Analysis on the Housing Market using Multiple Linear Regression

Yena Joo

December 2020

Analysis on Key Factors that affect the Housing Price and a causal inference of Renovation affecting the house price

Code and data supporting this analysis is available at: https://github.com/yenajoo/304final.git

Abstract

To predict the movement of the real estate market, the Housing Price Prediction dataset from Kaggle - a data science community with various public open data - is used to study and analyze some potential factors affecting the dwelling prices. The dataset includes information about dwellings sold between May 2014 and May 2015(cite), and only useful quantitative variables are selected through a simple data cleaning, as well as eliminating outliers/influential points in order to obtain a smooth analysis. To select the significant variables that affect the dwelling prices, AIC stepwise selection method[1] is used, and Multiple linear regression model is used to determine the relationship between each predictor variables and prices, as well as the Propensity score matching to determine the causal relationship between the renovated house and the house price. Through the analysis, we find significant positive relationships between the independent variables - living area(sqft) and condition of the house - and dependent variable house price, but the number of bedrooms and the year the house is built have a negative effect on the house price. The propensity matching score shows that the house renovation causes an increase in housing prices.

Keywords

housing price prediction, multiple linear regression, AIC selection method, VIF, model assumptions, treatment/control group, causal inference, propensity score matching

Introduction

The world has been overturned by an unexpected virus COVID-19, and it seems that the real estate market has been under the spotlight since the Covid-19 pandemic. People started to look for houses rather than apartments and condos. The spread of the virus caused the housing price to fluctuate by almost 40%. The real estate market is one of the most demanding markets for people since it is one of the most important components of people living. To have a better understanding of the housing prices, it is important to know what affect house price to increase.

Throughout the report, we are going to analyze some potential factors that affect housing prices to determine which characteristics of the houses have shown the most correlation to the housing price, and analyze

[1]*AIC stepwise selection method is explained in the Model section

the causal link between the houses that have been renovated and that have not been renovated using the propensity score matching between renovated houses and non-renovated houses, one being assumed as a "treatment" group, where they are expected higher price by renovation. The other houses would be in a "control" group, where the house has never been renovated.

To build a multiple linear regression model of housing prices by potential factors, AIC stepwise selection method is used, and the model is used to interpret the regression output to find relationships between the housing prices and potential factors. The housing price data will be used to investigate the relationship between housing prices and potential factors such as the number of bedrooms, bathrooms, sqft, etc. In the Methodology section, I describe the data and the model that is used to analyze the relationship. Results of the Difference in differences are provided in the Results section.

Data

The contents of the dataset contain house sale prices for King County, which includes Seattle, downloaded from a data science community Kaggle. It includes houses sold between May 2014 and May 2015 (cite). It consists of 21613 observations and 21 variables. However, through a simple data cleaning process, only 16227 observations are used, and 8 variables are included in the cleaned dataset. The target population of the analysis includes all sales of houses in the USA. The frame population is houses bought or sold within King County, including Seatle, as well as the sampled population. There is insufficient information given regarding the population of its own dataset.

Some key features of the dataset are that every data is quantitative values, which is perfect for linear regression. Also, all the components that are planned to use in building the regression model are the factors people actually consider when they plan to buy a house. Also, the reason for choosing the 2014-2015 dataset is because it shows the most stable housing prices which are before the pandemic.

Some weaknesses of the data are that this data obtains the real-estate market information in King County, which might have different standards of house sales in Toronto or Canada. The dataset represents the housing price in the USA rather than in Canada. Also, since the COVID-19 pandemic made the real-estate market to fluctuate, 2014-2015 house sales data might not fully represent the prediction of the house price made in the analysis.

