Multivariate Regression Modeling for Home Value Estimates with Evaluation using Maximum Information Coefficient

Gongzhu Hu, Jinping Wang, and Wenying Feng

Abstract Predictive modeling is a statistical data mining approach that builds a prediction function from the observed data. The function is then used to estimate a value of a dependent variable for new data. A commonly used predictive modeling method is regression that has been applied to a wide range of application domains. In this paper, we build multivariate regression models of home prices using a dataset composed of 81 homes. We then applied the maximum information coefficient (MIC) statistics to the observed home values (Y) and the predicted values (X) as an evaluation of the regression models. The results showed very high strength of the relationship between the two variables X and Y.

Key words: Predictive modeling, multivariate linear regression, hedonic price model, maximum information coefficient.

1 Introduction

Predictive modeling is a very commonly used method for estimating (or predicting) the outcome of an input data based on the knowledge obtained from a previous data set. It is to build a model (i.e. function f) from an observed data set $\mathbf{D}$ such that the model will predict the outcome of a new input $\mathbf{x}$ as $f(\mathbf{x})$ with the best

Gongzhu Hu
Department of Computer Science, Central Michigan University, Mt. Pleasant, MI 48859, USA.
e-mail: hu1g@cmich.edu

Jinping Wang
Graduate Biomedical Sciences, University of Alabama at Birmingham, Birmingham, Alabama.
e-mail: wangjp@uab.edu

Wenying Feng
Departments of Computing & Information Systems and Mathematics, Trent University, Peterborough, Ontario, Canada, K9J 7B8. e-mail: wfeng@trentu.ca

probability. The domain of $\mathbf{x}$ is a set of *predictors* or independent variables, while the outcome is a dependent variable. Various methods have been developed for predictive modeling, among them, multivariate linear regression is perhaps one of the most commonly used and relatively easy to build. The multivariate linear regression model is to express the dependent variable y as a linear function of p predictor variables x_i ($i = 1, \ldots, p$) and an error term ε:

$$y = c_0 + c_1 x_1 + \cdots + c_p x_p + \varepsilon$$

Note that if the the relationship between the dependent variable and the predictor variables is non-linear, we can create new variables for the non-linear terms and the regression model. For example, we can have

$$y = c_0 + c_1 x_1 + c_2 z_2 + c_3 z_3 + \varepsilon$$

where $z_2 = x_2^2$ and $z_3 = \ln x_3$. The linearity is actually between the dependent variable y and the coefficients c_i.

For a set of n data observations $\mathbf{x}$, the linear regression model can be expressed in matrix form:

$$\mathbf{y} = \mathbf{cX} + \mathbf{e}$$

The model is estimated by the least square measure that yields the coefficients $\mathbf{c}$ such that the predicted value

$$\hat{\mathbf{y}} = \mathbf{cX}$$

has the minimal sum of the squares of the errors $\mathbf{e} = |\mathbf{y} - \hat{\mathbf{y}}|$.

Predictive modeling has been widely used in many application areas, from business, economy, to social and natural sciences. In this paper, we apply multivariate linear regression to a specific economics application — estimating values of residential homes. This is not a new problem, neither is the regression method for solving the problem. The novel idea presented in this paper is to use the maximum information coefficient (MIC) [12], which is a new statistical measure published just a few months ago, to evaluate the regression models created. The MIC scores of the data set we used for the experiment showed that the regression models do have a very strong relationship with the observed home values. At the time of writing this paper, we are not aware of any published work using MIC as an evaluation measure for predictive models.

2 Estimate of Home Values

Home values are influenced by many factors. Basically, there are two major aspects:

- The environmental information, including location, local economy, school district, air quality, etc.

- The characteristics information of the property, such as lot size, house size and age, number of rooms, heating / AC systems, garage, and so on.

When people consider buying homes, usually the location has been constrained to a certain area such as not too far from the work place. With location factor pretty much fixed, the property characteristics information weights more in the home prices. There are many factors describing the condition of a house, and they do not weigh equally in determining the home value. In this paper, we present a modeling process for estimating home values using multivariate linear regression model based on the condition information of the dwellings in order to examine the key factors effecting their values. We also provide a general idea of figuring out if a transaction is a good deal based on the information provided.

