

m(T)) . Taylor expansion .You can write as follows

$$m(T) = a_0 + a_1(T_i - t) + a_2(T_i - t)^2 + \cdots + a_P(T_i - t)^P$$

By performing sequential differentiation we get

$$m^{'}(T) = a_1 + 2a_2(T_i - t) + 3a_3(T_i - t)^2 + \cdots + Pa_P(T_i - t)^{P-1}$$

$$m^{''}(T) = +2a_2 + 6a_3(T_i - t) + \cdots + P(P-1)a_P(T_i - t)^{P-2}$$

$$\vdots$$

$$m^{'P}(T) = \mathrm{P!}\, a_{\mathrm{P}}$$

Suppose if

$$m(t) = a_0 \implies m(t) = \alpha_0$$

$$m^{'}(t) = a_1 \implies m^{'}(t) = \alpha_1$$

$$m^{''}(t) = 2a_2 \implies \frac{m^{''}(t)}{2} = \alpha_2$$

$$\vdots$$

$$m^{'P}(t) = P!\, a_P \implies a_P = \frac{m^{'p}(t)}{P!} = \alpha_P$$

Doi: https://doi.org/10.54216/PMTCS.030104
Received: June 19, 2023 Revised: September 25, 2023 Accepted: December 27, 2023

$$m(T) = \alpha_0 + \alpha_1(T_i - t) + \alpha_2(T_i - t)^2 + \cdots + \alpha_P(T_i - t)^P$$

$$m(T) = \sum_{j=1}^{p} \alpha_j (T_i - t)^j \qquad \ldots (2-11)$$

$$Y_i = X_i^{'}\beta + m(T_i) + \varepsilon_i$$

$$Y_i = X_i^{'}\beta + \sum_{j=1}^{p} \alpha_j (T_i - t)^j + \varepsilon_i \qquad \ldots (2-12)$$

$K_h(T_i - t)$ By squaring both sides of the above equation, entering the sum and multiplying by [4,8]

$$\sum_{i=1}^{n} \varepsilon_i^2 * K_h(t - T_i) = \sum_{i=1}^{n} \left(Y_i - X_i^{'}\beta - \sum_{j=1}^{p} \alpha_j (T_i - t)^j \right)^2 * K_h(t - T_i) \quad ,$$
$$i = 1,2,\ldots,n\,; j = 1,2,\ldots,p$$

$\frac{1}{h} K\left(\frac{\cdot}{h}\right) = K_h(.)$ It is a kernel function, which is a real, symmetric, definite, continuous function, and its integral is equal to one $\int K(u)\, du = 1$, Referring to the Bandwidth Parameter, the aforementioned equation is converted into a matrix as follows:

$$\alpha = \left[\alpha_0, \alpha_1, \ldots, \alpha_p\right]^{'}$$

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \qquad T = \begin{bmatrix} 1 & (T_1 - t) \cdots \cdots (T_1 - t)^P \\ \vdots & \vdots \quad \vdots \quad \vdots \\ 1 & (T_n - t) \cdots \cdots (T_n - t)^P \end{bmatrix},$$

$$W = \begin{bmatrix} K_h(t - T_1) \cdots \cdots & 0 \\ \vdots & \ddots \quad \vdots \\ 0 & \cdots \cdots K_h(t - T_n) \end{bmatrix} = diag K_h(t - T_i)$$

After converting equation (2-12) into a matrix, it is written as follows:

$$Y = X\beta + T\alpha + \varepsilon \qquad \ldots (2-13)$$

The weighted least squares (WLS) method was used as follows:

$$\frac{\partial \varepsilon^{'} W \varepsilon}{\partial \alpha'} = -T^{'}\mathrm{W}(Y - X\beta - T\alpha) = 0$$

$$\therefore \hat{\alpha} = (T'WT)^{-1}T'W\left(Y - X\hat{\beta}_{LP}\right) \qquad \ldots (2-14)$$

As for α_j, its estimate is found using the least squares method, and we extract an estimate of β [9,10].