These are the first 6 observations in the cleaned dataset:

Table 1: Raw data output

price	bedrooms	bathrooms	sqft_living	floors	condition	yr_built	yr_renovated	treatment_group
221900	3	1.00	1180	1	3	1955	0	Not Renovated
538000	3	2.25	2570	2	3	1951	1991	Renovated
604000	4	3.00	1960	1	5	1965	0	Not Renovated
510000	3	2.00	1680	1	3	1987	0	Not Renovated
257500	3	2.25	1715	2	3	1995	0	Not Renovated
291850	3	1.50	1060	1	3	1963	0	Not Renovated

The dataset contains 9 variables. The included variables are the price of the house, bedrooms, bathrooms, living area(sqft), floors, condition, year built, year renovated, and treatment variable that is newly created. To briefly explain what each variable is, price_house is the price of the house, bedrooms, and bathrooms are the number of bedrooms and bathrooms in the house. sqft_living is the living area of the house in the sqft unit, and floors variable is the number of floors in the house. the condition indicates how good the condition of the house is, on a scale from 0 to 5. yr_built is the year the house was built, and yr_renovated is the year the house was renovated. If the house is never renovated, the value of yr_renovated is 0.

Table 2: Baseline Characteristics

	Not Renovated	Renovated	p	test
n	15929	298		
price (mean (SD))	515634.01 (295080.26)	793176.22 (514652.59)	<0.001	
bedrooms (mean (SD))	3.46 (0.82)	3.61 (0.93)	0.002	
bathrooms (mean (SD))	2.24 (0.68)	2.45 (0.82)	<0.001	
sqft_living (mean (SD))	2156.10 (813.38)	2480.32 (975.24)	<0.001	
floors (mean (SD))	1.55 (0.57)	1.45 (0.52)	0.002	
condition (mean (SD))	3.34 (0.59)	3.13 (0.36)	<0.001	
yr_built (mean (SD))	1984.24 (18.86)	1964.43 (10.83)	<0.001	
yr_renovated (mean (SD))	0.00 (0.00)	2001.99 (10.13)	<0.001	

Treatment_group is a dummy variable that has a value of 0 if the house is never renovated, and 1 if the house was renovated at least one time.

After reducing the number of variables, another thing to consider is to eliminate the outliers. This process could be done by looking at the scatter plot of the data.

Figure 1: Scatterplot of the Raw Data

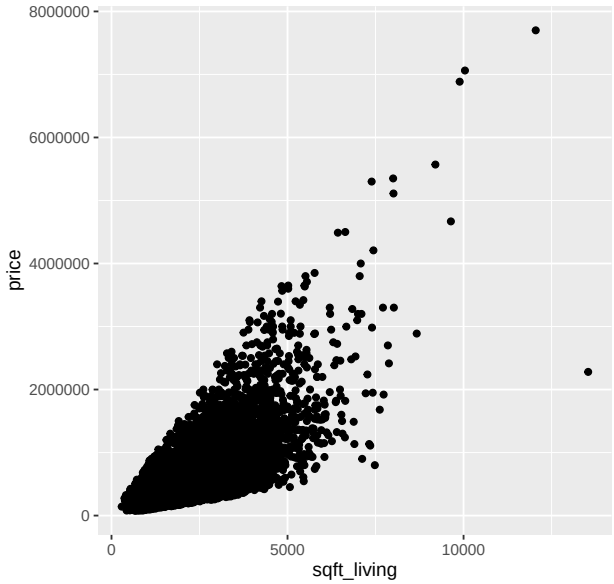
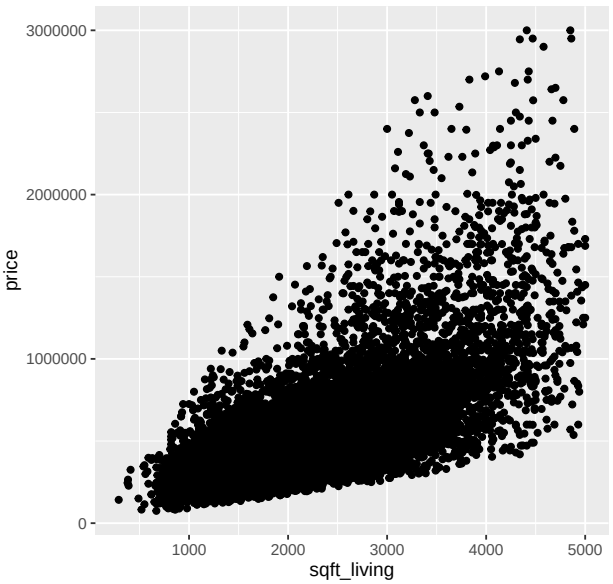


