Finding Accurate Price Generating Model for Real Estate Market

Alimzhan Gumran, Anel Salmenova, Akezhan Shakenov
Piotr Sebastian Skrzypacz
Nazarbayev University

July 25, 2024

Abstract

This study analyses the Real Estate Sales dataset and evaluates multiple regression models based on the given data, choosing the most accurate models. We identified the best predictor variables, using the least amount of predictors and returning a more accurate prediction of sales prices. In our work, using the criteria given in Chapter 9.3 [4], we tested all possible sets of predictor variables and found the best sets of predictors. The best subset models were found to use 4 and 5 predictors, and the 4-predictor model performed better on the validation set. Our model predicts sales price based on finished square feet, year built, presence of air conditioning, and style. We also compared our findings to the results of other researchers.

Keywords: Regression Analysis, Linear Models, Real Estate Sales, Model Comparison, Prediction Accuracy

1 Introduction

The primary goal of this report is to identify the most accurate regression models for predicting real estate sales prices. According to our textbook, the motivation behind the problem of determining the best models lies in efficient tax assessment. However, an in-depth analysis of house price prediction techniques is also beneficial for research and society since precise real estate prices enable better information in the real estate business, enhancing housing policies and house evaluation. Hence, the primary motivation for choosing the topic is creating better policies and offers for the market (both for buyers and sellers) and investigating and understanding the science behind the societal matter for the researchers [3, 7]. That is why we wanted to deepen into the selected problem by developing the best subset model and assessing our model's ability to predict, enhancing our knowledge of how to use the skills we learn on real-life issues. The study utilizes the book's Real Estate Sales data set and is built on the methodology presented in Case Study 9.31 [4].

While doing a literature review, we revealed that both traditional and advanced regression models have been employed for similar prediction tasks. However, there remains a need for a detailed comparison of these methodologies to understand their relative strengths and weaknesses. Nevertheless, this report presents the development of a linear model, assesses its performance, and compares it with the best regression model from Case Study 9.31 [4]. Further research needs to be done to understand more detailed benefits and risks between other models. The models' performance is compared to the best regression model from Case Study 9.31 in "Applied Linear Statistical Models" by Kutner, Nachtsheim, Neter, and Li.

2 Research Framework

2.1 Problem Overview

The project aims to forecast home prices using a dataset of residential sales from one city in the midwest in 2002. The dataset contains a range of characteristics, including the year of construction, lot size, style, completed square footage, number of bedrooms, air conditioning, and proximity to a road. Using these data variables, we needed to find the most effective way to use our criteria for all possible sets of 256 models and find the most accurate of them.

2.2 Methodology

We developed the most suitable subset model for sales price prediction using a randomly chosen subset of 300 observations. The remaining 222 observations became the validation set. As a part of data preparation, we removed those columns in the provided data in Appendix C. 7 [4] that were not stated in the problem and constructed matrices for our predictor and response variables using R programming language. Then, we wrote the code that uses our criteria from section 3.2 for all possible regression models. The most effective predictors' indexes were kept and then analyzed using the other 222 variables. In evaluation, we used statistical methods like Residual Sum of Squares, R-squared, adjusted R-squared, AIC (Akaike Information Criterion), Schwarz Bayesian Criterion, and Prediction Sum of Squares to determine which subset of predictors is the best and reach a balance between model complexity and prediction accuracy.

2.3 Modeling

A linear model is developed to predict real estate sales prices. We run the same algorithm for every model for each number of predictors 1 to 8. The algorithm includes multiple incorporated "while" loops we run using R programming language. In the outer loop, we fix the property by index and remove its corresponding column from the X matrix. In the inner loop, the fixed property is removed, and then we remove the other properties by turn so that the model includes removing the outer loop index and each of the indexes in the inner loop.

The depth of the loops depends on the number of predictors to be removed. The coefficient b for each model is calculated using matrix methods according to the formula given below:

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

3 Case Study

3.1 Data Description

The Real Estate Sales data set comprises various features related to property characteristics and sales prices. A random sample of 300 observations is selected for model building and evaluation. The data covers sales during the year 2002 in a certain city in the Midwest. The initial data in Appendix C.7 [4] included 11 predictor variables, but according to the problem statement, we needed only 8 of them. Therefore, we removed columns that were not required. The data that we used in the code was presented in matrix form. Below are the properties and their corresponding predictor variables

- X_{i1} - finished square feet
- X_{i2} - number of bedrooms
- X_{i3} - presence of air conditioning
- X_{i4} - presence of pool
- X_{i5} - year built
- X_{i6} - style
- X_{i7} - lot size
- X_{i8} - adjacency to a highway

3.2 Explanation of criteria

A comprehensive analysis of regression models with varying numbers of variables is carried out using our accompanying R script. Several criteria, including SSE, R^2, R_p^2, and others, were used to determine which model is the best. Below is the justification for each criterion used.

