Investigation of House Price in Eskisehir (Turkey) by Using Semiparametric Additive Regression Model

RABIA ECE OMAY, DURSUN AYDIN
Department of Statistics
Anadolu University
Eskisehir 26470
TURKEY

Abstract: - In this paper, different regression models are obtained to examine relation between house price and house features in Centrum of Eskisehir Province (Turkey). The statistical analyses of this paper indicate that some of explanatory variables affect the response variable parametrically and some of them nonparametrically. Therefore, obtained suitable model has both parametric and nonparametric variables and the model is semiparametric additive regression model. It has concluded that this model has given better results than linear model.

Key-Words: - Additive model, Semiparametric additive model, Smoothing splayn, Backfitting, Deviance

1 Introduction

In this paper, semiparametric additive regression model, which includes parametric and nonparametric components, is discussed [1], [4]. Smoothing spline is used for estimating such a model [3], [4]. Generally, choosing the suitable (optimal) smoothing parameters and estimating smoothing spline predictors by smoother matrix are main tools for additive models. The predictors for the smoothing spline are obtained by backfitting algorithm which is used frequently [3], [4]. For choosing smoothing parameters, Generalized Cross Validation method (GCV) which is automatic method is frequently used. On the other hands, in the additive models degrees of freedom measures for selecting the smoothing parameters are used instead of GCV, because GCV is less reliable and it's needed to select several smoothing parameters simultaneously [4].

The paper is structured as follows. In the next section, smoothing spline is discussed for additive and semiparametric models. Also estimating equations and backfitting algorithm is discussed. In the third section, house characteristics, which affect house price in Centrum of Eskisehir Province (Turkey), are examined. For this aim, linear model and semiparametric additive model are built. Statistical analyses made for this study indicate that two and more variables of the model affect response variable nonparametrically while most of the explanatory variables affect parametrically. Therefore, it is concluded that additive regression model with both parametric and nonparametric variables is suitable (best) regression model for data of house price. Then, it is also concluded that the best (suitable) semiparametric additive model better than linear model.

2 Estimating Equations for Additive Regression Models

An *additive regression model* [4] is defined by

$$y_i = \sum_{j=1}^{p} f_j(x_{ji}) + \varepsilon_i,\ i = 1,...,n,\ \ \varepsilon_i \sim N(0,\sigma^2)$$
$$\text{or}\ \ \mathbf{y} = \sum_{j=1}^{p} \mathbf{f}_j + \boldsymbol{\varepsilon} \qquad (1)$$

Where, f_j's are arbitrary univariate functions. Also $\mathbf{f}_j$ are the n-vectors $\mathbf{f}_j = (f_j(x_{j1}),..., f_j(x_{jn}))^T$, $j = 1,2,...,p$ with x_{ij} in the order y_i.

A *semiparametric additive regression model* is defined by

$$y_i = \mathbf{z}_i^T\boldsymbol{\beta} + f_1(x_{1i}) + ... + f_p(x_{pi}) + \varepsilon_i,\quad i = 1,2,...,n$$
$$\text{or}\quad \mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \sum_{j=1}^{p} \mathbf{f}_j + \boldsymbol{\varepsilon} \qquad (2)$$

with *p*-nonparametric components. Eq.(2) can be also sensed as additive regression model, because parametric part of Eq.(2) is summation of linear functions.

If Eq.(2) has only one f component, the model converts to *semiparametric regression model* [3].