Figure 2: Scatterplot of the Cleaned Data



As shown in Figure 1, The scatterplot looks densely distributed from price range 0 to about 3,000,000. The observations that have price larger than $3,000,000 are omitted from the dataset through the data cleaning. Also, houses with 0 bedrooms are omitted since houses with 0 bathrooms might be a potential estimate error, as well as houses built before 1950, since old houses are more likely to be rebuilt, and we would like to predict the future housing prices.

Also, the dataset is not the best choice for the causal inference since there is no information about the prices of the house before it was renovated. In this case, propensity score matching is a better method to use than the difference-in-differences method.

Model

Multiple Linear Regression model is chosen for the analysis, since it contains a lot of quantitative variables that are suitable for the linear regression. The predictor variables used in the model are number of bedrooms, number of bathrooms, living space in sqft.

Model Selection

There could be various potential factors affecting house prices. Therefore, it is critical to determine which variables should be included in the multiple linear regression. There are various ways to determine, but AIC stepwise selection method is going to be used in the model.

AIC is a short form of Akaike Information criterion, which quantifies the amount of information loss due to the simplification. AIC uses a model's maximum likelihood estimation as a measure of fit. Simply, smaller AIC shows improvement in model performance. Through the process of eliminating and adding the variables, it calculates and compares AIC in each step and determines the model with the lowest AIC which is the best fit for the data.

Formula for AIC is the following:

$$\text{AIC} = -2(log - likelihood) + 2k \quad (1)$$

Where k is the number of predictor variables included in the model, and the Log-likelihood estimate indicates a measure of goodness of fit for any model.

There are 3 methods, forward selection, backward elimination, and the combination of both. Forward selection starts from one variable, and iteratively adds one variable at a time to compare the AIC value until the AIC is the smallest. Backward elimination starts with a full model that includes every variable in the model, and iteratively removes the least contributive predictors until it reaches the lowest AIC value. Combination of both eliminates or adds potential contributive predictor variables, and stops when you have a model where all predictors are statistically significant.

Table 3: AIC model summary

term	estimate	std.error	statistic	p.value
sqft_living	260.58992	3.146406	82.821457	0
bedrooms	-58352.22288	2520.791520	-23.148373	0
condition	39273.67204	2936.306710	13.375194	0
bathrooms	49460.90775	3619.622863	13.664658	0
yr_built	-41.95124	6.475582	-6.478374	0

Using the AIC selection method, the following variables in Table 2 are selected: sqft_living, bedrooms, bathrooms, year built, and condition.

The equation for the regression is:

$$\text{Housing Price} = \hat{B_0} + \hat{B_1} * x_{bedrooms} + \hat{B_2} * x_{bathrooms} + \hat{B_3} * x_{sqftliving} + \hat{B_4} * x_{yrbuilt} + \hat{B_5} * x_{condition} \quad (2)$$

The equation above is going to be interpreted for the multiple linear regression in the Results section.

