Multi-linear Regression Model for Analyzing House Pricing in the USA

Mujing Chen, Jingjing Huang, Xiyi Gu and KevinZhang
STA302H1FLEC0101
Dr. Antonio Herrera Martin
June 17, 2024

Introduction:

During the pandemic, the global economic situation has been like a roller coaster, influencing real estate marketing significantly. Since the end of the pandemic, more people have rolled into real estate marketing and started analyzing the factors affecting the current house prices. In the past, multiple studies have explored various factors that influence house prices around the world utilizing multiple regression. For instance, in the article *The Use of Multiple Linear Regressions in Determining the Relationship between Housing Unit Price and Some Major Components in a Real Estate Building* by Paul Boye, he concludes that from his research, there is a "perfect relationship between housing unit price and some major components both within and around the real estate building" (15) in Ghana. In *A Case Study on Determination of House Selling Price Model Using Multiple Regression* by H. J. Zainodin and G. Khuneswari, they studied the predictors variables of floor area (square feet), number of rooms, age of house (years), number of bedrooms and number of bathrooms against house pricing using multiple regression (27). Similarly, in *Research on the Influencing Factors Affecting Beijing House Prices Using Linear Regression Model* by Mingwei Xu and Zhaojing Yang, they focused their study on influencing factors of house pricing, where they narrowed from 10 factors to 3 influential predictors (411). With consideration of previous studies, in our study, we plan on determining the significance of predictors such as the average income of residents, house age, number of rooms, number of bedrooms, and area population on the response variable of house pricing in the USA. We will be utilizing a multi-linear regression model to analyze the dataset because it allows us to see how the predictor variables collectively impact the response variable of house prices.

Method:

After loading the data into R studio, the model begins with data cleaning to ensure the integrity of the analysis. After, the cleaned dataset is then randomly split into two parts: training data with 70% of datasets, and testing data with 30% of datasets. This split ensures that the model can be trained on a substantial portion of the data while leaving a representative sample for validating the model's performance.

Next, the model is built from training data to analyze the effects of various predictors on house prices. Model 1 is obtained by fitting the model with all possible variables that would affect the house prices on Exploratory Data Analysis (EDA). It includes average area income, average area house age, average area number of rooms, average area number of bedrooms, and area population. If data points with p-values greater than 0.05, they should be removed since they are not significant. However, if the p-values are less than 0.05, they are considered significant to the model. The significant data forms model 2. Subsequently, all p-values of model 2 are rechecked to ensure their significance.

To find the best model, model 3 is built with fewer variables and compared with model 2 in a partial F test. If the p-value is less than 0.05, choose the more complex model. If the p-value is larger than 0.05, choose the model with fewer variables.

To ensure that the model does not suffer from multicollinearity, the Variance Inflation Factor (VIF) is then calculated in each model. Variables with a VIF of more than 5 are considered to have high multicollinearity and are evaluated for removal in the model. If the VIF is less than 5, it indicates that there is no significant multicollinearity in the regression model. Therefore, the predictors are not highly correlated with each other and can remain in the model.

After selecting the ideal model, the next step involves checking whether it satisfies the conditions for regression model conditions. By building a scatter plot, if data points are randomly scattered around a straight line, then there is a linear relation and it satisfies condition 1. Otherwise, it is unsatisfied. Next, plots are created for any numerical predictors in pairs. If a linear pairwise relationship exists within each pair, condition 2 is satisfied. Otherwise, it is unsatisfied.

Subsequently, a residual plot for predictors and fitted values is constructed, along with a QQ plot to assess whether the model satisfies the four assumptions of a linear model: linearity, constant variance, normality, and independence. If there is a linear relation between x and y, then the model satisfies linearity. If the residuals in the plot are evenly spread around the horizontal axis without any systematic increase or decrease, then the model has constant variance. If the residuals closely follow a straight diagonal line in the QQ plot, then the residuals are normally distributed. If there is no cluster in residual plots, then it is independent.

After confirming linearity, several statistical criteria are applied to the model. To measure the discrepancy between the data and the estimation model, minimizing the sum of squares of residuals is crucial. High values of Adjusted R-squared indicate a greater proportion of variance in the dependent variable explained by the independent variables. Additionally, checking the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) of the model is essential. Lower values of AIC and BIC suggest better models in terms of explaining the variability in the data.

Leverage points, outliers, and influential points are also necessary while checking the model. If the value of data is larger than the hat value, it is a leverage point. If the value is not between -2 and 2, it is an outlier. By checking Cook's distance, influential points can be identified.

Finally, for model validation, after confirming that all assumptions are satisfied with the training data, the preferred models are fitted with the testing data. To find minimal differences in the estimated regression coefficient, repeat the previous steps to check conditions, statistical criteria, multicollinearity, R-squared and Adjusted R-squared. By observing the data, if the characteristics of testing data are not similar to those of the training data, it indicates potential over-fitting or under-fitting. If they are similar, the final model with testing data is the ideal model and can be used for predictions.