$$\varepsilon'\varepsilon = (Y - X\beta - T\alpha)'(Y - X\beta - T\alpha)$$

$$\frac{\partial \varepsilon^{'} \varepsilon}{\partial \beta'} = -X'\left(Y - X\hat{\beta} - T\hat{\alpha}\right) = 0$$

$$X'X\hat{\beta} = X'Y - X'T\hat{\alpha} \qquad \ldots (2-15)$$

$$X'X\hat{\beta} = X'Y - X'T\left((T'WT)^{-1}T'W\left(Y - X\hat{\beta}_{LP}\right)\right)$$

$$X'X\hat{\beta} = X'Y - \left(X'T(T'WT)^{-1}T'WY - X'T(T'WT)^{-1}T'WX\hat{\beta}_{LP}\right)$$

$$X'(I - T(T'WT)^{-1}T'W)X\hat{\beta}_{LP} = X'(I - T(T'WT)^{-1}T'W)Y$$

Doi: https://doi.org/10.54216/PMTCS.030104
Received: June 19, 2023 Revised: September 25, 2023 Accepted: December 27, 2023

where

$\mathscr{R} = (I - T(T'WT)^{-1}T'W)$

$\hat{\beta}_{LP} = (X'\mathscr{R}X)^{-1}X'\mathscr{R}Y$ ... $(2-16)$

The arithmetic mean is obtained by using equation (2-16) as follows

$\hat{\beta}_{LP} = (X'\mathscr{R}X)^{-1}X'\mathscr{R}Y$

Substituting equation (2-13) into equation (2-16), it results:

$\hat{\beta}_{LP} = (X'\mathscr{R}X)^{-1}X'\mathscr{R}(X\beta + T\alpha + \varepsilon)$

$\hat{\beta}_{LP} = (X'\mathscr{R}X)^{-1}X'\mathscr{R}\,X\beta + (X'\mathscr{R}X)^{-1}X'\mathscr{R}T\alpha + (X'\mathscr{R}X)^{-1}X'\mathscr{R}\varepsilon$

$(X'\mathscr{R}X)^{-1}X'\mathscr{R}T\alpha = (X'\mathscr{R}X)^{-1}X'[(I - T(T'WT)^{-1}T'W)T\alpha]$

$(X'\mathscr{R}X)^{-1}X'\mathscr{R}T\alpha = (X'\mathscr{R}X)^{-1}X'[T\alpha - T\alpha] = 0$

$\therefore \hat{\beta}_{LP} = \beta + (X'\mathscr{R}X)^{-1}X'\mathscr{R}\varepsilon$ ... $(2-17)$

When the expectation is entered, the above equation becomes as follows

$\hat{\beta}_{LP} = \beta$

As for the variance, it can be found from equation (2-17), and it is as follows

$\therefore \hat{\beta}_{LP} - \beta = (X'\mathscr{R}X)^{-1}X'\mathscr{R}\varepsilon$

We square both sides of the equation

$$V - COV(\hat{\beta}_{LP}) = E(\hat{\beta}_{LP} - \beta)(\hat{\beta}_{LP} - \beta)' = (X'\mathscr{R}X)^{-1}\,X''\mathscr{R}\mathscr{R}X(X'\mathscr{R}X)^{-1}\sigma_{\varepsilon}^{2}$$

3. The practical aspect

Palm cultivation is considered one of the most important agricultural crops in Iraq because of its economic, nutritional and industrial importance. Palm cultivation is widespread in areas with a desert climate, and since Iraq's climate is semi-desert, it is a good place for palm cultivation. Date production is affected by several factors, including price and number of palm trees, along with climatic conditions such as humidity. Wind, amount of water, and temperature. Data were collected for the years from 1980 to 2020, so the sample size is (41), and the following variables were taken: the quantity of production, which represents the dependent variable Y, and the independent variables are the number of fruit palms X_1, which represents the first independent variable, and the price rate as the second independent variable. X_2, in addition to the dependence of the quantity of production on a nonparametric variable, which is the temperature variable t (1) Statistical metrics.

Table 1: shows the statistical measures for the variables included in the research.

