Model of Housing Price Prediction in Beijing by Multiple Linear Regression
Group 120 - STA 302 Final Project
December 2023

Contributions:

Names of Group Members	Contribution to Report
Tony Chen	Results, Discussion
Ziqing He	Introduction, Ethics
Yiqi Feng	Methods
ZiAng Liu	Ethics, Discussion

1. Introduction

This research analysis aims to study how the nine internal and external factors, including latitude, house active day on the market, number of people following the transaction, the community average prices, age, square area, and the floor of the building, and whether there is an elevator inside or is subway station near the building (categorical variables), affect the house price in Beijing in 2017 (response variable). In the previous study (Ma, et al., 2022) researchers used the data from second-hand houses with elevators installed in old residential areas in Beijing to infer a relationship between having elevators and the housing price. This motivates us if relationships could be moved to more general building types. In another study based on Hangzhou houses (Xiao, et al., 2019), they found four factors that mainly affect the house price: the number of house floors, the surrounding environment of the house, the average price of these whole communities, and the public attention of the house. As the results from the article, it can be found that the higher the floor, the better the environment near the house, the more focus on the house, the higher the average price of the community, the higher the price. Since Beijing and Hangzhou have identical geographical characteristics, we want to study whether this relationship can also be applied to houses in Beijing. Finally, according to the result of another article (Xu et al., 2017) in our study, the age and area of the house are also essential factors affecting the house price. The article pointed out that the older the house, the larger the area, the higher the housing price. In all three research papers, it is illustrated that there is a linear relationship between each predictor and response variable, so we decided to use multiple linear regression (MLR) models to study the effect of these predictors together on the Beijing house price. After model selection, we chose 7 variables (by removing the variables “house active day on the market” and “whether there is a subway station near the building”) in our final model to explore their influence on house prices.

2. Methods

2.1 Data Cleaning

The original data was sourced from Kaggle and fetched by the author based on the Chinese housing platform Lianjia.com from 2011 to 2017 (Kaggle, 2017). To make the data align with our research question, we chose the subset of traded houses in 2017 that are at least 130 square meters and removed all housing items that contain missing values in their associated variables. Before fitting our model, we made exploratory data analysis of different

plots including histograms, scatterplots, and bar plots (Figure A2. EDA plots) to the price and our interested variables.

In general, we sampled 1811 houses from our data subset as training data to fit the models and assigned the other 1812 houses to the test data for model validation. The preliminary model 1 is presented below:

Model1

$$y = \beta_0 + \beta_1 x_{lattitude\ difference} + \beta_2 x_{DOM} + \beta_3 x_{followers} + \beta_4 x_{community\ average} + \beta_5 x_{house\ age} + \beta_6 x_{square} + \beta_7 x_{floor} + \beta_8 I_{subway} + \beta_9 I_{elevator} + \epsilon$$

2.2 Assumption Assessment and Transformation Tools

Conditions need to be checked first, with the conditional mean response condition and the conditional mean predictor condition. Condition 1 holds since the points scatter randomly around the 45-degree diagonal straight line, as shown in the figure of response versus fitted. Condition 2 holds when there is a lack of curves or other linear patterns in the pairwise scatterplots between predictors. After two conditions holding, we made four more assumptions.

First, with the check of normality by whether there is significant deviation and wiggling from the diagonal line Q-Q plot, if the result is yes, then the violation of normality will be checked. Next, any systematic patterns of residual plots, especially curves or other functions of predictors that are non-linear patterns, indicate violations of linearity. As to the uncorrelated errors, the large clusters of many points or patterns across time will show a violation. For any violation of these three assumptions, we applied the Box-cox Power transformation which takes variables to suggest powers based on the Wald power estimates of predictors and the approximation of the maximum likelihood estimator for the response variable. After all the above assumptions are held, the constant variance assumption will be checked by whether there are any systematic patterns, especially a fanning pattern, that show a violation. If the model violates, a Transformation of Variance Stabliaze on the response will be used. With transformation tools, new linear models will be created, and we will check all conditions and assumptions again.

2.3 Hypothesis Tests: T-Test, ANOVA Test, and Partial F Test

ANOVA test is the first test to be used to check whether there is a significant linear relationship between response and at least one predictor. If an overall significant linear relationship exists, we will use individual T-test of slopes to measure the linear relationship between the response variable of house price per square and each predictor which is to test the null hypothesis if the individual slopes are equal to 0. As some predictors that have a non-significant relationship with the house price will be found, a reduced model (Model 3) is created without these specific predictors. However, before applying the full model and the reduced model to the Partial F test, an assessment of conditions assumptions should be taken first. Subsequently, the Partial F test will be applied to compare the full model of Model 2 and the reduced Model 3. The assumption of the test for the null hypothesis is changed to that the coefficients of the reduced predictor are equal to zero. With a p-value greater than 0.05, we fail to reject the null hypothesis and support the linear relationship between the house

price per square and any house active day on the market (DOM) or subway condition is non-significant, and remove the DOM and subway, they are kept in the model in otherwise.