Result:

Predictor variables (x_i)	**Description**
Avg.Area.Income	Average income of residents in the area
Avg.Area.House.Age	Average house age in the area
Avg.Area.Number.of.Rooms	Average number of rooms in the area
Avg.Area.Number.of.Bedrooms	Average number of bedrooms in the area
Area.Population	Area population in the area

Response variable (y)	**Description**
Price	Housing price (in $)

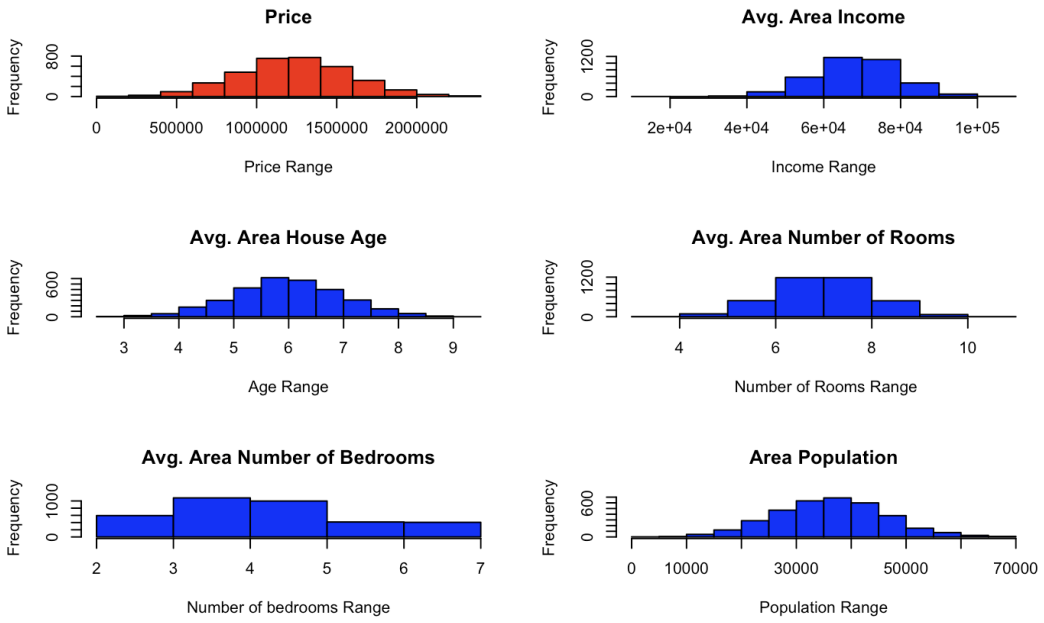

Figure 1: Histogram of response and each predictor

Before analyzing the data, we first constructed a histogram for our response (price) and predictors such that it indicates our response value is normally distributed. Then we cleaned the data such that all the missing values could be reviewed. Then we randomly split the data values into 70% of the training set and 30% of the testing set, therefore there are 3500 data as the training set and 1500 as the testing set from a total of 5000 data points. After splitting the data, the first model m1 is constructed using the following predictors: average income of the area, average house age, average number of rooms, average number of bedrooms and population size in the area.

By computing the global F test we get a result of 7512 and the p-value is less than 0.05, the model is statistically significant. However, the average number of bedrooms in the area has a p-value of 0.399 which is greater than 0.05, therefore this predictor is removed to obtain a reduced model m2, which contains the following predictors: average income of area, average house age, average number of rooms and population size in the area.

