HOUSE PRICE PREDICTION – URBAN INDIA

ECON F342

In this assignment, we estimate a linear model using OLS multiple regression to predict house prices in Urban India. We obtain a dataset of houses listed on popular property-dealing websites from over 256 Indian cities, and utilize this information to estimate the effect of different factors on house prices.

ASSIGNMENT-1

AMAN MAHAJAN	2018B3A70880H
AKSHAT GOYAL	2018B3A70864H
AYUSHMAN UPRETI	2018B3A30916H
KEDAR CHARKHA	2018B3A70912H
NIMISH PRABHUNE	2018B3AA0940H
RITIKA GARG	2018B3A70258H
SIDDARTH BALDWA	2018B3A70914H
VASTAV RATRA	2018B3A70903H
VIKAS SHEORAN	2018B3A70847H

QUESTION 1

Start the assignment with building a multiple regression model giving justification for the selection of the dependent & independent variables.

ANSWER 1[1]

$$y_{ics} = \beta_0 + \beta_1 * owner_{ics} + \beta_2 * builder_{ics} + \beta_3 * under_{construction_{ics}} + \beta_4 * rera_{ics} + \beta_5 * bhk_{no_{ics}} + \beta_6 * resale_{ics} + \beta_7 * tier2_{ics} + \beta_{71} * tier2_{ics} * square_{ft_{ics}} + \beta_8 * tier3_{ics} + \beta_{81} * tier3_{ics} * square_{ft_{ics}} + \beta_9 * coord_{ics} + \beta_{10} * d_{mark_{ics}} + \gamma_1 * avprice_{cs} + \gamma_2 * pm2_{cs} + \gamma_3 * cci_{cs} + \mu_1 * interestPop_s + \mu_2 * loanPop_s + \mu_3 * power_{availability_s} + \mu_4 * nhlength_s + \mu_5 * fixedcap_s + \mu_6 * inflation_s + \mu_7 * construction_{worker_{wage_s}} + \mu_8 * num_{factories_s} + \mu_9 * nsdp_s + \epsilon_{ics}$$

The price of a house depends on several factors. Some of these factors are directly related to the house itself – its size, its structure, age, and level of construction. In India, astrological factors like north-south facing and difference in depth of the front and back face of the house also influence the cost of the house. However, the price of the house is also determined by the neighbourhood, city and state in which it is located. This constitutes an implicit market – when we purchase a house, we are paying for the house and all the locational advantages that come with it. In accordance with this view, we estimate a linear model using OLS regression to predict house prices for Indian cities. Here, $\mathbf{y_{ics}}$, the dependent variable is the log of price of the house 'i' located in city 'c' and state 's'. We use a set of independent variables to control for various observables at house, city and state level. We explain these variables and the reasoning below:

	Variable Chosen	Description	Justification	Relevant Literature
1	**lprice2**	**Dependent (Study) Variable:** Natural Logarithm of House Prices	We wish to study the factors affecting the house prices in India, by estimating a linear regression model with some selected independent variables	
2	**Builder, Dealer**	These are dummy variables (base: owner)	The prices posted by the property dealer might	Literature on commission and rent-

[1] Data Cleaning:

We provide a brief review of data cleaning process. When we obtained the dataset from Kaggle, there was a major problem with the data. The column with latitude values was labelled as longitude and vice versa. So, as Step 1, we interchanged these names. Then, we obtained the latitudinal and longitudinal extent of India from the internet, and dropped all values that lied outside these boundaries. Third, we removed properties whose cost was greater than Rs.50 crore. Fourth, we removed properties whose acreage was higher than 99% percentile of the sample. Fifth, we removed almost 1570 properties whose city was not specified, and only state was mentioned. Sixth, we also removed a few properties that were located in cities whose names could not be matched to any Indian city in our list. As a result, we started with 29451 observations, but were left with 25865. In addition, another 878 observation were dropped in regression due to missing variables. The final regression had 24987 houses.

		which tell us who has posted the house on sale.	be inflated to account for his commissions.	seeking by middle men from various industries like insurance (Sudhak, 2009; Anagol et al., 2017)
3	**under_construction**	This is a dummy variable which tells us whether the house is fully constructed or if it is still under construction	The price of any house will depend on which stage of construction it is currently in. We expect the price of fully constructed houses to be higher than when it is still under construction	
4	**rera**	This is a dummy categorical variable which indicates whether the house falls under the areas specified in RERA Act	We expect the houses under RERA to be of higher value as the individuals can litigate against the builders who do not provide their houses on time. This ensures greater accountability of the builder.	Based on RERA literature (Pawar & Ahire, 2018; Gandhi et al., 2019 – now published JUE)
5	**bhk_no**	This variable indicates the number of bedrooms (BHK) in the house	Since the price of the house can be strongly expected to depend on the size of the house (measured by the BHK of the house), we expect a significant positive correlation between price and bhk_no	
6	**square_ft**	This is an indicator of the house's size based on total floor area (in square feet)	The house area in terms of square feet is a major determinant of the house size. We expect a significant positive coefficient when the regression is run	
7	**ready_to_move**	This is a dummy variable which is an indicator of whether the house is ready for the new owner to move in immediately after sale	Houses which are ready to move often demand more prices, ceteris paribus - since such a house would be fully constructed, may be partially furnished, and the society area (with amenities and other occupying residents) could be expected to be developed	
8	**resale**	This is a dummy variable which tells us whether the house is being resold or if it is being sold for the first time	A resold house is expected to fetch a lower price due to its depreciated value	
9	**tier2, tier3**	These are dummy variables (base: tier1) which indicates the tier	Houses in tier2 and tier3 cities are expected to be of lower price, due to	Property prices vary drastically between cities. Relevant papers:

		of the city in which the house is situated	lower cost of living, lower population density, and higher land availability	Tandel et al., (2019) provides summary of tier classification for India
10	**sqft_2**	This is an interaction term of the house size (in square feet) in the second tier cities (square_ft * tier2)	The per-sqft price of property would differ with the tier of the city in which it is located	
11	**sqft_3**	This is an interaction term of the house size (in square feet) in the third tier cities (square_ft * tier3)	The per-sqft price of property would differ with the tier of the city in which it is located	
12	**d_mark**	This is the indicator of distance of the house from the central market (major mall) of the city (in nautical miles). It has been calculated by the Latitude and Longitude values of the house and the mall, using the haversine formula	Proxy for distance of the house from the commercial centre/hub of the city. Locational factors can affect property prices.	Xu et al. (2016) explores impact of commercial and transit access on property prices. For broader set of amenities, see Tyravainen & Miettinen (2000).
13	**Coord**	This variable = x+y+x2+y2+xy+x2y+xy2 +x3+y3 Is a function of the latitude and longitude of a house.	The function is commonly used in literature to control for location-fixed that might determine house prices.	Variable borrowed fromanalysis in Gulzar & Pasquale (2016)
	CITY LEVEL FEATURES			
14	**avprice**	This is the average price of all houses listed in the city	This reflects the overall state of housing market in the city.	
15	**cci**	Construction Cost Index – Source is Construction Industry development Council. Wherever missing, we have put the average value of all cities in the state.	This reflects the costs of construction in the city.	Relevant literature: Wu et al. (2014), Sunde & Muzindutsi (2017)
16	**pm2**	This measures the ambient PM 2.5 concentration in the city. Source: NAMP and CPCB data.	Cities with high concentration of pollutants, ceteris paribus, will have lower property values due to the discomfort caused by pollution. Similar trends observed in China and USA.	Relevant literature: Chay & Greenstone (2005)
STATE LEVEL FACTORS				
17	**InterestPop**	This is an indicator of the state-wise interest payments (in rupees crores) per resident of the state	We expect that as the interest rate payment requirements increase, the loan seeking capacity and tendency would decrease. Hence demand for houses would reduce, thereby	

			reducing the price of the houses	
18	**loanPop**	This is an indicator of the amount of personal loans (in rupees crores) given by the state's commercial banks, per resident of the state	This is a measure of banks willingness to give personal loans. If larger loans are given per capita, then that would increase the liquidity (and hence marginal propensity to spend); thereby driving up house prices	
19	**Power_availability**	This variable indicates the per capita availability of power (kilowatt-hour) in the state in which the house is situated	Reflects the level of development of the state.	
20	**nh_length**	This variable accounts for the total length of National Highways (in km) in the state in which the house is situated	Reflects the level of development of the state.	Xu et al. (2016), other papers discussing transit access include: Duncan (2011), Zhang (2016)
21	**fixed_capital**	This is the state-wise fixed capital amounts (in rupees lakhs) in the state in which the house is situated	Reflects the level of development of the state.	
22	**inflation**	This is the state-wise average urban housing inflation (CPI %)	Reflects the macroeconomic conditions of the housing market	
23	**construction_worker_wage**	This is the state-wise average daily wage rates of male construction workers in rural regions (in rupees)	Cost of Construction	
24	**num_factories**	This is an indicator of the total number of registered factories in the state in which the house is situated in	Reflects the level of development of the state.	
25	**nsdp2**	This is the per capita state domestic product (in rupees) for the state in which the house is located in	Proxy for demand-side conditions in the housing market.	Coletta et al. (2019)

QUESTION 2

Calculate summary statistics of the variables & comment on the same.

ANSWER 2

Variable	Min	Median	Mean	Max	SD
Square_ft	3	1180.4	1309.2	9959.8	699.95
av_price	5.3	134.2	132.53	650	79.59
cci	141	144.6	144.2	147.9	1.92
pm 2.5	6.5	47.4	63.55	145.2	30.54
power_availability	256.7	1172.7	1177.9	2936.1	526.51
nh_length	293	6991	8052	16239	4387.8
fixed_capital	45161	13845026	21120233	67266206	14574724.6
inflation	2.3	7.6	7.791	12.9	2.4
construction_worker_wage	220.5	305	325.3	836.2	8451.10
num_factories	607	15830	16745	37787	3095224.32
d_mark	0	3.606	6.307	94.293	9.9
nsdp2	0.2223	1.3121	3.7219	252.697	20.95
interest_pop	0.1034	0.3019	0.3004	0.9404	0.133
loan_pop	0.3007	1.9838	1.9135	5.8944	1.13
price	0.25	59	87.66	1000	98.29

Tier	
Category	**Count**
1	10866
2	10069
3	4930

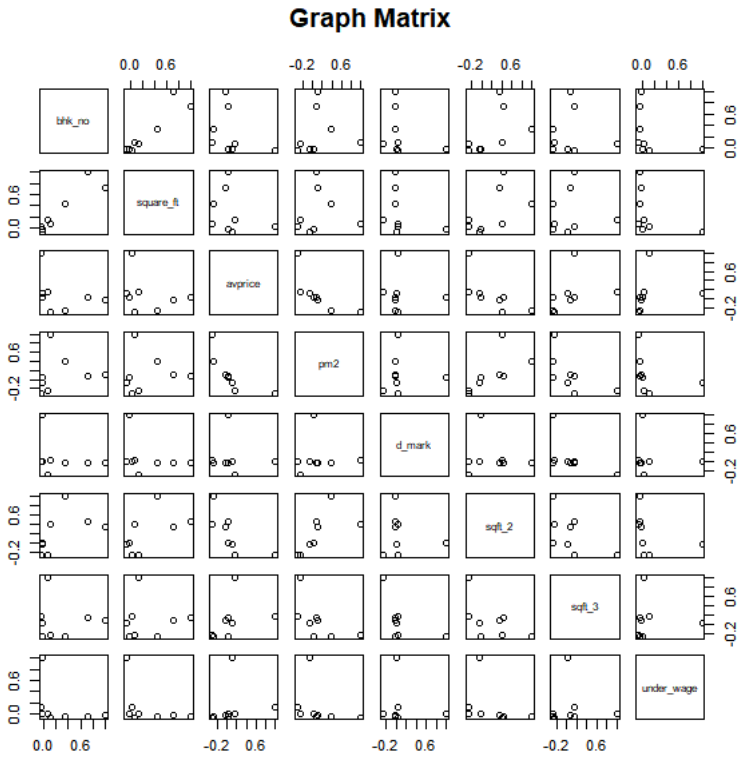
Comments:

- Number of observations: 25866
- Number of Variables: 25
- Most variables show positive skewness with *construction_worker_wage, d_mark* and *nsdp2* showing the highest skewness and *cci* showing the lowest
- High skewness of the variable *construction_worker_wage* is because workers in Kerala are paid much more than those in other states; thus, behaving as outlier
- The data contains a significantly large number of observations which are categorized as ready to move in (21420) when compared to those that are under construction or not yet ready to move in (4445).
- Similarly, observations pertaining to second hand homes are much more (24312) than that of newly constructed homes (1553)
- High variation in the variable bhk_no is observed with the values ranging from 2 to 20.

