STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 515–524

LOCALLY REGULARIZED LINEAR REGRESSION IN THE VALUATION OF REAL ESTATE

Mariusz Kubus [1]

ABSTRACT

Regression methods are used for the valuation of real estate in the comparative approach. The basis for the valuation is a data set of similar properties, for which sales transactions were concluded within a short period of time. Large and standardized databases, which meet the requirements of the Polish Financial Supervision Authority, are created in Poland and used by the banks involved in mortgage lending, for example.
We assume that in the case of large data sets of transactions, it is more advantageous to build local regression models than a global model. Additionally, we propose a local feature selection via regularization. The empirical research carried out on three data sets from real estate market confirmed the effectiveness of this approach. We paid special attention to the model quality assessment using cross-validation for estimation of the residual standard error.

Key words: large transactional data, local regression, feature selection, regularization, cross-validation.

1. Introduction

Objective and highly detailed valuation of real estate supports business decisions and the risk associated with these decisions. It affects the proper functioning of lending activities of the banks involved in mortgage financing of real estate. In newly amended Recommendation J, the Polish Financial Supervision Authority has stipulated the rules of collecting and processing data of real estate by banks. The recommendation pays special attention to the use of a reliable and standardized database. There are such databases in Poland, which are created and offered for commercial use. They include transaction prices, locations and property characteristics, which are obtained from several sources: notarial deeds, county authorities, valuations and vision of local experts. An example of this is the system for Analysis and Monitoring of Real Estate Market (AMRON),

[1] Department of Mathematics and Applied Computer Science, Opole University of Technology. E-mail: m.kubus@po.opole.pl

which was initiated by the Polish Bank Association. This interbank and standardized database contains a complex scope of information about Polish real estate market. Until now, AMRON have collected over a million records, whereas E-Valuer database offers information about real estate sale prices from across Poland. It contains over 400,000 records.

The legal basis for the valuation of real estate in Poland are: Real Estate Management Act of August 21, 1997, and Regulation of the Council of Ministers of September 21, 2004 on the valuation of real estate and preparing appraisal report. These regulations do not contain specific calculation procedures. They show only certain principles that should be followed. Some models of valuation are published by the Polish Federation of Real Estate Appraiser Associations in the Universal of National Valuation Rules and in the Interpretative Notes.

One of the four approaches to the valuation task is a comparative approach. The basis for the valuation is a data set of similar properties for which sale transactions were concluded within a short period of time (usually two years). Apart from the information about the transaction prices, real estate appraiser has also information on selected characteristics of the properties (so-called market features). It is assumed that they affect the level of prices. Thus, the valuation can be formulated as a regression task. The transaction database plays the role of the training set with multidimensional observations. The market features are the predictors, and price is the dependent variable. The goal is to predict the value of a new object. The Interpretative Note No. 1 "The application of a comparative approach in the valuation of real estate" specifies two of the three valuation methods: paired comparison and average price adjustment. The third one - the method of statistical analysis of the market – is not clarified, which gives the analyst freedom to use various econometric models. Recommendation J of the Polish Financial Supervision Authority, aimed at the banking sector, pays attention to the need of accurate and stabile models.

In the literature there are several propositions of application of the market statistical analysis. Undoubtedly, the most popular are classical multiple regression model and multiple simple regression model (Hozer 2008). Foryś (2010) applied linear ordering methods in order to select a subset of objects most similar to the valued real estate. Linear ordering methods were applied to construct the aggregated measure of attractiveness, which was used as the predictor in simple regression (Lis 2005; Doszyń 2012). Lis (2005) used information on the cluster structure of the data acquired by k-means method. The dummy variables which identified the clusters were included in the linear regression model. Mach (2012) investigated the impact of regional development on the price of a square meter of residential real estate using factor analysis, and multiple regression in the space of reduced dimension. Trzęsiok (2013) compared various nonparametric regression models on the data from the Warsaw housing market. In the literature on real estate valuation, propositions of neural networks application can also be found in (Lis 2001, Morajda 2005).