When *smoothing spline* is used to estimate Eq.(1), over all twice continuously differentiable functions $f_j \quad j=1,2,...,p$ *generalized penalized least-squares* criterion is considered as follows.

$$\sum_{i=1}^{n}\left\{ y_i-\sum_{j=1}^{p} f_j(x_{ij}) \right\}^2+\sum_{j=1}^{p}\lambda_j\int\left\{ f_j''(x) \right\}^2 dx \quad (3)$$

Each function f_j in Eq.(3) is penalized by a separate constant smoothing parameter

Analogously single-predictor case, Eq.(3) can be written as follows [4].

$$\left(\mathbf{y}-\sum_{j=1}^{p}\mathbf{f}_j \right)^T\left(\mathbf{y}-\sum_{j=1}^{p}\mathbf{f}_j \right)+\sum_{j=1}^{p}\lambda_j\mathbf{f}_j^T\mathbf{K}_j\mathbf{f}_j \quad (4)$$

Where the $\mathbf{K}_j$'s are *penalty matrices* for each predictor and they are defined analogously to the $\mathbf{K}$ for a single predictor. Eq.(4) is a square form with respect to $\mathbf{f}_j$, $j=1,2,...,p$ vectors. If we differentiate Eq.(4) with respect to the function $\mathbf{f}_j$, we obtain $np\times np$-system which is called *estimate equations* as follow [4].

$$\begin{pmatrix} \mathbf{I} & \mathbf{S}_1 & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \mathbf{S}_2 & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \mathbf{S}_p & \cdots & \mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_p \end{pmatrix}=\begin{pmatrix} \mathbf{S}_1\mathbf{y} \\ \mathbf{S}_2\mathbf{y} \\ \vdots \\ \mathbf{S}_p\mathbf{y} \end{pmatrix} \quad (5)$$

Where $\mathbf{S}_j=\mathbf{S}_{\lambda_j}=\left(\mathbf{I}+\lambda_j\mathbf{K}_j\right)^{-1}$ is the suitable *smoother matrix*. Eq.(5) shortly can be written as $\hat{\mathbf{P}}\mathbf{f}=\hat{\mathbf{Q}}\mathbf{y}$.

Eq.(5) reflect that each of $\hat{f}_k$ estimating function obtained by solving Eq.(3) is a cubic spline which is determined by linear smoother:

$$\hat{\mathbf{f}}_k=\mathbf{S}_k\left(\mathbf{y}-\sum_{j\neq k}\hat{\mathbf{f}}_j \right),\ k=1,2,...,p \quad (6)$$

Eq.(5) and Eq.(6) are equivalent systems. Let $\mathbf{f}_j^0$, $j=1,\ 2,\ ...,\ p$ be starting cases. Eq.(6) is suitable form of Eq.(5) in order to be performed *backfitting algorithm* which is obtained by *Gauss-Seidel* procedure. The backfitting algorithm can be written as follows.

$$\mathbf{f}_k^{(m+1)}=\mathbf{S}_k\left(\mathbf{y}-\sum_{j=1}^{k-1}\mathbf{f}_j^{(m+1)}-\sum_{j=k+1}^{p}\mathbf{f}_j^{(m)} \right) \quad (7)$$

$$k=1,2,...,p,\ m=0,1,2,...$$

For Eq.(2), semiparametric additive model, initial point of k in Eq(7) is zero: $k=0,1,2,...,p$. Therefore, matrix $\mathbf{S}_0=\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ is smoother matrix for parametric part. Also $\mathbf{f}_0=\mathbf{Z}\boldsymbol{\beta}$ is predictor of parametric term. Convergence of Eq(7) is discussed by A. Buja, T.J. Hastie and R.J Tibshirani (1989) and T.J. Hastie and R.J Tibshirani (1999) [1], [4].

2.1 Inferences

In this paper, the three main tools of inferences are deviance, degrees of freedoms and choosing smoothing parameters.