	Min	Max	Mean	Std
production quantity (y)	251440	931540	5.8847e+05	1.8586e+05
Price rate $(x1)$	191	984000	3.7219e+05	2.9339e+05
Preparation of fruit palms $(x2)$	1413700	159111900	1.5605e+07	2.3253e+07
Temperature average (t)	2.687	30.300	27.6877	4.2368

Doi: https://doi.org/10.54216/PMTCS.030104
Received: June 19, 2023 Revised: September 25, 2023 Accepted: December 27, 2023

When using simulation experiments to apply what was presented in the theoretical aspect, the following semi-parametric regression model was used in these experiments:

$$Y_i = sin(\pi t) + x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i$$

Assuming that Y represents a random variable and that the variable x_{i1}, x_{i2}

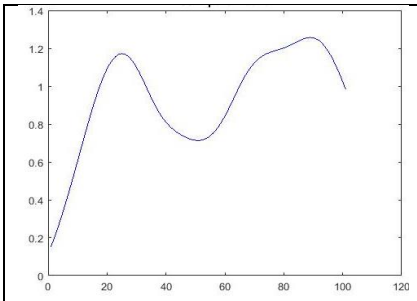
They were generated according to the standard normal distribution, while the variable t was generated according to the distribution. Two values were assumed for the features β_2, β_1she (0.5 , 0.25)و (-3 , 2) The error is normally distributed

$\varepsilon_i \sim N(0, \sigma^2)$ The simulation experiments carried out were carried out using three sample sizes $n =$ 30,500,100.

Table 2: shows the estimated values in case of different sample sizes.

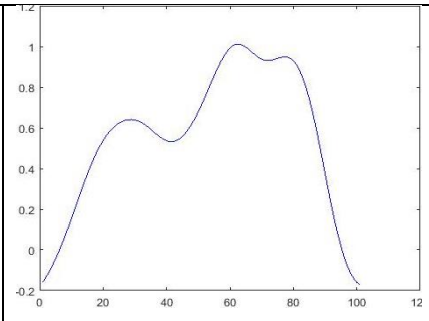
Models		PLM			OLS				
Estimated values	n	b_1	b_2	MSE	b_o	b_1	b_2	b_3	MSE
β_1=0.5	30	0.5970	0.3164	2.89e+03	1.7559	0.4298	-0.851	-0.0255	34.6056
β_2=0.25	50	0.5595	0.2456	2.10e+03	1.0569	0.4015	0.192	-0.2505	27.6355
	100	0.4583	0.2728	2.69e+03	0.4457	0.4856	0.277	0.1332	102.2379
β_1=2	30	1.9753	-3.0461	1.742e+03	0.9328	1.8866	-3.225	-0.3892	21.0481
β_2= -3	50	1.8902	-3.2589	3.25e+03	1.1472	2.0142	3.476-	0.0622	46.6195
	100	2.0316	-3.4133	4.253e+0	0.4279	1.9906	-3.030	0.2962	121.7104

Table (2) shows the estimated values of the parametric regression models with the semi-parametric regression models in the case of using sample sizes for a set of data generated by the method of simulation experiments, as we notice convergence with the default values with b_1 ,b_2 Also, for both models, the mean square error (MSE) was lower for the semi-parametric regression model.



