DE GRUYTER OPEN

Folia Oeconomica Stetinensia
DOI: 10.1515/foli-2016-0023

WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

ASSESSMENT OF PREDICTOR IMPORTANCE WITH THE EXAMPLE OF THE REAL ESTATE MARKET

Mariusz Kubus, Ph.D.

Opole University of Technology
Faculty of Production Engineering and Logistics
Institute of Mathematics and Physics
Department of Mathematics and Applied Computer Science
Sosnkowskiego 31, 45-272 Opole, Poland
e-mail: m.kubus@po.opole.pl

Received 15 November 2015, Accepted 16 November 2016

Abstract

Regression methods can be used for the valuation of real estate in the comparative approach. However, one of the problems of predictive modelling is the presence of redundant or irrelevant variables in data. Such variables can decrease the stability of models, and they can even reduce prediction accuracy. The choice of real estate's features is largely determined by an appraiser, who is guided by his/her experience. Still, the use of statistical methods of a feature selection can lead to a more accurate valuation model. In the paper we apply regularized linear regression which belongs to embedded methods of a feature selection. For the considered data set of real estate land designated for single-family housing we obtained a model, which led to a more accurate valuation than some other popular linear models applied with or without a feature selection. To assess the model's quality we used the leave-one-out cross-validation.

Keywords: valuation of real estate, multiple regression model, cross validation, feature selection, regularization

JEL classification: C52, R39

Introduction

The basis for real estate valuation in the comparative approach is the information about similar real estate transactions made on the local market within a short period of time (usually two years). In estimating the market value of real estate, apart from the transaction prices the characteristics of similar real estate are taken into account, assuming that they have significant impact on prices. All methods included in the comparative approach (paired comparison, average price adjustment and statistical analysis of the market) are based on the introduction of an adjustment of the average transaction price. The adjusting components in the corresponding formulas contain weighting coefficients for the features, which are defined in different ways depending on the method (Prystupa, 2001; Hozer, Kokot, Kuźmiński, 2002). In the literature there are several propositions of the application of the market statistical analysis. Probably the most popular are the 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 synthetic 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 the k-means method. Mach (2012) investigated the impact of regional development on the price of a square meter of residential real estate using a factor analysis, and multiple regression in the space of reduced dimension. In the literature on real estate valuation, propositions of the application of neural networks can also be found (Lis, 2001; Morajda, 2005).

The choice of features which characterize real estate is of key meaning in accurate valuation. In addition to the features that a real estate appraiser takes into account in accordance with laws and regulations, he/she has yet to choose a number of features specific to each market. The choice is determined by the experience of an appraiser and his knowledge of the local market. However, when applying an econometric model of valuation, one should be aware of redundant variables or irrelevant variables' impact on it. Such variables can decrease the stability of models, and they can even reduce prediction accuracy. Although the arbitrary choice of variables made by a real estate appraiser is made by his/her "know how", modelling on a reduced number of variables is worth attempting. Such an approach will either give the statistical confirmation of the optimality of the chosen feature subset, or will discover a more accurate valuation model. In the majority of the research on econometric valuation models, feature selection is made according to the criterion of a maximum correlation with a price and

minimum correlation between predictors at the same time. Moreover, the features are eliminated using tests of significance of the corresponding coefficients, or using stepwise regression.

In this paper, we propose an alternative approach to feature selection in the linear regression model. The model parameters are estimated by introducing to the optimized criterion a component which penalizes the complexity of the model (so called regularization). As a result, the absolute values of the estimated parameters become shrunk in relation to the corresponding estimates of the ordinary least squares method, and some of them are equal to zero. The empirical example refers to the real estate market intended for single-family housing, situated in two adjacent districts of Kraków.