Model Diagnostics

With the regression model created in the Model section, it is critical to keep in mind that the multiple linear regression analysis is well performed under the following assumptions:

1. Linearity: There should be a linear relationship between the dependent and independent variables.
   By showing the scatter plot above(Figure 2), it satisfies the linearity assumption. It can also be checked by the top-left graph in Figure 3, where the residual plot shows no visible pattern, and the red line is almost horizontal. Hence, the first assumption is satisfied

2. multivariate Normality: residuals of the regression should be normally distributed. This assumption may be checked by looking at a histogram or a Q-Q-Plot on the top right corner of Figure 3.
   Up to 2 theoretical quantiles, the data points perfectly trend the reference line and has a slight high tail at the end of the line. The data points that are off the line might indicate there exists some outliers, it should be into account that the result drawn from the regression model could be misleading or biased. This assumption is going to be dealt in the Weakness section. However, overall, most of the points fall approximately along the reference line, hence the second assumption is also satisfied.

3. No multicollinearity: Independent variables should not be highly correlated with each other.
   This assumption could be checked by using the Variance Inflation Factor(VIF) values. VIF indicates the value of how much the variances in the regression estimates are increased due to multicollinearity. The following table shows the VIFs of the model.

Table 4: VIF models

variables	VIF
bedrooms	1.595009
bathrooms	2.693725
living space(sqft)	2.293528
year built	1.790572
condition	1.245823

variables	VIF
bedrooms	1.514887
living space(sqft)	1.641526
year built	1.363597
condition	1.243540

The left table in Table 3 is the initial model from the AIC selection method. Two variables, bathrooms and living space(sqft) have relatively high VIF which indicates that multicollinearity is a problem for the two variables. Therefore, the variable bathrooms are omitted from the model as seen in the right table, and VIF values do not exceed 2 in the adjusted model. Hence, after eliminating one variable, the third assumption is also satisfied.

4. Homoscedasticity: Variance of error terms should be similar across the values of the independent variables.
   This assumption can be checked through the plot of standardized residuals vs predicted values, having points equally distributed across all values of independent variables(cite). To have the assumption satisfied, there should be no clear pattern in the distribution.

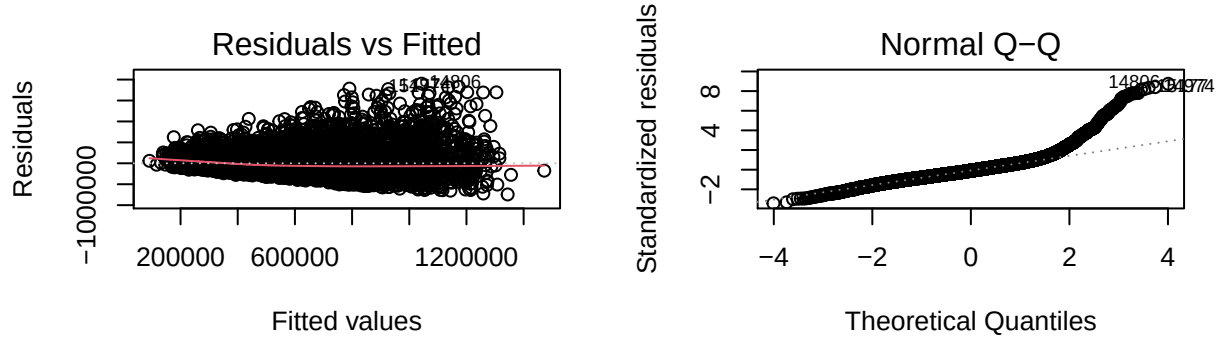
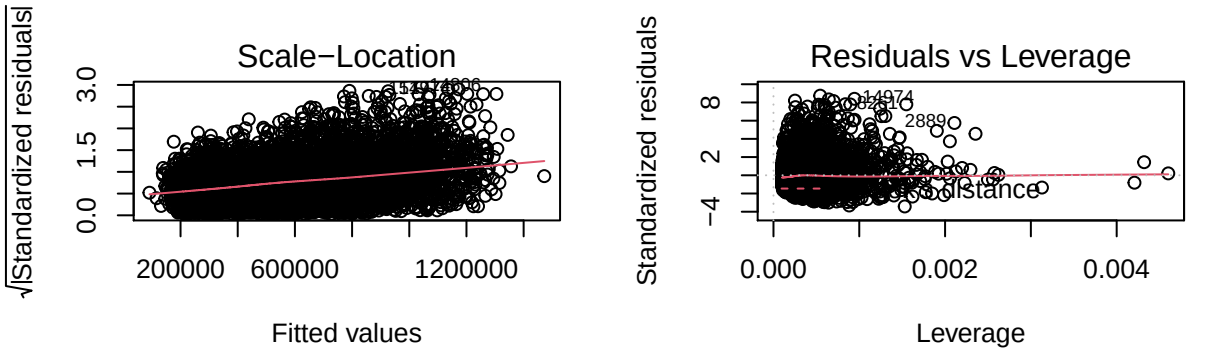


