Black Sea Journal of Agriculture
doi: 10.47115/bsagriculture.1077101

Open Access Journal
e-ISSN: 2618 – 6578

Research Article
Volume 5 - Issue 2: 160-166 / April 2022

SEMIPARAMETRIC REGRESSION MODELS AND APPLICABILITY IN AGRICULTURE

Esra YAVUZ[1]*, Mustafa ŞAHİN[2]

[1]*Şırnak University, Cizre Vocational School, Department of Accounting and Tax, 73200, Cizre, Şırnak, Turkey*
[2]*Kahramanmaras Sütçü Imam University, Faculty of Agriculture, Department of Agricultural Biotechnology, 46100, Kahramanmaraş, Turkey*

Abstract: Parametric regression models assume that the dependent variable is a linear relationship with the independent variables and the form of the relationship is known. Nonparametric regression methods are applied in cases where the relationship type is not known or assumptions cannot be provided. However, when there is more than one independent variable, some of the independent variables may be in a linear relationship with the dependent variable, while some may be in a nonlinear relationship. In order to model these variables, semiparametric regression models, which are a combination of parametric and nonparametric regression methods, are used. In this study parametric, nonparametric and semiparametric regression models, parametric estimates, fit statistical values of the models, confidence intervals and standard error values were calculated. As a result of the analysis, the parameters of the milking unit and the quarantine area among the parametric variables, the operation area, the ventilation area, the number of ventilation, the quarantine area, the infirmary area, the manure pit and the distance to the center among the non-parametric variables were found to be statistically very important (P<0.01). As a result, it was concluded that the correct definition of the variables (parametric and non-parametric) that are effective in determining the operating cost of agricultural enterprises and consequently the sales price, and the selection of the appropriate model are extremely important and that semiparametric models can be used easily in this field.

Keywords: Semiparametric, Regression, Agricultural businesses

***Corresponding author:** Şırnak University, Cizre Vocational School, Department of Accounting and Tax, 73200, Cizre, Şırnak, Turkey
E mail: yavuz7346@gmail.com (E. YAVUZ)
Esra YAVUZ https://orcid.org/0000-0002-5589-297X
Mustafa ŞAHİN https://orcid.org/0000-0003-3622-4543

Received: February 21, 2022
Accepted: March 18, 2022
Published: April 01, 2022

Cite as: Yavuz E, Şahin M. 2022. Semiparametric regression models and applicability in agriculture. BSJ Agri, 5(2): 160-166.

1. Introduction

Regression statistically analyzes the functional effect of independent variables on the dependent variable, based on a given or obtained data set. Regression analysis is an important method that is widely used in determining the relationship between variables. Regression analysis, which dates back to the 19th century, examines the conditional distribution of the dependent variable for certain values of the independent variables. It is used in many fields such as science, medicine, engineering and social sciences to determine and predict the relationships between variables (Aytaç, 1991; Alpar, 2003).

Linear regression analysis is examined on the assumptions that independent variables affect the dependent variable linearly and that the dependent variable has a normal distribution. Many theoretical and practical studies have been carried out for linear regression analysis, and the results of these studies provide a theoretical and practical basis for examining more complex regression models. When certain conditions are met, linear regression analysis yields appropriate results in solving practical estimation problems. However, in most estimation problems, some of the independent variables do not affect the dependent variable linearly. Thus, the need to examine regression models that are not fully linear and contain more complex correlations arises. Thus, regression analysis is examined in two different groups as parametric and non-parametric regression (Begun et al., 1983; Aneiros-Pérez, 2008).

The most important feature of parametric regression analysis is that the shape of the regression function is known beforehand. In addition, it is required to provide assumptions such as constant error variances for all values of the independent variable, normal distribution of error terms, no autocorrelation between error terms, and no multicollinearity between independent variables. If the assumptions are not provided, the results of the estimations made for the regression function cause misinterpretations. Thus, in case the assumptions of the model created by parametric regression analysis are not met, some adjustments can be made to provide assumptions. Thus, estimations can be made since necessary assumptions are provided (Buckley et al., 1988; Berry, 1993; Yatchew, 2003).

In non-parametric regression analysis, the shape of the function is not known beforehand. As in parametric regression analysis, important assumptions are not required. The only assumption is that the mean of the

error terms is zero and the variance is a finite number. Therefore, there is flexibility in determining the relationship between variables.