QUESTION 3

Prepare Graph Matrix for your model and discuss the results.

ANSWER 3



Apart from the obvious strong correlations like between bhk_no and square_ft (more the no. of bedrooms, more will be the size of the house in square ft.), we do not observe very strong correlations between any 2 variables in terms of magnitudes. Almost all variables have a positive correlation with square_ft except for d_mark and under_wage.

QUESTION 4

Estimate the model and interpret the coefficients of your variable appropriately.

ANSWER 4

Sr No.	Variable	Coefficient	Interpretation
1	builder	-0.138***	On average, the prices of houses posted by builders are 13.8% lower than those posted by owner.
2	dealer	-0.14 ***	On average, the prices of houses posted by dealers are 14% higher than those posted by owners.
3	under_construction	0.0035	*Not significant*
4	rera	0.03 ***	3% increase in price if the house falls under RERA act relative to which does not
5	bhk_no	0.278***	There is a 28% increase in the price of the house for every additional bedroom.
6	square_ft	0.0005 ***	0.05% increase in price if the area of the house increases by 1 square ft. An average house is roughly 1300 sqft.
7	tier2	0.036**	The average price of a house in tier 2 city is 3.6% higher than Tier 1 city. We attribute this result to (a) larger size of houses, and (b) bias due to over-representation of houses from Tier 1 cities.
8	sqft_2 (interaction term between tier 2 dummy and sqft)	-0.0001207 ***	Relative to tier 1 city, the per-sqft price of house in Tier 2 city is 0.01% lower. Roughly the size of 1000 sqft,

9	tier3	-0.008	*Insignificant*
10	sqft_3	-0.0001166 ***	Relative to Tier 1 city, the per-sqft price of the house is 0.011% lower in Tier 3 cities.
11	d_mark	-0.0031 ***	A 1 kilometer increase in distance from the commercial centre of the city decreases the cost of the house by 0.15% (the distance was computed in nautical miles)
12	coord	-0.000001136 ***	Location control variable, no meaning.
CITY AND STATE LEVEL VARIABLES – Since many of these variables are proxy for demand side conditions and level of development, the sign of the coefficient matters but the magnitude does not.			
13	avprice	0.0037 ***	Positive sign indicates that the price of the house increases with the average price of all houses in the city.
14	cci	0.01723 ***	Cities where construction is costlier have more expensive houses.
15	pm2	-0.002558 ***	All else constant, increase in pollution in a city leads to lower housing values.
16	interestPop	0.6799 ***	Housing values appreciate if the per-capita interest payments (proxy for the functioning of credit markets) is higher.
17	loanPop	-0.063***	As the number of loans issued per capita increases, the house price falls. The sign if the variable is counter-intuitive.
18	power_availability	-0.0004791 ***	The sign of the variable is counter-intuitive.

19	nh_length	0.000025 ***	Increase in the highway length (proxy for level of development of the state) increases housing prices.
20	fixed_capital	-0	*Insignificant.*
21	inflation	-0.0097 *	The sign is counter-intuitive
22	construction_worker_wage	-0.0004 ***	The sign is counter-intuitive. We believe that it is because of the inclusion of CCI in the regression. It absorbs the cost component of construction. The negative sign might reflect labour market conditions.
23	num_factories	0.000005831 ***	Positive Sign. Proxy for level of development of state.
24	nsdp2	0.04613 ***	Positive sign. States with greater per-capita NSDP have greater demand for houses and higher prices.

***** indicates significant at 99.9%, ** 99%, * 95%**

R Squared = 69.47%

QUESTION 5

Calculate all the model selection criteria (AIC, BIC, R2 & Adj, R2) for your model. Also, plot the predicted Y and comment on the accuracy of your model.

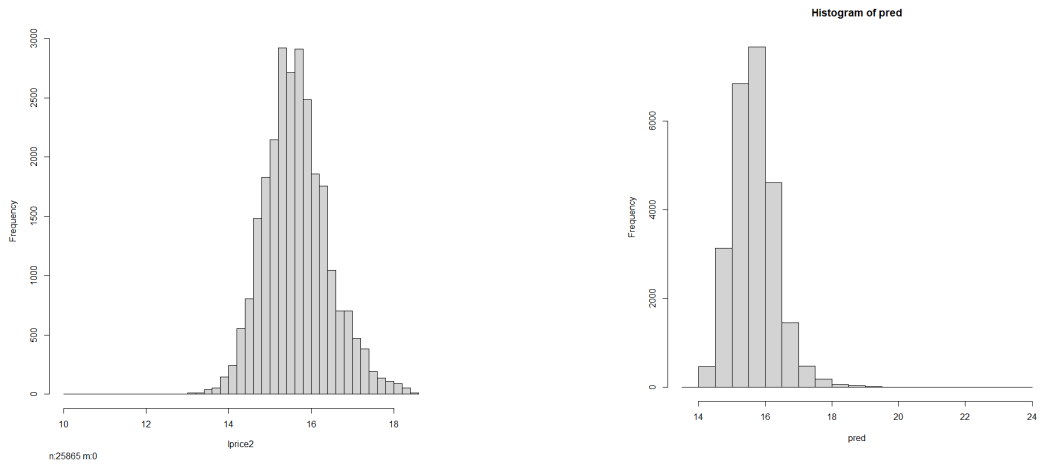
ANSWER 5

Models	df	AIC	BIC	R-squared	Adjusted R-squared
Complete Model	27	29543.7766	29763.18157	0.6947	0.6944
Without State Variables	16	32847.6135	32978.18387	0.6635	0.6634
Without City Variables	21	38572.4486	38743.09693	0.5616	0.5613
Without State and City Variables	10	46082.3393	46163.9458	0.4385	0.4385

The smaller the values of AIC and BIC, the closer is the model to reality. In our case, the complete model provides the smallest values. If we remove city-specific or state-specific variables, the AIC and BIC shoot up. The R-square of the regression falls considerably, especially if we drop the city-specific variables. On dropping both, the R-square falls by almost 30 percentage points. Hence, our hypothesis that the city and state level features are crucial determinants of housing prices is true.