3.2.1 R_p^2 and SSE_p criteria

In R_p^2 criterion, the better models have higher value because it represents the proportion of how much variation is lost due to using a particular set of predictor variables and its associated SSE. The less SSE is, the more accurate the model, and the higher the R_p^2.

The coefficient of determination R^2 for the pth model is defined as:

$$R_p^2 = 1 - \frac{\text{SSE}_p}{\text{SSTO}} \tag{1}$$

3.2.2 $R^2_{a,p}$ criterion

In the R^2_p criterion, since it does not decrease as the number of parameters increases, the $R^2_{a,p}$ criterion is used. It can decrease when new predictor is added when increase in R^2_p is so marginal that it is proportionately less than loss of extra degree of freedom, thus it becomes evident that the added predictor is odd.

The adjusted R^2 for the pth model is defined as:

$$R^2_{a,p} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}_p}{\text{SSTO}} = 1 - \frac{\text{MSE}_p}{\frac{\text{SSTO}}{n-1}} \tag{2}$$

3.2.3 C_p criterion

This criterion concerns the total difference between the fitted value and genuine mean for each level i. According to our textbook, if the set of predictor variables is correctly chosen, then C_p given by formula 3 is the estimator of the difference. And if C_p is unbiased, then its expected value is roughly equal to the line $C_p = p$. The sets of predictors that produce biased C_p fall higher than $C_p = p$ line.

Mallows' C_p criterion is defined as:

$$C_p = \frac{\text{SSE}_p}{\text{MSE}(X_1, \ldots, X_{p-1})} - (n - 2p) \tag{3}$$

3.2.4 AIC_p and SBC_p criteria

These criteria are used to identify when the model has excessive amounts of predictor variables. According to their formula, as we can notice, the fewer the values, the better the model is because these criteria are increased by $2p$ and $[ln(n)]p$. For more extensive set of observations SBC_p is more strict criteria.

The Akaike Information Criterion (AIC) for the pth model is defined as:

$$\text{AIC}_p = n \ln \text{SSE}_p - n \ln n + 2p \tag{4}$$

The Schwarz Bayesian Criterion (SBC) for the pth model is defined as:

$$\text{SBC}_p = n \ln \text{SSE}_p - n \ln n + [\ln n]p \tag{5}$$

3.2.5 $PRESS_p$ criterion

This criterion is concerned with the squared sum of differences between observed Y_i and $\hat{Y}_{i(i)}$. The $\hat{Y}_{i(i)}$ term means the Y_i term that is returned by regression function constructed without ith value. The closer each $\hat{Y}_{i(i)}$ to Y_i, the more accurate the model and the less the given difference. In other words, low $PRESS_p$ are returned by better models.

The Prediction Sum of Squares ($PRESS$) for the pth model is defined as:

$$\text{PRESS}_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)} \right)^2 \tag{6}$$

The results are summarized in Table 1.

3.3 Best Subsets Models

By following the best subsets algorithm, we have obtained the following table, which demonstrates the best values of each metric at every possible number of predictors $p-1$ (from 1 to 8).

p	R^2	R_a^2	C_p	AIC_p	SBC_p	$PRESS_p$
2	0.666	0.665	133.902	6769.370	6773.074	1.900e+12
3	0.695	0.693	98.940	6744.531	6751.938	1.753e+12
4	0.724	0.721	63.155	6716.361	6727.472	1.602e+12
5	0.729	0.725	58.233	6712.490	6727.306	1.577e+12
6	0.730	0.726	58.828	6713.284	6731.803	1.573e+12
7	0.731	0.725	60.315	6714.843	6737.066	1.584e+12
8	0.731	0.725	61.843	6716.436	6742.362	1.599e+12
9	0.731	0.724	63.835	6718.429	6748.060	1.621e+12

Table 1: Model Selection Criteria for Different Numbers of Predictors

3.4 Regression Model Validation

The 4-predictor model was favored by the $PRESS_p$ and R_a^2 metrics:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \beta_5 X_{i5} + \beta_6 X_{i6} \tag{7}$$