Studies on home prices have been going on for many years using various models. The traditional and standard model is the *hedonic pricing model* that says the prices of goods are directly influenced by external or environmental factors in addition to the characteristics of the goods. For housing market analysis, the hedonic price model [9] infers that the price of dwellings is determined by the internal factors (characteristics of the property) as well as external attributes. The method used in this model is hedonic regression that considers various combinations of internal and external predictors [1, 4, 13]. The predictors may be first-order or higher order (such as $Area^2$) so that the hedonic regression may be a polynomial function of the predictors [2, 7].

The regression method used in our work is in fact a variation of hedonic regression, except that we did not consider external factors in our modeling (the data set does not include such information). We did, however, consider different combinations of first-order and second-order attributes in the regression model. The attributes are given in Table 1, where *Value* is the dependent variable to be predicted, and the other are predictors including 11 first-order and 4 second-order variables. The given data set contains 81 homes.

3 Building of Regression Model

3.1 Best Subsets Procedure

Since there are quite a few attributes of home condition, best subsets analysis [8] was first performed to select the best indicators to build the appropriate model. This procedure finds best models with 1, 2, 3, and up to all n variables based on the χ^2 statistics. The Minitab output of the analysis is shown in Table 2.

Based on the rule that second order indicators cannot exist without first order indicators, the impossible models were marked out in gray shade and would not be considered for the further analysis. Three potential cases were selected based on the highest R-sq(adj), lowest Mallows Cp, and smallest S. These three cases are colored in Table 2.

Table 1 Attributes of a House

Variable			Description
Dependent		*Value*	Assessed home value
Predictor	first-order	*Acreage*	Area of lot in acres
		Stories	Number of stories
		Area	Area in square footage
		Exterior	Exterior condition, 1 = good / excellent, 0 = average / below
		NatGes	1 = natural gas heating system, 0 = other heating system
		Rooms	Total number of rooms
		Bedrooms	Number of bedrooms
		FullBath	Number of full bathrooms
		HalfBath	Number of half bathrooms
		Fireplace	1 = with, 0 = without
		Garage	1 = with, 0 = without
	second-order	*Area**2*	House area squared
		*Acreage**2*	Lot size squared
		*Stories**2*	Number of stories squared
		*Rooms**2*	Number of rooms squared

Table 2 Best subsets analysis of the variables
(Shaded rows are impossible and ignored. Rows in color are selected for modeling)

Vars	R-Sq	R-Sq (adj)	Mallows Cp	S	Acreage	Stories	Area	Exterior	Netgas	Rooms	Bedrooms	Fullbath	Halfbath	Fireplace	Garage	Area**2	Acreage**2	Stories**2	Roos**2
1	58.8	58.3	162.9	54006			X												
1	58.5	58.0	164.7	54211												X			
2	74.5	73.9	73.4	42748	X		X												
2	71.8	71.1	88.9	44929	X											X			
3	80.2	79.5	42.2	37902	X	X	X												
3	79.9	79.1	44.3	38251	X		X											X	
4	82.5	81.6	30.6	35842	X	X	X					X							
4	82.2	81.3	32.4	36143	X		X					X						X	
5	84.3	83.2	22.5	34226	X	X	X	X				X							
5	84.2	83.1	23.1	34352	X	X	X					X	X						
6	85.3	84.1	18.4	33303	X	X	X	X				X	X						
6	85.2	84.0	19.1	36433	X	X		X				X	X			X			
7	86.2	84.8	15.6	32572	X	X	X					X	X		X		X		
7	86.1	84.8	16.0	32647	X	X	X	X				X	X		X				
8	87.1	85.7	11.9	31604	X	X		X			X	X	X			X			X
8	87.0	85.6	12.6	31763	X	X		X				X	X		X	X	X		
9	87.8	86.3	9.9	30964	X	X		X			X	X	X			X	X		X
9	87.8	86.3	9.9	30965	X	X		X			X	X	X		X	X			X
10	88.7	87.1	6.8	30050	X	X		X			X	X	X		X	X	X		X
10	88.6	87.0	7.2	30130	X			X			X	X	X		X	X	X	X	X
11	88.8	87.0	8.4	30172	X	X		X		X	X	X	X		X	X	X		X
11	88.7	86.9	8.7	30230	X	X		X			X	X	X		X	X	X	X	X
12	88.8	86.8	10.3	30361	X	X		X		X	X	X	X		X	X	X	X	X
12	88.8	86.8	10.3	30372	X	X		X	X	X	X	X	X		X	X	X		X
13	88.8	86.6	12.1	30550	X	X		X	X	X	X	X	X		X	X	X	X	X
13	88.8	86.6	12.2	30573	X	X		X		X	X	X	X	X	X	X	X	X	X
14	88.8	86.5	14.0	30767	X	X		X	X	X	X	X	X	X	X	X	X	X	X
14	88.8	86.5	14.0	30768	X	X	X	X	X	X	X	X	X	X		X	X	X	X
15	88.8	86.3	16.0	30992	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