Figure 3: Model Diagnostics



The bottom left graph, which is the Scale-Location graph in Figure 3 determines whether the model satisfies the fourth assumption, by looking at the red line. A horizontal line with well-distributed points is a good indication of homoscedasticity(13). Therefore, the homoscedasticity assumption is also satisfied.

Results

The following is the summary of the multiple linear regression model using the selected predictor variables:

Table 5: Regression summary

term	estimate	std.error	statistic	p.value
(Intercept)	1228613.9495	213475.791438	5.755285	0
bedrooms	-56137.1984	2541.952084	-22.084287	0
sqft_living	290.2568	2.676011	108.466212	0
yr_built	-613.0295	105.363629	-5.818227	0
condition	22445.7958	3266.511068	6.871489	0

The regression equation using the estimates from Table n, the equation is the following:

$$\text{Housing Price} = 1228613.95 - 56137.20 * x_{bedrooms} + 290.26 * x_{sqftliving} - 613.03 * x_{yearbuilt} + 22445.80 * x_{condition} \quad (3)$$

.

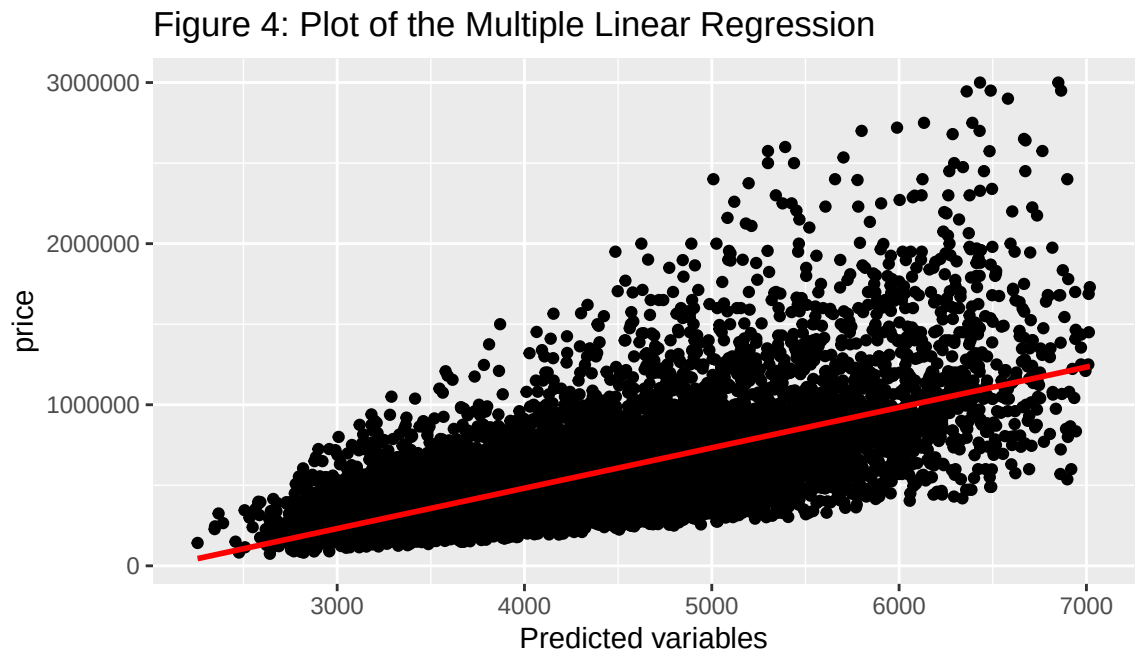
description for x variables are introduced in the previous section? explain it