Hence, we believe that our models are accurate.

We plot the predicted Y (log prices) below:



The left plot shows the actual distribution of log(price) in the dataset, and the right plot shows the predicted log(prices). As can be seen, the two distributions are very similar

QUESTION 6

For regression model fit in Question 1, use plots (wherever applicable) and formal tests for checking following OLS assumptions. Discuss the steps involved in the process with proper interpretation of the results

a. Heteroscedasticity
b. Multicollinearity
c. Normality of the error term
d. Omitted-Variable Bias

ANSWER 6

A. Test for Heteroscedasticity:

- *Definition: The variance of residuals changes with fitted values of the independent variable.*
- The presence of Heteroscedasticity is a violation of the assumptions for the Multiple Linear Regression Model
- The process of detecting and amending the problem of Heteroscedasticity is called Residual Analysis

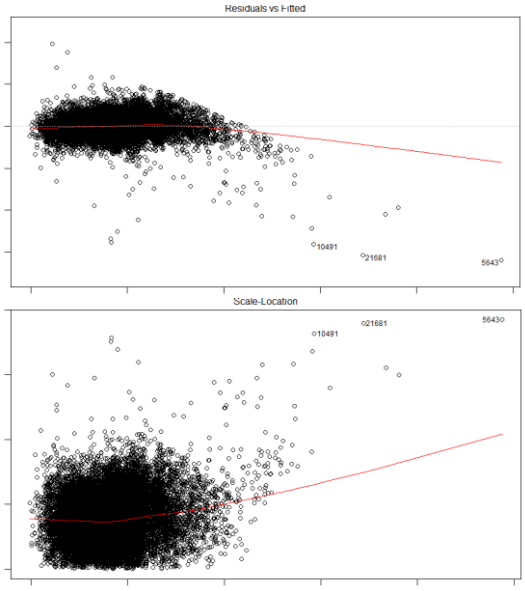
<u>How to Detect It?</u>
There are various methods to detect heteroscedasticity in a model. In this report, we make use of three methods.

1. Graphical Approach
2. Breush-Pagan Statistical Test
3. F Test for Heteroscedasticty

1. Graphical Approach

Here, we make use of two plots for analysis purposes.

(i) Residual vs Fitted Values
(ii) Standardized residuals vs Fitted Values



- The top graph shows residuals on the Y axis and fitted values on the X-axis. The only difference in the bottom plot is that Standardized residuals are plotted on the Y-axis.
- Now if there is no heteroscedasticity in the model, then the distribution of residuals should be completely random throughout the range of X-axis and we should get a straight red line.
- However, here in both plots the red line is curved and the variance of residuals seems to change with increasing fitted values. So, the inference here is that, heteroscedasticity does exist in the model

(B) Breush-Pagan Statistical Test

```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

              Data
-----------------------------------
Response : lprice2
Variables: fitted values of lprice2

       Test Summary
----------------------------
DF            =    1
Chi2          =    5981.6882
Prob > Chi2   =    0.0000
```

- ***Decision Rule :*** *If p-value < 0.05, reject the null hypothesis; otherwise not*
- Here, since p-value is coming out to be 0, we reject the null hypothesis and retain the alternative hypothesis that variance is not constant. Thus, heteroscedasticity exists in the model

(3) F Test for Heteroscedasticity

```
F Test for Heteroskedasticity
-----------------------------
Ho: Variance is homogenous
Ha: Variance is not homogenous

Variables: fitted values of lprice2

       Test Summary
----------------------------
Num DF      =    1
Den DF      =    24985
F           =    836.5434
Prob > F    =    5.875748e-181
```

- ***Decision Rule :*** *If p-value < 0.05, reject the null hypothesis; otherwise not*
- Here also, since p-value is coming out to be almost 0, we reject the null hypothesis and retain the alternative hypothesis that variance is not homogenous. Thus, heteroscedasticity exists in the model

B. NORMALITY OF RESIDUALS

- Normality of errors is one of the assumptions of the Multiple Linear Regression model.
- *Definition : The population error is independent of independent variables and normally distributed with zero mean and variance sigma squared*

How to Detect It?

- In this report we make use of two methods to detect normality of errors.
  - Graphical Approach
  - Shapiro Test for Normality

1. **Graphical Approach**
2. **Shapiro Test for Normality**

```
shapiro.test(reg1$residuals[0:5000])
```

- The shapiro.test command in R has a maximum limit of 5000 observations as it's input argument.
- Here we have considered the first 5000 observations to check if the residuals follow a normal distribution.
- Even though we are considering only 5000 observations out of 25866, they are enough for the normality test. The reason being, according to the Central Limit Theorem if the sample size is large enough then the distribution will approximate a normal distribution and 5000 observation are enough for estimating the normal distribution

```
        Shapiro-Wilk normality test

data:  reg1$residuals[0:5000]
W = 0.9329, p-value < 2.2e-16
```

- *Ho : Residuals are normally distributed;*
  Ha : Residuals are not normally distributed