2.4 Multicollinearity by VIF

Multicollinearity shows the probability of a relationship among more than two predictors and how much they impact the response. The measure of the variance inflation factor is computed to quantify how much larger the variance of a coefficient is due to multicollinearity. Generally, a cutoff of VIF > 5 is used to show severe multicollinearity of a coefficient (β), and any VIF >1 means some multicollinearity is present, while multicollinearity could result in incorrect conclusions in terms of significance.

2.5 Likelihood Measures of Goodness(AIC, BIC, AICc)

Akalke's Information Criteria (AIC) is a penalty for complexity; a corrected AIC is used when the number of observations is small or the probability of randomly chosen data points of the predictors. For the implication of the results, both smaller AIC and corrected AIC indicate a better model, Last, a Bayesian Information Criteria (BIC) penalize the model more stringent on complexity, and a smaller BIC represents a better model.

For a full picture, we use adjusted R-squared to find the largest or close to the largest value of the goodness of the model, while the three likelihood criteria are used to measure the smallest or close to the smallest value.

2.6 Problematic Observations

A Leverage observation is an observation that varies the distance from the center of the predictor space that may change the estimated mean response. Hat Matrix is used to calculate the distances between predictor values observed and the center of all predictors to measure the leverage, given how much impact the value of a particular response has on its mean response estimate compared to the other responses. Moreover, if the standardized residuals of points are outside the range between -4 and 4, they are identified as outliers.

Apart from this, influential points are also observed. First, for estimating all fitted values, we use Cook's Distance measure to calculate the difference in the estimated trend between our original model and a new model that fitted by using all observations minus one. The influence on the own fitted value is measured by DFFITS, which uses fewer observations and focuses on the change in the estimated trend value with the same comparison of the two models. Finally, DFBETAS is used to measure the influence of an observation on the estimated value of at least one coefficient, which can quantify the amount of influence and figure out how each coefficient changes with and without each observation.

2.7 Model Validation by Test Data

Validation is the process of checking the performance of the "best" model in an independent and separate dataset from the information used to build the model. One of the reasons is that there might be overfitting of data that only shows good prediction, and the validation dataset will be used to check the model's generalizability. The data will be split into training data and testing data, while we use the training data with previous methods to build our "best" model and use the testing data to fit a test "best" model to compare.

3. Results

3.1 Data Summary

Table1. Summary of Variables

Key Characteristics	latitude difference (degree)	DOM (days)	followers	price (CNY)	square (meter^2)	floor	house age (years)	community average price (CNY)
Min	0.06703	1.00	0.00	6272	130.0	3.00	1.00	24191.00
1st Quantile	0.34219	18.00	8.00	42184	138.4	8.00	9.00	43307.00
Median	0.40435	47.00	26.00	58103	140.7	15.00	13.00	58188.00
Mean	0.40609	69.14	41.49	61234	163.1	15.77	12.52	61240.00
Standard Deviation	0.1017058	73.33	53.61	23986.33	39.7	8.39	4.48	21678.29
3rd Quantile	0.4941	96.00	54.00	76180	172.0	23.00	16.00	75490.00
Max	0.67562	659.00	633.00	149551	532.2	63.00	52.00	152118.00
Range	0.60859	658.00	633.00	143279	402.2	60	51	127927
IQR	0.15191	78.00	46.00	33996	33.6	15	7	32183
Types	continuous numerical	discrete numerical	discrete numerical	continuous numerical	continuous numerical	discrete numerical	discrete numerical	continous numerical

Categorical Status	Elevator Frequency	Elevator Proportion	Subway Frequency	Subway Proportion
0 (No)	742	0.20480265	1596	0.440518907
1 (Yes)	2881	0.79519735	2027	0.559481093

Note:
1. For categorical variable of elevator. "0" means the house does not install an elevator, while "1" means it does.
2. For the categorical variable of subway, "0" represents that the house is nearby a subway, otherwise it is not.

Table 1 Notes: *In the data, the price unit is in Chinese yuan (CNY), and one CNY is exchanged for 5.3 Canadian dollars approximately. According to Table1 of the variable summary, the* ***response*** *variable of* ***price per square*** *indicates a higher mean at 61234 than the median at 58103, which implies that the price distribution has a slightly right-skewed pattern as shown by the (1) and (2) histograms in Figure 1 in both testing and training data sets.*