Semiparametric regression method is also called "partial linear regression models" because of the combination of parametric and non-parametric regression function and additive. If the independent variables are unrelated in the semiparametric regression model, the coefficients of the parametric variables of the model are estimated by applying the least squares method and the partial regression functions of the non-parametric variables are estimated by non-parametric methods such as spline. (Newey, 1989). While some assumptions are needed in parametric and non-parametric methods, research continues even if the assumptions are not fulfilled in the semiparametric regression method (Shi, 2009; Toprak, 2015).

In this study, three different regression methods as parametric, non-parametric and semiparametric regression methods, smoothing method in regression, roughness penalty approach and spline correction techniques together with estimation methods used in smoothing parameter are explained. Afterwards, semiparametric regression, which is the main subject of the study, was discussed and two different approaches, partial spline and backfitting algorithm, were examined in the estimation of the model. In addition, inferences regarding the semiparametric regression model were applied for both parametric and non-parametric regression methods.

2. Material and Methods

2.1. Material

The data used in this thesis belong to 60 agricultural livestock enterprises in Kahramanmaraş. The variables that are thought to be effective on the price of the barns, the price of the farm, the presence of the milking unit, the presence of the quarantine zone, the shelter area, the ventilation area, the number of ventilation, the presence of the quarantine area, the presence of the infirmary area, the presence of the manure pit, the presence of the birth unit and the distance to the city center are discussed. While some variables were included in the model in parametric form, some variables were included in non-parametric form.

In practice, semiparametric regression model, which is an additive model, was used to evaluate parametric and non-parametric variables. In the statistical evaluations, SAS 9.4 package program was used.

2.2. Methods

2.2.1. Regression analysis

Regression analysis is an analysis method used to measure the relationship between two or more variables that have a cause-effect relationship between them. In this analysis method, if the analysis is made using a single variable, it is called univariate regression, and if more than one variable is used, it is called multivariate regression analysis. Regression analysis is used to apply the existence of the relationship between the variables, the strength of the relationship if there is a relationship, and to make predictions or estimations about the subject by using this relationship. In the regression, one of the variables is considered as dependent and the others as independent variable (Hurvich and Tsai, 1989; Omay, 2007).

2.2.2. Parametric regression method

The parameter is the mean, ratio, variance, etc. that belong to the population. Parametric regression is to show the mean relationship between dependent and independent variables with a mathematical function and to express the parameters in this function clearly. Parametric regression assumes that the regression function is represented as a linear function of the arguments $x_1, x_2, \dots, x_q$. $E\ (y \mid X)$ explains the functional relationship of the mean distribution of y with X when the conditional expected value X is known (equation 1 and 2) (Speckman, 1988; Schimek, 2000).

$$E\ (y \mid X) = \boldsymbol{X}\beta \quad (1)$$

or

$$y = \boldsymbol{X}\beta + \ \varepsilon_i \quad (2)$$

shown in the form.

2.2.3. Non-parametric regression method

Non-parametric regression, simple non-parametric regression model, one of which is the dependent variable (y) and the independent variable (x) whose relationship with the dependent variable is unknown (equation 3),

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1,2, \dots, n \quad (3)$$

shown in the form. The main purpose of the nonparametric regression method is to estimate the unknown mean function $f(x_i)$ rather than estimating the parameters.

Although there are no limiting assumptions in the non-parametric regression method, it may have some features. It is difficult to make predictions when the number of independent variables is large. In addition, the resulting graphics are shown in a complex structure. As a result of these situations, the “dimensionality problem” arises. At the same time, it is difficult to handle discrete independent variables with non-parametric regression method and to interpret the effects of the y dependent variable with the increase in the number of independent variables. These difficulties can be eliminated by applying the semi-parametric regression method (Tezcan, 2011).

2.2.4. Semiparametric regression method

The most important advantage of non-parametric regression models is the absence of any assumptions about the functional form of the relationship between the dependent variable and the independent variables in regression models. The flexibility provided by non-

parametric regression models makes this model applicable. However, it is very difficult to calculate the smoothing process in this model. In addition, as the number of independent variables increases, the reliability of non-parametric estimates decreases gradually due to the size problem. Thus, when the functional form of the relationship is not known in the regression model, it may result in the absence of an important interpretation of the data in parametric and non-parametric regression models.

In order to overcome these problems, semiparametric regression model (semi-parametric regression model), which is regression models consisting of some parametric and some non-parametric variables, is applied (Schennach, 2004).