3.2 Regression model

Based on the Best Subsets analysis results, three regression models were built for the three selected cases:

$$\begin{aligned}\hat{V}_1 = &-104582 + 45216\,Acreage + 36542\,Stories + 67.4\,Area + 12242\,FullBath \\ &+ 16428\,HalfBath + 30480\,Garage - 4397\,Acreage^2 \qquad (M\text{-}1)\end{aligned}$$

$$\begin{aligned}\hat{V}_2 = &-101097 + 21512\,Acreage + 38141\,Stories + 71.2\,Area + 18580\,Exterior \\ &+ 12218\,FullBath + 14569\,HalfBath + 23999\,Garage \qquad (M\text{-}2)\end{aligned}$$

$$\begin{aligned}\hat{V}_3 = &-111721 + 42939\,Acreage + 38965\,Stories + 72.3\,Area + 18901\,Exterior \\ &- 6781\,Rooms + 12139\,Bedrooms + 9721\,FullBath + 21047\,HalfBath \\ &+ 24095\,Garage - 3919\,Acreage^2 \qquad (M\text{-}3)\end{aligned}$$

Notice that the third model (M-3) has fewer variables than as indicated in the row `Vars=14` of Table 2. This is because several non-significant indicators were removed (in the order of first removing least significant and second-order indicators).

The residuals versus fits plots for the three models are shown in Fig. 1, Fig. 2 and Fig. 3, respectively.

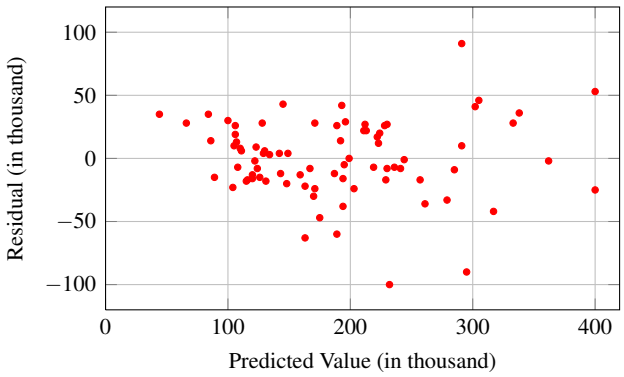

Fig. 1 Residuals versus fits plot of Model (M-1)

The figures show a fan-shaped pattern indicating that the diagnosis analysis revealed non-constant residual variances (residual error is not normally distributed), which is unacceptable. To alleviate the heterogeneity in the residual errors, a Box-Cox transform is applied to the dependent variable *Value*.

