UNIVERSITÉ PARIS 1

PANTHÉON SORBONNE

A case study of housing price in King County

Econometrics Project

Professor: Hélène Huber-Yahi

Group members: MAI Thu Quynh

Chanda LEANG

ZHU Peitong

Contents

I. Introduction

i. Choice of the topic

King County is the most popular county in Washington State. The county gradually converts into countryside and farming as we travel westward. Ten Fortune 500 companies, including Starbucks and Amazon, have their headquarters in King County. Microsoft and Amazon are two of the Big Five tech behemoths that are highly sought after by those looking for work in the tech industry. It is natural for those considering working for a Fortune 500 company in Seattle to be curious about the cost of housing in King County.

Therefore, it's important for people who want to buy houses to understand what their asset is worth. Homeowners and homebuyers can use house price predictions to assist them to decide whether to sell or buy a home at a specific price. However, determining the price of a house can be difficult because there are so many variables to consider.

In this paper, we are interested in the number of bathrooms, of the apartment's interior living space and the age of the house to the time it was sold, etc and we will investigate the best model for predicting house sale prices in King County, Washington, USA, using multiple regression and predictive methods.

ii. Methodology

This study systematically reviews the data for the dataset from Kaggle[1][1], containing listing attributes and prices for all houses sold in King County, Washington from May 2014 to May 2015, aiming to predict the price of properties in King County through creating a linear regression model.

The data was analyzed using an econometrics model and data management was performed by STATA. To investigate this statistically, we describe statistics and use histograms and plots first to check the data distribution. Second, building a sample linear regression model to analyze the determinants of house pricing in King County. Then looking for outliers and influential observations to improve the quality of the data analyzed, leading to more accurate conclusions. After that, we will diagnose heteroskedasticity and try to adjust it with GLS. Finally, we determine whether all chosen factors are significant and draw conclusions.

[1] https://www.kaggle.com/datasets/vallabhadattap/kingcountyhousing The data contains 21613 rows and 21 columns i.e. it contains 21613 observations with 21 variables. The data contains variables like unique home ID for sale, date of sale and 18 other home features.

II. Data exploration and setting-up of the model

i. Data presentation

The table below describes the interpretation of the variables in the dataset.

Variable	Description
price	The selling price of each house
bathrooms	Number of bathrooms[2]
sqft_living	Square footage of the apartment's interior living space
floors	Number of floors
waterfront	Whether the apartment was overlooking the waterfront or not[3]
view1	How good the view of the property was[4]
condition	The condition of the apartment[5]
grade	The quality level of construction and design[6]
age	How many years this house been built by the time of sale has
renovated	Has the house ever been renovated?[7]
dist_from_cent	How far the distance from the house to the center

First of all, in order to find the influencing factors that affect the price of a house, we considered different variables X_i. According to the whole market situation and customers' demands, we choose 21 variables, such as bathrooms, floors, etc, to analyze the relationships between variables we choose with house price.

Our data consists of 21 variables and total 21,613 observations >>100, which is a large set of data. We drop some variables that are not suitable for our analysis. Then, we generate a new variable called dist_from_center based on two other variables (latitude and longitude) (we will discuss in detail in section II.iii). For the "renovated" variable we consider it as a dummy variable, 1 for houses that have been renovated and 0 for houses that have never been renovated. We create a new variable called "age" which shows how long these houses were built up to the time they were sold. In the end, we obtained 11 variables as shown in the table below.

[2] Where .5 accounts for a room with a toilet but no shower.

[3] Dummy variable (1: The house has a view of rivers, lakes …; 0: No view of rivers, lakes, …)

[4] An index from 0 to 4 of how good the view is, 0 shows the house does not have a good view, and 4 represents a house that has a good view outside.

[5] An index from 1 to 5. The better the condition of the house, the higher the number represented.

[6] An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 has a high-quality level of construction and design.

