Application of linear regression on Beijing's housing price

Contributions:
Feng Kevin Wu: Drawing diagrams and plots in the result section and analyzing some of them.

Hanzhong Li: Writing the method part about how we approach our final model.

Zhongyuan Hu: Working on the discussion part. Mainly analyzing the reliability of the model with reality. Secondly checking its validation under the comparison of training data and test data.

Yi Qian: Concluding relation of linear regression and the background of housing price in Beijing.

Jiaxuan Jasmine Tian: Making plots for the 4 different models and analyzing some of them.

Introduction:

Our research revolves around a central question: How do factors such as building type, property size, number of living rooms, bathrooms, ladder ratio, presence of an elevator, kitchen, drawing room, and proximity to subway stations collectively influence housing prices in Beijing? For instance, studies such as Xiao et al. (2019) on the effects of floor level and landscape proximity on housing prices provide insights into the non-linear relationships that can exist between these variables and housing costs. Similarly, research by RBC (2013) and Xiao et al. (2017) delve into the multifaceted nature of housing market dynamics, highlighting how factors like population growth, economic conditions, and access to transportation hubs like subways significantly affect housing prices. By leveraging a linear regression model, we seek to quantify the impact of these variables on the price-per-square meter, offering insights into the relative importance of each factor. The choice of a linear regression model is informed by the continuous nature of our primary dependent variable, the price-per-square meter. This model allows us to understand the linear relationships between various predictors and housing prices. The beta coefficients of these predictors, obtained from the fitted model, will indicate the change in housing prices attributable to a unit change in each predictor, holding other factors constant. Our research is supported by a comprehensive review of peer-reviewed academic literature, providing a theoretical foundation and contextual understanding.

Methods:

For the beginning of the model construction, we imported the dataset into our RStudio. After that, just before we built our model, we did the data cleaning. In the process of data cleaning, we delete the useless data like the rows that have the value of NAs. When we cleaned our data, we divided the whole data set into two parts, the training data and the other is test data, since we want the chance of appearance the extreme value in the training and the testing data to be similar, we decided to let the amount of observations in the training and testing to be equal, which means the ratio of the observations in training and testing is 1:1. In summary, training data is 50% of the

cleaned dataset and test data is also 50% of the cleaned dataset. Then it is time for us to build our model.

According to the references in the introduction part, we know that many factors affect the unit price which means the price per unit square meter; for example, whether the subway is nearby or not. In one word, we select 9 variables for our first model, let us name it model1. After we build our model by using our training data, we first take a glance at the level of significance of our variables. If we find the variables that are not significant compared with other variables, we will delete that variable and then we get the new model named model2. Now with model 2, we want to use the stepwise selection method. We start with our model, then we calculate the AIC, after that, we first delete a predictor and check whether the AIC can be smaller. If it can be smaller, we will use the model that deleted the variable named model3; next, we repeat that procedure (which is the procedure of adding or deleting the predictor and then calculating the AIC) until we are in a situation that whether how we adjust the predictors in the model the AIC can not be smaller. Then we decided to use that model and give this model the name: model_final. Next step, we will check Condition 1, Condition 2, and all the assumptions we need to check relating to linearity, constant variances, uncorrelated errors and the QQ-plot normally check.

After checking the assumptions above, we will take a look at the scatter plots of our model_final, this is to check whether there are too many leverage points, outliers, and influential points. We adjusted our model a little bit by the analysis above and got the new model: model adjusted. Here we used the method of partial F test, whose H_0 is the beta of the predictor we deleted is zero. After we do the partial F test, we will get the conclusion of whether we can reject our null hypotheses. If we reject the null hypotheses we will use our model_final, if we do not have enough evidence to reject the null hypotheses, we will use our new model adjusted.

In the last step, we will use the knowledge we learned in module 10 to validate our model. We need to check that all the characters look similar in both the training data and the testing data. If they are similar we can conclude that our final model is good.