Semiparametric regression models are examined as a special case of additive models that generalize standard regression methods and provide an appropriate interpretation of the effect of each variable. Semiparametric regression model, where some of the variables are parametric and some of them are non-parametric variables (equation 4);

$$y_i = \alpha + f_1(x_1) + \cdots + f_j(x_j) + x_{j+1}\beta_1 + \cdots + x_k\beta_k + \varepsilon \quad (4)$$

shown in the form. The j variables in the semiparametric regression model have a non-linear effect on the dependent variable y and show the non-parametric part of the model. Other variables have a linear effect on the y dependent variable and show the parametric part of the model. In addition, there may be discrete variables such as dummy variables in the parametric part of the model. In the non-parametric part of the model, when there is more than one variable, the inconveniences of the non-parametric model will also be valid for these models. In order to eliminate these problems, the variables in the non-parametric part of the model are added to the model and a new model is created (Zhongyi and Baocheng, 2001).

2.2.5. Estimation of semiparametric regression models

Iterative algorithms are used in the estimation of semiparametric regression models and additive models. There are many algorithms developed for the estimation of these models and these algorithms are implemented in different computer software. R software and SAS software are the most preferred programs for analyzing these algorithms. When the independent variables are uncorrelated in the semiparametric regression model, it is quite easy to estimate the semiparametric regression models with many non-parametric variables. In other words, if the independent variables are unrelated, the coefficients of the parametric variables of the model are estimated by applying the least squares method, and the partial regression functions of the non-parametric variables are estimated by non-parametric methods such as spline. However, in semiparametric regression models, the parametric and non-parametric variables of the model may be related to each other. Thus, considering the relationships between the variables, different algorithms are needed. The most preferred among these algorithms are the Newton-Raphson algorithm and the backfitting algorithm (Mammadov, 2005; Liu et al., 2013).

3. Results and Discussion

It is a known fact that in determining the selling prices of agricultural livestock enterprises, the characteristics of the enterprise have an effect on the price. Knowing how these features affect the sales prices of the enterprises, determining the production cost, presenting the products in the supply-demand chain effectively and profitably will give the business owner important information about the sustainability of production and the future of the business. Because the presence of milking unit, quarantine zone, shelter area, ventilation area, number of ventilation, quarantine area, presence of infirmary area, presence of manure pit, presence of birthing unit and distance to the city center are directly or indirectly related to the efficiency of production and additional investment.

In this study, the results of the semiparametric regression model, which includes the multivariate parametric regression model, in which some of the variables affecting the firm price are linear, and the non-parametric regression model, which includes some nonlinear variables, were obtained. The data set used in the study belongs to 60 agricultural livestock enterprises in Kahramanmaraş province and its surroundings. SAS 9.4 package program was used in the analyzes for parametric regression, non-parametric regression and semiparametric regression models.

First of all, assuming that all independent variables have a linear effect on the selling price of the agricultural enterprise, the linear regression model expressed in equation 5 was defined.

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + x_8\beta_8 + x_9\beta_9 + \varepsilon \quad (5)$$

Data number (N), arithmetic mean ($\bar{x}$), standard deviation (S) for the price, area, ventilation area, ventilation number, quarantine area, quarantine zone, infirmary area, manure pit, milking unit, distance to the center variables of this model, median and minimum-maximum values are given in Table 1.

Estimated coefficients for the area, ventilation area, ventilation number, quarantine area, quarantine zone, infirmary area, manure pit, milking unit, distance to the center variables of this model, standard error values, t-calculus value, significance levels (P), determination coefficient, corrected coefficient of determination, sum of squares of error and deviation values are given in Table 2.

Table 1. Descriptive statistics values of variables

Variables	N	$\bar{x}$	S	Median	Min-Max
Price (y)	60	1913524	706285	1982230	955230-3150120
Area (x_1)	60	1627.07	616.77	1600	780-2790
Ventilation area(x_2)	60	283.40	1115.51	80	37-8635
Number of ventilation (x_3)	60	13.72	4.77	12.50	5-28
Quarantine zone (x_4)	60	0.58	0.49	1.00	0-1
Quarantine area(x_5)	60	26.82	23.82	40	0-80
Infirmary area (x_6)	60	47.58	24.71	45	0-142
Manure pit (x_7)	60	311.4	97.56	320	100-500
Milking unit (x_8)	60	0.68	0.46	1.00	0-1
Distance from center (x_9)	60	69.85	38.10	56	20-160