In model 2, the partial F statistics is 9391 and has a p-value that is less than 0.05, which also concludes that the model is statistically significant. In addition, the p-value of each predictor is less than 0.05, indicating that all predictors are significant in the model. The model is also a good fit given the adjusted R-squared is 0.9148, which is close to 1.

```
Analysis of Variance Table

Model 2: Price ~ Avg..Area.Income + Avg..Area.House.Age +
Avg..Area.Number.of.Rooms +
    Area.Population
Model 3: Price ~ Avg..Area.Income + Avg..Area.House.Age + Area.Population
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1   3495 3.6684e+13
2   3496 8.7815e+13 -1 -5.1132e+13 4871.5 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: Partial F test ANOVA between m2 and m3

A third model (m3) is constructed to test to compare with model 2, where we randomly removed more predictors in m3. In the third model, we chose predictors “the average income of the area”, “average house age”, and “population of the area” computed a partial F test and used ANOVA to compare the two models. Since both models m2 and m3 have p-values that are less than 0.05 from the partial F test, it shows that both models are statistically significant. Therefore, we chose the more complex model m2 since it has more predictors and set it as the ideal model.

Then, multicollinearity can be checked by finding the VIF of the model m2. Since all VIFs are less than 5 (VIF for Average income: 1.001194, VIF for Average House age: 1.000954, VIF for Average number of rooms 1.001451 and VIF for population: 1.001160), thus there is very low to no multicollinearity among the predictors in the model.

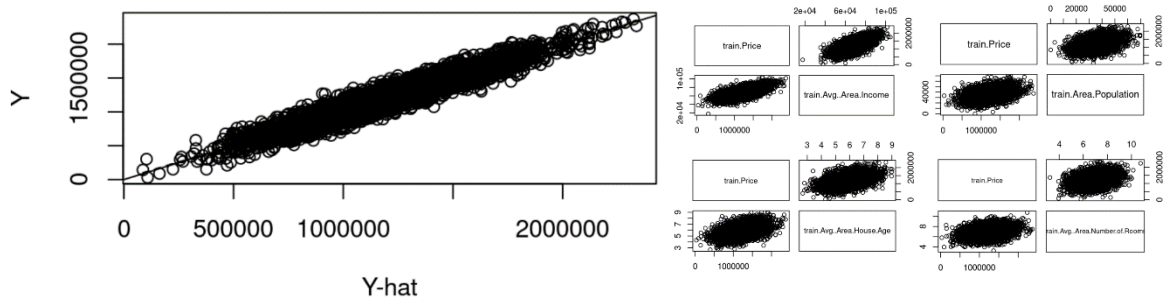

Figure 3: Pairwise scatter plot matrices for each predictor used in m2

After setting our ideal model, we then created a scatter plot for our actual values (y) and predicted values (y-hat) to see the linearity of the two values so condition 1 is satisfied.

By graphing the pairwise scatter plots between each predictor in model 2 and the response value (housing price), we can check if there are non-linear patterns that would violate the linear regression model. From the scatter plots, we can conclude that there are no apparent non-linear patterns in the pairwise scatter plots for each predictor used in model 2.

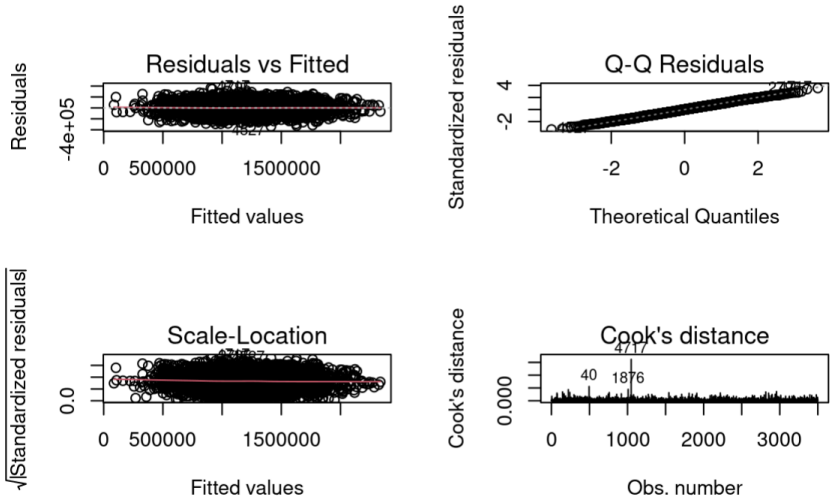

Figure 4: Linearity and homoscedasticity test through residuals analysis (training data)

From Figure 5 we designed two plots to test if the residuals are randomly scattered around the horizontal line at zero. Since this condition is met, it can be concluded that linearity and constant variance are met in the chosen model. Furthermore, there is a clear linear pattern forming a 45-degree line in the Q-Q Plot, which validates the assumption that

all residuals are following a normal distribution. From the graph of Cook's distance, we can see that there are three potential influential points: 40, 1876, and 4717. While checking the problematic points, we get 224 leverage points, 282 outliers, and 0 influential points. By looking at these points, there is no contextual reason to remove compared to a large number of training data.

Then we compute SSR, R-squared, Adjusted R-squared, AIC, AICc, and BIC using training data (detailed values please see Appendix to check if the model is well-fitted. They meet the requirements of a well-fitted model.

Last but not least, a new model (m4) is introduced where its predictors are composed of average income in the area, average housing age, average number of rooms and population in the studied area. It shares the same predictors as Model 2, but we used testing data instead of training data in Model 4 to ensure consistency in performance in the linear regression model.

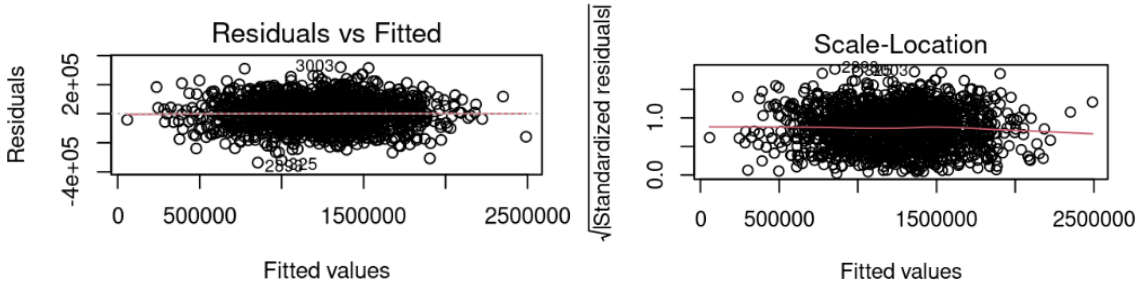

Figure 5: Residual analysis Linearity and homoscedasticity test through residuals analysis (testing data)

To ensure that the model's performance should be similar when using either testing data or training data, a scatter plot is used on model 4 to check if the output is similar to our model 2 with training data. Since both the scatter plots and the Q-Q plot resemble the previous plots from training data, model 2 is not overfitting and is applicable to both training and testing datasets.

Training Data:

MSE	RMSE	R^2