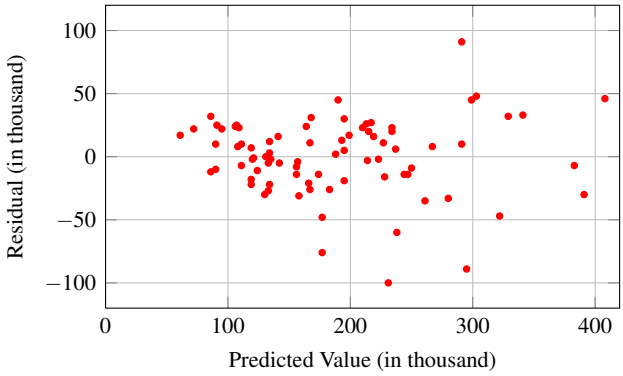

Fig. 2 Residuals versus fits plot of Model (M-2)

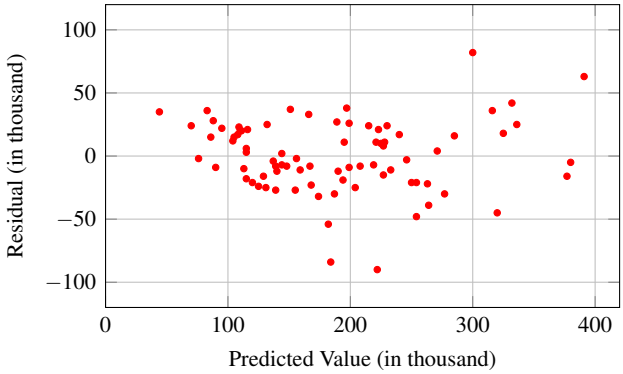

Fig. 3 Residuals versus fits plot of Model (M-3)

3.3 Box-Cox Transform

The Box-Cox procedure [3] provides a suggestion of the transformation on y:

$$y^{\lambda} = \begin{cases} (y^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log \lambda & \text{if } \lambda = 0 \end{cases}$$

After the transformation, the Box-Cox plot (λ versus standard deviation) is shown in Fig. 4.

From the plot, $\lambda = 0$, so the transformation on Y (*Value*) is

$$Value^{*} = \log(Value)$$

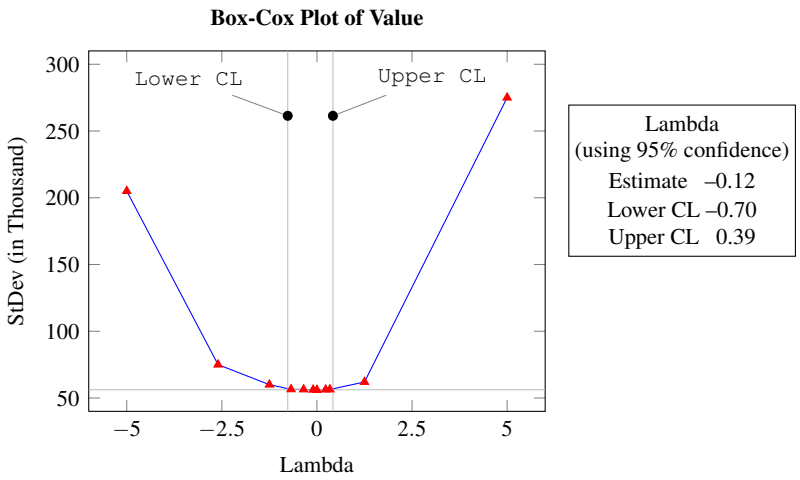

Fig. 4 The Box-cox analysis of Y (*Value*)

3.4 Redo Best Subsets Analysis

The Best Subsets analysis was restarted using the transformed Y value. This time, the best candidate model (with the highest R-sq(adj), lowest Mallows Cp and smallest S) is selected.

3.5 Rebuild the Linear Regression Model

Applying the regression procedure, we got the new model for $Value^*$ which is log(*Value*). Using $\hat{V}_4$ for the estimated *Value*, the model is expressed as

$$\begin{aligned}\log(\hat{V}_4) = {} & 9.99 + 0.311\,Acreage + 0.151\,Stories + 0.000305\,Area \\ & + 0.126\,Exterior + 0.115\,Rooms + 0.0556\,FullBath \\ & + 0.0816\,HalfBath + 0.163\,Garage - 0.0387\,Acreage^2 \\ & - 0.00548\,Rooms^2 \end{aligned} \quad (M\text{-}4)$$

3.6 Diagnostics of the New Model

From the result, we see that all the p-values are smaller than 0.05, which means all the predictors in the model are significant. So the next step would be the residual diagnostics. The residuals versus fits plot is shown as Fig. 5. And this time the plot is more close to random scatter plot.