[7] Dummy variable (1: The house is renovated …; 0: No renovation)

```
. sum price bathrooms sqft_living floors waterfront view1 condition grade age renovated dist_from_cent
```

Variable	Obs	Mean	Std. dev.	Min	Max
price	21,613	540088.1	367127.2	75000	7700000
bathrooms	21,613	2.114757	.7701632	0	8
sqft_living	21,613	2079.9	918.4409	290	13540
floors	21,613	1.494309	.5399889	1	3.5
waterfront	21,613	.0075418	.0865172	0	1
view1	21,613	.2343034	.7663176	0	4
condition	21,613	3.40943	.650743	1	5
grade	21,613	7.656873	1.175459	1	13
age	21,613	43.31782	29.37549	-1	115
renovated	21,613	.0422894	.2012532	0	1
dist_from_~t	21,613	.1788204	.0863229	.0030092	.9258285

ii. Descriptive statistics, Histograms and plots

```
. sum price, detail
```

price

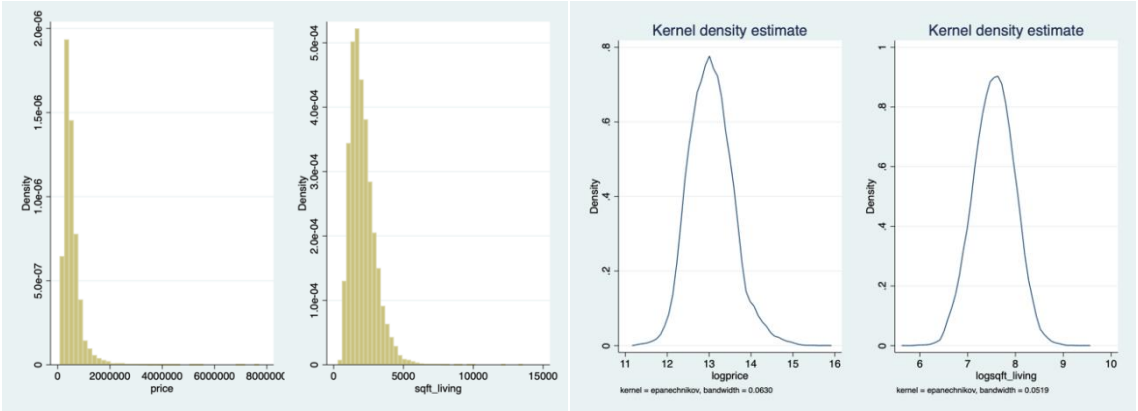
	Percentiles	Smallest		
1%	153500	75000		
5%	210000	78000		
10%	245000	80000	Obs	21,613
25%	321950	81000	Sum of wgt.	21,613
50%	450000		Mean	540088.1
		Largest	Std. dev.	367127.2
75%	645000	5570000		
90%	887000	6885000	Variance	1.35e+11
95%	1157200	7062500	Skewness	4.02379
99%	1965000	7700000	Kurtosis	37.57726

```
. sum sqft_living, detail
```

sqft_living

	Percentiles	Smallest		
1%	720	290		
5%	940	370		
10%	1090	380	Obs	21,613
25%	1427	384	Sum of wgt.	21,613
50%	1910		Mean	2079.9
		Largest	Std. dev.	918.4409
75%	2550	9890		
90%	3250	10040	Variance	843533.7
95%	3760	12050	Skewness	1.471453
99%	4980	13540	Kurtosis	8.241603

From the summary table of prices, we see that mean is about 540 thousand which is larger than the median which equals 450 thousand. We could say that the distribution is not a normal distribution, but instead, it is skewed to the right (Skewness>0). A similar interpretation applies to sqft_living. Look at the histogram below.