The others are ***predictor variables****. The variable of latitude difference has a similar mean and median, which demonstrates an approximately normal distribution but there exists an extreme mode of around 0.5 degrees of difference as shown by the (3) and (4) histograms in Figure A2. The* ***followers*** *range from 0 to 633 people, which shows that there are houses not supported by anyone on the platform, and a large standard deviation of about 53 people compared to the mean of about 41 people implies a relatively high spread of the number of followers for different traded houses. The variable of* ***square*** *represents the size of the house, and we investigate the medium-to-large houses in our research. We can observe that the mean of 163.1 m^2 is larger than the median of 140.7 m^2 and the corresponding right-skewed distribution as referred to in the (9) and (10) histograms of squares in Figure A2. The* ***DOM*** *(days of active on market) has a larger mean of about 69 days than the median of about 47 days, so the distribution pattern may be right skewed, and the standard deviation of 73 days also indicates a relatively high spread. In addition, scatterplots (17) and (18) in Figure A2 show that there is a decreasing variance pattern of DOM to the response price. The* ***floor*** *variable has a similar mean and median of value around 15, which implies that most on-sale houses are at medium floor level, and the standard deviation of about 8 floors is a relatively small spread. The distribution may be normal, and according to (15) and (16) histograms in Figure A2, the variance of the floor with respect to the response price is stable and mostly distributed lower than the 30th floor. The* ***house ages*** *have a similar mean and median of around 13 years, and the low spread indicated by the small standard deviation of 4.48 years indicates that the ages of houses are distributed normally as shown by histograms (5) and (6) in Figure A2. The* ***community's average price*** *measures the representative price level of a local community which is an important factor for real estate businesses to set the on-sale price. In the data, we observe that the mean community average at 62140 is larger than the median at 58188, and the distribution is skewed to the right as illustrated by histograms (7) and (8) in Figure A2. There are two binary categorical variables. For the variable* ***elevator****, we can see that most traded medium-to-large houses in Beijing in 2017 include elevator services as demonstrated by the high proportion of 79.5%. Moreover, about 55% of houses are built near the* ***subway*** *and the other 44% are built far away from a subway route.*

In conclusion, the response price per square has a right-skewed distribution, so there are potential outliers in price, and a large portion of observations of house prices are smaller than the mean. In addition, the latitude difference, floor, and house average variables are approximately normal, and the other variables are either skewed or not normally distributed. Furthermore, squares, DOM, and community average price are all skewed to the right, and we considered making a box-cox power transformation to improve normality for some responses and positive numerical predictors that are skewed in the Methods section. For categorical variables, most people have access to elevators in their buildings, and about half of households can live near a subway route.

3.2 Model Summary and Selection

Table2. Summary of Models
T-tests Hypothesis: H0: β = 0, Ha : β ≠ 0

Coefficients:	Train Model 1			Train Model 2			Train Model 3		
Variables	Estimate	Std. Error	p-value (Pr(>|t|))	Estimate2	Std. Error2	p-value (Pr(>|t|))2	Estimate3	Std. Error	p-value (Pr(>|t|))3
Intercept	14510	1548	< 2e-16	162.6	3.825	< 2e-16	162.0	3.794	< 2e-16
latitude difference	-9184	2186	0.00003	-28.43	4.509	0.000000000363	-29.45	4.471	0.0000000000588
followers	-23.5	4.199	0.000000026700	-0.04270	0.00866	0.000000896	-0.04914	0.007859	0.000000000502
community average price	1.02	0.01033	< 2e-16	0.001973	0.0000213	< 2e-16	0.001979	0.00002096	< 2e-16
house age	-224.5	49.84	0.000007	-0.413700	0.1028	0.0000598	-0.3842	0.1001	0.000128
square	-54.58	5.372	< 2e-16	-0.127900	0.01108	< 2e-16	-0.1318	0.01087	< 2e-16
floor	-93.87	30.9	0.002390	-3.969	1.244	0.00144	-3.493	1.203	0.003737
elevator	3550	679.1	0.0000001910	11.34	1.501	<2e-16	11.24	1.501	0.000000000000108
DOM	-7.498	3.326	0.0243	-0.012220	0.006863	0.0752			
subway	680.3	468	0.146420	1.2630	0.9673	0.19167			

Note: These coefficients are for train data and models

Coefficients:	Test Model 3		
Variables	Estimate	Std. Error	p-value (Pr(>|t|))
Intercept	159.2	3.722	< 2e-16
latitude difference	-16.91	4.262	0.0000758
followers	-0.065	0.008135	0.00000000000000237
community average price	0.001987	0.00002095	< 2e-16
house age	-0.2821	0.102	0.00573
square	-0.1642	0.01083	< 2e-16
floor	-3.115	1.197	0.00933
elevator	11.08	1.43	0.000000000000016

Note: These coefficients are for test data and model

Table 2 Notes: *Summary of tables shows the estimated slopes, associated standard errors, and p-values for each predictor in different models.*

By the individual T-tests of slopes, we can check if there was a significantly linear relationship between the associated predictors and the response of price per square for a house. Table 2 shows that the preliminary full Model 1 has mostly significant relationships between the predictors and the price except for the variable of the subway. The p-value for testing whether the slope of indicator subway involvement is equal to 0 is 1.4642 larger than the 0.05 significance cutoff, so the subway is not linearly significant with the house price and we considered deleting this variable later. Before forming the reduced model, we made a box-cox power transformation on the response price by a square root and the predictor floor by a cubic root based on the power transformation table and the maximum likelihood estimate that is close to 0.5 as shown in the likelihood plot in Figure 1.