Deviance: One way of testing adequacy of an interesting model and comparing the model is to compare the interesting model with *saturated model*, which has the maximum number of parameters that can be estimated. Let $l(\mathbf{b}_{\max},\mathbf{y})$ and $l(\mathbf{b},\ \mathbf{y})$ denote the maximum value of the likelihood function for the satturated model and interesting model, respectively. Where $\boldsymbol{\beta}_{\max}$ is parameter vector of saturated model and $\mathbf{b}_{\max}$ denotes the maximum likelihood estimator of $\boldsymbol{\beta}_{\max}$. Then deviance can be denoted as follows.

$$D(\mathbf{y};\mathbf{b})=2\left\{ l(\mathbf{b}_{\max};\mathbf{y})-l(\mathbf{b};\mathbf{y}) \right\} \quad (8)$$

Eq.(8) has an approximately chi-square distribution. It was called *deviance* by Nelder and Wedderburn (1972) [2].

The deviance plays the role of the *residual sum of squares* for generalized models and can be used for testing goodness-of-fit and comparing models [2]. It is evaluated that, model with smallest deviance value is the best model within the all present models.

For nonparametric and additive models, the deviance is used to assess models and their differences. The distribution theory is not developed, but chi-square distribution still is used to compare models as the reference distribution [4].

Degrees of freedoms: The *effective numbers of parameters* or *degrees of freedoms* (df) of a smoother are used to compare different smoothers and models. In actually, it is possible to choose the value of smoothing parameter by specifying the degrees of freedoms for the smooth [4]. For non-parametric regression model with one variable, degrees of

freedoms is the trace of a smoother matrix, say $\mathbf{S}_\lambda$: $df = tr(\mathbf{S}_\lambda)$. Analogously, for nonparametric model with more than one variable, *overall degrees of freedoms* can be defined as follows: $df = tr(\mathbf{R}_\lambda)$. Where $\mathbf{R}_\lambda$ is the smoother matrix that produces $\hat{\mathbf{f}}_+ = \mathbf{R}_\lambda \mathbf{y}$. $\hat{f}_+$ is sum of predictor vectors: $\hat{\mathbf{f}}_+ = \sum_{j=1}^{p} \hat{\mathbf{f}}_{\lambda_j}$ [4].

Selecting of smoothing parameters: Theoretically, selecting methods of smoothing parameters (GCV-Generalized Cross Validation, AIC-Akaike Information Criteria, etc.) that used for functions with single variable, can be also used for additive models. Classical selecting models, like GCV and AIC especially, are designed in order to select $\lambda_j, j = 1,..., p$ smoothing parameters [7]. In the additive model with p-terms, it is accepted that *p*-smoothing parameters λ_j make optimization of a criteria as GCV [4].

Practically, a suitable model is selected by changing degrees of freedoms, because selecting of two or more smoothing parameters simultaneously is too difficult and these parameters relation with degrees of freedoms directly.

3 Application

It is known that some explanatory variables affect house prices. In this application, it is examined that how these variables affect house prices. Shape of the variables on house prices (parametrically or nonparametrically) cannot be known. So, in this application, a suitable *semiparametric additive regression mode*l is constructed by using these variables and then if this model good or not is investigated

We collect data from information of 105 apartments for sale in Centrum of Eskisehir Province (Turkey) at 2006, May and June. The variables of the data set are defined as follows.

Price : House Price (YTL)
Age : Age of houses
Area : Using area of houses (m^2)
Garage : Dummy variable for garage
Elevator: Dummy variable for elevator
Park : Dummy variable for park
Erdgas : Dummy variable for erdgas
Doorkeeper: Dummy variable for doorkeeper
Car park: Dummy variable for car park

For these variables, suitable *semiparametric additive regression model* is constructed. On the other hand *linear regression model* that contains all variable parametrically is also constructed. Estimating results obtained from the semiparametric additive model are compared with the results of the linear model.

R and S-plus programmes are used for statistical analyses.

Details of semiparametric additive regression model: In constructing suitable semiparametric additive model, selection of smoothing parameter or equivalently degrees of freedom is one of the most problems. In this application, several models, whose nonparametric variables have different values of degrees of freedoms, are constructed. Then suitable (or the best) model is selected from between these models and the results for this selected suitable model are listed Table 1.