1. The Multiple regression model and its assessment

The application of the linear multiple regression model in the valuation of real estate falls within the scope of a market statistical analysis used in the comparative approach. The database of the similar real estate, which a real estate appraiser has, is a collection of historical data on transactions, called in regression the training set:

$$U = \left\{(x_1, y_1), ..., (x_N, y_N) : x_i \in X = (X_1, ..., X_p), y_i \in Y, i \in \{1, ..., N\}\right\} \tag{1}$$

It is used to estimate model parameters. A transaction price is the dependent variable Y, which is treated as a random variable. Objects in the U set are characterized by the explanatory variables X_1, ..., X_p (predictors), which are real estate characteristics in the task of valuation. It is important that the data came from a recent period of time, and that it represents a spatially homogeneous set of real estate (the same city, district, etc.). Generally, the regression task is to discover the impact of the predictors on the quantitative response variable Y (here price). Because of its simplicity a linear model is widely used:

$$y = b_0 + b_1 x_1 + ... + b_p x_p + \varepsilon \tag{2}$$

where ε is the random component, which is assumed to have a normal distribution with the mean equal to 0 and constant variance $\forall i \in \{1, ..., N\}$. Moreover, ε_i are independent of each other and independent of predictors. Despite the quite restrictive assumption about the linear relationship between the variables, this model is often successfully used in practice, and in the case of small samples it often turns out to be more accurate in the prediction than nonlinear – more complex – models. The regression model is used to predict the unknown value of Y (prices) for the new

object (real estate) $\boldsymbol{x}$ for which the features X_1, ..., X_p were observed. It is obvious that for the model to be useful, the values of variable Y should be predicted as accurately as possible. Thus, an important modelling stage is to assess the quality of a model. A widely used evaluation function is the mean square error, whose unbiased estimator has the form of:

$$MSE = \frac{1}{N-p-1}\sum_{i=1}^{N}\left(y_i - \hat{f}\left(\boldsymbol{x}_i\right)\right)^2 \tag{3}$$

or its square root called residual standard error, which is convenient for interpretation, because it is expressed in the same unit as a dependent variable.

However, the assessment of the predictive ability of a model, measured on the training set which was used for the estimation of model parameters, is too optimistic (Hastie, Tibshirani, Friedman, 2009, p. 228). A model that fits very well to the data does not guarantee the high ability of generalization, that is accurate prediction for new objects (out of a training sample), which is the primary objective of modelling. To estimate the prediction error, a 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 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 and based on training samples $U_i^{CV} = U - V_i$ (i = 1, ..., K), and the prediction errors are estimated and based on test samples V_i. Finally, the error is averaged. For really small training sets it is often assumed that the sets V_i are singletons (so called leave-one-out cross-validation). Then:

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

where $\hat{f}^{-i}\left(\boldsymbol{x}_i\right)$ is a fitted value of the model built on the basis of a training set without the i-th object. This measure is also used for the comparison of a model (Maddala, 2008, p. 531).

The classical approach to the estimation of the model parameters (2) is the ordinary least squares method (OLS). Estimators obtained in this way, under the Gauss-Markov theorem, are unbiased and they have the smallest variance in the class of linear and unbiased estimators (Maddala, 2008, pp. 228–229). When there are irrelevant variables in a data set, a model does not guarantee an accurate prediction for the new objects (out of the sample). In that case, a model reflects not only the systematic effect of predictors on the response, but also a noise. It has a theoretical foundation in the bias-variance trade-off and formally it is described in (Hastie et al., 2009, pp. 219–224). Generally, too complex models, that 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, that 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 is usually adopted as a measure of the linear model complexity. If interactions, or generally additional variables which are functions of the original predictors, are not taken into account, the complexity of the linear model is equivalent to the number of predictors. In an extreme case, a model without explanatory variables, where a prediction is made on the basis of an average value of the variable *Y*, is the simplest one, the errors will have zero variance, but the prediction will most likely be biased.[1] From this perspective, the essence of effective modelling in regression is the choice of an appropriate model complexity, which can be achieved by feature selection.

2. Feature selection

One of the problems of regression function modelling is an excess of information (redundant variables) or irrelevant information (irrelevant variables) for explaining the investigated phenomenon, in our case real estate prices. Such variables affect the reduction of the stability of the model, and they can even decrease prediction accuracy. It would seem that features describing real estate (explanatory variables) are carefully selected by experts, and undoubtedly they have an impact on transaction prices. However, many authors (Lis, 2005; Zeliaś, 2006; Bitner, 2007) point out the need of formal, statistical feature selection in the econometric valuation models.