Figure1. Box-cox Power Transformation to Normality

	Est Power Rounded	Pwr Wald
latitude difference	1.5156	1.52
DOM	0.2365	0.24
square	-3.6333	-3.63
floor	0.3293	0.33
house age	0.9074	0.91
community average price	-0.3018	-0.33

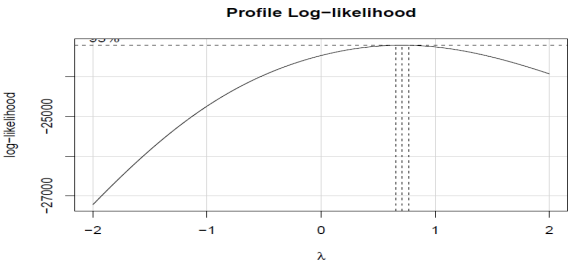


Figure 1. Notes: *The box-cox transformation table shows the estimated power of transformation and the actual power wald statistics (Pwr Wald) we choose to apply the transformation on corresponding variables.*
The plot of profile log-likelihood shows that the maximum likelihood estimate λ *of the expected house prices is close to 0.5.*

Model2

$$y^{1/2} = \beta_0 + \beta_1 x_{lattitude\ difference} + \beta_2 x_{DOM} + \beta_3 x_{followers} + \beta_4 x_{community\ average} + \beta_5 x_{house\ age} + \beta_6 x_{square} + \beta_7 x_{floor}^{1/3} + \beta_8 I_{subway} + \beta_9 I_{elevator} + \epsilon$$

Our transformed full model is recorded as Model 2 and we can see that subway involvement is still a nonsignificant variable and DOM becomes less statistically significant with a p-value of 0.0752 > 0.05 now. Besides, the ANOVA test for model 2 also supports that there exists at least one slope of predictors that are linearly significant since the small p-value

of the test rejects the null hypothesis, In the following stage, we applied the partial F test to see if at least one of the subway involvement and DOM exists a linear relationship with the price. Table 3 shows that the p-value of the partial F test is 0.09136 > 0.05, which implies that at least one of the subway involvement and DOM is not linearly significant with price and thus we decided to remove them and formed our reduced Model 3, and now Model 3 contains all linearly significant predictors since the p-values for each hypothesis t-test of slopes are smaller than 0.05 based on Table 2.

Model3

$$y^{1/2} = \beta_0 + \beta_1 x_{lattitude\ difference} + \beta_2 x_{followers} + \beta_3 x_{community\ average} + \beta_4 x_{house\ age}$$

$$+\beta_5 x_{square} + \beta_6 x_{floor}^{1/3} + \beta_7 I_{elevator} + \epsilon$$

Table3. ANOVA and Partical F Test for model 2 and model 3

ANOVA hypothesis
H0: All slopes of predictors equal to 0
Ha: At least one slope of predictors not equal to 0

Partial F hypothesis
H0: All slopes of k=2 removed predictors equal to 0
Ha: At least one slope of k=2 removed predictors not equal to 0

	Model 2(Full)			Model 3 (Reduced)		
Source	Regression	Residual	Total	Regression	Residual	Total
DF	9	1801	1811	7	1803	1811
Sum Squares	3620428.325	598432	4218860.325	3618836.325	600024	4218860.325
Mean Squares	402269.8139	332.2776235		516976.6179	332.7920133	
F value	1210			1552		
P value	< 2e-16			< 2e-16		

RSS drop (Residual Sum of Square)	1592.4
F value diff	2.3961
Df diff	2
p-value Pr(>F) diff	0.09136

Furthermore, we select our best model for predicting house prices by comparing the multicollinearity and likelihood criteria. According to Table A1, the multicollinearity of predictors is not very serious because the VIFs are usually between 1 and 2 for the predictors in 3 models. Moreover, the average VIFs are the smallest in Model 3 except for the increasing VIFs for the elevator status and the floor, and the other VIFs of predictors remain stable. What's more, the adjusted R^2 (coefficient of determination) is largest in Model 1 and smallest in Model 3 as illustrated by Table 4, but this may be because there are fewer predictors in Model 3 to explain the variation of price, which results in a higher proportion of RSS over SST. In addition, the adjusted R^2 shows a larger model (Models 2 and 1) preferred if extra predictors are linearly significant, but we have shown that the removed predictors are insignificant. Therefore, adjusted R^2 is not the prior consideration in this case, and we compare the AIC, BIC, and AICc based on Table 4 and find that there is a large drop in these criteria between Model 1 and 2, and Model 3 has the smallest BIC while the other two are similar to the ones of Model 2. Therefore, we choose Model 3 as the best model to predict house prices in terms of the largest number of linearly significant predictors related to the price and the smallest likelihood criteria.