In this paper we propose the application of locally estimated regression functions in the neighbourhood of the points that represent the valued real estate. This approach is dedicated to the large data sets of transactions. Additionally, we propose a local feature selection, assuming that the validity of the feature can differ in various regions of the feature space. Feature selection is made with a use of regularization in the estimation criterion. Unfortunately, we do not have access to the large data set from the Polish market. Thus, the proposition of locally regularized linear regression was verified on three publicly accessible data sets from the US market.

2. The estimates of the linear model coefficients

Suppose we are given a set of multivariate observations with known values of a quantitative dependent variable *Y* (training set):

$$U = \left\{(\boldsymbol{x}_1, y_1),...,(\boldsymbol{x}_N, y_N) : \boldsymbol{x}_i \in \boldsymbol{X} = (X_1,...,X_p), y_i \in Y, i \in \{1,...,N\}\right\}. \quad (1)$$

The goal of regression is to discover the impact of the predictors $X_1,...,X_p$ on the variable *Y*. Owing to its simplicity, a linear model is widely used:

$$f(\boldsymbol{x}) = b_0 + b_1 x_1 + ... + b_p x_p + \varepsilon, \quad (2)$$

where ε is the random component, which is assumed to have a normal distribution with mean equal to 0 and constant variance $\forall i \in \{1,...,N\}$. Moreover, ε_i are independent of each other and independent of predictors. The classical approach to the estimation of the linear model parameters is the ordinary least squares method (OLS). Estimators obtained in this way, under the Gauss-Markov theorem, are unbiased and have the smallest variance in the class of linear and unbiased estimators (Maddala 2008, pp. 228–229). A problem arises when there are irrelevant variables in the data, which do not affect the variable *Y*. Then the model does not guarantee an accurate prediction for new objects (out of training sample), which is the primary objective of modelling. It has a theoretical foundation in the bias-variance trade-off and formally it is described in (Hastie, Tibshirani, Friedman 2009, pp. 219–224). Generally, too complex models, which are well fitted to the training set, are characterized by a low bias and a high variance of the prediction error. On the other hand, too simple models, which do not extract all information from data, are characterized by a high bias and a low variance of the prediction error. The number of parameters to estimate (which is strongly related to the number of variables) is usually adopted as a measure of the linear model complexity. Thus, elimination of the irrelevant variables should improve the quality of the model. The market features in real estate valuation task are carefully selected by experts, and it would seem that they undoubtedly have an impact on the transaction prices. However, many authors indicate the need for

formal, statistical feature selection in the econometric models of valuation (Lis 2005; Zeliaś 2006; Bitner 2007).

The effect of feature selection can be achieved by regularization. A penalty component $P(\boldsymbol{b})$ for large absolute values of the parameters is imposed on the criterion used in the estimation task:

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \left(\sum_{i=1}^{N} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} \right)^2 + \lambda \cdot P(\boldsymbol{b}) \right). \tag{3}$$

Regularization decreases the absolute values of the estimates, and some of them are equal to zero. The core idea is to control the complexity of the model. The objective is to reach a compromise between the bias and the variance, and consequently to get a model with optimal generalization ability, which means the accuracy of prediction for new objects. Various forms of penalty component and the references are given in the Table 1.

Table 1. Penalty components in regularized linear regression

Ridge regression	LASSO	Elastic net
$P(\boldsymbol{b}) = \sum_{j=1}^{p} b_j^2$	$P(\boldsymbol{b}) = \sum_{j=1}^{p} \lvert b_j \rvert$	$P_\alpha(\boldsymbol{b}) = \sum_{j=1}^{p} \left(\alpha b_j^2 + (1-\alpha) \lvert b_j \rvert \right)$
(Hoerl, Kennard 1970)	(Tibshirani 1996)	(Zou, Hastie 2005)

Source: own work.

The regularization parameter λ defines the amount of the penalty, and, as a result, it controls the complexity of the model. Determination of the appropriate value of λ is a key task for the effective application of this method. In practice, a sequence of models corresponding to different values of λ is built, and then the optimal model is selected. Information criteria or prediction error estimated via cross-validation can be used as a model selection criterion. The empirical comparison of these criteria can be found in (Kubus 2013).