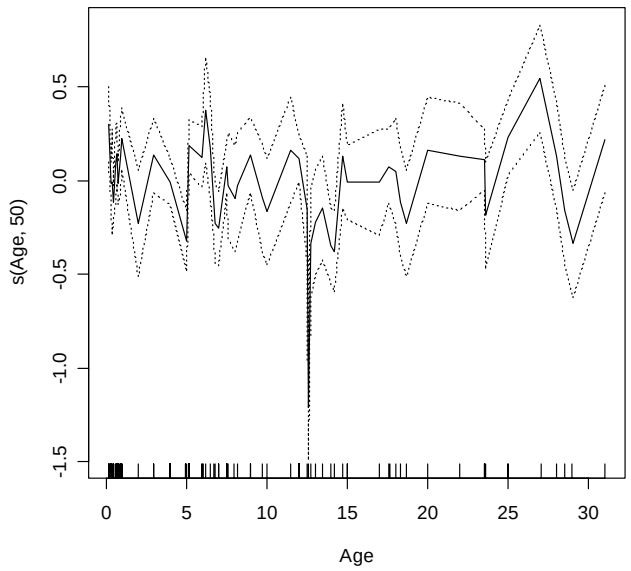
Table 1. Semiparametric additive regression model results.

	Parametric Part			
	Estimate	St. Er.	t-ist.	Pr (> $\lvert t \rvert$)
(Intercept)	10.1987	0.05999	170.01211	3.88587e-55
Garage	0.0796	0.03771	2.11203	4.14965e-02
Elevator	-0.1056	0.03772	-2.79872	8.10099e-03
Park	-0.1471	0.04394	-3.34829	1.87885e-03
Erdgas	0.0827	0.03105	2.66274	1.14053e-02
D.keeper	0.3488	0.03353	10.40233	1.54878e-12
Car park	-0.1548	0.03094	-5.00234	1.40050e-05
	Nonparametric Part			
	Sd Npar	Sd Npar	F	Pr (F)
s(Age)	1	49	2.8593	0.0006184***
s(Area)	1	9	3.6703	0.0023109**
Dependent var.: log (Price) Deviance = 0.7494 R^2=0.97097				

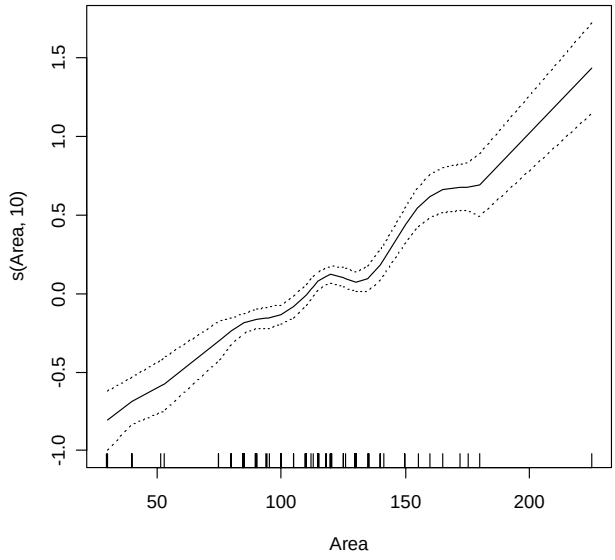
Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

When Table1 is considered, it's appeared that both parametric variables ("Garage", "Elevator", "Park", "Erdgas", "Doorkeeper" and "Car park") and nonparametric variables ("Age" and "Area") have statistically significant effects. Square of multiple correlation coefficient is determined as 0.9707 for the suitable semiparametric additive model. In other words, the suitable semiparametric additive model can explain 97.07% of the total variation in the data.

The nonparametric variables in nonparametric part of the suitable semiparametric additive model can be only displayed graphically, because they can't be expressed parametrically. Effects of these nonparametric variables are shown Figure1.