Table4. Adj. R^2, AIC, BIC, AICc Comparison

Measures	Model 1	Model 2	Model 3
Adjusted R-squared	0.8678597	0.8573656	train: 0.85714, test: 0.8581
AIC	32922	10524.59	10525.4
BIC	32977.01	10579.6	10569.41
AICc	32922.14	10524.73	10525.5

Table5. Identifying Problematic Observation

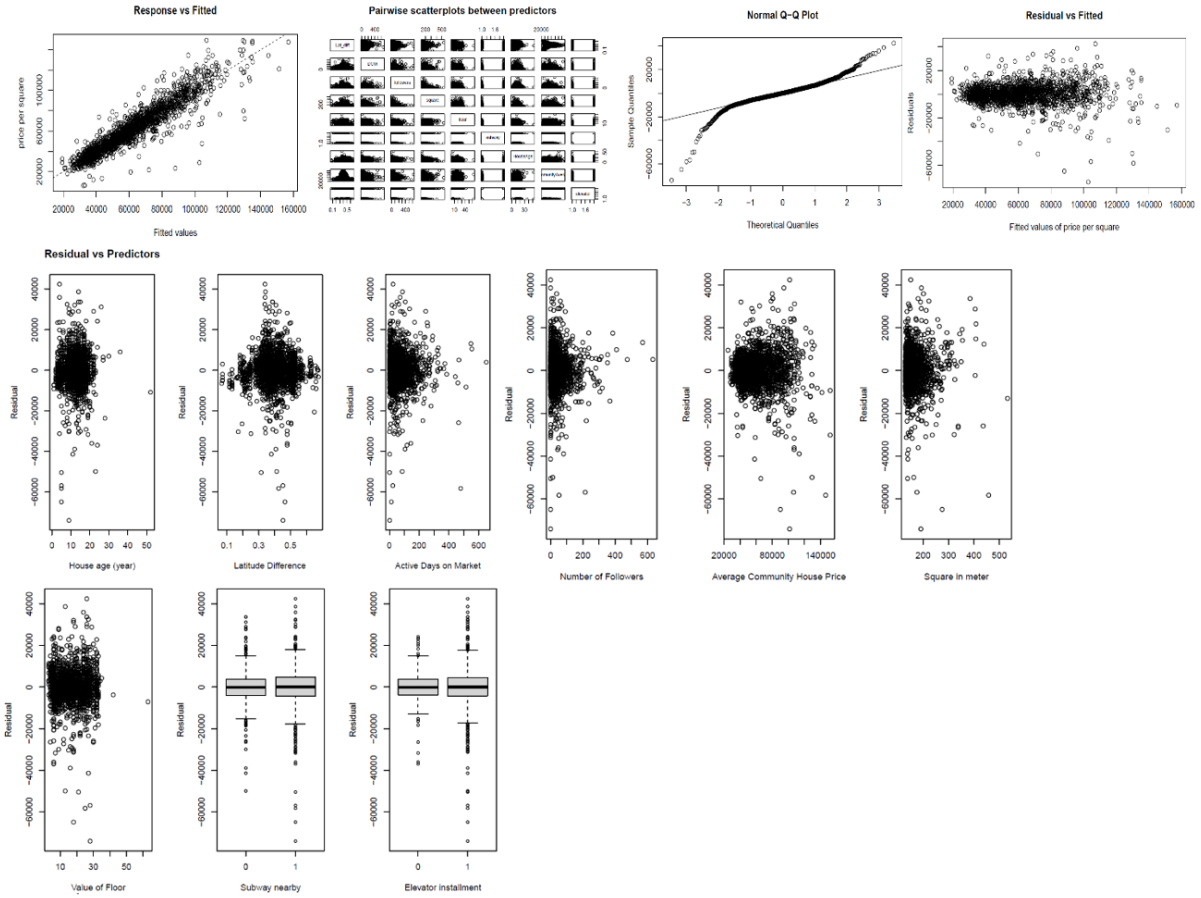
Type of Point	Training	Test
leverage	240	252
outliers	24	24
Influence (Cook)	0	0
Influence (DFFITS)	200	208
Influence (DFBETA)	Between 0-6	Between 0-6
outstanding observations	23, 604, 908, 909, 1314	23, 604, 908, 909, 1314

Table 4 & 5 Notes: *Table 4 compares the adjusted R^2 and likelihood measures (AIC, BIC, AICc) among different models. Table 5 identifies the problematic observations between the training data set for model selection and the testing dataset for model validation.*