Feature selection methods are currently classified into three groups: filters, wrappers and embedded methods (Guyon, Gunn, Nikravesh, Zadeh, 2006). The first one is maybe the most popular. Filters discard variables which have little chance to be useful, before an estimation of model parameters. In these methods, a criterion which evaluates the association between the predictors and a response is set. Note that a choice of criterion is strictly heuristic and there is no guarantee that the determined feature subset leads to a model with the highest accuracy of prediction. A criterion can evaluate the individual impact of the explanatory variable on the response (usually Pearson correlation coefficient is used in a valuation task), or it can evaluate the importance of the subset of predictors. In the second case, usually it is postulated that the predictors were maximally correlated with the response and minimally correlated with each other. One can use Hellwig's measure (Hellwig, 1969) or an algorithm that iterates two steps. In the first step, the predictor (say *X*) mostly correlated with the response *Y* is chosen. In the

[1] If only any predictor has a significant impact on the response Y, and the model specification is correct.

second step, other predictors correlated with X are discarded. The decision about introducing a variable into the model or about discarding a variable requires assuming the threshold values of the Pearson correlation coefficient.

In the second group of the feature selection methods (wrappers), the models are used which are built for various combinations of explanatory variables. Finally, the choice of optimal model in regard of some criterion (i.e. MSE or information criterion) identifies the optimal feature subset. This approach encounters a combinatorial complexity problem and various techniques of a heuristic search are used in practice. The most popular is stepwise regression. It can be applied in two directions, backward elimination or forward selection. In the first variant, the full set of features plays the role of a starting point. The variables which optimize the chosen criterion (in our work the Bayesian information criterion BIC) are removed in the following steps. In the second variant, an empty set of variables plays the role of the starting point, and the variable which optimizes the criterion is introduced into the model at every step.[2] Another popular technique is recursive feature elimination. This method discards variables corresponding to insignificant OLS estimators. The procedure is iterated until all the other coefficients are not significant. Another possibility is by removing only one variable at each step, the variable that minimizes the absolute value of the t statistic. This approach is not very time consuming in small data sets, and as a result can yield better results (see Table 2).

Finally, embedded methods perform simultaneously feature selection and the estimation of model parameters. Discarding some features is an integrated part of the learning algorithm. In the case of a linear regression model, this can be achieved by regularization.

3. Regularization in linear regression models

The core idea of regularization is the ability of controlling the complexity of the model. The objective is to get a state of compromise between the bias and the variance, and consequently to get the model with optimal generalization ability, which means the accuracy of the prediction for the new objects. This is achieved by imposing a penalty $P(\boldsymbol{b})$ for large absolute values of the parameters in the criterion used in the estimation:

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \left(\sum_{i=1}^{N} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} \right)^2 + \lambda \times P(\boldsymbol{b}) \right) \tag{5}$$

[2] There are several variants of stepwise regression. The review can be found in the work by Kubus (2014).

Regularization decreases the absolute values of the estimates, and some of them are equal to zero. This is equivalent to the feature selection and makes these methods attractive. There are several forms of penalty components proposed in the literature. Historically the first proposition was ridge regression (Hoerl, Kennard, 1970):

$$P(\boldsymbol{b}) = \sum_{j=1}^{p} b_j^2 \tag{6}$$

Then Tibshirani (1996) proposed LASSO:

$$P(\boldsymbol{b}) = \sum_{j=1}^{p} \left| b_j \right| \tag{7}$$

whereas Zou and Hastie (2005) proposed the penalty being their combination, so-called elastic net:

$$P_{\alpha}(\boldsymbol{b}) = \sum_{j=1}^{p} \left(\alpha b_j^2 + (1-\alpha) \left| b_j \right| \right) \tag{8}$$

Regularization parameter λ decides about the amount of the penalty, and as a result it controls the complexity of the model. Determining the appropriate value of lambda is a key task for the effective application of this method. In practice, a string of models corresponding to different values of λ is built, and then the optimal model is selected. As a model selection criterion one can use the prediction error estimated via cross-validation or information criteria. The empirical comparison of these criteria can be found in the work by (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, the LARS algorithm (Efron, Hastie, Johnstone, Tibshirani, 2004) is the most popular because of the low computational complexity. In the case of an 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 by being based on the current regression residuals, thus the previously unexplained variability of the response *Y*. In every step, updating the formula takes into account some predictors most correlated with the *Y*, while only one per iteration is introduced into the model.

