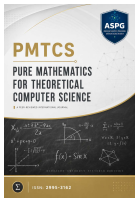
Pure Mathematics for Theoretical Computer Science

♦ ISSN: 2995-3162 Vol: 03, No. 01, 2024 ♦ PP. 16–20

Article DOI: https://doi.org/10.54216/PMTCS.030104



FULL LENGTH ARTICLE

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

Email: Kareemalataby28@gmail.com

Received: June 19, 2023 Revised: September 25, 2023 Accepted: December 27, 2023 ⋆ Corresponding author

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 semi-parametric 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].

2. THE THEORETICAL SIDE

The semi-parametric regression model, which was proposed by researchers Robinson in 1988 and Paul Speckman in 1988,

© 2024 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

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, it is called the partial linear regression function [3, 4, 5, 6]. The explanatory variables in the parametric part are binary or discrete explanatory variables, and they are often referred to as

$$\mathbf{X}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip}),$$

while the nonparametric part 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 was chosen. The average partial linear regression is written in the following formula:

$$E(Y_i \mid \mathbf{X}_i, T_i) = \mathbf{X}_i'\beta + g(T_i). \tag{2-1}$$

Here, Y_i is the dependent variable,

$$\mathbf{X}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip}), \qquad T_i = (t_{i1}, t_{i2}, \ldots, t_{ip}),$$
$$\beta = (\beta_1, \beta_2, \ldots, \beta_p)',$$

where β is the estimated feature vector and $g(T_i)$ is the anonymous bootstrap function.

2.1 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 a polynomial of order P, the Taylor expansion of $m(T)$ can be written as follows:

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

By performing sequential differentiation, we get

$$\begin{aligned} m'(T) &= \alpha_1 + 2\alpha_2(T_i - t) + 3\alpha_3(T_i - t)^2 \\ &\quad + \cdots + P\alpha_P(T_i - t)^{P-1}, \\ m''(T) &= 2\alpha_2 + 6\alpha_3(T_i - t) \\ &\quad + \cdots + P(P-1)\alpha_P(T_i - t)^{P-2}, \\ &\vdots \\ m^{(P)}(T) &= P!\,\alpha_P. \end{aligned}$$

Suppose that

$$\begin{aligned} m(t) &= \alpha_0, \\ m'(t) &= \alpha_1, \\ m''(t) &= 2\alpha_2 \quad \Longrightarrow \quad \frac{m''(t)}{2} = \alpha_2, \\ &\vdots \\ m^{(P)}(t) &= P!\,\alpha_P \quad \Longrightarrow \quad \alpha_P = \frac{m^{(P)}(t)}{P!}. \end{aligned}$$

Thus,

$$m(T) = \alpha_0 + \alpha_1(T_i - t) + \alpha_2(T_i - t)^2 + \cdots + \alpha_P(T_i - t)^P$$
$$m(T) = \sum_{j=1}^{p} \alpha_j (T_i - t)^j. \tag{2-11}$$

The semi-parametric regression model is then

$$Y_i = \mathbf{X}_i'\beta + m(T_i) + \varepsilon_i$$
$$Y_i = \mathbf{X}_i'\beta + \sum_{j=1}^{p} \alpha_j (T_i - t)^j + \varepsilon_i. \tag{2-12}$$

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

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

The scaled kernel

$$K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$$

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 matrix form as follows:

$$\alpha = [\alpha_0, \alpha_1, \ldots, \alpha_p]', \qquad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

$$\mathbf{T} = \begin{bmatrix} 1 & T_1 - t & \cdots & (T_1 - t)^P \\ \vdots & \vdots & \ddots & \vdots \\ 1 & T_n - t & \cdots & (T_n - t)^P \end{bmatrix},$$

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

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

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{T}\alpha + \varepsilon. \tag{2-13}$$

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

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

Therefore,

$$\widehat{\alpha} = (\mathbf{T}'\mathbf{W}\mathbf{T})^{-1}\mathbf{T}'\mathbf{W}(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{LP}). \tag{2-14}$$

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

$$\varepsilon'\varepsilon = (\mathbf{Y} - \mathbf{X}\beta - \mathbf{T}\alpha)'(\mathbf{Y} - \mathbf{X}\beta - \mathbf{T}\alpha),$$
$$\frac{\partial \varepsilon'\varepsilon}{\partial \beta'} = -\mathbf{X}'(\mathbf{Y} - \mathbf{X}\widehat{\beta} - \mathbf{T}\widehat{\alpha}) = 0.$$

Hence,

$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{T}\widehat{\alpha}. \tag{2-15}$$