3.3 Assumption Verification and Model Validation

Figure1. Multiple Residual Plots of Assumptions and Conditions Verification

Model 1



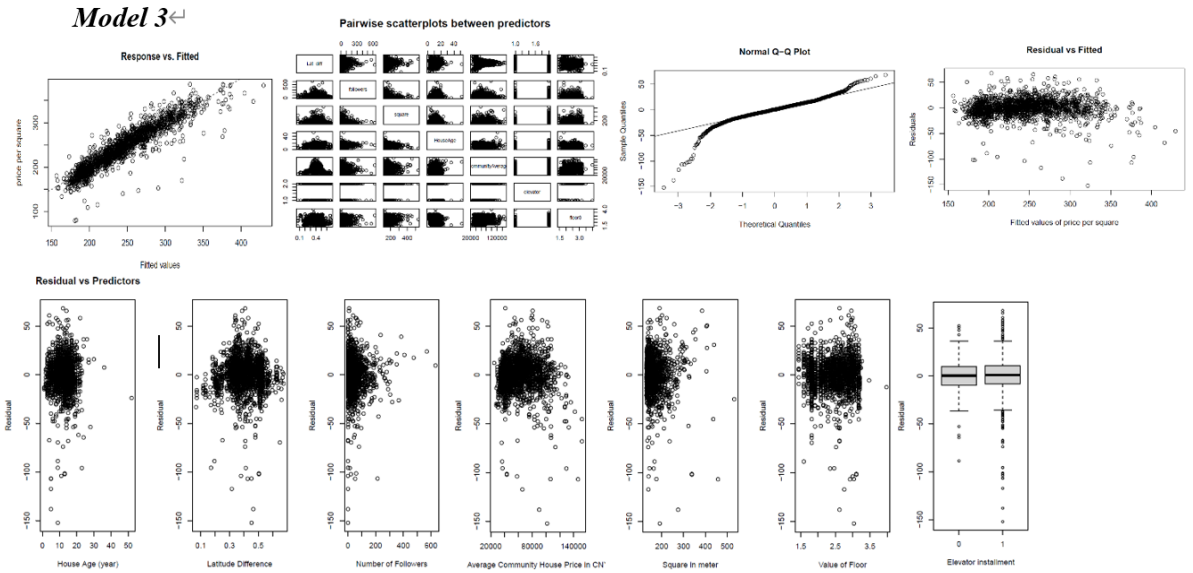


Figure 1 Notes: *These different types of residual plots check if the assumptions and conditions of our Model 1 and Model 3 are satisfied and matched between the training data and the testing data.*

According to conditions of MLR, condition 1 holds since the points scatter randomly around the 45-degree diagonal straight line, as shown in the figure of response versus fitted. Condition 2 holds when there is a lack of curves or other linear patterns in the pairwise scatterplots between predictors, which fit our residual plots of Models 1 and 3.

In addition, we use the residual plots to compare the results of assumption assessments in Models 1 and 3. For linearity, the points are scattered randomly around a horizontal line in the residuals vs fitted plots, and no distinct curves are found in residuals vs predictors plots, representing a more likely linear relationship between the house price per square and its predictors. For constant variances, the figures of residuals with predictors of DOM, number of followers, and square in the meter display a decreasing fanning pattern, and most of the data points concentrate on the left side, which shows that the constant variance assumption is violated. As to the uncorrelated errors, large single clusters appear in the residual plots for both Model 3 and 1, so this assumption is potentially violated. Finally, although normality is violated to some degree for both models, improvements in Model 3 display less deviation and a more fitting line compared to the full Model 1.

According to Table 2 and Table 4, we see that the number of significant predictors is the same, the adjusted R^2 and VIFs for each predictor are similar for both the test and train data. The residual plots of assumptions indicate comparable results in both datasets according to Figure 1 Model 3 and Figure A3. Table 5 also shows comparable results except that there are slightly more observations identified as leverages and influential on their fitted values but still around 10% of test data. However, some estimated slopes are out of 2 standard errors compared to the train data such as the estimated slopes of latitude difference and square, which implies a somehow different trend occurs in two data sets. Therefore, the Model 3 is not entirely validated.

4. Discussion

Estimated Model 3

$$y^{1/2} = 162 - 29.45x_{lattitude\ difference} - 0.04914x_{followers} + 0.001979x_{community\ average} - 0.3842x_{house\ age}$$
$$-0.1318x_{square} - 3.493x_{floor}^{1/3} + 11.24I_{elevator} + \hat{\epsilon}$$

According to the Estimated Model 3, the expected square root (power of ½) of housing prices in Beijing is 162 CNY per square when the latitude difference, followers, community average price, house ages, squares, and cubic root (power of ⅓) of the floor are 0 units and without an elevator installed in the building. For one unit increase in the latitude difference, followers, community average price, house ages, squares, and cubic root of floor, the expected square root of house prices changes in -29.45, -0.04914, +0.001979, -0.3842, -0.1318, and -3.493 CNY respectively when other predictors are fixed (which the "-" means decreasing, and "+" means increasing). Moreover, the expected square root of house prices increases by 11.24 CNY if an elevator is installed compared to the houses without elevators when other predictors are fixed. We have reached the same conclusion in the literature that houses with elevators and with higher community average prices would have higher house prices in Beijing while setting other predictors fixed (Ma, et al., 2022). However, our model disagrees with the result of the literature on the influence of latitude, house active day on the market, number of people following the transaction, age, the area, and the floor of the building on the housing price as these predictors have a negative correlation to the house price in our model, while in the relative literature, they have a positive correlation (Xiao, et al., 2019; Xu et al., 2017).

Some limitations in our research design may potentially affect the accuracy of our model prediction. In EDA analysis, we found that some numerical variables including squares and community average prices are skewed to the right which implies influences outliers and relatively high variability of data. We did not make a corresponding box-cox transformation because simple power transformation does not improve the normality of the final model substantially. Upon checking for problematic observations, we found 240 leverage points, 24 outliers, 0 influential observations on all fitted values, 208 that were influential on their fitted values, and more than 200 observations that were influential on the slopes (betas) from 0 to 6. We also have outstanding observations (23, 604, 908, 909, 1314) that are common in leverages, and outliers, and influential on their fitted values. We choose not to remove these outstanding observations since we want to ensure our model can predict prices for all medium-to-large houses in Beijing.