Table 2. Estimation results of the variables of the linear regression model

Variables	Coefficients	$S_{\bar{x}}$	t	P
Fixed	-30858.28	22117.62	-1.40	0.890
Area (x_1)	827.109	134.654	6.142	0.000**
Ventilation area (x_2)	-52.074	67.963	-0.766	0.447
Number of ventilation (x_3)	23792.965	12433.654	1.914	0.062
Quarantine zone (x_4)	12812.820	11259.863	1.138	0.261
Quarantine area (x_5)	-694084.866	510136.495	-1.361	0.180
Infirmary area (x_6)	3513.317	2719.605	1.292	0.202
Manure pit (x_7)	516.844	626.670	0.825	0.414
Milking unit (x_8)	-235931.494	122664.109	-1.923	0.060
Distance from center (x_9)	1171.255	1388.689	0.843	0.403
$R^2 = 0.820$ $\bar{R}^2 = 0.783$	F= 22.274 Error Sum of Squares=530700 S= 63944.903			

**the parameters are statistically very significant at the 0.01 significance level.

As seen in Table 2., while the effects of ventilation area, ventilation number, quarantine area, quarantine zone, infirmary area, manure pit, milking unit and distance to the center on the sales price of agricultural holdings were found to be statistically insignificant (P>0.05), the effect on the area variable was found to be insignificant. Effect was found to be very significant (P<0.01).

When multiple linear regression is applied, it is seen that the variables of ventilation area, quarantine area and milking unit have a negative effect on the sales price, while other variables have a positive effect. In addition, the sum of squares of the error and the deviation value were quite high. Thus, it has been seen that the linear parametric model is not sufficient to determine the variable effects on the selling prices of the agricultural enterprise. For this purpose, a semiparametric regression model was created for the variables.

Since the discrete variables included in the study do not affect the curvature of the function, in other words, they are included in the model parametrically since they do not need correction. On the other hand, other variables whose type of relationship with the dependent variable is not known precisely were included in the model as non-parametric part. In order to determine the appropriate semiparametric regression model, the semiparametric regression model in which both parametric and non-parametric variables are included and the relationship between the sales prices of agricultural enterprises and the characteristics of the agricultural enterprise is examined in order to see how the predictions of the model are interpreted, is defined with the equation 6.

$$y = \beta_0 + x_4\,\beta_1 + x_8\beta_2 + s(x_1) + s(x_2) + s(x_3) + s(x_5) + s(x_6) + s(x_7) + s(x_9) + \varepsilon \quad (6)$$

Parameter estimates, standard error values, Chi-square calculation value and significance levels (P) for the quarantine zone and milking unit variables of this model are given in Table3.

Table 3. Estimation results of parametric variables in semiparametric regression model

Variables	Coefficients	$S_{\bar{x}}$	χ^2	P
Fixed	13.569	0.048	77645.992	<0.001**
Milking unit (x_8)	-0.016	0.005	9.854	<0.001**
Quarantine zone (x_4)	-0.812	0.113	51.02	<0.001**

**= the parameters are statistically very significant at the 0.01 significance level.

When the parametric variables in the application are examined according to Table 3, it is seen that all the parametric variables in the model are statistically very significant (P<0.01). Among these variables, the milking unit and the quarantine zone negatively affect the price variable. Among the parametric variables, the variable that most negatively affects prices is the quarantine zone variable.

Table 4. Estimation results of nonparametric variables in semiparametric regression model

Component	EDF	F	P
Area (x_1)	6.8554.04	188.17	<.001**
Ventilation area (x_2)	6.05	68.22	<.001**
Number of ventilation (x_3)	5.01	110.35	<.001**
Quarantine area (x_5)	7.07	1162.24	<.001**
Infirmary area (x_6)	7.48	144.04	<.001**
Manure pit (x_7)	6.61	260.36	<.001**
Distance from center (x_9)	1.00	25.96	<.0002**

**the parameters are statistically very significant at the 0.01 significance level. EDF= effective degrees of freedom

The additive representation of the parametric and non-parametric regression models, the results of which are shown in Table 3 and Table 4, are shown in equation 7.

$$y = 13.569 + x_4(-0.812) + x_8(-0.016) + s(x_1) + s(x_2) + s(x_3) + s(x_5) + s(x_6) + s(x_7) + s(x_9) + \varepsilon \quad (7)$$

Equation 7 consists of two parts as parametric and non-parametric regression. Coefficient interpretations and inferences for these two sections are analyzed with separate methods. The interpretations and inferences for the parametric regression part of the semiparametric regression model are similar to the linear regression models. While the comments for the non-parametric regression part are analyzed with the help of graphics, the inferences are examined with the help of the F test.
Since the non-parametric part obtained contains many coefficients, in other words, it is obtained as a vector, it is not possible to express it parametrically, and thus non-parametric components can only be displayed with graphics (Turanlı and Bağdatlı, 2012). Therefore, the relationship between the price and the variables included in the model in non-parametric form is given in Figure 1 to Figure 7.