The task of parameter estimation in the ridge regression has a solution in a closed form, see, i.e. (Maddala 2008; Hastie *et al.* 2009). LASSO requires quadratic programming with linear constrains but approximate solutions are more practical and more commonly used. Presently, LARS algorithm (Efron, Hastie, Johnstone, Tibshirani 2004) is the most popular because of a low computational complexity. In the case of the elastic net it has been proven that the estimation task can be reformulated on the LASSO (Zou, Hastie 2005). In the following iterations of the LARS algorithm, the coefficients are updated based on the current regression residuals, thus previously unexplained variability of the response *Y*. In every step, the updating formula takes into account some predictors

most correlated with *Y*, while only one predictor per iteration is introduced into the model.

An important modelling stage is to assess the quality of th model. A widely used evaluation function is the mean square error, whose unbiased estimator has the form of:

$$MSE = \frac{1}{N - p - 1}\sum_{i=1}^{N}\left(y_i - \hat{f}(\boldsymbol{x}_i)\right)^2, \tag{4}$$

or its square root called residual standard error. However, the assessment of the predictive ability of a model, measured on the training set which was used for estimation of model parameters, is too optimistic (Hastie *et al.* 2009, p.228). A model fitted perfectly to the data does not guarantee a great ability to generalize, that is an accurate prediction for new objects (out of the training sample), which is the primary objective of modelling. To estimate the prediction error, the researcher should use a separate set of objects, from the same population, that did not take part in the learning stage. Cross-validation is quite a common strategy of the error estimation. The main idea of cross-validation is to reuse the learning sample many times. In this method, the training set *U* is split into *K* disjoint and approximately equinumerous subsets $V_1, V_2, ..., V_K$. Then *K* models ($\hat{f}_1, ..., \hat{f}_K$) are built based on training samples $U_i^{CV} = U - V_i$ ($i = 1, ..., K$), and the prediction errors are estimated based on test samples V_i. Finally, the error is averaged.

3. Local regression models

The assumption about the linear dependency between predictors and the dependent variable is very restrictive. There are numerous propositions of more flexible regression functions in the literature. A short review of non-parametric regression models with the examples of applications in R program is given in (Trzęsiok, Trzęsiok 2009). Some of these methods utilize the idea of local fitting, for example tree-based models. We take under consideration a slightly different approach, which also focus on local fitting. Consider a point $\boldsymbol{x} \in \boldsymbol{X} = (X_1, ..., X_p)$, which represents the valued real estate. Instead of building a global model (in all the domain) we estimate the linear regression function in the neighbourhood of $\boldsymbol{x}$. Additionally, one can use information about distances between points $\boldsymbol{x}_i \in U$ and a query point $\boldsymbol{x}$. Formally, a model of local regressions can be expressed by the formula:

$$\tilde{f}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}, \hat{\boldsymbol{b}}(\boldsymbol{x})), \tag{5}$$

where $\hat{f}$ is the linear model (2) fitted locally in the neighbourhood of $\boldsymbol{x}$. The parameter vector $\boldsymbol{b}$ is estimated with the use of weighted least squares:

$$\hat{\boldsymbol{b}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} K\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|}{h(\boldsymbol{x})}\right)\left(y_i - \hat{f}(\boldsymbol{x}_i, \boldsymbol{b})\right)^2 . \tag{6}$$

The weights are obtained by a kernel function $K(\cdot)$, where $h(\boldsymbol{x})$ is a width function that determines the width of the neighbourhood at $\boldsymbol{x}$. Various forms of kernel functions as well as various techniques of determining the width can be found in the literature (Loader 1999; Hastie *et al.* 2009). A simple solution is to use *k* nearest neighbours method, which adaptively sets the width due to the local density of objects from the training set. Still, one can use weights dependent on the distances from $\boldsymbol{x}$, or adopt equal weights for all objects from the neighbourhood of $\boldsymbol{x}$. Note that the models are fitted separately at each query point $\boldsymbol{x}$.