5. Ethics

In our model selection process, we use a manual selection method instead of automated selection. This is because the automated selection process has some disadvantages. Firstly, it would generate different best models, and these models may not be the best model if we calculate likelihood criteria. This has been reckless because it creates predictable errors. This error is not the target of the automated selection method but is caused by the algorithm of three automated selection methods. Secondly, the automated method can still run even if the initial model provided violates model assumptions. As a result, the output may be incorrect or

unreliable. In addition, it won't consider the background context but operate based on its program. In this way, it may get rid of some useful predictors even though it is proven that they have a significant impact on the prediction of the response variable. They have been negligent. Both situations can be avoided by taking reasonable measures. For example, we can manually select some important predictors based on the results of the literature, although the AIC or BIC of relative models may not meet requirements, and hence eliminated by automated selection. In the same way, we can avoid the problem of violating assumptions by manually checking whether the assumptions are satisfied each time a better model is made. These deficiencies can be largely avoided in manual selection, so we believe that manual selection is more ethical than the other.

References

Housing Price in Beijing. (2017). Kaggle, Retrieved December 10, 2023, from https://www.kaggle.com/datasets/ruiqurm/lianjia

Ma, S., Li, T., & Yang, Y. (2022) Housing price appreciation effects of elevator installation in old residential areas: Empirical evidence based on a multiperiod did model. *Advances in Civil Engineering*, *2022*, pp. 1–10. https://doi.org/10.1155/2022/7949252

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*, pp. 11–26. https://doi.org/10.1016/j.habitatint.2019.03.008.

Xu, Y., Zhang, Q., Zheng, S., & Zhu, G. (2018). House age, price and rent: Implications from land-structure decomposition. *The Journal of Real Estate Finance and Economics*, *56*(2), pp. 303–324. https://doi.org/10.1007/s11146-016-9596-6

Appendix

TableA1. VIFs of models

Variables	VIF Model 1	VIF Model 2	VIF Model 3
latitude difference	1.107203	1.107266	1.087082
followers	1.231656	1.230955	1.012340
community average price	1.174997	1.174934	1.135753
house age	1.196832	1.197293	1.132132
square	1.064141	1.063062	1.022698
floor	1.582990	1.854369	1.732299
elevator	1.696054	1.946794	1.945066
DOM	1.257761	1.258293	
subway	1.247889	1.251546	
Note: These coefficients are for train data and models.			

Variables	VIF Test Model 3
latitude difference	1.095989
followers	1.021392
community average price	1.150091
house age	1.137648
square	1.036421
floor	1.735263
elevator	1.928466
Note: These coefficients are for test data and model	

Table A1 Notes: *This table compares the VIFs of multiple predictors to different models.*