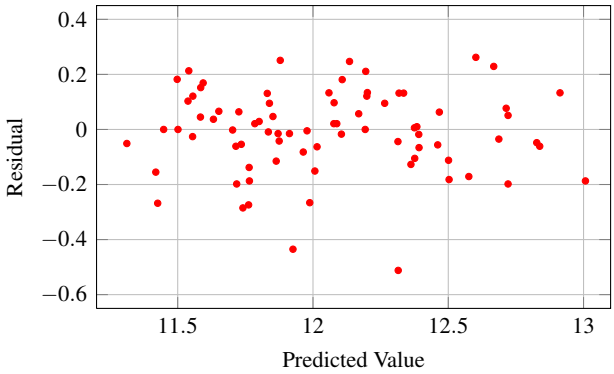

Fig. 5 Residuals versus fits plot of Model (M-4)

Finally, the Durbin-Watson test [5, 6] was used to test the autocorrelation between residuals. The Durbin-Watson value is 1.65277. Checking it against the critical values for $n = 81, k = 11$ from the Durbin-Watson table:

$$\alpha = 0.05, dL = 1.37434, dU = 1.92282$$

we see that $dL < 1.65277 < dU$, so the test is inconclusive meaning that there is no enough evidence to conclude whether there is a positive autocorrelation. However, the autocorrelation problem is not necessarily of a high concern because $1.65277 > 1$. Since $4 - 1.65277 = 2.34723 > dU$, there is no negative autocorrelation between residuals.

3.7 Discussion of the Results

From the above model diagnostics, we could conclude that all the predictors in the model are significant and there is no problem from the residual examinations. So it is a valid model. Since this is the one with the best results from the Best Subsets procedure, this model is confirmed as the final model of the linear regression analysis.

Except for the intercept predictor, there are ten predictors in the model. *NatGas*, *Fireplace* and *Bedrooms* are not in this model, which means that whether the heating system is using natural gas, whether there is a fireplace or not, and the number of bedrooms are not key factors influencing the home values. Furthermore, it appears (from this small data set), the predictor *Area* with a very small coefficient doesn't seem to have significant influence on home value. All the other predictors, especially *Acreage*, show a strong influence on the home values. So to conclude, the home values are closed related to lot size, number of stories, the exterior condition, number

of total rooms, number of full and half bathrooms, and with or without garages. Area also affects the home price but to a very limited degree.

However, we should also be aware that this is a simplified model from a small sample and we only considered the home condition information here. A more reliable analysis should include external and environmental factors such as geographic area, air quality, etc. as the hedonic price model suggested.

4 Model Evaluation using MIC

With these regression models established for predicting home values, it is often desirable to evaluate the goodness of the models. Of course, one way is to apply the models to additional data in similar home markets to see if the models generate satisfactory prediction accuracy. Since we do not have additional data for such goodness testing, we used a different approach for this purpose that relies on a statistic measure, called maximal information coefficient (MIC), to evaluate the relationship between the predicted values and the observed home values.

4.1 Maximum Information Coefficient

MIC was introduced very recently [12] as a new exploratory data analysis indicator that measures the strength of relationship between two variables. This measure is a statistic score between 0 and 1. It captures a wide range of relationships and is not limited to specific function types (such as linear as Pearson correlation does). A comprehensive companion article [11] provides detailed description of the theory and experimental results, as well as comparisons of MIC with other statistic measures. It shows that MIC gives similar scores to equally noisy relationships of different types. For example, the MIC scores for all the linear, cubic, exponential, sinusoidal, periodic/linear, parabolic relationships are 1, and for total random relationship is 0.

The basic idea of MIC is to compute mutual information on each cell of all grids on the x-y scatterplot up to the maximal grid resolution depending on the sample size. A characteristic matrix $M = (m_{x,y})$ is defined where each entry $m_{x,y}$ is the highest normalized mutual information of any x-by-y grid. MIC = $\max(m_{x,y})$ over (x,y) pairs such that $xy < B$, where $B = n^{\alpha}$ is the bound of grid size, n is the sample size (number of data points) and α is a parameter controlling the grid size. The MIC implementation described in [11] uses $\alpha = 0.6$ as the default value.