Assuming that the validity of the feature can differ in various regions of the domain, we propose to introduce the regularization to the criterion of estimation. The penalty component that we have chosen is the elastic net (Zou, Hastie 2005). We have also adopted the simplest system of weights, applying *k* nearest neighbours kernel function. Formally, it can be written as:

$$\hat{\boldsymbol{b}}^{\boldsymbol{l\text{-}EN}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} K_{k-NN}\left(\|\boldsymbol{x} - \boldsymbol{x}_i\|\right)\left(\left(y_i - \hat{f}(\boldsymbol{x}_i, \boldsymbol{b})\right)^2 + \lambda \sum_{j=1}^{p}\left(\alpha b_j^2 + (1-\alpha)|b_j|\right)\right) , \tag{7}$$

where $K_{k-NN}(\cdot)$ is the indicator function, which returns equal weights for objects from the neighbourhood of $\boldsymbol{x}$, and 0 otherwise.

4. Empirical study

To compare the effectiveness of the global and local modelling of a linear regression function we analyse three publicly available data sets from the real estate market.

The first one concerns housing values in suburbs of Boston (Harrison, Rubinfeld 1978). The market features that are taken into account are {*per capita crime rate by town; proportion of residential land zoned for lots over 25,000 sq. ft.; proportion of non-retail business acres per town; Charles River* dummy variable (= 1 if tract bounds river; 0 otherwise)*; nitric oxides concentration (parts per 10 million); average number of rooms per dwelling; proportion of owner-occupied units built prior to 1940; weighted distances to five Boston employment centres; index of accessibility to radial highways; full-value property-tax rate per $10,000; pupil-teacher ratio by town; 1000(Bk – 0.63)^2 where Bk is the*

proportion of blacks by town; % lower status of the population}. This data set contains 506 observations (https://archive.ics.uci.edu/ml/datasets.html).

The second data set consists of aggregated data from each of 20,460 neighbourhoods in California. It is available in the StatLib repository (http://lib.stat.cmu.edu/). The dependent variable is the median house value in each neighbourhood. The predictors are {*median income, housing median age, total rooms, total bedrooms, population, households, latitude, longitude*}.

The last data set comes from (Maddala 2008, p. 234-235), and it concerns sale prices of rural land in Florida (per acre). There are 67 multidimensional observations, which are characterized by market features/predictors {*proportion of acreage that is wooded; distance from parcel to Sarasota airport; distance from parcel to highway; acreage of parcel; month in which the parcel was sold*}. Although it is not large data set we decide to take it under consideration to have a wider field for the comparisons.

We use two estimation techniques in the global regression model as well as in local regressions. The ordinary least squares method is compared with regularization, where we apply the elastic net penalty component (Zou, Hastie 2005). The optimal value of the regularization parameter λ has been chosen with the use of Bayesian information criterion BIC. The number of nearest neighbours in the model of local regressions, estimated according to formula (6) or (7), has been set arbitrarily. It is $k = 30$ in Florida and Boston data set, and $k = 50$ in the third, larger data set. The residual standard errors for these models have been estimated via 10 fold cross-validation. The comparison is shown in the Table 2.

Table 2. Residual standard errors estimated via cross-validation for global and local fitting

Global linear model

	Florida	Boston	California
OLS	2537.71	4719.15	69581.07
Elastic net	2538.45	4887.26	73979.68

Local linear model

	Florida	Boston	California
OLS	2218.81	3719.58	64044.36
Elastic net	2154.19	3475.37	66036.35

Source: own calculations.

The prediction error of local regression models is lower in all the cases. In two out of three data sets (Florida and Boston) locally regularized regression has given better results. Note that the regularization has not affected the reduction of error in global models. This allows us to suppose that the validity of variables varies in different regions of the feature space.

5. Summary

We found the idea of the local fitting of regression function natural and especially suited to the comparative approach in the real estate valuation. Empirical examples investigated in the previous section confirmed our intuition. The core idea of the comparative approach is to use the information about similar properties that were sold in a short period of time. The use of a distance-based similarity measure in the feature space enables fully automatic and objective identification of such real estate. A special meaning is assigned to it in the analysis of large data sets. Our approach improved the accuracy of the valuation in two data sets. In the third one we obtained a slightly worse model, but still competitive for a global regression. Poorer performance of California housing data set seems to confirm the correct selection of the potential predictors made by experts. When there are no irrelevant variables in the data set, the unbiased OLS estimators are recommended. Note that locally regularized linear regression can be modified in several ways. One can utilize: various weighting systems, adaptive setting of the parameter k with the use of a validation set, various penalty components, polynomial regression function.

