Advanced Engineering Days

aed.mersin.edu.tr

Representing ordinal predictors in real estate valuation with multiple regression

Ümit Işıkdağ[*1]

[1]*Mimar Sinan Fine Arts University, Department of Informatics, Türkiye, umit.isikdag@msgsu.edu.tr*

Cite this study: Isikdag, U. (2022). Representing ordinal predictors in real estate valuation. 3rd Advanced Engineering Days, 1-4

Keywords
Real Estate
Valuation
Regression
Ordinal
Predictor

Abstract

There are several different valuation methods used in real estate valuation. Statistical methods such as multiple regression and machine learning techniques such as ANN and Ensemble Methods are among the key ones implemented in recent studies. In all these methods the types of predictor variables for estimation of the price/value range from nominal, ordinal to continuous. Although the representation of nominal and continuous variables are agreed in prediction models, the representation of ordinal variables can change in every model. This paper provides a comparison of the different representations of the same ordinal variable in models. The findings indicate that representation of ordinal variables as continuous can be beneficial when the ordinal variable has many levels.

Introduction

Real estate valuation has been a very active topic of research since many years. Real estate valuation has been done using several methods depending on the purpose of the valuation study. The valuation can be defined as the determination of amount for which the property will transact on a particular date [1]. There are several concepts related with the value, these are price, worth and the value. As explained by [1] price is the observable exchange price in the open market, the value is an estimation of the price that would be achieved if were the property to be sold in the market; and worth is a specific individual's perception of the capital sum that he/she would be prepared to pay for the stream of benefits that she/he expects to be produced by the property. Value models concentrate on predicting the price of the property, the value is considered as the final result. The purpose of the valuation studies is generating reports for sale purposes, accounting purposes in companies, for calculation of loans at banks, for finding a minimum price for auctions, for insurance purposes, for taxation purposes and for compulsory purchase purposes [1]. There are several different valuation methods used in real estate valuation. [2] summarizes these methods into 2 groups as traditional methods and advanced methods. The traditional methods include comparable method, investment/income method, profit method, development residual method, contractor's method/cost method, multiple regression method, stepwise regression method. The advanced methods include, artificial neural networks (ANNs), hedonic pricing method, spatial analysis methods, fuzzy logic and time series methods such as autoregressive integrated moving average (ARIMA). As explained in [3] the real estate valuation efforts are usually focused on two aspects, first one is known as individual valuation and is concentrated on determining the price of a specific (or focused) real estate. The second approach is known as mass valuation, where the value of a group of real estates are predicted. The comparable (sales comparison), contractor's method/cost method, income method, profit methods are suitable for individual valuation. However mass appraisal benefits more from statistical and machine learning approaches including regression-based methods, artificial neural networks, spatial analysis methods. Examples of the use of statistical and machine learning methods from the recent literature include the use of who implemented multiple regression analysis, hedonic models, ANNs, Ensemble Methods, Quantile and Semi-Log Regression [4-7]. Furthermore [7] presents a comprehensive overview of real estate appraisal methods in the era of big data.