4.2 Model evaluation using MIC

We applied MIC to the linear regression models generated for our study on the home value data, along with the relationships between the home value and each of the 15 variables (11 first order and 4 second order variables). That is, we treated each of the models as a new "variable" for the purpose of calculating MIC scores. The results are given in Table 3.

Table 3 MIC results of the relationship between home value (Y) and the predictor variables (X)

Y	X	MIC (strength)	Linear regression (p)
Value	$\hat{V}_2$	0.9319	0.9315
	$\hat{V}_3$	0.9258	0.9399
	$\hat{V}_1$	0.8562	0.9306
	$\hat{V}_4$	0.7484	0.5963
log(Value)	$\log(\hat{V}_4)$	0.7484	0.7927
Value	Area	0.5939	0.7668
	Area**2	0.5939	0.7648
	Acreage	0.5597	0.6078
	Acreage**2	0.5597	0.5312
	FullBath	0.5178	0.6216
	Rooms	0.4852	0.6267
	Rooms**2	0.4852	0.5748
	Fireplace	0.4476	0.5497
	HalfBath	0.3730	0.4386
	Bedrooms	0.3333	0.5806
Value	Exterior	0.3002	0.1242
	Natgas	0.2560	0.1481
	Stories	0.2036	0.2536
	Stories**2	0.2036	0.2505
	Garage	0.1900	0.1962

In this table, the MIC scores were computed for pairs of (X,Y) variables, where X is a model or predictor and Y is home *Value*. We divide these variables into three sections as shown in Table 3. The first part are the variables $\hat{V}_i, i = 1,2,3,4$, representing the predicted values from the regression models. The middle part are ten variables with MIC score higher than the critical value (0.31677) at $p = 0.05$ for sample size $n = 80$. The table of critical values is given in [10]. The bottom section are those variables with MIC below the critical value. We also included the linear regression p-value showing the linearity of the X and Y variables (note: this is linear regression on (X,Y), not the linear regression we used to build the models).

From the results, we have the following observations.

(1) It was expected the regression models $\hat{V}_i$ would produce strong relationships with the home value, and this was confirmed with the MIC measures that are quite high (about 0.75 to 0.93).
(2) The high MIC scores along with the high linear regression p-values given in Table 3 imply that the models indeed are good estimates of the home values.
(3) The MIC scores of the models $\hat{V}_i$ are much higher than the scores of each individual variable. The lowest score (0.7484) for model $\hat{V}_4$ is 26% higher than the highest score (0.5939) for the individual variable (*Area*), indicating that the regression models for estimating home value are much reliable than individual variables, even though some of the individual variables do have strong influence on the home value on their own.
(4) For a pair of first-order and second-order variables (such as *Acreage* and *Acreage**2*), the linear regression p-value of (*Value*, *Acreage*) is 0.6078 while the p-value of (*Value*, *Acreage**2*) is 0.5312, indicating that linear regression approach does not fully capture the functional relationship between a first-order variable X and its second-order version X^2. In contrast, MIC yields identical score because one variable is simply a square function of the other so that the strength of the relationship between home value and X is the same as that between home value and X^2.
(5) The MIC scores for the two versions of our final model $\log(\hat{V}_4)$ and $\hat{V}_4$ are the same, while the linear regression p-value of (home *Value*, $\log(\hat{V}_4)$) is much higher than the p-value of (home *Value*, $\hat{V}_4$). This shows that the Box-Cox transform was helpful for removing the non-Gaussian problem in the residual errors, and that MIC is a better indicator of non-linear relationships between (*Value*, Y).
(6) The final model $\log(\hat{V}_4)$ (or $\hat{V}_4$) we obtained from the analysis (went through diagnosis and transformation) has noticeably weaker relationship with home value than the initial three models $\hat{V}_1, \hat{V}_2, \hat{V}_3$. It is unclear at this time what is the cause of such phenomenon. It may well be due to the small sample size or some other factors that we will explore.