Result:
For our dataset, the following tables are the summary of some important independent variables.

elevator <dbl>	num_lines <int>
0	66628
1	92748

Table 1: the number of elements in elevator variable

subway <dbl>	num_lines <int>
0	64136
1	95240

Table 2: the number of elements in subway variable

buildingType <dbl>	num_lines <int>
1	41815
2	33
3	30253
4	87275

Table3: the number of elements in building type variable

```
     square          livingRoom     drawingRoom       kitchen      
 Min.   :  7.37   Min.   :0.000   Min.   :0.000   Min.   :0.0000  
 1st Qu.: 57.74   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0000  
 Median : 73.64   Median :2.000   Median :1.000   Median :1.0000  
 Mean   : 82.59   Mean   :2.006   Mean   :1.143   Mean   :0.9917  
 3rd Qu.: 97.63   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.0000  
 Max.   :640.00   Max.   :7.000   Max.   :5.000   Max.   :3.0000  
    bathRoom      ladderRatio      
 Min.   :0.000   Min.   :       0  
 1st Qu.:1.000   1st Qu.:       0  
 Median :1.000   Median :       0  
 Mean   :1.182   Mean   :      63  
 3rd Qu.:1.000   3rd Qu.:       0  
 Max.   :6.000   Max.   :10009400  
```

Figure 1: summary of other predictors

These variables are chosen since they have a great correlation with our response predictors, which is housing price, according to the reference. Moreover, they do not have a lot of Na, which may generate bias for the dataset.
Based on this dataset, we use a function to generate the original model as M1, which can be explained as the following equation.

$$\hat{y}_{price} = \hat{\beta}_0 + \hat{\beta}_1 x_{square} + \hat{\beta}_2 x_{livingroom} + \hat{\beta}_3 x_{drawingroom} + \hat{\beta}_4 x_{kitchen} + \hat{\beta}_5 x_{bathroom} + \hat{\beta}_6 x_{BT-plate-and-tower} + \hat{\beta}_7 x_{BT-plate} + \hat{\beta}_8 x_{BT-tower} + \hat{\beta}_9 x_{ladderRatio} + \hat{\beta_{10}} x_{Elevator-Yes} + \hat{\beta_{11}} x_{Subway-Yes}$$

Then we start to check if this model satisfies the assumption for multiple regression,

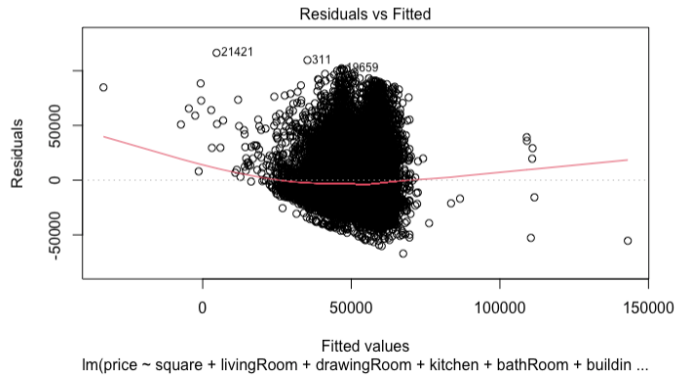

Figure 2: Residual plot against fitted value:

The data all gathered near zero for this plot, which satisfies the assumption.

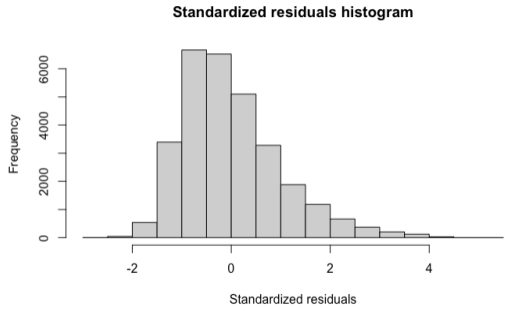

Figure 3: Standardized residual against the fitted value in histogram

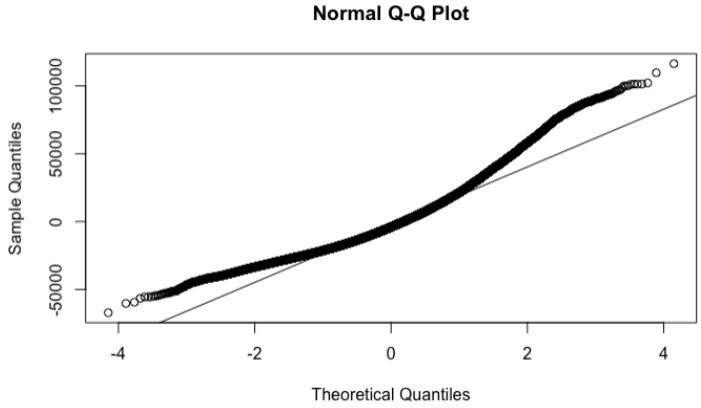

Figure 4: QQplot for M1
The QQ plot looks like a line, which satisfies the assumption