In Table 4, as observed from the output, the price of the house increases as the number of a living area(in sqft), and condition increase. As 1 sqft increase of the house, the house price is expected to increase about $290, and if the house condition rating is 1 unit higher, the house has a /$22445 higher expected price.

However, an increase in the number of bedrooms, and year built have a negative effect on the housing prices. As the house has one more bedroom, the price of the house would likely to fall approximately $56137. Also, a house that is built 1 year before is expected to have a higher price(approximately $613 more)than the newly built house.

Also, every variable has a very significant value at a 1% significance level, as seen in Table 3, p-value column. The p-values are rounded to 0 which indicates every p-value is significantly small. Therefore, the regression model suggests that every factor has a significant linear relationship to the housing price.

Now, here is the visual image of how the model looks like:



As seen in Figure 4, The combined predictor variables have positive effect to the housing price.

Causal inference

if yr_renovated == 0, treatment = 0 The propensity score is the probability of treatment assignment conditional on observed baseline characteristics that is derived from the fitted regression model.(cite,15) The probability is constructed based on the values of independent variables before the treatment.

Propensity score matching involves assigning some probability to each observation. We construct that probability based on the observation's values for the independent variables, at their values before the treatment. That probability is our best guess at the probability of the observation being treated, regardless of whether it was treated or not. For instance, if 18-year-old males were treated but 19-year-old males were not, then as there is not much difference between 18-year-old males and 19-year-old males our assigned probability would be fairly similar. We can then compare the outcomes of observations with similar propensity scores.

- Treatment causes the outcomes.

-the renovated year variable in the propensity score regression has a small p-value, indicating that it is a significant predictor variable.

One advantage of propensity score matching is that is allows us to easily consider many independent variables at once, and it can be constructed using logistic regression. The same variables used in multiple linear regression model are used to construct the multiple logistic regression to find out the effect of the predictor variables on renovation.

- for propensity score matching you need it to be binary (in this course) because we only covered binary outcome regression (logistic) and if your treatment is continuous you wont be able to "match".
- one-to-one or pair matching is done based on the characteristics covered in the logistic regression model, in which pairs of treated and untreated subjects are formed, such that matched subjects have similar values of the propensity score. -> Based on the characteristics covered in the logistic regression model in which pairs of treated and untreated subjects are created, one-to-one or pair matching is done such that matched subjects have similar propensity score values. After matching the data, the treatment effect can be estimated by comparing the treated and control group. Here, treated subject is the house that is renovated, and the control group is the house that is not renovated.

If the outcome is continuous (e.g., a depression scale), the effect of treatment can be estimated as the difference between the mean outcome for treated subjects and the mean outcome for untreated subjects in the matched sample. Once the effect of treatment has been estimated in the propensity score matched sample, the variance of the estimated treatment effect and its statistical significance can be estimated.

The dataset is reduced to 596 observations after matching the variables. Using this new dataset, a propensity score regression is created in the following table:

price	bedrooms	bathrooms	sqft_living	floors	condition	yr_built	yr_renovated	.fitted	cnts
439990	4	2.50	1540	2	3	2014	0	0.0001756	1
268000	4	3.00	1840	2	5	1988	2013	0.0001803	1
295500	3	2.50	1410	2	3	2014	0	0.0001851	1
649950	3	2.50	1500	2	3	2014	0	0.0002060	1
394950	2	1.00	1131	3	3	2011	0	0.0002201	1
239950	3	1.75	1600	1	3	2014	0	0.0002321	1

After .... , the following table is the 'effect' of being treated on average spend in the 'usual' way(rohan, cite).