The C_p, AIC_p and SBC_p metrics favored the 5-predictor model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_8 X_{i8} \tag{8}$$

To choose our final model, we tested both models on the validation set, and their performance was as follows:

p	R^2	R_a^2	C_p	AIC_p	SBC_p	$PRESS_p$
6	0.720	0.712	22.925	4987.023	5007.439	1.287e+12
5	0.720	0.713	21.377	4985.454	5002.467	1.285e+12

Table 2: Performance on the Validation Set

The table favors the 4-predictor model (8). Thus, our final model relies on style, finished square feet, the presence of air conditioning, and the year built to predict a house's price.

4 Comparison with existing models

It is known that a lot of research has been done regarding a given topic. For instance, the paper written by Ozgur *et al.* (2016) [5] focused on predicting the prices of sold houses in North West Indiana in 2014 using multiple linear regression models too. Like many similar papers, it is looking for the best fitting linear regression model to identify how some factors (such as square footage, number of bathrooms, presence of a finished basement, type of home, etc.) may influence price generation. It also examines various predictor combinations to get the most relevant price prediction model. The study used almost the same statistical measures as we did for model performance evaluation. The application of their presented model lies in facilitating the price setting for homeowners, allowing them to understand the fairness and competitiveness of their price and providing accurate evaluations for real estate agents, quickening their job. As we can see, rationales like policy formulation, consumer and seller guidance, and business decision-making were also considered in our paper and others, too [6, 2, 1]. Despite the similarity in many aspects, there are still differences in the results. For example, if the results of our model rely more on style, finished square feet, presence of air conditioning, and year built to predict the price of a house, the results of Ozgur *et al.* (2016) model mostly depend on Homeowner Association (HOA) fee. In general, our literature review suggests that the concepts, key aims, statistical measures, and predictors were more or less the same since supply and demand for housing are an economically universal principle, and the goal of getting the maximum for a lower price by achieving a balance among dealers is a common goal for all market groups.

5 Conclusions

The study demonstrates that our modeling method can be used to determine accurate sets of predictors for Real Estate Market prices. Using the criteria in chapter 9.3 of our textbook proved usefulness for predictors evaluation according to our validation data. For the given dataset our results show that there are two models that are the most predict sales prices the most accurately. First uses finished square feet, presence of air conditioning, year built and style as its predictors. Second uses the same predictors and lot size. The literature review assisted us in understanding the current situation of real estate market pricing, gaining foundational knowledge of the topic, and focusing on research questions. In further comparison with our linear modeling, we found out that the examination of the chosen subject would follow similar primary goals, statistical measures, predictor variables, and ideas because such a general topic observes common goals and seeks profitable equilibrium for all markets over the globe. However, some papers, as tested on various data sets, may differ in their model reliance due to the priorities connected to geographical, cultural, and socioeconomic matters.

References

[1] Ningyan Chen. House price prediction model of zhaoqing city based on correlation analysis and multiple linear regression analysis. *Wireless Communications and Mobile Computing*, 2022(1):9590704, 2022.

[2] Faqiang Cui. Quantitative study on factors affecting the price of residential real estate multiple linear regression model. In *Journal of Physics: Conference Series*, volume 1629, page 012071. IOP Publishing, 2020.

[3] Margot Geerts, Jochen De Weerdt, et al. A survey of methods and input data types for house price prediction. *ISPRS International Journal of Geo-Information*, 12(5):200, 2023.

[4] Michael H Kutner, Christopher J Nachtsheim, John Neter, and William Li. *Applied linear statistical models*. McGraw-hill, 2005.

[5] Ceyhun Ozgur, Zachariah Hughes, Grace Rogers, and Sufia Parveen. Multiple linear regression applications in real estate pricing. *International Journal of Mathematics and Statistics Invention (IJMSI)*, 4(8), 2016.

[6] P Priya, G Mahalakshmi, K Manojkumar, and V Kumararaja. Property price prediction and possibility prediction in real estate using linear regression. *Adv Appl Math Sci*, 21(7):3805–3819, 2022.

[7] Jincheng Zhou, Tao Hai, Ezinne C Maxwell-Chigozie, Afolake Adedayo, Ying Chen, Celestine Iwendi, and Zakaria Boulouard. Effective house price prediction using machine learning. In *International Conference on Advances in Communication Technology and Computer Engineering*, pages 425–436. Springer, 2023.