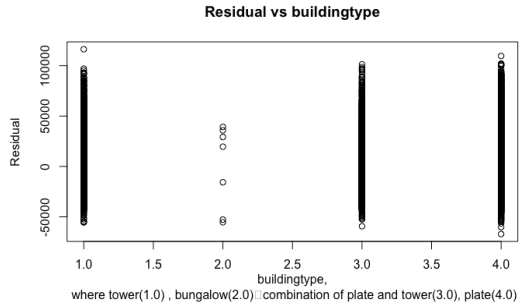

Figure 5: Residual plot for the building type

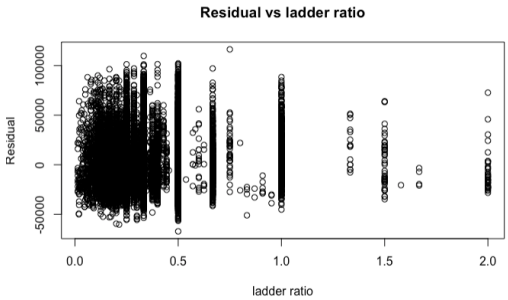

Figure 6: Residual plot for the ladder ratio

By checking the residual plot for the building type and ladder ratio, we found two problems. For building types, there are few type 2, which leads to a non-constant variance. For ladder ratio, there are many outliers in the plot, which influence the accuracy of our model. Therefore, removing these two variables may produce a better model.
M1: the original model.
M2: ladder ratio is deleted from M1
M3: ladder ratio and building type are deleted from M1
M4: building type is deleted from M1
Therefore, to find our final model, we decided to compare these four models using the knowledge we learned in the lecture.
We made plots of Cook's distance to these four models to illustrate the influence of the observation, where Cook's distance is used to estimate the influence of a data point when performing a least-squares regression analysis, showing the influence of each observation on the fitted response values.
For M1,

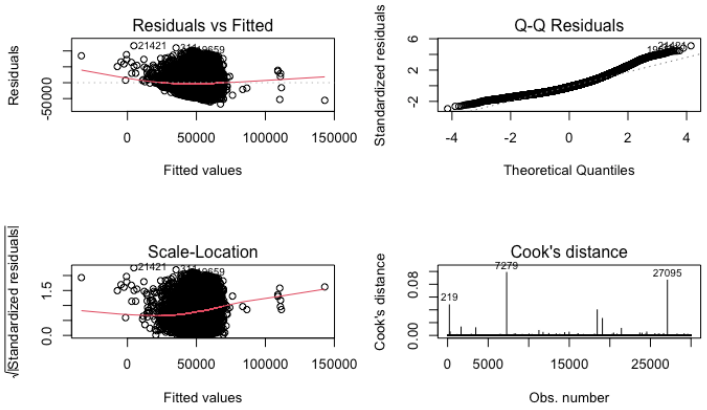

Figure 7 : Residual plot, QQ plot, Standardized residual against the fitted value in histogram and plot of Cook's distance for M1

Several spikes stand out above the rest, particularly the ones labelled with observation numbers like 219, 7279, and 27095. These spikes suggest these points have a higher influence on the model's parameters. The Cook's Distance plot here suggests the model has a few observations that might be disproportionately influencing the regression results, which should be investigated further to understand their impact on the model. The influential points have a higher Cook distance compared to the rest, indicating a greater potential to alter the model fit.