Substitution gives

$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{T}(\mathbf{T}'\mathbf{W}\mathbf{T})^{-1}\mathbf{T}'\mathbf{W}(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{LP}),$$
$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{Y} - \big(\mathbf{X}'\mathbf{T}(\mathbf{T}'\mathbf{W}\mathbf{T})^{-1}\mathbf{T}'\mathbf{W}\mathbf{Y} - \mathbf{X}'\mathbf{T}(\mathbf{T}'\mathbf{W}\mathbf{T})^{-1}\mathbf{T}'\mathbf{W}\mathbf{X}\widehat{\beta}_{LP}\big),$$

Writing

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

the last expression becomes

$$\mathbf{X}'\mathscr{R}\mathbf{X}\widehat{\beta}_{LP} = \mathbf{X}'\mathscr{R}\mathbf{Y},$$

and

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

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

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

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

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

Furthermore,

$$\mathscr{R}\mathbf{T}\alpha = \big(\mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{W}\mathbf{T})^{-1}\mathbf{T}'\mathbf{W}\big)\mathbf{T}\alpha = \mathbf{T}\alpha - \mathbf{T}\alpha = 0,$$
$$(\mathbf{X}'\mathscr{R}\mathbf{X})^{-1}\mathbf{X}'\mathscr{R}\mathbf{T}\alpha = 0.$$

Therefore,

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

When the expectation is entered, the above equation becomes

$$E(\widehat{\beta}_{LP}) = \beta.$$

As for the variance, it can be found from equation (2-17):

$$\widehat{\beta}_{LP} - \beta = (\mathbf{X}'\mathscr{R}\mathbf{X})^{-1}\mathbf{X}'\mathscr{R}\varepsilon.$$

By squaring both sides of the equation,

$$\text{V-COV}(\widehat{\beta}_{LP}) = E\Big[(\widehat{\beta}_{LP} - \beta)(\widehat{\beta}_{LP} - \beta)'\Big] = (\mathbf{X}'\mathscr{R}\mathbf{X})^{-1}\mathbf{X}'\mathscr{R}'\mathscr{R}\mathbf{X}(\mathbf{X}'\mathscr{R}\mathbf{X})^{-1}\sigma_\varepsilon^2.$$

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.

3.1 Statistical Metrics

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

Table 1. Statistical measures for the variables included in the research.

Variable	Min	Max	Mean	Std
Production quantity (y)	251440	931540	5.8847e+05	1.8586e+05
Price rate (x_1)	191	984000	3.7219e+05	2.9339e+05
Preparation of fruit palms (x_2)	1413700	159111900	1.5605e+07	2.3253e+07
Temperature average (t)	2.687	30.300	27.6877	4.2368

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

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

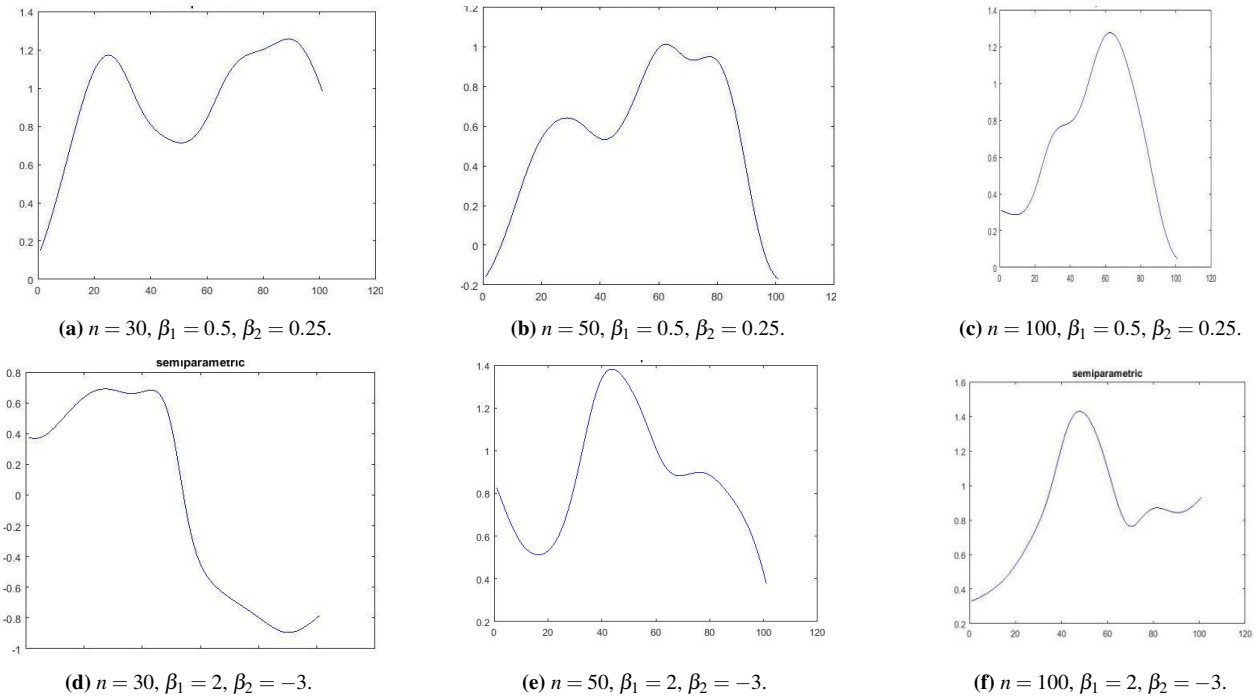
Assuming that Y represents a random variable, the variables x_{i1} and x_{i2} were generated according to the standard normal distribution, while the variable t was generated according to the distribution. Two parameter settings were assumed: $(\beta_1, \beta_2) = (0.5, 0.25)$ and $(\beta_1, \beta_2) = (2, -3)$. The error is normally distributed,