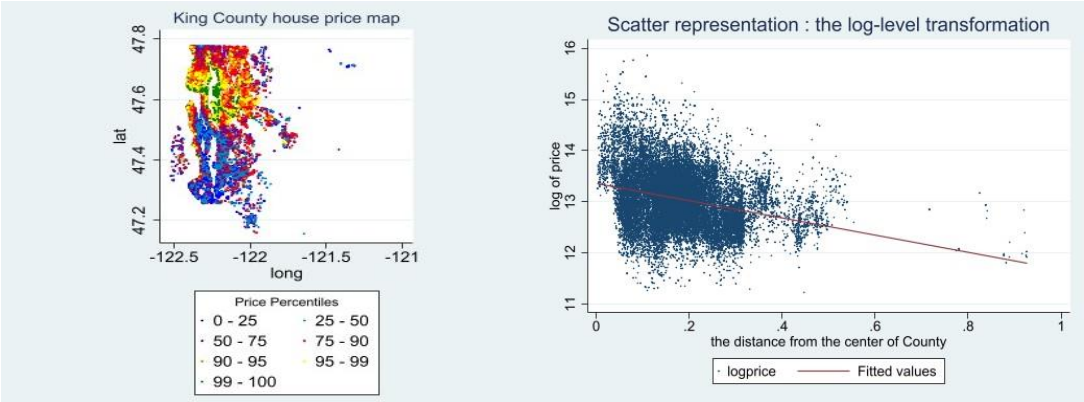


We use log on two variables “price” and “living square” because our data is heavily skewed and the regression line is heavily sensitive to extreme points. After putting the log, our distribution becomes more symmetric (Bell-shaped).

iii. Distance from the center

We obtain dist_from_cent variable by finding the median of latitude and longitude, then calculate the house distance from the center by using the following formula:

$$dist\_from\_cent = \sqrt{\big(lat(i) - var(lat)\big)^2 + \big(lot(i) - var(lot)\big)^2}$$



The graph on the left allows us to observe that the prices of houses depend on their variation from the center. The top 5% highest prices are close to the county center (green dots).

The scatterplot on the right shows us the negative relationship between the log of price and the distance from the center. Meaning that the further away from the center, the lower the price of the houses.

III. Description of the model

i. Simple linear regression model

In this project, we want to analyze the determinants of house pricing in King County by calculating the value. We start by running a simple linear regression model:

$$log(price_i) = \beta_0 + \beta_1 * logsqft\_living_i + \varepsilon_i$$

```
. reg logprice logsqft_living
```

Source	SS	df	MS	Number of obs	=	21,613
				F(1, 21611)	=	18079.14
Model	2730.80804	1	2730.80804	Prob > F	=	0.0000
Residual	3264.28677	21,611	.151047465	R-squared	=	0.4555
				Adj R-squared	=	0.4555
Total	5995.09481	21,612	.277396577	Root MSE	=	.38865

logprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
logsqft_living	.836771	.0062233	134.46	0.000	.8245729	.848969
_cons	6.729916	.047062	143.00	0.000	6.637671	6.822161

From the regression result, our model yields $R^2 \sim 0.46$, this means that the model explains only 46% of the variance of the outcome (logprice) can be predicted from the independent variable (logsqft_living) and is globally significant (Prob > F = 0.0000). Since we use log in both dependent and independent variables, we interpret the coefficient as elasticity. So, if square_living increases by 1%, the price of the house will increase by 0.83%. This result is statistically significant at 5% level (p-value < 5%).

ii. Multiple linear regression model

We introduce the idea of multiple linear regression because a simple linear regression leads to omitted variable biased.

$$log(price_i) = \beta_0 + \beta_1 * bathrooms_i + \beta_2 * logsqft\_living_i + \beta_3 * floors_i + \beta_4 * waterfront_i + \beta_5 * view1_i + \beta_6 * condition_i + \beta_7 * grade_i + \beta_8 * age_i + \beta_9 * renovated_i * + \beta_{10} * dist\_from\_cent_i + \varepsilon_i$$