For M2,

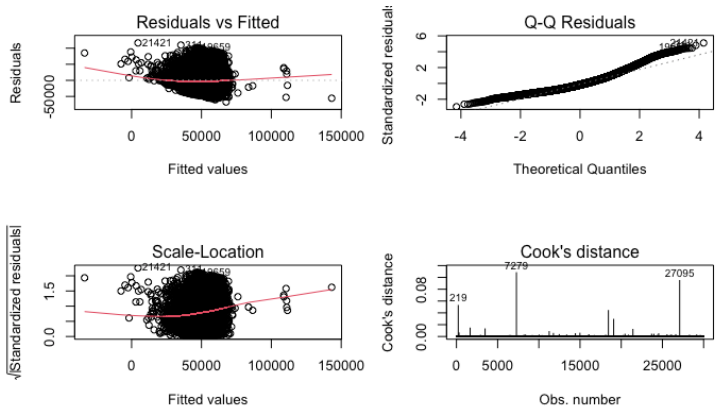

Figure 8 : Residual plot, QQ plot, Standardized residual against the fitted value in histogram and plot of Cook's distance for M2

The Cook's Distance plot for M2 looks similar to the M1's, suggesting that the removed variable had little to no effect on the influence of individual observations in the regression model. To decide our final model precisely, we still need to check other diagnostic measures such as adjusted R-squared, AIC, BIC, etc., to see if they have changed significantly with the removal of the variable.

For M3,

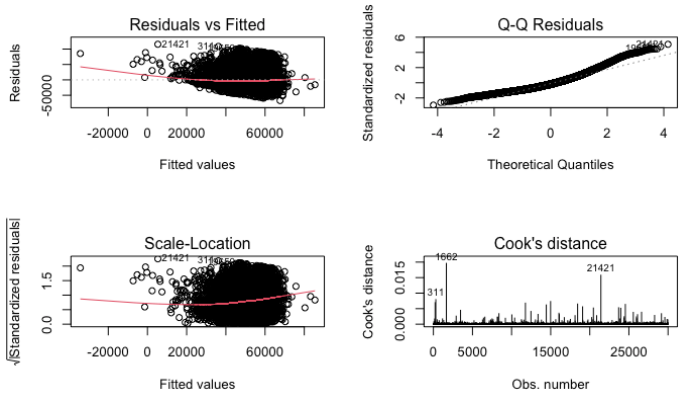

Figure 9 : Residual plot, QQ plot, Standardized residual against the fitted value in histogram and plot of Cook's distance for M3

The Cook's Distance plot for M3 indicates a few potentially influential observations. The labelled points, such as observations 311, 1662, and 21421, have noticeably higher Cook's distances compared to the rest, suggesting they have more influence on the model's coefficients. The comparison between the plot for M3 and M1/2 shows that the regression model corresponding to the first 2 plots may be more sensitive to the removal of its influential observations than this one. This is because the influential points in the first two plots have a higher Cook distance, indicating a greater potential to alter the model fit.

For M4,

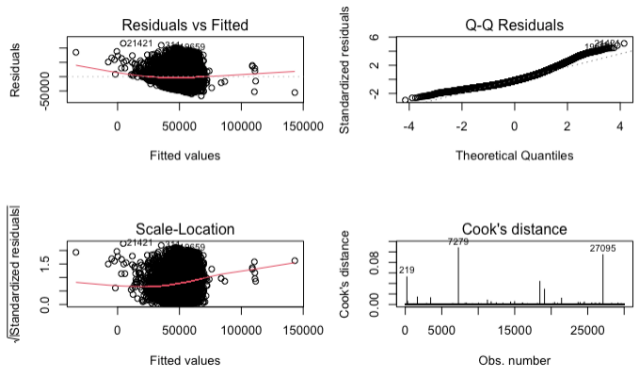

Figure 10 : Residual plot, QQ plot, Standardized residual against the fitted value in histogram and plot of Cook's distance for M4

The points labelled with observation numbers (219, 7279, and 27095) are significantly higher than the others, suggesting that they are potential outliers or have high leverage and may disproportionately influence the regression model's coefficients. The plot with higher Cook's distances indicates a model that is more sensitive to specific data points. Decisions about whether to keep or remove these points should be made in the context of domain knowledge and the objectives of the analysis.