$$\varepsilon_i \sim N(0, \sigma^2).$$

The simulation experiments were carried out using three sample sizes: $n = 30$, 50, and 100. Table 2 shows the estimated values for the different sample sizes.

Table 2. Estimated values for different sample sizes.

Models		PLM			OLS				
(β_1, β_2)	n	b_1	b_2	MSE	b_0	b_1	b_2	b_3	MSE
$(0.5, 0.25)$	30	0.5970	0.3164	2.89e+03	1.7559	0.4298	−0.851	−0.0255	34.6056
	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
$(2, -3)$	30	1.9753	−3.0461	1.742e+03	0.9328	1.8866	−3.225	−0.3892	21.0481
	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



(a) $n = 30$, $\beta_1 = 0.5$, $\beta_2 = 0.25$. **(b)** $n = 50$, $\beta_1 = 0.5$, $\beta_2 = 0.25$. **(c)** $n = 100$, $\beta_1 = 0.5$, $\beta_2 = 0.25$.

(d) $n = 30$, $\beta_1 = 2$, $\beta_2 = -3$. **(e)** $n = 50$, $\beta_1 = 2$, $\beta_2 = -3$. **(f)** $n = 100$, $\beta_1 = 2$, $\beta_2 = -3$.

Figure 1. Drawings for each of the sample sizes and default values in the 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 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 is shown in Table 3.

Table 3. Estimated parameter values and mean square errors extracted from real-world data.

Models	Parameters	MSE
OIS	$\beta_0 = 551577.674$ $\beta_1 = 0.001$ $\beta_2 = 0.106$ $\beta_3 = -76.432$	6.9558e+44
PLM	$h = 0.0800$ $\beta_1 = 0.0796$ $\beta_2 = 0.0011$	4.3569e−14

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 non-linear, it behaves nonparametrically; the mean square error is 4.3569e−14. We note that the PLM estimation method outperformed OIS because it gave a lower mean square error.

4. CONCLUSION

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.

REFERENCES

[1] P. L. Gozalo, "A consistent model specification test for nonparametric estimation of regression function models," *Econometric Theory*, vol. 9, pp. 451–477, 1993.

[2] A. K. Han, "Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator," *Journal of Econometrics*, vol. 35, pp. 303–316, 1987.

[3] J. Fox, *An R and S-Plus Companion to Applied Regression*. Sage, 2002.

[4] W. González-Manteiga and R. Cao, "Testing the hypothesis of a general linear model using nonparametric regression estimation," *Test*, vol. 2, pp. 161–188, 1993.

[5] T. Juhl, "A nonparametric test of the predictive regression model," *Journal of Business & Economic Statistics*, vol. 32, pp. 387–394, 2014.

[6] K. Jeong, W. K. Härdle, and S. Song, "A consistent nonparametric test for causality in quantile," *Econometric Theory*, vol. 28, pp. 861–887, 2012.

[7] T. Choi and M. J. Schervish, "On posterior consistency in nonparametric regression problems," *Journal of Multivariate Analysis*, vol. 98, pp. 1969–1987, 2007.

[8] R. L. Eubank, C.-S. Li, and S. Wang, "Testing lack-of-fit of parametric regression models using nonparametric

regression techniques," *Statistica Sinica*, pp. 135–152, 2005.

[9] P. L. Gozalo and O. B. Linton, "Testing additivity in generalized nonparametric regression models with estimated parameters," *Journal of Econometrics*, vol. 104, pp. 1–48, 2001.

[10] J. X. Zheng, "A consistent nonparametric test of parametric regression models under conditional quantile restrictions," *Econometric Theory*, vol. 14, pp. 123–138, 1998.