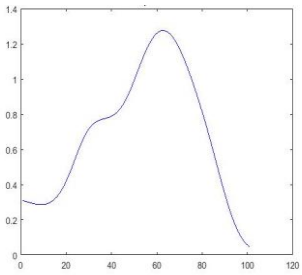
$n = 30$

β_1=0.5,β_2=0.25



$n = 50$

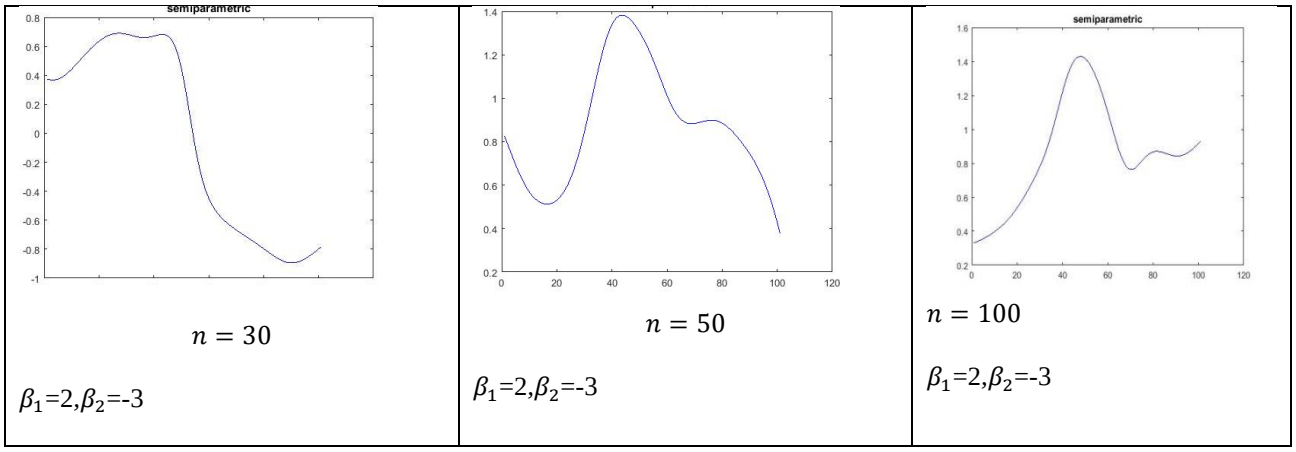
β_1=0.5,β_2=0.25



$n = 100$

β_1=0.5,β_2=0.25

Doi: https://doi.org/10.54216/PMTCS.030104
Received: June 19, 2023 Revised: September 25, 2023 Accepted: December 27, 2023



$n = 30$

β_1=2,β_2=-3

$n = 50$

β_1=2,β_2=-3

$n = 100$

β_1=2,β_2=-3

Figure 1: shows the drawings for each of the sample sizes and default values in the case of simulation experiments.

We notice from Figure (1) that the curve of the semi-parametric regression function varies depending on the sample sizes and the default values of the parameters in the case of simulation experiments.

If real data is used to estimate the semi-parametric regression function and compare it with the ordinary regression function, the mean square error will be in Table 3.

Table 3: shows the estimated parameter values and mean square errors extracted from real-world data

Models		parameters	MSE
OIS		β_o=551577.674	6.9558e+44
		β_1=0.001	
		β_2=0.106	
		β_3=-76.432	
PLM	h=0.0800	β_1=0.0796	4.3569e-14
		β_2=0.0011	

We note from Table (3) the estimated values of the parameters using the ordinary least squares method in the event that the variables are assumed to behave parametrically, as the mean square error is 6.9558e+44. However, if we make the first variable and the second variable linear, and the third variable is non-linear, it behaves Nonparametric behavior, the mean square error is 4.3569e-14, and we note that the PLM estimation method outperformed OIS because it gave a lower mean square error.

4. Conclusions

From what was mentioned in the practical aspect, it is clear that the semi-parametric regression model has flexibility in describing data in the case of simulation experiments and in the case of real data compared to the semi-parametric model. We also note that the semi-parametric regression model has shown better results in the case of using sample sizes for a set of data. Generated by the method of simulation experiments. Also, the semi-parametric regression model is better in the prediction process because it has the lowest mean square error compared to the usual least squares method.

Doi: https://doi.org/10.54216/PMTCS.030104
Received: June 19, 2023 Revised: September 25, 2023 Accepted: December 27, 2023

References

[1] Gozalo PL. A consistent model specification test for nonparametric estimation of regression function models. Econometric Theory 1993;9:451–77.
[2] Han AK. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. Journal of Econometrics 1987;35:303–16.
[3] Fox J. An R and S-Plus companion to applied regression. Sage; 2002.
[4] González-Manteiga W, Cao R. Testing the hypothesis of a general linear model using nonparametric regression estimation. Test 1993;2:161–88.
[5] Juhl T. A nonparametric test of the predictive regression model. Journal of Business & Economic Statistics 2014;32:387–94.
[6] Jeong K, Härdle WK, Song S. A consistent nonparametric test for causality in quantile. Econometric Theory 2012;28:861–87.
[7] Choi T, Schervish MJ. On posterior consistency in nonparametric regression problems. Journal of Multivariate Analysis 2007;98:1969–87.
[8] Eubank RL, Li C-S, Wang S. Testing lack-of-fit of parametric regression models using nonparametric regression techniques. Statistica Sinica 2005:135–52.
[9] Gozalo PL, Linton OB. Testing additivity in generalized nonparametric regression models with estimated parameters. Journal of Econometrics 2001;104:1–48.
[10] Zheng JX. A consistent nonparametric test of parametric regression models under conditional quantile restrictions. Econometric Theory 1998;14:123–38.

Doi: https://doi.org/10.54216/PMTCS.030104
Received: June 19, 2023 Revised: September 25, 2023 Accepted: December 27, 2023