(a)



(b)

Figure 1: (a) Changing of house prices with respect to their ages and 95% confidence interval. **(b)** Changing of house prices with respect to their areas and 95% confidence interval.

Additionally, linear regression model, which has all explanatory variables parametrically, is also constructed. Then, these two models are compared according to their deviance and R^2 criteria. Several variables ("Age", "Garage", "Elevator", "Car park"), which have significant effects in the suitable semiparametric additive model, have not significant effects in the linear model. Especially, it is recognized that, "Age" variable has significant effect nonparametrically in the semiparametric additive model; however, it has no effect in linear model. Because of these insignificant variables, backward model selection method can be used for a better new linear model (reduced model) with significant variables. Backwards model selection method is performed by repeatedly removing the single term with highest p-value, above some threshold (e.g. 0.05), and than refitting the resulting reduced model, until all terms have significant p-values [6]. Results of the reduced linear model are shown in Table 2.

Table 2. Reduced linear regression model results.

	Estimate	St. Error	t-value	Pr (> $\|t\|$)
(Int.)	10.22687	0.07239	141.267	< 2e-16 ***
Area	0.00808	0.00065	12.437	< 2e-16 ***
Park	-0.12625	0.05774	-2.186	0.031147 *
Erdgas	0.14808	0.04096	3.616	0.000473 ***
D.keeper	0.20502	0.04324	4.742	7.12e-06 ***
R^2= 0.7082 Deviance= 3.669647				

Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

In Table 2, the reduced linear model contains only "Area", "Park", "Erdgas" and "Doorkeeper" variables. On the other hands, the suitable semiparametric additive model contains also "Age", "Garage", "Elevator", "Car park" variables as well as "Area", "Park", "Erdgas" and "Doorkeeper" variables. Therefore, if we construct the linear model instead of the semiparametrik additive regression model, we maybe ignore some variables with significant effect.

Afterwards, R^2 and deviance values of both models are compared, if the semiparametric additive model is better than the reduced linear model or not. The results of the comparison are listed Table 3.

Table 3. Some results of semiparametric additive and reduced linear model.

Model	R^2	Deviance
Semiparametric Additive Model	0.97097	0.74940
Linear Model	0.70820	3.66965

Table3 shows that, R^2 value for semiparametric additive model (0.97097) is bigger than R^2 value for reduced linear model (0.70820) while deviance value of semiparametric additive model is less than the other model. Therefore, we can say that, the semiparamteric additive model is better than the reduced linear model.

4 Conclusion

In this paper, relationship between house price and apartment features is examined by using linear and semiparametric additive regression models. It's concluded that, semiparametric additive model with both nonparametric and parametric variables gives

better results than the linear model. This conclusion emphasize that, estimates which base on a method like smoothing spline, are better than the traditional methods, as a linear regression. Additionally, It's seen that one of the most subject for regression analysis is to construct a suitable semiparametric additive regression model by specifying nature of explanatory variables which affect response variables.

References:

[1] Buja, A., Hastie, T. and Tibshirani, R., Linear Smoothers and Additive Models, *The Annals of Statistics*, Vol.17, No.2, 1989, pp.453-555.

[2] Dobson, A.J., *An Introduction to Generalized Linear Models*, Chapman&Hall/CRC, 2002.

[3] Green, P.J. and Silverman, B.W., *Nonparametric Regression and Generalized Linear Models*, Chapman&Hall, 1994.

[4] Hastie, T.J. and Tibshirani, R.J., *Generalized Additive Models*, Chapman&Hall/CRC, 1999.

[5] Schimek, M.G., *Smoothing and Regression*, John Wiley&Sons, 2000.

[6] Wood, S.N., *Generalized Additive Models: An Introduction with R*, Chapman&Hall, 2006.

[7] Wood, S.N., Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties, *Journal of the Royal Statistical Society: Series B*, Vol.62,Part 2, pp., 2000, 413-428.