4.3 MIC as a Variable Selection Tool

It light of the power of MIC for detecting the relationship between two variables, it is possible to use MIC as a measure for selecting the "best subset" of predictors for building the regression model. The idea is to use the m predictors X_i in the middle tier of Table 3, for which the MIC scores are higher than the critical value. Select incrementally the set of variables $\{X_1\}, \{X_1, X_2\}, \ldots, \{X_1, \cdots, X_m\}$ where X_1 has the highest MIC score. A regression model $\hat{V}_j$ is built using each set of the variables, and compare the MIC scores of (*Value*, $\hat{V}_j$). At the time of writing this paper, we have applied the idea to the 7 first-order variables to build a regression model $\hat{V}_5$ and all 10 first-order and second-order variables for $\hat{V}_6$. The result is given in Table 4.

Table 4 MIC result of home value (Y) and $\hat{V}_5, \hat{V}_6$ (X)

Y	X	MIC (strength)	Linear regression (p)
Value	$\hat{V}_5$	0.8296	0.9133
Value	$\hat{V}_6$	0.7861	0.8776

It shows that the regression models $\hat{V}_5$ and $\hat{V}_6$ built with the predictors selected based on MIC are similar to the models using the original Best Subset procedure.

5 Conclusion

Multivariate regression has been widely used for years applying to almost all the areas in our lives. In this paper, we presented a process of building a multivariate regression model for a simplified problem of estimating housing prices. This process involves five steps: (a) apply the Best Subsets procedure to select the variables; (b) build linear regression models from the selected variables; (c) conduct diagnostics to find if the residual errors are normally distributed; (d) apply Box-Cox transformation to fix the non-Gaussian residual problem found in the diagnosis; and (e) restart the analysis to build the final model. This is a typical process of building regression models that may apply to many applications where predictive modeling is the goal.

For model evaluation, rather than testing the model against new data sets with known target values to find the experimental accuracy, we used the newly introduced statistic measure, maximum information coefficient, to find the strength of the relationship between the model and the target value. The test we conducted, although for a small data sample, revealed that the MIC measure may be a viable metric for evaluation of the models built. MIC may also be used for variable selection. We are currently exploring ways of using MIC as a tool for model building.

References

1. Anselin, L., Lozano-Gracia, N.: Errors in variables and spatial effects in hedonic house price models of ambient air quality. Tech. Rep. Working Paper 2007–1, University of Illinois, Urbana-Champaign (2007)
2. Bitter, C., Mulligan, G.F., Dall'erba, S.: Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method. Journal of Geographical Systems (1), 7–27 (2007)
3. Box, G.E.P., Cox, D.: An analysis of transformations. Journal of the Royal Statistical Society **26**(2), 211–252 (1964)
4. Campbell, J., Stefano, G., Pathak., P.: Forced sales and house prices. National Bureau of Economic Research Working Paper (2009)

5. Durbin, J., Watson, G.: Testing for serial correlation in least squares regression, I. Biometrika **37**, 409–428 (1950)
6. Durbin, J., Watson, G.: Testing for serial correlation in least squares regression, II. Biometrika **38**, 159–179 (1951)
7. Fik, T.J., Ling, D.C., Mulligan, G.F.: Modeling spatial variation in housing prices: A variable interaction approach. Real Estate Economics **31**(4), 423–464 (2003)
8. Furnival, G.M., Wilson, Jr., R.W.: Regression by leaps and bounds. Technometrics **16**(4), 499–511 (1974)
9. Lentz, G.H., Wang, K.: Residential appraisal and the lending process: A survey of issues. Journal of Real Estate Research **15**(1-2), 11–40 (1998)
10. MINE: Maximal Information Nonparametric Exploration: P-Value Tables. www.exploredata.net/Downloads/P-Value-Tables
11. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Supporting online material for detecting novel associations in large data sets. www.sciencemag.org/cgi/content/full/334/6062/1518/DC1
12. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. Science (6062), 1518–1524 (2011)
13. Sheppard, S.: Hedonic analysis of housing markets. Handbook of Regional and Urban Economics pp. 1595–1635 (1999)