REFERENCES

BITNER, A., (2007). Konstrukcja modelu regresji wielorakiej przy wycenie nieruchomości [Construction of the multiple regression model in real estate valuation], Acta Scientiarum Polonorum, Administratio Locorum, 6 (4), pp. 59–66.

DOSZYŃ, M., (2012). Ekonometryczna wycena nieruchomości [Econometric evaluation of real estate], Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego, No. 26, pp. 41–52, Szczecin.

EFRON, B., HASTIE, T., JOHNSTONE, I., TIBSHIRANI, R., (2004). Least Angle Regression, ,,Annals of Statistics" 32 (2), pp. 407–499.

FORYŚ, I., (2010). Wykorzystanie metod taksonomicznych do wyboru obiektów podobnych w procesie wyceny lokali mieszkalnych [The multivariate analysis using to the choice the similar object in the housing valuation process], Studia i Materiały Towarzystwa Naukowego Nieruchomości, Vol. 18, No. 1, pp. 95–105, TNN, Olsztyn.

HARRISON, D., RUBINFELD, D. L., (1978). Hedonic prices and the demand for clean air, J. Environ. Economics & Management, Vol. 5, 81–102.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer, New York.

HOERL, A. E., KENNARD, R., (1970). Ridge regression: biased estimation for nonorthogonal problems, „Technometrics" 12, pp. 55–67.

HOZER, J., (ed.), (2008). Wycena nieruchomości [Real estate valuation], KEiS US, IADiPG w Szczecinie, Szczecin.

KUBUS, M., (2013). On model selection in some regularized linear regression methods, Acta Universitatis Lodziensis, Folia Oeconomica 285, pp. 115–223.

LIS, C., (2001). Sieci neuronowe a masowa wycena nieruchomości [Neural networks and the mass valuation of real estate], Zeszyty Naukowe US, No 318, Prace Katedry Ekonometrii i Statystyki, Szczecin.

LIS, C., (2005). Ekonometryczne modele cen transakcyjnych lokali mieszkalnych [Econometric models transaction prices of residential premises], Zeszyty Naukowe US, No. 415, Prace Katedry Ekonometrii i Statystyki, No. 16, Szczecin.

LOADER, C., (1999). Local Regression and Likelihood, Springer, New York.

MACH, Ł., (2012). Determinanty ekonomiczno-gospodarcze oraz ich wpływ na rozwój rynku nieruchomości mieszkaniowych [Economic determinants and their impact on development of residential real estate market], Ekonometria, 4 (38), pp. 106–116.

MADDALA, G. S., (2008). Ekonometria [Econometrics], PWN, Warszawa.

MORAJDA, J., (2005). Wykorzystanie perceptronowych sieci neuronowych w zagadnieniu wyceny nieruchomości [The use of perceptrons neural networks in the issue of real estate valuation], Zeszyty Naukowe Małopolskiej Wyższej Szkoły Ekonomicznej w Tarnowie, 7, pp. 101–108.

TIBSHIRANI, R., (1996). Regression shrinkage and selection via the lasso, J.Royal. Statist. Soc. B., 58, pp. 267–288.

TRZĘSIOK, J., TRZĘSIOK, M., (2009). Nieparametryczne metody regresji [Nonparametric regression methods], [in:] M. Walesiak, E. Gatnar (eds), Statystyczna analiza danych z wykorzystaniem programu R [Statistical data analysis with a use of R program], Wydawnictwo Naukowe PWN, Warszawa, pp. 156–192.

TRZĘSIOK, M., (2013). Wycena rynkowej wartości nieruchomości z wykorzystaniem wybranych metod wielowymiarowej analizy statystycznej [Real estate market value estimation based on multivariate statistical analysis], Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, No. 278, Taksonomia 20, Klasyfikacja i analiza danych – teoria i zastosowania, pp. 188–196.

ZELIAŚ, A., (2006). Kilka uwag na temat doboru zmiennych występujących na rynku nieruchomości [Several remarks about the methods of selecting variables occurring on the real estate market], Zeszyty Naukowe US, No 450, Prace Katedry Ekonometrii i Statystyki, No. 17, pp. 685–696, Szczecin.

ZOU, H., HASTIE, T., (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society Series B, 67 (2), pp. 301–320.