4. Empirical example

The real estate market intended for single family housing, situated in two adjacent districts of the city of Krakow, will be the subject of analysis. The dataset consists of information from 23 transactions which were made within 29 months in the period 2005–2007 (Czaja, Ligas, 2010). Estate price (dependent variable) is expressed in (PLN/m^2). The first potential predictor considered in the analysis is the *time of transaction*. It is measured in months in the first transaction in the dataset. Other predictors are features, which determine the attractiveness of the real estate: *location, designt, communication, neighbourhood, management, shape of the plot*. These characteristics were formulated on the basis of legal documentation and site inspection. Experts assessed the real estate according to the above features, giving them points on a scale from 1 to 5. Five points being the highest grade.

The coefficients of a linear model were estimated using quadratic loss function with elastic net (8). The optimal value of the regularization parameter λ was chosen with the use of the bayesian information criterion BIC. Figure 1 depicts how the values of this criterion change with the number of variables introduced into the model. The minimum was reached for five variables and this model was adopted. Table 1 shows the estimated coefficients and the validity of the variables. The variables *design* and *shape of the plot* were eliminated. The most important in determining prices were: *time* and *location*.

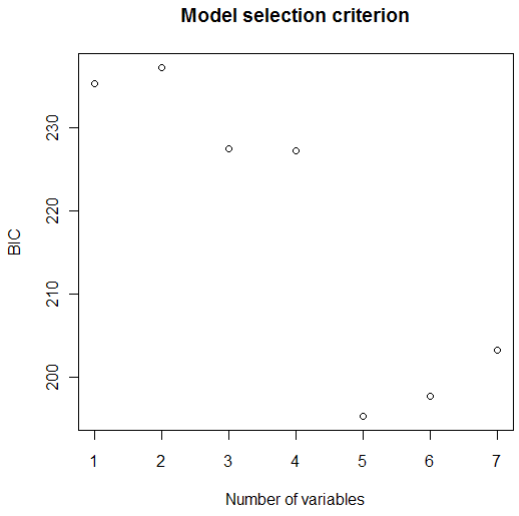

Figure 1. Bayesian information criterion for the models with various numbers of predictors

Source: own calculations.

Table 1. The coefficients of the linear model estimated using quadratic loss function with elastic net, and validity of predictors (expressed in %)

	Time	Location	Communication	Neighbourhood	Management
Coefficients	7.9242	49.5192	29.4834	22.3566	34.4029
Validity	39.4	22.0	14.7	8.1	15.7

Source: own calculations.

To assess the accuracy of the proposed model, the leave-one-out cross-validation was applied. The residual standard error was equal to 61.83 (PLN/m^2). Table 2 shows the comparison of errors obtained for the model with classical OLS estimators, and shrunken estimators (result of regularization). Some feature selection methods also were taken into consideration. Among them, we applied a filter which maximizes the correlation with the price and minimizes the correlations between predictors (MaxC-MinI) in the same time. Other applied feature selection methods were the stepwise regression and recursive feature elimination. The last one was investigated in two variants: removing subsets of irrelevant variables (RE) or removing only one variable in the iteration (RE-1). The second-accuracy was the MaxC-MinI filter with OLS estimation on the reduced set of predictors. It yielded an of error 63.19 (PLN/m^2) which was greater than the valuation proposed in this paper of 1.36 (PLN/m^2). Note that the greatest error 81.34 (PLN/m^2) was obtained applying recursive feature elimination, which seems to be quite a popular feature selection technique. It was almost 20 (PLN/m^2) greater than error of the proposed model.