10481054994	102377	0.9148798

Testing Data:

MSE	RMSE	R^2
9599960572	97979.39	0.9251403

Discussion:

Through our research, we could statistically see the impacts and significance of predictors, such as the average income of residents, house age, number of rooms, and area population on house pricing in the USA. Since we aim for a model that is a combination of good predictions and descriptive results, we use model 2. Our final model is:

$$\widehat{price} = -2622000 + 21.49 * Avg.\ Area\ Income + 165700 * Avg.\ Area\ House\ Age + 120600 * Avg.\ Area\ Number\ of\ Rooms + 15.11 * Area\ Population.$$

Our model does answer our research question as it showcases the statistical significance of each of our chosen predictor variables. Based on the model, the most influential factor is the average area house age. In addition, as the average area house age increases by one unit and other variables remain constant, the house pricing would increase by 165700 units on average. This reflects a modern preference for newer homes that meet present standards in terms of design, safety, and facilities. In the current real estate market, newer houses are often associated with better construction quality, modern designs, and updated facilities, making them more desirable and thus more expensive. The equation also shows that the least influential factor is the area population. As the area population increases by one unit and other variables remain constant, the house pricing would increase by 15.11 units. While population density can affect housing demand, it appears to be a less critical factor compared to others in our model. Although area population can drive demand for housing, other factors such as income and house features tend to play a more significant role in the determination of house prices. As stated by H. J. Zainodin and G. Khuneswari, various factors such as "a person's willingness/readiness to buy a house, income status, and the facilities around the housing areas can also affect the house selling price." (H. J. Zainodin, G. Khuneswari, 2009, p43)

Similar to previous research on the topic of house prices, our research confirms that variables such as income, house age, the number of rooms, and area population as significant predictors and factors of house pricing. However, due to the limited size of our dataset, our conclusions may not fully represent the entire US housing market, which restricts the generalizability of our findings. Additionally, outliers and leverage points can impact the accuracy and quality of our model. Our study highlights the complex factors that influence house prices in the USA. By focusing on average income, house age, number of rooms, and area population, we have pinpointed some key factors of market trends. However, since the housing market is multifaceted, continuous research is essential to adapt to evolving trends. Understanding these factors not only aids buyers in making informed decisions, but also identifies areas where government intervention may be necessary to ensure equitable access to housing. As the market evolves, so must our analytical approaches, ensuring that they remain relevant and reflective of current realities.

Appendix

Training Data:

SSres	Rsq	Rsq_adj	AIC	AIC_c	BIC
3.668369e+13	9.148798e-01	9.147823e-01	8.076292e+04	8.076295e+04	8.080389e+04

References

Boye, P. (2012). The Use of Multiple Linear Regressions in Determining the Relationship between Housing Unit Price and Some Major Components in a Real Estate Building. Scottish Journal of Arts, Social Sciences and Scientific Studies. https://www.researchgate.net/publication/332423615_The_Use_of_Multiple_Linear_Regressions_in_Determining_the_Relationship_between_Housing_Unit_Price_and_Some_Major_Components_in_a_Real_Estate_Building

Moriya, B. (2021, June 18). *USA housing*. Kaggle. https://www.kaggle.com/datasets/bhavinmoriya/usa-housing

Xu, M., & Yang, Z. (1970, January 1). *Research on the influencing factors affecting Beijing house prices using linear regression model*. SpringerLink. https://link.springer.com/chapter/10.1007/978-981-99-6441-3_37

Zhu, Q. (2014, November 15). *A case study on determination of house selling price model using multiple regression*. Academia.edu. https://www.academia.edu/9318805/A_Case_Study_on_Determination_of_House_Selling_Price_Model_Using_Multiple_Regression