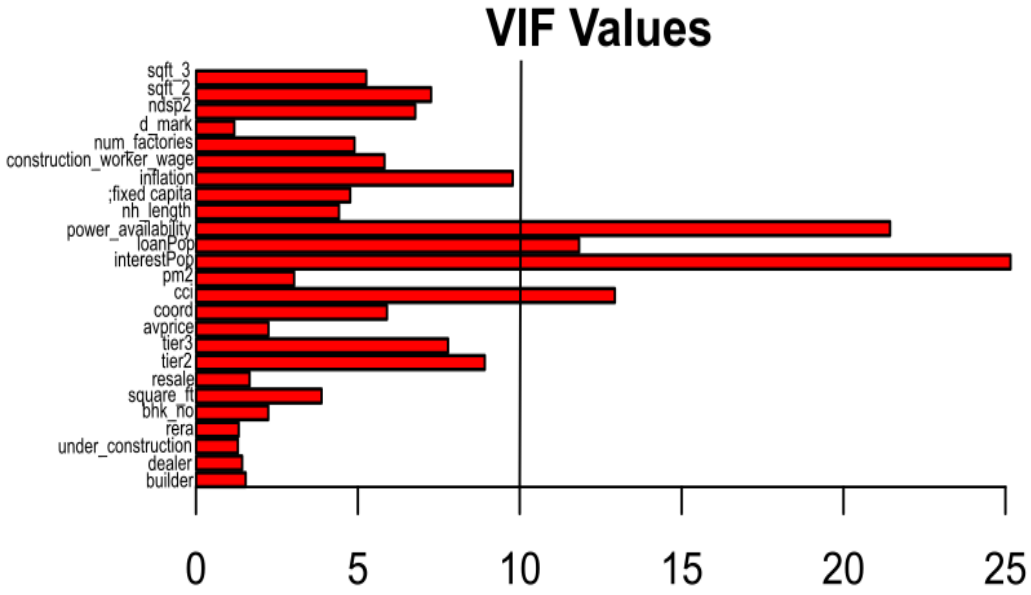
- *Decision Rule : If p-value < 0.05, then reject the null hypothesis; otherwise not*
- Here, since p-value is less than 0.05, we reject the null hypothesis that residuals are normally distributed.

C. MULTICOLLINEARITY

- Multicollinearity is part of an important assumption in the Multiple Linear Regression model.
- Multicollinearity occurs when the independent variables are correlated with each other in the model.
- In OLS estimation of coefficients, there should be no perfect correlation between two independent variables, partial correlations are allowed

How to Detect it?

- VIF (Variance Inflation Factor) is an index used to identify the correlation between independent variables.
- If VIF>10, then we infer that multicollinearity exists in the model.



The mean VIF of the model comes out at 7.90625, which is less than 10. Other than that,

- Variables such as cci, interestPop, loanPop, power_availability, under_wage have high values of VIF, showing that variables are highly collinear with other independent variables.
- Dropping these variables may solve the multicollinearity problem further, but doing so will reduce the predicting power of the model.
- Apart from that, since the mean VIF is less than 10, we can avoid dropping these variables.

D. OMITTED VARIABLE BIAS

- Omitted variable bias occurs when we ignore a variable which actually belongs to the true population model.
- Omitting relevant and correlated variables (with independent variables) from the model leads to misspecification of the model.

How to Detect it?

- To detect if there is any omitted variable bias in the model, we use the Ramsey Reset Test

- Ho : The model does not have Omitted variable bias
  Ha : The model does have Omitted variable bias

- ***Decision rule: If p-value > 0.05, we fail to reject null hypothesis.***

```
> library(lmtest)
> resettest(reg1, power=2,  type='fitted')

        RESET test

data:  reg1
RESET = 1912.1, df1 = 1, df2 = 24959, p-value < 2.2e-16
```

Here since p-value is close to 0, we cannot reject the null hypothesis and thus infer that Omitted variable bias does exist in the model.

QUESTION 7

Based on results of Question 6, use the remedies to address the issues identified and alter your model suitably. Please discuss each step involved in the process. No marks will be given without proper interpretation and discussion of the steps.

ANSWER 7

(A). Rectification of Heteroscedasticity

Box-Cox Transformation

- Box Cox transformation is a parametric power transformation technique used to alleviate heteroscedasticity from the multiple linear regression model
- The parameter 'lambda' is at the core of the Box-Cox transformation and its range is from -5 to +5
- Among all the values between -5 to +5, the most optimal lambda value is chosen which best approximates the distribution of the dependent variable to the normal distribution.
- The transformation has the following form if all Y values are positive.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

- In our model, the R code for the Box-Cox transformation yields estimated lambda value of -0.8.
- After transforming the Y variable according to the formula in the above figure, we rerun the regression model.
- After re-running the regression model, we apply the Breush Pagan test again. Even in this case, the p-value comes out to be close to 0. Thus we can infer that even after the Box-Cox transformation, heteroscedasticity still exists in the model.

```
Box-Cox Transformation

25865 data points used to estimate Lambda

Input data summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.13   15.10   15.59   15.64   16.10   18.42

Largest/Smallest: 1.82
Sample Skewness: 0.371

Estimated Lambda: -0.8
```

```
        studentized Breusch-Pagan test

data:  reg_BC
BP = 713.46, df = 25, p-value < 2.2e-16
```

Estimating standard errors using the coeftest() function:

- Since we know that our model suffers from heteroscedasticity, we want to obtain heteroscedasticity robust standard errors and their corresponding t values.
- In R the function coeftest from the lmtest package can be used in combination with the function vcovHC from the sandwich package to do this.

```
t test of coefficients:

                           Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)              1.2466e+01  7.4961e-01  16.6303 < 2.2e-16 ***
dealer                  -2.4483e-01  2.7349e-02  -8.9521 < 2.2e-16 ***
under_construction       4.2389e-02  3.9473e-02   1.0739 0.2828869
rera                     4.3207e-02  6.9608e-03   6.2073 5.477e-10 ***
bhk_no                   2.7992e-01  2.2564e-02  12.4055 < 2.2e-16 ***
square_ft                5.3247e-04  2.6960e-05  19.7502 < 2.2e-16 ***
resale                   6.0311e-02  1.7248e-02   3.4968 0.0004717 ***
tier2                    1.5503e-02  3.2076e-02   0.4833 0.6288647
tier3                   -2.3522e-02  5.2462e-02  -0.4484 0.6538989
avprice                  3.9648e-03  6.4731e-05  61.2499 < 2.2e-16 ***
coord                   -1.1356e-06  6.4549e-08 -17.5929 < 2.2e-16 ***
cci                      1.7233e-02  5.4212e-03   3.1789 0.0014802 **
pm2                     -2.5578e-03  1.7423e-04 -14.6809 < 2.2e-16 ***
interestPop              8.2248e-01  1.1670e-01   7.0477 1.866e-12 ***
loanPop                 -8.4548e-02  9.5817e-03  -8.8239 < 2.2e-16 ***
power_availability      -4.7908e-04  2.7429e-05 -17.4660 < 2.2e-16 ***
nh_length                2.9981e-05  1.3387e-06  22.3953 < 2.2e-16 ***
fixed_capital           -1.5359e-10  4.0146e-10  -0.3826 0.7020383
inflation               -8.5277e-03  3.8229e-03  -2.2307 0.0257095 *
construction_worker_wage -3.8276e-04  8.6390e-05  -4.4306 9.436e-06 ***
num_factories            4.8344e-06  8.0208e-07   6.0273 1.690e-09 ***
d_mark                  -3.2158e-03  3.2093e-04 -10.0204 < 2.2e-16 ***
nsdp2                    4.5774e-02  5.3149e-03   8.6124 < 2.2e-16 ***
sqft_2                  -1.2066e-04  2.4115e-05  -5.0033 5.674e-07 ***
sqft_3                  -1.1658e-04  4.4208e-05  -2.6370 0.0083694 **
data3$under_wage        -5.9767e-05  1.2201e-04  -0.4898 0.6242527
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- As expected, we see larger standard error values after accounting for the heteroscedasticity by using robust standard errors

High Breakdown And High Efficiency Robust Linear Regression: estimating standard errors using lmrob() function

- By default, the lmRob function automatically chooses an appropriate algorithm to compute a final robust estimate with high breakdown point and high efficiency.
- Breakdown point of an estimator can be described as the proportion of incorrect observations (e.g. arbitrarily large observations) an estimator can handle before giving an incorrect (e.g., arbitrarily large) result
- Thus, this method is expected to calculate more accurate estimators, and in extension, standard errors
- The final robust estimate is computed based on an initial estimate with high breakdown point

```
Residuals:
      Min        1Q    Median        3Q       Max
-8.247995 -0.240639 -0.005169  0.236983  3.900408

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.590e+01  6.443e-01  24.671  < 2e-16 ***
dealer                   -1.826e-01  2.413e-02  -7.568 3.93e-14 ***
under_construction        1.999e-02  3.353e-02   0.596    0.551
rera                      2.464e-02  6.017e-03   4.096 4.22e-05 ***
bhk_no                    2.105e-01  7.526e-03  27.965  < 2e-16 ***
square_ft                 7.159e-04  1.675e-05  42.746  < 2e-16 ***
ready_to_move                    NA         NA      NA       NA
resale                    1.244e-02  1.510e-02   0.824    0.410
tier2                     9.638e-03  2.042e-02   0.472    0.637
tier3                    -1.680e-01  2.704e-02  -6.212 5.31e-10 ***
avprice                   3.526e-03  5.875e-05  60.025  < 2e-16 ***
coord                    -1.623e-06  7.172e-08 -22.638  < 2e-16 ***
cci                      -1.049e-05  4.541e-03  -0.002    0.998
pm2                      -2.224e-03  1.430e-04 -15.556  < 2e-16 ***
interestPop               7.744e-01  9.097e-02   8.512  < 2e-16 ***
loanPop                  -2.214e-01  9.710e-03 -22.800  < 2e-16 ***
power_availability       -9.549e-04  3.299e-05 -28.949  < 2e-16 ***
nh_length                 4.672e-05  1.221e-06  38.261  < 2e-16 ***
fixed_capital            -5.467e-10  3.380e-10  -1.617    0.106
inflation                -2.921e-02  3.010e-03  -9.707  < 2e-16 ***
construction_worker_wage -1.202e-03  8.728e-05 -13.774  < 2e-16 ***
num_factories             1.618e-05  7.730e-07  20.931  < 2e-16 ***
d_mark                   -2.396e-03  2.592e-04  -9.244  < 2e-16 ***
nsdp2                     2.404e-01  8.815e-03  27.272  < 2e-16 ***
sqft_2                   -1.622e-04  1.513e-05 -10.719  < 2e-16 ***
sqft_3                   -2.936e-05  2.200e-05  -1.335    0.182
data3$under_wage          4.839e-05  1.009e-04   0.480    0.632
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 0.3509
  (878 observations deleted due to missingness)
Multiple R-squared:  0.7641,    Adjusted R-squared:  0.7639
Convergence in 20 IRWLS iterations
```

(B) **Rectification of Normality of Residuals:**

From the previous question, we have observed that the Shapiro Wilk test on our first 5000 (out of over 25000) sample data-points has shown that the residuals are not normally distributed. Thus, we try to increase the p-value generated through some Transformations on the dependent variable, in order to correct for this. The results of the transformation are as follows:

- **Transformations:**

1. Logarithmic transformation

```
        Shapiro-Wilk normality test

data:  reg_Norm_Log$residual[0:5000]
W = 0.92505, p-value < 2.2e-16
```

2. Square-root transformation

```
        Shapiro-Wilk normality test

data:  reg_Norm_Sqrt$residuals[0:5000]
W = 0.88493, p-value < 2.2e-16
```

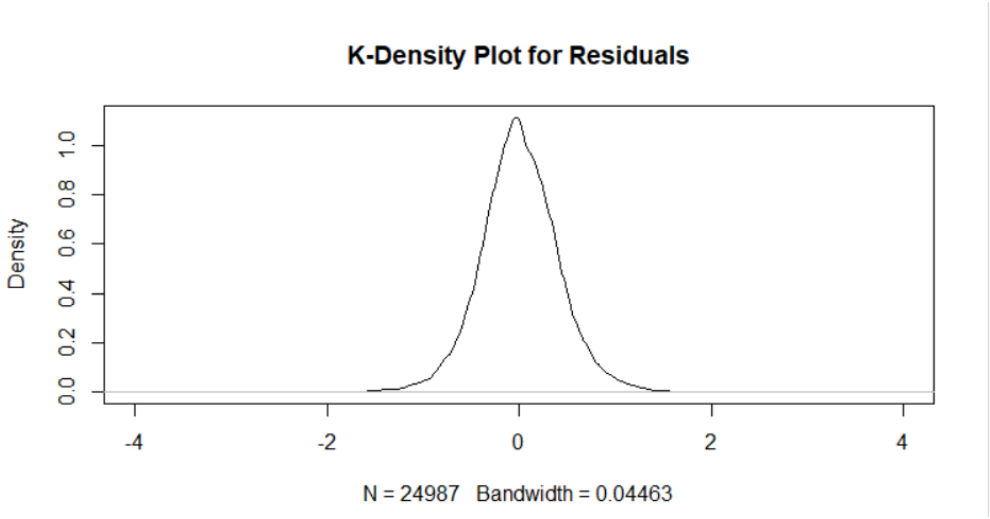
3. Inverse transformation