If there is multicollinearity then the OLS quantity cannot be estimated and R-squared can also be overestimated. Therefore, we have to check Multicollinearity in case independence variables are correlated.

```
. correlate logprice bathrooms logsqft_living floors waterfront view1 condition grade age renovated dist_from_cent
(obs=21,613)
```

	logprice	bathro~s	logsqf~g	floors	waterf~t	view1	condit~n	grade	age	renova~d	dist_f~t
logprice	1.0000										
bathrooms	0.5508	1.0000									
logsqft_li~g	0.6749	0.7613	1.0000								
floors	0.3106	0.5007	0.3676	1.0000							
waterfront	0.1746	0.0637	0.0793	0.0237	1.0000						
view1	0.3465	0.1877	0.2467	0.0294	0.4019	1.0000					
condition	0.0396	-0.1250	-0.0481	-0.2638	0.0167	0.0460	1.0000				
grade	0.7036	0.6650	0.7437	0.4582	0.0828	0.2513	-0.1447	1.0000			
age	-0.0806	-0.5064	-0.3497	-0.4896	0.0261	0.0535	0.3607	-0.4474	1.0000		
renovated	0.1141	0.0503	0.0508	0.0063	0.0933	0.1041	-0.0601	0.0140	0.2248	1.0000	
dist_from_~t	-0.2767	0.0141	0.0036	0.0473	-0.0186	-0.0469	-0.0929	-0.0452	-0.2482	-0.0509	1.0000

```
( 1)  age = 0
( 2)  renovated = 0
( 3)  condition = 0

      F(  3, 21602) =  912.53
           Prob > F =    0.0000
```

We concern about variable age, renovated, and condition which have a low correlation with the dependent variable. So, we do the Fisher test to check if we should keep or drop these variables. We cannot remove them because the Prob > F = 0.0000 is less

than 0.05 (means our added coefficients improve the model)

```
. reg logprice bathrooms logsqft_living floors waterfront view1 condition grade age renovated dist_from_cent
```

Source	SS	df	MS	Number of obs	=	21,613
				F(10, 21602)	=	4544.98
Model	4063.6626	10	406.36626	Prob > F	=	0.0000
Residual	1931.43221	21,602	.089409879	R-squared	=	0.6778
				Adj R-squared	=	0.6777
Total	5995.09481	21,612	.277396577	Root MSE	=	.29901

logprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bathrooms	.0595448	.0046446	12.82	0.000	.0504411	.0686484
logsqft_living	.3551582	.0087004	40.82	0.000	.3381047	.3722117
floors	.0775418	.0047148	16.45	0.000	.0683005	.0867831
waterfront	.3496484	.0257348	13.59	0.000	.2992064	.4000905
view1	.0614396	.0030568	20.10	0.000	.0554482	.0674311
condition	.0357691	.003448	10.37	0.000	.0290108	.0425273
grade	.2130326	.0028473	74.82	0.000	.2074517	.2186136
age	.0043594	.0001003	43.48	0.000	.0041629	.0045559
renovated	.0304798	.0107949	2.82	0.005	.0093211	.0516385
dist_from_cent	-1.164757	.0248638	-46.85	0.000	-1.213492	-1.116022
_cons	8.372464	.0531058	157.66	0.000	8.268373	8.476556

The above graph shows that all the independent variables are significant within the confidence interval. To test for multicollinearity, the VIF of each variable is first calculated after running linear regression.

```
. vif
```

Variable	VIF	1/VIF
logsqft_li~g	3.30	0.302848
bathrooms	3.09	0.323323
grade	2.71	0.369318
age	2.10	0.476870
floors	1.57	0.638266
view1	1.33	0.753967
condition	1.22	0.821759
waterfront	1.20	0.834535
renovated	1.14	0.876538
dist_from_~t	1.11	0.898054
Mean VIF	1.88	

According to this graph, we know that the largest VIF=3.3 <<5, hence, no need to worry about the existence of multicollinearity. Where:

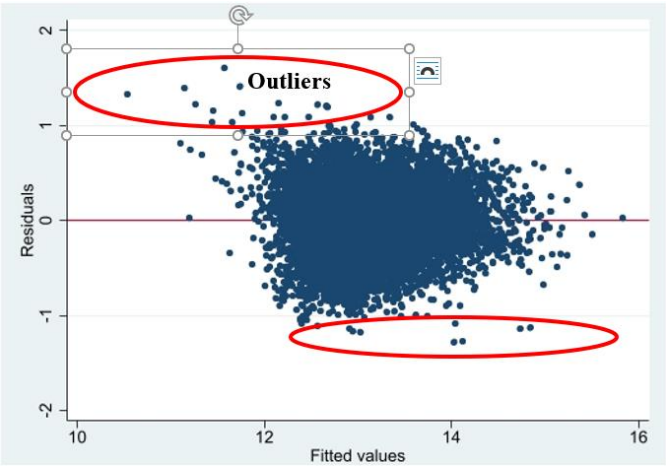
$$vif = \frac{1}{1 - R^2}$$

Afterwards, the variables were estimated by multiple linear regression

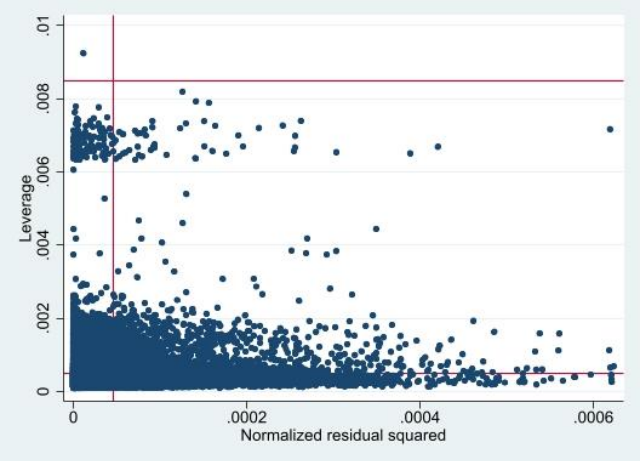
The regression results at the top of the table show that the model explains 67.78% of the variation in logprice and the p-value(P>|t|) of all regression coefficients is less than 0.05. Therefore, all independent variables are significant at the 5% level. The table

shows that the coefficient estimates for logsqft_living is 0.3552%, i.e. the price growth rate of the square footage of the apartments' interior living space is 0.3552%. Meaning that it's a significant factor for the house price in the King County. The house price growth rates for an additional unit of waterfront and the quality level of houses, are 34.96% and 21.3%, respectively. Significantly, the coefficient estimates for dist_from_cent is negative (~ -1.165). This indicates that the further the house is from the city center, the lower the price, and vice versa, the higher the price is in the city center, which is reasonable. Moreover, if the house is close to the waterfront, the price will increase by about 34.96% and 3.05% if it has been renovated.

IV. Outlier Detection



Looking the graph, we realize that the scatter of this plot is not uniform. There are quite a lot of outliers that lies an abnormal distance from other values. But in our case, we just drop the outliers with a large residual compared to other observations and are not influential data point. We find 0.09% of observations (19 spots) located in the elimination area.



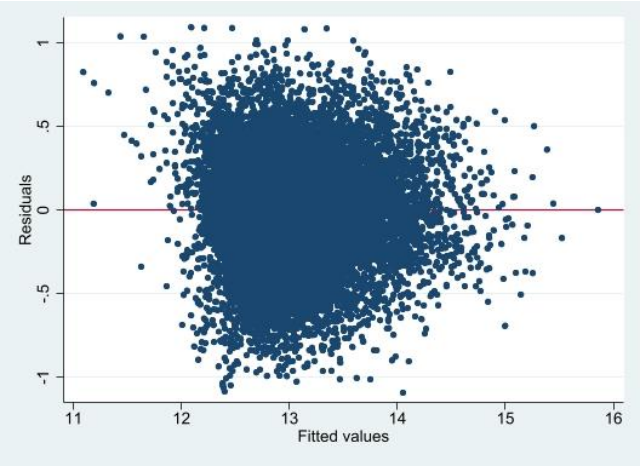
Normally, we can remove the points which are far away from the average residuals based on the leverage axis and normalized residual squared axis, but in our case, we see the number of these points is negligible compared to the sample size, so we will keep them.