	(1)
(Intercept)	560325.532 **
	(216678.427)
bedrooms	-54993.961 ***
	(2525.826)
sqft_living	286.160 ***
	(2.672)
yr_built	-286.146 **
	(106.904)
condition	28778.107 ***
	(3271.794)
treatment_group	193578.540 ***
	(12938.493)
N	16227
R2	0.491
logLik	-222345.766
AIC	444705.532

*** p < 0.001; ** p < 0.01; * p < 0.05.

Discussion

Summary

The goal of the analysis is to find the significant factors that affect the price of houses. Using the 2014-2015 housing price dataset from Kaggle, only quantitative data is used to show the accurate multiple linear regression. Through the data cleaning process in the Data Section, some data that are outliers or influential points are eliminated as well as unrealistically big data. Then, in the Model Section, to determine which variables to include in the regression, the AIC stepwise selection method is used and a linear regression model is built. After building a linear model, model diagnostics is done by showing the four assumptions of the regression - Linearity, Normality, Homogeneity of residuals variance, and Independence of the Error terms - using q-q plot, residual plot, and VIF models in the Model Section.

after the model selection, multiple linear regression and a propensity score matching is used to determine the correlation as well as the causal effect of the significant factors on house prices. Multiple linear regression along with the Propensity score matching shows both correlation and causal relationship between the price of the house and the observed variables.

The predictor variables that contribute to change in housing prices are the number of bedrooms, living area in sqft, year built, and condition of the house as shown in Table 5. Every predictor variable has a significantly small p-value which indicates that it is a significant predictor variable. In the Result section, Figure 6 shows

that there exist a linear relationship between the selected factors and the independent variable housing price. Among the predictor variables, the living area and condition of the house show a positive effect on the housing price, whereas the number of bedrooms and year built show a negative effect on the price.

The result from Propensity score matching, which could be found in Table 6, shows that renovation has a positive causal effect. This means it is more likely for the house to be priced higher in the real estate market when the house is renovated. To connect to real life, there are various TV shows where people renovate their homes and sell their houses with a jump of the house price.

Conclusions

The Multiple Linear Regression analysis shows that houses that are in good condition are more likely to be sold at a higher price, and as the house size is bigger, the price increases. However, even though the number of the bedroom shows a negative coefficient for the regression estimate, by looking at the scatter plot in figure N, it is hard to tell that the number of bedrooms has a negative effect on the housing price.

Weaknesses

Every data analysis contains some weaknesses. There are a few weaknesses the analysis includes. First, the dataset contains housing prices of a limited area which is King County, and the data is from 2014-2015. Since the data does not represent the Canadian housing price or the recent housing price, the analysis might not be the most accurate way to predict the housing prices.

Also, the AIC stepwise selection method does not consider interaction terms, there might exist some relationships between the independent variables. This might also lead to an omitted variable bias, where the omitted variables should be correlated with the dependent variable, and correlated with the explanatory variables included in the model. There might be an important variable that would affect the model, but it is hard to figure out since the variable might be missing in our data set, or might be impossible to measure. Since the location of the house is also important, missing the location variable might have affected the model.

Another weakness is that the multiple linear regression model does not fully satisfy the model assumption introduced in the Model section. Multivariate Normality, which shows the normal distribution of residuals of the regression is not satisfied. By looking at the QQ plot, it is noticeable that the data points do not trend the theoretical line, and the points at the upper tail of the data seem to jump and have higher values than the theoretical line, telling that the data might contain a gap in the values. This suggests that our model does not satisfy the normality assumption on the error terms. Therefore, we need to take into account that the result drawn from the regression model could be misleading or biased.