```
        Shapiro-Wilk normality test

data:  reg_Norm_inv$residuals[0:5000]
W = 0.17128, p-value < 2.2e-16
```

Thus, we can see that despite transforming the variable, the p-value is still much lower than 0.05; thereby indicating that the residuals are still not following normal distribution.

However, we must note that this test is only considering 1/5th of our actual sample size, and the Central Limit Theorem states that for sufficiently large sample sizes (>30), the residuals will approach the Normality Distribution. This can be seen graphically by the following plot as well:



Thus, there is no need for any correction to account for the normality of the residual terms

(C) **Rectification of Multicollinearity**

As discussed in Q6, there is no need to drop any variables to account for multicollinearity, since the mean vif value for the entire model (7.90635) is below the standardized critical value (vif=10). However, we see a disproportionate distribution of vif values for the independent variables; hence if multicollinearity consideration is the priority, then we can drop the variables *loan_pop*, *power_availabilty*, *under_wage*, *under_construction* and *ready_to_move.* The resultant vif test can be seen below:

```
Call:
imcdiag(mod = reg_mult, vif = 10)


All Individual Multicollinearity Diagnostics Result

                             VIF    TOL         Wi         Fi Leamer     CVIF Klein  IND1   IND2
dealer                    1.4649 0.6826   580.3586   610.9282 0.8262  -3.6254     0 5e-04 0.4764
rera                      1.1695 0.8550   211.6414   222.7893 0.9247  -2.8944     0 7e-04 0.2176
bhk_no                    2.2274 0.4490  1532.1042  1612.8059 0.6700  -5.5122     0 4e-04 0.8272
square_ft                 3.7837 0.2643  3474.9204  3657.9574 0.5141  -9.3639     1 2e-04 1.1044
resale                    1.5766 0.6343   719.8095   757.7246 0.7964  -3.9018     0 5e-04 0.5490
tier2                     7.2400 0.1381  7789.4022  8199.6991 0.3716 -17.9175     1 1e-04 1.2938
tier3                     6.8056 0.1469  7247.1702  7628.9058 0.3833 -16.8425     1 1e-04 1.2805
avprice                   2.0462 0.4887  1305.9898  1374.7811 0.6991  -5.0640     0 4e-04 0.7675
coord                     3.9048 0.2561  3626.0716  3817.0704 0.5061  -9.6636     1 2e-04 1.1167
cci                      12.5908 0.0794 14468.8097 15230.9360 0.2818 -31.1597     1 1e-04 1.3819
pm2                       2.8460 0.3514  2304.3855  2425.7661 0.5928  -7.0433     0 3e-04 0.9737
interestPop              11.8462 0.0844 13539.2661 14252.4299 0.2905 -29.3168     1 1e-04 1.3744
nh_length                 3.5750 0.2797  3214.3289  3383.6397 0.5289  -8.8473     1 2e-04 1.0812
fixed_capital             4.4060 0.2270  4251.7292  4475.6837 0.4764 -10.9040     1 2e-04 1.1604
inflation                 9.7650 0.1024 10941.3903 11517.7143 0.3200 -24.1664     1 1e-04 1.3474
construction_worker_wage  2.6291 0.3804  2033.5687  2140.6844 0.6167  -6.5064     0 3e-04 0.9301
num_factories             4.7098 0.2123  4630.9898  4874.9213 0.4608 -11.6559     1 2e-04 1.1824
d_mark                    1.1772 0.8495   221.1937   232.8448 0.9217  -2.9133     0 7e-04 0.2260
nsdp2                     5.0587 0.1977  5066.4512  5333.3201 0.4446 -12.5192     1 2e-04 1.2044
sqft_2                    7.2202 0.1385  7764.6521  8173.6454 0.3722 -17.8685     1 1e-04 1.2932
sqft_3                    5.1915 0.1926  5232.2440  5507.8459 0.4389 -12.8479     1 2e-04 1.2120

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

tier2 , nsdp2 , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.6851

* use method argument to check which regressors may be the reason of collinearity
===================================
```

(D) **Rectification of Omitted Variable Bias:**

We know that Omitted variable bias occurs when relevant and correlated variable terms occur in the residual term. Since our model has 25 independent variables, we are unable to come up with any more mutually uncorrelated, measurable variables that could reduce this bias.

However, we know if not accounted for, the omitted variable bias threatens the accuracy of our estimates, by allowing for overestimation/underestimation of regression coefficients. Thus, to reduce the effect of the bias, we introduce three transformed variables derived from the pre-existing independent variables: bhk_sqr (squared value of the number of bedrooms), dmark_sqr (squared value of the distance from the market), and sqft_sqr (squared value of the floor area of the house).

Rationale: adding the square of these variables accounts for the fact that the rate of increase of the house prices with respect to these variables is expected to be proportional to the variable itself.

After including these 3 variables, we run the Ramsey's Reset test, to see that the P value of the hypothesis (0.8096) has shot above the critical P-value (0.05 - at a 95% CI). Thus, we fail to reject the null hypothesis, and we can conclude that we have reduced the omitted variable bias in the model.