Material and Method

As indicated by the literature most modern methods of real estate appraisal involves dealing with statistical and mostly with machine learning techniques. In these techniques the estimation of real estate prices is done following the "regression" family of the estimation and machine learning methods, as the price is a variable having a continuous scale. There are many factors that has an impact on the price including the size, the location, attributes of the unit (if the real estate is a property) such as number of rooms (living rooms, bathrooms …), existence of different rooms and facilities inside the apartment, scenery, the level of the apartment, the orientation, condition, distance to transport, nearby facilities, schools, markets, health facilities, socio-economic status of the neighborhood and many more. Although the dependent variable, price, has a continuous scale, the independent variables (factors/predictors) that are used to estimate the dependent variable ranges from dichotomous, nominal to ordinal. In statistical and machine learning models, it is agreed that dichotomous variables can be used without any modification, but nominal variables having levels of 2+ need to be converted into (n-1) dummy variables, where n denotes to number of categories in the nominal variable. This process is called the dummy encoding of the nominal variable. In fact, in the case of ordinal variables (variables representing ordered values) there is no such agreement on how to represent them in statistical and machine learning models. One view advocates that these variables "are not continuous in nature" and should be represented, as if they were nominal, and similar to nominal variables, and dummy encoding should be applied to the ordinal variables as well and generated dummy variables should be used as independent variables (IVs) in the model in lieu of the ordinal variable. In fact, once this dummification operation is done, the ordered nature of the data is lost completely. A solution to tackle with that through using dummy variables has been proposed in [8] and this encoding technique is known as staircase encoding. The final view on the use of ordinal variables is treating them as continuous variables in the models. In this study in order to test all these approaches we have collected data on real estate prices in a neighborhood of Istanbul from a well-known real estate web site. The data was collected manually by a human operator. The raw data contained 253 rows, and included 3 fields, 'price', 'sqm' and 'bl'. The price variable represents the price in terms of Turkish Lira, the sqm represents the gross floor area of the apartment, and 'bl' represents the number of bedrooms + living rooms of the apartment. The scale of price and sqm are continuous, while the scale of 'bl' can be considered both as nominal and ordinal, the variable has 14 categories/levels. In this research, we have ordered 'bl' in ascending order by taking b as prior and l as secondary important factor. As a result, we generated a new ordinal variable 'bl_ordinal', and we also kept the raw data in a variable named 'bl_nominal'. On the other hand, we have implemented a natural log transformation to the price variable and generated a new variable named 'lnPrice'. The natural log transformation helped to remove the excess kurtosis and positive skewness of data and ensured the normality of the variable. Later we have implemented both dummy and staircase encoding to the ordinal variable through the program code developed during the study. The staircase and dummy variables generated as a result of this process were 'sc1-sc13' and 'd1-d13' (26 variables in total). The code used in the generation of these variables are provided in [9]. Once the data processing is complete, we have conducted 3 multiple regression analyses. The models proposed and tested is provided below:

$$lnPrice = \beta_0 + \beta_1 sqm + \beta_2 d1 + \ldots\ldots\ldots\ldots + \beta_{14} d13$$
$$lnPrice = \beta_0 + \beta_1 sqm + \beta_2 sc1 + \ldots\ldots\ldots\ldots + \beta_{14} sc13$$
$$lnPrice = \beta_0 + \beta_1 sqm + \beta_2 bl\_ordinal$$

Results

The results of the multiple regression analysis are provided below. In Model 1(Fig.1), the coefficients of all dummy variables (d1-d13) with exception of 'd8' are statistically insignificant, while the 'β_0' and the coefficient of 'sqm' is statistically significant. On the other hand, in Model 2 (Fig.1), the coefficients of all staircase encoded variables (sc1-13) are statistically insignificant, while the 'β_0' and the coefficient of 'sqm' is statistically significant. In Model 3 (Fig.2), coefficients of all variables were statistically significant. The F value in regression is the result of a test where the null hypothesis is that all of the regression coefficients are equal to zero. The F-test of overall significance shows whether the linear regression model provides a better fit to the data than a model that contains no independent or no significant independent variables. In all 3 models, the null hypotheses of F tests are rejected, showing that all 3 models have a predictive capability.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.84178	0.161728	85.58676	0.0000
D1	0.280101	0.177205	1.580658	0.1153
D2	0.396599	0.410958	0.965060	0.3355
D3	0.174794	0.166891	1.047355	0.2960
D4	0.239222	0.333915	0.716416	0.4744
D5	0.262840	0.179025	1.468172	0.1434
D6	0.490448	0.258933	1.894114	0.0594
D7	0.257455	0.269074	0.956817	0.3396
D8	0.483212	0.204093	2.367600	0.0187
D9	0.356382	0.266651	1.336509	0.1827
D10	0.402968	0.271059	1.486642	0.1384
D11	0.328184	0.293921	1.116570	0.2653
D12	0.057534	0.379933	0.151432	0.8798
D13	0.849007	0.533621	1.591029	0.1129
SQM	0.004273	0.000921	4.641457	0.0000

R-squared	0.439528	Mean dependent var	14.69719
Adjusted R-squared	0.406420	S.D. dependent var	0.490995
S.E. of regression	0.378283	Akaike info criterion	0.951329
Sum squared resid	33.91418	Schwarz criterion	1.161414
Log likelihood	-104.8674	Hannan-Quinn criter.	1.035863
F-statistic	13.27556	Durbin-Watson stat	2.149335
Prob(F-statistic)	0.000000		