To decide which model is the “best” model for our prediction of the research, we need to calculate each model’s Adjusted R^2, AIC, BIC and corrected AIC values to make the final decision.

	Adjusted R^2	AIC	BIC	corrected AIC
M1	0.1098343	602240.9	602340.6	602240.9
M2	0.1098363	602239.8	602331.2	602239.8
M3	0.1079643	602299.8	602366.3	602299.8
M4	0.1079399	602301.6	602376.4	602301.6

Table 4: The summary of Adjusted R^2, AIC, BIC and corrected AIC for M1, M2, M3, M4

By comparing the adjusted R^2, AIC, BIC and corrected AIC, we found that the M2 is a better model. Therefore, we chose M2 as our final model.

Discussion:

After the data processing, we have our final model which is the relationship between house price and predictors(square, number of living rooms, number of drawing rooms, number of kitchens, number of bathrooms, type of buildings, whether elevator included, and whether this house is close to the subway). According to our model, we could find that with the increase of living room, kitchen and bathroom, and with the condition of the elevator included and close to the subway, the house price tends to increase. By the way, being close to the subway also adds value

to this house. Conversely, more drawing rooms lead to a lower house price. Moreover, despite whatever kind of building, the price will drop absolutely.

After comparing the final model under training data and test data. Firstly, all of these differences between coefficients are less than 2 times of standard error correspondingly. Secondly, both of them have similar adjusted R square. Finally, the significance of each predictor is unchanged. All these three criteria show validation of final model passes.

With the conclusion above, this model most likely predicts the house price in Beijing. Beijing is one of the most popular cities in China, as the capital of China. Around 21893095 people live here, most people consider functions instead of comfort when they purchase houses. According to the research, in China, house prices would be affected by the Two-Child policy.(2022, Na Li, Rita Yi Man Li, Jotikasthira Nuttapong). This makes sense, people are encouraged to have more children in a family, those who have a small house originally, demand a house with more functions, like more living rooms, bathrooms and kitchen, because these are essentials to live for a bigger family. Since functional houses are in high demand, the price of these houses with large areas most likely will drop due to the supply-demand relationship. Moreover, both elevators and subways provide more convenience to residents. To sum up with this model, most factors which will affect house prices are considered, and with a close relationship with reality.

But the model cannot be perfect, there must be some limitations and undesirable predictions. In our model house price was indicated to drop with both types of houses. However, towers are most likely located in the city centre, and the house prices in the city centre and outside of the centre are around 11,024.29 ¥ and 5,361.06 ¥ respectively(2023, NUMBEO). This shows a converse direction to our prediction, and this will affect our prediction accuracy. The possible reason for this lingering issue is the lack of data. Because the data comes from a real estate agency, these houses are all for sale, which does not include those houses not for sale. Those houses are listed for sale for many reasons, for example, bad traffic, bad building conditions, or other personal issues. These houses can not represent the whole house market overall.

To sum up, the model is fitted to assist people, not to make decisions for people, there are far more unexpected factors than those represented in the model.

Reference

1. (N.d.). *Property Management*. https://doi.org/10.1108/pm
2. *Property prices in Beijing*. Property Prices in Beijing. Price per square feet/square meter in Beijing. (n.d.). https://www.numbeo.com/property-investment/in/Beijing#:~:text=Buy%20Apartment%20Price%0A%0A,8%2C361.20
3. RBC. (2013, November). *Priced out: Understanding the factors affecting home prices in the GTA. - RBC*. Priced Out: Understanding the factors affecting home prices in the GTA. http://www.rbc.com/community-sustainabilit y/_assets-custom/pdf/Priced-Out-RBC-Pemb ina.pdf

4. Xiao, Y., Chen, X., Li, Q., Yu, X., Chen, J., & Guo, J. (2017, November 15). *Exploring determinants of housing prices in Beijing: An enhanced hedonic regression with open access poi data. MDPI.* https://www.mdpi.com/2220-9964/6/11/358
5. Xiao, Y., Hui, E. C. M., & Wen, H. (2019). Effects of floor level and landscape proximity on housing price: A hedonic analysis in Hangzhou, China. *Habitat International*, *87*, 11–26. https://doi.org/10.1016/j.habitatint.2019.03.008