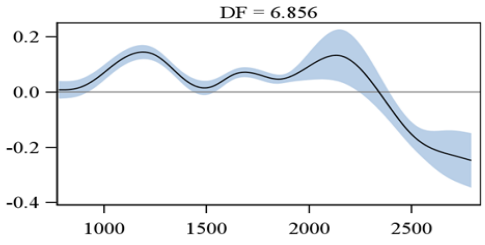

Figure 1. Area and S(A) graph.

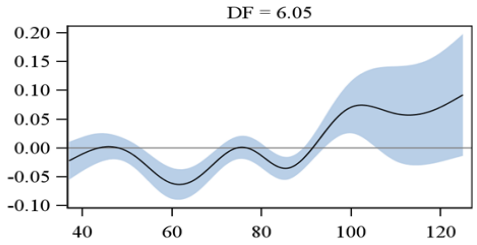

Figure 2. Ventilation area and S(VA) graph.

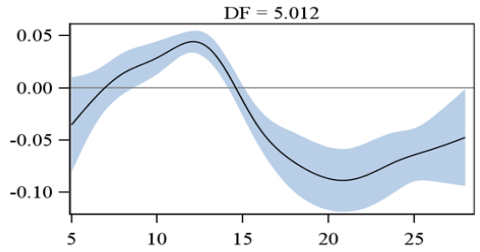

Figure 3. Number of vents and S(NV) graph.

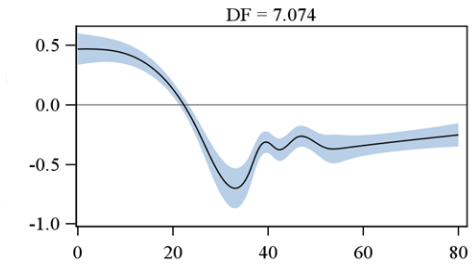

Figure 4. Quarantine area and S(QA) graph.

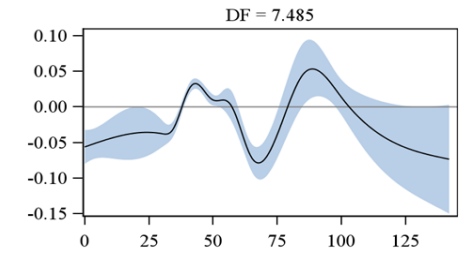

Figure 5. Infirmary area and S(IA) graph.

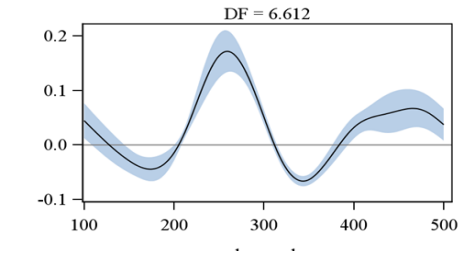

Figure 6. Quarantine area and S(QA) graph.

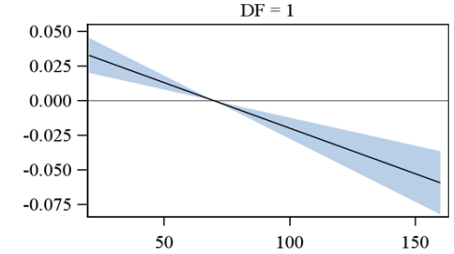

Figure 7. Distance to the center and S(DC).

When the graphs in the figure (Figure 1, 2, 3, 4, 5, 6 and 7) are examined, it can be observed that there is a non-linear relationship between the price and the variables included in the model in non-parametric form. Regarding the estimation of the semiparametric model, the values of spline values on the vertical axis and non-parametric variables on the horizontal axis were obtained. That is, it shows how the coefficient estimates change in response to the change in the value of each nonparametric variable. The shaded areas in the figure (Figure 1, 2, 3, 4, 5, 6 and 7) indicate that it is in the 95% confidence interval band.

4. Conclusion

In practice, the variables that should be included in the model by smoothing were examined and since there were both parametric and non-parametric variables in the model, it was found appropriate to apply semiparametric regression analysis. The most important feature of the semiparametric regression model is that it can examine the relationship between the dependent variable and the independent variables with statistical tests. In other words, it decides whether to include a variable in the model by smoothing it, linearizing it, or linearizing it by transforming methods. It also shows which model is suitable by comparing the models. In addition to modeling with the semiparametric regression method, determining the structures of the variables using this method also provides the best estimates.