V. An analysis of heteroscedasticity

One of the underlying assumptions of linear regression is that the model is homoscedasticity. In this part, we will check if our model meets this assumption, in case that it is heteroskedasticity we need to correct our model.

To be more specific, at the first step, the scatter plot is used to roughly examine whether there is heteroskedasticity and non-linearity in a linear regression model. This is also the way to detect the outliers as we did before. Secondly, a statistical test is needed to rigorously determine the heteroskedasticity. The third step is to adjust the model-based standard errors. Finally, the Feasible generalized least squares (FGLS) is used to deal with the heteroskedasticity.

i. Graph of residuals against fitted values



Since the residuals can be considered as the realized values of the perturbation terms, the existence of heteroskedasticity can be roughly examined by the fluctuations of the residuals, which can be seen in the scatter plot of the residuals and the fitted values, which is the most intuitive method. From the graph, there is a hint of heteroscedasticity, since the variance of residuals is getting smaller as the fitted value increases (i.e. an unequal scatter of the error term).

To check with more certainty whether our model is heteroscedasticity or not, we consider Breusch-Pagan test and White-test in following steps.

ii. Breusch-Pagan Test

"The null hypothesis for this test is that the error variances are all equal. (homoscedasticity).
The alternative hypothesis is that the error variances are not equal. More specifically, as Y increases, the variances increase (or decrease). (heteroscedasticity).

$$H0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma^2$$
$$H1: \sigma_1^2 \neq \sigma_2^2 \neq \ldots \neq \sigma^2$$

A small chi-square value (along with an associated small p-value) indicates the null hypothesis is true (i.e., that the variances are all equal)."

```
. estat hettest, rhs iid

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: All independent variables

H0: Constant variance

   chi2(10) = 308.18
Prob > chi2 = 0.0000
```

The figure above shows that the probability value of the chi-square statistic is 0.000 (less than our chosen significance value of 0.05) Therefore the null hypothesis of constant variance can be rejected at 5% level of significance. It implies the presence of heteroscedasticity in the dependent variable logprice.

iii. White test

The BP test assumes that the conditional variance function is linear and is only a first-order approximation to the conditional variance function, possibly ignoring the higher-order terms. For this reason, the White test adds all quadratic terms to the auxiliary regression of the BP test.

```
. estat imtest, white

White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

   chi2(63) = 822.95
Prob > chi2 = 0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	822.95	63	0.0000
Skewness	545.75	10	0.0000
Kurtosis	145.43	1	0.0000
Total	1514.13	74	0.0000

Similar to the results of the Breusch-Pagan test, in White test, we have the prob > chi2 = 0.000 << 0.05. Therefore, the null hypothesis of constant variance is rejected at 5% level of significance. The implication from all findings is that there is heteroscedasticity in the residuals.

iv. Heteroscedasticity robust standard errors

```
. reg logprice bathrooms logsqft_living floors waterfront view1 condition grade age renovated dist_from_cent, robust