Table 2. Residual standard errors estimated according to formula (4)

Method	Residual standard errors expressed in (PLN/m^2)
OLS	66.82
Elastic net	61.83
Ridge regression	69.10
LASSO	67.49
Stepwise regression	75.34
RE + OLS	81.34
RE-1 + OLS	75.34
MaxC-MinI + OLS	63.19

Source: own calculations.

Conclusions

One of the key issues in the problem of real estate valuation is the set of market features that characterize the objects. It plays a prominent role in the accuracy of the valuation. As mentioned before, the choice of the market features is largely determined by the real estate appraiser, who is guided by his/her experience. However, it is always a subjective choice. Formal statistical methods of feature selection can lead to more accurate valuation models; therefore it is worth using them. In this paper we proposed the application of regularized linear regression with an elastic net penalty (Zou, Hastie, 2005). For the considered data set of land real estate designated for single-family housing we obtained a model, which led to a more accurate valuation than some other popular linear models applied with or without a feature selection. It is worth noting that the applied method, in addition to the feature selection, gave estimates with smaller variances than OLS estimators. This property also influenced the resulting accuracy of the valuation.

References

Bitner, A. (2007). Konstrukcja modelu regresji wielorakiej przy wycenie nieruchomości. *Acta Scientiarum Polonorum, Administratio Locorum*, *6* (4), 59–66.

Czaja, J., Ligas, M. (2010). Zaawansowane metody analizy statystycznej rynku nieruchomości. *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, *18* (1), 7–19.

Doszyń, M. (2012). Ekonometryczna wycena nieruchomości. *Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego*, *26*, 41–52.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, *32* (2), 407–499.

Foryś, I. (2010). Wykorzystanie metod taksonomicznych do wyboru obiektów podobnych w procesie wyceny lokali mieszkalnych. *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, *18* (1), 95–105.

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (ed.) (2006). *Feature Extraction: Foundations and Applications.* New York: Springer.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inferance, and Prediction*. New York: Springer.

Hellwig, Z. (1969). Problem optymalnego wyboru predykant. *,Przegląd Statystyczny*, *3–4*.

Hoerl, A.E., Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Hozer, J. (ed.) (2008). *Wycena nieruchomości.* Szczecin: KEiS US, IADiPG w Szczecinie.

Hozer, J., Kokot, S., Kuźmiński, W. (2002). *Metody analizy statystycznej rynku w wycenie nieruchomości.* Warszawa: Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych.

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

Kubus, M. (2014). Discriminant stepwise procedure. *Acta Universitatis Lodziensis, Folia Oeconomica*, *3* (302), 151–159.

Lis, Ch. (2001). Sieci neuronowe a masowa wycena nieruchomości. *Zeszyty Naukowe Uniwersytetu Szczecińskiego*, *318*, *Prace Katedry Ekonometrii i Statystyki*.

Lis, Ch. (2005). Ekonometryczne modele cen transakcyjnych lokali mieszkalnych. *Zeszyty Naukowe Uniwersytetu Szczecińskiego*, *415*, *Prace Katedry Ekonometrii i Statystyki*, *16*.

Mach, Ł. (2012), Determinanty ekonomiczno-gospodarcze oraz ich wpływ na rozwój rynku nieruchomości mieszkaniowych. *Ekonometria*, *4* (38), 106–116.

Maddala, G.S. (2008). *Ekonometria*. Warszawa: Wydawnictwo Naukowe PWN.

Morajda, J. (2005). Wykorzystanie perceptronowych sieci neuronowych w zagadnieniu wyceny nieruchomości. *Zeszyty Naukowe Małopolskiej Wyższej Szkoły Ekonomicznej w Tarnowie, 7*, 101–108.

Prystupa, M. (2001). *Wycena nieruchomości przy zastosowaniu podejścia porównawczego.* Warszawa: Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych.

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

Zeliaś, A. (2006). Kilka uwag na temat doboru zmiennych występujących na rynku nieruchomości. *Zeszyty Naukowe Uniwersytetu Szczecińskiego*, *450*, *Prace Katedry Ekonometrii i Statystyki*, *17*, 685–696.

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