The fact that there are many investment and environmental factors that determine the costs and therefore sales prices of agricultural enterprises clearly reveals how important the correct modeling is. Because, when the variables examined here are taken into the model parametrically, an erroneous result emerges that many variables known to be very effective on cost and selling price have an insignificant effect. In the semiparametric model, on the other hand, inclusion of some of the variables in the parametric and non-parametric form of the variables, which are known to be important in practice, turned out to be statistically very important.

As a result, it can be said that the correct definition of the variables (parametric and non-parametric) and the selection of the appropriate model are extremely important in determining the operating price of agricultural enterprises and accordingly the sales price, and it can be said that semi-parametric models can be easily used in this area.

Author Contributions

E.Y.: initiated the research idea, developed, organized, analyzed and interpreted the data and wrote the manuscript. M.S.: supervised the research, suggested the research methods, structured the paper and edited the manuscript.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Approval

Ethical approval is not required, because this article does not contain any studies with human or animal subjects. Also the data used in the study were obtained from Agriculture and Rural Development Support Institution and the permission and approval of the institution were obtained (Date: April 01, 2004, Approval number: E70805362-622.03-49453).

Acknowledgments

This study was produced from a doctoral thesis.

References

Alpar R. 2003. Introduction to applied multivariate statistical methods: I, Nobel, Ankara, Turkey.

Aneiros-Pérez G, Vieu P. 2008. Nonparametric time series prediction: A semi-functional partial linear modeling. J. Multivariate Anal, 99(5): 834-857.

Aytaç M. 1991. Applied non-parametric statistical tests. Uludag University Press, Bursa, Turkey.

Begun J, Hall W, Huang W, Wellner J. 1983. Information and asymptotic efficiency in parametric-nonparametric models. Annals of Stat, 11: 432-452.

Berry WD. 1993. Understanding Regression Assumptions, Vol. 92. SAGE Publications, London, UK, pp: 104.

Buckley MJ, Eagleson GK, Silwerman GK. 1988. The Estimation of residual variance in nonparametric regression. Biometrika, 75(2): 189-199.

Hurvich CM, Tsai CL. 1989. Regression and time series model selection in small samples. Biometrika, 76(2): 297-307.

Liu J, Zhang R, Zhao W. 2013. A robust and efficient estimation method for single index models. J Multivariate Anal, 122: 226-238.

Mammadov M, Yüzer AF, Aydın D. 2005. Splayn correction regression and correction parameter selection. 4th Statistics Congress proceedings and poster abstracts book, Belek-Antalya, September 25-28, 2005, pp: 148-149.

Newey WK. 1989. The Asymptotic variance of semiparametric estimators. princeton university. Econometric Res Program Memo, No: 346.

Omay RE. 2007. Roughness Penalty approach in regression. PhD Thesis, Anadolu University, Institute of Science and Technology, Department of Statistics, Eskisehir, Turkey, pp: 129.

Speckman P. 1988. Kernel smoothing in partially linear model. J Royal Stat Soc B, 50: 413-436.

Schennach SM. 2004. Nonparametric regression in the presence of measurement error. Econometric Theory, 20: 1046-1093.

Schimek MG. 2000. Estimation and inference in partially linear models with smoothing splines. J Stat Plan Infer, 91: 525-540.

Shi X. 2009. Applications of nonparametric and semiparametric methods in economics and finance. PhD Thesis, Economics in the Graduate School of Binghamton University, New York.

Tezcan N. 2011. Non-parametric regression analysis. Atatürk Univ J Econ Admin Sci, 25: 341-352.

Toprak S. 2015. Semi-parametric regression models with measurement errors. PhD Thesis, Dicle University, Institute of Science, Department of Mathematics, Diyarbakır, Turkey, pp: 98.

Turanlı M, Bağdatlı KS. 2012. Determining the factors affecting

the flat prices in the site by semiparametric regression analysis. Istanbul Commerce Univ J Soc Sci, 11(21): 383-402.

Yatchew A. 2003. Semiparametric regression for the applied econometrician. Cambridge University Pres, Cambridge, UK, pp: 213.

Zhongyi Z, Baocheng W. 2001. Dianostic and influence analysis for semiparametric nonlinear regression models. Acta Math Appl Sinica, 24(4): 568-581.