Linear regression                               Number of obs     =     21,594
                                                F(10, 21583)      =    4260.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6821
                                                Root MSE          =     .29685
```

logprice	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
bathrooms	.0616758	.0047952	12.86	0.000	.0522769	.0710747
logsqft_living	.3529573	.0089221	39.56	0.000	.3354693	.3704453
floors	.0778404	.0045193	17.22	0.000	.0689823	.0866984
waterfront	.3559071	.0274984	12.94	0.000	.3020081	.409806
view1	.0610874	.0029852	20.46	0.000	.0552362	.0669386
condition	.03573	.0036496	9.79	0.000	.0285765	.0428836
grade	.2153292	.0028462	75.66	0.000	.2097504	.2209079
age	.0044351	.0001008	44.00	0.000	.0042375	.0046326
renovated	.0276481	.0112037	2.47	0.014	.005688	.0496081
dist_from_cent	-1.171146	.0278379	-42.07	0.000	-1.22571	-1.116582
_cons	8.364478	.0545982	153.20	0.000	8.257461	8.471494

The above table shows the regression result after correction in the heteroscedasticity tests. As a consequence, the problem of heteroscedasticity is no longer an issue. This result presents the same coefficients but robust standard errors different from the standard errors in previous regression model. For example, for the "waterfront" variable, its robust standard error is ~ 0.0275 instead of 0.0256 as before; or for the "bathrooms" variable, robust standard error is ~ 0.0048 while before a correction, its value is ~ 0.0046.

v. FGLS

We use FGLS to find the accurate weight and reweight the data. Here, we estimate the model by multiplying all variables and the constant by 1/sqrt(hhat).

Source	SS	df	MS	Number of obs	=	21,594
				F(11, 21583)	>	99999.00
Model	147987556	11	13453414.2	Prob > F	=	0.0000
Residual	72049.3046	21,583	3.33824328	R-squared	=	0.9995
				Adj R-squared	=	0.9995
Total	148059605	21,594	6856.51593	Root MSE	=	1.8271

logpricestar	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bathroomsstar	.05925	.0043644	13.58	0.000	.0506954	.0678046
logsqft_livingstar	.3718656	.00848	43.85	0.000	.3552443	.388487
floorsstar	.078681	.0043977	17.89	0.000	.0700612	.0873007
waterfrontstar	.3629654	.0266074	13.64	0.000	.3108129	.4151179
viewstar	.0588559	.0027019	21.78	0.000	.0535601	.0641518
conditionstar	.0356331	.0034389	10.36	0.000	.0288925	.0423736
gradestar	.2083371	.0026696	78.04	0.000	.2031044	.2135698
agestar	.0042168	.0000984	42.87	0.000	.004024	.0044096
renovatedstar	.030483	.0109624	2.78	0.005	.008996	.0519701
dist_from_centstar	-1.254943	.0240342	-52.21	0.000	-1.302052	-1.207834
constantstar	8.304168	.0517866	160.35	0.000	8.202663	8.405674

Compared to the unweighted estimate, The R-squared of our FGLS model is 99.95%, and we can say that our model fits us very well our data. It shows that this model reliable after an autocorrection using FGLS estimation and all the coefficients remain statistically significant in our weighted model.

vi. Endogeneity

In this part, we discuss the possibility that our model might have an endogeneity issue where independent variable and error terms are correlated. For example, one of our independent variable distances from the center could correlate to whether there is a subway or not. If this is the case, the coefficient of this independent variable is not the true impact on the house price, but in fact, it is the mixture of the true coefficient plus some biases. This could lead to a misleading result. However, we do not have information about the subway in our data so we cannot further discuss this issue.

VI. Conclusion

Finding an appropriate regression model plays an important role in predicting house prices. In this report, based on available data on house prices and factors affecting house prices from the data collector's point of view, we have selected the correlated variables and conducted several tests to find the optimal regression model for predicting house prices in King County. The results obtained on a large number of 21,594 observations and 11 variables (including 10 independent variables) are as follows.

The regression model fits for all the variables considered. It is better than the model with lack of some of variables since all the p-value associated with each variable are less than significant level 0.05.

Except for the distance to the center, which shows a negative coefficient, the other variables all have a positive effect on house prices. In which the price of house increases by 0.37% for every 1% increase the square footage of the apartment’s interior living

space. While the variables "dist_from_cent", “grade” and "waterfront" and have a higher correlation level than other variables. It means in addition to core factors such as geographical distance from the center or the interior square of the house, the quality of the construction and design and water view are factors that have a great impact on house prices in King County. These results make perfect sense in practice.

The model regression with the autocorrection by using FGLS gives better results than the original model. It has limited heteroscedasticity and improves R-squared indicator very well. It reveals that 99.95% of data fit the regression model.

VII. References

1. Data from Kaggle.com
2. Code from lectures & Stata.com