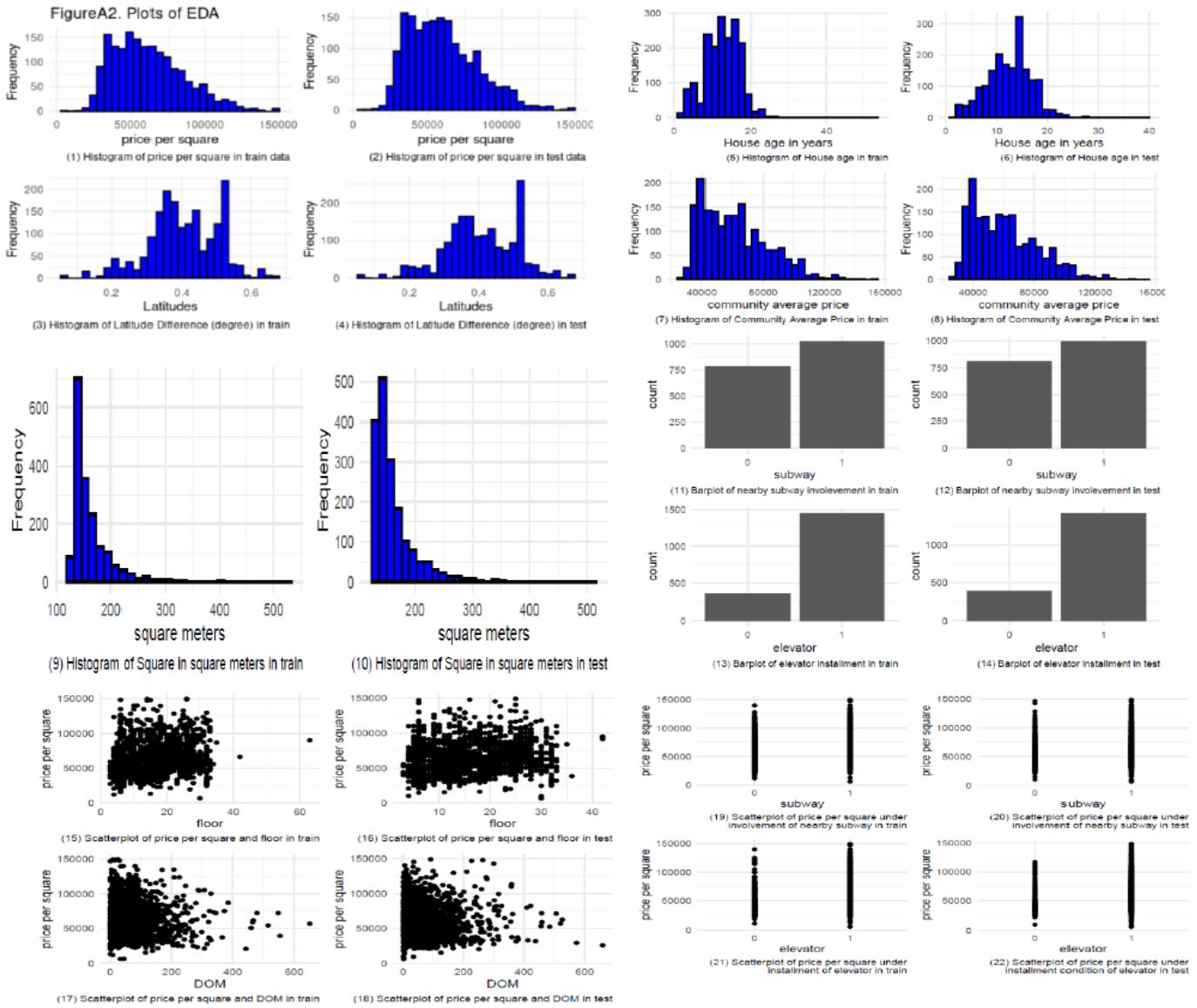


Figure A2 Notes: *These are a combination of EDA plots used to explore the variables' trends and features, including histograms to describe the skewness and center of data, bar plots to describe the proportion of categorical variables, and scatterplots to describe the trends between house price and the variables.*

Figure A3. Multiple Residual Plots of Assumptions and Conditions Verification for Test Data

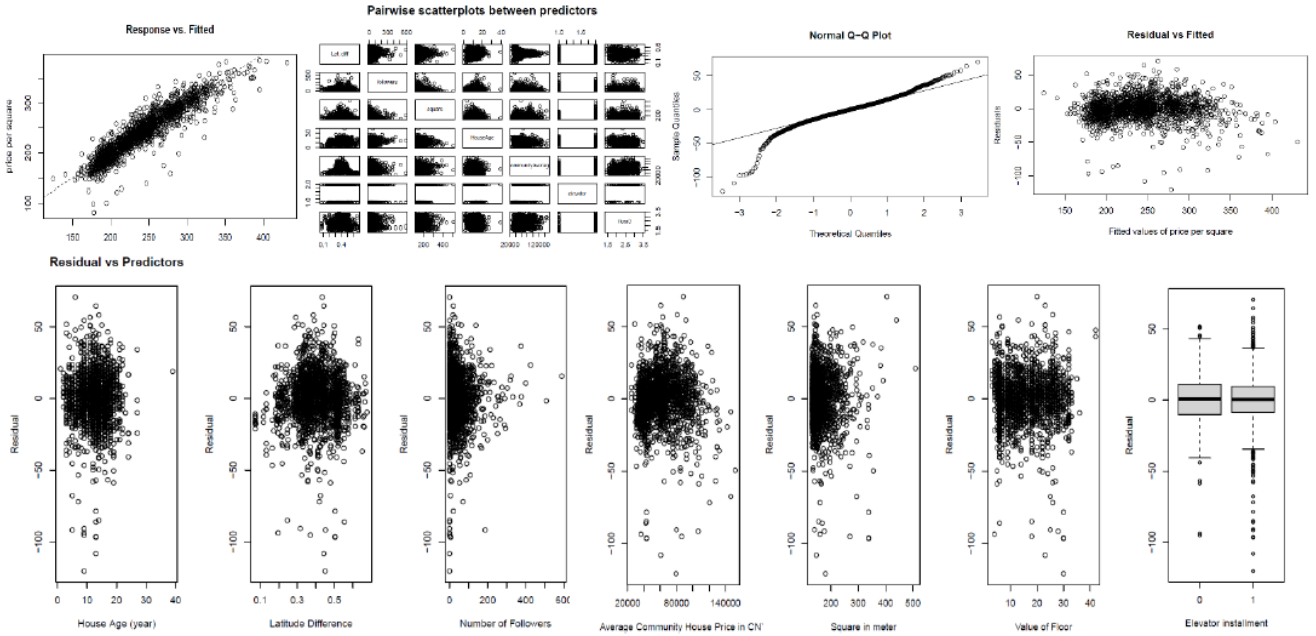


Figure A3 Notes: *These different types of residual plots check if the assumptions and conditions of our final* ***Model 3*** *are satisfied and matched in terms of the testing data.*