```
> library(lmtest)
> resettest(reg_OVB, power=2,  type='fitted')

        RESET test

data:  reg_OVB
RESET = 0.058031, df1 = 1, df2 = 24956, p-value = 0.8096
```

QUESTION 8

Test at least two joint hypothesis with linear combinations of regression coefficients from your model and comment on the results

ANSWER 8

We test the following hypotheses:

a. H_o: $\beta_7 = \beta_8 = 0$ H_a: at-least one of them is non-zero

b. H_0: $\beta_{71} - \beta_{81} = 0$ H_a: the two coefficients are different from each other

For a, the results are as follows:

```
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1  24963 4680.4
2  24961 4679.0  2    1.4445 3.8531 0.02123 *
```

Hence, we reject the null hypothesis that both the coefficients are zero. Thus, the average price of the houses is dependent on the tier of the city in which it is located.

For b, the results are as follows:

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1  24962 4679.2
2  24961 4679.0  1   0.26316 1.4039 0.2361
```

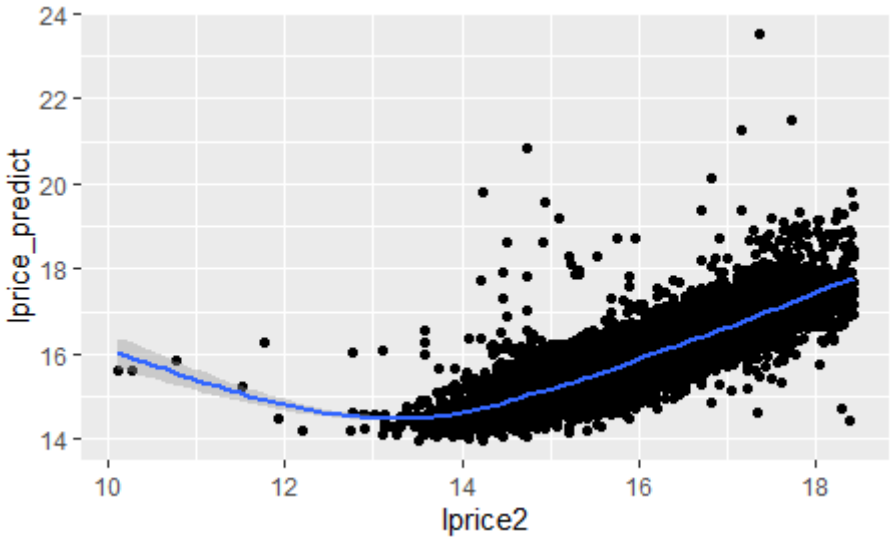
Here, we cannot reject the null hypothesis that both coefficients are equal. This indicates that for Tier 2 and Tier 3 cities, the percentage change in the price of the house for a unit change in the area of the house is the same. This is not what we expected – in general, the cost per square feet of house in Tier 2 cities should be higher than Tier 3. We attribute the result to a statistical illusion.

Here, we have classified cities as Tier 2 and Tier 3 on the basis of the classification given in Census 2011. However, many cities have grown tremendously between 2011 and 2020 (year in which the house prices have been recorded). This growth in urban landscape is seen for all inter-Census periods. For example, according to the Census of 2001, Gurgaon was not even a Census Town (it was classified as a rural area). By 2011, the city's population had grown by 1200%, and is now classified as a Tier 2 city. We imagine that similar classification issue permeate our dataset too.

QUESTION 9

Lastly, make prediction based on your model and comment on the result.

ANSWER 9



We draw a graph between the predicted log prices and the actual log prices from the regression. We find the relationship shown above. When we carry out a linear regression, the slope of the line is 0.7, and a linear hypothesis test rejects the hypothesis that the slope is equal to one.

We attribute this to the outliers in the regression. While cleaning data, we had removed most of the outliers (property prices>50 crore),, still some exaggerated property values with large areas remain. In the presence of these outliers, our predicted log price values show some bias.

REFERENCES

Anagol, S., Cole, S., & Sarkar, S. (2017). Understanding the advice of commissions-motivated agents: Evidence from the Indian life insurance market. *Review of Economics and Statistics*, *99*(1), 1-15.

Chay, K. Y., & Greenstone, M. (2005). Does air quality matter? Evidence from the housing market. *Journal of political Economy*, *113*(2), 376-424.

Coletta, M., De Bonis, R., & Piermattei, S. (2019). Household debt in OECD countries: the role of supply-side and demand-side factors. *Social Indicators Research*, *143*(3), 1185-1217.

Duncan, M. (2011). The impact of transit-oriented development on housing prices in San Diego, CA. *Urban studies*, *48*(1), 101-127.

Gandhi, S., Tandelb, V., Tabarrokc, A., & Ravid, S. (2019). Too Slow for the Urban March: Litigations and Real Estate Market in Mumbai, India.

Gulzar, S., & Pasquale, B. J. (2017). Politicians, bureaucrats, and development: Evidence from India. *American Political Science Review*, *111*(1), 162-183.

Pawar, S., & Ahire, H. (2018). Study of Impact in Construction Project Due to Introduction of Rera. *Asian Journal For Convergence In Technology (AJCT)*, *4*(3).

Sadhak, H. (2009). *Life Insurance in India: Opportunities, Challenges and Strategic Perspective*. SAGE Publications India.

Sunde, T., & Muzindutsi, P. F. (2017). Determinants of house prices and new construction activity: an empirical investigation of the Namibian housing market. *The Journal of Developing Areas*, *51*(3), 389-407.

Tandel, V., Hiranandani, K., & Kapoor, M. (2019). What's in a definition? A study on the suitability of the current urban definition in India through its employment guarantee programme. *Journal of Asian Economics*, *60*, 69-84.

Tyrväinen, L., & Miettinen, A. (2000). Property prices and urban forest amenities. *Journal of environmental economics and management*, *39*(2), 205-223.

Wu, J., Deng, Y., & Liu, H. (2014). House price index construction in the nascent housing market: the case of china. *The Journal of Real Estate Finance and Economics*, *48*(3), 522-545.

Xu, T., Zhang, M., & Aditjandra, P. T. (2016). The impact of urban rail transit on commercial property value: New evidence from Wuhan, China. *Transportation Research Part A: Policy and Practice*, *91*, 223-235.

Zhang, X., Liu, X., Hang, J., Yao, D., & Shi, G. (2016). Do urban rail transit facilities affect housing prices? Evidence from China. *Sustainability*, *8*(4), 380.