Next Steps

For the next steps, it would be a great idea to compare the housing prices after COVID-19 and do a causal inference pre & post analysis, to determine whether the COVID-19 affects the housing prices. In this analysis, the treatment variable would be houses that got affected by the COVID-19. Using the dataset that contains prices of condos, town houses, and detached houses would lead to an interesting result.

Also, as addressed in the Weakness section, the location variable could be added with housing prices data of Toronto, or Canada, and analyze the regression model including the location. It would make a big difference when the location variable is added, since location affects the housing prices significantly as we can see in our real lives(urban/suburban, Toronto/other minor cities, etc).

It also would be a good idea to do a survey on what people consider when buying a house. Which factor comes in first? We could compare the data of the survey and the regression model performed in the analysis, and determine whether the regression model actually fits into the real life housing price prediction.

Reference

1. Propensity Score methods: Austin, Peter C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," May 2011. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.

2. Kableone function: "Bsurial/Bernr Source: R/kableone_nonnorm.R." Accessed December 22, 2020. https://rdrr.io/github/bsurial/bernr/src/R/kableone_nonnorm.R.

3. Side by Side plot: "Side-by-Side Plots with ggplot2," September 1, 1958. https://stackoverflow.com/questions/1249548/side-by-side-plots-with-ggplot2.

4. Regression Model assumption: Editor, Minitab Blog. "How to Identify the Most Important Predictor Variables in Regression Models." Accessed December 22, 2020. https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models.

5. Normal QQ plot: "How to Interpret a QQ Plot," July 1, 1963. https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot.

6. Model Diagnostics: Kassambara. "Linear Regression Assumptions and Diagnostics in R: Essentials," March 11, 2018. http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.

7. Exponential notation: Matt BannertMatt. "Force R Not to Use Exponential Notation (E.g. e+10)?," March 1, 1961. https://stackoverflow.com/questions/9397664/force-r-not-to-use-exponential-notation-e-g-e10.

8. AIC model selection: "AIC or p-Value: Which One to Choose for Model Selection?," May 1, 1960. https://stats.stackexchange.com/questions/9171/aic-or-p-value-which-one-to-choose-for-model-selection.

9. Baseline Characteristics Table: Rich, Benjamin. "Baseline Characteristics Table," November 25, 2020. https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html.

10. House price prediction: Shree. "House Price Prediction," August 26, 2018. https://www.kaggle.com/shree1992/housedata.

11. Information about real estate: Steve Huebl. April 27, 2020. "COVID-19's Impact on Real Estate: Toronto Home Sales Down 69%," April 27, 2020. https://www.canadianmortgagetrends.com/2020/04/covid-19s-impact-on-real-estate-toronto-home-sales-down-69/.

12. ggplot: Susanejohnston. "A Quick and Easy Function to Plot Lm() Results with ggplot2 in R," April 23, 2015. https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-lm-results-in-r/.

13. Tableone R package: "Tableone r Package." Accessed December 22, 2020. http://rstudio-pubs-static.s3.amazonaws.com/13321_da314633db924dc78986a850813a50d5.html.

14. AIC selection: Tripathi, Ashutosh. "What Is StepAIC in R?," June 16, 2019. https://medium.com/@ashutosh.optimistic/what-is-stepaic-in-r-a65b71c9eeba.

15. AIC selection 2: "Variable Selection in Multiple Regression." Accessed December 22, 2020. https://www.jmp.com/en_in/statistics-knowledge-portal/what-is-multiple-regression/variable-selection.html.

16. Model diagnostics: Zach. "The Four Assumptions of Linear Regression," January 8, 2020. https://www.statology.org/linear-regression-assumptions/.

17. Model diagnostics: Zhang, Zhongheng. "Variable Selection with Stepwise and Best Subset Approaches," April 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4842399/.

18. Model Diagnostics: "Assumptions of Multiple Linear Regression," March 10, 2020. https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/.