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.84178	0.161728	85.58676	0.0000
SC1	0.280101	0.177205	1.580658	0.1153
SC2	0.116498	0.390114	0.298625	0.7655
SC3	-0.221805	0.380975	-0.582202	0.5610
SC4	0.064427	0.283586	0.227189	0.8205
SC5	0.023618	0.274019	0.086192	0.9314
SC6	0.227608	0.193131	1.178519	0.2398
SC7	-0.232993	0.250719	-0.929301	0.3537
SC8	0.225757	0.172819	1.306314	0.1927
SC9	-0.126830	0.157424	-0.805654	0.4213
SC10	0.046586	0.230916	0.201746	0.8403
SC11	-0.074784	0.256647	-0.291389	0.7710
SC12	-0.270650	0.309841	-0.873514	0.3833
SC13	0.791473	0.479088	1.652041	0.0999
SQM	0.004273	0.000921	4.641457	0.0000

R-squared	0.439528	Mean dependent var	14.69719
Adjusted R-squared	0.406420	S.D. dependent var	0.490995
S.E. of regression	0.378283	Akaike info criterion	0.951329
Sum squared resid	33.91418	Schwarz criterion	1.161414
Log likelihood	-104.8674	Hannan-Quinn criter.	1.035863
F-statistic	13.27556	Durbin-Watson stat	2.149335
Prob(F-statistic)	0.000000		

Figure 1. Multiple Regression Analysis Results Models [1,2]

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.95263	0.061490	226.9091	0.0000
BL_ORDINAL	0.037358	0.017333	2.155327	0.0321
SQM	0.003820	0.000733	5.207539	0.0000

R-squared	0.412958	Mean dependent var	14.69719
Adjusted R-squared	0.408243	S.D. dependent var	0.490995
S.E. of regression	0.377701	Akaike info criterion	0.902408
Sum squared resid	35.52194	Schwarz criterion	0.944425
Log likelihood	-110.7034	Hannan-Quinn criter.	0.919315
F-statistic	87.58017	Durbin-Watson stat	2.064853
Prob(F-statistic)	0.000000		

Figure 2. Multiple Regression Analysis Results Models [3]

Discussion and Conclusion

The results of the assumptions testing of the Model 3 were as follows. Normality of residuals, Jarque-Bera:3.02 ($p>0.05$) (Ho: Normality), Breusch-Godfrey Serial Correlation LM Test: Obs*R-squared :1.25($p>0.05$) (Ho: No serial correlation), White Test: Obs*R-squared :7.17($p>0.05$) (Ho: Homoskedasticity), VIF:3.45 (<5 for both independent variables showing no signs of multicollinearity). In the Model 3, where the ordinal variable was treated as a continuous variable, all coefficients of the model were found significant, in addition all assumptions of the multiple linear regression were met. The results show that treating an ordinal variable as a continuous one provides a statistically significant model, where all assumptions of multiple linear regression are met. On the other hand, models where the ordinal variable is represented with dummy encoding and staircase encoding provided insignificant coefficients for all dummy /staircase variables with an exception for one dummy variable. This insignificance might be result of imbalance between the levels of the ordinal variable (thus in the number of dummy variables generated for each level). In addition, as the variable in focus had 14 levels (13 dummy/staircase variables), some of the dummy/staircase variables might have a very minor contribution in explaining the variance in the dependent variable, especially in the case of linear regression. This in turn might have resulted in the insignificance of the coefficients of dummy/staircase variables. The results have shown that an 'ordinal variable with many levels' can be interpreted/included as a continuous variable in Multiple Regression Analysis based Real Estate valuation. The case provided is a single example only, and more research and testing are required to reach a more comprehensive and general conclusion on the subject.

References

1. French, N. (2004). The valuation of specialised property: A review of valuation methods. Journal of Property Investment & Finance.
2. Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. Journal of Property Investment & Finance.

3. Sisman, S & Aydinoglu, A. C. (2022) Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis, Land Use Policy,119,106167
4. Benjamin, J., Guttery, R., & Sirmans, C. F. (2004). Mass appraisal: An introduction to multiple regression analysis for real estate valuation. Journal of Real Estate Practice and Education, 7(1), 65-77.
5. Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772-1778.
6. Torres-Pruñonosa, J., García-Estévez, P., & Prado-Román, C. (2021). Artificial Neural Network, Quantile and Semi-Log Regression Modelling of Mass Appraisal in Housing. Mathematics, 9(7), 783.
7. Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., & Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. Land, 11(3), 334.
8. Walter, S. D., Feinstein, A. R., & Wells, C. K. (1987). Coding ordinal independent variables in multiple regression analyses. American Journal of Epidemiology, 125(2), 319-323.
9. https://github.com/uisikdag/ordinal_pred